Data Science / Statistical Learning

Sommersemester 2021

Der Ergebnisbericht der Simulationsstudie

Thema: Vorhersage der Wahrscheinlichkeit der Retoure bei der Bestellung im Online Shop

Studierende: Thien Hoa Dang, Doan

I. Einleitung

In der vorliegenden Arbeit geht es um das Vorgehen, die Umsetzung und die Ergebnisse der statistischen Simulationsstudie. Das Hauptziel ist die Erforschung eines statistischen Modells zum Prognostizieren der Wahrscheinlichkeit, dass ein bestimmter Artikel der Bestellung zurückgeschickt wird, im Moment der Produktbestellung beim Online-Shop.

Der Kontext dieser Datensätze in dieser Studie bezieht sich auf einen Online-Schuhe-Shop. In der Realität ist es möglich, dass beim Schuhe-Shop es verschiedene Produktarten außer Schuhe gibt. Dennoch wird in dieser Simulation nur Schuhe als Produkte des Shops in Betracht gezogen, um die Struktur der Produktkategorie zu vereinfachen. Auf Grund dieser Annahme werden alle Produkte in vier Hauptproduktkategorien eingeteilt, nämlich „Damen“, „Herren“, „Jungen“ und „Mädchen“. Ferner beinhaltet jede Kategorie unterschiedliche Subkategorien z.B. „Flache Schuhe“, „Sandalen“, „Sneaker“, „Stiefel“ usw. Da jedes Schuhmodell in verschiedenen Größen erhältlich sein kann, werden Produkte mit gleichem Modell aber unterschiedlicher Größe als unterschiedliche Produkte in dem Datensatz betrachtet. Darüber hinaus wird diese Datei in Bezug auf die Zeitspanne als die Verkaufsrekorde in einem Jahr konzipiert.

Angesicht des großen Umfangs der möglichen zu berücksichtigenden Variablen in der Realität, basieren die ausgewählten erklärenden Variablen auf der Studie von Urbanke et al. (2015) über Produktrückgabe im E-Commerce. Allerdings ist die ursprüngliche Liste von Urbanke et al. noch umfangreich mit insgesamt 24 Merkmale, deshalb beruhend die ausgewählten Variablen weiter auf einer Selbsteinschätzung zur Priorisierung und Durchführbarkeit. Darüber hinaus werden diese Variablen unter einigen wesentlichen statistischen und inhaltlichen Annahmen ausgegangen, damit können die Daten im Modell weiterführend generierbar und interpretierbar ermöglicht werden. Die Liste der zehn in der Studie berücksichtigten Variablen, einschließlich der erklärenden und nicht erklärenden Variablen, wird im zweiten Teil dieser Arbeit aufgeführt und die detaillierten Annahmen der entsprechenden Variablen sind auch dargestellt.

Die weitere Struktur dieser Studie folgt der Vorgaben der Aufgabenstellung, die aus drei Abschnitte bestehen. Im ersten Teil wird die Vorgehensweise zur Erzeugung der Grundgesamtheit von Artikel vorgestellt und die Ergebnisse werden ausführlich illustriert. Der zweite Teil widmet sich dem Aufbau des statistischen Models und der Untersuchung der Abhängigkeit vom Stichprobenumfang. Nach der Erstellung eines Models unter Einsatz von der logistischen Regression, wird es mit vielfältige Stichprobensets von den verschiedenen Umfängen geprüft, um den Einfluss des Stichprobenumfangs zu ermitteln. Im Fokus des dritten Abschnitts steht die Optimierung der Parametrisierung des Models. Mehrere Modelle mit verschiedene Komplexitätsstufe werden auf Trainingsdaten analysiert und auf Validierungsdaten geprüft. Damit kann ein optimales Modell mit guter Prognosefähigkeit erreicht werden.

II. Erzeugung der Grundgesamtheit

Dieser Abschnitt beschreibt zuerst die Variablen und die dazugehörigen Annahmen. Daraus folgen die Ergebnisse der Erstellung dieser statistischen Voraussetzungen in R durch die graphischen Darstellungen. Die betrachteten Attribute werden in drei Kategoried eingeteilt, wie Urbanke et al. vorschlug, nämlich „Produkt-Ebene“, „Warenkorb-Ebene“ und „Kunde-Ebene“. Im Folgenden wird der Tabelle von einem Überblick der Variablen gezeigt. Diese erklärenden Variablen werden nach der Studie von Urbanke et al. (2015) ausgewählt, weil sie großen Einfluss auf die Wahrscheinlichkeit der Rücksendung haben können. Die andere werden davon ausgegangen, dass es keinen direkten statistischen Zusammenhang zwischen ihnen und den zu erklärenden Variablen gibt.

Erklärende Variablen	Nicht erklärende Variablen
Produkt-Ebene	
Produktpreis	Produktgröße
Produktkategorie	
Warenkorb-Ebene	
Anzahl der Produkte im Warenkorb mit der gleichen Subkategorie	Gesamtmenge der bestellten Produkte
Anzahl der Subkategorien aller Produkte im Warenkorb	Anzahl der Kategorien aller Produkte im Warenkorb
	Stunde der Bestellung
Kunde-Ebene	
Bisherige Retourenquote	Kundenregion

Tabelle 1: Überblick von der ausgewählten Variablen

Wie es aus der vorstehenden Liste entnommen werden kann, sind zwei Arten von Behandlungen erforderlich, um diese Variablen zu erzeugen. Bei der ersten handelt es sich um eine direkte Generierung der Pseudozufallszahl (z.B. Produktpreis etc.), wohingegen die zweite die Durchführung der Aggregation von vorhandenen Attributen zur Berechnung erfordert (z.B. Anzahl der Produkte im Warenkorb mit der gleichen Subkategorie etc.). Alle diese Variablen - „Anzahl der Produkte im Warenkorb mit der gleichen Subkategorie“, „Anzahl der Subkategorien aller Produkte im Warenkorb“, „Anzahl der Kategorien aller Produkte im Warenkorb“ basieren auf die bestellte Menge jedes Produkts im Warenkorb. Um die Pseudozufallszahlen zu generieren, muss man im Wesentlichen den Datentyp, die Verteilung und mögliche Grenzwerte dieser Variable identifizieren. Der Angaben aller obengenannten Attribute werden im folgenden Abschnitt behandelt.

Weil ein Produkt in diesem Datensatz als ein Schuhmodell mit spezifischer Größe definiert wird, muss es zuerst die Daten von Schuhmodell generieren. Jedes Modell gehört zu einer Kategorie („Damen“/ „Herren“/ „Jungen“ / „Mädchen“) und einer entsprechenden Subkategorie (z.B.

Sandalen, Sneaker, High Heels etc.). Ferner hat jedes einen bestimmten Preis. Normalerweise ist die Anzahl der Modelle der Damen erheblich höher als die Anzahl der Modelle der anderen. Daher wird es davon ausgegangen, dass der Anteil der 2000 Modelle durch alle vier Kategorien wie im Folgenden lautet, 0.6D: 0.3H: 0.035J: 0.065M. Weil dieses Attribut als nominale Variable behandelt werden muss, wird dieser Wert jedes Produkts in der Grundgesamtheit zur Dummy-Variablen umwandelt.

Anhand von diesem Anteil wird zum Ersten das nominale Merkmal „Kategorie“ erzeugt. Danach wird das zweite nominale Attribut „Subkategorie“ in Anlehnung an den Wert der entsprechenden Kategorie erstellt. Bei diesem Attribut wird kein besonderer Anteil der Subkategorie angewendet.

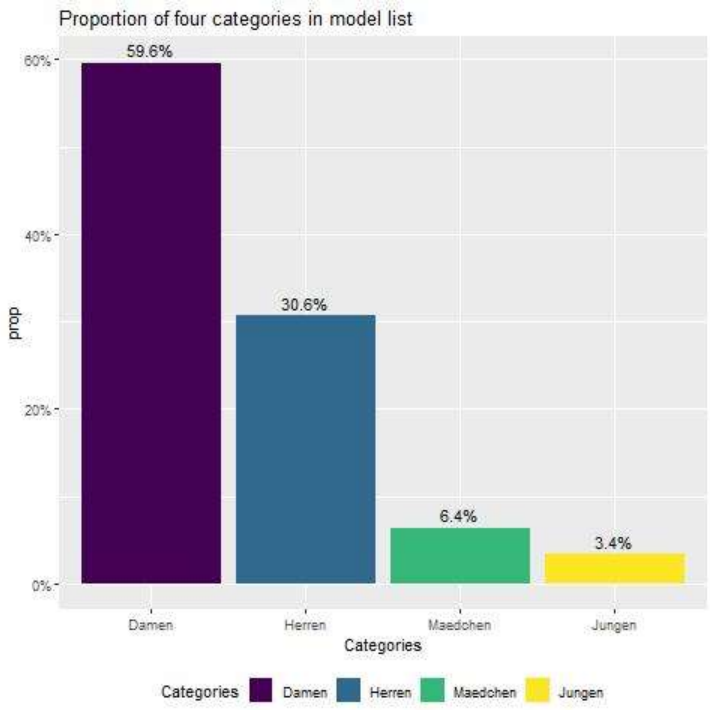

Abbildung 1: Der Anteil der Kategorie in der gesamten Modellliste

Die nächste erforderliche Variable ist der Modellpreis. Der Preis ist eine kontinuierliche numerische Variable aber darf nicht negativ sein, und die Dichte eines bestimmten Bereichs ist wesentlich höher als die des anderen. Dazu sollte die Verteilung des Preises nicht symmetrisch sein. Auf Grund von diesen Voraussetzungen, werden die Werte anhand von log-Verteilung generiert. Des Weiteren ist der Wertebereich von Produkten in jeder Kategorie auch abweichend. Deshalb hat jede Produktkategorie eigene log-Verteilung. In Tabelle 2 werden die Einstellung für jede Verteilung jeder Kategorie gezeigt und die tatsachliche Verteilung der erstellten Zahlen in R in Abbildung 2 dargestellt.

Kategorie	Mean	SD
Damen	100	15
Herren	120	20
Jungen	70	10
Mädchen	70	10

Tabelle 2: Die Einstellung der log-Verteilung jeder Kategorie

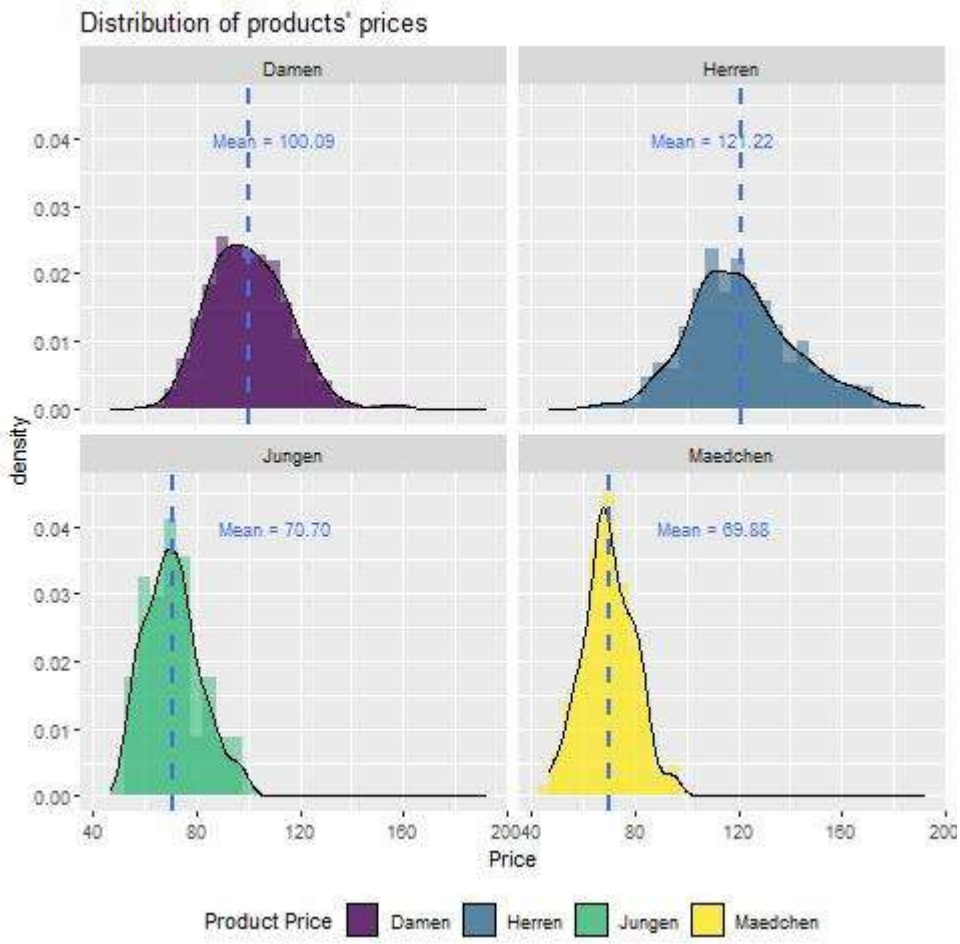

Abbildung 2: Die tatsächliche Verteilung des Modellpreis jeder Kategorie

Bei der Produktgröße wird keine besondere Verteilung verwendet aber auf jede Kategorie wird eigen Intervall angewendet. Der Wert wird als diskrete numerische Variable behandelt. Die Schuhgröße von Damen ist im Bereich von [35 – 43], und die von Herren ist im Bereich von [39 – 50] und die von Kindern ist im Bereich von [19 – 38].

Vor der Erstellung der Bestellungsdaten werden die Angaben der Kunden auch benötigt. Als Einschränkung des Datenschutzes in der Realität wird bei der Analyse auf die persönlichen Informationen verzichtet. In dieser Simulation werden nur Kundenregion und bisherige Retourenquote benutzt. Es ist davon ausgegangen, dass die Proportion der 65000 Kunden in den 16 Bundesländern fast gleich ist. Diese nominale Variable wird mithilfe einer Liste von 16 Bundesländern generiert und in der vorbereiteten Grundgesamtheit wird der Regionswert von

jedem Kunden zu Dummy-Variablen transformiert. Der Anteil von generiert in R Kunden in 16 Bundesländer wird in Abbildung 3 illustriert.

Eine der am schwierigsten zu generierenden Variablen ist die bisherige Retourenquote. Um die genaueste Rate für jeden Kunden zu erreichen, sollte diese Rate nach jeder Bestellung aktualisiert werden. Diese Logik erfordert jedoch riesig Ressourcen für die Verarbeitung. In dieser Simulation wird sie schon probiert und jede Stunde können nur circa 6.000 Einträge verarbeitet werden. Aufgrund der begrenzten verfügbaren Zeit und Ressourcen muss ein anderer Ansatz gewählt werden, um ein Gleichgewicht zwischen der Echtzeitgenauigkeit und der optimaler Ressourcenbenutzung zu erreichen. Es wird davon ausgegangen, dass diese Quote als eine einjährige historische Retourenquote betrachtet wird und nur jährlich aktualisiert wird. Denn es wird erwartet, dass es keine extreme Verhaltensänderung bei treuen Kunden gibt und bei Neukunden können diese Daten nicht so viele Informationen über das Verhaltensmuster dieses Kunden bringen. Die Werte der Quote liegen im Bereich [0 – 1] und die Dichte des Intervalls [0 - 0.4] sollte höher als die des Überrests, weil es erwartet wird, sollte eine Retourenquote eines normalen Kunden unter 40% liegen. Deshalb wird diese Quote zufällig anhand der Beta-Verteilung mit $\alpha = 1$ und $\beta = 6$ generiert. Die Verteilung der generierten Quote in R wird in Abbildung 4 veranschaulicht.

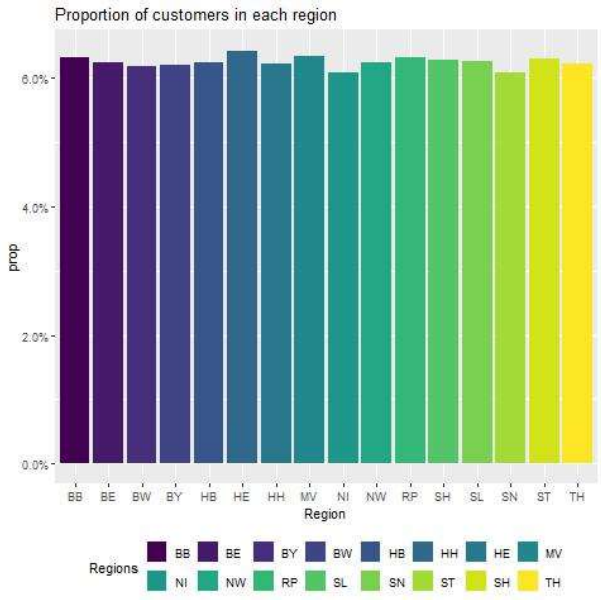

Abbildung 3: Der Anteil der Kunden in 16 Bundesländer

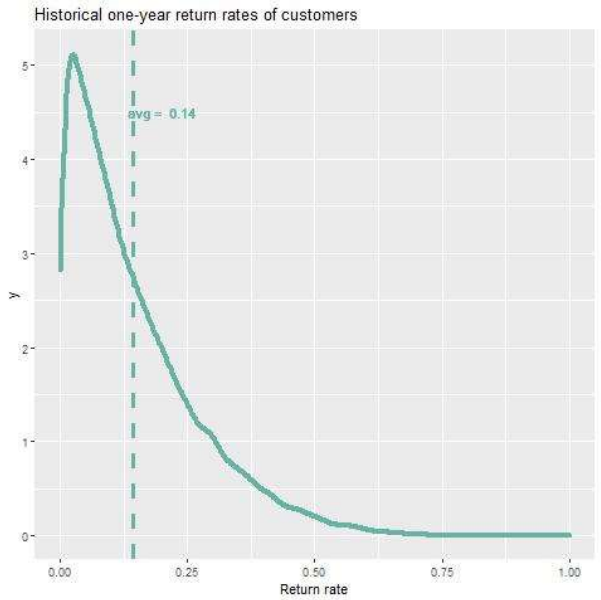

Abbildung 4: Die Verteilung der bisherigen Retourenquote

Jeder Eintrag in diesem Satz wird als einen Artikel im Warenkorb einer Bestellung betrachtet. Um die Analyse weiter zu vereinfachen und interpretierbar zu machen, kann die bestellte Menge eines Artikels nur eins sein. Die bestellte Menge eines Produkts in einer Bestellung kann mehr als eins aber jedes Stück wird als ein Artikel eingetragen. Durch diesen Wert werden andere Variable inklusive der Anzahl der Produkte im Warenkorb mit der gleichen Subkategorie, der Anzahl der

Subkategorien aller Produkte im Warenkorb und der Anzahl der Kategorien aller Produkte im Warenkorb. Allerdings muss vor der Erstellung der Produktliste jeder Bestellung die Gesamtmenge der bestellten Produkte mithilfe der Poisson-Verteilung generiert werden. Da wird es davon ausgegangen, dass jeder Kunde jährlich durchschnittliche vier Bestellung hat und jede Bestellung beinhaltet durchschnittlich vier Produkte. Danach werden bei jeder Bestellung eine Probe von Produktliste zufällig genommen und der bestellten Menge jedes Produkts wird eins zugewiesen. Abbildung 5 und Abbildung 6 stellen die Verteilung der jährlichen Anzahl der Bestellung und die Verteilung der durchschnitten Mengen jeder Bestellung jedes Kunden dar. Dann in Abbildung 7, 8 und 9 werden die Proportionen der drei abgeleiteten Variablen auch gezeigt.

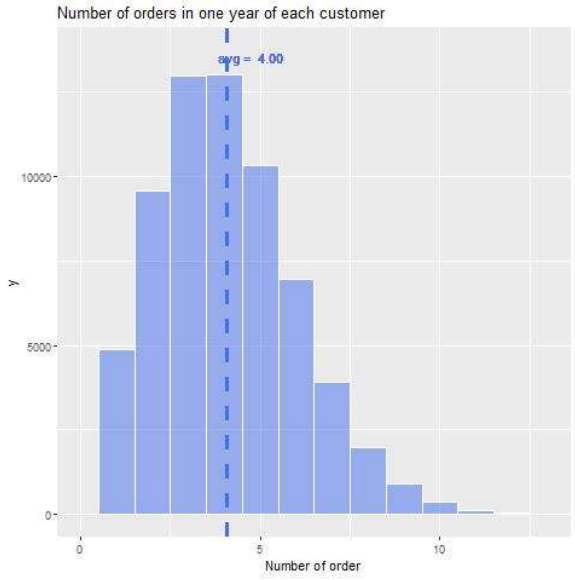

Abbildung 5: Die jährliche Anzahl der Bestellung jedes Kunden

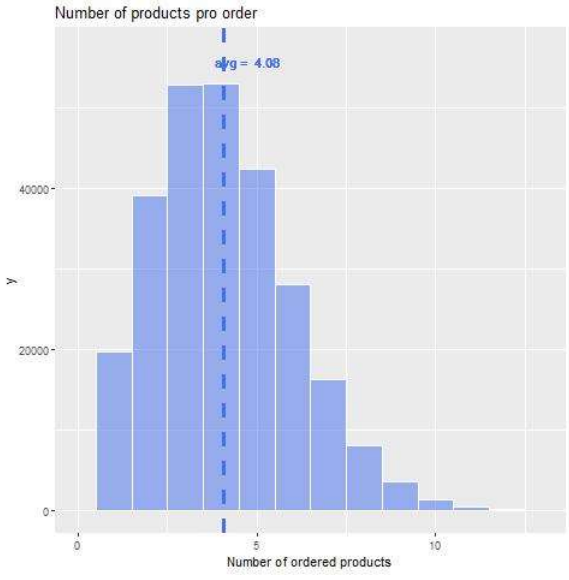

Abbildung 6: Die Menge jeder Bestellung in einem Jahr

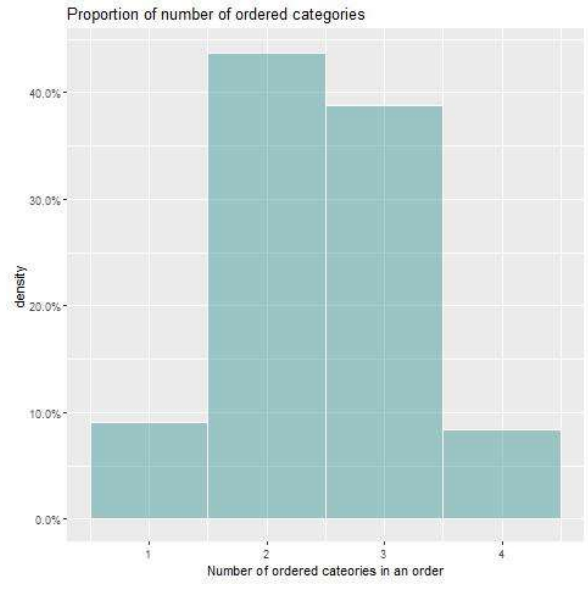

Abbildung 7: Der Anteil der Anzahl der Kategorien aller Produkte im Warenkorb

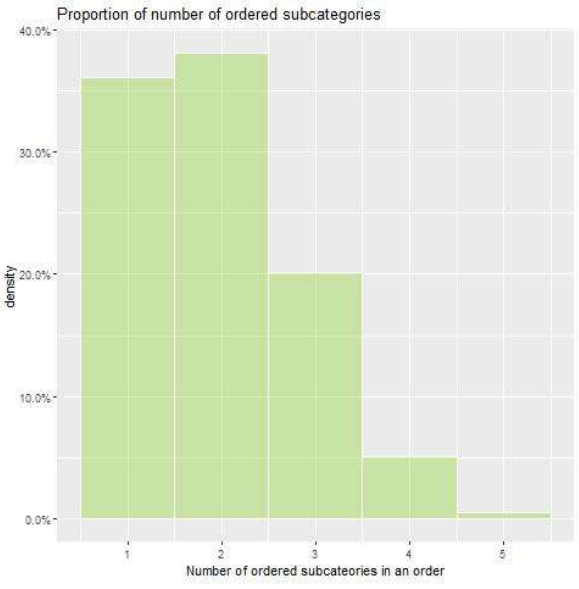

Abbildung 8: Der Anteil der Anzahl der Subkategorien aller Produkte im Warenkorb

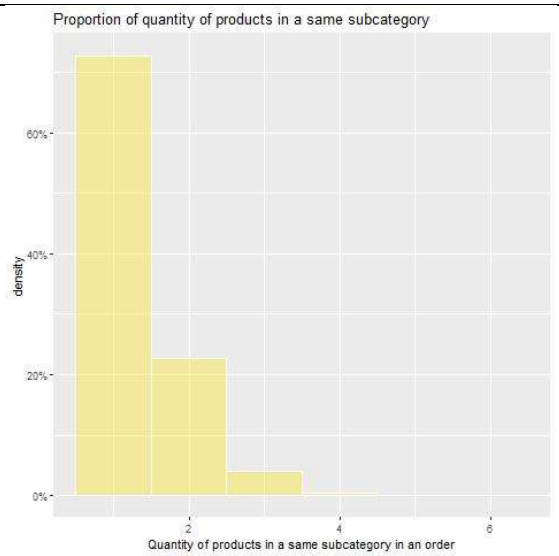

Abbildung 9: Der Anteil der Anzahl der Produkte im Warenkorb mit der gleichen Subkategorie

Bezüglich des Attributs - die Stunde der Bestellung – wird auf Grund der Annahme, dass der größte Teil des Umsatzes im Zeitraum von 19 bis 24 Uhr, es für jedes Zeitintervall eins eigenen Gewichts angewendet. Nämlich ist die Wahrscheinlichkeit jeder Stunde von 00 bis 07 1%, die jeder Stunde von 08 bis 13 ist 1.5%, und die jeder Stunde von 14 bis 18 ist 3%, während ist die jeder Stunde von 19 bis 23 ist 13.8%.

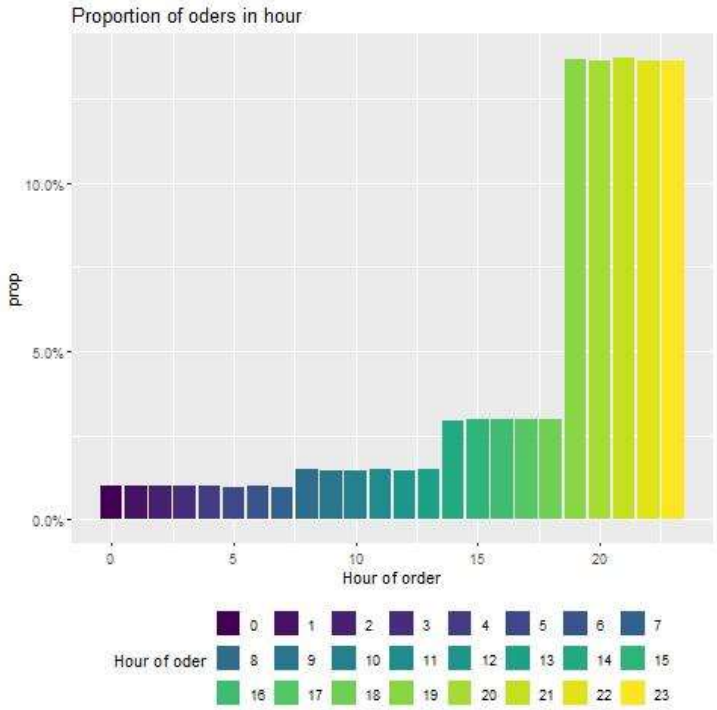

Abbildung 10: Der Anteil der Bestellung in Stunde

Der folgende Abschnitt beschäftigt sich mit der Einstellung der β- Werte, damit können die Anhängigkeiten zwischen zu erklärender Variablen und erklärenden Variablen erstellt werden.

Tabelle 3 zeigt die eingestellte Werte der Koeffizienten der erklärenden Variablen. Diese Werte werden basierend auf dem geschätzten Chancenverhältnisse berechnet z.B. falls es erwartet wird, die Wahrscheinlichkeit, dass ein Produkt zurückgesendet wird, wenn es zur Damenkategorie gehört, ist 7.4-mal höher als in den anderen Fällen, wird Beta geschätzt als $\ln 7.4$. In diesem Modell wird der Produktkategorie „Jungen“ als die repräsentative Grundlage, daher werden nur die β-Werte für die anderen übrigen Kategorie zugewiesen.

Variable	**Geschätzte Änderung des Odd-Ratio**	**Koeffizient β**
Produktpreis	0,98	$\beta_{ProdPreis} = \ln 0{,}98 \approx -0{,}02$
Anzahl der Produkte im Warenkorb mit der gleichen Subkategorie	2,8	$\beta_{AnzahlGleichSubkat} = \ln 2{,}8 \approx 2{,}00$
Anzahl der Subkategorien aller Produkte im Warenkorb	1,35	$\beta_{AnzahlSubkat} = \ln 1{,}35 \approx 0{,}30$
Bisherige Retourenquote	7,4	$\beta_{BisherQuote} = \ln 7{,}4 \approx 2{,}00$
Produktkategorie-Damen	7,4	$\beta_{DamenKat} = \ln 7{,}4 \approx 2{,}00$
Produktkategorie-Herren	2,8	$\beta_{HerrenKat} = \ln 2{,}8 \approx 1{,}00$
Produktkategorie-Mädchen	1,65	$\beta_{MädchenKat} = \ln 1{,}65 \approx 0{,}50$
Kreuzabhängigkeit zwischen der Anzahl der Produkte mit der gleichen Subkategorie und der Anzahl der Subkategorien aller Produkte	1,35	$\beta_{MixGleich+InsSubcat} = \ln 1{,}35 \approx 0{,}30$

Tabelle 3: Eingestellte Koeffizienten für die Wahrscheinlichkeitsformel

Es wird davon ausgegangen, dass der Einflüsse der bisherige Retourenquote und der Anzahl der Produkte mit der gleichen Subkategorie in einem Warenkorb größer als der der anderen Variablen, weil sie auf ein Muster des Probier- und Rückgabeverhaltens hindeuten können. Außerdem hat Kategorie Damen auch großen Einfluss. Weil es in dieser Kategorie eine wesentliche große Auswahl an Modellen gibt, im Vergleich zu den anderen Kategorien. Damit ist die Wahrscheinlichkeit der Retoure des Produkts in dieser Kategorie deutlich höher. Außerdem wegen der Denkweise, dass der niedrigere Preis der Produkte und die festen Versandkosten pro Paket den Kunden dazu bringen, mehrere Chancen für eine "Probe und Rückgabe" zu kaufen. Daher ist es möglich, dass, je höher der Preis, desto niedriger das Chancenverhältnis ist, aber diese Auswirkung ist nicht so kräftig gegenüber der der übrigen erklärenden Variablen. Hinzu kommt noch die Synergieeffekt zwischen der Anzahl der Produkte mit der gleichen Subkategorie und der Anzahl der Subkategorien aller Produkte im Warenkorb. Diese Beziehung ist sinnvoll, denn wenn ein Kunde viele verschiedene Produkte in verschiedenen Subkategorien kauft und auch in jeder

Subkategorie viele Produkte verschiedener Modelle oder des gleichen Modells, aber mit verschiedenen Größen gekauft werden, kann dies auf eine hohe Wahrscheinlichkeit hinweisen, dass es sich um ein typisches Probier- und Rückgabeverhalten handelt. Somit ergibt sich ein möglicher Synergieeffekt oder eine Kreuzabhängigkeit zwischen beiden Variablen.

In dieser Formel wird dem y-Achsenabschnitt der Wert -3.12 zugewiesen. Er wird durch „Trial-and-Error“ Verfahren ausgewählt, um die Wahrscheinlichkeit zu verteilen und den Anteil von 0 und 1 in einem vernünftigen Verhältnis zu halten. In diesem Falls sollte die erwartete Wahrscheinlichkeit circa 35% sein, weil das als ein angemessener Wert der Retourenquote bei normalem Online-Shop betrachtet wird. Die Verteilungen der erzeugten Wahrscheinlichkeit und des entsprechenden Retoure-Status des Produkts werden in Abbildung 8 und 9 im Folgenden veranschaulicht.

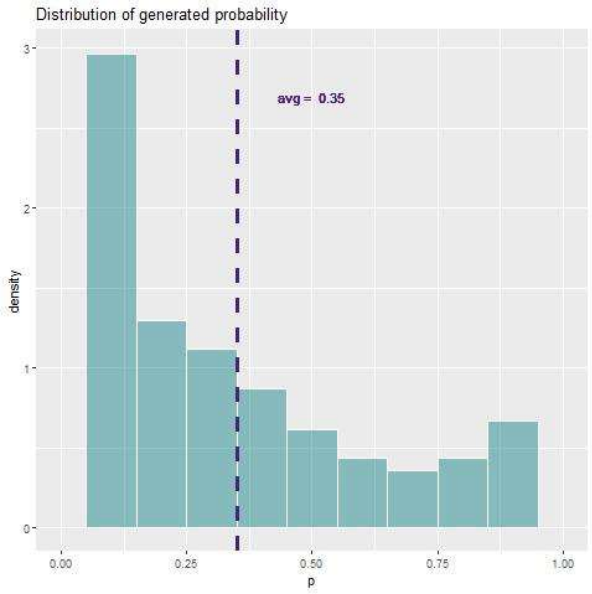

Abbildung 11:Die Verteilung der erzeugten Wahrscheinlichkeit

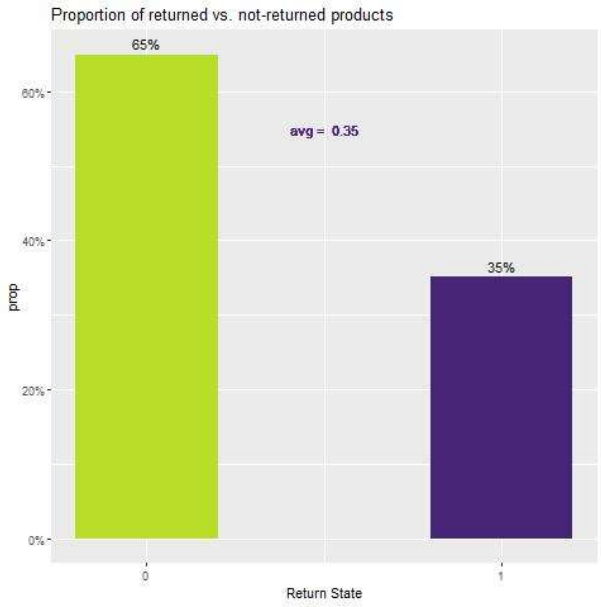

Abbildung 12: Das erzeugte Verhältnis der zurückgeschickten Produkte

Im folgenden Abbildung 13 werden eine kurze deskriptive Statistik der endgültig generierten Grundgesamtheit mit mehr als 1.000.000 Einträge dargestellt. Allerdings sind die Werte von solchen kategorischen Attributen oder Zeitmarke nur „NA / - Inf“, weil diese fundamentalen Statistiken nur zu numerischen Merkmale passen.

	vars	n	mean	sd	min	max	range	se
return_state	1	1081231	0.35	0.48	0.00	1.00	1.00	0.00
p	2	1081231	0.35	0.26	0.01	1.00	0.99	0.00
hour_of_order	3	1081231	17.93	5.55	0.00	23.00	23.00	0.01
cust_prev_return_rate	4	1081231	0.14	0.12	0.00	0.89	0.89	0.00
cust_region	5	1081231	NaN	NA	Inf	-Inf	-Inf	NA
prod_size	6	1081231	38.98	6.55	19.00	50.00	31.00	0.01
mod_preis	7	1081231	101.84	24.18	46.58	191.25	144.67	0.02
prod_num_per_order	8	1081231	5.00	1.99	1.00	16.00	15.00	0.00
qty_same_subcat	9	1081231	1.32	0.58	1.00	6.00	5.00	0.00
num_subcat_ord	10	1081231	1.96	0.90	1.00	5.00	4.00	0.00
num_cat_ord	11	1081231	2.47	0.77	1.00	4.00	3.00	0.00
mod_subcat1	12	1081231	NaN	NA	Inf	-Inf	-Inf	NA

Abbildung 13: Die deskriptive Statistik der erzeugten Grundgesamtheit

III. Simulation der Anwenderperspektive

III.1 Das Verfahren des Identifizierens eines optimalen Modells

Das folgende Kapitel beschäftigt sich zuerst mit der Bearbeitung von einer zufälligen Stichprobe, die angewendet wird, um das zur Grundgesamtheiterzeugung benutztes Modell herauszufinden. Danach beschreibt es die Ergebnisse von der Untersuchung der Abhängigkeit der Qualität von Schätzern vom Stichprobeumfang.

In einem ersten Schritt der Simulation, wird ein Stichprobe im Umfang von n = 10.000 aus der Grundgesamtheit gezogen. Dann wird sie in zwei Teilen eingeteilt. Der Teil mit 75% der Einträge wird als Trainingsdaten benutzt, um das implizierte Modell zu untersuchen und der andere Teil mit übrige 2.500 Einträge wird als Validierungsdaten verwendet, um die Prädiktionsfähigkeit jedes untersuchten Modells zu bewerten. Damit können das richtige Modell identifiziert werden. In dieser Arbeit werden in aller Modelle die p-Werte zu dem Grenzwert $\alpha = 0.05$ verglichen, damit die Hypothese $\beta = 0$ angenommen oder verworfen wird.

Der Zusammenhang zwischen der Retourenquote und zehn betrachteten Variablen, und auch die Abhängigkeit zwischen diesen zehn Variablen werden vorläufig durch der Korrelationsmatrix geprüft. In der Abbildung 14 weisen die gezeigten p-Werte mit dem regulären ausgewählten Grenzwert $\alpha = 0.05$ die möglichen Zusammenhangen wischen die Quote und acht anderen Variablen außer der Stunde der Bestellung und der Region der Kunden hin. Darüber hinaus werden die Korrelationen inmitten der Produktgröße, des Produktpreis, der anderen Variablen bezüglich des Inhaltes der Bestellung und auch der Produktkategorie in Abbildung 15 deutlich angezeigt. Dies bedeutet, dass diese Variablen hochkorreliert sind. Dazu führt es auf den Einfluss der Auswahl der erklärenden Variablen in dem Modell zu. Die Mitwirkung vom Prädiktor können insignifikant sein, obwohl er und die zu erklärende Variable korreliert sind. Diese Beziehung wird ausführlich in verschiedenen Modellen einschließlich des Tests der Multikollinearität und der Kreuzabhängigkeit weiter untersucht. Darüber hinaus kann die Unabhängigkeit der Retourenquote von der Region des Kunden durch diese Korrelationen auch bewiesen werden.

n= 7500

P

	return_state
return_state	
hour_of_order	0.3911
cust_prev_return_rate	0.0000
prod_size	0.0006
mod_preis	0.0000
prod_num_per_order	0.0000
qty_same_subcat	0.0000
num_subcat_ord	0.0000
num_cat_ord	0.0000
mod_subcat1_Damen	0.0000
mod_subcat1_Herren	0.0000
mod_subcat1_Jungen	0.0000
mod_subcat1_Maedchen	0.0000
cust_region_BB	0.3613
cust_region_BE	0.5641
cust_region_BW	0.2284
cust_region_BY	0.6331
cust_region_HB	0.5631
cust_region_HE	0.2930
cust_region_HH	0.6984
cust_region_MV	0.2087
cust_region_NI	0.9839
cust_region_NW	0.2010
cust_region_RP	0.8837
cust_region_SH	0.5220
cust_region_SL	0.9758
cust_region_SN	0.0720
cust_region_ST	0.3989
cust_region_TH	0.9230

Abbildung 14: p-Werte der Korrelationen zwischen der zu erklärenden Variablen und den 10 betrachteten Variablen

n= 7500

P

	hour_of_order	cust_prev_return_rate	prod_size	mod_preis	prod_num_per_order	qty_same_subcat	num_subcat_ord	num_cat_ord
hour_of_order		0.6085	0.2799	0.8325	0.5115	0.1821	0.8044	0.6936
cust_prev_return_rate	0.6085		0.8309	0.9339	0.3290	0.6372	0.0588	0.6083
prod_size	0.2799	0.8309		0.0000	0.8356	0.0000	0.0000	0.0000
mod_preis	0.8325	0.9339	0.0000		0.5049	0.0000	0.0000	0.0000
prod_num_per_order	0.5115	0.3290	0.8356	0.5049		0.0000	0.0000	0.0000
qty_same_subcat	0.1821	0.6372	0.0000	0.0000	0.0000		0.0000	0.0000
num_subcat_ord	0.8044	0.0588	0.0000	0.0000	0.0000	0.0000		0.0000
num_cat_ord	0.6936	0.6083	0.0000	0.0000	0.0000	0.0000	0.0000	
mod_subcat1_Damen	0.4153	0.1961	0.7640	0.0000	0.0543	0.0000	0.0000	0.0000
mod_subcat1_Herren	0.7334	0.3617	0.0000	0.0000	0.1666	0.5797	0.0000	0.0086
mod_subcat1_Jungen	0.0618	0.7419	0.0000	0.0000	0.6358	0.0000	0.0000	0.0000
mod_subcat1_Maedchen	0.5190	0.3551	0.0000	0.0000	0.5451	0.0000	0.0000	0.0000
cust_region_BB	0.4153	0.5239	0.2608	0.4721	0.6842	0.9458	0.9997	0.0734
cust_region_BE	0.1171	0.1145	0.3555	0.8275	0.6553	0.1331	0.3718	0.0697
cust_region_BW	0.8215	0.2454	0.0379	0.5628	0.0580	0.7540	0.0009	0.7052
cust_region_BY	0.7129	0.4309	0.6376	0.1809	0.8290	0.1009	0.7225	0.6719
cust_region_HB	0.5241	0.1376	0.3044	0.9380	0.0376	0.4022	0.2329	0.5773
cust_region_HE	0.3782	0.3064	0.6920	0.4146	0.0204	0.8108	0.2238	0.0192
cust_region_HH	0.9434	0.0336	0.7159	0.2892	0.8190	0.9818	0.6527	0.9926
cust_region_MV	0.2122	0.8579	0.8991	0.8991	0.2563	0.8308	0.9064	0.4998
cust_region_NI	0.0253	0.8356	0.8152	0.5135	0.2537	0.4472	0.0819	0.1865
cust_region_NW	0.7051	0.2094	0.8027	0.8383	0.4520	0.9383	0.6700	0.6026
cust_region_RP	0.4561	0.3024	0.1865	0.6571	0.7635	0.8918	0.3793	0.1021
cust_region_SH	0.8458	0.0262	0.5258	0.1490	0.8511	0.6800	0.3845	0.3416
cust_region_SL	0.2132	0.0070	0.1401	0.2383	0.8677	0.0900	0.1108	0.5418
cust_region_SN	0.0102	0.5472	0.7992	0.2071	0.1201	0.5118	0.4264	0.5307
cust_region_ST	0.8208	0.1726	0.6146	0.5898	0.0040	0.3597	0.0023	0.9327
cust_region_TH	0.3622	0.0428	0.1473	0.1295	0.5886	0.1080	0.5017	0.5901

Abbildung 15: p-Werte der Korrelationen inmitten der 10 betrachteten Variablen

Das erste Modell wird mit aller zehn möglichen Variablen anhand der logistischen Regression versuch. Die Zusammenfassung des Ergebnisses dieses Modells wird in Abbildung 16 wiederabgegeben. Dadurch kann man bestimmt es einsehen, dass nur die Koeffizienten dieser Variablen inklusive des Produktpreis (mod_preis), der bisherigen Retourenquote(cust_prev_return_rate), der Anzahl der Produkte im Warenkorb mit der gleichen Subkategorie(qty_same_subcat), der Anzahl der Subkategorien aller Produkte im Warenkorb

(num_subcat_ord) und der Produktkategorie (mod_subcat1_Damen, mod_subcat1_Herren und mod_subcat1_Maedchen) signifikant sind, während die der anderen Attribute insignifikant sind. Darüber hinaus werden durch das Ergebnis der VIF Funktion (Variance Inflation Factor), die in Abbildung 17 gezeigt wird, die Multikollinearität zwischen manchen Variablen bewiesen. Dieses Resultat ist übereinstimmend mit dem Ergebnis der Korrelationsmatrix und es ist nachvollziehbar, weil es möglich ist, dass jede Produktkategorie eigenes Set des Produktpreis und eigenes Set der Produktgröße hat. Deswegen werden aufgrund dieser kombinierten Ergebnisse im nächsten Modell nur diese fünf Variablen mit signifikanten p-Werte beinhaltet.

```
Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          -4.187132   0.380532 -11.003  < 2e-16 ***
prod_num_per_order    0.031048   0.024658   1.259  0.20798
prod_size             0.005816   0.008692   0.669  0.50344
mod_preis            -0.021746   0.001848 -11.769  < 2e-16 ***
cust_prev_return_rate 2.072650   0.232548   8.913  < 2e-16 ***
hour_of_order         0.002569   0.005286   0.486  0.62691
qty_same_subcat       1.538763   0.061867  24.872  < 2e-16 ***
num_cat_ord           0.020637   0.052175   0.396  0.69245
num_subcat_ord        0.630164   0.045833  13.749  < 2e-16 ***
mod_subcat1_Damen     2.230478   0.218635  10.202  < 2e-16 ***
mod_subcat1_Herren    1.127067   0.252490   4.464 8.05e-06 ***
mod_subcat1_Maedchen  0.670948   0.204590   3.279  0.00104 **
cust_region_BE        0.214274   0.144899   1.479  0.13920
cust_region_BY       -0.065501   0.143969  -0.455  0.64913
cust_region_HB        0.141197   0.143374   0.985  0.32471
cust_region_HE       -0.122211   0.145226  -0.842  0.40006
cust_region_HH       -0.062154   0.144274  -0.431  0.66661
cust_region_MV       -0.100028   0.139935  -0.715  0.47472
cust_region_NI        0.018762   0.139080   0.135  0.89269
cust_region_NW       -0.009795   0.138545  -0.071  0.94364
cust_region_RP       -0.181618   0.142455  -1.275  0.20234
cust_region_SH       -0.208370   0.141761  -1.470  0.14160
cust_region_SL        0.167219   0.139145   1.202  0.22946
cust_region_SN        0.234685   0.137111   1.712  0.08696 .
cust_region_ST       -0.115941   0.147736  -0.785  0.43258
cust_region_TH       -0.059655   0.147566  -0.404  0.68602
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9750.7  on 7499  degrees of freedom
Residual deviance: 7198.3  on 7474  degrees of freedom
AIC: 7250.3
```

Abbildung 16: Das Modell der logistischen Regression mit 10 Variablen

```
> vif(var10_log_mod)
   prod_num_per_order            prod_size            mod_preis cust_prev_return_rate       hour_of_order       qty_same_subcat         num_cat_ord
             2.680636             2.928587             1.823695             1.016117             1.006460             1.294697             1.923068
      num_subcat_ord     mod_subcat1_Damen    mod_subcat1_Herren  mod_subcat1_Maedchen      cust_region_BE      cust_region_BY      cust_region_HB
             1.946594            13.431360            14.847960             4.241383             1.374834             1.381399             1.385869
      cust_region_HE      cust_region_HH      cust_region_MV      cust_region_NI      cust_region_NW      cust_region_RP      cust_region_SH
             1.372670             1.379543             1.415278             1.419051             1.423415             1.392973             1.401585
      cust_region_SL      cust_region_SN      cust_region_ST      cust_region_TH
             1.422444             1.435887             1.356853             1.356630
```

Abbildung 17: Das Ergebnis des Tests der Multikollinearität im Modell mit aller 10 Variablen

Die Abbildung 17 und 18 präsentieren die Ergebnisse des zweiten Modells. In diesem Modell sind alle fünf Koeffizienten signifikant und der Maximalwert der VIF bliebt circa 10. Deshalb können alle diese Variablen im Modell einbezogen werden. Da es jedoch noch die Korrelationen inmitten dieser Variablen wie die vorgenannte Korrelationsmatrix hindeutet, sollte ein erweitertes Modell mit Kreuzabhängigkeiten untersucht werden. Beruhend auf die Korrelationen werden der Zusammenhang unter dem Produktpreis, den anderen Variablen bezüglich der Bestellung und der Produktkategorie als Kreuzabhängigkeiten zusätzlich eingeschlossen. Allerdings wird die Mitwirkung jeder eigenen erklärenden Variablen verteilt und es führt auf die Bedeutungslosigkeit

der entsprechenden Koeffizienten zu. Dies wird durch das Resultat des dritten Modells im Folgenden illustriert.

```
Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          -3.837785   0.234600 -16.359  < 2e-16 ***
mod_preis            -0.021465   0.001838 -11.677  < 2e-16 ***
cust_prev_return_rate  2.065083   0.231424   8.923  < 2e-16 ***
qty_same_subcat        1.560819   0.056506  27.622  < 2e-16 ***
num_subcat_ord         0.666722   0.035123  18.982  < 2e-16 ***
mod_subcat1_Damen      2.203016   0.196069  11.236  < 2e-16 ***
mod_subcat1_Herren     1.159511   0.210636   5.505  3.7e-08 ***
mod_subcat1_Maedchen   0.648994   0.203940   3.182  0.00146 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9750.7  on 7499  degrees of freedom
Residual deviance: 7223.3  on 7492  degrees of freedom
AIC: 7239.3
```

Abbildung 18: Das Modell der logistischen Regression mit 5 erklärenden Variablen

```
> vif(var5_log_mod)
           mod_preis cust_prev_return_rate       qty_same_subcat         num_subcat_ord
            1.811967              1.012308              1.086772              1.146173
    mod_subcat1_Damen    mod_subcat1_Herren  mod_subcat1_Maedchen
           10.847311             10.381548              4.232985
```

Abbildung 19: Das Ergebnis des Tests der Multikollinearität im Modell mit 5 Variablen

```
                                                    Estimate  Std. Error     z value     Pr(>|z|)
(Intercept)                                       21.7660547 548.0103153  0.03971833 9.683177e-01
cust_prev_return_rate                              2.0775358   0.2320059  8.95466842 3.407582e-19
mod_preis                                         -0.4863733   9.8158299 -0.04954989 9.604811e-01
num_subcat_ord                                   -26.2456450 547.9477470 -0.04789808 9.617975e-01
qty_same_subcat                                  -22.7418892 547.9621507 -0.04150266 9.668952e-01
mod_subcat1_Damen                                -23.7003940 548.0127369 -0.04324789 9.655039e-01
mod_subcat1_Herren                               -26.5827171 548.0155887 -0.04850723 9.613120e-01
mod_subcat1_Maedchen                             -24.9085980 548.0339357 -0.04545083 9.637480e-01
mod_preis:num_subcat_ord                           0.4888822   9.8150495  0.04980945 9.602742e-01
mod_preis:qty_same_subcat                          0.4467251   9.8152573  0.04551333 9.636981e-01
num_subcat_ord:qty_same_subcat                    26.7179458 547.9197950  0.04876251 9.611086e-01
mod_preis:mod_subcat1_Damen                        0.4784224   9.8158428  0.04873982 9.611266e-01
num_subcat_ord:mod_subcat1_Damen                  27.1613114 547.9482570  0.04956912 9.604658e-01
qty_same_subcat:mod_subcat1_Damen                 23.8291549 547.9632494  0.04348678 9.653135e-01
mod_preis:mod_subcat1_Herren                       0.4808289   9.8158501  0.04898495 9.609313e-01
num_subcat_ord:mod_subcat1_Herren                 28.1796886 547.9488908  0.05142759 9.589848e-01
qty_same_subcat:mod_subcat1_Herren                25.4574525 547.9639334  0.04645826 9.629450e-01
mod_preis:mod_subcat1_Maedchen                     0.4722443   9.8161016  0.04810915 9.616293e-01
num_subcat_ord:mod_subcat1_Maedchen               26.7295825 547.9573191  0.04878041 9.610943e-01
qty_same_subcat:mod_subcat1_Maedchen              25.6959448 547.9754780  0.04689251 9.625989e-01
mod_preis:num_subcat_ord:qty_same_subcat          -0.4862728   9.8147164 -0.04954527 9.604848e-01
mod_preis:num_subcat_ord:mod_subcat1_Damen        -0.4960331   9.8150521 -0.05053800 9.596937e-01
mod_preis:qty_same_subcat:mod_subcat1_Damen       -0.4488541   9.8152628 -0.04573022 9.635253e-01
num_subcat_ord:qty_same_subcat:mod_subcat1_Damen -26.5321580 547.9200609 -0.04842341 9.613788e-01
mod_preis:num_subcat_ord:mod_subcat1_Herren       -0.4995727   9.8150538 -0.05089862 9.594063e-01
mod_preis:qty_same_subcat:mod_subcat1_Herren      -0.4573269   9.8152637 -0.04659344 9.628373e-01
num_subcat_ord:qty_same_subcat:mod_subcat1_Herren -27.5380140 547.9202548 -0.05025916 9.599159e-01
mod_preis:num_subcat_ord:mod_subcat1_Maedchen     -0.4896995   9.8151598 -0.04989216 9.602083e-01
mod_preis:qty_same_subcat:mod_subcat1_Maedchen    -0.4770735   9.8154130 -0.04860452 9.612345e-01
num_subcat_ord:qty_same_subcat:mod_subcat1_Maedchen -27.7898246 547.9255759 -0.05071825 9.595500e-01
mod_preis:num_subcat_ord:qty_same_subcat:mod_subcat1_Damen  0.4875513   9.8147178  0.04967553 9.603810e-01
mod_preis:num_subcat_ord:qty_same_subcat:mod_subcat1_Herren 0.4943000   9.8147181  0.05036314 9.598330e-01
mod_preis:num_subcat_ord:qty_same_subcat:mod_subcat1_Maedchen 0.5071951 9.8147854  0.05167664 9.587864e-01
```

Abbildung 20: Das Modell mit 5 Hauptvariablen und zusätzlichen Kreuzabhängigkeiten

Basierend auf dieses Ergebnis werden mehrere neuen Modelle mit der Anpassung von möglicher Kreuzabhängigkeit versucht. Darüber hinaus werden nicht nur die p-Werte der Koeffizienten geprüft, sondern die Leistungsfähigkeit dieser Modelle auf der Validierungsdaten werden durch die MSE Werte (Mean Squared Error) auch ausgewertet. Insgesamt gibt es acht getestete Modelle mit verschiedenen Kombinationen von möglichen Interaktionsterms zwischen vier grundsätzlichen

erklärenden Variablen. Die Variable – bisherige Retourenquote – ist unabhängig von den anderen nach der Korrelationsmatrix, deshalb wird sie nicht in Kreuzabhängigkeit einbezogen. Zufolge der Ergebnisse dieser acht Modelle gibt es nur ein Modell, dessen alle Koeffizienten signifikant sind und dessen MSE auf Validierungsdaten im Vergleich zu der anderen sieben Modelle auch minimal ist, während in anderen Modellen die Interaktionsterms insignifikant sind. Das ist das Modell mit der Kreuzabhängigkeit zwischen der Anzahl der Produkte mit der gleichen Subkategorie und der Anzahl der Subkategorien aller Produkte im Warenkorb. Es wird in Abbildung 28 als „CrQtyNum“ bezeichnet. Diese Resultate werden im Folgenden von Abbildung 21 bis Abbildung 28 gezeigt.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.697005059	0.932398639	-3.9650477	7.338128e-05
cust_prev_return_rate	2.070968720	0.231359278	8.9513105	3.512869e-19
mod_preis	-0.016277981	0.009182454	-1.7727267	7.627401e-02
num_subcat_ord	0.901011723	0.478382774	1.8834535	5.963893e-02
qty_same_subcat	1.255858431	0.616994049	2.0354466	4.180596e-02
mod_subcat1_Damen	2.217484825	0.203302639	10.9073096	1.063575e-27
mod_subcat1_Herren	1.161176650	0.217811933	5.3310975	9.762098e-08
mod_subcat1_Maedchen	0.662083722	0.203141953	3.2592171	1.117201e-03
mod_preis:num_subcat_ord	-0.005611676	0.004490647	-1.2496364	2.114324e-01
mod_preis:qty_same_subcat	-0.002083022	0.005660749	-0.3679764	7.128908e-01
num_subcat_ord:qty_same_subcat	-0.077693498	0.336306397	-0.2310200	8.172993e-01
mod_preis:num_subcat_ord:qty_same_subcat	0.003334460	0.003066738	1.0872986	2.769049e-01

Abbildung 21: Das Modell mit der Kreuzabhängigkeiten zwischen dem Produktpreis und anderen Variablen bezüglich des Inhaltes der Bestellung

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.96571955	1.894729164	-2.0930271	3.634673e-02
cust_prev_return_rate	2.07291112	0.231564477	8.9517665	3.498384e-19
mod_preis	-0.02142029	0.001850082	-11.5780226	5.326043e-31
num_subcat_ord	1.01052202	1.549732976	0.6520620	5.143612e-01
qty_same_subcat	1.73963412	1.641061722	1.0600662	2.891145e-01
mod_subcat1_Damen	3.37212329	1.908690861	1.7667205	7.727505e-02
mod_subcat1_Herren	1.05900548	1.935139313	0.5472503	5.842068e-01
mod_subcat1_Maedchen	1.17221738	2.018975234	0.5806002	5.615100e-01
num_subcat_ord:qty_same_subcat	-0.35579410	1.389367612	-0.2560835	7.978864e-01
num_subcat_ord:mod_subcat1_Damen	-0.81087035	1.554083129	-0.5217677	6.018321e-01
qty_same_subcat:mod_subcat1_Damen	-0.85043634	1.652015261	-0.5147872	6.067017e-01
num_subcat_ord:mod_subcat1_Herren	-0.36500591	1.560925421	-0.2338394	8.151096e-01
qty_same_subcat:mod_subcat1_Herren	-0.31937311	1.658922631	-0.1925184	8.473362e-01
num_subcat_ord:mod_subcat1_Maedchen	-0.44860459	1.617392368	-0.2773629	7.815015e-01
qty_same_subcat:mod_subcat1_Maedchen	-0.76911867	1.729887973	-0.4446061	6.566045e-01
num_subcat_ord:qty_same_subcat:mod_subcat1_Damen	0.66355745	1.392232164	0.4766141	6.336370e-01
num_subcat_ord:qty_same_subcat:mod_subcat1_Herren	0.51374357	1.394602953	0.3683798	7.125901e-01
num_subcat_ord:qty_same_subcat:mod_subcat1_Maedchen	0.62955410	1.437568379	0.4379298	6.614371e-01

Abbildung 22:Das Modell mit der Kreuzabhängigkeiten zwischen der Produktkategorie und anderen Variablen bezüglich des Inhaltes der Bestellung

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.57282718	0.603388830	-5.92126834	3.194681e-09
cust_prev_return_rate	2.06328994	0.231464940	8.91404956	4.920182e-19
mod_preis	-0.02155983	0.001843446	-11.69539249	1.345636e-31
num_subcat_ord	0.66817370	0.035168056	18.99944965	1.723421e-80
qty_same_subcat	1.34098817	0.475481864	2.82027196	4.798296e-03
mod_subcat1_Damen	1.97644682	0.598076776	3.30467074	9.508804e-04
mod_subcat1_Herren	0.75654747	0.610446150	1.23933530	2.152213e-01
mod_subcat1_Maedchen	0.67953304	0.641370131	1.05950216	2.893711e-01
qty_same_subcat:mod_subcat1_Damen	0.19500598	0.481379348	0.40509836	6.854052e-01
qty_same_subcat:mod_subcat1_Herren	0.31618399	0.484319467	0.65284179	5.138583e-01
qty_same_subcat:mod_subcat1_Maedchen	-0.01026277	0.513325251	-0.01999272	9.840492e-01

Abbildung 23: Das Modell mit der Kreuzabhängigkeiten zwischen der Produktkategorie und der Anzahl der Produkte im Warenkorb mit der gleichen Subkategorie

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.82453608	0.52718234	-7.2546741	4.026300e-13
cust_prev_return_rate	2.06686085	0.23162316	8.9233774	4.522804e-19
mod_preis	-0.02147728	0.00184178	-11.6611520	2.013002e-31
qty_same_subcat	1.57069677	0.05688323	27.6126486	7.843950e-168
num_subcat_ord	0.64702714	0.38134157	1.6967128	8.975097e-02
mod_subcat1_Damen	2.36203592	0.51863582	4.5543247	5.255413e-06
mod_subcat1_Herren	0.75876828	0.53636331	1.4146536	1.571701e-01
mod_subcat1_Maedchen	0.27415836	0.56391196	0.4861723	6.268450e-01
num_subcat_ord:mod_subcat1_Damen	-0.06130037	0.38352418	-0.1598344	8.730115e-01
num_subcat_ord:mod_subcat1_Herren	0.19979351	0.38755693	0.5155204	6.061894e-01
num_subcat_ord:mod_subcat1_Maedchen	0.25813548	0.41143547	0.6274021	5.303957e-01

Abbildung 24: Das Modell mit der Kreuzabhängigkeiten zwischen der Produktkategorie und der Anzahl der Subkategorien aller Produkte im Warenkorb

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.229834843	0.412807173	-7.824076	5.113976e-15
cust_prev_return_rate	2.060526309	0.231510259	8.900367	5.566224e-19
mod_preis	-0.027611145	0.003923038	-7.038206	1.947312e-12
qty_same_subcat	1.109979007	0.258414244	4.295348	1.744198e-05
num_subcat_ord	0.669247936	0.035178913	19.024122	1.076760e-80
mod_subcat1_Damen	2.233682739	0.196315240	11.378040	5.379490e-30
mod_subcat1_Herren	1.187388476	0.210880810	5.630614	1.795690e-08
mod_subcat1_Maedchen	0.659630601	0.203067161	3.248337	1.160816e-03
mod_preis:qty_same_subcat	0.004264326	0.002394285	1.781043	7.490536e-02

Abbildung 25: Das Modell mit der Kreuzabhängigkeiten zwischen dem Produktpreis und der Anzahl der Produkte im Warenkorb mit der gleichen Subkategorie

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.9272017414	0.401603773	-9.7787969	1.388511e-22
cust_prev_return_rate	2.0640012928	0.231470948	8.9168914	4.795597e-19
mod_preis	-0.0204505246	0.004124635	-4.9581416	7.117067e-07
num_subcat_ord	0.7164043184	0.184480444	3.8833618	1.030221e-04
qty_same_subcat	1.5600558567	0.056567135	27.5788379	1.996530e-167
mod_subcat1_Damen	2.1888618548	0.202697262	10.7986750	3.492064e-27
mod_subcat1_Herren	1.1452225963	0.216936027	5.2790798	1.298343e-07
mod_subcat1_Maedchen	0.6456643560	0.204335124	3.1598305	1.578610e-03
mod_preis:num_subcat_ord	-0.0004797772	0.001748541	-0.2743872	7.837871e-01

Abbildung 26: Das Modell mit der Kreuzabhängigkeiten zwischen dem Produktpreis und der Anzahl der Subkategorien aller Produkte im Warenkorb

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-3.182442	0.280845	-11.332	< 2e-16	***
mod_preis	-0.021381	0.001843	-11.598	< 2e-16	***
cust_prev_return_rate	2.070610	0.231016	8.963	< 2e-16	***
qty_same_subcat	1.034685	0.137848	7.506	6.10e-14	***
num_subcat_ord	0.309328	0.092907	3.329	0.00087	***
mod_subcat1_Damen	2.233719	0.195466	11.428	< 2e-16	***
mod_subcat1_Herren	1.183244	0.210194	5.629	1.81e-08	***
mod_subcat1_Maedchen	0.664559	0.202962	3.274	0.00106	**
qty_same_subcat:num_subcat_ord	0.278969	0.068004	4.102	4.09e-05	***

Abbildung 27: Das Modell mit der Kreuzabhängigkeiten zwischen der Anzahl der Produkte mit der gleichen Subkategorie und der Anzahl der Subkategorien aller Produkte im Warenkorb

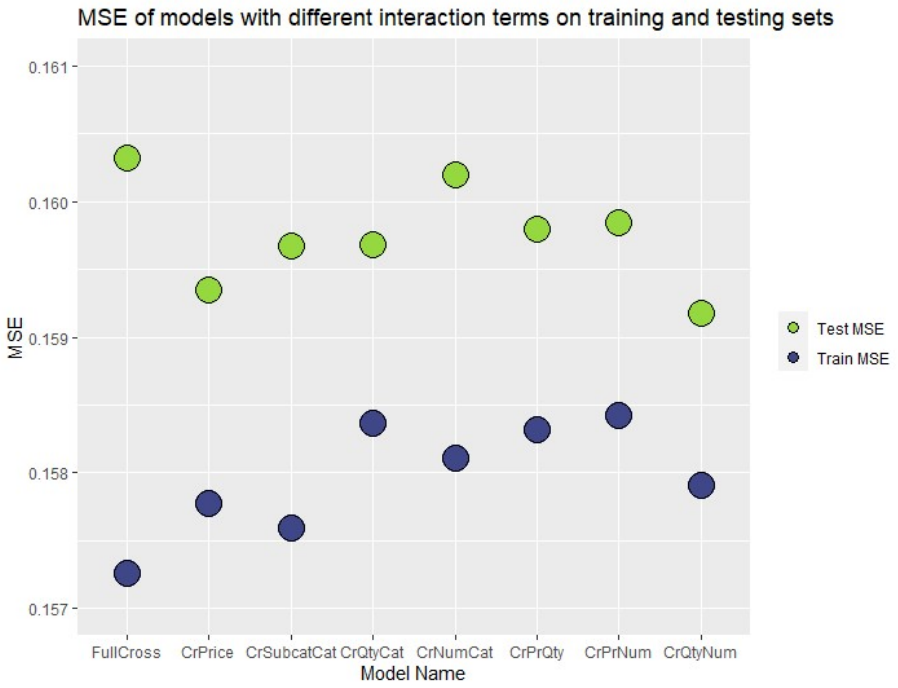

Abbildung 28: Die MSEs der Modelle mit Interaktionsterms bei Trainingsdaten und Validierungsdaten

Die Leistung dieses Modells wird auch zu der der zweiten bisherigen untersuchten Modelle verglichen. Auch wenn es als das optimale Modell zu sein scheint, wird für einen umfassenden Analyseprozess auch ein nächstes Modell mit polynomialen Abhängigkeiten geprüft. Denn das Problem, dass die Korrelation in der Korrelationsmatrix signifikant ist, aber der Korrelationskoeffizient in der linearen Regression nicht, kann darauf hinweisen, dass die Beziehung zwischen diesen Variablen nicht linear, sondern in einer anderen Form ist. Allerdings macht das Ergebnis des Modells mit polynominalen Terms in Abbildung 29 sichtbar, dass es keine bessere Anpassung an diese Daten bieten kann. Diese polynominalen Terme sind meist insignifikant und die Leistung bei den Validierungsdaten ist auch nicht besser als die der anderen.

```
Coefficients:
                                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                                       -2.304e+00  1.123e+00  -2.053  0.04008 *
poly(mod_preis, 3)1                               -4.565e+01  4.058e+00 -11.249  < 2e-16 ***
poly(mod_preis, 3)2                                3.608e+00  3.715e+00   0.971  0.33138
poly(mod_preis, 3)3                                9.180e-01  3.252e+00   0.282  0.77770
poly(cust_prev_return_rate, 3)1                    2.207e+01  2.485e+00   8.879  < 2e-16 ***
poly(cust_prev_return_rate, 3)2                   -8.916e-01  2.427e+00  -0.367  0.71330
poly(cust_prev_return_rate, 3)3                    3.546e+00  2.403e+00   1.476  0.14002
poly(qty_same_subcat, 3)1                          1.071e+02  4.241e+02   0.252  0.80069
poly(qty_same_subcat, 3)2                          4.807e+01  8.798e+02   0.055  0.95643
poly(qty_same_subcat, 3)3                          1.443e+01  8.394e+02   0.017  0.98628
poly(num_subcat_ord, 3)1                           5.799e+01  2.038e+02   0.285  0.77603
poly(num_subcat_ord, 3)2                           2.996e+00  2.797e+02   0.011  0.99145
poly(num_subcat_ord, 3)3                          -2.907e-01  2.014e+02  -0.001  0.99885
mod_subcat1_Damen                                  2.272e+00  2.149e-01  10.573  < 2e-16 ***
mod_subcat1_Herren                                 1.211e+00  2.250e-01   5.380 7.44e-08 ***
mod_subcat1_Maedchen                               6.540e-01  2.032e-01   3.218  0.00129 **
poly(qty_same_subcat, 3)1:poly(num_subcat_ord, 3)1  1.702e+03  7.176e+04   0.024  0.98108
poly(qty_same_subcat, 3)2:poly(num_subcat_ord, 3)1 -1.834e+03  1.416e+05  -0.013  0.98967
poly(qty_same_subcat, 3)3:poly(num_subcat_ord, 3)1 -5.145e+03  1.460e+05  -0.035  0.97189
poly(qty_same_subcat, 3)1:poly(num_subcat_ord, 3)2  2.125e+03  9.753e+04   0.022  0.98262
poly(qty_same_subcat, 3)2:poly(num_subcat_ord, 3)2  1.398e+03  1.912e+05   0.007  0.99416
poly(qty_same_subcat, 3)3:poly(num_subcat_ord, 3)2 -3.099e+03  1.990e+05  -0.016  0.98758
poly(qty_same_subcat, 3)1:poly(num_subcat_ord, 3)3 -2.287e+02  7.446e+04  -0.003  0.99755
poly(qty_same_subcat, 3)2:poly(num_subcat_ord, 3)3 -4.079e+03  1.514e+05  -0.027  0.97850
poly(qty_same_subcat, 3)3:poly(num_subcat_ord, 3)3 -7.019e+03  1.491e+05  -0.047  0.96245
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Abbildung 29: Das Modell mit polynomialen Abhängigkeiten und Kreuzabhängigkeiten

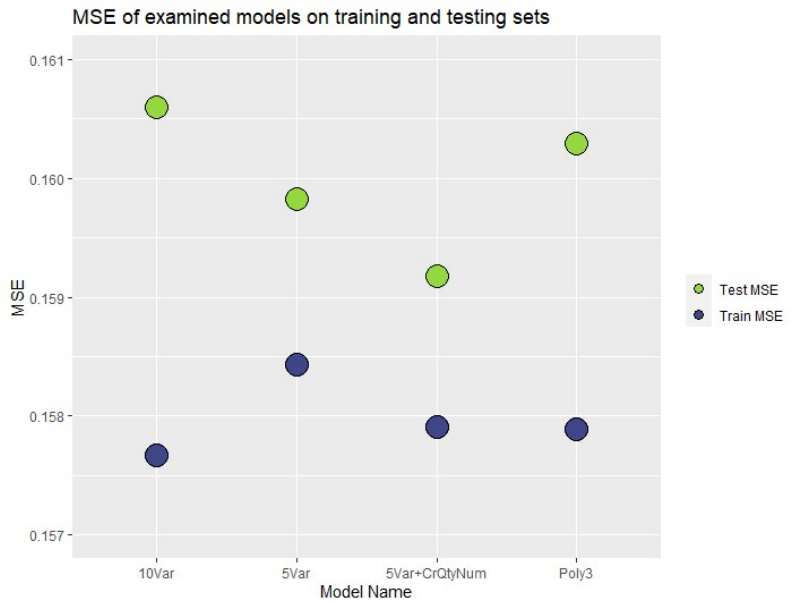

Abbildung 30: Die MSEs der untersuchten Modelle bei Trainingsdaten und Validierungsdaten

Das gezeigt Ergebnis in Abbildung 30 hat zur Folge, dass das Modell mit der Kreuzabhängigkeit zwischen der Anzahl der Produkte mit der gleichen Subkategorie und der Anzahl der Subkategorien aller Produkte im Warenkorb ein optimales ist. Es wird betrachtet als das benutzte Modell beim Datenerzeugung mit allen angemessenen einbezogen Variablen.

III.2 Die Analyse der Abhängigkeit vom Stichprobenumfang

Der folgende Abschnitt beschreibt die Analyse des Einflusses vom Stichprobenumfang auf die Qualität des geschätzten Modellparameters. In dieser Untersuchung wird der Koeffizient der Variable „bisherige Retourenquote“ im obengenannten optimalen Modell im Teil III.1 ausgewählt, zu beobachten. Nach dem Training dieses optimalen Modells auf je 1.000 zufällige Stichprobe vom Umfang 1.000, 10.000 und 50.000 nacheinander mithilfe der unabhängigen und zufälligen Stichprobenauswahl von der Grundgesamtheit, gibt es für jeden Stichprobenumfang eine eigene Verteilung dieses geschätzten Modellparameters.

Weiterhin werden sie als Graphen der Verteilungsdichte in Abbildung 31 darunter abgebildet. Alle drei Verteilungen haben ähnliche Erwartungswert, circa 2, aber die Standardabweichungen sind erheblich unterschied. Die Grafik verdeutlicht die Verhältnisse zwischen dem Stichprobenumfang und der simulierten Standardabweichung. Sie visualisiert, dass die Dichte des Schätzwerts aus dem größten Stichprobenumfang, der sehr nahe beim Mittelwert liegt, sehr hoch ist und die entsprechende Verteilung mit dünnen Verteilungsenden ist. Daraus kann man schließen, dass je größer der Stichprobenumfang ist, desto kleiner ist die Standardabweichung der geschätzten Koeffizienten. Hinzu kommt der statistische Überblick dieser Verteilungen in Tabelle 4, damit kann

die Unterschiede zwischen der Standardabweichung jeder Verteilung durch konkrete numerische Werte klar dargestellt werden.

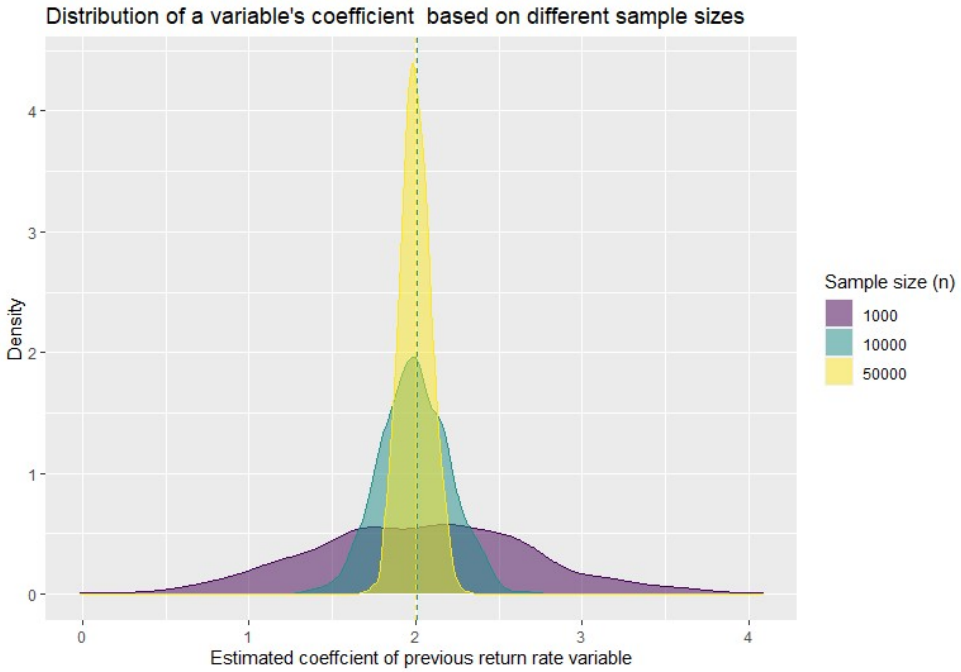

Abbildung 31: Die Verteilung der geschätzten Koeffizienten bei jedem Stichprobenumfang

Stichprobenumfang	**n = 1.000**	**n = 10.000**	**n = 50.000**
Erwartungswert	2,030	1,9939	1,9993
Varianz	0,4288	0,0428	0,0081
Standardfehler	0,6548	0,2070	0,0901

Tabelle 4: Die Eigenschaften der Verteilung des geschätzten Koeffizienten

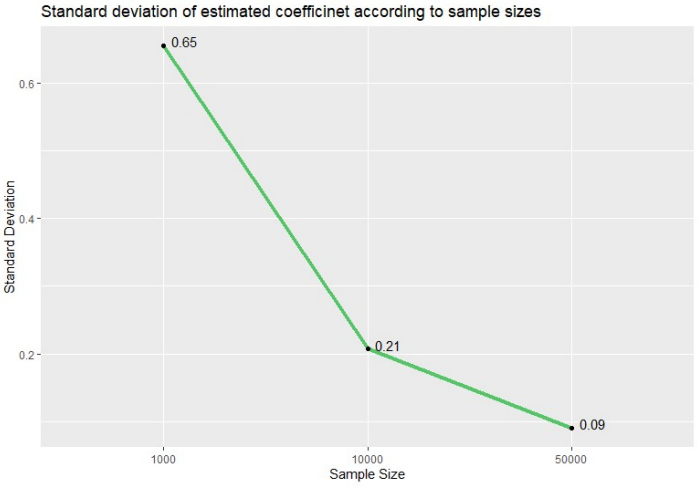

Abbildung 32: Der simulativ Standardfehler des Modellparameters für die unterschiedlichen Stichprobengrößen

Angesichts dieser Ergebnisse liegt die Schlussfolgerung nahe, dass nach dem starken Gesetz der großen Zahlen die simulierte Standardabweichung in Abhängigkeit von den Stichprobengrößen sich entwickelt (Feller, 1968). Weil aus der Abbildung 31 es sich eindeutig ableiten lässt, dass mit zunehmender Stichprobengröße die Varianz definitiv abnimmt und der Schätzwert gegen den Mittelwert konvergiert.

Im Teil III.1 wird das optimale Modell durch den Trainingsdaten nur mit dem Umfang n = 7.500 trainiert, deshalb sollte noch ein Trainieren auf der ganzen Stichprobe vom Umfang n = 10.000 durchgeführt werden, um einen vergleichbaren Standardfehler zu erreichen. Damit kann die Schlussfolgerung auf den Zusammenhang zwischen dem unter Einzeldatensatz ermittelten Standardfehler des Modellparameters mit der simulativ ermittelten Standardabweichung des Modellparameters für die unterschiedlichen Stichprobenumfänge gezogen werden. Der unter Einzeldatensatz ermittelten Standardfehler des Modellparameters ist annähernd gleich zur Wurzel der simulativ ermittelten Standardabweichung für die entsprechende gleiche Stichprobengröße n = 10.000. Dieses Ergebnis wird durch den gezeigten Wert in Abbildung 33 und auch den Wert in zweiten Spalten in Tabelle 4 gedeutet.

```
Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                -3.066359   0.239304 -12.814  < 2e-16 ***
mod_preis                  -0.022021   0.001586 -13.886  < 2e-16 ***
cust_prev_return_rate       2.205199   0.202079  10.913  < 2e-16 ***
qty_same_subcat             1.065624   0.121911   8.741  < 2e-16 ***
num_subcat_ord              0.371296   0.082873   4.480 7.45e-06 ***
mod_subcat1_Damen           2.029063   0.153513  13.218  < 2e-16 ***
mod_subcat1_Herren          1.011744   0.166599   6.073 1.26e-09 ***
mod_subcat1_Maedchen        0.461242   0.162426   2.840  0.00452 **
qty_same_subcat:num_subcat_ord  0.265586   0.060221   4.410 1.03e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 12978.4  on 9999  degrees of freedom
Residual deviance:  9534.6  on 9991  degrees of freedom
AIC: 9552.6
```

Abbildung 33: Das trainiert auf einer Stichprobe der Größe n =10.000 optimale Modell

IV. Optimierung der Parametrisierung

Der folgende Teilabschnitt konzentriert sich auf die Untersuchung der Modellkomplexität, dadurch kann ein optimales Modell mit angemessenem Komplexitätsgrad festgestellt werden. Die Hauptvorgehensweise ist die Überprüfung von Prognosefähigkeit der untersuchten Modelle auf der Validierungsdaten durch den MSE.

IV.1 Logistische Regression Modell mit nur einer erklärenden Variablen

Als Erstes wird ein Modell der Logistische Regression mit nur einer erklärenden Variablen ausgewählt. Das Verfahren der Auswahl fängt mit dem Modell, das alle möglichen zehn Variablen beinhaltet, an. Danach werden nur die Variablen, deren geschätzten Koeffizienten signifikant sind, im nächsten Schritt behalten. Zufolge des präsentierten Ergebnisses in Abbildung 34 bleiben diese fünf Variablen einschließlich des Produktpreis (mod_preis), der bisherigen

Retourenquote(cust_prev_return_rate), der Anzahl der Produkte im Warenkorb mit der gleichen Subkategorie(qty_same_subcat), der Anzahl der Subkategorien aller Produkte im Warenkorb (num_subcat_ord) und der Produktkategorie (mod_subcat1_Damen, mod_subcat1_Herren und mod_subcat1_Maedchen) bestehen.

```
Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          -3.447957   0.462511  -7.455 9.00e-14 ***
prod_num_per_order   -0.029672   0.030110  -0.985  0.32440
prod_size            -0.013738   0.010620  -1.294  0.19581
mod_preis            -0.020099   0.002269  -8.857  < 2e-16 ***
cust_prev_return_rate  1.908893   0.289745   6.588 4.45e-11 ***
hour_of_order         0.005173   0.006697   0.773  0.43981
qty_same_subcat       1.699258   0.076402  22.241  < 2e-16 ***
num_cat_ord           0.029925   0.064086   0.467  0.64054
num_subcat_ord        0.678624   0.056172  12.081  < 2e-16 ***
mod_subcat1_Damen     1.956196   0.243856   8.022 1.04e-15 ***
mod_subcat1_Herren    0.990307   0.291835   3.393  0.00069 ***
mod_subcat1_Maedchen  0.358714   0.227166   1.579  0.11432
cust_region_BE        0.050315   0.175937   0.286  0.77489
cust_region_BY        0.121851   0.174038   0.700  0.48384
cust_region_HB        0.113544   0.175246   0.648  0.51704
cust_region_HE       -0.046351   0.173030  -0.268  0.78879
cust_region_HH       -0.031238   0.178416  -0.175  0.86101
cust_region_MV        0.037138   0.171450   0.217  0.82851
cust_region_NI       -0.054494   0.179819  -0.303  0.76185
cust_region_NW       -0.196036   0.180761  -1.085  0.27814
cust_region_RP       -0.057971   0.178338  -0.325  0.74514
cust_region_SH        0.074333   0.180032   0.413  0.67969
cust_region_SL       -0.013613   0.181164  -0.075  0.94010
cust_region_SN        0.139670   0.170507   0.819  0.41270
cust_region_ST        0.196617   0.174198   1.129  0.25903
cust_region_TH        0.030415   0.176427   0.172  0.86313
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6470.7  on 4999  degrees of freedom
Residual deviance: 4792.9  on 4974  degrees of freedom
AIC: 4844.9
```

Abbildung 34: Das Logistische Regression Modell mit 10 Variablen auf Trainingsdaten mit 5000 Einträge

Laut der geschätzten Koeffizienten sind erwartungsgemäß die Einflüsse der „cust_prev_return_rate“, „qty_same_subcat“ und „“mod_subcat1_Damen wesentlich größer als die der anderen. Daher sollte die gewählte erklärende Variable unter diesen dreien sein. Allerdings sollte die Begründung der Auswahl die Prognoseleistung auf Validierungsdaten sein. Im nächsten Schritt werden fünf Modelle jeweils mit einer erklärenden Variablen erstellt und die MSEs der Modelle auf Validierungsdaten werden miteinander vergleichen, damit das optimale Modell mit einer erklärenden Variablen bestimmt wird. Es ist auch möglich, dass nur drei Modelle mit drei obengenannten potenziellen Variablen miteinander verglichen werden. Dieses Verfahren wird mit fünf Modelle durchgeführt, um eine umfangreiche Analyse zu erreichen.

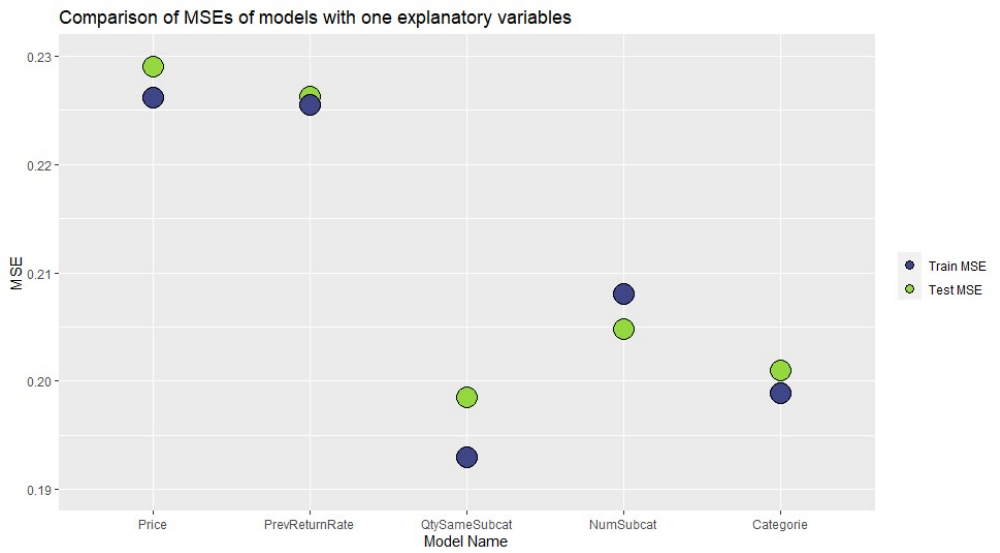

Abbildung 35: Die MSEs auf Validierungsdaten der ein-erklärende Variable Logistische Regression Modelle

Aus der Abbildung 35 kann es erkannt werden, dass das Modell mit der Variablen „qty_same_subcat“ niedrigsten MSE auf Validierungsdaten hat. Deshalb wird dieses Modell bevorzugt ausgewählt.

IV.2 Logistische Regression mit zwei erklärenden Variablen

Das Verfahren der Auswahl des zwei Variablen Modell ist ähnlich wie der zweite Schritt im Teil IV.1 aber werden vier Modelle geprüft. In jedem Modell muss die Variable „qty_same_subcat“ einbezogen werden, weil nach dem Ergebnis des Teil IV.1 das Modell mit dieser Variablen besser Vorhersageleistung auf Validierungsdaten im Vergleich zu anderen hat.

Die Abbildung 36 vergleicht die Leistung der Vorhersage dieser Modelle und sie zeigt, dass der MSE des mit einer zusätzlichen Variablen Kategorie Modells unter vier untersuchten Modellen minimal ist. Deswegen wird dieses Modell mit zwei erklärenden Variablen, der Anzahl der Produkte im Warenkorb mit der gleichen Subkategorie(qty_same_subcat) und der Produktkategorie (mod_subcat1_Damen, mod_subcat1_Herren und mod_subcat1_Maedchen), ausgewählt. Dieses Ergebnis stimmt auch mit der Erwartung überein, die auf der Größe des Koeffizienten beruht.

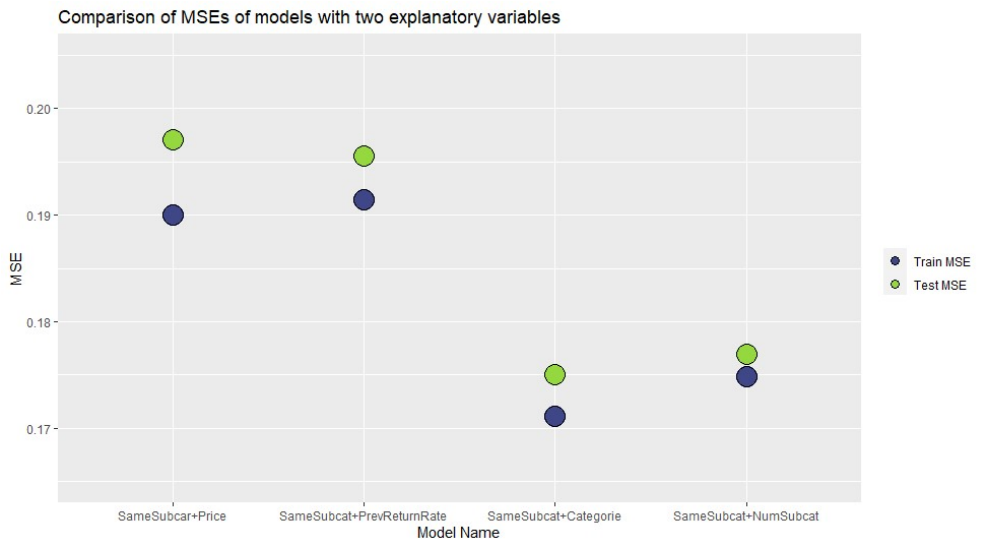

Abbildung 36: Die MSEs auf Validierungsdaten der zwei-erklärenden Variablen Logistische Regression Modelle

IV.3 Der Leistungsvergleich zwischen sechs untersuchten Modellen

Im folgenden Abschnitt wird der Vergleich zwischen sechs untersuchten Modellen einschließlich des logistische Regression mit einer erklärenden Variablen Modells (identifiziert in Teil IV.1), des logistische Regression mit zwei erklärenden Variablen Modells (obengenannt in Teil IV.2), des optimales Modells – das beim Grundgesamtheiterzeugung benutzt wurde, des logistische Regression mit fünf erklärenden Variablen und zusätzlich Kreuzabhängigkeiten und polynominalen Abhängigkeiten, des KNN Modells mit k = 100 und des KNN Modells mit k = 1 beschrieben. Bei jedem Modell werden beide MSE auf Trainingsdaten und Validierungsdaten berechnet und diese werden mit dem minimalen MSE der Validierungsdaten verglichen. Das Ergebnis dieses Vergleichs wird in folgender Abbildung 37 veranschaulicht.

Die klare Erkenntnis kann man dieser Grafik entnehmen ist, dass wenn ein Modell wirklich optimal ist, bleibt der MSE auf der Validierungsdaten sehr nähe von den minimalen MSE. Es weist auf die Erreichung einer minimalen Varianz und eines minimalen Bias im Modell hin. Eine weitere Dimension ist, wenn ein Modell erheblich einfach ist und es kann nicht alle signifikante Abhängigkeiten beinhalten, wird die Vorhersageleistung auf den neuen Daten wichtig beeinträchtigt, weil das Bias des Modells mächtig ist und das Modell „unterparametrisiert“ ist. Das ist der Fall des ersten Modells, das nur eine erklärende Variable einbezieht. An dieser Stelle ist hinzuzufügen, dass wenn relevante Faktoren weiter im Modell involvieren können, wird die Prognosefähigkeit des Modells verbessert. Dies wird anhand der Fälle des logistische Regression Modell mit zwei erklärenden Variablen und des optimalen Modells mit insgesamt sechs Variablen bewiesen. Die MSE auf Trainingsdaten und Validierungsdaten dieser beiden Modelle sind jeweils niedriger als die des ersten Modells.

Andererseits muss man dabei jedoch berücksichtigen, ob die zusätzlichen hinzugefügten Faktoren wirklich nützliche Informationen oder nur Störung einbringen. Als Beispiel kann das Modell mit zusätzlich mehrere polynominalen Abhängigkeiten und auch mit Kreuzabhängigkeiten gelten: das überparametrisierte Modell haben deutlich mehr Variablen als das optimale Modell aber der MSE des Modells auf Trainingsdaten ist annähernd gleich und das Vorhersageleistung auf Validierungsdaten wird nicht verbessert. Dies bedeutet, dass ein kompliziertes Modell nicht automatisch gleich wie ein Modell mit gut Vorhersagefähigkeit auf neuen Daten ist. Die tatsächliche Leistung auf Validierungsdaten könnte im Gegensatz zu solcher Erwartung stehen. Das ist das Problem der Überparametrisierung (James et al., 2013). Das Ergebnis von dem Modell KNN mit k=1 kann diese Ansicht erwiesen. Zwar kann das Modell perfekt zu den Trainingsdaten passen und es ergibt sich keinen Fehler, aber die Vorhersagen für neuen Daten sind erheblich unpräzise. Im Gegensatz dazu ist der MSE auf Trainingsdaten des Modells mit k = 100 höher aber der MSE auf Validierungsdaten ist deutlich niedriger.

Hier ist außerdem noch der Gedanke aufzugreifen, dass der MSE des KNN Modells mit k = 100 auf Trainingsdaten höher als die der anderen logistischen Regression Modelle ist und dies kann hindeuten, dass dieses Modells unterparametrisiert ist. Daher könnte dieses Modell nicht generalisiert werden, um zu prognosezieren. Das ist vor allem auf die wichtigste Überlegung der ausgewählte Wert von k bei der Modellentwicklung zurückzuführen.

Die Komplexität der Parametrisierung ist einer der wichtigsten Themen bei der Untersuchung von optimalen Modellen. Unterparametrisiert Modelle sind nicht gültig zu generalisieren, deswegen können sie auch nicht benutzt werden, um mit neuen Daten vorherzusagen. Überparametrisiert Modelle sind nicht nur mit großem Zeit- und Mittelaufwand verbunden, sondern auch ihre Prognose sind nicht aussagekräftig, obwohl ihre Leistung mit Trainingsdaten gut ist. Angesichts dieser Ergebnisse liegt die Schlussfolgerung nahe, dass die Vorgehensweise von der Datenaufteilung in Trainings- und Validierungsdaten in Data Science sehr hilfreich ist. Damit können die Leistungen zwischen verschiedenen Modellen mit verschiedener Komplexität verglichen werden. Beim Verfahren der Untersuchung eines optimalen Modells ist es erforderlich, eine umfangreiche Analyse mit vielfältigen Modellarten und bei jeder Art mit verschiedene Komplexitätsstufe durchzuführen. Dadurch können sowohl ein Umfang der Vorhersageleistung von Modellen als auch eine entsprechende Rangfolge von Komplexität der Parametrisierung festgestellt werden. Demzufolge kann ein optimales Modell mit angemessener erwarteter Prognoseleistung identifiziert werden.

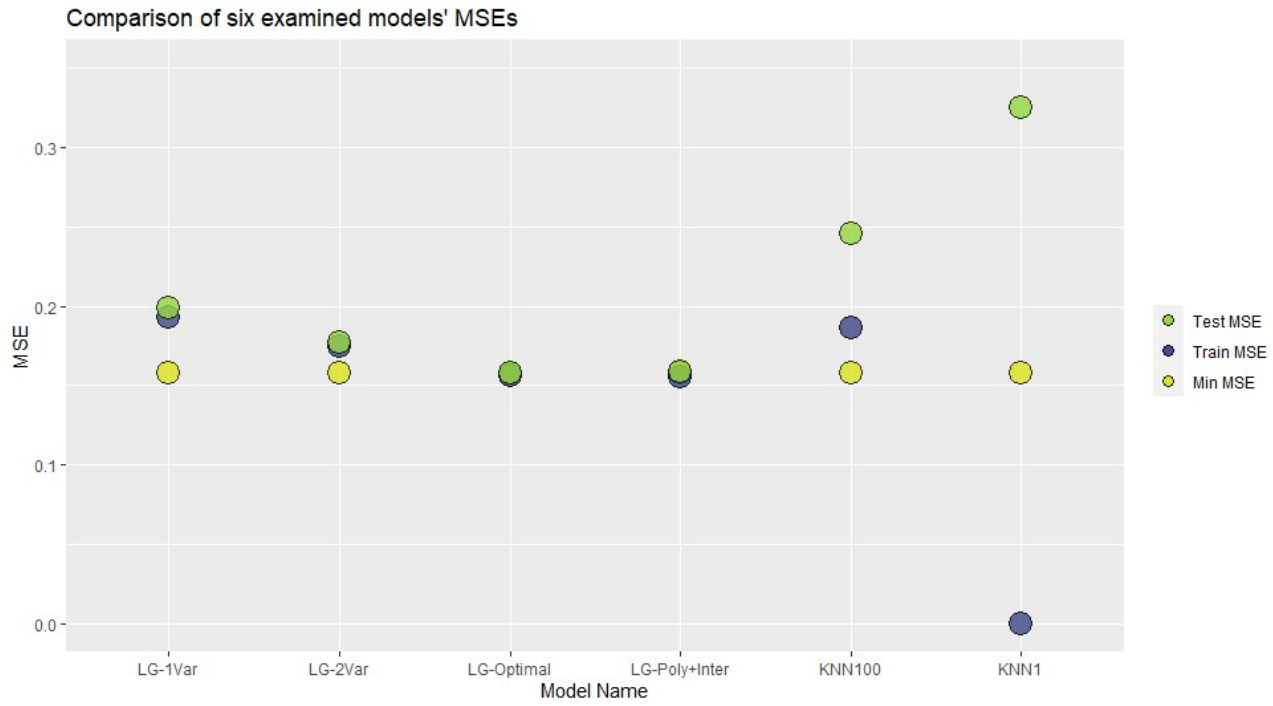

Abbildung 37: Der Vergleich der MSE zwischen sechs untersuchten Modellen

V. Bibliographie

[1]. P. Urbanke, J. Kranz und L. Kolbe, „Predicting Product Returns in E-Commerce: The Contribution of Mahalanobis Feature Extraction “, Thirty Sixth International Conference on Information System, Forth Worth, 2015.

[2]. Feller, W. "The Strong Law of Large Numbers." §10.7 in “An Introduction to Probability Theory and Its Applications”, Vol. 1, 3rd ed. New York: Wiley, 1968

[3]. G. James, D. Witten, T. Hastie, R. Tibshirani, „An Introduction to Statistical Learning with Applications in R”, New York : Springer Science+Business Media, 2013.