

Maschinelle Früherkennung abbruchgefährdeter Studierender

-

Konzeption, Systemvergleich und Evaluation

Dissertation zur Erlangung des Doktorgrades (Dr. rer. oec.)

eingereicht an der Fakultät für Wirtschaftswissenschaft der Bergischen Universität
Wuppertal

von

Dipl.-Kfm. (FH) Johannes Berens, M.Sc.

Erste Gutachterin: Prof. Dr. Kerstin Schneider

Zweite Gutachterin: Prof. Dr. Christine Volkmann

Die Dissertation kann wie folgt zitiert werden:

urn:nbn:de:hbz:468-20210318-114945-9

[<http://nbn-resolving.de/urn/resolver.pl?urn=urn%3Anbn%3Ade%3Ahbz%3A468-20210318-114945-9>]

DOI: 10.25926/7jv7-2s67

[<https://doi.org/10.25926/7jv7-2s67>]

Inhalt

I Einleitung	1
II Konzeption eines Frühwarnsystems	4
1. Studienabbruch und Frühwarnsysteme	4
1.1. Educational Data Mining	6
1.2. Relevanz des Zeitpunkts der Früherkennung	10
2. Konzeption eines maschinellen Frühwarnsystems	12
2.1. Variablenauswahl	13
2.2. Namensbasierte Imputation des Migrationshintergrundes	15
2.2.1. Herkunftsregionen in der Namensdatenbank	17
2.2.2. Zusammenstellung des Diktionärs	18
2.2.3. Namensanalyse	20
2.2.4. Evaluation des imputierten Migrationshintergrundes	22
2.2.5. Zwischenfazit	26
2.3. Beschreibung der Daten	28
2.3.1. Grund des Abbruchs	33
2.3.2. Zeitpunkt des Abbruchs	35
2.3.3. Demografie der Studierenden	36
2.3.4. Hochschulzugangsberechtigung	39
2.3.5. Akademische Leistungen	41
2.4. Feature Engineering	43
2.5. Verfahrensauswahl und Methodenbeschreibung	46
2.5.1. Regressionsanalyse	46
2.5.2. Neuronale Netze	47
2.5.3. Entscheidungsbäume	48
2.5.4. AdaBoost	50
2.5.5. Wahl des Identifikationsschwellenwerts	51
2.5.6. Gütekriterien eines Klassifikators	52
3. Die Ergebnisse des Früherkennungssystems FragSte	54
3.1. Lern- und Testkohorten UNI und FH	54
3.2. Identifikationsschwellenwert	55
3.3. Logit Model	57
3.4. Bagging with Random Forest und Neural Network	58
3.5. AdaBoost	62
3.6. Einfluss der Testkohorte auf die Prognosegenauigkeit	64
4. Zwischenfazit: Das Früherkennungssystem FragSte	68
III Gegenüberstellung verschiedener Arten von Frühwarnsystemen	70
1. Frühwarnsystem PASST?!	71
2. Frühwarnsystem SVM	72
3. Die Prognosegüte der Frühwarnsysteme	73
3.1. Prognosegenauigkeiten von FragSte	73
3.2. Prognosegenauigkeiten von PASST?!	74
3.3. Prognosegenauigkeiten von SVM	78

3.4. Diskussion der Ergebnisse	80
4. Zwischenfazit	82
IV Prognoserelevanz einzelner Informationen	83
1. Variablenbeschreibung und Methodenwahl	84
1.1. Variablenbeschreibung	84
1.2. Information Gain	85
1.3. Gain Ratio	86
1.4. Correlation Based Feature Selection	87
1.5. Faktoranalyse	88
2. Ergebnisse	90
2.1. Regressionsanalyse	90
2.2. Verzicht auf demografische Informationen	96
2.3. Information Loss	97
2.4. Information Gain der Variablen	99
2.5. Gain Ratio der Variablen	103
2.6. Auswahl der relevanten Variablen durch CFS	108
2.7. Faktoranalyse	111
2.8. Prognosegenauigkeit bei reduzierter Information	125
3. Zwischenfazit und Diskussion: Informationsgehalt von Variablen	127
V Anpassungen des Frühwarnsystems	129
1. Frühsterkennung innerhalb der ersten vier Monate	132
1.1. Feature Engineering	132
1.2. Einfluss des Prognosezeitpunkts im ersten Semester	134
1.3. Prognosegenauigkeit	135
2. Prognose des Vorlesungserfolgs	136
3. Prognosen anhand digitalem akademischen Verhalten	138
3.1. Feature Engineering	138
3.2. Prognoserelevanz digitalen akademischen Verhaltens	140
3.3. Prognosegenauigkeit	143
4. Zwischenfazit	144
VI Fazit und Ausblick	145
Literaturverzeichnis	149
Anhang:	161

Tabellenverzeichnis:

Tabelle 1. Studienabbruch und Data mining	9
Tabelle 2. Nach §3 HStatG zu erhebende Studierendendaten	14
Tabelle 3: Namensanalyse ausgesuchter Beispielnamen.....	21
Tabelle 4: Identifikationsquote der Herkunftsländer bei ausländischen Studierenden.....	22
Tabelle 5: Identifikationsquote bei ausländischen Studierenden.....	23
Tabelle 6: Ergebnis Namensanalyse SOEP-Test	26
Tabelle 7: Zusammenfassung Ermittlungsgenauigkeit des Migrationshintergrunds	27
Tabelle 8. Ethnische Zusammensetzung der Studierenden.....	28
Tabelle 9. Verbleibquote der Immatrikulationskohorten 2009-2017 zum WiSe 2017/18.....	29
Tabelle 10. Beschreibung der Daten: UNI	31
Tabelle 11. Beschreibung der Daten: FH.....	32
Tabelle 12. Gründe des Abbruchs	34
Tabelle 13. Abbruchquoten nach Migrationshintergrund.....	37
Tabelle 14. Hochschulzugangsberechtigung UNI	39
Tabelle 15. Hochschulzugangsberechtigung FH.....	40
Tabelle 16. Studierende ohne CP.....	43
Tabelle 17. Variablenbeschreibung	45
Tabelle 18. Anzahl ehemaliger Studierender in Lern- und Testkohorten.....	55
Tabelle 19. Prognosegenauigkeit des Logitmodells (tatsächlichen Abbruchquote).....	57
Tabelle 20. Prognosegenauigkeit des Logitmodells (durchschnittliche Abbruchquote)	58
Tabelle 21. Prognosegenauigkeit des BRF (tatsächliche Abbruchquote).....	59
Tabelle 22. Prognosegenauigkeit des BRF (durchschnittliche Abbruchquote)	59
Tabelle 23. Prognosegenauigkeit des neuronalen Netzes (tatsächliche Abbruchquote)	59
Tabelle 24. Prognosegenauigkeit des neuronalen Netzes (durchschnittliche Abbrüchen)	60
Tabelle 25. Prognosegenauigkeit des AdaBoost (tatsächliche Abbruchquote)	63
Tabelle 26. Prognosegenauigkeit des AdaBoost (durchschnittliche Abbruchquote).....	64
Tabelle 27 Prognosegenauigkeit des BRF — in Abhängigkeit der Testkohorte	65
Tabelle 28. Abhängigkeit der Prognosegenauigkeit von der Betrachtungsdauer	66
Tabelle 29: Prognosegüte FragSte	73
Tabelle 30: Prognosegüte PASST?!.....	74
Tabelle 31: Identifikationsquoten PASST?!.....	75
Tabelle 32: Genauigkeit der Merkmale 1 (<2 Prüfungen im ersten Semester) und 2 (<30CP in zwei Semestern) des Studienfortschritts in PASST?!.....	76
Tabelle 33: Merkmal 3 - Genauigkeit des zweiten Wiederholungsversuchs in PASST?!.....	77
Tabelle 34: Merkmal 4 - Genauigkeit des zwei oder mehrmaligen Rücktritts in PASST?!	78
Tabelle 35: Prognosegüte SVM.....	79
Tabelle 36: Identifikationsquoten SVM	79
Tabelle 37: Merkmal 1 - Genauigkeit des Studienfortschritts (>9CP) in SVM.....	80
Tabelle 38. Beispielrechnung Information Gain	85
Tabelle 39. Effekte von demografischen und Leistungsdaten auf Studienabbrüche (UNI)....	94
Tabelle 40. Effekte von demografischen und Leistungsdaten auf Studienabbrüche (FH)	95
Tabelle 41. Prognosegenauigkeit unter ausschließlicher Verwendung von Leistungsdaten..	96
Tabelle 42. Prognosegenauigkeit des AdaBoost bei tatsächlicher Abbruchquote	97
Tabelle 43. Information Loss	98
Tabelle 44. Information Gain – UNI	101
Tabelle 45. Information Gain – FH.....	102

Tabelle 46. Gain Ratio – UNI.....	105
Tabelle 47. Gain Ratio – FH.....	107
Tabelle 48. Relevante Informationen nach Semester - UNI.....	109
Tabelle 49. Relevante Informationen nach Semester - FH.....	110
Tabelle 50: Faktoren des reduzierten Variablensetting der UNI.....	123
Tabelle 51: Faktoren des reduzierten Variablensetting der FH.....	124
Tabelle 52. Prognosegüte ausgesuchter Prognosemodelle der Kohorten 2010-2012.....	125
Tabelle 53: Prognosegenauigkeit unter Berücksichtigung aller Studienabbrecher.....	131
Tabelle 54: Tage bis Semesterende.....	133
Tabelle 55: Variablen der Prüfungsanmeldung.....	133
Tabelle 56: Information Gain der Prüfungsanmeldungen.....	134
Tabelle 57. Benchmark Prognosegenauigkeiten FragSte im nullten Semester der UNI.....	135
Tabelle 58. Prognosegenauigkeiten Frühstwarnsystem.....	136
Tabelle 59. Genauigkeit der Erfolgsprognose einer Zweitsemesterklausur.....	137
Tabelle 60: Variablen des digitalen akademischen Verhaltens.....	139
Tabelle 61: Information Gain Skript.....	140
Tabelle 62: Information Gain Übungen.....	141
Tabelle 63: Information Gain Online Tests.....	141
Tabelle 64: Information Gain weiteres Verhalten.....	141
Tabelle 65: Prognosegenauigkeit digitalen akademischen Verhaltens.....	143
Tabelle 66. Faktoranalyse UNI Semester 0.....	161
Tabelle 67. Faktoranalyse UNI Semester 1.....	161
Tabelle 68. Faktoranalyse UNI Semester 2.....	162
Tabelle 69. Faktoranalyse UNI Semester 3.....	163
Tabelle 70: Faktoranalyse UNI Semester 4.....	164
Tabelle 71: Faktoranalyse FH Semester 0.....	165
Tabelle 72: Faktoranalyse FH Semester 1.....	165
Tabelle 73: Faktoranalyse FH Semester 2.....	166
Tabelle 74: Faktoranalyse FH Semester 3.....	167
Tabelle 75: Faktoranalyse FH Semester 4.....	168

Abbildungsverzeichnis:

Abbildung 1. Studienabbrüche im Semesterverlauf	35
Abbildung 2. Altersverteilung der Studierenden bei Immatrikulation	36
Abbildung 3. Abbruchquoten nach Alter bei Immatrikulation	37
Abbildung 4. Abbruchquote und Migrationshintergrund	38
Abbildung 5. Abbruchquote nach HZB-Note der UNI	41
Abbildung 6. Abbruchquote nach Durchschnittsnote	42
Abbildung 7. Prognosegenauigkeit in Abhängigkeit des Identifikationsschwellenwertes	56
Abbildung 8. ROC und AUC ROC — UNI	61
Abbildung 9. ROC und AUC ROC — FH	62
Abbildung 10. Faktoranalyse UNI Semester 0	113
Abbildung 11. Faktoranalyse UNI Semester 1	114
Abbildung 12. Faktoranalyse UNI Semester 4	116
Abbildung 13. Faktoranalyse FH Semester 0	117
Abbildung 14. Faktoranalyse FH Semester 1	119
Abbildung 15. Faktoranalyse FH Semester 4	121
Abbildung 16. ROC UNI	126
Abbildung 17. ROC FH	127
Abbildung 18. Zeitpunkt der Verfügbarkeit von Leistungsdaten im Folgesemester	129
Abbildung 19. Information Gain und Korrelation nach Vorlesungswochen	142

I Einleitung

Studienabbrüche sind mit erheblichen Kosten für Studierende, Hochschulen und die Gesellschaft verbunden: Ungeachtet der von dem Studienabbrecher erlangten Bildungsrenditen, kann ein Studienabbruch als Fehlinvestition interpretiert werden und wird von den Studienabbrechern selbst meist als Scheitern und persönliches Versagen empfunden (Larsen et al., 2013; Schröder & Daniel, 1998). So ist es nur folgerichtig, dass die Bemühungen Studienabbrüche zu verhindern verstärkt werden (Gaebel et al., 2012). Ausdruck dieser Bemühungen ist auch, dass im Zukunftsvertrag „Studium und Lehre stärken“ vom 06.07.2019 in §3 III die Hochschulfinanzierung künftig zu 20% an den Studienanfängern, zu 20% an der Zahl der Absolventen und zu 60% an der Zahl der Absolventen in Regelstudienzeit bemisst. Um die Zahl der Studienabbrecher zu senken, kann Unterstützung angeboten werden. Diese Unterstützungsangebote beinhalten in einem ersten Schritt häufig ergebnisoffene Beratung. Ziel kann sein, den Studienabbruch zu verhindern oder wenn nötig auch zu beschleunigen und Alternativen aufzuzeigen. Beide Entscheidungsoptionen können durch die Hochschulen unterstützt und begleitet werden. In aller Regel unterhalten Hochschulen hierzu viele Unterstützungsangebote für Studierende, die sich in fachliche, persönliche und psychologische Angebote unterscheiden lassen. Während fachliche Angebote bspw. mathematische Fähigkeiten oder wissenschaftliches Schreiben trainieren, werden in persönlichen Beratungsangeboten die Selbstmotivation oder Selbstorganisation adressiert. Allerdings können Hochschulen diese Unterstützungsangebote in der Regel nicht gezielt abbruchgefährdeten Studierenden anbieten – es fehlt an der Information, welcher Studierende abbruchgefährdet ist. Dazu können Frühwarnsysteme genutzt werden (Falk & Marschall, 2019; Villano et al., 2018; Brandstätter et al., 2006). Weiterhin kann ein Frühwarnsystem dazu verwendet werden den Erfolg von Maßnahmen zu evaluieren oder es kann als Grundlage für bildungswissenschaftliche Erhebungen genutzt werden. Durch eine Früherkennung ist es möglich abbruchgefährdete Studierende gezielt zu befragen, ohne dass hilfsweise Informationen zur Abbruchneigung genutzt werden müssen.

In der vorliegenden Arbeit wird ein an allen deutschen Hochschulen einsetzbares System zur „Früherkennung abbruchgefährdeter Studierender“ (FragSte) konzipiert, implementiert und evaluiert. Die Prognosegenauigkeit des konzipierten Frühwarnsystems wird getestet, mit anderen Frühwarnsystemen verglichen und die verfügbaren Informatio-

nen auf Prognoserelevanz getestet. Hierauf aufbauend werden alternative Systeme entwickelt und getestet, die entweder durch datensparsamere Konzeption effizienter sind oder frühere Prognosen des Studienerfolgs beziehungsweise von Prüfungsergebnissen erlauben. Diese Arbeit kann auch als Leitfaden und Entscheidungshilfe zur Konzeption und Implementation von Frühwarnsystemen insbesondere für Hochschulen verstanden werden – der Praxisverwertbarkeit wird ein hoher Stellenwert eingeräumt.

FragSte wird an zwei mittelgroßen nordrheinwestfälischen Hochschulen durchgeführt: Einer staatlichen Universität (UNI) mit etwa 23.000 Studierenden und 90 verschiedenen Bachelor-Studiengängen sowie einer privat refinanzierten Fachhochschule (FH) mit etwa 6.500 Studierenden, 26 Bachelorstudiengängen und Studiengebühren in Höhe von etwa 400,- € pro Monat. Es wurden eine staatliche Universität und eine private, gebührenfinanzierte Fachhochschule gewählt, um Hinweise für die Anwendbarkeit der Modelle in verschiedenen Kontexten und für die Übertragbarkeit der Ergebnisse zu erhalten.

Zur Analyse von Studienabbrüchen muss zwischen Schwund und Abbruch unterschieden werden. Studierende, die das Hochschulsystem ohne Studienabschluss verlassen, werden als Studienabbrecher bezeichnet. Studierende, die eine Hochschule ohne Studienabschluss verlassen, ihr Studium aber an einer anderen Hochschule weiterverfolgen, zählen zum Schwund einer Hochschule (Brandstätter et al., 2006). Da aus Sicht der Hochschulen eine Differenzierung zwischen Schwund und Abbruch nicht möglich ist, werden in der vorliegenden Analyse alle Studierenden, die die Hochschule ohne Abschluss verlassen, als Studienabbrecher bezeichnet. Studienfachwechsler innerhalb der Hochschule verlassen hingegen die Hochschule nicht und werden daher nicht als Studienabbrecher gezählt.

Die Arbeit gliedert sich wie folgt:

In Teil II wird ein Frühwarnsystem (FragSte) entwickelt. Dabei wird die Auswahl der verwendeten Informationen unter Berücksichtigung des Datenerhebungsaufwands und des Anspruchs der Übertragbarkeit auf andere Hochschulen vorgenommen (Teil II, Kapitel 2). Auch die Prognosemethode soll die Übertragbarkeit auf andere Hochschulen erlauben. In Teil II Kapitel 3 werden die Prognosegenauigkeiten einzelner Verfahren und des Frühwarnsystems für verschiedene Prognosezeitpunkte für beide Hochschulen dargestellt. Ebenfalls wird in Teil II (Abschnitt 2.2) eine automatisierte, namensbasierte Imputation des Migrationshintergrundes konzipiert, getestet und in das Frühwarnsystem eingefügt. Wesentliche Inhalte von Teil II wurden in Berens et al. (2019) veröffentlicht.

In Teil III wird das entwickelte Frühwarnsystem mit den Frühwarnsystemen PASST?! (Schulze-Stocker et al., 2017) und SVM (Westerholt et al., 2018) verglichen. Hierbei werden insbesondere die Prognosegenauigkeiten in Relation der benötigten Daten zu verschiedenen Zeitpunkten im Studium betrachtet. Dadurch wird die Effizienz der Systeme in Relation zueinander verdeutlicht. Auszüge von Teil III wurden in Berens & Schneider (2019) veröffentlicht.

In Teil IV werden die genutzten Studierendendaten auf Prognoserelevanz untersucht, einerseits um Erkenntnisse hinsichtlich der Indikatoren des Studienabbruchs zu erhalten und andererseits, um irrelevante Informationen zugunsten eines datensparsameren Systems zu entfernen. Es wird geprüft, ob eine sparsamere Datenauswahl – bis hin zu einem reinem ECTS Monitoring - vergleichbare Prognoseergebnisse aufweist. Inhalte von Teil IV wurden in Schneider et al. (2019) veröffentlicht.

In einem letztem Teil V werden aufbauend auf den vorangegangenen Teilen Systeme entwickeln, die bereits in den ersten Studienmonaten Prognosen des Studienerfolgs oder auch den Erfolg spezifischer Prüfungen möglichst früh und genau erlauben.

II Konzeption eines Frühwarnsystems

Im Folgenden wird die Konzeption des Frühwarnsystems FragSte beschrieben. Dabei wird auf die Auswahl der verwendeten Daten, eine Imputation des Migrationshintergrundes von Studierenden, dem feature engineering und der Verfahrensauswahl eingegangen. Wesentliche Inhalte des Teils II sind wie folgt veröffentlicht worden:

Berens, J., Schneider, K., Görtz, S., Oster, S. & Burghoff, J. (2019) Early Detection of Students at Risk - Predicting Student Dropouts Using Administrative Student Data from German Universities and Machine Learning Methods. *Journal of Educational Data Mining*, 11(3), 1-41

1. Studienabbruch und Frühwarnsysteme

Die maßgeblichen Erklärungsmodelle von Studienabbrüchen sind international überwiegend durch soziologische Sichtweisen dominiert (Sarclotti & Müller, 2011; Heublein, 2014; Larsen et al., 2013). Ausgehend von Tinto's (1975) „student integration model“, hat die soziale und akademische Integration des Studierenden eine zentrale Bedeutung. Pascarella und Terenzini (1979) übernehmen den Gedanken der Integration und erweitern das Modell um die Unterscheidung zwischen erzwungenen und freiwilligen Abbrüchen. Bean (1983) hingegen relativiert die Bedeutung der Integration als Abbruchursache und fügt in seinem „student attrition model“ die studentische Zufriedenheit als zentrale Variable für dieses Ereignis ein. Allen drei Ansätze ist gemein, dass sie den Prozess des Abbruchs im amerikanischen Kontext betrachten (Larsen et al., 2013, S. 46). Aber auch hinsichtlich der deutschen Hochschullandschaft werden bei Fragestellungen des studentischen Erfolges der Studierendenzufriedenheit eine zentrale Bedeutung beigegeben (Gold, 1988; Harvey & Green, 2000; Wiers-Jenssen et al., 2002; Schiefele & Jacob-Ebbinghaus, 2006; Brandstätter et al., 2006; Bargel et al., 2008). Insbesondere Apenburg (1980) und Ströhlein (1983) ordnen die Abbruchtendenz in theoretische Konstrukte der Studierendenzufriedenheit ein. Darauf aufbauend zeigen Winteler (1984) und Meulemann (1991) empirisch, dass die Studienzufriedenheit mit der Studienabbruchtendenz in hohem Maße korreliert. Brandstätter et al. (2006) weisen hierüber hinaus in einer Längsschnittstudie durch Analyse von Prüfungsergebnissen und Befragungen eine kausale Wechselwirkung nach. Hiernach ist die Unzufriedenheit gleichzeitig Ursache wie Begleiterscheinung des Studienabbruchs (Brandstätter et al., 2006, S. 126).

Erschwerend kommt hinzu, dass die Ursachen von Studienabbrüchen nach herrschender Meinung weniger auf eine Ursache zurückzuführen sind, sondern vielmehr einen Prozess mit multikausalem Ursachenkonglomerat mit individueller Intensitätskombination darstellen (Heublein et al., 2017; Burrus et al., 2013; Larsen et al., 2013, S. 45; Gold, 1988). Auch hinsichtlich der deutschen Hochschullandschaft folgen Heublein et al. (2017; 2014) dem prozessualen Modellaufbau und identifizieren sieben Motive des Studienabbruches: *Leistungsanforderung, Finanzen, Prüfungsversagen, Motivation, Studienbedingungen, berufliche Neuorientierung und Krankheit*. Die Motive wirken dabei wechselseitig und in unterschiedlichen Ausprägungen; eine Gewichtung wird nicht vorgenommen. Das Modell von Thiel et al. (in Blüthmann, 2012) versucht den Studienerfolg zu erklären. Dabei werden basierend auf Modellen der Schul- und Unterrichtsqualität und dem Angebots- und Nutzungsmodell von Helmke (2009) in einer Mehrebenenstruktur die Lebensbedingungen, die Studienbedingungen, die Eingangsvoraussetzungen und als Ergebnis dieser Dimensionen der Lernprozess betrachtet. Von einer Betrachtung vollständiger Erklärungsmodelle losgelöst, werden bei Fragestellungen des studentischen Erfolges in erheblichen Teilen der Literatur der Studierendenzufriedenheit eine zentrale Bedeutung beigemessen (Gold, 1988; Harvey & Green, 2000; Wiers-Jenssen et al., 2002; Schiefele & Jacob-Ebbinghaus, 2006; Bargel et al., 2008).

Zusammengefasst kommen der studentischen Integration, gesellschaftlich wie akademisch, der Zufriedenheit, der persönlichen Lebenssituation und den institutionellen Rahmenbedingungen zentrale Bedeutung zu.

Üblicherweise werden in der deutschen Hochschullandschaft Erkenntnisse zu Studienabbrüchen und Studienerfolgen durch Befragungen ermittelt (Blüthmann, 2012; Heublein et al., 2017; Gold, 1988; Ströhlein, 1983). Larsen et al. (2013, S. 66) fasst zusammen:

“In contrast to the studies in the [data driven] British research approach they [studies in the German approach] rely heavily on extensive questionnaire surveys as their primary data source.”

Jedoch sind Befragungen von Studierenden – vor ihrem Abbruch – in der Ursachenforschung von Studienabbrüchen nur sehr eingeschränkt verwertbar. In ex-ante Befragungen muss die abhängige Variable des Studienabbruchs mit dem formulierten Wunsch des Abbruchs ersetzt werden. Dieser ist aber in der Literatur als Prädiktor des tatsächlichen Abbruchs umstritten. Während Gold (1988, S. 34) und Bean (1982) der formulierten

Absicht, das Studium abzubrechen, einen großen Einfluss auf den tatsächlichen Abbruch bescheinigen, attestiert Georg (2008, S. 202), dass der Zusammenhang zwischen Abbruchneigung und Handlungsvollzug theoretisch wie empirisch ungeklärt ist und lediglich spekulative Überlegungen angestellt werden können. Auch Brandstätter et al. (2006) konnten empirisch keinen Zusammenhang zwischen der durch den Studierenden eingeschätzten Wahrscheinlichkeit des künftigen Abbruches und dem tatsächlichen Studienabbruch nachweisen. Aber auch in ex-post Befragungen ist fraglich, ob die im Anschluss an den Studienabbruch wahrgenommenen Ursachen bereits während des Studiums hätten formuliert werden können und ob sie geeignet sind, einen drohenden Abbruch zu prognostizieren. Es besteht die Gefahr, dass Befragte nach einem gesellschaftlich stigmatisierten Studienabbruch im Wege der Reduktion kognitiver Dissonanzen (Festinger, 1957), die Einschätzung der persönlichen und institutionellen Ursachen ihren subjektiven Befindlichkeiten angleichen (Georg, 2008).

Sekundärdatenanalysen, die im Gegensatz zu den umfragebasierten Erhebungen die Studierendendaten betrachten, unterscheiden sich von den vorgenannten Untersuchungen ganz erheblich. Anders als durch Umfragedaten, kann durch die Auswertung administrativer Studierendendaten die Wahrscheinlichkeit des Abbruchs ermittelt werden (Arulampalam et al., 2005; Danilowicz-Gösele et al., 2017). Von maßgeblichem Einfluss ist dabei die akademische Leistung des Studierenden und die Leistungsbandbreite der Kohorte. Auch der Note der HZB wird als Prädiktor der Studienergebnisse in der Literatur häufig eine besondere Stellung eingeräumt (Trapmann et al., 2007; Brandstätter et al., 2006; Gold et al., 2005).

Mit zunehmend mehr verfügbaren Daten an Hochschulen und Entwicklungen im Bereich des *Data mining* werden zur Prognose künftiger Studienabbrüche zunehmend maschinelle Verfahren eingesetzt – es entstand eine noch immer vergleichsweise junge Disziplin des *educational data mining* (Baker & Yacef, 2009).

1.1. Educational Data Mining

Eines der wesentlichen Ziele des educational data mining ist die Prognose von Lernerfolgen (Baker, 2010; Baker & Yacef, 2009). Dabei können Lernerfolge einzelner Aufgaben und Übungen, Modulprüfungen oder des gesamten Studiums untersucht und prognostiziert werden (Asif et al., 2017). Beispielsweise Feng et al. (2006) prognostizierten Kursnoten anhand eines intelligenten Nachhilfesystems (intelligent tutoring system) unter Verwendung der Information, wie viel Unterstützung ein Studierender benötigte, um eine

Aufgabe zu lösen. Strecht et al. (2015) und Barber and Sharkey (2012) hingegen prognostizierten den Erfolg einer Vorlesung, beziehungsweise eines Studienprogramms. Eine der ersten veröffentlichten Studien, die den Studienerfolg prognostizierten, stammt von Kotsiantis et al. (2003). In dieser werden die vorhandenen Studierendaten des Fernstudiengangs Informatik der Hellenic Open University in einen Lern- und eine Testdatensatz aufgeteilt und mittels Verfahren des maschinellen Lernens auf ihre Prognosefähigkeit des Studienerfolgs untersucht. In einer Kombination von demografischen Studierendaten und erster akademischer Leistungen konnten 70% erfolgreicher Studienabschlüsse richtig prognostiziert werden. Zum Einsatz kommen Entscheidungsbäume, neuronale Netze, naives bayes Verfahren, logistische Regressionanalyse, support vector machines und instant learning algorithms. Spätere Studien folgen überwiegend diesem Aufbau und der Verfahrensauswahl. So zum Beispiel die Studien von Kotsiantis et al. (2003), Xenos (2004), Minaei-Bidgoli et al. (2004), Nghe et al. (2007), Dekker et al. (2009), Zhang et al. (2010), Bayer et al. (2012), Er (2012) und Yukselturk et al. (2014), Sara et al. (2015) und Santana et al. (2015). Im Vergleich dieser Studien lässt sich feststellen, dass

- die Studien trotz vergleichbarer Verfahren durch sehr unterschiedliche Variablen, Untersuchungsobjekte und Stichprobengrößen nur eingeschränkt vergleichbar sind,
- weit überwiegend mindestens demografischen Studierendaten und Leistungsdaten verwendet werden konnten,
- die Prognosegenauigkeit mit zunehmend späteren Betrachtungszeitpunkt im Studium zunimmt,
- die Prognosegenauigkeit gleicher Verfahren zwischen den Studien in Abhängigkeit der verfügbaren Daten stark variiert und
- der Unterschied zwischen der Prognosegenauigkeit der Verfahren innerhalb einer Studie meist nur wenige Prozentpunkte beträgt.

Vieles deutet darauf hin, dass die Prognosegenauigkeit primär von der Aussagekraft der verwendeten Informationen abhängt und nur sekundär vom Verfahren. Die Art der verfügbaren und verwendbaren Studierendaten hängt jedoch stark von juristischen und institutionellen Rahmenbedingungen ab. So haben üblicherweise Fernstudiengänge, insbesondere Online-Studiengänge bessere administrative Informationen über ihre Studierenden als Präsenzstudiengänge. In vielen Studien zu Fernstudiengängen lassen sich erhebliche Teile des studentischen Verhaltens elektronisch durch die Häufigkeit und Regelmäßigkeit von An- und Abmeldungen im System, zu Kursen, zu Literaturdatenbanken,

Skriptservern und zu Leistungsüberprüfungen abbilden. Weiterhin können die studentischen Leistungen dezidiert bis zu einzelnen Antworten in speziellen Klausuren ausgewertet werden (Minaei-Bidgoli et al., 2004).

Aber auch in Präsenzstudiengängen können zusätzliche mögliche Einflussgrößen auf den Studienerfolg die Prognosegenauigkeit erhöhen. So lagen Kotsiantis et al. (2003) neben den üblichen demografischen Studierendaten (Alter, Geschlecht, Art der Hochschulzugangsberechtigung, Wohnort) noch weitere Informationen über die Studierenden vor, unter anderem die Anzahl der Kinder, Familienstand, der Beschäftigungsgrad, die Art der Beschäftigung und der Erfahrungsgrad mit Computern. Die üblichen Leistungsdaten (Anzahl und Ergebnis der Prüfungen) konnten um die Teilnahmebereitschaft optionaler „face-to-face Consulting meetings“ ergänzt werden.

Zhang et al. (2010) hingegen legten ihrer Datenselektion Tinto's integration modell (1975) zugrunde und nutzen diejenigen Informationen über ihre Studierenden, die die gesellschaftliche und akademische Integration des Studierenden bestmöglich beschreibt. Hierfür wurden die Leistungsdaten, die Anmeldungen in Online-Lern-Plattformen und Skriptservern, der Nutzungsgrad der Universitätsbibliothek, das Online-Lese-Verhalten der Online-Bibliothek sowie das Aktivitätslevel in universitätseigenen Foren ausgewertet. Insbesondere das Lernverhalten und die studentische Interaktion mit Lehrenden konnte damit abgebildet werden. Andere Bestandteile von Tintos Modell, wie die persönliche Entwicklung des Studierenden, das Fachinteresse und die gesellschaftliche integration konnten nicht beschrieben werden. Die gesellschaftliche Integration in das universitäre Umfeld adressieren hingegen Bayer et al. (2012). Hierzu werteten sie das Verhalten und die Verbindungen von 775 Studierenden und ihrer direkten Kommilitonen in sozialen Netzwerken aus. Es wurden die besuchten Nutzerprofile, Austausch von Dateien, Emails, markierte Beiträge und Beteiligungen an Gruppendiskussionen berücksichtigt und die Stärke der Verbindung gewichtet. Es zeigte sich, dass aktivere und vernetztere - also integriertere - Studierende auch die erfolgreichereren waren. Weiterhin konnte die Prognosegenauigkeit basierend auf demografischen Studierendendaten und akademischen Leistungen durch Hinzufügen des Sozialverhaltens um etwa 5 Prozentpunkte auf 72% im ersten Semester gesteigert werden. Eine gute Darstellung fortgeschrittener Methoden des educational data mining bieten Romero und Ventura (2010) sowie Peña-Ayala (2014).

Tabelle 1. Studienabbruch und Data mining

Autoren	Journal	Jahr	Titel	Untersuchungsobjekt	n	Verfahren	Accuracy
Santana, M.; Costa, E.; Neto, B.; Silva, I.; Rego, J. Alagoa, Brasilien	Working Paper	2015	A Predictive Model for Identifying Students with Dropout Profiles in Online Courses	Distance Learners	1.800	Naives Bayes AD (J48) Support Vector Machine Neuronal Net (MP)	85,5% 86,5% 92% 90,9%
Sara, N.; Halland, R.; Igel, C.; Alstrup, S. Copenhagen	European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Pp. 22-24, Bruges (Belgium)	2015	High-School Dropout Prediction Using Machine Learning: A Danish Large-scale Study	High School Students	72.598	Naives Bayes Support Vector Machine Random Forrest CART	85,6% 90,4% 93,5% 89,8%
Yukselturk, E.; Ozekes, S.; Kilic, Y. Turkey	European Journal of Open, Distance and e-Learning. Vol. 17, No. 1	2014	Predicting Dropout Student: An Application of Data Mining Methods in an Online Education Program	Distance Learners	189	K-NN Decision Tree Naives Bayes Neuronal Net	87% 79,7% 76,8% 73,9%
Bayer, J.; Bydzovska, H. Geryk, J.; Obsivac, T.; Popelinsky, L. Czech Republik	Proceedings of the 5th International Conference on Educational Data Mining	2012	Predicting drop-out from social behaviour of students	Bachelor Students of Applied Informatics	775	ZeroR Naives Bayes Support Vector Machines (SMO) IB1 lazy learner OneR PART Decision Tree (J48)	50% - 77% 68% - 79% 72% - 88% 66% - 91% 62% - 84% 70% - 91% 70% - 87%
Er, E. Turkey	International Journal of Machine Learning and Computing. Vol. 2, No. 4	2012	Identifying At-Risk Students Using Machine Learning Techniques: A Case Study with IS 100	students	625	K-Star Decision Tree C4.5 Naives Bayes Decision scheme	60% - 82% 60% - 81% 60% - 79% 65% - 85%
Zhang, Y.; Oussena, S.; Clark, T.; Kim, H. UK	Proceedings of the 12th International Conference on Enterprise Information Systems,	2010	Use Data Mining to Improve Student Retention in Higher Education - a Case Study	Students	4.223	Naive Bayes Support Vector Machine Decision Tree	89,5 83,5% 81,3%
Dekker, G.; Pechenizkiy, M.; Vleeshouwers, J. Netherland	Educational Data Mining	2009	Predicting Students Drop Out: A Case Study	Students in first semester	648	CART Decision Tree C4.5 Bayes Net Logit Jrip Random Forrest OneR	79% 80% 75% 79% 77% 79% 75%
Nghe, N. T.; Janecek, P.; Haddaway, P. Thailand	Frontiers in Education Conference-Global Engineering: Knowledge Without Borders, Opportunities Without Passports, FIE'0, 37th Annual IEEE	2007	A Comparative analysis of techniques for predicting academic performance	Undergraduate and Postgraduate Students	20.492 Undergraduate CTO 936 Graduate AIT	Decision Tree (DT) Bayesian Network	72%-94% 70%-93%
Minaei-Bidgoli, B.; Kortemeyer, G.; Punch, W. F. USA	Proceedings of the Seventh IASTED International Conference on Computers and Advanced Technology in Education	2004	Enhancing Online Learning performance: An Application of Data Mining Methods	Students in Online-Courses	2.174	Quadratic Bayesian classifier 1-nearest neighbor (1-NN) k-nearest neighbor (k-NN) Parzen-Window	52% - 72% 62% - 77% 55% - 79% 54% - 70%
Kotsiantis, S.; Pierrakeas, C.; Pintelas, P. Greece	International Conference on Knowledge-Based and Intelligent Information and Engineering Systems. pp. 267-274, Springer Berlin/Heidelberg	2003	Preventing Student Dropout in Distance Learning Using Machine Learning Techniques	Distance Learners, Hellenic Open University	345	Naives Bayes Decision Tree C4.5 neural network (Backpropagation) Support Vector Machines (SMO) k-nearest neighbor (k-NN) Maximum Likelihood Estimation	61% - 83% 63% - 84% 63% - 84% 58% - 83% 60% - 78% 60% - 84%

1.2. Relevanz des Zeitpunkts der Früherkennung

Grundsätzlich kommen alle Studien zu dem Ergebnis, dass die Prognosen in späteren Semestern genauer werden, da zu einem späteren Zeitpunkt mehr Informationen über die Studienleistungen vorliegen und die Kohorten zunehmend aus gut zu prognostizierenden künftigen Absolventen bestehen. Jedoch senkt ein später Prognosezeitpunkt den Nutzen eines Frühwarnsystems. Der Nutzen eines Frühwarnsystems hängt maßgeblich davon ab, wie früh im Studium Prognosen möglich sind. Je später im Studium prognostiziert wird, desto mehr Studienabbrecher haben die Hochschule bereits verlassen und können nicht mehr erreicht werden und desto mehr Zeit haben die noch in der Kohorte befindlichen künftigen potentiellen Studienabbrecher bereits investiert: Das erste Studienjahr hat dabei einen besonderen Einfluss auf den akademischen Werdegang (Brahm et al., 2016) und korreliert stärker mit dem Studienerfolg als andere Studienjahre (Arcidiacono et al., 2016; Arnold & Pistilli, 2012; Stinebrickner & Stinebrickner, 2014; 2008; Barefoot et al., 2005).

Die Abbruchneigung von Studierenden sollte also früh erkannt werden, um im Prozess des Studienabbruchs möglichst früh Unterstützung anbieten zu können (Heublein et al., 2017; Wilson et al., 2016; Arulampalam et al., 2005). Frühe Interventionen erweisen sich zudem in der Regel wirksamer als späte (Bourn, 2007), insbesondere dann, wenn sie zielgerichtet abbruchgefährdeten Studierenden angeboten werden (Zhang et al., 2014). Nach Seidmann (1996) hängt der studentische Erfolg abbruchgefährdeter Studierender von der frühen Identifikation und der frühen, intensiven sowie kontinuierlichen Intervention ab:

“Retention = Early Identification + (Early + Intensive + Continuous) Intervention”

Idealerweise richten sich Unterstützungsangebote bereits in der Studieneingangsphase zielgerichtet an die Studierenden (Ahles et al., 2016, S. 129; Blüthmann et al., 2008). Dafür müssen aber auch die für die Prognose notwendigen Informationen verfügbar sein. Diese setzen sich an deutschen Hochschulen aus den Informationen, die bei Immatrikulation erhoben werden und Leistungsdaten zusammen. Da aber an Hochschulen in Deutschland erste Prüfungen in der Regel erst am Ende des ersten Semesters abgelegt werden, liegen Leistungsdaten frühestens zu Beginn des zweiten Semesters vor – unter Berücksichtigung einer angemessenen Korrekturzeit, Eingabezeit ins System und anschließender Datenaufbereitung und Datenexport tendenziell Mitte bis Ende des zweiten Semesters. Zu einem früheren Zeitpunkt sind nur Informationen verfügbar, die bei der

Immatrikulation erhoben wurden. Prognosen, die allerdings nur auf im Anmeldeprozess erhobenen persönlichen Daten beruhen, könnten weniger akzeptiert sein, als ein auf akademischen Leistungen beruhendes Frühwarnsystem.

Sofern Prognosen bereits früh im Studium mit hinreichender Genauigkeit möglich sind, bietet ein Frühwarnsystem die Möglichkeit des Monitorings einzelner Studierendengruppen, Studiengängen oder ganzer Kohorten. Damit stellt das Frühwarnsystem einerseits einen guten Ausgangspunkt der deskriptiven Forschung dar und bietet andererseits Hochschulführungen wichtige Erkenntnisse über die eigene Hochschule. Es kann für die strategische, taktische und operative Hochschulentwicklung wichtige Informationen und Entscheidungshilfen zur Verfügung stellen. So ist es beispielsweise möglich, Veränderungen im Studienverlauf, den Einfluss von Eintrittsbarrieren oder Studiengebühren zu messen oder die Effizienz von Interventionsmaßnahmen oder Hilfsprogrammen zu beobachten. Auch im Wege der Hochschulsteuerung kann das Frühwarnsystem eingesetzt werden. Beispielsweise ermöglicht ein Frühwarnsystem eine effiziente Maßnahmenallokation von Hilfsprogrammen und Interventionen. In der Regel werden an einer Hochschule eine Vielzahl von Maßnahmen zur Reduktion von Studienabbrüchen vorgehalten, die allerdings nur der Grundgesamtheit der Studierenden angeboten werden können und anschließend selbstreferenziell wahrgenommen werden.

Die bestehenden Maßnahmenlandschaften gegen Studienabbrüche der Hochschulen unterscheiden sich in Interventionen, die die Grundgesamtheit einer konkreten Studierendengruppe ansprechen (z.B. durch Tutorien, Brückenkurse, eine optimierte Eingangsphase oder allgemein eine verbesserte Didaktik) und in jene, die gezielt einzelne Studierende mit Unterstützungsbedarf ansprechen (in Form von fachlicher, persönlicher oder psychologischer Beratung und Hilfe, bspw. durch Studienberatung, soziale Dienste, Mentoring oder Workshops). Insbesondere die zweite Gruppe der unterstützenden Maßnahmen adressiert zwar zielgenau die defizitäre Situation der Studierenden, sie kommt den unterstützungsbedürftigen Studierenden aber möglicherweise nicht passgenau zugute, wenn das Angebot nicht bekannt ist oder nicht auf die intendierte Nachfrage stößt. Die Studierenden müssen die Maßnahme kennen und das Angebot auch wahrnehmen. Die selbstreferenzielle Maßnahmenteilnahme setzt wiederum ein Problembewusstsein und einen gewissen Handlungswillen des Studierenden voraus. Um abbruchgefährdete Studierende sowohl von den bestehenden Maßnahmen zielgerichtet in Kenntnis zu setzen, als auch bei der Problemidentifikation zu unterstützen und gegebenenfalls über die

Notwendigkeit der Handlung zu sensibilisieren, ist es notwendig abbruchgefährdete Studierende zu identifizieren.

2. Konzeption eines maschinellen Frühwarnsystems

Das Frühwarnsystem „FragSte“ soll auf Grundlage administrativer Studierendendaten möglichst früh im Studium einen drohenden Abbruch prognostizieren. Dabei können die Merkmale und Schwellenwerte abbruchgefährdeter Studierender entweder gesetzt werden oder anhand historischer Studierendendaten endogen ermittelt werden. In dem vorliegenden Teil II wird ein Frühwarnsystem konzipiert, in dem anhand exmatrikulierter Abbrecher und Absolventen die Merkmale gefunden werden, die eine frühzeitige Identifikation abbruchgefährdeter Studierender ermöglichen. Hierfür bieten sich insbesondere maschinelle Verfahren des überwachten Lernens (supervised learning) an.

Für die Untersuchung werden Studierendendaten ehemaliger Bachelor-Studierenden der Immatrikulationsjahre 2007 bis 2017 genutzt. Master- und andere Studiengänge werden nicht betrachtet. An der UNI umfassen die elf Kohorten 20.706 exmatrikulierte Studierende, davon 7.475 Absolventen und 13.231 Studienabbrecher. An der FH sind es 11.209 exmatrikulierte Studierende, davon 6.934 Absolventen und 4.275 Studienabbrecher. Bei Interpretation der Zahlen muss berücksichtigt werden, dass Studienabbrecher deutlich überpräsentiert sind. Damit ist beispielsweise eine Abbruchquote nicht ermittelbar: Zum Zeitpunkt des Datenexports im Oktober 2017, weisen die Immatrikulationskohorten 2015-2017 eine Beobachtungsdauer von weniger als sechs Semestern auf. Da nur exmatrikulierte Studierende berücksichtigt werden – bei noch immatrikulierten Studierenden ist der Studienaussgang noch ungewiss – setzen sich die Kohorten 2015 – 2017 aus Studienabbrechern und denjenigen seltenen Absolventen zusammen, die einen Abschluss in weniger als der Regelstudienzeit erlangten. Erst wenn eine Immatrikulationskohorte eine nur noch geringe Anzahl noch eingeschriebener Studierender (Verbleibquote) aufweist, können relative Kennzahlen, wie eine Abbruchquote, ermittelt und verglichen werden. Auch ein Vergleich der Hochschulen ist mit vorgenannten Zahlen nur schlecht möglich, da Studienabbrüche und -abschlüsse an der FH (ggf. durch die zu zahlenden Studiengebühren motiviert) früher stattfinden, als an der UNI und sich daher die Kohortenzusammensetzung im Studienverlauf anders verändert, als an der UNI.

Da an den betrachteten Hochschulen andere Erstabschlüsse als der Bachelor nicht mehr angeboten werden, werden auch nur historische Studierendendaten von Bachelor-

Studiengängen betrachtet. Das Frühwarnsystem kann jedoch problemlos auf Master-Studiengänge ausgeweitet werden. Ein Vergleich mit anderen Verfahren erfolgt in Teil III.

2.1. Variablenauswahl

Zur Konzeption eines universell einsetzbaren Frühwarnsystems, können nur standardisierte Daten berücksichtigt werden, die bereits an allen Hochschulen gleichartig erhoben werden und wurden. Anderenfalls ist das Frühwarnsystem nicht auf andere Hochschulen übertragbar. Als standardisierte und bundesweit verfügbare Datengrundlage wird daher nur auf Studierendendaten zurückgegriffen, die nach dem Hochschulstatistikgesetz zu erheben und speichern sind. Anders als die Hochschulgesetze der Länder normiert das Hochschulstatistikgesetz einen bundesweiten Standard zu erhebender Studierendendaten. Weiterhin hat der für die vorliegende Erhebung maßgebliche § 3 HStatG letztmalig in 1997 eine Änderung erfahren (BGBl. I, 1997, Artikel 7. Nr. 87, S. 3158). Für alle hier nach eingeführten Bachelor-Studiengänge müssen damit an allen deutschen Hochschulen die gleichen Datenfelder vorliegen.

Sollten an Hochschulen nach oder in Übereinstimmung mit den jeweiligen Hochschulgesetzen der Länder und den hochschulspezifischen Ordnungen weitere Studierendendaten erhoben werden, die dafür geeignet sind, den Studienerfolg zu prognostizieren, soll das Frühwarnsystem hierum erweiterbar sein. Ein Beispiel hierfür stellt die Note der Hochschulzugangsberechtigung dar, die nach herrschender Meinung ein gut geeigneter Prädiktor ist (Trapmann et al., 2007; Brandstätter & Farthofer, 2002; Schmidt-Atzert, 2005; Gold et al., 2005).

Öffentliche und staatlich anerkannte private Hochschulen haben nach §3 Hochschulstatistikgesetz die folgenden Studierendendaten zu erhebenden, speichern und regelmäßig zu melden.

Tabelle 2. Nach §3 HStatG zu erhebende Studierendendaten

Demografische Daten	Leistungsdaten
Matrikelnummer	Name der Prüfung
Vor- und Nachname	Leistungspunkte der Prüfung
Heimat- und Semesteranschrift	Prüfungsdatum
Geschlecht	Ergebnis der Prüfung
Geburtsort	evtl. Rücktritt der Prüfung
Geburtsjahr	evtl. trotz Anmeldung „nicht teilgenommen“
Krankenversicherung	
Staatsangehörigkeit	
Immatrikulationsdatum	
Hochschul- und Fachsemester	
Studiengang	
Abschlussart	
Art der HZB	
Ort der HZB	
Exmatrikulationsdatum	
Exmatrikulationsgrund	

Die Verwendung der in §3 HStatG normierten Studierendendaten im Frühwarnsystem stellt eine wesentliche Einschränkung zugunsten der universellen Einsetzbarkeit und Effizienz dar. Die in der Literatur üblicherweise mit einem Studienabbruch verbundenen Umstände und Gründe, lassen sich überwiegend nicht in den verwendeten Studierendendaten abbilden. Weder die Studienzufriedenheit - ungeachtet der Frage ob sie eine eigenständige Determinante ist, ein Moderator oder ein Mediator - noch die finanzielle oder familiäre Situation, die Motivation, die Sorgfalt bei der Studienplatzwahl, das Studien- und Fachinteresse, die berufliche Neigung, die akademische oder gesellschaftliche studentische Integration oder der Gesundheitszustand des Studierenden ist Bestandteil administrativer Studierendendaten im Sinne des §3 HStatG und damit im Frühwarnsystem auch nicht verwendbar. Lediglich die demografischen Studierendendaten und die akademischen Leistungen sind bekannt. Jedoch wird die zentrale Stellung der akademischen Leistungen als Prädiktor des Studienabbruchs immer wieder betont. Larsen et al. (2013) zitieren dazu in einem Literaturüberblick acht neuere Studien, die die Einschätzung von Tinto (1975) stützen: *“With respect to grade performance, many studies have shown it to be the single most important factor in predicting persistence in college.”* (in Larsen et al., 2013, S. 115-116).

Inwieweit die verfügbaren administrativen Daten die Genauigkeit des Frühwarnsystems einschränkt, hängt davon ab, ob und wie schnell sich die vorgenannten Einflussgrößen auf die akademischen Leistungen auswirken. In jedem Fall kann in einer reinen Betrachtung der akademischen Leistungsdaten ein drohender Studienabbruch erst dann

prognostiziert werden, wenn er sich in den Leistungsdaten ankündigt. Bei Verwendung demografischer Informationen – eine Korrelation mit dem Studienergebnis vorausgesetzt – sind Prognosen bereits früher möglich. Auch bei demografischen Informationen ist das Frühwarnsystem auf Studierendendaten im Sinne des §3 HStatG beschränkt. Das schließt aber auch alle Informationen ein, die aus den Studierendendaten nach §3HStatG imputiert werden können. Zur Betrachtung von Studienerfolg ist der der Migrationshintergrund eine wichtige Information, insbesondere dann, wenn der Studienerfolg von Studierenden aus bildungsfernen Schichten betrachtet werden soll (OECD, 2018; 2016). Allerdings ist aus den Studierendendaten nach §3 HStatG nur ein Migrationshintergrund der ersten Generation ermittelbar. Studierende mit Migrationshintergrund der zweiten und dritten Generation lassen sich vorerst nicht von Studierenden ohne Migrationshintergrund unterscheiden.

2.2. Namensbasierte Imputation des Migrationshintergrundes

Hochschulen sind in der Regel nur die Staatsbürgerschaft, der Ort der Hochschulzugangsberechtigung und der Geburtsort der Studierenden bekannt. Damit können ausländische Staatsbürger durch den Ort der Hochschulzugangsberechtigung in Bildungsausländer und Bildungsinländer unterschieden werden. Deutsche Staatsbürger, die im Ausland geboren sind, können als Migranten der ersten Generation angenommen werden. Migrationshintergründe der zweiten oder folgenden Generationen sind aus diesen Daten jedoch nicht ermittelbar. Aus diesem Grund wird, basierend auf Vor- und Zunamen automatisiert ein möglicher Migrationshintergrund imputiert. Deutschen Staatsbürgern mit Geburtsort in Deutschland, deren Vor- und Zuname auf einen Migrationshintergrund schließen lassen, werden als Migranten der ersten oder zweiten Generation bezeichnet.

Zum Zwecke der Analyse administrativer Studierendendaten ist es nur sinnvoll Verfahren zur Vorselektion einer Population zu betrachten. Verfahren, die eine Befragung einer Spezialpopulation ohne vorherige Selektion ermöglichen in dem der Befragte einen Folgebefragten empfiehlt, wie „respond driven sampling“ (Salganik & Heckathorn, 2004) oder „snowball sampling“ (Goodman, 1961; Salentin, 1999), sind nicht zielführend.

Sofern Vor- und Zunamen vorliegen, wird in der Literatur häufig auf eine Namenserkennung („Onomastik“) zurückgegriffen. In Betrachtung von Einzelfällen erreicht die Einordnung eines Muttersprachler eine akzeptable Güte (Salentin, 1999). Bei größeren Fallzahlen bedarf es automatisierter Verfahren. Dabei wird von der geographischen Verbrei-

tung spezieller Vor- und Zunamen auf die ethnische Zugehörigkeit geschlossen. In mehreren Studien wurden dieses Verfahren getestet und als hinreichend genau erachtet (Borhek, 1970; Diehl, 2005; Fernández et al., 2006; Farwick, 2009; McGarrigle & Kearns, 2009). Das bedeutet jedoch nicht, dass es mittels onomastischer Verfahren möglich ist, jede Person mit Migrationshintergrund zu identifizieren oder dass Fehlidentifikationen ausgeschlossen werden können. Zu diesen kommt es immer dann, wenn Namen in vielen Ländern gleichermaßen verbreitet sind oder Personen ohne Migrationshintergrund ausländische Namen tragen. Diese Einschränkung sind vertretbar, wenn es sich um wenige Fälle handelt und weder eine systematische Verzerrung, noch disproportionale Ausfälle bestimmter Merkmalsträger beobachtbar sind (Salentin & Wilkening, 2003). Alternativ zur Onomastik kann die ethnische Zugehörigkeit durch den Wohnort geschlussfolgert werden. Hierbei wird die messbare ethnische residentielle Segregation genutzt, um über den Wohnort die Wahrscheinlichkeit abzuleiten, dass ein spezifischer Abstammungshintergrund vorliegt. Allerdings werden ethnische Zuordnungen, die auf dem Wohnort beruhen, deutlich kritischer als die Ergebnisse der Onomastik gesehen (Strubelt et al., 2008, S. 6).

Als gängiges Verfahren der automatisierten Herkunftsermittlung durch die Kombination des Vor- und Zunamens hat sich das Verfahren von Humpert und Schneiderheinze (2000; 2002) etabliert (Haug et al., 2009; Flaig, 2008; Hanhörster, 2013; Salentin, 1999; Salentin & Wilkening, 2003; Schenk et al., 2007; Berger et al., 2004). Dieses Verfahren wurde ursprünglich entwickelt, um eine telefonische Umfrage gezielt an Migranten richten zu können. Dafür ist es notwendig, im Vorfeld der Befragung diejenigen Namen im Telefonbuch auszuwählen, bei denen von einem Migrationshintergrund auszugehen ist. Grundlage des Verfahrens stellt eine von Humpert und Schneiderheinze erstellte Namensdatenbank von 40.000 Vornamen und 500.000 Nachnamen dar, in der den Namen eine Sprache zugeordnet wird. Diese Namensdatenbank baut auf verschiedenen Namenslexika auf und wurde sukzessive durch ausgewertete Telefonverzeichnisse erweitert. Anschließend wurden zur gezielten Befragung von Personen mit Migrationshintergrund, Namen aus dem Telefonbuch mit dem erstellten Diktionär abgeglichen.

Im Folgenden wird ein Verfahren, dass Personen mit Migrationshintergrund ermittelt, entwickelt und getestet. Für die vorliegende Untersuchung genügt es grundsätzlich, wenn die Herkunft nach geographischen Regionen ausgewiesen wird. In ausgesuchten Fällen sollen jedoch spezielle Ethnien untersucht und miteinander verglichen werden.

Hierzu sind in einem ersten Schritt in Abschnitt 2.2.1 die Länder der Vornamen- und Nachnamensdatenbank (Michael, 2007; Michael, 2016) zu Regionen zu aggregieren. In einem zweiten Schritt ist in Abschnitt 2.2.2 aus der Häufigkeit, mit der ein Name in einem Land vorkommt die Häufigkeit zu ermitteln, mit dem ein Name aus den jeweiligen Regionen stammt, wenn er in Deutschland auftritt. Aus diesem erstellen Diktionär wird in Abschnitt 2.2.3 der Migrationshintergrund von Personen durch Analyse des Vor- und Nachnamens imputiert. In Abschnitt 2.2.4 wird das Verfahren durch Abgleich mit Namenslisten mit bekannten Migrationshintergrund evaluiert.

2.2.1. Herkunftsregionen in der Namensdatenbank

Angelehnt an die Methodik von Humpert und Schneiderheinze (2000; 2002) werden im Folgenden zwei Namensdatenbanken von 316.077 Vor- und 752.864 Zunamen (Michael, 2007; Michael, 2016) aufbereitet, um Personen mit Migrationshintergrund identifizieren zu können. In der Datenbank ist die prozentuale Verbreitung eines Namens innerhalb eines Landes für 145 Ländern weltweit angegeben.

Allerdings zeigt sich eine Unterscheidung in 145 Länder als zu feingranular. Nur für Namen, die eindeutig einem Land zugeordnet werden können, kann so ein Herkunftsland ermittelt werden. Die meisten Namen treten jedoch in mehr als einem Land auf.

Beispielsweise kommen die Namen David und Maria in 88 von 145 Ländern vor, Ali immerhin in 66 Ländern.

Daher werden die 145 Einzelländer zu Regionen aggregiert. Dabei wird die Gliederung der Vereinten Nationen zugrunde gelegt (United Nations Statistics Division, 2016) und um die häufigen Herkunftsorte der Migranten in Deutschland (Statistisches Bundesamt, 2015) und den Herkunftsstaaten ausländischen Studierenden in Deutschland (Statistisches Bundesamt; DZHW-Berechnungen, 2015) modifiziert. Wenn eine durch die Vereinten Nationen definierte Region in Deutschland nur selten als Migrationsherkunft auftritt, wird sie solange weiter zusammengefasst, bis die Mengenverhältnisse etwa ausgeglichen sind.

Beispielsweise stammen 631.000 Migranten in Deutschland aus Afrika, hiervon 306.000 aus Nordafrika (Statistisches Bundesamt, 2015). Aus diesem Grund genügt eine Unterscheidung zwischen Nord- und Rest-Afrika. Die Unterteilung der Vereinten Nationen in Nord-, West-, Ost-, Zentral- und Südafrika würde die Komplexität erhöhen und die Genauigkeit der Untersuchung senken.

Folgende 11 Regionen werden gebildet:

- Nordamerika
- Mittel- & Südamerika
- Nord- & Westeuropa
- Südeuropa
- Osteuropa
- Nordafrika
- Rest-Afrika
- Westasien
- Ost- & Südostasien
- Südasien
- Australasien

Zusätzlich zu den Regionen sind Länder gesondert zu betrachten, aus denen besonders viele Studierende stammen. Im Bundesdurchschnitt bilden türkische, italienische und russische Studierende die am häufigsten vertretenden Studierenden mit Migrationshintergrund (Heublein & Burkhart, 2013, S. 21). Geringe Abweichungen zur Studie von Heublein und Burkhart (2013) zeigen die betrachteten Hochschulen hinsichtlich der chinesischen und kroatischen Studierenden. Erste sind geringfügig stärker und zweitens geringfügig schwächer vertreten. Daher wird China als Herkunftsland zusätzlich gesondert betrachtet.

2.2.2. Zusammenstellung des Diktionärs

Innerhalb der Namensbank ist die Verteilung eines Namens innerhalb eines Landes angegeben (Michael, 2007; Michael, 2016). Bei geschlechtsspezifischen Namen ist das jeweilige Geschlecht angegeben. Geschlechtsneutrale Namen sind ohne Geschlecht angegeben. Namen, deren männliche und weibliche Form unterschiedlich häufig oder in unterschiedlichen Ländern auftreten, sind in der männlichen und in der weiblichen Form mit den jeweiligen Häufigkeiten innerhalb der Länder aufgeführt.

So tritt beispielsweise der weibliche Vorname „Ali“ ausschließlich in Nordeuropa auf, während das männliche Pendant in West-Asien, insbesondere der Türkei, stark verbreitet ist.

Aus der Verteilung eines Namens innerhalb eines Landes wird unter Berücksichtigung der Einwohnerzahlen der Länder (The World Bank Group, 2016) die absolute Anzahl der Namensträger in jedem Land gebildet und mit der weltweiten Anzahl verglichen. Das stellt die Herkunftswahrscheinlichkeit bei einer ortsungebundenen Betrachtung dar und ist zur Ermittlung eines Migrationshintergrundes in Deutschland noch nicht zielführend:

Beispielsweise wird die Ermittlung eines Migrationshintergrundes für den Namen „Johannes“ dadurch erschwert, dass durch die hohe Einwohnerzahl Afrikas mehr afrikanische Namensträger als europäische vorliegen. In einer weltweiten Betrachtung ist es damit korrekt, dass 52,47% aus Rest-Afrika stammen – nicht aber in einer Betrachtung innerhalb von Deutschland.

Zur Imputation eines Migrationshintergrundes von Studierenden an einer deutschen Hochschule wird die Wahrscheinlichkeit benötigt, dass ein Name aus einem speziellen Land stammt, wenn der Name in Deutschland aufgetreten ist. Hierzu ist das Verhältnis zu berücksichtigen, in dem Menschen aus den jeweiligen Regionen nach Deutschland migrieren.

Beispielsweise leben in der Türkei 78,67 Millionen Menschen und in Deutschland 2,85 Millionen Türken. Das entspricht einem Anteil von 3,62%. In den USA leben 321,42 Millionen Menschen, aber nur 137.000 Amerikaner in Deutschland, das entspricht 0,043%.

Das Verhältnis aus der Einwohnerzahl eines Landes und den in Deutschland lebenden Migranten dieses Landes wird im Folgenden als Migrationsverhältnis *MV* bezeichnet.

Unter der Annahme, dass der Name keinen Einfluss auf die Migrationsbereitschaft hat, jeder Namensträger also gleich wahrscheinlich migriert, sind in dem Migrationsverhältnis eines Landes auch ihre Namensträger nach Deutschland migriert. Das bedeutete, dass die Anzahl eines Namens eines Landes (AN_{Land}) multipliziert mit dem Migrationsverhältnis *MV*, die Anzahl der in Deutschland lebenden Migranten abbildet. Auch hierzu ein Beispiel:

Der Name Anna wird weltweit von 10.269.500 Menschen getragen. Davon werden 306.462 (2,98%) Menschen das Herkunftsland Deutschland und 401.770 (3,91%) Menschen das Herkunftsland USA zugewiesen. In einer weltweiten Betrachtung ist das Herkunftsland USA wahrscheinlicher. Da aber nur 0,043% der Amerikaner in Deutschland leben, kann davon ausgegangen werden, dass auch nur 0,043% der 401.770 (172) amerikanischen „Anna“ in Deutschland leben. Gleiches gilt für den weltweit geläufigsten Namen Ali. Ihn tragen 12.765.970 Menschen. Hiervon leben 1.179.990 (9,2%) in der Türkei und 1.889.250 (14,8%) in Pakistan. In einer ortsungebundenen Betrachtung ist es damit wahrscheinlicher, dass der Namensträger Ali aus Pakistan stammt. Da aber den 188.924.874 Einwohnern Pakistans nur 94.000 Migranten in Deutschland gegenüberstehen, ergibt sich

ein Migrationsverhältnis von 0,00049% für Pakistan (Türkei 3,62%). Es ist also davon auszugehen, dass in Deutschland 42.765 Namensträger aus der Türkei aber nur 940 aus Pakistan stammen.

Abschließend wird aus der absoluten Anzahl der Namensträger (AN_{Land}) multipliziert mit dem Migrationsverhältnis MV , die Anzahl der in Deutschland lebenden Menschen aus einer Region mit diesem Namen gebildet. Aus der der Anzahl der Namensträger aus einem speziellen Land und der Anzahl aller Namensträger, die sich in Deutschland aufhalten, kann anschließend die Wahrscheinlichkeit der Herkunft gebildet werden.

$$\text{Herkunftswahrscheinlichkeit Land } x = \frac{AN_{Land\ x} * MV_{Land\ x}}{\sum_{n=1}^{n=145} AN_{Land\ n} * MV_{Land\ n}}$$

Unter Berücksichtigung des Migrationsverhältnisses stammen $AN_{Dt} = 306.462$ Anna aus Deutschland und $AN_{nicht\ DT} = 120.031$ Anna stammen aus anderen Regionen. Das entspricht einer Wahrscheinlichkeit von etwa

$$\frac{AN_{DT}}{AN_{DT}+AN_{nichtDT}} = \frac{306.462}{306.462+120.031} = 71,86\%,$$

dass kein Migrationshintergrund vorliegt. In besonderer Betrachtung beispielsweise italienischer Anna sind von $AN_{IT} 1.216.041$ insgesamt $MV_{IT} = 1,27627\%$, demnach 15.520 italienische Anna nach Deutschland migriert. Hieraus ergibt sich eine Wahrscheinlichkeit von

$$\frac{AN_{IT}*MV_{IT}}{\sum_{n=1}^{n=145} AN_{Land\ n}*MV_{Land\ n}} = \frac{15.520}{426.493} = 3,66\%,$$

dass eine Anna in Deutschland aus Italien stammt. Namen hingegen, denen das Herkunftsland Deutschland nicht zugewiesen wird, haben eine Migrationswahrscheinlichkeit von 100%. In Deutschland lebende Menschen namens Ali, stammen zu 67,68% aus der Türkei, zu 26,44% aus den übrigen Gebieten Westasiens, zu 11,3% aus Nordafrika und zu 8,56% aus Südasien.

2.2.3. Namensanalyse

In der Migrationsanalyse werden Vor- und Zunamen jeweils mit den Vor- und Nachnamendiktionären abgeglichen und die Wahrscheinlichkeit eines Migrationshintergrundes ermittelt. Als Region oder Land wird die wahrscheinlichste Region beziehungsweise das wahrscheinlichste Land ausgewählt.

Die Trägerin des weiblichen Vornamens Simone hat mit einer Wahrscheinlichkeit von 99,08% keinen Migrationshintergrund. Ist der Träger hingegen ein Mann, so stammt

er mit einer Wahrscheinlichkeit von 97% aus Italien und zu 100% aus Südeuropa. Bei weniger eindeutig verbreiteten Namen, sinkt die Genauigkeit der Herkunftsbestimmung. Beispielsweise ergibt sich für den Namen Nadia zwar eine Migrationswahrscheinlichkeit von 100%, der Name stammt aber unter anderem zu einer Wahrscheinlichkeit von 13% aus Nord + Westeuropa, zu 23% aus Südeuropa, zu 38% aus Osteuropa oder zu 14% aus Nordafrika.

Da der Vornamendiktionär in keiner Verbindung zum Nachnamendiktionär steht, werden der Vorname und Nachname getrennt voneinander ausgewertet.

Die Migrationsanalyse der Namen Kerstin Schneider und Johannes Berens ergeben einen vernachlässigbaren Migrationshintergrund. Beide Namen stammen aus Nord+West-Europa, speziell Deutschland. Dem Namen Hung Lai, ist hingegen ein Migrationshintergrund zuzurechnen. Weniger eindeutig ist die Herkunft. Der Nachname Lai ist auch in Italien geläufig, so dass sich im Mittel zwar eine noch überwiegende Wahrscheinlichkeit für die Region Ost+Südost-Asien ergibt, als Einzelland aber Italien ausgewählt wird.

Tabelle 3: Namensanalyse ausgesuchter Beispielnamen

Vorname	Nachname	Ergebnis Vorname	Ergebnis Nachname	Gesamtergebnis	Alternatives Ergebnis Region	Alternatives Ergebnis Land	Nordamerika	Mittel- & Südamerika	Nord- & Westeuropa	Südeuropa	Osteuropa	Nordafrika	Rest-Afrika	Westasien	Ost- & Südostasien	Süd-asien	Australasien
Kerstin	Schneider	0,18%	0,37%	0,28%	Nord- & Westeuropa	Deutschl.	0,04%	0,02%	99,89%	0,01%	0,04%	0,00%	0,00%	0,01%	0,00%	0,00%	0,00%
Johannes	Berens	0,46%	0,27%	0,36%	Nord- & Westeuropa	Deutschl.	0,03%	0,00%	99,93%	0,00%	0,00%	0,00%	0,04%	0,00%	0,00%	0,00%	0,00%
Hung	Lai	100,00%	100,00%	100,00%	Ost- & Südostasien	Italien	0,61%	0,05%	2,18%	25,17%	0,00%	0,00%	0,14%	0,00%	71,59%	0,00%	0,13%

Wenn der zu untersuchenden Vor- oder Nachnamen ein Doppelname ist, ist die Wahrscheinlichkeit hoch, dass die jeweilige Namenskombination nicht im Diktionär enthalten ist, auch wenn es sich um die Kombination zweier geläufiger Namen handelt. Erschwerend kommt hinzu, dass jeder Doppelname durch einen Bindestrich oder ein Leerzeichen verbunden sein kann. Daher wird ein Doppelname, wenn er im Diktionär nicht enthalten ist, in seine Bestandteile aufgeteilt und jeder Einzelname überprüft. Die Analyseergebnisse der Einzelnamen liegen als prozentual ausgegebene Wahrscheinlichkeiten vor und können daher als Durchschnitt gemittelt werden.

Namen die im Diktionär nicht buchstabengenau enthalten sind, können nicht zugeordnet werden. Ist zumindest ein Namensbestandteil bekannt, so ist das Ergebnis stellvertretend zu übernehmen. Insbesondere das Zusammenspiel aus Vor- und Zunamen reduziert die Anzahl unbekannter Namen erheblich. In Untersuchung von 16.193 Studierenden, ergeben sich 129 unbekannte Fälle. In näherer Betrachtung dieser 129 unbe-

kannter Namen durch Muttersprachler (Salentin, 1999) ist festzuhalten, dass jedem unbekanntem Namen ein Migrationshintergrund zugesprochen werden kann. Beispiele¹ sind *Todik Erdogmus*, *Beatrisa Izgi*, *Pempe Bytyci* oder *Sintiya Qerkini*.

2.2.4. Evaluation des imputierten Migrationshintergrundes

Im Folgenden wird das Verfahren zur namensbasierten Imputation des Migrationshintergrundes an unterschiedlichen Populationen getestet, die Testergebnisse vorgestellt, auf Plausibilität geprüft und diskutiert.

Die vorliegenden Studierendendaten beider Hochschulen beinhalten die Namen von Studierenden mit ausländischer Staatsbürgerschaft. Da für diese der Migrationshintergrund bekannt ist, können diese Namen zur Evaluation der Migrationsanalyse genutzt werden.

Tabelle 4: Identifikationsquote der Herkunftsländer bei ausländischen Studierenden

Regionen	UNI		FH	
	n=	korrekt	n=	korrekt
Nordamerika	7	14,29%	2	0,00%
Mittel- & Südamerika	25	24,00%	13	30,77%
Nord & Westeuropa	82	69,51%	88	65,91%
Südeuropa	467	58,03%	338	61,29%
Osteuropa	356	73,03%	13	55,17%
Nordafrika	270	53,70%	91	50,00%
Rest-Afrika	106	29,25%	36	38,46%
Westasien	597	81,74%	815	86,59%
Ost- & Südostasien	346	96,24%	143	84,51%
Südostasien	95	57,89%	40	57,50%
Australasien	1	0,00%	0	0,00%
Länder				
Italien	135	67,81%	102	72,73%
Russische Föderation, auch Russland	93	35,00%	33	32,76%
Türkei	493	84,14%	761	95,23%
China	278	84,40%	123	81,63%

Vorteilhaft ist, dass diejenigen Regionen und Länder, aus denen größere Anteile ausländischer Studierenden stammen, in der Regel auch gut zuzuordnen sind. So sind an beiden Hochschulen Studierende aus Westasien, insbesondere aus der Türkei, sowie Ost- und Südostasien, insbesondere aus China, häufiger vertreten als andere Herkunftsregionen und mit Genauigkeiten jenseits von 80% auch relativ genau klassifizierbar. Studierende aus Südeuropa hingegen sind weniger gut klassifizierbar. Insgesamt ergibt sich eine hohe Identifikationsquote:

¹ Randomisiert veränderte Zusammensetzung

Tabelle 5: Identifikationsquote bei ausländischen Studierenden

	UNI	FH
Testdatensatz ausländische Staatsbürger	2.351	1.653
fälschlicherweise nicht identifiziert	141	88
nicht gefunden	77	53
identifizierter Migrationshintergrund	2.205	1.512
Identifikationsquote	93,8%	91,47%
Annahme: unbekannt = Migrationshintergrund	2.210	1.565
Identifikationsquote	94%	94,68%

Die Identifikationsquote kann dadurch verbessert werden, dass unbekanntes Namenskombinationen ein Migrationshintergrund zugeordnet wird.

In einem zweiten Test werden Namen von 1.458 Studierender mit deutscher Staatsbürgerschaft, aber ausländischen Geburtsort analysiert. Aufgrund des ausländischen Geburtsorts ist anzunehmen, dass es sich um Migranten der ersten Generation handelt. Etwas anderes gilt nur dann, wenn die Studierenden als Kind deutscher Eltern während eines Auslandsaufenthalts geboren wurden.

Da der Geburtsort in den vorliegenden Daten ohne Angabe eines Ortes nur als ausländisch angegeben ist, beschränkt sich der Test auf die reine Identifikation. Anders als im vorangegangenen Test ergibt der Test eine deutlich geringere Identifikationsquote von 54,4%. Das bedeutet, dass 665 von 1.458 Namen fälschlicherweise kein Migrationshintergrund zugeordnet wurde. In Betrachtung der fehlinterpretierten Namen, sind für die deutlich geringere Genauigkeit vier Ursachen zu vermuten:

- Ein in Deutschland sehr geläufiger Namensbestandteil trifft auf einen unbekanntes ausländischen Namensbestandteil. In diesen Fällen bildet das Migrationsergebnis des bekannten Teils, das Ergebnis der gesamten Analyse. Beispiele² sind *Sascha Trifunovic*, *Norbert Kantrzonki*, *Pelewe Sido* oder *Sara Shekoomand*. Es liegen 118 Namenskombinationen eines bekannten geläufigen Vornamens bei unbekanntes Nachnamen vor und 10 in umgekehrter Konstellation. Demnach ist das bei 128 von 665 fehlidentifizierten Namen der Fall
- Wenn beide Namensbestandteile bekannt sind, jedoch ein unterdurchschnittlich verbreiteter Namensbestandteil auf einen in Deutschland verbreiteten trifft,

² Randomisiert veränderte Zusammensetzung. Unbekannte Bestandteile in kursiv.

so ergibt sich im Mittel eine unterdurchschnittliche Migrationswahrscheinlichkeit. Zugunsten der sichereren Betrachtung wird dem Namen erst ab einer überwiegenden Wahrscheinlichkeit ein Migrationshintergrund zugesprochen. So zum Beispiel³ bei Namen wie *Svetlana Gottselig*, *Ekaterina Essert*, *Feliks Goltsmann*, Erika *Baran* oder Patrick *Murawski*. Das bei 86 der 665 fehlidentifizierten Namenskombinationen der Fall.

- Beide Namensbestandteile sind in Deutschland hinreichend verbreitet. Beispiele sind³: Stefanie Neubert, Timo Zill, Jürgen Maile, Bernhard Schüller, Elisabeth Krass. Das ist in 451 von 665 fehlidentifizierten Namen der Fall.

Die beschriebenen drei möglichen Erklärungen für die fehlidentifizierten Namen umfassen 665 der zu erklärenden, fehlidentifizierten 665 Namen. Wobei die verfahrensbedingte Verfälschung, die sich aus der Mittelung zweier konträr identifizierter Namen ergibt, mit 86 Fällen den geringsten Anteil hat.

Besondere Betrachtung bedarf die in diesem Test vorkommende Spezialpopulation von Studierenden, deren Vor- und Zuname entsprechend des Diktionärs in Deutschland hinreichend verbreitet sind. In einer Kontrolle durch deutsche Muttersprachler werden den Namenskombinationen kein Migrationshintergrund zugesprochen. Bei diesen 451 der 1.458 getesteten Namenskombinationen (31%) kann vermutet werden, dass es um Migranten der ersten Generation aus grenznahen deutschsprachigen Gebieten oder Spätaussiedler handelt. Letztere sind deutsche Volkszugehörige und stammen aus osteuropäischen Staaten (Bundesamt für Migration und Flüchtlinge, 2017). Da sie die deutsche Staatsbürgerschaft verfügen und im Ausland geboren sind, entsprechen Spätaussiedler der Testselektion.

Eine weitere Ursache des hohen Anteils von Namenskombinationen, die auch durch Muttersprachler nicht als ausländisch identifiziert werden, können „angeglichene“ Namen sein. Nach Art. 47 EGBGB haben Ausländer, die die deutsche Staatsbürgerschaft erhalten, die Möglichkeit ihren Namen anzugleichen. Werden hierbei ausländische Namen zu einer deutschen Entsprechung angeglichen, kann in einer namensbasierten Migrationsanalyse weder im Diktionärsabgleich noch durch Muttersprachler ein Migrationshintergrund ermittelt werden. In diesen Fällen ist ein Migrationshintergrund aus dem Namen

³ Randomisiert veränderte Zusammensetzung

nicht ermittelbar. Werden aber ausländische Namen nur dem Klang und der Schreibweise, aber ohne deutsche Entsprechung, angeglichen, so sprechen Muttersprachler noch immer einen Migrationshintergrund zu, Verfahren mit Diktionärabgleich jedoch nicht. Da diese „angeglichenen“ Schreibweisen nur in Deutschland auftreten, werden sie richtigerweise nur Deutschland zugeordnet, jedoch fälschlicherweise hieraus ein fehlender Migrationshintergrund geschlussfolgert.

Beispielsweise dürfen Namen wie Aleksandr oder Krystyna in Alexander oder Christina geändert werden. Für diese kann in einer namensbasierten Migrationsanalyse weder im Diktionärabgleich noch durch Muttersprachler ein Migrationshintergrund zugewiesen werden. Werden aber Namen wie Šmok zu Schmok oder Dmitri zu Dimitri, so findet ein Angleich ohne deutsche Entsprechung statt. Daher sprechen Muttersprachlern den Namen unerheblich der Schreibweise einen Migrationshintergrund zu, Analysen mit Diktionärabgleich nach Angleichung nicht.

In den zwei vorangegangenen Tests konnte nur Namen ausländischer Staatsbürger oder im Ausland geborener Studierender auf Identifikation getestet werden. Ob Namen ohne Migrationshintergrund richtig identifiziert werden, kann nicht getestet werden. Die ausgewiesene Genauigkeit kann damit auch Folge einer überhöhten Zuweisung sein. Daher werden im Folgendem Namen getestet, bei denen der Migrationshintergrund im Wege einer Befragung ermittelt wurde. Hierfür werden Befragungsergebnisse des sozioökonomischen Panels (SOEP) genutzt. Aus datenschutzrechtlichen Gründen beschränkt sich der Test auf eine Analyse der Vornamen. Als Grundlage der Verfahrensevaluation werden die Vornamen von 16.192 Studierenden der FH um Doppelnamen, Bindestrichnamen und Duplikate bereinigt. Die verbleibenden 2.608 Vornamen auf einen Migrationshintergrund geprüft und die Analyseergebnisse den Umfrageergebnissen des SOEP gegenübergestellt. Dabei wird überprüft, ob die Migrationsanalyse die im SOEP erfragten Migrationshintergründe korrekt identifiziert.

Von den 2.608 Vornamen sind 1.010 Namen im SOEP nicht enthalten. Die verbleibenden 1.598 Vornamen bestehen aus 838 Namen mit eindeutigem Migrationshintergrund und 760 Namen mit überregionaler Verbreitung, also ohne eindeutigen Hintergrund. Da in diesem Fall ein im SOEP mehrfach enthaltener Name sowohl mit als auch ohne Migrationshintergrund aufgeführt ist, erfolgt die Auswahl des Namens und des dazugehörigen Migrationshintergrundes randomisiert.

Tabelle 6: Ergebnis Namensanalyse SOEP-Test

SOEP \ Imputation	Ohne Migrationshintergrund	Mit Migrationshintergrund
Ohne Migrationshintergrund	509	129
mit Migrationshintergrund	151	768

Von den im SOEP enthaltenen zu überprüfenden 1.598 Einzelnamen, wurden in 638 Fällen ein bestehender und in 919 ein fehlender Migrationshintergrund erfragt. Im Rahmen der Analyse wurden von 919 fehlenden Migrationshintergründen 897 korrekt zugeordnet und von 638 bestehenden 660 zugesprochen. Das ergibt 1.277 (82%) korrekt identifizierte Namen.

In dem vorliegenden Test werden aus datenschutzrechtlichen Gründen nur die Vornamen betrachtet. Damit stehen die Vorteile, die sich aus der Kombination der Analyseergebnisse des Vor- und Zuname ergeben, nicht zur Verfügung. Das senkt die Genauigkeit. Werden vergleichsweise im ersten Test nur die Vornamen von ausländischen Staatsbürger betrachtet, sinkt die Genauigkeit von 94,68% (unter der Annahme, dass unbekannte Namen ausländisch sind) auf 88,08%.

2.2.5. Zwischenfazit

In einer Analyse aller ausländischen Staatsbürger, die zwischen 2007 und 2016 ihr Studium begannen, ergeben sich hohe Genauigkeiten. Gleichwohl zeigt die zweite Verfahrensevaluation mit Studierenden, die Migranten der ersten Generation darstellen, dass die Genauigkeit des ersten Tests nicht durch eine stark überhöhte Identifikationsneigung erreicht wurde. Anderenfalls wäre nicht der Hälfte der untersuchten Namen ein fehlender Migrationshintergrund zugesprochen worden. Vielmehr zeigt sich im zweiten Test eine Tendenz zur gehemmten Identifikation. Bei 451 Namenskombinationen, die auch durch einen Muttersprachler nicht als ausländisch identifiziert werden, weist die namensbasierte Analyse keinen Migrationshintergrund aus. Werden diese 451 Namen im zweiten Test nicht berücksichtigt, verbleibt eine Fehlidentifikation von 214 Namen bei 1.007 betrachteten Studierenden. Das relativiert die Genauigkeit bei Verwendung auf 78,75%. Damit ist

noch immer eine Tendenz zur Zuweisung eines fehlenden Migrationshintergrundes erkennbar. Zur sichereren Betrachtung von Studierenden mit Migrationshintergrund ist das wünschenswert.

Tabelle 7: Zusammenfassung Ermittlungsgenauigkeit des Migrationshintergrunds

Test	Genauigkeit
Ausländische Staatsbürger	94,68%
im Ausland Geborene	54% (78,75%)
Vergleich mit SOEP	82%

In einem letzten Test wurden die Ergebnisse der Migrationsanalyse von 1.598 Einzelnamen mit im sozioökonomischen Panel erfragten Migrationshintergründen verglichen. Anders als in vorangegangenen Tests wurde eine Population analysiert, die sich nicht nur aus Menschen mit Migrationshintergründen zusammensetzt, sondern auch aus solchen ohne einen Migrationshintergrund. Damit ist es möglich nicht nur die absolute Genauigkeit zu ermitteln, sondern auch zwischen einer Über- und Unteridentifikation zu differenzieren. Es ergibt sich eine Genauigkeit von 82%, wobei Personen mit Migrationshintergrund mit einer Genauigkeit von 85,62% und solche ohne mit 77,12% identifiziert wurden. Es liegt damit eine tendenzielle Unteridentifikation vor.

Tabelle 8. Ethnische Zusammensetzung der Studierenden

Regionen	UNI			FH		
	ausl. Staatsbürger	Deutsche mit Migrationshintergrund	Anteil	ausl. Staatsbürger	Deutsche mit Migrationshintergrund	Anteil
Nordamerika	8	46	0,20%	0	41	0,25%
Mittel- & Südamerika	27	133	0,60%	15	88	0,64%
Northern & Western Europa	103	1.102	4,52%	88	778	5,35%
Südeuropa	615	324	3,52%	341	255	3,68%
Osteuropa	433	419	3,19%	87	322	2,53%
Nordafrika	296	137	1,62%	90	113	1,25%
Rest-Afrika	136	116	0,94%	39	61	0,62%
Westasien	748	697	5,41%	812	1.008	11,24%
Ost- & Südostasien	392	323	2,68%	142	65	1,28%
Südasien	116	165	1,05%	40	201	1,49%
Australasien	3	2	0,02%	0	1	0,01%
Länder						
Italien	174	153	1,23%	102	103	1,27%
Russland	93	154	0,93%	33	143	1,09%
Türkei	620	641	4,73%	761	1.000	10,88%
China	278	274	2,07%	123	26	0,92%
Deutschland	23.757	0	71,07%	14.537	0	71,07%
Zusammenfassung						
Anzahl Studierende		26.686			16.192	
Nicht identifizierte Namen		234			147	
Stud. mit Mig.hintergrund		7.721			11.510	
Deutsche		18.574			4.684	
Anteil		28,93%			28,93%	

In Tabelle 8 werden die Ergebnisse beider Hochschulen zusammengefasst. Für beide Hochschulen ergibt sich ein Anteil Studierender mit Migrationshintergrund von 28,9%. Auch die prozentuale Verteilung der Herkunftsländer ist an beiden Hochschulen vergleichbar. Nennenswerte Unterschiede ergeben sich einzig im Anteil chinesischer und türkischstämmiger Studierender.

2.3. Beschreibung der Daten

Im Folgenden wird die Zusammensetzung der Immatrikulationskohorten 2007 bis 2012 beider Hochschulen dargestellt. An dieser Stelle erfolgt lediglich ein Überblick über die Hochschulen zum Zwecke der Konzeption des Frühwarnsystems. Eine vertiefte Betrachtung der Studierendendaten findet in Teil IV statt.

Dabei beziehen sich die deskriptiven Statistiken auf Studierende der Immatrikulationskohorten 2007 bis 2012, die bis zum Zeitpunkt des Datenexports 2017 exmatrikuliert waren – sie ihr Studium in dem Zeitraum also entweder abgeschlossen oder abgebrochen haben. Eine deskriptive Darstellung von Immatrikulationskohorten nach 2012 ist nicht

sinnvoll, da in diesen Studienabbrecher überpräsentiert sind. Obwohl die Regelstudienzeit eines Bachelorstudienganges an beiden Hochschulen 6 Semester beträgt, sind nach 6 Semestern noch erhebliche Anteil der Studierenden immatrikuliert (vgl. Tabelle 9). Da Studienabbrüche durchschnittlich deutlich früher stattfinden, als Studienabschlüsse, führen kurze Beobachtungszeiträume zu Kohorten, die sich vornehmlich aus Studienabbrechern zusammensetzen, während der Großteil künftiger Absolventen noch immatrikuliert ist und daher nicht Teil der „gelabelten“ Studierendendaten sein kann. Erst bei einer geringen Verbleibquote, sind auch Studierende Teil der Studierendendaten, die weder zu schnellen und eindeutigen Studienabbrechern oder Absolventen gezählt werden können.

In Tabelle 9 werden die Verbleibquoten der Immatrikulationskohorten von Bachelorstudiengängen beider Hochschulen dargestellt.

Tabelle 9. Verbleibquote der Immatrikulationskohorten 2009-2017 zum WiSe 2017/18

Jahr	UNI	FH
2009	8,23%	2,63%
2010	11,74%	2,42%
2011	16,26%	4,76%
2012	20,56%	6,94%
2013	30,57%	16,07%
2014	48,38%	35,72%
2015	58,79%	64,12%
2016	76,96%	83,39%
2017	97,72%	95,27%

Während an der UNI noch 11,74% der Immatrikulationskohorte 2010 im WiSe 2017/2018 immatrikuliert sind, sind es an der FH nur 2,42%. Das bedeutet, dass eine an beiden Hochschulen geltende Regelstudiendauer eines Bachelorstudienganges von sechs Semestern für die vorliegende Analyse kein geeigneter Maßstab ist. Eine Betrachtung exmatrikulierter Studierender nach Vollendung des sechsten Semesters, würde an der UNI etwa die Hälfte und an der FH etwa ein Drittel der Immatrikulationskohorte vernachlässigen – diese sind noch immatrikuliert. Die Unterschiede in den Verbleibquoten charakterisieren große – und für die Abbruchprognose relevante – Unterschiede zwischen der staatlichen Universität und der privaten Fachhochschule. Die längere Verweildauer der UNI ist, mindestens teilweise, den erhobenen Studiengebühren der FH in Höhe von etwa 400,- € monatlich geschuldet. An der FH besteht ein zusätzlicher monetärer Anreiz, das Studium frühzeitig zu beenden. An der UNI hingegen werden seit dem WiSe

2011/2012 keine Studiengebühren, sondern nur noch ein ASTA- und Mobilitätsbeitrag erhoben. Dieser ist auch von den Studierenden der privaten FH zusätzlich zu den Studiengebühren in vergleichbarer Höhe zu entrichten.

Tabelle 10 und Tabelle 11 zeigen noch weitere Unterschiede zwischen den Hochschulen: Die Zahl der an der UNI eingeschriebenen Studierenden ist in der Regel deutlich höher als an der FH. Darüber hinaus ist die Einschreibung an der FH auf nur einen Studiengang beschränkt. Daher ist eine Analyse nach Studiengängen möglich. Das ist bei der UNI anders. An der UNI waren von 20.707 eingeschriebenen Studenten der Kohorten 2007 bis 2017 insgesamt 11.193 Studenten in zwei oder mehr Studiengängen, 10.467 in drei oder mehr Studiengängen und 2.770 in vier oder mehr Studiengängen eingeschrieben. Insbesondere Studierende in Kombi-Bachelor Studiengänge sind in mindestens drei Studiengängen eingeschrieben, so auch in Lehramtsstudiengängen. Studiert ein Studierender beispielsweise Deutsch und Mathematik auf Lehramt, ist er in zwei unterschiedlichen Fachbereichen immatrikuliert – zusätzlich muss ein Optionalbereich gewählt werden. Weiterhin ist es Studierenden möglich, durch geschickte Mehrfachimmatrikulation Prüfungen in dem Studiengang abzulegen, der die einfacheren Rahmenbedingungen (Anzahl möglicher Prüfungswiederholungen, Anmeldefristen oder auch Pflichtanmeldungen) vorsieht und zu einem späteren Zeitpunkt die Anrechnung auf den Wunschstudiengang zu beantragen. Dadurch erfolgt eine Konsolidierung der Prüfungsleistungen erst zu einem späteren Zeitpunkt. Aber auch ein Studiengangswechsel bedeutet nicht, dass eine Exmatrikulation stattfindet. Da die Anzahl an Studiengängen nicht beschränkt ist und viele gleichzeitige Studiengänge keine Nachteile jedoch gelegentlich Vorteile bieten, kommt bei einem Studiengangswechsel lediglich ein neuer Studiengang hinzu. Eine differenzierte Betrachtung nach Studiengängen ist damit an der UNI nicht möglich.

Auch in ihren Zulassungsvoraussetzungen unterscheiden die Hochschulen sich. Im Allgemeinen sind die Voraussetzungen für ein Studium an einer FH weniger restriktiv als an einer Universität; dies gilt sowohl für die Note der Hochschulzugangsberechtigung (z.B. gibt es keinen Numerus clausus) als auch für die Art der Hochschulzugangsberechtigung. Infolgedessen ist die Zusammensetzung der Studierenden an den Hochschulen unterschiedlich.

Tabelle 10. Beschreibung der Daten: UNI

Mittelwert (Standardabweichung) der Immatrikulationskohorte	2007	2008	2009	2010	2011	2012
Geschlecht (1=männlich; 0=weiblich)	0,34	0,43	0,40	0,43	0,51	0,46
Alter bei Immatrikulation	21,24 (3,15)	21,84 (3,75)	21,86 (3,56)	21,93 (3,72)	22,28 (4,38)	22,60 (4,67)
Migrationshintergrund 1. Generation (1=ja; 0=nein)	0,08	0,11	0,10	0,09	0,07	0,08
Migrationshintergrund 2.-3. Generation (1=ja; 0=nein)	0,19	0,17	0,19	0,18	0,17	0,19
Ort der HZB (1= Ort der HS; 0=woanders)	0,14	0,19	0,18	0,18	0,21	0,22
allgemeine HZB (1=ja; 0=nein)	0,97	0,95	0,95	0,94	0,94	0,94
FHR (1=ja; 0=nein)	0,00	0,01	0,00	0,01	0,01	0,01
fgHR (1=ja; 0=nein)	0,01	0,01	0,01	0,02	0,01	0,01
Ausländische HZB (1=ja; 0=nein)	0,03	0,04	0,03	0,03	0,04	0,03
Note HZB	2,87 (0,82)	2,85 (1,00)	2,79 (0,97)	2,71 (0,87)	2,68 (0,92)	2,61 (0,89)
Krankversicherung (1=privat; 0=öffentlich)	0,06	0,08	0,06	0,05	0,07	0,06
Anzahl Studiengänge	3,06 (1,85)	2,62 (1,87)	2,67 (1,76)	2,71 (1,97)	2,32 (1,68)	2,20 (1,51)
Akademische Vorleistungen (1=ja; 0=nein)	0,17	0,25	0,30	0,39	0,38	0,44
Hochschulsemester bei Einschreibung	1,63 (4,45)	2,47 (5,32)	2,65 (5,01)	3,15 (5,09)	2,63 (4,50)	3,02 (5,13)
Durchschnittsnote pro Semester	2,46 (0,55)	2,49 (0,59)	2,45 (0,56)	2,50 (0,56)	2,51 (0,58)	2,49 (0,58)
Durchschnitts-CP pro Semester	12,91 (16,44)	17,18 (26,92)	18,33 (29,29)	19,80 (30,12)	15,38 (22,38)	15,22 (23,87)
Ohne Prüfungen	0,19	0,20	0,22	0,19	0,24	0,31
Anzahl angemeldeter aber nicht wahrgenommener Prüfungen	0,18 (0,62)	0,40 (1,56)	0,38 (1,44)	0,44 (1,28)	0,43 (1,16)	0,45 (1,35)
Anzahl nicht bestandener Prüfungen	0,44 (1,02)	0,63 (1,78)	0,62 (1,88)	0,75 (1,90)	0,59 (1,24)	0,64 (1,77)
Beobachtungen:	2.637	1.846	2.215	2.170	2.860	2.674

Hinweis: Die dargestellten Leistungsdaten stammen aus dem ersten Semester

An der UNI sind Frauen in den meisten Jahren überrepräsentiert, was wahrscheinlich durch eine große Geistes- und Kulturwissenschaftliche Fakultät an der UNI erklärt wird (vgl. Tabelle 10). Das Durchschnittsalter bei der Immatrikulation liegt zwischen 21

und 22,6 Jahren. Zwischen 24% und 29% der Studierenden haben einen Migrationshintergrund. Der Anteil der im Ausland geborenen Studierenden liegt zwischen 7% und 11%. Hinsichtlich der Migration zeigt sich zumindest kein eindeutiger Trend. Die überwiegende Mehrheit der Studierenden stammt aus einer anderen Stadt als der Stadt der Hochschule. Die Durchschnittsnote der Hochschulzugangsberechtigung liegt zwischen 2,6 und 2,9. Zwischen 5% und 8% der Studierenden sind privat krankenversichert und die durchschnittliche Anzahl der bestandenen Prüfungen pro Semester liegt zwischen 0,44 und 0,75.

Tabelle 11. Beschreibung der Daten: FH

Durchschnitt (Standardabweichung) der Immatrikulationskohorte	2007	2008	2009	2010	2011	2012
Geschlecht (1=männlich; 0=weiblich)	0,66	0,68	0,65	0,64	0,65	0,62
Alter bei Immatrikulation	22,27 (2,99)	23,95 (3,33)	24,40 (3,22)	24,36 (3,04)	24,01 (2,76)	23,97 (2,39)
Migrationshintergrund 1. Generation (1=ja; 0=nein)	0,13	0,15	0,13	0,10	0,08	0,09
Migrationshintergrund 2.-3. Generation (1=ja; 0=nein)	0,20	0,15	0,18	0,17	0,18	0,18
Ort der HZB (1= Ort der HS; 0=woanders)	0,26	0,34	0,34	0,30	0,30	0,32
allgemeine HZB (1=ja; 0=nein)	0,56	0,49	0,48	0,49	0,51	0,49
FHR (1=ja; 0=nein)	0,40	0,44	0,48	0,46	0,44	0,44
fgHR (1=ja; 0=nein)	0,01	0,00	0,01	0,02	0,02	0,05
Ausländische HZB (1=ja; 0=nein)	0,04	0,06	0,03	0,02	0,02	0,02
Akademische Vorleistungen (1=ja; 0=nein)	0,32	0,34	0,34	0,34	0,34	0,34
Durchschnittsnote pro Semester	2,37 (0,55)	2,32 (0,51)	2,32 (0,53)	2,28 (0,53)	2,24 (0,51)	2,28 (0,53)
Durchschnitts-CP pro Semester	12,78 (11,15)	16,25 (10,84)	18,77 (10,71)	19,11 (11,17)	19,98 (11,67)	19,69 (11,63)
Ohne Prüfungen	0,35 (0,00)	0,18 (0,00)	0,11 (0,00)	0,09 (0,00)	0,09 (0,00)	0,09 (0,00)
Anzahl angemeldeter aber nicht wahrgenommener Prüfungen	0,40 (0,57)	0,50 (0,80)	0,21 (0,42)	0,25 (0,46)	0,25 (0,43)	0,23 (0,49)
Anzahl nicht bestandener Prüfungen	0,17 (0,34)	0,30 (0,45)	0,62 (0,79)	0,56 (0,69)	0,50 (0,62)	0,54 (0,68)
Beobachtungen:	193	1.175	1.423	1.343	1.358	1.563

Hinweis: die dargestellten Leistungsdaten stammen aus dem ersten Semester

Im Vergleich dazu zeigt die deskriptive Statistik der FH in Tabelle 11 erhebliche Unterschiede. Männliche Studierende sind an der FH überrepräsentiert, das Einschreibungsalter ist höher und es studieren mehr ausländische Studierende. Weniger Studenten verfügen über die allgemeine Hochschulzugangsberechtigung (AHR), dafür deutlich mehr über die Fachhochschulreife (FHR). Es liegen keine automatisiert exportierbaren Informationen über die Note der Hochschulzugangsberechtigung und die Art der Krankenversicherung vor. Die durchschnittliche Anzahl der bestandenen Prüfungen pro Semester liegt zwischen 0,17 und 0,62 und ist damit niedriger als bei der UNI.

2.3.1. Grund des Abbruchs

Das Hochschulstatistikgesetz unterscheidet verschiedene Gründe der Exmatrikulation. Diese sind:

- Bestandene Abschlussprüfung
- Aufgabe oder Unterbrechung des Studiums
- Hochschulwechsel
- Endgültig nicht bestandene Prüfung
- Sonstiges

Die Ausprägung „Bestandene Abschlussprüfung“ sowie „endgültig nicht bestandene Prüfung“ sind Folge eines Prüfungsergebnisses, die ohne weiteres Zutun des Studierenden in einem Rechtsakt der Hochschule münden. Die verbleibenden Gründe sind die Folge einer Exmatrikulation durch den Studierenden. Auch die zwangsweise Exmatrikulation wegen fehlender Rückmeldung kann als Exmatrikulation durch den Studierenden durch Unterlassen bewertet werden. In aller Regel ist das ein Verwaltungsakt, der angedroht wird und durch nachträgliche Rückmeldung leicht heilbar ist. Nur wenn weder auf die Androhung noch die Exmatrikulation keine Reaktion des Studierenden erfolgt, bleibt der Studierende exmatrikuliert. Ob das aus dem Wunsch der Beendigung, finanzieller Not oder anderen Gründen geschieht, ist nicht ersichtlich. Macht der Studierende bei der Exmatrikulation Angaben zu seinen Gründen, werden diese eingetragen. Ob diese in Einklang mit dem späteren Lebensweg des Studierenden stehen, wird nicht überprüft. Aus diesem Grund können nur erfolgreiche Studienabschlüsse, leistungsbedingte Exmatrikulationen und sonstige Abbrüche unterschieden werden. Tabelle 12 fasst die im Studierendenmanagement-System hinterlegten Gründe des Abbruchs beider Hochschulen zusammen.

Tabelle 12. Gründe des Abbruchs

Exmatrikulationsgrund	UNI		FH	
fehlende Rückmeldung	6.245	40,15%	10	0,26%
Hochschulwechsel	3.508	22,55%	89	3,05%
Aufgabe/Unterbrechung des Studiums	2.050	13,18%	2.954	76,65%
sonstige Gründe	1.675	10,77%	537	13,93%
endgültig NICHT bestandene Prüfung	1.438	9,25%	218	5,66%
1.Studiengebühr nicht gezahlt! (Rückzieher)	475	3,05%	0	0,00%
Streichung durch die Hochschule	94	0,60%	1	0,03%
Wehr-/Zivildienst	62	0,40%	1	0,03%
Gesamt	10.062	100,00%	3.854	100,00%

Deutliche Unterschiede bestehen in der Häufigkeit des Grundes „fehlende Rückmeldung“ und „Aufgabe/Unterbrechung des Studiums“. Während an der UNI 40,15% der Studienabbrecher das Studium beenden, indem sie sich nicht zurückmelden, sind es an der FH 0,26%. Dafür geben 76,65% der FH-Studierenden ihr Studium auf, oder unterbrechen es⁴. Dieser Unterschied ergibt sich auch aus einem zusätzlichen privatrechtlichen Verhältnis zwischen der privaten, gebührenfinanzierten Fachhochschule und ihren Studierenden. Der geschlossene Studienvertrag muss bis sechs Wochen vor Semesterende gekündigt werden, anderenfalls verlängert er sich um ein Semester.

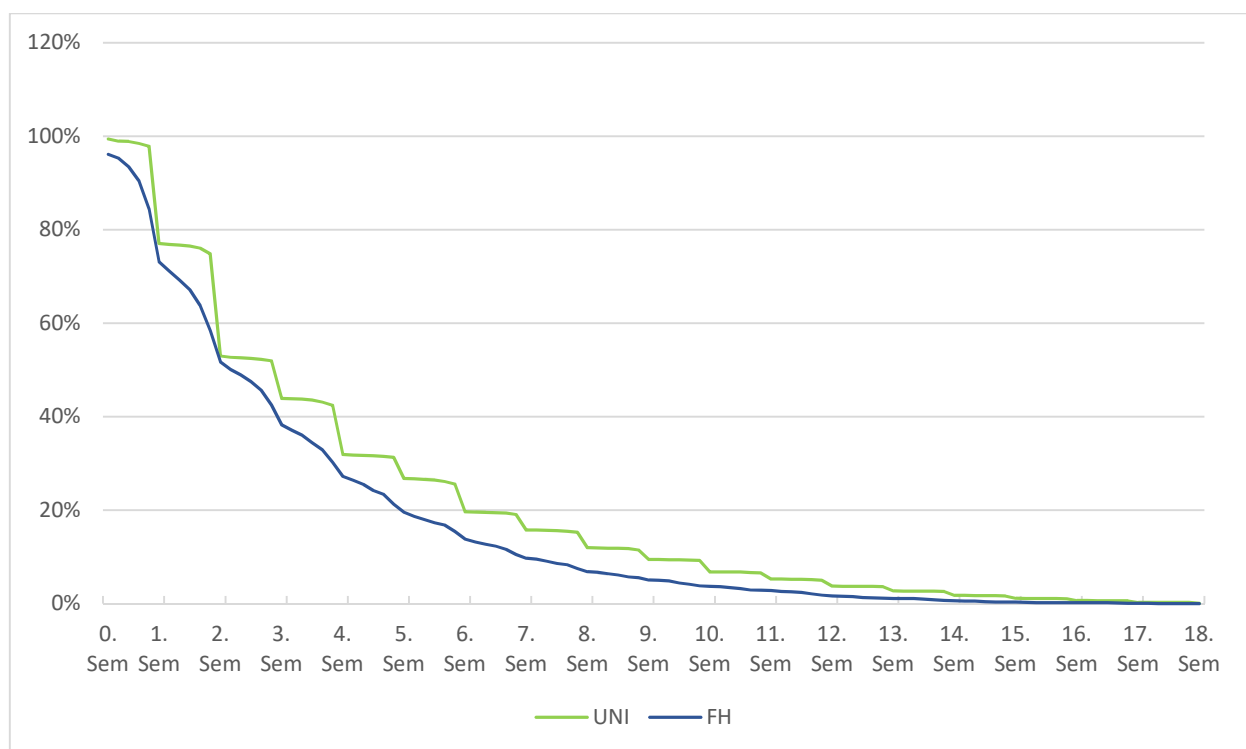
Die Angabe „endgültig nicht bestandene Prüfung“ ist an beiden Hochschulen jedoch gleich zu interpretieren. Hierbei handelt es sich um Studierende, die nach der geltenden Bachelor-Prüfungsordnung die maximale Anzahl an Wiederholungsprüfungen nicht bestanden haben. Es zeigt sich, dass an beiden Hochschulen nur ein kleiner Teil der Studierenden, die das Studium abbrechen, ihr Studium durch leistungsbedingte Zwangsexmatrikulation verlassen (UNI: 9,25% und FH 5,66%). Das bedeutet, dass eine letzte Wiederholungsprüfung (in der Regel der dritte Versuch) von den Studierenden entweder bestanden wird, oder die Hochschule bereits vorher verlassen wurde. Hierbei ist zu beachten, dass an der UNI ein dritter Prüfungsversuch frühestens im dritten Semester angetreten werden kann. Unter Berücksichtigung üblicher Korrekturzeiten und Verarbeitungszeiten in der Verwaltung ist mit einer Konsequenz erst im vierten Semester zu rechnen. Studienabbrecher, die das Studium vor dem vierten Semester verlassen, werden

⁴ An dieser Stelle soll nicht diskutiert werden, dass die „Aufgabe/Unterbrechung des Studiums“ nicht ein Abbruchgrund, sondern die Folge eines Abbruchgrundes ist. Jeder Abbruchgrund führt zu diesem Status. Die auswählbaren Gründe sind offensichtlich nicht trennscharf und bei jedem Studienabbruch können mindestens zwei Felder gleichermaßen angegeben werden.

also in der Regel nicht aufgrund einer „endgültig nicht bestandenen Prüfung“ exmatrikuliert.

2.3.2. Zeitpunkt des Abbruchs

Für ein Frühwarnsystem ist der Zeitpunkt an den Abbruchprognosen vorgenommen werden sollen, von essentieller Bedeutung. Soll mittels der Früherkennung Abbruchgefährdeten Hilfe angeboten werden, sollte das so früh wie möglich geschehen. Einerseits sind späte Interventionen weniger wirkungsvoll als frühe, zielgerichtete Angebote; andererseits werden Studienabbrecher, die die Hochschule bereits verlassen haben nicht mehr erreicht (vgl. Teil II, Abschnitt 1.2). Abbildung 1 zeigt die verbleibenden Studienabbrecher im Semesterverlauf.



Hinweis: Monatsgenaue Ermittlung verbleibender Studienabbrecher der Immatrikulationskohorten 2007 bis 2012
Abbildung 1. Studienabbrüche im Semesterverlauf

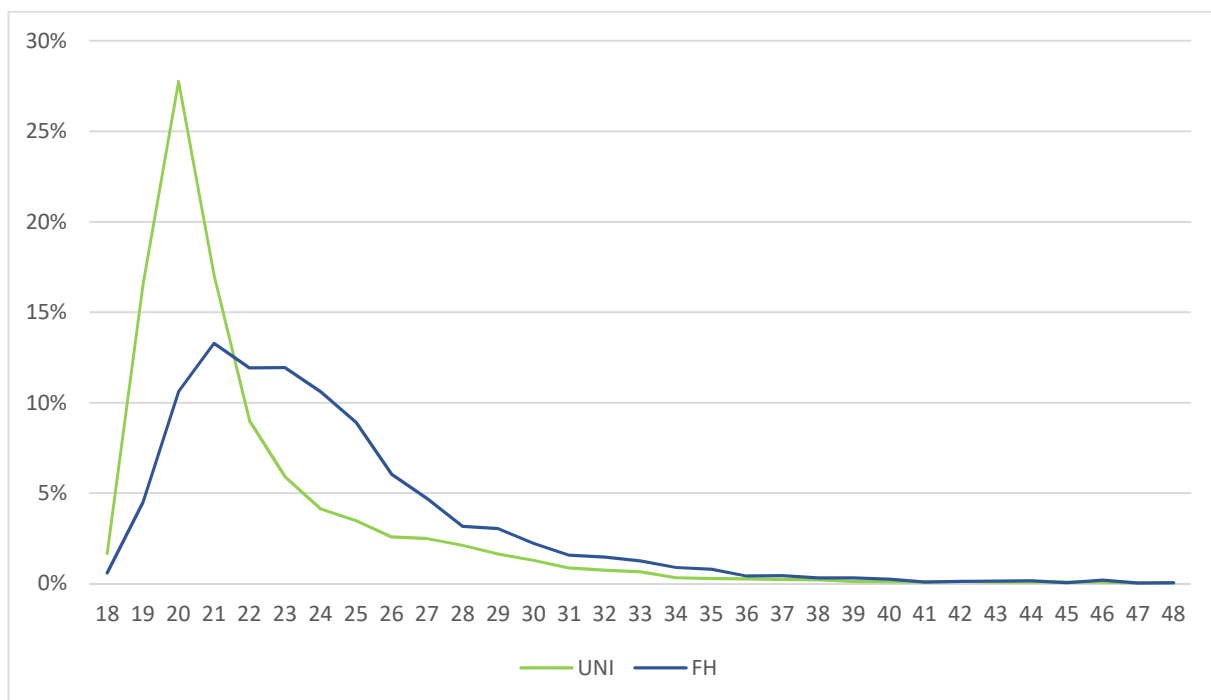
Erwartungsgemäß finden Abbrüche an der gebührenpflichtigen FH früher als an der UNI statt, allerdings ist der zeitliche Abbruchverlauf beider Hochschulen vergleichbar. An beiden Hochschulen verlassen die ersten Studierenden die Hochschule bereits vor dem ersten Vorlesungstag. Innerhalb des ersten Semesters verlassen bereits 23% (UNI) beziehungsweise 27% (FH) die Hochschule, im zweiten Semester folgen weitere 24% und 21% (UNI und FH). Etwa die Hälfte aller Studienabbrecher erreicht das dritte Semester damit nicht. Nach drei Semestern sind 44% (UNI) beziehungsweise 38% (FH) und nach

vier Semestern 31% und 27% aller Studienabbrecher noch immatrikuliert. Ein Frühwarnsystem, das nach dem dritten oder vierten Semester abbruchgefährdete Studierende identifiziert, erreicht einen Großteil der Abbrecher nicht mehr – sie haben die Hochschule bereits verlassen.

Im Folgenden wird der Zusammenhang zwischen Studienabbruch und ausgewählten demografischen Variablen beschrieben.

2.3.3. Demografie der Studierenden

Das Alter bei Immatrikulation ist ein Indikator für die Lebensphase des Studierenden. Mit zunehmendem Alter nimmt die Wahrscheinlichkeit zu, dass der Studierende bereits über vorangegangene Ausbildungen verfügt, Berufstätig ist und möglicherweise bereits familiäre Verpflichtungen eingegangen ist. Während an der UNI die Studierenden bei Einschreibung etwa 21 Jahre alt waren und die Varianz eher gering ist, ist die Altersverteilung an der FH breiter gefächert. Ein Grund hierfür ist der größere Anteil berufsbegleitender Studierender. Abbildung 2 zeigt die Altersverteilung der Studierenden an beiden Hochschulen.

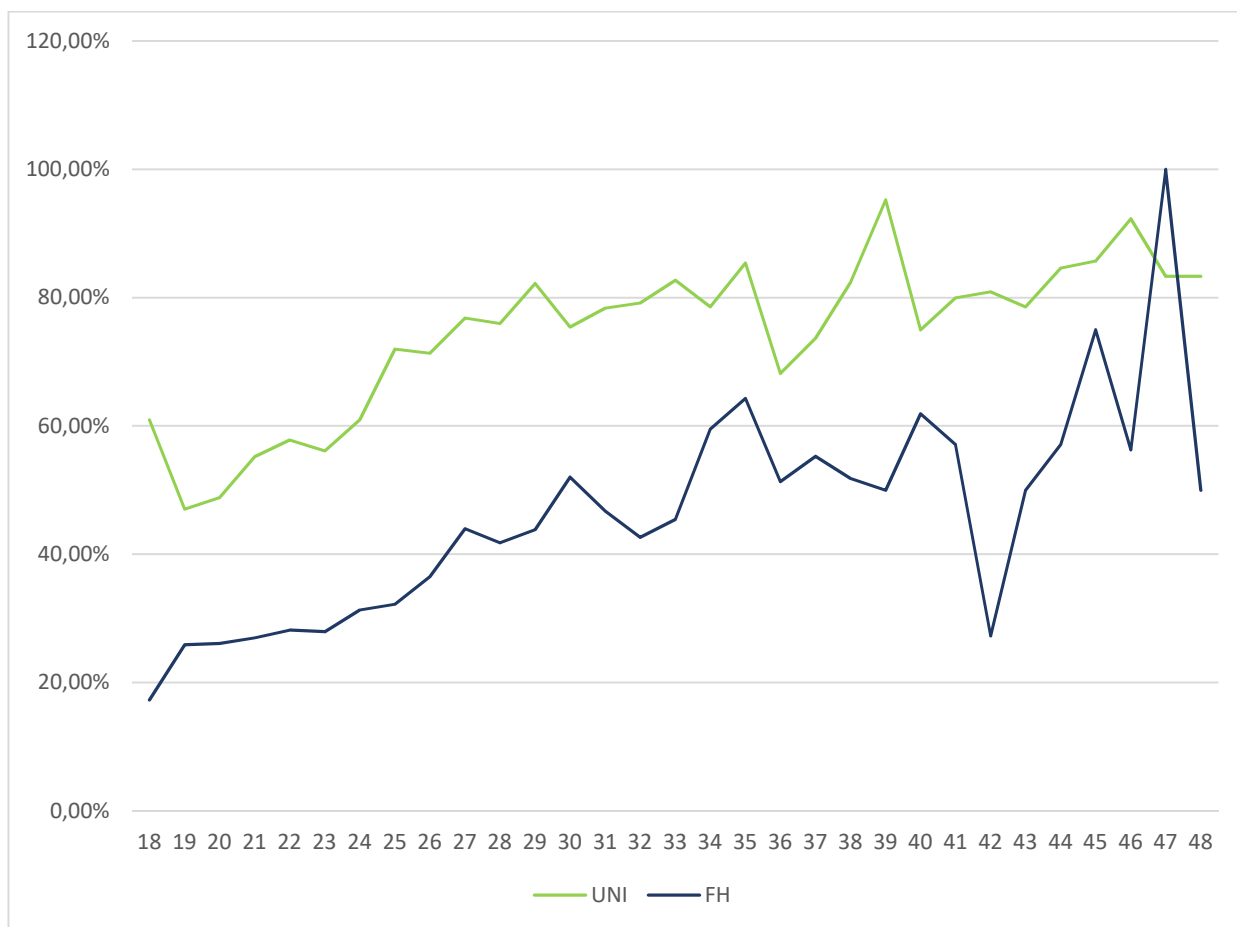


Hinweis: Altersverteilung Immatrikulationskohorten 2007 bis 2012

Abbildung 2. Altersverteilung der Studierenden bei Immatrikulation

Abbildung 3 zeigt die Abbruchquoten Studierender, die bei Immatrikulation zwischen 18 und 48 Jahre alt waren. An beiden Hochschulen nimmt die Rate der Studienabbrüche mit zunehmendem Alter zu. Zu berücksichtigen ist jedoch, dass nur wenige Studierende

bei Einschreibung jünger als 19 oder älter als 30 Jahre waren – die Fallzahlen sind entsprechend klein.



Hinweis: Abbruchquote der Immatrikulationskohorten 2007 bis 2012 nach Alter bei Immatrikulation
 Abbildung 3. Abbruchquoten nach Alter bei Immatrikulation

In gesonderter Betrachtung des Migrationshintergrundes lassen sich aus den verfügbaren Daten verschiedene Unterscheidungen treffen. Ausländische Studierende können entsprechend einer deutschen HZB in Bildungsinländer und Bildungsausländer unterschieden werden. Weiterhin kann der Geburtsort betrachtet werden. Tabelle 13 zeigt die Abbruchquoten nach Migrationshintergrund.

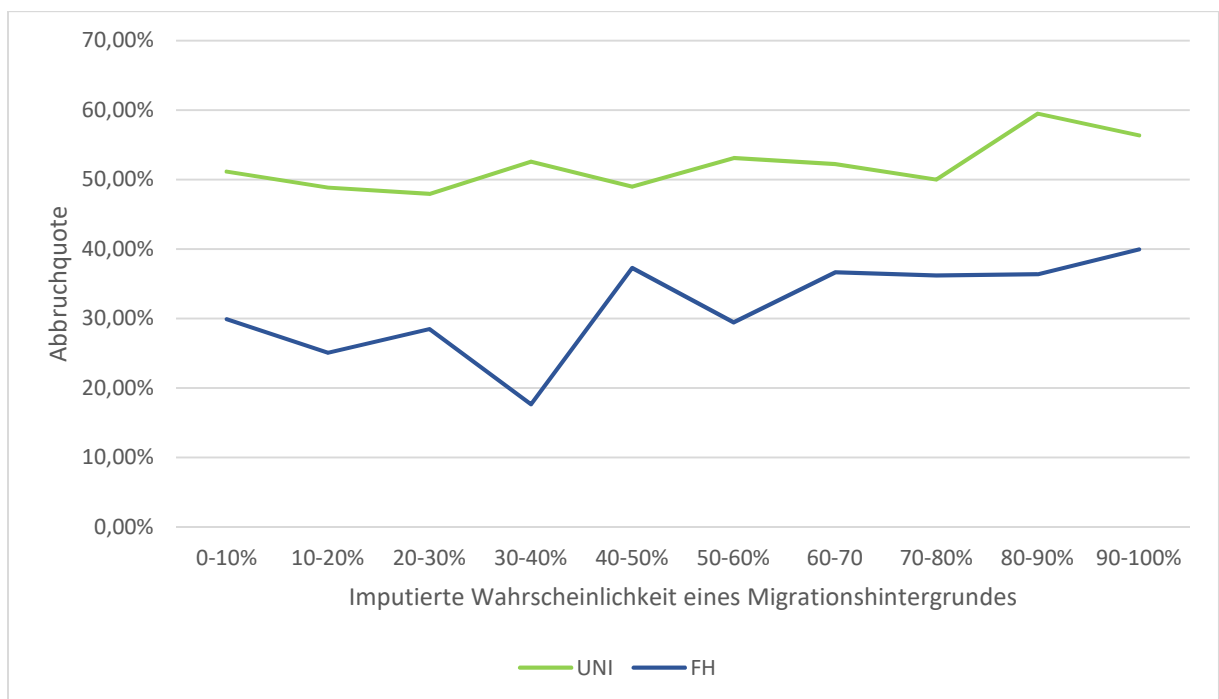
Tabelle 13. Abbruchquoten nach Migrationshintergrund

Migrationshintergrund	UNI	FH
Studierende ohne Migrationshintergrund ¹	51,61%	30,13%
Ausländischer Geburtsort	56,91%	45,02%
Bildungsinländer	59,74%	44,16%
Bildungsausländer	65,91%	74,76%

Hinweis:¹In Deutschland geborene deutsche Staatsbürger mit deutscher HZB, bei denen die imputierte Migrationswahrscheinlichkeit unter 10% (keine Veränderung bei 0% oder 5%) liegt.

In Deutschland geborene deutsche Staatsbürger, die in Deutschland die HZB erworben und keine imputierte Migrationswahrscheinlichkeit zugerechnet wurde, weisen an der UNI eine Abbruchquote von etwa 51,6% und an der FH von etwa 30% auf. Studierende, die im Ausland geboren wurden, brachen an der UNI in etwa 57% (2.063) und an der FH in 45% (1.164) der Fälle ab. In vergleichbarer Häufigkeit (UNI: 59,74% | FH: 44,16%) brachen auch Bildungsinländer, also Studierende mit ausländischer Staatsbürgerschaft, die in Deutschland die HZB erworben, ab (UNI: 862 | FH: 625). Bildungsausländer, also Studierende mit ausländischer Staatsbürgerschaft, die die HZB im Ausland erworben, brachen deutlich häufiger ab, an der UNI in etwa 66% (296) und an der FH in etwa 75% (210).

In Deutschland geborene deutsche Staatsbürger, die in Deutschland die HZB erworben, können anhand des imputierten Migrationshintergrundes – in dem Fall der zweiten und dritten Generation – betrachtet werden. In Abbildung 4 wird die Wahrscheinlichkeit des Abbruchs in Abhängigkeit eines Migrationshintergrundes angegeben.



Hinweis: Abbruchquoten der Immatrikulationskohorten 2007 bis 2012
 Abbildung 4. Abbruchquote und Migrationshintergrund

Abbildung 4 zeigt an beiden Hochschulen einen moderaten Anstieg der Abbruchquote von etwa 10 Prozentpunkten. Die Namensanalyse hat nur wenigen Studierenden eine Wahrscheinlichkeit zwischen 40% und 90% zugerechnet, daher sind diese Bereiche

sehr vorsichtig zu interpretieren. Von insgesamt 13.396 (FH: 6.315) in Deutschland geborenen, deutschen Studierenden mit deutscher HZB ergab sich eine imputierte Migrationswahrscheinlichkeit der zweiten bzw. dritten Generation von:

- unter 10% bei 9.103 (FH: 4.757) Studierenden und
- über 90% bei 1.322 (FH: 528) Studierenden.

Wohingegen in den Bereichen zwischen:

- zwischen 10% und 49% insgesamt 1.245 (FH: 539) Studierende und
- zwischen 50% und 90% insgesamt 1.726 (FH: 491) Studierenden eingeordnet wurden.

2.3.4. Hochschulzugangsberechtigung

Tabelle 10 und Tabelle 11 zeigen deutliche Unterschiede zwischen den Hochschulen. Fachhochschulen haben weit weniger restriktive Aufnahmebedingungen als Universitäten. Während der weit überwiegende Teil der Studierenden der UNI über die allgemeine Hochschulreife verfügen, sind es an der FH weniger als die Hälfte. In Tabelle 14 und Tabelle 15 werden der Zusammenhang zwischen Abbruchquoten und verschiedener Hochschulzugangsberechtigung beider Hochschulen dargestellt.

Tabelle 14. Hochschulzugangsberechtigung UNI

Hochschulzugangsberechtigung	Anteil	Abbruchquote
Gymnasium (AHR)	61,82%	47,29%
Gesamtschule (AHR)	15,84%	65,91%
allg.Hochschulr.Kolleg NRW entspr.Eintr.and.Ld.	6,92%	61,20%
allg.Hochschulreife im Ausland erworben (AHR)	2,21%	75,24%
fachgebundene Hochschulreife i.Ausld.erworben	2,44%	66,02%
Fachhochschulreife Kollegschule NRW entspr.Ld.	1,81%	77,91%
Fachhochschulr.Fachoberschule(incl.Abendform)	1,26%	78,66%
FHReife d.Berufsfachschule	0,99%	80,95%
allg.Hochschulreife Studienkolleg	0,73%	78,26%
Fachhochschulreife Gymnasium	0,69%	75,00%

Hinweis: Es werden die häufigsten 10 von 44 Arten der HZB exmatrikulierter Studierende (19.027) der Immatrikulationskohorten 2007 bis 2012 dargestellt.

Über drei Viertel (77,66%) der Studierenden der UNI haben die allgemeine Hochschulreife an einem Gymnasium oder einer Gesamtschule erworben. Der verbleibende Anteil verteilt sich auf 42 weitere Arten der HZB. In Betrachtung der Abbruchquoten zeigt sich der hohe Stellenwert einer gymnasialen AHR. Studierenden, die die AHR an einem

Gymnasium erworben haben, haben ihr Studium um 18,6 Prozentpunkte häufiger erfolgreich beendet, als Kommilitonen, die die AHR an einer Gesamtschule erworben. Das ist insofern bemerkenswert, da formell zwischen der AHR beider Schulen kein Unterschied besteht, die Abbruchquoten jedoch auf eine fehlende Gleichartigkeit hindeuten. Eine noch höhere Abbruchquote ergibt sich für die verschiedenen Formen der FHR oder für ausländische Abschlüsse.

Tabelle 15. Hochschulzugangsberechtigung FH

Hochschulzugangsberechtigung	Anteil	Abbruchquote
Gymnasium (AHR)	39,40%	21,55%
Abendgymnasium, Kolleg (FHR)	30,91%	41,38%
Gesamtschule (AHR)	5,92%	27,86%
Gymnasium (FHR)	5,54%	42,57%
Abendgymnasium/Kolleg (AHR)	4,50%	23,06%
Berufsfachschule (FHR)	2,96%	39,69%
Allgemeine Hochschulreife, Ausland	2,64%	70,51%
Fachoberschule (FHR)	1,92%	32,94%
Beruflich Qualifizierte (fgHR)	1,22%	50,00%
Beruflich Qualifizierte (AHR)	1,15%	53,92%

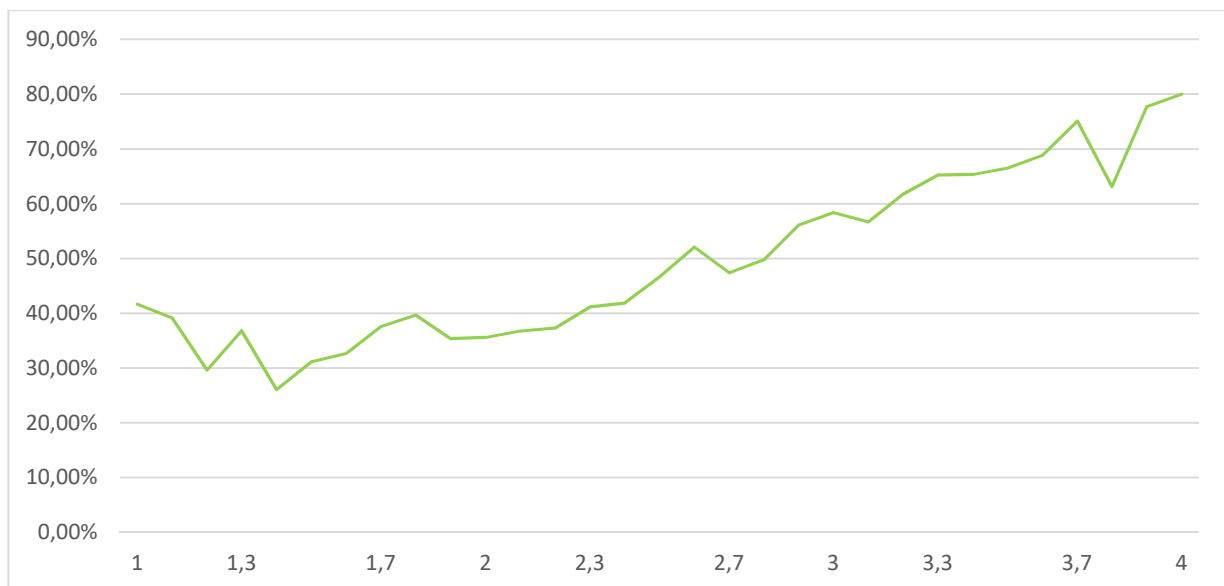
Hinweis: Es werden die häufigsten 10 von 33 HZB exmatrikulierter Studierende (8.858) der Immatrikulationskohorten 2007 bis 2012

Zwei Drittel der Studierenden der FH haben die AHR am Gymnasium oder die Fachhochschulreife an einem Abendgymnasium beziehungsweise Kolleg erworben. Ein weiteres Viertel besuchte die Gesamtschule (AHR), das Gymnasium bis zur Jahrgangsstufe 12 (FHR) oder ein Abendgymnasium beziehungsweise Kolleg (AHR). Die verbleibenden 15 % der Studierenden verteilen sich auf 28 weitere Arten der HZB. Die niedrigste Abbruchquote (21,55%) weist die gymnasiale AHR auf, die AHR des Abendgymnasiums bzw. Kollegs (23,06%) und der AHR Gesamtschule (27,86%) ist geringfügig höher. Deutlich höhere Abbruchzahlen weisen Studierende mit der FHR als HZB auf. Allerdings bestehen auch zwischen den Arten der FHR Unterschiede. Studierende, die ein Abendgymnasium bzw. Kolleg oder das Gymnasium mit der FHR verließen, brechen etwa 10% häufiger ab, als Studierende, die die FHR auf der Fachoberschule erworben. Besonders hohe Abbruchquoten weisen Studierende mit im Ausland erworbener HZB aus.

Im Vergleich beider Hochschulen zeigt sich an der FH eine heterogenere Verteilung der HZB. An beiden Hochschulen brechen Studierende mit gymnasialer HZB am seltensten ab. Jedoch ist an der UNI entscheidender an welcher Schulform die AHR erworben

wurde, während die Unterschiede zwischen den deutschen Studierenden mit AHR an der FH geringer ist. Deutliche Unterschiede hingegen ergeben sich an der FH zwischen Studierenden mit AHR und FHR, wobei auch Studierende mit unterschiedlichen FHR untereinander abweichende Abbruchquoten aufweisen. Bei beiden Hochschulen brechen Bildungsausländer am häufigsten das Studium ab.

Neben der Art und dem Ort der HZB, wird der Note der HZB als Prädiktor der Studienergebnisse nach herrschender Meinung eine besondere Stellung eingeräumt (Trapmann et al., 2007; Brandstätter & Farthofer, 2002; Schmidt-Atzert, 2005; Gold et al., 2005). Abbildung 5 stellt die Abbruchquote nach HZB-Note der UNI dar; an der FH wird die HZB-Note nicht erhoben. Dabei zeigt sich, dass die niedrigste Abbruchquote Studierende mit einer HZB Note von 1,2 bis 1,4 aufweisen. Studierende mit einer geringeren Durchschnittsnote haben im Durchschnitt eine höhere Abbruchquote. Ab einer Note von 3,4 beenden etwa zwei Drittel der Studierenden ihr Studium nicht erfolgreich.



Hinweis: Abbruchquoten der Immatrikulationskohorten 2007 bis 2012 der UNI nach HZB Noten

Abbildung 5. Abbruchquote nach HZB-Note der UNI

2.3.5. Akademische Leistungen

Die Gründe von Studienabbrüche werden als multikausal beschrieben. Einer der angeführten Gründe sind Leistungsprobleme, aber auch andere Gründe gehen häufig mit Leistungsproblemen einher (Heublein et al., 2017). Akademische Leistungen können quantitativ in CP pro Semester und qualitativ in Noten gemessen werden. Weitere Indikatoren akademischer Leistung sind die Anzahl nicht bestandener Prüfungen, die Anzahl

der Rücktritte von Prüfungsleistungen und die Anzahl angemeldeter, aber nicht teilgenommener Prüfungen.



Hinweis: Absolventen und Studienabbrecher der Immatrikulationskohorten 2007 bis 2012.

Abbildung 6. Abbruchquote nach Durchschnittsnote

Für beide Hochschulen ergibt sich eine sehr ähnliche Verteilung der Abbruchquote, die auch der Verteilung der Abbruchquote der HZB-Noten in Abbildung 5 ähnelt. Studierende, die bis zur Exmatrikulation eine Durchschnittsnote von 1,0 erzielt hatten, haben an beiden Hochschulen weit überwiegend das Studium vorzeitig beendet. Hierzu ist anzumerken, dass die Fallzahlen sehr klein sind. Zum Beispiel weisen an der UNI 3 Absolventen und 21 Abbrecher eine Durchschnittsnote von 1,0 auf. Während Studienabbrecher diese Durchschnittsnote bereits mit einer Prüfung – auch einer angerechneten – erzielen können, fließen bei Absolventen notwendigerweise alle Prüfungen eines Studiums in die Durchschnittsnote ein. An der FH hatten 2 Absolventen eine Durchschnittsnote von 1,0 und 23 Studienabbrecher. Das Gleiche gilt für das untere Ende des Notenspektrums. Eine Durchschnittsnote schlechter als 3,5 ist an beiden Hochschulen sehr selten. An der UNI absolvierten 5 und an der FH 3 Studierende mit einer schlechteren Durchschnittsnote als 3,59. Dem standen 477 Studienabbrecher an der UNI und 87 Studienabbrecher an der FH im selben Notenbereich gegenüber.

Aussagekräftiger ist der Notenbereich zwischen 1,5 und 3,5. An beiden Hochschulen liegen hier Absolventen (UNI: 6.684 | FH: 5.131) und Studienabbrecher (UNI: 3.831 | FH: 1.734) in großer Zahl vor. Abbildung 6 zeigt, dass mit zunehmend schlechterer Durchschnittsnote die Abbruchquote zunimmt.

Zu beachten ist, dass alle Absolventen der betrachteten Kohorten mit einer Durchschnittsnote vertreten sind, nicht aber jeder Studienabbrecher. An der UNI brachen 3.363 und an der FH 963 Studierende ihr Studium ohne eine erfolgreiche Prüfung ab. Eine reine Betrachtung der Noten vernachlässigt daher einen großen Anteil der Studienabbrecher. Bei Betrachtung der durch die Studierenden erlangten CP werden inaktive Studierende hingegen berücksichtigt. Tabelle 16 weist den Anteil und die Abbruchquote von Studierenden ohne CP aus.

Tabelle 16. Studierende ohne CP

Studiendauer in Semester:		1	2	3	4
UNI	Abbruchquote ohne CP	83,45%	89,73%	92,90%	93,82%
	Ohne CP / alle Studierende	36,77%	30,86%	27,57%	26,63%
	Ohne CP / alle Abbrecher	62,30%	57,46%	54,53%	53,71%
FH	Abbruchquote ohne CP	84,42%	96,15%	98,27%	98,47%
	Ohne CP / alle Studierende	15,50%	11,52%	11,02%	10,91%
	Ohne CP / alle Abbrecher	38,67%	33,00%	32,39%	32,11%

Hinweis: Immatrikulationskohorten 2007 bis 2012

Studierende, die im ersten Semester keinen CP erwarben, brachen in 83,45% (UNI) beziehungsweise 84,42% (FH) ihr Studium ab. Wenn auch im zweiten Semester keine CP erworben wurden, steigt die Abbruchquote auf 89,73% (UNI) beziehungsweise 96,15% (FH). Das bedeutet, dass in den sieben Immatrikulationskohorten beider Hochschulen nur 10,27% (UNI) beziehungsweise 3,85% (FH) derer, die in den ersten beiden Semestern ohne CP blieben, ihr Studium noch erfolgreich beendeten.

An beiden Hochschulen nimmt der Anteil Studierender ohne CP innerhalb der Studierendenschaft mit zunehmender Studiendauer ab. Der Anteil Studierender ohne CP an allen Studienabbrechern hingegen ist jeweils relativ konstant und liegt ab dem zweiten Semester an der Uni bei etwa der Hälfte und an der FH bei etwa einem Drittel.

2.4. Feature Engineering

Das zu konzipierende Frühwarnsystem FragSte beschränkt sich auf Studierenden-daten die nach §3 HStatG erhoben werden. Damit ist sichergestellt, dass das Frühwarnsystem an allen deutschen Hochschulen ohne Modifikation einsetzbar ist, den einsetzenden Hochschulen kein zusätzlicher Erhebungs- und Verwaltungsaufwand entsteht, das

Frühwarnsystem nicht von freiwilligen Teilnahmen von Studierenden abhängig ist und dass historische Studierendendaten exmatrikulierter Studierender zum Anlernen und Justieren des Frühwarnsystems vorhanden sind. Hierzu sind die vorhandenen Informationen in maschinenlesbare Variablen umzuwandeln. Dieser Vorgang wird als Featureentwicklung – englisch *feature engineering* – bezeichnet. Variablen werden im Kontext maschinellen Lernens auch als *Feature* bezeichnet.

Die Featureentwicklung ist für die Prognosegenauigkeit von entscheidender Bedeutung. Während ein Teil der nach §3 HStatG zu erhebenden Informationen unverändert verwendet werden können, sind andere anzupassen um ihren Informationsgehalt maschinell verwerten zu können. Um bei kohortenübergreifender Betrachtung beispielsweise die Variable „Alter bei Immatrikulation“ zu erhalten, sind das Geburtsjahr und das Immatrikulationsdatum zu verknüpfen. Weiterhin kann die Verknüpfung von Immatrikulationsdatum und Semesteranfang Aufschluss darüber geben, wie rechtzeitig sich Studierende um den Studienplatz an der Betrachtungshochschule bemüht haben. Hierzu ist die Anzahl von Tagen zwischen Immatrikulation und Semesteranfang zu bilden.

Auch die Information, ob der Studierende an seinem Heimatort studiert oder zugezogen ist, kann relevant sein (Dekker et al., 2009). An deutschen Hochschulen wird in der Regel zwischen einer Semesteranschrift und einer Heimatanschrift unterschieden. In diesem Fall kann ermittelt werden, ob der Studierende zum Zwecke des Studiums umgezogen ist, ob er über größere Strecken pendelt oder ob er am Heimatort studiert. Die Betrachtung der Krankenversicherung der Studierenden kann Aufschluss über den sozioökonomischen Hintergrund geben (Danilowicz-Gösele et al., 2017). Sofern Studierende privat krankenversichert sind, ist entweder der Studierende selbst oder die Eltern selbstständig, verbeamtet oder mit einem Einkommen jenseits der Beitragsbemessungsgrenze privat versichert.

Allerdings sind beim Feature Engineering auch die Besonderheiten einer Hochschule zu berücksichtigen. Während beispielsweise an der UNI etwa ein Drittel der Studierenden in mehr als einem Studiengang eingeschrieben sind, ist an der FH jedem Studierender nur ein einziger Studiengang zugewiesen (siehe Teil II, Abschnitt 2.3). Aus diesem Grund wird die Art des Studiengangs als Variable nur an der FH und nicht an der UNI verwendet. Jedoch kann an beiden Hochschulen die durchschnittliche Abbruchquote der gewählten Studiengänge ermittelt und berücksichtigt werden. Da bei mehrfach einge-

schriebenen Studierenden der primäre Studiengang nicht ermittelbar ist, muss die durchschnittliche Abbruchquote der gewählten Studiengänge hilfsweise ungewichtet gemittelt werden.

Tabelle 17 beschreibt die aus den Studierendendaten im Sinne des §3 HStatG erstellten Variablen. Eine differenzierte Betrachtung des Informationsgehalts einzelner Variablen erfolgt in Teil IV.

Tabelle 17. Variablenbeschreibung

Variable	Ausprägung
Demografie	
Geschlecht	Männlich oder weiblich
ImmtrAlter	Alter bei Einschreibung
KV	Gesetzliche oder private Krankenversicherung
StBAusland	Ausländische Staatsbürgerschaft
Geburtsort	Geburtsort in Deutschland oder im Ausland
Migration	Wahrscheinlichkeit eines Migrationshintergrundes aufgrund des Vor- und Nachnamens
Land	Türkei, Russland, China, Italien, Deutschland oder ein anderes Land
Region	Nordamerika, Mittel- & Südamerika, Nord- & Westeuropa, Südeuropa, Osteuropa, Nordafrika, Rest-Afrika, Westasien, Ost- & Südostasien, Südasien, Australasien
Schulbildung:	
HZB	Art der HZB: AHR Gymnasium, AHR Gesamtschule, FHR, fgHR, Ausland, Hochschul-Abschluss
HZBNote	UNI: Note der HZB
Kreis der HZB	Entspricht der Ort der HZB dem Ort der Hochschule, handelt es sich um einen heimatnah Studierenden
Frühere Hochschulbildung:	
akad. Vorleistungen	War bereits an einer Hochschule immatrikuliert
VorSem	Anzahl Vorsemester
Bei Einschreibung	
StartSem	Start im Winter- oder Sommersemester
AnmeldeZeit	Differenz in Tagen zwischen Anmeldung und Studienstart (nur UNI)
StGAbbruchquote	Durchschnittliche Schwundquote der gewählten Studiengänge (nur UNI)
Studienform	Vollzeit oder berufsbegleitend
Studiengang	Anzahl der Studiengänge (UNI); gewählter Studiengang (FH)
Akad. Leistungen.	
Schein1-9	Bestehen der am stärksten mit Studienerfolg korrelierten Prüfungen 1-9 des Studiengangs im Betrachtungssemester
Leistungen 1. Sem.	
CPSem1	Erworbene ECTS Punkte im ersten Semester
NoteSem1	Durchschnittsnote im ersten Semester
NTSem1	Anzahl der Rücktritte (UNI) Nicht teilgenommen (FH) von Prüfungen im ersten Semester
AnzahlScheinSem1	Anzahl bestandener Prüfungen im ersten Semester
NBSem1	Anzahl nicht bestandener Prüfungen im ersten Semester
Leistungen 2. Sem.	
CPSem2	Erworbene ECTS Punkte im zweiten Semester
NoteSem2	Durchschnittsnote im zweiten Semester
NTSem2	Anzahl der Rücktritte (UNI) Nicht teilgenommen (FH) von Prüfungen im zweiten Semester
AnzahlScheinSem2	Anzahl bestandener Prüfungen im zweiten Semester
NBSem2	Anzahl nicht bestandener Prüfungen im zweiten Semester
Leistungen 3.-x Sem.	
Vgl. mit 1 & 2 Sem	Vgl. mit 1 & 2 Sem
Label	Studienabbruch oder Studienerfolg

2.5. Verfahrensauswahl und Methodenbeschreibung

Die Prognose von Bildungsergebnissen ist ein Schwerpunkt der sich rasch entwickelnden Disziplin des „educational data mining“ (Baker & Yacef, 2009). Vorliegend soll der Studienausgang von Studierenden prognostiziert werden. Dabei hängt Wahl des Prognoseverfahrens wesentlich von der Verfügbarkeit administrativer Studierendendaten ab. Liegen Daten von ehemaligen Studierenden vor (*labeled data*), können durch die Analyse historischer Studierendendaten die Merkmale und Merkmalskombinationen potentieller Absolventen und Studienabbrecher identifiziert werden. Hierfür bieten sich Verfahren des überwachten Lernens (*supervised learning*) an, einer Teildisziplin des maschinellen Lernens (Witten et al., 2016). Da die zu erklärende Variable nur die Werte 0 = Studienerfolg und 1 = Studienabbruch annehmen kann, liegt ein binäres Klassifikationsproblem vor. Zur binären Klassifikation können traditionelle ökonomische Verfahren wie die logistische Regression, aber auch maschinelle Verfahren des überwachten Lernens wie neuronale Netze oder Entscheidungsbäume eingesetzt werden (Baker & Yacef, 2009).

Mit dem Ziel Zusammenhänge zwischen und Einflüsse der einzelnen Variablen auf den Studienabbruch zu identifizieren, werden im ersten Schritt Regelmäßigkeiten (Koeffizienten, Gewichte, Regeln) anhand eines Lerndatensatzes abgeleitet, um im zweiten Schritt bisher unbekannte Beispiele (Studierende) hinsichtlich des Studienerfolgs oder –abbruchs zu klassifizieren, d.h. den drohenden Studienabbruch möglichst genau zu prognostizieren. Als Lerndatensatz werden dabei historische Daten von Absolventen und Studienabbrechern genutzt. Somit können nur Informationen verwendet werden, die bereits über einen längeren Zeitraum erhoben wurden. Später aufgenommene Informationen mit einem potentiell hohen Erklärungsgehalt – wie beispielsweise Self-Assessments – können erst für die Prognose verwendet werden, wenn die Studierenden mit durchlaufendem Self-Assessment mindestens die Regelstudiendauer erreicht haben.

Im Folgenden werden die Regressionsanalyse, neuronale Netze und Entscheidungsbäume beschrieben. Hierauf aufbauend wird eine gewichtete Verfahrensauswahl durch den AdaBoost beschrieben. Abschließend wird die Ermittlung der Genauigkeit binärer Klassifikatoren vorgestellt.

2.5.1. Regressionsanalyse

Einer der klassischen Ansätze einen statistischen Zusammenhang zwischen einer abhängigen Variable und mehreren unabhängigen Variablen zu messen, ist das Regressionsverfahren. In der Ökonometrie wird besonders häufig die logistische Regression,

kurz Logit-Modell, zur Modellierung von Wahrscheinlichkeiten binärer Auswahlprobleme verwendet. Übertragen auf den vorliegenden Untersuchungsgegenstand hat das vorliegend verwendete Logit-Modell folgende Form:

$$P(y_{it} = 1|x_i, z_{it}) = \Lambda(\beta_0 + \beta_1 x_i + \beta_2 z_{it}),$$

wobei i und t den Studierenden und das Semester repräsentieren. Λ stellt die logistische Funktion dar. Die unabhängige Variable y_{it} ist binär: 0 = Studienerfolg und 1 = Studienabbruch. Die demografische Information x_i ist zeitinvariant, wohingegen Leistungsdaten z_{it} im Semesterverlauf variieren.

2.5.2. Neuronale Netze

Das Forschungsgebiet der „künstlichen Intelligenz“ (KI) ist von der Hirnforschung inspiriert und unternimmt seit Beginn der 1950er Jahre den Versuch, Struktur und Funktionsweise des menschlichen Gehirns zu modellieren. Nach anfänglich großen Erfolgen des Einsatzes erster Neurocomputer zur Lösung von Mustererkennungsproblemen, zeigten Minsky und Papert (1969), dass die damals verwendeten Trainingsalgorithmen nur linear separable Funktionen berechnen konnten. In Weiterentwicklung des von Werbos (1974) erstmals in neuronalen Netzen eingesetzten Algorithmus zur Fehlerrückübertragung, entwickelten Rummelhart und Siple (1974) das Verfahren der „Backpropagation“ weiter und belebten das bis zu diesem Zeitpunkt stagnierende Forschungsgebiet der künstlichen Intelligenz neu. Heutzutage werden insbesondere höherdimensionale neuronale Netze (Multilayer Perceptrons, MLP) entwickelt, die neben der Erkennung von Mustern, Bildverarbeitung und Spracherkennung auch zur Optimierung von Prozessen, Kontrolle von Systemen sowie Diagnose und Vorhersage unterschiedlichster Problemstellungen und Bereichen eingesetzt werden.

Künstliche neuronale Netze bestehen aus einem Netzwerk von Neuronen. Diese sind in verschiedenen Schichten (Layern) angeordnet. In einer ersten Schicht, dem Input-Layer, wird jede Variable durch eine Neurone berücksichtigt. Von den Neuronen der Inputlayer fließen Informationen über gewichtete Verbindungen, so genannten Kanten, hin zu Neuronen einer oder mehrerer Zwischenschichten (hidden Layer), bis hin zu einer Schicht mit Ergebnisneuronen. Dabei wird der Output einer Neurone über die Kante zu einem gewichteten Input einer anderen Neurone. Dabei sind positive und negative Gewichtungen und eine Gewichtung von 0 möglich. Ein positives Gewicht hat eine verstär-

kenden, ein negatives Gewicht eine hemmenden und ein Gewicht von 0 gar keinen Einfluss auf die Folge Neurone. Eine gute Übersicht über die Konzeption und den Aufbau neuronaler Netze bietet Dreyfus (2005).

In dem in FragSte konzipierten neuronalen Netz entspricht die Zahl der Inputneuronen der Anzahl der Variablen der jeweiligen Hochschulen und des jeweiligen Semesters. Die zweischichtige hidden layer verwendet 16 in der ersten und 8 Neuronen in der zweiten Schicht. Die Ergebnisschicht besteht aus einer binären Ausgabeneurone (0 = Studienerfolg und 1 = Studienabbruch).

In einem ersten Schritt wird das neuronale Netz mittels gelabelter Studierenden trainiert (Mucherino et al., 2009). Dabei werden die Inputneuronen mit dem Lerndatensatz initialisiert. Dieser besteht aus den Variablen und dem Label \tilde{y}_{it} (Abbrecher oder Absolvent). Alle anderen Neuronen der hidden layer, werden randomisiert mit einem Wert zwischen -1 und 1 versehen. Während der Trainingsphase werden die über die Kanten zwischen den Neuronen fließenden Informationen gewichtet. Wie die Gewichtung vorgenommen wird, hängt von der eingesetzten Lernregel ab. Vorliegend wird der als überlegen angesehene "Backpropagation" Algorithmus genutzt (Rumelhart et al., 1986). Durch einen dreistufigen Lernaufbau ist dieser in der Lage, die Modifikation der Gewichte auch innerhalb der Hidden-Layer vornehmen zu können. In einem ersten Schritt werden aus den Inputneuronen das Ergebnis der Outputneurone errechnet. In einem zweiten Schritt wird das Label und der errechnete Output verglichen und der Fehler e_t bestimmt. Die Fehlerfunktion besteht aus der Summe der quadrierten Fehler:

$$e_t = \sum_i (\tilde{y}_{it} - y_{it})^2$$

Überschreitet der Fehler einen definierten Schwellenwert, werden die Fehlerterme rückwärts bis zurück zur Inputlayer ausgebreitet. Mit Hilfe dieser Fehlerterme werden über viele Iterationen hinweg, die Gewichte so lange verändert, bis der Fehler die definierte Schwelle unterschreitet (Rumelhart & Siple, 1974; Werbos, 1974). Mittels dieser erlernten Gewichte ist ein neuronales Netz in der Lage, den Studienerfolg \tilde{y}_{it} innerhalb des ungelabelten Testdatensatzes zu prognostizieren.

2.5.3. Entscheidungsbäume

Alle Verfahren der Entscheidungsbäume gehen nach einem rekursiven Top-down-Prinzip vor und bestehen aus einer Aneinanderreihung von „Wenn-dann“ Abfragen der

verfügbaren Variablen. Diese Abfrage findet an sogenannten Knoten statt. Der erste Knoten eines Baumes wird Wurzelknoten genannt. An jedem Knoten wird der vorliegende Variablenwert mit einem zuvor ermittelten Schwellenwert verglichen (beispielsweise ob „ECTS pro Semester“ größer als 15 CP ist). In Abhängigkeit des Ergebnisses folgt ein weiterer Knoten, ein sogenannter „Sohnknoten“, an dem wiederum eine Variable mit einem Schwellenwert verglichen wird (beispielsweise ob im ersten Semester mehr als zwei Prüfungen bestanden wurden). Knoten, auf die keine Sohnknoten mehr folgen, stellen die Blätter des Entscheidungsbaumes dar und bilden damit das Ergebnis. Im vorliegenden Anwendungsfall findet eine binäre Klassifikation in Studienabbrecher und Absolventen statt. Mit welcher Variable an einem Knoten separiert wird, und auf welche Weise der Schwellenwert bestimmt wird, hängt von der Art des eingesetzten Entscheidungsbaum-Algorithmus ab. Häufig verwendete Algorithmen sind der CART (Classification and Regression Trees) und C4.5 Algorithmus. Der CART Algorithmus gehört zur Klasse der binären Entscheidungsbaumalgorithmen und verwendet zur Wahl der optimalen Verzweigung die Berechnung der minimalen Fehlerquadratrate (Breimann, 1996; 2001). Die Attributauswahl wird zur Maximierung des Informationsgehalts durch Entropie bestimmt. Der C4.5 Algorithmus ist eine leistungsfähigere Variante des ID3 Algorithmus (Quinlan, 1986). Er behebt die Einschränkung des ID3 auf vollständige und fehlerfreie Daten sowie diskreter Merkmalsausprägungen und wiederholt den Vorgang des Baumerzeugens und –beschneidens (pruning) mehrfach. Eine Übersicht der am häufigsten verwendeten Algorithmen findet sich in Schapire und Freund (2012) oder in Sammut und Webb (2017).

Für die vorliegende Studie wird der C4.5 Algorithmus genutzt (Hall et al., 2009). Dieser nutzt den Information Gain der Variablen zur Auswahl und Reihenfolge abzufragender Variablen zur rekursiven Konzeption des Entscheidungsbaumes. Der Information Gain gibt Aufschluss darüber, um wieviel die Entropie der abhängigen Variable durch Hinzufügen der Betrachtungsvariable abnimmt.

Entscheidungsbäume stellen flexible, nichtparametrische maschinelle Verfahren dar, die große Stichproben benötigen und zum Overfitting tendieren. Hierbei erfolgt eine Überanpassung an die Lerndaten, Muster werden „auswendig gelernt“. Um die Varianz zu verringern und die Prognosegenauigkeit zu verbessern, wird der Meta-Learning-Algorithmus Bagging angewendet (Bootstrap-Aggregation). Dabei werden mehrere randomisierte Versionen des Baums durch Bootstrapping der Lerndaten und anschließender Mittelung erstellt. Um zu verhindern, dass möglicherweise viele ähnliche Bäume entstehen,

wird ein Teil der Variablen randomisiert Knoten zugeordnet. Hieraus entstehen unterschiedliche, nicht korrelierte Bäume. Das Ergebnis stellt einen randomisierten Wald von Entscheidungsbäumen, einen Random Forrest, dar (Breimann, 1996; 2001). In einem zusätzlichen Schritt wurde der Lernphase des Random Forrest eine zusätzliche Anwendung des Bagging-Verfahrens auf die Lerndaten vorgeschaltet. Das angewendete Verfahren wird daher hier als Bagging Random Forrest (BRF) bezeichnet.⁵

2.5.4. AdaBoost

Das zu konzipierende Frühwarnsystem soll möglichst effizient und auf andere Hochschulen übertragbar sein. Da bereits in der Literatur unterschiedliche Verfahren die jeweils besten Ergebnisse lieferten, ist bei Implementierung eines Frühwarnsystems eine aufwendige Verfahrensevaluation und -auswahl vorzunehmen. Um das zu vermeiden, werden die Prognosen der logistischen Regression, des neuronalen Netzes und des BRF unter Verwendung eines Boosting-Algorithmus zu einem robusten, gewichteten Ergebnis zusammengefügt. Boosting-Algorithmen werten den Einfluss der einzelnen Methoden (weak classifier) aus und führen die Ergebnisse zu einem einzigen (strong) Klassifikator zusammen. In dem vorliegenden Beitrag wird der von Freund und Schapire (1997; 1999) entwickelte adaptive Boosting (AdaBoost) Algorithmus eingesetzt.

Der AdaBoost ist ein Algorithmus für das Ensemble-Lernen. Hierbei wird ein Kollektiv von Prädiktoren gebildet um ein gewichtetes Kollektivmittelwert zu bilden. Das erlaubt es, dass Ausreißer einzelner Verfahren bei Prognose des Studienerfolgs einzelner Studierender durch das Kollektiv ausgeglichen werden. In einem ersten Schritt werden die errechneten Prognosen der logistischen Regression, des neuronalen Netzes und des BRF zunächst zu einem gleich gewichteten Mittel zusammengefügt. Bei jeder Iteration des Algorithmus werden die einzelnen Gewichte entsprechend der Verteilung so angepasst, dass der resultierende Klassifikator den kleinstmöglichen Fehlerwert hat. Die Prognose des AdaBoost ergibt sich dann aus der Summe der gewichteten Einzelprognosen der logistischen Regression, des neuronalen Netzes und des BRF. Im Ergebnis weist die Prognose eine höhere Genauigkeit auf, als die Einzelverfahren. Weiterhin erübrigt sich durch die Verwendung des AdaBoost die Bestimmung des jeweiligen hochschulspezifischen und unter Umständen semesterabhängig genauesten Verfahren – das Ergebnis

⁵ Von allen getesteten Entscheidungsbäumen (u.a. C4.5, M5p, CART, Decision Stump, RepTree) mit allen getesteten Meta-Learning-Algorithmen (u.a. Bagging, random subspace, random committee, classification via regression, random forest) bieten der BRF und C4.5 die robustesten Ergebnisse. Ergebnisse sind auf Anfrage erhältlich.

des AdaBoost wird entweder gegen das genaueste Verfahren konvergieren oder dieses übertreffen.

2.5.5. Wahl des Identifikationsschwellenwerts

Jeder Klassifikator schätzt für jeden Studenten eine Abbruchwahrscheinlichkeit mit einer Zahl zwischen 0 (Absolvent) und 1 (Abbrecher). Daher wird für eine automatisierte Früherkennung ein Schwellenwert benötigt, ab dem Studierende als gefährdet definiert werden – also als Studienabbrecher klassifiziert werden. Die Wahl des Schwellenwerts ist für die Verwendbarkeit eines Frühwarnsystems von entscheidender Bedeutung. Je sensibler der Schwellenwert, desto mehr Studierende werden als Studienabbrecher identifiziert. Damit erhöht sich der Anteil der künftigen Studienabbrecher, die zu Recht identifiziert wurden. Gleichzeitig erhöht sich aber auch die Anzahl künftiger Absolventen, die zu Unrecht als Studienabbrecher identifiziert wurden. Es ist demnach abzuwägen, ob besonders viele der künftigen Studienabbrecher gefunden werden sollen oder die Identifikationen besonders glaubwürdig sein sollen (Swets, 1988; Gleason & Dynarski, 2002; Bowers et al., 2013). Knowles (2015, S. 23) fasst zusammen:

“Where do the indicators draw the line between false-alarm and true classification of students and is the resulting student group the group that schools should serve?”

Eine mögliche Lösung ist es, den Schwellenwert anhand der durchschnittlichen Abbruchquote früherer Kohorten zu bestimmen. Da aber Abweichungen zwischen der Abbruchquote der getesteten Kohorte und der zugrunde gelegten durchschnittlichen Abbruchquote vergangener Kohorten zu Abweichungen der Genauigkeitsmaße führen, wird der Schwellenwert zu Evaluationszwecken anhand der tatsächlichen Abbruchquote der Kohorte bestimmt. D.h. der Schwellenwert wird so gewählt, dass die Anzahl der identifizierten Studierenden der Anzahl künftiger Studienabbrecher in der Testkohorte entspricht. Anderenfalls ist es nicht möglich Genauigkeitseinbußen durch Schwellenwertabweichungen von daten- und verfahrensbasierten Genauigkeitseinbußen zu unterscheiden. Wenn das Frühwarnsystem auf aktuell immatrikulierte Kohorten angewendet werden soll, demnach die Zahl der Studienabbrecher der Kohorte noch unbekannt ist, dann kann der Schwellenwert auf der durchschnittlichen Abbruchquote früherer Kohorten basieren.

In den folgenden Abschnitten werden die Prognosegenauigkeiten für beide Schwellenwerte dargestellt. Der erste Schwellenwert basiert auf der tatsächlichen Anzahl der Abbrecher, der zweite auf der durchschnittlichen Abbruchquote vorangegangener Kohorten.

2.5.6. Gütekriterien eines Klassifikators

Der Studienerfolg wird als binäre Variable erfasst. Bei Prognosen von Studienerfolg und Studienabbruch können Studierende:

- als Abbrecher prognostiziert werden und abbrechen (richtig positiv)
- als Absolvent prognostiziert werden, aber abbrechen (falsch negativ)
- als Abbrecher klassifiziert werden und absolvieren (falsch positiv) und
- als Absolvent klassifiziert werden und absolvieren (richtig negativ).

Diese vier Möglichkeiten werden in der nachfolgenden Wahrheitsmatrix (*confusion matrix*) dargestellt.

	Prognose: Abbruch	Prognose: Erfolg
Studienabbruch	richtig positiv (t_p)	falsch negativ (f_n)
Studienerfolg	falsch positiv (f_p)	richtig negativ (t_n)

Wie in der Signaltheorie und Analyse der Genauigkeit diagnostischer Systeme üblich (Swets, 1988; Zweig & Campbell, 1993), kann die Prognosegüte binärer Klassifikatoren anhand der Gütemaße Accuracy, Specificity, Recall und Precision bestimmt werden.

$$\text{Accuracy: } \frac{t_p + t_n}{t_p + f_p + f_n + t_n} \quad \text{Recall: } \frac{t_p}{t_p + f_n}$$

$$\text{Precision: } \frac{t_p}{t_p + f_p} \quad \text{Specificity: } \frac{t_n}{t_n + f_p}$$

In der Beurteilung der Prognosegüte eines Frühwarnsystems gibt die Accuracy an, wie hoch der Anteil richtiger Prognosen an allen Prognosen ist. Eine Differenzierung nach Studienerfolg oder Studienabbruch entfällt. Für die vorliegende Fragestellung ist daher die Accuracy als Gütemaß nicht immer geeignet. Wenn beispielsweise eine Kohorte eine Abbruchquote von 10% aufweist, dann hat ein Frühwarnsystem, das 100% der Studierenden als Absolventen klassifiziert, eine Genauigkeit von 90%, obwohl im vorliegenden Fall kein Abbrecher korrekt prognostiziert wurde. Daher sind auch Gütemaße zur Beurteilung heranzuziehen, die nach Studienerfolg und -abbruch differenzieren. Specificity, gibt den Anteil der korrekt prognostizierten künftigen Absolventen an und ist daher für die vorliegende Fragestellung nicht relevant. Recall und Precision hingegen geben Aufschluss über die Prognosegüte der Verfahren für nicht erfolgreiche Studierende. Der Recall gibt den Anteil künftiger Studienabbrecher an, der zutreffend als Studienabbrecher erkannt wurde, während die Precision den Anteil richtiger Abbruchprognosen in Relation zu allen prognostizierten Abbrüchen darstellt. Ein Recall von 80% bedeutet, dass 80% aller zukünftigen

Studienabbrecher als solche erkannt wurden, wohingegen bei einer Precision von 80% auch 80% aller prognostizierten Studienabbrecher ihr Studium tatsächlich abbrechen.

Das Verhältnis der Gütemaße *Recall* und *Precision* ist abhängig von dem verwendeten Schwellenwert. Für FragSte werden zwei Schwellenwerte vorgestellt (vgl. Teil II, Abschnitt 2.4.5). Der erste Schwellenwert basiert auf der tatsächlichen Anzahl der Abbrecher, der zweite auf der durchschnittlichen Abbruchquote vorangegangener Kohorten.

Da unter Verwendung der tatsächlichen Anzahl der Abbrecher die Anzahl der tatsächlichen Abbrecher der Anzahl der prognostizierten Abbrecher entspricht, entspricht auch die Anzahl der falsch als Absolvent prognostizierten Studierenden der Anzahl der falsch als Abbrecher prognostizierten Studierenden; also ist $f_p = f_n$. Demnach sind auch die Gütemaße Recall und Precision identisch. Wird hingegen ein Schwellenwert basierend auf der durchschnittlichen Abbruchquote vorangegangener Kohorten genutzt, sind Recall und Precision davon abhängig, ob die tatsächliche Abbruchquote geringer oder höher als die durchschnittliche ist. Ist sie geringer, werden mehr Abbrecher identifiziert, ist sie höher, werden weniger Abbrecher identifiziert. Im ersten Fall findet eine Überidentifikation statt, durch die auch mehr Abbrecher korrekt identifiziert werden; t_p nimmt zu, was zu einem höheren Recall führt. Gleichzeitig werden allerdings auch mehr Absolventen zu Unrecht als Abbrecher klassifiziert; f_p nimmt zu, wodurch die Precision und damit die Glaubwürdigkeit einer Identifikation sinkt.

Zur weiteren Beurteilung der Prognosegenauigkeit wird die ROC-Kurve (Receiver Operating Characteristics) für die UNI und FH dargestellt. Die ROC-Kurve stammt aus der Signalentdeckungstheorie und kann dazu eingesetzt werden, ein Zwei-Klassen-Klassifizierungsproblem zu evaluieren (Bowers et al., 2013). Dabei wird der Recall auf der Ordinate und $1 - \frac{t_n}{t_n + f_p} = 1 - specificity$ auf der Abszisse dargestellt. Damit wird bei Wahl des Schwellenwertes zu treffende Kompromiss zwischen Trefferquote (richtig positiv) und Falschpositivquote illustriert. Eine bestmögliche Prognosemethode ergibt beispielsweise den Punkt $(x,y)=(0,1)$, was 100% Recall (keine falschen Negative) und 100% Specificity (keine falschen Positive) bedeutet. Eine zufällige Schätzung wird durch eine 45°-Linie dargestellt (50% Falschnegative und 50% Falschpositive). Je weiter sich die ROC Kurve der linken oberen Ecke annähert, desto höher ist die Gesamtgenauigkeit des Tests (Zweig & Campbell, 1993). Um die Performance der Klassifikation in einem Einzelwert darzustellen, wird zusätzlich Fläche unter der ROC-Kurve bestimmt (AUC ROC). Der AUC ROC

liegt zwischen 0 und 1, wobei Werte unter 0,5 die 45°-Referenzlinie unterschreiten und damit weniger genaue Vorhersagen implizieren als zufälliges Schätzen (Bradley, 1997).

3. Die Ergebnisse des Früherkennungssystems FragSte

In Abhängigkeit der Zeitpunkte an denen Informationen an deutschen Hochschulen verfügbar werden, ergeben sich unterschiedliche Prognosezeitpunkte: Bei Immatrikulation sind demografische Informationen sowie Informationen zum Schulabschluss und eventuell vorangegangener Studien verfügbar. Bis erste akademische Leistungen verfügbar werden, sind Prognosen nur auf Basis dieser Nicht-Leistungsdaten möglich. Da Prüfungsergebnisse erst im Laufe des Folgesemesters in die EDV aufgenommen werden, ist auch die Verwendung dieser Daten für Prognosen nicht vorher möglich. Beispielsweise sind Prognosen auf Basis der Erstsemesterleistungen frühestens Mitte des zweiten Semesters möglich. Gleiches gilt für alle Folgesemester. Im Folgenden werden Prognosen mit der Anzahl der vollendeten Semester gekennzeichnet. Prognosen basierend auf Erstsemesterleistungen werden als „1. Sem“ gekennzeichnet – finden aber im zweiten Semester statt. Damit sind Prognosen zwischen der Immatrikulation und dem Abschluss des ersten Semesters als „0. Sem“ beschrieben.

Im Folgenden wird zuerst die Aufteilung der Studierenden in Lern- und Testdaten und die Identifikationsschwelle beschrieben. Anschließend werden die Ergebnisse der einzelnen Verfahren vorgestellt. Das Logit-Modell wurde mit Stata, die Entscheidungsbäume mit WeKa (Frank et al., 2016), das neuronale Netz mit Membrain (Jetter, 2017) und der AdaBoost erneut mit Stata errechnet.

3.1. Lern- und Testkohorten UNI und FH

Das vorliegend vorgestellte Frühwarnsystem FragSte konnte an den eingangs beschrieben Hochschulen – UNI und FH – getestet werden. Dabei werden maschinellen Verfahren an ehemaligen Bachelor-Studierenden mit der Immatrikulationsjahre 2007 bis 2017 angelernnt und an jeweils einer Testkohorte eines Immatrikulationsjahrganges getestet, die nicht Teil der Lerndaten war. Als Testkohorte der privaten Fachhochschule wurden das Sommer- und Wintersemester des Jahres 2012 und der staatlichen Universität das Sommer und Wintersemester des Jahres 2010 ausgewählt. Das waren zum Zeitpunkt des Datenexports die aktuellsten Jahrgänge, in denen ein Großteil der immatrikulierten Stu-

dierenden ihr Studium beendet hatten und somit verlässliche Aussagen über den Studienerfolg der getesteten Studierenden möglich ist (siehe Tabelle 9). Die frühere Immatrikulationskohorte der Universität ist den vergleichsweise längeren Studiendauern geschuldet.

An der staatlichen Universität umfassten die Lerndaten der Kohorten 2007 bis 2009 und 2011 bis 2017 insgesamt 18.574 ehemalige Bachelor-Studierende und die Testkohorte 2010 insgesamt 2.127 ehemalige Bachelor-Studierende zu Beginn des ersten Semesters. An der privaten Fachhochschule umfassten die Lerndaten der Kohorten 2007 bis 2011 und 2013 bis 2017 insgesamt 9.721 ehemalige Studierende und die Testkohorte 2012 insgesamt 1.488 ehemalige Bachelor-Studierende zu Beginn des ersten Semesters. Da im Studienverlauf immer mehr Studierende die Hochschulen verlassen, nimmt die Anzahl ehemaliger Studierender in Lern- wie Testdatensatz mit höheren Betrachtungssemestern ab. In den ersten sechs Semestern sind es weit überwiegend Studienabbrecher die die Hochschulen verlassen. Tabelle 18 zeigt die Anzahl ehemaliger Studierender in Lern- und Testkohorten im Verlauf der ersten fünf Semester.

Tabelle 18. Anzahl ehemaliger Studierender in Lern- und Testkohorten

Sem.	UNI		FH	
	Lernkohorten 2007-2017	Testkohorte 2010	Lernkohorten 2007-2017	Testkohorte 2012
0	18.574	2.127	9.721	1.488
1	15.081	1.765	8.658	1.392
2	11.312	1.582	7.803	1.294
3	9.984	1.494	7.244	1.228
4	8.556	1.341	6.753	1.158

Hinweis: Die jeweilige Lernkohorte war nicht Teil der Lerndaten

3.2. Identifikationsschwellenwert

Jedes Verfahren prognostiziert die Abbruchwahrscheinlichkeit für jeden Studierenden des Testdatensatzes mit einer Zahl zwischen 0 (Absolvent) und 1 (Abbrecher). Abbildung 7 stellt die Genauigkeit einer Prognose in Abhängigkeit des Abstandes des Prognosewertes zum Identifikationsschwellenwert dar. Der Identifikationsschwellenwert stellt den Wert dar, ab dem ein Studierender als abbruchgefährdet gilt.

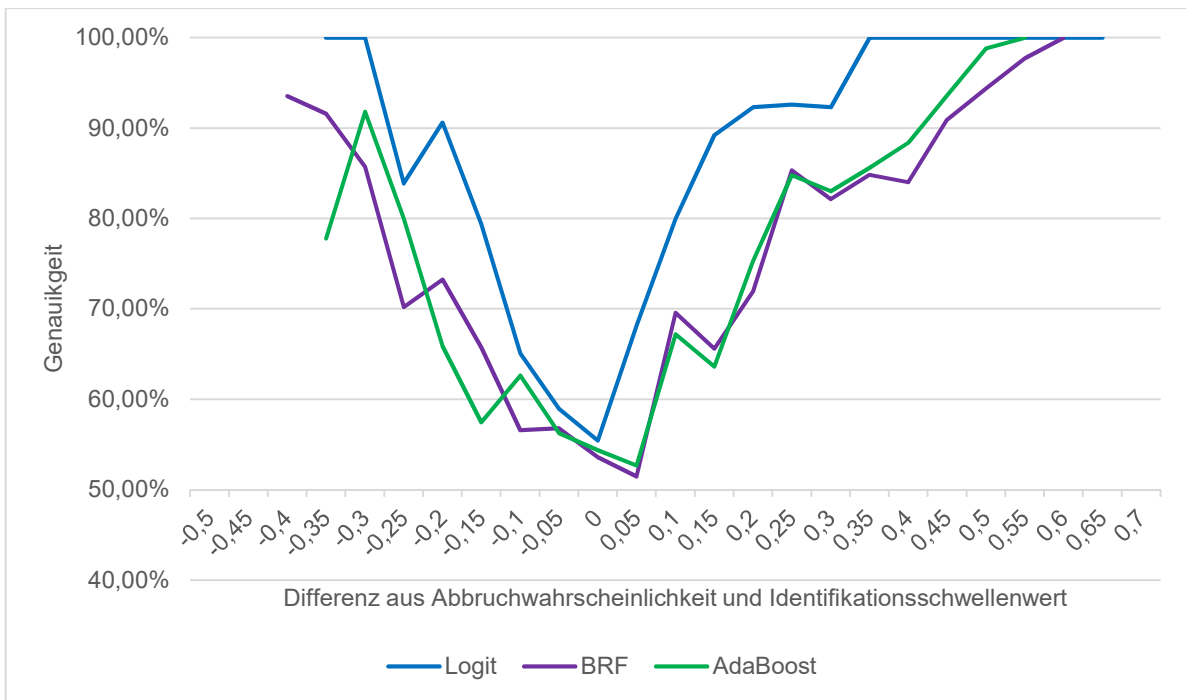


Abbildung 7. Prognosegenauigkeit in Abhängigkeit des Identifikationsschwellenwertes

Es zeigt sich, dass Prognosen nahe 0 oder 1 eine hohe Genauigkeit aufweisen und Prognosen, die sich in der Nähe des Identifikationsschwellenwerts bewegen, unsicher sind. Dies gilt für alle Klassifikatoren, jedoch übertrifft der AdaBoost den Logit und den BRF, wenn auch nicht über den gesamten Beobachtungsbereich.

Aber auch die Kohortenzusammensetzung hat einen Einfluss auf die Prognosegenauigkeit. Die Abbruchquote der Testkohorte bestimmt in der vorliegenden Untersuchung zu Evaluationszwecken den Identifikationsschwellenwert (vgl. Teil II, Abschnitt 2.4.5). Dieser wird so gewählt, dass die Anzahl tatsächlicher Studienabbrecher mit der Anzahl identifizierter Studienabbrecher übereinstimmt. Die geringsten Genauigkeiten ergeben sich ceteris paribus für Kohorten mit einem ausgeglichenen Verhältnis von Absolventen und Studienabbrecher. Mit zunehmendem Ungleichgewicht, nimmt die Wahrscheinlichkeit „richtig“ klassifizierter Studierender zu. Sollte eine Kohorte gänzlich aus künftigen Abbrechern oder Absolventen bestehen, ist eine Fehlklassifikation nicht mehr möglich. Die Kohortenzusammensetzung aus Studienabbrechern und Absolventen ist hochschulspezifisch, aber auch vom Zeitpunkt der Prognose abhängig. Da Studienabbrüche durchschnittlich früher eintreten als Studienabschlüsse, sinkt der Anteil künftiger Studienabbrechern innerhalb einer Kohorte mit zunehmend späterem Betrachtungszeitpunkt.

Damit ist zu erwarten, dass die Prognosegenauigkeit mit zunehmend späterem Betrachtungszeitpunkten auch aufgrund der Kohortenzusammensetzung zunimmt und frühe Prognosen ungenauer sind als späte. Weiterhin liegen zu späteren Betrachtungszeitpunkten mehr Informationen über das Studierverhalten vor.

3.3. Logit Model

Tabelle 19 stellt die Prognosegenauigkeit des Logit Modells dar. Wie erwartet, nimmt die Qualität der Prognose mit zunehmend späterem Betrachtungszeitpunkten zu. Dies gilt für alle Gütemaße. So steigt beispielsweise der Recall (wie viele künftige Studienabbrecher identifiziert werden) an der UNI von etwa 71% im ersten Semester auf 80% im vierten Semester. An der FH betrug der Recall für das 1. und 4. Semester 69% bzw. 78%.

Tabelle 19. Prognosegenauigkeit des Logitmodells (tatsächlichen Abbruchquote)

Logit	UNI					FH				
	0. Sem	1. Sem.	2. Sem.	3. Sem.	4. Sem.	0. Sem	1. Sem.	2. Sem.	3. Sem.	4. Sem.
Accuracy ^a	63,47	76,15	81,80	86,55	89,56	66,53	82,67	88,47	91,36	93,76
Recall ^a	61,67	70,80	74,05	78,71	79,94	48,95	68,68	74,30	76,68	78,26
Precision ^a	61,74	71,09	74,05	79,20	79,94	48,95	69,23	74,56	77,03	78,26
Anzahl Absolventen	1.112	1.039	1.027	1.015	992	976	976	974	969	961
Anzahl Abbrecher	1.015	726	555	479	349	476	380	284	223	161
korrekt klassifizierte Absolventen	724	830	883	916	922	733	860	902	918	926
falsch klassifizierte Abbrecher	388	209	144	99	70	243	116	72	51	35
korrekt klassifizierte Abbrecher	626	514	411	377	279	233	261	211	171	126
falsch klassifizierte Absolventen	389	212	144	102	70	243	119	73	52	35
korrekt klassifizierte Absolventen ^a	65,11	79,88	85,98	90,25	92,94	75,10	88,11	92,61	94,74	96,36
falsch klassifizierte Abbrecher ^a	34,89	20,12	14,02	9,75	7,06	24,90	11,89	7,39	5,26	3,64
korrekt klassifizierte Abbrecher ^a	61,67	70,80	74,05	78,71	79,94	48,95	68,68	74,30	76,68	78,26
falsch klassifizierte Absolventen ^a	38,33	29,20	25,95	21,29	20,06	51,05	31,32	25,70	23,32	21,74

Hinweis: ^a Werte in Prozent.

Wenn die tatsächliche Anzahl der Studienabbrecher unbekannt ist – beispielsweise wenn das Frühwarnsystem im Realeinsatz zur Prognose von Studienabbrüchen an Studierenden eingesetzt wird - muss die Identifikationsschwelle auf den durchschnittlichen Abbruchquoten früherer Kohorten basieren (vgl. Tabelle 20). Dies wird voraussichtlich zu einer Verringerung der Prognosegenauigkeit führen: Im vorliegenden Fall liegt die Genauigkeit (Recall) für die UNI zwischen 73% (84%) im ersten Semester und 82% (92%) im

vierten Semester. Das heißt, im Vergleich der Ergebnisse, die auf der tatsächlichen Abbruchquote basieren, nimmt die Genauigkeit erwartungsgemäß ab. Jedoch ergeben sich für den Recall bessere Werte. Diese sind aber der unterdurchschnittlichen Abbruchquote der Testkohorte geschuldet. Durch eine Identifikationsschwelle, die der durchschnittlichen Abbruchquote entspricht, werden mehr Studierende als Studienabbrecher identifiziert als Studienabbrecher in der Kohorte enthalten sind – es findet eine Überidentifikation statt. Die besseren Werte des Recalls gehen damit zu Lasten aller anderen Gütemaße.

Tabelle 20. Prognosegenauigkeit des Logitmodells (durchschnittliche Abbruchquote)

Logit	UNI					FH				
	0. Sem.	1. Sem.	2. Sem.	3. Sem.	4. Sem.	0. Sem.	1. Sem.	2. Sem.	3. Sem.	4. Sem.
Accuracy ^a	63,19	72,75	76,36	81,33	81,73	66,87	80,53	86,96	91,11	93,05
Recall ^a	69,06	83,61	87,57	90,81	91,69	49,47	63,36	68,52	74,68	73,18
Precision ^a	59,91	62,64	61,44	64,93	59,70	48,95	72,37	78,17	79,37	81,37
Anzahl Absolventen	1.112	1.039	1.027	1.015	992	976	976	974	969	961
Anzahl Abbrecher	1.015	726	555	479	349	476	380	284	223	161
korrekt klassifizierte Absolventen	643	677	722	780	776	738	817	872	909	913
falsch klassifizierte Abbrecher	469	362	305	235	216	243	105	62	46	30
korrekt klassifizierte Abbrecher	701	607	486	435	320	233	275	222	177	131
falsch klassifizierte Absolventen	314	119	69	44	29	238	159	102	60	48
korrekt klassifizierte Absolventen ^a	57,82	65,16	70,30	76,85	78,23	75,61	83,71	89,53	93,81	95,01
falsch klassifizierte Abbrecher ^a	42,18	34,84	29,70	23,15	21,77	51,05	27,63	21,83	20,63	18,63
korrekt klassifizierte Abbrecher ^a	69,06	83,61	87,57	90,81	91,69	48,95	72,37	78,17	79,37	81,37
falsch klassifizierte Absolventen ^a	30,94	16,39	12,43	9,19	8,31	24,39	16,29	10,47	6,19	4,99

Hinweis: ^a Werte in Prozent.

3.4. Bagging with Random Forest und Neural Network

In einem zweiten Schritt werden Verfahren des maschinellen Lernens angewandt. In Übereinstimmung mit verwandten Beiträgen (vgl. Tabelle 1), sind die Prognoseergebnisse aus dem Logit-Regressionsmodell, dem neuronalen Netzwerk und BRF sehr vergleichbar. Ebenfalls bestätigen sich die robusten Ergebnisse des BRF. Diese Methode übertraf die Prognosegenauigkeit der anderen Methoden um 0,88 - 2,93% (UNI) und 0,88 - 1,03% (FH).

Tabelle 21. Prognosegenauigkeit des BRF (tatsächliche Abbruchquote)

BRF	UNI					FH				
	0. Sem	1. Sem.	2. Sem.	3. Sem.	4. Sem.	0. Sem	1. Sem.	2. Sem.	3. Sem.	4. Sem.
Accuracy ^a	65,02	78,41	82,43	86,95	88,96	63,36	82,52	88,79	91,36	93,58
Recall ^a	63,55	73,83	75,14	79,96	79,08	44,17	68,77	75,09	76,79	77,30
Precision ^a	63,30	73,73	74,87	79,46	78,63	44,54	68,95	75,35	77,13	78,26
Anzahl Absolventen	1.112	1.039	1.027	1.015	992	976	976	974	969	961
Anzahl Abbrecher	1.015	726	555	479	349	476	380	284	223	161
korrekt klassifizierte Absolventen	738	848	887	916	917	708	857	903	917	924
falsch klassifizierte Abbrecher	374	191	140	99	75	264	118	70	51	35
korrekt klassifizierte Abbrecher	645	536	417	383	276	212	262	214	172	126
falsch klassifizierte Absolventen	370	190	138	96	73	268	119	71	52	37
korrekt klassifizierte Absolventen ^a	66,37	81,62	86,37	90,25	92,44	72,54	87,81	92,71	94,63	96,15
falsch klassifizierte Abbrecher ^a	33,63	18,38	13,63	9,75	7,56	55,46	12,09	24,65	22,87	21,74
korrekt klassifizierte Abbrecher ^a	63,55	73,83	75,14	79,96	79,08	44,54	68,95	75,35	77,13	78,26
falsch klassifizierte Absolventen ^a	36,45	26,17	24,86	20,04	20,92	27,46	31,32	7,29	5,37	3,85

Hinweis: ^a Werte in Prozent.

Tabelle 22. Prognosegenauigkeit des BRF (durchschnittliche Abbruchquote)

BRF	UNI					FH				
	0. Sem	1. Sem.	2. Sem.	3. Sem.	4. Sem.	0. Sem	1. Sem.	2. Sem.	3. Sem.	4. Sem.
Accuracy ^a	65,35	74,84	77,24	80,79	81,43	63,71	82,45	87,52	91,44	92,87
Recall ^a	71,43	86,36	89,01	90,40	91,40	44,61	66,28	69,30	75,31	72,38
Precision ^a	61,86	64,51	62,30	64,24	59,29	44,33	76,05	80,28	80,72	81,37
Anzahl Absolventen	1.112	1.039	1.027	1.015	992	976	976	974	969	961
Anzahl Abbrecher	1.015	726	555	479	349	476	380	284	223	161
korrekt klassifizierte Absolventen	665	694	728	774	773	714	829	873	910	911
falsch klassifizierte Abbrecher	447	345	299	241	219	265	91	56	43	30
korrekt klassifizierte Abbrecher	725	627	494	433	319	211	289	228	180	131
falsch klassifizierte Absolventen	290	99	61	46	30	262	147	101	59	50
korrekt klassifizierte Absolventen ^a	59,80	66,79	70,89	76,26	77,92	73,16	84,94	89,63	93,91	94,80
falsch klassifizierte Abbrecher ^a	40,20	33,21	29,11	23,74	22,08	55,67	23,95	19,72	19,28	18,63
korrekt klassifizierte Abbrecher ^a	71,43	86,36	89,01	90,40	91,40	44,33	76,05	80,28	80,72	81,37
falsch klassifizierte Absolventen ^a	28,57	13,64	10,99	9,60	8,60	26,84	15,06	10,37	6,09	5,20

Hinweis: ^a Werte in Prozent.

Tabelle 23. Prognosegenauigkeit des neuronalen Netzes (tatsächliche Abbruchquote)

Neural network	UNI					FH				
	0. Sem	1. Sem.	2. Sem.	3. Sem.	4. Sem.	0. Sem	1. Sem.	2. Sem.	3. Sem.	4. Sem.
Accuracy ^a	62,53	72,75	81,54	85,27	86,35	66,67	82,49	88,70	92,79	94,76
Recall ^a	60,69	47,80	70,09	73,28	72,78	49,13	68,51	74,44	80,00	80,42
Precision ^a	60,75	77,28	75,53	79,23	74,27	49,13	68,51	74,44	80,00	80,42
Anzahl Absolventen	1.112	1.039	1.027	1.015	992	976	976	974	969	961
Anzahl Abbrecher	1.015	726	555	479	349	476	380	284	223	161
korrekt klassifizierte Absolventen	714	937	901	923	904	707	826	870	892	897
falsch klassifizierte Abbrecher	398	102	126	92	88	233	114	68	41	28
korrekt klassifizierte Abbrecher	616	347	389	351	254	225	248	198	164	115
falsch klassifizierte Absolventen	399	379	166	128	95	233	114	68	41	28
korrekt klassifizierte Absolventen ^a	64,21	90,18	87,73	90,94	91,13	75,21	87,87	92,75	95,61	96,97
falsch klassifizierte Abbrecher ^a	35,79	9,82	12,27	9,06	8,87	24,79	12,13	7,25	4,39	3,03
korrekt klassifizierte Abbrecher ^a	60,69	47,80	70,09	73,28	72,78	49,13	68,51	74,44	80,00	80,42
falsch klassifizierte Absolventen ^a	39,31	52,20	29,91	26,72	27,22	50,87	31,49	25,56	20,00	19,58

Hinweis: ^a Werte in Prozent.

Tabelle 24. Prognosegenauigkeit des neuronalen Netzes (durchschnittliche Abbrüchen)

Neural Network	UNI					FH				
	0. Sem	1. Sem.	2. Sem.	3. Sem.	4. Sem.	0. Sem	1. Sem.	2. Sem.	3. Sem.	4. Sem.
Accuracy ^a	63,38	73,14	76,86	78,78	79,34	66,74	81,34	87,54	91,65	92,51
Recall ^a	69,06	80,99	84,86	87,06	83,67	49,23	64,27	68,59	74,12	68,42
Precision ^a	60,12	63,64	62,55	62,05	57,03	48,69	74,03	80,45	82,44	81,82
Anzahl Absolventen	1.112	1.039	1.027	1.015	992	976	976	974	969	961
Anzahl Abbrecher	1.015	726	555	479	349	476	380	284	223	161
korrekt klassifizierte Absolventen	647	703	745	760	772	710	791	840	874	871
falsch klassifizierte Abbrecher	465	336	282	255	220	235	94	52	36	26
korrekt klassifizierte Abbrecher	701	588	471	417	292	223	268	214	169	117
falsch klassifizierte Absolventen	314	138	84	62	57	230	149	98	59	54
korrekt klassifizierte Absolventen ^a	58,18	67,66	72,54	74,88	77,82	75,53	84,15	89,55	93,68	94,16
falsch klassifizierte Abbrecher ^a	41,82	32,34	27,46	25,12	22,18	25,00	10,00	5,54	3,86	2,81
korrekt klassifizierte Abbrecher ^a	69,06	80,99	84,86	87,06	83,67	48,69	74,03	80,45	82,44	81,82
falsch klassifizierte Absolventen ^a	30,94	19,01	15,14	12,94	16,33	50,22	41,16	36,84	28,78	37,76

Hinweis: ^a Werte in Prozent.

Unter Verwendung der durchschnittlichen Abbruchquote als Identifikationsschwellenwert beträgt die Genauigkeit des BRF (Neuronales Netzwerk) im ersten Semester 75% (73%) und steigt im vierten Semester auf 81% (79%) an. Da die Abbruchquote in der Testkohorte unterdurchschnittlich ist, ist ein hoher Recall auf Kosten einer reduzierten Precision zu erwarten. Das bestätigt sich. Der Recall des BRF (Neuronales Netzwerk) beträgt 86% (81%) im ersten Semester und 91% (83%) im vierten Semester.

Um die diagnostische Qualität unserer Klassifikatoren weiter zu beurteilen, wird der ROC für jedes Verfahren, beide Hochschulen und für die ersten vier Semester dargestellt und der AUC ROC ermittelt (Abbildungen 8 und 9). Dabei zeigt sich, dass alle Methoden einer zufälligen Schätzung deutlich überlegen sind. Zudem nimmt der AUC-ROC mit zunehmend späteren Betrachtungszeitpunkt im Studium zu, das bedeutet die Prognosegenauigkeit nimmt zu. Darüber hinaus sind Unterschiede in den Ergebnissen der Verfahren je nach Hochschule und Semester festzustellen. Aus diesem Grund werden die Prognosen des neuronalen Netzes, des BRF und des Logit-Modells mittels des AdaBoost-Algorithmus kombiniert.

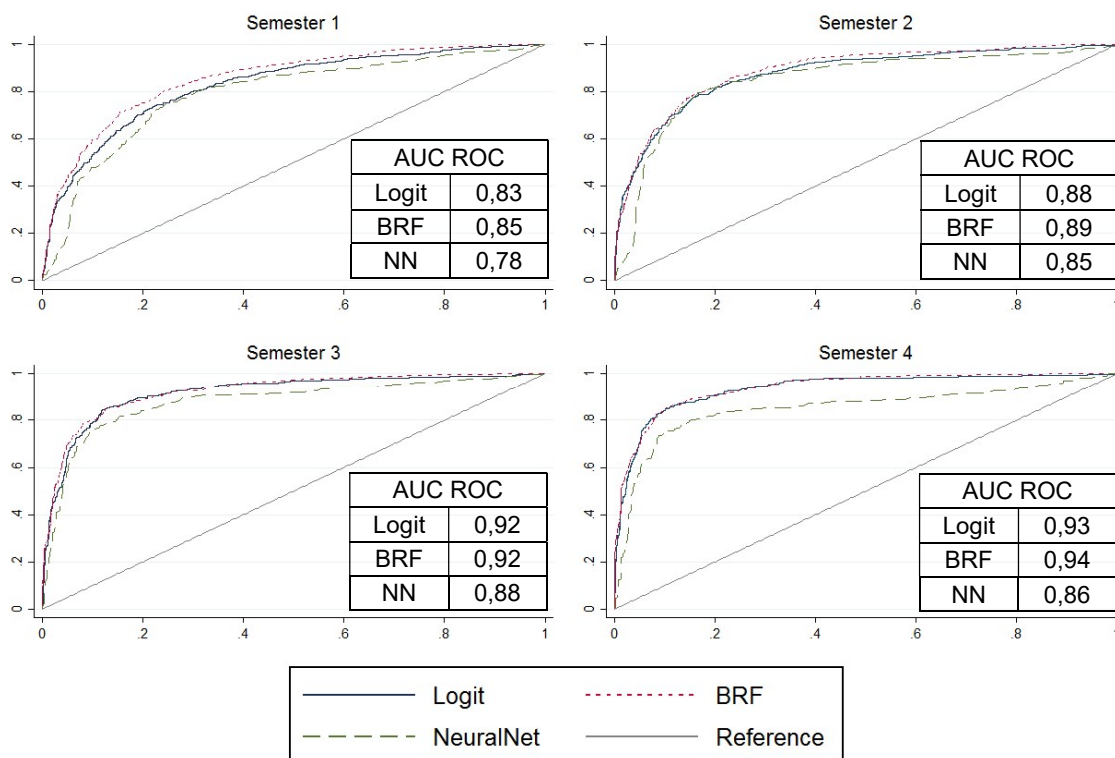


Abbildung 8. ROC und AUC ROC — UNI

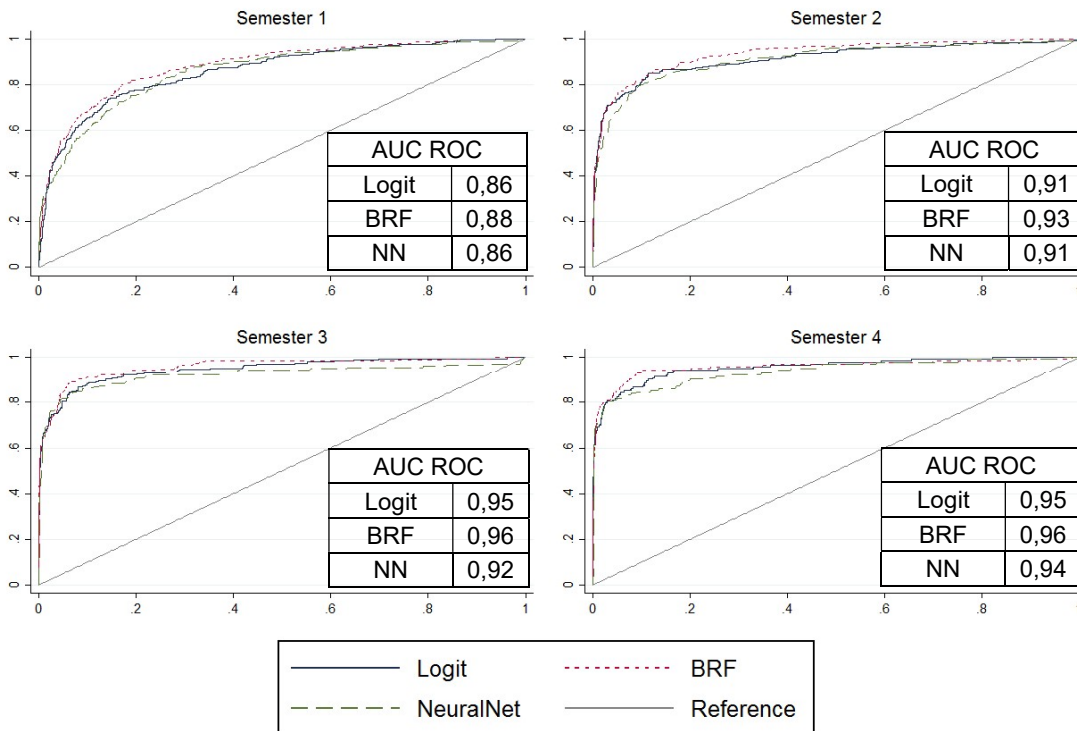


Abbildung 9. ROC und AUC ROC — FH

3.5. AdaBoost

Tabelle 25 fasst die Prognosegenauigkeit des AdaBoost zusammen. Der AdaBoost stellt eine Kombination aller Prognosen aller eingesetzter Klassifikatoren als gewichtetes Mittel dar. Das Ergebnis des AdaBoost wird dabei stets gegen das beste Ergebnis konvergieren oder diesem überlegen sein (Freund & Schapire, 1999).

Für die UNI und FH ergeben sich erhebliche Unterschiede in der Prognostizierbarkeit von Studienabbrüchen - Accuracy, Recall und Precision weichen stellenweise stark voneinander ab. Allerdings zeigen sich auch hochschulübergreifende Gemeinsamkeiten. Mit zunehmend späteren Betrachtungszeitpunkt im Studium werden die Prognosen genauer. Einerseits liegen zu einem späteren Zeitpunkt zunehmend mehr Informationen über die akademische Leistung vor und zum anderen ändert sich die Kohortenzusammensetzung im Studienverlauf. Die Kohorten setzen sich stärker aus künftigen Absolventen zusammen.

Tabelle 25. Prognosegenauigkeit des AdaBoost (tatsächliche Abbruchquote)

AdaBoost	UNI					FH				
	0. Sem.	1. Sem.	2. Sem.	3. Sem.	4. Sem.	0. Sem.	1. Sem.	2. Sem.	3. Sem.	4. Sem.
Accuracy ^a	67,65	78,53	82,43	87,62	89,63	67,17	84,49	89,70	93,50	95,51
Recall ^a	65,81	73,83	74,95	80,58	79,94	49,78	72,10	76,69	81,95	83,22
Precision ^a	66,20	73,93	74,95	80,75	80,17	49,89	72,10	76,69	81,95	83,22
Anzahl Absolventen	1.112	1.039	1.027	1.015	992	976	976	974	969	961
Anzahl Abbrecher	1.015	726	555	479	349	476	380	284	223	161
korrekt klassifizierte Absolventen	771	850	888	923	923	711	839	876	896	901
falsch klassifizierte Abbrecher	341	189	139	92	69	230	101	62	37	24
korrekt klassifizierte Abbrecher	668	536	416	386	279	228	261	204	168	119
falsch klassifizierte Absolventen	347	190	139	93	70	229	101	62	37	24
korrekt klassifizierte Absolventen ^a	69,33	81,81	86,47	90,94	93,04	75,64	89,26	93,39	96,03	97,41
falsch klassifizierte Abbrecher ^a	30,67	18,19	13,53	9,06	6,96	24,47	10,74	6,61	3,97	2,59
korrekt klassifizierte Abbrecher ^a	65,81	73,83	74,95	80,58	79,94	49,78	72,10	76,69	81,95	83,22
falsch klassifizierte Absolventen ^a	34,19	26,17	25,05	19,42	20,06	50,00	27,90	23,31	18,05	16,78

Hinweis: ^a Werte in Prozent.

Im ersten Semester, noch vor der ersten Prüfung, verließen etwa 21% aller Studienabbrecher in der Stichprobe die UNI und 28,5% die FH. Die Prognosegenauigkeit liegt für beide Hochschulen bei etwa 68%, jedoch werden an UNI und FH abbruchgefährdete Studierende nicht gleich gut prognostiziert. An der FH werden erfolgreiche Studenten besser vorhergesagt als an der Uni; während an der UNI gefährdete Studenten besser vorhergesagt werden als an der FH.

Im Verfahrensvergleich sind die durch den AdaBoost kombinierten Prognosen den Prognosen der einzelnen Verfahren stets überlegen. Das ist insofern vorteilhaft, da damit nicht nur die Prognosegenauigkeit verbessert werden kann, sondern eine auf Hochschule und Semester basierende Auswahl eines Einzelverfahrens zugunsten des AdaBoost entfallen kann.

Tabelle 26. Prognosegenauigkeit des AdaBoost (durchschnittliche Abbruchquote)

AdaBoost	UNI					FH				
	0. Sem	1. Sem.	2. Sem.	3. Sem.	4. Sem.	0. Sem	1. Sem.	2. Sem.	3. Sem.	4. Sem.
Accuracy ^a	67,28	75,35	78,07	82,13	82,18	67,10	83,18	88,95	92,88	93,45
Recall ^a	73,20	86,91	89,91	92,28	92,55	49,78	67,15	71,25	77,19	71,35
Precision ^a	63,67	64,98	63,16	65,77	60,26	49,34	77,35	83,83	85,85	85,31
Anzahl Absolventen	1.112	1.039	1.027	1.015	992	976	976	974	969	961
Anzahl Abbrecher	1.015	726	555	479	349	476	380	284	223	161
korrekt klassifizierte Absolventen	688	699	736	785	779	712	803	848	881	876
falsch klassifizierte Abbrecher	424	340	291	230	213	232	82	43	29	21
korrekt klassifizierte Abbrecher	743	631	499	442	323	226	280	223	176	122
falsch klassifizierte Absolventen	272	95	56	37	26	228	137	90	52	49
korrekt klassifizierte Absolventen ^a	61,87	67,28	71,67	77,34	78,53	75,74	85,43	90,41	94,43	94,70
falsch klassifizierte Abbrecher ^a	38,13	32,72	28,33	22,66	21,47	24,68	8,72	4,58	3,11	2,27
korrekt klassifizierte Abbrecher ^a	73,20	86,91	89,91	92,28	92,55	49,34	77,35	83,83	85,85	85,31
falsch klassifizierte Absolventen ^a	26,80	13,09	10,09	7,72	7,45	49,78	37,85	33,83	25,37	34,27

Hinweis: ^a Werte in Prozent.

Wird der Identifikationsschwellenwert für die Testkohorte der UNI anhand der durchschnittlichen Abbruchquote anstelle der tatsächlichen Abbruchquote bestimmt, beträgt die AdaBoost-Genauigkeit 75% (Recall: 87%) im ersten Semester und 82% (93%) im vierten Semester. Im Vergleich der Prognosegenauigkeit beider Identifikationsschwellenwerte ergibt sich erneut eine geringere Accuracy und Precision bei erhöhtem Recall bei Verwendung der durchschnittlichen Abbruchquote. Die eine unterdurchschnittliche Abbruchquote der Testkohorte führt zu einer Überidentifikation.

3.6. Einfluss der Testkohorte auf die Prognosegenauigkeit

In vorherigen Abschnitten wurde die Prognosegenauigkeit verschiedener Verfahren, zu verschiedenen Zeitpunkten im Studium an zwei Hochschulen dargestellt. Dabei wurden die Studierendendaten exmatrikulierter Studierender der Immatrikulationskohorten 2007 bis 2017, abzüglich der Testkohorte, als Lerndatensatz verwendet. Als Testkohorten wurden die Immatrikulationskohorten 2010 (UNI) beziehungsweise 2012 (FH) verwendet. Immatrikulationskohorten späterer Jahrgänge sind als Testkohorte ungeeignet, da in diesen der Anteil noch immatrikulierter Studierender die Kohortenzusammensetzung verändert. Da Studienabbrüche durchschnittlich deutlich früher

eintreten als Studienabschlüsse, führen hohe Verbleibquoten zu einer Überrepräsentation von Studienabbrechern. In Tabelle 9 werden die Verbleibquoten der Immatrikulationskohorten von Bachelorstudiengängen beider Hochschulen dargestellt

In Tabelle 27 werden die Prognosegenauigkeiten für beide Hochschulen für fünf Zeitpunkte im Studium jeweils für die Testkohorten 2010, 2011, 2012 und 2013 dargestellt. Da nur der Einfluss der Verbleibquote der Testkohorte betrachtet werden soll, sind lediglich die Genauigkeiten der Testkohorten untereinander zu vergleichen. Um den Testaufwand zu begrenzen, genügt es, wenn ein Verfahren exemplarisch für alle Verfahren, verwendet wird. Daher wird der BRF verwendet. Die Ergebnisse des BRF haben sich in der Literatur und in der vorliegenden Studie als hinreichend robust gezeigt.

Tabelle 27 Prognosegenauigkeit des BRF — in Abhängigkeit der Testkohorte

BRF		UNI					FH				
		0. Sem.	1. Sem.	2. Sem.	3. Sem.	4. Sem.	0. Sem.	1. Sem.	2. Sem.	3. Sem.	4. Sem.
Accuracy	2010	65,02	78,41	82,43	86,95	88,96	61,75	82,03	88,00	90,28	92,89
	2011	59,35	81,07	83,99	87,63	88,71	62,31	80,46	87,45	90,37	93,92
	2012	69,07	83,02	86,90	89,25	90,19	63,36	82,52	88,79	91,36	93,58
	2013	74,41	86,53	88,65	89,57	89,86	64,80	82,78	88,93	92,99	95,98
Recall	2010	63,55	73,83	75,14	79,96	79,08	42,02	65,69	69,08	67,97	69,91
	2011	66,88	81,73	82,01	84,80	83,22	39,90	61,54	69,23	70,93	77,94
	2012	76,79	86,18	87,39	88,54	87,06	44,17	68,77	75,09	76,79	77,30
	2013	84,09	91,15	91,33	91,28	89,50	50,32	71,07	75,00	80,43	84,92

Hinweis: Das Jahr stellt das Immatrikulationsjahr der Testkohorte dar. Alle Ergebnisse in Prozent

Die Prognosegenauigkeit des BRF nimmt, mit wenigen Ausnahmen, mit zunehmend späterer Testkohorte in Abhängigkeit der Verbleibquote stetig zu. Das gilt für alle fünf Prognosezeitpunkte und für beide Hochschulen. An der FH ergeben sich in den Jahren 2010-2012 nur geringe Veränderungen der Verbleibquote und auch die Genauigkeiten nehmen nur verhalten zu. Anders hingegen die Entwicklung der Verbleibquote und der Genauigkeit an der UNI. Die Verbleibquote und die Veränderung der Prognosegenauigkeit sind – je nach Semester, Hochschule und Gütemaß – zwischen $r=0,80$ und $r=0,98$ korreliert (Cohen, 2013). Wenn ein Zusammenhang zwischen Verbleibquote und Prognosegenauigkeit besteht, dann verändert sich die Prognosegenauigkeit bei Veränderung der Verbleibquote. In dem Fall führt eine „künstliche“ Erhöhung der Verbleibquote zu besseren Prognoseergebnissen. Das kann getestet werden, indem eine Testkohorte mit geringer Verbleibquote – d.h. großem zeitlichen Ab-

stand zwischen Immatrikulationsjahr und Jahr des Datenexports – sukzessive um Studierende mit später Betrachtungszeitpunkte bereinigt wird. Dadurch wird der Betrachtungszeitraum künstlich verkürzt und die Verbleibquote erhöht. Dem Test wird die Immatrikulationskohorte 2010 der UNI nach dem ersten Semester zugrunde gelegt. Demnach werden Nichtleistungsdaten und Erstsemesterprüfungen berücksichtigt. Die Kohorte verfügt zum Zeitpunkt des Datenexports in 2017 nach einem Betrachtungszeitraum von 14 Semestern über eine Verbleibquote von 11,74%. Wird die Betrachtungsdauer auf 12 Semester reduziert, sind alle Abbrecher und Absolventen mit einer Studiendauer von mehr als 11 Semestern dem Datensatz zu entziehen – sie wären bei einer Betrachtung nach 12 Semestern noch immatrikuliert gewesen. Nachfolgend werden Prognosen basierend auf den Informationen der Semester 4-14 in zwei Semesterschritten betrachtet. Da die Entwicklung der Verbleibquote und der Prognosegenauigkeit insbesondere nahe der Regelstudiendauer von hohem Interesse ist, wird zusätzlich das siebte Semester berücksichtigt.

Tabelle 28. Abhängigkeit der Prognosegenauigkeit von der Betrachtungsdauer

BRF	UNI-Kohorte 2010 - Semester 1						
	14 Sem.	12 Sem.	10 Sem.	8 Sem.	7 Sem.	6 Sem.	4 Sem.
Accuracy ^a	78,30	79,30	79,70	81,11	82,77	83,80	86,08
Recall ^a	73,59	75,21	76,54	80,54	84,26	87,79	92,06
Precision ^a	73,69	75,32	76,66	80,69	84,43	87,97	92,31
Verbleibquote ^a	11,74	15,44	24,27	37,76	47,18	55,52	67,39
Anzahl Absolventen	1.039	975	829	584	410	236	47
Anzahl Abbrecher	727	702	635	555	502	475	378
korrekt klassifizierte Absolventen	847	801	680	476	331	178	17
falsch klassifizierte Abbrecher	191	173	148	107	78	57	29
korrekt klassifizierte Abbrecher	535	528	486	447	423	417	348
falsch klassifizierte Absolventen	192	174	149	108	79	58	30
korrekt klassifizierte Absolventen ^a	81,52	82,15	82,03	81,51	80,73	75,42	36,17
falsch klassifizierte Abbrecher ^a	18,38	17,74	17,85	18,32	19,02	24,15	61,70
korrekt klassifizierte Abbrecher ^a	73,59	75,21	76,54	80,54	84,26	87,79	92,06
falsch klassifizierte Absolventen ^a	26,41	24,79	23,46	19,46	15,74	12,21	7,94

Hinweis: ^a Werte in Prozent.

Prognosen von Testkohorten mit kurzer Betrachtungsdauer und hoher Verbleibquote führen zwar zu guten Prognosen, die aber irreführend sind. Die Abhängigkeit von Prognosegenauigkeit und Verbleibquote erklärt sich aus der Zusammensetzung der Testkohorten nach einer jeweiligen Betrachtungsdauer. Verfahren des „*supervised learning*“ benötigen gelabelte Daten. Das bedeutet, dass nur ehemalige Studierende in Lern- und Testdatensätzen enthalten sind. Des Weiteren finden Studienabbrüche früher statt als Studienabschlüsse. Das bedeutet, dass Kohorten mit einer Beobachtungsdauer von weniger als sechs Semester sich aus Studienabbrechern und wenigen atypisch schnellen Absolventen⁶ zusammensetzen – die künftigen Absolventen in und jenseits der Regelstudienzeit sind noch immatrikuliert und nicht Teil der Daten. Mit zunehmend längeren Beobachtungsdauern kommen erst typische Absolventen und späte Abbrecher und bei noch längeren Beobachtungsdauern Langzeitstudierende hinzu. Das bedeutet, dass mit zunehmend längeren Beobachtungsdauern zunehmend indifferente Studierende im Testdatensatz enthalten sind, deren Studienergebnis zu Studienbeginn – hier der Prognosezeitpunkt im des zweiten Semester – schlecht prognostizierbar ist.

Das stellt einen Nachteil von Frühwarnsystemen basierend auf Verfahren des „*supervised learning*“ dar. Es werden erhebliche Beobachtungsdauern benötigt, bis auch „langsame“ Studienabbrecher und Absolventen Teil der Lern- und Testdaten sind. Erschwerend kommt hinzu, dass nicht nur das Label – Abschluss oder Abbruch – lange Beobachtungsdauern benötigt, sondern alle verwendeten Informationen bereits seit Beginn des Beobachtungszeitraums vorliegen müssen. Wenn neu verfügbare Informationen der Datenselektion hinzugefügt werden sollen, beispielsweise Auswertungen von Self-Assessments als Indikator der Identifikation, können diese Informationen erst mit zunehmend längerer Beobachtungsdauer sinnvoll verwendet werden. Die hierfür notwendige Beobachtungsdauer ist hochschulspezifisch. An der hier betrachteten FH genügen wesentlich kürzere Zeiträume als an der UNI (siehe Tabelle 9).

⁶ Als atypisch schnelle Absolventen sind Studierende zu verstehen, die das Studium vollständig an der Betrachtungshochschule in weniger als der Regelstudienzeit absolvieren (selten) oder die sich Studienleistungen von anderen Hochschulen an der Betrachtungshochschule anrechnen lassen (häufig). Studiengangswechsler sind hierunter jedoch nicht zu verstehen, da die Anzahl der Semester die der Studierende an der Betrachtungshochschule immatrikuliert ist betrachtet wird, und nicht die Semester innerhalb eines Studiengangs.

4. Zwischenfazit: Das Früherkennungssystem FragSte

In der vorliegenden Studie wird ein Prognosesystem zur Früherkennung von Studienabbrüchen entwickelt und getestet. Das Prognosesystem basiert auf administrativen Daten nach §3 HStatG. Diese Studierendendaten liegen an allen deutschen Hochschulen vor. Das System ist übertragbar und an allen deutschen Hochschulen einsetzbar.

Neben der Verwendung traditioneller Regressionsanalysen werden maschinelle Lernalgorithmen zur Vorhersage von Abbrüchen angewendet. Besonderer Vorteil dieser Verfahren ist, dass sie nicht auf einen komplexen Modellaufbau angewiesen sind und sich durch regelmäßig wiederholte Lernphasen an Entwicklungen innerhalb der Hochschule anpassen. Allerdings zeigt sich auch, dass in Abhängigkeit von der Studiendauer der Hochschule teils erhebliche Beobachtungszeiträume ergeben, bis eine Immatrikulationskohorte eine ausreichend geringe Verbleibquote aufweist, um sie gut als Lerndaten verwenden zu können.

Anstatt aus einer Reihe von Verfahren für alle Hochschulen und Prognosezeitpunkte ein Verfahren auszuwählen, wird der AdaBoost-Algorithmus verwendet. Mittels des AdaBoost werden die Prognosen verschiedener Verfahren kombiniert und zu einem Ergebnis zusammengefügt, das mindestens gegen das jeweils genaueste Verfahren konvergiert. Dies reduziert die Nachteile, die mit der Anwendung einer einzelnen Methode verbunden sind, sowie die Nachteile, die sich aus der Heterogenität der Studiengänge und der Zusammensetzung der Studierenden an den verschiedenen Hochschulen ergeben.

Im vorliegenden Beitrag wird das Frühwarnsystem anhand der Studierendendaten einer staatlichen und privaten, gebührenfinanzierten Universität entwickelt und getestet. Trotz der großen Unterschiede beider Hochschultypen ist es möglich, Studienabbrecher an beiden Hochschulen zu prognostizieren. Durch die vorangegangene Lernphase berücksichtigt das System selbstjustierend die Besonderheiten einer Hochschule.

Die im Frühwarnsystem FragSte genutzte Datenselektion steht im Spannungsfeld zwischen Potential und Aufwand. Die Verwendung der nach §3 HStatG normierten Studierendendaten sichert die Übertragbarkeit des Frühwarnsystems auf andere deut-

sche Hochschulen und reduziert den Erhebungsaufwand. Es werden nur bereits vorhandene Studierendendaten genutzt und es müssen keine neuen Informationen beispielsweise im Rahmen einer Vollerhebung erhoben werden. Allerdings limitiert diese Datenselektion auch das Frühwarnsystem. Die nach §3 HStatG vorliegenden Informationen sind zum Zwecke der Studierendenverwaltung festgelegt worden, nicht zur Abbruchsprädiktion. Von den nach herrschender Meinung sieben Determinanten des Studienabbruchs: „*Leistungsanforderung, Finanzen, Prüfungsversagen, Motivation, Studienbedingungen, berufliche Neuorientierung und Krankheit*“ (Heublein et al., 2017) sind sechs Determinanten in den Studierendendaten nicht enthalten – lediglich das Prüfungsversagen ist abbildbar. Trotzdem sind Studienabbrüche bereits früh prognostizierbar und die Genauigkeit der Ergebnisse verbessert sich mit zunehmend späteren Prognosezeitpunkt im Studium.

Die Analyse zeigt, dass der Zeitpunkt von großer Bedeutung bei der Vorhersagegenauigkeit des Frühwarnsystems ist. Ein Früherkennungssystem, das nur demographische Daten verwendet, die bei der Einschreibung verfügbar sind, prognostiziert bereits 66% der Studienabbrecher in der UNI; die Vorhersagegenauigkeit steigt nach dem vierten Semester auf 80%. Die entsprechenden Zahlen für die FH betragen 50% zum Zeitpunkt der Immatrikulation und 83% nach dem vierten Semester.

Basierend auf dem Frühwarnsystem FragSte können abbruchgefährdete Studierende bereits früh im Studium an deutschen Hochschulen identifiziert werden. Damit ist es möglich Unterstützung zielgerichtet anzubieten um Studienabbrüche zu verhindern oder bei Abbruchentscheidung des Studierenden zu assistieren.

III Gegenüberstellung verschiedener Arten von Frühwarnsystemen

Plant eine Hochschule ein Frühwarnsystem einzuführen, ist zunächst zu entscheiden, welche Informationen über die Studierenden und die Studienverläufe für eine gute Prognose des Studienerfolgs erforderlich sind. Dabei kann zum einen auf Befragungsdaten zurückgegriffen werden. Diese sind wichtig, um die Gründe für einen intendierten Studienabbruch zu verstehen, sie sind aber weniger geeignet, um Studienabbruchgefährdete einer Hochschule zu identifizieren, zu beraten oder zu unterstützen (Georg, 2008, S. 202). Zum anderen können auch administrative Studierendendaten, die nach HStatG erhoben werden müssen, für Frühwarnsysteme eingesetzt werden. Entscheidet sich eine Hochschule für die Einführung eines auf administrativen Daten beruhenden Frühwarnsystems, ist zunächst zu prüfen, ob Daten ehemaliger Studierender (*labeled data*) vorliegen und verwendet werden können. Liegen Daten von ehemaligen Studierenden nicht vor, müssen die Merkmale und Schwellenwerte eines Frühwarnsystems hilfsweise anders bestimmt werden. Üblicherweise werden hierfür die erreichten ECTS-Punkte pro Semester und die Anzahl nicht bestandener Prüfungen pro Semester betrachtet und es wird geprüft, ob zuvor definierte Schwellenwerte überschritten werden. Die dafür festzusetzenden Kriterien können sowohl aus empirischen Abbruchsstudien als auch aus Empfehlungen zum Studienverlauf abgeleitet werden.

Unabhängig davon, ob Daten ehemaliger Studierender vorliegen, ist in einem zweiten Schritt abzuwägen, welche Informationen über Studierende und ihre Studienverläufe zur Abbruchprognose genutzt werden sollen. Grundsätzlich besteht dabei ein Trade-off zwischen der Prognosegenauigkeit und den Datenanforderungen. Es gilt, je komplexer die verwendeten Informationen und die eingesetzten Verfahren, desto präziser sind die Ergebnisse. Diese Genauigkeit geht aber auch mit einem höheren wiederkehrenden Aufwand für die Datenerhebung und Datenaufbereitung einher. Für die Beurteilung eines Frühwarnsystems kann die Prognosegenauigkeit damit nicht alleiniges Kriterium sein. Ein besonders genaues, aber datenintensives Frühwarnsystem wird möglicherweise an den Hochschulen nicht eingesetzt, wenn es mit einem sehr hohen und wiederkehrenden Aufwand verbunden ist.

In dem vorliegenden Teil III werden zwei alternative, für deutsche Hochschulen erstellte, Frühwarnsysteme vorgestellt, deren Konzeption veröffentlicht wurde. Beide Veröffentlichungen erschienen erst nach Konzeption von FragSte. Die zwei Frühwarnsysteme sind:

1. SVM (Westerholt et al., 2018) und
2. PASST?! (Schulze-Stocker et al., 2017)

Im Folgenden werden SVM und PASST?! mit dem in Teil II entwickelten Frühwarnsystem FragSte verglichen. Dabei werden die Prognosegenauigkeit, der Zeitpunkt der Prognose und die Datenanforderung als Kriterien angeführt. Wesentliche Ergebnisse des Teils III sind wie folgt veröffentlicht worden:

Berens, J. & Schneider, K. (2019) Drohender Studienabbruch: Wie gut sind Frühwarnsysteme. *Qualität in der Wissenschaft (QIW)*, 13 (3+4), 102-08.

1. Frühwarnsystem PASST?!

Das Frühwarnsystem PASST?! nutzt ausschließlich einzelne Leistungskennzahlen. Für jede betrachtete Kennzahl wird ein Schwellenwert festgesetzt, der abbruchgefährdete Studierende identifiziert. Dabei genügt es, wenn ein Schwellenwert überschritten ist.

Das erste an einer deutschen Hochschule flächendeckend eingesetzte Frühwarnsystem PASST?! der TU Dresden arbeitet mit fünf Leistungsindikatoren (Schulze-Stocker et al., 2017). Als abbruchgefährdet gelten Studierende, die

- 1.) im ersten Semester nicht mindestens zwei Prüfungsleistungen bestanden haben,
- 2.) in den vorangegangenen zwei Semestern in der Summe weniger als 30 Leistungspunkte erbracht haben,
- 3.) eine erste Wiederholungsprüfung nicht bestanden haben und damit vor der zweiten und damit letzten Wiederholungsprüfung stehen,
- 4.) im vorangegangenen Semester mindestens drei Mal von Prüfungsleistungen zurückgetreten sind bzw. haben sich drei Mal abgemeldet haben oder
- 5.) die Regelstudienzeit um zwei oder mehr Semester überschreiten.

Da ausschließlich Leistungsindikatoren eingesetzt werden, ist eine Warnung oder Intervention erst nach Bekanntgabe aller Prüfungsergebnisse des ersten Semesters möglich. Der früheste Zeitpunkt ist damit etwa Mitte des zweiten Semesters. Während Merkmal 1 nur im zweiten Semester für das erste Semester überprüft werden kann, wird Merkmal 2 ab dem vollendeten zweiten Semester geprüft. Merkmal 3 kann – sofern Prüfungen nur einmal im Semester angeboten werden - erstmalig nach dem vollendeten dritten Semester erfüllt werden. Merkmal 5 kann frühestens im neunten Semester zur Warnung führen.

Jedes Merkmal wird an der TU Dresden separat überprüft und behandelt. Sollte ein Studierender mehrere Merkmale erfüllen, werden auch mehrere Warnungen per Email versandt. Von insgesamt 4.464 zwischen 2016 und 2018 verschickten Emails, war zu

- 36,76% Merkmal 1,
- 31,99% Merkmal 2,
- 9,72% Merkmal 3,
- 15,43% Merkmal 4 und
- 6,09% Merkmal 5

der auslösende Grund (Schulze-Stocker et al., 2018).

Da ein Frühwarnsystem getestet und verglichen werden soll, das erlaubt, Hilfe möglichst früh anzubieten, wird Merkmal 5 im Folgenden nicht weiter betrachtet.

2. Frühwarnsystem SVM

Auch das Studienverlaufsmonitoring (SVM) der RWTH Aachen und Hochschule Niederrhein (Westerholt et al., 2018, S. 17) nutzt ausschließlich Leistungsdaten. Hier werden Studierende zu einer Beratung eingeladen, wenn sie eine Wiederholungsprüfung nicht bestanden haben oder weniger als 30% der im Semester vorgesehenen ECTS Punkte erwerben oder die Regelstudienzeit um das 1,5fache überschreiten. An der RWTH werden zusätzlich Studierende eingeladen, die zu den 10% der Notenbesten gehörten oder in einem MINT-Studiengang im 1. und 2. Fachsemester eingeschrieben sind. Da SVM nicht ausschließlich abbruchgefährdete Studierende adressiert, ist ein Vergleich von SVM mit FragSte und PASST?! nur eingeschränkt möglich. Im Folgenden werden daher nur die ersten beiden Bedingungen berücksichtigt.

3. Die Prognosegüte der Frühwarnsysteme

Im Folgenden werden die Frühwarnsysteme PASST?! und SVM mit FragSte verglichen. Dabei werden die drei System auf die UNI und FH angewendet. Anders als im vorherigen Teil II werden die Ergebnisse an beiden Hochschulen für jeweils drei Testkohorten (2010-2012) ausgegeben um den in Teil II, Abschnitt 3.6 dargestellten Einfluss der Verbleibquote (siehe Tabelle 9) im Systemvergleich verschiedener Früherkennungen zu berücksichtigen.

3.1. Prognosegenauigkeiten von FragSte

In Teil II, Kapitel 3 ist die Prognosegenauigkeit detailliert dargestellt. Um aber den Einfluss unterschiedlicher Frühwarnsysteme auf unterschiedliche Verbleibquoten zu prüfen (siehe Teil II, Abschnitt 3.5), werden im Folgenden die Ergebnisse für die Immatrikulationskohorten 2010, 2011 und 2012 dargestellt. Dabei wird eine Prognose zum Zeitpunkt der Immatrikulation (Sem 0) sowie der Semester 1-4 erstellt. Es werden zu jedem Zeitpunkt alle zum entsprechenden Zeitpunkt an den Hochschulen verfügbaren Informationen verwendet. Beispielsweise kann die Prognose nach dem ersten Semester auf alle Informationen zum Zeitpunkt der Immatrikulation und die akademischen Leistungen des ersten Semesters zurückgreifen.

Tabelle 29: Prognosegüte FragSte

Sem:	2010			2011			2012			
	Acc	Rec	Prec	Acc	Rec	Prec	Acc	Rec	Prec	
UNI	0	72,4	70,8	71,1	74,8	77,3	77,4	77,4	82,7	83,0
	1	78,2	73,5	73,6	81,5	82,0	82,2	84,3	87,1	87,2
	2	81,5	73,6	73,7	83,9	81,7	81,9	87,0	87,3	87,4
	3	87,6	80,4	80,6	86,8	83,6	83,7	89,1	88,1	88,3
	4	89,0	78,6	78,8	89,4	83,9	84,1	90,4	86,9	87,1
FH	0	64,8	46,3	46,7	65,2	45,3	45,5	67,2	50,1	50,5
	1	80,9	63,8	63,8	80,0	61,3	61,9	81,6	67,8	67,8
	2	86,0	64,3	64,3	86,5	67,9	67,9	86,8	71,4	71,7
	3	89,1	64,8	64,8	89,7	70,7	70,7	91,5	78,3	78,3
	4	91,7	65,5	65,5	93,4	77,6	77,6	93,9	79,8	80,2

Hinweis: Alle Werte in Prozent. Acc=Accuracy; Rec=Recall; Prec=Precision. Die Semesterzahl bezeichnet die Anzahl vollendeter Semester. Das Folgesemester stellt den Prognosezeitpunkt dar.

Die Ergebnisse in Tabelle 29 zeigen sowohl Gemeinsamkeiten als auch Unterschiede zwischen beiden Hochschulen. Prognosen zu späteren Zeitpunkten im Studium sind zunehmend genauer, da zum einen Leistungsdaten verfügbar sind und zum anderen die Testkohorten homogener werden, weil frühe Studienabbrecher bereits ausgeschieden sind. Dies trifft insbesondere auf die private FH zu, denn hier finden

Studienabbrüche in der Regel in den ersten beiden Semestern und damit früher als an der UNI statt. Ein möglicher Grund hierfür könnten die Studiengebühren an der privaten FH sein (vgl. Teil II, Abschnitt 3.6).

Die Prognose von Studienabbrechern im ersten Semester, wenn nur Informationen zum Zeitpunkt der Immatrikulation vorliegen, ist für die FH deutlich ungenauer als für die UNI. An beiden Hochschulen zeigt sich aber auch der hohe prognostische Wert akademischer Leistungen. Wenn erstmalig Prüfungsergebnisse – auch in Form von nicht bestandenen Leistungen – verfügbar werden, erhöht das die Prognosegenauigkeit deutlich.

3.2. Prognosegenauigkeiten von PASST?!

Nach dem ersten Semester liefert das Frühwarnsystem bereits gute Ergebnisse und mit zunehmender Semesterzahl steigt, wie auch in FragSte, die Prognosegüte. Tabelle 30 zeigt die Ergebnisse von PASST?! angewendet auf die in Teil II, Abschnitt 2.3 und 2.4 beschriebenen Studierendendaten beider Hochschulen.

Tabelle 30: Prognosegüte PASST?!

Sem:	2010			2011			2012			
	Acc	Rec	Prec	Acc	Rec	Prec	Acc	Rec	Prec	
UNI	0	-	-	-	-	-	-	-	-	
	1	75,0	61,7	73,3	74,1	60,7	84,8	75,1	63,6	93,5
	2	75,5	83,8	61,0	81,2	86,2	75,2	86,7	88,6	86,0
	3	79,1	89,6	62,1	84,3	89,3	76,1	88,5	89,5	86,4
	4	80,7	91,7	58,2	85,3	86,9	73,8	87,8	88,9	80,4
FH	0	-	-	-	-	-	-	-	-	
	1	77,5	37,5	61,9	77,1	39,7	59,2	77,9	42,8	68,0
	2	71,5	74,6	38,3	75,0	75,5	44,5	72,3	77,7	44,5
	3	86,3	64,2	54,8	88,1	67,6	65,8	88,7	69,6	71,7
	4	90,5	70,6	59,2	90,7	76,3	66,3	93,7	76,4	81,4

Hinweis: Alle Werte in Prozent. Acc=Accuracy; Rec=Recall; Prec=Precision. Die Semesterzahl bezeichnet die Anzahl vollendeter Semester. Das Folgesemester stellt den Prognosezeitpunkt dar.

Zwischen dem ersten und zweiten Semester wechselt das Verhältnis von Recall und Precision. Während basierend auf den Leistungsdaten des ersten Semesters mit $Recall < Precision$ zu selten gewarnt wird, wird nach dem zweiten Semester mit $Recall > Precision$ eher zu häufig gewarnt. Dieser Wechsel ist den Kriterien, die eine Warnung auslösen, geschuldet. Nach dem zweiten Semester wird nicht mehr die Anzahl bestandener Prüfungen überprüft, sondern ob im ersten und zweiten Semester in Summe mindestens 30 ECTS Punkte erreicht wurden. Das hat Konsequenzen für die Anzahl der Identifikationen und somit für die Prognosegüte des Frühwarnsystems. Tabelle 31 zeigt für die ersten vier vollendeten Semester die Zahl der Abbrecher und

Identifikationen. Nur wenn die Anzahl der Abbrecher der Zahl der Identifikationen entspricht, ergibt sich $Recall = Precision$.

Tabelle 31: Identifikationsquoten PASST?!

HS	Kohorte	Sem	Absolventen	Abbrecher	Identifikationen	Identifikationsquote
UNI	2010	1	1039	726	611	84,16%
UNI	2010	2	1027	555	762	137,30%
UNI	2010	3	1015	479	691	144,26%
UNI	2010	4	992	349	550	157,59%
UNI	2011	1	1233	1319	943	71,49%
UNI	2011	2	1171	934	1071	114,67%
UNI	2011	3	1157	783	919	117,37%
UNI	2011	4	1137	572	673	117,66%
UNI	2012	1	879	1382	940	68,02%
UNI	2012	2	837	888	915	103,04%
UNI	2012	3	829	724	750	103,59%
UNI	2012	4	806	479	530	110,65%
FH	2010	1	874	312	189	60,58%
FH	2010	2	874	213	415	194,84%
FH	2010	3	871	159	186	116,98%
FH	2010	4	864	119	142	119,33%
FH	2011	1	889	315	211	66,98%
FH	2011	2	885	237	402	169,62%
FH	2011	3	883	188	193	102,66%
FH	2011	4	874	152	175	115,13%
FH	2012	1	995	397	250	62,97%
FH	2012	2	993	301	526	174,75%
FH	2012	3	988	240	233	97,08%
FH	2012	4	980	178	167	93,82%

Auch für das dritte und vierte Semester neigt das System für die hier betrachteten Hochschulen zu einer Überidentifikation. Das heißt, es werden viele künftige Studienabbrecher korrekt identifiziert, aber es werden zu viele Studierende als mögliche Abbrecher angesprochen. Das ist dann unproblematisch, wenn auch künftige Absolventen während des Studiums eine Phase der Unsicherheit durchlaufen, in der Unterstützung hilfreich ist. Wenn aber systematisch zu häufig gewarnt wird, kann das die Akzeptanz und Glaubwürdigkeit des Früherkennungssystems gefährden.

PASST?! arbeitet mit fünf Indikatoren, bei denen es zur Identifikation genügt, wenn ein Merkmal erfüllt ist (siehe Teil III, Kapitel 1). Tabelle 32 stellt die Anzahl der Identifikationen, der korrekt Identifizierten sowie der Gütemaße Accuracy, Recall und Precision für die Merkmale 1 und 2 dar. Gleiches wird in den Tabelle 33 und Tabelle 34 für die Merkmale 3 und 4 dargestellt.

Tabelle 32: Genauigkeit der Merkmale 1 (<2 Prüfungen im ersten Semester) und 2 (<30CP in zwei Semestern) des Studienfortschritts in PASST?!

Merkm al	HS	Kohorte	Sem	Abbrecher	Identifi- kationen	Korrekt Identifizierte	Accuracy	Recall	Precision
1	UNI	2010	1	726	607	446	75.01%	61.43%	73.48%
3	UNI	2010	2	555	677	440	77.75%	79.28%	64.99%
3	UNI	2010	3	479	590	407	82.93%	84.97%	68.98%
3	UNI	2010	4	349	446	293	84.41%	83.95%	65.70%
1	UNI	2011	1	1319	938	796	73.94%	60.35%	84.86%
3	UNI	2011	2	934	1003	782	82.28%	83.73%	77.97%
3	UNI	2011	3	783	820	672	86.65%	85.82%	81.95%
3	UNI	2011	4	572	568	461	87.24%	80.59%	81.16%
1	UNI	2012	1	1382	938	878	75.06%	63.53%	93.60%
3	UNI	2012	2	888	867	772	87.77%	86.94%	89.04%
3	UNI	2012	3	724	684	624	89.70%	86.19%	91.23%
3	UNI	2012	4	479	455	405	90.35%	84.55%	89.01%
1	FH	2010	1	312	177	108	76.98%	34.62%	61.02%
3	FH	2010	2	213	291	131	77.74%	61.50%	45.02%
3	FH	2010	3	159	182	101	86.50%	63.52%	55.49%
3	FH	2010	4	119	126	79	91.15%	66.39%	62.70%
1	FH	2011	1	315	205	123	77.24%	39.05%	60.00%
3	FH	2011	2	237	310	159	79.59%	67.09%	51.29%
3	FH	2011	3	188	179	122	88.52%	64.89%	68.16%
3	FH	2011	4	152	156	111	91.62%	73.03%	71.15%
1	FH	2012	1	397	245	167	77.87%	42.07%	68.16%
3	FH	2012	2	301	384	194	77.05%	64.45%	50.52%
3	FH	2012	3	240	226	165	88.93%	68.75%	73.01%
3	FH	2012	4	178	157	132	93.87%	74.16%	84.08%

An der UNI werden Prüfungen überwiegend einmal im Semester angeboten, an der FH werden Klausuren zweimal pro Semester und alle anderen Prüfungsformen einmal im Semester angeboten. Daher ist es nur an der FH und auch nur für Klausuren möglich im zweiten Semester bereits den dritten Versuch anzutreten. Daher werden in Tabelle 33 für die UNI nur die Semester 3 und 4 und an der FH die Semester 2 bis 4 ausgewiesen.

Tabelle 33: Merkmal 3 - Genauigkeit des zweiten Wiederholungsversuchs in PASST?!

HS	Ko- horte	Sem	Abbre- cher	Identifi- kationen	Korrekt Identifizierte	Accuracy	Recall	Precision
UNI	2010	1	726					
UNI	2010	2	555					
UNI	2010	3	479	198	106	68.88%	22.13%	53.54%
UNI	2010	4	349	170	79	73.08%	22.64%	46.47%
UNI	2011	1	1319					
UNI	2011	2	934					
UNI	2011	3	783	335	239	67.01%	30.52%	71.34%
UNI	2011	4	572	236	169	72.50%	29.55%	71.61%
UNI	2012	1	1382					
UNI	2012	2	888					
UNI	2012	3	724	231	181	61.82%	25.00%	78.35%
UNI	2012	4	479	147	92	65.60%	19.21%	62.59%
FH	2010	1	312					
FH	2010	2	213	182	60	74,70%	28,17%	32,97%
FH	2010	3	159	221	50	72.82%	31.45%	22.62%
FH	2010	4	119	186	36	76.30%	30.25%	19.35%
FH	2011	1	315					
FH	2011	2	237	132	50	76,02%	21,10%	37,88%
FH	2011	3	188	178	51	75.35%	27.13%	28.65%
FH	2011	4	152	160	38	77.00%	25.00%	23.75%
FH	2012	1	397					
FH	2012	2	301	204	80	73,34%	26,58%	39,22%
FH	2012	3	240	243	76	73.05%	31.67%	31.28%
FH	2012	4	178	220	52	74.61%	29.21%	23.64%

Merkmal 3 erreicht nennenswerte Anteile der Studienabbrecher nicht und auch der Anteil der richtigen Identifikationen ist gering. Es werden etwa ein Fünftel bis ein Drittel aller Studienabbrecher korrekt identifiziert. Zudem ist der Zeitpunkt im 3. oder 4. Semester sehr spät.

Im Gegensatz dazu ist es durch Merkmal 4, Anzahl der Rücktritte im vorangehenden Semester, möglich alle vier vollendeten Semester zu betrachten. Tabelle 34 stellt die Genauigkeit dar.

Tabelle 34: Merkmal 4 - Genauigkeit des zwei oder mehrmaligen Rücktritts in PASST?!

HS	Kohorte	Sem	Abbrecher	Identifikationen	Korrekt Identifizierte	Accuracy	Recall	Precision
UNI	2010	1	726	11	8	59.15%	1.10%	72.73%
UNI	2010	2	555	18	15	65.68%	2.70%	83.33%
UNI	2010	3	479	31	22	68.81%	4.59%	70.97%
UNI	2010	4	349	26	18	74.72%	5.16%	69.23%
UNI	2011	1	1319	24	21	49.02%	1.59%	87.50%
UNI	2011	2	934	57	45	57.20%	4.82%	78.95%
UNI	2011	3	783	50	39	61.08%	4.98%	78.00%
UNI	2011	4	572	48	32	67.47%	5.59%	66.67%
UNI	2012	1	1382	18	16	39.50%	1.16%	88.89%
UNI	2012	2	888	27	24	49.74%	2.70%	88.89%
UNI	2012	3	724	47	36	54.99%	4.97%	76.60%
UNI	2012	4	479	34	22	63.50%	4.59%	64.71%
FH	2010	1	312	19	16	74.79%	5.13%	84.21%
FH	2010	2	213	24	16	81.14%	7.51%	66.67%
FH	2010	3	159	11	6	84.66%	3.77%	54.55%
FH	2010	4	119	24	9	87.28%	7.56%	37.50%
FH	2011	1	315	14	10	74.34%	3.17%	71.43%
FH	2011	2	237	16	10	79.23%	4.22%	62.50%
FH	2011	3	188	24	15	83.01%	7.98%	62.50%
FH	2011	4	152	26	12	84.99%	7.89%	46.15%
FH	2012	1	397	13	10	71.98%	2.52%	76.92%
FH	2012	2	301	10	7	77.05%	2.33%	70.00%
FH	2012	3	240	186	161	91.53%	67.08%	86.56%
FH	2012	4	178	157	122	92.14%	68.54%	77.71%

Wenn ein Studierender durch Merkmal 4 als Abbrecher klassifiziert wurde, dann ist das im ersten Semester in 72% bis 89% der Fälle korrekt (Precision). In den Folge semestern nimmt die Precision an beiden Hochschulen stetig ab. Jedoch ist Anzahl der Identifikationen nur sehr gering. An beiden Hochschulen melden sich nur wenige Studierende mehrmals wieder von den Prüfungen ab.

3.3. Prognosegenauigkeiten von SVM

Auch bei SVM steigt die Prognosegüte mit der Semesteranzahl. Durch die in SVM gewählten Schwellenwerte sind Recall und Precision für die UNI allerdings näher beieinander als bei PASST?!. Im Vergleich beider Hochschulen zeigt sich aber, dass die in SVM gewählten Schwellenwerte deutlich besser abbruchgefährdete Studierenden der UNI als der FH identifizieren. Für die FH neigt SVM zu einer im Studienverlauf zunehmenden Überidentifikation.

Tabelle 35: Prognosegüte SVM

Sem:	2010			2011			2012		
	Acc	Rec	Prec	Acc	Rec	Prec	Acc	Rec	Prec
UNI 0	-	-	-	-	-	-	-	-	-
UNI 1	74,1	67,9	68,7	76,3	70,6	81,1	81,9	78,2	90,9
UNI 2	79,5	77,3	68,4	82,7	79,9	80,8	85,6	80,5	90,5
UNI 3	82,3	84,1	68,2	85,0	80,7	81,9	86,0	81,8	87,4
UNI 4	83,1	83,1	63,5	87,0	80,2	80,8	88,0	81,6	85,6
FH 0	-	-	-	-	-	-	-	-	-
FH 1	71,6	57,7	46,8	74,8	56,2	51,6	75,9	62,0	57,1
FH 2	82,9	60,6	55,8	84,5	61,6	63,8	84,5	67,1	66,4
FH 3	82,1	73,0	45,1	85,1	68,1	56,1	86,4	79,2	61,9
FH 4	84,9	74,8	43,0	86,5	78,3	52,9	85,9	82,0	52,7

Hinweis: Alle Werte in Prozent. Acc=Accuracy; Rec=Recall; Prec=Precision. Die Semesterzahl bezeichnet die Anzahl vollendeter Semester. Das Folgesemester stellt den Prognosezeitpunkt dar.

Anders als bei PASST?! kommt es nach dem ersten Semester nicht zu einem Merkmalswechsel. SVM überprüft als Indikator des Studienfortschritts die Anzahl der im vorangegangenen Semester erzielten CP. Das zeigt sich auch in der Identifikationsquote des ersten und zweiten Semesters.

Tabelle 36: Identifikationsquoten SVM

HS	Ko-horte	Sem	Absolventen	Abbre-cher	Identifikationen	Identifikati-onsquote
UNI	2010	1	1039	726	718	98,90%
UNI	2010	2	1027	555	627	112,97%
UNI	2010	3	1015	479	591	123,38%
UNI	2010	4	992	349	457	130,95%
UNI	2011	1	1233	1319	1148	87,04%
UNI	2011	2	1171	934	923	98,82%
UNI	2011	3	1157	783	772	98,60%
UNI	2011	4	1137	572	568	99,30%
UNI	2012	1	879	1382	1189	86,03%
UNI	2012	2	837	888	790	88,96%
UNI	2012	3	829	724	677	93,51%
UNI	2012	4	806	479	457	95,41%
FH	2010	1	874	312	385	123,40%
FH	2010	2	874	213	231	108,45%
FH	2010	3	871	159	257	161,64%
FH	2010	4	864	119	207	173,95%
FH	2011	1	889	315	343	108,89%
FH	2011	2	885	237	229	96,62%
FH	2011	3	883	188	228	121,28%
FH	2011	4	874	152	225	148,03%
FH	2012	1	995	397	431	108,56%
FH	2012	2	993	301	304	101,00%
FH	2012	3	988	240	307	127,92%
FH	2012	4	980	178	277	155,62%

Tabelle 37: Merkmal 1 - Genauigkeit des Studienfortschritts (>9CP) in SVM

HS	Kohorte	Sem	Abbrecher	Identifi- kationen	Korrekt Identifizierte	Accuracy	Recall	Precision
UNI	2010	1	726	718	493	74.05%	67.91%	68.66%
UNI	2010	2	555	560	408	81.10%	73.51%	72.86%
UNI	2010	3	479	508	375	84.14%	78.29%	73.82%
UNI	2010	4	349	363	256	85.09%	73.35%	70.52%
UNI	2011	1	1319	1148	931	76.29%	70.58%	81.10%
UNI	2011	2	934	885	726	82.57%	77.73%	82.03%
UNI	2011	3	783	677	581	84.64%	74.20%	85.82%
UNI	2011	4	572	506	429	87.13%	75.00%	84.78%
UNI	2012	1	1382	1189	1081	81.91%	78.22%	90.92%
UNI	2012	2	888	755	698	85.68%	78.60%	92.45%
UNI	2012	3	724	618	562	85.96%	77.62%	90.94%
UNI	2012	4	479	415	371	88.17%	77.45%	89.40%
FH	2010	1	312	385	180	71.59%	57.69%	46.75%
FH	2010	2	213	147	106	86.38%	49.77%	72.11%
FH	2010	3	159	147	93	88.35%	58.49%	63.27%
FH	2010	4	119	111	71	91.05%	59.66%	63.96%
FH	2011	1	315	343	177	74.75%	56.19%	51.60%
FH	2011	2	237	171	128	86.45%	54.01%	74.85%
FH	2011	3	188	146	110	89.36%	58.51%	75.34%
FH	2011	4	152	133	101	91.91%	66.45%	75.94%
FH	2012	1	397	431	246	75.86%	61.96%	57.08%
FH	2012	2	301	215	174	87.02%	57.81%	80.93%
FH	2012	3	240	186	161	91.53%	67.08%	86.56%
FH	2012	4	178	157	122	92.14%	68.54%	77.71%

Im Vergleich der Prognosegenauigkeit von SVM in Tabelle 35 mit der Prognosegenauigkeit basierend auf Merkmal 1 in Tabelle 37 zeigen sich insbesondere Studienbeginn nur geringfügige Unterschiede. Erst in höheren Semestern (FH: Semester 2 | UNI: Semester 3) kann der zweite Abbruchsindikator, das erstmalige Nichtbestehen einer Wiederholungsprüfung, die Prognose verbessern. Der zweite Abbruchsindikator ist identisch mit Merkmal 3 in PASST?! (vgl. Tabelle 33). Damit stellt SVM zu Studienbeginn ein Single-Prädiktor-Verfahren dar.

3.4. Diskussion der Ergebnisse

Die Ergebnisse in Tabelle 30 und Tabelle 35 zeigen auch für PASST?! und SVM, dass die Ergebnisse für Immatrikulationskohorten jüngerer Jahrgänge besser als die Ergebnisse der älteren Kohorten sind. Weiterhin ergeben sich im Vergleich zu Tabelle 29 in FragSte bessere Gütemaße als für PASST?! und SVM. Frühwarnsysteme, die auf einzelnen Leistungsindikatoren beruhen, neigen zu Fehlern, da die Anzahl der als abbruchgefährdet identifizierten Studierenden in der Regel von der Zahl tatsächlicher Abbrecher abweicht. Zudem verwenden PASST?! und SVM ausschließlich ausgewählte Leistungsdaten. Daher sind Prognosen im ersten Semester nicht möglich und die anderen in FragSte eingesetzten Informationen können die Prognose verbessern.

Bei allen Verfahren werden die Prognosen mit weiteren verfügbaren Leistungsdaten genauer. Doch auch nach Berücksichtigung der Prüfungsergebnisse des ersten Semesters zeigen die Frühwarnsysteme deutliche Unterschiede, die auf unterschiedliche Schwellenwerte zurückzuführen sind. Während in SVM geprüft wird, ob mehr als 30% der vorgesehen ECTS Punkte erreicht wurden, wird in PASST?! geprüft, ob mehr als zwei Prüfungen bestanden wurden. Das führt in PASST?! an beiden Hochschulen im Vergleich zu FragSte und SVM zu einem geringeren Recall. Sofern eine eher zurückhaltende Ansprache abbruchgefährdeter Studierender nicht erwünscht ist, zeigt SVM, dass bereits nach dem ersten Semester die erreichten ECTS-Punkte ein sehr guter Indikator sind.

Nach dem zweiten Semester wird in PASST?! geprüft, ob in den beiden vorangegangenen Semestern in Summe mindestens 30 ECTS Punkte erreicht wurden. Dieser Wechsel in den Kriterien führt von einer Unter- zu einer Überidentifikation, die jedoch nach den Semestern drei und vier nicht mehr zu beobachten ist. Dafür neigt SVM nach den Semestern drei und vier an der FH zur Überidentifikation - nicht aber an der UNI. Diese Ergebnisse sprechen dafür, dass die Wahl der Indikatoren und der Schwellenwerte erwartungsgemäß einen großen Einfluss auf die Prognosegüte hat, der auch von der betrachteten Hochschule abhängt.

SVM ist nahezu ein Single-Prädiktor Verfahren, mit dem adäquate Genauigkeiten erreicht werden können, sofern der Schwellenwert richtig gewählt ist. Im Vergleich von Tabelle 32 und Tabelle 37 zeigt sich, dass der Schwellenwert von weniger als 10 CP pro Semester (SVM) an der UNI bessere Ergebnisse erzielt, als ein Schwellenwert von durchschnittlich 15 CP über zwei Semester (PASST?!). An der FH hingegen ist der von PASST?! gewählte Schwellenwert erfolgreicher. Weiterhin sind die gewählten Schwellenwerte in unterschiedlichen Semestern unterschiedlich erfolgreich. D.h. die Merkmale eines Studienabbrechers verändern sich mit dem Betrachtungszeitpunkt. Das bedeutet, die Schwellenwerte müssen an die jeweilige Hochschule und den Betrachtungszeitpunkt angepasst werden. Das ist aber ohne Analyse historischer Daten nur schwer möglich.

Ein weiterer Nachteil besteht darin, dass vor dem zweiten Semester noch keine Analyse möglich ist und die Prognosen zu jedem Zeitpunkt im Studienverlauf weniger

genau sind als „erlernte“ Schwellenwerte. Diese Einschätzung beruht auf der Annahme, dass die Prognosegüte ein geeignetes Bewertungskriterium ist. Hochschulen können aber auch andere Ziele verfolgen, die eine systematische Über- oder Unteridentifikation rechtfertigen.

FragSte kann abbruchgefährdete Studierende zwar durchweg besser identifizieren als die beiden anderen Modelle, die Voraussetzung für den Einsatz von FragSte ist jedoch die Nutzung administrativer Studierendendaten aktueller und exmatrikulierter Studierender. Diese Bedingung ist ggf. nicht an allen Hochschulen erfüllt. Wenn diese Daten jedoch verwendet werden können, automatisieren und optimieren Methoden des maschinellen Lernens das Frühwarnsystem für jede Hochschule.

4. Zwischenfazit

In Teil III werden drei bereits an deutschen Hochschulen eingesetzte Prognosemodelle verglichen. Dabei ist kein Verfahren grundsätzlich überlegen, denn eine höhere Genauigkeit oder die Möglichkeit früher Prognosen gehen stets mit einem höheren Aufwand einher. Es zeigt sich aber, dass Modelle des maschinellen Lernens wie in FragSte unter Verwendung aller verfügbaren Informationen zu jedem Zeitpunkt im Verlauf des Studiums die beste Prognosegüte aufweisen. Auch bevor Leistungsdaten verfügbar sind, sind erste Prognosen basierend auf Informationen bei Immatrikulation möglich. Zudem ist das Modell flexibel und die Besonderheiten von Hochschulen, Studiengängen und sich im Zeitverlauf ändernden Bedingungen werden automatisch berücksichtigt. Wenn Daten ehemaliger Studierender verfügbar sind und weder die Kosten der Datenaufbereitung oder die Rechenzeit ein Hindernis darstellen, ist diese Form der Modellierung zu empfehlen. Sind Daten exmatrikulierter Kohorten jedoch nicht verfügbar, können Modelle wie PASST?! und SVM eine gute Alternative sein.

IV Prognoserelevanz einzelner Informationen

In der Konzeption von Frühwarnsystemen ist zu entscheiden, welche Informationen über die Studierenden und die Studienverläufe für eine gute Prognose des Studienerfolgs erforderlich sind. Grundsätzlich ist zwischen Prognosegenauigkeit und Datenanforderungen abzuwägen: je komplexer die verwendeten Informationen und die eingesetzten Verfahren, desto genauer sind die Prognosen. Diese Präzision geht aber auch mit einem höheren wiederkehrenden Aufwand für die Datenerhebung und Datenaufbereitung einher. Ein besonders genaues Frühwarnsystem wird möglicherweise an den Hochschulen nicht eingesetzt, wenn es mit einem sehr hohen Datenaufwand verbunden ist. Auch datenschutzrechtliche und ethische Bedenken können Argumente gegen hohe Datenanforderungen sein (Zafar et al., 2017). In wie weit die Verwendung demografischer Variablen als Ausdruck sozialer Verantwortung der Hochschule ethisch vertretbar ist oder als Eingriff in die studentische Selbstbestimmung grundsätzlich abzulehnen ist, ist nicht Gegenstand der vorliegenden Arbeit, aber es kann geprüft werden ob demographische Variablen die Leistung eines Frühwarnsystems erheblich verbessern oder ob sie verzichtbar sind.

In dem vorliegenden Teil IV wird geprüft, welche Informationen zu welchem Zeitpunkt im Studium einen besonders hohen Informationsgehalt haben und ob ein datensparsameres Frühwarnsystem mit wenigen Variablen vergleichbare Prognosegenauigkeiten erzielen kann. Es soll ein effizientes Prognosemodell entwickelt werden, das ohne hohe Anpassungskosten von Hochschulen für die Evaluation und die Allokation von Unterstützungsmaßnahmen eingesetzt werden kann.

Teil IV gliedert sich wie folgt: Zunächst werden in Kapitel 1 die Variablen kurz zusammengefasst und Verfahren beschrieben, die dazu geeignet sind den Informationsgehalt der eingesetzten Variablen zu bestimmen. Anschließend wird in Kapitel 2 der Informationsgehalt analysiert. Aufbauend darauf wird ein Frühwarnsystem entwickelt, das nicht mehr alle verfügbaren Studierendendaten verwendet, sondern nur Variablen mit einem hohen Informationsgehalt einsetzt. Abschließend wird die Prognosegenauigkeit datenreduzierter Frühwarnsysteme und eines reinen ECTS-Monitorings mit der Genauigkeit von FragSte verglichen. Eine Diskussion und Einordnung der Ergebnisse folgt in Kapitel 3.

Wesentliche Ergebnisse des Teil IV sind wie folgt veröffentlicht worden:

Schneider, K., Berens, J. & Burghoff, S. (2019) Drohende Studienabbrüche durch Frühwarnsysteme erkennen: Welche Informationen sind relevant? Zeitschrift für Erziehungswissenschaften, 22(1), 1121-46

1. Variablenbeschreibung und Methodenwahl

1.1. Variablenbeschreibung

Anspruch von FragSte ist es, weitestgehend automatisiert abbruchgefährdete Studierende früh zu erkennen und den Hochschulen Anhaltspunkte für mögliche Interventionen und Unterstützungsangebote zu bieten. Es wird dabei nur auf Studierenden-daten zurückgegriffen, die nach § 3 HStatG standardisiert und bundesweit zu erheben und zu speichern sind. Aus diesen können die für die Prognose eingesetzten Variablen konstruiert werden (vgl. Tabelle 17). Die Variablen können wie folgt gegliedert werden:

- **Demografische Variablen:** Diese Variablen geben Aufschluss über das Geschlecht, Alter, Herkunft, Migrationshintergrund und die Art der Krankenversicherung (gesetzlich oder privat).
- **Schulbildung:** Bekannt ist die Art und Note der Hochschulzugangsberechtigung. Bei ‚Art der Hochschulzugangsberechtigung‘ wird die allgemeine Hochschulreife nach der Schulform, an der diese erlangt wurde, unterschieden.
- **Frühere Hochschulbildung:** Aus der Anzahl der Hochschulsemester kann geschlossen werden, ob und wie lange die Studierenden bereits immatrikuliert waren.
- **Zeitpunkt der Einschreibung:** Es sind Informationen zur Wahl des Studiengangs oder der Studiengänge und Wahl der Studienform (Vollzeit oder berufsbegleitend) verfügbar. Durch den Zeitpunkt der Immatrikulation ist bekannt, wie frühzeitig Studierende sich für ein Studium an der Hochschule entschieden haben.
- **Akademische Leistungen:** Für jedes Semester werden die Credit Points (CP), die Noten, die Anzahl (nicht) bestandener Prüfungen und die Anzahl der Rücktritte /Abmeldungen erhoben. Zudem ist bekannt, ob die neun am stärksten mit dem Studienerfolg korrelierten Prüfungen (Schein1-9) bestanden wurden.

1.2. Information Gain

Da insbesondere bei der Verwendung demografischer Informationen auch ethische Fragen bedeutsam sein können (Zafar et al., 2017) und auch die Datenaufbereitung in der Praxis sehr aufwändig ist, ist zu prüfen, ob alle verfügbaren Informationen einen wichtigen Beitrag für ein Frühwarnsystem leisten oder ob die Liste der eingesetzten Variablen gekürzt werden kann. Für die Entwicklung von Maßnahmen gegen den Studienabbruch aber auch für weiterführende wissenschaftliche Studien zum besseren Verständnis der Gründe und des Prozesses kann die Auswahl relevanter Variablen hilfreich sein. Zur Bewertung des prognostischen Wertes einzelner Variablen wird im Folgenden das Konzept des Information Gain verwendet (Quinlan, 1986). Der Information Gain erfasst, wie sehr die Entropie der abhängigen Variablen durch eine unabhängige Variable reduziert wird. So könnte der Wert der Information über das Geschlecht der Studierenden für die Prognose des Studienabbruchs berechnet werden. Dazu wird zunächst die Entropie der abhängigen Variable (Abbruch: ja/nein) bestimmt und mit der bedingten Entropie verglichen, wenn eine erklärende Variable (z.B. das Geschlecht) berücksichtigt wird. Der Information Gain ist dann die Reduktion der Entropie der abhängigen Variablen durch die Berücksichtigung der erklärenden Variablen x :

$$\text{Information Gain}(y, x) = H(y) - H(y|x)$$

Das folgende Beispiel erläutert die Berechnung des Information Gain für die Variable Geschlecht. Männliche und weibliche Studierende verteilen sich auf Absolventen und Abbrecher wie folgt:

Tabelle 38. Beispielrechnung Information Gain

		Absolvent/in	Abbrecher/in	gesamt
Geschlecht	männlich	2.990	7.313	10.303
	weiblich	4.693	7.352	12.045
Gesamt		7.683	14.665	22.348

Im vorliegenden Anwendungsfall ist der Studienabbruch (ja/nein) die abhängige Variable. Das Maß für die Entropie, H , ist

$$H(y) = -\sum_i p(y_i) \cdot \log_2 p(y_i),$$

wobei y_i die Ausprägungen der Variablen y sind und $p(y_i)$ ist die relative Häufigkeit, mit der die Variable y den Wert y_i annimmt.

Damit ist die Entropie H

$$H(\text{Abbruch}) = H(7683; 14665) = -(0,34 \log_2 0,34) - (0,66 \log_2 0,66) = 0,928$$

Wird die Information zum Geschlecht berücksichtigt, ergibt sich eine bedingte Entropie:

$H(\text{Abbruch}|\text{Geschlecht})$

$$\begin{aligned} &= P(\text{männlich}) \cdot H(\text{Abbruch}|\text{männlich}) + P(\text{weiblich}) \cdot H(\text{Abbruch}|\text{weiblich}) \\ &= (10303/22348) \cdot H(2990; 7313) + (12045/22348) \cdot H(4693; 7352) \\ &= 0,461 \cdot 0,869 + 0,539 \cdot 0,965 = 0,920 \end{aligned}$$

Aus der Differenz der Entropie der abhängigen Variable und der bedingten Entropie ergibt sich der Information Gain als Differenz beider Werte, im vorliegenden Beispiel also 0,008. Der Informationsgehalt der Variable Geschlecht ist also tendenziell gering.

Nach Ermittlung des Information Gain für alle Variablen in allen Semestern wird eine Rangliste der unabhängigen Variablen nach Wichtigkeit erstellt (Quinlan, 1986). Diese gibt Aufschluss darüber, welche Informationen einen hohen prognostischen Wert haben. Die Werte des Information Gain liegen zwischen 0 (kein prognostischer Wert) und 1 (aus der betrachteten Variable kann die Ergebnisvariable direkt bestimmt werden).

1.3. Gain Ratio

Der Gain Ratio basiert auf dem Information Gain, berücksichtigt jedoch zusätzlich die Entropie innerhalb der unabhängigen Variablen selbst (Quinlan, 1986). Dabei gilt:

$$\text{GainRatio}(y, x) = \frac{H(y) - H(y|x)}{H(x)}$$

Demnach:

$$\text{GainRatio} = \frac{\text{InfoGain}}{\text{Entropie der unabhängigen Variable}}$$

Der Gain Ratio berücksichtigt die Anzahl und Größe der „Sohnknoten“ in die eine Variable einen Datensatz unterteilt (Hall et al., 2009; Quinlan, 1986). Das bedeutet,

dass der Information Gain durch die Beachtung der Entropie der Aufteilung der Instanzen korrigiert wird. Dadurch werden Variablen mit geringerer Entropie favorisiert, da neben dem Informationsgewinn (Information Gain) auch der mit der Teilung einhergehende Informationsverlust berücksichtigt wird.

Der Information Gain ist gut geeignet den prädiktiven Informationszuwachs einer Variable zu bewerten. Allerdings wird nicht berücksichtigt, ob hierdurch auch die Prognose besser wird. Mehrwertige Variablen, die zu Überanpassung (overfitting) führen können, werden durch den Information Gain zumindest nicht sanktioniert (Karegowda et al., 2010). Daher liefert das Gain Ratio wertvolle Informationen zur Zusammenstellung einer effizienten Variablenauswahl. Weiterhin kann das Gain Ratio im Vergleich zum Information Gain dazu beitragen Variablen zu identifizieren, deren Prognoseverwendbarkeit hinter ihrer Prognoserelevanz zurückbleibt.

1.4. Correlation Based Feature Selection

Der Information Gain beschreibt den Informationsgehalt einer Variable und das Gain Ratio das Verhältnis aus Information Gain und Entropie der unabhängigen Variable und berücksichtigt daher zusätzlich zum Information Gain die Entropie der Variable selbst. Bei dieser Einordnung der Prognoserelevanz und Prognoseverwertbarkeit werden keine weiteren unabhängigen Variablen betrachtet. Es ist nicht bekannt, ob die verwendeten Variablen redundant sind. Ein effizientes Frühwarnsystem würde bei vergleichbarer Genauigkeit nur aus Variablen mit komplementären Informationen bestehen. Die Analyse des Information Gain und des Gain Ratio trennt zwar relevante von nicht relevanten Informationen, reduziert aber nicht die Anzahl der Variablen. Das im Folgenden eingesetzte Correlation-Based Feature Selection Verfahren (CFS) ist eine Teilmengen-Bewertungsfunktion und bestimmt eine Auswahl an Variablen, die mit der Ergebnis-Variablen möglichst hoch korrelieren, aber möglichst wenig mit anderen gewählten Variablen korrelieren (Hall, 1998).

Dazu wird schrittweise eine Teilmenge S an unabhängigen Variablen gewählt. Die Auswahl der Variablen in jedem Schritt k wird durch die Bewertungsfunktion

$$M_{Sk} = \frac{k \cdot r_{cf}}{\sqrt{k+k \cdot (k-1) \cdot r_{ff}}}$$

geleitet (Hall, 1998). r_{cf} steht dabei für die durchschnittliche Korrelation der abhängigen Variable zu den in Schritt k gewählten unabhängigen Variablen und r_{ff} für die durchschnittliche Korrelation der unabhängigen Variablen untereinander. Sollte die Bewertungsfunktion M_{Sk} in dem Schritt $k + 1$ keine bessere Teilmenge als die derzeitige Teilmenge S ermitteln können, terminiert der Algorithmus und die derzeitige Teilmenge S wird ausgegeben (Hall, 1998).

Ziel des Verfahrens ist es, sowohl redundante als auch irrelevante Informationen aus den Daten herauszufiltern. Verfahren des maschinellen Lernens können zwar auch mit irrelevanten und redundanten Informationen umgehen, aber neben reduzierten Rechenzeiten und weniger Problemen durch Überanpassung (Overfitting), erleichtert die Fokussierung auf relevante Informationen die Interpretation der Modelle.

1.5. Faktoranalyse

Die Faktoranalyse gehört zu den explorativen, strukturentdeckenden Verfahren der Datenanalyse und wird auf Datensätze mit korrelativer Struktur angewendet (Backhaus et al., 2016, S. 386). Wie die CFS dient die Faktoranalyse dazu Datensätze auf eine geringere Anzahl latenter Variablen zu reduzieren. Das Verfahren ist dabei aber nicht nur datenreduzierend, sondern auch dimensionsreduzierend. Im Gegensatz zum CFS-Verfahren werden Variablen zu Variablengruppen – sogenannten Faktoren – zusammengefasst, die jeweils ähnliche Informationen beinhalten. Dabei werden miteinander korrelierte Variablen gruppiert und von anderen Variablen getrennt und somit Beziehungszusammenhänge in Datensätzen strukturiert. Neben dem Ziel der Datenstrukturierung kann die Faktoranalyse auch zur Datenreduzierung eingesetzt werden, indem die Ausprägung der Faktoren, die Faktorwerte, ermittelt werden.

Zur Durchführung einer Faktoranalyse ist in einem ersten Schritt der Datensatz auf Eignung zu prüfen. Backhaus et al. (2016) schlagen hierzu verschiedene Verfahren vor:

- Durch Ermittlung der Korrelationen der unabhängigen Variablen untereinander, unter Berücksichtigung der Signifikanz, können erste Aussagen über die „Bündelungsfähigkeit“ der Variablen getroffen werden.

- Alternativ kann ein Bartlett-Test durchgeführt werden. Hierbei wird die Hypothese, dass der Datensatz unkorreliert ist, getestet. Allerdings setzt der Bartlett-Test eine Normalverteilung voraus (Bartlett, 1937).
- Eine weitere Möglichkeit der Prüfung des Datensatzes auf Eignung ist die Anti-Image-Kovarianz-Matrix, bei der nach Guttman (1953) die Varianz einer Variable in zwei Teile unterteilt wird: Einen Teil der durch die verbleibenden Variablen erklärt werden kann (Image) und einen der es nicht kann, da er von anderen Variablen unabhängig ist (Anti-Image). Ein Datensatz ist nur dann für eine Faktoranalyse geeignet, wenn das Anti-Image entsprechend klein ist. Ein Datensatz wird für eine Faktoranalyse als geeignet angesehen, wenn weniger als 25% der Nicht-Diagonal-Elemente der Anti-Image-Kovarianz-Matrix Werte von $>0,09$ annehmen (Dziuban, 1974).
- Auf Basis der Anti-Image-Korrelationsmatrix wird eine von 0 bis 1 skalierte Prüfgröße, „measure of sampling adequacy“ (MSA) für jede Variable ermittelt, aus der das Kaiser-Meyer-Olkin-Kriterium (KMO) für den Datensatz gebildet wird. Das KMO kann durch Herauslösen von Variablen mit niedriger MSA verbessert werden. MSA von $<0,5$ werden als nicht akzeptabel und Werte zwischen $>0,5$ und 1 als „miserable“ bis „marvelous“ angesehen (Kaiser & Rice, 1974). Nach herrschender Meinung stellt das KMO das derzeit beste Verfahren dar (Backhaus et al., 2016; Dziuban, 1974).

Werden durch die Eignungsprüfung Variablen identifiziert, die sich nicht zur Faktoranalyse eignen, können diese sukzessive ausgeschlossen werden.

Eignet sich der Datensatz zur Faktoranalyse, sind in einem nächsten Schritt die Faktoren zu extrahieren. Die zwei wichtigsten Verfahren der Faktorenextraktion stellen die Hauptkomponentenanalyse und die Hauptachsenanalyse dar (Cleff, 2015). Der wesentliche Unterschied beider Verfahren, besteht darin, dass die Hauptkomponentenanalyse eine nahezu vollständige Darstellung der Varianz einzelner Variablen durch den Faktor unterstellt. Die Hauptachsenanalyse betrachtet die Varianz hingegen differenzierter und teilt die Varianzen von Variablen in zwei Komponenten. Einem Varianzteil der sich durch die Varianz aller Variablen bestimmt und einem variablenspezifischen Teil. Der Faktor erklärt nur den ersten Varianzteil. Während die Hauptkomponentenanalyse versucht, jede Variable zum Zwecke der Datenreduktion möglichst exakt

abzubilden, soll in der Hauptachsenanalyse der Zusammenhang zwischen den einzelnen Variablen zur Identifikation latenter Variablen erklärt werden.

Anschließend ist zu entscheiden wie viele Faktoren gebildet werden sollen. Dabei kann die Bestimmung exogen durch Expertise des Anwenders oder auch endogen geschehen. In der Regel wird hierzu das Kaiser-Kriterium, zuweilen auch Kaiser-Guttman-Kriterium genannt, angewandt. Hiernach werden alle Faktoren verwendet, deren Eigenwert >1 ist. Der Eigenwert stellt den Varianzerklärungsbeitrag eines Faktors im Vergleich zur Varianz aller Variablen dar.

Abschließend sind die einzelnen Variablen den gebildeten Faktoren zuzuordnen. Dabei sind die Ladungen der Variablen auf die Faktoren zu interpretieren. Grundsätzlich kann der Schwellenwert einer Ladung zur Zuordnung frei gewählt werden, allerdings muss dies einheitlich gestehen. Das ist immer dann problematisch, wenn eine gleichhohe Ladung auf mehrere Faktoren vorliegt. Um die Ergebnisse besser interpretieren zu können, wird die Faktormatrix rotiert. In der Regel erfolgt eine rechtwinklige Rotation, auch Varimax Rotation genannt.

Aufgrund der in den jeweiligen Schritten zu treffenden Entscheidungen, ist eine Faktoranalyse nicht als ein Verfahren, sondern als eine Vielzahl unterschiedlicher Verfahren zu betrachten. Um die Wechselwirkung der administrativen Studierendendaten untereinander zu untersuchen, wird in einem ersten Schritt der Datensatz auf Eignung mittels des KMO-Kriteriums geprüft. Um zusätzlich zur Betrachtung der Wechselwirkung auch die durch die CFS getroffene Datenreduktion vergleichen zu können, wird eine Hauptkomponentenanalyse durchgeführt. Die Anzahl der Faktoren wird nach dem Kaiser-Kriterium bestimmt und die Ergebnisse werden Varimax-rotiert.

2. Ergebnisse

2.1. Regressionsanalyse

Um ein besseres Verständnis über die Aussagekraft der verwendeten Informationen zu erhalten, werden in einem ersten Schritt die Lerndaten mittels einer logistischen Regression untersucht. Tabelle 39 zeigt die Odds Ratio des Logit models aus den ersten vier Semestern der UNI (vgl. Tabelle 39, Spalten 1 bis 5). Ziel ist es, Korrelationen zwischen den erklärenden und der abhängigen Variable aufzuzeigen um

relevante Prädiktoren des Studienabbruchs zu identifizieren. Bei Interpretation der Ergebnisse ist zu berücksichtigen, dass lediglich Korrelationen und keine Kausalzusammenhänge bestimmt werden (Zafar et al., 2017). Zum Zwecke der Abbruchsprädiktion ist das ausreichend - eine Identifikation kausaler Zusammenhänge ist nicht Ziel dieser Arbeit.

Da die zu erklärende Variable bei der Prognose von Studienabbrüchen nur die Werte 0 = Studienerfolg und 1 = Studienabbruch annehmen kann, liegt ein binäres Klassifikationsproblem vor. Da lediglich exmatrikulierte Studierende Teil der Lern- und Testdaten sind, die zum Betrachtungszeitpunkt noch eingeschrieben waren, sinkt die Anzahl der Beobachtungen mit zunehmendem späterem Zeitpunkt im Studium. Die Anzahl der Beobachtungen in Tabelle 39 sinkt um 47% vom ersten (12.728) auf das vierte Semester (6.693), demnach waren noch 53% der Studierenden des ersten Semesters eingeschrieben. Daraus folgt, dass die Koeffizienten in den Spalten nicht direkt vergleichbar sind, da sich die Größe und Zusammensetzung der Stichprobe jedes Semester ändert.

Einem ersten Schritt wird die Passung des Regressionsmodells mittels des Akaike-Informationskriterium untersucht (*AIC*). Der *AIC* beträgt 15.658 (Tabelle 39, Spalte (1)), dabei werden nur die zum Zeitpunkt der Einschreibung verfügbaren demografischen Informationen verwendet. Die Einbeziehung der Leistungsdaten des ersten Semesters reduziert den *AIC* auf 11.988 (Tabelle 39, Spalte (2)). Der *AIC* sinkt auf 7.448 im zweiten Semester und 4.178 im vierten Semester. So verbessert sich erwartungsgemäß die Passung des Modells mit fortschreitendem Semester.

Spalte 1 beinhaltet ausschließlich demografische Informationen:

- Bei Immatrikulation ist Chance eines Studienabbruchs bei männliche Studierende etwa 60% höher als bei weiblichen.
- Das Alter bei der Immatrikulation ist positiv mit dem Studienabbruch korreliert.
- Studierende mit Migrationshintergrund haben zum Zeitpunkt der Einschreibung eine höhere Abbruchwahrscheinlichkeit als zu Studierende ohne Migrationshintergrund (Vergleichsgruppe).

- Migranten der ersten Generation haben eine höhere Abbruchwahrscheinlichkeit, als Migranten der zweiten oder späteren Generationen. Die Abbruchwahrscheinlichkeit zum Zeitpunkt der Immatrikulation ist bei Migranten der ersten Generation um 22% und bei Einwanderern der zweiten Generation um 46% höher als bei Studierenden ohne Migrationshintergrund.
- Studierende, die ihr Studium mit der Fachhochschulreife aufnehmen, haben eine höhere Abbruchwahrscheinlichkeit als Studierende mit einer allgemeinen Hochschulreife.
- Der Effekt der Note der Hochschulzugangsberechtigung ist negativ und statistisch signifikant.
- Der Koeffizient der Dummy-Variablen für die private Krankenversicherung ist nur von geringer Bedeutung.
- Es ist deutlich wahrscheinlicher, dass Studierende mit akademischen Vorlesungen (Quereinstieg) an einer UNI graduieren.

Die meisten der demographischen Variablen verlieren ihre statistische Signifikanz, wenn für Leistungsdaten des ersten Semesters kontrolliert wird. Das Geschlecht in Spalte 1 hat beispielsweise einen großen und signifikanten Effekt. Dieser Effekt nimmt mit zunehmenden Semester ab und unter Verwendung der Informationen aus dem vierten Semester, wird nur noch ein geringer Effekt in Spalte 5 ausgewiesen. Auch der Migrationshintergrund ist nicht mehr signifikant, sobald Leistungsdaten verfügbar sind. Die zum Zeitpunkt der Immatrikulation verfügbaren umfangreichen Studierenden-daten sind daher nur für die Identifizierung von gefährdeten Studierenden zu Beginn des Studiums wertvoll, da bereits nach dem ersten Semester die Leistungsdaten die relevanteren Informationen bilden. Dieses Ergebnis steht im Einklang mit vergleichbaren Untersuchungen (Stinebrickner & Stinebrickner, 2014; 2012). Eine Ausnahme bildet die Dummy-Variable für die private Krankenversicherung, bei der eine signifikante Korrelation noch im vierten Semester geschätzt wird. Selbst wenn akademische Leistungen kontrolliert wird, haben Studierende mit privater Krankenversicherung eine geringere Abbruchwahrscheinlichkeit als ihre gesetzlich versicherten Kommilitonen. Pri-

vat versicherte Studierende kommen entweder aus einkommensstarken Familien, haben verbeamtete Eltern oder sind selbst einkommensstark, verbeamtet oder selbstständig. Der sozioökonomische Hintergrund erklärt die Abbruchneigung auch noch in späteren Semestern. Leistungsvariablen (Durchschnittsnote, nicht bestandene und nicht angetretene Prüfung) sind negativ mit dem Studienerfolg korreliert (Spalten 2-7). Wenig überraschend sind nicht bestandene und nicht teilgenommene Prüfungen gute Abbruchsprädiktoren. Es zeigt sich, dass die Aussagekraft akademischer Leistungen auch noch in den Folgesemestern besteht. Dies gilt auch für die Leistungsvariablen im ersten Semester. So haben Studierende, die nach schlechter Leistung im ersten Semester nicht abbrechen, immer noch eine höhere Wahrscheinlichkeit, ihr Studium nicht abzuschließen. Insbesondere die Anzahl der Credit Points (CP) ist ein statistisch signifikanter Prädiktor für die abhängige Variable.

Tabelle 39. Effekte von demografischen und Leistungsdaten auf Studienabbrüche (UNI)

Abhängige Variable: Studienabbruch (1=ja;0=nein); Logistische Regression (Odds Ratio)					
	(1)	(2)	(3)	(4)	(5)
	0. Sem.	1. Sem.	2. Sem.	3. Sem.	4. Sem.
Geschlecht (1=männlich; 0=weiblich)	1,612** (0,000)	1,309** (0,000)	1,313** (0,000)	1,288** (0,000)	1,200* (0,035)
Alter bei Einschreibung	1,076** (0,000)	1,048** (0,000)	1,070** (0,000)	1,046** (0,000)	1,077** (0,000)
Migrationshintergrund 1. Gen. (1=Ja; 0=Nein)	1,462** (0,000)	1,083 (0,222)	1,054 (0,300)	1,113 (0,959)	1,072 (0,860)
Migrationshintergrund 2. Gen. (1=Ja; 0=Nein)	1,222** (0,000)	1,074 (0,410)	1,082 (0,661)	1,004 (0,429)	1,019 (0,661)
Ort der HZB (1= Hochschulstadt; 0=andere)	1,337** (0,000)	1,084 (0,170)	1,143+ (0,076)	1,199* (0,036)	1,332** (0,005)
FHR (1=Ja; 0=Nein)	2,033** (0,003)	1,611+ (0,085)	1,132 (0,683)	0,878 (0,714)	0,888 (0,776)
fgHR (1=Ja; 0=Nein)	0,822 (0,432)	1,173 (0,612)	0,891 (0,738)	0,897 (0,786)	0,903 (0,828)
Ausländische HZB 1=Ja; 0=Nein)	1,055 (0,695)	0,869 (0,387)	0,944 (0,769)	0,887 (0,589)	0,968 (0,898)
Note HZB	1,368** (0,000)	1,056+ (0,081)	1,000 (0,996)	0,947 (0,220)	0,944 (0,281)
Krankenversicherung (1=privat; 0=gesetzlich)	0,875+ (0,097)	0,694** (0,000)	0,775+ (0,042)	0,670** (0,004)	0,718+ (0,044)
Anzahl Studiengänge	0,879** (0,000)	0,934** (0,000)	0,978 (0,292)	1,048+ (0,070)	1,087** (0,005)
Quereinsteiger	0,399** (0,000)	0,570** (0,000)	0,474** (0,000)	0,498** (0,000)	0,462** (0,000)
Anzahl Semester an vorherigen Hochschule	1,089** (0,000)	1,061** (0,000)	1,059** (0,000)	1,047** (0,000)	1,030+ (0,052)
Durchschnittsnote aktuelles Semester		1,658** (0,000)	1,381** (0,000)	1,193** (0,006)	1,352** (0,000)
Durchschnitts-CP aktuelles Semester		0,948** (0,000)	0,932** (0,000)	0,936** (0,000)	0,942** (0,000)
Ohne Prüfungen		17,142** (0,000)	5,988** (0,000)	3,470** (0,000)	4,180** (0,000)
Anzahl nicht angemeldeter aber nicht angetretener Prüfungen (NE)		1,464** (0,000)	1,272** (0,000)	1,183** (0,000)	0,993 (0,892)
Anzahl nicht bestandener Prüfungen (NB)		1,380** (0,000)	1,252** (0,000)	1,175** (0,000)	1,274** (0,000)
Konstante	0,231** (0,000)	0,177** (0,000)	0,120** (0,000)	0,262** (0,000)	0,175** (0,000)
Leistungen vorheriger Sem.:					
Vorherige CP und Noten			JA	JA	JA
Vorherige Anzahl von Prüfungen			JA	JA	JA
Vorherigen Anzahl „NE“			JA	JA	JA
Vorherigen Anzahl „NB“			JA	JA	JA
Wichtige Prüfungen		JA	JA	JA	JA
AIC	15.657,54	11.988,42	7.448,05	5.757,76	4.178,26
N	12.728	12.728	9.228	8.015	6.693

Hinweis: + $p < 0,1$; * $p < 0,05$; ** $p < 0,01$. Standardfehler in Klammern.

Tabelle 40 zeigt die Ergebnisse der Logit-Schätzung der FH. Die Zahl der Beobachtungen sinkt von 7.077 im ersten Semester auf 5.448 Studierende im vierten Semester. Ähnlich wie bei der UNI verbessert sich die Modellpassung in den Folgesemestern. Es ist zu vermuten, dass die an der FH erhobenen Studiengebühren zu beschleunigten Abbruchentscheidungen führt. Die Ergebnisse sind vergleichbar mit den Ergebnissen aus der UNI – insbesondere in Bezug auf die Stärke und Richtung der Koeffizienten der leistungsbezogenen Daten.

Tabelle 40. Effekte von demografischen und Leistungsdaten auf Studienabbrüche (FH)

Abhängige Variable: Studienabbruch (1=ja;0=nein); Logistische Regression (Odds Ratio)					
	(1)	(2)	(3)	(4)	(5)
	0. Sem.	1. Sem.	2. Sem.	3. Sem.	4. Sem.
Geschlecht (1=männlich; 0=weiblich)	1,576** (0,000)	1,279** (0,002)	1,064 (0,556)	0,968 (0,808)	0,865 (0,387)
Alter bei Einschreibung	1,045** (0,000)	1,030** (0,001)	1,031** (0,006)	1,035* (0,016)	1,072** (0,000)
Migrationshintergrund 1. Gen. (1=Ja; 0=Nein)	1,602** (0,976)	0,904 (0,002)	0,795 (0,001)	0,827 (0,000)	0,763 (0,045)
Migrationshintergrund 2. Gen. (1=Ja; 0=Nein)	1,003 (0,001)	0,777** (0,387)	0,692** (0,139)	0,569** (0,323)	0,712* (0,248)
Ort der HZB (1= Ort der Hochschule; 0=andere)	1,197* (0,045)	1,006 (0,932)	0,868 (0,138)	0,892 (0,342)	0,856 (0,306)
FHR (1=Ja; 0=Nein)	2,136** (0,000)	1,314** (0,000)	1,122 (0,228)	1,194 (0,139)	1,040 (0,795)
fgHR (1=Ja; 0=Nein)	3,148** (0,000)	2,143** (0,000)	1,865* (0,035)	1,980* (0,083)	1,922 (0,199)
Ausländische HZB (1=Ja; 0=Nein)	5,365** (0,000)	3,388** (0,000)	1,414 (0,280)	0,788 (0,544)	0,320* (0,017)
akademische Vorleistungen	1,468** (0,000)	1,345** (0,000)	1,212** (0,000)	1,142* (0,046)	1,110 (0,202)
Durchschnittsnote des Semesters		2,141** (0,000)	1,357** (0,001)	1,352** (0,010)	0,999 (0,995)
Durchschnitt CP des Semester		0,974** (0,008)	0,901** (0,000)	0,893** (0,000)	0,915** (0,000)
Ohne Prüfungen		17,281** (0,000)	10,948** (0,000)	3,620** (0,002)	2,645* (0,038)
Anzahl nicht angemeldeter aber Nicht angetretener Prüfungen (NE)		1,333** (0,000)	1,281** (0,000)	1,140* (0,027)	1,330** (0,000)
Anzahl nicht bestandener Prüfungen		1,339** (0,000)	1,082* (0,040)	1,108* (0,024)	1,230** (0,000)
Konstante	-0,055 (0,108)	-0,065* (0,070)	0,461** (0,000)	0,670** (0,000)	0,704** (0,000)
Vollzeit / Teilzeit / dual	JA				
Leistungen vorheriger Sem.:					
Vorherige CP und Noten			JA	JA	JA
Vorherige "NE"			JA	JA	JA
Vorherige verpasste Prüfungen			JA	JA	JA
Vorherigen nicht bestandene Prüf.		JA	JA	JA	JA
Wichtige Scheine			JA	JA	JA
AIC	3.983,58	6.108,78	3.719,80	2.525,38	1.703,21
N	7.077	7.077	6.329	5.847	5.448

Hinweis: + $p < 0,1$; * $p < 0,05$; ** $p < 0,01$. Standardfehler in Klammern.

2.2. Verzicht auf demografische Informationen

Aufbauend auf den Erkenntnissen der Regressionsanalyse wird dargestellt, in wie weit demografische Studierendendaten für eine Prognose verzichtbar sind. Hierfür werden die Teil II, Kapitel 3 durchgeführten Prognosen erneut durchgeführt, allerdings ausschließlich mit Leistungsdaten. Dabei zeigen sich bereits für das erste Leistungssemester respektable Genauigkeiten, wie Tabelle 40 darstellt.

Tabelle 41. Prognosegenauigkeit unter ausschließlicher Verwendung von Leistungsdaten

AdaBoost	UNI					FH				
	0. Sem.	1. Sem.	2. Sem.	3. Sem.	4. Sem.	0. Sem.	1. Sem.	2. Sem.	3. Sem.	4. Sem.
Accuracy ^a	-	76,60	81,42	86,61	88,52	-	83,64	90,45	92,44	94,76
Recall ^a	-	71,49	73,51	79,12	77,94	-	69,89	78,20	79,02	80,42
Precision ^a	-	71,49	73,51	79,12	77,94	-	70,87	78,49	79,02	80,42
Anzahl Absolventen	1.112	1.039	1.027	1.015	992	976	976	974	969	961
Anzahl Abbrecher	1.015	726	555	479	349	476	380	284	223	161
korrekt klassifizierte Absolventen	-	833	880	915	915	-	836	881	890	897
falsch klassifizierte Abbrecher	-	207	147	100	77	-	109	58	43	28
korrekt klassifizierte Abbrecher	-	519	408	379	272	-	253	208	162	115
falsch klassifizierte Absolventen	-	206	147	100	77	-	104	57	43	28
korrekt klassifizierte Absolventen ^a	-	80,17	85,69	90,15	92,24	-	88,94	93,92	95,39	96,97
falsch klassifizierte Abbrecher ^a	-	19,92	14,31	9,85	7,76	-	11,60	6,18	4,61	3,03
korrekt klassifizierte Abbrecher ^a	-	71,49	73,51	79,12	77,94	-	69,89	78,20	79,02	80,42
falsch klassifizierte Absolventen ^a	-	28,37	26,49	20,88	22,06	-	28,73	21,43	20,98	19,58

Hinweis: ^a Werte in Prozent.

Verglichen mit Prognosen unter Verwendung demografischer Informationen – zur besseren Lesbarkeit hier noch einmal dargestellt – stellt die Prognostizierbarkeit von Studienabbrüchen vor Mitte des zweiten Semesters den erheblichsten Unterschied dar.

Tabelle 42. Prognosegenauigkeit des AdaBoost bei tatsächlicher Abbruchquote

AdaBoost	UNI					FH				
	0. Sem.	1. Sem.	2. Sem.	3. Sem.	4. Sem.	0. Sem.	1. Sem.	2. Sem.	3. Sem.	4. Sem.
Accuracy ^a	67,65	78,53	82,43	87,62	89,63	67,17	84,49	89,70	93,50	95,51
Recall ^a	65,81	73,83	74,95	80,58	79,94	49,78	72,10	76,69	81,95	83,22
Precision ^a	66,20	73,93	74,95	80,75	80,17	49,89	72,10	76,69	81,95	83,22

Hinweis: ^a Werte in Prozent. Verkürzte Darstellung

Prognosen vor Mitte des zweiten Semesters sind nur unter Verwendung von demografischen Informationen möglich. Hiernach werden an beiden Universitäten die Prognosen unter Verwendung von demographischen als auch Leistungsdaten nur geringfügig verbessert. Die Verwendung von demografischen Daten der Studierenden erscheint nur dann sinnvoll, wenn keine Leistungsdaten verfügbar sind und Hilfsangebote bereits zu Beginn Studiums ausgesprochen werden sollen. Sobald nach dem ersten Semester Leistungsdaten verfügbar sind, ergänzen demographische Daten das Prognosemodell nur noch geringfügig. Nach dem ersten Semester beträgt der Prozentsatz der korrekt vorhergesagten Abbrecher in der UNI 71% bei alleiniger Verwendung von akademischen Leistungsdaten und 74% bei Verwendung von Demographie- und Leistungsdaten.

2.3. Information Loss

Sinnverwandt zur Betrachtung des vorangegangenen Kapitels wird im Folgenden die Prognoserelevanz jeder einzelnen Variable als Genauigkeitsreduktion nach Hinwegnahme ceteris paribus ermittelt. Für beide Hochschulen und für fünf Betrachtungszeitpunkte wurden die Datensätze mit je nach Semester 18 bis 45 Variablen jeweils um eine Variable reduziert und die Auswirkung auf die Accuracy dargestellt. Negative Zahlen stellen einen Genauigkeitszuwachs dar. Werte größer als 0,5% sind zur besseren Lesbarkeit markiert.

Tabelle 43. Information Loss

Semester:	UNI					FH				
	0	1	2	3	4	0	1	2	3	4
StgAbbruchquote	3,65%	0,16%	0,19%	0,20%	0,45%					
StartSem	0,37%	-0,23%	-0,06%	-0,03%	0,06%	-0,20%	-0,36%	0,18%	0,18%	0,09%
Studiengang	1,90%	0,92%	-0,08%	0,38%	-0,14%	0,96%	-0,05%	0,19%	0,09%	0,15%
ImmtrAlter	-0,48%	0,21%	-0,04%	0,14%	-0,01%	0,31%	0,09%	0,32%	0,71%	0,15%
Geschlecht	0,71%	0,07%	-0,05%	0,14%	-0,25%	0,07%	-0,03%	0,64%	0,56%	0,50%
AnmeldeZeit	0,79%	0,72%	0,06%	0,25%	0,01%					
HZB	1,27%	-0,02%	-0,19%	-0,08%	-0,13%	0,47%	0,10%	0,41%	0,47%	0,40%
HZBNote	1,26%	-0,01%	-0,21%	0,05%	-0,14%					
VorSem	0,38%	0,25%	-0,08%	-0,10%	0,00%	0,03%	0,02%	0,32%	0,23%	0,20%
Kreis der HZB	0,53%	0,20%	-0,35%	-0,09%	-0,10%	-0,33%	0,23%	-0,36%	0,14%	0,10%
Land	0,48%	-0,20%	-0,28%	-0,02%	-0,13%	-0,08%	-0,32%	-0,04%	0,47%	0,00%
Migration	0,24%	-0,16%	-0,23%	-0,19%	-0,15%	-0,07%	0,10%	0,01%	0,14%	0,05%
Akad. Vorleistung	1,49%	0,30%	0,20%	-0,10%	0,09%	-0,05%	-0,03%	0,18%	0,28%	0,20%
StBAusland	0,36%	0,04%	0,15%	-0,07%	0,04%	-0,05%	0,04%	-0,03%	-0,32%	0,00%
Region	0,08%	-0,08%	-0,16%	-0,06%	-0,22%	-0,42%	-0,20%	0,45%	-0,01%	0,05%
GeburtsortAusland	0,34%	-0,02%	-0,12%	-0,02%	0,14%	-0,28%	0,12%	0,00%	0,04%	0,10%
Studienform	0,43%	0,14%	0,07%	-0,19%	0,09%	-0,25%	-0,30%	0,18%	0,19%	0,19%
KV	0,00%	-0,07%	-0,14%	-0,02%	-0,17%	-0,15%	-0,54%	0,09%	0,28%	0,10%
CPSem1		0,95%	-0,14%	0,12%	-0,29%		0,71%	-0,10%	0,37%	0,31%
AnzahlScheinSem1		0,11%	-0,19%	0,01%	-0,07%		-0,04%	0,28%	0,66%	0,30%
NoteSem1		0,62%	-0,04%	0,10%	-0,22%		0,55%	0,23%	0,38%	-0,05%
NTSem1		0,17%	0,17%	0,15%	0,02%		0,12%	-0,05%	0,19%	0,30%
NBSem1		0,31%	-0,12%	0,15%	0,10%		0,00%	0,14%	0,09%	0,20%
Schein1		0,27%	-0,16%	0,08%	-0,09%		0,04%	-0,09%	0,33%	0,60%
Schein6		0,04%	-0,21%	-0,07%	-0,05%		-0,21%	0,45%	-0,24%	0,40%
Schein2		-0,15%	-0,03%	0,01%	0,04%		0,08%	0,27%	0,28%	0,10%
Schein3		0,06%	-0,41%	0,12%	0,00%		-0,08%	0,05%	0,18%	0,20%
Schein5		-0,09%	-0,28%	0,04%	-0,01%		-0,17%	0,31%	0,28%	0,30%
Schein4		-0,08%	-0,16%	-0,01%	-0,12%		-0,08%	0,46%	0,47%	0,30%
AnzahlScheinSem2			-0,26%	-0,01%	0,21%			0,04%	0,19%	-0,05%
CPSem2			-0,15%	-0,04%	-0,10%			0,09%	0,04%	0,30%
NoteSem2			-0,05%	-0,11%	-0,21%			0,00%	0,28%	0,30%
NBSem2			-0,37%	0,21%	0,18%			0,27%	-0,01%	0,10%
NTSem3			0,03%	-0,09%	0,03%			0,32%	0,47%	0,15%
AnzahlScheinSem3				0,29%	-0,08%				-0,01%	0,59%
CPSem3				0,37%	0,31%				0,75%	0,50%
NoteSem3				-0,03%	-0,13%				0,46%	0,25%
NBSem3				-0,01%	-0,17%				0,04%	0,30%
NTSem3				0,07%	-0,32%				0,47%	0,20%
4CPSem					-0,14%					0,53%
AnzahlScheinSem4					0,19%					0,11%
NoteSem4					-0,06%					0,40%
NTSem4					-0,06%					0,50%
NBSem4					0,13%					0,10%

Wie im vorangegangenen Abschnitt zeigen sich Unterschiede zwischen beiden Hochschulen. Welche Variablen besondere Aussagekraft haben, ist hochschulabhängig. Von hoher Bedeutung an beiden Hochschulen sind die akademischen Leistungen des vorangegangenen Semesters gemessen in CP und Note oder die Anzahl nicht teilgenommener Prüfungen des Semesters. Demografische Informationen verlieren nach vollendetem ersten Semester schnell an Bedeutung.

Nur innerhalb der ersten Studienmonate reduziert der Verzicht auf eine Variable die Genauigkeit nennenswert. Bereits nach dem ersten Semester, wenn Leistungsdaten verfügbar werden, sinkt die Accuracy nur noch um weniger als einen Prozentpunkt. Das bedeutet, dass durch korrelierende Informationen der Informationsgehalt einzelner Variablen nahezu redundant ist.

2.4. Information Gain der Variablen

Für Prognosen in der frühen Studienphase kann nicht auf Leistungsdaten zurückgegriffen werden, ab Mitte des zweiten Semesters scheinen demografische Informationen entbehrlich. Daher ist zu prüfen, wie hoch der Informationsgehalt des Studienfortschritts und aller Nichtleistungsdaten im Vergleich und zu unterschiedlichen Zeitpunkten im Studienverlauf sind.

Zur Bewertung des Wertes einzelner Variablen wird der in Teil IV, Abschnitt 1.2 beschriebene Information Gain verwendet (Quinlan, 1986). Nach Ermittlung des Information Gains für alle Variablen in allen Semestern wird eine Rangliste der unabhängigen Variablen nach Wichtigkeit erstellt (Quinlan, 1986) (vgl. Tabelle 44 & Tabelle 45). Dabei zeigt sich, dass sich die Rangfolge der Variablen im Studienverlauf nur wenig verändert. Demografische Variablen und Informationen zum Zeitpunkt der Immatrikulation sind jedoch wichtig und alternativlos, wenn frühe Prognosen gewünscht sind. Ihre Bedeutung nimmt allerdings mit verfügbar werdenden Leistungsdaten sehr schnell ab. Die jeweils aktuellen akademischen Leistungen – gemessen in ECTS-Punkten, Durchschnittsnote und Anzahl bestandener Prüfungen des Betrachtungssemesters – bieten zu jedem Zeitpunkt den höchsten Informationsgewinn. Dies spricht für ein – evtl. durch Noten angereichertes – ECTS-Monitoring nach einer Studieneingangsphase. Wenn aber abbruchgefährdeten Studierenden Hilfe angeboten werden soll, bevor Leistungsdaten vorliegen (d.h. vor Mitte des zweiten Semesters), können nur demografische Informationen und Informationen zu schulischen Leistungen, voruniversitären Entscheidungen und akademischer Vorerfahrung verwendet werden. Im Folgenden wird geprüft, welche dieser Variablen einen hohen Informationsgehalt für einen drohenden Studienabbruch haben.

Tabelle 44 zeigt den Information Gain verfügbarer administrativer Studierendendaten an der UNI. Zu Beginn des Studiums sind die Variablen, die den Studieneinstieg beschreiben, besonders informativ. Hierunter fällt an der UNI die Art und Anzahl der

gewählten Studiengänge (Rang 1-2 von 18), wie groß der zeitliche Abstand zwischen Studienbewerbung und Studienstart (Rang 3) war und ob ein Studienstart zum Winter- oder Sommersemester gewählt wurde (Rang 5). Aus den verfügbaren demografischen Variablen sind lediglich das Alter bei Einschreibung (Rang 4) und das Geschlecht (Rang 7) bedeutsam. Weniger wichtig sind die Variablen zur ethnischen Herkunft (Rang 10, 12, 14-16) oder die Art der Krankenversicherung (gesetzlich oder privat) als möglicher Indikator für den sozioökonomischen Hintergrund der Studierenden (Rang 18). Die Herkunft der Studierenden wird anhand der Staatsbürgerschaft, des Geburtsortes und, sofern beides deutsch ist, anhand einer Namensanalyse ermittelt (vgl. Teil II und V). Für die Prognose sind diese Informationen aber von untergeordneter Bedeutung, wesentlich wichtiger sind Informationen zu Entscheidungen vor und bei Studienstart.

Zu prüfen ist die Bedeutung schulischer Leistungen sowie eventuelle vorangegangene akademische Erfahrungen. Die Art und Note der Hochschulzugangsberechtigung sind in etwa gleich wichtig und nehmen zusammen mit dem Geschlecht einen mittleren Rang (Rang 6-8) ein. Ob akademische Erfahrungen vorliegen und wie viele Semester (Rang 9 und 11) der Studierende bereits in einem vorangegangenen Studium eingeschrieben war, hat einen eher geringen Erklärungsgehalt.

Tabelle 44. Information Gain – UNI

Semester	0		1		2		3		4		Variable
	Rang	InfoGain	Rang	InfoGain	Rang	InfoGain	Rang	InfoGain	Rang	InfoGain	
StgAbbruchquote	1	0,117	4	0,113	7	0,112	10	0,100	13	0,083	StgAbbruchquote
Studiengang	2	0,045	7	0,033	17	0,030	23	0,030	24	0,029	Studiengang
AnmeldeZeit	3	0,042	6	0,044	9	0,055	15	0,047	21	0,052	AnmeldeZeit
ImmtrAlter	4	0,034	8	0,033	11	0,051	16	0,047	18	0,056	ImmtrAlter
StartSem	5	0,017	16	0,010	21	0,018	30	0,012	31	0,014	StartSem
HZBNote	6	0,017	10	0,023	12	0,042	18	0,045	22	0,051	HZBNote
Geschlecht	7	0,015	13	0,018	18	0,029	22	0,031	25	0,029	Geschlecht
HZB	8	0,015	14	0,017	20	0,024	24	0,025	28	0,024	HZB
VorSem	9	0,008	17	0,009	22	0,018	26	0,015	30	0,016	VorSem
Land	10	0,004	22	0,004	28	0,005	33	0,005	38	0,005	Land
Akad. Vorleistung	11	0,003	23	0,003	25	0,006	31	0,008	36	0,008	Akad. Vorleistung
Region	12	0,002	25	0,003	29	0,004	35	0,003	41	0,003	Region
Kreis der HZB	13	0,002	21	0,004	27	0,005	32	0,006	37	0,006	Kreis der HZB
Migration	14	0,002	24	0,003	26	0,006	34	0,004	39	0,005	Migration
StBAusland	15	0,001	26	0,001	30	0,002	36	0,003	40	0,004	StBAusland
GeburtsortAusland	16	0,000	27	0,000	32	0,001	37	0,001	42	0,002	GeburtsortAusland
Studienform	17	0,000	28	0,000	33	0,000	38	0,000	43	0,000	Studienform
KV	18	0,000	29	0,000	34	0,000	39	0,000	44	0,000	KV
CPSem1			1	0,216	4	0,203	7	0,180	10	0,156	CPSem1
NoteSem1			2	0,200	6	0,171	9	0,153	12	0,135	NoteSem1
AnzahlScheinSem1			3	0,196	5	0,183	8	0,164	11	0,144	AnzahlScheinSem1
Schein1			5	0,054	8	0,061	12	0,067	17	0,061	Schein1
Schein2			9	0,027	31	0,002	13	0,062	15	0,069	Schein2
Schein3			11	0,019	10	0,053	11	0,075	14	0,070	Schein3
NBSem1			12	0,019	19	0,024	25	0,025	29	0,018	NBSem1
NTSem1			15	0,010	24	0,012	29	0,013	34	0,011	NTSem1
Schein6			18	0,008	15	0,034	14	0,048	20	0,052	Schein6
Schein4			19	0,005	13	0,041	17	0,046	16	0,066	Schein4
Schein5			20	0,004	14	0,037	19	0,043	19	0,054	Schein5
CPSem2					1	0,309	4	0,279	7	0,237	CPSem2
AnzahlScheinSem2					2	0,303	5	0,274	8	0,236	AnzahlScheinSem2
NoteSem2					3	0,294	6	0,252	9	0,218	NoteSem2
NBSem2					16	0,030	21	0,032	26	0,026	NBSem2
NTSem2					23	0,012	27	0,015	33	0,013	NTSem2
CPSem3							1	0,326	5	0,271	CPSem3
AnzahlScheinSem3							2	0,324	4	0,273	AnzahlScheinSem3
NoteSem3							3	0,302	6	0,242	NoteSem3
NBSem3							20	0,035	23	0,033	NBSem3
NTSem3							28	0,013	32	0,014	NTSem3
CPSem4									1	0,312	4CPSem
AnzahlScheinSem4									2	0,306	AnzahlScheinSem4
NoteSem4									3	0,278	NoteSem4
NBSem4									27	0,024	NBSem4
NTSem4									35	0,011	NTSem4

Tabelle 45 fasst den Information Gain für die FH zusammen. Abweichungen in den verwendeten Variablen zwischen den Hochschulen sind auch durch die Nichtverfügbarkeit einiger Informationen an der FH zu erklären. Da Studienplätze an der FH nicht in Abhängigkeit von der Note der HZB vergeben werden, wird diese Information nicht erhoben. Auch wird nicht das tatsächliche Datum der Immatrikulation erfasst, sondern nur der Semesterstart. Damit ist es für die FH nicht möglich, den zeitlichen

Abstand zwischen Anmeldung und Studienstart zu ermitteln. Da jedoch an der FH anders als an der Uni alle Studierenden nur in jeweils einem Studiengang immatrikuliert sind, kann der Studiengang selbst als Information genutzt werden.

Tabelle 45. Information Gain – FH

Semester	0		1		2		3		4		Variable
	Rang	InfoGain	Rang	InfoGain	Rang	InfoGain	Rang	InfoGain	Rang	InfoGain	
	1	0,041	6	0,043	9	0,038	16	0,037	18	0,028	HZB
	2	0,034	8	0,037	12	0,033	17	0,031	19	0,024	Studiengang
	3	0,021	13	0,021	15	0,022	20	0,021	20	0,022	ImmatriAlter
	4	0,017	12	0,022	17	0,020	21	0,020	24	0,016	Geschlecht
	5	0,012	16	0,012	19	0,015	23	0,016	27	0,014	Land
	6	0,011	19	0,011	25	0,010	30	0,009	32	0,008	Migration
	7	0,010	17	0,012	22	0,012	27	0,012	33	0,008	Region
	8	0,010	20	0,011	23	0,011	28	0,011	31	0,009	StBAusland
	9	0,009	18	0,011	21	0,013	25	0,013	29	0,014	Akad. Vorleistung
	10	0,008	22	0,007	28	0,005	35	0,003	39	0,003	Studienform
	11	0,007	21	0,008	24	0,011	29	0,010	28	0,014	VorSem
	12	0,006	23	0,007	26	0,006	31	0,007	36	0,005	Geburtsort
	13	0,006	24	0,006	27	0,006	33	0,006	37	0,004	VorVersuch
	14	0,006	25	0,005	29	0,004	34	0,004	38	0,004	Kreis der HZB'
	15	0,002	26	0,001	30	0,003	36	0,002	41	0,002	StartSem
	16	0,000	27	0,000	32	0,000	37	0,000	43	0,000	HZBNote
	17	0,000	28	0,000	33	0,000	38	0,000	42	0,000	KV
			1	0,204	4	0,159	8	0,122	10	0,096	CPSem1
			2	0,182	5	0,143	9	0,114	12	0,087	AnzahlScheinSem1
			3	0,132	6	0,103	10	0,085	13	0,065	NoteSem1
			4	0,070	7	0,064	6	0,137	9	0,123	Schein1
			5	0,062	31	0,001	14	0,055	16	0,041	Schein2
			7	0,041	8	0,039	11	0,081	14	0,065	Schein3
			9	0,028	14	0,022	13	0,071	17	0,040	Schein5
			10	0,026	13	0,024	19	0,022	23	0,020	NTSem1
			11	0,025	18	0,020	22	0,018	26	0,014	NBSem1
			14	0,017	10	0,036	12	0,075	22	0,021	Schein4
			15	0,014	11	0,034	15	0,048	15	0,050	Schein6
					1	0,250	3	0,201	6	0,153	CPSem2
					2	0,229	5	0,179	8	0,131	AnzahlScheinSem2
					3	0,170	7	0,134	11	0,095	NoteSem2
					16	0,021	18	0,023	21	0,022	NTSem3
					20	0,015	24	0,014	25	0,015	NBSem2
							1	0,265	3	0,201	CPSem3
							2	0,251	4	0,187	AnzahlScheinSem3
							4	0,189	7	0,138	NoteSem3
							26	0,012	30	0,013	NTSem3
							32	0,006	34	0,006	NBSem3
									1	0,233	4CPSem
									2	0,222	AnzahlScheinSem4
									5	0,179	NoteSem4
									35	0,006	NTSem4
									40	0,003	NBSem4

Wie auch an der UNI sind das Geschlecht (Rang 4 von 14) und das Alter bei Immatrikulation (Rang 3) relevante Informationen und wichtiger als die ethnische Herkunft (Rang 5-8). Anders als an der UNI ist der Erklärungsgehalt der Herkunft jedoch höher als die akademische Vorerfahrung (Rang 9 bis 13). An der FH stellt nicht die Wahl des Studiengangs (Rang 2) die bedeutsamste Information dar, sondern die Art

der Hochschulzugangsberechtigung (Rang 1). Nicht bedeutend ist hingegen, ob das Studium im Sommer- oder Wintersemester (Rang 15) begonnen wurde. In Gegenüberstellung beider Hochschulen ergeben sich demnach Rangunterschiede zwischen den Variablen, die durch die Systemunterschiede zwischen den Hochschulen zu erklären sind.

Die Analyse zeigt, dass ein Frühwarnsystem mit reduzierter Variablenanzahl an die jeweilige Hochschule angepasst werden muss. Der Erklärungsgehalt der Variablen hängt stark von der betrachteten Hochschule bzw. der Art der Hochschule ab. Hervorzuheben ist beispielsweise die unterschiedliche Bedeutung der Art der HZB für die Prognose des Studienabbruchs. Während an der UNI weit überwiegend nur Studierende mit der allgemeinen Hochschulreife am Gymnasium oder an der Gesamtschule immatrikuliert sind, kann ein Studium an der FH auch bei Vorliegen der Fachhochschulreife, der fachgebundenen Hochschulreife oder unabhängig vom Schulabschluss bei einer abgeschlossenen Berufsausbildung nach drei Jahren Berufserfahrung aufgenommen werden. Daher weist die Art der HZB eine deutlich höhere Variation auf als an der UNI und der Erklärungsgehalt ist deutlich höher für die private FH als für die UNI.

Wie erwartet, kommt der akademischen Leistung gemessen am Studienfortschritt in ECTS-Punkten pro Semester an beiden Hochschulen erhebliche Bedeutung zu. Ebenfalls relevant sind die Anzahl der bestandenen Prüfungen und die Durchschnittsnote des betrachteten Semesters (Rang 2 und 3). Ob die Studierenden sich hingegen von Prüfungen abgemeldet haben oder der Prüfung unentschuldigt ferngeblieben sind, ist weniger bedeutsam.

2.5. Gain Ratio der Variablen

In Betrachtung der akademischen Leistungen beider Hochschulen nach dem ersten Semester in Tabelle 44 und Tabelle 45 ist festzuhalten, dass dem Studienfortschritt in ECTS-Punkten pro Semester erheblich Bedeutung zukommt. Ebenfalls bedeutsam ist die Anzahl der bestandenen Prüfungen und die Durchschnittsnote des Betrachtungssemesters (Rang 2 und 3). Ob der Studierende sich hingegen von Prüfungen abgemeldet hat, oder der Prüfung unentschuldigt ferngeblieben ist, hat nur einen vernachlässigbaren prognostischen Wert. Der Information Gain beschreibt das Verhältnis zwischen der unabhängigen und abhängigen Variable, nicht aber ob die abhängige

Variable sich gut für maschinelle Prognosen eignet. Daher berücksichtigt das Gain Ratio ebenfalls die Entropie der unabhängigen Variable. Insbesondere die Verwendung mehrwertiger nominaler Variablen neigt zum Overfitting und verschlechtern die Prognosegüte. Der Vergleich zwischen Information Gain und Gain Ratio kann dazu beitragen, Informationen zu identifizieren, die von hoher Relevanz sind, aber in vorliegender Operationalisierung noch nicht optimal zur Prognose beitragen oder sie durch Overfittung sogar verschlechtern.

Tabelle 46 und Tabelle 47 zeigen den Gain Ratio der Variablen beider Hochschulen. Hiernach weisen die verfügbaren Studierendeninformationen an der UNI folgende Rangfolge auf:

1. Akademische Leistungen in CP und Note des Betrachtungssemesters
2. Akademische Leistungen in CP und Note vorangegangener Semesters und in absteigender Reihenfolge die Leistungen früherer Semester. Die am weitesten zurückliegenden Informationen, weisen den geringsten Gain Ratio auf.
3. Voruniversitäre Entscheidungen des Studienfachs und Anmeldeverhalten
4. Alter und Geschlecht
5. Schulische Leistungen
6. Akademische Vorerfahrung
7. Herkunft

An der FH:

1. Akademische Leistungen in CP und Note des Betrachtungssemesters
2. Akademische Leistungen in CP und Note vorangegangener Semesters und in absteigender Reihenfolge die Leistungen früherer Semester. Die am weitesten zurückliegenden Informationen, weisen den geringsten Gain Ratio auf.
3. Schulische Leistungen
4. Akademische Vorerfahrung
5. Alter und Geschlecht
6. Herkunft
7. Voruniversitäre Entscheidungen des Studienfachs

Tabelle 46. Gain Ratio – UNI

Semester	0		1		2		3		4		Variable
	Rang	Gain Ratio	Rang	Gain Ratio	Rang	Gain Ratio	Rang	Gain Ratio	Rang	Gain Ratio	
StgAbbruchquote	1	0,038	5	0,040	9	0,043	14	0,040	17	0,043	StgAbbruchquote
StartSem	2	0,031	9	0,021	12	0,038	19	0,029	20	0,037	StartSem
Studiengang	3	0,024	10	0,019	23	0,017	29	0,014	32	0,016	Studiengang
ImmtrAlter	4	0,016	12	0,016	15	0,029	20	0,027	22	0,032	ImmtrAlter
Geschlecht	5	0,015	11	0,019	14	0,029	17	0,031	23	0,029	Geschlecht
HZB	6	0,013	14	0,016	18	0,022	22	0,023	27	0,023	HZB
AnmeldeZeit	7	0,013	16	0,013	21	0,018	26	0,017	26	0,023	AnmeldeZeit
HZBNote	8	0,007	19	0,010	22	0,017	25	0,018	30	0,021	HZBNote
VorSem	9	0,006	20	0,007	25	0,014	30	0,012	35	0,013	VorSem
Kreis der HZB	10	0,003	21	0,005	26	0,007	31	0,008	38	0,008	Kreis der HZB'
Land	11	0,003	23	0,004	29	0,005	35	0,005	40	0,006	Land
Migration	12	0,003	22	0,004	31	0,004	34	0,006	39	0,008	Migration
Akad. Vorleistung	13	0,003	24	0,003	27	0,006	32	0,008	36	0,009	Akad. Vorleistung
StBAusland	14	0,003	26	0,002	28	0,005	33	0,007	37	0,009	StBAusland
Region	15	0,003	25	0,003	30	0,004	36	0,003	41	0,004	Region
GeburtsortAusland	16	0,000	27	0,000	32	0,002	37	0,002	42	0,003	GeburtsortAusland
Studienform	17	0,000	28	0,000	33	0,000	38	0,000	43	0,000	Studienform
KV	18	0,000	29	0,000	34	0,000	39	0,000	44	0,000	KV
CPSem1			1	0,079	4	0,070	7	0,064	10	0,063	CPSem1
AnzahlScheinSem1			2	0,073	6	0,066	9	0,059	11	0,060	AnzahlScheinSem1
NoteSem1			3	0,058	5	0,069	8	0,061	12	0,055	NoteSem1
Schein1			4	0,047	7	0,044	13	0,041	21	0,035	Schein1
Schein6			6	0,036	8	0,043	12	0,046	18	0,041	Schein6
Schein2			7	0,031	17	0,022	11	0,046	14	0,047	Schein2
Schein3			8	0,028	10	0,042	10	0,055	13	0,049	Schein3
NTSem1			13	0,016	20	0,019	23	0,021	31	0,018	NTSem1
NBSem1			15	0,015	19	0,020	24	0,020	33	0,015	NBSem1
Schein5			17	0,013	13	0,037	16	0,039	19	0,039	Schein5
Schein4			18	0,011	11	0,039	15	0,039	15	0,047	Schein4
AnzahlScheinSem2					1	0,123	4	0,110	9	0,091	AnzahlScheinSem2
CPSem2					2	0,113	5	0,102	6	0,105	CPSem2
NoteSem2					3	0,088	3	0,117	7	0,100	NoteSem2
NBSem2					16	0,023	21	0,024	29	0,021	NBSem2
NTSem3					24	0,014	27	0,017	16	0,043	NTSem3
AnzahlScheinSem3							1	0,136	3	0,116	AnzahlScheinSem3
CPSem3							2	0,126	4	0,110	CPSem3
NoteSem3							6	0,088	8	0,099	NoteSem3
NBSem3							18	0,029	24	0,029	NBSem3
NTSem3							28	0,015	34	0,015	NTSem3
4CPSem									1	0,122	4CPSem
AnzahlScheinSem4									2	0,119	AnzahlScheinSem4
NoteSem4									5	0,106	NoteSem4
NTSem4									25	0,026	NTSem4
NBSem4									28	0,022	NBSem4

Im Vergleich von Tabelle 44 mit Tabelle 46 verändert sich die Rangfolge der meisten Variablen an der UNI nur geringfügig. Die größten Veränderungen der Rangfolge ergeben sich nach dem ersten und zweiten Semester. Zwischen den drei Betrachtungspunkten zu Semester „0“ bis „2“ verändern sich die Kohorten durch frühe Studienabbrecher am stärksten. Bei folgenden Variablen ist der Rang der Variablen nach Bewertung durch das Gain Ratio geringer als nach dem Kriterium Information Gain:

- AnmeldeZeit
- HZBNote
- Studiengang

Wohingegen folgende Variablen durch das Gain Ratio einen höheren Rang zugewiesen bekommen, als durch den Information Gain:

- NTSem2 und 4
- Schein 2 und 6
- StartSem

Während die binäre Variable StartSem zu allen Zeitpunkten wichtiger wurde, ist die Veränderung der Scheine 2 und 6 nach den Semestern zwei und drei und der nicht teilgenommenen Prüfungen im zweiten erst nach dem vierten Semester besonders hoch.

Tabelle 47. Gain Ratio – FH

Semester	0		1		2		3		4		Variable
	Rang	Gain Ratio	Rang	Gain Ratio	Rang	Gain Ratio	Rang	Gain Ratio	Rang	Gain Ratio	
	1	0,024	9	0,025	16	0,023	20	0,022	23	0,018	HZB
	9	0,011	17	0,016	20	0,014	27	0,011	33	0,010	Studiengang
	6	0,014	18	0,015	21	0,013	26	0,012	29	0,012	ImmaterAlter
	4	0,018	12	0,023	17	0,021	21	0,020	24	0,017	Geschlecht
	5	0,017	16	0,016	19	0,016	22	0,017	26	0,015	Land
	7	0,014	19	0,013	23	0,012	28	0,011	32	0,011	Migration
	8	0,013	20	0,012	25	0,011	29	0,011	36	0,009	Region
	3	0,021	11	0,023	15	0,023	19	0,024	21	0,022	StBAusland
	11	0,008	21	0,010	22	0,012	25	0,013	27	0,014	Akad. Vorleistung
	10	0,009	22	0,008	27	0,006	35	0,004	40	0,003	Studienform
	12	0,006	23	0,008	26	0,010	32	0,009	31	0,012	VorSem
	14	0,005	24	0,007	28	0,006	33	0,006	38	0,005	Geburtsort
	2	0,022	10	0,025	13	0,026	18	0,026	20	0,023	VorVersuch
	13	0,006	25	0,006	29	0,004	34	0,004	39	0,004	Kreis der HZB'
	15	0,002	26	0,001	31	0,003	36	0,002	41	0,002	StartSem
	16	0,000	27	0,000	32	0,000	37	0,000	43	0,000	HZBNote
	17	0,000	28	0,000	33	0,000	38	0,000	42	0,000	KV
			3	0,062	4	0,056	11	0,053	15	0,041	CPSem1
			2	0,066	5	0,056	13	0,048	16	0,037	AnzahlScheinSem1
			5	0,046	9	0,038	16	0,034	17	0,033	NoteSem1
			1	0,070	7	0,046	1	0,174	1	0,167	Schein1
			4	0,052	30	0,004	15	0,037	18	0,032	Schein2
			7	0,042	8	0,039	10	0,054	12	0,044	Schein3
			8	0,035	14	0,024	6	0,084	14	0,042	Schein5
			6	0,044	6	0,048	14	0,046	13	0,042	NTSem1
			13	0,021	18	0,018	23	0,016	28	0,013	NBSem1
			15	0,019	10	0,038	8	0,077	22	0,021	Schein4
			14	0,020	11	0,033	12	0,051	10	0,056	Schein6
					1	0,106	5	0,084	7	0,069	CPSem2
					2	0,102	7	0,083	9	0,062	AnzahlScheinSem2
					3	0,075	9	0,066	11	0,048	NoteSem2
					12	0,026	17	0,033	19	0,031	NTSem3
					24	0,011	30	0,011	30	0,012	NBSem2
							2	0,111	4	0,085	CPSem3
							3	0,091	6	0,076	AnzahlScheinSem3
							4	0,086	8	0,064	NoteSem3
							24	0,015	25	0,016	NTSem3
							31	0,009	34	0,010	NBSem3
									2	0,115	4CPSem
									3	0,097	AnzahlScheinSem4
									5	0,080	NoteSem4
									37	0,007	NTSem4
									35	0,010	NBSem4

Im Vergleich von Tabelle 45 und Tabelle 47 ergibt für folgende Variable an der FH eine abweichende Prognoseverwendbarkeit, die anders als an der UNI auf alle Zeitpunkte zutrifft. Folgende Variablen bekommen durch das Gain Ratio einen niedrigeren Rang zugewiesen, als durch den Information Gain:

- Studiengang
- ImmatrAlter
- NTSem1

Folgende Variablen bekommen durch das Gain Ratio einen höheren Rang zugewiesen, als durch den Information Gain:

- VorVersuch
- StBAusland
- HZB

2.6. Auswahl der relevanten Variablen durch CFS

Ziel des CFS-Verfahrens ist es, sowohl redundante als auch irrelevante Informationen aus den Daten herauszufiltern. Verfahren des maschinellen Lernens können zwar auch mit irrelevanten und redundanten Informationen umgehen, aber neben reduzierten Rechenzeiten und weniger Problemen durch Überanpassung (Overfitting) erleichtert die Fokussierung auf relevante Informationen die Interpretation der Modelle und reduziert datenschutzrechtliche und ggf. auch ethische Bedenken. Zudem könnte der Aufwand der Datenbereitstellung geringer werden.

Wenn abbruchgefährdete Studierende bereits zu Beginn des Studiums aktiv unterstützt werden sollen, werden folgende Variablen in die Prognose aufgenommen:

- Aus dem Bereich der Studienfachwahl und des Anmeldeverhaltens werden für die UNI die Abbruchquoten der gewählten Studiengänge, die Anzahl der Studiengänge sowie der Zeitpunkt der Anmeldung berücksichtigt.
- Aus den demographischen Informationen werden das Geschlecht und das Alter bei Immatrikulation ausgewählt und auch die Art der HZB wird als relevante Variable aufgenommen. Für die FH sind Alter und Geschlecht, die Art der HZB, das gewählte Studienfach, frühere Studienerfahrungen und die Herkunft der Studierenden relevante Informationen.

Sobald Leistungsdaten verfügbar sind, werden persönliche Kontextinformationen weniger wichtig. Das gilt für die FH in stärkerem Maße als für die UNI. Die gewählten wichtigsten Variablen variieren zwar zwischen den Semestern, die durchschnittliche Abbruchquote der gewählten Studiengänge ist an der UNI jedoch durchgängig Teil der CFS-Variablen. Hervorzuheben ist die Bedeutung der Art der HZB, also die Schulform, an der die Hochschulzugangsberechtigung erworben wurde. Die Note der HZB, obwohl für die UNI verfügbar, ist selbst zu Studienbeginn nicht Teil eines durch das CFS-Verfahren bestimmten effizienten Frühwarnsystems. Dieser Befund entspricht nicht den Erwartungen an den hohen prognostischen Wert der HZB-Note für den Studien-erfolg, muss aber vorsichtig interpretiert werden, insbesondere mit Blick auf die externe Validität der Ergebnisse (Trapmann et al., 2007; Brandstätter et al., 2006; Gold et al., 2005). Jedoch könnte das Ergebnis ein Hinweis auf die Nichtvergleichbarkeit von HZB Noten im Vergleich der Schulformen sein. Das Ergebnis des CFS-Verfahrens ist in Tabelle 48 und Tabelle 49 dargestellt.

Tabelle 48. Relevante Informationen nach Semester - UNI

Semester:	0	1	2	3	4
HZB	X	X		X	
StgAbbruchquote	X	X	X	X	X
Studiengang	X				
ImmtrAlter	X		X		X
Geschlecht	X				
VorSem				X	
StartSem	X				
AnmeldeZeit	X				
CPSem1		X	X	X	X
NoteSem1		X	X	X	
NTSem1		X	X	X	
CPSem2			X	X	X
NoteSem2			X	X	X
NTSem3					X
CPSem3				X	X
NoteSem3				X	X
4CPSem					X
NoteSem4					X
AnzahlScheinSem1		X			
AnzahlScheinSem2			X	X	X
AnzahlScheinSem3				X	X
AnzahlScheinSem4					X
Schein1		X			
Schein2			X		
Schein3				X	X
Anzahl der Variablen	(7)	(7)	(9)	(13)	(14)

Tabelle 49. Relevante Informationen nach Semester - FH

Semester:	0	1	2	3	4
HZB	X	X		X	
StBAusland	X		X		X
Studiengang	X				
ImmatrAlter	X				
Geschlecht	X				
VorVersuch	X				
Land	X				
Akad. Vorleistung					X
StartSem			X		
CPSem1		X	X	X	
NTSem1		X	X	X	X
NoteSem1		X			
CPSem2			X	X	X
NoteSem2			X		
CPSem3				X	X
NoteSem3				X	
4CPSem					X
NoteSem4					X
AnzahlScheinSem1		X			
AnzahlScheinSem2			X	X	
AnzahlScheinSem3				X	X
AnzahlScheinSem4					X
Schein1		X	X	X	X
Schein2		X			X
Schein3		X			
Schein4				X	
Schein5				X	
Schein6					X
Anzahl der Variablen	(7)	(8)	(8)	(11)	(12)

Variablen, die die Studienleistungen beschreiben, sind hingegen erwartungsgemäß unter den relevanten Variablen zu finden. So werden jeweils die aktuellsten Studienleistungen gemessen in ECTS-Punkten, Durchschnittsnote, Anzahl bestandener Prüfungen und ob die am höchsten mit Studienerfolg korrelierten Prüfungen des Studienganges bestanden wurden vom CFS-Verfahren ausgewählt. ECTS-Punkte und Durchschnittsnoten der Semester sind in allen Semestern relevant und neuere Informationen ersetzen ältere. Aber auch Studienleistungen der Vergangenheit werden berücksichtigt. Nicht berücksichtigt werden hingegen an beiden Hochschulen die Anzahl nicht bestandener Prüfungen pro Semester. Das gleiche gilt für die Anzahl der angemeldeten, aber nicht teilgenommenen Prüfungen in höheren Semestern. Allerdings ist die Anzahl der nicht teilgenommenen Prüfungen im ersten Semester auch noch in höheren Semestern eine weitere wichtige Information. So wird beispielsweise bei der Prognose des Studienerfolgs im vierten Semester die Anzahl abgemeldeter bzw. nicht teilgenommener Prüfungen des ersten Semesters unter den relevanten Variablen geführt. Dem Studienverhalten im ersten Semester kommt damit an beiden Hochschulen

ein besonderer Stellenwert zu. In diesem Kontext ist auch der stetig sinkende Information Gain in Tabelle 44 und Tabelle 45 sowie die sinkende Precision des Merkmals 4 in Tabelle 34 zu deuten. Ob angemeldete Prüfungen angetreten werden, oder nicht, ist im ersten Semester ein Prädiktor für den Studienerfolg, nicht aber in den Folgesemestern.

2.7. Faktoranalyse

Die nach Informationsgehalt vergebene Reihenfolge, insbesondere des Gain Ratio, bestimmt maßgeblich die Auswahl eines durch CFS reduzierten Variablenauswahl. Variablen mit besonders hohem Informationsgehalt werden in der Regel berücksichtigt. Allerdings wird von dieser Reihenfolge zuweilen abgewichen. Eine Erklärung hierfür ist, dass vorliegende Variablen Substitute sind. In dem Fall ist ihr prognostischer Wert redundant und eine zusätzliche Berücksichtigung erhöht die Genauigkeit des Frühwarnsystems nicht – im Gegenteil, es erhöht sich die Dimensionalität und Komplexität.

Um die Wechselwirkung der Variablen untereinander zu untersuchen, wird eine Faktoranalyse durchgeführt. Bei dieser Faktoranalyse wird stets nur die höchste Ladung als Zuordnung zu einem Faktor ausgegeben. Werte unter 0,4 werden nicht berichtet.

Tabelle 66 bis Tabelle 75 im Anhang zeigen die Einteilung der vorliegenden Studierendendaten des Semesters null bis vier beider Hochschulen sowie die MSA der Variablen und das KMO-Kriterium der Datensätze. Die KMO-Kriterien der UNI liegen zwischen 0,64 („mediocre“) im nullten Semester und 0,84 („meritorious“) im vierten Semestern und die KMO-Kriterien der FH zwischen 0,64 („mediocre“) im zweiten Semester und 0,73 („middling“) im vierten Semester.

An der UNI werden im nullten Semester sechs, im ersten neun und im vierten Semester elf Faktoren gebildet. Hiervon bilden fünf Faktoren demografische Variablen ab, die über die Semester 0 bis 4 unverändert bleiben:

- Herkunft
- Schulbildung
- Akademische Historie
- Beruflicher Hintergrund
- Zeitpunkt der Studienaufnahme

Die Studienfachwahl hingegen erweist sich nur zu Studienbeginn als eigenständiger Faktor. Sobald in den darauffolgenden Semestern Leistungsdaten hinzukommen, wird dieser Faktor nicht mehr gebildet. Die Leistungsdaten werden zu folgenden Faktoren zusammengefügt:

- Studienfortschritt des ersten Semesters sowie die Durchschnittsnoten aller Semester
- Studienfortschritt (CP und Anzahl Prüfungen) der Semester 2-4
- Anzahl zurückgetretener Prüfungen in den Semestern 1-4
- Anzahl nicht bestandener Prüfungen in den Semestern 1-4
- Anzahl bestandener besonders relevanter Prüfungen

Für eine möglichst frühzeitige Identifikation sind die Semester null und eins besonders relevant. Insbesondere da nur zu Studienbeginn noch alle Studienabbrecher, allen Absolventen gegenübergestellt werden können. Bereits im ersten und zweiten Semester verändert sich die Kohortenzusammensetzung durch frühe Abbrecher erheblich. Wohingegen das Zusammenspiel von Leistungsdaten untereinander und die Bedeutung der Erstsemesterleistungen im späteren Studienverlauf nur in hohen Semestern überprüft werden können. Daher erfolgt eine besondere Betrachtung der Semester null, eins und vier.

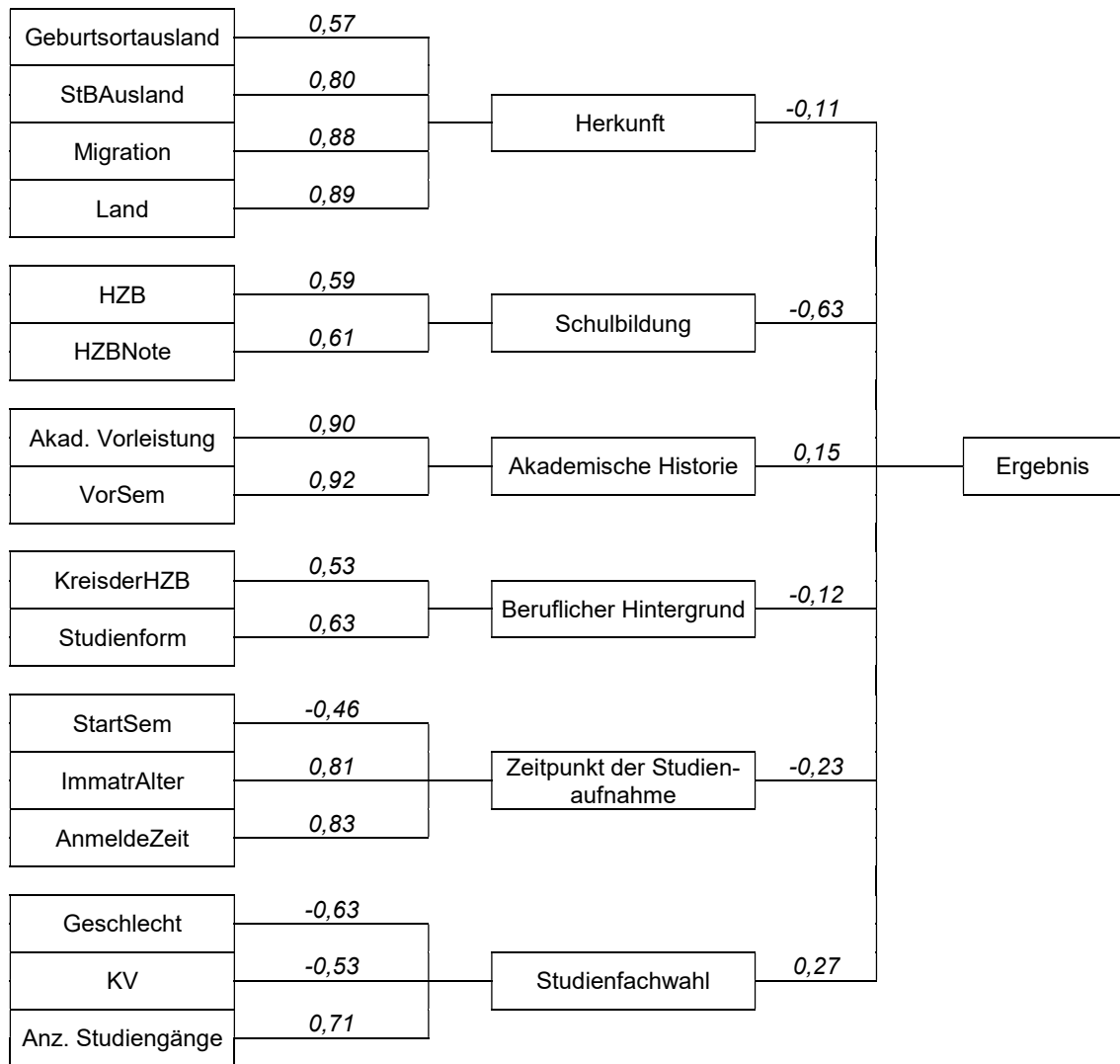


Abbildung 10. Faktoranalyse UNI Semester 0

Im ersten Semester stellt die Schulbildung, bestehend aus Schultyp (im Wesentlichen die AHR des Gymnasiums oder der Gesamtschule, vgl. Teil II, Abschnitt 2.3.4) und HZB Note den wichtigsten Faktor dar. Deutlich schwächer ist der Einfluss der Studienfachwahl. Die Anzahl der Studiengänge ist insofern mit dem Geschlecht korreliert, als dass viele als Kombibachelor angebotene Studiengänge von Frauen studiert werden. Der Faktor „Zeitpunkt der Studienaufnahme“ beinhaltet, in welchem Alter das betrachtete Bachelorstudium, im Sommer- oder Wintersemester angetreten wurde und wie groß der zeitliche Abstand zwischen Immatrikulation und Semesterstart war.

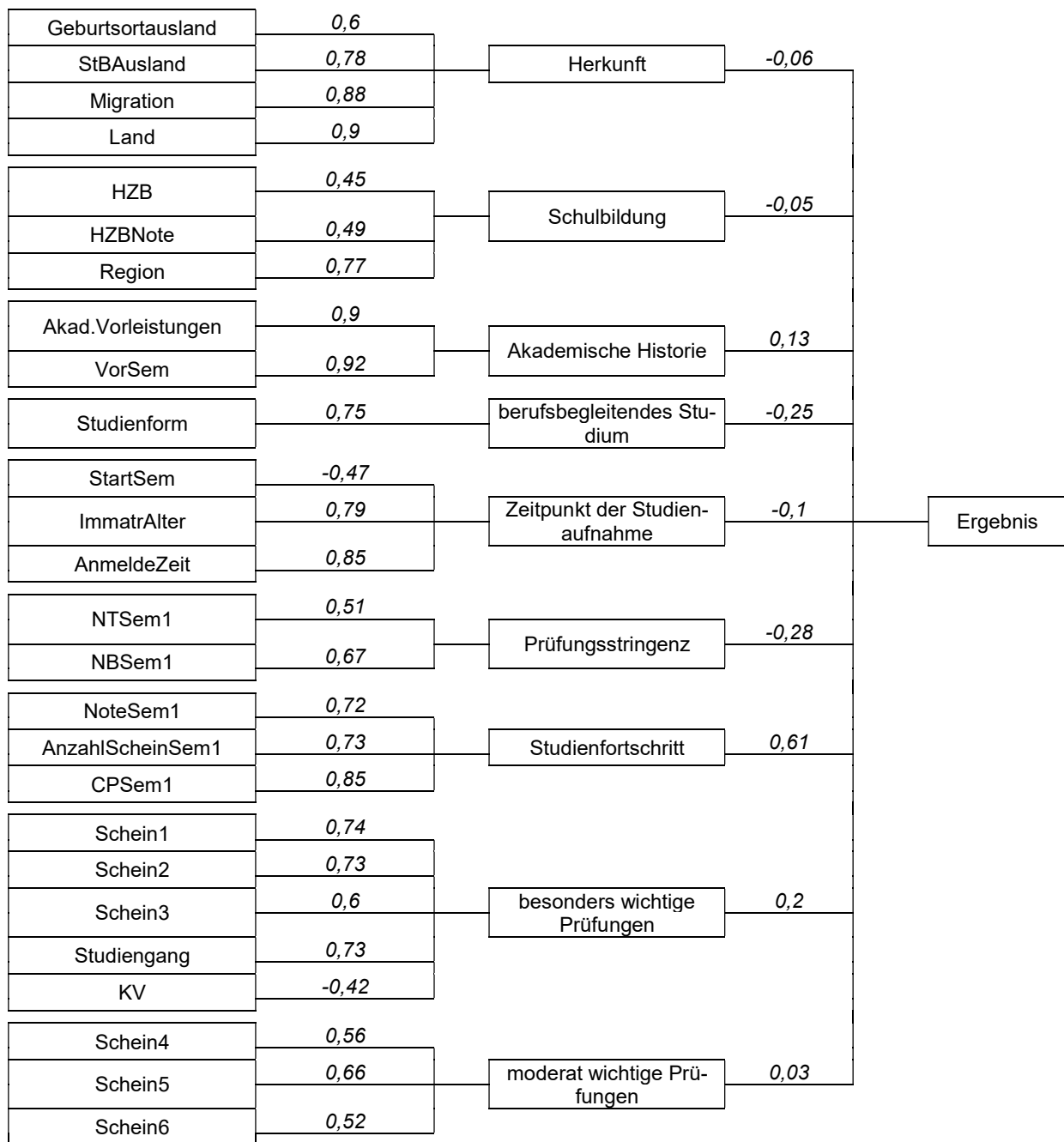
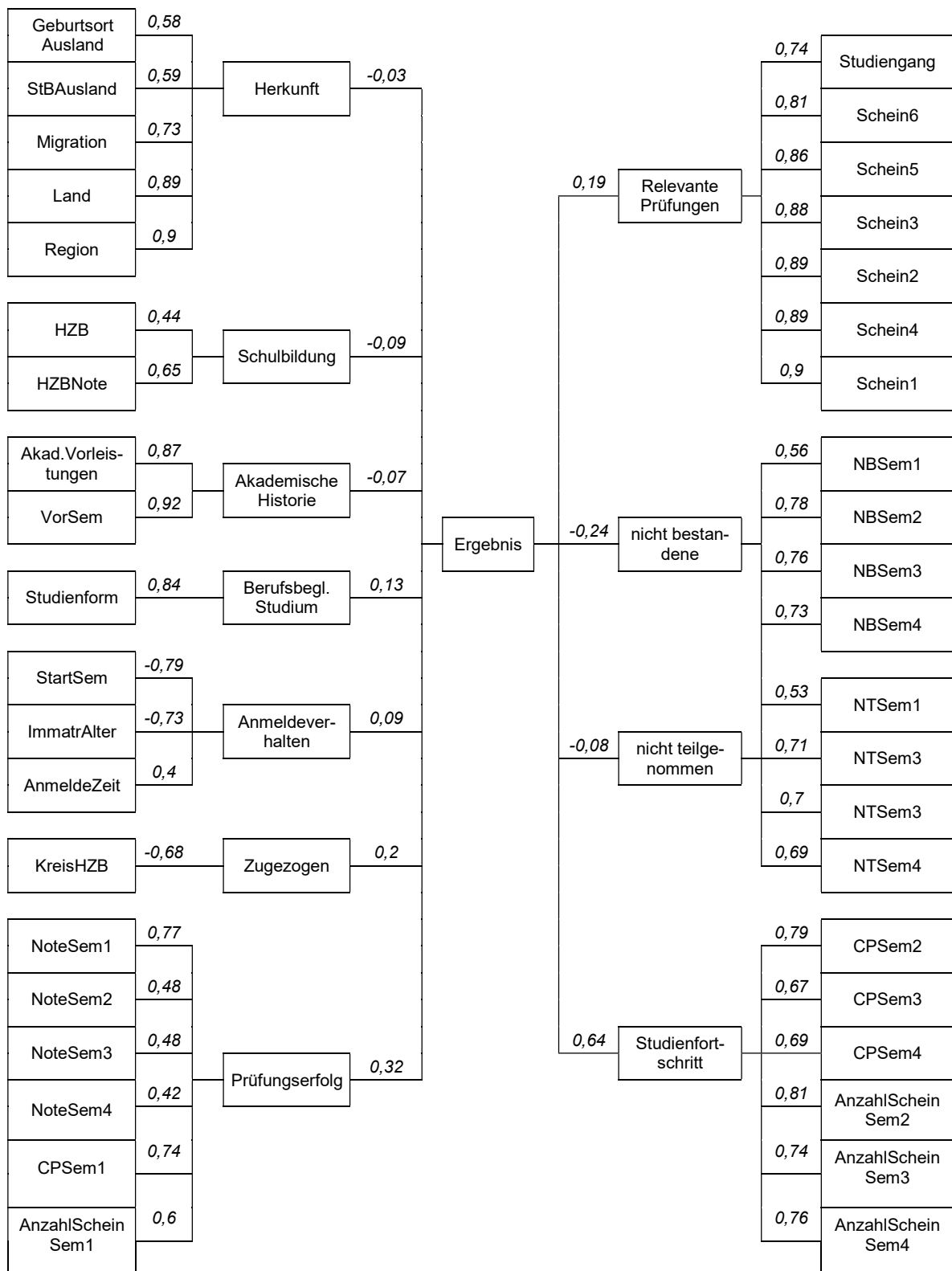


Abbildung 11. Faktoranalyse UNI Semester 1

Mit nun verfügbaren Leistungsdaten verlieren demografische Faktoren, insbesondere die Schulbildung deutlich an Bedeutung. Der bedeutsamste Faktor ist der Studienfortschritt, gemessen in CP, der Anzahl an Prüfungen und der Durchschnittsnote des ersten Semesters. In einem weiteren Faktor wird die Stringenz, mit der angemeldete Prüfungen bestanden werden, dargestellt. Dieser Faktor setzt sich zusammen aus der Anzahl angemeldeter aber nicht angetretener Prüfungen und der Anzahl angemeldeter sowie angetretener, jedoch nicht bestandener Prüfungen. Auch in den Semestern eins bis drei erfolgt eine vergleichbare Einteilung in Faktoren. Dabei wird der

Studienfortschritt, Indikatoren der Prüfungsstringenz und bestandene, besonders relevanten Prüfungen in getrennte Faktoren eingeordnet. Dem Studienfortschritt des ersten Semesters kommt insofern eine besondere Bedeutung zu, als dass er in den Semestern 2-4 nicht mit dem Studienfortschritt der Folgesemester in einen gemeinsamen Faktor geordnet wird, sondern auch weiterhin den Notendurchschnitten des aktuellen und vorangegangenen Semesters zugeordnet bleibt.

Eine Sonderstellung nimmt der Ort der HZB ein, dieser wird in Abhängigkeit des Betrachtungssemesters einmal dem Faktor der Wahl eines berufsbegleitenden Studiums (Semester 0), der Anzahl nicht bestandener oder zurückgetretener Prüfungen (Semester 1) oder eigenständig (Semester 4) zugeordnet. Auch das Geschlecht und die Krankenversicherung werden in den Betrachtungssemestern 0 bis 4 nicht immer demselben Faktor zugeordnet.



Hinweis: Um alle Faktoren und Variablen auf einer Seite darzustellen, werden diese auch rechts der abhängigen Variable „Ergebnis“ dargestellt.

Abbildung 12. Faktoranalyse UNI Semester 4

Für die FH ergeben sich geringe Abweichungen zur Uni der nicht leistungsba-
sierten Faktoren:

- Migrationshintergrund
- Herkunft
- Vollzeit oder berufsbegleitende Studierende
- Akademische Historie
- Wahl des Studienganges

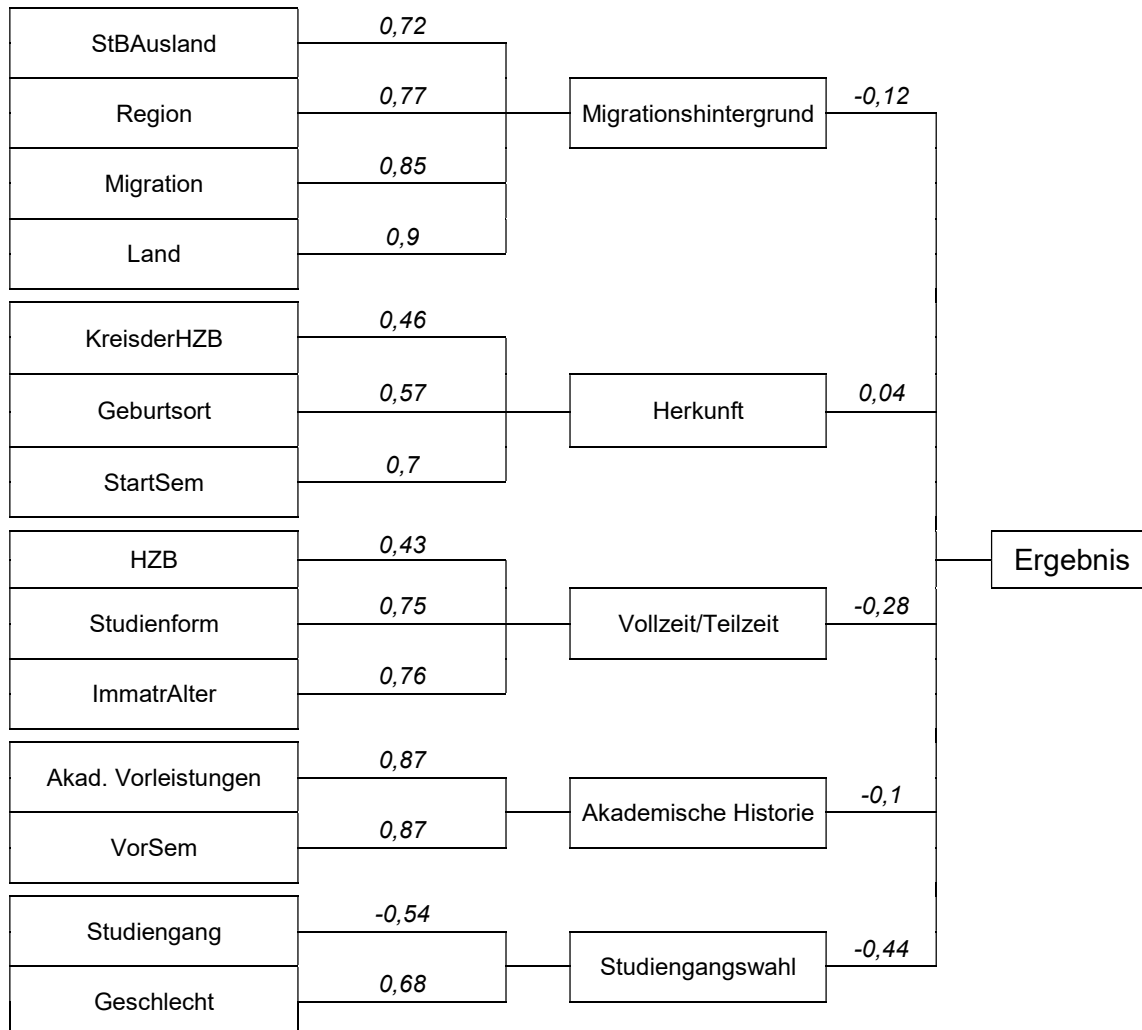


Abbildung 13. Faktoranalyse FH Semester 0

Faktor drei setzt sich im nullten und ersten Semester aus der Art der HQB, der Studienform (Vollzeit oder berufsbegleitend) und dem Alter bei Immatrikulation zusammen. Die Zusammensetzung von Faktor 3 erklärt sich aus den Besonderheiten der FH. Anders als an der UNI können alle Studiengänge der FH auch berufsbegleitend studiert werden. Durchschnittlich sind berufsbegleitende Studierende zu Studienbeginn älter als Vollzeitstudierende. Weiterhin - ebenfalls anders als an der UNI - genügt

dem FH-Studium als Hochschulzugangsberechtigung eine dreijährige Berufserfahrung nach abgeschlossener Berufsausbildung. Beruflich qualifizierte Studierende sind weit überwiegend berufstätig und wählen die berufsbegleitende Studienform.

Von gemäßigtem Einfluss ist die Studiengangswahl und die Entscheidung ob in Vollzeit oder berufsbegleitend studiert werden soll. Die verbleibenden Faktoren sind von geringem Einfluss. Dieses Ergebnis steht im Einklang mit vorangegangenen Ergebnissen. Die Prognosegenauigkeit der FH in Semester 0 ist deutlich geringer als an der UNI oder der Folgesemester an beiden Hochschulen. Dies überrascht nicht, denn an der FH liegen zu Studienbeginn weniger relevante Informationen als an der UNI vor.

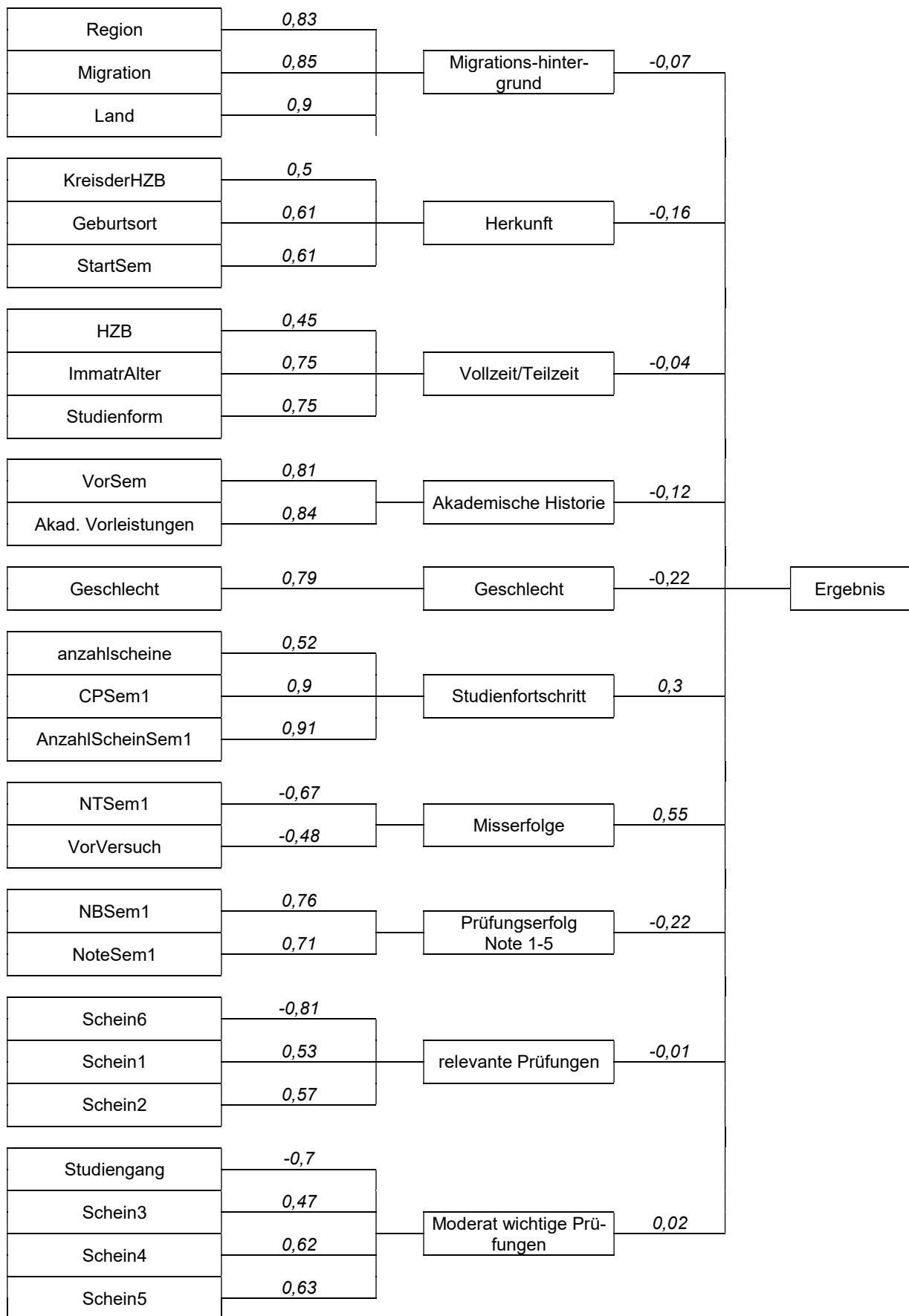


Abbildung 14. Faktoranalyse FH Semester 1

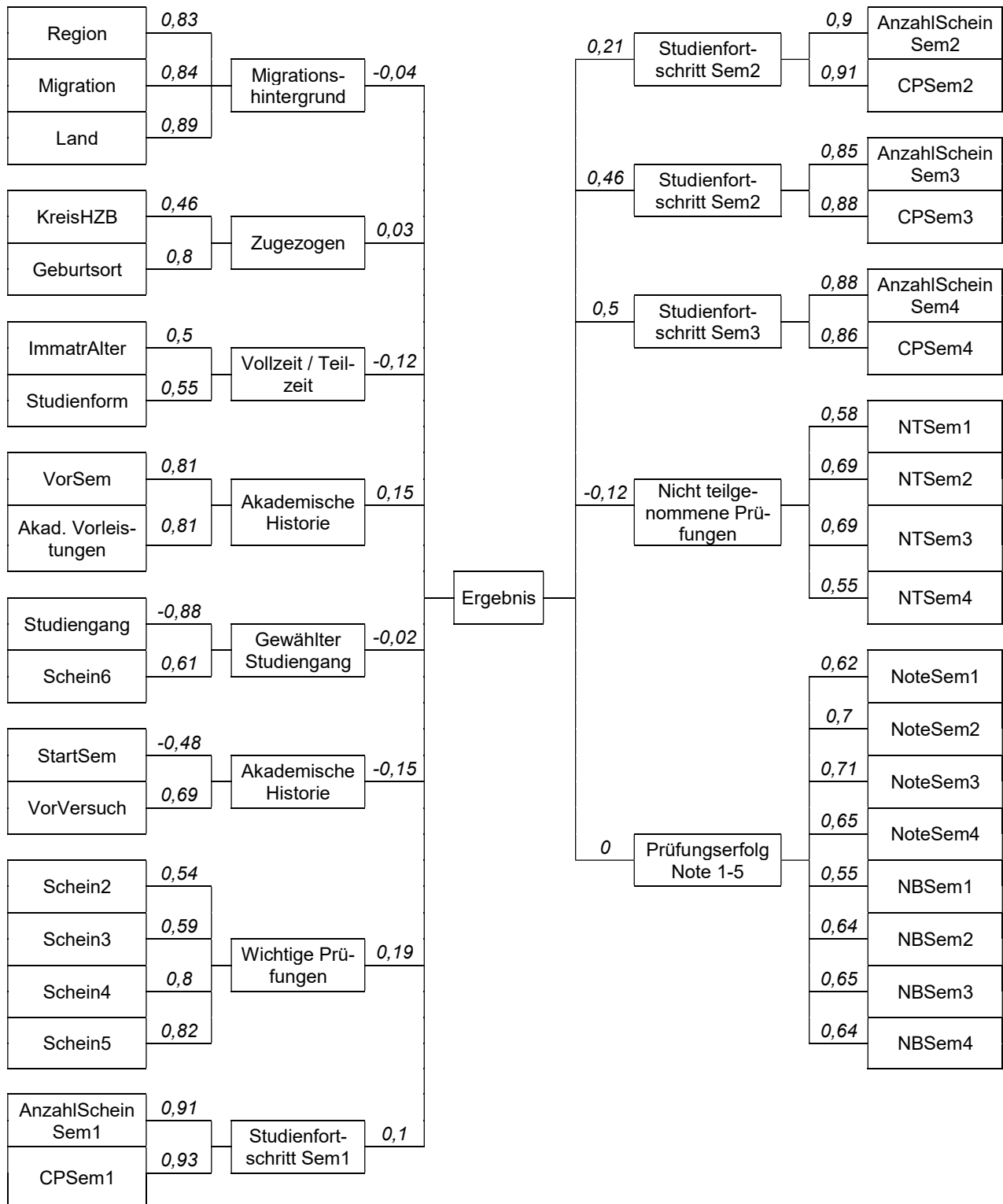
Sobald nach dem ersten Semester erste Leistungsdaten verfügbar werden, bilden diese den wichtigsten Faktor. Allerdings sind es nicht Indikatoren des Studienfortschritts, sondern Indikatoren des Misserfolgs bestehend aus der Anzahl nicht teilgenommener Prüfungen und der Anzahl vorheriger Studienversuche an der gleichen Hochschule.

Auch an der FH nimmt das Geschlecht eine Sonderstellung ein. Während im nullten Semester eine Zuordnung zum Faktor „Wahl des Studienganges“ erfolgt – die FH bietet männer- wie frauendominierte Studiengänge an – lädt der Studiengang im ersten Semester stärker auf bestandene relevante Prüfungen und das Geschlecht bildet einen eigenen Faktor.

In Semester vier der FH ergeben sich folgende leistungsorientierte Faktoren:

- Studienfortschritt des ersten Semesters
- Studienfortschritt des zweiten Semesters
- Studienfortschritt des dritten Semesters
- Studienfortschritt des vierten Semesters
- Anzahl angemeldeter aber nicht angetretener Prüfungen aller Semester
- Durchschnittsnoten und nicht bestandener Prüfungen aller Semester
- Anzahl bestandener besonders relevanter Prüfungen

Der Studienfortschritt, gemessen in ECTS Punkten und der Anzahl bestandener Scheine, werden nach Semester getrennt in eigene Faktoren eingeordnet. Einen semesterübergreifenden Faktor bilden die Durchschnittsnoten der Semester zusammen mit der Anzahl nicht bestandener Prüfungen. Die Anzahl der nicht teilgenommenen Prüfungen der Semester hingegen bildet einen eigenen Faktor.



Hinweis: Um alle Faktoren und Variablen auf einer Seite darzustellen, werden diese auch rechts der abhängigen Variable „Ergebnis“ dargestellt.

Abbildung 15. Faktoranalyse FH Semester 4

Auch an der FH haben Informationen über den jüngsten Studienfortschritt den höchsten Einfluss. Bemerkenswert ist jedoch, dass der Prüfungserfolg, gemessen an

den Durchschnittsnoten aller Semester, sowie der Anzahl nicht bestandener Prüfungen, nur einen sehr geringen Einfluss hat.

An beiden Hochschulen bilden die gebildeten Faktoren der nicht leistungsbasier-ten Informationen den Werdegang der Studierenden ab. Der geografischen Herkunft des Studierenden folgt die schulische Laufbahn, mit der Schulwahl sowie einer erziel-ten Abschlussnote. Hierauf folgt ein dem Betrachtungsstudium eventuell vorangegan-genes Studium. Dabei ist auch die Anzahl der Semester in einem vorangegangenen Studium bekannt. Anschließend folgen die Entscheidungen bei Studienstart. Hierunter fällt in welchem Alter die Studienentscheidung realisiert wird, wie frühzeitig die Bewer-bung um einen Studienplatz eingegangen ist und ob das Studium im Sommer- oder Wintersemester aufgenommen werden soll. Anders als bei leistungsbasier-ten Informa-tionen ist es nicht so, dass jüngere Informationen einen höheren prognostische Wert aufweisen. Werden aber leistungs-basierte Informationen hinzugefügt, nimmt die La-dung nicht leistungs-basierter Faktoren erheblich ab. Eine besonders hohe Ladung ha-ben auch in dieser Betrachtung der Studienfortschritt, gemessen in CP pro Semester. Abweichungen zwischen beiden Hochschulen ergeben sich beim Studienfortschritt in-sofar, als dass an der UNI der Studienfortschritt des ersten Semesters und der da-rauffolgenden Semester keinen gemeinsamen Faktor bilden. An beiden Hochschulen ist die Einteilung nicht bestandener oder nicht angetretener Prüfungen identisch. Diese bilden über alle Semester hinweg jeweils einen Faktor.

In Gegenüberstellung der durch CFS getroffenen Variablenauswahl mit den ge-bildeten Faktoren ergibt sich für die UNI für die Semester null bis vier (siehe Tabelle 48 und Tabelle 66 bis Tabelle 70), dass die getroffene Variablenauswahl etwa die Hälfte aller Faktoren mit mindestens einer Variable berührt. Im nullten Semester wer-den Variablen aus den wichtigsten drei der sechs gebildeten Faktoren ausgewählt. Der am stärksten mit der abhängigen Variable verbundene Faktor, die Schulbildung (F2), wird mit der Art der HZB abgebildet, während der zweitwichtigste Faktor (F6) mit dem Geschlecht und dem Studiengang vertreten wird. Der dritte Faktor (F5) repräsentiert die bei Studienbeginn getroffenen Entscheidungen (Studienstart im WiSe/SoSe, zeit-licher Abstand zwischen Bewerbung und Studienstart und Alter bei Immatrikulation).

Auch im ersten Semester werden Variablen aus etwa der Hälfte an Faktoren ausgewählt, in diesem Fall vier von neun Faktoren. Der wichtigste Faktor, der Studienfortschritt (F6) im ersten Semester, ist mit allen Variablen - Note, Anzahl Scheine und ECTS Punkten – vertreten. Der zweitwichtigste Faktor (F7), die akademischen Schlechtleistungen, wird mit der Anzahl der Rücktritte repräsentiert. Die Faktoren drei und vier (F8 & F2) verlassen die Rangfolge nach Ladung auf das Ergebnis und beinhalten Informationen über das Bestehen der relevantesten Prüfung sowie der Schulbildung. Die Schulbildung wird erneut nur mit der Art, nicht der Note der HZB vertreten.

Im vierten Semester werden von elf Variablen aus fünf Faktoren in ein reduziertes Setting aufgenommen. Wie auch im ersten Semester bildet der Studienfortschritt (F7 & F8) den wichtigsten Faktor. Aus beiden Faktoren werden die CP aller Semester, die Anzahl der Prüfungen und die Durchschnittsnote der Semester zwei bis vier ausgewählt. Der dritte Faktor (F11) beinhaltet, ob besonders relevante Prüfungen des Studiums bestanden wurden. Der vierte im reduzierten Setting angesprochene Faktor (F5) beinhaltet das Anmeldeverhalten.

Ein nach CFS reduziertes Variablensetting setzt sich damit an der UNI aus den folgenden Faktoren zusammen:

Tabelle 50: Faktoren des reduzierten Variablensetting der UNI

Sem 0	Sem 1	Sem 2	Sem 3	Sem 4
Schulbildung	Schulbildung	Anmeldeverhalten	Schulbildung	Zeitpunkt der Studienaufnahme
Zeitpunkt der Studienaufnahme	Prüfungs-stringenz	Studienleistung Semester 1	akademische Historie	Prüfungserfolg
Studienfachwahl	Studienfortschritt	Studienleistung Semester 2	Studienleistung Semester 1	Studienfortschritt Semester 2-4
	relevante Prüfungen	Rücktritte	Studienleistung Semester 2-3	Nicht teilgenommene
		relevante Prüfungen	Rücktritte	relevante Prüfungen
			relevante Prüfungen	

Auch an der FH werden, durch die getroffene Variablenauswahl (siehe Tabelle 49 und Tabelle 71 bis Tabelle 75), etwa die Hälfte der gebildeten Faktoren angesprochen.

Tabelle 51: Faktoren des reduzierten Variablensetting der FH

Sem 0	Sem 1	Sem 2	Sem 3	Sem 4
Studiengangswahl	Vollzeit / Teilzeit	StartSemester	Schulbildung	Akad. Historie
Vollzeit / Teilzeit	Prüfungserfolg Note 1-5	Prüfungserfolg Note 1-5	Studienfortschritt	Studienfortschritt
Akad. Historie	Studienfortschritt	Studienfortschritt	Nicht teilgenommene Prüfungen	Prüfungserfolg Note 1-5
	Relevante Prüfungen	Nicht teilgenommene Prüfungen	relevante Prüfungen	Nicht teilgenommene Prüfungen
		relevante Prüfungen		relevante Prüfungen

Der hohe Stellenwert des akademischen Studienfortschritts zeigt sich an beiden Hochschulen. Alle im Betrachtungssemester gebildeten Faktoren des Studienfortschritts, sind durch mindestens eine Variable enthalten. Sofern Studienleistungen vergangener Semester vorliegen, werden auch diese aufgenommen. Auch die Durchschnittsnote des jeweiligen Semesters wird, sofern verfügbar, um die Noten vergangener Semester ergänzt, jedoch nur an der UNI. An der FH erfährt die Durchschnittsnote eine geringere Beachtung und es wird nur die aktuelle Note aufgenommen. Der durch die aktuelle Note angesprochene Faktor beinhaltet an der FH auch nicht bestandene Prüfungen – nicht jedoch an der UNI. An dieser stellen diese einen eigenen Faktor dar. Teil eines reduzierten Variablensettings sind nicht bestandene Prüfungen an beiden Hochschulen nicht - im Gegensatz zu nicht angetretenen Prüfungen an beiden Hochschulen. Die Anzahl der nicht angetretenen Prüfungen im aktuellen und vorangegangenen Semestern bilden einen eigenen Faktor, der an beiden Hochschulen in allen Betrachtungssemestern Berücksichtigung findet. Dabei kommt an beiden Hochschulen der gewissenhafte Antritt zu Prüfungen früher Semesters höhere Aufmerksamkeit zu, als der späteren Semester. An beiden Hochschulen werden nicht angetretene Prüfungen des ersten Semesters in allen Betrachtungssemester herangezogen. Eine Ausnahme bildet lediglich das vierte Semester der UNI, hier werden nicht angetretene Prüfungen des zweiten Semesters eingesetzt.

2.8. Prognosegenauigkeit bei reduzierter Information

Abschließend wird die Prognosegüte von FragSte unter Verwendung aller verfügbaren Informationen mit der Prognosegenauigkeit unter Verwendung der durch das CFS-Verfahren gewählten Variablen getestet. Da dem Studienfortschritt in ECTS-Punkten pro Semester eine besondere Bedeutung zukommt und Hochschulen angehalten sind, ein ECTS-Monitoring aufzubauen, wird außerdem die Prognosegüte eines reinen ECTS-Monitorings ermittelt. Hierbei werden die in FragSte eingesetzten Prognoseverfahren ausschließlich mit ECTS-Punkten bis zum jeweiligen Fachsemester durchgeführt. Das reine ECTS-Monitoring wird abschließend um die Durchschnittsnote pro Semester ergänzt, damit auch das Leistungsniveau der Studierenden abgebildet werden kann. Tabelle 52 fasst die Ergebnisse zusammen.

Tabelle 52. Prognosegüte ausgesuchter Prognosemodelle der Kohorten 2010-2012

Sem:	FragSte			CFS			ECTS			ECTS + Note			
	Acc	Rec	Prec	Acc	Rec	Prec	Acc	Rec	Prec	Acc	Rec	Prec	
UNI	0	74,86	76,94	77,18	71,71	74,02	74,20	-	-	-	-	-	-
	1	81,37	80,86	80,97	80,52	80,04	80,15	78,11	76,39	78,11	78,23	77,80	77,86
	2	84,15	80,86	81,00	84,04	80,73	80,89	83,76	80,35	80,46	83,54	80,28	80,36
	3	87,80	84,04	84,17	87,01	82,93	83,10	86,29	82,12	82,15	86,83	82,72	82,85
	4	89,55	83,13	83,31	89,05	82,11	82,34	88,56	81,43	81,57	89,12	82,53	82,53
FH	0	65,74	47,24	47,59	65,34	46,49	46,95	-	-	-	-	-	-
	1	80,85	64,27	64,47	79,53	61,70	62,08	79,20	57,25	62,44	79,44	60,51	62,14
	2	86,42	67,89	67,97	85,54	65,59	65,85	85,18	65,04	65,04	85,03	64,54	64,72
	3	90,13	71,29	71,29	89,78	70,01	70,37	89,13	67,92	68,73	89,38	69,05	69,29
	4	92,97	74,32	74,47	92,46	72,08	72,90	91,88	70,69	70,69	91,87	70,63	70,77

Hinweis: Die Werte bilden die durchschnittlichen Werte über alle drei Immatrikulationskohorten ab.

In direkter Gegenüberstellung zeigt sich, dass eine Prognose nach dem ersten Semester basierend auf den CFS-Variablen im Vergleich zu FragSte nur geringfügig schlechter ist. Eine wichtige Ausnahme ist dabei die Prognose zum Zeitpunkt der Immatrikulation bzw. während des ersten Semesters. Zu diesem Zeitpunkt liegen an der UNI etwa 3 %-Punkte zwischen der Prognosegenauigkeit von FragSte und dem CFS-Modell. An der FH fällt der Unterschied deutlich geringer aus. Der Zeitpunkt, an dem erstmalig Leistungsdaten verfügbar sind, ist frühestens zu Beginn bzw. in der Mitte des zweiten Semesters. Wenn also Erst- und Zweitsemester Unterstützung erhalten sollen, dann erhöht die Nutzung aller Informationen zum Zeitpunkt der Immatrikulation die

Prognosegenauigkeit. Nach dem zweiten Semester ist hingegen ein reines ECTS Monitoring kaum ungenauer als FragSte. Auch zeigt sich, dass ein um detailliertere Leistungsdaten ergänztes ECTS Monitoring (ECTS + Note) die Genauigkeit nur geringfügig erhöht. Dieses Ergebnis bestätigen die ROC-Kurve und der AUC.

UNI

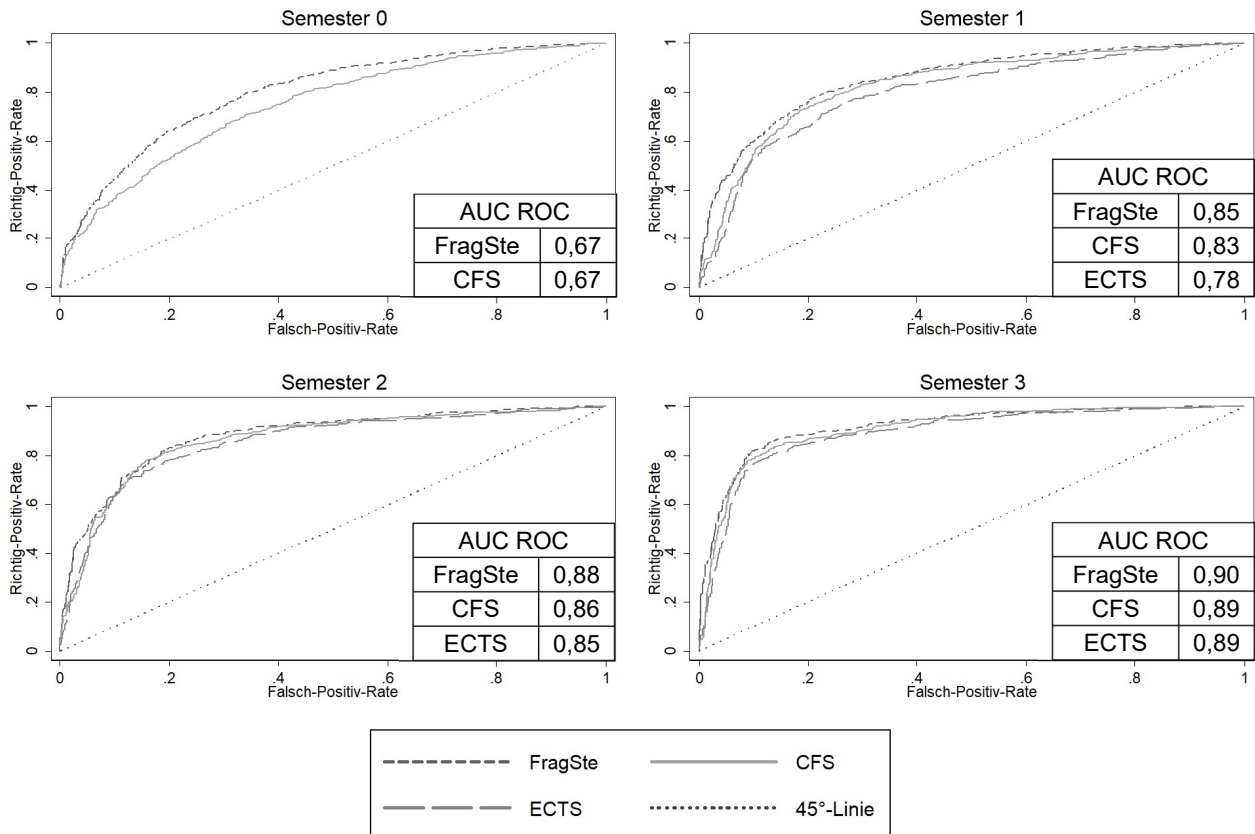


Abbildung 16. ROC UNI

FH

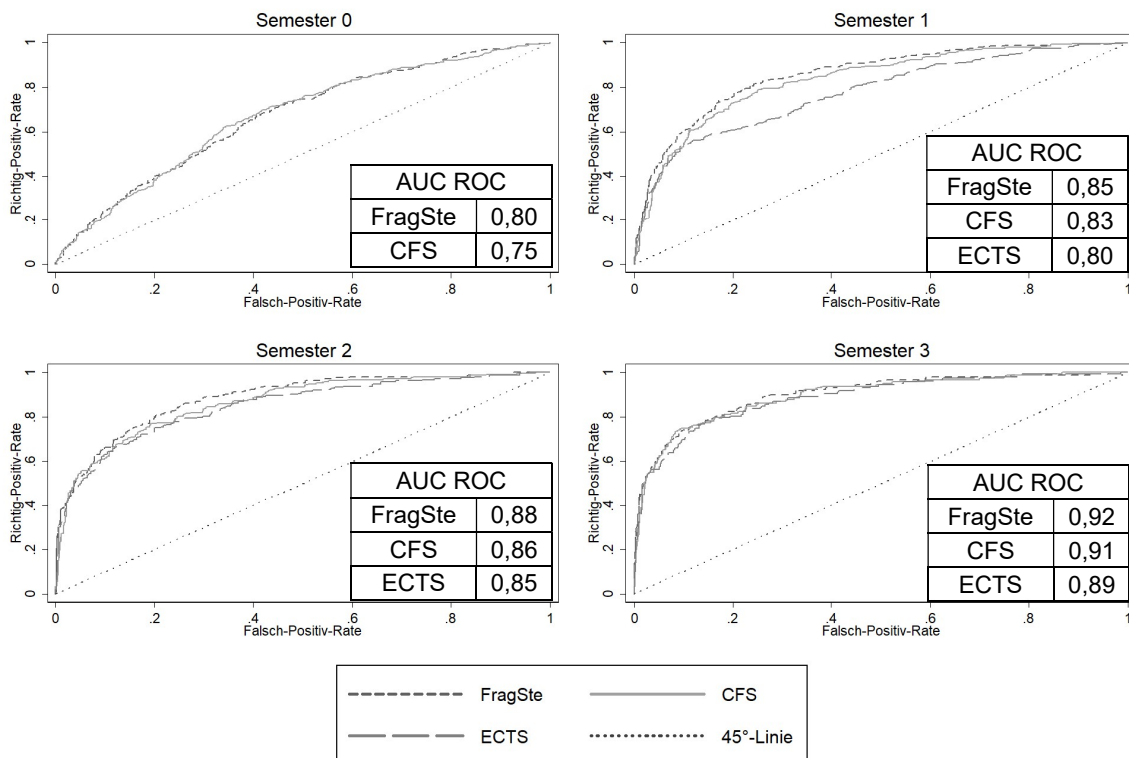


Abbildung 17. ROC FH

Die ROC-Kurven und der AUC in Abbildung 16. ROC UNI und Abbildung 17 illustrieren zudem, dass alle Verfahren zu allen Zeitpunkten deutlich genauer sind als eine zufällige Zuweisung (45°-Linie). Weiterhin zeigt sich, dass die Unterschiede in den Genauigkeiten im Studienverlauf abnehmen. Ab dem zweiten Semester sind kaum nennenswerte Unterschiede zwischen den ROC-Kurven der drei betrachteten Frühwarnsysteme erkennbar. Da jedoch die Leistungsdaten des zweiten Semesters in der Regel erst Mitte des dritten Semesters erfasst werden, ist ein ECTS-basiertes Frühwarnsystem nicht zielführend.

3. Zwischenfazit und Diskussion: Informationsgehalt von Variablen

Bei Konzeption und Implementierung eines Frühwarnsystems an einer Hochschule ist die Auswahl relevanter Informationen essentiell. Ohne relevante Informationen sind Prognosen nicht möglich. Aber auch die Verwendung von zu vielen Informationen ist nicht sinnvoll. Während Verfahren des maschinellen Lernens durchaus mit großen Datensätzen umgehen kann, ist die Datenreinigung eine ressourcenintensive Aufgabe.

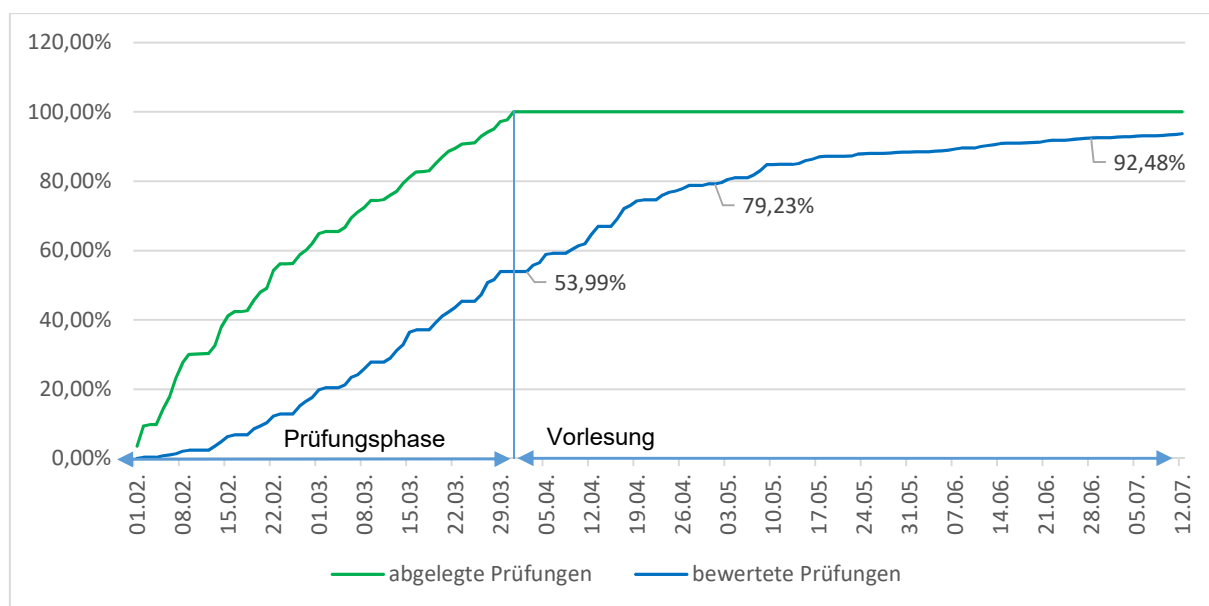
Aus Gründen der Effizienz sowie aus ethischen Gründen, lohnt es sich eine Einschränkung der erforderlichen Variablen zu erwägen. Insbesondere die Verwendung demografischer Daten zur Abbruchprädiktion könnte bei den Studierenden Bedenken aufkommen lassen, wenn sie aufgrund ihres Geschlechts oder ihrer ethnischen Zugehörigkeit und unabhängig von ihrem akademischen Potenzial als potenzielle Abbrecher eingestuft werden. Darüber hinaus können die Datenschutzbestimmungen die Umsetzung erschweren, insbesondere wenn das System demographische Informationen verwendet. Ein datensparsameres Frühwarnsystem könnte die Akzeptanz an den Hochschulen erhöhen. Aufbauend auf FragSte wurde daher geprüft, ob vergleichbar genaue Prognosen auch mit weniger datenintensiven Modellen erreicht werden können.

Die Analysen zeigen, dass die genutzten administrativen Studierendendaten zum Teil substitutiv sind, es liegen ggf. redundante Informationen vor. Werden einzelne Variablen nicht verwendet, sinkt die Genauigkeit nur in den ersten Studienmonaten nennenswert. Von besonderer Prognoserelevanz sind Leistungsdaten des vorangegangenen Semesters. Demzufolge werden bei Zusammenstellung einer datensparsameren Variablenauswahl durch CFS vornehmlich Leistungsdaten ausgewählt. Diese datensparsameren Variablenzusammenstellungen erweisen sich in höheren Semestern als vergleichbar genau. Anders fällt die Bewertung jedoch aus, wenn eine Prognose zu einem sehr frühen Zeitpunkt erfolgen soll. Dies ist eine wichtige Erkenntnis, denn nur durch frühe Interventionen kann ggf. verhindert werden, dass sich nicht erfolgreiches Studierverhalten in Folgesemestern wiederholt oder ein Studienabbruch hinausgezögert wird. Unterstützungsangebote für abbruchgefährdete Studierende bereits im ersten Semester wären somit sinnvoll. Zu diesem frühen Zeitpunkt sind jedoch noch keine Leistungsdaten verfügbar. Daher sind demografische Informationen, Informationen zur schulischen und akademischen Laufbahn vor der Immatrikulation und Entscheidungen zum Zeitpunkt der Immatrikulation von großem Wert für Frühwarnsysteme. Um die Prognosegenauigkeit für die ersten Studienmonate weiter zu erhöhen und dabei möglichst auf demografische Informationen verzichten zu können, bedarf es weiterer Informationen, die an den Hochschulen vorliegen. Zu prüfen wäre beispielsweise, ob das Anmeldeverhalten zu Prüfungen oder auch die Nutzung eventueller Online-Angebote die Prognosegenauigkeit verbessern können.

V Anpassungen des Frühwarnsystems

Studienabbrüche müssen früh erkannt werden. Der Nutzen eines Hilfsangebots hängt von der Rechtzeitigkeit ab (siehe Teil II, Abschnitt 1.2, insbesondere Abbildung 1). Studierende, die ihr Studium vor der Prognose abbrechen, können nicht mehr erreicht und unterstützt werden. An beiden Hochschulen brechen mehr als 23% der Studienabbrecher ihr Studium im ersten Semester ab. Weitere 24% und 21% (UNI und FH) der ursprünglich Immatrikulierten folgen im zweiten Semester. Etwa die Hälfte aller Studienabbrecher erreicht das dritte Semester nicht.

Weiterhin zeigt sich, dass zwischen Prüfungszeitpunkt und erstmaliger Verwendbarkeit der Information im Frühwarnsystem erhebliche Zeiträume liegen können. Diese sind Korrekturzeiten, der Erfassung der Daten durch die Verwaltung, dem Datenexport, der Datenaufbereitung und abschließenden Modellierung des Systems (Lernphase) geschuldet.



Hinweis: Es werden die Prüfungen des WiSe 2018 der UNI dargestellt.

Abbildung 18. Zeitpunkt der Verfügbarkeit von Leistungsdaten im Folgesemester

Zur Prognose genügt es nicht, wenn Teile der abgelegten Prüfungen verfügbar sind. Innerhalb der Lernphase „erlernen“ die Verfahren anhand vollständig vorliegender Leistungsdaten, Studienabbrecher von Absolventen zu unterscheiden. Dabei weisen die maschinellen Verfahren der „Studiengeschwindigkeit“, gemessen in CP, eine hohe Bedeutung zu (siehe Teil IV, Abschnitt 2.2.). Wird die Zahl der erarbeiteten CP

des vorangegangenen Semesters zu niedrig ausgewiesen, weil Daten noch nur unvollständig vorliegen, führt das zu einer zu hoch ausgewiesenen Abbruchwahrscheinlichkeit. Prognosen, basierend auf den Erstsemesterleistungen sind frühestens Mitte des zweiten Semesters verfügbar und erreichen etwa 30-40% der Studienabbrecher nicht mehr. Sofern sich die in Abbildung 1 dargestellten Studienabbrüche zu Ende des zweiten Semesters bereits in der Umsetzung befinden, werden etwa 50% der Studienabbrecher nicht mehr erreicht.

Sofern Leistungsdaten im Verlauf des zweiten Semesters verfügbar werden und Prognosen ermöglichen, benötigt auch der Prozess des Hilfsangebots sowie der Vermittlung in entsprechende Maßnahmen Zeit. Eine Teilnahme an Hilfsangeboten wird daher frühestens gegen Ende des zweiten, wahrscheinlicher zu Beginn des dritten Semesters stattfinden.

Die aufgrund hoher Prognoserelevanz sinnvoller zu verwendenden Leistungsdaten, ihre verzögerte Verfügbarkeit, die Reaktionszeit der Studierenden und die prozessualen Ablaufzeiten in der Verwaltung führen dazu, dass erst spät prognostiziert und Hilfe erst gegen Mitte des Studiums allokiert werden kann. Vor dem Hintergrund der in Abbildung 1 dargestellten zeitlichen Verlauf von Studienabbrüchen, werden an beiden Hochschulen 50% der Studierenden nicht mehr erreicht. Das bedeutet, dass Frühwarnsysteme eher eine deutlich geringere Wirksamkeit entfalten können, als die vorgestellten Genauigkeiten ausweisen. Es werden die Genauigkeiten hinsichtlich der Prognostizierbarkeit der in der Kohorte verbleibenden Studienabbrecher ausgewiesen, nicht hinsichtlich aller Studienabbrecher. In der vorliegenden Studie und in verwandten Beiträgen wird die Prognostizierbarkeit von Studienabbrechern innerhalb einer Kohorte zu einem Zeitpunkt getestet. Alle vor diesem Zeitpunkt exmatrikulierten Studierenden sind nicht mehr Teil der Kohorte. Zur Beurteilung der technischen Leistungsfähigkeit eines Frühwarnsystems ist das korrekt. Wenn aber die Leistungsfähigkeit eines Frühwarnsystems zum Zwecke der Abbruchprävention evaluiert werden soll, dann sind Studierende, die sich vor dem Prognosezeitpunkt exmatrikulierten problematisch – das Frühwarnsystem konnte diese Studierende vor dem Studienabbruch nicht identifizieren. Tabelle 53 stellt die sich ergebenden Genauigkeiten der Frühwarnsysteme dar, wenn jedem Test die Anzahl aller Studienabbrecher der Immatrikulationskohorte anstelle der Anzahl der noch in der Kohorte enthaltenen Studienabbrecher zum Zeitpunkt der Prognose zugrunde gelegt wird.

Tabelle 53: Prognosegenauigkeit unter Berücksichtigung aller Studienabbrecher

Hochschule	Sem	FragSte (AdaBoost)			PASST?!			SVM		
		ACC	REC	PREC	ACC	REC	PREC	ACC	REC	PREC
UNI	0	67,65%	65,81%	66,20%	-	-	-	-	-	-
	1	66,70%	51,59%	73,93%	63,72%	43,12%	73,32%	62,90%	47,45%	68,66%
	2	63,12%	40,04%	74,95%	57,84%	44,75%	61,02%	60,89%	41,29%	68,42%
	3	63,73%	37,15%	80,75%	57,55%	41,29%	62,08%	59,88%	38,79%	68,19%
	4	59,18%	26,85%	80,17%	53,27%	30,80%	58,18%	54,90%	27,91%	63,46%
FH	0	68,11%	51,37%	51,26%	-	-	-	-	-	-
	1	78,30%	57,36%	69,74%	72,87%	34,41%	68,00%	70,92%	49,80%	57,08%
	2	77,55%	45,73%	75,35%	62,88%	47,37%	44,49%	73,50%	40,89%	66,45%
	3	76,52%	37,77%	78,92%	73,48%	33,81%	71,67%	71,59%	38,46%	61,89%
	4	73,86%	27,25%	78,88%	73,61%	27,53%	81,44%	67,50%	29,55%	52,71%

Hinweis: FH = Kohorte 2012, UNI = Kohorte 2010

Wird beispielsweise die erste Prognose im dritten Semester mit einem der drei Frühwarnsysteme durchgeführt, so sind etwa 57-63% der Prognosen an der UNI richtig (FH 72-77%). Jedoch können an der UNI nur (noch) 40-45% der Abbrecher identifiziert werden. An der FH sind es 41-47%.

Der Prognosezeitpunkt an dem die meisten abbruchgefährdeten Studierenden identifiziert werden (Recall) steht im Spannungsfeld zwischen verfügbaren, relevanten Informationen und der Anzahl noch verbleibender Studienabbrecher. Je später der Prognosezeitpunkt im Studienverlauf ist, desto mehr Informationen liegen vor, aber desto weniger abbruchgefährdete Studierende verbleiben.

Für PASST?! ergibt sich der Beginn des dritten Semesters, wenn die Leistungsdaten des zweiten Semesters verfügbar sind, als idealer Prognosezeitpunkt. Das ist allerdings dem Wechsel des Merkmals (vgl. Teil III, Abschnitt 3.2) geschuldet. Für FragSte und SVM ist der jeweils erstmögliche Prognosezeitpunkt auch der mit der besten Abbrecheridentifikation (Recall). Eine Ausnahme stellt FragSte an der FH dar, hier ist es nicht Semester 0 sondern Semester 1, zu dem die besten Prognosen möglich sind. Das ist der geringen Prognoserelevanz demografischer Variablen an der FH geschuldet.

Da zu Studienbeginn die Abbrecheridentifikation (Recall) am höchsten ist und auch Interventionen zu Studienbeginn die höchste Wirkung entfalten (vgl. Teil IV, Abschnitt 1.2): ist es wünschenswert genaue Prognosen bereits im ersten Semester erstellen zu können.

1. Frühsterkennung innerhalb der ersten vier Monate

Die Prognosegenauigkeit des Frühwarnsystems innerhalb der ersten Studienmonate kann nur dadurch optimiert werden, dass weitere relevante Informationen hinzugefügt werden. Wie in Teil IV gezeigt, sind Prüfungsleistungen von besonderer Prognoserelevanz. Zwar sind Prüfungsleistungen selbst noch nicht verfügbar, aber es liegen schon Informationen darüber vor, ob und wie früh sich ein Studierender für welche Prüfungen angemeldet hat. Da die FH den Anmeldezeitpunkt nicht im Campusmanagementsystem erfasst, ist diese Information nur an der UNI verfügbar ist. Daher können die folgenden Untersuchungen auch nur für die UNI durchgeführt werden.

Im Folgenden soll überprüft werden, ob unter Berücksichtigung von Prüfungsanmeldungen genauere Prognosen innerhalb der ersten Studienmonate möglich sind und ab wann.

1.1. Feature Engineering

An der UNI wird der Name, das Prüfungsdatum und das Prüfungsanmeldungsdatum erhoben. Weiterhin ist bekannt ob es sich um eine beschränkt wiederholbare oder eine unbeschränkt wiederholbare Prüfung handelt. Das beinhaltet folgende potentiell prognoserelevante Informationen:

- Zeitlicher Abstand zwischen Prüfung und Anmeldung
- Zeitlicher Abstand zwischen Prüfungsanmeldung und Semesterende
- Anzahl wichtiger Prüfungen (vgl. Teil II, Abschnitt 2.1 und Tabelle 17)
- Anzahl angerechneter wichtiger Prüfungen aus vorherigem Studium
- Anzahl angemeldeter Prüfungen zu unterschiedlichen Zeitpunkten innerhalb des ersten Semesters

Die Prüfungsteilnahme ist je nach Studiengang, Fachbereich und Hochschule unterschiedlich geregelt. Prüfungen können anmeldepflichtig oder anmeldefrei sein. Für anmeldepflichtige Prüfungen besteht eine definierte Frist, die aber ebenfalls sehr unterschiedlich sein kann. Das Anmeldeverhalten der Studierenden wird durch die jeweiligen Regelungen und Fristen bestimmt. Damit hängt auch die Prognoserelevanz von den jeweiligen Regelungen und Fristen ab. Um nicht alle Regelungen und alle

Fristen aller Studiengänge und Fachbereiche einer Hochschule im Frühwarnsystem hinterlegen und pflegen zu müssen, wird anhand des Prüfungsanmeldeverhaltens der Studierenden geprüft, ob eine Anmeldepflicht bestand. Hierbei wird die Annahme getroffen, dass eine frühzeitige Prüfungsanmeldung des größten Teils der Prüfungsteilnehmer aufgrund einer Ausschlussfrist geschieht und nicht freiwillig.

Die Zeitpunkte einer Frühsterkennung innerhalb des ersten Semesters können in Tagen bis zum Semesterende angegeben werden:

Tabelle 54: Tage bis Semesterende

Tage	WiSe	SoSe
0	01.04.	01.10
15	17.03.	16.09.
30	02.03.	01.09.
45	16.02.	17.08.
60	01.02.	02.08.
75	17.01.	18.07.
90	02.01.	03.07.
105	18.12.	18.06.
120	03.12.	03.06.
135	18.11.	19.05.
150	03.11.	04.05.
165	19.10.	19.04.
180	04.10.	04.04.

Hieraus können die Variablen der Prüfungsanmeldung in Tabelle 55 gebildet werden. Zusätzlich stehen alle in Tabelle 17 ausgewiesenen demografischen Variablen zur Verfügung.

Tabelle 55: Variablen der Prüfungsanmeldung

Variable	Beschreibung
B15...B150	Anzahl beschränkt wiederholbarer Prüfungen, die die angegebene Anzahl von Tagen vor Semesterende angemeldet waren
U15...U150	Anzahl unbeschränkt wiederholbarer Prüfungen, die die angegebene Anzahl von Tagen vor Semesterende angemeldet waren
A15...A150	Anzahl aller Prüfungen, die die angegebene Anzahl von Tagen vor Semesterende angemeldet waren
V1...V9	Anzahl der neun am stärksten mit Studienerfolg korrelierten Prüfungen des Studiengangs im Betrachtungssemester aus einem vorherigen Studium
E1...E9	Anzahl der neun am stärksten mit Studienerfolg korrelierten Prüfungen des Studiengangs die zum Betrachtungszeitpunkt angemeldet waren
DummyBPrüf	Dummy Anmeldepflicht
AnmeldeVorlauf	Durchschnittlicher Abstand zwischen den Prüfungsanmeldungen und Prüfung

Hinweis: Die Zahlen 15 bis 150 kennzeichnen die Anzahl von Tagen zwischen Prüfung und Semesterende.

1.2. Einfluss des Prognosezeitpunkts im ersten Semester

Zu welchem Zeitpunkt Prognosen möglich sind, ergibt aus der Prognoserelevanz der Variablen A15 bis A150, B15 bis B150 und U15 bis U150. Je größer der Abstand zum Semesterende, desto weniger Prüfungsanmeldungen sind bereits im System erfasst. Zur Ermittlung der Prognoserelevanz wird der Information Gain verwendet (siehe Teil IV, Abschnitt 1.2). Tabelle 56 weist den Information Gain der Variablen der Prüfungsanmeldung aus. Zur besseren Einordnung ist der Information Gain demografischer Variablen ebenfalls angegeben.

Tabelle 56: Information Gain der Prüfungsanmeldungen

Prüfungsanmeldungen			Wichtige Prüfungen			Demografische Informationen			Nicht relevant		
Variable	Rang	Info Gain	Variable	Rang	Info Gain	Variable	Rang	Info Gain	Variable	Rang	Info Gain
A30	1	0,150	AnmeldVorlauf	20	0,095	StgAbbruchquote	19	0,098	V5	52	0,000
B30	2	0,150	DummyB-Pruef	25	0,037	AnmeldeZeit	21	0,054	V2	53	0,000
A45	3	0,146	E1	24	0,039	ImmaterAlter	22	0,043	U120	57	0,000
B45	4	0,146	E3	29	0,019	VorSem	23	0,040	U135	58	0,000
A60	5	0,146	E2	31	0,009	Anzahl StG	26	0,026	U195	59	0,000
B60	6	0,146	E5	32	0,009	HZBNote	27	0,024	U60	60	0,000
A75	7	0,146	E4	34	0,008	StartSem	28	0,021	U75	61	0,000
B75	8	0,146	E6	35	0,007	HZB	30	0,014	U90	62	0,000
A90	9	0,145	E8	36	0,005	Geschlecht	33	0,008	U45	63	0,000
B90	10	0,145	E9	37	0,004	Akad.Vorleist	41	0,002	U30	64	0,000
A105	11	0,144	E7	38	0,004	Migration	42	0,002	U15	65	0,000
B105	12	0,144	V7	39	0,003	Land	43	0,002	U180	66	0,000
B120	13	0,139	V8	40	0,002	Region	45	0,001	U165	67	0,000
A120	14	0,139	V4	44	0,002	Kreis der HZB'	46	0,001	U150	68	0,000
A135	15	0,137	V3	47	0,001	StBAusland	51	0,001	U105	69	0,000
B135	16	0,137	V9	48	0,001	Studienform	54	0,000			
A150	17	0,135	V1	49	0,001	GeburtsortAusland	55	0,000			
B150	18	0,135	V6	50	0,001	KV	56	0,000			

Der Information Gain nimmt mit zunehmend größeren Abstand zum Semester langsam und stetig ab. Zur Ermittlung eines Prognosezeitpunkts ist das vorteilhaft. Als Prognosezeitpunkt kann damit ein Zeitpunkt kurz vor Semestermitte erwogen werden. Das ist früh genug im Semester, um noch im ersten Semester intervenieren zu können.

Ebenfalls von relevantem Einfluss ist die Dummy-Variable in wie weit Prüfungen anmeldepflichtig waren und die durchschnittliche Anzahl an Tagen, die zwischen Prü-

fungsanmeldung und Prüfung lagen (AnmeldVorlauf) sowie die bereits bekannten relevanten demografischen Variablen (vgl. Tabelle 44). Von mittlerer Relevanz ist die Information welche wichtigen Prüfungen angemeldet wurden (E1-E9). Von vernachlässigbarer Relevanz ist die Anzahl angemeldeter unbeschränkt wiederholbarer Prüfungen U15 bis U150 und das nicht, weil die Information in den Variablen A15-A150 und B15-B150 implizit enthalten ist. Der Information Gain beurteilt nur den Einfluss einer Variable auf die Entropie der zu erklärenden Variable. Ebenfalls nur sehr geringen Einfluss hat die Information welche wichtigen Prüfungen V1 bis V9 bereits vor dem ersten Semester – demnach in einem vorangegangenen Studium - erworben wurde.

1.3. Prognosegenauigkeit

Im Folgenden wird die Prognosegenauigkeit eines Frühwarnsystems getestet, das 105 Tage vor Semesterende prognostiziert, alle eingangs beschriebenen demografischen und Prüfungsanmeldevariablen nutzt und den in Teil II, Abschnitt 2.4, beschriebenen Verfahrensaufbau besitzt. Der zeitliche Abstand von 105 Tagen zum Semesterende ist operativen Gegebenheiten der Betrachtungshochschule geschuldet. Im Wintersemester liegt dieser Zeitpunkt Mitte Dezember und ermöglicht Prognosen zu Beginn des Folgejahres. Maßnahmen oder Interventionen können so noch vor Ende der Anmeldephase und insbesondere vor der Klausurphase implementiert werden.

Benchmark ist das Frühwarnsystem FragSte unter Nutzung der demografischen Variablen. Anders als in vorherigen Teilen ist Zeitpunkt des Datenexports November 2018, da das Datum von Prüfungsanmeldungen nicht Teil des Exports 2017 ist⁷.

Tabelle 57. Benchmark Prognosegenauigkeiten FragSte im nullten Semester der UNI

Testkohorte	AdaBoost			Neural Net			Regression			BRF		
	Acc	Rec	Prec	Acc	Rec	Prec	Acc	Rec	Prec	Acc	Rec	Prec
2010	69,81	62,31	62,51	61,31	51,87	51,93	65,47	57,14	57,08	70,21	62,97	62,97
2011	71,18	68,83	68,93	61,51	40,69	63,16	67,27	64,86	64,63	69,68	67,39	67,24
2012	74,24	76,64	76,84	67,31	70,16	70,68	70,72	73,63	73,58	73,02	75,83	75,58
2013	75,92	78,99	79,33	69,77	73,87	73,92	71,61	75,52	75,47	75,07	78,50	78,45

Hinweis: alle Werte in Prozent.

⁷ Mit der veränderten Datengrundlage ergeben sich von bisherigen Ergebnisse abweichende Genauigkeiten in Tabelle 57. Die veränderte Datengrundlage ist einem Systemumstellungsprozess innerhalb der Hochschulverwaltung geschuldet.

Unter Verwendung des Prüfungsanmeldeverhaltens ergeben sich die in Tabelle 58 dargestellten Genauigkeiten.

Tabelle 58. Prognosegenauigkeiten Frühwarnsystem

Testkohorte	AdaBoost			Neural Net			Regression			BRF		
	Acc	Rec	Prec	Acc	Rec	Prec	Acc	Rec	Prec	Acc	Rec	Prec
2010	75,14	69,13	69,36	64,65	56,20	56,38	70,87	64,24	63,96	73,38	67,28	67,06
2011	75,65	73,84	73,89	65,18	61,19	63,03	73,08	71,12	71,12	75,55	73,77	73,77
2012	80,29	82,17	82,38	69,05	71,27	72,66	77,49	79,79	79,79	79,59	81,70	81,65
2013	82,30	84,73	84,78	73,02	76,77	76,81	79,73	82,62	82,52	82,44	84,91	84,86

Hinweis: alle Werte in Prozent.

Im Vergleich des Frühwarnsystem mit dem Benchmark FragSte ergeben sich etwa 5-6 Prozentpunkte bessere Ergebnisse in allen Gütemaßen. Das entspricht etwa der Genauigkeit von FragSte-Prognosen basierend auf Erstsemesterprüfungen. Damit ist es möglich, Prognose etwa drei Monate nach Studienstart mit der Genauigkeit durchzuführen, die zuvor erst Mitte bis Ende des zweiten Semesters möglich war.

2. Prognose des Vorlesungserfolgs

Eins der wesentlichen Ziele des educational data mining ist die Prognose von Lernerfolgen (Baker, 2010; Baker & Yacef, 2009). Es können die Ergebnisse unterstützender Maßnahmen wie Tutorien, das Bestehen einer Klausur oder eines Studiums prognostiziert werden (Asif et al., 2017). In den bisherigen Teilen wurde stets der Studienerfolg prognostiziert. Die Kenntnis der Abbruchwahrscheinlichkeit der Studierenden kann Hochschulverwaltungen dabei unterstützen, Ressourcen zielgerichtet zu allokalieren und auch die Wirksamkeit der Maßnahmen zu evaluieren.

Neben zentralen abbruchspräventiven Bemühungen kann es sinnvoll sein, alternativ oder flankierend dezentrale Lösungen anzubieten. Eine dezentrale Lösung kann sein, dass in der Vorlesung zielgerichtete Unterstützung für diese Vorlesung angeboten wird. In wie weit dezentral angebotene Hilfe der zentral angebotenen Hilfe überlegen ist, kann an dieser Stelle nicht überprüft werden. Aber mit Hilfe der Früherkennung wird sie überprüfbar.

Für die Konzeption einer Früherkennung innerhalb einer Vorlesung, ist in einem ersten Schritt das Label zu verändern. Weiterhin ist die Zusammenstellung der Studie-

rendendaten zu verändern. Es ist nicht mehr eine Kohorte zu einem Zeitpunkt im Studium – beispielsweise nach dem zweiten Semester – zu betrachten, sondern jetzt werden die Prüfungsteilnehmer einer Klausur betrachtet. Diese selektieren sich aus unterschiedlichen Fachsemestern. Dementsprechend liegen indifferente Leistungshistorien der Teilnehmer vor. Im Gegenzug können den in Tabelle 17 beschriebenen Variablen weitere, prüfungsspezifische Informationen hinzugefügt werden. Diese umfassen das Fachsemester sowie die Anzahl bisheriger erfolgloser und erfolgreicher Versuche (Verbesserungsversuche) die Prüfung zu bestehen. Weitere Änderungen sind nicht notwendig. Der in Teil II, Abschnitt 2.4 beschriebene Verfahrensaufbau kann unverändert bleiben.

In Tabelle 59 werden die Genauigkeiten der Prognosen eines Klausurergebnisses im zweiten Semesters exemplarisch dargestellt. Die Lerndatensatz umfasst insgesamt 2.706 Studierende mehrerer Klausurzeitpunkte. Hiervon haben 1.010 die Klausur nicht bestanden. Der Testdatensatz besteht aus 677 Teilnehmer wovon 258 die Klausur nicht bestanden haben.

Tabelle 59. Genauigkeit der Erfolgsprognose einer Zweitsemesterklausur

AdaBoost	UNI
	Klausur 2. Semester
Accuracy ^a	83,46
Recall ^a	77,91
Precision ^a	78,21
Anzahl „bestanden“	420
Anzahl „nb“	258
korrekt klassifiziert „bestanden“	331
falsch klassifiziert „nb“	87
korrekt klassifizierte „nb“	169
falsch klassifizierte „bestanden“	89
korrekt klassifizierte „bestanden“ ^a	78,81
falsch klassifizierte „nb“ ^a	20,71
korrekt klassifizierte „nb“ ^a	65,50
falsch klassifizierte „bestanden“ ^a	34,50

Hinweis: ^a Werte in Prozent. nb = nicht bestanden

Die Erfolgsprognose der exemplarisch getesteten Vorlesung entspricht etwa den Studienabbruchprognosen von FragSte nach dem dritten oder vierten Semester. Die vorliegend getestete Prüfung ist sehr gut prognostizierbar. Allerdings ist eine Prognose frühestens Ende der Vorlesung möglich, da die Klausurteilnehmer erst durch Klausur

suranmeldung bekannt werden. Und wenn in Erst- und Zwisemesterprüfungen Klausuranmeldungen als Prädiktoren genutzt werden, liegen diese Informationen ebenfalls erst mit der Anmeldephase vor. Zur Prognose eines Vorlesungserfolgs müssen die Vorlesungsteilnehmer möglichst früh bekannt sein und weitere Prädiktoren ermittelt werden.

3. Prognosen anhand digitalem akademischen Verhalten

Mit zunehmender Digitalisierung der Lehre stehen zunehmend mehr nutzbaren Informationen zur Verfügung. Beispielsweise Feng et al. (2006) prognostizierten Kursnoten anhand eines intelligenten Nachhilfesystems (intelligent tutoring system) unter Verwendung der Information, wie viel Unterstützung ein Studierender benötigte, um eine Aufgabe zu lösen.

Sofern im Rahmen einer Vorlesung flankierend Skripte, Übungen, Tests oder auch Tutorials angeboten werden, können diese zusätzlichen Informationen zur Prognose des Studienerfolgs oder des Vorlesungserfolgs genutzt werden. Das ist im Kontext dieser Arbeit besonders reizvoll, da diese Informationen bereits in den ersten Wochen einer Vorlesung vorliegen können und damit auch bereits zu Beginn des ersten Semesters Informationen über das akademische Verhalten zugänglich werden.

Da die digitalen Möglichkeiten von den Dozierenden an der UNI sehr unterschiedlich genutzt werden, ist eine hochschulweite Betrachtung nicht möglich. Im Folgenden wird eine Vorlesung exemplarisch ausgewertet und die Prognostizierbarkeit des Klausurerfolgs untersucht. Besonderes Interesse erfährt dabei der Prognosezeitpunkt.

3.1. Feature Engineering

Im Rahmen einer Vorlesung im SoSe 2019 wurde ein Kurs auf der Online-Plattform der UNI erstellt, in dem die Vorlesungsskripte und die Unterlagen zur Übungsteilnahme jeweils wöchentlich zur Verfügung gestellt wurden. Zusätzlich wurde ein durch Notenanrechnung incentivierter wöchentlicher Kurztest eingestellt, den die Studierenden jeweils nur eine Woche bearbeiten konnten. Diese Online-Tests wurden in der ersten Vorlesungsstunde mittels „Test 0“ erläutert. Test 0 hat zwar keinen inhaltlichen Vorlesungsbezug, gibt aber Aufschluss darüber, wer in der ersten Vorlesung anwesend war.

Hierüber hinaus wurden unter anderem eine Lehrevaluation und die Wahl in eine der angebotenen Übungen sowie verschiedene Informationen über die Online-Plattform angeboten. Weiterhin wurde ein öffentliches Forum zum Austausch sowie ein anonymes Forum für Fragen angeboten. Für alle Angebote ist in der Online-Plattform für jeden Kursteilnehmer erfasst, ob und wann er das ein Angebot annahm. Für die Kurztests ist zusätzlich bekannt, wie viel Zeit der Studierende zur Lösung benötigte und wie viele Punkte erreicht wurden. Aus den vorhandenen Informationen wurden die in Tabelle 60 dargestellten Variablen gebildet:

Tabelle 60: Variablen des digitalen akademischen Verhaltens

Variable	Beschreibung
Kap1...13	Binäre Angabe ob das jeweilige Teilskript heruntergeladen wurde.
DownKap1...13	Tage zwischen Vorlesungsbeginn und Download
Ueb1...12	Binäre Angabe ob die jeweiligen Übungsaufgaben heruntergeladen wurden
DownUeb1...12	Tagen zwischen Vorlesungsbeginn und Download
Test0...10	Binäre Angabe ob am Test teilgenommen wurde
DownTest0...10	Tage zwischen Vorlesungsbeginn und Test
PunkteTest1...10	Anzahl der im jeweiligen Test erarbeiteten Punkte
BDauer1...10	Bearbeitungsdauer des Tests in Minuten
LehrEva	Binäre Angabe ob an der Lehrevaluation teilgenommen wurde
LehrEvaDat	Tage zwischen Vorlesungsbeginn und Evaluationsteilnahme
TeilnForum	Binäre Angabe ob das Forum genutzt wurde
TeilnAnonymForum	Binäre Angabe ob das anonyme Forum genutzt wurde
ÜbungsTermine	Tage zwischen Vorlesungsbeginn und Test
WahlTutorium	Bearbeitungsdauer des Tests in Minuten
AltKlausur 15&17	Binäre Angabe ob die Altklausuren 2015 & 2017 heruntergeladen wurden
DownAltKlausur 15&17	Tage zwischen Vorlesungsbeginn und Download
Sport	Binäre Angabe der Abstimmungsteilnahme (ob in der Pause ein kurzes Sportprogramm angeboten werden soll)
Ankündigung	Binäre Angabe, ob Ankündigungen verfolgt wurden
Label	Binäre Angabe, ob die Prüfungen bestanden wurde

Sollten Leistungsdaten verfügbar sein, sind auch diese potentiell verwendbar - wie auch in vorangegangener Untersuchung der Prognostizierbarkeit von Prüfungserfolgen in Teil V, Kapitel 2. Vorliegend wird darauf verzichtet, da das Ziel möglichst frühe Prognosen sind und zu Beginn der Vorlesungen im ersten und zweiten Semesters noch keine Leistungsdaten vorliegen.

3.2. Prognoserelevanz digitalen akademischen Verhaltens

Für in Tabelle 60 beschriebenen Variablen wird die Prognoserelevanz in den Tabelle 61 bis Tabelle 64 mittels des Information Gains (vgl. Teil IV, Abschnitt 1.2) dargestellt.

Im Vergleich zu vorangegangenen Untersuchungen ergeben sich erhebliche Unterschiede in Größe und Zusammensetzung der Population. Die betrachteten Studierenden sind bei Datenexport noch eingeschriebene Studierende am Anfang ihres Studiums und ebenfalls weit überwiegend einem Studiengang zuzuordnen. Von 750 Studierenden, haben 262 Studierende die Prüfung bestanden, bei 131 Studierenden war der Versuch nicht erfolgreich und 358 Studierende partizipierten an der Vorlesung, haben sich aber nicht für die Klausur angemeldet. Die abhängige Variable ist binär mit „bestanden“ und „nicht bestanden“ kodifiziert. Studierende, die an der Vorlesung partizipierten, nicht aber an der Klausur, sind als „nicht bestanden“ gewertet. In den 15 Vorlesungswochen wurden Skripte, Übungsaufgaben, Online-Tests und weitere Angebote und Ankündigungen angeboten.

Tabelle 61: Information Gain Skript

Woche	Kapitel		Downloadzeitpunkt			
	Variable	Rang	InfoGain	Variable	Rang	InfoGain
1	Kap1	110	0,0224	DownKap1	106	0,0266
1	Kap2	97	0,0398	DownKap2	92	0,0465
2	Kap3	83	0,0584	DownKap3	49	0,0759
3	Kap3a	64	0,0682	DownKap3a	63	0,0690
3	Kap4	74	0,0632	DownKap4	70	0,0656
5	Kap5	84	0,0567	DownKap5	76	0,0611
6	Kap6	41	0,0798	DownKap6	16	0,1065
7	Kap7	30	0,0910	DownKap7	24	0,0965
8	Kap8	26	0,0938	DownKap8	17	0,1063
8	Kap8n	61	0,0700	DownKap8n	57	0,0720
9	Kap9	32	0,0883	DownKap9	31	0,0907
10	Kap10-12	89	0,0522	DownKap10-12	90	0,0522
10	Kap10-12n	38	0,0813	DownKap10-12n	39	0,0813
12	Kap11	94	0,0463	DownKap11	93	0,0463
13	Kap12	87	0,0549	DownKap12	79	0,0592
14	Kap13	91	0,0480	DownKap13	88	0,0538

Tabelle 62: Information Gain Übungen

Woche	Übung		Downloadzeitpunkt			
	Variable	Rang	Info-Gain	Variable	Rang	Info-Gain
2	Ueb1	98	0,0394	DownUeb1	86	0,0559
3	Ueb2	62	0,0694	DownUeb2	19	0,0995
4	Ueb3	59	0,0708	DownUeb3	37	0,0826
5	Ueb4	47	0,0764	DownUeb4	44	0,0791
6	Ueb5	42	0,0797	DownUeb5	25	0,0958
7	Ueb6	56	0,0727	DownUeb6	46	0,0766
8	Ueb7	15	0,1067	DownUeb7	12	0,1193
9	Ueb8	14	0,1118	DownUeb8	13	0,1118
12	Ueb9	55	0,0735	DownUeb9	54	0,0735
13	Ueb10	69	0,0661	DownUeb10	53	0,0740
14	Ueb11	60	0,0704	DownUeb11	48	0,0759
15	Ueb12	52	0,0742	DownUeb12	45	0,0788

Tabelle 63: Information Gain Online Tests

Woche	Test		Bearbeitungszeitpunkt			Bearbeitungsdauer			Punkte im Test			
	Variable	Rang	Info-Gain	Variable	Rang	Info-Gain	Variable	Rang	Info-Gain	Variable	Rang	Info-Gain
1	Test0	108	0,0240	DownTest0	101	0,0386	BDauer0	102	0,0341	PunkteTest0	-	-
2	Test1	71	0,0645	DownTest1	23	0,0975	BDauer1	72	0,0645	PunkteTest1	2	0,2970
4	Test2	34	0,0834	DownTest2	35	0,0834	BDauer2	36	0,0834	PunkteTest2	4	0,2883
5	Test3	66	0,0672	DownTest3	65	0,0672	BDauer3	67	0,0672	PunkteTest3	6	0,2711
6	Test4	51	0,0749	DownTest4	50	0,0749	BDauer4	40	0,0799	PunkteTest4	3	0,2907
7	Test5	20	0,0976	DownTest5	21	0,0976	BDauer5	22	0,0976	PunkteTest5	1	0,3097
8	Test6	27	0,0910	DownTest6	29	0,0910	BDauer6	28	0,0910	PunkteTest6	5	0,2858
9	Test7	77	0,0604	DownTest7	78	0,0604	BDauer7	73	0,0638	PunkteTest7	10	0,2280
10	Test8	81	0,0590	DownTest8	82	0,0590	BDauer8	58	0,0709	PunkteTest8	8	0,2520
12	Test9	95	0,0447	DownTest9	96	0,0447	BDauer9	11	0,1225	PunkteTest9	7	0,2666
13	Test10	100	0,0391	DownTest10	99	0,0391	BDauer10	18	0,1022	PunkteTest10	9	0,2377

Tabelle 64: Information Gain weiteres Verhalten

Weiteres Verhalten			
Woche	Variable	Rang	InfoGain
1	WahlTutorium	104	0,0327
1	TermineTutorien	114	0,0177
5	Sport	85	0,0564
13	DownAltKlausur15	33	0,0851
13	AltKlausur15	43	0,0795
13	DownAltKlausur17	75	0,062
13	AltKlausur17	80	0,059
13	Ankündigung	115	0,0144
14	LehrEvaDat	111	0,0187
14	LehrEva	112	0,0187
15	TeilnameForum	116	0,0058
15	TeilnameAnoFor	118	0,004

Im Vergleich der Variablen, die das Herunterladen von Skripten, Übungen und die Teilnahme an Online-Tests beschreiben, zeigen die erzielte Punkte in den wöchentlichen Online-Tests den höchsten Informationsgehalt. Die erreichten Punkte in den zehn Online-Kurztests stellen die zehn relevantesten Variablen dar. Alle anderen Variablenarten stehen etwa gleichrangig zueinander. Hinsichtlich der Skripte und

Übungen ist der Zeitpunkt des Downloads entscheidender als die Tatsache, dass sie heruntergeladen wurde. Während die Bearbeitung von Online-Tests und Zeitpunkt wie Dauer der Bearbeitung in den ersten zehn Wochen in der jeweiligen Woche etwa den gleichen Informationsgehalt aufweisen, ist in den Wochen 10 bis 13 die Bearbeitungsdauer zunehmend relevanter als der Zeitpunkt oder die Tatsache, dass der Test bearbeitet wurde.

Im Zeitverlauf ergeben sich innerhalb jeder Variablenart große Unterschiede in der Relevanz. Mit fortschreitenden Vorlesungswochen nimmt der Informationsgehalt der Variablen erst zu und im späteren Verlauf wieder ab. Die höchste Relevanz weisen Variablen auf, die das Verhalten in der siebten von insgesamt 15 Vorlesungswochen beschreiben. Hierzu gehören alle Informationen zu Test 5, Übung 6 und Kapitel 7.

Abbildung 19 zeigt den durchschnittlichen Information Gain aller Variablen der jeweiligen Vorlesungswoche. Zusätzlich wird die Korrelation mit der abhängigen Variable nach Pearson abgebildet. Es wird der erst ansteigende und anschließend wieder fallende Informationsgehalt im Semesterverlauf illustriert.

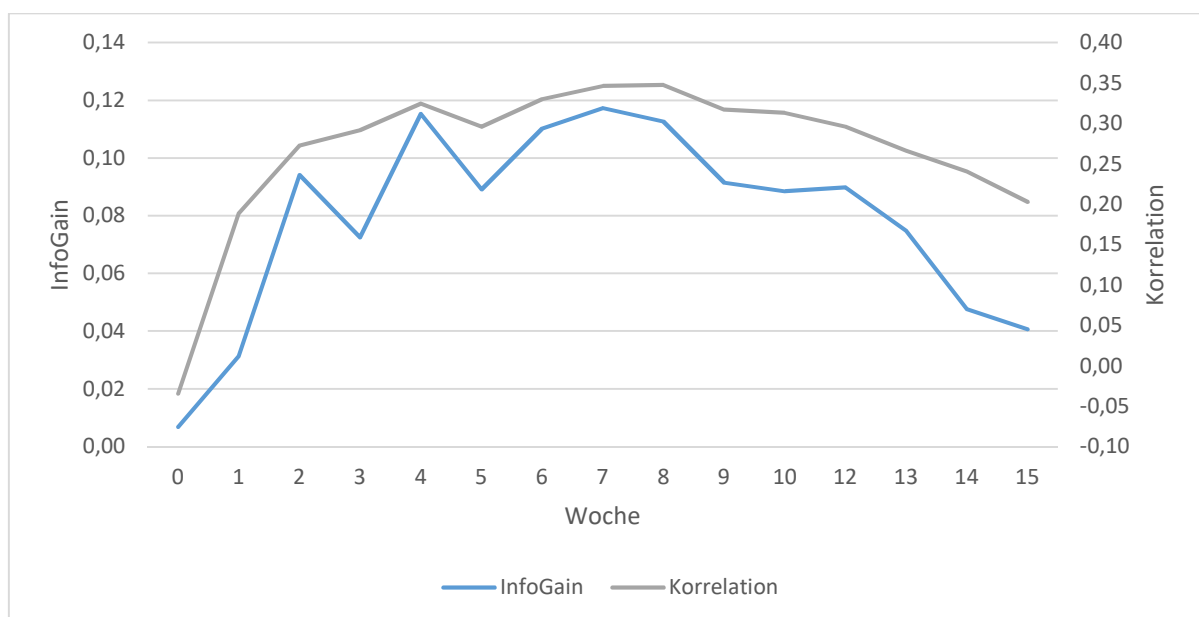


Abbildung 19. Information Gain und Korrelation nach Vorlesungswochen

Der Verlauf des Informationsgehaltes und der Korrelation deutet darauf hin, dass das Aktivitätsniveau der Studierenden zu Beginn und zu Ende einer Vorlesung weniger relevant ist, als zwischen der vierten und achten Vorlesungswoche.

Die schnelle Anstieg des Informationsgehalts in den ersten Vorlesungswochen und die hohe Relevanz der Leistungen zu Mitte des Semesters sind hinsichtlich einer frühen Prognostizierbarkeit des Prüfungserfolgs vorteilhaft. Bereits zur Semesterhälfte liegen die wichtigsten Prädiktoren vor.

3.3. Prognosegenauigkeit

Aus der Grundgesamtheit von 750 Studierenden werden 600 Studierende randomisiert dem Lerndatensatz und 150 Studierende dem Testdatensatz zugeteilt. Zu beachten ist, dass die folgenden Ergebnisse nur bedingt mit vorangegangenen Ergebnissen vergleichbar sind. Einerseits führt der geringe Anteil erfolgreicher Prüfungsteilnehmer 262 (35%), zu einem Ungleichgewicht, das die Gütemaße positiv verschiebt (vgl. Teil II, Abschnitt 2.4.6). Im Gegenzug werden die Genauigkeiten dadurch verringert, dass nur wenige Studierende in Lern- und Testdaten vielen Variablen gegenüberstehen.

Die in Tabelle 61 beschriebenen Variablen werden für 15 Prognosezeitpunkte der Vorlesung in jeweils 15 Lern- und Testdatensätze zusammengefügt. Tabelle 65 stellt die Prognosegenauigkeiten der 15 Zeitpunkte dar.

Tabelle 65: Prognosegenauigkeit digitalen akademischen Verhaltens

AdaBoost	Uni		
	Acc	Rec	Prec
Vor der Vorlesung	58,04 %	66,20 %	66,20 %
Nach Woche 1	59,82 %	67,61 %	67,61 %
Nach Woche 2	83,04 %	85,92 %	85,92 %
Nach Woche 3	79,46 %	83,10 %	83,10 %
Nach Woche 4	81,25 %	84,51 %	84,51 %
Nach Woche 5	84,82 %	87,32 %	87,32 %
Nach Woche 6	84,82 %	87,32 %	87,32 %
Nach Woche 7	84,82 %	87,32 %	87,32 %
Nach Woche 8	81,25 %	84,51 %	84,51 %
Nach Woche 9	83,04 %	85,92 %	85,92 %
Nach Woche 10	83,04 %	85,92 %	85,92 %
Nach Woche 11	81,25 %	84,51 %	84,51 %
Nach Woche 12	83,04 %	85,92 %	85,92 %
Nach Woche 13	84,82 %	87,32 %	87,32 %
Nach Woche 14	83,04 %	85,92 %	85,92 %

Dabei zeigt sich nach der zweiten Woche, wenn die ersten Übungsunterlagen verfügbar werden und der erste vorlesungsbezogene Kurztest angeboten wird, ein erheblicher Anstieg der Genauigkeiten. Das bedeutet, dass bereits zu diesem frühen Zeitpunkt Prognosen mit hoher Genauigkeit möglich sind.

4. Zwischenfazit

Die Genauigkeit von Prognosen wird durch die verfügbaren Daten bestimmt. Der Zeitpunkt an den Informationen verfügbar werden, bestimmt damit den erstmöglichen Prognosezeitpunkt. Innerhalb der ersten Studienmonate liegen noch keine Prüfungsergebnisse, dafür aber die Prüfungsanmeldungen vor. Mittels dieser sind 105 Tage vor Semesterende Prognosen (etwa Mitte Dezember bzw. Mitte Juni), etwa auf Niveau von Prognosen Mitte bis Ende des zweiten Semesters, möglich. Durch die Hochschule zentral angebotene Hilfsangebote können damit etwa ein halbes Jahr früher angeboten werden, als wenn auf Leistungsdaten des ersten Semesters gewartet wird.

Neben zentral angebotener Hilfe kann alternativ oder flankierend auch Unterstützung innerhalb von Vorlesungen oder den dazugehörigen Übungen und Tutorien zielgerichtet angeboten werden. Hierzu bedarf es der Prognose eines Prüfungserfolgs. Unter Verwendung der in Teil II vorgestellten Informationen sind Prognosen des Prüfungserfolgs der exemplarisch getesteten Vorlesung auf vergleichsweise hohem Niveau möglich – allerdings erst innerhalb der Klausuranmeldephase. Vorher liegen die relevanten Informationen noch nicht vor. Frühere zielgerichtete vorlesungsbegleitende Unterstützungsangebote sind wünschenswert, aber nur zu realisieren, wenn die Vorlesungsteilnehmer bekannt sind und weitere Prädiktoren gefunden werden können. Hierzu kann das Online-Verhalten auf Skriptservern und Online-Angeboten genutzt werden. Durch Beitritt zu einem Onlinekurs sind die Teilnehmer und die Zeitpunkte und Intensität verschiedener Aktivitäten – wie Skriptdownload oder Teilnahme an Tests – bekannt. Besonders relevant für die Prognose sind dabei die im Rahmen von Online-Tests erzielten Punkte und das Aktivitätslevel zwischen der vierten und achten Vorlesungswoche. Die Prognose einer exemplarisch analysierten Vorlesung zeigt vielversprechende Prognosen bereits im ersten Vorlesungsdrittel.

VI Fazit und Ausblick

Studienabbrüche sind ein wichtiges Thema in der Bildungspolitik. Die für die Ausbildung der Studenten aufgewendeten Ressourcen sind für den Studierenden, die Hochschule und die Gesellschaft nur von begrenztem wirtschaftlichen Wert, wenn sie nicht mit einem Studienabschluss beendet werden. Frühwarnsysteme können in diesem Kontext einen wertvollen Beitrag leisten. Mittels der frühzeitigen Identifikation abbruchgefährdeter Studierende kann Hilfe zielgerichtet angeboten und effizienter allokiert werden. Bislang war es nur möglich, Hilfe allgemein der Grundgesamtheit der Studierenden, bspw. durch Homepage und Poster, in der Hoffnung anzubieten, dass Studierende mit erhöhter Abbruchneigung selbstreferenziell teilnehmen. Weiterhin erlaubt die Kenntnis der Abbruchwahrscheinlichkeit teilnehmender Studierender die Evaluation abbruchspräventiver Maßnahmen, beispielsweise zur Maßnahmenoptimierung und auch zur innerhochschulischen Ressourcenallokation. Maßnahmen können entsprechend ihres Zeitpunkts in Maßnahmen vor dem Studium und Maßnahmen im Studium unterteilt werden. Maßnahmen vor dem Studium adressieren in der Regel die Passung zwischen Studienbewerber und Studienangebot, beispielsweise durch intensivere Studieneingangsberatung, Self-Assessments oder Vorsemester. Maßnahmen im Studium adressieren Studierende mit erhöhter Abbruchwahrscheinlichkeit mit dem Ziel der Studienfortführung oder eines beschleunigten Abbruchs und Aufnahme eines alternativen Bildungswegs. Ohne Kenntnis des Wirkungsgrades der teilweise sehr personalintensiven Angebote kann keine Priorisierung nach Effizienz vorgenommen werden. Über die „innerbetrieblichen“ Bemühungen eines bestmöglichen Mitteleinsatzes hinaus, gewinnt die Effizienz der Abbruchprädiktion zur Abbruchprävention zusätzliche Bedeutung, wenn auch die Einnahmenseite künftig zu 80 % durch die Zahl der Absolventen bestimmt wird (§3 III Zukunftsvertrag „Studium und Lehre“ vom 06.07.2019).

In der vorliegenden Arbeit wurde ein Frühwarnsystem konzipiert, das es erlaubt, abbruchgefährdete Studierende möglichst früh im Studium zu identifizieren. Damit das Frühwarnsystem auf andere Hochschulen übertragbar ist, erlernen maschinelle Algorithmen anhand administrativer historischer Studierendendaten nach §3 HStatG die Merkmale, die Studienabbrecher von Absolventen an der Betrachtungshochschule unterscheiden. In Teil III wird gezeigt, dass weniger aufwendige Verfahren wie PASST?! und SVM, geringere Prognosegenauigkeiten aufweisen und statische Schwellenwerte

nicht an jeder Hochschule gleich gute Ergebnisse liefern können. Weiterhin ist es nicht möglich, die Genauigkeit des Frühwarnsystems zu evaluieren oder die Schwellenwerte an der Hochschule einzusetzen – es sei denn anhand der Analyse historischer Studierendendaten. Dann aber erübrigt sich der Vorteil des geringeren Aufwands. Das aufwendigere aber selbstjustierende Frühwarnsystem FragSte nutzt zur Analyse historischer Studierendendaten eine Kombination verschiedener maschineller Verfahren, die abschließend mittels des AdaBoost Algorithmus zu einem Prognoseergebnis verschmolzen werden. Damit entfällt eine vorzunehmende Verfahrensauswahl. Der Aufwand des Frühwarnsystems kann weiterhin dadurch verringert werden, dass eine datensparsamere Datenselektion verwendet wird. Sofern diese datensparsamere Auswahl vergleichbare Ergebnisse ermöglicht, reduziert sich hierdurch der Datenreinigungs- und Datenaufbereitungsaufwand. Daher wurden in Teil IV die eingesetzten Variablen auf ihren Informationsgehalt überprüft und geprüft, ob eine datensparsamere Datenselektion vergleichbare Ergebnisse erlaubt. Dabei zeigt sich, dass der Prognosezeitpunkt von entscheidender Bedeutung ist. Prognosen, die in Mitten eines Bachelor-Studiums erstellt werden, genügt eine datensparsamere Datenauswahl – bis hin zu einem reinem ECTS-Monitoring. Da aber alle Studierenden, die ihr Studium bis zum Prognosezeitpunkt bereits abgebrochen haben, nicht mehr erreicht werden können und allgemein davon ausgegangen wird, dass die Wirkung von Hilfsangeboten und Maßnahmen mit zunehmend späteren Zeitpunkt abnimmt, ist eine späte Prognose von geringerem Wert als eine frühe. Für besonders frühe Prognosen genügt eine datensparsamere Auswahl jedoch nicht. Weiterhin zeigt sich, dass bedingt durch Prüfungskorrektur, Eingabe ins System, Export, Datenreinigung und -aufbereitung die Prognose basierend auf Leistungsdaten erheblich verzögert werden. Prognosen, basierend auf den Erstsemesterprüfungen, können nicht vor Mitte des zweiten Semesters vorgenommen werden. Daher wird in Teil V geprüft, ob früher verfügbare Informationen den Informationsgehalt von Prüfungsleistungen zugunsten früherer Prognosen substituieren können. Dabei wird gezeigt, dass Art, Anzahl und Datum von Prüfungsanmeldungen einen adäquaten Ersatz darstellen – es sind Prognosen Ende des ersten Semesters auf Niveau des zweiten Semesters möglich. Damit enden jedoch die Möglichkeiten der Studierendendaten nach §3 HStatG. Wenn Prognosen zu Studienbeginn noch früher und genauer getroffen werden sollen, müssen relevante Informationen ei-

gens dafür erhoben werden. Vielversprechende Datenquellen können sich aus der Digitalisierung von Lehrinhalten ergeben. Die Nutzung von Informationen aus Lernplattformen ist im Vergleich zu Befragungsdaten mit geringem Aufwand realisierbar und beschreibt das studentische Verhalten innerhalb einer Veranstaltung. Es wird gezeigt, dass Prognosen in den ersten acht Studienwochen bereits eine hohe Genauigkeit aufweisen – allerdings nur exemplarisch anhand einer Vorlesung. Ob das Ergebnis auf andere Vorlesungen übertragbar ist, kann nicht geprüft werden. Es ist wahrscheinlich, dass Studiengangs- und Hochschulspezifika einen Einfluss auf die Prognosegenauigkeit haben. Sollten die vorhandenen Variablen nicht mit dem zu prognostizierenden Studienergebnis korreliert sein, ist eine Prognose nicht möglich. Das ist ebenfalls zur Einordnung der Ergebnisse – insbesondere der des Prognosegehalts einzelner Variablen – relevant: Zur Prädiktion genügen korrelierte Merkmale. In wie weit die ermittelten Prädiktoren für den Studienabbruch kausal sind, muss in weitergehender Forschung adressiert werden.

Auch wie identifizierte abbruchgefährdete Studierende effizient informiert werden können, ist von besonderer praktischer Bedeutung und in der Folgeforschung zu ermitteln. Hierbei ist sowohl die Formulierung wie das genutzte Medium abzuwägen. Die Ansprache kann persönlich, postalisch, per Email oder auch über eine hochschuleigene Lösung wie ein App erfolgen. Dabei kann die Formulierung unterstützend oder drängend sowie persönlich oder amtlich erfolgen.

Im Kontext der Ansprache abbruchgefährdeter Studierende ist auch zu entscheiden, durch wen die Ansprache und Unterstützung erfolgen soll. In Frage kommen zentrale Ansprache und Unterstützung, beispielsweise durch eine zentrale Studienberatung, oder dezentrale Lösungen auf Ebene von Vorlesungen oder Fachbereichen. Es ist zu untersuchen, ob die Wirkweisen zentraler Hinweise und Angebote von dezentralen abweichen.

Ebenfalls wichtiges Thema, das in der vorliegenden Studie ausgeklammert wurde, ist die Unterscheidung zwischen Schwund und Abbruch. Studierende, die wegen schlechter Leistungen das Studium aufgeben, werden ebenso als Studienabbrecher behandelt, wie Studierende, die mit herausragenden Leistungen die Hochschule verlassen, um ihr Studium an einer anderen Hochschule fortzusetzen. In einem ersten Schritt ist fehlende Differenzierung unproblematisch, da beide Gruppen aus Sicht der

Hochschule das Studium abbrechen. Wenn die Hochschulfinanzierung künftig auch maßgeblich an der Zahl erfolgreicher Studierender bemessen wird, haben Hochschulen ein weiter steigendes Interesse an erfolgreichen Absolventen. Daher kann es auch aus diesem Grund für die Hochschulen wichtig sein, besonders leistungsstarke und wechselwillige Studierende an der Hochschule zu halten und diese rechtzeitig zu identifizieren. In einer Folgestudie sollten daher potentielle Hochschulwechsler und Abbrecher stärker differenziert werden.

Auch das politisch immer wieder diskutierte Problem der Scheinstudierenden wird im vorliegenden Beitrag nicht thematisiert. Als Scheinstudierende können immatrikulierte Studierende bezeichnet werden, die keinen Studienabschluss (mehr) anstreben und sich durch akademische Inaktivität auszeichnen. Motive können Privilegien des Studierendenstatus sein, wie die Nutzung des Semestertickets oder der Bezug von Kindergeld. Scheinstudierende mit Hilfe der vorliegenden Daten zu identifizieren, ist jedoch nicht trivial, da z.B. auch Krankheit, Familienzeiten oder finanzielle Probleme zu Inaktivität im Studium führen können. Schwerwiegender aus Sicht der Hochschule ist aber die Schwierigkeit, inaktive Studierende, die nie einen Studienabschluss erwerben wollten, von inaktiven Studierende zu unterscheiden, die erst im Verlauf des Studiums zu inaktiven Studierenden wurden. Diese Unterscheidung wäre jedoch für die Allokation von Unterstützungsangeboten an den Hochschulen bedeutsam.

Literaturverzeichnis

- Ahles, L., Köstler, U., Vetter, N. & Wulff, A. (2016) *Studienabbrüche an deutschen Hochschulen. Stand der Thematisierung und strategische Ansatzpunkte*. Baden-Baden: Nomos Verlagsgesellschaft.
- Apenburg, E. (1980) *Untersuchungen zur Studienzufriedenheit in der heutigen Massenuniversität*. Frankfurt a.M: Lang.
- Arcidiacono, P., Aucejo, E., Maurel, A. & Ransom, T. (2016) College Attrition and the Dynamics of Information Revelation. *NBER Working Papers - National Bureau of Economic Research*.
- Arnold, K. & Pistilli, M. (2012) Course Signals at Purdue: Using Learning Analytics to Increase Student Success. *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, 267-70.
- Arulampalam, W., Naylor, R.A. & Smith, J.P. (2005) Effects of in-class variation and student rank on the probability of withdrawal: cross-section and time-series analysis for UK university students. *Economics of Education Review*, 24, 251-62.
- Asif, R., Merceron, A., Ali, S.A. & Haider, N.G. (2017) Analyzing undergraduate students' performance using educational data mining. *Computers & Education*, 113, 177-94.
- Backhaus, K., Erichson, B., Plinke, W. & Weiber, R. (2016) *Multivariate Analysemethoden*. 14th ed. Heidelberg: Springer Verlag.
- Baker, R.S.J. (2010) Statistical Data Mining Tutorials. In B. McGaw, P. Peterson & E. Baker, eds. *International encyclopedia of education*. UK: Elsevier. 112-18.
- Baker, R.S.J.d. & Yacef, K. (2009) The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, 1(1), 3-16.
- Barber, R. & Sharkey, M. (2012) Course correction: using analytics to predict course success. *LAK '12 Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, 259-62.
- Barefoot, B.O., Gardner, J.N., Cutright, M., Morris, L.V., Schroeder, C.C., Schwartz, S.W., Siegel, M.J. & Swing, R.L. (2005) *Achieving and sustaining institutional excellence for the first year of college*. San Francisco, CA.: Jossey-Bass.

- Bargel, T., Multrus, F. & Ramm, M. (2008) *Studiensituation und studentische Orientierungen*. Bonn: BMBF.
- Bartlett, M. (1937) Properties of sufficiency and statistical tests. *Proceedings of the Royal Statistical Society Series A*, 160, 268-82.
- Bayer, J., Bydzovská, H., Géryk, J., Obsivac, T. & Popelinsky, L. (2012) Predicting Drop-Out from Social Behaviour of Students. *Proceedings of the 5th International Conference on Educational Data Mining*, 103-09.
- Bean, J. (1982) Student attrition, intentions, and confidence: Interaction effects in a path model. *Research in Higher Education*, 17(4), 291-32.
- Bean, J.P. (1983) The Application of a Model of Turnover in Working Organizations to the Student Attrition Process. *The Review of Higher Education*, 6, 129-48.
- Berens, J. & Schneider, K. (2019) Drohender Studienabbruch: Wie gut sind Frühwarnsysteme. *Qualität in der Wissenschaft (QIW)*, 13 (3+4), 102-08.
- Berens, J., Schneider, K., Görtz, S., Oster, S. & Burghoff, J. (2019) Early Detection of Students at Risk - Predicting Student Dropouts Using Administrative Student Data from German Universities and Machine Learning Methods. *Journal of Educational Data Mining*, 11(3), 1-41.
- Berger, M., Galonska, C. & Koopmans, R. (2004) Political Integration by a Detour? Ethnic Communities and Social Capital of Migrants in Berlin. *Journal of Ethnic and Migration Studies*, 30, 491-507.
- Blüthmann, I. (2012) Individuelle und studienbezogene Einflussfaktoren auf die Zufriedenheit von Bachelorstudierenden. *Zeitschrift für Erziehungswissenschaft*, 15, 273-303.
- Blüthmann, I., Lepa, S. & Thiel, F. (2008) Studienabbruch und -wechsel in den neuen Bachelorstudiengängen. *Zeitschrift für Erziehungswissenschaft*, 11(3), 406-29.
- Borhek, J.T. (1970) Ethnic-Group Cohesion. *American Journal of Sociology*, 76, 33-46.
- Bourn, J. (2007) *Staying the course: The retention of students in higher education*. London: National Audit Office.

- Bowers, A.J., Sprott, R. & Taff, T.A. (2013) Do We Know Who Will Drop Out? A Review of the Predictors of Dropping out of High School: Precision, Sensitivity and Specificity. *The High School Journal*, 77-100.
- Bradley, A.P. (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7), 1145-59.
- Brahm, T., Jenert, T. & Wagner, D. (2016) The crucial first year: a longitudinal study of students' motivational development at a Swiss Business School. *Higher Education*, 73(3), 459-78.
- Brandstätter, H. & Farthofer, A. (2002) Studienerfolgsprognose – konfigurativ oder linear additiv? *Zeitschrift für Differentielle und Diagnostische Psychologie*, 23, 381-91.
- Brandstätter, H., Grillich, L. & Farthofer, A. (2006) Prognose des Studienabbruchs. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 38(3), 121-31.
- Breimann, L. (1996) Bagging Predictors. *Machine Learning*, 24(2), 123-40.
- Breimann, L. (2001) Random Forests. *Machine Learning*, 45(1), 5-32.
- Bundesamt für Migration und Flüchtlinge. (2017) *Bundesamt für Migration und Flüchtlinge* [Online]. Verfügbar unter: <http://www.bamf.de/DE/Migration/Spaetaussiedler/spaetaussiedler-node.html>. [Zugegriffen am 02. November 2019].
- Burrus, J., Elliott, D., Brennemann, M. & Markle, R. (2013) Putting and Keeping Students on Track: Toward a Comprehensive Model of College Persistence and Goal Attainment. *ETS Research Report Series*, 1, 1-61.
- Cleff, T. (2015) *Deskriptive Statistik und moderne Datenanalyse*. Wiesbaden: Springer Gabler.
- Cohen, J. (2013) *Statistical power analysis for the behavioral sciences*. New York: Routledge.
- Danilowicz-Gösele, K., Lerche, K., Meya, J. & Schwager, R. (2017) Determinants of students success at university. *Education Economics*, 25(5), 513-32.

- Dekker, G.W., Pchenenizkiy, M. & Vleeshouwers, J.M. (2009) Predicting Students Drop out: A Case Study. In T. Barnes, M. Desmarais, C. Romero & S. Ventura, eds. *Proceedings of the 2nd International Conference on Educational Data Mining*. Cordoba, Spain. 41-50.
- Diehl, C. (2005) Der Integrationssurvey des Bundesinstituts für Bevölkerungsforschung. In S. Haug & C. Diehl, eds. *Aspekte der Integration. Eingliederungsmuster und*. Wiesbaden: Verlag für Sozialwissenschaften. 11-22.
- Dreyfus, G. (2005) *Neural Networks: Methodology and Applications*. 2nd ed. Heidelberg: Springer-Verlag.
- Dziuban, C.S.E. (1974) When is a Correlation Matrix Appropriate for Factor Analysis. *Psychological Bulletin*, 81(6), 358-61.
- Er, E. (2012) Identifying At-Risk Students Using Machine Learning - Techniques: A Case Study with IS 100. *International Journal of Machine Learning and Computing, Vol 2, No 4*, 476-80.
- Falk, S. & Marschall, M. (2019) Studienabbruch - Was können Hochschulen tun? Forschungsstand und Interventionsmöglichkeiten. *QiW*, 1, 23-27.
- Farwick, A. (2009) *Segregation und Eingliederung. Zum Einfluss der räumlichen Konzentration von Zuwanderern auf den Eingliederungsprozess*. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Feng, M., Heffernan, N. & Koedinger, K.R. (2006) Predicting state test scores better with intelligent tutoring systems: developing metrics to measure assistance required. *International Conference on Intelligent Tutoring Systems*, 31-40.
- Fernández, O., Rother, N. & Braun, M. (2006) Stichprobenziehung für Migrantenpopulationen in fünf Ländern: Eine Darstellung des methodischen Vorgehens im PIONEUR-Projekt. *ZUMA-Nachrichten*, 30, 72-88.
- Festinger, L. (1957) *A theory of cognitive dissonance*. Stanford: Row, Peterson and Company.
- Flaig, B. (2008) *Lebenslagen und Milieus der Heidelberger mit Migrationshintergrund. Heidelberger Migrantenstudie*. Heidelberg: Stadt Heidelberg – Amt für Stadtentwicklung und Statistik.

- Frank, E., Hall, M.A. & Witten, I.H. (2016) *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques"*. Morgan Kaufmann.
- Freund, Y. & Schapire, R.E. (1999) A short Introduction to Boosting. *Journal of Japanese Society for Artificial Intelligence*, 771-80.
- Gaebel, M., Hauschildt, K., Mühleck, K. & Smidt, H. (2012) *Tracking Learners' and Graduates' Progression Paths. TRACKIT*. Brussels, Belgium: EUA Publications.
- Georg, W. (2008) Individuelle und institutionelle Faktoren der Bereitschaft zum Studienabbruch. *ZSE Zeitschrift für Soziologie der Erziehung und Sozialisation*, (2), 191-206.
- Gleason, P. & Dynarski, M. (2002) Do we know whom to serve? Issues in using risk factors to identify dropouts. *Journal of Education for Students Placed At Risk*, 25-41.
- Gold, A. (1988) *Studienabbruch, Abbruchneigung und Studienerfolg. Vergleichende Bedingungsanalysen des Studienverlaufs*. Frankfurt a.M.
- Gold, A. & Souvignier, E. (2005) Prognose der Studierfähigkeit - Ergebnisse aus Längsschnittanalysen. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 37, 214-22.
- Goodman, L. (1961) Snowball Sampling. *The Annals of Mathematical Statistics*, 148-70.
- Guttman, L. (1953) Image Theory for the Structure of Quantitative Vanates. *Psychometrika*, 18(4), 277-96.
- Hall, M. (1998) *Correlation-based Feature Selection for Machine Learning*. Hamilton, New Zealand.
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, R. & Witten, I. (2009) The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1).
- Hanhörster, H. (2013) *Türkeistämmige Eigentümer in Migrantenvierteln*. Dortmund: Springer.
- Harvey, L. & Green, D. (2000) Qualität definieren. Fünf unterschiedliche Ansätze. *Zeitschrift für Pädagogik*, 41, 17-39.

- Harvey, L. & Green, D. (2000) Qualität definieren. Fünf unterschiedliche Ansätze. *Zeitschrift für Pädagogik*, 41, 17-39.
- Haug, S., Müssig, S. & Stichs, A. (2009) *Muslimisches Leben in Deutschland. Im Auftrag der Deutschen Islam Konferenz*. Nürnberg: Forschungsberichte des Bundesamts für Migration und Flüchtlinge.
- Helmke, A. (2009) *Unterrichtsqualität und Lehrerprofessionalität – Diagnose, Evaluation und Verbesserung des Unterrichts*. Seelze-Velber: Klett/Kallmeyer.
- Heublein, U. (2014) Student Drop-out from German Higher Education Institutions. *European Journal of Education. Research, Development and Policy*, 49(4), 497-513.
- Heublein, U. & Burkhart, S. (2013) *Bildungsinländer 2011 - Daten und Fakten zur Situation von von ausländischen Studierenden*. Bonn.
- Heublein, U., Ebert, J., Hutzsch, C., Isleib, S., König, R., Richter, J. & Woisch, A. (2017) *Zwischen Studierenerwartung und Studienwirklichkeit. Ursachen des Studienabbruchs, beruflicher Verbleib der Studienabbrecherinnen und Studienabbrecher und Entwicklung der Studienabbruchquote an deutschen Hochschulen*. Hannover: DZHW.
- Humpert, A. & Schneiderheinze, K. (2000) Stichprobenziehung für telefonische Zuwanderumferagen. *ZUMA-Nachrichten*, 47, 36-64.
- Humpert, A. & Schneiderheinze, K. (2002) *Stichprobenziehung für telefonische Zuwandererumfragen. Praktische Erfahrungen und Erweiterung der Auswahlgrundlage*. Münster: Waxmann.
- Jetter, T. (2017) *Membrain NN* [Online]. Verfügbar unter: <https://www.membrain-nn.de/> [Zugegriffen am 03. April 2019].
- Kaiser, H.F. & Rice, J. (1974) Little Jiffy, Mark IV. *Educational and Psychological Measurement*, 34, 111-17.
- Karegowda, A.G., Manjunath, A.S. & Jayaram, M.A. (2010) Comparative study of attribute selection using gain ratio and correlation based feature selection. *International Journal of Information Technology and Knowledge Management*, 2(2), 271-77.

- Knowles, J. (2015) Of Needles and Haystacks: Building an Accurate Statewide Dropout Early Warning System in Wisconsin. *Journal of Educational Data Mining*, 7(3), 18-67.
- Kotsiantis, S.B., Pierrakeas, C.J. & Pintelas, P.E. (2003) Preventing Student Dropout in Distance Learning - Using Machine Learning Techniques. *KES 2003*, 267-74.
- Larsen, M.L., Kornbeck, K.P., Kristensen, R.M., Larsen, M.R. & Sommersel, H.B. (2013) Dropout Phenomena at Universities: What is Dropout? Why does Dropout Occur? What Can be Done by the Universities to Prevent or Reduce it? A systematic review. *Danish Clearinghouse for Educational Research*.
- McGarrigle, J. & Kearns, A. (2009) Living Apart? Place, Identity and South Asian Residential Choice. *Housing Studies*, 24, 451-75.
- Meulemann, H. (1991) Zufriedenheit und Erfolg in der Bildungslaufbahn. Ein Längsschnitt vom Gymnasium bis zum Studienabschluss. *Zeitschrift für Sozialisationsforschung und Bildungssoziologie*, 11, 215-38.
- Michael, J. (2007) Anredebestimmung anhand des Vornamens. *c't*, 17/2007, 182-83.
- Michael, J. (2016) Name Quality Pro (to be published). (*available from the author; mail to: namequality.pro@gmail.com*).
- Minaei-Bidgoli, B., Kortemeyer, G. & Punch, W.F. (2004) Enhancing Online Learning performance: An Application of Data Mining Methods. *Proceedings of the Seventh IASTED International Conference on Computers and Advanced Technology in Education*.
- Minsky, M. & Papert, S. (1969) Perceptrons: An Introduction to Computational Geometry. *Institute of Technology: Massachusetts*.
- Mucherino, A., Papajorgji, P.J. & Pardalos, P.M. (2009) k-Nearest Neighbor Classification. *Data Mining in Agriculture. Springer Optimization and Its Applications*, 34, 109-13.
- Nghe, N.T., Janecek, P. & Haddaway, P. (2007) A Comparative analysis of techniques for predicting academic performance. *Frontiers in Education Conference-Global Engineering: Knowledge Without Borders, Opportunities Without Passports, FIE'0, 37th Annual IEEE*.

- OECD. (2016) Immigrant background, student performance and students' attitudes towards science. In *PISA 2015 Results (Volume I): Excellence and Equity in Education*. Paris: OECD Publishing.
- OECD. (2018) *Equity in Education: Breaking Down Barriers to Social Mobility*. Paris: OECD Publishing.
- Pascarella, E.T. & Terenzini, P.T. (1979) Interaction Effects in Spady's and Tinto's Conceptual Models of College Dropout. *Sociology of Education*, 52, 197-210.
- Peña-Ayala, A. (2014) Review: Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications*, 1432-62.
- Quinlan, J.R. (1986) Induction of Decision Trees. *Machine Learning* 1, 81-106.
- Romero, C. & Ventura, S. (2010) Educational Data Mining: A Review of the State of the Art. *IEEE Transactions on Systems Man and Cybernetics Part C (Applications and Reviews)*, 601-18.
- Rumelhart, D.E., Hinton, G.E. & Williams, R.J. (1986) Learning representations by back-propagating errors. *Nature*, 323, 533-36.
- Rummelhart, D.E. & Siple, P. (1974) Learning internal representations by error propagation. *Cambridge, MA: Bradford Books*.
- Salentin, K. (1999) Die Stichprobenziehung bei Zuwandererbefragungen. *ZUMA-Nachrichten*, 23, 115-35.
- Salentin, K. & Wilkening, F. (2003) Ausländer, Eingebürgerte und das Problem einer realistischen. *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 55, 279-98.
- Salganik, M. & Heckathorn, D. (2004) Sampling and Estimation in Hidden Populations using Respondent-Driven Sampling. *Sociological Methodology*, 34, 193-239.
- Sammut, C. & Webb, G. (2017) *Encyclopedia of Machine Learning and Data Mining*. New York: Springer US.
- Santana, M., Costa, E., Neto, B., Silva, I. & Rego, J. (2015) High-School Dropout Prediction Using Machine Learning: A Danish Large-scale Study. *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. 22-24, Bruges (Belgium)*.

- Sara, N.-B., Halland, R., Igel, C. & Alstrup, S. (2015) High-School Dropout Prediction Using Machine Learning: A Danish Large-scale Study. *ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence*, 319-24.
- Sarcletti, A. & Müller, S. (2011) Zum Stand der Studienabbruchforschung. Theoretische Perspektiven, zentrale Ergebnisse und methodische Anforderungen an künftige Studien. *Zeitschrift für Bildungsforschung*, 235-48.
- Schapire, E. & Freund, Y. (1997) A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Science*, 55, 119-39.
- Schapire, R.E. & Freund, Y. (2012) *Boosting - Foundations and Algorithms*. Massachusetts: Institute of Technology.
- Schenk, L., Ellert, U. & Neuhauser, H. (2007) Kinder und Jugendliche mit Migrationshintergrund in Deutschland - Methodische Aspekte im Kinder- und Jugendgesundheitsurvey. *Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz*, 590-99.
- Schiefele, U. & Jacob-Ebbinghaus, L. (2006) Studienzufriedenheit. *Zeitschrift für Pädagogische Psychologie*, 20, 199-212.
- Schmidt-Atzert, L. (2005) Prädiktion von Studienerfolg bei Psychologiestudenten. *Psychologische Rundschau*, 56, 131-33.
- Schneider, K., Berens, J. & Burghoff, J. (2019) Drohende Studienabbrüche durch Frühwarnsysteme erkennen: Welche Informationen sind relevant? *Zeitschrift für Erziehungswissenschaften (ZfE)*, 22(5), 1121-46.
- Schröder, M. & Daniel, H.D. (1998) Studienausstieg. Eine annotierte Bibliographie (1975-1997). *Herausgegeben vom Wissenschaftlichen Zentrum für Berufs- und Hochschulforschung der Universität Gesamthochschule Kassel*, 54.
- Schulze-Stocker, F., Cesca, S. & Pelz, R., 2018. Das PASST?-Programm der TU Dresden – Evaluation von & Forschung zu Monitoring und Beratung. *DeGEval* 13.09.2018.

- Schulze-Stocker, F., Schäfer-Hock, C. & Pelz, R. (2017) Weniger Studienabbruch durch Frühwarnsysteme – Das Beispiel des PASST?-Programms an der TU Dresden. *Zeitschrift für Beratung und Studium*, 26-32.
- Seidmann, A. (1996) Spring Retention Revisited: $RET = E ID + (E + I + C) IV$. *College and University*, 71, 18-20.
- Statistisches Bundesamt; DZHW-Berechnungen. (2015) Studierendenstatistik.
- Statistisches Bundesamt. (2015) Bevölkerung und Erwerbstätigkeit. Bevölkerung mit Migrationshintergrund – Ergebnisse des Mikrozensus 2015.
- Stinebrickner, T. & Stinebrickner, R. (2008) The Effect of Credit Constraints on the College Drop-Out Decision: A Direct Approach Using a New Panel Study. *American Economic Review*, 98, 2163-84.
- Stinebrickner, T. & Stinebrickner, R. (2012) Learning about Academic Ability and the College Dropout Decision. *Journal of Labor Economics*, 32, 707-48.
- Stinebrickner, T. & Stinebrickner, R. (2014) Academic Performance and College Dropout: Using Longitudinal Expectations Data to Estimate a Learning Model. *Journal of Labor Economics*, 32, 601-44.
- Strecht, P., Cruz, L., Soares, C., Mendes-Moreira, J. & Abreu, R. (2015) A Comparative Study of Classification and Regression Algorithms for Modelling Students' Academic Performance. *International Educational Data Mining Society*.
- Ströhlein, G. (1983) *Bedingungen des Studienabbruchs*. Frankfurt am Main: Europäische Hochschulschriften, Reihe XI.
- Strubelt, W., Gatzweiler, H.P. & Kaltenbrunner, R. (2008) Migration/Integration und Stadtteilpolitik. *Bundesamt für Bauwesen und Raumordnung*, 34.
- Swets, J.A. (1988) Measuring the Accuracy of Diagnostic Systems. *American Association for the Advancement of Science*, 1285-93.
- The World Bank Group. (2016) [Online]. Verfügbar unter: http://data.worldbank.org/indicator/SP.POP.TOTL?year_high_desc=true [Zugegriffen am 10. August 2016].
- Tinto, V. (1975) Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research*, 45, 89-125.

- Trapmann, S., Hell, B., Weigand, S. & Schuler, H. (2007) Die Validität von Schulnoten zur Vorhersage des Studienerfolgs - eine Metaanalyse. *Zeitschrift für pädagogische Psychologie*, 21, 11-27.
- United Nations Statistics Division. (2016) *Composition of macro geographical (continental) regions, geographical sub-regions, and selected economic and other groupings* [Online]. Verfügbar unter: <http://millenniumindicators.un.org/unsd/methods/m49/m49regin.htm> [Zugegriffen am 06. Dezember 2016].
- Villano, R., Harrison, S., Lynch, G. & Chen, G. (2018) Linking early alert systems and student retention: a survival analysis approach. *Higher Education*, 903-820.
- Werbos, P. (1974) *Beyond regression: New tools for prediction and analysis in the behavioral sciences*. Cambridge, MA: Harvard University.
- Westerholt, N., Lenz, L., Stehling, V. & Isenhardt, I. (2018) *Beratung und Mentoring im Studienverlauf - Ein Handbuch*. Münster: Waxmann.
- Wiers-Jenssen, J., Stensaker, B. & Groggaard, J.B. (2002) Student satisfaction: towards an empirical deconstruction of the concept. *Quality in Higher Education*, 8, 183-95.
- Wilson, K., Murphy, K., Pearson, A., Wallace, B., Reher, V., Buys, N. & Buys, N. (2016) Understanding the early transition needs of diverse commencing university students in a health faculty: informing effective intervention practices. *Higher Education*, 41(6), 1023-40.
- Winteler, A. (1984) *Pfadanalytische Validierung eines konzeptionellen Schemas zum Studienabbruch*. München: Institut für Empirische Pädagogik, Pädagogische Psychologie und Bildungsforschung der Universität München.
- Witten, I.H., Eibe, F. & Hall, M.A. (2016) *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.
- Xenos, M. (2004) Prediction and assessment of student behaviour in open and distance education in computers using Bayesian networks. *Computers & Education Journal*, 345-59.

- Yukselturk, E., Ozekes, S. & Türel, Y.K. (2014) Predicting dropout student: an application of data mining methods in an online education program. *European Journal of Open, Distance and e-Learning*, 17(1), 118-33.
- Zafar, M.B., Valera, I., Rdriguez, M.G. & Gummadi, K.P. (2017) Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. *WWW '17 Proceedings of the 26th International Conference on World Wide Web*, 1171-80.
- Zhang, Y., Oussena, S., Clark, T. & Kim, H. (2010) Use Data Mining to Improve Student Retention in Higher Education - a Case Study. *Proceedings of the 12th International Conference on Enterprise Information Systems, Volume 1, DISI, Funchal, Madeira, Portugal, June 8 - 12, 2010*.
- Zhang, Y., Qiang, F., Quddus, M. & Davis, C. (2014) An Examination of the Impact of Early Intervention on Learning Outcomes of At-Risk Students. *Research in Higher Education Journal*, 26.
- Zweig, M.H. & Campbell, G. (1993) Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, 39(4), 561-77.

Anhang:

Tabelle 66. Faktoranalyse UNI Semester 0

	F1	F2	F3	F4	F5	F6	MSA&KMO
Geburtsortausland	0,57						0,87
StBAusland	0,80						0,84
Migration	0,88						0,68
Land	0,89						0,69
HZB		0,59					0,62
HZBNote		0,61					0,69
Akad. Vorleistung			0,90				0,51
VorSem			0,92				0,48
KreisdHKB				0,53			0,61
Studienform				0,63			0,56
StartSem					-0,46		0,82
ImmtrAlter					0,81		0,54
AnmeldeZeit					0,83		0,55
Geschlecht						-0,63	0,59
KV						-0,53	0,62
Studiengang						0,71	0,59
Region							0,72
Ergebnis	-0,11	-0,63	0,15	-0,12	-0,23	0,27	0,64

Tabelle 67. Faktoranalyse UNI Semester 1

	F1	F2	F3	F4	F5	F6	F7	F8	F9	MSA&KMO
Geburtsortausland	0,60									0,87
StBAusland	0,78									0,84
Migration	0,88									0,68
Land	0,90									0,71
HZBNote		0,45								0,77
HZB		0,49								0,69
Region		0,77								0,57
Akad. Vorleistung			0,90							0,52
VorSem			0,92							0,47
Studienform				0,75						0,87
StartSem					-0,47					0,59
ImmtrAlter					0,79					0,57
AnmeldeZeit					0,83					0,64
NoteSem1						0,72				0,79
AnzahlScheinSem1						0,73				0,68
CPSem1						0,85				0,70
NTSem1							0,51			0,76
NBSem1							0,67			0,77
Schein1								0,74		0,74
Schein2								0,73		0,84
Schein3								0,60		0,84
Studiengang								0,73		0,90
KV								-0,42		0,78
Schein4									0,56	0,77
Schein5									0,66	0,87
Schein6									0,52	0,83
Ergebnis	-0,06	-0,05	0,13	-0,25	-0,10	0,61	-0,28	0,20	0,03	0,73

Tabelle 68. Faktoranalyse UNI Semester 2

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	MSA&KMO
Geburtsortausland	0,61										0,87
StBAusland	0,78										0,84
Migration	0,89										0,69
Land	0,90										0,72
HZBNote		0,42									0,78
HZB		0,55									0,71
Region		0,71									0,53
Akad. Vorleistung			0,87								0,56
VorSem			0,92								0,47
StartSem				-0,46							0,91
ImmtrAlter				0,74							0,64
AnmeldeZeit				0,81							0,58
NoteSem1					0,69						0,80
AnzahlScheinSem1					0,73						0,65
CPSem1					0,77						0,74
NoteSem2						0,57					0,87
AnzahlScheinSem2						0,82					0,71
CPSem2						0,87					0,73
NBSem1							0,74				0,76
NBSem2							0,79				0,58
KV								0,48			0,83
NTSem2								0,54			0,73
Studienform								0,55			0,64
NTSem1								0,56			0,81
Studiengang									0,69		0,87
Schein6									0,72		0,95
Schein5									0,81		0,92
Schein4									0,83		0,92
Schein3									0,84		0,91
Schein1									0,87		0,89
KreisderHZB										0,81	0,82
Geschlecht											0,88
Schein2											0,87
Ergebnis	-0,05	-0,02	0,08	-0,19	0,31	0,63	-0,25	-0,06	0,20	-0,01	0,79

Tabelle 69. Faktoranalyse UNI Semester 3

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	MSA&KMO
Geburtsortausland	0,61										0,87
StBAusland	0,78										0,84
Migration	0,88										0,70
Land	0,90										0,73
HZBNote		0,44									0,79
HZB		0,54									0,72
Region		0,74									0,53
Studienform			0,44								0,63
KreisderHZB			0,68								0,85
StartSem				-0,45							0,92
ImmatrAlter				0,72							0,69
AnmeldeZeit				0,82							0,68
AnzahlScheinSem1					0,63						0,61
CPSem1					0,75						0,79
NoteSem1					0,77						0,86
NoteSem2						0,47					0,91
NoteSem3						0,48					0,91
CPSem3						0,76					0,72
AnzahlScheinSem3						0,78					0,75
AnzahlScheinSem2						0,80					0,66
CPSem2						0,81					0,73
NBSem1							0,65				0,87
NBSem3							0,77				0,81
NBSem2							0,82				0,71
NTSem3								0,59			0,84
NTSem1								0,66			0,85
NTSem2								0,70			0,78
Akad. Vorleistung									0,87		0,58
VorSem									0,92		0,48
Studiengang										0,72	0,92
Schein6										0,74	0,97
Schein5										0,82	0,95
Schein4										0,84	0,95
Schein3										0,86	0,95
Schein2										0,88	0,93
Schein1										0,89	0,93
Ergebnis	-0,03	-0,04	-0,05	-0,16	0,29	0,66	-0,25	-0,07	0,08	0,20	0,82

Tabelle 70: Faktoranalyse UNI Semester 4

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	MSA& KMO
Geburtsortausland	0,58											0,88
Region	0,59											0,85
StBAusland	0,73											0,84
Migration	0,89											0,74
Land	0,90											0,76
HZBNote		0,44										0,75
HZB		0,65										0,70
Akad. Vorleistung			0,87									0,62
VorSem			0,92									0,85
Studienform				0,84								0,60
AnmeldeZeit					-0,79							0,48
ImmatrAlter					-0,73							0,65
StartSem					0,40							0,71
KreisderHZB						-0,68						0,93
NoteSem1							0,77					0,93
NoteSem2							0,48					0,92
NoteSem3							0,48					0,92
NoteSem4							0,42					0,69
CPSem1							0,74					0,81
AnzahlScheinSem1							0,60					0,88
CPSem2								0,79				0,76
CPSem3								0,67				0,75
CPSem4								0,69				0,78
AnzahlScheinSem2								0,81				0,78
AnzahlScheinSem3								0,74				0,74
AnzahlScheinSem4								0,76				0,69
NTSem1									0,53			0,89
NTSem3									0,71			0,85
NTSem3									0,70			0,86
NTSem4									0,69			0,85
NBSem1										0,56		0,89
NBSem2										0,78		0,88
NBSem3										0,76		0,85
NBSem4										0,73		0,79
Studiengang											0,74	0,94
Schein6											0,81	0,95
Schein5											0,86	0,95
Schein3											0,88	0,96
Schein2											0,89	0,94
Schein4											0,89	0,96
Schein1											0,90	0,94
Ergebnis	-0,03	-0,09	-0,07	0,13	0,09	0,20	0,32	0,64	-0,08	-0,24	0,19	0,84

Tabelle 71: Faktoranalyse FH Semester 0

	F1	F2	F3	F4	F5	MSA&KMO
StBAusland	0,72					0,81
Region	0,77					0,78
Migration	0,85					0,82
Land	0,90					0,73
KreisderHQB		0,46				0,83
Geburtsort		0,57				0,57
StartSem		0,70				0,54
HQB			0,43			0,74
Studienform			0,75			0,56
ImmatrAlter			0,76			0,62
Akad. Vorleistung				0,87		0,58
VorSem				0,87		0,58
Studiengang					-0,54	0,52
Geschlecht					0,68	0,67
Ergebnis	-	0,04	-0,28	-0,10	-0,44	0,71
	0,12					

Tabelle 72: Faktoranalyse FH Semester 1

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	MSA&KMO
Region	0,83										0,78
Migration	0,85										0,75
Land	0,90										0,69
KreisderHQB		0,50									0,68
Geburtsort		0,61									0,85
StartSem		0,61									0,58
HQB			0,45								0,55
ImmatrAlter			0,75								0,77
Studienform			0,75								0,68
VorSem				0,81							0,64
Akad. Vorleistung				0,84							0,69
Geschlecht					0,79						0,65
anzahlscheine						0,52					0,75
CPSem1						0,90					0,57
AnzahlScheinSem1						0,91					0,57
NTSem1							-0,67				0,77
VorVersuch							-0,48				0,76
NBSem1								0,76			0,82
NoteSem1								0,71			0,59
Schein6									-0,81		0,64
Schein1									0,53		0,31
Schein2									0,57		0,82
Studiengang										-0,70	0,70
Schein3										0,47	0,38
Schein4										0,62	0,63
Schein5										0,63	0,54
Ergebnis	-0,07	-0,16	-0,04	-0,12	-0,22	0,30	0,55	-0,22	-0,01	0,02	0,65

Tabelle 73: Faktoranalyse FH Semester 2

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	MSA& KMO
Region	0,82											0,78
Migration	0,85											0,75
Land	0,89											0,69
KreisderHZB		0,55										0,85
Geburtsort		0,78										0,58
StartSem			0,80									0,49
ImmtrAlter				0,46								0,70
VorSem				0,84								0,69
Akad. Vorleistung				0,85								0,67
Studiengang					-0,62							0,30
Schein3					0,60							0,45
Geschlecht					0,62							0,70
Schein6						0,53						0,69
AnzahlScheinSem1						0,91						0,58
CPSem1						0,91						0,57
NTSem2							-0,52					0,70
NTSem1							-0,52					0,75
Studienform							0,65					0,64
NBSem1								0,63				0,71
NoteSem1								0,69				0,72
NBSem2								0,69				0,71
NoteSem1								0,75				0,67
AnzahlScheinSem2									0,90			0,62
CPSem2									0,91			0,62
Schein1										-0,59		0,52
Schein5										0,50		0,50
Schein4										0,61		0,59
Schein2											0,84	0,35
HZB												0,78
VorVersuch												0,72
Ergebnis	-0,07	0,01	-0,06	-0,23	-0,11	0,19	0,26	-0,06	0,58	-0,09	0,11	0,64

Tabelle 74: Faktoranalyse FH Semester 3

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	MSA& KMO
Geburtsortausland	0,61										0,79
StBAusland	0,78										0,77
Migration	0,88										0,70
Land	0,90										0,52
HZBNote		0,44									0,56
HZB		0,54									0,63
Region		0,74									0,73
Studienform			0,44								0,34
KreisderHZB			0,68								0,72
StartSem				-0,45							0,67
ImmatrAlter				0,72							0,70
AnmeldeZeit				0,82							0,56
AnzahlScheinSem1					0,63						0,55
CPSem1					0,75						0,71
NoteSem1					0,77						0,68
NoteSem2						0,47					0,68
NoteSem3						0,48					0,92
CPSem3						0,76					0,72
AnzahlScheinSem3						0,78					0,72
AnzahlScheinSem2						0,80					0,82
CPSem2						0,81					0,83
NoteSem2							0,65				0,82
NBSem3							0,77				0,82
NBSem2							0,82				0,83
NTSem3								0,59			0,80
NTSem 1								0,66			0,82
NTSem2								0,70			0,82
Akad. Vorleistung									0,87		0,78
VorSem									0,92		0,37
Studiengang										0,72	0,51
Schein6										0,74	0,76
Schein5										0,82	0,78
Schein4										0,84	0,67
Schein3										0,86	0,82
Schein2										0,88	0,86
Schein1										0,89	0,79
Geschlecht											0,77
KV											0,70
Ergebnis	-0,03	-0,04	-0,05	-0,16	0,29	0,66	-0,25	-0,07	0,08	0,20	0,71

Tabelle 75: Faktoranalyse FH Semester 4

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	MSA& KMO
Region	0,83													0,79
Migration	0,84													0,77
Land	0,89													0,71
KreisderHZB		0,46												0,85
Geburtsort		0,80												0,54
ImmatrAlter			0,50											0,70
Studienform			0,55											0,67
VorSem				0,81										0,74
Akad. Vorleistung				0,81										0,76
Studiengang					-0,88									0,37
Schein6					0,61									0,75
StartSem						-0,48								0,46
VorVersuch						0,69								0,71
AnzahlScheinSem1							0,91							0,55
CPSem1							0,93							0,54
AnzahlScheinSem2								0,90						0,69
CPSem2								0,91						0,69
AnzahlScheinSem3									0,85					0,70
CPSem3									0,88					0,69
AnzahlScheinSem4										0,88				0,71
CPSem4										0,86				0,69
NTSem1											0,58			0,74
NTSem2											0,69			0,83
NTSem3											0,69			0,80
NTSem4											0,55			0,79
NoteSem1												0,62		0,86
NoteSem2												0,70		0,86
NoteSem3												0,71		0,86
NoteSem4												0,65		0,88
NBSem1												0,55		0,84
NBSem2												0,64		0,87
NBSem3												0,65		0,85
NBSem4												0,64		0,88
Schein2													0,54	0,43
Schein3													0,59	0,52
Schein4													0,80	0,76
Schein5													0,82	0,67
Ergebnis	-0,04	0,03	-0,12	0,15	-0,02	-0,15	0,10	0,21	0,46	0,50	-0,12	0,00	0,19	0,73