



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, 15. Januar 2019

---

## Inhaltsverzeichnis

<b>1</b>	<b>Geschäftsverständnis</b>	<b>3</b>
1.1	Beschreibung der Situation . . . . .	3
1.1.1	Problembeschreibung . . . . .	3
1.2	Situationsbewertung . . . . .	3
1.2.1	Beschreibung der gelieferten Daten . . . . .	3
1.2.2	Risikofaktoren . . . . .	3
1.2.3	Erfolgskriterien . . . . .	4
1.3	Projektplanung . . . . .	4
<b>2</b>	<b>Datenverständnis</b>	<b>5</b>
2.1	Datensammlung . . . . .	5
2.1.1	Vernachlässigbare Felder . . . . .	5
2.1.2	Erfolgsversprechende Felder . . . . .	6
2.2	Datenbeschreibung . . . . .	6
2.2.1	Datenmenge . . . . .	6
2.2.2	Untersuchung der Daten . . . . .	6
2.2.3	Konsistenz . . . . .	6
<b>3</b>	<b>Vorbereitung der Daten</b>	<b>7</b>
3.1	Bereinigung der Daten . . . . .	7
3.2	Erstellung neuer Felder . . . . .	7
3.3	Bewertung der Felder . . . . .	9
<b>4</b>	<b>Modellierung</b>	<b>9</b>
4.1	Wahl der Modellierungsknoten . . . . .	10
4.2	Test-Design . . . . .	10
4.2.1	Merkmalauswahl . . . . .	10
4.2.2	Fehlklassifizierungskosten . . . . .	10
4.2.3	Boosting . . . . .	11
4.2.4	Umsetzung . . . . .	11
4.3	Getestete Modellierungsknoten . . . . .	12
4.3.1	CHAID . . . . .	12
4.3.2	C5.0 . . . . .	13
4.3.3	Random-Trees . . . . .	14
4.3.4	XGBoost-Baum . . . . .	15
<b>5</b>	<b>Evaluierung und Bereitstellung</b>	<b>16</b>
<b>6</b>	<b>Tabellenverzeichnis</b>	<b>17</b>
<b>7</b>	<b>Abbildungsverzeichnis</b>	<b>18</b>

<i>INHALTSVERZEICHNIS</i>	2
<b>Abbildungsverzeichnis</b>	<b>18</b>
<b>8 Literaturverzeichnis</b>	<b>19</b>

# 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 Wahrscheinlichkeit 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.

Der abschließenden Vergleich kann über den Realdatensatz Datei dmc2010\_real.txt geführt werden, welcher das reale Kaufverhalten dokumentiert.

## 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 unberücksichtigt.)

### 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 9310,50€ im Testdatensatz und 8547,50€ im class-Datensatz<sup>1</sup> liegt ein erfolgreiches Datamining vor

#### Bemerkung:

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

	Wiederkäufer <sup>2</sup>	Wiederkäufer
Gutschein	1.5	-5
Gutschein	0	0

Tabelle 1: Zugelieferte Entscheidungsmatrix

Ein an eine Person, die von alleine nicht wieder bestellt, ausgegebener Gutschein bedeutet für den Onlineshop einen Gewinn von 1,50 €. Jeder Kunde der fälschlicherweise einen Gutschein bekommt, ergibt einen Verlust in Höhe des Wertes des Gutscheins, also 5 €. Als Referenzwert für die Modellbildung wird folgende Situation zugrunde gelegt: Jeder Kunde erhält einen Gutschein. 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

$$(26377 * 1.50 - 6051 * 5) \text{ €} = 9310,50 \text{ €} \quad (1)$$

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:

<sup>1</sup> Dieser Werte wird ebenfalls mit Gl.(1) aus dmc2010\_real.txt ermittelt.

<sup>2</sup> Das überstrichene Symbol ist hier und im folgenden als mathematisches Komplement zu verstehen.

- Geschäftsverständnis: ca zwei Wochen
- Datenverständnis: ca zwei Wochen
- Datenvorbereitung: ca zwei Wochen
- Modellbildung: ca eineinhalb Wochen
- Evaluierung und Bereitstellung: ca eine halbe Woche

## 2 Datenverständnis

### 2.1 Datensammlung

Einige Feler des Datensatzes sind weniger erfolgsversprechend als andere. Da bereits eine Beschreibung der einzelnen Felder 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<sup>3</sup> 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 wurden die im folgenden beschriebenen identifiziert:

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

---

<sup>3</sup> Die Entfernung erfolgte mittels eines Typ-Knotens im SPSS Modeler.

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 zu schlechten Datenqualität entfernt. Spezifische Artikelinformation wie Versandkosten („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 uninteressant 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

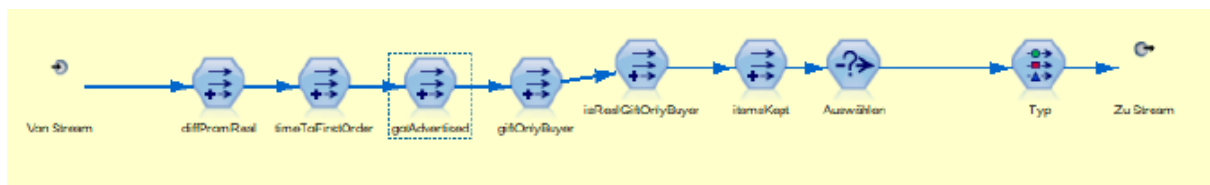


Abbildung 1: Erstellte Felder

Durch studieren der Situation konnten neue Felder aus bestehenden Feldern gebildet<sup>4</sup> 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
```

<sup>4</sup> Der Ableiten-Knotens des SPSS Modelers bietet die Möglichkeit neue Felder mittels minimaler Programmierung zu erstellen.



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

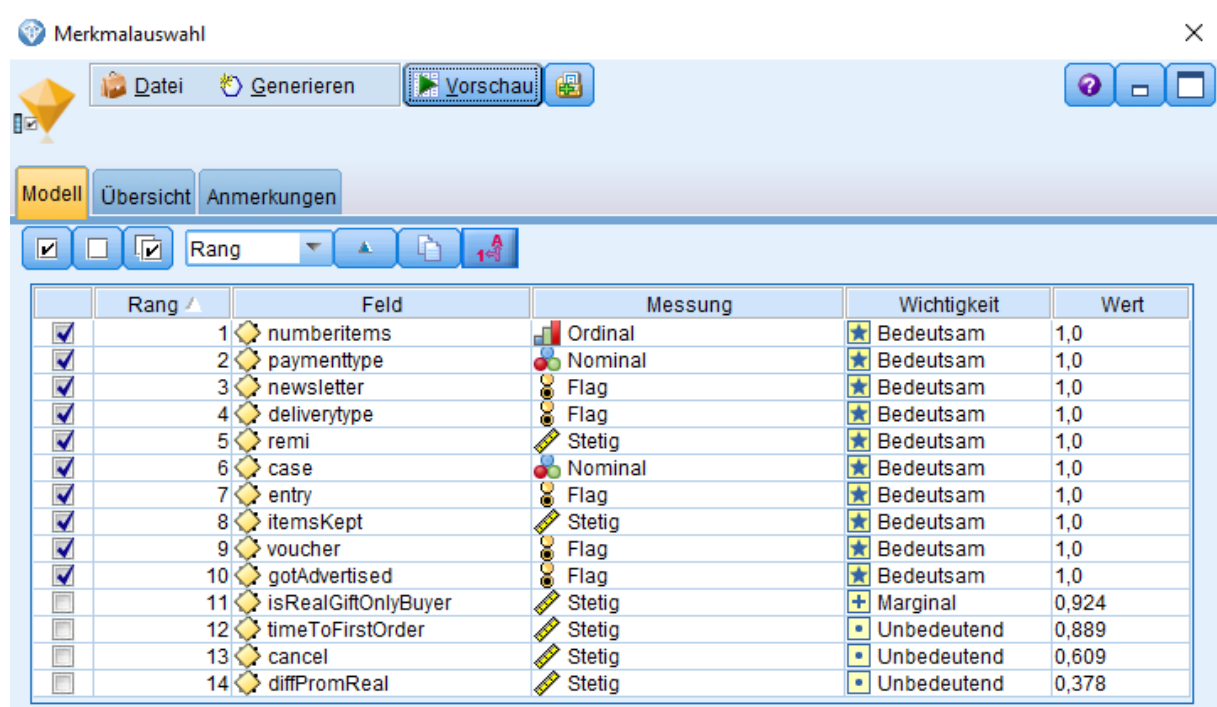
Nachdem neue Felder erstellt wurden, muss der Einfluss dieser auf die Modellierung bewertet werden. Hierfür wurde nach der Datenvorbereitung ein Knoten zur Merkmalauswahl angefügt.

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:



	Rang	Feld	Messung	Wichtigkeit	Wert
<input checked="" type="checkbox"/>	1	numberitems	Ordinal	★ Bedeutsam	1,0
<input checked="" type="checkbox"/>	2	paymenttype	Nominal	★ Bedeutsam	1,0
<input checked="" type="checkbox"/>	3	newsletter	Flag	★ Bedeutsam	1,0
<input checked="" type="checkbox"/>	4	deliverytype	Flag	★ Bedeutsam	1,0
<input checked="" type="checkbox"/>	5	remi	Stetig	★ Bedeutsam	1,0
<input checked="" type="checkbox"/>	6	case	Nominal	★ Bedeutsam	1,0
<input checked="" type="checkbox"/>	7	entry	Flag	★ Bedeutsam	1,0
<input checked="" type="checkbox"/>	8	itemsKept	Stetig	★ Bedeutsam	1,0
<input checked="" type="checkbox"/>	9	voucher	Flag	★ Bedeutsam	1,0
<input checked="" type="checkbox"/>	10	gotAdvertised	Flag	★ Bedeutsam	1,0
<input type="checkbox"/>	11	isRealGiftOnlyBuyer	Stetig	+ Marginal	0,924
<input type="checkbox"/>	12	timeToFirstOrder	Stetig	■ Unbedeutend	0,889
<input type="checkbox"/>	13	cancel	Stetig	■ Unbedeutend	0,609
<input type="checkbox"/>	14	diffPromReal	Stetig	■ Unbedeutend	0,378

Abbildung 2: Merkmalauswahl

## 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

Anfänglich wurden alle von SPSS Modeler zu Verfügung gestellten Modellierungsknoten betrachtet, 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. Da jeder Modellierungsknoten über bestimmte Einstellungen verfügt, wurden diese in den jeweiligen Tests berücksichtigt und sollen daher im folgenden kurz besprochen werden. Die Informationen sind [2] entnommen.

### 4.2.1 Merkmalauswahl

Der Merkmalauswahlknoten sichtet die Eingabefelder, um auf der Grundlage einer Reihe von Kriterien (z. B. dem Prozentsatz der fehlenden Werte) zu entscheiden, ob diese entfernt werden sollen. Anschließend erstellt er eine Wichtigkeitsrangfolge der verbleibenden Eingaben in Bezug auf ein angegebenes Ziel.

### 4.2.2 Fehlklassifizierungskosten

Fehlklassifizierungskosten sind im Grunde Gewichtungen, die auf bestimmte Ergebnisse angewendet werden. Diese Gewichtungen werden in das Modell mit einbezogen und können die Vorhersage (als Schutz gegen kostspielige Fehler) ändern.

Im Falle einer Verbesserung der Gewinnprognose durch Verwendung der Fehlklassifizierungskosten wird überprüft ob sich die Prognose durch Modifikation der Werte weiter verbessert.

### 4.2.3 Boosting

Diese Option erhöht die Modellgenauigkeit. Das Boosting funktioniert so, dass mehrere Modelle in einer Folge erstellt werden. Das erste Modell wird auf die übliche Weise erstellt. Anschließend wird ein zweites Modell erstellt, bei dem besonders die Datensätze berücksichtigt werden, bei denen es im ersten Modell zu Fehlklassifizierungen kam. Das dritte Modell wird in Bezug auf die im zweiten Modell enthaltenen Fehler erstellt usw. Zum Schluss werden die Fälle klassifiziert, indem der gesamte Modellsatz auf ihnen angewendet wird, wobei ein gewichtetes Voting-Verfahren genutzt wird, um die einzelnen Vorhersagen zu einer Gesamtvorhersage zu kombinieren. Das Boosting kann die Genauigkeit eines Entscheidungsbaummodells signifikant verbessern, macht aber auch ein längeres Training notwendig.

Sofern der zu testende Modellierungsknoten über diese Option verfügt, wird die Gewinnprognose in Abhängigkeit von Boosting betrachtet.

### 4.2.4 Umsetzung

Das jeweilige Modell generiert auf dem Trainingsdatensatz mit target90 als Ziel ein Nugget. Diese Nugget wird auf den class-Datensatz angewandt um dort das Verhalten des Feldes target90 zu prognostizieren (\$R-Target90\$), welches mit dem entsprechenden Feld des dmc2010\_real Datensatzes verglichen werden kann.

Zu diesem Zweck wird die Entscheidungstabelle modifiziert: Der Wert 0 im Feld \$R-

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

Tabelle 2: Modifizierte Entscheidungstabelle

target90\$ prognostiziert, dass ein Kunde keine erneute Bestellung tätigt.

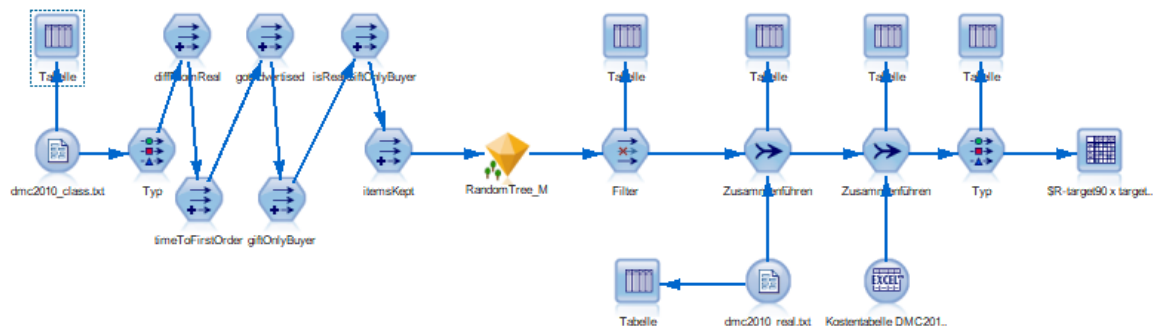


Abbildung 3: Modellierung des Testdesigns

## 4.3 Getestete Modellierungsknoten

### 4.3.1 CHAID

#### Erklärung:

Der CHAID-Knoten generiert Entscheidungsbäume unter der Verwendung der  $\chi^2$  - 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.

Betrachtet wird das Verhalten des Gewinnes unter Merkmalauswahl und Boosting auf dem class-Datensatz.

	Boosting	$\overline{\text{Boosting}}$
Merkwahlauswahl	8548.50 €	8575 €
$\overline{\text{Merkwahlauswahl}}$	8697 €	8548.50 €

Tabelle 3: Vorhersage der CHAID-Modellierung



Abbildung 4: CHAID-Modellierung mit Merkmalauswahl

#### Auswirkung der Merkmalauswahl:

Die Verwendung der Merkmalauswahl produziert ein schlechteres Ergebnis als ohne diese, d.h. die Merkmalauswahl für dieses Modell wurde unpassend gewählt.

#### Auswirkung der Fehlklassifizierung:

Bei Baummodellen kann eine Gewichtung der Fehlklassifikationen vorgenommen werden. Diese Gewichtung dient der Anpassung, indem eine Fehlklassifizierung höhere Kosten bedeutet. Ohne vorgeschaltete Merkmalauswahl und ohne Boosting bei einer Fehlklassifizierung mit dem Faktor 3.0 prognostiziert der CHAID-Algorithmus einen Gewinn von 8,548.50 € was zu keiner Verbesserung führt.

### 4.3.2 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 werden für den class-Datensatz 9055 € Gewinn vorhergesagt, also eine deutliche Verbesserung gegenüber der Vorhersage des CHAID-Knotens.

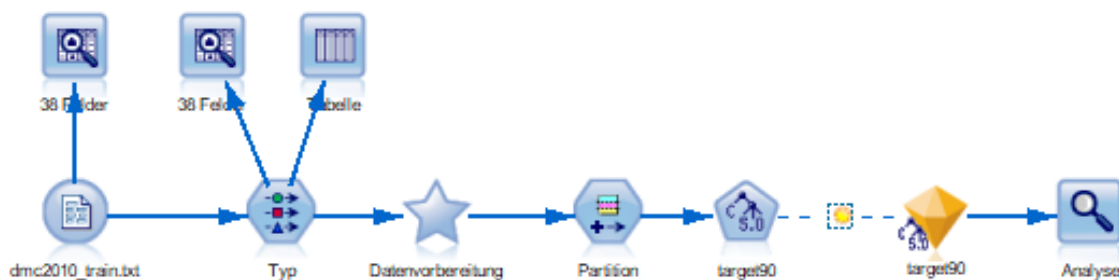


Abbildung 5: C5.0-Modellierung ohne Merkmalauswahl

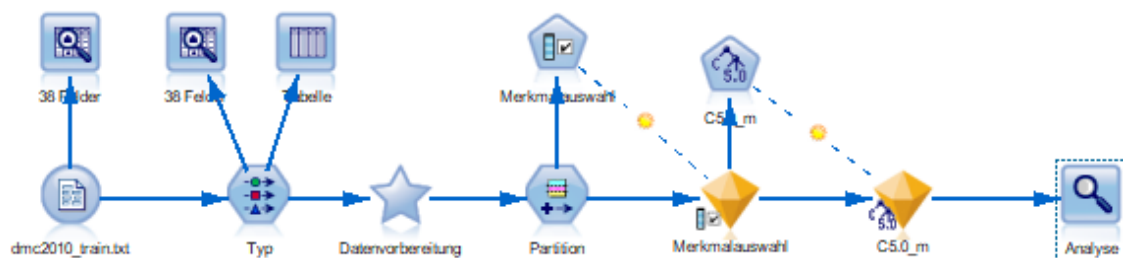


Abbildung 6: C5.0-Modellierung mit Merkmalauswahl

#### Auswirkung der Merkmalauswahl:

Im Gegensatz zum CHAID-Algorithmus, verbessert eine vorgeschaltete Merkmalauswahl hier den vorhergesagten Gewinn: Ohne vorgeschaltete Merkmalauswahl erzielt der Algorithmus 8548.50 € Gewinn, mit dieser sogar 8732 € auf dem class-Datensatz.

**Auswirkung der Fehlklassifizierung:**

Bei einer Fehlklassifizierung mit dem Faktor 3.0 liefert der C5.0-Algorithmus eine Prognose von 9028.50 €, für den Gewinn auf dem class-Datensatz, mit vorgeschalteter Merkmalauswahl erhält man sogar einen Gewinn von 9055.00 € .

**4.3.3 Random-Trees**

Dieser Knoten erweist sich hinsichtlich der Modellierung als am besten geeignet.

**Erklärung:**

Der Algorithmus entwickelt ein Modell, welches sich aus verschiedenen Entscheidungsbäumen zusammensetzt. Random Trees entspricht in der grundlegenden Vorgehensweise 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 maximal 10143.50 € Gewinn vorhergesagt werden:

Fehlklassifizierungskosten	–	1.0	3.0
Merkmalauswahl	10113.50 €	10143.50 €	10143.50 €
Merkmalauswahl	10109 €	10109 €	10109 €

Tabelle 4: Vorhersage der Random-Trees-Modellierung

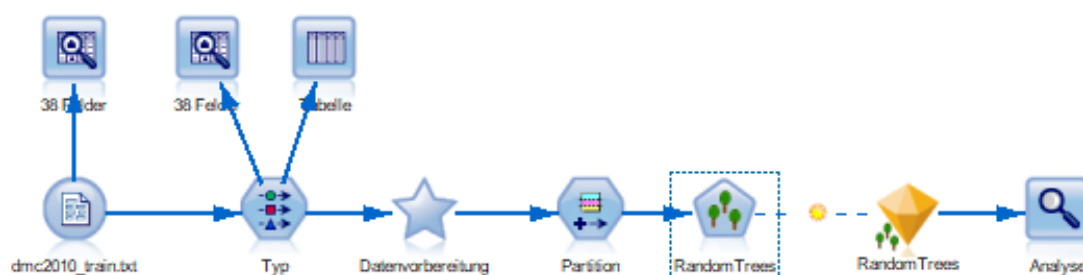


Abbildung 7: Random-Trees Modellierung ohne Merkmalauswahl

**Auswirkung der Merkmalauswahl und der Fehlklassifizierungskosten:**

Bei Verwendung der Merkmalauswahl liefert der Random Trees-Algorithmus leicht bessere Ergebnisse. Daran lässt sich erkennen, dass die Merkmalauswahl für dieses Modell

sinvoll gewählt wurde. Eine zusätzliche Verwendung der Fehlklassifizierungskosten verbesserte das Ergebnis ebenfalls, eine Erhöhung der Gewichtung mit dem Faktor 3.0 lieferte allerdings keine weitere Verbesserung.



Abbildung 8: Random-Trees-Modellierung mit Merkmalauswahl

#### 4.3.4 XGBoost-Baum

##### Erklärung:

Dieser Modellierungsknoten ist die erweiterte Implementierung eines Gradienten-Boosting-Algorithmus mit einem Baummodell als Basismodell. Boosting-Algorithmen lernen iterativ schwache Klassifikationsmerkmale und fügen diese einem endgültigen starken Klassifikationsmerkmal hinzu.

Für den vorliegenden class-Datensatz liefert der Algorithmus eine maximal Gewinnprognose von 8910.50 € :

	Prognose
Merkmalauswahl	8910.50 €
Merkmalauswahl	8849.50 €

Tabelle 5: Vorhersage der XGBoost-Baum-Modellierung



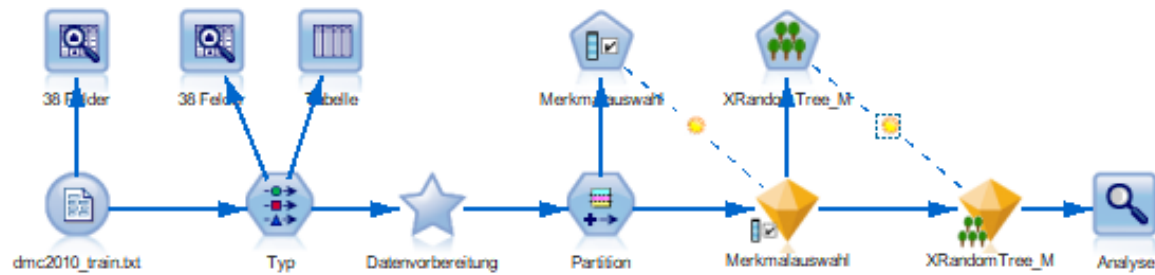


Abbildung 9: XGBoost-Baum -Modellierung mit Merkmalauswahl

**Auswirkung der Merkmalauswahl:**

Die vorgeschaltete Merkmalauswahl liefert nur eine geringe Verbesserung der Gewinnprognose um 61 €, die gewählten Felder besitzen also vermutlich nur eine geringe Korrelation zu dem Kaufverhalten (Der XGBoost-Baum Algorithmus stellt keine Information über die Modellbildung zu Verfügung).

## 5 Evaluierung und Bereitstellung

Die Evaluierung erfolgte durch Anwendung der entwickelten Modelle auf den class-Datensatz. Der jeweilige Gewinn wurde mit Hilfe der modifizierten Entscheidungsmatrix wie in Kapitel 4.2 beschrieben, ermittelt.

Modellierungsknoten	–	CHAID	C5.0	Random Trees	XGBoost-Baum
Errechnete Prognose	8548.50 €	8697 €	9055 €	10758.50 €	8910.50 €
Gewinn	0 euro	148.50 €	506.50 €	2210 €	362 €
relativer Gewinn	0%	1.74%	5.93%	25.85%	4.23%

Tabelle 6: Gewinnübersicht auf dem class-Datensatz

## 6 Tabellenverzeichnis

### Tabellenverzeichnis

1	Zugelieferte Entscheidungsmatrix . . . . .	4
2	Modifizierte Entscheidungstabelle . . . . .	11
3	Vorhersage der CHAID-Modellierung . . . . .	12
4	Vorhersage der Random-Trees-Modellierung . . . . .	14
5	Vorhersage der XGBoost-Baum-Modellierung . . . . .	15
6	Gewinnübersicht auf dem class-Datensatz . . . . .	17

## 7 Abbildungsverzeichnis

### Abbildungsverzeichnis

1	Erstellte Felder . . . . .	7
2	Merkmalauswahl . . . . .	9
3	Modellierung des Testdesigns . . . . .	11
4	CHAID-Modellierung mit Merkmalauswahl . . . . .	12
5	C5.0-Modellierung ohne Merkmalauswahl . . . . .	13
6	C5.0-Modellierung mit Merkmalauswahl . . . . .	13
7	Random-Trees Modellierung ohne Merkmalauswahl . . . . .	14
8	Random-Trees-Modellierung mit Merkmalauswahl . . . . .	15
9	XGBoost-Baum -Modellierung ohne Merkmalauswahl . . . . .	16
10	XGBoost-Baum -Modellierung mit Merkmalauswahl . . . . .	16

## 8 Literaturverzeichnis

### Literatur

- [1] *CRISP-DM: Ein Standard-Prozess-Modell für Data Mining* <https://statistik-dresden.de/archives/1128>.
- [2] *IBM SPSS Modeler 18.1 Modellierungsknoten* <ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/18.1/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 15. Januar 2019

Michael Krasser