HOCHSCHULE COBURG

Hochschule für angewandte Wissenschaften
Coburg

Institut für Informatik

Projektbericht Datamining

Michael Krasser

Betreuer: Dr. Detlef Bittner
Abgabe des Berichts: 22 Januar 2019

Coburg, 8. Januar 2019

Inhaltsverzeichnis

1 Geschäftsverständnis

1.1 Beschreibung der Situation

Das zu untersuchende System beschreibt einen Onlineshop dessen Produktpalette aus Medien wie z.B. CDs, Bücher, Hörbücher, ebooks und ebook-Readern besteht. Um gegenüber großen Online-Anbietern wettbewerbsfähig zu sein muß der Onlineshop regelmäßig Maßnahmen zur Aquise und Kundenbindung ergreifen. Eine Möglichkeit zur Kundenbindung besteht in einer Gutscheinausstellung.

1.1.1 Problembeschreibung

Viele Kunden tätigen meist nur eine Bestellung im Onlineshop. Es kann keine Vorhersage darüber getroffen werden ob ein Kunde eine weitere Bestellung tätigen oder sein Interesse an den Produkten des Onlineshops erlischt. Das Ziel dieser Arbeit besteht in der Bestimmung der Loyalität des einzelnen Kunden anhand von Merkmalen die sich aus den beigelegten Datensätzen erschließen lassen. Ist die Wahrscheinlichkeit sehr gering, dass ein Kunde einen weiteren Einkauf tätigt, erscheint es sinnvoll diesen mittels eines Gutscheins an den Onlineshop zu erinnern. Ist dagegen die Wahrscheinlicht eines erneuten Einkaufs eher hoch, so verursacht die Zusendung eines Gutscheins nur unnötige Kosten.
Um dieses Problem zu lösen, soll mittels des CRISP-DM Modells [1] und des Datamining-Tools SPSS von IBM eine Bewertung durchgeführt werden, welche die Loyalität des Kunden bestimmen soll. Hierfür wurden zwei *.txt-Dateien über Kundendaten geliefert. Als Lösung wird eine Abbildung der Kundennummer auf das zu erwartende Kaufverhalten gefordert.

1.2 Situationsbewertung

1.2.1 Beschreibung der gelieferten Daten

Bei den zur Verfügung gestellten Daten handelt es sich um Auszüge aus den Bestelldaten einer Kundendatenbank. Es gibt Spalten für das Datum einer Lieferung, Accounterstellung und der ersten Bestellung, Domain und Kundennummer des Kunden und eine Reihe von Flags, die das Kaufverhalten des Kunden beschreiben: Es wird registriert ob gebrauchte oder importierte Artikel gekauft wurden, wie die Ware zum Kunden gelangt ist, wie der Kunde bezahlt hat, ob die Lieferung zurückgeschickt oder gecancelt wurde und wie viele Artikel der Kunde auf einmal gekauft hat.

1.2.2 Risikofaktoren

Das Risiko des Onlineshops liegt in der Versendung zu vieler Gutscheine. Diese Gutscheine können dann auch Kunden erreichen, die auch ohne diese wieder eine Bestellung getätigt hätten. In diesem Fall verliert der Shop pro Bestellung 5 €. Je nachdem, wie

viele der Gutscheine unnötig ausgegeben werden, kann sich diese Vorgehensweise als unrentabel erweisen. (Einfluß des falsch positiven und des falsch negativen Fehlers bleiben unberücksichtig.)

1.2.3 Erfolgskriterien

Es konnten drei zentrale Erfolgskriterien für das Projekt identifiziert werden:

- Entscheidung ob ein Kunde einen Gutschein erhält
- Maximierung des Gewinns im Datensatz dmc2010 class.txt
- Ab einem Gewinn von 9.310,50€ im Testdatensatz und 8547,50€ im class-Datensatz liegt ein erfolgreiches Datamining vor

Bemerkung:
Zur Berechnung des Gewinnes (bzw. des Verlustes) wurde folgende Entscheidungsmatrix vorgeschrieben:

	Kein Wiederkäufer	Wiederkäufer
Kein Gutschein	0	0
Gutschein	1.5	-5

Jeder Gutschein, der an eine Person ausgegeben wurde, die ansonsten nicht wieder bestellen würde, ergibt einen Gewinn von 1,50 € für den Onlineshop. Jeder Kunde der fälschlicherweise einen Gutschein bekommt, bedeutet einen Verlust in Höhe des Wertes des Gutscheins, also 5 €.
Als Referenzwert für die Modellbildung wird folgende Situation zugrunde gelegt: Jedem Kunden wird ein Gutschein zugesand.Unter der Annahme, dass jeder dieser Kunden auch wieder eine Bestellung tätigt, erhält man im Testdatensatz (26377 Kunden tätigten keine weitere Bestellung innerhalb von 90 Tagen, 6051 aber schon) einen Verlust von

$$(26377 * 1.50 - 6051 * 5)\,€ = 9310,50\,€$$

Diese Werte müssen in der Modellbildung übertroffen werden.

1.3 Projektplanung

Für die Projektdurchführung wurde zu Projektbeginn der Aufwand in Wochen für die jeweiligen Projektschritte des CRISP-DM Modells geschätzt:

- Geschäftsverständnis: ca eine Woche
- Datenverständnis: ca eine Woche
- Datenvorbereitung: ca zweieinhalb Wochen
- Modellbildung: ca zwei Wochen

- Evaluierung: ca eine halbe Woche
- Bereitstellung: ca eine halbe Woche

2 Datenverständnis

2.1 Datensammlung

Einige Attribute des Datensatzes sind weniger erfolgsversprechend als andere. Da bereits eine Beschreibung der einzelnen Attribute zugeliefert wurde, kann an dieser Stelle eine Vorsortierung der Spalten erfolgen. Um versehentliche Korrelationen mit unwichtigen Daten im Ergebnis der Untersuchung zu vermeiden, sollten diese vor der Betrachtung aussortiert werden.

2.1.1 Vernachlässigbare Felder

Unwichtige Felder lassen sich anhand verschiedener Merkmale identifizieren. Bei einer auffallend schlechten Datenqualität sollte das Feld aussortiert werden(fehlende Einträge, Ausreisser). Außerdem existieren Felder, anhand deren Beschreibung bereits erkannt werden kann, dass diese keinen Einfluß auf das Ergebnis haben können. Als solche Felder wurdendie im Folgenden beschriebenen identifiziert:

- „salutation“
- „domain“
- „model“
- „invoicepostcode“
- „delivpostcode“
- „advertisingdatacode“
- „points“
- „shippingcosts“
- „weight“
- „used“

Die Anrede („salutation“) wurde zu Beginn als relevant eingestuft, in einem späteren Analyseschritt aber entfernt: Email-Domain oder Anrede eines Kunden haben sicher wenig bis gar keinen Einfluß auf das Kaufverhalten des Kunden. Die Bedeutung des Feldes „model“ konnte nicht schlüssig geklärt werden, daher wurde es auf Grund des vernachlässigbaren Prädikatoreinflusses entfernt. Die Rechnungsadresse („invoicepostcode“) und die Lieferadresse („delivpostcode“) wurden auch verworfen da hier keine Rückschlüsse auf das Kaufverhalten des Kunden erkennbar sind. Außerdem erwies sich die Datenqualität des Feldes „invoicepostcode“ als schlecht wegen vieler NULL-Werte. Der Werbecode („advertisingdatacode“) wurde auf Grund schlechter Datenqualität entfernt. Punkte eingelöst („points“) schien zunächst relevant, da es Aufschluss über die Empfänglichkeit des Kunden für Werbeaktionen gibt, erwies sich aber als nicht hilfreich und wurde in einem späteren

Schritt auf Grund der der zu schlechtene Datenqualität entfernt. Spezifische Artikelinformation wir Versankosten („shippingcosts“) Gewicht („weight“) und Second Hand („used“) wurden auch entfernt, da sich hieraus keinerlei Information ableiten ließ.
Zusätzlich wurden die Produktkategorien w0 - w10 gleich zu Beginn als uniteressant für die Modellbildung eingestuft.

2.1.2 Erfolgsversprechende Felder

Bewertet wurde die Datenqualität, die Datenplausibilität und der Vorteil, den die Daten für die Auswertung ergeben könnten. Das Feld „datecreated“ kann in Verbindung mit dem Feld „date“ dazu genutzt werden, um die Zeit zwischen der Accounterstellung und der ersten Bestellung zu berechnen. Daher werden beide Attribute beibehalten. Auch aus den Feldern „cancel“, „deliverydatepromised“, „deliverydatereal“ etc. können zusammengesetzte Felder ermittelt werden. Zusätzlich sollten die Spalten „gift“, „voucher“ und „newsletter“ für die Auswertung verwendet werden, da sie Aufschluss über das Kaufverhalten des Kunden geben: Durch das Feld „newsletter“ kann beispielsweise erkannt werden, ob der Kunde generell über Interesse an den Produkten des Shops verfügt.

2.2 Datenbeschreibung

2.2.1 Datenmenge

Im Trainingsdatensatz und dem Vorhersagedatensatz befinden sich 32.428 bzw. 32.427 Datensätze. Datenmengen dieser Größe sind für eine vollständige Analyse im Sinne von Big Data nicht ausreichend, was sich auch in der späteren Modellierung der Daten zeigte.

2.2.2 Untersuchung der Daten

Jede Zeile der Dateien liefert Informationen über eine Bestellung. Jede Bestellung lässt sich eindeutig einem Kunden mittels des Feldes „customernumber“ zuordnen. Anhand der angegebenen Werte der Bestellung lässt sich feststellen, ob die versprochene Lieferzeit eingehalten wurde, welche Artikel bestellt wurden und wie sich diese zusammensetzen. Weitere Daten sind indirekt enthalten und lassen sich aus vorhandenen Daten ermitteln.

2.2.3 Konsistenz

Felder mit schlechter Datenqualität (fehlende Werte, NULL-Werte, viele Ausreisser,...) sollten unter Verwendung eines Filters entfernt werden. Hierzu gehören die Felder „ points „und „advertisingdatacode“ (viele fehlende Werte), „invoicepostcode“ und „deliverydatereal“ (NULL-Werte). Jedoch mussten diese beiden Felder zunächst für die Datenvorbereitung zur Erzeugung neuer Attribute beibehalten werden.

3 Vorbereitung der Daten

3.1 Bereinigung der Daten

Die Bereinigung der Daten erfolgte in zwei Schritten:
Zunächst wurden alle NULL-Werte aussortiert und danach Extremwerte (Ausreißer). Die Wertemenge einiger Felder enthielt viele NULL-Werte welche bei der späteren Modellierung hinderlich sein könnten. Entsprechend wurden diese Werte (z.B bei „deliverydatereal“ und „deliverydatepromised“) mit einem Auswahl-Knoten des SPSS Modelers entfernt, was eine deutliche Reduzierung der Datenmenge nach sich zog: Etwa 7000 Datensätze wurden entfernt. Unrealistische Werte wurden ebenfalls mit dem Auswahl-Knoten entfernt.

3.2 Erstellung neuer Felder

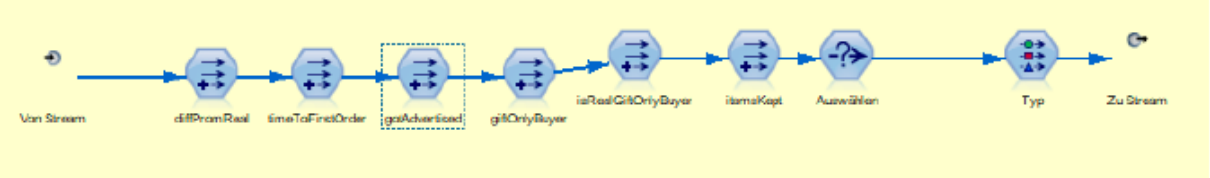

Abbildung 1: Erstellte Felder

Durch studieren der Situation konnten neue Felder aus bestehenden Feldern gebildet werden:

```
diffPromReal = deliverydatepromised - deliverydatereal
```

Differenz aus versprochenem und tatsächlichem Lieferdatum: Liegt das tatsächliche Lieferdatum nach dem versprochenen ist die Wahrscheinlichkeit recht groß den Kunden zu verlieren.

```
timeToFirstOrder = datecreated - date
```

Zeitdauer zwischen Registrierung des Kunden und seiner ersten Bestellung.

```
gotAdvertised = newsletter OR voucher
```

Aufschluss ob Kunde an Werbemaßnahmen teilnimmt.

```
giftOnlyBuyer = gift AND (numberitems == 1)
```

Aufschluss ob Kunde ein Geschenk bestellt hat.

```
isRealGiftOnlyBuyer = (giftOnlyBuyer == 1) AND (date == datecreated)
```

Aufschluss ob Kunde sich nur für die Bestellung eines Geschenks registriert hat.

```
itemsKept = numberitems - remi - cancel
```

Aufschluss über die Anzahl der Artikel die der Kunde bestellt und auch behalten hat.

3.3 Bewertung der Felder

Nachdem neue Felder erstellt wurden, muss der Einfluss dieser auf die Modellierung bewertet werden. Hierfür wurde nach der Datenvorbereitung ein Knoten zur Merkmalauswahl angefügt.
Dabei ergab sich, dass 10 der 19 Felder als bedeutsam (100 %) eingestuft wurden, ein weiteres Feld besaß eine Korrelation von über 90 %, die restlichen wurden als unbedeutsam (Korrelation von unter 90%) eingestuft. Korrelationswerte sollten lediglich als Richtlinie betrachtet werden, da die Gewichtung vieler Attribute bei 100% lag. Unter den 10 bedeutsamen Feldern befanden sich auch einige der generierten:

„itemsKept“ „gotAdvertised“

Damit stehen für die Modellierung zehn Felder zu Verfügung.

4 Modellierung

Als Grundlage für die Modellierung werden die unter 1.2.3 ermittelten Referenzwerte verwendet. Ziel der Modellierung ist eine Maximierung des Gewinns, also ein Modell zu entwickeln, dessen Gewinn die berechneten Gewinne übertrifft.

4.1 Wahl der Modellierungsknoten

Die Erklärungen zu den einzelnen Knoten sind [2] entnommen und können dort nachgelesen werden. Hinsichtlich der Struktur der Aufgabe kann man sich auf Baummodelle bei der Wahl der Knoten im SPSS Modeler beschränken. Hierfür wurden folgende Modellierungsknoten in Betracht gezogen:

- CHAID
- C5.0
- Random-Trees
- XGBoost-Baum
- Tree-AS
- C&R-Tree

Anfänglich wurden alle von SPSS Modeler zu Verfügung gestellten Moellierungsknoten in Betracht gezogen, aber nicht alle lieferten brauchbare Ergebnisse: Die Support Vector Machine (SVM) lieferte viele NULL-Werte.
Hinsichtlich der Validierung wurde zusätzlich überprüft, ob sich die getroffene Merkmalauswahl positiv auf den erzielten Gewinn auswirken konnte. Dabei konnte erkannt werden, welche Felder welchen Einfluß auf die Modellierung besitzen.

4.2 Test-Design

Ziel der Testausführung ist eine automatisierte Auswertung der verwendeten Modellierungsknoten aus den Datensätzen: Das jeweilige Modell-Nugget erstellt das Feld $R-Target90$ aus dem class Datensatz welches anschließend mit dem realen Feld target90 verglichen werden soll. Zu diesem Zweck wird die Entscheidungstabelle modifiziert:

target90	$R-target90$	Betrag
0	0	1.5
0	1	0
1	0	-5
1	1	0

Der Wert 0 bedeutet in diesem Kontext, dass ein Gutschein versendet wird.

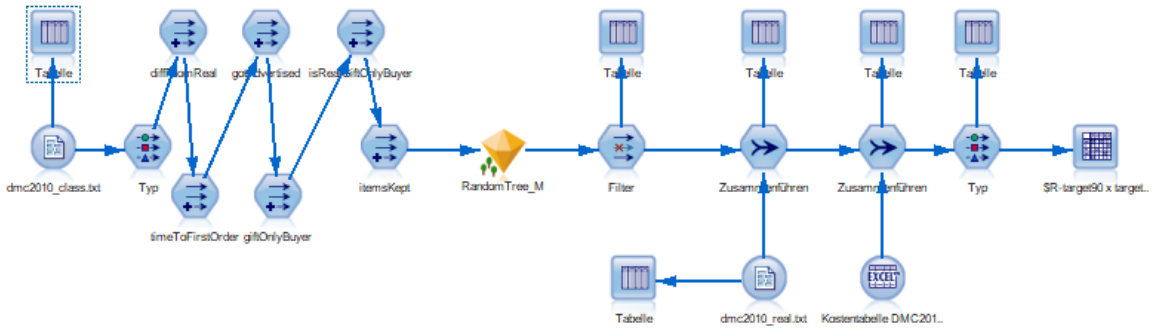

Abbildung 2: Testdesign

4.2.1 CHAID

Der CHAID-Knoten generiert Entscheidungsbäume unter der Verwendung der χ^2 - Verteilung. Im Gegensatz zum C&R-Baum und QUEST können mit diesem Modell auch nicht binäre Bäume generiert werden, d.h. Bäume mit mehr als zwei Verzweigungen. Berücksichtigt wurde das Verhalten des Gewinnes unter Merkmalauswahl und Boosting auf dem class Datensatz.

	Boosting	$\overline{\text{Boosting}}$
Merkwahlauswahl	8,548.50 €	8,575.00 €
$\overline{\text{Merkwahlauswahl}}$	8,697.00 €	8,548.50 €

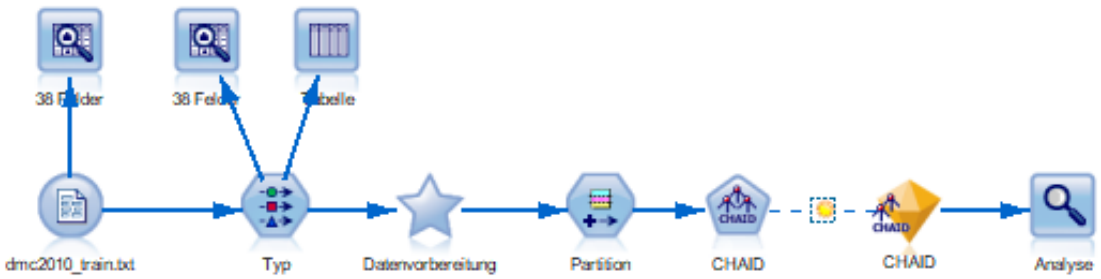

Abbildung 3: CHAID - ohne Merkmalauswahl

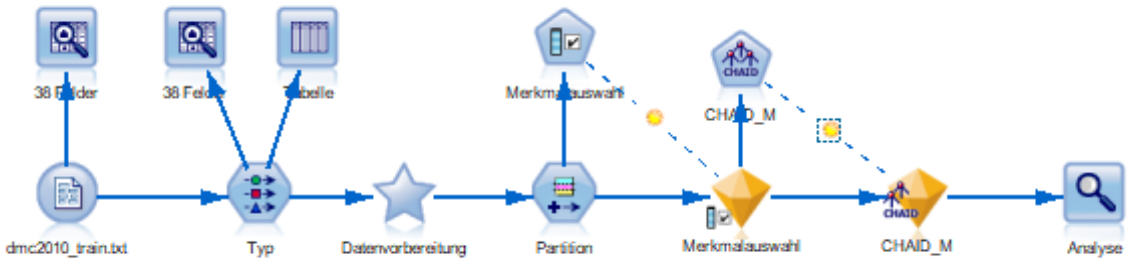

Abbildung 4: CHAID - mit Merkmalauswahl

Bemwerkenswert hierbei ist, dass Verwendung der Merkmalauswahl ein schlechteres Ergebnis produziert als ohne diese.

4.2.2 C5.0

Erklärung:

Der C5.0-Algorithmus, erstellt einen Entscheidungsbaum oder ein Regelset. Ein C5.0-Modell teilt die Stichprobe auf der Basis des Felds auf, das den maximalen Informationsgewinn liefert. Jede durch die erste Aufteilung definierte Teilstichprobe wird anschließend wieder aufgeteilt, üblicherweise auf der Grundlage eines anderen Felds. Der Prozess wird so lange fortgesetzt, bis die Unterstichproben nicht weiter aufgeteilt werden können. Zum Schluss werden die Aufteilungen der untersten Ebene noch einmal untersucht, wobei solche entfernt oder reduziert werden, die nicht wesentlich zum Wert des Modells beitragen.

Mit dem C5.0-Algorithmus konnten auf dem class Datensatz 9,055.00 € Gewinn errechnet werden, also eine deutliche Verbesserung gegenüber der Vorhersage des CHAID-Knotens.

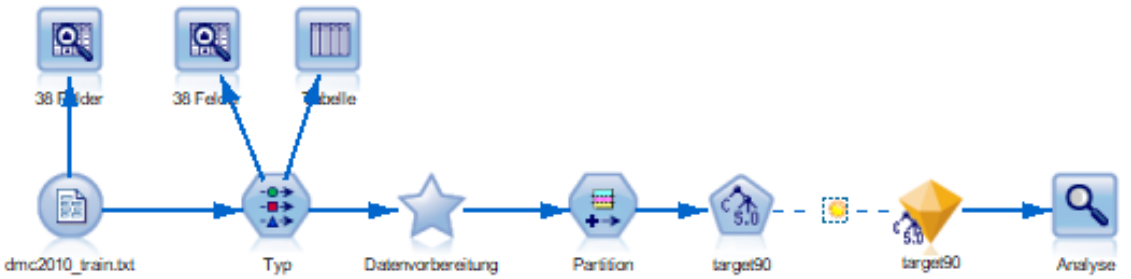

Abbildung 5: C5.0 - ohne Merkmalauswahl

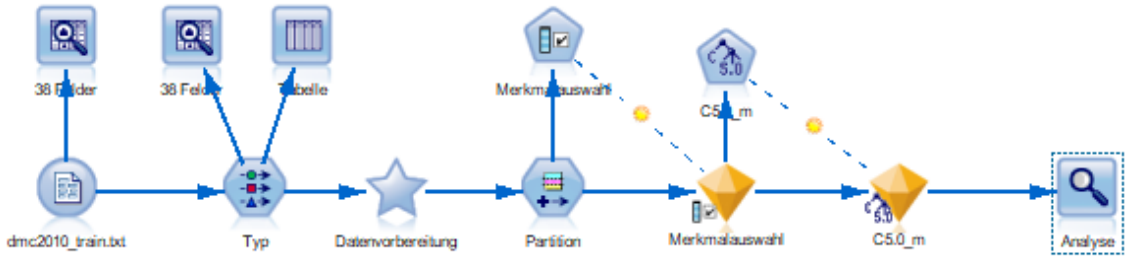

Abbildung 6: C5.0 - mit Merkmalauswahl

Auswirkung der Merkmalauswahl:
Im Gegensatz zum CHAID-Algorithmus, verbessert eine vorgeschaltete Merkmalauswahl hier den vorhergesagten Gewinn: Ohne vorgeschaltete Merkmalauswahl erziehlt der Algorithmus 8,548.50 € Gewinn, mit dieser sogar 8,732.00 € auf dem class Datensatz.

Auswirkung der Fehlklassifizierung:
Bei Baummodellen kann eine Gewichtung der Fehlklassifikationen vorgenommen werden. Diese Gewichtung dient der Anpassung, indem eine Fehlklassifizierung höhere Kosten bedeutet. Bei einer Fehlklassifizierung mit dem Faktor 3.0 berechnet der C5.0-Algorithmus einen Gewinn von 9,028.50 €, mit vorgeschalteter Merkmalauswahl erhält man sogar einen Gewinn von 9,055.00 € auf dem class Datensatz.

4.2.3 Random-Trees

Erklärung:

Dieser Knoten erwies sich als bester Modellierungsknoten. Der Algorithmus entwickelt ein Modell, welches sich aus verschiedenen Entscheidungsbäumen zusammensetzt. Random Trees entspricht in der grundlegenden Vorghensweise der des C&R-Baums, erweitert diesen jedoch um zwei Punkte: Zum einen wird in diesem Modell Bagging verwendet, um ein Overfitting des Datensatzes zu vermeiden. Zum anderen wird für jede Aufteilung des Baums lediglich eine Stichprobe der Input-Werte zur Errechnung der Unreinheit benutzt. Mittels dieses Modells konnten auf dem class Datensatz 10,143.50 €Gewinn vorhergesagt werden.

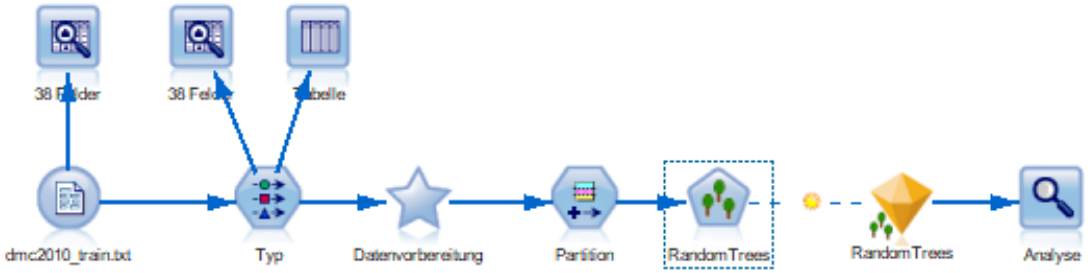

Abbildung 7: Random Trees Modellierung

Auswirkung der Merkmalauswahl und der Fehlklassifizierungskosten:
Unter Verwendung der Merkmalauswahl liefert der Random Trees-Algorithmus leicht bessere Ergebnisse, eine zusätzliche Verwendung der Fehlklssifizierungskosten konnte auch bei Gewichtung mit 3.0 keine Verbesserung mehr bewirken.

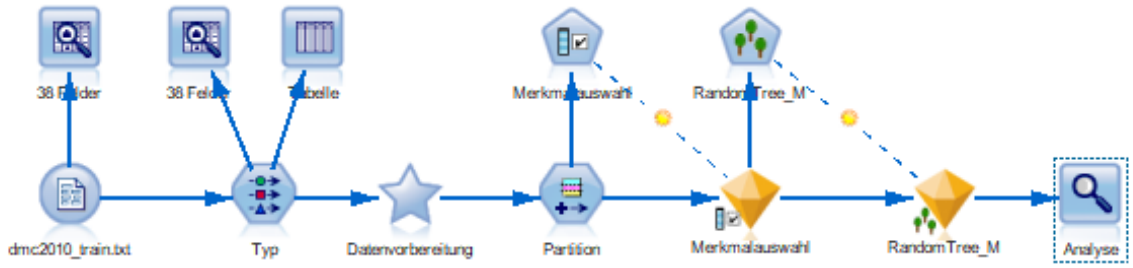

Abbildung 8: Random Trees Modellierung mit Merkmalauswahl

4.2.4 XGBoost-Baum

4.2.5 Tree-AS

4.2.6 C&R-Tree

5 Abbildungsverzeichnis

Abbildungsverzeichnis

6 Literaturverzeichnis

Literatur

[1] *CRISP-DM: Ein Standard-Prozess-Modell für Data Mining* https://statistik-dresden.de/archives/1128.

[2] *IBM SPSS Modeler 18.0 Modellierungsknoten* ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/18.0/de/ModelerModelingNodes.pdf

Erklärung:

Die vorliegende Projektarbeit wurde am Institut für Informatik der Hochschule Coburg nach einem Thema von Herrn Dr. Detlef Bittner erstellt. Hiermit versichere ich, dass ich diese Arbeit selbstständig angefertigt und dazu nur die angegebenen Quellen verwendet habe.

Coburg, den 8. Januar 2019

Michael Krasser