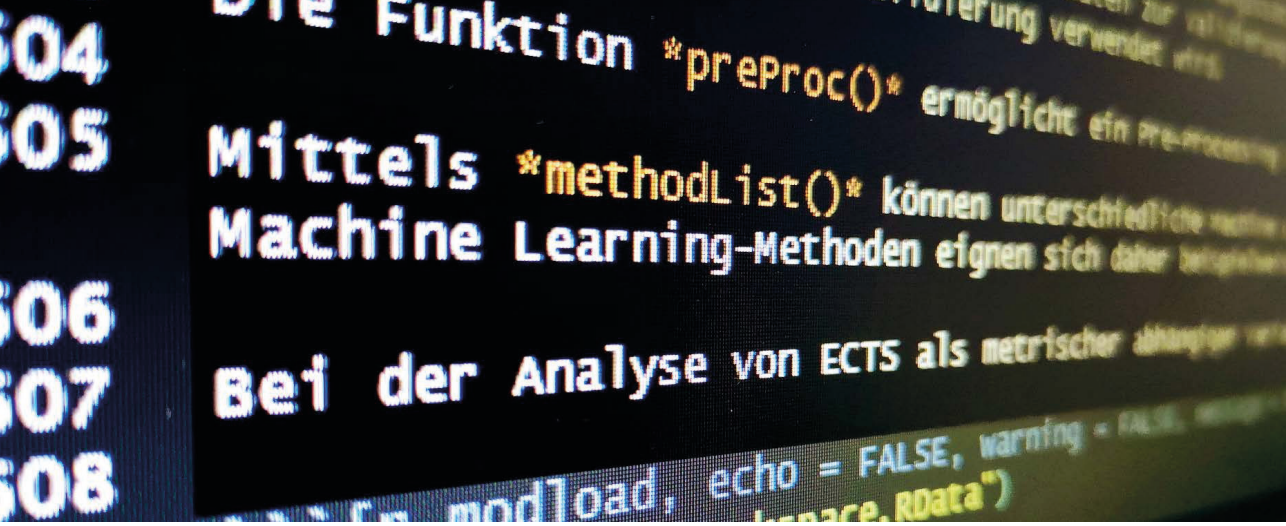
Bartok, Larissa; Spörk, Julia; Gleeson, Robin; Krakovsky, Maria; Ledermüller, Karl

Anwendung statistischer und Machine-Learning-Methoden für Fragestellungen zu Studienerfolg. Erfahrungen in den Projekten "Learning Analytics - Studierende im Fokus" und "PASSt - Predictive Analytics Services für Studienerfolgsmanagement"

Münster ; New York : Waxmann 2024, 77 S.



Larissa Bartok, Julia Spörk, Robin Gleeson,
Maria Krakovsky, Karl Ledermüller

Anwendung statistischer und Machine-Learning-Methoden für Fragestellungen zu Studienerfolg

WAXMANN

Larissa Bartok, Julia Spörk, Robin Gleeson,
Maria Krakovsky, Karl Ledermüller

Anwendung statistischer und Machine-Learning-Methoden für Fragestellungen zu Studienerfolg

Erfahrungen in den Projekten „Learning Analytics – Studierende im Fokus“ und „PASSt – Predictive Analytics Services für Studienerfolgsmanagement“

Waxmann 2024
Münster · New York

Diese Publikation wurde durch die Ausschreibung zur „Digitalen und sozialen Transformation in der Hochschulbildung“ vom österreichischen Ministerium für Bildung, Wissenschaft und Forschung (BMBWF) ermöglicht. Nähere Informationen zur Ausschreibung und den zugehörigen Projekten finden sich unter:
https://www.bmbwf.gv.at/Ministerium/Presse/Digitale-soziale-Transformation-HS.html

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.dnb.de abrufbar.

Print-ISBN 978-3-8309-4883-4
E-Book-ISBN 978-3-8309-9883-9
https://doi.org/10.31244/9783830998839

Waxmann Verlag GmbH, 2024
Steinfurter Straße 555, 48159 Münster

www.waxmann.com
info@waxmann.com

Umschlaggestaltung: Anne Breitenbach, Münster
Umschlagabbildung: Robin Gleeson
Satz: satz&sonders, Dülmen

Inhalt

Danksagung

Im Rahmen der Digitalisierungsinitiative des Bundesministeriums für Bildung, Wissenschaft und Forschung (BMBWF) wurden die beiden Projekte „PASSt: Predictive Analytics Services für Studierendenerfolgsmanagement“ und „Learning Analytics – Studierende im Fokus“ kofinanziert. Die beiden thematisch ähnlichen Projekte wurden zu dem inhaltlichen Cluster „Learning Analytics“ zusammengeführt. Der vorliegende Erfahrungsbericht fasst die wesentlichen Erkenntnisse aus der Zusammenarbeit der gemeinsamen Arbeitsgruppe des Clusters zusammen und stellt diese einer breiteren Leser*innenschaft zur Verfügung.

Unser besonderer Dank gilt den Kolleg*innen der Projektteams, insbesondere der Projektleitung von PASSt, Dr. Shabnam Tauböck, und Learning Analytics, Dr. Martin Ebner, welcher auch als Leiter des projektübergreifenden Clusters fungierte. Die offene Kooperation sowie der Austausch von Ideen und Erfahrungen waren eine wesentliche Grundlage für den Erfolg der Zusammenarbeit. Unsere Vorhaben sowie die vorliegende Publikation waren nur möglich, weil viele engagierte Personen zum Gelingen beigetragen haben. Insbesondere bedanken wir uns für das außerordentlich hilfreiche und detaillierte Feedback von vielen Kolleg*innen aller beteiligter Universitäten. Ebenso möchten wir uns herzlich bei unserem Kollegen René Krempkow für die wertschätzende und hilfreiche Rückmeldung sowie die Gelegenheit, unsere Überlegungen zu diskutieren, bedanken, ebenso für sein Vorwort.

Dem BMBWF möchten wir für die Möglichkeit danken, im Rahmen der beiden aus der Ausschreibung kofinanzierten Projekte zu kooperieren. Die Zusatzfinanzierung durch das BMBWF war die Grundvoraussetzung für die erfolgreiche Durchführung der beiden Projekte und die damit verbundenen Weiterentwicklungen der beteiligten Universitäten.

Wir danken darüber hinaus allen Beteiligten für ihre Beiträge, die konstruktive Zusammenarbeit und ihr Engagement und hoffen, dass dieser Erfahrungsbericht auch anderen Projekten und Initiativen im Bereich Learning Analytics als wertvolle Ressource und zur kollegialen Weiterentwicklung des Themenfeldes dienen wird.

Vorwort

Mehr Studienerfolg mit Machine Learning, KI & Co.?!

René Krempkow

Seit November 2022 ist mit ChatGPT4 die Artificial Intelligence (AI) und das Machine Learning endgültig in der öffentlichen Wahrnehmung angekommen. Ausschlaggebend dafür war die medial sehr erfolgreiche Veröffentlichung des Large-Language-Models GPT 3–5 von Open.AI. An den Hochschulen und in der Wissenschaft wurde sich mit dem Themenkreis von einschlägigen Fachdisziplinen bereits länger befasst und dies auch (mehr oder weniger) öffentlich thematisiert. Dabei wurde nicht nur die aktuell öffentlich besonders diskutierte textgenerierende bzw. generative AI in den Fokus genommen, sondern auch die analytische AI, die hier im Vordergrund stehen soll.

Die Nutzung solcher Ansätze für Fragestellungen zu Studienerfolg wurde aber bisher nur recht selten thematisiert. In jüngerer Zeit gab es zwar durchaus einige Veröffentlichungen (z. B. Spörk et al., 2021; Schmidt, 2023; Lübcke et al., 2023), sie sind bisher jedoch eher die Ausnahme als die Regel in Publikationen, die sich mit Studienerfolg befassen.[1]

Auch Lübcke et al. (2021) schätzen ein, dass der derzeitige Forschungsstand zu Digitalen Studienassistenzsystemen, die sich regelbasierter KI-Methoden

1 Über die erwähnten Beiträge hinaus wurde dies in einschlägigen Publikationsorganen im deutsch-sprachigen Raum in den letzten Jahren noch kaum thematisiert, die Ergebnisse entsprechender geförderter Projekte erscheinen allerdings naturgemäß meist erst zu deren Ende (vgl. z. B. Daniel et al., 2019; Krempkow et al., 2021; Falk et al., 2022; Krempkow, 2022). Ähnlich gilt dies für Sammelwerke (z. B. Bornkessel, 2018; Neugebauer et al., 2021). Künftig ist aber eine vermehrte Thematisierung zu erwarten, darauf deutet z. B. ein Call for Papers der Zeitschrift für Hochschulentwicklung – ZFHE hin (Lübcke et al., 2021).Beispielsweise an der Humboldt-Universität zu Berlin gab es bereits vor einigen Jahren eine Nutzung von Ansätzen analytischer AI und Machine Learning zusätzlich zu Regressionsanalysen und Diskussionen der jeweiligen Vor- und Nachteile. Dies wurde aber nicht über einen kleinen Kreis hinaus bekannt, da es keine Veröffentlichung in einschlägigen Publikationsorganen gab und es dem Verfasser auch nur durch den berufsbiografischen Zufall seiner zeitweisen Tätigkeit dort bekannt ist. Publizität kann neben einschlägigen Zeitschriften, Buchbänden und Online zwar grundsätzlich auch über die Teilnahme und den Austausch in einschlägigen Tagungen erreicht werden. Jedoch ist dies erfahrungsgemäß flüchtiger. Zudem sind die finanziellen Hürden höher, da dies meist teurer ist. Teilweise sind erhebliche Kostenanteile für Tagungen privat finanziert (vgl. z. B. Janson & Rathke, 2023, S. 132).

und Ansätzen maschinellen Lernens bedienen, noch überschaubar ist. Dies bedeutet nicht, dass es keine Versuche zur Nutzung gab.[2]

Doch ist ein Austausch von Erfahrungen über regionale Bezüge hinaus, mögliche Nachnutzungen und Weiterentwicklungen auf der Basis vorhandenen Wissens ohne ausreichende Publizität naturgemäß mit größeren Schwierigkeiten behaftet. Er findet dementsprechend seltener bzw. kaum statt.[3]

Hier setzt die vorliegende Publikation eines Erfahrungsberichtes zweier einschlägiger Projekte dankenswerterweise an. Denn die Art der Aufbereitung erlaubt es, die nachfolgend in der Publikation vorgestellten Modelle auch an anderen Hochschulen anzuwenden, sofern die Gelingensbedingungen erfüllt sind. Sie richtet sich hierbei einerseits an Personen, die für die inhaltliche Anwendung der Modelle verantwortlich zeichnen, und andererseits an Personen, die mit der technischen Implementierung in der Hochschule vertraut sind. Dieser Ansatz erscheint nach vorliegenden Erfahrungen und Studien gut geeignet, die Nutzung von empirischen Daten und Ergebnissen empirischer Analysen auch für Studienerfolg und Lehrentwicklung zu fördern, da dies auch gut ermöglicht, Narrative aus quantitativen Ergebnissen zu formulieren (Isett & Hicks, 2020) und Stakeholder*innen einzubinden (Wegner et al., 2023; Janson et al., 2024).

Förderlich für ihre künftig stärkere Nutzung dürfte auch die übersichtliche Systematisierung von Analytics-in-Higher-Education als Schaubild sein, wie sie in dieser Publikation zu Beginn des zweiten Kapitels vorgestellt wird. Darin erfolgt einerseits eine hierarchische Stufenordnung nach der Komplexität der verwendeten analytischen Methoden, andererseits eine die praktische Anwendung vereinfachende Zuordnung zu beschreibenden Modellen und Prognosemodellen. Hierbei erläutern die Autor*innen auch anhand von Beispielen die unterschiedlichen Stärken der Modelle, welche in den jeweiligen Anwendungskontexten zu berücksichtigen sind: So können beschreibende Modelle eher anhand von Daten aus Vergangenheit oder Gegenwart Informationen anschaulich aufbereitet darstellen (descriptive analytics) bzw. helfen, Ursachen für identifizierte Sachverhalte und Probleme zu finden (diagnostic analytics). Prognosemodelle können dagegen eher helfen, Wahrscheinlichkeiten für künf-

2 Beispielsweise an der Humboldt-Universität zu Berlin gab es bereits vor einigen Jahren eine Nutzung von Ansätzen analytischer AI und Machine Learning zusätzlich zu Regressionsanalysen und Diskussionen der jeweiligen Vor- und Nachteile. Dies wurde aber nicht über einen kleinen Kreis hinaus bekannt, da es keine Veröffentlichung in einschlägigen Publikationsorganen gab und es dem Verfasser auch nur durch den berufsbiografischen Zufall seiner zeitweisen Tätigkeit dort bekannt ist.

3 Publizität kann neben einschlägigen Zeitschriften, Buchbänden und Online zwar grundsätzlich auch über die Teilnahme und den Austausch in einschlägigen Tagungen erreicht werden. Jedoch ist dies erfahrungsgemäß flüchtiger. Zudem sind die finanziellen Hürden höher, da dies meist teurer ist. Teilweise sind erhebliche Kostenanteile für Tagungen privat finanziert (vgl. z. B. Janson & Rathke, 2023, S. 132).

tige Entwicklungen vorherzusagen, z. B. individuellen oder kollektiven Studienfortschritt (predictive analytics), oder Auswirkungen von Veränderungen zu simulieren, z. B. eines Curriculums (presciptive analytics).

Die Autor*innen beschränken sich aber nicht auf die Systematisierung und Beschreibung von Modellen, sondern geben einen Überblick und eine anwendungsbezogene kurze Erläuterung von häufig zur Modellierung von Studienerfolg verwendeten statistischen Methoden auf jeweils ein bis zwei Seiten und mit Abbildungen veranschaulicht. Anschließend stellen sie kurz und prägnant die Datenbasis vor, die sie für zwei im Folgenden ausführlicher erläuterte und dokumentierte Anwendungsszenarien zugrunde legen. Die gut nachvollziehbare und aufgrund der abgedruckten sowie online bereitgestellten R-Codes nachnutzbare Dokumentation der beiden Anwendungsszenarien ist das eigentliche Kernstück des Erfahrungsberichtes. Hier kommt ein besonders glücklicher (bzw. aufgrund gelungener Zusammenarbeit in besonders geeigneter Weise geschaffener) Umstand zum Tragen: Denn es erfolgt durch den engen Erfahrungsaustausch zweier an unterschiedlichen Hochschulen derselben Stadt bereits existenter Projekte eine Art konzertierte Aktion bezüglich Datenbasis und Analyseansätzen. So ist es möglich, an konkreten Anwendungsbeispielen in vergleichbarer Weise die unterschiedlichen Stärken von beschreibenden Modellen und Prognosemodellen zu veranschaulichen. Damit werden zugleich auch die unterschiedlichen Bedingungen an verschiedenen Hochschulen anhand der beiden Anwendungsszenarien konstruktiv genutzt, um zu zeigen, wie eine Übertragbarkeit bzw. Adaption von Ansätzen gemäß unterschiedlichen Zielen möglich sein kann.

So zeigt das erste Anwendungsszenario der Universität Wien, wie mittels Regressionsanalysen die Frage beantwortet werden kann, welche Faktoren in einem bestimmten Studienprogramm Einfluss auf den Studienerfolg haben (gemessen in Anzahl Kreditpunkte ECTS im Studienjahr). Die Ergebnisse werden für Stakeholder*innen in Form von Diagrammen aufbereitet. Er wird anschaulich gezeigt, dass neben den ECTS im Vorjahr (mit erwartungsgemäß größtem Effekt) das Alter bei Studienbeginn, Auslandssemester und Doppelstudium signifikante Effekte haben. Dieselben Faktoren gelten in diesem Anwendungsbeispiel auch, wenn man anstelle von ECTS die Einflussfaktoren auf Prüfungs(in)aktivität im Studienjahr analysiert.

Das zweite Anwendungsszenario der Wirtschaftsuniversität Wien zeigt, wie mittels ähnlicher Daten die Anzahl der ECTS und die Prüfungs(in)aktivität im Folgejahr prognostiziert werden können. Die Verwendung dieser Prognose hängt von den Zielen der Hochschule ab. Im Beispiel wird eine Zuordnung zu Clustern gezeigt, um zielgerichtet Maßnahmen für unterschiedliche Gruppen entwickeln und anbieten zu können. Ein zielgerichtetes Informations- und Serviceangebot für Personen in einer relativ homogenen Gruppe (etwa bezüglich

Altersstruktur, Erwerbstätigkeit) könnten dann dabei helfen, Studienerfolg und -zufriedenheit zu erhöhen.

Da viele Hochschulen im deutschsprachigen Raum in den letzten Jahren Studienverlaufsanalysen einführten bzw. noch einführen, haben die in den Anwendungsszenarien vorgestellten Ansätze großes Potenzial für eine Nutzung an weiteren Hochschulen. Hierbei könnte es hilfreich sein, mögliche Ansatzpunkte für spätere Maßnahmen bereits bei der Planung bzw. Weiterentwicklung von Datenerhebungen mitzudenken. Beispielsweise wäre, wenn vermutet werden kann, dass ein Einflussfaktor auf den Studienerfolg die Erwerbstätigkeit von Studierenden ist (bzw. ein De-Facto-Teilzeitstudium)[4], deren Erfassung nützlich, um eine möglichst hohe Erklärungskraft bei beschreibenden Modellen bzw. eine möglichst hohe Vorhersagekraft bei Prognosemodellen erreichen zu können. Ähnlich gilt dies für die soziale Herkunft bzw. Bildungsherkunft, die ebenfalls Einfluss haben kann.[5]

Auf längere Sicht könnten solche Modelle – beschreibend oder prognostisch – über ihren konkreten Nutzen für einzelne Studienprogramme und Hochschulen hinaus auch einen Systemnutzen stiften: Sie sind gute Beispiele für Verknüpfungsmöglichkeiten mit einer empirisch informierten Lehrentwicklung an Hochschulen, die insgesamt zu einer Kultur stärkerer Evidenzorientierung bei Entscheidungen an Hochschulen und damit einer noch systematischeren Förderung von Studienerfolg beitragen könnten. Darüber hinaus könnte – analog zu erfolgreichen Beispielen im Schulbereich[6] – mit Hilfe solcher Modelle bei adäquater Formulierung von Zielen und entsprechenden Maßnahmen nicht nur eine gezieltere Förderung von bestimmten Studierendengruppierungen erfolgen. Vielmehr könnte dies auch für Hochschulen geschehen, die sich z. B. aufgrund ihres Profils, ihrer geografischen Lage oder ihres Rekrutierungspotenzials in besonderer Weise mit deren Förderung befassen (müssen). Hierfür gibt es bereits Beispiele – bisher v. a. außerhalb deutschsprachiger Länder, so in

4 Für Hochschulen in Deutschland ist dies ein bedeutsamer Einflussfaktor. Daneben gilt dies fächerübergreifend auch für die Einflussfaktoren Geschlecht und Note der Hochschulzugangsberechtigung, sowie in mehreren Fächergruppen und großen Fächern zudem für Elternschaft, Auslandsaufenthalte und einen Berufsabschluss vor dem Studium (Krempkow, 2020a).

5 Dies zeigte sich in hochschulspezifischen Analysen mit denselben Modellen (Krempkow, 2020b).

6 So gelang es dem Stadtstaat Hamburg in den vergangenen zehn Jahren, bei den Schülerleistungen in mehreren Leistungsvergleichen zur Spitzengruppe aufzuschließen bzw. sich teilweise sogar ganz vorn zu platzieren, bei zugleich hierfür schwieriger Sozialstruktur. Als maßgeblicher Grund hierfür wird der strategisch-langfristige Ansatz einer datenbasierten Schulentwicklung und das Förderprogramm „starke Schulen" für Schulen mit sozialstrukturell bedingtem Förderbedarf genannt, was parteiübergreifend Anerkennung finde und als Vorbild ähnlicher Vorhaben diene (Wiarda, 2024).

Finnland, UK, Australien –, das Potenzial dafür ist aber im deutschsprachigen Raum ebenfalls vorhanden.[7] Angesichts veränderter Studierendenpopulationen könnte dies nicht nur ein wichtiger Beitrag zu den individuellen Zukunftschancen von Studierenden sein, sondern auch zur Zukunftsfähigkeit der Hochschulstudien insgesamt beitragen.

7 Dies legen Reviews für diese Länder und Empirie nahe (vgl. Sörlin, 2007; Harris, 2007; Krempkow, 2015).

Abstract

Auf statistischen Modellen aufbauende Analytics-Instrumente können dabei helfen, mehr über den Lern- und Studienerfolg von Studierenden herauszufinden oder Prognosen auf aggregierter oder individueller Ebene zu erstellen. Zwei Projekte, die sich dieser und ähnlicher Problemstellungen widmen, wurden vom österreichischen Bundesministerium für Bildung, Wissenschaft und Forschung im Rahmen der Ausschreibung „Digitale und soziale Transformation in der Hochschulbildung" finanziert. Die beiden Projekte „Learning Analytics – Studierende im Fokus" und „PASSt – Predictive Analytics Services für Studienerfolgsmanagement" wurden zur Generierung von Synergieeffekten im Rahmen des Clusters Learning Analytics konzeptionell verzahnt, indem generische Kernprobleme gemeinsam bearbeitet und Lessons learned diskutiert wurden. Im Rahmen des Clusters wurde im Juli 2020 die gemeinsame Arbeitsgruppe „Variablendefinition und Modellbildung" gebildet, um den projektübergreifenden Wissensaustausch bei der Anwendung von statistischen Modellen im Bereich von (Learning) Analytics zu fördern. Die Erkenntnisse der Arbeitsgruppe werden in dieser Arbeit vorgestellt.

Dieser Erfahrungsbericht will einen Überblick zum methodischen Instrumentarium geben, das Anwender*innen bei der Modellierung von Studienerfolg zur Verfügung steht. Dabei werden nicht nur theoretische Überlegungen zur Modellierung diskutiert, sondern auch konkret illustriert, wie entlang von beschreibenden oder prädiktiven Anwendungsszenarien modellbasierte Analytics-Instrumente eingesetzt werden können. Im Bericht werden beispielhaft Analysen mittels reproduzierbarem Programmcode (inklusive eigens entwickelter R-Funktionen) in der Open-Source-Programmiersprache R (R Core Team, 2022) anschaulich dargestellt.

Entlang der beiden definierten Anwendungsszenarien der Universitäten (Beschreibung vs. Prognose) wird zudem beschrieben, inwiefern sich bestimmte statistische Verfahren für bestimmte Zielsetzungen besser eignen als andere. Wesentlich ist hier die Feststellung, dass passgenauere Modelle mit hoher Treffsicherheit oft mit einer eingeschränkten Interpretierbarkeit einzelner Einflussfaktoren einhergehen (Blackbox-Problem). Es wird anhand zweier Anwendungsszenarien mit anonymisierten Datensätzen der Universität Wien und der Wirtschaftsuniversität Wien gezeigt, wie sich unterschiedliche Zielsetzungen auf die Modell- und Methodenwahl auswirken können. Dabei stellt diese Arbeit keinen Anspruch auf Vollständigkeit aller potenziellen Anwendungsszenarien und Zielsetzungen, sondern will in erster Linie die Erfahrungen der Arbeitsgruppe widerspiegeln. Zusätzlich zu dem beschriebenen Ziel hält der

Erfahrungsbericht aus der Literatur und in unserer Praxis bestätigte Gelingensbedingungen für die Implementation von Higher Education Analytics an der Hochschule fest. Dazu gehören Bedingungen wie eine gute Datenqualität, Datenmanagement, Data Governance und Expertise an der Hochschule.

Weiters folgt eine erfahrungsbasierte Einschätzung, welche Bedingungen für die Implementierung an Hochschulen als Voraussetzung geschaffen werden müssen. In der Conclusio wird die Gesamteinbettung der beschriebenen Modelle zu Studienerfolg und Prüfungs(in)aktivität und deren Grenzen diskutiert.

1 Einleitung

1.1 Was ist das Ziel dieses Erfahrungsberichts?

Dieses Dokument verfolgt mehrere Ziele:

- Es informiert anhand unserer Erfahrung über Möglichkeiten und Grenzen von Prognosemodellen zu Studienerfolg und Prüfungs(in)aktivität.
- Es gibt einen Überblick über das methodische Instrumentarium, das Anwender*innen zur Verfügung steht.
- Es schafft einen konkreten Einblick in die direkte Anwendung mit der Hilfe der Software R (R Core Team, 2022).

Es wird anhand zweier Anwendungsszenarien mit anonymisierten Datensätzen der Universität Wien und der Wirtschaftsuniversität Wien gezeigt, wie Prognosemodelle im Kontext der Modellierung von Studienerfolg oder Prüfungs(in)aktivität konkret umgesetzt werden können und welche Methoden sich gut eignen, um die zugrundeliegende Fragestellung zu beantworten. Die Anwendungsfälle sind mit *rmarkdown* (Allaire et al., 2021) umgesetzt und enthalten sowohl den Codeinput, der notwendig ist, um statistische Prognosemodelle zu implementieren, als auch deren Ergebnisse. Diese Art der Aufbereitung erlaubt es, sofern die Gelingensbedingungen an der jeweiligen Hochschule erfüllt sind, die vorgestellten Modelle auch an anderen Hochschulen anzuwenden.

Dieses Dokument richtet sich einerseits an Personen, die mit der technischen Implementierung in der Hochschule vertraut sind, andererseits an Personen, die für die inhaltliche Anwendung der Modelle verantwortlich zeichnen. Daher enthält es neben zwei nachvollziehbaren Beispielen mit dem Programmiercode (für die technische Implementierung) auch inhaltliche Information zur Interpretation und Nutzung der Ergebnisse (zur inhaltlichen Implementierung). Auch Praktiker*innen können die Interpretationshilfen an ihrer Institution implementieren um Entscheidungsträger*innen die Interpretation zu vereinfachen.

1.2 Warum werden statistische Prognosemodelle im Hochschulkontext angewandt und welche Fragestellungen können beantwortet werden?

Hochschulen haben großes Interesse daran, den Studienerfolg ihrer Studierenden empirisch zu quantifizieren und besser zu verstehen, um passgenaue

Unterstützungsmaßnahmen ableiten zu können. Auch kann es darum gehen, Studienerfolg differenziert vorhersagen zu können. Die Gründe dafür sind vielfältig und reichen von Aussagen zur Studierbarkeit im Allgemeinen bis hin zur Diversitätsgerechtigkeit eines Studienprogramms (Buß, 2019) oder auch zur Vorhersage des individuellen Studienabbruchrisikos (Beaulac & Rosenthal, 2018). Auch können derartige Prognosen für die Planung von Studienplätzen Anwendung finden. In dieser Arbeit wird unterschieden zwischen dem Anwendungsfeld der Prognose (beispielsweise zur Planung, durch Schätzung von Dropoutrisiken oder Erfolgsfaktoren) und dem Anwendungsfeld der Erklärung oder Beschreibung (beispielsweise zur Quantifizierung der Einflussfaktoren auf den Studienerfolg in einem Studienprogramm).

Diese aktuellen und praktischen Problemstellungen können mit Hilfe von Prognose- und Beschreibungsmodellen (wie sie in der multivariaten Statistik angewandt werden) zielgerichtet analysiert werden. Die verwendeten Modelle versuchen durch die *Reduktion von Komplexität und der Fokussierung auf relevante und verfügbare Wirkgrößen Zusammenhänge zu erklären und zu beschreiben, zukünftige Entwicklungen zu prognostizieren und durch die Anwendung am Modell Entscheidungsgrundlagen zu liefern* (Stoetzer, 2020, S. 151 f.). Regressions- bzw. Prognosemodelle können – wie eingangs beschrieben – verwendet werden um unterschiedliche Fragestellungen zu beantworten. Welche Studierenden haben erhöhtes Risiko hinsichtlich eines Studienabbruchs? Welche Faktoren beeinflussen diese Risiken? Welche Faktoren bedingen bzw. hemmen einen zügigen Studienfortschritt? Wie identifizieren wir relevante Zielgruppen für Serviceleistungen? Welche Zielgruppen haben spezifischen Unterstützungsbedarf? Aus diesen *exemplarischen Fragestellungen*, die allesamt unserer akademischen Praxis entnommen wurden, wird bereits ersichtlich, dass es die Hochschulen bzw. ihre Funktionsträger*innen und Stakeholder selbst sind, die diese Fragen formulieren (müssen). Jedes Prognosemodell soll von einer zuvor entwickelten Fragestellung geleitet bzw. an diese angepasst werden: Erst die Fragestellung, dann das Modell, das helfen soll, diese zu beantworten. Die zwei im vorliegenden Beitrag skizzierten Anwendungsszenarien folgen diesem Ansatz.

Häufig werden im Zusammenhang mit Prognose und Beschreibung unterschiedliche Regressionsmodelle oder andere Ansätze aus dem Bereich des Machine Learning (wie z. B. Random Forest) verwendet, die dabei helfen, komplexe Sachverhalte als Modell formal beschreibbar und/oder prognostizierbar zu machen. Klassische Ansätze wie OLS-Regressionen oder logistische Regressionen werden dabei öfters angewandt (siehe z. B. Krempkow, 2020a). Im Rahmen der Arbeitsgruppe wurde daher die Frage aufgeworfen, *welchen Mehrwert Machine Learning Ansätze bei der Modellierung von Studienerfolg* bieten. Erfahrungen in beiden Anwendungsszenarien bestätigen die aus der Theorie abgeleitete Annahme, dass *passgenauere Modelle mit hoher Treffsicherheit auch in diesem*

Kontext mit einer eingeschränkten Interpretierbarkeit einzelner Einflussfaktoren einhergehen. Dies soll anhand zweier exemplarischer Fragestellungen illustriert und nachvollziehbar dargestellt werden. Dieses Dokument widmet sich zunächst der inhaltlichen Einbettung der Prognose von Studienerfolg und den definierten Anwendungsszenarien, bevor die verwendeten statistischen Methoden beschrieben werden. Den Kern dieser Arbeit stellen die beiden Szenarien inklusive R-Code und Output und den eigens programmierten R-Funktionen dar, die insbesondere für Praktiker*innen hilfreich sein können. Abschließend weist diese Arbeit insbesondere auf die Grenzen in der Anwendung von Prognosemodellen im Themengebiet Studienerfolg hin und bietet eine auf Erfahrung in den Projekten und theoretischer Überlegungen ausgearbeitete Entscheidungshilfe zur Anwendung der Methodik.

2 Einbettung in die Hochschule: Beschreibung und Prognose

2.1 Prognosemodelle als Teil von Analytics an Hochschulen

In ihrem Rahmenmodell zu „Analytics in Higher Education" unterscheiden Van Barneveld et al. (2012) zwischen Academic Analytics und Learning Analytics. Während sich Academic Analytics grundsätzlich an Entscheidungsträger*innen richten, die mit entsprechend aufbereiteten Daten und Auswertungen zur eigenen Bildungseinrichtung versorgt werden sollen, verwenden Learning Analytics-Tools spezifisch Daten aus Lehr- und Lernsettings mit dem Ziel, Studierenden die Möglichkeit zu geben ihr Lernverhalten anzupassen (vgl. Leitner et al., 2019). Die Abgrenzung ist allerdings nicht immer einfach, weshalb in der Literatur auch oft von Learning & Academic Analytics die Rede ist (vgl. Hochschulforum Digitalisierung, 2015).

Im vorliegenden Beitrag wird das von Norris und Baer (2013) geprägte Konzept, das auf die begriffsklärende Arbeit von Van Barneveld et al. (2012) aufbaut, verwendet. Van Barneveld et al. (2012) versuchen in ihrem Artikel: „Analytics in Higher Education: Establishing a Common Language" den „Wildwuchs an Definitionen" rund um Analytics aufzugreifen, zu strukturieren und zu vereinfachen und etablieren den Begriff „Analytics in Higher Education", dem wir konzeptionell und definitorisch folgen.

In der Literatur gibt es unterschiedliche Abgrenzungen der Anwendungsfelder von Analytics. In der Regel werden Anwendungsfelder in eine hierarchische Stufenordnung nach der Komplexität der verwendeten analytischen Methoden gebracht.

Abbildung 2.1 gibt einen Überblick über unterschiedliche Anwendungsfelder von Analytics in Higher Education und reiht Anwendungsfelder entlang unterschiedlicher Analyseniveaus (siehe auch Bartok et al., 2022, angelehnt an Davenport & Harris, 2007). Davenport und Harris, 2007 unterteilen Anwendungsfelder von Analytics generell in Descriptive, Diagnostic, Predictive und Prescriptive Analytics. Unterschiedliche Autor*innen wie Van Barneveld et al. (2012) oder Norris und Baer (2013) entwickelten darauf aufbauend Modelle, um die Anwendungsfelder von Learning Analytics nach Komplexität und Ausrichtung der Analyse zu strukturieren.

Descriptive Analytics bereiten unterschiedliche Informationen in anschaulicher Weise auf und informieren über die Daten, die die Vergangenheit oder die Gegenwart (Real Time) betreffen. Diese Daten werden nach Van Barneveld et al. (2012) im Datenmanagement sowie in Academic, Learning und Business Analytics in Hochschulen in Reports oder Dashboards aufbereitet und von Uni-

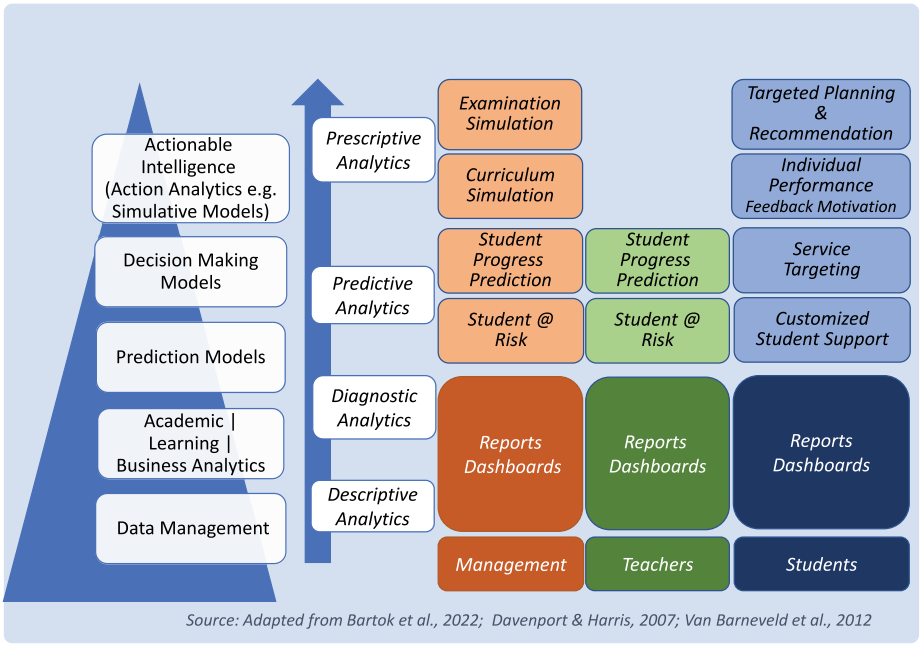


Abb. 2.1 Anwendungsfelder von Learning Analytics-Projekten

versitätsmanagement, Lehrenden sowie Studierenden verwendet. Diagnostic Analytics fokussieren stärker auf Kausalitäten, helfen also dabei, Ursachen und Gründe für identifizierte Sachverhalte bzw. Probleme zu finden. In Academic, Learning oder Business Analytics wird durch unterschiedliche Methoden wie etwa Regressionsmodelle versucht, Wirkungen und Einflussfaktoren zu identifizieren und deren Einfluss wird in Beschreibungsmodellen analysiert. Diese Kausalanalysen (bspw. welche Ursachen für Prüfungsinaktivität verantwortlich sind) werden in der Regel ebenso in Reports und Dashboards dargestellt.

Predictive Analytics helfen – basierend auf vorhandenen Daten – Wahrscheinlichkeiten von zukünftigen Entwicklungen vorherzusagen. Mit Hilfe von Modellen wie beispielsweise Regressionsmodellen wird ermittelt, welche und wie viele Studierende bspw. mit höherer Wahrscheinlichkeit abbruchgefährdet sind oder ein hohes Risiko haben, nicht prüfungsaktiv zu sein (students at risk). Darüber hinaus ist es möglich, individuellen oder kollektiven Studienfortschritt zu prognostizieren (student progress prediction). Prädiktive Modelle können in der Umsetzung dabei helfen, zielgerichtete Maßnahmen für Prüfungsinaktivität oder Studienabbruch zu entwickeln. Beispiele für solche zielgerichteten Maßnahmen umfassen Informationssysteme für Studierende, Mentoringprogramme oder Beratungsangebote für Studierende (siehe Bartok et al., 2023), Studienverlaufsanalysen für interne Ressourcenplanung, Zielgruppenanalyse für Unterstützungsangebote / service targeting und Ähnliches. Prescriptive Analytics

simulieren zu treffende Entscheidungen oder Veränderungen von Variablen anhand eines Modells. Diese Simulationsergebnisse können dabei helfen, Ergebnisse von Entscheidungen oder Veränderungen von Rahmenbedingungen in einem Modell vorab zu testen. Beispielsweise können Prescriptive Analytics in Higher Education dabei helfen, Auswirkungen von Veränderungen des Curriculums im Modell zu simulieren, bevor diese Veränderungen tatsächlich umgesetzt werden (curriculum oder examination simulation). Auf Studierendenebene können Recommendersysteme (recommendation systems) dabei helfen, Empfehlungen für den eigenen Studienverlauf in strukturoffenen Curricula zu erhalten und die eigene Entwicklung im Vergleich zu anderen Studierenden einzuordnen. Für unseren Erfahrungsbericht sprechen wir in weiterer Folge vereinfachend von beschreibenden (descriptive & diagnostic) Modellen und Prognosemodellen (predictive & prescriptive models).

2.2 Gelingensbedingungen von Analytics-Projekten in Hochschulen

Norris und Baer (2013) ergänzen zu Davenport und Harris (2007) vor allem, wie Abbildung 2.2 zeigt, dass speziell unter Analytics-Projekten unterschiedliche

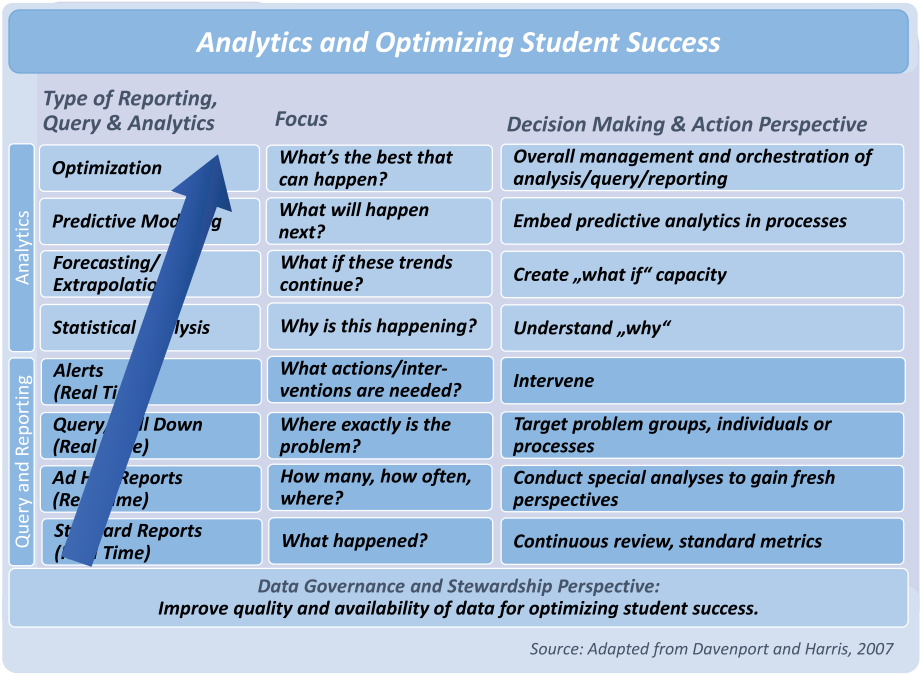
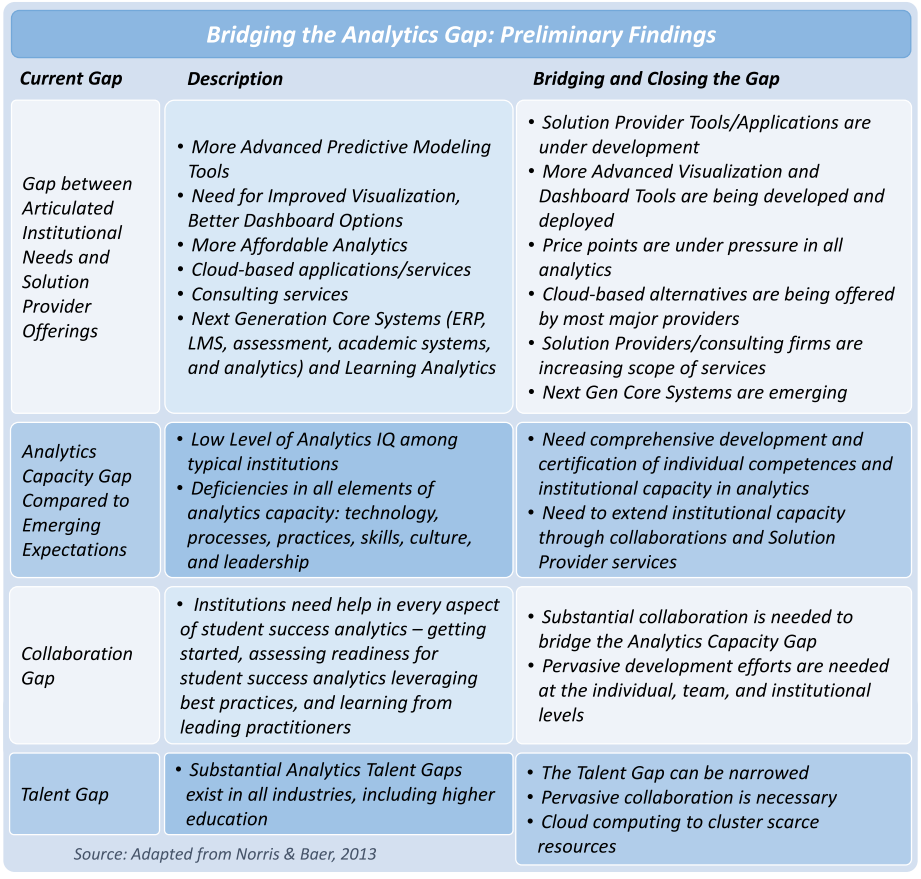


Abb. 2.2 Analytics und Unterstützung von Studienerfolg (Norris & Baer, 2013, S. 8)

Gelingensbedingungen relevant sind. Obwohl die Diskussion, welche Lücken bei der Verwendung von Analytics in Hochschulen existieren, schon lange besteht (bspw. Davenport/Harris-Famework, 2007), ist die Relevanz des Themas ungebrochen hoch. Das zeigt sich nicht nur in Projekten im Zusammenhang mit Predictive Analytics, sondern auch bspw. im Zusammenhang mit Open Data.

Die Data Governance and Stewardship Perspective hebt die Notwendigkeit verfügbarer und gut gewarteter Daten hervor. Dies stellt einen ersten Meilenstein dar, den es als Hochschule zu erreichen gilt, bevor Analytics angewandt werden können. Norris und Baer (2013) gehen in ihren Erläuterungen über die Governance and Stewardship Perspektive hinaus und untersuchen in ihrem Beitrag „Building Organizational Capacity" – basierend auf Expert*inneninterviews – relevante Dimensionen für Analytics, in denen Defizite (Gaps) im Hochschulsystem identifiziert wurden.



Bridging the Analytics Gap: Preliminary Findings

Current Gap	Description	Bridging and Closing the Gap
Gap between Articulated Institutional Needs and Solution Provider Offerings	• More Advanced Predictive Modeling Tools • Need for Improved Visualization, Better Dashboard Options • More Affordable Analytics • Cloud-based applications/services • Consulting services • Next Generation Core Systems (ERP, LMS, assessment, academic systems, and analytics) and Learning Analytics	• Solution Provider Tools/Applications are under development • More Advanced Visualization and Dashboard Tools are being developed and deployed • Price points are under pressure in all analytics • Cloud-based alternatives are being offered by most major providers • Solution Providers/consulting firms are increasing scope of services • Next Gen Core Systems are emerging
Analytics Capacity Gap Compared to Emerging Expectations	• Low Level of Analytics IQ among typical institutions • Deficiencies in all elements of analytics capacity: technology, processes, practices, skills, culture, and leadership	• Need comprehensive development and certification of individual competences and institutional capacity in analytics • Need to extend institutional capacity through collaborations and Solution Provider services
Collaboration Gap	• Institutions need help in every aspect of student success analytics – getting started, assessing readiness for student success analytics leveraging best practices, and learning from leading practitioners	• Substantial collaboration is needed to bridge the Analytics Capacity Gap • Pervasive development efforts are needed at the individual, team, and institutional levels
Talent Gap	• Substantial Analytics Talent Gaps exist in all industries, including higher education	• The Talent Gap can be narrowed • Pervasive collaboration is necessary • Cloud computing to cluster scarce resources

Source: Adapted from Norris & Baer, 2013

Abb. 2.3 Bridging the Analytics Capacity Gap - Vorläufige Erkenntnisse (Norris & Baer, 2013, S. 44)

In Abbildung 2.3 werden vier Dimensionen bzw. Gaps von Norris und Baer (2013) identifiziert.

Der erste Gap beschreibt die nicht vorhandene Produktlandschaft im Zusammenhang mit Analytics-Lösungen auf Hochschulebene. Dieser Gap ist auch im Kontext der europäischen Hochschullandschaft gegeben. Im Rahmen der beiden Projekte „LA – Studierende im Fokus“ sowie „PASSt“ wurde das Open Source Programm (R Core Team, 2022) verwendet und eigene Skripte und Funktionen entwickelt, um die benötigten Analysen durchzuführen und die Ergebnisse grafisch aufzubereiten. Der hier ausgewiesene Programmcode soll daher Hochschulen unterstützen, ähnliche Analysen auf Basis ihrer eigenen Fragestellungen durchzuführen. Wie bereits weiter oben beschrieben ist eine der zentralen Herausforderungen im Zusammenhang mit Analytics im Hochschulkontext die sehr heterogene Datenstruktur, die im Regelfall aus mehr als einem relevanten datenführenden System besteht. Analysen, wie die in diesem Erfahrungsbericht dargelegten, sind somit in jedem Fall an die Datenstruktur der jeweiligen Institution und an die jeweiligen Fragestellungen anzupassen.

Der zweite von Norris und Baer (2013) identifizierte Gap beschreibt das Spannungsfeld zwischen gering ausgeprägter institutioneller „Analytics Capacity“ im Vergleich zu den steigenden Erwartungen für Analytics an Hochschulen. Fachadäquate Aus- und Fortbildungen bzw. Zertifikate im Zusammenhang mit Analytics an Hochschulen sind – wie von Norris und Baer (2013) beschrieben – nicht verfügbar. Kollaborationen zwischen Universitäten innerhalb der einzelnen Projekte, aber auch projektübergreifend im Rahmen des Clusters, halfen auf mehreren Ebenen, Kompetenzen und Erfahrungen im Zusammenhang mit Analytics aufzubauen.

Die als „Collaboration Gap“ bezeichnete Lücke baut auf dem „Analytics Capacity Gap“ auf und identifiziert die Problemlage, dass Kollaborationen innerhalb und außerhalb der Hochschulen nicht zuletzt notwendig sind, um den „Analytics Capacity Gap“ zu schließen. Wie weiter oben beschrieben, halfen beide Projekte, die Kollaboration zwischen den beteiligten Universitäten zu fördern. Um das Thema Analytics im Hochschulbereich und darüber hinaus (bspw. Kooperation mit Unternehmen) voranzutreiben und gewonnene Erkenntnisse auch zu verbreiten, ist das Schließen des Collaboration Gaps hilfreich. Solche Kollaborationen sollten auch in Zukunft gestärkt werden, um gering ausgeprägter „Analytics Capacity“ entgegenzuwirken.

Allgemein identifizierten Norris und Baer (2013) auch einen „Talent Gap“, also einen generellen Mangel an Personen mit ausgeprägten Analytics Skills am Arbeitsmarkt. Dieser Gap ist sicherlich auch relevant im österreichischen Kontext. Wie von Norris und Baer (2013) ausgeführt, ist diese vorhandene Lücke jedenfalls breiter zu sehen und nicht unmittelbar in Zusammenhang mit Analytics-Projekten zu lösen, weil der Gap den gesamten Arbeitsmarkt betrifft und nur mittelfristig gelöst werden kann. Auch in aktuellen Arbeiten im Kon-

text von Higher Education Analytics werden die genannten Gaps aufgegriffen und analysiert inwieweit Gelingensbedingungen an Hochschulen derzeit nicht gegeben sind (Wegner et al., 2023).

In diesem Kontext seien auch internationale Frameworks für die Transformation von Prozessen an Hochschulen erwähnt (Borden & Jin, 2022). Somit ist aus der Literatur, aber auch aus den Erfahrungen aus der Arbeitsgruppe ersichtlich, dass es Grundvoraussetzungen an Hochschulen gibt, um Projekte im Zusammenhang mit Learning- und Academic Analytics – im Folgenden zusammenfassend als Higher Education Analytics – sinnvoll durchführen und in die Prozesse einer Hochschule implementieren zu können.

3 Statistische Methoden: Methodische Vorgehensweisen bei der Beschreibung und Prognose von Studienerfolg

Dieses Kapitel soll einen Überblick über häufig zur Modellierung von Studienerfolg verwendete statistische Methoden geben. Dabei stellt das Kapitel keineswegs den Anspruch auf Vollständigkeit und soll lediglich auf bestehende Quellen hinweisen. Eine tiefergehende Auseinandersetzung mit den jeweiligen Methoden unter Verwendung der angegebenen oder auch weiterer Literatur ist für Praktiker*innen jedenfalls empfohlen.

Vor der Methodenauswahl müssen noch einige weitere Punkte berücksichtigt werden, die hier kurz illustriert werden sollen. Je nach Zielsetzung der Analysen, müssen die zu verwendenden Daten theoriegeleitet ausgewählt werden. Auch die Rahmenbedingungen an der Hochschule (Verfügbarkeit der Daten) spielen hierbei eine Rolle. Soll für die Analysen eine Aufteilung in einen Trainings- und einen Validierungsdatensatz erfolgen, muss sicher gestellt sein, dass die Daten in einer zufälligen Anordnung vorliegen oder nachträglich randomisiert bzw. randomisiert in die Datensätze aufgeteilt werden.

Grundsätzlich sollte theoriegeleitet festgelegt werden, welche Daten und Variablen verwendet werden sollen (sowohl Auswahl der Variablen als auch das Skalenniveau). Für die Auswahl an Variablen und eine Einteilung siehe auch Petri (2021), Krempkow et al. (2021) oder Daniel et al. (2019). Die Variablen und die verwendete Population können theoriegeleitet eingeschränkt werden – in jedem Fall sollte zielgerichtet, auf die jeweilige Fragestellung abgestimmt, eine Auswahl getroffen werden. Je nach verwendeten Variablen kann es besser sein, eine Variable zu normalisieren oder Ausreißer aus dem Datensatz auszuc schließen. In den Voraussetzungen zur Verwendung der jeweiligen Modelle finden sich darüber hinaus weitere Einschränkungen oder Kriterien, die für eine Anwendung erfüllt sein müssen. So kann es Variablen geben, deren Prädiktoren zu hoch miteinander korrelieren, um gemeinsam in das Modell mit aufgenommen werden zu können. Zum Beispiel entwickeln sich das Alter der Studierenden und das Studienalter (gemeint ist hier, im wievielten Semester bzw. Studienjahr Studierende sind – oft auch als Studiendauer bezeichnet – vgl. Krempkow, 2020a) häufig parallel, was zu Multikollinearität im Modell und daher zur Schwierigkeiten bei der Parameterschätzung führen kann (vgl. ausführlicher bspw. Fahrmeir et al., 2007, oder Gelman et al., 2020).

Auf die Verfügbarkeit, die sinnvolle Anwendbarkeit der Modelle (bspw. sehr kleine Studiengänge) und Kontrolle der Qualität der Daten wird hier nicht weiter eingegangen, da dies für jede Universität, für jede Abteilung, die sich damit beschäftigt, jeweils unterschiedliche technische Prüfungen, Aufgaben oder Fragen bedeutet.

Neben Theorie und Methodenwahl sind auch Fragen zur Definition der Grundgesamtheit relevant, die innerhalb des jeweiligen analytischen Projekts bedacht werden müssen. Exemplarische Leitfragen hierzu wären bspw.:

- Wie ist das Vorgehen bei Incomings, also Studierenden, die nur in diesem Semester an der Universität sind?
- Sollen Studierende, die sich kurz vor dem Studienabschluss befinden, denen beispielsweise nur noch 8 ECTS fehlen, in der Population sein?
- Bei welchen Fragestellungen verbessert sich die Qualität der Ergebnisse, wenn Studierendengruppen wie beispielsweise Studierende im ersten Semester inkludiert werden, wo sollten Datensätze ausgeschlossen werden?
- Wie möchte man Studierende miteinbeziehen, die zwei parallele Studien an der selben Universität betreiben, im letzten Studienjahr aber nur in einem davon prüfungsaktiv waren?
- Ist für die Zielsetzung die zeitliche Einteilung in Studienjahre geeignet oder sollte jedes Semester einzeln betrachtet werden?
- Ist für die Fragestellung das abgeschlossene Studienjahr relevant oder sollten auch Daten aus vorhergehenden Studienjahren miteinbezogen werden?

Die hier vorgestellten Modelle werden in der klassischen Inferenzstatistik wie auch im Machine Learning verwendet. Je nach Fragestellung passt das eine oder andere Modell besser. Im vorliegenden Bericht werden einige Modelle, die dem Machine Learning zuzuordnen sind, verwendet und ihre Prognosegüte gegenübergestellt, um das am besten geeignete Modell für die Daten auszuwählen. Modelle mit höherer Prognosegüte liefern treffsicherere Prognosen, mit denen je nach Zielsetzung weitergearbeitet werden kann. Die weitere Arbeit mit den Daten, bspw. ob ausgehend von den Prognosedaten ein hierarchisches Clustering angestrebt wird, um herauszufinden, welche Personengruppen studieninaktiv sind, oder ob bspw. abgeleitet werden soll, welche Kurse oder Programme unterstützt oder begleitet werden sollen, ist nicht Teil dieser Arbeit. Im Folgenden werden die Modelle und Methoden, die in den Anwendungsfällen eingesetzt wurden, vorgestellt. Ebenfalls nicht beschrieben sind Methoden aus der deskriptiven, beschreibenden Statistik, die jedenfalls zusätzlich eingesetzt werden sollten.

3.1 Lineare Regression

Die lineare Regression dient zur Vorhersage einer intervallskalierten abhängigen Variablen durch eine oder mehrere unabhängige, erklärende Variablen. Dabei wird ein linearer Zusammenhang angenommen. Die Parameter des Modells werden so geschätzt, dass die Fehlerquadratsumme der Residuen minimal wird. Das Minimierungsproblem wird bei diesem klassischen Ansatz über die Methode der kleinsten Quadrate (Ordinary Least Squares [OLS] Methode) gelöst (siehe bspw. Gelman et al., 2020).

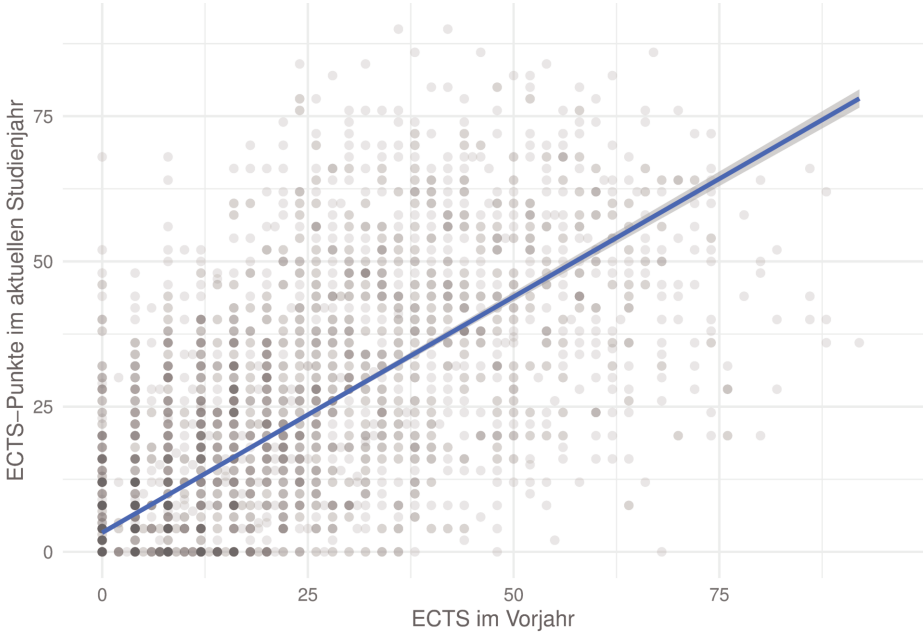


Abb. 3.1 Beispiel lineare Regression

In der Abbildung 3.1 wird ein linearer Zusammenhang zwischen der Anzahl der ECTS-Punkte, die Studierende im Vorjahr erreicht haben und der Anzahl der ECTS-Punkte, die im aktuellen Studienjahr erreicht wurden, angenommen und dargestellt. Die Farbintensität der Punkte erhöht sich mit der Anzahl der Personen, die in der selben Position sind. Für die Durchführung einer OLS-Regression müssen einige Annahmen als erfüllt gelten: Zur Verwendung des linearen Regressionsmodells müssen die Residuen unkorreliert sein. Sie haben einen Erwartungswert von null, konstante Varianz (Annahme der Varianzhomogenität), und sind normalverteilt (bspw. in Rasch et al., 2021, S. 108 ff.) In dieser Darstellung sieht man die Regressionsgerade auf der Punktwolke. Das Bestimmtheitsmaß R^2 gibt den Anteil der durch die Prädiktoren erklärten Va-

rianz an und liegt hier unter Verweundung nur eines Prädiktors bei ungefähr 57 %.

Die Methode kann sowohl beschreibenden als auch Prognosemodellen zugeordnet werden. In den Berechnungen werden neben den ECTS-Punkten aus dem aktuellen bzw. vorigen Studienjahr weitere erklärende Variablen verwendet. Die lineare Regressionsanalyse ist üblicherweise einfach verständlich und erfreut sich nicht zuletzt auch aufgrund der unmittelbaren Interpretierbarkeit der standardisierten oder unstandardisierten Regressionskoeffizienten großer Beliebtheit. Die Annahme der Linearität ist natürlich restriktiv: Voraussetzungsverletzungen und Ausreißer in den Daten (siehe methodische Voraussetzungen) können großen Einfluss auf die Ergebnisse haben.

Für ausführlichere Informationen zur Methode sei auf Standardlehrbücher der Statistik, mit Aspekt auf die Wahl der Software beispielhaft auf Hatzinger et al. (2011), Field et al. (2012), Lantz (2019) oder Gelman et al. (2020) verwiesen. Die Anwendung einer typischen, linearen Regressionsanalyse im Hochschulkontext ist im Anwendungsbeispiel I illustriert.

3.2 (Boosted) Logistische Regressionen

Die logistische Regression kann durch die Verallgemeinerung der linearen Regression beschrieben werden. Die logistische Regression dient zur Vorhersage einer nominalskalierten (dichotomen) abhängigen Variable durch mehrere unabhängige Variablen. Die logistische Regression zählt in der Statistik zu den generalisierten (verallgemeinerten), linearen Modellen. Hier wird das allgemeine lineare Modelle so erweitert, dass nun die abhängige Variable über eine (nichtlineare) Verknüpfungs- (oder Link) Funktion in linearem Zusammenhang zu den Prädiktoren steht. Auch die Residuen müssen hier nicht einer Normalverteilung folgen, sondern können einer Verteilung aus der Familie der Exponentialfamilie folgen. Die vorherzusagende Variable in diesen Modellen ist die Prüfungsaktivität (siehe Definition nach §12 Absatz 4 Satz 1a Universitätsgesetz, 2002), die bei unseren Daten die Ausprägungen 0 oder 1 hat. Die Schätzung der Regressionsparameter findet üblicherweise über sogenannte Maximum Likelihood-Schätzungen unter Zuhilfenahme eines iterativen Algorithmus statt.

Bei der Verwendung dieser Methode innerhalb des Machine Learnings wird eine Variante mit Boosting-Ansatz verwendet: Innerhalb des Trainingsdatensatzes werden die Daten mit dem Label prüfungsaktiv bzw. prüfungsinaktiv versehen. Diese Klassifikation passiert im Lernalgorithmus, der ein Modell generiert, welches die Kennzeichnung für weitere Punkte (aus dem Evaluierungsdatenset) vornimmt. Im Evaluierungsdatenset werden falsch klassifizierte Datenpunkte identifiziert und unterschiedlich stark gewichtet (boosting), um eine gute Anpassung für neue Daten zu erzielen (Schapire, 1990; Burkov, 2019). Für

das Boosting wird in dieser Arbeit der LogitBoost-Algorithmus von Friedman et al. (2000) verwendet.

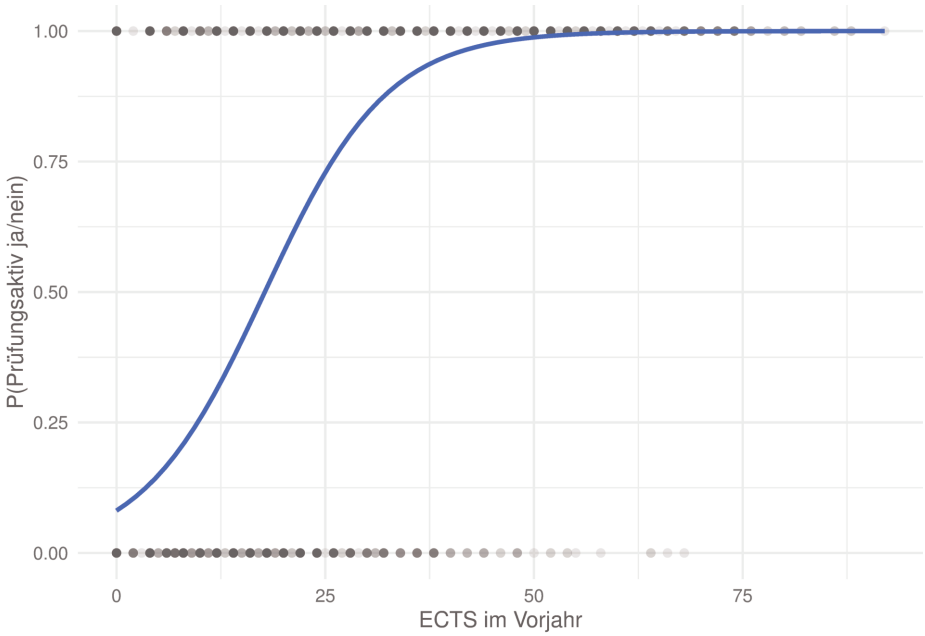


Abb. 3.2 (Boosted) logistische Regression

Es gibt einen starken Zusammenhang zwischen den ECTS-Punkten, welche im Vorjahr erbracht wurden, und der Prüfungsaktivität. In der Abbildung erkennt man, dass Personen, die im Vorjahr wenige ECTS-Punkte hatten, eher prüfungsinaktiv sind, während Studierende mit mehreren ECTS-Punkten eher prüfungsaktiv sind. In Modellen mit mehreren Prädiktoren eignen sich zur Quantifizierung des Effekts der Einflussgrößen als Visualisierung und zur Interpretation besonders die Odds Ratio (Chancenverhältnisse).

Ausführliche Erklärungen zur logistischen Regression finden sich in Lehrbüchern der Statistik, beispielsweise Field et al. (2012), Hatzinger et al. (2011) oder Menard (2002), die Anwendung der Methode im Bereich Machine Learning, Boosting Ansätze und Alternativen sind zum Beispiel in Lantz (2019) und Schapire (1990) anschaulich aufbereitet.

3.3 Generalized additive models (GAM)

Generalisierte additive Modelle stellen eine Verallgemeinerung der klassischen linearen Regressionsmodelle dar und identifizieren automatisch geeignete Transformationen der Prädiktoren. Diese Modelle können nicht-lineare

und nicht-monotone Beziehungen zwischen abhängiger und unabhängigen Variablen berücksichtigen. Sie erweitern das lineare Standardmodell, indem sie nichtlineare Funktionen in jeder der Variablen zulassen. Diese verallgemeinerten Modelle sind als GAM bekannt (Hastie & Tibshirani, 1990). Um einen geeigneten Weg zwischen Überanpassung (Overfitting) durch zu beliebig glatte Funktionen und zu wenig bzw. zu schlechter Anpassung zu finden, wird für jede glatte Funktion ein Strafterm eingefügt, der mit einem Glättungsparamenter multipliziert wird. Die Schätzung der Glättungsparameter kann in R an die Daten angepasst werden. Neben dem Vorteil der besseren Anpassung ist in diesem Modell auch die Voraussetzung der Normalverteilung der Residuen gelockert.

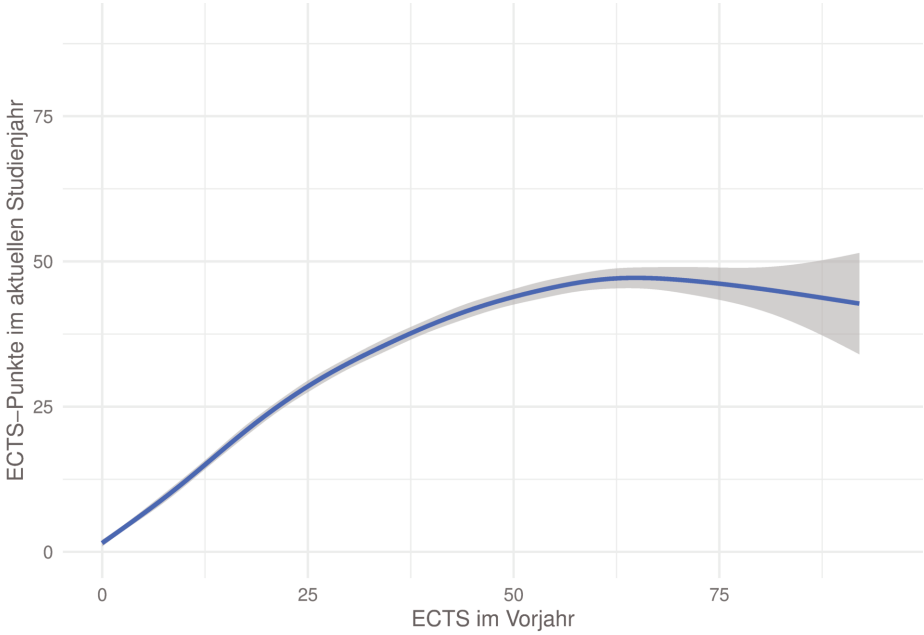


Abb. 3.3 Generalisiertes, additives Modell

Die Methode wird sowohl in Beschreibungs- als auch in Prognosemodellen mit und ohne Machine Learning Einsatz angewandt. Wie in Abbildung 3.3 ersichtlich wird durch das Smoothing eine bessere (in diesem Fall nicht-lineare) Passung an die Daten erreicht. Der in der Abbildung grau eingefärbte Bereich markiert dabei das (bei abnehmender Datengrundlage) nach oben hin breiter werdende Konfidenzintervall.

Um generalisierte additive Modelle genauer in den Blick zu nehmen, seien hier exemplarisch Guisan et al. (2002), Wood (2006), Marx und Eilers (1998) und Ruppert et al. (2003) erwähnt.

3.4 Random Forest

Die Methode Random Forest (Breiman, 2001) kombiniert eine Vielzahl von Entscheidungsbäumen durch eine Art Mehrheitsentscheidung. Aus dem Datensatz wird zunächst ein kleiner und zufälliger Teil der Daten verwendet um einen Entscheidungsbaum zu erstellen. Die Methode wählt also aus dem theoriegeleitet ausgewählten Variablenset zunächst zufällig einzelne Variablen für erste Berechnungen aus. Dieser Schritt wird mehrmals (im Beispiel dieses Berichts 100 mal) wiederholt. Beim Erstellen jedes Baums wird ein Prädiktor anhand des mitberechneten Reinheits- bzw. Unreinheitsmaßes (in der Abbildung als Prädiktionsfehler bezeichnet) gewählt, welcher die Erstellung des nächsten Baumes modifiziert. Die so erzeugten Bäume stellen in Summe gemeinsam einen Random Forest dar. Entscheidungsbäume (Decision Trees) stellen Entscheidungsregeln dar, die mit Hilfe eines nichtparametrischen Algorithmus erstellt werden. Sie eignen sich sowohl für Klassifizierungsaufgaben (bei einer dichotomen abhängigen Variablen) als auch für Regressionsaufgaben (metrische abhängige Variablen). Die sogenannten Decision Trees bestehen aus einem Wurzelknoten, auf den weitere Knoten mit jeweils mindestens zwei Blättern folgen. Die Knoten stellen jeweils eine Entscheidungsregel (bspw. Aktivität auf der Lehr- und Lernplattform) und die Blätter die Ausprägungen (ja oder nein) dar. Dabei wird (basierend auf dem Gini-Index) jeder Knoten so gewählt, dass sich möglichst deutliche Unterscheidungen der Studierendengruppen zwischen den Blättern ergeben (Knotenreinheit).

Die Darstellung eines Random Forest für die Daten dieses Berichts wäre aus Platzgründen nicht sinnvoll, daher wird exemplarisch eine grafische Darstellung eines einzelnen Entscheidungsbaums mit einer kleinen Auswahl von Variablen verwendet. Die Illustration dient dem besseren Verständnis, wie die Methode Random Forest arbeitet.

Am Wurzelknoten sieht man, ob eine Person prüfungsaktiv ist (dunkler Teil der Box, bezeichnet mit 1) oder ob sie prüfungsinaktiv ist (heller Teil der Box, 0). Einen wesentlichen Einfluss (signifikant) hat nach dieser Grafik die Information, ob Personen mit der Lernplattform gearbeitet haben. Von den Personen, die nicht auf der Lernplattform aktiv waren (kleinergleich 0, linker Ast), sind sehr wenige prüfungsaktiv. Für Studierende, welche eine Lernplattformaktivität größer 0 haben, ist die nächste wichtige Unterscheidung die Schulform, die vor dem Studium besucht wurde. Wurde eine allgemeine oder berufsbildende höhere Schule besucht, ist relevant, ob bereits mindestens ein Auslandssemester absolviert wurde. Bei Studierenden mit Berufsreifeprüfung oder „andere“, erkennt man, dass etwas weniger als die Hälfte dieser Personen prüfungsaktiv ist.

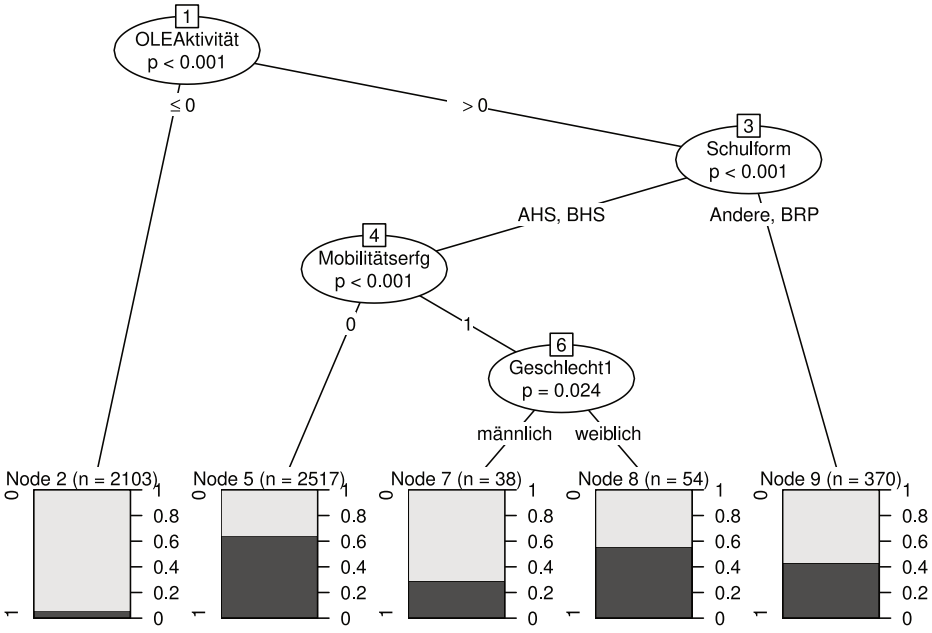


Abb. 3.4 Random Forest Plot – Beispiel eines Decision Trees

Die Methode Random Forest (Breiman, 2001) kombiniert eine Vielzahl solcher, aufeinander aufbauend „besser“ (also mit höherer Reinheit) werdender Entscheidungsbäume durch eine Art Mehrheitsentscheidung. Durch die Kombination und das Training, welches parallel über einen Algorithmus bei der Erstellung geschieht, soll Zufallsergebnissen entgegengewirkt werden. Wichtig ist, dass die Bäume untereinander eine sehr niedrige Korrelation aufweisen. Dadurch sollen Schwächen der einzelnen Entscheidungsbäume durch unterschiedliche Zufallsstichproben und Zufallsteilmengen an Prädiktoren bei jeder Verzweigung ausgeglichen werden. Gleichzeitig wird auch das Risiko des Overfittings (Überanpassung an die Trainingsdaten) minimiert.

Die Darstellung 3.5 aus Nawar und Mouazen (2017) veranschaulicht das Vorgehen, innerhalb des Machine Learnings Entscheidungsbäume aus den Daten zu generieren, diese zu selektieren und eine geeignete Auswahl als Random Forest zu sammeln. Random Forest kann sowohl für Klassifikation (binäre abhängige Variable) als auch für Regression (metrische abhängige Variable) verwendet werden. Geeignet sind Random Forests, wenn für alle Klassen Daten im Trainingsdatensatz vorhanden sind. Die Methode ist für die Prognose von neuen Klassen oder Werten ungeeignet.

Um in die Welt der Decision Trees und Random Forests einzutauchen, empfiehlt sich Literatur aus dem Machine Learning Bereich. Erwähnt seien hier Lantz (2019), Liu et al. (2012) und Schell (2022).

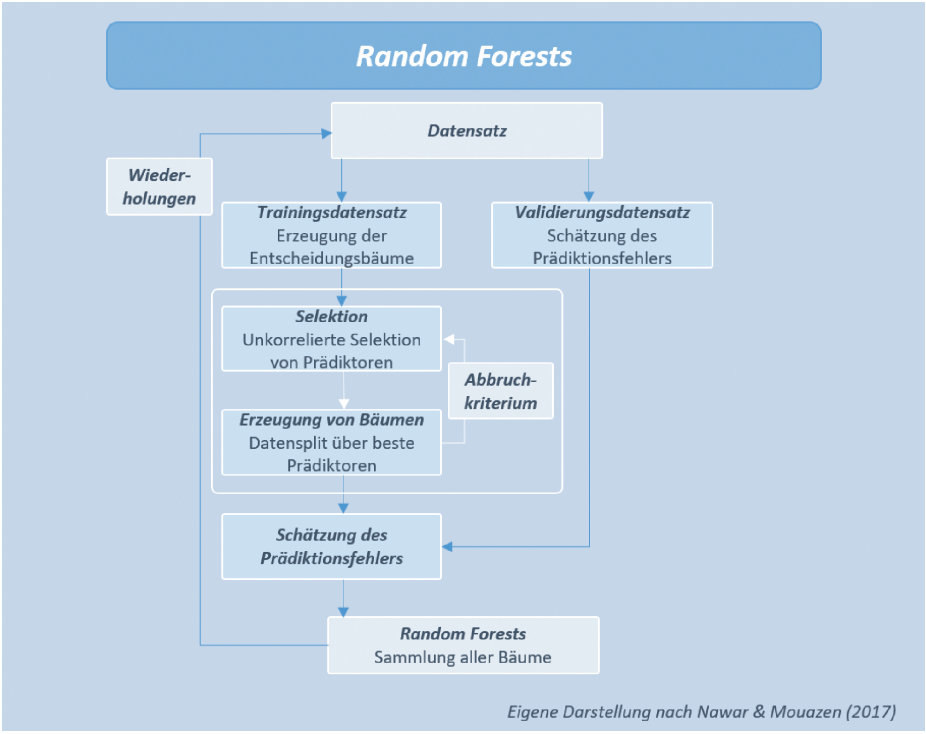


Abb. 3.5 Random Forests

3.5 Gradient-Boosting-Machine-Modelle (GBM)

GBM-Modelle basieren ebenfalls auf Entscheidungsbäumen, welche jedoch sequentiell unter Verwendung von Informationen aus den vorherigen Bäumen aufeinander aufgebaut werden. Während bei Random Forests viele unterschiedliche Bäume in die finale Prognose herangezogen werden, die gegenseitig ihre Schwächen ausgleichen, wird im Gradient Boosting Machine Modell jeder neue Baum so gebaut, dass der Prognosefehler aus dem vorherigen Baum im neuen Baum geringer ist. In Abbildung 3.6 wird dieser sich wiederholende Kreislauf dargestellt. Durch die Iterationen wird schrittweise ein besseres und schlussendlich ein passendes Modell generiert.

Hintergründe und ausführliche Information über Gradient Boosting Machine Modelle findet man in Machine Learning Literatur, beispielhaft Kuhn und Johnson (2013), Hastie et al. (2009) und Schell (2022).

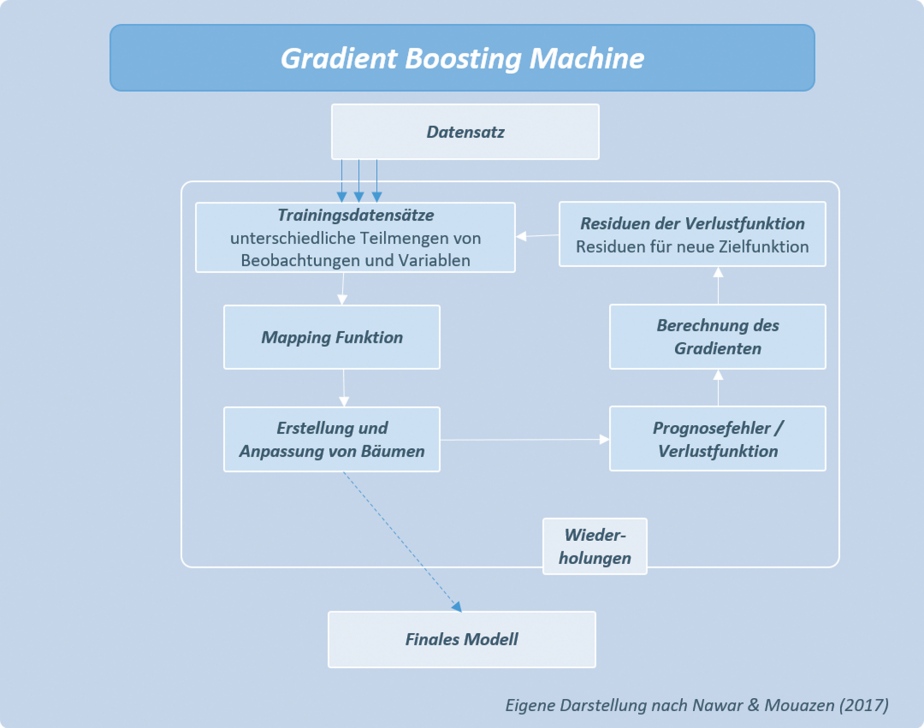


Abb. 3.6 Gradient Boosting Machine-Modelle

3.6 Support Vector Machine (SVM)

Die Support Vector Machine versucht, Objekte mithilfe von Trennungslinien oder -ebenen zu teilen und die Daten so zu separieren. Diese Ebenen werden so gewählt, dass zwischen den verschiedenen Klassen ein möglichst großer freier Bereich ist. Zur Optimierung wird der Abstand zwischen Support Vektoren (Margins) maximiert. Im 2-dimensionalen Bereich wird eine Trennlinie gezogen, im 3-dimensionalen Bereich eine Trennfläche eingezogen und bei 4 oder mehr Dimensionen eine sogenannte Hyperebene. Durch die Anwendung des Kernel-Tricks lässt sich die Methode auch bei nicht-linearen Entscheidungsgrenzen einsetzen: Hierfür werden die Trennungsvektoren in eine höhere Dimension transformiert.

Abbildung 3.7 zeigt eine Trennfläche eingefügt in eine Punktwolke aus den Studierendendaten Alter, ECTS-Punkte im Vorjahr und der Verwendung der Lernplattform. Hierbei wird noch kein Ergebnis interpretiert, es soll hier veranschaulicht werden, wie die Methode die Trennung der Daten vornimmt, um über viele solcher Schritte an ein an die Daten angepasstes Modell zu gelangen, mit dessen Hilfe die Prognosen getroffen werden können.

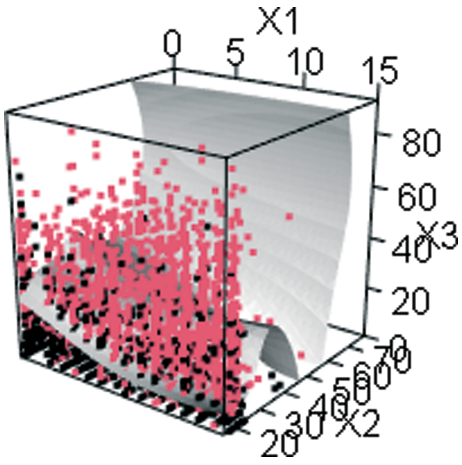


Abb. 3.7 Darstellung einer Trennfläche SVM

Die Methode wird in allgemeiner Machine Learning Literatur eingehend betrachtet, neben Lantz (2019) und Schell (2022) seien hier mit Blick auf die Anwendung Karatzoglou et al. (2004), Karatzoglou et al. (2006) und Boehmke und Greenwell (2019) erwähnt.

In diesem Kapitel wurden die verwendeten Methoden kurz dargestellt und jeweils auf weiterführende Literatur verwiesen. Die erläuterten Methoden stellen zwar einen breiten Überblick über in der Literatur bzw. in ähnlichen Projekten verwendeten *Status quo* dar, jedoch kann selbstverständlich kein Anspruch auf Vollständigkeit erhoben werden. Algorithmusbasierte Methoden, die eher dem Machine Learning zugeschrieben werden können, gehen häufig mit einer höheren Prognosegüte einher, wohingegen modellbasierte Methoden (wie die OLS-Regression) restriktiver sind und sich dafür aber einer besseren Interpretierbarkeit erfreuen. Beides ist häufig gleichzeitig nicht erreichbar (Kuhn & Johnson, 2013). Im Kontext von Prognosemodellen kann auch, sofern genügend Beobachtungen vorliegen, auf Modelle aus dem Bereich der Zeitreihenanalyse zurückgegriffen werden (Loder, 2023). Ebenfalls können die dargestellten Methoden mit Simulationsbasierten Ansätzen wie Agent-based Modelling kombiniert werden. Aufgrund der verschiedenen möglichen Zielsetzungen im Hochschulkontext sollte zunächst abgewogen werden, ob die primäre Zielsetzung eher Prognose von Studienerfolg zum Ziel hat oder eher insbesondere (lineare) Wirkungszusammenhänge untersucht und Einflussfaktoren quantifiziert werden sollen.

Im Folgenden werden die beiden Herangehensweisen anhand von Anwendungsszenarien näher erläutert und beschriebene Modelle anhand dieser Szenarien angewandt.

4 Methodik

4.1 Anwendungsszenarien von Prognosemodellen anhand zweier exemplarischer Fragestellungen

Dieses Kapitel soll nun praxisnah und exemplarisch die im vorherigen Kapitel beschriebenen, statistischen Verfahren anhand zweier Anwendungszenarien vorstellen. Dabei werden Vor- und Nachteile bei der Anwendung der verschiedenen Methoden (auch im Zusammenhang mit unterschiedlichen Zielgruppen) diskutiert, praktische R-Funktionen vorgestellt und Ergebnisse aus den Anwendungen erläutert.

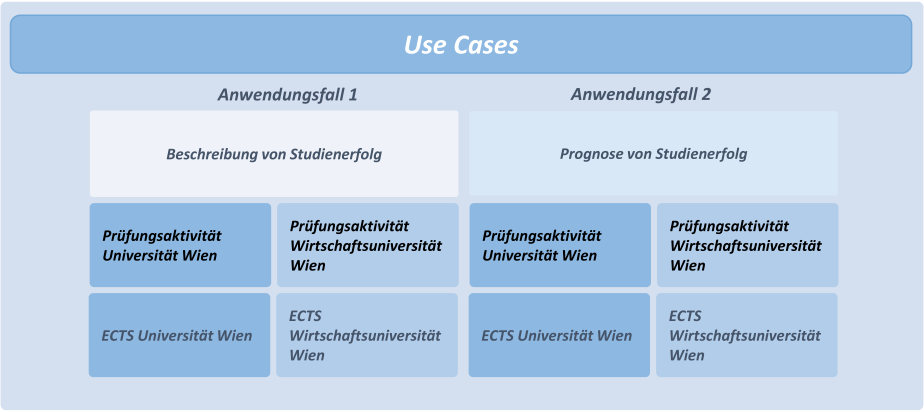


Abb. 4.1 Anwendungszenarien

Die beiden Anwendungsszenarien ergeben sich einerseits aus der Methodik, andererseits aus dem Pool an potenziellen Fragestellungen, die mit Hilfe von Higher Education Analytics beantwortet werden sollen. Eine beispielhafte Fragestellung aus dem ersten Anwendungsszenario „Beschreibung von Studienerfolg“ wäre beispielsweise die Fragestellung „Welche Faktoren beeinflussen die Prüfungsaktivität im Studienprogramm XY?“. Eine beispielhafte Fragestellung zur Prognose von Studienerfolg wäre „Wie viele prüfungsaktive Studierende können im Studienprogramm XY im nächsten Semester erwartet werden?“. Anhand dieser beider Fragestellungen soll die Anwendung verschiedener Verfahren sowie deren Vor- und Nachteile und Grenzen demonstriert werden.

Die Illustration der beiden Anwendungsfälle (vgl. Abbildung 4.1) werden mit der statistischen Programmiersprache R (R Core Team, 2022) durchgeführt.

Als Teile eines zunächst internen R-Pakets zur Aufbereitung von Studierendendaten und zur Visualisierung von Regressionsergebnissen wurden Funktionen geschrieben, deren Anwendung im Ergebnisteil beschrieben werden und die dem Anhang entnommen werden können. Das R-Paket wird laufend durch – unserer Erfahrung nach – nützliche Funktionen ergänzt und wird nach Projektende auf Github zur Verfügung gestellt: https://github.com/larissabartok/LAViennaLAVienna. Neben nützlichen Funktionen zur Visualisierung von Modellergebnissen, werden im R-Paket auch deskriptive Visualisierungen zum Thema Studienerfolg und Diversität vorgeschlagen.

4.2 Datengrundlage und Variablen

Als Datengrundlage dienen Daten aus verschiedenen Bachelorstudienprogrammen der Wirtschaftsuniversität Wien (WU) und der Universität Wien. Die zur Verfügung stehenden Variablen und Datengrundlagen unterscheiden sich zwischen den beiden Universitäten und wurden soweit wie möglich aneinander angepasst. Aufgrund der unterschiedlichen Datengrundlagen und Variablen können die Ergebnisse zwischen den Universitäten nicht verglichen werden. An beiden Universitäten werden exemplarisch Daten aus dem Studienjahr 2019/20 verwendet. Hierbei handelt es sich um reine Beispieldaten. Das Ziel ist hier nicht die Ergebnisse aus den Modellen zu diskutieren, noch erheben die Modelle mit ihren inkludierten Variablen einen Anspruch auf Vollständigkeit.

Definition von Studienerfolg und unabhängige Variablen

So unterschiedlich wie die Fragestellungen sind auch die (fachspezifischen) Definitionen von Studienerfolg (siehe z. B. Bartok et al., 2021), wobei Hochschulen im deutschsprachigen Raum neben Studienzufriedenheit, Persönlichkeitsentwicklung und Berufsfähigkeit vor allem den Studienverlauf und hier in erster Linie den Studienabschluss als relevantes Kriterium betrachten (Meyer-Guckel & Jorzik, 2015, S. 17). Das ist nicht weiter verwunderlich, da der (zeitnahe) Studienabschluss bzw. eine hohe Prüfungsaktivität für öffentlich finanzierten Hochschulen unmittelbar budgetrelevant und somit bei Leistungsvereinbarungen zwischen Hochschulen und Ministerium von zentraler Bedeutung sind (§ 12 Abs. 4 Z 1 lit. a UG 2002). Ähnlich gilt dies auch für etliche Bundesländer in Deutschland – vgl. bspw. Krempkow (2007). In diesem Beitrag werden zwei Definitionen von Studienerfolg verwendet:

- Erreichte ECTS im Studienjahr
- Prüfungsaktivität (mindestens 16 ECTS-Punkte oder positiv beurteilte Studienleistungen im Umfang von acht Semesterstunden pro Studienjahr)

Aus Demonstrationsgründen wurden aus der Literatur einzelne, relevante unabhängige Variablen ausgewählt. In der Praxis ist die Liste der Prädiktoren vermutlich länger und sollte jedenfalls theoriegeleitet und in weiterer Folge auch modellbasiert erfolgen. Variablen können modellbasiert ausgeschlossen werden, wenn sie Modelvoraussetzungen verletzen (siehe Multikollinearität) oder im Anwendungsszenario II, wenn sie nicht zur Vorhersagegüte beitragen.

Abbildung 4.2 enthält eine Übersicht der in diesen Beispielen verwendeten unabhängigen und abhängigen Variablen inklusive der Definitionen, die verwendet wurden und ihre Ausprägungen. Übersichtstabellen wie diese haben sich in der Kommunikationen zwischen den Projektpartner*innen und auch in der Kommunikation mit Stakeholder*innen als hilfreich erwiesen.

Datengrundlage Beschreibung und Prognose von Studienerfolg			
	Variable	Definition	Ausprägung
AV	**Studienerfolg**	**Prüfungsaktivität:** Wurden im Studienjahr zumindest 16 ECTS erworben?	dichotom
		Erbrachte ECTS der positiven Prüfungsleistungen im Studienjahr	metrisch
Unabhängige Variablen: Demografische Merkmale	Alter	Alter in Jahren; berechnet anhand des Geburtsdatums	metrisch
	Geschlecht	Geschlecht in 3 Kategorien; von den Studierenden angegeben	kategorial
	Schulform	Schulform, innerhalb der die Studienzulassung erworben wurde	kategorial
	Staatsbürger-schaft	Staatsbürgerschaft in 2 Kategorien; von den Studierenden angegeben	dichotom
Unabhängige Variablen: Veränderbare Merkmale	Erbrachte ECTS im Vorjahr	Erbrachte ECTS der positiven Prüfungsleistungen im Vorjahr	metrisch
	Prüfungsaktives Zweitstudium	Zulassung zu einem weiteren Studium in beiden Semestern	dichotom
	Mobilitäts-erfahrung	Wurde bisher im Rahmen des Studiums mind. ein Auslandssemester absolviert?	dichotom
	OLE-Aktivität	Aktive Verwendung der Lernplattform (nur an WU verfügbar)	dichotom

Abb. 4.2 Variablenübersicht

Das Alter zum Studienbeginn wurde anhand des Geburtsdatums berechnet und in Jahren in die Modelle mitaufgenommen. Das prüfungsaktive Zweitstudium ist so definiert, dass es an der jeweiligen Universität zumindest ein weiteres Studium gibt, zu dem die/der Studierende im Beobachtungszeitraum inskribiert ist und in diesem prüfungsaktiv ist. Bezüglich der ECTS-Punkte im Vorjahr wurden nur positive Prüfungsleistungen inkl. angerechnete Prüfungsleistungen berücksichtigt. Die OLE-Aktivität (Online Learning Environment) wurde dichotom operationalisiert und wird aufgrund der Datenverfügbarkeit nur in den Modellen der WU berücksichtigt. Bzgl. der Staatsbürgerschaft haben wir uns hier in diesem Beispiel ebenfalls für eine Dichotomisierung entschieden (inländische (Österreich) vs. ausländische Staatsbürgerschaft). Die Variable Mobilitätserfahrung wurde ebenfalls als dichotome Variable berücksichtigt (wurde im Rahmen des Studiums mindestens ein Auslandssemester absolviert? Ja vs. nein). Zusätzlich zu aus der Literatur bekannten, relevanten Einflussgrößen können und sollten auch strukturelle Merkmale des Studiums mitmodelliert werden (z. B. curriculare Veränderungen, verschiedene Studienzweige etc.), wobei insbesondere facheinschlägige Spezialist*innen wie Studiendekan*innen (Studienprogrammleitungen) hier wertvolle Beiträge liefern können.

In den folgenden Kapiteln werden nun die beispielhaften Analysen einmal anhand der metrischen und einmal anhand der dichotomen Definition von Studienerfolg gezeigt. Dabei wurden für Anwendungszenario I und II die Modelle wie in Abbildung 4.3 definiert. Dargestellt sind die verwendeten unabhängigen Variablen (UV) auf der linken und die beiden Defitionen der abhängigen Variablen (AV) auf der rechten Seite.

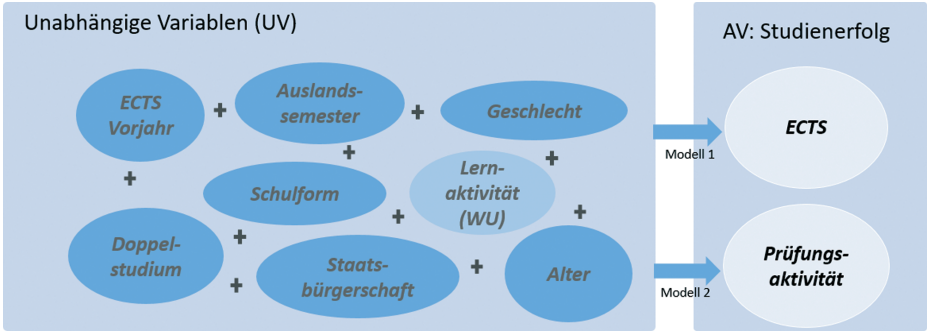


Abb. 4.3 Modelle

5 Durchführung der Analysen anhand zweier Anwendungsszenarien

In Anwendungsszenario I soll zunächst exemplarisch die Frage beantwortet werden, welche Faktoren in einem bestimmten Studienprogramm Einfluss auf Studienerfolg haben. In Anwendungsszenario II wird gezeigt, wie Studienerfolg (sowohl metrisch als ECTS pro Jahr, als auch dichotom als Prüfungsaktivität) prognostiziert werden kann.

Beide Problemstellungen können durch die Verwendung von Regressionsmodellen gelöst werden. Anwendungsszenario I beschreibt dabei den Zusammenhang zwischen Variablen bzw. verfolgt einen modellbasierten Ansatz (Descriptive & Diagnostic Analytics/Beschreibungsmodell) während Anwendungsszenario II die Prognose von zukünftigen Größen im Fokus hat und für diesen Zweck (vor der Beschreibung von inhaltlichen Zusammenhängen) das optimale Verfahren (predictive & prescriptive models / Prognosemodell) ausgewählt werden muss.

5.1 Anwendungsszenario I: Beschreibung von Studienerfolg

Zunächst soll die Frage beantwortet werden, welche Faktoren an der Universität Wien in einem bestimmten Studienprogramm einen Einfluss auf Studienerfolg haben. Zuerst wird in diesem Kapitel beispielhaft die Analyse und Interpretation einer OLS-Regression dargestellt, indem als Zielvariable die ECTS im Studienjahr definiert worden ist. Anschließend wird die Durchführung und Interpretation einer logistischen Regression vorgestellt.

5.1.1 OLS-Regression (abhängige Variable: Anzahl ECTS)

Bevor die Analyse durchgeführt wird, müssen die Daten entsprechend aufbereitet werden. An der Universität Wien werden die Daten aus dem Data Warehouse exportiert und mit Hilfe einer eigens dafür entwickelten R-Funktion für jedes Studienprogramm separat aufbereitet. Für diesen Erfahrungsbericht wurden exemplarisch Daten aus einer Studienrichtung verwendet. Einzelne Schritte der Datenaufbereitung werden nicht dargestellt, da sie sich für jede Universität individuell gestalten. Anschließend wird die übliche Vorgehensweise zur Überprüfung der Voraussetzungen zur Durchführung der OLS-Regression illustriert.

Insgesamt wurden in dieser Analyse die Daten von 882 Studierenden verwendet. 527 Personen gaben an, weiblich zu sein, 355 männlich und 0 divers.

5.1.1.1 Methodische Voraussetzungsüberprüfung

Wie bereits im theoretischen Kapitel erwähnt muss zunächst überprüft werden, ob die Voraussetzungen zur Anwendung dieses statistischen Verfahrens vorliegen. Die sogenannten Residuen müssen zum Beispiel annähernd einer Normalverteilung folgen und Ausreißer müssen identifiziert und ggf. entfernt werden. Um die Voraussetzungen zu überprüfen, muss zunächst ein Modell gefitted werden. Für eine OLS-Regression wird die Funktion *lm()* verwendet, auch die Verwendung der verallgemeinerten Funktion *glm()* ist möglich.

```
lm1 <- lm(ECTS_2019_20 ~ ALTER_STUDIENBEGINN + GESCHLECHT + MOBILITAET +
        DOPPELSTUDIUM + ECTS_2018_19 + SCHULFORM1 +
        STAATSANGEHOERIGKEIT_CODE_AT,  data = daten_bw)
```

Zunächst bietet sich die Verwendung der Funktion *plot()* an, der eine umfangreichere Outlier-Diagnostik folgen sollte. Gerade dem zweiten Punkt sollte bei der Analyse von ECTS besonders große Beachtung geschenkt werden, weil hier durch Anrechnungen etc. schnell hohe Werte auftreten können, die unter Umständen zur Schwierigkeiten bei der Modellierung führen können. Wird die Funktion ohne Spezifikation des Funktionsarguments *which* ausgeführt, werden die wichtigsten Plots nacheinander dargestellt. Abbildung 5.1 kann verwendet werden, um die Voraussetzung der Linearität zu überprüfen. Je linearer die Beziehung zwischen den Residuen und den Fitted Values, desto eher ist die Voraussetzung erfüllt. Im Beispiel kann die Voraussetzung als gegeben angenommen werden.

```
plot(lm1, which = 1, sub.caption = "... of the model defined above")
```

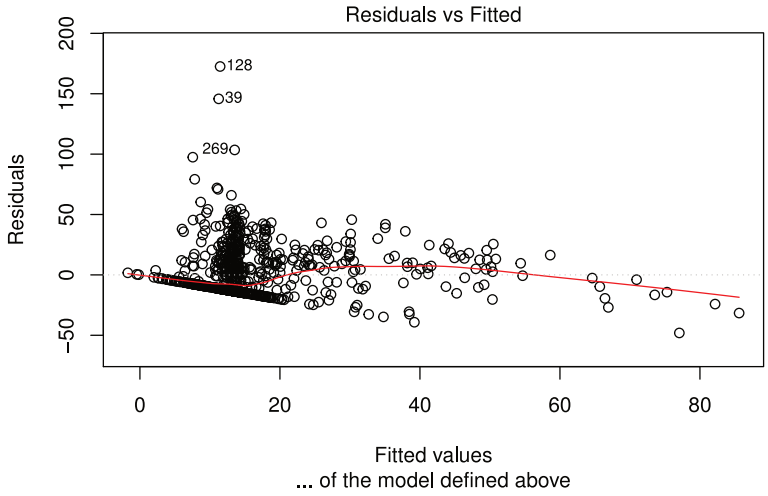


Abb. 5.1 Plot Residuen und Fitted Values

```
plot(lm1, which = 1, sub.caption = "... of the model defined above")
```

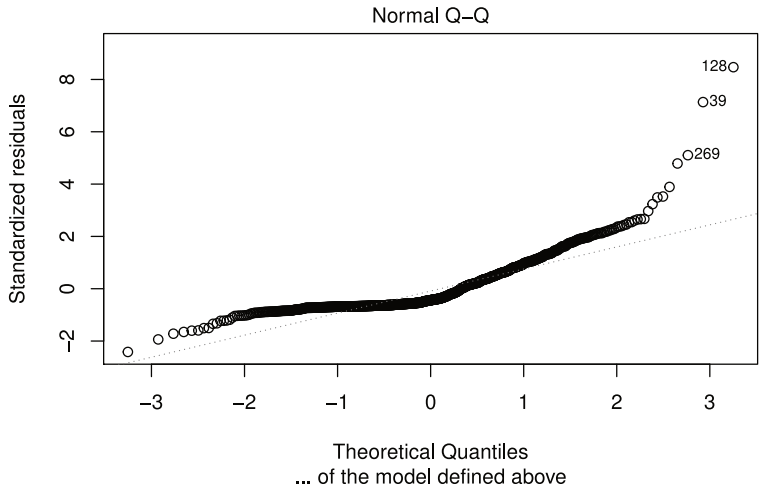


Abb. 5.2 QQ-Plot der Residuen

Der QQ-Plot in Abbildung 5.2 wird in der Statistik generell verwendet, um zu überprüfen ob eine Variable annähernd einer Normalverteilung folgt. Hier kann er eingesetzt werden, um die Normalität der Residuen zu überprüfen. Diese Annahme ist insbesondere relevant, wenn das Modell für die Vorhersage einzelner Datenpunkte verwendet werden soll (Gelman et al., 2020). Auffällig sind hier wieder dieselben Ausreißer wie in Abbildung 5.1. Hier könnte überlegt werden, die einzelnen Ausreißer aus der Analyse auszuschließen, eine Transformation der Variablen (wie z. B. logarithmieren) vorzunehmen oder ein robusteres Modell zu wählen, für das normalverteilte Residuen nicht notwendig sind.

Der Plot in Abbildung 5.3 überprüft die Annahme der Homoskedastizität. Auch diese Annahme ist besonders dann wichtig, wenn das Modell auch zur Vorhersage verwendet werden soll (Gelman et al., 2020). Wird das Modell nur zur Erklärung herangezogen, hat diese Annahme, ebenso wie die Annahme normalverteilter Residuen, eine nachgeordnete Bedeutung. Hier sollte die (rote) Linie annähernd einer horizontalen Linie folgen. Hier kann die Voraussetzung als gegeben angenommen werden. Sollte Heteroskedastizität vorliegen, ist es von Vorteil weighted least squares anstelle einer OLS-Regression zu verwenden. Sollten nicht-linare Zusammenhänge gefunden werden, sollten nicht-lineare Modelle (oder algortihmusbasierte Ansätze wie bspw. Machine Learning) in Betracht gezogen werden – Linearität stellt die wichtigste Voraussetzung dar.

```
plot(lm1, which~= 3, sub.caption~= "... of the model defined above")
```

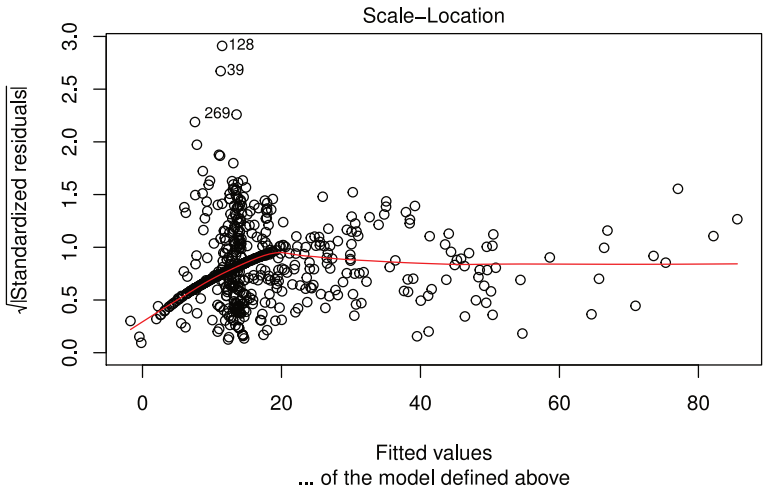


Abb. 5.3 Plot zur Homoskedastizität

```
plot(lm1, which~= 4, sub.caption~= "... of the model defined above")
```

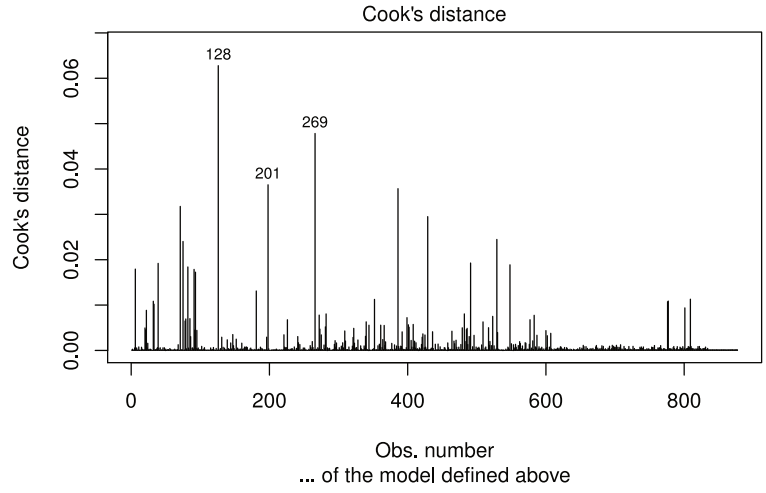


Abb. 5.4 Plot Cook's Distance

```
plot(lm1, which~= 5, sub.caption~= "... of the model defined above")
```

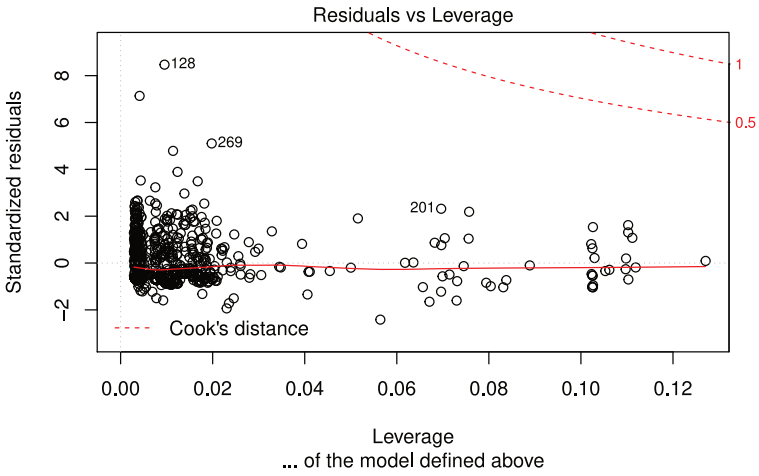


Abb. 5.5 Plot Einfluss von Extremwerten

Ausreißer, die hier als auffällig identifiziert wurden, können anschließend noch einer genaueren Leverage-Point Diagnostik unterzogen werden und müssen ggf. ausgeschlossen werden. Im Beispiel wurde keine Beobachtung identifiziert, die sowohl als Ausreißer als auch als Leverage Point identifiziert wurde. Nähere Informationen zur Outlier-Diagnostik, die hier absichtlich kurz gehalten wurde, können in bereits zitierten Standardwerken oder beispielsweise bei Fahrmeir et al. (2007) nachgelesen werden.

5.1.1.2 Durchführung der Analysen und Berichterstellung inkl. neuer Funktionen

Eine Übersicht der Ergebnisse kann mit der Funktion *summary()* aufgerufen werden. Um den Output für Stakeholder*innen übersichtlich darzustellen und einen Vergleich der Stärke der Einflussfaktoren zu ermöglichen, müssen standardisierte Regressionskoeffizienten berechnet werden. Das kann zum Beispiel über die Funktion *beta()* des R-Pakets *reghelper* erfolgen. Um die standardisierten Regressionskoeffizienten für dieses Anwendungsszenario zu berechnen und grafisch darzustellen, wurde im Projekt *Learning Analytics – Studierende im Fokus* die Funktion *betaplot()* geschrieben. Mit ihrer Hilfe können sowohl standardisierte als auch unstandardisierte Regressionskoeffizienten inkl. Konfidenzintervalle übersichtlich grafisch dargestellt werden. Alle verwendeten und neu geschriebenen Funktionen werden nach Abschluss der Projekts hier zur Verfügung gestellt: https://github.com/larissabartok/LAViennaLAVienna. Das

Paket wird laufend weiterentwickelt und die Neuerungen unter angegebenem Pfad zur Verfügung gestellt.

Bei der Funktion *betaplot()* ist es auch möglich, die standardisierten Regressionskoeffizienten anhand ihrer Größe, alphabetisch oder nach der Reihenfolge, wie sie in das Modell eingegangen sind, zu ordnen. Außerdem werden in Zukunft verschiedene Methoden der Standardisierung der Regressionskoeffizienten möglich sein. Zum Beispiel sollte eine andere Art der Standardisierung gewählt werden, wenn standardisierte Koeffizienten, die auf Basis von Dummy-Variablen gebildet werden, mit denen von metrischen Variablen verglichen werden sollen (Gelman, 2008).

Anstelle von einzelnen Nachbearbeitungen werden die unabhängigen Variablen (oder Faktorstufen) und die abhängige Variable auf Basis eines zuvor zu erstellenden, genau definierten .CSV-Files beschriftet. Dafür muss ein entsprechendes Beschriftungsfile in einem beliebigen Ordner, dessen Pfad der Funktion übergeben wird, abgespeichert werden. Zunächst sollen die Ergebnisse der Regressionsanalyse übersichtlich dargestellt werden. Das erfolgt mit Hilfe der Funktion *summ()* aus dem R-Paket *jtools*. Zunächst werden die benötigten R-Pakete geladen:

```
# Laden des Learning-Analytics R-Pakets, zuvor muss es von Github
# installiert werden
library(LAVienna)

# Laden des R-Pakets zur Tabellengestaltung:
# Zur Verwendung der Funktion summ()
library(jtools)
```

Anschließend werden die unstandardisierten Regressionskoeffizienten aus dem Objekt des linearen Modells extrahiert und übersichtlich in einer Tabelle dargestellt. Die unstandardisierten Regressionskoeffzienten können verwendet werden, um direkt den Einfluss der unabhängigen Variablen (in deren Einheiten) auf die abhängige Variable zu interpretieren. Mit Hilfe der Funktion *namenplot()* können die neuen Beschriftungen, unabhängig von ihrer Reihenfolge im Modell, im Output übernommen werden. Besonders wichtig ist hier die Interpretation des adjustierten R^2 für Stakeholder*innen: Je näher bei 1, umso mehr Varianz der abhängigen Variable (des Studienerfolgs) kann durch die unabhängigen Variablen im Modell erklärt werden. In unserem Beispiel können ungefähr 20 Prozent der Varianz der ECTS im Studienjahr 2019/2020 durch die Variablen im Modell erklärt werden.

```
# Extrahieren unabhängiger Variablen aus dem Objekt des linearen Modells
unabl <- variablen_lm(lml)

# Beschriftung der UV ändern
```

```
bsg_neu1 <- NULL
for(i in 1:length(unabl)){
 beschriftung1 <- namenplot(unabl[i], lm = lm1,
                            file = "./Daten/VariablenbeschriftungenUW.csv")
 bsg_neu1 <- c(bsg_neu1, beschriftung1)
}

summ1 <- jtools::summ(lm1)
rownames(summ1$coeftable)[2:length(rownames(summ1$coeftable))] <- bsg_neu1

# Bezeichnung der AV ändern
attr(summ1, which = "dv") <- variablenname(attr(summ1, which = "dv"),
                            file = "./Daten/VariablenbeschriftungenUW.csv")
summ1
```

Observations	877 (5 missing obs. deleted)
Dependent variable	38
Type	OLS linear regression

Tab. 5.1 Beschreibung Regressionsmodell – OLS-Regression

F(10,866)	22.04
R^2	0.20
Adj. R^2	0.19

Tab. 5.2 Modellfit Regressionsmodell – OLS-Regression

5.1.1.3 Interpretation für Stakeholder*innen

Für Stakeholder*innen ist üblicherweise die Interpretation der sogenannten standardisierten Regressionskoeffzienten interessant, da diese ermöglichen, den Einfluss einzelner Variablen vergleichend zu betrachten. Bei der folgenden Darstellung (Abbildung 5.6) sind die standardisierten Regressionskoeffizienten auf der x-Achse anhand eines Punktes aufgetragen. Das heißt, diese geben den Einfluss der jeweiligen sogenannten unabhängigen Variablen auf die abhängige Variable (in diesem Fall: ECTS im Studienjahr 2019/2020) an. Zudem ist die Schwankungsbreite mit Hilfe eines 95 Prozent Konfidenzintervalles visualisiert. Ist der Punkt rot, bedeutet das, dass die jeweilige Einflussgröße unter Berücksichtigung aller aufgelisteter Variablen im Modell einen signifikanten Einfluss auf Studienerfolg nach ECTS hat. Liegt der Punkt nahe an der strichlierten 0-Linie, kann man von einem kleinen Effekt ausgehen – ist er hingegen weit entfernt, von einem großen Effekt. Weiters muss die Richtung des Effekts interpretiert werden: Ein negativer Effekt (linke Seite der Grafik) bedeutet: Umso niedriger die Ausprägung in der jeweiligen Variablen, umso größer ist der Studienerfolg nach ECTS.

	Est.	S.E.	t val.	p
(Intercept)	35.38	6.98	5.07	0.00
Alter Studienbeginn	−0.49	0.19	−2.56	0.01
Geschlecht: w (Vergleich: m)	0.33	1.42	0.23	0.81
Auslandssemester: Nein (Vergleich: Ja)	−16.88	5.28	−3.20	0.00
Nein (Vergleich: Ja)	4.98	2.13	2.34	0.02
ECTS im Studienjahr 18/19	0.58	0.05	12.71	0.00
Schulform Matura: Andere (Vergleich: AHS)	2.39	2.56	0.93	0.35
Schulform Matura: BHS (Vergleich: AHS)	4.32	2.45	1.77	0.08
Schulform Matura: BRP (Vergleich: AHS)	7.89	6.68	1.18	0.24
Schulform Matura: Studienberechtigung (Vergleich: AHS)	1.08	6.90	0.16	0.88
Staatsbürgerschaft dichotom: Nein (Vergleich: Ja)	−2.76	2.46	−1.12	0.26

Standard errors: OLS

Tab. 5.3 Unstandardisierte Regressionskoeffizienten – OLS-Regression

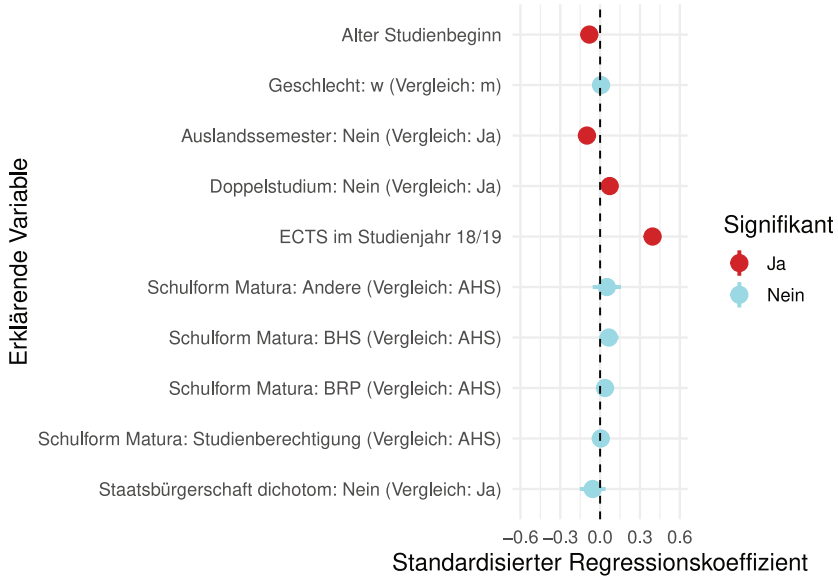


Abb. 5.6 Betaplot – OLS-Regression

Beispiel Interpretation für die Variable „ECTS im Studienjahr 2018/2019“: Die ECTS im Vorjahr haben einen signifikanten, positiven und großen Einfluss auf Studienerfolg nach ECTS. Umso mehr ECTS ein*e Studierende*r im Vorjahr

erbringt, umso besser ist sein/ihr Studienerfolg nach ECTS. Im Vergleich zu den anderen Variablen im Modell stellt diese Einflussgröße die stärkste dar.

5.1.2 Logistische Regression (abhängige Variable: Prüfungs(in)aktivität)

5.1.2.1 Durchführung der Analysen

Zur Modellierung der Prüfungsaktivität wurden logistische Regressionsmodelle verwendet und die selben Daten wie im vorherigen Modell herangezogen. Allerdings wurde die Variable der erbrachten ECTS-Punkte der Definition für Prüfungsakivität folgend dichotomisiert. Wie bereits oben in der Methodenbeschreibung illustriert, kommt hier die Funktion *glm()* zur Anwendung. Als Link-Funktion wurde hier im Beispiel eine Logit-Funktion verwendet. Als Visualisierung für Stakeholder*innen eigenen sich hier insbesondere die Odds Ratio. Dafür wurde die Funktion *oddsplot()* geschrieben. Auch diese Funktion wird laufend erweitert und zukünftig werden weitere Funktionen (wie bspw. das Ordnen der Koeffizienten der Größe nach) zur Verfügung stehen.

```
glm1 <- glm(INAKTIV ~ ALTER_STUDIENBEGINN + GESCHLECHT +
        MOBILITAET +   DOPPELSTUDIUM + ECTS_2018_19 +
        SCHULFORM1 + STAATSANGEHOERIGKEIT_CODE_AT,
        data~= daten_bw,  family~= "binomial")
```

```
# Extrahieren unabhängiger Variablen aus dem Objekt des linearen Modells
unab1 <- variablen_lm(lm1)

# Beschriftung der UV ändern
bsg_neu1 <- NULL
for(i in 1:length(unab1)){
 beschriftung1 <- namenplot(unab1[i], lm = lm1,
                        file = "./Daten/VariablenbeschriftungenUW.csv")
 bsg_neu1 <- c(bsg_neu1, beschriftung1)
}

summ1 <- jtools::summ(lm1)
rownames(summ1$coeftable)[2:length(rownames(summ1$coeftable))] <- bsg_neu1

# Bezeichnung der AV ändern
attr(summ1, which = "dv") <- variablenname(attr(summ1, which = "dv"),
                        file = "./Daten/VariablenbeschriftungenUW.csv")
summ1
```

Observations	877 (5 missing obs. deleted)
Dependent variable	60
Type	Generalized linear model
Family	binomial
Link	logit

Tab. 5.4 Beschreibung Regressionsmodell – Logistische Regression

$\chi^2(10)$	175.89
Pseudo-R^2 (Cragg-Uhler)	0.25
Pseudo-R^2 (McFadden)	0.15
AIC	1007.64
BIC	1060.18

Tab. 5.5 Modellfit Regressionsmodell – Logistische Regression

5.1.2.2 Berichterstellung inkl. neuer Funktionen und Interpretation für Stakeholder*innen

Die Odds Ratio (Chancenverhältnisse) können der Grafik entnommen werden. Eine Erhöhung einer unabhängigen Variable (um eine Einheit), geht bei Odds Ratios > 1 mit einer erhöhten, bei Odds Ratios < 1 mit einer verringerten Wahrscheinlichkeit für das Auftreten der betrachteten Ausprägung der abhängigen Variable einher. Im Datenbeispiel würde dies bedeuten: Die Odds von Studierenden im betrachteten Studienprogramm prüfungsaktiv im Folgejahr zu sein werden bei Studierenden ohne Nebenstudium als 2,5 Mal höher eingeschätzt als bei Studierenden mit Nebenstudium an der jeweiligen Universität.

	Est.	S.E.	z val.	p
(Intercept)	2.17	1.05	2.06	0.04
Alter Studienbeginn	−0.08	0.03	−2.71	0.01
Geschlecht: w (Vergleich: m)	0.30	0.16	1.86	0.06
Auslandssemester: Nein (Vergleich: Ja)	−2.58	0.86	−3.01	0.00
Doppelstudium: Nein (Vergleich: Ja)	0.95	0.28	3.38	0.00
ECTS im Studienjahr 18/19	0.08	0.01	7.60	0.00
Schulform Matura: Andere (Vergleich: AHS)	0.50	0.30	1.70	0.09
Schulform Matura: BHS (Vergleich: AHS)	0.43	0.28	1.54	0.12
Schulform Matura: BRP (Vergleich: AHS)	1.08	0.70	1.55	0.12
Schulform Matura: Studienberechtigung (Vergleich: AHS)	0.79	0.77	1.02	0.31
Staatsbürgerschaft dichotom: Nein (Vergleich: Ja)	−0.37	0.28	−1.31	0.19

Standard errors: MLE

Tab. 5.6 Unstandardisierte Regressionskoeffizienten – Logistische Regression

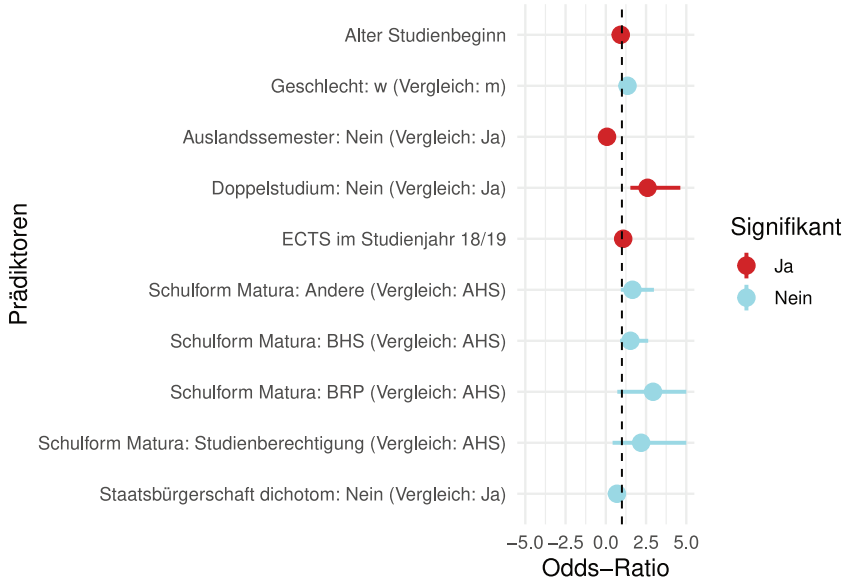


Abb. 5.7 Odds-Ratio Plot

5.2 Anwendungsszenario II: Prognose von Studienerfolg

Im folgenden Anwendungsszenario werden zwei Variablen auf Basis der Daten aus dem vergangenen Studienjahr prognostiziert. Verwendet werden diesmal Daten der WU. Begonnen wird mit der Prüfungsaktivität, welche als dichotome Variable die Ausprägungen 0 und 1 bzw. ja und nein hat. Wir beginnen hier nun mit der dichotomen, abhängigen Variablen, da Im Anschluss wird gezeigt, wie sich die Anzahl der ECTS-Punkte prognostizieren lässt. Die Prognose ist hinsichtlich des Prozesses zwar ähnlich, jedoch muss innerhalb der statistischen Verfahren und Kennzahlen beachtet werden, dass es sich einmal um eine dichotome und im anderen Fall um eine metrische Variable handelt. Die grundlegenden Schritte zur Prognose gelten für beide zu prognostizierenden Variablen und sind in folgender Abbildung 5.8 im Überblick zur Orientierung dargestellt.

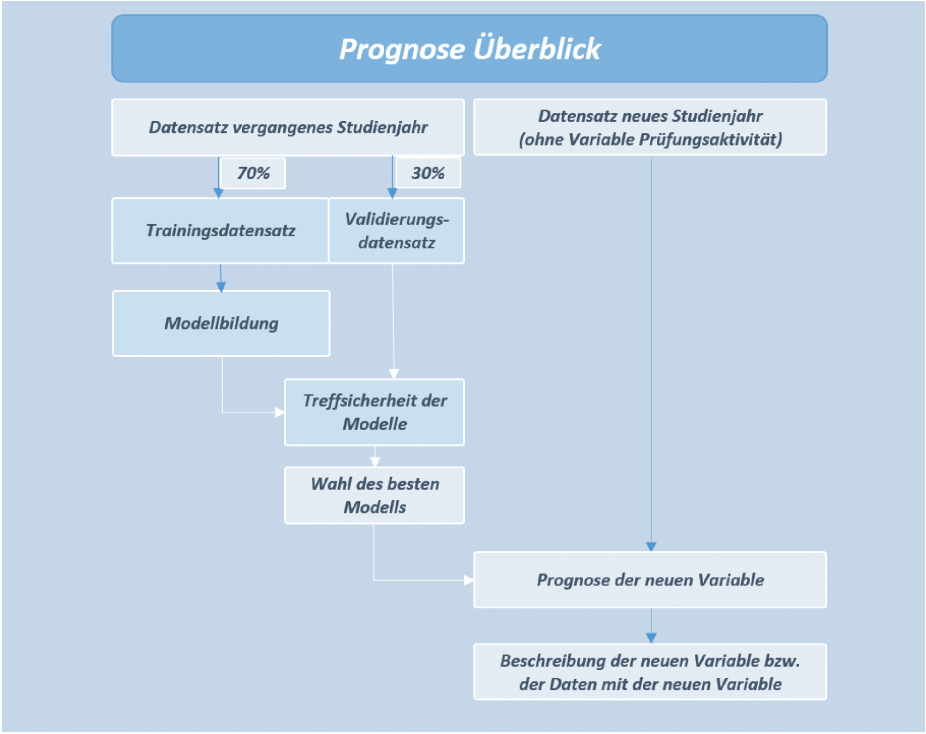


Abb. 5.8 Überblick der Prognoseberechnungen

Auf die Unterschiede und einzelnen Schritte wird in den praxisnahen Beispielen genauer eingegangen.

5.2.1 Abhängige Variable: Prüfungs(in)aktivität

Das folgende Kapitel stellt exemplarisch den Code für die Vorhersage von Prüfungsaktivität dar. In diesem Kapitel liegt der Fokus in der Vorhersage von Studienerfolg, um z. B. die Prüfungsaktivität für das nächste Studienjahr in einem Studienprogramm vorherzusagen. Nach der Auswahl der verwendeten Variablen wird auf Grund des Datensatzes des vorhergehenden, abgeschlossenen Studienjahres eine neue Variable „prüfungsaktiv ja/nein" des aktuell laufenden Studienjahres prognostiziert. Dieser Schritt erfolgt zum Beispiel im November oder Dezember, wenn aus dem vergangenen Studienjahr alle Daten vorhanden sind und aus dem neuen Studienjahr die Inskriptionsdaten und damit einhergehend die erklärenden Variablen, welche auch im vergangenen Studienjahr verwendet wurden, vorhanden sind.

Die Daten aus dem vergangenen Studienjahr und aus dem aktuellen Studienjahr unterscheiden sich also nur dadurch, dass im aktuellen Studienjahr keine Werte für „prüfungsaktiv ja/nein" vorhanden sind. Um dies zu erreichen,

wurde in den hier verwendeten Daten die Aktivität auf dem Online Learning Evironment (OLE) (welche Auskunft darüber gibt, ob Studierende auf der Lernplattform aktiv sind) im alten Datensatz angeglichen. Die Variable OLE-Aktivität verändert sich im Laufe des Studienjahres: Zu Beginn jedes Studienjahres sind Studierende auf der Lernplattform nicht aktiv, im Laufe des Studienjahres wird eine Vielzahl der Studierenden die Plattform nutzen und damit auf den Status OLE-aktiv gesetzt. Für die Angleichung kann man entweder für beide Datensätze die Variable im Dezember erheben, oder die Daten aus dem aktuellen Studienjahr werden hochgerechnet oder prognostiziert.

Wie in der Übersicht dargestellt beginnt die Prognose mit der Aufteilung des Datensatzes aus dem vergangenen Studienjahr.

5.2.1.1 Aufteilung in Trainings- und Validierungsdatensatz

Hierfür muss der Gesamtdatensatz zunächst randomisiert in einen Trainingsdatensatz und einen Validierungsdatensatz eingeteilt werden. Die Randomisierung ist für die Modellbildung wichtig, da der Trainingsdatensatz für die Modellbildung herangezogen wird und dieser aus 70 % der Daten des letzten Semesters besteht. Wären die Daten zum Beispiel nach Studienerfolg von sehr gut nach sehr schlecht geordnet, so könnte das Modell schlechten Studienerfolg nicht richtig identifizieren, da diese Daten für die Modellbildung und das Training der Modelle nicht zur Verfügung stünden.

Im Trainingsdatensatz werden verschiedene Modelle trainiert. Mittels Validierungsdatensatz, der aus den anderen 30 % der Daten des vergangenen Studienjahres besteht, auch Out Of Bag-Datensatz genannt, bewertet das Modell seine eigene Leistung.

Im vorliegenden Fall werden 70 % Daten randomisiert dem Trainingsdatensatz zugeordnet. Damit bleiben 30 % für den Validierungsdatensatz. Die prozentuelle Aufteilung kann frei gewählt werden. In kürzlich publizierten Artikeln mit ähnlichen Methoden findet sich häufig ein Trainingsdatensatz zwischen 60 % und 80 % (vgl. Mooney et al., 2021; Lee & Kim, 2022).

```
set.seed(666) train0 <- sample(nrow(dat0_modell1), 0.7*nrow(dat0_modell1),
        replace~= F)
TrainSet0 <- dat0_modell1[train0,]
ValidSet0 <- dat0_modell1[-train0,]
```

Im Beispiel arbeiten wir zum Zwecke der Demonstration mit einem Datensatz ohne fehlenden Werten. Der Umgang mit fehlenden Werten (Imputationsverfahren) ist nicht Teil des Umfangs dieses Erfahrungsberichts und stellt ein eigenes Themengebiet dar (vgl. hierzu für einen Überblick bspw. Lüdtke et al., 2007).

5.2.1.2 Modellformulierung

Folgende Modelle sollen in unserem Beispiel für die Berechnung der dichotomen Variable verwendet werden: logistische Regression, Random Forests, logistische Regression mit Boosting-Ansatz, Support Vector Machine-Modelle und Gradient Boosting Machine-Modelle.

Die Berechnung wird mit Hilfes der R-Packages *caret* Kuhn (2021) und *caretEnsemble* Deane-Mayer und Knowles (2019) durchgeführt. *Caret* bzw die Funktion *caretList()* ermöglicht es, mehrere Machine Learning-Modelle gleichzeitig mit denselben Parametern zu berechnen und anschließend zu vergleichen.

Hierfür kann zunächst die Funktion *trainControl()* verwendet werden, um die Art des Resamplings festzulegen. Statt Bootstrapping (default-Einstellung) wird hier in unserem Beispiel 10-fache Kreuzvalidierung verwendet. Die Modelle werden dann am selben Trainingsdatensatz mit denselben Resampling-Parametern trainiert. Beim Bootstrapping werden Datensätze durch „Ziehen mit Zurücklegen“ gebildet, somit kann ein Datensatz eine Beobachtung öfters oder gar nicht enthalten. Bei der Kreuzvalidierung wird sichergestellt, dass alle Stichproben in den Trainings- und Validierungsssätzen (test-sets) erscheinen, sodass alle Datensätze entweder für den einen oder den anderen Zweck verwendet werden. Dabei wird der Datensatz in diesem Fall in 10 zufällig gewählte Teilmengen aufgeteilt und eine Teilmenge der Daten zur Validierung des Modells verwendet, welches mittels der anderen neun Teilmengen trainiert wird. Dieser Vorgang wird in diesem Fall 10 Mal wiederholt, wobei jede Teilmenge einmal zur Validierung verwendet wird.

```
train_control0 <- trainControl(method="repeatedcv", number=10, repeats=3,
    classProbs=TRUE,
    index=createResample(TrainSet0$Studienaktiv, 25),
    savePredictions~= TRUE)
```

Die Funktion *preProc()* ermöglicht ein Pre-Processing im Zuge der Modellformulierung, so werden in diesem Beispiel die Daten skaliert und standardisiert. Durch die Prüfung der Voraussetzungen zur Verwendung der linearen und logistischen Funktion im vorhergehenden Abschnitt (bspw. der Umgang mit Ausreißern oder der Voraussetzung der Homoskedastizität) und die Skalierung und Standardisierung der Daten mit Hilfe der Pre-Processing Funktion werden die Voraussetzungen für die verwendeten Modelle als erfüllt angenommen und die Modelle können ohne weitere, genauere Überprüfung der Voraussetzungen je nach Modell berechnet werden.

Mittels *methodList()* oder *tuneList()* können unterschiedliche Machine Learning-Methoden gewählt werden. In diesem Beispiel wird als abhängige Variable die Prüfungsaktivität und somit eine dichotome Variable gewählt. Als Machine Learning-Methoden eignen sich daher beispielweise (Boosted) Logitische Regres-

sionen, Random Forest, Support Vector Machine (SVM) oder Gradient Boosting Machine (GBM). Wenn die Logistische Regression ohne Boosting-Ansatz verwendet werden soll, bietet es sich an, die Modelle in der Funktion *tuneList()* anzugeben, um sie zu spezifizieren. Werden die Modelle ohne weitere Spezifikation verwendet, können Sie einfach über die Funktion *methodList()* angegeben werden.

Anmerkung für Prognose der ECTS: Bei der Analyse von ECTS als metrischer abhängiger Variable können anstatt logistischer Regressionen beispielsweise lineare Regressionen oder Generalized Additive Models gerechnet werden.

```
modellist0_pruefungsaktiv <- caretList(
    Studienaktiv~., data=TrainSet0,
    trControl=train_control0,
    preProc = c("center", "scale"),
    tuneList = list( glm=caretModelSpec(method='glm', family='binomial'),
                     rf=caretModelSpec(method='rf'),
                     LogitBoost=caretModelSpec(method='LogitBoost'),
                     svmLinear=caretModelSpec(method='svmLinear'),
                     gbm=caretModelSpec(method='gbm')))
```

5.2.1.3 Treffsicherheit der Modelle

Um ein treffsicheres Modell aus der vorangehenden Vielfalt zu ermitteln, dienen hier im Fall dichotomer prognostizierter Daten folgende Kennzahlen zur Treffsicherheit:

- *Receiver Operating Characteristics-Kurven (ROC-Kurven)*: Hierfür werden in einem Diagramm die Falsch-Positiv-Rate (1-Spezifität) auf der x-Achse und die Richtig-Positiv-Rate (Sensitivität) auf der y-Achse dargestellt.
- *Accuracy (Gesamttrefferquote)*: Die Gesamttrefferquote kann als Anteil der korrekten Vorhersagen an allen getroffenen Prognosen verstanden werden (Gesamtheit der richtigen Vorhersagen dividiert durch Gesamtzahl der Vorhersagen).
- *Cohens Kappa*: vergleicht die beobachtete Genauigkeit mit einer erwarteten Genauigkeit.

Nach dem Modelltraining wird im Validierungsdatensatz, also in den verbleibenden 30 % des Datensatzes aus dem vergangenen Studienjahr, ermittelt, wie treffsicher die Modelle prognostizieren. Dazu werden automatisch die prognostizierten Werte mit den vorhandenen Werten verglichen.

Abbildung 5.9 vergleicht die Treffsicherheit der Modelle am Validierungsdatensatz. Für den vorliegenden Fall wird dies mittels ROC-Kurven dargestellt. Je weiter sich die Kurve eines Prognosemodells an die obere linke Ecke der Grafik annähert, desto besser ist somit die Trefferquote. Die Fläche unter der ROC-Kurve wird als AUC (Area under the Curve) bezeichnet und kann maximal

1 sein – wobei ein Wert unter 0,5 bedeuten würde, dass eine Zufallszuweisung bessere Vorhersagen erzielen würde. Eine AUC von 0,5 würde sich als Diagonale im Plot zeigen.

Mit folgenden Anweisungen wird der ROC-Plot dargestellt und die AUC-Werte werden berechnet.

```
modelpreds0 <- lapply(modellist0_pruefungsaktiv, predict,
newdata=ValidSet0)

modelpreds0 <- data.frame(modelpreds0)
modelpreds0_num <- sapply(modelpreds0, as.numeric)

colAUC(modelpreds0_num,ValidSet0$Studienaktiv, plotROC = T)
```

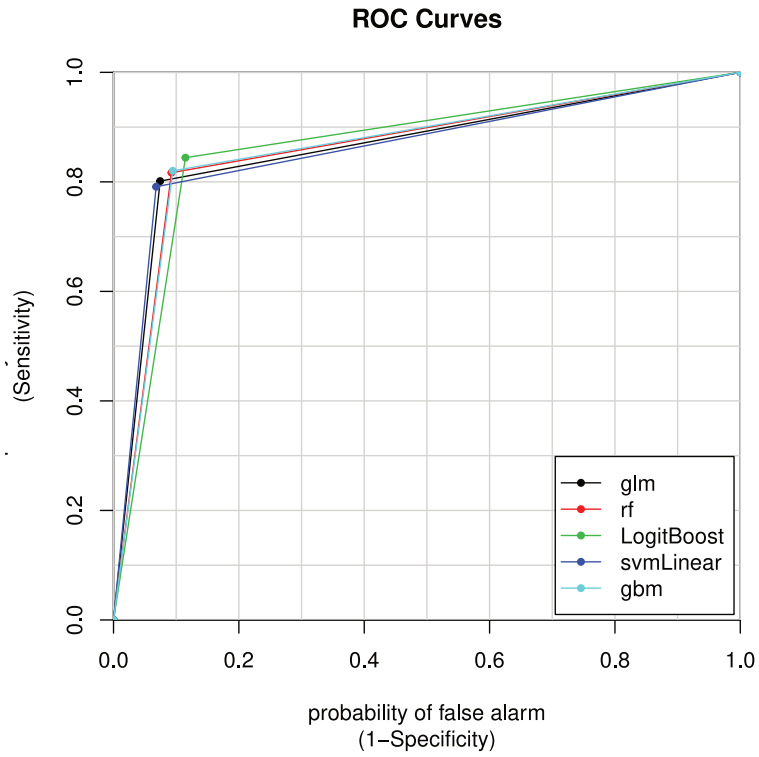


Abb. 5.9 ROC-Vergleich


```
modelpreds0 <- lapply(modellist0_pruefungsaktiv, predict,
newdata=ValidSet0)

modelpreds0 <- data.frame(modelpreds0)
modelpreds0_num <- sapply(modelpreds0, as.numeric)
```

	GLM	RF	Logit-Boost	SVMLi-near	GBM
A vs. B	0.863	0.862	0.865	0.862	0.863

Tab. 5.7 AUC

Auch die beiden anderen Kennzahlen Accuracy und Cohens Kappa können ohne größeren Aufwand übersichtlich dargestellt werden.

Die Wahl zu Receiver Operating Characteristics-Kurven als wichtiges Entscheidungskriterium wird im Bereich Machine Learning häufig verwendet, da die Darstellung der richtigen, positiven Vorhersagen (y-Achse) zu den falschen, positiven Vorhersagen (x-Achse) sehr verständlich ist und ein gutes Modell natürlich möglichst viele richtige Vorhersagen (und möglichst wenige falsche Vorhersagen) haben soll. Cohens Kappa ist ein Maß für die Übereinstimmung zwischen den tatsächlichen Werten, die ja im Validierungsdatensatz vorhanden sind und dem Klassifikator, also den Vorhersagen des jeweiligen Modells. Je höher der Grad der Übereinstimmung (d. h. Genauigkeit) ist, desto besser geeignet ist das Modell. Die Gesamttrefferquote (Accuracy) ist ebenso eine einfach interpretierbare Kennzahl. Sie kann im demonstrierten Beispiel problemlos als Vergleichskennzahl angewandt werden, da alle Modelle mit den selben Daten trainiert wurden und die Prognosen auf die selben Validierungsdaten erfolgten.

Weiterführende Informationen zu den verwendeten Kennzahlen finden sich bspw. in Ben-David (2008).

```
results0 <- resamples(modellist0_pruefungsaktiv) dotplot(results0)
```

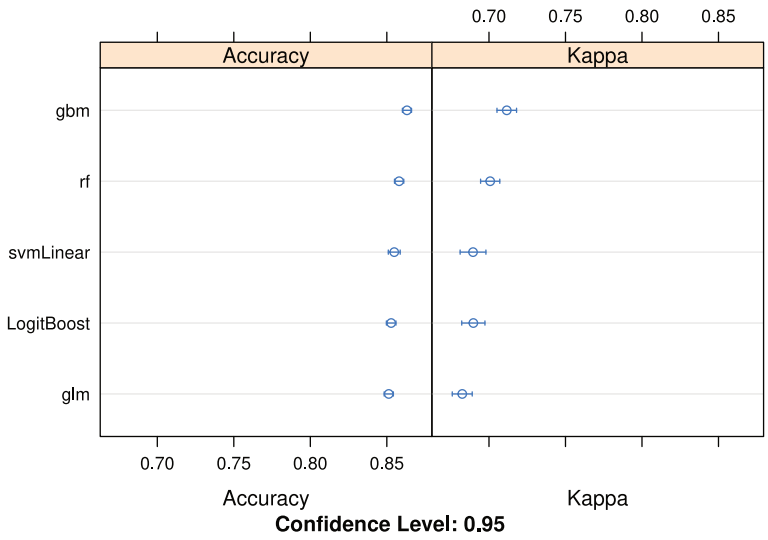


Abb. 5.10 Kennzahlen-Vergleich

Aus Abildung 5.10 geht hervor, dass rund 86 % aller getroffenen Vorhersagen korrekt waren, womit die Gesamttrefferquote allgemein sehr gut ist.

5.2.1.4 Auswahl des besten Modells

Um die weiteren Analysen automatisiert mit dem Modell mit bester Treffsicherheit durchführen zu können, kann dieses bspw. anhand des höchsten AUC-Wertes ausgewählt werden.

```
AUC <- as.data.frame(t(colAUC(modelpreds0_num,ValidSet0$Studienaktiv)))
AUC <- setDT(AUC, keep.rownames~= TRUE)[]
```

```
Ergebnis <- as.data.frame(AUC[AUC[, .I[AUC[,2] == max(AUC[,2])]]]) i <-
Ergebnis[1,1] # bestes Modell
```

In diesem Fall stellt das logistische Regressionsmodell mit Boosting-Ansatz das Modell mit der höchsten Treffsicherheit dar. Wie angemerkt wäre es auch denkbar, aufgrund des Kappa das Gradient Boosting Machine Modell für die Vorhersagen des kommenden Studienjahres zu verwenden, wir haben uns in diesem Beispiel anhand der ROC-Kurven entschieden.

In einzelnen Fällen („glm“, „rf“, „gbm“) gibt es die Möglichkeit die Stärke des Einflusses der einzelnen Variablen für die Modellierung darzustellen. Nicht alle Modelle arbeiten auf diese Weise, weshalb diese Betrachtung auch nur bei ausgewählten Modellen sinnvoll ist. Als Beispiel wird hier die geschätzte Relevanz der Variablen für das GBM Modell dargestellt. Das Diagramm zeigt, welche Variablen für das GBM Modell den stärksten Einfluss hatten. Es wird aber auch deutlich gezeigt, welche Variablen keinen Einfluss auf das Modell hatten. Grund für diesen kleinen Exkurs ist, dass diese Information bereits helfen kann, um Fragestellungen zur Prüfungsaktivität besser zu beleuchten.

Abbildung 5.11 zeigt die Variable Importance, d. h. wie groß der Einfluss der gewählten unabhängigen Variablen auf die Vorhersage ist – exemplarisch für das GBM Modell. Als Maß wird bei Klassifikationsproblemen der Gini Index für die Homogenität in den einzelnen Blattregionen (Node Impurity) verwendet und analysiert, wie sich diese bei Weglassen einer der unabhängigen Variablen verändern würde (Daniya et al., 2020). Inspiration für die Gestaltung des Variable Importance Plots gab hierfür die R-Bloggers-Community, insbesondere Lares (2018). Auch hier können natürlich die Odds-Ratio (logistische Regression) wieder grafisch dargestellt werden, wie im vorigen Anwendungsszenario illustriert.

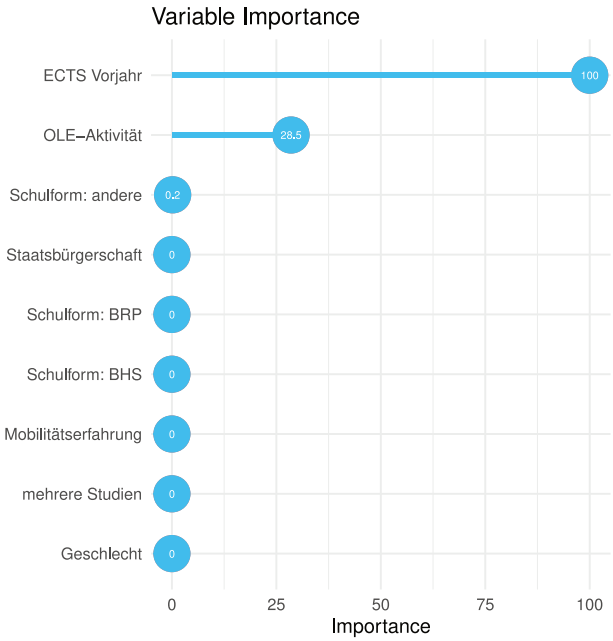


Abb. 5.11 Importance Plot

5.2.1.5 Vorhersage der Prüfungsaktivität auf neue Daten

Um das Modell nun Prognosewerte für neue Daten erstellen lassen zu können, müssen die neuen Daten gleich aufgebaut sein wie der Trainings- bzw. Validierungsdatensatz und alle im Trainingsdatensatz verwendeten Variablen müssen vorhanden sein. Nur die Variable Prüfungsaktiv ist natürlich noch nicht vorhanden. Mit Hilfe des folgenden Codes werden die durch das beste Modell vorhergesagten Werte in einer neuen Variable ‚Vorhersage' dem Datensatz hinzugefügt, dieser wird als Datensatz ‚predictions' gespeichert. Im zweiten Teil wird der Datensatz ‚predictions' als csv-File gespeichert.

```
newdata <- read.csv("./Daten/newdata.csv", sep=";")

  ## Vorhersage
  predictions <- cbind(newdata,
                       Vorhersage = predict(modellist0_pruefungsaktiv[[i]],
                       newdata = newdata, interval = 'prediction'))

  ## Erstellen des CSV
  write.csv(predictions, "./Daten/Predictions.csv", row.names = F)
```

Welche weitere Vorgehensweise mit den prognostizierten Daten erfolgt, obliegt der Hochschule und hängt von der Fragestellung und dem Ziel der Organisation ab. Welche Analysen und Darstellungen mit den neu gewonnenen

Daten erfolgt, ist daher nicht Teil dieses Dokuments. In diesem Bericht wird daher lediglich exemplarisch über *kmeans* ein Clusterplot mit drei Bereichen dargestellt. Darunter befindet sich ein Violinplot, mit dessen Hilfe man erkennen kann, in welchem Studiensemester sich viele aktive Studierende befinden.

Clusterverfahren können dazu verwendet werden, Studierende unterschiedlichen, in sich homogenen Gruppen, zuzuordnen, da die Clusterverfahren Ähnlichkeitsmaße zwischen Objekten (in diesem Fall Studierenden) berechnen. Eine Hochschule kann aufgrund dieser Gruppeneinteilung zielgerichtete Maßnahmen für unterschiedliche Gruppen entwickeln und anbieten.

Beispielsweise könnte eine Gruppe in hohem Ausmaß erwerbstätig sein, eine gewisse Altersstruktur oder andere Personenmerkmale aufweisen. Ein zielgerichtetes Informations- und Serviceangebot für Personen in einer homogenen Gruppe (z. B. zielgerichtete Kommunikation von für diese Gruppe besonders relevanten Angeboten/Informationen/Vernetzungsmöglichkeiten) könnten dann dabei helfen, die Studienaktivität oder Studierendenzufriedenheit zu erhöhen. Der Violinplot zeigt, dass Studierende im analysierten Studienprogramm mit einem Studienalter zwischen 4 und 5 Semestern besonders hohe Prüfungsaktivität aufweisen. Personen mit höherem Studienalter sind im Beispiel häufiger inaktiv. Derartige Darstellungen können dabei helfen, Personen aber auch Teile des Studiums zu identifizieren, die Zielgruppe einer Unterstützungsmaßnahme sein können.

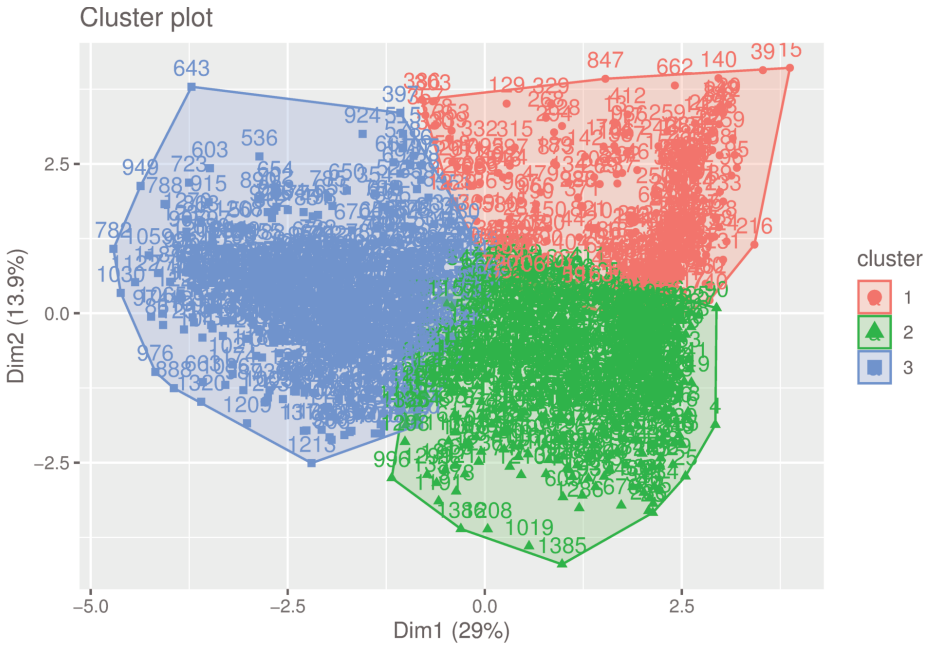


Abb. 5.12 kmeans Cluster und Violin Plot

Abb. 5.13 kmeans Cluster und Violin Plot

5.2.2 Abhängige Variable: Anzahl ECTS

Im folgenden Beispiel wird der Studienerfolg in Form von ECTS-Punkten (als metrische Variable) vorhergesagt. Die Fragestellung „Wie viele ECTS-Punkte werden Studierende im kommenden Studienjahr (im Durchschnitt) erreichen?“ klingt ähnlich der Prognose der Prüfungsaktivität im vorhergehenden Abschnitt, hier werden aber die ECTS-Punkte prognostiziert, ohne die Variable vorher zu dichotomisieren. Eine Person, die im Studienjahr 12 ECTS-Punkte erreicht, ist auf Grund der niedrigen Anzahl nicht als prüfungsaktiv zu werten, jedoch nimmt diese Person an Lehrveranstaltungen und Prüfungen teil. Diese Unterscheidung kann auf Basis der dichotomen Variable Prüfungsaktivität nicht getroffen werden. Auch bei dieser Prognose liegt hohe Priorität auf der Qualität der verwendeten Daten, denn ein Modell kann nur dann gut für die Prognosen funktionieren, wenn es an relevanten Testdaten trainiert wurde. Wie in der Übersicht 5.8 dargestellt, beginnt die Prognose mit der Aufteilung des Datensatzes aus dem vergangenen Studienjahr, dieser beinhaltet jetzt auch die Anzahl der erreichten ECTS-Punkte im vergangenen Studienjahr.

5.2.2.1 Aufteilung in Trainings- und Validierungsdatensatz

Auch in diesem Fall wurde eine Verteilung in 70 % Trainingsdaten und 30 % gewählt, die Daten wurden vorher randomisiert, Studierende wurden also wiederum zufällig einem der beiden Datensätze zugeteilt. Der Datensatz der WU, welcher die Anzahl der ECTS-Punkte enthält, wurde als dat0_modell2 gespeichert. Um die Konzentration auf die Methode zu lenken, wurden die gleichen erklärenden Variablen verwendet, die auch schon in der Prognose zur Prüfungsaktivität enthalten waren.

```
set.seed(666) # Für Reproduzierbarkit
train0 <- sample(nrow(dat0_modell2), 0.7*nrow(dat0_modell2), replace = F)
TrainSet0_ECTS <- dat0_modell2[train0,]
ValidSet0_ECTS <- dat0_modell2[-train0,]
```

Beobachtungen, die fehlende Daten enthalten, werden in unseren Anwendungsfällen nicht verwendet.

```
# Löschen von Beobachtungen mit fehlenden Daten
TrainSet0_ECTS <- na.omit(TrainSet0_ECTS)
ValidSet0_ECTS <- na.omit(ValidSet0_ECTS)
```

5.2.2.2 Modellformulierung

Folgende Modelle sollen in unserem Beispiel für die Berechnung der metrischen Variable verwendet werden: Generalisierte Additive Modelle, Lineare Modelle, Random Forests, Support Vector Machine Modelle und Gradient Boosting Machine-Modelle.

Die Berechnung wird wiederum mit Hilfes des R-Packets *caret* (Kuhn, 2021) und *caretEnsemble* (Deane-Mayer & Knowles, 2019) durchgeführt.

Vor der Prediction werden die Parameter für das Training festgelegt. Hier wird die Funktion *trainControl()* verwendet, welche mit 10-facher Kreuzvalidierung und drei Wiederholungen spezifiziert wird. Die interessierende Variable ist nicht mehr die Prüfungsaktivität, sondern „ECTSimSJ“ – die Anzahl der ECTS-Punkte im Studienjahr.

```
train_control0 <-trainControl(method="repeatedcv", number=10, repeats=3,
                  index=createResample(TrainSet0_ECTS$ECTSimSJ, 25),
                  savePredictions = TRUE)
```

Mit der Funktion *caretList()* werden gleichzeitig alle gewählten Modelle trainiert und anschließend deren Prognosegüte verglichen. Die Funktion *preProc()* ermöglicht ein Pre-Processing im Zuge der Modellformulierung, so werden in diesem Beispiel die Daten skaliert und standardisiert. Die einzelnen Modelle

werden in der *methodList()* aufgezählt, da hier keine weiteren Spezifizierungen notwendig sind. Die Berechnung über die Funktion *caretList()* kann, je nach technischen Gegebenheiten, mehr Zeit in Anspruch nehmen.

5.2.2.3 Treffsicherheit der Modelle

In den Prognosen metrischer Variablen wird für die Entscheidung, welches Modell am besten geeignet ist, die Prognosegüte ermittelt. Dies ist vergleichbar mit den ROC-Kurven, Accuracy und Cohens Kappa aus den dichotomen Modellen in den Fragen der Prüfungs(in)aktivität. Hier bei den metrischen Variablen werden folgende Kennzahlen ermittelt:

- R^2: setzt die Varianz der vorhergesagten Werte ins Verhältnis zur beobachteten Varianz
- *Mean Absolute Error (MAE)*: Mittelwert der absoluten Differenzen vorhergesagter und beobachteter Werte
- *Root Mean Square Error (RMSE)*: Wurzel aus dem gemittelten Fehlerquadrat

```
# Verteilung der vorhergesagten Werte & Treffsicherheit
modelpreds_ects <- lapply(modellist0_ects, predict,
newdata=ValidSet0_ECTS)
modelpreds_ects <- data.frame(modelpreds_ects)
modelpreds_ects_num <- sapply(modelpreds_ects, as.numeric)
resultsects <- resamples(modellist0_ects)

dotplot(resultsects)
```

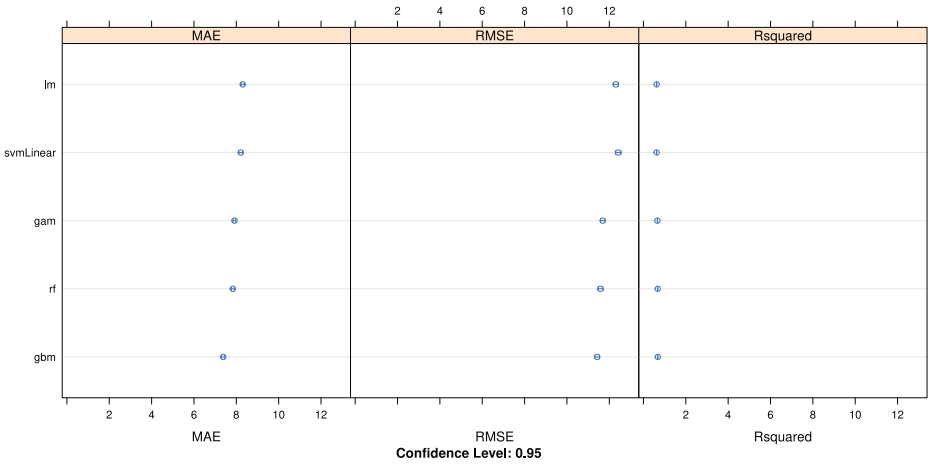


Abb. 5.14 Kennzahlenvergleich

Für eine Darstellung der Kennzahlen werden die Resultate in eine übersichtliche Form gebracht, mit Hilfe eines Punkte-Plots (siehe Abbildung 5.14) erlangt man eine gute Übersicht.

In diesem Beispiel wurde anhand des Mean Absolute Errors das beste Modell identifiziert. Der Mittelwert der (absoluten) Differenz der prognostizierten Werte zu den tatsächlichen Werten aus dem Validierungsdatensatz soll natürlich möglichst klein sein.

5.2.2.4 Auswahl des besten Modells

Gradient Boosting Machine liefert in diesem Anwendungsfall das Modell mit der besten Vorhersage.

Gerade wenn die Kennzahlen nahe aneinander liegen, hilft es für die Wahl des besten Modells, die Werte zahlenmäßig zu betrachten, was beispielsweise mit folgender Tabelle erreicht wird.

```
xtable::xtable(MAE0, caption="Mean Absolute Error")
## Direkt in R können die Kennzahlen auch über den Befehl:
## summary(resultsects) angezeigt werden.
```

	Model	Mean Absolute Errors
1	LM	8.30
2	RF	7.84
3	GAM	7.91
4	SVMLinear	8.21
5	GBM	7.38

Tab. 5.8 Mean Absolute Error

In Abbildung 5.15 wird deutlich, dass neben den im Vorjahr erreichten ECTS insbesondere auch das Studienalter entscheidend für die Prognose des GBM (Gardient Boosting Machine) waren. Zur Visualisierung der (standardisierten) Regressionskoeffizienten der linearen Regression können auch diese wie im Anwendungsfall I, visualisiert werden um den Einfluss der einzelnen Variablen grafisch darzustellen.

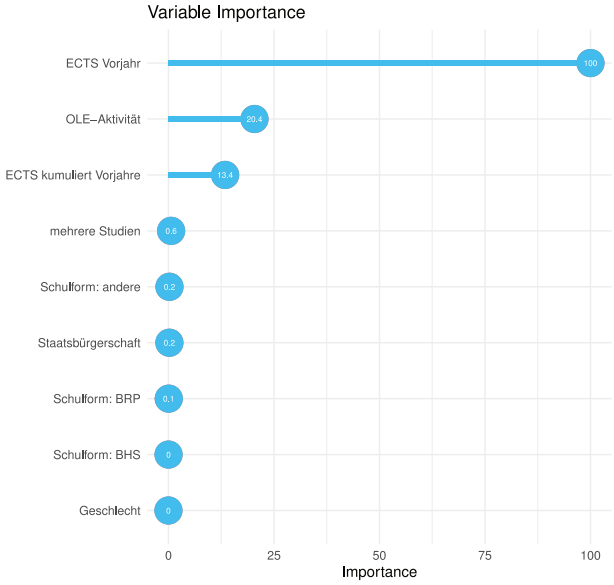


Abb. 5.15 Importance Plot

5.2.2.5 Vorhersage der Prüfungsaktivität auf neue Daten

Mit diesem Modell – erstellt und trainiert durch die Daten des vergangenen Studienjahres – werden nun die neuen Studierendendaten bzw. die Anzahl der erreichten ECTS-Punkte im neuen Studienjahr prognostiziert. Dazu wird das beste Modell verwendet und auf den neuen Datensatz angewandt.

```
newdatamet <- read.csv("./Daten/newdatamet.csv", sep=";")

  ## Vorhersage
  predictionsmet <- cbind(newdatamet,
                          Vorhersage = predict(modellist0_ects[[j0]],
                          newdata = newdatamet, interval = 'prediction'))

  ## Erstellen des CSV
  write.csv(predictionsmet, './Daten/Predictions_ects.csv', row.names = F)
```

Welche weiteren Analysen folgen, ist abhängig von der genauen Fragestellung.

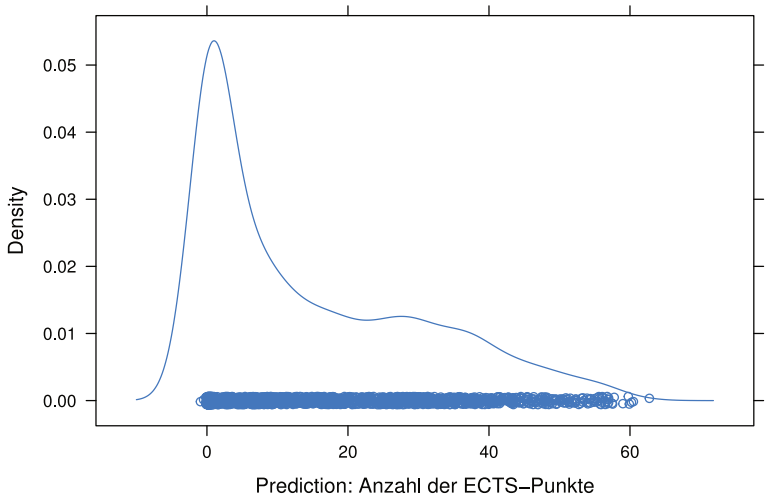


Abb. 5.16 Density Plot

In der Abbildung 5.16 ist die Dichteverteilung von Studierenden nach ihrer Anzahl der vorhergesagten ECTS-Punkte ersichtlich. Die Dichte wurde mithilfe eines Glättungsparameters berechnet, der dafür verantwortlich ist, dass in der Darstellung auch negative ECTS-Werte möglich wären, die in der Realtität natürlich nicht existieren. Unter der Dichtekurve befinden sich 100 % der Studierenden. Durch die Dichtedarstellung vermeidet man einerseits, auf einem Prognosewert zu verharren ohne die Varianz der Daten zu berücksichtigen, andererseits ist aber auch sofort erkennbar, in welchem ECTS-Punkte-Bereich sich viele Studierende befinden.

6 Lessons learned und Limitationen

Dieses Dokument gibt – basierend auf der Arbeit der Arbeitsgruppe „Variablendefinition und Modellbildung“ im Cluster Learning Analytics (Projekt „PASSt – Predictive Analytics Services für Studierendenerfolgsmanagement“ & Projekt „Learning Analytics – Studierende im Fokus“) – einen Einblick in das Themenfeld Model-Based-Analytics mit Fokus auf Studierendenerfolg. Dabei werden nach einer Diskussion der vorhandenen Literatur und möglicher, zu verwendender Methoden zwei analytische Anwendungsszenarien entlang dem verwendeten Programmiercode beschrieben und diskutiert. Erste Ergebnisse: Die zwei beschriebenen Anwendungsszenarien zeigen konkret, wie Learning Analytics dazu verwendet werden können, Einflussvariablen auf den Studienerfolg zu identifizieren und deren Wirkung zu beschreiben (Anwendungsszenario I) bzw. zukünftige Werte im Zusammenhang mit Studienerfolg als ECTS-Punkte [metrisch] oder Prüfungsaktivität [dichotom] zu prognostizieren (Anwendungsszenario II). Bevor mit entsprechenden Analytics-Projekten begonnen werden kann, müssen die dafür notwendigen Voraussetzungen an Institutionen geschaffen werden.

In Anlehnung an das Rahmenmodell von Norris und Baer (2013) zeigte sich sehr deutlich, dass an einer Hochschule Gelingensbedingungen für die Implementierung von Analytics-Projekten im Allgemeinen und Learning Analytics im Speziellen vorliegen müssen. Im Rahmen der Arbeitsgruppe haben wir folgende relevante Fragestellungen identifiziert:

- Ist die Hochschule mit den notwendigen technischen Strukturen und Ressourcen für das Projekt ausgestattet? (Datenverfügbarkeit, Datenstruktur, Tools, Anwendungen)
- Sind die rechtlichen Rahmenbedingungen für das Projekt an der Hochschule geklärt?
- Ist die Hochschule mit den notwendigen budgetären Ressourcen für das jeweilige Analytics-Projekt ausgestattet?
- Gibt es Personen an der Hochschule, die Analytics-Projekte entwickeln und technisch implementieren können?
- Gibt es Personen an der Hochschule, die Analytics-Projekte prozessual implementieren und betreuen können (Dateninterpretation, Weiterentwicklung an die Bedarfe der Hochschule…)?
- Gibt es innerhalb der Hochschule die notwendige Kooperation zwischen den zuständigen Einheiten und Akteuren um Analytics-Projekte zu entwickeln?
- Wurden konkrete Ziele formuliert, die mit dem Projekt erreicht werden sollen?

Gemäß Norris und Baer (2013) zählen eben zu den Gelingensbedingungen das Vorliegen der relevanten Infrastruktur im Datenmanagement (Datenverfügbarkeit, Datenzugriff) sowie die Verfügbarkeit von Expert*innen mit einschlägigem Know-how in der Entwicklung und in der Interpretation der verwendeten statistischen Modelle (Talent Gap). Der ebenfalls von Norris und Baer (2013) postulierte Collaboration Gap wird auf institutioneller Ebene durch die laufenden Projekte und eben auch diese Arbeitsgruppe bereits ein Stück weit geschlossen.

Die Umsetzung an weiteren Hochschulen bedarf folglich einer Evaluation der oben angeführten Voraussetzungen, bevor mit den in dieser Arbeit gezeigten Analysen begonnen werden kann. Die im Projekt aufgegriffenen Gelingensbedingungen spiegeln sich ebenfalls in neueren Arbeiten im internationalen Feld wider (Wegner et al., 2023). Hier ergeben sich in Zukunft eventuell auch Parallelen zu Open Data und der Notwendigkeit, Kooperationen und Shared Services aufzubauen.

Einmal mehr hat sich sehr deutlich gezeigt, dass zur Entwicklung, aber auch zum Betrieb von Analytics-Projekten das Festlegen eines zu lösenden Problems bzw. eines zu erreichenden Ziels im Vorfeld wesentlich ist. Unterschiedliche Ziele werden sich durch die verschiedene Ausrichtung des jeweiligen Projekts auf die Art der Modellbildung auswirken. In diesem Erfahrungsbericht sind zwei exemplarische Anwendungsfälle mit unterschiedlichen Zielsetzungen geschildert, welche eine unterschiedliche Vorgehensweise erfordern und zur Folge haben können, dass methodische Entscheidungen unterschiedlich ausfallen. So kann eine Zielsetzung etwa auf eine Anwendung klassischer Methoden (wie z. B. OLS-Regressionen) hinauslaufen oder einen Fokus auf Modelle mit höherer Prognosegüte (z. B. Random Forest) legen, deren Ergebnisse jedoch schwieriger an Stakeholder*innen zu kommunizieren sind. Eine exemplarische Zielsetzung im ersten Fall wäre das Ausfindigmachen von Faktoren, die einen Einfluss auf Studienerfolg in einem bestimmten Studienprogramm haben. Beispielhafte Zielsetzungen im zweiten Fall könnten die Vorhersage der Anzahl aktiver Studierender im nächsten Semester oder das Bestimmen von Drop-out-Risiken sein. Auch ein Clustering unterschiedlicher Studierendengruppen auf Basis relevanter Personenmerkmale wäre damit möglich, um in weiterer Folge bedarfsspezifische Angebote bereitzustellen.

Diese Entscheidungen zur Zielsetzung sind bereits in der konzeptionellen Phase des Projekts verbindlich zu treffen.

Diese Ziele leiten dann auch bei der notwendigen Definition der Population bzw. populationsbeschreibender Parameter an, die als unabhängige Variablen ins Modell aufgenommen werden müssen. Diese „theoriegeleitete Variablenauswahl“ ist projektbezogen zu treffen: Eine Standardlösung gibt es hier ebenso wenig wie für die Definition der abhängigen Variablen, denn auch der Studienerfolg wird bislang nur über Arbeitsdefinitionen operationalisiert, eine all-

gemeingültige Definition existiert in der Literatur nicht. Von der „theoriegeleiteten Variablenauswahl" klar abzugrenzen, ist in weiterer Folge die „modellgeleitete Variablenauswahl", die anhand statistischer Imperative erfolgt und zum Ausschluss zu stark miteinander korrelierter oder nicht signifikanter Variablen führen kann. Auch der Ausschluss der unabhängigen Variablen ist abhängig von der Zielsetzung. So würde man bei Interesse an Prädiktion nur Variablen im Modell behalten, die einen Beitrag zur Prognose der abhängigen Variablen leisten. Ist man aber zum Beispiel daran interessiert, den Einfluss einer bestimmten Variablen „statistisch zu kontrollieren" oder etwa auch den Nicht-Einfluss einer vermeintlich wichtigen Variablen aufzuzeigen, dann könnte selbst eine Variable mit geringem Erklärungswert im Modell behalten werden.

Es muss aber auch die Verwertungslogik des Projekts in der Konzeptionsphase berücksichtigt werden, also die geplante Rückmeldung und Darstellung der Ergebnisse an Stakeholder*innengruppen. So sind die in Anwendungsszenario I gezeigten linearen Modelle, die in detaillierten und automatisch generierten Berichten eingebettet worden sind, besonders geeignet für die Stakeholder*innenkommunikation. Sie ermöglichen es, diverse Personengruppen mit unterschiedlicher Vorbildung auch im (technischen) Detail über einzelne Einflussfaktoren auf Facetten von Studienerfolg (Prüfungsaktivität, Studienabbruch, Notenschnitt…) zu informieren und ggf. zu beraten. Allerdings zeigen diese linearen Modelle zuweilen eine schlechtere Prognosegüte als Supervised-Machine-Learning-Verfahren (Anwendungsszenario II), deren Interpretation bereits ein gewisses Spezialist*innenwissen voraussetzt. Diese Modelle eignen sich somit insbesondere für prognostische Zielsetzungen, konkrete Vorhersagen über kommende Semester oder gar auf der Ebene einzelner Studierender, um zum Beispiel frühzeitig individuellen Unterstützungsbedarf zu erkennen, knappe Ressourcen zielgerichtet einsetzen und Studierende bestmöglich fördern zu können. Auf Basis eines hierarchischen Clusterings könnte man zum Beispiel auch studien(in)aktive Personengruppen identifizieren oder Lehrveranstaltungen und Studienprogramme mit besonderem Unterstützungsbedarf ermitteln.

Mit dem in diesem Dokument inkludiertem R-Code für Praktiker*innen an Hochschulen möchten wir einen Beitrag dazu leisten, die Lücke zwischen dem herrschenden institutionellen Bedarf und angebotenen technischen Lösungen („gap between institutional needs and solution provider offerings" nach Norris und Baer, 2013) ein Stück weit zu schließen.

Abschließend kann festgestellt werden, dass die Ausrichtung des jeweiligen Projekts bzw. die konkret verfolgte Problemstellung und somit (exemplarisch) die Wahl der Methode, die Ergebnisdarstellung und Kommunikation mit Stakeholder*innen und die Einbettung innerhalb der Hochschule notwendige Rahmenbedingungen für den technisch-analytischen Kern von Analytics-Projekten darstellen. Eine One-size-fits-all-Lösung für derartige Analytics-Projekte

kann es aufgrund unterschiedlicher Problemstellungen und unterschiedlicher Rahmenbedingungen an den Hochschulen nicht geben. Den Erfahrungen dieser Arbeitsgruppe folgend sollte eine hochschulübergreifende Zusammenarbeit zwischen Expert*innen zu diesem Thema unterstützt werden, da beide Projekte (PASSt und LA – Studierende im Fokus) durch den methodischen Austausch nachhaltig profitiert haben und die in dieser Arbeit präsentierten Erkenntnisse durch die gewinnbringende Zusammenarbeit entstanden sind. Die Zusammenarbeit über Hochschulen, Projekte und Netzwerke hinweg sollte daher mit einer langfristigen Perspektive aktiv gefördert werden.

Literatur

Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., Chang, W., & Iannone, R. (2021). *rmarkdown: Dynamic Documents for R*. https://github.com/rstudio/rmarkdown

Bartok, L., Gleeson, R., & Kriegler-Kastelic, G. (2021). The impact of individual factors on definitions of academic success at an austrian university. *Studierbarkeit und Studienerfolg: Zwischen Konzepten, Analysen und Steuerungspraxis*, *4*, 119. https://doi.org/10.3217/zfhe-16-04/07

Bartok, L., Hubert, M., Gleeson, R., & Kriegler-Kastelic, G. (2023). Eine datengestützte Peer-Beratung zur Unterstützung individueller Studienziele. *Zeitschrift für Hochschulentwicklung*, *18*(3), 109–135. https://doi.org/10.21240/zfhe/18-03/06

Bartok, L., Ledermüller, K., & Tauböck, S. (2022). *Dealing with the impact of new technologies: The role of learning analytics*. LOTUS Policy Dialogue Workshop on Leading Digitalisation, WU Vienna.

Beaulac, C., & Rosenthal, J. S. (2018). Predicting university students' academic success and choice of major using random forests. *ArXiv E-Prints*. https://doi.org/10.1007/s11162-019-09546-y

Ben-David, A. (2008). About the relationship between ROC curves and Cohen's Kappa. *Engineering Applications of Artificial Intelligence*, *21*(6), 874–882. https://doi.org/10.1016/j.engappai.2007.09.009

Boehmke, B., & Greenwell, B. (2019). *Hands-on Machine Learning with R*. Chapman; Hall/CRC. https://doi.org/10.1201/9780367816377

Borden, V. M., & Jin, S. (2022). An Evidence-Based Framework for Transforming Higher Education Programs and Processes. In B. Broucker, R. Pritchard, R. Krempkow & C. Milsom (Hrsg.), *Transformation fast and slow* 117–134. Brill. https://doi.org/10.1163/9789004520912_007

Bornkessel, P. (2018). *Erfolg im Studium. Konzeptionen, Befunde und Desiderate*. wbv. https://doi.org/10.3278/6004654w

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Burkov, A. (2019). *Machine Learning kompakt: Alles, was Sie wissen müssen*. MITP-Verlagsgesellschaft.

Buß, I. (2019). The relevance of study programme structures for flexible learning: An empirical analysis. *Zeitschrift für Hochschulentwicklung*, *14*(3), 303–321. https://doi.org/10.3217/zfhe-14-03/18

Daniel, H.-D., Krempkow, R., & Schmidt, U. (2019). *Studienerfolg und Studienabbruch*. Bielefeld: Universitatsverlag Webler.

Daniya, T., Geetha, M., & Kumar, K. S. (2020). Classification and regression trees with Gini index. *Advances in Mathematics Scientific Journal*, *9*(10), 8237–8247. https://doi.org/10.37418/amsj.9.10.53

Davenport, T., & Harris, J. (2007). *Competing on analytics: The new science of winning*. Harvard Business Press. https://doi.org/10.5860/choice.44-6322

Deane-Mayer, Z. A., & Knowles, J. E. (2019). *CaretEnsemble: Ensembles of caret models*. https://CRAN.R-project.org/package=caretEnsemble

Fahrmeir, L., Kneib, T., & Lang, S. (2007). *Regressionsmodelle*. Springer.

Falk, S., Kercher, J., & Zimmermann, J. (2022). Internationale Studierende in Deutschland: Ein Überblick zu Studiensituation, spezifischen Problemlagen und Studienerfolg. *Beiträge zur Hochschulforschung*, *2*(3), 2022. https://doi.org/10.46685/DAADStudien.2022.05

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. Sage Publications.

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting (with discussion and a rejoinder by the authors). *The Annals of Statistics*, *28*(2), 337–407. https://doi.org/10.1214/aos/1016218223

Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine*, *27*(15), 2865–2873. https://doi.org/10.1002/sim.3107

Gelman, A., Hill, J., & Vehtari, A. (2020). *Regression and other stories*. Cambridge University Press. https://doi.org/10.1017/9781139161879

Guisan, A., Edwards Jr, T. C., & Hastie, T. (2002). Generalized linear and generalized additive models in studies of species distributions: Setting the scene. *Ecological Modelling*, *157*(2–3), 89–100. https://doi.org/10.1016/S0304-3800(02)00204-1

Harris, K.-L. (2007). A critical examination of a recent performance-based incentive fund for teaching excellence in australia. In B. Longden & K.L. Harris (Hrsg.), *Funding Higher Education: A Question of Who Pays*, 62–78.

Hastie, T., & Tibshirani, R. (1990). *Generalized Additive Models*. Chapman; Hall. https://doi.org/10.1214/ss/1177013604

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Vol. 2). Springer. http://www-stat.stanford.edu/ tibs/ElemStatLearn/

Hatzinger, R., Hornik, K., & Nagel, H. (2011). *R: Einführung durch angewandte Statistik*. Pearson Deutschland.

Hochschulforum Digitalisierung. (2015). *Diskussionspapier – 20 Thesen zur Digitalisierung der Hochschulbildung, Arbeitspapier Nr. 14*. Berlin: Hochschulforum Digitalisierung.

Isett, K. R., & Hicks, D. (2020). Pathways from research into public decision making: Intermediaries as the third community. *Perspectives on Public Management and Governance*, *3*(1), 45–58. https://doi.org/10.1093/ppmgov/gvz020

Janson, K., Krempkow, R., & Thiedig, C. (2024). *Was beeinflusst die Nutzung von Daten für die Qualitätsentwicklung der Lehre? – Ein Zwischenbericht des Projekts NuDHe*. Evaluation an Hochschulen im Spannungsfeld zwischen Wissenschaftlichkeit und Pragmatismus. (Dokumentation Frühjahrstagung des Arbeitskreis Hochschulen der DeGEval, 15.-16.05.2023, Goethe-Universität Frankfurt/Main.); Berlin: DUZ Verlags- und Medienhaus. (angenommen zur Veröffentlichung).

Janson, K., & Rathke, J. (2023). Weiterbildung im Wissenschaftsmanagement: Was muss man wissen, über das Wissen? In R. Krempkow, K. Janson, S. Harris-Huemmert, J. Rathke, E. Höhle & M. Hölscher (Hrsg.), *Berufsfeld Wissenschaftsmanagement*, 121–147. Bielefeld: Universitätsverlag Webler. https://doi.org/10.53183/9783946017301

Karatzoglou, A., Meyer, D., & Hornik, K. (2006). Support vector machines in R. *Journal of Statistical Software*, *15*, 1–28. https://doi.org/10.18637/jss.v015.i09

Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2004). kernlab-an S4 package for kernel methods in R. *Journal of Statistical Software*, *11*(9), 1–20. https://doi.org/10.18637/jss.v011.i09

Krempkow, R. (2007). *Leistungsbewertung, Leistungsanreize und die Qualität der Hochschullehre*. UVW, Univ.-Verl. Webler.

Krempkow, R. (2015). Can performance-based funding enhance diversity in higher education institutions? In R. M. O. Pritchard, M. Klumpp & U. Teichler (Hrsg.), *Diversity and excellence in higher education*, 231–244. Rotterdam: SensePublishers. https://doi.org/10.1007/978-94-6300-172-4_13

Krempkow, R. (2020a). Determinanten der Studiendauer – individuelle oder institutionelle Faktoren? Sekundärdatenanalyse einer bundesweiten Absolvent(inn)enbefragung. *Zeitschrift für Evaluation*, *19*(1), 37–67. https://3 doi.org/10.31244/zfe.2020.01.03

Krempkow, R. (2020b). Was beeinflusst die Studiendauer? Die Rolle individueller und institutioneller Faktoren. *Qualitätssicherung im Student Life Cycle*, 27–42.

Krempkow, R. (2022). *Studieneinstieg und Studien(einstiegs-)erfolg*. Universitätsverlag Webler.

Krempkow, R., Vettori, O., & Buss, I. (2021). Studierbarkeit und Studienerfolg: Zwischen Konzepten, Analysen und Steuerungspraxis. *Zeitschrift für Hochschulentwicklung – ZFHE*, *17*(4).

Kuhn, M. (2021). *Caret: Classification and regression training*. https://CRAN.R-project.org/package=caret

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer. https://doi.org/10.1007/978-1-4614-6849-3

Lantz, B. (2019). *Machine learning with R: Expert techniques for predictive modeling*. Packt publishing ltd.

Lares, B. (2018). *Machine Learning Results in R: one plot to rule them all*. R-Bloggers.com. https://www.r-bloggers.com/2018/07/machine-learning-results-in-r-one-plot-to-rule-them-all/

Lee, C., & Kim, H. (2022). Machine learning-based predictive modeling of depression in hypertensive populations. *PLoS One*, *17*(7), e0272330. https://doi.org/10.1371/journal.pone.0272330

Leitner, P., Ebner, M., Ammenwerth, E., Andergassen, M., Csanyi, G., Gröblinger, O., Kopp, M., Reichl, F., Schmid, M., Steinbacher, P. et al. (2019). *Learning Analytics: Einsatz an österreichischen Hochschulen*. Graz: Verein Forum neue Medien in der Lehre Austria (fnma).

Liu, Y., Wang, Y., & Zhang, J. (2012). New machine learning algorithm: Random forest. In B. Liu, M. Ma & J. Chang (Hrsg.), *Information Computing and Applications*, 246–252. https://doi.org/10.1007/978-3-642-34062-8_32

Loder, A. K. F. (2023). Predicting the Number of Active Students: A Method for Preventive University Management. *Journal of College Student Retention: Research, Theory & Practice*, 1–20. https://doi.org/10.1177/15210251231201394

Lübcke, M., Schrumpf, J., Schurz, K., Seyfeli-Özhizalan, F., Thelen, T., Wannemacher, K., & Weber, F. (2021). Mit digitalen Studienassistenzsystemen durchs Studium. Call for Papers. *Zeitschrift für Hochschulentwicklung – ZFHE*, *19*(4).

Lübcke, M., Schrumpf, J., Seyfeli-Özhizalan, F., & Wannemacher, K. (2023). Künstliche Intelligenz zur Studienindividualisierung. Der Ansatz von SIDDATA. In T. Schmohl, A. Watanabe & K. Schelling (Hrsg.), *Künstliche Intelligenz in der Hochschulbildung. Chancen und Grenzen des Ki-gestützten Lernens und Lehrens*, 213–226. https://doi.org/10.14361/9783839457696-012

Lüdtke, O., Robitzsch, A., Trautwein, U., & Köller, O. (2007). Umgang mit fehlenden werten in der psychologischen forschung. *Psychologische Rundschau*, *58*(2), 103–117. https://doi.org/10.1026/0033-3042.58.2.103

Marx, B. D., & Eilers, P. H. (1998). Direct generalized additive modeling with penalized likelihood. *Computational Statistics & Data Analysis*, *28*(2), 193–209. https://doi.org/10.1016/S0167-9473(98)00033-4

Menard, S. (2002). *Applied logistic regression analysis*. Sage. https://doi.org/10.4135/9781412983433

Meyer-Guckel, V., & Jorzik, B. (2015). Studienerfolg – Schlaglichter auf einen blinden Fleck der Exzellenzdebatte. In C. Berthold, B. Jorzik & V. Meyer-Guckel (Hrsg.), *Handbuch Studienerfolg. Strategien und Maßnahmen: Wie Hochschulen Studierende erfolgreich zum Abschluss führen*, 6–12. Stifterverband.

Mooney, C., Eogan, M., Ni Ainle, F., Cleary, B., Gallagher, J. J., O'Loughlin, J., & Drew, R. J. (2021). Predicting bacteraemia in maternity patients using full blood count parameters: A supervised machine learning algorithm approach. *International Journal of Laboratory Hematology*, *43*(4), 609–615. https://doi.org/10.1111/ijlh.13434

Nawar, S., & Mouazen, A. M. (2017). Comparison between random forests, artificial neural networks and gradient boosted machines methods of on-line vis-nir spectroscopy measurements of soil total nitrogen and total carbon. *Sensors*, *17*(10), 2428. https://doi.org/10.3390/s17102428

Neugebauer, M., Daniel, H.-D., & Wolter, A. (2021). *Studienerfolg und Studienabbruch*. Springer. https://doi.org/10.1007/978-3-658-32892-4_10

Norris, D. M., & Baer, L. L. (2013). Building organizational capacity for analytics. *Educause Learning Initiative*, *2013*, 7–56. https://library.educause.edu/resources/2013/2/building-organizational-capacity-for-analytics. https://doi.org/10.1145/2330601.2330612

Petri, P. S. (2021). Study success – a multilayer concept put under the microscope. *Zeitschrift für Hochschulentwicklung*, *16*(4), 59–78. https://doi.org/10.3217/zfhe-16-04/04

Rasch, B., Friese, M., Hofmann, W., & Naumann, E. (2021). *Quantitative Methoden 1: Einführung in die Statistik für Psychologie, Sozial- & Erziehungswissenschaften* (5. Aufl.). Springer Berlin Heidelberg Imprint: Springer. https://doi.org/10.1007/978-3-662-63282-6

R Core Team. (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric regression*. Cambridge university press. https://doi.org/10.1017/CBO9780511755453

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning*, *5*(2), 197–227. https://doi.org/10.1007/BF00116037

Schell, U. (2022). *Maschinelles lernen mit r: Daten aufbereiten und verarbeiten mit h2o und keras*. Carl Hanser Verlag. https://doi.org/10.3139/9783446472440.fm

Schmidt, S. (2023). Künstliche Intelligenz und Qualitätsmanagement an Hochschulen. *Qualität in der Wissenschaft – QiW*, *17*(3), 99–101.

Sörlin, S. (2007). Funding diversity: Performance-based funding regimes as drivers of differentiation in higher education systems. *Higher Education Policy*, *20*, 413–440. https://doi.org/10.1057/palgrave.hep.8300165

Spörk, J., Ledermüller, K., Krikawa, R., Wurzer, G., & Tauböck, S. (2021). Analyse von Studierbarkeit mittels Prognose- und Simulationsmodellen. *Zeitschrift für Hochschulentwicklung – ZFHE*, *16*(4), 163–182. https://doi.org/10.3217/zfhe-16-04/09

Stoetzer, M.-W. (2020). *Regressionsanalyse in der empirischen Wirtschafts- und Sozialforschung Band 2*. Springer. https://doi.org/10.1007/978-3-662-61438-9

Van Barneveld, A., Arnold, K. E., & Campbell, J. P. (2012). Analytics in higher education: Establishing a common language. *EDUCAUSE Learning Initiative*, *1*(1), l–ll. https://library.educause.edu/resources/2012/1/analytics-in-higher-education-establishing-a-common-language. https://doi.org/10.1515/cercles-2011-0001

Wegner, A., Thiedig, C., Janson, K., & Krempkow, R. (2023). *Von der Evidenz zum Impact? – Ein systematischer Überblick zu Gelingensbedingungen der Nutzung von Evidenz im Hochschul- und Forschungssektor*. Jahrestagung der DeGEval 2023: Gesellschaft für Evaluation e. V. https://www.degeval.org/veranstaltungen/jahrestagungen/magdeburg-2023/dokumentation/

Wiarda, J.-M. (2024). *Hamburgs Bildungssenator Ties Rabe tritt zurück*. Wissenschaftsblog: Bildung und Politik. www.jmwiarda.de/blog/bildung-und-politik/

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*. chapman; hall/CRC. https://doi.org/10.1201/9781315370279

Dortmunder Symposium der Empirischen Bildungsforschung

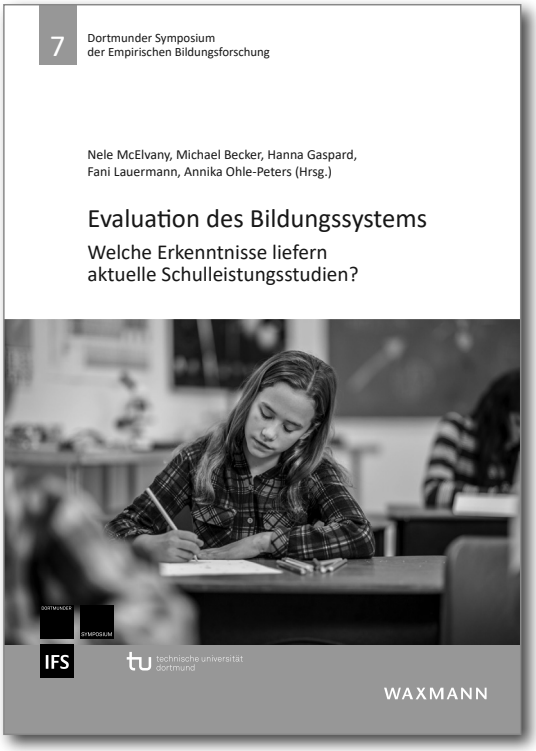
BAND 7

Nele McElvany, Michael Becker, Hanna Gaspard, Fani Lauermann, Annika Ohle-Peters (Hrsg.)

Evaluation des Bildungssystems

Welche Erkenntnisse liefern aktuelle Schulleistungsstudien?

2024, 122 Seiten, br., 29,90 €,
ISBN 978-3-8309-4834-6
E-Book: 26,99 €,
ISBN 978-3-8309-9834-1



Der siebte Band der Herausgeberreihe „Dortmunder Symposium der Empirischen Bildungsforschung“ widmet sich (internationalen) Schulleistungsstudien und wie deren Ergebnisse für die Optimierung von Bildungssystemen nutzbar gemacht werden können. Die Erfassung von Leistung und motivationalen Orientierungen Lernender sowie von institutionellen Rahmenbedingungen schulischen Lehrens und Lernens liefert reichhaltige Daten, die insbesondere (internationale) Vergleiche zwischen Bildungssystemen und Trendanalysen ermöglichen. Die sich daraus ergebenden Chancen für die Weiterentwicklung von Bildungssystemen werden in diesem Band von verschiedenen Disziplinen der Empirischen Bildungsforschung beleuchtet und diskutiert.

WAXMANN
www.waxmann.com
info@waxmann.com