Data Mining CUP 2010
Revenue maximisation by intelligent couponing

Fakultät für Mathematik und Informatik
der Friedrich-Schiller-Universität Jena

Projektarbeit

vorgelegt von

Lukas Schmauch, 152412
Thomas Friedrich, XXXXXX

im Januar 2020

Prüfer: Prof. Dr. Martin Bücker
Modul: Big Data

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Aufgabenstellung

		Vorhergesagt	
		kein Wiederkäufer(0)	Wiederkäufer(1)
Tatsächlich	kein Wiederkäufer(0)	1.5	0
	Wiederkäufer(1)	-5	0

2 Datensatz und Preprocessing

2.0.1 Der Datensatz

Merkmal	Beschreibung
customernumber	individuelle Kundennummer
date	Datum der ersten Bestellung
salutation	Anrede des Kunden bzw. Firmenkunde
title	Titel vorhanden oder nicht
domain	Domain des Email Providers
datecreated	Datum der Accounterstellung
newsletter	wurde der Newsletter abboniert
model	nicht spezifiziert (Werte: 1,2,3)
paymenttype	gewählter Zahlungstyp
deliverytype	Versandart
invoicepostcode	Rechnungsadresse
delivpostcode	Lieferadresse
voucher	wurde ein Gutschein eingelöst
advertising	Werbecode
case	Wert der bestellten Produkte
numberitems	Anzahl der bestellten Artikel
gift	wurde die Geschenkoption verwendet
entry	Zugang zum Shop durch einen Partner oder nicht
points	wurden Punkte eingelöst
shippingcosts	sind Versandkosten angefallen
deliverydatepromised	versprochenes Lieferdatum
deliverydatepromised	tatsächliches Lieferdatum
weight	Gewicht der Bestellung
remi	Anzahl zurückgesendeter Artikel
cancel	Anzahl stornierter Artikel
used	Anzahl gebrauchter Artikel
w0	Anzahl bestellter gebundener Bücher
w1	Anzahl bestellter Taschenbücher
w2	Anzahl bestellter Schulbücher
w3	Anzahl bestellter eBooks
w4	Anzahl bestellter Hörbücher
w5	Anzahl heruntergeladener Hörbücher
w6	Anzahl bestellter Filme
w7	Anzahl bestellter Musikartikel
w8	Anzahl bestellter Hardwareartikel
w9	Anzahl bestellter importierter Artikel
w10	Anzahl sonstige bestellte Artikel
target90	Zielvariable: Folgebestellung innerhalb von 90 Tagen oder nicht

Tabelle 2.1: Übersicht aller Merkmale

2.0.2 Die Zielvariable

Die Zielvariable target90 gibt an, ob ein Kunde innerhalb von 90 Tagen erneut eine Bestellung beim Online-Händler getätigt hat oder nicht. Es handelt sich um eine nominalskalierte Variable, welche als Ganzzahl codiert worden ist. Das Klassenlabel ist in Trainings- und Testdatensatz für alle Einträge angegeben. Deswegen handelt es sich hierbei um überwachtes Lernen. Dadurch, dass ausschließlich zwei Ausprägungen möglich sind, bezeichnet man die Aufgabenstellung als binäre Klassifikation.

In Abbildung 2.1 ist die Verteilung der Zielvariable für Trainings- und Testdaten dargestellt. Es wird deutlich, dass in beiden Datensätzen eine ähnliche Verteilung vorliegt. Rund 80 % der Einträge nehmen die Ausprägung 0 an. Das heißt, dass 80 % der Kunden nach der Erstbestellung nicht innerhalb von 90 Tagen erneut bestellen. Dagegen sind 20 % der Kunden als Wiederkäufer etikettiert. Diese Verteilung zeigt die Unbalanciertheit der beiden Klassen.

Die fehlende Gleichverteilung kann im Zuge der Modellbildung und Klassifikation zu falschen Ergebnissen führen. Das begründet sich darin, dass die meisten Algorithmen implizit eine Gleichverteilung der Klassen annehmen. Es gibt deutlich weniger Trainingsbeispiele, um die Eigenschaften der unterrepräsentierten Klasse zu erlernen. Das kann zu Klassifikatoren führen, welche lediglich die überrepräsentierte Klasse vorhersagen. Diese Beobachtung führt dazu, dass im späteren Verlauf des Modellbildungsprozesses Resampling-Strategien angewendet werden.

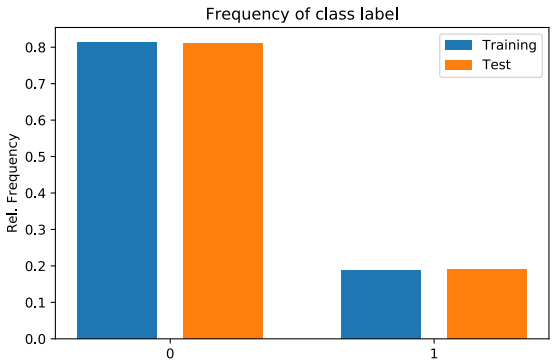

Abbildung 2.1: Klassenverteilung in Trainings- und Testdaten

2.0.3 Feature Engingeering

Ein wichtiger Schritt der Datenvorverarbeitung ist einerseits das Löschen irrelevanter Features. Andererseits wird beim Feature Engineering versucht neue, aussagekräftigere Merkmale aus den bestehenden zu erzeugen oder bestehende so zu bearbeiten, dass sie bestmöglich für den nachgeschalteten Klassifikationsalgorithmus geeignet sind. Im Folgenden werden die neu konstruierten Merkmale beschrieben:

Konstruierte Attribute

accountdur:

books:

nobooks:

itemseff:

Aus der Anzahl der bestellten Artikel, abzüglich der stornierten und zurückgegebenen Artikel, wird die Anzahl der tatsächlich gekauften Artikel erstellt (itemseff). Dieses Merkmal gibt an, wie viele Artikel effektiv durch den Kunden gekauft worden sind.

Modifizierte Attribute

OneHotEncoding:

Algorithmen des maschinellen Lernens können nicht unmittelbar mit kategorialen Merkmalen arbeiten. Damit trotzdem Modelle mit diesen Features trainiert werden können, müssen die Ausprägungen der Merkmale zunächst in einen ganzzahligen Wert codiert werden. Damit die Merkmale anschließend nicht als numerisches Merkmal interpretiert werden ist One Hot Encoding nötig.

Beim One Hot Encoding wird aus den codierten Merkmalen ein Binärer Vektor erstellt. Die Ausprägungen werden dadurch repräsentiert, dass nur die Spalte des Merkmals den Wert 1 annimmt und die anderen Spalten 0 werden. Besonders deutlich wird dies im Teil Modeling. Die dort trainierten Entscheidungsbäume (CART) nutzen nur binäre Splits. Die binären Splits treffen Entscheidungen anhand von „größer gleich oder kleiner gleich“ Beziehungen. Auf die nachfolgenden kategorialen Variablen wird deshalb One Hote Encoding angewendet, um richtige Ergebnisse zu gewährleisten.

salutation:

Die Anrede des Kunden nimmt drei Ausprägungen an. Ein Kunde wird als männlich, weiblich oder als Firmenkunde erfasst. Das Merkmal ist bereits im Datensatz als Ganz-

zahl repräsentiert. Aufgrund des nominalen Skalenniveaus und der Überschreitung von zwei Ausprägungen wird das Merkmal durch One Hot Encoding transformiert.

model:

Die Bedeutung des Features model ist nicht genauer spezifiziert. Wir können dennoch nicht daraus schließen, dass das Merkmal unbedeutsam für die Klassifikationsgüte ist. Das Merkmal nimmt ebenfalls drei Ausprägungen an. Die Ausprägungen sind bereits als Ganzzahl codiert und deshalb ist nur noch die Vektorisierung des Merkmals durch One Hot Encoding nötig. Die Transformation erzeugt drei neue Spalten in unserer Merkmalsmatrix. Inwiefern das Merkmal tatsächlich Relevanz hat, wird bei einer Feature Selection im Evaluationsteil bewertet.

paymenttype:

Der Zahlungstyp besitzt die vier, bereits codierten Ausprägungen Zahlung auf Rechnung, Barzahlung, Zahlung mit dem bestehenden Account und per Kreditkarte. Das Merkmal wird ebenfalls in die Merkmalsmatrix aufgenommen und zuvor mit One Hot Encoding transformiert.

Gelöschte Attribute

deliverydatereal und deliverydatepromised:

datecreated und date:

Aus dem Datum der Accounteröffnung datecreated und dem Datum der Erstbestellung date wird das Merkmal accountdur konstruiert. Dieses gibt die Anzahl der Tage von der Accounteröffnung bis zur ersten Lieferung an. Das neue Merkmal hat zur Folge, dass die beiden genannte Merkmale gelöscht werden.

customernumber:

Die Kundennummer wird gelöscht, weil sie aufgrund ihrer Individualität keine Gruppierung bezüglich der beiden Klassen ermöglicht.

invoicepostcode und delivpostcode:
domain:

points:

Das nominalskalierte Merkmal points gibt an, ob bei der Bestellung Punkte eingelöst worden sind. In der Phase des Data Understandings hat sich gezeigt, dass dieses Merkmal ausschließlich den Wert 0 annimmt. Es wird deshalb aus der Merkmalsmatrix entfernt.

title:

gift:

2.0.4 Resampling

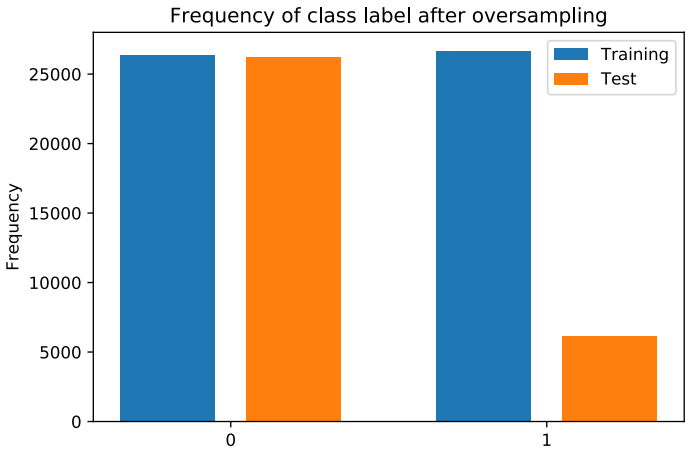

Abbildung 2.2: Klassenverteilung nach Oversampling der Minderheitsklasse

3 Modeling und Evaluation

3.1 Modeling

3.1.1 Entscheidungsbaum Klassifikator

3.1.2 Random Forests

3.2 Evaluation

3.2.1 Das Evaluationskriterium

3.2.2 Backward Feature Elimination

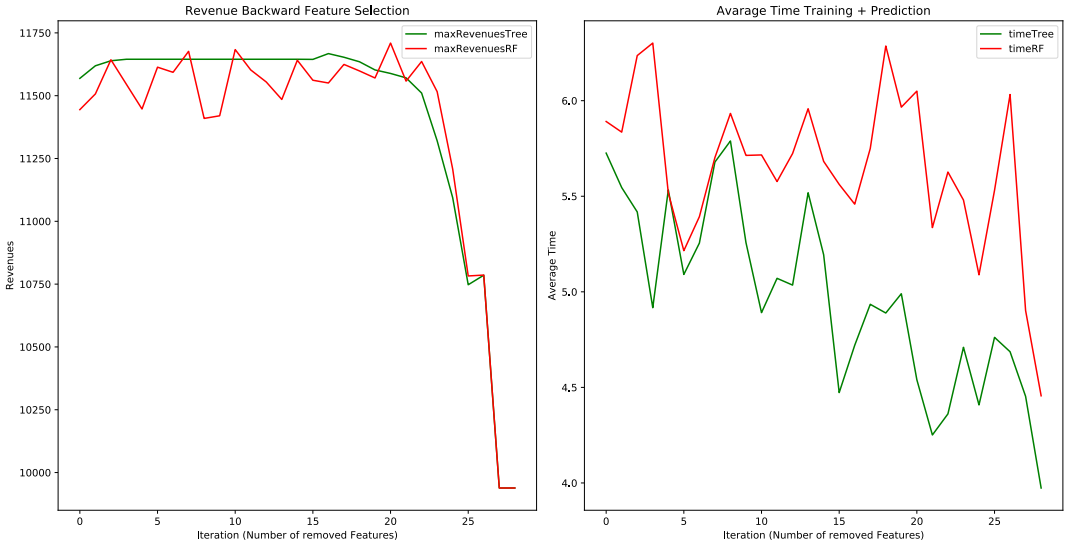

Abbildung 3.1: Umsatz in Anbhängigkeit zur Anzahl der entfernten Merkmale

Rev.	**CART**	**Iter.**	**Rev.**	**RF**	**Rev.**	**AdaBoost**
903.0	w9	1	857.0	numberitems	918.5	books
910.5	nobooks	2	962.5	paymenttype0	918.5	paymenttype1
910.5	paymenttype0	3	931.5	books	918.5	paymenttype1
910.5	paymenttype1	4	980.0	paymenttype2	918.5	paymenttype3
910.5	paymenttype2	5	997.0	model2	918.5	model1
910.5	paymenttype3	6	958.0	advertising	918.5	model2
910.5	model1	7	932.0	model3	918.5	salutation0
910.5	model2	8	951.0	salutation1	918.5	salutation1
910.5	model3	9	940.5	model1	918.5	advertising
910.5	salutation0	10	928.5	entry	918.5	entry
910.5	salutation1	11	899.5	deliverydiff	918.5	w4
910.5	salutation2	12	940.0	accountdur	918.5	w8
910.5	voucher	13	944.5	w10	918.5	itemseff
910.5	advertising	14	934.0	voucher	918.5	deliverydiff
910.5	numberitems	15	964.0	w1	916.5	nobooks
911.0	books	16	910.5	nobooks	936.0	case
911.0	entry	17	949.0	salutation2	937.5	w5
911.0	cancel	18	976.5	paymenttype3	922.0	w1
911.0	used	19	921.5	salutation0	928.5	deliverytype
914.0	weight	20	962.5	w0	957.0	w10
914.0	w0	21	964.0	w4	957.0	paymenttype0
914.0	w6	22	914.0	itemseff	941.0	w6
914.0	w3	23	931.5	shippingcosts	939.0	w9
918.5	w4	24	912.5	used	933.0	used
918.5	w8	25	935.0	w2	926.5	w2
918.5	deliverydiff	26	903.5	w6	907.0	numberitems
918.5	w7	27	915.0	paymenttype1	932.5	voucher
918.5	accountdur	28	913.5	w8	954.5	w0
914.0	case	29	942.0	cancel	951.0	salutation2
917.5	w5	30	947.5	w9	968.0	shippingcosts
913.0	w2	31	934.0	w7	921.5	model3
907.0	shippingcosts	32	960.0	w5	920.0	cancel
889.0	w1	33	909.5	case	908.0	w7
839.0	itemseff	34	899.5	w3	896.0	w3
865.0	w10	35	865.0	weight	824.0	weight
812.5	deliverytype	36	812.5	deliverytype	812.5	accountdur
759.0	remi	37	759.0	remi	759.0	remi
759.0	newsletter	38	759.0	newsletter	759.0	newsletter

Tabelle 3.1: Ergebnisse Backward Feature Elimination

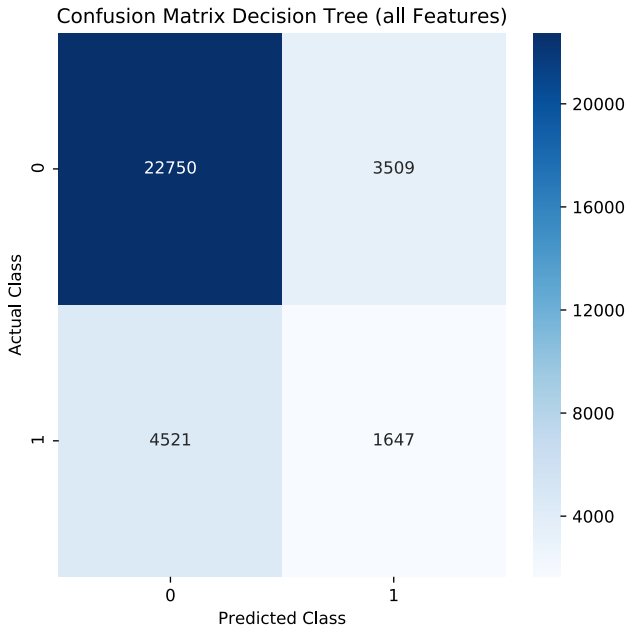

Abbildung 3.2: Konfusionsmatrix Decision Tree mit allen Merkmalen

4 Zusammenfassung