Methoden zur Daten-Synthetisierung

Bachelorthesis der Fakultät für Mathematik, Informatik und Statistik

Institut für Statistik

der

Ludwig-Maximilians-Universität München

Eleftheria Papavasiliou

betreut von

Prof. Dr. Christian Heumann

Oktober 2020

Zusammenfassung

Das Sammeln und Veröffentlichen von Daten war noch nie so einfach wie jetzt. Damit einhergehend werden jedoch zunehmend Verletzungen der Privatsphäreregelungen gefördert. Ein neben herkömmlichen Anonymisierungsmethoden alternativer Weg, der immer mehr Zuspruch erlangt, ist die Datensynthetisierung. Hierbei werden statistische Modelle gebaut, die sensible Werte in den Daten mithilfe der anderen Werte schätzen sollen. Ein in dieser Arbeit vorgestelltes Beispiel für ein solches statistisches Modell bilden Entscheidungsbäume. Dieses und weitere Verfahren wurden an drei verschiedenen Datenbeispielen, die sich hauptsächlich an der Anzahl ihrer Variablen unterscheiden, hinsichtlich ihres Nutzens, ihres Reidentifikationsrisikos und ihrer Güte bei der logistischen Regressionsmodellierung getestet. Dabei konnten nahezu alle Verfahren zufriedenstellende Ergebnisse liefern und aufzeigen, dass Anonymisierung auch anders geht. Es konnte festgestellt werden, dass die Synthetisierung mit zunehmender Variablenanzahl erschwert wird und lineare Modellierungsansätze eher ungeeignet sind.

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

1 Einführung

Statistische Ämter und andere Institutionen sammeln immer größere Mengen an Daten für ihre Analysen aber auch um sie der Öffentlichkeit zur Verfügung zu stellen. Nie zuvor war das Thema *Big Data* und die damit einhergehenden Möglichkeiten aber auch Gefahren von so großer Relevanz wie heute. Ein breiter Zugang zu solchen Daten hat große Vorteile und führt beispielsweise zu Fortschritten in der Forschung und Verbesserungen bei der Politikgestaltung. Aber auch die Möglichkeit für Studenten, Datenanalysefähigkeiten zu erlernen und ihre Gesellschaft durch die gegebenen Daten besser verstehen und an ihr teilhaben zu können, sind positive Nebeneffekte bei öffentlich zugänglichen Daten.

Regierungsbehörden stehen jedoch zunehmend unter Druck, den Zugang zu Daten zu beschränken, da ihre Vertraulichkeit bedroht ist. Der uneingeschränkte Zugriff auf diese Daten ist deswegen oftmals nur für ausgewählte Mitarbeiter der jeweiligen Institution erlaubt. Es hat sich gezeigt, dass selbst das Entfernen offensichtlicher Kennungen wie Namen oder Adressen nicht ausreicht, um Anonymität gewährleisten zu können. Auch klassische Störungsmethoden wie Aggregation, Rekodierung, Austausch von Datensätzen, Unterdrückung sensibler Werte oder Hinzufügen von zufälligem Rauschen können die Identifizierung einzelner Personen nicht wie erwartet verhindern (15, 7).

Dies wurde bereits 1997 festgestellt, als die damalige Doktorandin des MIT Latanya Sweeney, die Krankenakten des Gouverneurs William Weld von Massachusetts fand, der während eines öffentlichen Auftritts zusammengebrochen war. Sie verwendete die verfügbare Postleitzahl und das Geburtsdatum von Weld, um die Datenbank der Massachusetts Group Insurance Commission (GIC), die anonymisiert gewesen sein sollte, nach seinen Unterlagen zu durchsuchen, und bestätigte die Identität anhand von Wählerregistrierungsunterlagen aus Cambridge (2). Weiter noch zeigt Sweeney in ihrer Arbeit (19), dass 97% der Wähler eindeutig identifiziert werden können, wenn Informationen über das Geburtsdatum und die Postleitzahl vorhanden sind. Personen, die solche Rückschlüsse aus den Daten ziehen, werden *Eindringlinge* bzw. *Intruders* genannt. Sie kombinieren Informationen mehrerer frei zugänglicher Datensätze, indem sie gemeinsame *key variables* (*Schlüsselvariablen*) wie das Geburtsdatum abgleichen und können dadurch einzelne Individuen identifizieren (7).

Ziel der Datenanonymisierung ist es also dieses Reidentifikationsrisiko zu senken, während der Nutzen der Daten, wie beispielsweise Korrelationen zwischen den einzelnen Variablen, nicht verloren gehen soll (5, S.59 ff.).

Ein immer mehr an Beliebtheit erlangendes Vorgehen, das die oben genannten klassischen Anonymisierungsmethoden langsam ablöst, ist die Datensynthetisierung. Die daraus resultierenden Daten werden neben *synthetischen Daten* auch *Surrogatdaten* oder *artifizielle Daten* genannt (6). Hierbei handelt es sich um ein statistisches Verfahren, das auf der Idee der multi-

plen Imputation beruht, welche fehlende Werte in Datensätzen durch plausible Werte ersetzt. Dies wird durch die Konstruktion von Modellen ermöglicht. Rubin hatte bereits 1993 die Idee, sensible Werte durch dieses Verfahren zu ersetzen (14).
Nach dem Erhalt dieser Surrogatdaten können artifizielle Datensätze an die Öffentlichkeit weitergegeben werden, mit denen man statistische Analysen wie beispielsweise Regressionen wie gewohnt durchführen kann, die zu den selben Ergebnissen führen sollten, wie Analysen auf Basis der Originaldaten.
Diese These zu untersuchen wird Ziel dieser Arbeit sein. Dafür werden zunächst die grundlegenden theoretischen Hintergründe hinter der Methodik der Datensynthese erläutert. Anschließend sollen drei unterschiedliche Datensätze auf verschiedene Weisen synthetisiert werden und hinsichtlich ihres Reidentifikationsrisikos und ihres Nutzens verglichen werden. Zuletzt werden die artifiziellen Datensätze genutzt, um ein logistisches Regressionsmodell zu trainieren. Die Güte des Modells soll dann mit der auf den Originaldaten basierenden verglichen werden.

2 Datensynthese

Bei der Datensynthese handelt es sich um ein Verfahren, mit dem eine künstliche Repräsentation eines Originaldatensatzes erstellt werden kann. Der neue Datensatz enthält also keine realen Personen mehr. Die Grundidee besteht darin, die beobachteten Werte durch Stichproben aus geeigneten Wahrscheinlichkeitsverteilungen zu ersetzen, damit die wesentlichen statistischen Eigenschaften der Originaldaten erhalten bleiben.
Dafür werden Modelle erstellt, die die Originaldaten so gut wie möglich erklären. Aus diesen Modellen werden neue Daten generiert, die die wichtigsten statistischen Eigenschaften des Originaldatensatzes enthalten. Diese Modelle können dabei parametrischer Natur sein, beispielsweise Regressionsmodelle aber auch aus dem Machine-Learning stammen, was sich in letzter Zeit sehr bewährt hat. Grund hierfür ist vor allem, dass bei letzteren Verfahren Muster in den Datensätzen automatisch erkannt werden ohne dass Verteilungsannahmen getroffen werden oder konkrete Fragestellungen vorliegen müssen. Insbesondere können hier auch nicht-lineare Zusammenhänge anders als bei einfachen Regressionsmodellen leicht aufgedeckt werden. Dadurch wird auch der Modellierungsaufwand deutlich verringert (6).
Ein Beispiel für ein nicht parametrisches Verfahren, das in dieser Arbeit eine tragende Rolle spielen wird, sind Klassifikations- und Regressionsbäume, genannt *CART*, die in Kapitel 3 vorgestellt werden.
Benutzt man zur Modellierung einen parametrischen Ansatz, schätzt man eine Variable X_1, indem man eine Regression fittet, welche X_1 als Zielvariable Y_1 behandelt und alle anderen Variablen $X_2, \ldots, X_n$ als Prädiktorvariablen X_i. Dieses Vorgehen wiederholt man für jede Variable, die synthetisiert werden soll. Dabei sollte beachtet werden, dass auch die Prädiktorvariablen X_i in den nachfolgenden Regressionen durch die bereits ersetzten Werte Y_i geupdatet werden. Möchte man beispielhaft die vierte Variable eines Datensatzes ersetzen und hat die drei ersten bereits ersetzt, modelliert man eine Regression, die X_4 als Zielvariable behandelt. Die Prädiktorvariablen setzen sich dann aus den Variablen Y_1, Y_2, Y_3, die bereits ersetzt wurden und den Variablen $X_5, \ldots, X_n$, die nachfolgend noch ersetzt werden müssen, zusammen (7, 15).
Natürlich kann bei diesem Verfahren einfach der Reihe nach vorgegangen werden. Allerdings ist es sinnvoll sich im Vornherein Gedanken darüber zu machen, in welcher Reihenfolge die Variablen synthetisiert werden sollten, um einen möglichst großen Nutzen mit einem gleichzeitig geringem Reidentifikationsrisiko zu garantieren. Eine mathematisch fundierte Theorie dazu ist jedoch nicht vorhanden. Ein Ansatz bestünde darin verschiedene Reihenfolgen auszuprobieren und zu beobachten, welche die zufriedenstellensten Ergebnisse liefert. Eine andere Möglichkeit wäre die Variablen nach ihrer Menge an sensiblen Werten zu ordnen und diejenigen Variablen als letztes zu synthetisieren, die die meisten sensiblen Werte enthalten, da das

Reidentifikationsrisiko umso geringer ist je später im Prozess die Variablen ersetzt wurden. Alternativ kann man die Variablen auch so ordnen, dass die Laufzeit des Algorithmus minimiert wird. Die Wahl der Reihenfolge ist also sehr subjektiv und variiert je nach Situation (7).

Benutzt man ein nicht parametrisches Verfahren funktioniert das Vorgehen analog. Allerdings wird dann anstatt einer Regression für die einzelnen Variablen beispielsweise ein Entscheidungsbaum erstellt.

3 Entscheidungsbäume

Eine Möglichkeit, um die einzelnen Variablen vorhersagen zu können, sind Entscheidungsbäume. Hierbei betrachtet man beispielsweise im binären Fall die Fragestellung, ob ein Objekt Klasse 1 oder Klasse 2 zugeordnet werden soll. Dafür stellt man nacheinander Fragen, die mit *Ja* oder *Nein* beantwortet werden können. Es handelt sich also anders als bei anderen Verfahren um einen sequenziellen Entscheidungsfindungsprozess.
Auf Basis welchen Merkmals die Fragen jeweils gestellt werden, hängt von der Antwort der vorherigen Frage ab. Ziel ist es immer dasjenige Merkmal zu wählen, das im nächsten Schritt das beste Ergebnis hinsichtlich einer vorher definierten Metrik liefert. Durch das Teilen will man also erreichen, Knoten zu erhalten, die in sich möglichst homogen und untereinander möglichst heterogen in Bezug auf die Zielvariable sind.
Das Ergebnis stellt dann im kategoriellen Fall ein sogenannter Klassifikationsbaum dar, wohingegen bei metrischen Zielvariablen von Regressionsbäumen die Rede ist.

3.1 Klassifikationsbäume

Ein Baum wird aus Trainingsdaten erzeugt, für die die Werte der Zielvariable bereits bekannt sind. Die Trainingsdaten mit p Einflussgrößen bestehen aus N Observationen (x_i, y_i), für $i = 1, 2, \ldots, N$, mit $x_i = (x_{i1}, \ldots, x_{ip})$. Zunächst wird davon ausgegangen, dass alle Variablen binär sind.
Der Baum besteht aus Ästen und Knoten, welche wiederum in Entscheidungsknoten, Endknoten und einem Wurzelknoten unterteilt werden können. Letzterer ist dabei der oberste Knoten, in welchem die gesamte Stichprobenziehung enthalten ist.
Die Wahrscheinlichkeit, dass ein Objekt im Wurzelknoten zur Klasse 1 gehört, wird mit p_{t_1} bzw. p_t für einen beliebig anderen Knoten bezeichnet. Ist diese Wahrscheinlichkeit größer als 0.5 ordnet man das Objekt der Klasse 1 zu, ist sie kleiner der Klasse 2. Eine zufällige Zuordnung ist dann unumgänglich, wenn $p_t = 0.5$.
Besonders am Wurzelknoten ist, dass hier p_{t_1} der relativen Häufigkeit der Klasse 1 in der Stichprobe entspricht. Würde man also nur den Wurzelknoten für die Entscheidungsfindung benutzen ohne andere Merkmale mit einzubeziehen, würde man alle Objekte der Klasse zuordnen, welche in der Stichprobe am häufigsten auftritt, weil man dadurch den kleinsten Fehler machen würde. Dies ist natürlich nicht sinnvoll, vor allem dann nicht, wenn die zwei Klassen in der Stichprobe annähernd gleich oft repräsentiert sind. Deswegen wird der Baum nach dem Wurzelknoten weitergeführt, indem man ihn in zwei disjunkte Teilmengen teilt, wodurch man zu den Entscheidungsknoten gelangt.
Hier wird erstmals eine *Ja-Nein*-Frage gestellt. In der Regel geht man bei Beantwortung der Frage mit *Ja* im linken Ast des Baumes zum nächsten Knoten, während man bei einem *Nein*

den rechten Ast wählt. Dieses Prozedere führt man so lange weiter bis man ein Stoppkriterium erreicht und damit am Endknoten angelangt ist (9, S.391-398). Im Endknoten befinden sich dann idealerweise nur Objekte einer Klasse, sodass man neue Observationen, die diesen Weg im Baum gehen, dieser Klasse mit einer Fehlerwahrscheinlichkeit von 0 zuordnen kann. In der Praxis führt man die Bäume jedoch nicht so weit, dass am Ende alle Objekte richtig zugeordnet werden. Grund hierfür ist, dass es sich beim Trainingsdatensatz nicht um die gesamte Population handelt und die Struktur des Baumes stark von der gewählten Stichprobe abhängt, was dazu führen kann, dass neue Observationen unter Umständen bei zu großen Bäumen falsch klassifiziert werden (9, S.391-398). Dieses Problem ist als *overfitting* bekannt. Andererseits darf der Baum auch nicht zu klein sein. Verfahren, mit denen man die Größe des Baumes bestimmen kann werden später betrachtet.
Beinhaltet ein Endknoten also nicht nur Objekte einer Klasse, ordnet man neue Observationen, die an diesen Endknoten gelangen, normalerweise der Klasse zu, die im Endknoten am häufigsten vorkommt. Die Entscheidungsregel entspricht im Endknoten also der selben wie im Wurzelknoten bzw. der anderen Entscheidungsknoten.
Bei der Synthetisierung von Daten mithilfe von Klassifikationsbäumen wird hingegen ein etwas anderes Verfahren vorgeschlagen, um zeitgleich zum größten Datennutzen eine niedrige Reidentifikationswahrscheinlichkeit gewährleisten zu können. In diesem Verfahren wählt man zufällig einen der Werte aus dem Endknoten als Vorhersagewert. Beispielsweise kann dies durch bayesianischen Bootstrap erzielt werden. Aus diesem Grund ist es vor allem bei der Datensynthetisierung sehr wichtig, dass im Endknoten genügend Beobachtungen enthalten sind, um einen validen Zufallsprozess garantieren zu können (7).

3.1.1 Die richtige Merkmalswahl im Entscheidungsknoten

Eine zentrale Frage bei der Konstruktion eines Klassifikationbaums ist, welche Variable in welchem Entscheidungsknoten wie oft benutzt werden sollen. Für die Beantwortung dieser Frage gibt es mehrere Ansätze. Der erste Ansatz, der hier vorgestellt werden soll, beruht auf der Arbeit von Breiman et al. (3), der die Anwendung von Klassifikationsbäumen erstmals populär gemacht hat. Die Idee dahinter ist, dass man die Unsicherheit der Entscheidung für oder gegen Klasse 1 als Unreinheitsmaß verwendet und dann dasjenige Merkmal für den Entscheidungsknoten wählt, das die größte Verminderung des Unreinheitsmaßes liefern kann. Um die Unsicherheit der Entscheidung berechnen zu können, muss man zunächst eine Zufallsvariable Y definieren

$$Y = \begin{cases} 1, & \text{falls die Observation zur Klasse 1 gehört} \\ 0, & \text{falls die Observation zur Klasse 2 gehört} \end{cases}$$

und ihr eine Verteilungsannahme unterstellen. Es ist naheliegend, dass Y Bernoulli-verteilt ist, wobei der Parameter p_t entspricht. Es ist bekannt, dass sich die Varianz oder Unsicherheit

einer bernoulli-verteilten Zufallsgröße als Produkt der Erfolgswahrscheinlichkeit p_t und ihrer Gegenwahrscheinlichkeit $1 - p_t$ ergibt, die hier nicht bekannt sind, sondern geschätzt werden. Da in diesem Fall wie bereits erwähnt das Unreinheitsmaß einfach der Unsicherheit entspricht, ist dieses gegeben durch:

$$i(t) = p_t(1 - p_t)$$

Für die Wahl des richtigen Merkmals kann man nun für alle Variablen jeweils das Unreinheitsmaß berechnen und dann das Merkmal wählen, das dieses Maß am meisten verringert. Wichtig ist hierbei zu verstehen, dass das richtige Merkmal nicht unbedingt perfekt hinsichtlich der Unreinheit des gesamten Baumes ist, sondern lediglich optimal für genau diesen Split. Weiterführend kann man auch überprüfen, ob die Verminderung beim Teilen eines Knotens in zwei weitere (rechts und links), so groß ist, dass sie das Splitten des Knotens überhaupt rechtfertigt oder der Knoten einen Endknoten bildet. Die Verringerung kann wie folgt berechnet werden:

$$\Delta(s, t) = i(t) - p_{t_L} i(t_L) - p_{t_R} i(t_R)$$

Dabei steht s für die Regel, die im Knoten verwendet wurde, um die Stichprobe in zwei weitere Knoten zu teilen und t_L bzw. t_R jeweils für den linken bzw. rechten Knoten, die sich durch Splitten von t ergeben haben. Die Unreinheitsmaße für die Unterknoten t_L und t_R werden mit den Anteilen p_{t_L} bzw. p_{t_R} der in ihnen enthaltenen Observationen gewichtet.
Eine andere Möglichkeit das Unreinheitsmaß zu definieren stellt die Entropie

$$i(t) = -p_t ln(p_t) - (1 - p_t) ln(1 - p_t)$$

dar. Des Weiteren gibt es noch die Möglichkeit die Unreinheit durch die Devianz zu messen. Diese wird definiert durch:

$$i(t) = -2 \left[n_{1t} \, ln(p_t) \; + \; (n - n_{1t}) \, ln(1 - p_t) \right]$$

Dabei gibt n_{1t} die Anzahl der Observationen im t-ten Knoten an, die zu Klasse 1 gehören. Analog zum ersten Verfahren, wird immer dasjenige Merkmal gewählt, welches die größte Verminderung des Unreinheitsmaßes (hier: Entropie bzw. Devianz) liefert (S.391-398 9, 17, S.373-384).

3.1.2 Klassifikationsbäume bei nicht-binären Merkmalen

In der Praxis sind nicht nur Datensätze mit ausschließlich binären Variablen von Interesse, sondern auch solche die kategorielle Variablen mit mehr als zwei Ausprägungen oder metrische Merkmale beinhalten. Schauen wir uns zunächst den mehrkategoriellen Fall an. Hier kann man wieder zwei Szenarien unterscheiden. Entweder ist die Zielvariable, die wir vorhersagen

wollen mehrkategoriell oder es liegt eine binäre Zielvariable vor, aber die Merkmale, die zur Entscheidungsfindung mit einbezogen werden, haben mehr als zwei Kategorien.
Die Problematik des letzteren Szenarios ist leicht zu lösen. Man behandelt die einzelnen Kategorien einer nominalen Variable quasi als eigene Variablen. Wenn man also eine *Ja-Nein*-Frage stellt, heißt *Ja*, dass die Observation in dieser Kategorie ist und *Nein*, dass sie in einer der anderen *k-1* Kategorien einzuordnen ist. Natürlich kann man auch zwei oder mehrere Ausprägungen der k möglichen Ausprägungen zu einer Variable zusammenfassen.
Das heißt, dass man bei der Überlegung welches Merkmal zur Entscheidungsfindung herangezogen werden soll, alle Kombinationen der k Ausprägungen betrachten muss und dann analog zum binären Fall, das Merkmal bzw. die Kombination von Ausprägungen wählt, die die größte Verminderung des Unreinheitsmaßes bieten kann. Sehr zeitaufwendig ist dieses Vorgehen, wenn ein Merkmal viele Ausprägungen aufweist und damit einhergehend es sehr viele verschiedene Kombinationen gibt. Ein möglicher Lösungsansatz ist, in jedem Knoten für jede mögliche Ausprägung a_k, $k = 1, \ldots, K$ der Variablen x_i die Häufigkeiten über alle Observationen der jeweiligen Zielvariablen zu berechnen, die die Ausprägung a_k annehmen. Die Häufigkeiten der einzelnen Ausprägungen können dann der Größe nach geordnet werden, sodass sich die mehrkategorielle Variable zu einer ordinalen umformen lässt (17). Dadurch müssen viel weniger Kombinationen betrachtet werden. Doch dazu später mehr.
Wenn das Zielmerkmal mehrkategoriell ist, ordnet man eine Observation derjenigen Klasse zu, die im Endknoten am häufigsten vertreten ist. Beim speziellen Fall der Datensynthese zieht man wieder analog zum binären Fall zufällig aus dem Endknoten.
Dabei ändert sich das Unreinheitsmaß, basierend auf der Varianz zum Gini-Index:

$$i(t) = \sum_{i=1}^{k} p_{it} \left(1 - p_{it}\right) = 1 - \sum_{i=1}^{k} p_{it}^2$$

Wobei p_{it} die Wahrscheinlichkeit darstellt, dass eine Observation im Knoten t zur i-ten Klasse gehört. Dieses Unreinheitsmaß wird auch in dem in R implementierten *CART*-Algorithmus benutzt und ist somit die Basis für die in dieser Arbeit vorgestellten Ergebnisse. Die Formel für die Entropie wird zu:

$$i(t) = -\sum_{i=1}^{k} p_{it}\, ln(p_{it})\,.$$

Beinhaltet der zu untersuchende Datensatz metrische Variablen, muss man einen Schwellenwert $c \in \mathbb{R}$ festlegen, anhand dessen man die metrische Variable dichotomisiert. Sodass sie die Form $X \leq c$ und $X > c$ erlangt. Das Vorgehen ist nach dieser Umformung dann identisch wie bisher (S.391-398 9, 17, S.373-384).
Abbildung 1 zeigt ein Beispiel für einen Entscheidungsbaum. Dafür wird der in 7 vorgestellte Einkommensdatensatz genutzt. Er enthält unter anderem die Variable *sex*, welche das

Geschlecht der befragten Person angibt. Dieses Merkmal bildet im Entscheidungsbaum die Zielvariable.

Man kann erkennen, dass im ersten Schritt diejenige Variable zur Entscheidungsfindung gewählt wird, die den Beziehungsstatus der Person angibt. Dies ist eine mehrkategorielle Variable. Wie oben erklärt wird hier die Ausprägung *verheiratet* (*Mrr*) als eine Variable betrachtet und die restlichen drei Ausprägungen bilden den Gegenpol zu dieser Variable. Der rechte Ast beantwortet die Frage, ob die befragte Person verheiratet ist oder nicht mit *Ja* und der linke Ast mit *Nein*.

Anschließend sieht man, dass der rechte Ast nicht weiter unterteilt wird und somit einen Endknoten bildet. Von den 305 Personen, die in diesem Endknoten sind und laut Entscheidungsregel als Männer identifiziert werden, sind tatsächlich 272 Personen männlich.

Der linke Ast wird noch weiter unterteilt. Das Vorgehen ist dabei identisch wie beim ersten Split. Man sieht, dass die Frage nach dem Beziehungsstatus erneut auftritt, was auch zeigt, dass eine Variable nicht nur einmal zur Entscheidungsfindung herangezogen werden kann. Auch metrische Variablen treten auf. Hier kann man gut die Dichotomisierung erkennen, beispielsweise bei der Variablen *age*. Personen über 33 Jahren landen im linken Ast, Personen, die jünger sind im rechten.

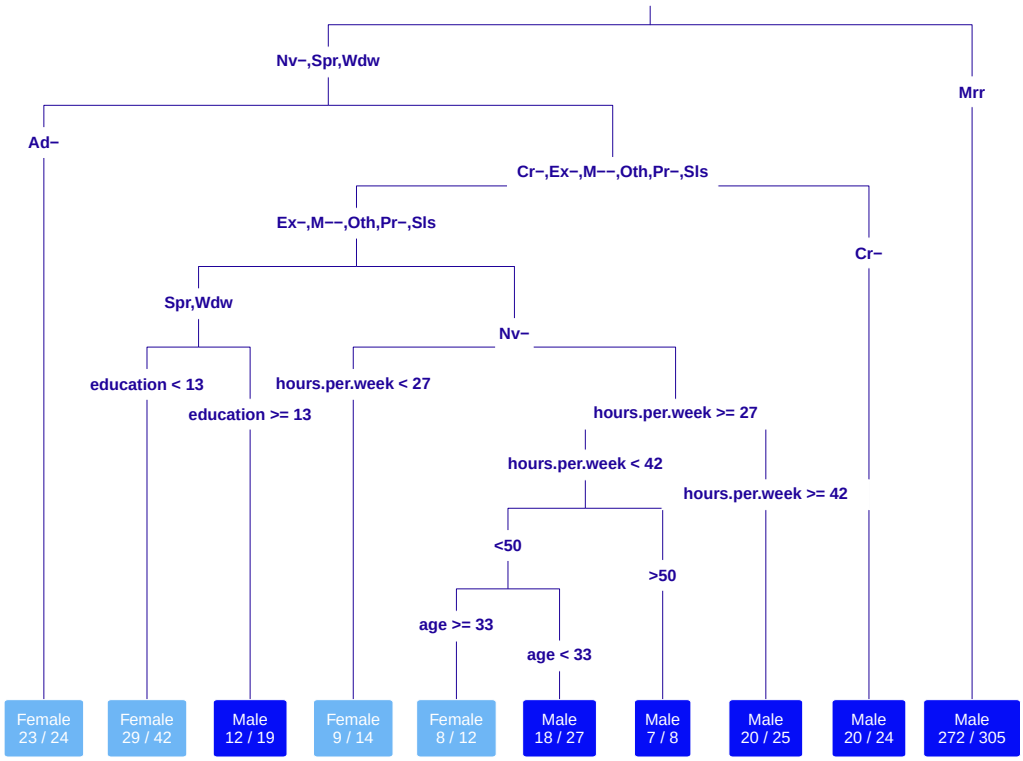


Abbildung 1: Beispiel für einen Entscheidungsbaum für den Einkommensdatensatz mit der Zielvariable Geschlecht

3.1.3 Die richtige Baumgröße

Um die richtige Größe des Baumes festzulegen, wird zuerst ein Maximalbaum konstruiert und anschließend wieder gekürzt. Wie weit er gekürzt wird, entscheidet eine geeignete Metrik, die beispielsweise so

$$min\,(R(\tilde{T})\,+\,\alpha\,|\tilde{T}|) \tag{1}$$

aussehen kann. $R(T)$ wird die Fehlklassifikationsrate genannt und gibt die Chance an, dass eine Observation fälschlicherweise in einen Endknoten i gelangt.

$$R(T) = (1 - \max_{i=1,\ldots,k} p_{it})p_{\tilde{t}} \tag{2}$$

$p_{\tilde{t}}$ ist dabei die Wahrscheinlichkeit, dass eine Observation im Endknoten t landet. $R(\tilde{T})$ ist folglich die gesamte Fehlklassifikation des Baumes, die sich durch Summierung der einzelnen Fehlklassifikationsraten ergibt. Diese soll minimiert werden, was mit steigender Größe des Baumes automatisch geschieht. Deswegen bestraft der zweite Summand mit einer geeigneten Gewichtungskonstanten α die Größe des Baumes, die dem Problem des *Overfittings* entgegenwirkt. $|\tilde{T}|$ gibt dabei die Anzahl der Endknoten an. Kleine Werte von α bestrafen die Größe des Baumes gering, während bei größeren α Werten Bäume mit wenigen Endknoten favorisiert werden. Um den perfekten Wert von α zu finden, kann beispielsweise ein Kreuzvalidierungsprozess benutzt werden.
Alternativ kann auch die Devianz zur Bestimmung der Baumgröße herangezogen werden. Hierbei lautet die Entscheidungsregel, dass ein Knoten immer dann einen Endknoten darstellt, wenn seine Devianz kleiner als 1% der Devianz des Wurzelknotens ist. Eine noch einfachere Möglichkeit wäre eine minimale Anzahl an Beobachtungen zu bestimmen, die in einem Endknoten enthalten sein darf, sodass alle Knoten, die weniger als diese Anzahl enthalten, gestrichen werden. Breiman et al. schlagen beispielsweise eine Grenze von 5 vor (S.391-398 9, 17, S.373-384). Letztere Methode ist bei der Datensynthese mithilfe von Entscheidungsbäumen sehr beliebt, da wie bereits angemerkt damit der Zufallsprozess gut kontrolliert werden kann. Zudem wurde gezeigt, dass das Problem des *Overfittings* bei der Datensynthese keine tragende Rolle spielt.

3.2 Regressionsbäume

Regressionsbäume verhalten sich sehr ähnlich zu Klassifikationsbäumen. Das Vorgehen kann hier weitgehendst beibehalten werden. Unterschiede ergeben sich bei der Bestimmung des Unreinheitsmaßes $i(t)$ und der daraus resultierenden Berechnung des Zielvariablenwertes. Anders als beim Klassifikationsproblem minimiert man hier nicht die Varianz, sondern die Fehlerquadratsumme

$$\sum_{i=1}^{N}(y_i - \hat{f}(x_i))^2 \tag{3}$$

mit

$$\hat{f}(x_i) = \sum_{e=1}^{E} c_e \mathbb{1}(x_i \in R_e) \tag{4}$$

E gibt dabei die Anzahl der Endknoten an und R_e den Raum des jeweiligen Endknotens. c_e stellt den Prädiktionswert für den jeweiligen Endknoten dar, den es geeignet zu schätzen gilt. Es lässt sich zeigen, dass 3 genau dann minimal wird, wenn c_e dem Mittelwert über alle Beobachtungen in diesem Endknoten entspricht (10, S.307 ff.). Folglich sagt man die Zielvariabe vorher, indem man über die Werte des jeweiligen Endknoten mittelt. Die Verringerung des Unreinheitsmaßes lässt sich dann wie folgt berechnen:

$$\Delta(s,t) = i(t) - i(t_L) - i(t_R)$$

Anders als bei Klassifikationsbäumen werden $i(t_L)$ bzw. $i(t_R)$ nicht gewichtet. Für die Trennung von einem Knoten in zwei Unterknoten muss also ein Split gefunden werden, der

$$\sum_{x_i \in R_{t_L}} (y_i - c_{t_L})^2 + \sum_{x_i \in R_{t_R}} (y_i - c_{t_R})^2$$

minimiert.

Das Kosten-Komplexitätskriterium zur Bestimmung der Baumgröße ändert sich zu (10, S.307 ff.):

$$min \sum_{e=1}^{|T|} \sum_{x_i \in R_e} (y_i - c_e)^2 \, + \, \alpha \left|T\right| .$$

Die Güte des Modells kann durch den MSE bestimmt werden:

$$\frac{1}{N} \sum_{i=1}^{N}(y_i - \hat{f}(x_i))^2 \tag{5}$$

Wie auch bei den Klassifikationsbäumen weicht die Vorhersage für das Synthetisieren vom klassischen Vorgehen ab. Auch bei Regressionsbäumen werden zufällig Werte aus den Endknoten gezogen. Allerdings ist dies bei der Benutzung von metrischen Variablen problemati-

scher als bei kategoriellen. Deswegen kann man, um eine noch höhere Sicherheit zu erreichen, mithilfe eines Gaußschen Kerndichteschätzers eine Dichte aus den Daten des Endknotens erzeugen aus der dann zufällig gezogen wird (7). Damit senkt man das Reidentifikationsrisiko deutlich.

3.3 Bewertung von Entscheidungsbäumen

Entscheidungsbäume sind in vielen Anwendungen anderen Verfahren, wie beispielsweise den Regressionsmodellen, vorzuziehen. Grund dafür ist, dass es sich bei Entscheidungsbäumen um ein nicht parametrisches Verfahren handelt. Das heißt, es müssen keine Verteilungsannahmen für die Zielvariable getroffen werden.
Außerdem können Entscheidungsbäume anders als Regressionsmodelle auch nicht-lineare Zusammenhänge einfach erkennen. Hinzu kommt, dass fehlende Werte für dieses Verfahren nicht problematisch sind und somit nicht unbedingt gelöscht oder ersetzt werden müssen.
Grund hierfür ist, dass die Prädiktoren univariat betrachtet werden. Das führt dazu, dass bei der Suche nach der richtigen *Ja-Nein*-Frage, die auf einem Merkmal basiert, dessen Wert für eine konkrete Beobachtung fehlt, diese Observation für den Entscheidungsprozess ignoriert wird. Allerdings kann sie für den nächsten Entscheidungsprozess, der auf einem anderen Merkmal basiert wieder herangezogen werden, sodass man trotz eines fehlenden Wertes immer noch einen globalen Nutzen aus dieser Beobachtung ziehen kann. Wurde bereits eine geeignete *Ja-Nein*-Frage gewählt stellt sich die Frage, welchem Ast die Beobachtung mit dem fehlenden Wert folgt. Dafür werden Ersatzsplits ausgewählt, die die Werte des für die Entscheidung herangezogenen Merkmals am besten vorhersagen. Anhand dessen kann dann die Observation mit dem fehlenden Wert dem linken oder rechten Knoten zugeordnet werden (18).
Des Weiteren können Entscheidungsbäume gut mit einer hohen Anzahl an Variablen umgehen und beinhalten einen automatischen Variablenselektionsmechanismus.
Zuletzt muss noch angeführt werden, dass einzelne Bäume eine anschauliche Interpretation zulassen, in der man genau beobachten kann, wie der Entscheidungsprozess erzeugt wurde. Dies wurde mit Abbildung 1 deutlich.
Wie bei jeder statistischen Methode weisen auch Entscheidungsbäume Nachteile auf, die nicht ignoriert werden sollten. So werden beispielsweise metrische Variablen vom Algorithmus den kategoriellen vorgezogen. Grund hierfür ist, dass eine metrische oder zumindest geordnete Variable mit k verschiedenen Werten $k-1$ Möglichkeiten bietet sie in der Form $X \leq c$ aufzuteilen. Eine nominale Variable dagegen hat für die selbe Anzahl an Werten k, die in diesem Fall natürlich ungeordnet sind, $2^{(k-1)} - 1$ mögliche Teilungen. Das führt dazu, dass in der Regel metrische Merkmale, eine höhere Chance haben ausgewählt zu werden (13).
Ein weiterer Nachteil ist, dass durch das Nutzen von Zufallsstichproben meistens instabile Bäume erzeugt werden. Ändert sich die Stichprobe nur leicht, kann dies schon zu komplett

anderen Ergebnissen bei der Entscheidungsfindung führen.Es gibt jedoch eine Reihe von Ansätzen, die dieses Problem beheben können. Ein Weg dafür ist *Bagging*. Die Idee dahinter ist, mehrere verschiedene Bäume zu erzeugen und dann ein Objekt der Klasse zuzuordnen, welche bei den meisten Bäumen gewählt worden ist. Dafür werden aus der gegebenen Zufallsstichprobe mehrere Bootstrap-Stichproben gezogen auf denen dann das Verfahren durchgeführt werden kann (17, S.373-384).

4 Random Forest

Eine andere Möglichkeit, um das Problem der Instabilität bei einzelnen Entscheidungsbäumen zu umgehen, stellt das Verfahren des *Random Forest* dar. Wie beim *Bagging* werden auch hier mehrere verschiedene Bäume erzeugt und deren Ergebnisse dann gemittelt. Die Diversität dieser Bäume ist dabei noch höher, denn es werden nicht nur verschiedene *Bootstrap* Stichproben für die Prädiktion benutzt, sondern auch unterschiedliche Unterräume des Merkmalsraums. Das heißt, dass bei jedem Split eine zufällig ausgewählte Menge an Prädiktionsvariablen gewählt wird, aus der dann das für diese *Ja-Nein*-Frage geeignetste Merkmal gewählt wird. Dadurch haben *schwächere* Variablen, die bei Betrachtung des gesamten Merkmalraums durch *stärkere* Merkmale verdrängt worden wären, auch die Möglichkeit in den Entscheidungsprozess einzufließen und bisher unentdeckte Zusammenhänge aufzudecken.
Das führt dazu, dass auf lokaler Ebene suboptimale Splits durchgeführt werden, die aber helfen können das globale Ergebnis des Baumensembles zu verbessern. Grund hierfür ist, dass durch die sequenzielle Vorgehensweise der optimale Split bedingt auf alle bisherigen Splits gewählt wird aber unabhängig von den noch bevorstehenden.
Unter anderem deswegen liefert *Random Forest* im Allgemeinen bessere Ergebnisse als *Bagging*. Ein anderer Grund für die bessere Performance ist, dass die obere Grenze für den Generalisierungsfehler eines Baumensembles von der Korrelation zwischen den einzelnen Bäumen abhängt, die bei *Random Forest* aufgrund der unterschiedlichen Merkmalsunterräume natürlich kleiner ist als beim *Bagging*, welches immer den gesamten Prädiktionsraum betrachtet (18).
Beachtet werden sollte bei der Durchführung des nachfolgenden Algorithmus, dass durch das zufällige Wählen der Prädiktoren, die Ergebnisse bei mehrmaligem Ausführen voneinander abweichen und der Zufallsprozess deswegen durch einen *seed* festgelegt werden sollte.
Algorithmus:

1. Ziehe aus den Trainingsdaten eine *Bootstrap* Stichprobe vom Umfang $\frac{2}{3}N$ (7)
2. Erzeuge für diese Stichprobe einen Entscheidungsbaum T_b
   (a) Starte mit allen Beobachtungen im Wurzelknoten
   (b) Wähle zufällig $m = \sqrt{p}$ der p möglichen Variablen aus (7)
   (c) Teile den Wurzelknoten anhand der geeignesten Variable $\in m$ in zwei Unterknoten
   (d) Wiederhole Schritt (b) und (c) bis zu einer minimalen Endknotengröße
3. Wiederhole Schritt 1 für jede der $b = 1, \ldots, B$ *Bootstrap* Stichproben
4. Sage für eine neue Beobachtung x_{neu} die kategorische Zielvariable folgendermaßen vor-

her:

$$\hat{f}(x_{neu}) = arg\,\underset{k}{max} \sum_{b=1}^{B} \mathbb{1}(\hat{f}^b(x_{neu}) = k)$$

Dabei gibt k die Kategorien der Zielvariablen an und $\hat{f}^b(x_{neu})$ die Prädiktionsfunktion von Baum T_b.

$$\hat{f}^b(x_{neu}) = arg\,\underset{k}{max} \sum_{x_i \in \{R_e | x_{neu} \in R_e\}} \mathbb{1}(y_i = k)$$

mit R_e: Raum des jeweiligen Endknotens.
oder für eine metrische Variable:

$$\hat{f}(x_{neu}) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}^b(x_{neu})$$

$\hat{f}^b(x)$ wurde in 4 definiert.

5. Berechne die Fehlklassifikationsrate durch Mitteln der einzelnen Fehlklassifikationsraten $R(T_b)$ 2 bzw. bestimme das Mittel der MSEs 5 zum Beurteilen der Modellgüte

Auch beim Verfahren des *Random Forest* ändert sich die Vorhersage der Zielvariablen im Falle der Datensynthetisierung. Für kategorielle Daten wird jeder Baum bis zum Endknoten geführt und für jeden Eintrag Y_{ij} der Zielvariable Y_i der Vorhersagewert notiert. Diese Vorhersagewerte werden dann genutzt, um eine multinomiale Verteilung zu erzeugen, aus der dann der Wert der Zielvariable des *Forest* gezogen wird.
Anschaulich kann man sich das Vorgehen wie folgt vorstellen. Angenommen man hat 500 einzelne Bäume gefittet, die als Ziel haben die Herkunft einer neuen Person vorherzusagen. Von diesen 500 Bäumen sagen 300 eine europäische Herkunft vorher, 100 eine afrikanische, 50 eine asiatische und wiederum 50 sind unter sonstige einzuordnen. Dann wird die multinomiale Verteilung so gebildet, dass $p(europian) = 0,6$, $p(african) = 0,2$ und $p(asian) = p(sonstige) = 0,1$ gilt. Aus dieser wird anschließend zufällig ein Wert gezogen.
Ähnlich verhält es sich mit metrischen Variablen. Auch hier zieht man zufällig einen Wert aus den Vorhersagen der einzelnen Bäume, allerdings muss man möglicherweise wie bereits in 3.2 erläutert mithilfe einer Gaußschen Kerndichteschätzung eine Dichte erzeugen, aus der gesamplet werden kann, um das Reidentifikationsrisiko zu senken (7).
Neben dem herkömmlichen Weg der Prädiktion, aus der man anschließend auch die erwartete Fehlerrate des Baumensembles berechnen kann, gibt es noch eine weitere Möglichkeit, genannt *out-of-bag* Vorhersage. Durch das Nutzen von *bootstrap* Stichproben, werden einige Datenpunkte $z_i = (x_i, y_i)$ in einzelnen Bäumen nicht aufgenommen. Diese Observationen dienen dann als eingebauter Testdatensatz, mit dem die Prädiktionsgüte berechnet werden kann. Das führt im Vergleich zu der eher optimistischen Fehlerberechnung bei der normalen

Prädiktion zu realistischeren Einschätzungen (18).
Analog dazu auch die Berechnung des MSEs für metrische Zielgrößen. Vor allem bei der Datensynthese ist dieses Vorgehen sehr beliebt, weil dadurch auch das Reidentifikationsrisiko sinkt, da der wahre Wert bei der Vorhersage dann nicht gezogen werden kann.
Anders als bei der Konstruktion eines einfachen Baumes kürzt man die Bäume bei *Random Forest* in der Regel nicht. Durch die Mittelung über die einzelnen Bäume führt dieses Verfahren weniger zu *overfitting*. Der Nachteil bei der Anwendung von *Random Forest* ist, dass nun die leichte Interpretierbarkeit, die Entscheidungsbäume so beliebt machen, entfällt. Denn es ist nicht möglich einen „mittleren“ Baum zu fitten, der die Ergebnisse auf eine geeignete Weise zusammenfasst. Infolgedessen kann man nicht einfach ablesen, welche Variablen wann und wie oft für die Entscheidungsfindung benutzt wurden und es werden alternative Metriken gebraucht, um die Wichtigkeit eines Merkmals bestimmen zu können. Ein sehr naiver Ansatz wäre einfach zu messen, wie oft eine Variable in einem Baum verwendet wurde. Weitere komplexere Ansätze können beispielsweise in (18) nachgelesen werden, sind jedoch für diese Arbeit nicht von besonderem Interesse und werden deswegen hier nicht weiter beschrieben.

5 Logistische Regression

Wie bereits angesprochen wird nach der Synthetisierung der Datensätze ein logistisches Modell gefittet, das im Anschluss mit dem logistischen Modell basierend auf den Originaldaten verglichen werden soll. Dieser Abschnitt dient dem theoretischen Verständnis dieser Modellierung.

Der Unterschied zwischen der logistischen Regression und der einfachen, bzw. multiplen Regressionsanalyse besteht darin, dass die abhängige Variable binär ist, also nur zwei Ausprägungen aufweist. An die unabhängigen Variablen hingegen werden keine Anforderungen gestellt. Im Gegensatz zur linearen Regressionsanalyse, bei der die Kleinste-Quadrate-Schätzung angewandt wird, greift die logistische Regression auf die Maximum-Likelihood-Schätzung zurück. Durch das Regressionsmodell wird die Eintrittswahrscheinlichkeit von Y vorhergesagt, somit ist das Eintreten eines Ereignisses sehr unwahrscheinlich bei Werten nahe 0, während das Eintreten bei Werten nahe 1 sehr plausibel ist. Die logistische Regressionsfunktion ist wie folgt definiert:

$$P(Y = 1) = \frac{1}{1 + e^{-z}}$$

Der Logit wird durch z dargestellt, der ein lineares Regressionsmodell darstellt: $z = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \cdots + \beta_k * x_k + \epsilon$. x_k bezeichnen die unabhängigen Variablen und β_k die Regressionskoeffizienten. ϵ stellt den Fehlerterm dar.

Bei der Interpretation gilt, dass ein positives Vorzeichen des Regressionskoeffizienten bedeutet, dass ein Anstieg der betreffenden unabhängigen Variablen bewirkt, dass die Wahrscheinlichkeit für Y=1 steigt. Wohingegen ein negatives Vorzeichen eine Abnahme der Wahrscheinlichkeit bedeutet. Der Zusammenhang kann mithilfe von *Odds* genauer interpretiert werden. Die Berechnung erfolgt durch in Relation setzen der Wahrscheinlichkeit, dass ein Ereignis eintritt, und der Wahrscheinlichkeit des Nichteintretens. Das Chancenverhältnis, bzw. das relative Risiko oder die Chance, ist folgendermaßen definiert:

$$Odds(\pi) = \frac{\pi}{1 - \pi} = \frac{P(Y = 1)}{1 - P(Y = 1)} = \frac{P(Y = 1)}{P(Y = 0)}$$

Um den Regressionskoeffizienten zu interpretieren, werden *Odds Ratios* verwendet. Diese bezeichnen das Verhältnis der *Odds* nach dem Anstieg von x um eine Einheit und der *Odds* vor dem Anstieg von x um eine Einheit.

$$Odds\ Ratio = exp(\beta) = \frac{Odds\ nach\ dem\ Anstieg\ um\ eine\ Einheit}{Odds\ vor\ dem\ Anstieg\ um\ eine\ Einheit} = \frac{Odds_{nach}}{Odds_{vor}}$$

Dabei bezeichnet β den Regressionskoeffizenten. Durch die *Odds Ratio* einer unabhängigen Variable wird die Veränderung der relativen Wahrscheinlichkeit von Y=1 angegeben, wenn diese Variable um eine Einheit steigt, unter der Bedingung, dass alle anderen Variablen im

Modell konstant gehalten werden. Die *Odds Ratio* ist somit der Faktor, um den sich die *Odds* verändern bei Anstieg der Variable um eine Einheit. Somit ergibt sich für die Interpretation, dass bei einem *Odds Ratio* größer 1 die Wahrscheinlichkeit steigt. Bei einem *Odds Ratio* gleich 1 bleibt die Wahrscheinlichkeit gleich und bei einem *Odds Ratio* kleiner 1 sinkt die Wahrscheinlichkeit (1).

6 Reidentifikationsrisiko

Die theoretischen Hintergründe und die praktische Implementierung des Reidentifikationsrisikos stützt sich auf das von Hittmeir, Ekelhart und Mayer veröffentllichte Paper *Utility and Privacy Assessments of Synthetic Data for Regression Tasks*(11).

Es werden zwei verschiedene Möglichkeiten vorgestellt, das Reidentifikationsrisiko zu bewerten. Das erste Verfahren ist relativ intuitiv. Die Idee ist, dass man die Zeilen des Originaldatensatzes mit den Zeilen des synthetisierten Datensatzes einzeln vergleicht und nach identischen Zeilen, also Objekten sucht.

Die Grundannahme ist hierbei, dass die Privatsphäre dann gefährdet ist, wenn Personen im Originaldatensatz ebenfalls im synthetischen Datensatz zu finden sind. Allerdings sind nicht nur identische Zeilen von Interesse sondern auch solche mit großen Ähnlichkeitsstrukturen, da auch diese möglicherweise genug Informationen enthalten können, um zu einer Verletzung der Privatsphäreregelungen zu führen.

Infolgedessen berechnet man mittels des *k-nearest-neighbours* Verfahren für jede Zeile im synthetischen Datensatz den nächsten Nachbarn im Originaldatensatz. Dies tut man mithilfe der euklidischen Distanz:

$$d(s,o) = ||o - s||_2 = \sqrt{\sum_{i=1}^{p}(o_i - s_i)^2}$$

s gibt dabei eine Zeile im synthetischen Datensatz an und o eine im originalen. p sind die einzelnen Variablen einer Zeile, also die Spalten. Diese Distanzberechnung wird auf alle Zeilenkombinationen angewendet. Anschließend wird die kleinste Distanz gesucht, welche dem nächsten Nachbarn entspricht.

Für metrische Variablen ist diese Berechnung sehr einfach. Etwas komplizierter wird es für kategorische Variablen. Diese müssen zunächst *dummy*-codiert werden, um eine Berechnung möglich zu machen, danach ist das Vorgehen identisch.

Das hier vorgestellte Verfahren setzt $k = 1$, berechnet also nur den ersten nächsten Nachbarn. Theoretisch könnte man auch mehr als nur einen Nachbarn berechnen, um weitere Analysen durchzuführen. Dies ist einer der Unterschiede zum nachfolgenden Ansatz.

Hierbei handelt es sich um das sogenannte *CAP*. Man stellt die Grundannahme, dass der in 2 vorgestellte *Eindringling* bzw. *Intruder* bereits gewisse Teilinformationen des Datensatzes kennt. Dies sind im folgenden einzelne Variablen, die der *Intruder* beispielsweise aus anderen Studien rausfiltern konnte. Sie werden wie in 2 bereits angeschnitten *Schlüsselvariablen* bzw. *key variables* genannt.

Basierend auf diesen Zusatzinformationen versucht der *Intruder* über ein spezifisches Individuum, weitere Informationen zu extrahieren, die meistens sehr sensibel sind. Diese nennt man Zielvariablen oder *targets*.

Für das *CAP* Verfahren werden zusätzlich einige Annahmen über das Verhalten des *Intruders* getroffen. So sucht er zunächst nach exakten Übereinstimmungen der zugrundeliegenden *keys* im synthetischen Datensatz und schätzt die Zielvariablen bei metrischen Variablen als Mittelwert über die Übereinstimmungen und bei kategoriellen Variablen über den häufigsten Wert (*majority vote*).
Allerdings ist dieser Ansatz besonders bei stetigen Schlüsselvariablen problematisch, da exakte Übereinstimmungen kaum auftreten. Daher will man diese Basisidee weiterführen und nicht nur exakte Übereinstimmungen für die Vorhersage berücksichtigen, sondern auch solche Observationen mit einbeziehen, welche ähnliche Werte aufzeigen.
Anders als im vorherigen Verfahren möchte man hier aber nicht nur einen nächsten Nachbarn festlegen. Stattdessen wird ein gewichteter *knn*-Algorithmus benutzt, welcher alle Datenpunkte berücksichtigt und so möglichst objektiv ist.
Das Gewicht wird invers zur berechneten Distanzmetrik gewählt. Auch hier wurde das euklidische Distanzmaß verwendet. Allerdings wurden die Distanzen zwischen den Zeilen natürlich nicht bzgl. aller Spalten berechnet, sondern nur bzgl. jener, die den Schlüsselvariablen entsprechen. Die Gewichtung führt dazu, dass Beobachtungen, die einen kleineren Abstand aufweisen ein größeres Gewicht bei der Vorhersage erhalten. Diese Vorhersage entspricht in diesem Fall einem gewichteten Mittelwert bzw. einem gewichteten *majority vote*.
Die Berechnung der Gewichte gestaltet sich wie folgt. Zunächst wird der Vektor der relativen Abstände invertiert und anschließend normalisiert. Das Resultat ist ein Wahrscheinlichkeitsvektor. Je höher die Wahrscheinlichkeit ist, desto größer ist die Ähnlichkeit zu den Schlüsselvariablen des Originaldatensatzes.
Hierzu zwei anschauliche Beispiele. Angenommen es gibt lediglich vier Observationen. Bei der stetigen Variablen ergibt sich basierend auf den Schlüsselvariablen des Individuums, welches für den *Intruder* von Interesse ist, folgender inverse Distanzvektor: $c(0.5, 0.2, 0.2, 0.1)$ und die beobachteten Werte im synthetischen Datensatz sind $c(2, 5, 5, 5)$. Der einfache Mittelwert würde 4.25 betragen, während der gewichtete Mittelwert, der sich als Skalarprodukt aus den zwei Vektoren zusammensetzt, 3.5 ergibt.
Bei der kategorischen Variable summiert man lediglich für jede Kategorie die Werte des Wahrscheinlichkeitsvektors separat und bestimmt die Kategorie mit dem höchsten Wert. Ist der Wahrscheinlichkeitsvektor also beispielsweise wie folgt gegeben $c(0.3, 0.4, 0.25, 0.05)$ und die dazu passenden Werte im synthetischen Datensatz *c(a,b,a,c)*, dann ergibt sich für die Ausprägung a mit 0.55 die größte Wahrscheinlichkeit. Damit ist a der plausibelste Wert für diese Zielvariable.
Um das Reidentifikationsrisiko zu beurteilen, wird genau dieses Verfahren für randomisierte Ziel- und Schlüsselvariablen angewandt. Man erzeugt also zufällig verschiedene Szenarien, die eintreten könnten. Diese unterscheiden sich nicht nur anhand der Auswahl der Ziel- und

Schlüsselvariablen, sondern auch an deren Anzahl. Allerdings werden zu Vergleichszwecken für alle Synthetisierungsmethoden die gleichen Szenarien erstellt. Je nach Zielvariable wird dann entweder der *mse* oder die *accuracy* als Metrik für den Erfolg bzw. Misserfolg des Verfahrens benutzt.

7 Praktische Anwendung

7.1 Vorstellung der Datensätze

Alle im Folgenden vorgestellten Datensätze sind bis auf kleine Änderungen, die für eine adäquate Modellierung notwendig waren, auf (8) zu finden. Es wurden Stichproben in der Form gezogen, dass alle drei Datensätze vergleichbar groß sind und sich der spätere Vergleich zwischen ihnen auf die unterschiedliche Variablenanzahl fokussieren kann.

7.1.1 Bluttransfusionen

Der erste Datensatz umfasst Informationen zu 748 Blutspendern des Bluttransfusions Servicecenters der Stadt Hsin-Chu in Taiwan. Es wurden Informationen dazu gesammelt, wie lange die letzte (*Last*) bzw. aller erste (*First*) Spende her ist und wie oft insgesamt Blut gespendet wurde (*Frequency*). Außerdem wurde noch eine binäre Variable (*march2007*) erhoben, die 1 ist, falls die befragte Person im März Blut gespendet hat und 0, wenn nicht. Sie stellt im Folgenden die Zielvariable dar. Aufgrund extrem rechtsschiefer Verteilungen wurden die Prädiktionsvariablen *log*-transformiert, wobei *First* und *Last* zusätzlich um eine Einheit nach rechts verschoben wurden, um $log(0)$ Werte zu vermeiden.
Inhaltlich stellt dies kein Problem dar, da eine Verschiebung um eine Einheit nach rechts einfach einer Verschiebung um einen Tag entspricht. Zusätzlich wurde nur eine Stichprobe des Datensatzes benutzt, um das Auftreten von *unbalanced data* zu umgehen, was später bei der logistischen Regression Probleme verursacht hätte, da es sehr viel mehr Personen gibt, die im März kein Blut gespendet haben als solche, die gespendet haben.
Abbildung 2 zeigt, dass die Variable *First* keinen großen Einfluss auf die Blutspende hat. Sowohl für die Spender, als auch für die Nichtspender scheint die Variable symmetrisch verteilt zu sein.
Anders verhält es sich mit der Variable *Last*. Umso länger die Blutspende her ist, desto wahrscheinlicher ist es, dass man im März kein Blut spendet. Vor allem unter den Spendern scheint die Variable sehr linksschief verteilt zu sein, dh. auch unter den Spendern ist die letzte Blutspende für den Großteil länger her, wenn auch nicht so lange wie bei den Nichtspendern.
Bei der Häufigkeit der Blutspende im rechten Boxplot lässt sich erkennen, dass Spender im Schnitt auch in der Vergangenheit häufiger gespendet haben auch wenn der Unterschied hier nicht so deutlich ist wie bei *Last*. Besonders interessant ist hier, dass die Box bei den Nichtspendern keinen unteren *Whisker* aufweist, was darauf hindeutet, dass es sehr viele Personen unter den Nichtspendern gibt, die bisher nur ein mal gespendet haben ($log(1) = 0$). Später wird überprüft werden, ob diese Zusammenhänge nach der Synthetisierung erhalten bleiben.

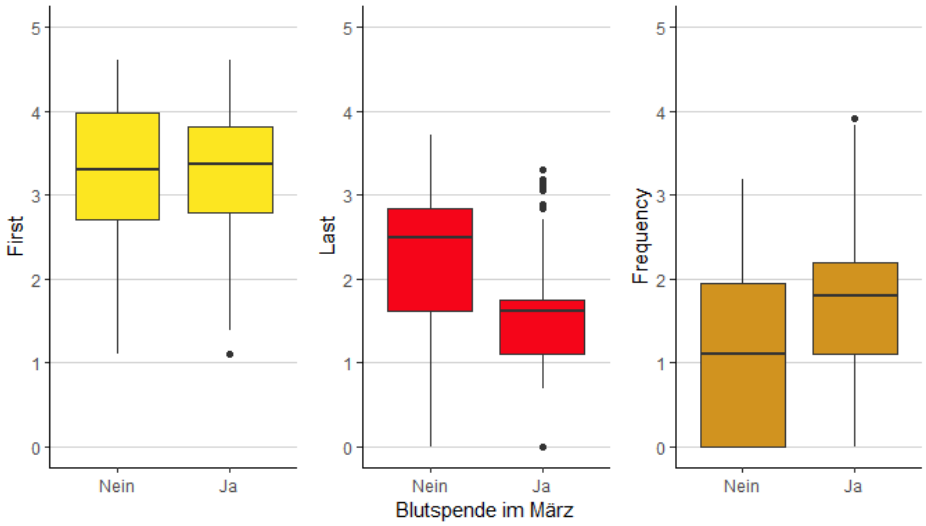


Abbildung 2: Blutspende im März, abhängig von den Variablen *First, Last, Frequency*

7.1.2 Einkommen

Die Gewinnung der Daten des zweiten Datensatzes wurde von Barry Becker aus der Volkszählungsdatenbank von 1994 durchgeführt. Hier ist von Interesse vorherzusagen, ob eine Person mehr oder weniger als 50 tausend US-Dollar im Jahr verdient.
Einflussfaktoren sind beispielsweise das Alter, der Grad der Schulbildung, das Geschlecht oder der Beziehungsstatus. Insgesamt werden zehn Variablen betrachtet. Auch hier wird aus dem Originaldatensatz gesamplet, um die Problematik der *unbalanced data* zu beheben.
Abbildung 3 zeigt, dass Personen, die über 50 tausend verdienen im Schnitt älter sind. Fast alle Befragten, die weniger verdienen, sind unter 50 Jahre alt. Bei der Schulbildung lassen sich ähnliche Resultate erkennen. Je höher die Schulbildung desto plausibler ein Einkommen über 50 tausend. Interessant hierbei ist, dass der Median bei den Hochverdienern genau bei 12 liegt, was einem Abiturabschluss entspricht.
Außerdem ist eine extrem schiefe Verteilung bei den Niedrigverdienern zu beobachten. Der Median liegt hier direkt am Ende der Box, was bedeutet, dass ein Großteil der Niedrigverdiener nur bis zur neunten Klasse die Schule besucht hat. Bei den Arbeitsstunden pro Woche ist verwunderlich, dass die Mediane für beide Einkommensgruppen beinahe identisch sind. Die Stunden bei den Hochverdienern streuen zwar etwas mehr nach oben, bei den Niedrigverdienern hingegen gibt es viel mehr Ausreißer nach oben (und auch nach unten). Hier kann also nicht geschlussfolgert werden, dass eine erhöhte Arbeitsstundenanzahl ein höheres Einkommen indiziert.

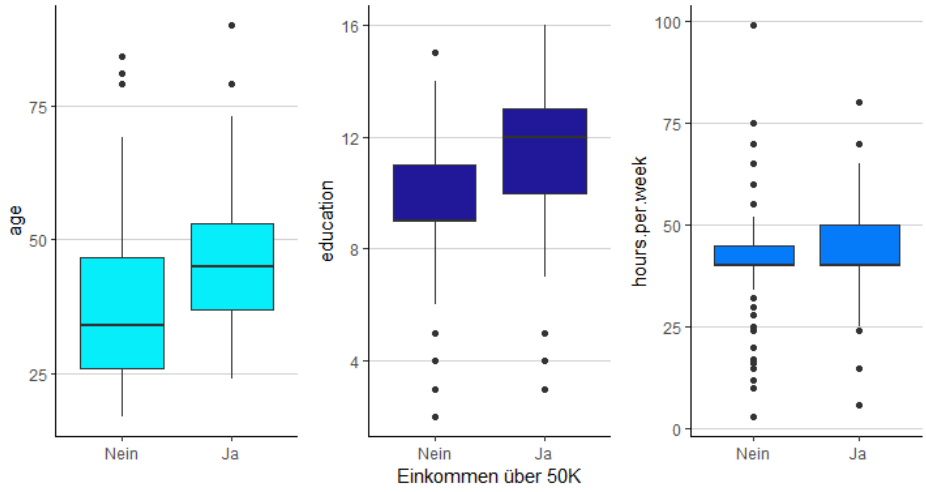


Abbildung 3: Einkommen über 50K, abhängig von den numerischen Variablen Alter, Dauer der Schuldbildung und Arbeitsstunden pro Woche

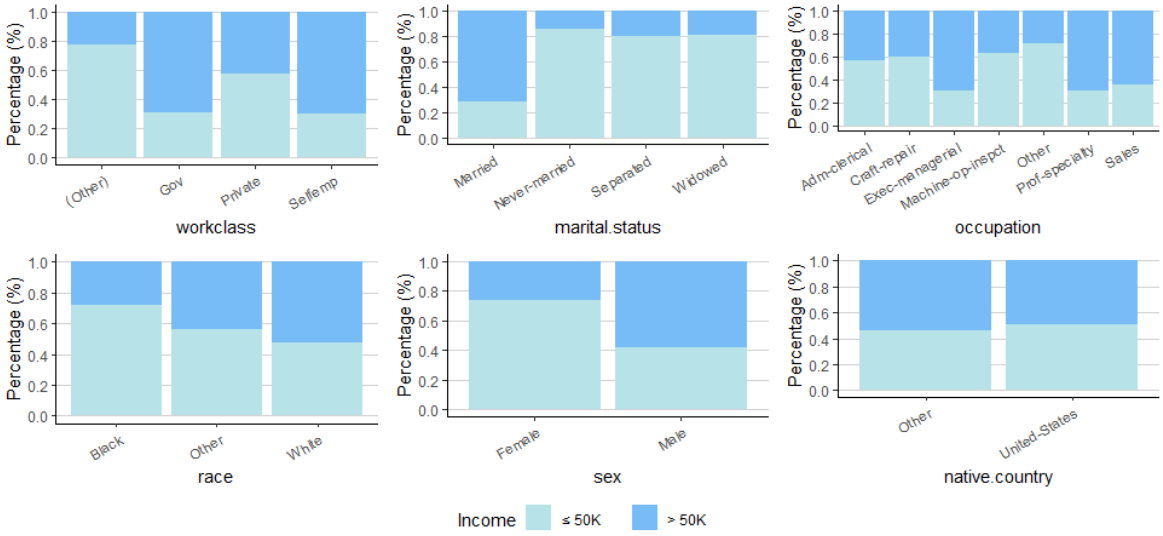


Abbildung 4: Einkommen über 50K, abhängig von den kategoriellen Variablen Arbeiterklasse, Beziehungsstatus, berufliche Tätigkeit, Rasse, Geschlecht und Heimatland

Die Histogramme in 4 zeigen das prozentuale Verhältnis der Hochverdiener zu den Niedrigverdienern in den einzelnen Kategorien der einzelnen Variablen. Man sieht beispielsweise, dass verheiratete Personen eher Hochverdiener sind, was natürlich auch damit zusammenhängen könnte, dass diese Variable mit dem Alter korreliert.
Interessant ist außerdem, dass es sehr viel mehr Frauen gibt, die weniger als 50 tausend Dollar verdienen als welche, die über 50 tausend verdienen. Bei Männer ist das Gegenteil zu beobachten, wobei hier gesagt werden muss, dass das Verhältnis noch relativ ausgeglichen ist.
Das Herkunftsland scheint keine Rolle für das Einkommen zu spielen. Hier liegt in beiden Kategorien eine annähernde Gleichverteilung vor. Arbeitet eine Person beim Staat oder ist selbstständig erhöht das ihre Chance ein Hochverdiener zu sein.

Bei der beruflichen Tätigkeit zeigen vor allem Personen, die im Verkauf oder im Management tätig sind hohe Anteile an Hochverdienern.
Zuletzt fällt bei der Betrachtung der Variablen *race* auf, dass das Verhältnis bei Personen mit weißer Hautfarbe ziemlich genau 50/50 beträgt, wohingegen die Kategorie *Black* einen sehr hohen Anteil an Niedrigverdienern aufweist.

7.1.3 Schulleistungen

Der letzte Datensatz, der hier untersucht werden soll, betrachtet die Schulleistungen von portugiesischen Schüler in der Sekundarausbildung, welche ungefähr unserer Realschulbildung entspricht, im Grundkurs Mathematik. Es wird untersucht wie sich demografische, soziale und schulbezogene Merkmale auf das Bestehen bzw. Nichtbestehen der Schüler auswirkt.
Dabei werden die Noten von 0 bis neun zur Zielkategorie *Nichtbestanden* (kodiert mit 0) zusammengefasst und die Noten von 10 bis 20 zur Zielkategorie *Bestanden* (kodiert mit 1). Der zur Analyse verwendete Datensatz umfasst 395 Schüler und 31 Variablen.
Beispiele für die Prädiktionsmerkmale sind der tägliche Alkoholkonsum, das Schulbildungsniveau der Eltern und deren berufliche Tätigkeit aber auch die Anfahrtszeit zur Schule und die Anzahl der Familienmitglieder. Nachfolgend werden der Übersicht wegen nur die interessantesten Merkmale näher beleuchtet.
Der Datensatz beinhaltet zwei metrische Variablen, die in Abbildung 5 visualisiert sind. Beim Alter kann man erkennen, dass unter den Schülern, die bestanden haben, ein großer Anteil genau 16 Jahre alt ist, was dem Regelalter für diesen Schulabschluss entspricht. Diejenigen, die nicht bestanden haben sind im Schnitt älter. Hier fällt die Verteilung ziemlich symmetrisch aus.
Bei den Fehltagen zeigen beide Boxplots eine relativ symmetrische Verteilung auf, die Streuung unter den bestandenen Schülern ist hier aber deutlich geringer. Allgemein kann man aber für beide Gruppen sagen, dass die meisten Schüler nicht mehr als 10 Tage gefehlt haben. Ausreißer gibt es für beide Variablen in beiden Gruppen nur nach oben.
Das erste, was bei den kategoriellen Variablen 6 auffällt, ist, dass in fast allen Kategorien der Anteil derjenigen, die bestanden haben, deutlich höher ist als derjenigen, die durchgefallen sind. Das deutet wieder auf *unbalanced data* hin. Allerdings liegen hier zu wenig Observationen vor, um eine geeignete Unterstichprobe ziehen zu können. Deswegen werden für spätere Analysen aus den *failed* Beobachtungen Bootstrapstichproben gezogen werden, die den Datensatz künstlich so vergrößern, dass die Anteile einigermaßen ausgeglichen sind. Für die explorative Analyse wurde aber der Originalsatensatz beibehalten.
Die wohl größte visuelle Ausnahme stellt die Variable *higher* dar, die angibt ob die Person eine weiterführende Schule besuchen will oder nicht. Hier ist als einziges Beispiel in der Kategorie *no* ein höherer Anteil an Schüler zu sehen, die durchgefallen sind. Weiter ist zu erkennen, dass

bei der Variable *schoolsup*, die die Frage beantwortet, ob man ergänzende Nachhilfestunden bezieht, unter denjenigen, die Nachhilfe beziehen ein ungefährer Anteil von 55% das Schuljahr nicht bestanden haben. Diejenigen, die keine Nachhilfe in Anspruch nehmen, bestehen dagegen im Schnitt zu ca. 70%. Mit dem Vergleich der zwei Balken sollte man jedoch vorsichtig sein, da oftmals die eine Gruppe in der Stichprobe viel kleiner vertreten ist als die andere und einen Vergleich somit nicht rechtfertigt.
Besonders auffallend ist dies bei der Variable *Medu*, welche das Schulbildungsniveau der Mutter angibt. Hier könnte man zu der Schlussfolgerung verleitet werden, dass vor allem Personen, deren Mutter keine Schulbildung hat, mit hoher Wahrscheinlichkeit bestehen, während diese Wahrscheinlichkeit für Personen, deren Mütter nur eine Grundschulbildung genossen haben, geringer ist. Derartige Schlussfolgerungen sind in der Regel nicht richtig. Grund hierfür ist, dass im Datensatz nur drei Personen enthalten sind, deren Mütter keine Schulbildung haben. Das reicht nicht aus, um solche Vergleiche zu machen.
Auffallend ist noch, dass bei dem Merkmal *guardian* der Anteil der durchgefallenen Kinder besonders dann hoch ist, wenn der Vormund jemand anderes ist als die Eltern. In dieser Kategorie sind ungefähr die Hälfte der Schüler durchgefallen. Bei der Variable *Fjob* ist der Anteil der Bestandenen besonders groß, wenn der Vater Lehrer ist und etwas niedriger, wenn er arbeitslos ist.
Zuletzt soll noch die Variable *romantic* betrachtet werden. Hier erkennt man, dass unter den Personen, die in einer romantischen Beziehung sind ca. 40% das Schuljahr nicht bestanden haben, während es bei den Singles nur grob 35% sind. Der Unterschied ist hier eher gering.

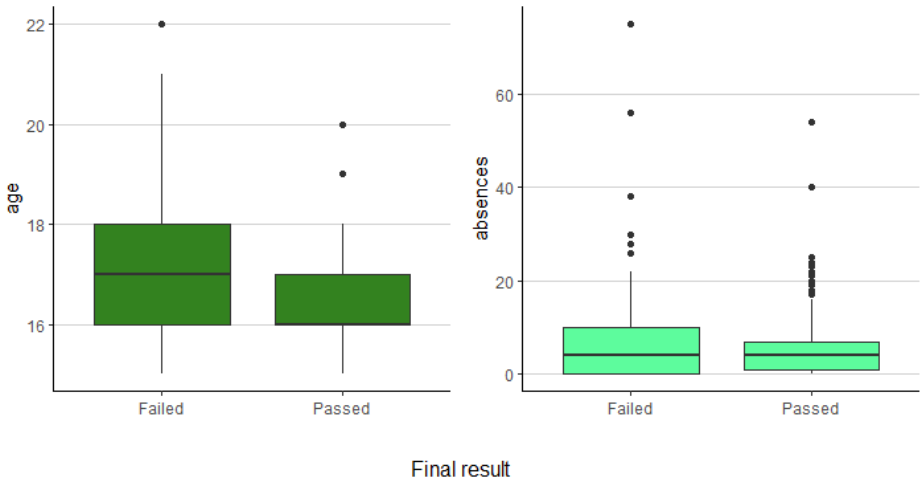


Abbildung 5: Schuljahresendleistung, abhängig vom Alter und den Fehltagen

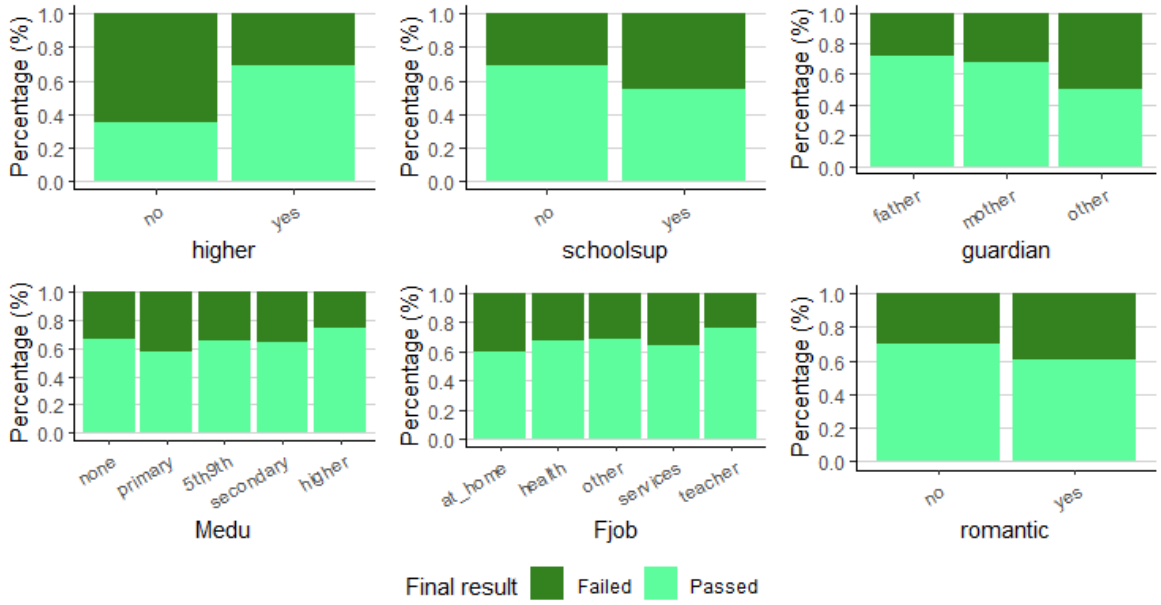


Abbildung 6: Schuljahresendleistung, abhängig vom Wunsch nach höherer Schulbildung, dem Erhalt von Nachhilfestunden, dem Vormund, der Schulausbildung der Mutter, dem Job des Vaters und dem Beziehungsstatus

7.2 Nutzen der synthetischen Datensätze

Nach der explorativen Analyse der Daten wurden die Datensätze mithilfe des *synthpop* Packets synthetisiert. Neben den in Kapitel 3 und 4 ausführlich beschriebenen Methoden *CART* und *Random Forest* wurden auch einige andere im Paket vorimplementierte Methoden zum Vergleich verwendet.

Eine davon ist die Methode *sample*, die zufällige Werte aus den Daten zieht, um die Zielvariable vorherzusagen. Die anderen Methoden basieren auf Regressionsmodellen und sind somit parametrischer Natur. Neben den herkömmlichen Modellen der logistischen Regression und der multiplen Regression *norm* wurden auch Abwandlungen dieser benutzt, wie *normrank* und *polyreg*, die möglicherweise auch etwas kompliziertere Zusammenhänge aufdecken könnten.

Wie in 2 angesprochen, musste noch eine Synthetisierungreihenfolge festgelegt werden, die in dieser Arbeit zufällig gewählt wurde. Des Weiteren wurde für die metrischen Variablen zur Prädiktion zusätzlich ein Gaußscher Kerndichteschätzer 3 verwendet, um ein niedrigeres Reidentifikationsrisiko zu erlangen.

Nach dem Erhalt der Surrogatdaten soll nun ihr Nutzen analysiert werden. Dafür vergleicht man großteils visuell den Originaldatensatz mit dem synthetischen.

Dabei spielen natürlich die Häufigkeitsverteilungen in den einzelnen Variablen aber auch die Korrelationen zwischen den Variablen und die Häufigkeiten innerhalb einer Variable bedingt auf eine andere Variable eine tragende Rolle.

7.2.1 Bluttransfusionen

Beim Bluttransfusionsdatensatz soll als erstes die spätere Zielvariable *march2007* betrachtet werden, die angibt, ob eine befragte Person im März Blut gespendet hat oder nicht. Sie ist in diesem Datensatz die einzige kategorische Variable und ihre Häufigkeitsverteilung wird deswegen als Balkendiagramm dargestellt.
Abbildung 7 zeigt bei den nicht parametrischen Modellen, dass die synthetischen Datensätze die Spender im Vergleich zum Originaldatensatz eher unterschätzen, wobei alle drei Methoden *CART*, *Random Forest* und *sample* nahezu identische Ergebnisse liefern.
Genau das Gegenteil ist bei den parametrischen Modellen zu beobachten. Hier werden die Spender eher überschätzt. Auch hier liefern die beiden Methoden identische Ergebnisse. Die Abweichungen sind für alle Modelle eher gering, was zu einem großen Nutzen führt. Aus diesem Graphen können keine Rückschlüsse darüber gezogen werden, welche Methode die beste ist.
Betrachtet man nachfolgend die anderen im Datensatz auftretenden Variablen, deren Verteilung dargestellt in Dichtefunktionen ist, lassen sich mehr Unterschiede erkennen. Bei den nicht parametrischen Modellen erhält man noch ziemlich homogene Ergebnisse 8. Alle drei Methoden scheinen den Originaldatensatz gut nachzuahmen. Bei der Variable *First* trifft das *Random Forest*-Verfahren die Dichte nahezu perfekt, bei der Variable *Last* scheint *sample* sehr gut zu performen, was doch etwas erstaunlich ist, da hier nur zufällig Werte gezogen wurden ohne irgendeine Form von Modell zu benutzen und bei der Variable *Frequency* liegt *CART* mit seiner Schätzung am besten.
Bei den parametrischen Modellen fällt gleich auf, dass die normale lineare Regression *norm* die Dichte der Originaldaten vor allem bei der Variable *Last* überhaupt nicht treffend schätzt, was möglicherweise an einem nicht-linearen Zusammenhang liegen könnte, während *norm-rank*, welche zusätzlich die Randverteilungen in die Schätzung mit einfließen lässt, deutlich zufriedenstellendere Ergebnisse liefert und bei der Variablen *Last* sogar besser abschneidet als die nicht parametrischen Modelle. Insgesamt kann man nach Betrachtung der Häufigkeitsverteilungen sagen, dass alle Synthetisierungmethoden bis auf *norm* gute Ergebnisse liefern.

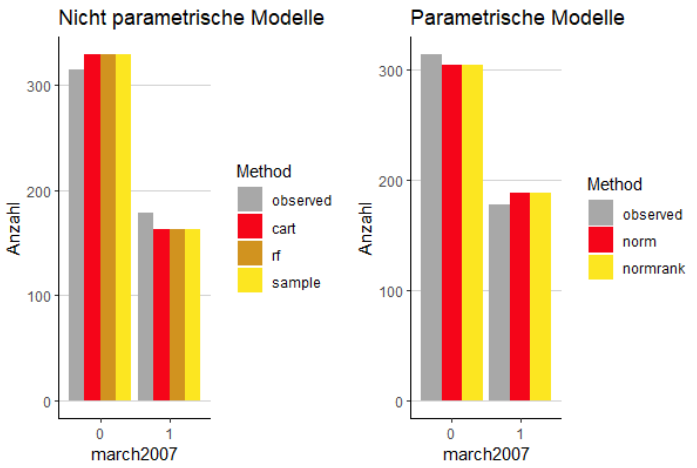


Abbildung 7: Häufigkeiten der synthetischen Datensätze im Vergleich zum Originaldatensatz in den beiden Gruppen von march2007: Blut gespendet und kein Blut gespendet. Getrennt dargestellt für parametrische und nicht parametrische synthetisierung Methoden

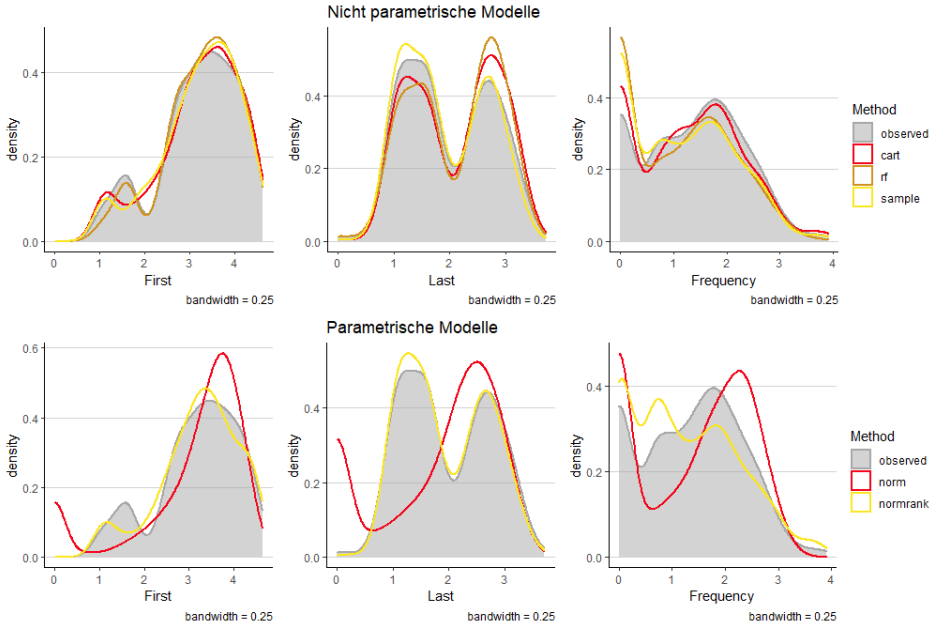
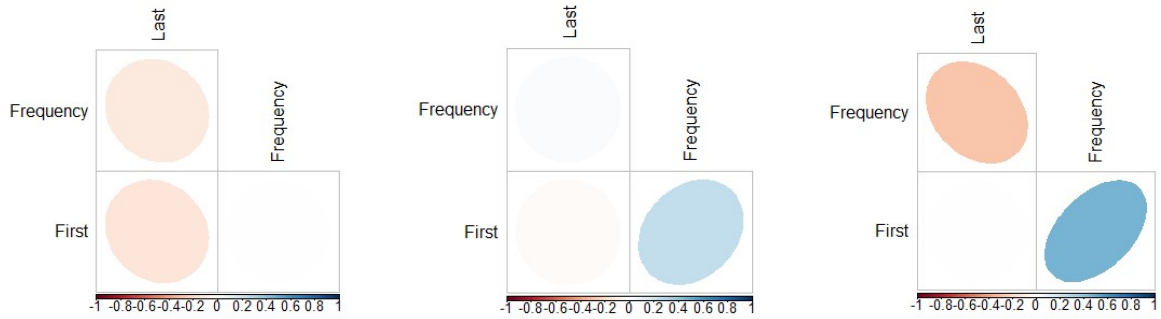


Abbildung 8: Dichtefunktionen der synthetischen Datensätze im Vergleich zum Originaldatensatz in den erklärenden Variablen First, Last, Frequency. Getrennt dargestellt für parametrische und nicht parametrische Synthetisierungsmethoden

Weiterführend sollen nun die Korrelationen zwischen den metrischen Variablen näher beleuchtet werden. Dafür wurden zunächst die Korrelationen zwischen den Variablen in den synthetischen Datensätzen berechnet, die dann von den Korrelationen im Originaldatensatz abgezogen wurden, sodass die nachfolgenden Grafiken 9, 10 die Differenzen der Korrelationen zeigen.

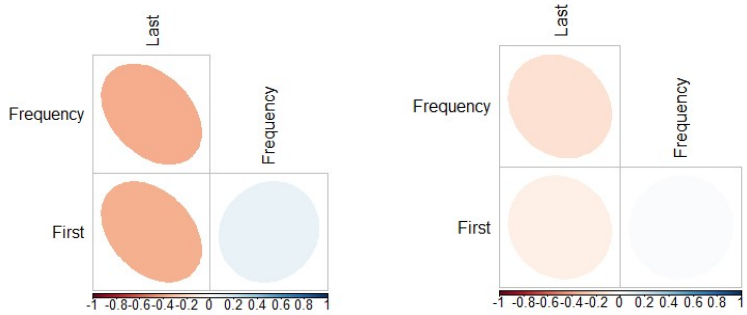
Diese sind leichter zu interpretieren, denn eine Differenz von null sagt aus, dass die Korrelationen im synthetischen wie im originalen Datensatz identisch sind und somit dieser synthetische Datensatz einen hohen Nutzen darstellt.

Bei den nicht parametrischen Modellen sieht man, dass vor allem *Random Forest* sehr niedrige Differenzen aufweist. Lediglich bei der Korrelation zwischen den Variablen *First* und *Frequency* ist ein kleiner Unterschied zwischen den Datensätzen zu erkennen. Auch *CART* weist nur kleine Unterschiede auf. *Sample* dagegen hat sich bei den Korrelationen zwischen *Frequency* und *Last* als auch bei *First* und *Frequency* deutlich vertan und ist folglich auch eher ungeeignet als Synthetisierungverfahren.
Bei den parametrischen Modellen zeigt *norm* wie bereits erwartet die deutlich schlechtesten Ergebnisse. Im Vergleich dazu sind die Differenzen bei *normrank* denen von *CART* sehr ähnlich.



(a) *CART* (b) *Random Forest* (c) *sample*

Abbildung 9: Differenz der Korrelationen zwischen den Variablen des Originaldatensatzen und des synthetischen Datensatzes mittels nicht parametrischer Methoden für den Bluttransfusionsdatensatz



(a) *norm* (b) *normrank*

Abbildung 10: Differenz der Korrelationen zwischen den Variablen des Originaldatensatzen und des synthetischen Datensatzes mittels parametrischer Methoden für den Bluttransfusionsdatensatz

Als letztes soll noch analysiert werden, wie sich die Mittelwerte der metrischen Variablen in den zwei Gruppen der kategorischen Variable bei den verschiedenen Surrogatdaten verändern. Die Ergebnisse sind in den Tabellen 1, 2, 3, 4, 5, 6 abgebildet. Allgemein kann man sagen, dass alle Verfahren ähnliche Ergebnisse liefern wie der Originaldatensatz und keine beson-

ders interessanten Abweichungen zu betrachten sind. Die Methode *normrank* liegt mit ihren Schätzungen am nächsten. *Norm* hingegen zeigt die größten Abweichungen. Diese Ergebnisse stützen das, was man schon in den vorherigen Analysen feststellen konnte.
Zusammenfassend kann also gesagt werden, dass die Methode *norm* für alle betrachteten Statistiken, die zur Evaluierung des Nutzens herangezogen wurden, am schlechtesten abgeschnitten hat. Allerdings wäre die Aussage, dass parametrische Verfahren folglich ungeeignet sind, um diesen Datensatz zu synthetisieren falsch, da die Methode *normrank* mit die besten Ergebnisse geliefert hat. *CART* und *Random Forest* gehören auch zu den stärkeren Verfahren. Sie weisen nur geringfügige Unterschiede auf, wobei *Random Forest* minimal besser ist. *Sampling* bildet das Mittelfeld.

march2007	first	last	freq
0	3.20	2.17	1.15
1	3.21	1.63	1.66

Tabelle 1: Kreuztabelle der Mittelwerte des Originaldatensatzes

march2007	first	last	freq
0	3.26	2.30	1.13
1	3.17	1.71	1.52

Tabelle 2: Kreuztabelle der Mittelwerte des synthetisierten Datensatz mittels der Methode CART

march2007	first	last	freq
0	3.27	2.29	0.889
1	3.26	1.67	1.46

Tabelle 3: Kreuztabelle der Mittelwerte des synthetisierten Datensatz mittels der Methode Random Forest

march2007	first	last	freq
0	3.25	1.90	1.08
1	3.18	1.95	1.16

Tabelle 4: Kreuztabelle der Mittelwerte des synthetisierten Datensatz mittels der Methode sampling

march2007	first	last	freq
0	3.11	2.08	1.25
1	3.06	1.44	1.62

Tabelle 5: Kreuztabelle der Mittelwerte des synthetisierten Datensatz mittels der Methode norm

march2007	first	last	freq
0	3.23	2.15	1.03
1	3.23	1.62	1.52

Tabelle 6: Kreuztabelle der Mittelwerte des synthetisierten Datensatz mittels der Methode normrank

7.2.2 Einkommen

Will man den Nutzen der artifiziellen Daten für den Einkommensdatensatz evaluieren, geht man ähnlich vor wie zuvor. Auch hier werden zunächst Häufigkeitsverteilungen und Dichtefunktionen für die einzelnen Variablen betrachtet. Folgend mit dem Vergleich der Korrelationsdifferenzen. Allerdings werden zuletzt nicht die Mittelwerte der metrischen Variablen in den einzelnen Kategorien der kategorischen Variablen berechnet, da die Anzahl der metrischen Variablen sehr gering ist und somit nur nachrangig von Bedeutung, während sehr viele nominale Variablen vorhanden sind, die diese Berechnungen sehr umfangreich machen lassen würden. Stattdessen betrachtet man eine Kontingenztabelle, die das paarweise Auftreten zweier Kategorien veranschaulicht.
Aufgrund der Vielzahl der kategorischen Variablen werden wir uns nachfolgend für die Balkendiagramme auf ein paar ausgewählte Variablen beschränken, die die größten Unterschiede gezeigt haben und somit für die Evaluierung am interessantesten sind.
Beginnend mit den nicht parametrischen Modellen 11 kann gesagt werden, dass *Random Forest* fast immer am schlechtesten die Häufigkeiten der Variablen vorhersagt. Dies wird vor allem bei der Variable *occupation* deutlich. Der Anteil im *executive manager*-Bereich wird hier beispielsweise deutlich überschätzt, während er im *prof-speciality*-Bereich unterschätzt wird. Aber auch die anderen Variablen zeigen diese Muster.
Sampling hat teilweise auch kleine Schwächen bei der Prädiktion während *CART* meistens sehr nah an den beobachteten Werten liegt und somit den besten Nutzen hinsichtlich dieser Statistik bietet.
Das für diesen Datensatz ausgewählte parametrische Modell *polyreg* 12 zeigt meist zufriedenstellende Ergebnisse. Lediglich bei der Variablen *occupation* sind größere Abweichungen zu beobachten. Ansonsten sind die Ergebnisse mit *CART* am besten.

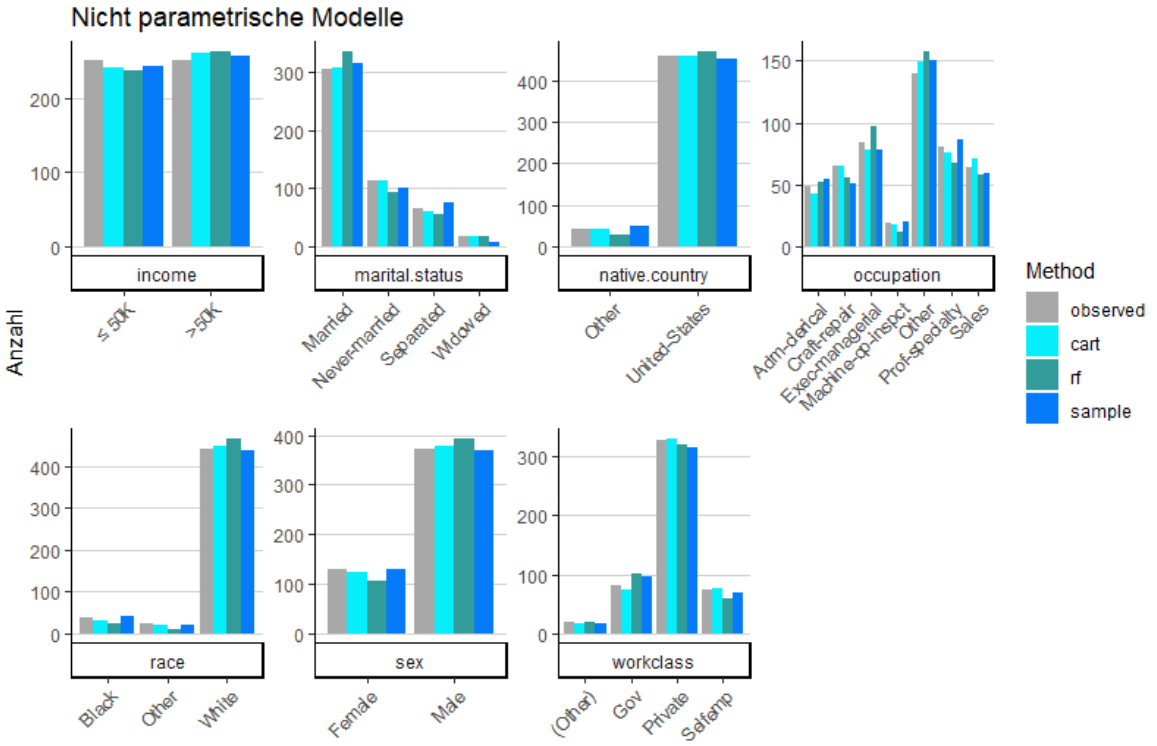


Abbildung 11: Häufigkeiten der synthetischen Datensätze im Vergleich zum Originaldatensatz für die kategorischen Variablen bei nicht parametrischen Modellen für den Einkommensdatensatz

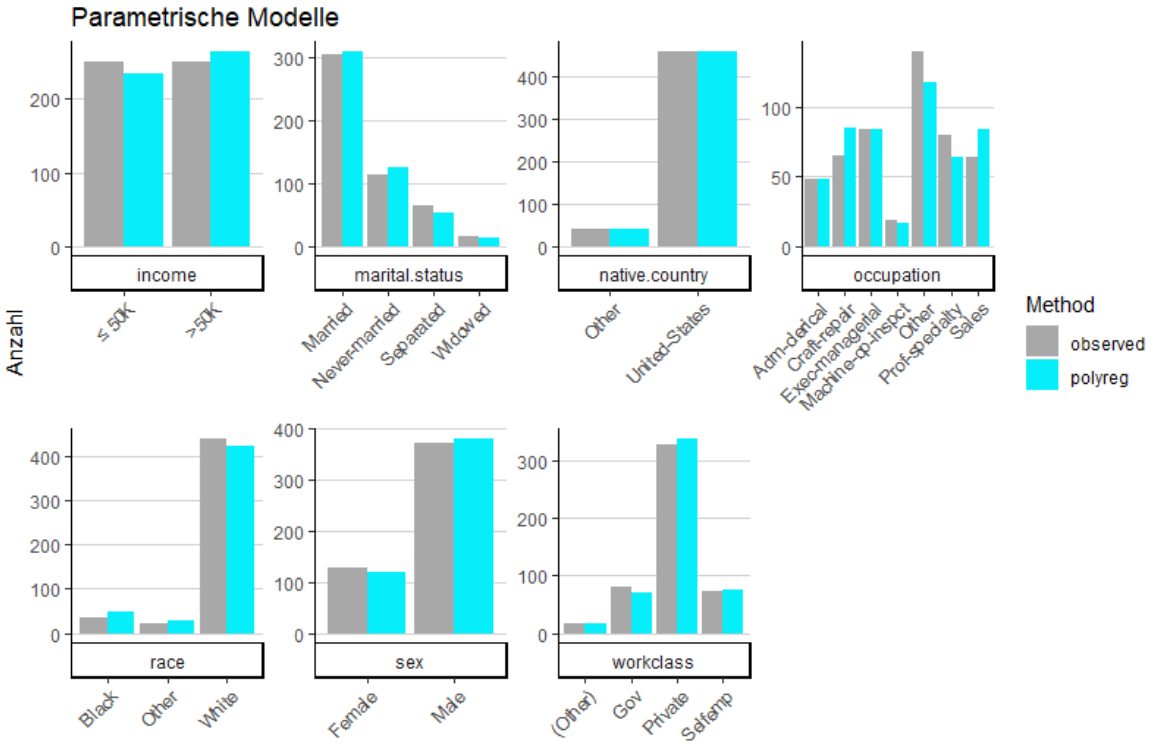


Abbildung 12: Häufigkeiten der synthetischen Datensätze im Vergleich zum Originaldatensatz für die kategorischen Variablen bei parametrischen Modellen für den Einkommensdatensatz

Betrachtet man die Dichten der drei metrischen Variablen 13 ist unschwer zu erkennen, dass *polyreg* die schlechtesten Ergebnisse liefert und besonders für die Variablen *education* und *hours per week* keine brauchbaren Ergebnisse hervorbringen kann.
Anders ist das Ergebnis bei den nicht parametrischen Verfahren zu interpretieren. Hier sieht man sehr gute Anpassungen. Besonders bei der Variablen *hours per week* gibt es kaum Abweichungen von den Originaldaten für alle verwendeten Methoden. Ansonsten scheint man mit *sample* sehr gute Anpassungen zu erhalten, während *Random Forest* vor allem bei *education* Schwierigkeiten aufweist. *CART* bildet das Mittelfeld, wobei gesagt werden muss, dass hier nur von minimalen Abweichungen die Rede ist.

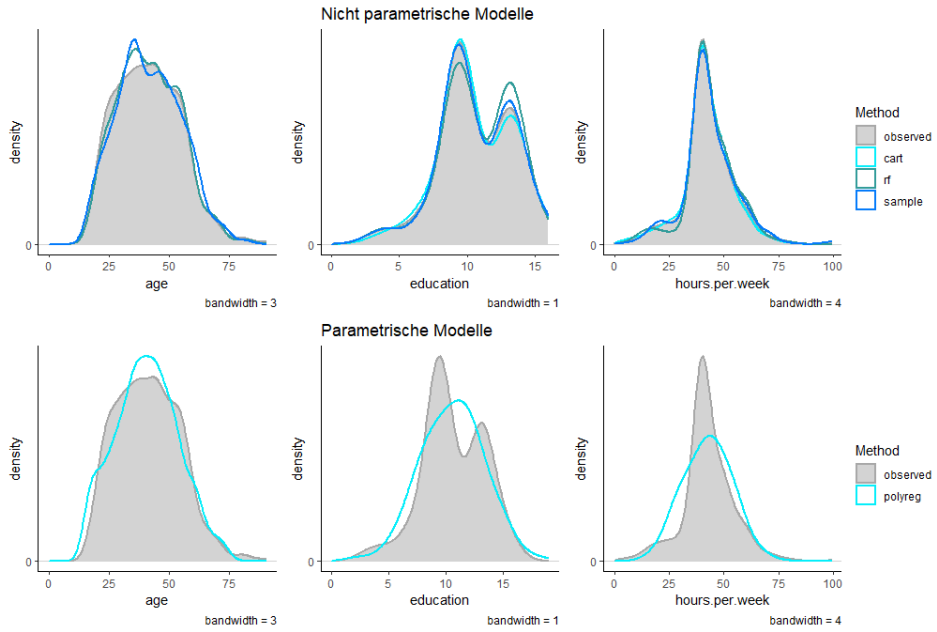
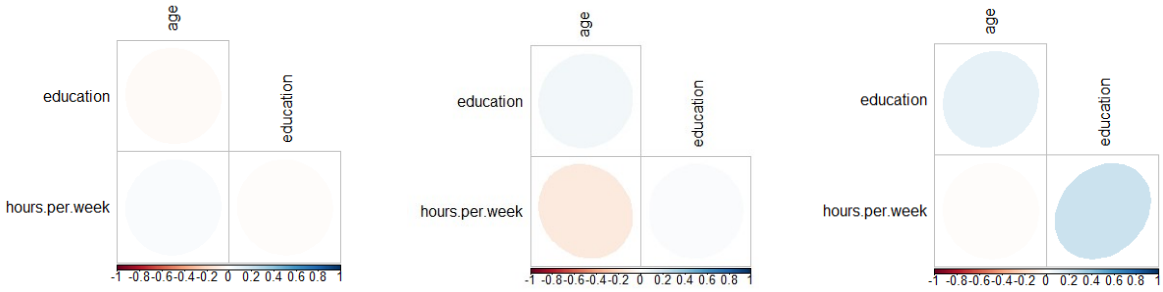


Abbildung 13: Dichtefunktionen der synthetischen Datensätze im Vergleich zum Originaldatensatz für die metrischen Variablen. Getrennt dargestellt für parametrische und nicht parametrische Synthetisierungsmethoden für den Einkommensdatensatz

Geht man nun weiter und analysiert die Differenzen der Korrelationen zwischen den metrischen Variablen, sind sie bei *CART* nahezu null, was auf einen sehr guten Nutzen hindeutet. Auch bei *polyreg* sind kaum Unterschiede zwischen den Originaldaten und den synthetisierten zu sehen.

Random Forest weist etwas stärkere Unterschiede auf, wobei diese auch noch sehr gering sind. Nur die Korrelation zwischen *hours per week* und *age* scheint etwas größer zu sein als vorhergesehen. Sogar *sample*, welches im vorherigen Datensatz große Probleme bei den Schätzungen der Korrelationen gezeigt hat, schneidet gut ab. Jedoch deutlich schlechter als vergleichsweise *CART* oder *polyreg* 14 15.



(a) *CART* (b) *Random Forest* (c) *sample*

Abbildung 14: Differenz der Korrelationen zwischen den Variablen des Originaldatensatzen und des synthetischen Datensatzes mittels nicht parametrischer Methoden für den Einkommensdatensatz

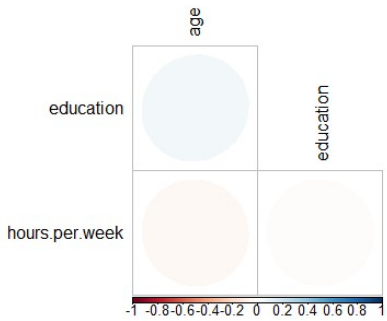


Abbildung 15: Differenz der Korrelationen zwischen den Variablen des Originaldatensatzen und des synthetischen Datensatzes mittels der Methode *polyreg* für den Einkommensdatensatz

Nicht zu vergessen ist jedoch, dass der Datensatz zehn Variablen umfasst, von denen nur drei metrisch sind. D.h, dass man mithilfe der Korrelationsgrafiken nicht wirklich Schlüsse darauf schließen kann, wie gut die einzelnen Synthetisierungsmethoden global sind, da sie nur einen kleinen Anteil des Datensatzes ausmachen.

Dafür werden im Folgenden noch Kontingenztabellen näher beleuchtet, die das paarweise Auftreten zweier Kategorien von den einzelnen Variablen aufzeigt. Hierbei werden ähnlich wie bei der Betrachtung der Korrelationsdifferenzen nicht die absoluten Anzahlen der jeweiligen Pärchenkombinationen dargestellt, sondern es werden die Anzahlen ins Verhältnis zu den Anzahlen im Originaldatensatz gesetzt, sodass man erkennen kann, ob verschiedene Kombinationen öfter oder weniger oft als ursprünglich beobachtet, auftreten. Um eine Vergleichbarkeit zwischen den einzelnen Synthetisierungsverfahen gewährleisten zu können, wurde die Skala auf 0-4 gezwungen. Alle Werte, die größer sind als vier, sind grau eingefärbt und deuten auf eine enorme Abweichung hin.

Zunächst wird das Verfahren *CART* zur Analyse herangezogen. Abbildung 16 zeigt viele weiße Flächen, was auf einen guten Nutzen hinweist, da weiß einem $1:1$ Verhältnis entspricht. Ansonsten ist die Farbe lila etwas mehr repräsentiert als grün, was ein Anzeichen dafür ist, dass die Anzahlen im synthetischen Datensatz eher unterschätzt werden. Besonders auffällig ist hier die Kategorie *Other 2* der Variablen *race*. Hier scheinen oftmals falsche Anzahlen geschätzt zu werden, unabhängig von der Kombinationsvariablen.

Die *Random Forest* Methode 17 zeigt deutlich mehr farbige Flächen. Auch hier wird die Anzahl eher unterschätzt. Die Kategorie *Other 2* scheint wie oben besonders problematisch. Die Kategorie *Black* der selben Variablen *race* hat auch Schwierigkeiten mit der Schätzung, diese sind bereits bei *CART* zu erkennen gewesen, jedoch hier noch stärker. Auffällig ist außerdem der besonders große Wert bei der Kombination *Adm-clerical* mit *Selfemployed*, der im synthetischen Datensatz vier mal so groß ist wie im Originaldatensatz.

Ein durchaus anderes Ergebnis zeigt das *sampling* 18 Verfahren. Hier sind vorwiegend grüne Flächen zu beobachten. Außerdem gibt es vier grau eingefärbte Werte, die größere Verhältnisse als 4 aufweisen. Dies sind enorme Abweichungen von Werten, die das bis zu 13-fache des

Originaldatensatzes schätzen und somit keine brauchbaren Analysen ermöglichen. Hier lassen sich anders als bei den zwei vorherigen Verfahren keine eindeutigen Strukturen feststellen. Von der Falschschätzung betroffen sind vor allem die Kategorien links von der Kategorie *widowed*. Auch wenn nicht unbedingt gesagt werden kann, dass sehr viel weniger Flächen weiß geblieben sind, als bei den anderen Verfahren, kann man doch deutlich sehen, dass die Farben und somit die Verhältnisse stärker von der 1 abweichen als bisher. Das lässt die *sample* Methode weniger geeignet wirken als die Methoden *CART* und *Random Forest*.
Zuletzt soll noch das parametrische Verfahren *polyreg* 19 untersucht werden. Es zeigt ähnliche Ergebnisse wie die *CART* Methode und zählt somit zu den besten Verfahren. Auch hier bleibt die Kategorie *Other 2* mit am problematischten. Grund hierfür kann sein, dass die Kategorie zu wenig Beobachtungen enthalten hat, um eine gute Vorhersage gewährleisten zu können.
Insgesamt hat sich *CART* als die für diesen Datensatz geeignetste Methode hinsichtlich des Nutzens bewährt, da sie bei allen drei Statistiken zu den besten gehört hat. *Polyreg* hatte sehr große Schwierigkeiten bei der Vorhersage der Dichten der metrischen Variablen, war aber ansonsten auch immer ein Vorreiter und kann somit auch als durchaus geeignetes Verfahren für diesen Datensatz betrachtet werden. Vor allem im Hinblick auf die Tatsache, dass die metrischen Variablen hier eher eine Ausnahme bilden. *Sampling* und *Random Forest* haben etwas schlechter abgeschnitten, sowohl bei den Histogrammen, als auch bei den Kontingenztabellen stellen diese Verfahren das Schlusslicht dar.

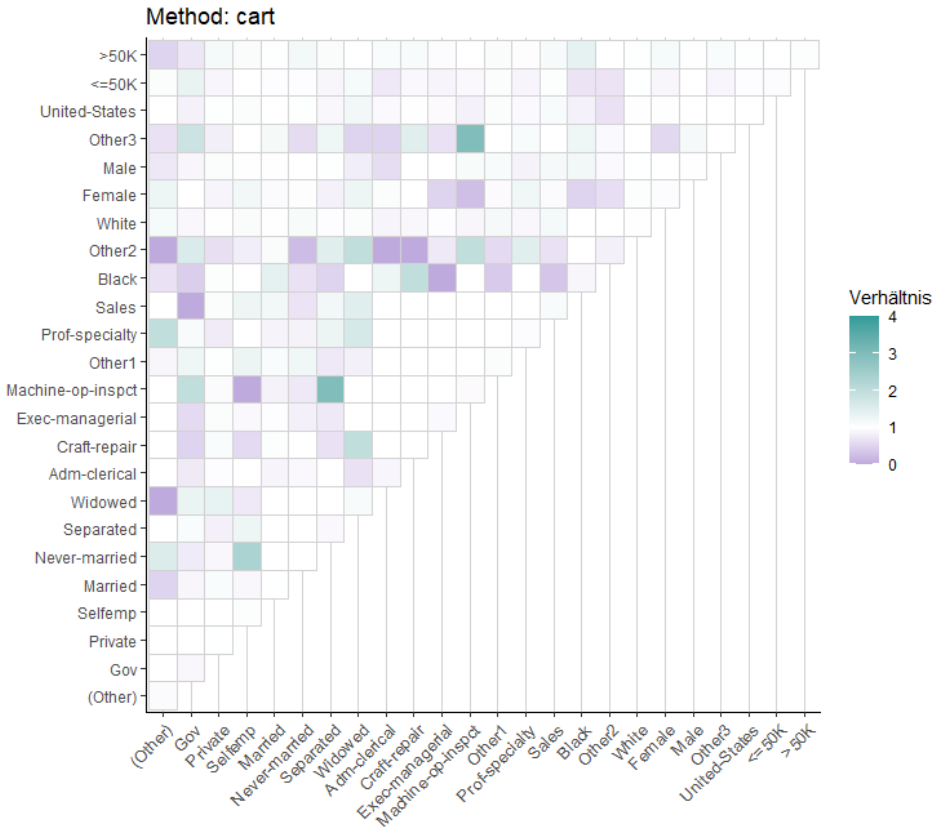


Abbildung 16: Kontigenztabelle für die kategorischen Variablen mit den Verhältnissen des Auftretens im Originaldatensatz im Bezug auf den synthetischen Datensatz mittels der Methode *CART* für den Einkommensdatensatz

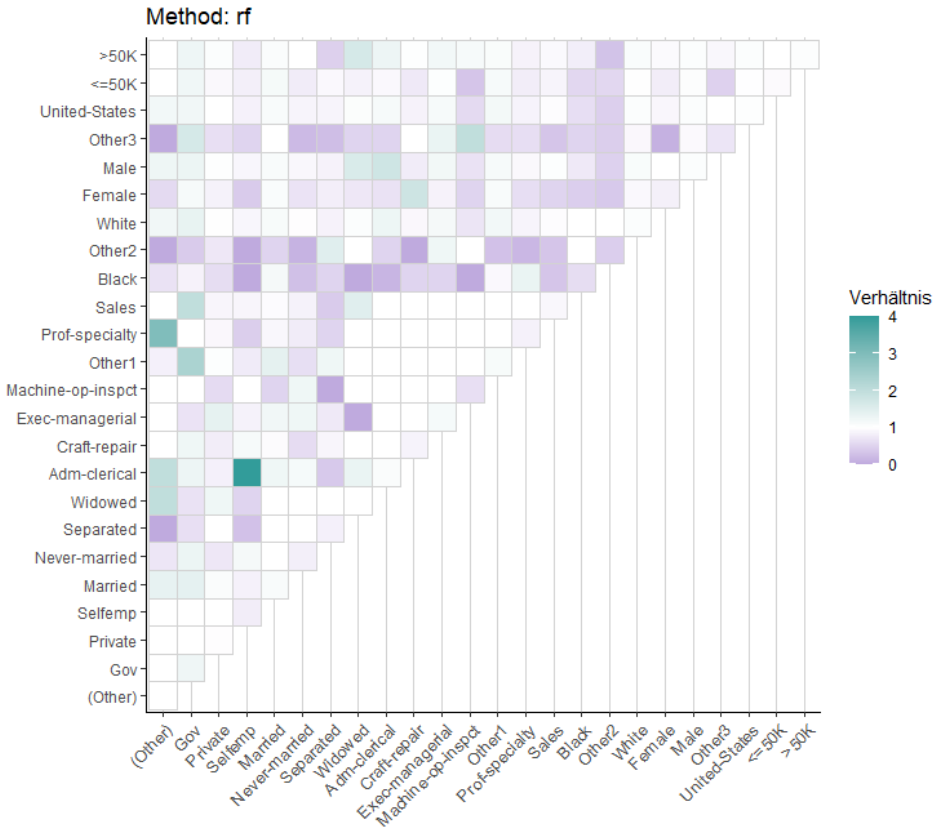


Abbildung 17: Kontigenztabelle für die kategorischen Variablen mit den Verhältnissen des Auftretens im Originaldatensatz im Bezug auf den synthetischen Datensatz mittels der Methode *Random Forest* für den Einkommensdatensatz

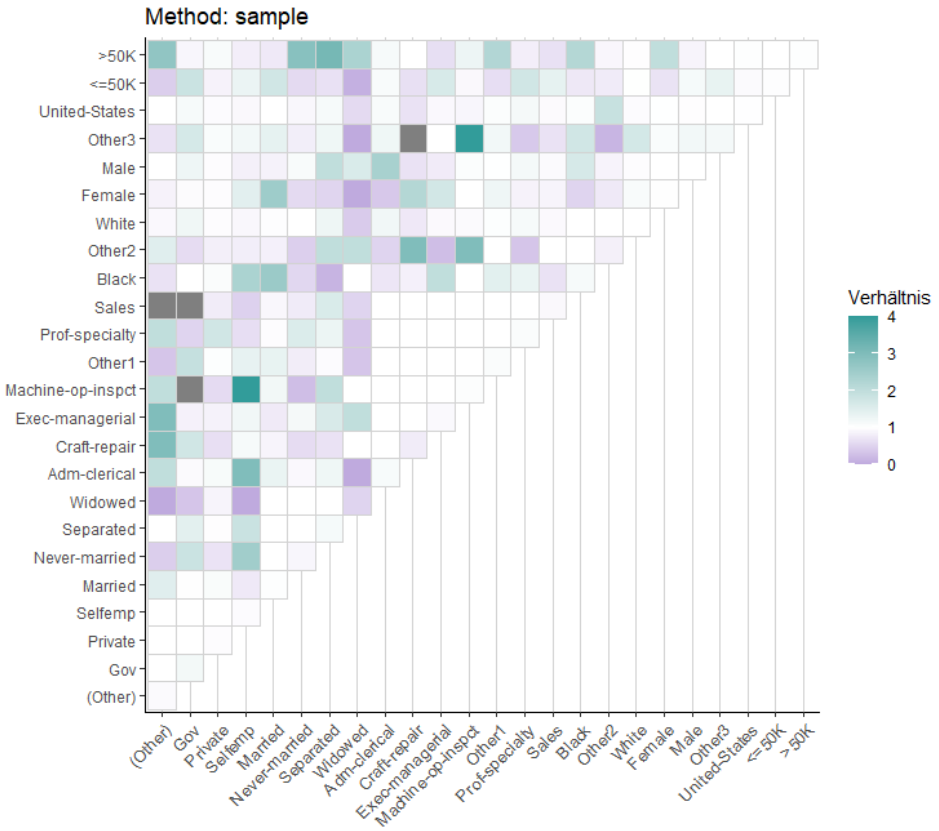


Abbildung 18: Kontigenztabelle für die kategorischen Variablen mit den Verhältnissen des Auftretens im Originaldatensatz im Bezug auf den synthetischen Datensatz mittels der Methode *sample* für den Einkommensdatensatz

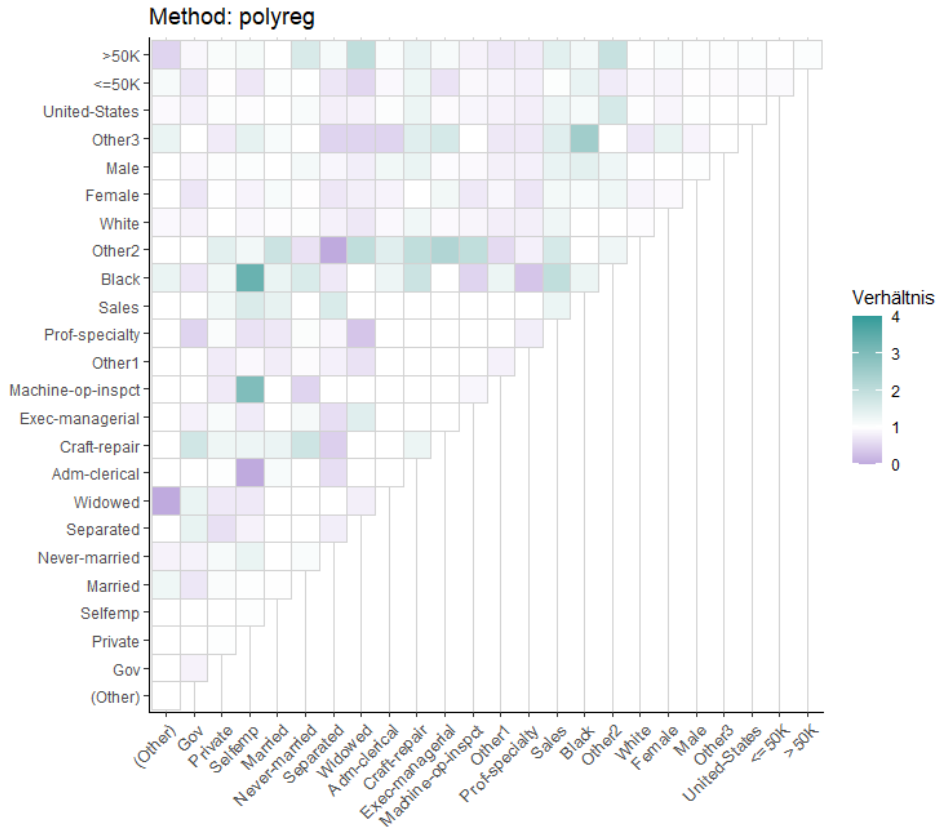


Abbildung 19: Kontigenztabelle für die kategorischen Variablen mit den Verhältnissen des Auftretens im Originaldatensatz im Bezug auf den synthetischen Datensatz mittels der Methode *polyreg* für den Einkommensdatensatz

7.2.3 Schulleistungen

Zur Evaluierung des Nutzens des letzten Datensatzes, der die Schulleistungen portugiesischer Schüler einer Realschulabschlussklasse behandelt, geht man vor wie beim vorherigen Datensatz. Auch hier sind die kategorischen Variablen vorherrschend, weswegen analog zum Einkommensbeispiel wieder Kontingenztabellen, die die paarweisen Kombinationen zwischen unterschiedlichen Kategorien betrachten, visualisiert sind. Wie bisher werden zusätzlich Häufigkeitsverteilungen bzw. Dichten für die metrischen Variablen beleuchtet werden. Aufgrund der Vielzahl an Variablen (31) werden auch hier nur ein paar ausgewählt, um die Häufigkeitsverteilungen der einzelnen Synthetisierungsverfahren zu beurteilen. Bei den Kontingenztabellen wurde der Übersicht halber auf das Hinzufügen der Variablennamen verzichtet. Hier sollen nur allgemein Strukturen analysiert werden, ohne auf eine spezielle Kategorie einer bestimmten Variablen einzugehen. Des Weiteren wurde auf eine Visualisierung der Korrelationen verzichtet, da der Datensatz nur zwei metrische Variablen enthält.
Die Differenz der Korrelationen zwischen diesen zwei Variablen des synthetischen Datensatzes und des Originaldatensatzes ist für jeden Surrogatdatensatz in nachfolgender Tabelle zu sehen

7.

	CART	Random Forest	sample	polyreg
Corr(age, absences)	0.038	0.125	0.138	0.057

Tabelle 7: Differenz der Korrelation zwischen den metrischen Variablen des Originaldatensatzen und der des synthetischen Datensatzes

Ähnlich wie beim vorherigen Datensatz zeigen die Verfahren *CART* und *polyreg* sehr geringe Differenzen auf, was dafür spricht, dass sich hier die Korrelationsstrukturen kaum verändert haben.
Die Werte für die Methoden *Random Forest* und *sample* sind mehr als doppelt so hoch aber immer noch gering. Auch hier gilt, der Datensatz enthält lediglich zwei metrische Variablen, weswegen diese nicht zu stark in die Gesamtevaluierung einfließen sollten.
Betrachtet man nun die Histogramme der nicht parametrischen Modelle 20 erkennt man zunächst, dass verglichen mit den vorherigen Datenbeispielen die Vorhersagen für alle Methoden etwas schlechter sind. Ein eindeutig bestes Verfahren kann hier nicht gefunden werden.
Für die Variablen *address* und *health* liegt *Random Forest* mit seinen Schätzungen deutlich daneben. Bei den Variablen *Fedu* und *Walc* weisen die Methoden *CART* und *sample* deutliche Schwächen auf. Auch das parametrische Verfahren *polyreg* 21 hat Schwierigkeiten mit der Vorhersage der Variable *Fedu*. *Health*, *traveltime* und *activities* scheinen dagegen recht gut prädiktiert worden zu sein.

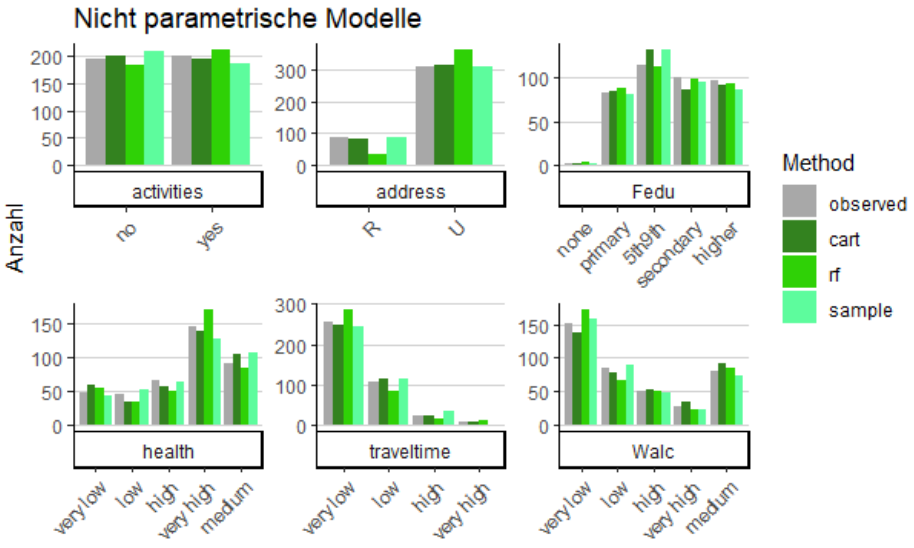


Abbildung 20: Häufigkeiten der synthetischen Datensätze im Vergleich zum Originaldatensatz für die kategorischen Variablen bei nicht parametrischen Modellen für den Schulleistungsdatensatz

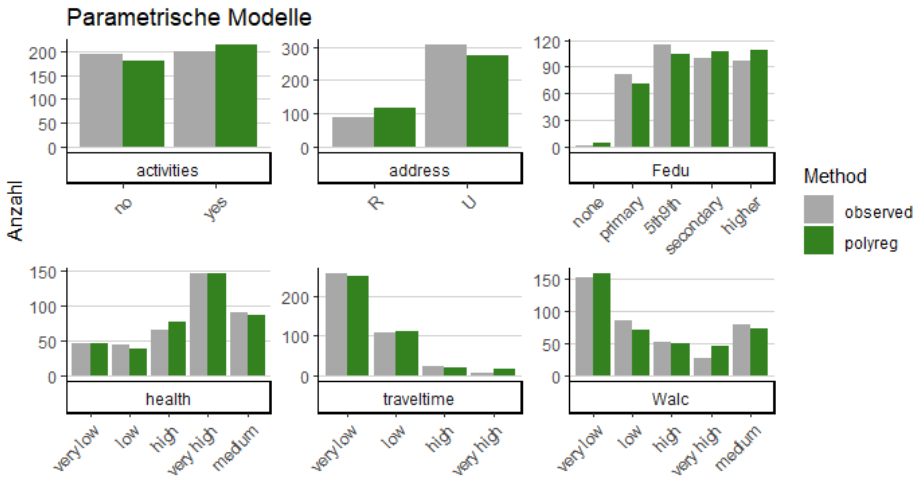


Abbildung 21: Häufigkeiten der synthetischen Datensätze im Vergleich zum Originaldatensatz für die kategorischen Variablen bei parametrischen Modellen für den Schulleistungsdatensatz

Die Dichtefunktionen liefern auch ähnliche Ergebnisse wie zuvor 22. *Polyreg* ist deutlich schlechter als die nicht parametrischen Modelle, allerdings muss erwähnt werden, dass die Schätzungen für diesen Datensatz deutlich besser sind als beim Einkommensbeispiel.
Bei den nicht parametrischen Methoden überlagern sich die Dichten für die Variable *age* komplett, kleine Unterschiede sind bei den *absences* zu sehen. Insgesamt passen sich die unterschiedlichen Verfahren gut an, wobei *sample* die glatteste Kurve erzeugt, was den Originaldaten am nächsten kommt.

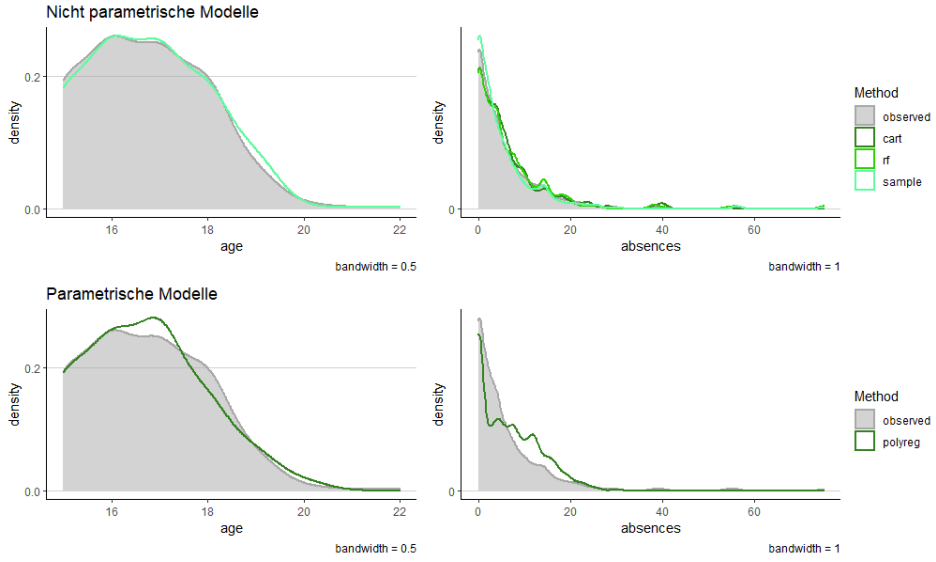


Abbildung 22: Dichtefunktionen der synthetischen Datensätze im Vergleich zum Originaldatensatz für die metrischen Variablen. Getrennt dargestellt für parametrische und nicht parametrische Synthetisierungsmethoden für den Schulleistungsdatensatz

Zum Schluss sollen noch die Kontingenztabellen 23, 24, 25, 26 betrachtet werden. Das Verfahren *CART* weist keine Strukturen auf. Die Fehlschätzungen sind mehr oder weniger über den

gesamten Bereich zufällig verteilt. Die Über- bzw. Unterschätzungen sind ziemlich ausgeglichen. Anders als beim Einkommensdatensatz gibt es hier jedoch einige grau gefärbten Werte welche eine verhältnismäßige Abweichung von mehr als 4 kennzeichnen.
Die Methode *Random Forest* scheint die meisten Pärchenkombinationen eher zu unterschätzen, da die lila Farbe präsenter ist. Hier sind klarere Strukturen zu erkennen. Es gibt einige Variablen, die für alle Kombinationspartner schlecht eingeschätzt werden. Dies erkennt man an den fast durchgezogenen lila Linien. Die grau eingefärbten Flächen sind hier hingegen nur gering repräsentiert, was zumindest lokal betrachtet zu besseren Vorhersagen und somit einem besseren Nutzen führt.
Sample hingegen zeigt große Ähnlichkeiten zu *CART*. Auch hier ist keine deutliche Struktur zu erkennen und das Verhältnis von Überschätzung zu Unterschätzung scheint weitestgehend ausgeglichen. Auffällig sind die vielen Ausreißer, welche hohe Verhältniszahlen bedeuten. Hier ist meistens die Rede von mehr als dem zehnfachen. Vor allem die zwei Variablen im unteren Eck scheinen viele dieser Ausreißer zu beinhalten.
Ganz entgegengesetzt zum *Random Forest* sind beim letzten Verfahren *polyreg* deutlich mehr grüne Flächen zu erkennen, was auf eine Überschätzung der Anzahlen für die verschiedenen Kombinationen hindeutet. Auch hier sind deutliche Strukturen zu erkennen. Vor allem im linken Bereich zeichnet sich eine dunkelgrüne fast durchgezogene Linie deutlich ab. Auch horizontal sind diese Linien sichtbar. Anders als bei *Random Forest* sind hier einige Ausreißer vorhanden.
Abschließend kann man sagen, dass *CART* die meisten hellen Stellen aufweist, während *Random Forest* kaum Ausreißer enthält. Aus diesen Gründen sind diese beiden Verfahren hinsichtlich dieser Statistik zu favorisieren.
Zusammenfassend kann gesagt werden, dass dieses Datenbeispiel etwas schlechtere Ergebnisse geliefert hat, was möglicherweise an der hohen Anzahl an Variablen liegen kann. Um dies jedoch mit Sicherheit sagen zu können, müssten mehr Analysen durchgeführt werden, was an diesem Punkt zu weit führen würde. Auch die Vermutung, dass metrische Variablen durch nicht parametrische Verfahren besser vorhergesagt werden, benötigt weitere Analysen.
Da durch einen höheren Nutzen automatisch das Reidentifikationsrisiko zu steigen droht, reicht eine Evaluierung des Nutzens allein nicht aus, um Rückschlüsse darauf ziehen zu können, welches Synthetisierungsverfahren das beste ist. Deswegen soll im nächsten Beispiel auch das Reidentifikationsrisiko für die einzelnen Surrogatdatensätze analysiert und bewertet werden.

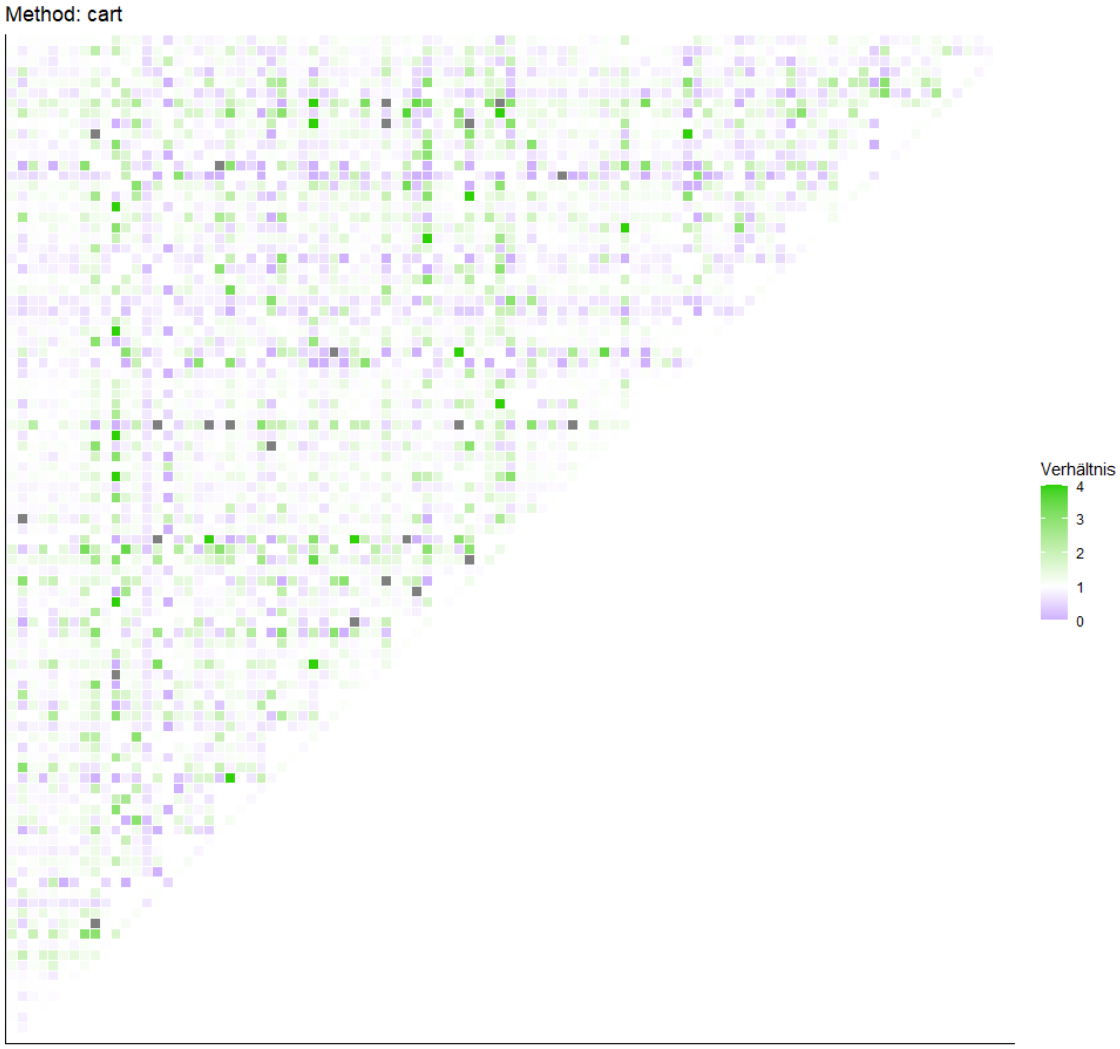


Abbildung 23: Kontigenztabelle für die kategorischen Variablen mit den Verhältnissen des Auftretens im Originaldatensatz im Bezug auf den synthetischen Datensatz mittels der Methode *CART* für den Schulleistungsdatensatz

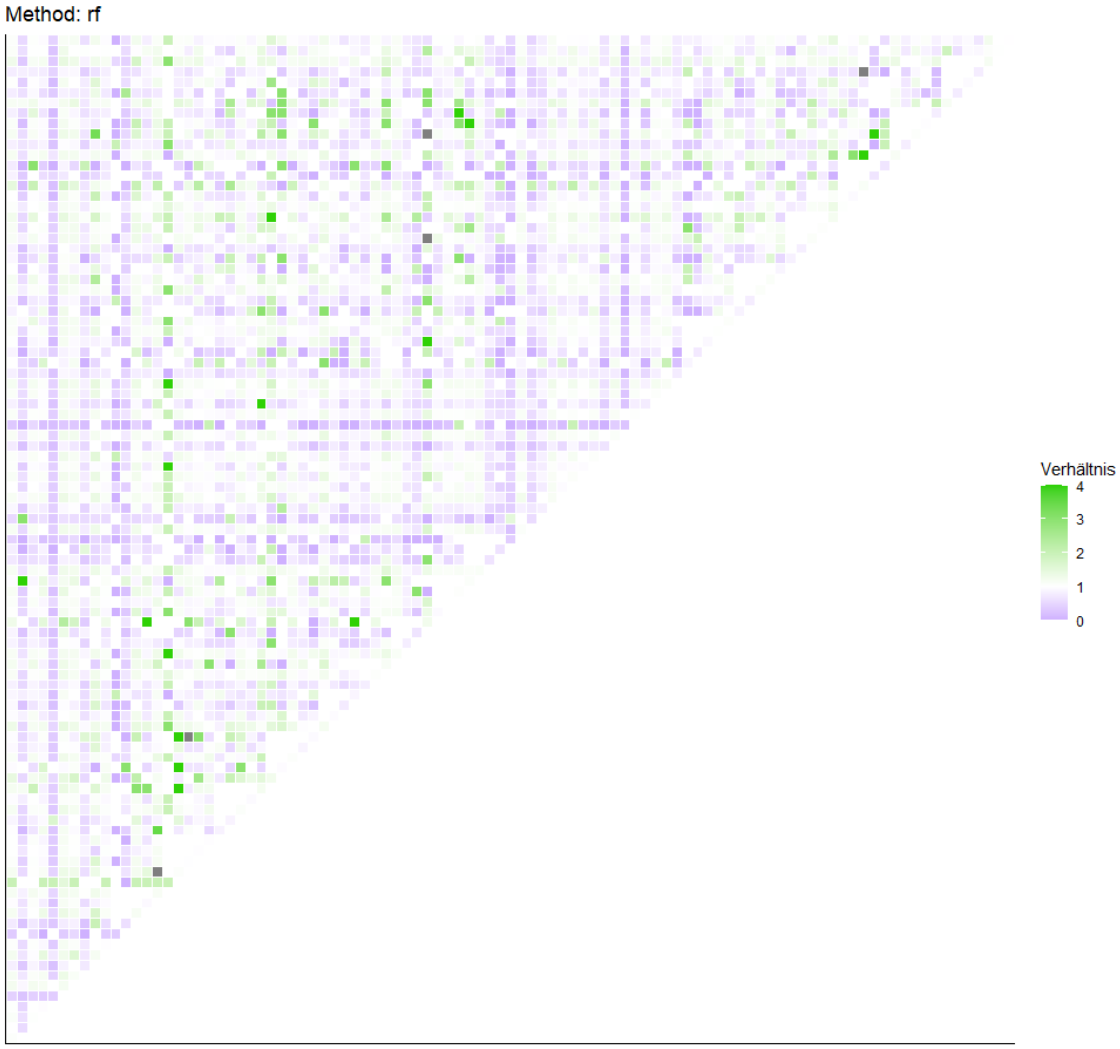


Abbildung 24: Kontigenztabelle für die kategorischen Variablen mit den Verhältnissen des Auftretens im Originaldatensatz im Bezug auf den synthetischen Datensatz mittels der Methode *Random Forest* für den Schulleistungsdatensatz

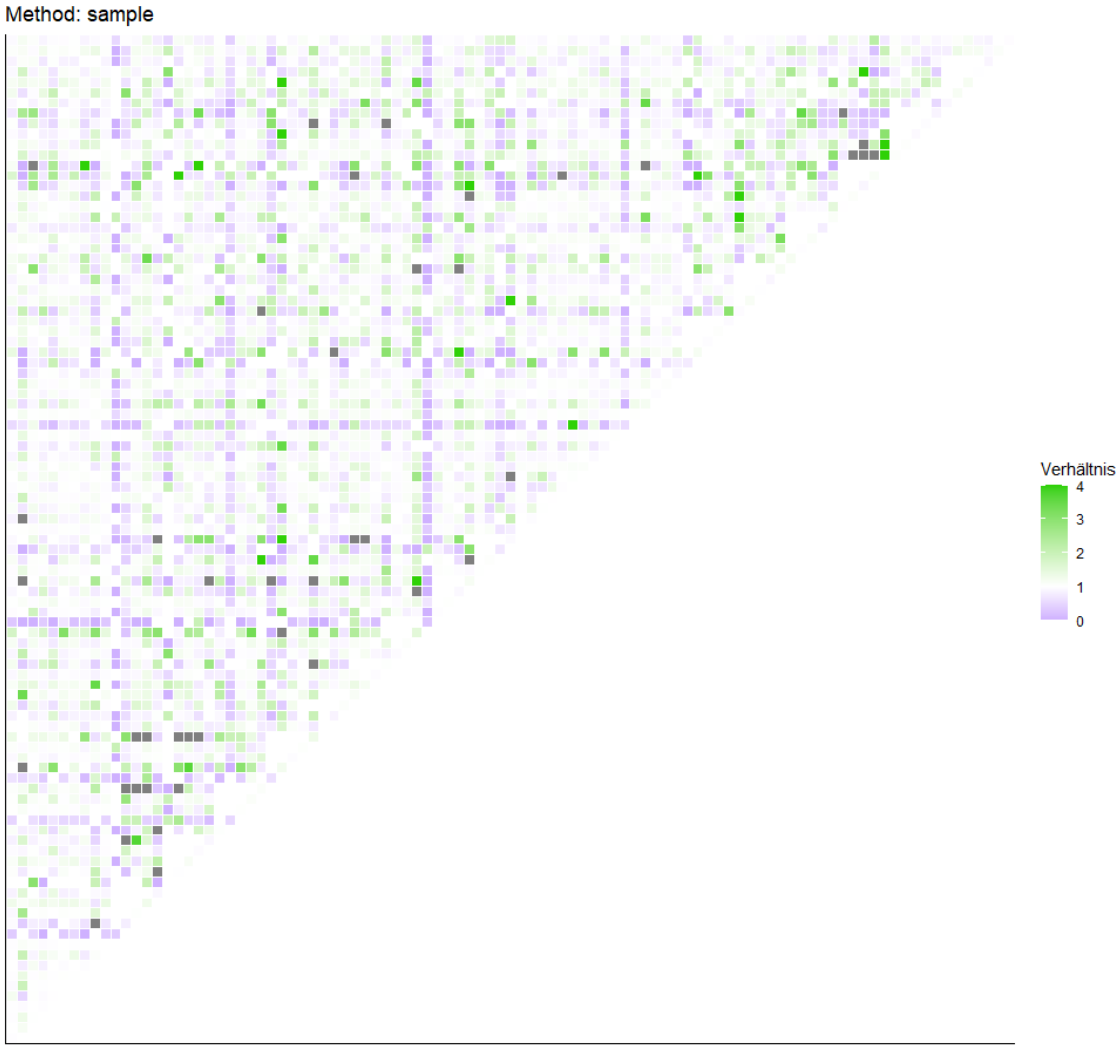


Abbildung 25: Kontigenztabelle für die kategorischen Variablen mit den Verhältnissen des Auftretens im Originaldatensatz im Bezug auf den synthetischen Datensatz mittels der Methode *sample* für den Schulleistungsdatensatz

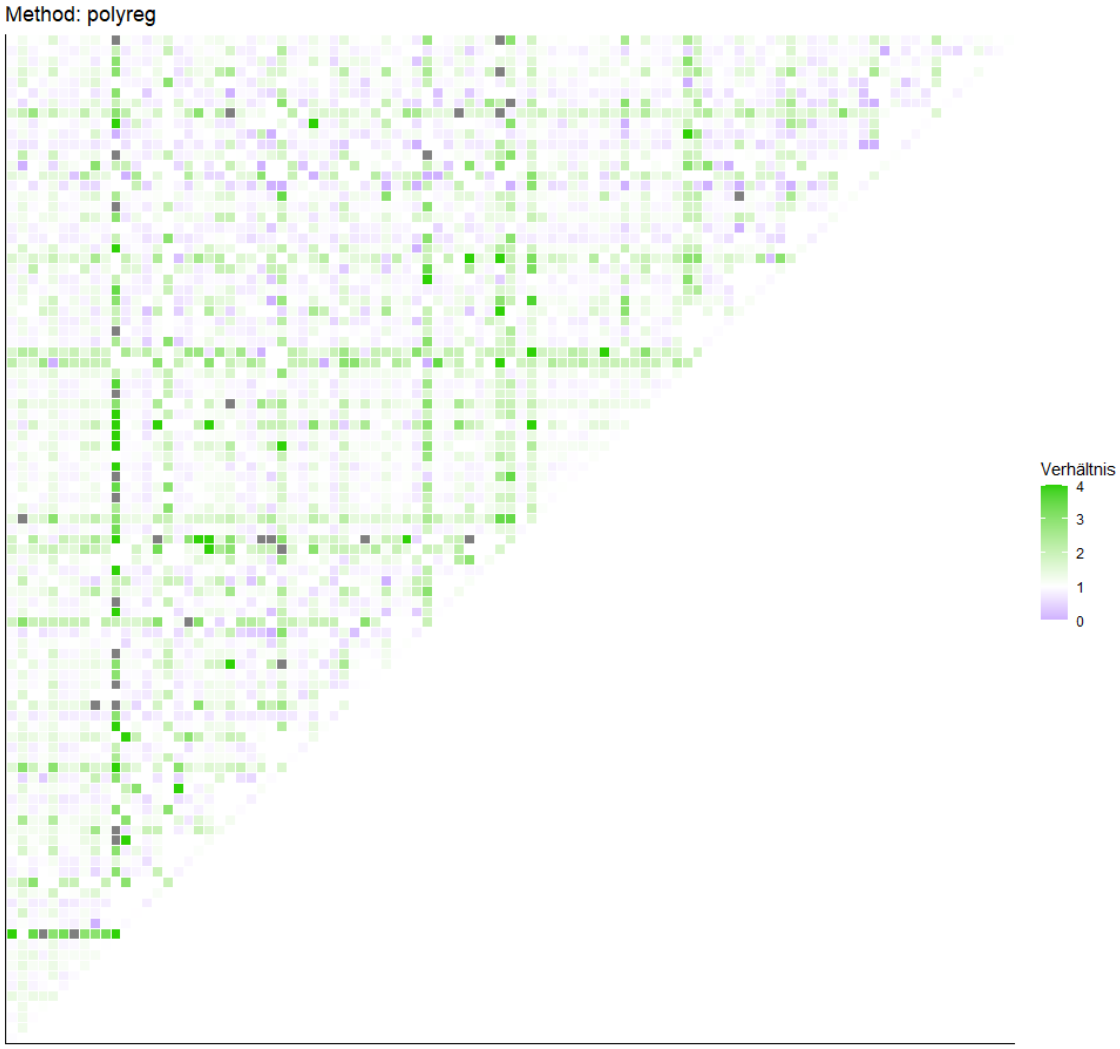


Abbildung 26: Kontigenztabelle für die kategorischen Variablen mit den Verhältnissen des Auftretens im Originaldatensatz im Bezug auf den synthetischen Datensatz mittels der Methode *polyreg* für den Schulleistungsdatensatz

7.3 Reidentifikationsrisiko der synthetischen Datensätze

In diesem Abschnitt soll das Reidentifikationsrisiko für die einzelnen Synthetisierungsverfahren der drei Beispieldatensätze evaluiert werden. Dafür werden die in 6 vorgestellten Methoden benutzt.

Zu Beginn wird der nächste Nachbar gesucht. Hierbei spricht ein hoher Wert der Distanzmetrik für ein niedriges Reidentifikationsrisiko und somit einer guten Privatsphäre, die erzielt werden soll. Nachfolgend werden die *CAP* Ergebnisse vorgestellt. Bei diesem Verfahren spricht ein niedriger *mse* bzw. eine hohe *accuracy* für ein hohes Reidentifikationsrisiko. Es muss infolgedessen davon ausgegangen werden, dass es dem *Intruder* mithilfe der Schlüsselvariablen möglich sein wird sensible Werte einzelner Individuen zu schätzen.

7.3.1 Bluttransfusionen

nearest-neighbour	CART	Random Forest	sample	norm	normrank
min	0.079	0.041	0.014	0.031	0.037
mean	0.242	0.348	0.409	0.372	0.286

Tabelle 8: Der kleinste Nachbar im gesamten Datensatz (min). Der über alle Zeilen im synthetischen Datensatz gemittelte kleinste Nachbar (mean) für den Bluttransfusionsdatensatz

Tabelle 8 zeigt einerseits den im gesamten synthetischen Datensatz kleinsten nächsten Nachbarn zum Originaldatensatz, andererseits das Mittel über alle kleinsten Nachbarn des synthetisierten Datensatzes zum Originaldatensatz. Ersteres ist besonders dann interessant, wenn sich im Datensatz beispielsweise eine Person des öffentlichen Lebens befindet, nach der gezielt gesucht werden könnte. Letzteres ist hingegen eine bessere Metrik zur Evaluierung des globalen Reidentifikationrisikos und wird deshalb bei der Analyse im Vordergrund stehen.

Betrachtet man die in der Tabelle abgebildeten Werte, fällt zunächst auf, dass *sample* die deutlich höchste Distanz aufweist. *Norm* und *Random Forest* folgen. *CART* und *normrank* weisen die geringsten Werte auf und sind somit am schlechtesten geeignet, um einen hohen Privatsphäreschutz garantieren zu können.

Ein kurzer Blick auf die minimalen Distanzen zeigt jedoch, dass *sample* vergleichsweise eine sehr kleine Distanz aufweist und somit von dem Mittelwert, bei dem die Methode den höchsten Wert aufgewiesen hat, stark abweicht. Dies ist ein Zeichen dafür, dass die *sample* Methode stark streut und nicht besonders robust ist. *Random Forest* und *CART* zeigen hier die höchsten Werte.

Aus diesem Grund scheint *Random Forest* hinsichtlich des Reidentifikationsrisikos ein geeignetes Synthetisierungsverfahren zu sein, da dieses Verfahren bei beiden Metriken die mit am besten Ergebnisse liefern konnte.

Nachdem nun der einfache Zeilenvergleich vorgenommen wurde, sollen jetzt die Ergebnisse des *CAP* Verfahrens näher beleuchtet werden. Diese sind in 27 abgebildet.
Die erste Zeile zeigt die *accuracy* für die kategorischen Variablen, was in diesem Fall nur die Variable *march2007* ist. Die zweite Zeile stellt den *mse* für die numerischen Merkmale dar. Links wird mit einer einzigen Schlüsselvariablen gestartet. Die Anzahl erhöht sich nach rechts. Für jeden Plot sind die Ergebnisse für alle Synthetisierungsverfahren dargestellt.
Dazugesagt werden muss, dass beispielsweise in der ersten Zeile im ersten Plot zwar nur ein Punkt für jedes Verfahren zu sehen ist, tatsächlich aber mehrere verschiedene *Ein-key-Szenarien* getestet wurden, die sich hier überschneiden. Betrachtet man beispielsweise in der zweiten Zeile den ganz rechten Plot, kann man bei der *Random Forest* Methode zwei Punkte für die Zielvariable *Last* erkennen.
Insgesamt wurden für jede dargestellte Anzahl an Schlüsselvariablen zehn zufällige Szenarien getestet, die sich teilweise auch überschneiden.
Kommt man nun zur Interpretation der Grafik, kann man über die nominale Variable nicht viele spannenden Aussagen treffen. Die Anzahl an Schlüsselvariablen scheint für die Reidentifikationswahrscheinlichkeit keine Rolle zu spielen, genauso wenig wie das gewählte Verfahren. Für alle möglichen Szenarien wird eine ungefähre *accuracy* von 64% vorhergesagt. Das heißt, dass der Intruder mithilfe einer, zwei oder drei Variablen, die Zielvariable *march2007* zu 64% richtig vorhersagen werden kann. Da die Zielvariable nur zwei Ausprägungen besitzt, ob eine Person im März 2007 Blut gespendet hat oder nicht und die Trefferwahrscheinlichkeit dadurch durch einfaches Raten bereits 0.5 beträgt ist der Wert 0.64 nur mäßig hoch.
Betrachtet man nun die *mse* Ergebnisse, sieht man zunächst, dass die Werte stark von der betrachteten Zielvariablen abhängen. Dies liegt unter anderem daran, dass die Zielvariablen unterschiedliche Wertebereiche haben und die Abweichungen somit unterschiedlich groß sind. Analysiert man die Variable *Last*, die mit Abstand die kleinsten *mse* Werte aufweist, was zum höchsten Reidentifikationsrisiko führt, erkennt man wiederum nur kleine Unterschiede bei den verschiedenen Synthetisierungsverfahren, genauso wie bei den verschiedenen Anzahlen an Schlüsselvariablen.
Normrank und *sample* haben etwas geringere Werte, während *norm*, *CART* und *Random Forest* sich durch minimal höhere Werte abzeichnen. Allerdings sind die Unterschiede so gering, dass dies auch zufällige Schwankungen darstellen könnten.
Auch bei den anderen Zielvariablen sind ähnliche Strukturen hinsichtlich der Anzahl der Schlüsselvariablen vorhanden.
Während *Last* im Wertebereich von 0.6 liegt, zeigt *Frequency* Werte um 0.9 und *First* Werte um 0.8, die sich bei Änderung der Anzahl der *keys* nicht ändern. Allerdings sind die Unterschiede zwischen den verschiedenen Synthetisierungsverfahren hier etwas deutlicher zu sehen. Vor allem bei der Variablen *Frequency* hebt sich *Random Forest* mit seinen deutlich höheren

Werten von den anderen Verfahren ab. Auch *sample* deutet hier auf ein etwas niedrigeres Reidentifikationsrisiko hin.

Verknüpft man die Ergebnisse der zwei Statistiken für das Reidentifikationsrisiko, kann gesagt werden, dass *Random Forest* das wohl geeignetste Verfahren ist. *Normrank* schneidet dagegen im Vergleich etwas schlechter ab als die anderen Verfahren.

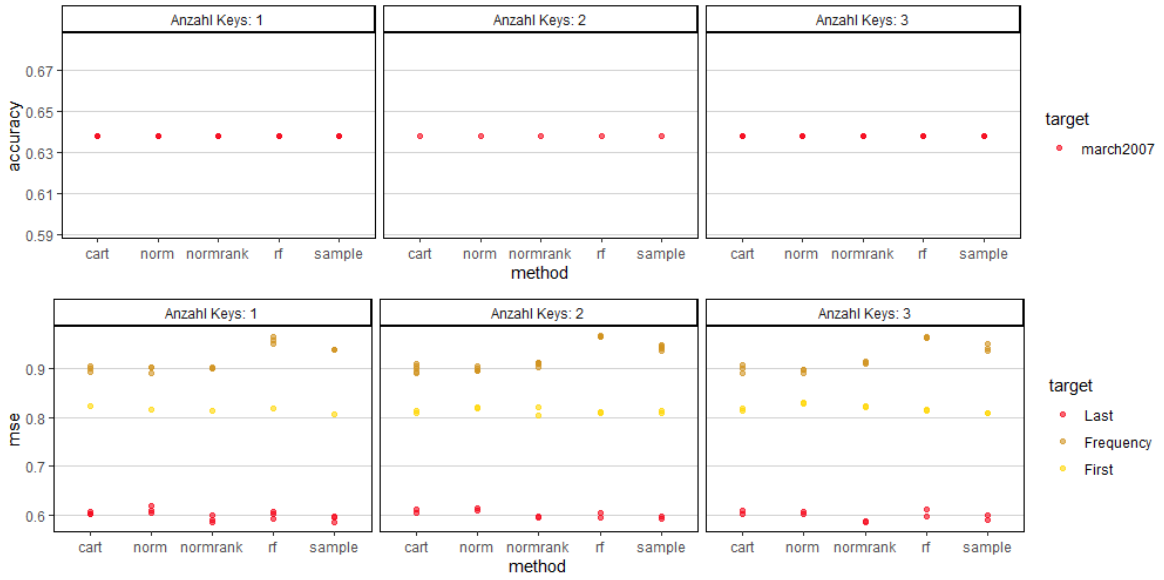


Abbildung 27: Vergleich der Güte der Vorhersage zwischen den verschiedenen synthetischen Datensätze und zwischen verschiedenen Anzahlen an key Variablen für den Bluttransfusionsdatensatz

7.3.2 Einkommen

nearest-neighbour	CART	Random Forest	sample	polyreg
min	0.043	0.084	0.062	0.083
mean	0.858	0.898	1.304	1.092

Tabelle 9: Der kleinste Nachbar im gesamten Datensatz (min). Der über alle Zeilen im synthetischen Datensatz gemittelte kleinste Nachbar (mean) für den Einkommensdatensatz

Tabelle 9 zeigt, dass *sample* die mit Abstand größte Distanz aufweist gefolgt von *polyreg*. *CART* und *Random Forest* weisen sehr ähnliche Werte auf und bilden das Schlusslicht. Auch hier lassen sich bei den minimalen Distanzen neben *CART* auch bei *sample* sehr kleine Distanzen erkennen. Dies unterstützt die beim vorherigen Datenbeispiel aufgestellte Vermutung, dass die *sample* Methode stark streut und nicht besonders robust zu sein scheint. Aus diesem Grund ist eindeutig das *polyreg* Verfahren zu favorisieren, welches für beide Metriken mit die besten Werte liefert.

Ob auch das *CAP* Verfahren zu dem selben Ergebnis führt, kann mithilfe von 28 beantwortet werden.

Beginnt man auch hier zunächst mit den kategorischen Variablen, kann man eine bisher nicht erschienene rosa Linie erkennen. Diese zeigt die *accuracy* über alle Zielvariablen gemittelt an. Anhand dieser ist unschwer zu erkennen, dass auch hier keine Unterschiede zwischen den Synthetisierungsverfahren zu erkennen sind. Für die verschiedenen Anzahlen an Schlüsselvariablen sieht man hingegen unterschiedliche Ergebnisse. Während eine Steigung der *accuracy* bei Erhöhung der Anzahl der Schlüsselvariablen von eins auf drei für den naiven Beobachter zu erwarten war, erstaunt die Abnahme der *accuracy* bei Erhöhung der *keys* auf fünf umso mehr.
Des Weiteren scheint sich die Variable *occupation* mit einem sehr geringen Reidentifikationsrisiko deutlich von den anderen Zielvariablen hervorzuheben. Grund hierfür könnte beispielsweise die hohe Anzahl an verschiedenen Kategorien innerhalb der Variablen sein. Im Gegensatz dazu scheint die Vorhersage der Zielvariablen *native country* für den *Intruder* sehr leicht zu machen sein.
An diesem Punkt sollte noch hinzugefügt werden, dass nicht nur die Schlüsselvariablen für jedes Szenario zufällig gewählt wurden, sondern auch die Zielvariablen. Dadurch kann es vorkommen, dass einzelne Zielvariablen in einem Plot vorhanden sind, während sie in anderen fehlen, da sie durch den Zufallsmechanismus nicht ausgewählt wurden. Dies tritt beispielsweise bei der eben betrachteten Variable *native country* auf, die in dem *5-key-Szenario* allem Anschein nach gar nicht gewählt wurde. Das ist auch die Erklärung dafür, dass die mittlere *accuracy*, dargestellt mit der rosa Linie, hier deutlich niedriger ist als in den zwei linken Plots. Ein direkter Vergleich zwischen diesen Linien ist also nicht unbedingt ratsam.
Allerdings zeigt dieses Ergebnis, dass die Anzahl der Schlüsselvariablen für den *Intruder* nicht unbedingt von zentralem Interesse ist. Vielmehr sind geringe Ausprägungsmöglichkeiten der Zielvariable für die Vorhersage sensibler Werte relevant.
Auch bei Betrachtung der metrischen Zielvariablen sind keine eindeutigen Ergebnisse zu sehen. Während *Random Forest* für die Variable *education* die besten Werte erzielt, schneidet das Verfahren für die variable *our per week* und *age* mit am schlechtesten ab.
Zusammenfassend kann gesagt werden, dass für den Einkommensdatensatz kein Verfahren zu favorisieren ist.

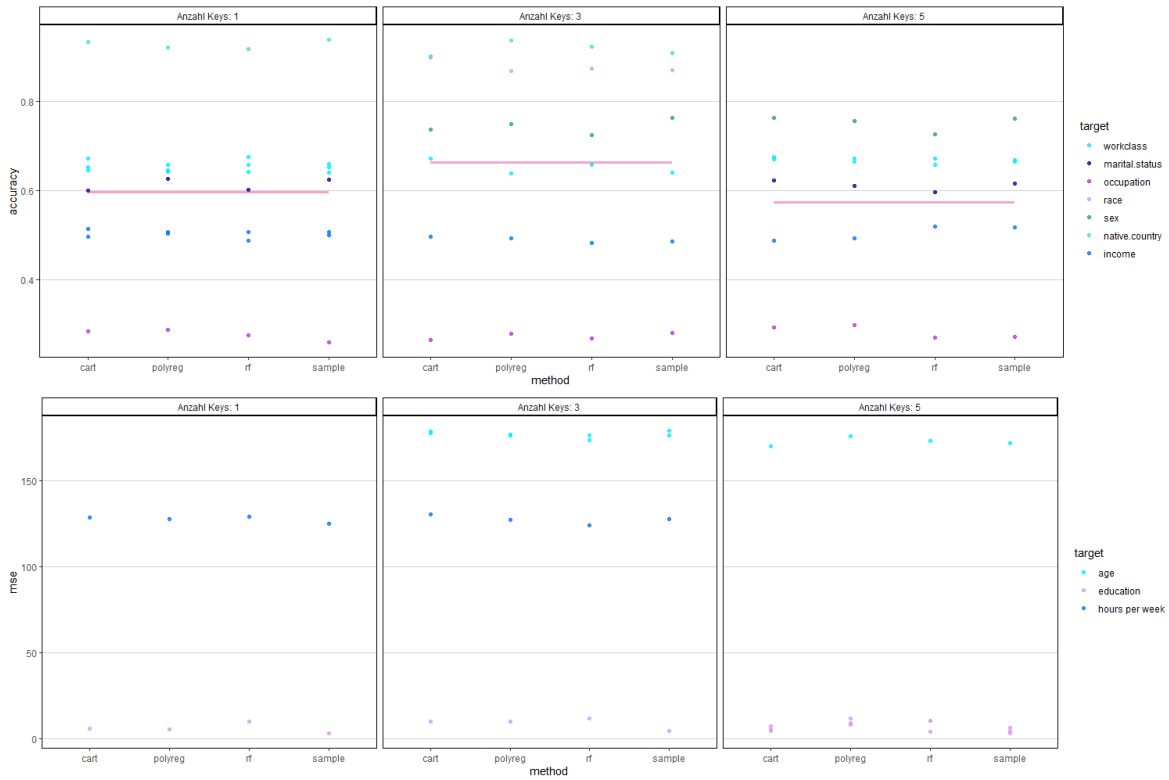


Abbildung 28: Vergleich der Güte der Vorhersage zwischen den verschiedenen synthetischen Datensätze und zwischen verschiedenen Anzahlen an key Variablen für den Einkommensdatensatz

7.3.3 Schulleistungen

nearest-neighbour	CART	Random Forest	sample	polyreg
min	1.999	1.463	2.469	1.925
mean	3.241	3.163	3.462	3.347

Tabelle 10: Der kleinste Nachbar im gesamten Datensatz (min). Der über alle Zeilen im synthetischen Datensatz gemittelte kleinste Nachbar (mean) für den Schulleistungsdatensatz

Zuletzt soll nach das Reidentifikationsrisiko für den Datensatz, der sich mit portugiesischen Schulleistungen auseinandersetzt, evaluiert werden.

Tabelle 10 zeigt ähnliche Ergebnisse wie das vorherige Datenbeispiel. Auch hier weisen die Verfahren *sample* und *polyreg* bei den Mittelwerten die besten Werte auf. Unterschiede ergeben sich hingegen bei der Betrachtung der minimalen Werte. Anders als zuvor bildet das *sampling* Verfahren hier die Spitze. *Random Forest* lässt im Gegensatz dazu das größte Reidentifikationsrisiko vermuten. Hier ist eindeutig die *sample* Methode zu favorisieren.

Auffällig bei diesem Datenbeispiel ist, dass die Werte im Allgemeinen sehr viel höher sind als bei den vorherigen Datenbeispielen, wodurch man schlussfolgern könnte, dass die Privatsphäre umso besser wird, desto mehr Variablen der Datensatz enthält.

Dies ist in diesem Fall auch plausibel, da die Wahrscheinlichkeit eine identische Zeile im synthetischen Datensatz zu finden automatisch sinkt, wenn mehr Variablen miteinander verglichen werden müssen. Des Weiteren ist das zur Berechnung verwendete Distanzmaß der euklidische Abstand, welcher nur positive Werte annehmen kann und somit bei Erhöhung der Variablenanzahl per Definition steigen muss.
Dies führt dazu, dass man die Ergebnisse der einzelnen Datensätze in diesem Fall nicht ohne Weiteres direkt miteinander vergleichen kann und derartige Schlüsse wie oben nicht ziehen kann.
Möglicherweise widersprechen die Ergebnisse des *CAP* Verfahrens 29 sogar oben angeführter These. Zunächst sollen hierfür die mittleren *accuracy* Berechnungen analysiert werden.
Hierfür betrachten wir wie bisher die rosa gekennzeichnete Linie. Während bei einer einzigen Schlüsselvariable noch kein Trend zu erkennen ist, sieht man bei drei Schlüsselvariablen einen leichten Knick nach unten bei dem *polyreg* Verfahren, welches hier das niedrigste Reidentifikationsrisiko ausweist. Dieses Ergebnis ist für die weiteren Szenarien jedoch nicht übertragbar. Für fünf Schlüsselvariablen bildet die Linie einen leichten Bogen, sodass *CART* und *sample* minimal bessere Ergebnisse liefern. Auch bei zehn *keys* scheint vor allem *sample* schlechte Vorhersagen und somit ein niedriges Reidentifikationsrisiko bieten zu können. Bei 15 Schlüsselvariablen ist wieder *CART* an der Spitze.
Allerdings sind auch hier die Abweichungen sehr gering, weswegen die Ergebnisse nicht überinterpretiert werden sollten. Allgemein befinden sich die mittlere *accuracy* im Bereich von 0.6 und ist somit der vom Einkommensbeispiel sehr ähnlich und etwas geringer als beim Bluttransfusionsbeispiel. Auf die Angabe der Namen der einzelnen Zielvariablen wurde hier der Übersicht halber verzichtet.
Bei den *mse* Betrachtungen sind ebenfalls keine eindeutigen Strukturen zu erkennen. Allerdings gilt auch hier wieder, dass nur zwei von 31 Variablen im Datensatz metrisch sind und sie somit nur eine untergeordnete Rolle für die Evaluierung des Reidentifikationsrisikos spielen.
Dass das Reidentifikationsrisiko bei diesem Datensatz wie durch die obige Tabelle anfangs vermutet niedriger ist als bei den anderen Datenbeispielen, konnte nicht bestätigt werden, weswegen die genannte *Trade-off* Problematik doch anzuzweifeln ist. Allerdings ist es auch nicht deutlich höher.

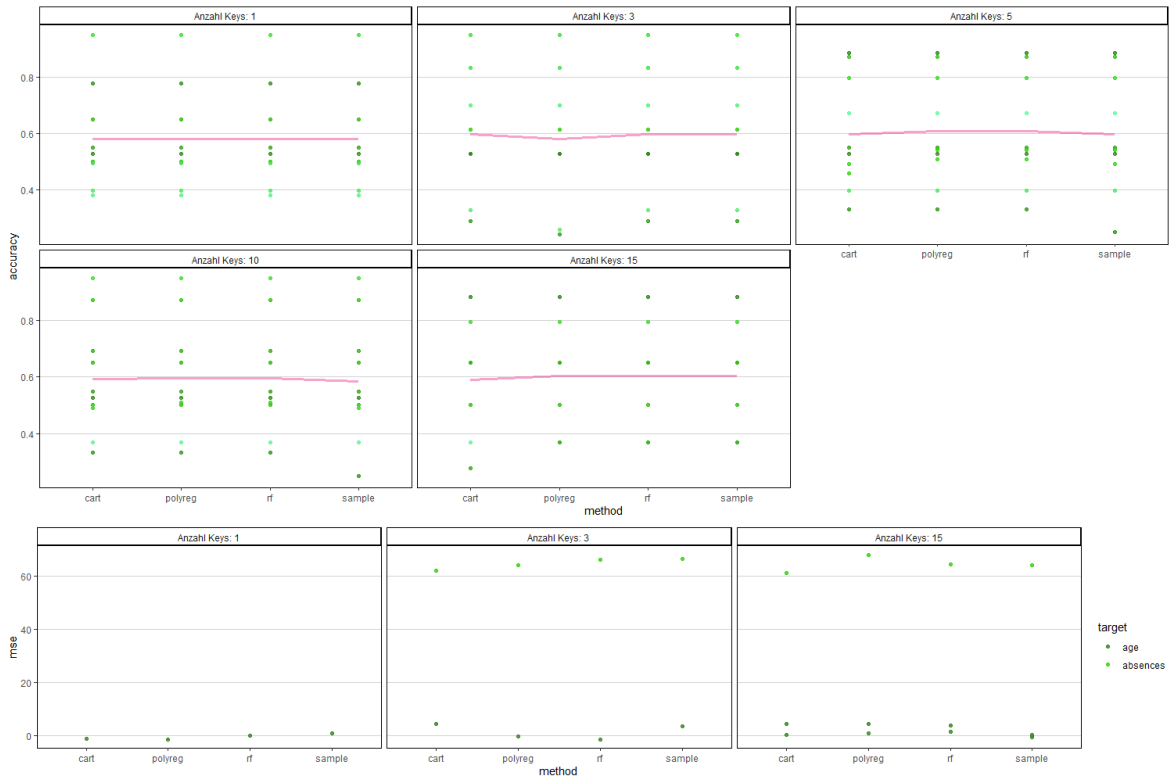


Abbildung 29: Vergleich der Güte der Vorhersage zwischen den verschiedenen synthetischen Datensätze und zwischen verschiedenen Anzahlen an key Variablen für den Schulleistungsdatensatz

Insgesamt lässt sich nicht sagen, dass ein Synthetisierungsverfahren für die Datenbeispiele hinsichtlich des Reidentifikationsrisikos besonders geeignet oder ungeeignet ist. Außerdem ist man zu der Erkenntnis gekommen, dass in erster Linie die gewählte Zielvariable und nicht die Anzahl der Schlüsselvariablen von großer Bedeutung für den *Intruder* sind.
Hinzu kommt, dass das Reidentifikationsrisiko enorm sinkt, wenn die Anzahl an Ausprägungen der Zielvariablen zunimmt. Es bleibt also zu hoffen, dass sensible Informationen mit einer hohen Anzahl an Ausprägungsmöglichkeiten einhergehen. Sind diese jedoch gering müssen möglicherweise weitere Anonymisierungsverfahren angewandt werden, um die Privatsphäre der befragten Personen gewährleisten zu können.

7.4 Vergleich der logistischen Regressionen

Nachdem in den vorherigen Kapiteln die synthetischen Datensätze hinsichtlich ihres Nutzens und ihres Reidentifikationsrisikos verglichen wurden, soll in diesem Kapitel ein statistisches Modell mit diesen Datensätzen gefittet werden. Dieses Modell soll auf den synthetischen Daten trainiert werden, um die Werte der Zielvariablen im Originaldatensatz vorherzusagen. Für eine robuste Evaluierung der Ergebnisse nutzt man Kreuzvalidierung. Die Güte des Modells soll dann nicht absolut, sondern relativ zu der Güte des Modells, welches auf den Originaldaten trainiert wurde, beurteilt werden.

Der Grund für die Notwendigkeit der Nutzung der Kreuzvalidierung ist, dass ohne diese die selben Daten sowohl zur Schätzung der Parameter als auch zur Beurteilung der Modelanpassung verwendet werden, was aufgrund von *overfitting* zu verfälschten Ergebnissen führen kann. Für das Originalmodell ist diese Problematik offensichtlich. Bei dem Modell, welches auf den Daten des synthetischen Datensatzes trainiert wurde, wird zusätzlich die Annahme getroffen, dass die Daten denen im Originaldatensatz sehr ähnlich sind und somit auch die nahezu selben Daten zur Schätzung wie zur Evaluierung genutzt werden.
Um dieser Problematik zu entgehen, wird Kreuzvalidierung genutzt, welche zum Ziel hat, die Berechnungen der Anpassungsgüte unabhängig von den Schätzungen der Parameter durchzuführen.
Dafür teilt man den vorhandenen Datensatz, welcher zunächst dem Originaldatensatz entspricht, in einen Trainingsdatensatz, der für die Parameterschätzung benutzt werden wird, und in einen Testdatensatz auf, der dafür verwendet werden wird, die Modellanpassung zu bewerten (16, S.110 ff.).
Um noch realistischere Ergebnisse zu erhalten, tut man dies nicht einmalig, sondern wiederholt den Vorgang k-mal, indem man den Datensatz anstatt in 2, in k zufällig erzeugte, gleich große Teildatensätze teilt, wobei immer ein Teildatensatz dem Testdatensatz entspricht und die restlichen Teildatensätze die Trainingdaten bilden, damit jeder Datenpunkt *k-1* mal als Trainingsdatenpunkt fungiert und ein mal als Testdatenpunkt (12, S.176-186).
Im vorliegenden Fall entspricht $k = 5$. Für jeden der fünf Testdatensätze wird dann die Modellgüte berechnet.
Anschließend mittelt man über die erhaltenen Ergebnisse und erhält einen realistischen Wert für die Güte der Modellanpassung der Originaldaten. Nachfolgend soll noch die Güte der Modellanpassung der synthetischen Datensätze bestimmt werden.
Der Testdatensatz entspricht dabei wieder einem der fünf gebildeten Teildatensätze des Originaldatensatzes. Der Trainingsdatensatz wird aus den Teildatensätzen des synthetischen Datensatzes erzeugt. Dabei wird derjenige Teildatensatz, der dem Analogon im Originaldatensatz entspricht, welcher gerade den Testdatensatz bildet, ausgespart.
Wird beispielsweise der dritte Teildatensatz der Originaldaten als Testdatensatz benutzt, so bilden der erste, der, zweite, der vierte und der fünfte Teildatensatz der Surrogatdaten den Trainingsdatensatz. Anschließend wird das Modell wie gewohnt trainiert und evaluiert.
Aufgrund dessen, dass in allen drei behandelten Datenbeispielen die Zielvariable kategorischer Natur ist, wird das Modell die logistische Regression 3.2 sein.

7.4.1 Bluttransfusionen

Wie schon öfters erwähnt, soll beim Bluttransfusionsbeispiel vorhergesagt werden, ob bzw. wie wahrscheinlich eine befragte Person im März 2007 Blut spendet oder nicht. Die Zielvariable

ist also *march2007*, wobei 0 dafür steht, dass die Person kein Blut gespendet hat, während 1 für eine Bluttransfusion steht. Die Prädiktorvariablen sind entsprechend *Last*, *Frequency* und *First*.

Die erste Tabelle 11 zeigt den AUC-Wert des Originalmodells auc_o und den AUC-Wert des auf den Surrogatdaten trainierten Modells auc_s für die verschiedenen Verfahren. Der AUC-Wert kann hier als Wahrscheinlichkeit interpretiert werden, dass ein Blutspender auch tatsächlich als solcher klassifiziert wird. Diese beträgt hier 74%.

Alle Verfahren bis auf *normrank* zeigen sehr ähnliche *AUCs* wie das Original. *Normrank* liegt mit einem Wert von 52% deutlich daneben. Hier käme die Vorhersage einfachem Raten gleich. Dies entspricht nicht den Ergebnissen des Originalmodells, weswegen dieses Verfahren als ungeeignet eingestuft werden kann.

Die zwei rechten Spalten der Tabelle geben an, wie viel Prozent der geschätzten Koeffizienten signifikant sind, wobei als Signifikanzniveau 0.05 gewählt wurde. Man kann erkennen, dass im originalen Modell alle Koeffizienten signifikant waren. Sehr große Abweichungen von diesem Ergebnis zeigen die Verfahren *norm* und *sample*. Vor allem *sample*, das nur eine durchschnittliche Signifikanz von 5% vorhersagt, scheint Probleme damit zu haben, die Originalschätzung nachzuahmen.

Ein nicht signifikanter Koeffizient im synthetischen Modell heißt dabei nicht, dass er einen anderen Wert hat als der gleiche Koeffizient im Originalmodell. Allerdings würde ein Analyst, der nur die synthetischen Daten vorliegen hat, möglicherweise gesondert darauf aufmerksam machen, dass seine Ergebnisse nicht signifikant sind, was nicht den Ergebnissen im Originalmodell entspricht und somit trotzdem eine Abweichung darstellt.

Die zweite Tabelle 12 zeigt die mittlere verhältnismäßige Änderung der Werte der einzelnen Koeffizienten zum Originaldatensatz und ob es bei der Vorhersage der Koeffizienten in den fünf Testdatensätzen zu einem Vorzeichenwechsel im Vergleich zu den Werten im Originalmodell kam. Ist das Verhältnis 1 bedeutet dies, dass der Wert des betrachteten Koeffizienten identisch zu dem im Originalmodell ist.

Hier muss darauf hingedeutet werden, dass die geschätzten Werte im originalen Modell bereits recht gering sind, weswegen eine hohe Verhältniszahl nicht unbedingt einer enormen Abweichung gleichkommt. Nichtsdestotrotz kann man einige Unterschiede bzgl. dieser Verhältnisse zwischen den Verfahren sehen.

Beim Intercept zeigt die Methode *norm* beispielsweise sehr hohe Abweichungen. Beim Koeffizienten der Variable *Frequency* ist *sample* sehr schlecht. Allgemein kann gesagt werden, dass $CART$ im Schnitt hier die besten Ergebnisse liefert.

Dieses Resultat lässt sich auch auf die Analyse der Vorzeichenwechsel übertragen. Bei $CART$ wird in keinem *Fold* bei keinem Koeffizienten das Vorzeichen vertauscht, was einer entgegengesetzten Wirkung der Zunahme des Wertes einer Variablen entspricht. Doch auch *normrank*

und *rf* können dieses Ergebnis aufweisen. *Sample* hingegen verschätzt sich bei dem Intercept und der Variable *Last* in jedem *Fold*.
Zusammenfassend kann gesagt werden, dass *CART* die besten Ergebnisse liefert und somit als Synthetisierungsverfahren favorisiert werden kann. Sowohl beim *AUC*, als auch bei der Signifikanz, den Verhältnissen und den Vorzeichenvergleichen konnte sich das Verfahren durch seine realitätsnahen Ergebnisse hervorheben. *Sample* und *norm* sind dagegen am ungeeignetsten, um den Originaldatensatz in der Modellierung nachzuahmen. Auch hier gilt, dass sie in allen betrachteten Metriken einheitlich am schlechtesten waren. *Random Forest* und *normrank* bilden das Mittelfeld.

method	auc_o	auc_s	sig_o	sig_s
CART	0.74	0.74	1	0.85
norm	0.74	0.73	1	0.45
normrank	0.74	0.52	1	0.80
rf	0.74	0.72	1	0.60
sample	0.74	0.74	1	0.05

Tabelle 11: Vergleich der Güte und des Signifikanzanteils der Koeffizienten des Originalmodells mit denen der synthetischen Modellen für den Bluttransfusionsdatensatz

method	verh_{Int}	verh_{Last}	verh_{Freq}	verh_{First}	vorz_{Int}	vorz_{Last}	vorz_{Freq}	vorz_{First}
CART	1.08	1.39	0.70	0.64	0	0	0	0
norm	0.12	1.25	0.36	0.20	5	0	0	5
normrank	1.12	1.58	0.69	0.47	0	0	0	0
rf	0.75	1.57	0.53	0.20	0	0	0	0
sample	0.43	0.20	0.11	0.22	5	5	0	0

Tabelle 12: Vergleich der zum Originalmodell verhältnismäßigen Koeffizientenänderung und der Vorzeichenänderung zwischen den synthetischen Modellen für den Bluttransfusionsdatensatz

7.4.2 Einkommen

Der zweite Datensatz befasst sich mit der Frage, ob eine befragte Person mehr oder weniger als 50 tausend US-Dollar im Jahr verdient. Die Zielvariable ist demnach *income*, die den Wert 1 annimmt, falls die Person mehr als 50 tausend verdient und 0, wenn nicht. Es gibt neun erklärende Variablen mit denen die Zielvariable im Folgenden vorhergesagt wurde.
Tabelle 13 gibt zunächst Auskunft über die Güte der Modelle. Das Originalmodell weist einen *AUC* von 85% auf, was bedeutet, dass eine Person, die mehr als den genannten Schwellenwert verdient auch tatsächlich vom Modell als solche identifiziert wird. Somit ist dieses Modell

bzgl. dieser metrik etwas besser als das Modell des ersten Datenbeispiels.
Die Güten der Surrogatdaten stimmen im Großteil mit denen des Originalmodells überein. Lediglich die Methode *sample* weist einen deutlich schlechteren AUC Wert auf. Hier gilt wieder, dass das Modell nur minimal besser zur Vorhersage der Zielvariablen geeignet ist, als einfaches Raten. Dieses Ergebnis ist nicht zufriedenstellend und schließt das Verfahren *sample* somit als geeignetes Synthetisierungverfahren in diesem Fall aus. Interessant ist auch, dass die anderen Verfahren bessere $AUCs$ aufweisen als das Originalmodell. Jedoch sind diese Unterschiede vernachlässigbar gering.
Ähnliche Ergebnisse zeigen die Signifikanzvergleiche. Auch hier gilt, während im Originalmodell lediglich 32% der Koeffizienten signifikant sind, sind es in den meisten Methoden um die 40%. Eine Ausnahme bildet hier das Verfahren *polyreg*, welches nur 4% der Koeffizienten signifikant schätzt. Das ist deutlich geringer als das Originalmodell und deutet somit auf Probleme bei der Schätzung mittels dieses Verfahrens hin. Am ähnlichsten zum Originalmodell ist hier die Methode *Random Forest*.
Betrachtet man sich nun die zweite Tabelle 14 an, ist gleich zu erkennen, dass die verhältnismäßige Änderung der Koeffizienten im Schnitt deutlich höher ist als im ersten Datenbeispiel. Vor allem *CART* und *Random Forest* heben sich mit Werten über neun bzw. sechs deutlich von den anderen Verfahren ab. *Polyreg* dagegen scheint vergleichsweise gute Schätzungen erlangt zu haben. Auch *sample* deutet zunächst auf ein gutes Modell hin.
Dies ändert sich hingegen, wenn man die Vorzeichenwechsel näher beleuchtet. Hier werden bei der zehnten und 13ten Variable in allen fünf *Folds* die Vorzeichen entgegengesetzt zum Original geschätzt, bei der vierten Variable sind es immer noch vier *Folds*, die falsche Vorzeichen vorhersagen.
Lediglich die 14te Variable liegt mit zwei *Folds* nicht deutlich daneben. Gute Ergebnisse hingegen zeigen die Verfahren *CART* und *polyreg*.
Es kann argumentiert werden, dass auch wenn die Verhältnisse der Koeffizienten bei *CART* nicht stimmen, zumindest die Richtung des Effekts des Koeffizienten richtig geschätzt wird. Sind die Werte der Koeffizienten zusätzlich nahe null ist eine große verhältnismäßige Änderung zum Originalmodell außerdem nicht gravierend.
Zuletzt soll angemerkt werden, dass die Tabelle nicht alle geschätzten Koeffizienten zeigt, da dies sehr unübersichtlich werden würde. Es wurden zufällig vier Koeffizienten gewählt und näher analysiert.
Zusammenfassend kann sich kein Verfahren deutlich hervorheben. Je nachdem auf was man besonders Wert legt sind unterschiedliche Methoden zu favorisieren. *Polyreg* ist dabei bis auf den Signifikanzvergleich immer mit das beste Verfahren. *Sample* muss vor allem bei der Modellgüte und den Vorzeichenvergleichen viel einbüßen, während *Random Forest* und *CART* vorwiegend bei den Koeffizientenvergleichen häufig daneben liegen.

method	auc_o	auc_s	sig_o	sig_s
CART	0.85	0.86	0.32	0.49
rf	0.85	0.86	0.32	0.36
sample	0.85	0.55	0.32	0.40
polyreg	0.85	0.87	0.32	0.04

Tabelle 13: Vergleich der Güte und des Signifikanzanteils der Koeffizienten des Originalmodells mit denen der synthetischen Modellen für den Einkommensdatensatz

method	$verh_{10}$	$verh_{13}$	$verh_{14}$	$verh_4$	$vorz_{10}$	$vorz_{13}$	$vorz_{14}$	$vorz_4$
CART	9.70	2.79	3.25	0.51	0	0	4	1
rf	2.26	1.98	0.24	6.09	2	0	3	1
sample	1.64	0.75	1.58	2.36	5	5	2	4
polyreg	1.77	0.84	1.3	1.70	1	0	1	1

Tabelle 14: Vergleich der zum Originalmodell verhältnismäßigen Koeffizientenänderung und der Vorzeichenänderung zwischen den synthetischen Modellen für den Einkommensdatensatz

7.4.3 Schulleistungen

Das letzte Datenbeispiel behandelt die Frage ob ein zufällig gewählter Schüler der Abschlussklasse einer portugiesischen Realschule die Klasse bestanden hat oder nicht. Bestanden wird mit einer 1 gekennzeichnet und durchgefallen mit einer 0. Es gibt 30 Prädiktoren, die auf diese Frage eine Antwort geben sollen.

Wie in Abschnitt 7.1.3 bereits angesprochen handelt es sich bei diesem Datenbeispiel um *unbalanced data*, da zum Glück der Großteil der Klasse das Jahr bestanden hat. Aufgrund der geringen Observationsanzahl konnten jedoch nicht wie in den vorherigen Datenbeispielen Unterstichproben des Originaldatensatzes gezogen werden, um das Verhältnis von Einsen und Nullen auszugleichen. Stattdessen wurde der Datensatz künstlich mithilfe des Bootstrapverfahrens vergrößert. Dieses Verfahren beinhaltet natürlich einen Zufallsprozess, weswegen es auch bei den AUC- und Signifikanzwerten des Originaldatensatzes zu leicht unterschiedlichen Ergebnissen gekommen ist. Diese Unterschiede sind jedoch minimal und müssen nicht weiter betrachtet werden. Der Hinweis dient nur dem besseren Verständnis der nachfolgenden Tabellen. Außerdem wurden auch hier der Übersicht halber nur 4 Koeffizienten zur näheren Betrachtung zufällig gewählt 15.

Das vorliegende Modell weist mit einem durchschnittlichen AUC von 62% die deutlich schlechteste Modellgüte auf. Hier werden nur 62% der bestandenen Schüler auch tatsächlich als solche klassifiziert. Dieses Ergebnis kann Anzeichen dafür sein, dass zu viele Prädiktovariablen in das Modell aufgenommen wurden und durch Variablenselektionsmechanismen reduziert werden

müssen.

Allerdings ist dies nicht Thema dieser Arbeit, weswegen darauf verzichtet wurde. Von zentralem Interesse ist der Vergleich der Modelle und nicht das Erreichen eines möglichst guten Originalmodells.

Vergleicht man also das Originalmodell mit den auf den Surrogatdaten erzeugten Modellen 1516, sieht man, dass die Modellgüte bei den Verfahren *CART* und *polyreg* überschätzt wird. *Sample* unterschätzt sie wieder und *Random Forest* trifft die Güte nahezu perfekt. Jedoch sind die Abweichungen hier für alle Verfahren minimal. Insgesamt können alle Methoden zufriedenstellende Ergebnisse vorweisen.

Der Prozentsatz der Signifikanzen beträgt bei dem Originalmodell im Schnitt ca. 25%, was erneut darauf hinweist, dass möglicherweise zu viele Variablen in das Modell aufgenommen wurden. *Sample* scheint diese Problematik nicht aufzuzeigen, da hier fast doppelt so viele Variablen signifikant zu sein scheinen. Auch *Random Forest* hat Schwierigkeiten mit dieser Schätzung. Hier sind es deutlich zu wenig Variablen.

Betrachtet man anschließend die verhältnismäßige Änderung der Koeffizientenschätzungen im Vergleich zum Originalmodell springt ein Wert besonders ins Auge. *Polyreg* schätzt den Wert der zehnten Variable 818 mal so groß wie das Originalmodell. Erstaunlich ist, dass die Schätzungen für die anderen Koeffizienten dafür recht gut sind. Doch nicht nur *polyreg* weist bei dem zehnten Koeffizienten Probleme auf. Auch die anderen Verfahren zeigen hier deutlich zu hohe Werte an. Insgesamt kann gesagt werden, dass vor allem *Random Forest* gute Schätzungen vorweisen kann.

Die Betrachtung der Vorzeichenwechsel sieht für dieses Verfahren hingegen nicht sehr gut aus. Hier zeichnet sich vor allem *polyreg* durch seine gute Schätzung aus.

Insgesamt sind die Ergebnisse im Vergleich zu den anderen Datenbeispielen deutlich schlechter. Wieder wirft das die Vermutung auf, dass umso mehr Variablen ein Datensatz besitzt, desto schwerer seine Ersetzung durch einen synthetischen Datensatz wird. Auch hier kann kein Verfahren als klarer Sieger hervorgehen. Sogar wenn man eine Priorisierung der vorgestellten Metriken vornimmt, zeichnet sich kein Verfahren deutlich ab.

method	auc_o	auc_s	sig_o	sig_s
CART	0.65	0.69	0.26	0.23
rf	0.62	0.63	0.32	0.21
sample	0.60	0.54	0.24	0.43
polyreg	0.62	0.69	0.22	0.17

Tabelle 15: Vergleich der Güte und des Signifikanzanteils der Koeffizienten des Originalmodells mit denen der synthetischen Modellen für den Schulleistungsdatensatz

method	verh53	verh10	verh45	verh56	vorz53	vorz10	vorz45	vorz56
CART	1.12	3.59	3.22	0.45	1	2	3	2
rf	0.02	6.60	0.11	1.01	2	3	3	1
sample	0.59	13.50	1.42	1.06	2	0	4	5
polyreg	1.03	818	1.39	0.25	0	3	0	3

Tabelle 16: Vergleich der zum Originalmodell verhältnismäßigen Koeffizientenänderung und der Vorzeichenänderung zwischen den synthetischen Modellen für den Schulleistungsdatensatz

7.5 Fazit

Der nachfolgende Abschnitt dient der Zusammenfassung der in Kapitel 7 erlangten Erkenntnisse. Dabei soll besonders im Vordergrund stehen, welche Synthetisierungsmethode für das jeweilige Datenbeispiel am besten geeignet ist und wie sich die Beispiele dabei voneinander unterscheiden.

Beginnend mit dem Bluttransfusionsbeispiel konnte festgestellt werden, dass bezüglich des Nutzens *CART*, *Random Forest* und *normrank* die besten Ergebnisse geliefert haben, während *norm* eher enttäuschend war.

Das Reidentifikationsrisiko wird dagegen zwar ebenfalls durch *Random Forest* am meisten verringert, allerdings weist *normrank* hier die schlechtesten Ergebnisse auf. Damit zeigt *normrank* hier deutlich die erwähnte *Trade-Off* Problematik zwischen Nutzen und Privatsphäre auf. *Random Forest* beweist hingegen, dass eine Erhöhung des Nutzens mit einer gleichzeitigen Senkung des Reidentifikationsrisikos möglich ist.

Bei der logistischen Regressionsmodellierung konnte sich vor allem das Verfahren *CART* mit seinen guten Ergebnissen hervorheben. Im Gegensatz dazu, haben die Methoden *norm* und *sample* ein weiteres Mal gezeigt, dass sie zur Synthese dieses Datenbeispiels nicht vorrangig geeignet sind. Möchte man sich für dieses Datenbeispiel auf ein Verfahren zur Erzeugung artifizieller Daten festlegen, sollte dies *Random Forest* oder *CART* sein.

Weiterführend sollen die Ergebnisse des Einkommenbeispiels zusammengefasst werden. Bei der Nutzenevaluierung konnte *CART* die realistischten Ergebnisse erzeugen, während *sample* und *Random Forest* größere Abweichungen vom Originaldatensatz gezeigt haben. *Polyreg* hat zwar insgesamt ebenfalls zufriedenstellende Ergebnisse erzeugen können, jedoch liegt dies daran, dass der Großteil der Variablen kategorisch war. Für metrische Variablen ist das Verfahren ungeeignet.

Bei der Analyse des Reidentifikationsrisikos ist kein Verfahren als deutlich bestes hervorgegangen. Auch bei näherer Betrachtung der Modellierungsresultate, lässt sich nicht klar sagen, welches Verfahren am geeignetsten ist. Allerdings weist beispielsweise *polyreg* meistens sehr gute Ergebnisse auf.

Insgesamt kann also gesagt werden, dass die Entscheidung, welches Verfahren das beste ist mit zunehmender Variablenanzahl erschwert wird. Dies wird im dritten Datenbeispiel ebenfalls deutlich. Nichtsdestotrotz sprechen tendenziell die Verfahren *CART* und *polyreg* minimal mehr dafür verwendet zu werden.
Abschließend werden die Synthetisierungsverfahren für das Schuldatenbeispiel verglichen. Beginnend mit dem Nutzen, kann gesagt werden, dass auch hier *polyreg* für die Prädiktion metrischer Variablen am schlechtesten gewesen ist. Besonders bei den Kontigenztabellen heben sich wieder *CART* und *Random Forest* durch ihre positiven Ergebnisse hervor.
Bei der Betrachtung des Reidentifikationsrisikos konnte zwar nicht festgestellt werden, welches Verfahren am geeignetsten ist, jedoch haben die Ergebnisse gezeigt, dass nicht die Anzahl der Schlüsselvariablen sondern deren Wahl entscheidend ist.
Auch die Modellierung war bei keinem Verfahren deutlich besser oder schlechter. Die Entscheidung ist hier noch subjektiver als beim Beispiel davor, was wieder zeigt, dass die Evaluierung der Surrogatdaten mit steigender Variablenanzahl schwerer wird. Je nach Variable ist das ein oder andere Verfahren besser geeignet was im Schnitt zu gleich guten Ergebnissen führt.
Zusammenfassend kann gesagt werden, dass im Besonderen *CART* und *Random Forest* für alle Datenbeispiele zur Synthetisierung herangezogen werden können. Vor allem *norm* scheint eher ungeeignet, was vermutlich daran liegt, dass dieses Verfahren nur lineare Zusammenhänge aufdecken kann. Außerdem führt eine Erhöhung der Variablenanzahl nicht notwendigerweise zu einem hohen Nutzen oder einem niedrigeren Reidentifikationsrisiko. Ganz im Gegenteil scheint der Nutzen der Daten mit steigender Variablenanzahl zu sinken. Die Modellierung hat gezeigt, dass eine Erhöhung der Variablenanzahl den Verfahren mehr Spielraum für Fehler bietet. Demzufolge sind besonders kleine bis mittlere Datensätze einfach zu synthetisieren.

8 Schluss

Die Resultate dieser Arbeit haben gezeigt, dass die Synthetisierung von Daten insbesondere für Modellierungszwecke eine durchaus zufriedenstellende Alternative zu klassischen Anonymisierungsverfahren darstellen kann.
Das Vorgehen dabei ist nicht allzu schwer zu erlernen und in den meisten Programmen bereits zum Teil implementiert.
Enthält der Datensatz besonders sensible Werte kann es hingegen manchmal nicht ausreichend sein, diesen nur zu synthetisieren. In diesem Fall müssen weitere Verfahren wie beispielsweise *differential privacy* angewandt werden, um einen ausreichenden Schutz für die Studienteilnehmer gewährleisten zu können. Allerdings muss hinzugefügt werden, dass dies vermutlich auch bei klassischen Anonymisierungsverfahren notwendig sein wird.
Ob die Datensynthese tatsächlich klassische Anonymisierungsverfahren übertrifft, kann man mit weiteren Vergleichen, die sich auf die Unterschiede zwischen den herkömmlichen Methoden und der Synthetisierung fokussieren werden, beantworten. Insbesondere das Reidentifikationsrisiko bei den Surrogatdaten kann dann besser beurteilt werden. In dieser Arbeit ist ein Vergleich zwischen den einzelnen Synthetisierungsverfahren bzgl. des Reidentifikationsrisikos möglich. Allerdings kann aufgrund fehlender Vergleichswerte, die beispielsweise klassische Störmethoden bieten könnten, keine globale Aussage darüber getroffen werden, wie gut die Privatsphäre tatsächlich geschützt ist.
Die Modellierung mit den Surrogatdaten führt großteils zu sehr ähnlichen Ergebnissen wie die Modellierung mit den Originaldaten. Hier müssten Analysten, die die artifiziellen Daten erzeugen, möglicherweise verschiedene Verfahren vergleichen, um die für diesen Datensatz passendste Methode wählen zu können.
Es bleibt zu hoffen, dass sich in Zukunft mehr Analysten trauen, statistische Modelle zur Anonymisierung zu nutzen, als nur einfache Störungsmethoden.

Literatur

[1] Behnke, J.: *Logistische Regressionsanalyse: Eine Einführung.* Springer-Verlag, 2014

[2] Berkeley, School of I.: *Keeping Secrets: Anonymous Data Isn't Always Anonymous.* `https://ischoolonline.berkeley.edu/blog/anonymous-data/`, 2014. – Eingesehen am 17.09.2020

[3] Breiman, L. ; Friedman, J. ; Olshen, R. ; Stone, C.: *Classification and Regression Trees.* Belmont : Wadsworth, 1984

[4] Cortez, P. ; Silva, A.: Using data mining to predict secondary school student performance. (2008)

[5] Domingo-Ferrer, J. ; Montes, F.: *Privacy in Statistical Databases: UNESCO Chair in Data Privacy, International Conference, PSD 2018, Valencia, Spain, September 26–28, 2018, Proceedings.* Bd. 11126. Springer, 2018

[6] Drechsler, J. ; Jentzsch, N.: Synthetische Daten: Innovationspotential und gesellschaftliche Herausforderungen. (2018)

[7] Drechsler, J. ; Reiter, J.: An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. In: *Computational Statistics & Data Analysis* (2011), S. 3232–3243

[8] Dua, D. ; Graff, C.: *UCI Machine Learning Repository.* `http://archive.ics.uci.edu/ml`. Version: 2017

[9] Handl, A. ; Kuhlenkasper, T.: *Multivariate Analysemethoden: theorie und praxis mit R.* Springer-Verlag, 2017

[10] Hastie, T. ; Tibshirani, R. ; Friedman, J.: *The elements of statistical learning: data mining, inference, and prediction.* Springer Science & Business Media, 2009

[11] Hittmeir, M. ; Ekelhart, A. ; Mayer, R.: Utility and Privacy Assessments of Synthetic Data for Regression Tasks. In: *2019 IEEE International Conference on Big Data (Big Data)* IEEE, 2019, S. 5763–5772

[12] James, G ; Witten, D ; Hastie, T ; Tibshirani, R: *An introduction to statistical learning.* Bd. 112. Springer, 2013

[13] Loh, W.: Classification and regression trees. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1 (2011), Nr. 1, S. 14–23

[14] Nowok, B. ; Dibben, C. ; Raab, G.: Recognising real people in synthetic microdata: risk mitigation and impact on utility. (2017)

[15] Nowok, B. ; Raab, G. ; Dibben, C. u. a.: synthpop: Bespoke creation of synthetic data in R. In: *J Stat Softw* 74 (2016), Nr. 11, S. 1–26

[16] Ruppert, D ; S.Matteson, D: *Statistics and Data Analysis for Financial Engineering*. Springer, 2011

[17] Schlittgen, R.: *Multivariate Statistik*. Walter de Gruyter, 2011

[18] Strobl, C. ; Malley, J. ; Tutz, G.: *An introduction to recursive partitioning*

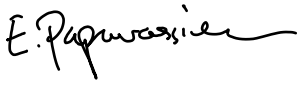
[19] Sweeney, L.: Computational disclosure control for medical microdata: The Datafly system. In: *Record Linkage Techniques 1997: Proceedings of an International Workshop and Exposition*, 1997, S. 442–453

Selbständigkeitserklärung

Ich versichere hiermit, die vorliegende Arbeit mit dem Titel

Methoden zur Daten-Synthetisierung

selbständig verfasst zu haben, keine anderen als die angegebenen Quellen und Hilfsmittel verwendet zu haben und die den benutzten Quellen entnommenen Passagen als solche kenntlich gemacht zu haben.



Eleftheria Papavasiliou

München, den 16. Oktober 2020