TECHNISCHE HOCHSCHULE MITTELHESSEN

CAMPUS FRIEDBERG

MND
Mathematik, Naturwissenschaften und Datenverarbeitung

Bachelorarbeit

Künstliche Intelligenz zur Optimierung von Marketingkampagnen: Steigerung von Reaktionsraten und Rentabilität

zur Erlangung des akademischen Grades

Bachelor of Science (B.Sc.)

vorgelegt dem

Fachbereich Mathematik, Naturwissenschaften und Datenverarbeitung

der Technischen Hochschule Mittelhessen

Mouad Lakhroufi

im Dezember 2023

Referent: Prof. Dr. Frank Kammer
Korreferent: Prof. Dr. Harald Ritz

Abstract

In der vorliegenden Bachelorarbeit wird die Optimierung von Marketingkampagnen durch den Einsatz von Künstlicher Intelligenz und Data-Mining-Methoden untersucht. Es wird ein ausgewählter Datensatz herangezogen, der detaillierte Informationen über personalisiertes Marketing und individuelles Kundenverhalten bereitstellt. Ziel ist es, zu ergründen, wie durch präzise und zielgerichtete Kundenansprachen die Effektivität von Direktmarketing-Aktionen gesteigert werden kann. Im Fokus stehen verschiedene Vorhersagemodelle, basierend auf Methoden des maschinellen Lernens, die darauf abzielen, Kundenreaktionen auf Marketingmaßnahmen vorherzusagen. Dies soll zu einer Erhöhung der Reaktionsraten und einer Steigerung der Rentabilität der Kampagnen beitragen. Ergänzend wird eine RFM-Analyse (Recency, Frequency, Monetary) durchgeführt, um Kunden nach ihrem Kaufverhalten zu segmentieren. Diese Segmentierung dient der Entwicklung effizienterer, personalisierter Marketingstrategien, welche den Kundennutzen maximieren und zum Unternehmenserfolg beitragen. Die Kombination dieser analytischen Ansätze eröffnet neue Perspektiven für die Verbesserung von Marketingstrategien und erweitert das Verständnis in diesem Bereich.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

1.1 Einführung in das Thema

Die Einführung von Künstlicher Intelligenz (KI) in die globale Wirtschaft markiert eine Ära der Transformation mit weitreichenden Auswirkungen. Eine Prognose von Price Waterhouse Cooper (PwC) deutet darauf hin, dass KI bis zum Jahr 2030 über 15 Billionen US-Dollar zur Weltwirtschaft beitragen und das lokale Wirtschaftswachstum um bis zu 26% steigern könnte. Diese Prognosen unterstreichen die zunehmende Relevanz der KI in diversen Wirtschaftsbereichen und signalisieren eine Änderung der traditionellen Geschäftsmodelle [vgl. Pric2017].

Insbesondere im Marketing zeichnet sich ab, dass Künstliche Intelligenz erhebliche Veränderungen herbeiführen wird. Laut einer Studie von McKinsey & Company wird erwartet, dass das Marketing, neben dem Vertrieb, der Unternehmensbereich sein wird, in dem KI den größten finanziellen Einfluss ausüben wird [vgl. Marr2022]. Diese Vorhersage spiegelt die transformative Kraft der KI in der Marketingbranche wider und zeigt auf, wie KI-basierte Technologien das Potenzial haben, Marketingstrategien grundlegend neu zu gestalten.

Die Implementierung von KI in Marketingstrategien umfasst den Einsatz von Technologien wie Datensammlung, datengetriebener Analyse, Natural Language Processing (NLP) und Maschinellem Lernen (ML). Diese Technologien ermöglichen es Unternehmen, ein tieferes Verständnis für das Kundenverhalten zu entwickeln und ihre Marketingmaßnahmen effizienter zu gestalten. Durch den Einsatz von KI können Marketingteams von der präzisen Zielgruppenansprache, über die Optimierung von Werbebudgets, bis hin zur Vorhersage von Markttrends profitieren. Dies führt zu einer signifikanten Steigerung der Marketingeffizienz und -effektivität und ermöglicht es Unternehmen, in einem zunehmend wettbewerbsintensiven Marktumfeld erfolgreich zu agieren.

1.2 Problemstellung und Zielsetzung

Marketingkampagnen spielen eine zentrale Rolle in der modernen

Geschäftswelt, insbesondere bei der Gewinnung und Bindung von Kunden. Eine spezifische Herausforderung besteht in der präzisen Ausrichtung dieser Kampagnen auf die Kunden. Oftmals werden sämtliche Kundengruppen gleichartig angesprochen, einschließlich derer, die voraussichtlich negativ reagieren könnten. Dieser generalisierte Ansatz führt zu Ineffizienzen in den Marketingkampagnen, gekennzeichnet durch unzureichende Segmentierung, erhöhte Kosten und das Risiko eines Scheiterns der gesamten Kampagne.

Vor diesem Hintergrund ist das primäre Ziel dieser Arbeit, eine optimierte Strategie für Marketingkampagnen zu entwickeln, die eine differenzierte und zielgerichtete Kundenansprache ermöglicht. Ziel ist es, durch die Verbesserung der Präzision und Effektivität von Marketingkampagnen mittels KI-gestützter Techniken und Data-Mining-Methoden sowohl die Reaktionsraten als auch die Rentabilität der Kampagnen zu steigern. Diese Arbeit richtet sich dabei auf die zentrale Forschungsfrage: „Wie verbessern KI-gestützte Techniken und Data-Mining-Methoden die Präzision und Effektivität von Marketingkampagnen?“

Die Beantwortung dieser Forschungsfrage wird es ermöglichen, konkrete Ansätze zur Verbesserung von Marketingstrategien zu identifizieren und umzusetzen, um die Herausforderungen der modernen Marketinglandschaft effektiv zu meistern.

1.3 Vorgehensweise und Gliederung

In dieser Bachelorarbeit wird zur Erreichung der gesetzten Ziele ein speziell ausgewählter Datensatz von der Plattform „Kaggle“ verwendet. Dieser Datensatz, der eine umfassende Abdeckung relevanter Aspekte des personalisierten Marketings sowie eine detaillierte Abbildung des individualisierten Kundenverhaltens bietet, bildet die Grundlage der Untersuchung.

Der Beginn des Forschungsprozesses liegt im detaillierten Einblick in den Datensatz, wie in Kapitel 2 „Datenverständnis“ beschrieben. Dieses Kapitel befasst sich mit einer gründlichen Analyse der verfügbaren Daten, um deren Merkmale und potenzielle Einflüsse auf das personalisierte Marketing vollständig zu verstehen.

In Kapitel 3 liegt der Fokus auf der Datenaufbereitung, einem entscheidenden Bestandteil des Data-Mining-Prozesses. Dieses Kapitel behandelt die wesentlichen Schritte der Datenvorbereitung, die notwendig sind, um die Daten für die anschließende Analyse und Entwicklung von Vorhersagemodellen optimal aufzubereiten. Zu diesen Schritten gehören das Ersetzen fehlender Werte und das Entfernen von Duplikaten, um die Qualität und Genauigkeit der Daten zu gewährleisten. Darüber hinaus werden in diesem Kapitel fortgeschrittene Techniken des Feature Engineerings und der Feature-Extraktion erörtert. Diese umfassen die Transformation numerischer Merkmale und die Kodierung kategorialer Merkmale. Ein weiterer wichtiger Aspekt ist die gezielte Auswahl relevanter Features, die für die präzise Modellierung von Marketingstrategien von Bedeutung sind. Zusätzlich werden Strategien im Umgang mit unausgewogenen Datenverteilungen vorgestellt, um mögliche Verzerrungen in den Modellergebnissen zu minimieren. Die adäquate Behandlung und präzise Repräsentation von Daten sind in diesem Kontext von entscheidender Bedeutung, um die Verlässlichkeit und Effektivität der entwickelten Vorhersagemodelle zu gewährleisten.

Weiterführend befasst sich Kapitel 4 mit der Entwicklung und Bewertung verschiedener auf maschinellem Lernen basierender Klassifikationsmodelle. Diese Modelle nutzen Kundendaten, um vorherzusagen, welche Kunden die nächste Marketingkampagne annehmen und welche sie ablehnen werden. Die Evaluierung dieser Modelle umfasst nicht nur die technische Leistungsfähigkeit, sondern auch ihre Anwendbarkeit im Kontext von Marketingstrategien. Zusätzlich wird eine Clustering-Analyse unter Verwendung von RFM-Metriken (Recency, Frequency, Monetary) durchgeführt, um eine effektive Kundensegmentierung und gezielte Zielgruppenidentifikation zu ermöglichen. Diese Segmentierung ist von entscheidender Bedeutung, da sie es ermöglicht, Marketingkampagnen präziser und kundenorientierter zu gestalten, indem Verhaltensmuster und Präferenzen innerhalb der Kundenbasis identifiziert werden.

Das letzte Kapitel widmet sich der strategischen Marketinganalyse und der Bewertung der Ergebnisse. Es wird untersucht, inwieweit die implementierten Modelle zur Steigerung der Reaktionsraten und zur Reduzierung der Marketingkosten beitragen. Diese Analyse zielt darauf ab, den

Unternehmensgewinn zu maximieren und die Effizienz von Marketingkampagnen zu verbessern. Hierbei wird die praktische Anwendung der entwickelten Modelle im Rahmen strategischer Marketingentscheidungen beleuchtet.

2 Datenverständnis: Ein tiefer Einblick in den Datensatz

Das Datenverständnis stellt einen essenziellen Schritt in der Entwicklung von KI-basierten Vorhersagemodellen dar. Ein fundierter Überblick über die Beschaffenheit, Struktur und die potenziellen Einschränkungen eines Datensatzes kann die Richtung zukünftiger Analysen erheblich beeinflussen. In diesem Kapitel wird ein tiefgreifender Einblick in den ausgewählten Datensatz geboten. Mittels Grafiken und Tabellen werden verschiedene Aspekte des Datensatzes hervorgehoben und analysiert. Dies legt das Fundament für die nachfolgende Datenaufbereitung und bietet wertvolle Erkenntnisse, die zur Formulierung von Hypothesen und zur gezielten Modellierung in den weiteren Kapiteln beitragen können. Es ist essenziell, die in den Daten verborgenen Muster und Zusammenhänge zu erkennen, um zielgerichtete Entscheidungen im weiteren Verlauf der Arbeit treffen zu können.

2.1 Auswahl des Datensatzes

Ein spezifischer Datensatz, bezogen von der renommierten Online-Plattform „Kaggle“, bildet die Grundlage dieser Untersuchung [vgl. Sald2020]. „Kaggle“ steht in der Datenwissenschaft als verlässliche Quelle für vielfältige Datensätze aus unterschiedlichen Disziplinen. Der in dieser Arbeit verwendete Datensatz bietet detaillierte Informationen über Kundenverhalten im Kontext von Marketingkampagnen. Aufgrund seiner umfangreichen Merkmale, die sowohl das Kundenverhalten als auch die Kaufgewohnheiten abbilden, wurde er als besonders relevant für das vorliegende Marketingproblem identifiziert. Das Hauptziel dieses Datensatzes ist die Optimierung der Effektivität von Marketingkampagnen und die Fähigkeit, Vorhersagen über Kundenreaktionen auf unterschiedliche Produktangebote zu treffen.

2.2 Beschreibung des Datensatzes

Der Datensatz mit 2240 Einträgen bietet eine umfassende Übersicht über Kundeninteraktionen und Kaufverhalten. Er ermöglicht eine detaillierte Analyse verschiedener Aspekte der Kundenbeziehung, einschließlich ihrer Reaktionen auf Marketingmaßnahmen, demografischen Informationen, Kaufgewohnheiten und -Vorlieben sowie wirtschaftlichen Daten wie Kosten der Kundenansprache und Umsatz. Diese Vielfalt und Struktur des Datensatzes sind besonders wertvoll für die Erarbeitung fundierter Einsichten, die zur Optimierung zukünftiger Marketingstrategien beitragen können.

2.3 Merkmalsbeschreibung

Die folgende Tabelle bietet eine umfassende Beschreibung der einzelnen Merkmale des Datensatzes:

Merkmal	**Merkmalstyp**	**Daten-Typ**	**Beschreibung**
ID	Kategorial	int64	Eindeutige Identifikationsnummer für jeden Kunden.
Year_Birth	Numerisch	int64	Geburtsjahr des Kunden.
Education	Kategorial	object	Bildungsniveau des Kunden.
Marital_Status	Kategorial	object	Familienstand des Kunden.
Kidhome	Numerisch	int64	Anzahl der kleinen Kinder im Haushalt des Kunden.
Teenhome	Numerisch	int64	Anzahl der Teenager im Haushalt des Kunden.
Income	Numerisch	float64	Jährliches Haushaltseinkommen des Kunden.
DtCustomer	Datum	object	Datum der Registrierung des Kunden beim Unternehmen.
Recency	Numerisch	int64	Tage seit dem letzten Einkauf des Kunden.
MntFishProducts, MntMeatProducts, MntFruits, MntSweetProducts, MntWines, MntGoldProds	Numerisch	int64	Ausgaben für verschiedene Produktkategorien in den letzten zwei Jahren.

NumDealsPurchases, NumCatalogPurchases, NumStorePurchases, NumWebPurchases	Numerisch	int64	Anzahl der Einkäufe über verschiedene Kanäle.
AcceptedCmp1-5	Boolesch	int64	Reaktion auf verschiedene Marketingkampagnen.
Response	Boolesch	int64	Reaktion auf das letzte Marketingangebot (Zielvariable).
Complain	Boolesch	int64	Beschwert sich der Kunde in den letzten 2 Jahren?
Z_CostContact	Numerisch	int64	Kosten für den Kontakt mit einem Kunden
Z_Revenue	Numerisch	int64	Einnahmen nach Annahme des Angebots durch den Kunden

Tabelle 1: Detaillierte Übersicht der Merkmale

2.4 Deskriptive Statistische Analyse der Kundenmerkmale

	Dt_Customer	Year_Birth	Income	Kidhome	Teenhome	Recency
count	2240	2240	2216	2240	2240	2240
mean	2013-07-10 10:01:42.857142784	1968.806	52247.251	0.444	0.506	49.109
std	-	11.984	25173.077	0.538	0.545	28.962
min	2012-07-30 00:00:00	1893	1730	0	0	0
25%	2013-01-16 00:00:00	1959	35303	0	0	24
50%	2013-07-08 12:00:00	1970	51381.500	0	0	49
75%	2013-12-30 06:00:00	1977	68522	1	1	74
max	2014-06-29 00:00:00	1996	666666	2	2	99

Tabelle 2: Statistische Zusammenfassung der demografischen Daten, Familienstruktur und Kundenbindung im Kundenverhaltens-Datensatz.

Die Tabelle 2 liefert eine umfassende statistische Analyse verschiedener numerischer Kundenmerkmale, darunter Verhaltens- und Demografiedaten. Sie umfasst die Anzahl der Datenpunkte („count“), wobei beispielsweise beim Merkmal „Income“ Daten von 24 Kunden fehlen. Der Durchschnittswert („mean“) gibt Aufschluss über allgemeine Trends, wie das Durchschnittsjahr der Geburt (1968) oder das durchschnittliche Einkommen (52.247,25), dessen leicht rechtsschiefe Verteilung durch den Unterschied zwischen Mittelwert und Median (51.381,50) ersichtlich wird.

Die Standardabweichung („std“) verdeutlicht die Streuung der Daten um den Mittelwert, wobei ein hoher Wert auf eine breite Einkommensverteilung hinweist. Der kleinste Wert („min“) in jedem Merkmal kann auf mögliche Dateneingabefehler aufmerksam machen, wie ein Geburtsjahr von 1893. Quartilwerte (25%, 50%, 75%) bieten Einblick in die Datenverteilung, wobei der Median (50%) besonders hilfreich zur Bestimmung des Verteilungszentrums ist. Der höchste Wert („max“) wird ebenso festgehalten.

Diese statistischen Daten sind wertvoll für das Verständnis der Kundenmerkmale und -verhalten und geben wichtige Hinweise für die erforderliche Datenaufbereitung und -analyse.

	MntWines	MntFruits	MntMeatProducts	MntFishProducts	MntSweetProducts	MntGoldProds
count	2240	2240	2240	2240	2240	2240
mean	303.936	26.302	166.950	37.525	27.063	44.022
std	336.597	39.773	225.715	54.629	41.280	52.167
min	0	0	0	0	0	0
25%	23.750	1	16	3	1	9
50%	173.500	8	67	12	8	24
75%	504.250	33	232	50	33	56
max	1493	199	1725	259	263	362

Tabelle 3: Statistische Zusammenfassung der Ausgaben der Kunden für verschiedene Produktkategorien in den letzten 2 Jahren

Die Tabelle 3 gibt detailliert Auskunft über die Ausgaben der Kunden in verschiedenen Produktkategorien über die letzten 2 Jahre. Es ist auffällig, dass die Kunden am meisten für Wein ausgeben, wobei der Durchschnitt bei 303.936 Einheiten liegt. Dies wird gefolgt von den Ausgaben für Fleischprodukte mit einem Durchschnitt von 166.950 Einheiten. Diese Werte sind im Vergleich zu anderen Produktkategorien deutlich höher. Des Weiteren zeigt die maximale Ausgabensumme von 1725 Einheiten bei den Fleischprodukten, dass einige Kunden in dieser Kategorie besonders hohe Beträge in den letzten zwei Jahren ausgegeben haben. Die Standardabweichung, besonders bei Wein und Fleischprodukten, deutet auf eine große Varianz in den Kaufgewohnheiten hin, was auf unterschiedliche Präferenzen der Kunden hindeutet. Es ist ebenfalls interessant zu beobachten, dass der Mindestwert in jeder Kategorie 0 beträgt, was darauf hindeutet, dass

einige Kunden in bestimmten Kategorien überhaupt keine Käufe getätigt haben. Insgesamt legen diese Statistiken nahe, dass Wein und Fleischprodukte bei den Kunden besonders beliebt sind, es jedoch auch eine signifikante Diversität in den Kaufgewohnheiten gibt.

	NumDealsPurchases	NumWebPurchases	NumCatalogPurchases	NumStorePurchases	NumWebVisitsMonth	Z_CostContact	Z_Revenue
count	2240	2240	2240	2240	2240	2240	2240
mean	2.325	4.085	2.662	5.790	5.317	3	11
std	1.932	2.779	2.923	3.251	2.427	0	0
min	0	0	0	0	0	3	11
25%	1	2	0	3	3	3	11
50%	2	4	2	5	6	3	11
75%	3	6	4	8	7	3	11
max	15	27	28	13	20	3	11

Tabelle 4: Statistische Zusammenfassung der Kaufgewohnheiten über verschiedene Kanäle und Online-Interaktionen der Kunden

Die Tabelle zeigt verschiedene Kaufkanäle und Kundenerfahrungen auf. Auffällig ist die hohe Anzahl von Einkäufen in physischen Geschäften, durchschnittlich 5.790 pro Kunde, was die Wichtigkeit von Ladengeschäften hervorhebt. Online-Interaktionen sind ebenfalls signifikant, mit durchschnittlich 5.317 Webseitenbesuchen pro Monat pro Kunde. Dies deutet auf ein starkes Online-Engagement hin, obwohl nicht alle Besuche zu Käufen führen.

Die konstanten Werte von „Z_CostContact“ und „Z_Revenue“, jeweils 3 und 11, tragen nicht zur Vorhersagekraft des Modells bei und werden daher in der Modellierung ausgeschlossen. Außerdem zeigen die Daten, dass manche Kunden gewisse Kanäle nicht nutzen, was durch den Mindestwert von 0 in einigen Kategorien sichtbar wird. Dies deutet auf unterschiedliche Präferenzen und Verhaltensweisen in verschiedenen Kundensegmenten hin.

Die Analyse kategorischer Merkmale enthüllt wichtige Kundencharakteristika. Jeder der 2240 Kunden hat eine einzigartige ID. Ein Großteil der Kunden, 1127, hat einen Hochschulabschluss („Graduation“), und die Mehrheit ist verheiratet. Auffällig ist die geringe Akzeptanzrate der Marketingkampagnen „AcceptedCmp1“ bis „AcceptedCmp5“ sowie „Response“, da die meisten Kunden diese abgelehnt haben. Zudem haben fast alle Kunden, 2219 von 2240, keine Beschwerden geäußert, was auf hohe Kundenzufriedenheit oder geringe Reklamationsneigung hindeutet. Diese Erkenntnisse beleuchten die generellen Kundeneigenschaften und zeigen Unterschiede in der

Kampagnenakzeptanz, was wertvolle Ansätze für Zielgerichtete Marketingstrategien bietet.

	count	unique	top	freq
ID	2240	2240	0	1
Year_Birth	2240	59	1976	89
Education	2240	5	Graduation	1127
Marital_Status	2240	8	Married	864
AcceptedCmp1	2240	2	0	2096
AcceptedCmp2	2240	2	0	2210
AcceptedCmp3	2240	2	0	2077
AcceptedCmp4	2240	2	0	2073
AcceptedCmp5	2240	2	0	2077
Complain	2240	2	0	2219
Response	2240	2	0	1906

Tabelle 5: Statistische Zusammenfassung kategorischer Merkmale

2.5 Erkundung und Visualisierung der Datenstrukturen: Explorative Datenanalyse

Die explorative Datenanalyse (EDA) ist ein unverzichtbarer Schritt im Datenanalyseprozess. Sie dient dazu, Muster, Beziehungen und Auffälligkeiten in den Daten visuell zu erkennen, bevor komplexere statistische Methoden angewendet werden. Diese eingehende Erkundung ermöglicht es, ein grundlegendes Verständnis der Daten, ihrer Verteilungen und Zusammenhänge zu erlangen und Hypothesen für weiterführende Analysen zu formulieren [vgl. SSPP2019, S. 4727]. In dieser Arbeit werden sowohl univariate als auch multivariate Analysen durchgeführt, wobei die Datenstruktur mit der Python-Bibliothek „Pandas“ organisiert werden. Für die Visualisierung der Analysen werden „Matplotlib“ und „Seaborn“ eingesetzt. Diese Tools ermöglichen effiziente und umfassende Datenexplorationen und sind essenziell für das effektive Erkennen von Trends, Mustern und Anomalien.

2.5.1 Ausreißererkennung

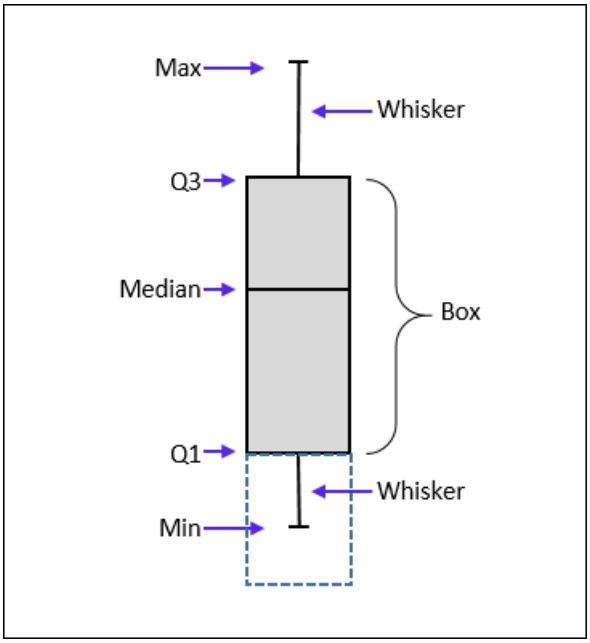

Abbildung 1: Struktur und Bestandteile eines Boxplots.

Quelle : https://biostats.w.uib.no/9-how-to-draw-a-boxplot

Ein Boxplot, auch als Whisker-Plot bezeichnet, bietet eine grafische Darstellung zur Beschreibung der Verteilung von Datenwerten. Das zentrale Rechteck im Boxplot repräsentiert die mittleren 50% der Daten, wobei die obere und untere Grenze den dritten (Q3) bzw. ersten Quartil (Q1) der Daten darstellen. Innerhalb dieses Rechtecks zeigt eine horizontale Linie den Median der Daten, der die Daten in eine obere und eine untere Hälfte teilt. Die „Whisker“ des Plots erstrecken sich von Q1 und Q3 bis zum letzten Datenpunkt innerhalb eines Abstandes von 1,5-mal dem Interquartilbereich (IQR). Datenpunkte, die außerhalb dieser Whisker-Grenzen liegen, werden als Ausreißer bezeichnet und sind ungewöhnlich hohe oder niedrige Werte im Vergleich zum Rest des Datensatzes. Somit ermöglicht der Boxplot nicht nur eine Einschätzung der Datenverteilung, sondern hilft auch, potenzielle Ausreißer schnell zu identifizieren [vgl. Nuzz2016, S. 269].

Aus den erstellten Boxplots für verschiedene Merkmale lässt sich erkennen, dass in den Spalten „Year_Birth“, „Income“, „MntWines“, „MntFruits“ und vielen weiteren Ausreißer vorhanden sind. Abbildung 2 illustriert beispielhaft die Merkmale „Year_Birth“ und „Income“. Die Daten für das Geburtsjahr in der Abbildung zeigen mehrere Ausreißer. Die Mehrheit der Kunden wurde zwischen 1960 und Ende der 1970er Jahre geboren, mit einem Median um das Jahr 1970. Drei Kunden, die vor dem Jahr 1900 geboren wurden, stechen besonders hervor. Bei der Darstellung des Einkommens sind ebenfalls Ausreißer erkennbar. Das Einkommen der meisten Kunden bewegt sich zwischen 35.000 und 70.000, wobei der Median bei etwa 50.000 liegt. Es gibt jedoch eine kleine Gruppe von Einkommenswerten nahe 160.000 und einen auffälligen Ausreißer mit einem Einkommen von über 600.000.

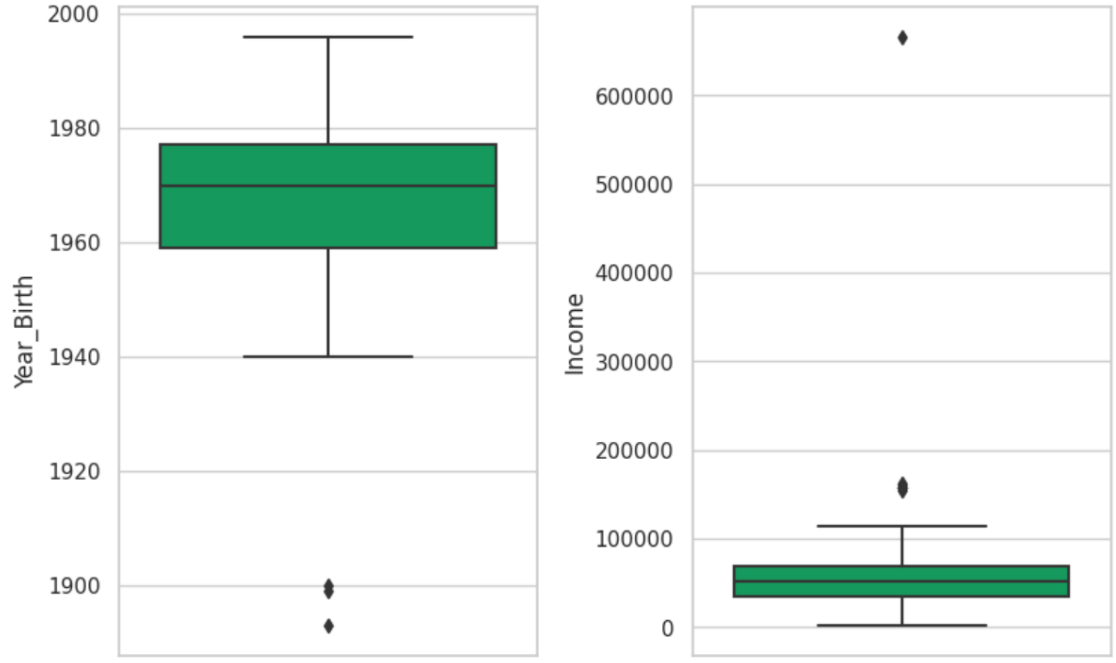

Abbildung 2: Boxplots der Merkmale „Year_Birth“ und „Income“

Aufgrund der in den Daten festgestellten Ausreißer sind geeignete Korrekturmaßnahmen erforderlich.

2.5.2 Verteilungsanalyse

Histogramme sind grafische Darstellungen, die die Verteilung eines Datensatzes visualisieren. Sie teilen die Daten in eine Reihe von Bins oder Intervallen auf und zeigen die Anzahl der Datenpunkte in jedem Bin an. Durch diese Art der Darstellung wird es möglich, die zugrunde liegende Häufigkeitsverteilung der Daten zu erfassen, wie zum Beispiel deren Zentralität, Streuung und das Vorhandensein von Modi. Innerhalb dieses Kontextes geben Histogramme wichtige Einsichten in die Verteilung einzelner Merkmale. Sie können beispielsweise aufzeigen, ob die Datenverteilung symmetrisch um den Mittelwert ist, was auf eine Normalverteilung hindeuten würde, oder ob es eine Rechtsschiefheit (viele Datenpunkte unter dem Mittelwert mit höheren Ausreißern) oder eine Linksschiefheit (viele Datenpunkte über dem Mittelwert mit niedrigeren Ausreißern) gibt. Diese Erkenntnisse über die Form der Verteilung können für spätere Analysen und Modellierungsansätze entscheidend sein.

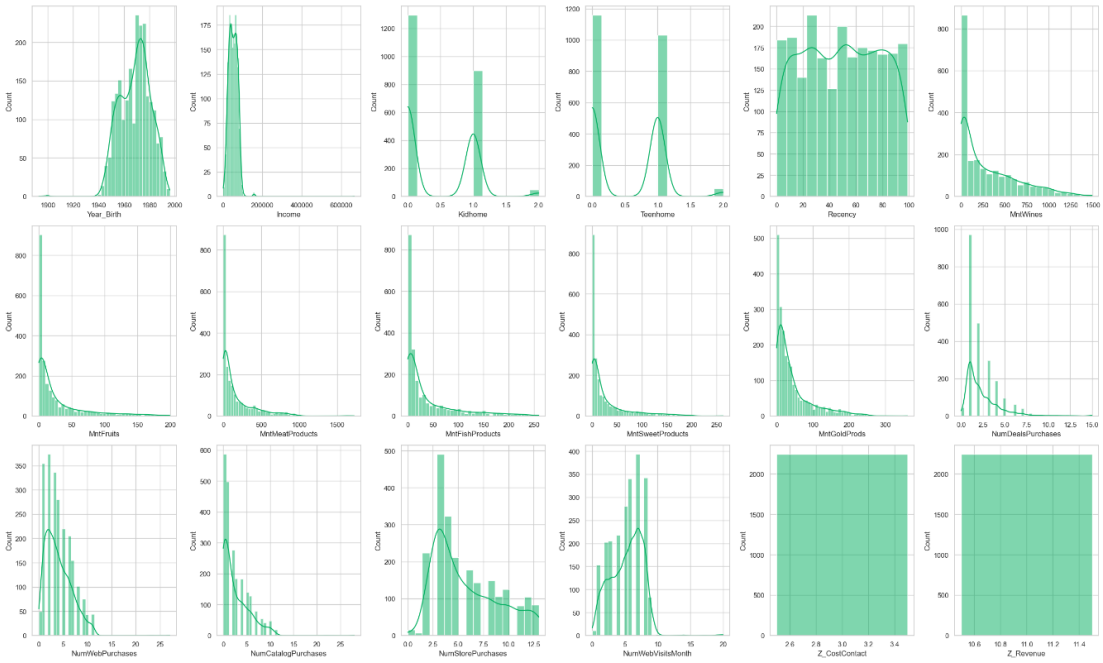

Abbildung 3: Histogramm-Übersicht der Kundenmerkmale

In der Verteilungsanalyse dieses Datensatzes ergeben sich verschiedene Verteilungsmuster, die für die anschließende Datenverarbeitung und Analyse relevant sind. Einige Variablen, darunter „Recency“, „Year_Birth“ und „NumWebVisitsMonth“, nähern sich einer normalen Verteilung an, was auf eine symmetrische Verteilung der Datenwerte um den Mittelwert hindeutet. Andere, wie „Z_CostContact“ und „Z_Revenue“, zeigen eine gleichförmige Verteilung mit lediglich einem einzigen Wert und verweisen somit auf eine konstante Eigenschaft.

Mehrere Variablen weisen eine positiv schiefe Verteilung auf, darunter Einkommen und Ausgaben für verschiedene Produkte sowie verschiedene Kaufarten. Dies bedeutet, dass ein Großteil der Daten unterhalb des Durchschnitts liegt, während höhere Werte seltener vorkommen. Zusätzlich zeigen die Merkmale „Kidhome“ und „Teenhome“ eine bimodale Verteilung, was auf das Vorhandensein zweier dominanter Werte in diesen Daten hinweist.

Aus diesen Erkenntnissen ergibt sich die Notwendigkeit, Daten mit positiver Schiefe einer logarithmischen Transformation zu unterziehen, um eine normalere Verteilung zu erzielen. Diese Anpassung verbessert die Genauigkeit bei der Anwendung statistischer Modelle und stellt sicher, dass die Analyseergebnisse verlässlich sind.

2.5.3 Analyse der kategorialen Variablen

Die Analyse kategorialer Merkmale mittels Countplots dient der visuellen Darstellung der jeweiligen Häufigkeiten der Kategorien. Abbildung 4 visualisiert ausgewählte kategoriale Merkmale des Datensatzes.

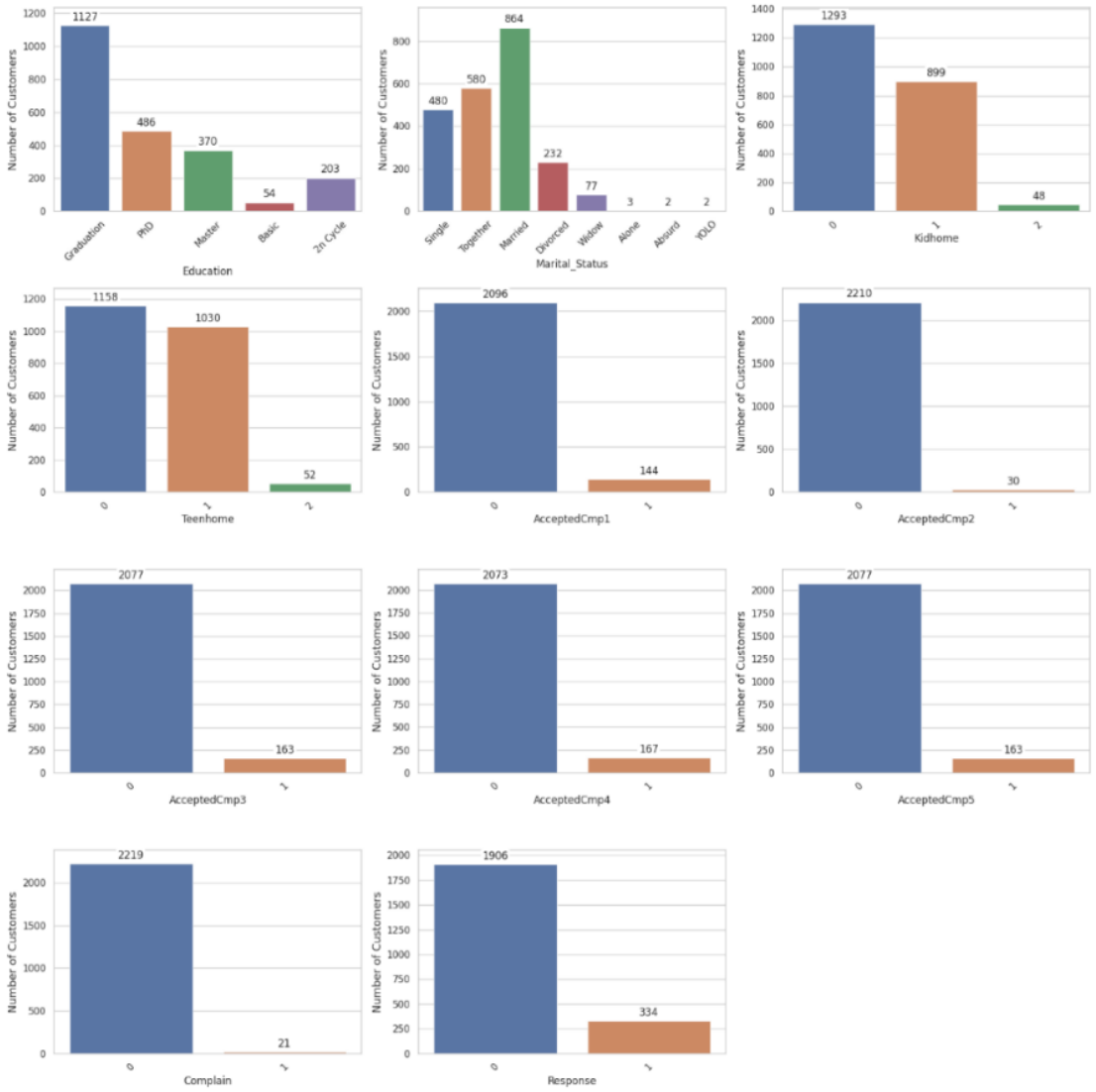

Abbildung 4: Darstellung kategorialer Merkmalsverteilungen durch Countplots

Bei genauerer Betrachtung der Spalten „Education“ und „Marital_Status“ werden bestimmte Unklarheiten und Überlappungen sichtbar. Beispielsweise scheinen die Begriffe „2n Cycle“ und „Master“ in der Kategorie „Education“ ebenso synonym zu sein wie „Single“ und „Alone“ in der „Marital_Status“ Kategorie. Des Weiteren zeichnet sich in der „Education“-Kategorie ein signifikanter Anteil an Kunden ab, die unter „Graduation“ fallen, während in der „Marital_Status“-Kategorie die meisten Kunden als verheiratet gelten. Betrachtet man die Spalten „Kidhome“ und „Teenhome“, wird ersichtlich, dass

eine Mehrheit der Kunden weder Kinder noch Teenager im Haushalt hat. Ebenso offenbaren die Spalten von „AcceptedCmp1“ bis „AcceptedCmp5“, ergänzt durch „Complain“ und „Response“, eine dominierende Präsenz von Werten, die auf das Fehlen einer Antwort oder Beschwerde hindeuten. Dies verweist auf eine signifikante Unausgewogenheit in der Zielvariablen „Response“.

Angesichts dieser Beobachtungen sollten während der Datenaufbereitungsphase spezifische Maßnahmen ergriffen werden, um Redundanzen zu beseitigen, die durch überlappende und synonyme Kategorien entstanden sind.

2.5.4 Korrelationen zwischen Merkmalen

In der Datenanalysephase ist das Erkennen und Verstehen der Beziehungen zwischen verschiedenen Variablen eines Datensatzes von zentraler Bedeutung. Hierfür bietet sich die Heatmap als effiziente Methode zur Visualisierung dieser Beziehungen an. Sie stellt Werte innerhalb einer Matrix durch unterschiedliche Farbtöne dar und liefert so eine übersichtliche Interpretation der Datenstruktur.

Für diesen Datensatz zeigt die Heatmap Korrelationen zwischen ausgewählten Merkmalen. Die „heatmap“-Funktion der „Seaborn“-Bibliothek wurde verwendet, um die Heatmap zu erstellen. In dieser Darstellung symbolisieren warme Farbtöne eine positive Korrelation und kühle Farbtöne hingegen eine negative Korrelation. Ergänzend werden genaue Korrelationswerte für jedes Merkmalpaar angegeben. Hierbei impliziert ein Korrelationskoeffizient nahe +1 eine starke positive Beziehung, während ein Wert nahe -1 eine negative Beziehung anzeigt. Ein Wert um 0 weist auf eine geringe oder fehlende Korrelation hin. Die Intensität der Farben in der Heatmap reflektiert die Stärke der Korrelationen, wodurch schnell erkannt werden kann, welche Merkmale in starker Beziehung zueinanderstehen. Solche Informationen sind entscheidend, um die Datenstruktur zu verstehen und die nächsten Schritte in der Datenverarbeitung und Modellierung zu leiten.

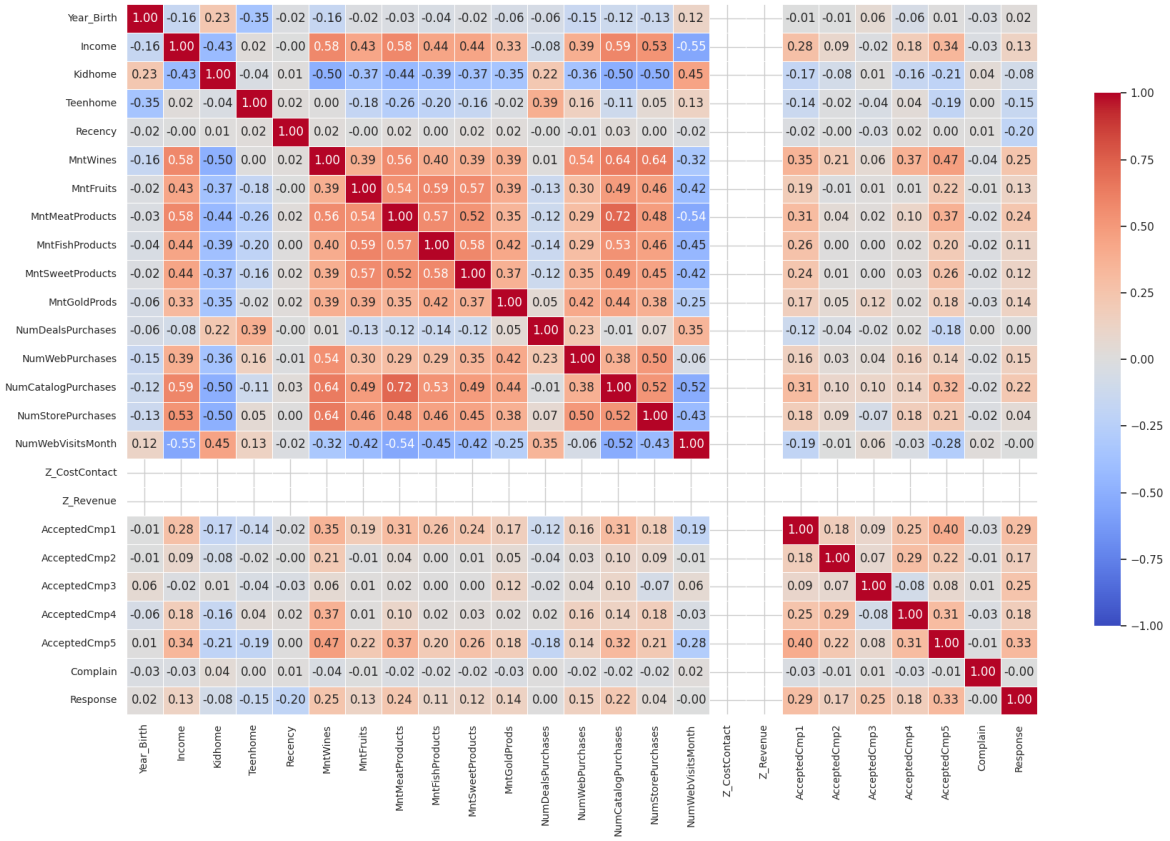

Abbildung 5: Visuelle Darstellung der Korrelationsbeziehungen mittels Heatmap

Die Heatmap in Abbildung 5 zeigt deutlich, welche Merkmale am stärksten mit der Zielvariablen korrelieren. Dies ist von besonderer Bedeutung, um zu verstehen, welche Faktoren die Kundenreaktion am meisten beeinflussen. Interessanterweise weist die Spalte „Response“ die stärkste Korrelation mit „AcceptedCmp5“ auf, was auf Ähnlichkeiten zwischen den beiden Kampagnen hindeutet. Dies könnte bedeuten, dass die Strategien oder Angebote in diesen Kampagnen ähnliche Präferenzen bei den Kunden hervorrufen. Weiterhin zeigen die Daten, dass Kunden besonders an Weinen „MntWines“ und Fleischprodukten „MntMeatProducts“ interessiert sind. Dies spiegelt sich in den positiven Korrelationswerten wider. Die bevorzugte Einkaufsmethode ist der Katalogkauf „NumCatalogPurchases“, was darauf hindeutet, dass Katalogkäufe für diese Kategorie von Produkten effektiver sein könnten.

Abseits der direkten Korrelation zur Zielvariablen gibt es auch interessante Muster bei den Korrelationen zwischen den Merkmalen selbst. Ein höheres Einkommen korreliert mit einer Zunahme von Käufen, vor allem bei Wein- und Fleischprodukten. Zudem besteht eine positive Korrelation zwischen höherem Einkommen und Käufen über verschiedene Vertriebskanäle. Im Gegensatz

dazu steht eine negative Korrelation zwischen der Anzahl der Webseitenbesuche und dem Einkommen, was darauf hindeutet, dass Kunden mit höherem Einkommen weniger dazu neigen, Webseiten ohne Kaufabsicht zu besuchen.

Des Weiteren neigen Kunden, die Wein erwerben, auch zum Kauf von Fleischprodukten, was auf eine mögliche Verbindung im Kaufverhalten in diesen Produktkategorien hinweist.

Bei den Angebotseinkäufen „NumDealsPurchases“ zeigt sich, dass mehr Rabatte zu häufigeren Webseitenbesuchen führen. Allerdings korrelieren diese häufigen Besuche negativ mit Käufen über Kataloge oder in physischen Geschäften, was auf eine Präferenz der Kunden für Online-Angebote.

Das Verständnis dieser Zwischenmerkmalskorrelationen ist entscheidend, um die Dynamik und Vorlieben der Kunden zu verstehen. Es kann auch dazu beitragen, gezielte Marketingstrategien zu entwickeln, die auf den erkannten Mustern und Beziehungen basieren.

2.5.5 Kategoriale Einflussfaktoren auf Kundenreaktionen

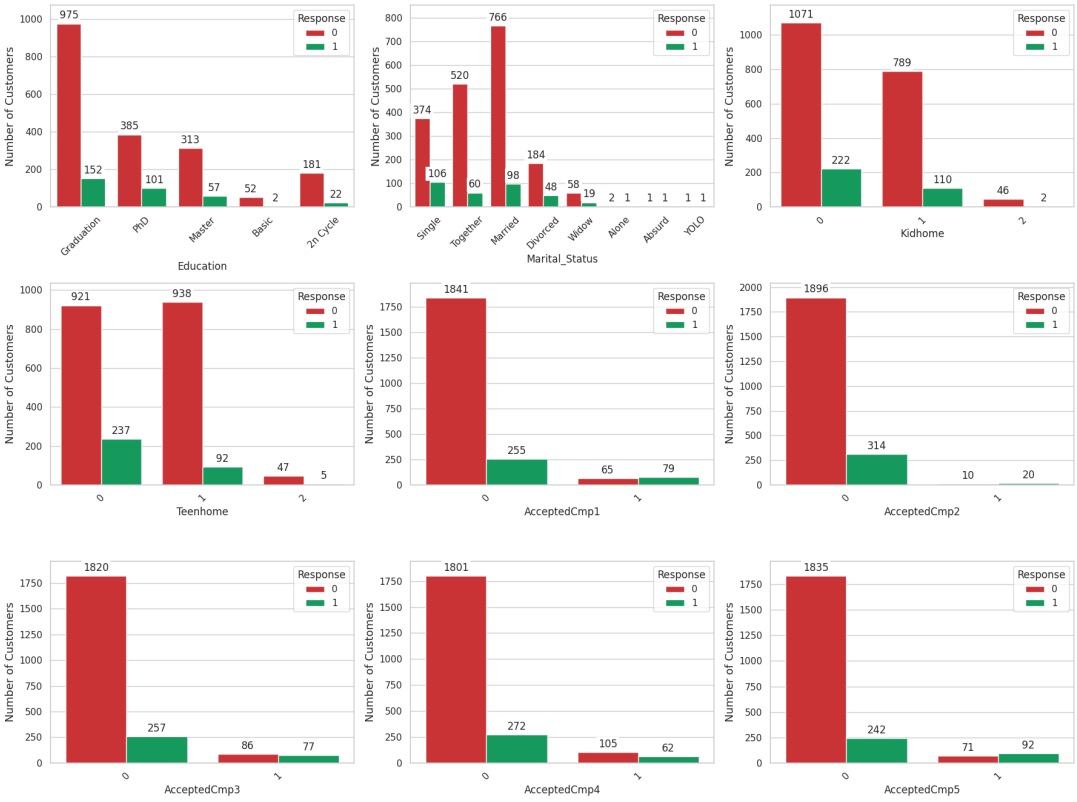

Abbildung 6: Kundenreaktionen auf neueste Kampagne nach demografischen und Verhaltenskategorien

Die Reaktion von Kunden auf Marketingkampagnen kann stark variieren, je nachdem, welche persönlichen oder demographischen Merkmale sie aufweisen. Das Verständnis dieser Unterschiede ist für die Optimierung von Marketingstrategien entscheidend. Im Folgenden wird, basierend auf den Grafiken in Abbildung 6, die Resonanz von Kunden auf die neueste Marketingkampagne in Bezug auf verschiedene Kategorien analysiert und interpretiert.

Bei der Betrachtung der Bildungsabschlüsse fällt auf, dass Kunden mit einem höheren Bildungsniveau, insbesondere mit einem „PhD“ oder „Master“, eine stärkere Tendenz zur Reaktion aufzeigen. Dies deutet darauf hin, dass die Kampagneninhalte möglicherweise gezielt auf dieses Bildungsniveau ausgerichtet sind. Es ist bemerkenswert, dass Kunden, die einen Abschluss der Kategorie „Graduation“ besitzen, trotz ihrer quantitativen Bedeutung eine vergleichsweise geringe Reaktionsrate aufweisen.

Der Familienstand beeinflusst ebenfalls das Antwortverhalten. „Single“- und „Divorced“-Personen reagieren häufiger auf die Kampagnen als Verheiratete oder in einer Partnerschaft lebende Personen. Die Präsenz von Kindern oder Teenagern im Haushalt wirkt sich tendenziell negativ auf die Antwortquote aus.

Die Reaktionsmuster auf vorherige Kampagnen unterstreichen, dass Kunden, die bereits auf frühere Aktionen reagierten, auch eine höhere Wahrscheinlichkeit haben, auf aktuelle Kampagnen zu antworten. Diese Beobachtung deutet auf eine spezifische Kundensegmentierung hin, die besonders empfänglich für Marketingmaßnahmen ist.

Zusammenfassend zeigt sich, dass die Resonanz auf Marketingmaßnahmen stark von den individuellen Merkmalen der Kunden abhängt. Es wird deutlich, wie essenziell es ist, die Zielgruppe genau zu kennen und Kampagnen entsprechend auszurichten.

In diesem Kapitel wurde ein tiefgreifendes Verständnis des bereitgestellten Datensatzes erlangt. Durch detaillierte Analysen, unterstützt durch Grafiken und Tabellen, wurden wesentliche Charakteristika und Muster hervorgehoben. Dies legt den Grundstein für die nächste Phase: die Daten-Aufbereitung im nächsten Kapitel. Dort wird der Fokus auf der Optimierung und Vorbereitung

der Daten für nachfolgende Modellierungsansätze liegen, basierend auf den Erkenntnissen aus dieser Untersuchung.

3 Datenaufbereitung: Methoden und Techniken zur Modellierungsvorbereitung

Die Gewährleistung der Datenintegrität und -qualität ist von zentraler Bedeutung für die Verlässlichkeit von Vorhersagemodellen, die auf künstlicher Intelligenz basieren. Die effiziente Aufbereitung dieser Daten bildet den Schlüssel zur Optimierung von Marketingkampagnen.

Im vorangegangenen Kapitel wurden wichtige Qualitäts- und Strukturaspekte des Datensatzes untersucht. Es wurde ersichtlich, dass neben dem Vorhandensein fehlender Werte und Ausreißern auch eine Reihe von Merkmalen eine positiv schiefe Verteilung aufweisen. Solche Inkonsistenzen und Anomalien können die Leistungsfähigkeit und Genauigkeit des Vorhersagemodells beeinträchtigen und bedürfen daher einer sofortigen Korrektur.

Das aktuelle Kapitel widmet sich der methodischen Bewältigung dieser Herausforderungen. Dies umfasst Techniken des Feature Engineerings und der Feature Extraktion, die Transformation numerischer Merkmale, die Kodierung kategorialer Merkmale, die gezielte Auswahl relevanter Features sowie Strategien im Umgang mit unausgewogenen Datenverteilungen. Ziel ist es, den Datensatz umfassend für die Modellierung von Marketingstrategien aufzubereiten und mögliche Verzerrungen in den Ergebnissen des Modells zu minimieren.

3.1 Behandlung ungültiger Werte

Die adäquate Behandlung und präzise Repräsentation von Daten sind essenziell für die Effektivität und Genauigkeit datengetriebener Modelle. Ein wichtiger Aspekt dabei ist der Umgang mit ungültigen oder inkonsistent repräsentierten Daten.

Eine erste Herausforderung stellt die Spalte „DtCustomer“ dar, welche das Beitrittsdatum der Kunden angibt. Die ursprüngliche Form dieses Datums liegt nicht im optimalen Format vor und wird daher in das standardisierte „pandas

datetime“-Format umgewandelt. Diese Umwandlung ermöglicht eine effizientere Weiterverarbeitung und präzisere Analyse.

Ein weiterer Optimierungsschritt betrifft die Spalte „Marital_Status“. Bei genauer Betrachtung dieser Daten wird deutlich, dass bestimmte Kategorien semantisch ähnliche Bedeutungen haben. Daher werden die Kategorien „Widow“, „Alone“, „Absurd“ und „YOLO“ zu einer allgemeineren Kategorie „Single“ zusammengefasst. Ebenso wird die Kategorie „Together“ als „Married“ reklassifiziert, um eine klarere Struktur zu gewährleisten.

Ein ähnlicher Konsolidierungsprozess wird für die „Education“-Spalte durchgeführt. Die Kategorien „2n Cycle“ und „Master“ werden als inhaltlich verwandt erkannt. Daher wird beschlossen, die „2n Cycle“-Einträge zu entfernen und durch die „Master“-Kategorie zu ersetzen.

Durch diese gezielten Anpassungen können Inkonsistenzen im Datensatz effektiv behoben und eine kohärente, homogenere Datenstruktur erzielt werden.

3.2 Datenaufteilung

Die Datenverarbeitung transformiert Rohdaten in ein für die Modellierung geeignetes Format. Ein kritischer Punkt in diesem Prozess ist die Verhinderung von Datenleckagen. Datenleckagen treten auf, wenn Informationen aus dem Testdatensatz in den zur Modellbildung verwendeten Trainingsdatensatz einfließen. Ein solches Eindringen von Informationen kann zu verzerrten oder falschen Leistungsschätzungen führen, wenn Vorhersagen für neue Daten getroffen werden [vgl. Brow2020, S. 5].

Deshalb sollte die Datenaufbereitung ausschließlich auf dem Trainingsdatensatz basieren. Dies impliziert, dass jegliche Parameter, Koeffizienten oder im Rahmen der Datenaufbereitung generierte Modelle lediglich auf Grundlage des Trainingsdatensatzes entwickelt werden sollten. Nach dieser Anpassung kann die Aufbereitungsmethode dann auf den Trainings- sowie den Testdatensatz angewendet werden.

Die Aufteilung des Datensatzes in ein Trainings- und ein Testset erfolgt im Verhältnis von 75 zu 25 Prozent. Diese Aufteilung zielt darauf ab, eine solide

Basis für die nachfolgende Modellbewertung zu schaffen und zu gewährleisten, dass die entwickelten Modelle auf unbekannten Daten angemessen generalisieren können.

3.3 Behandlung fehlender Datenwerte

Fehlende Werte in Daten sind eine häufige Herausforderung und können aus nicht erfassten Beobachtungen oder Datenverlust resultieren. Viele maschinelle Lernalgorithmen unterstützen keine Datensätze mit fehlenden Werten, was eine gezielte Behandlung dieser Leerstellen notwendig macht [vgl. ebd, S. 66]. In der vorliegenden Datenanalyse wurde identifiziert, dass die Spalte „Income“ 24 fehlende Einträge aufweist. Davon befinden sich 20 im Trainingsdatensatz und 4 im Testdatensatz, was 1,07% der Gesamtdaten entspricht.

Grundsätzlich gibt es zwei Ansätze zum Umgang mit fehlenden Daten: die Löschung und die Imputation. Die Löschung von Datenpunkten mit fehlenden Werten kann jedoch zum Verlust wertvoller Informationen führen, besonders wenn andere Merkmale dieser Datenpunkte relevant sind. Imputation ersetzt fehlende Daten durch berechnete Werte und lässt sich in Single-Imputation und Multiple-Imputation unterteilen. Während die Single-Imputation einfach und schnell ist, kann sie ungenaue Schätzungen liefern, da nur ein Wert für jede fehlende Beobachtung eingesetzt wird. Für komplexere Datensätze ist oft Multiple Imputation vorzuziehen [vgl. Patr2002, S. 78 f.].

Ein etabliertes Verfahren hierfür ist die „Multiple Imputation by Chained Equations“ (MICE). MICE erstellt mehrere vervollständigte Datensätze und kombiniert diese anschließend, um Schätzungen und Standardfehler zu erhalten. MICE nutzt vorhandene Daten aus anderen Merkmalen als Prädiktoren und führt mehrere Iterationen von Modelltrainings durch, um die fehlenden Werte zu schätzen. Multiple-Imputation-Methoden wie MICE berücksichtigen die Wechselwirkungen zwischen verschiedenen Merkmalen eines Datensatzes, was besonders bei der „Income“-Spalte von Bedeutung ist, die starke Korrelationen mit anderen Merkmalen zeigt.

Für die „Income“-Spalte wird die Multiple-Imputation-Methode mittels MICE angewendet. Hierbei kommt das „miceforest“-Paket zum Einsatz, das Random

Forests als Vorhersagemodell verwendet. Zunächst wird eine künstliche Amputation im Trainingsdatensatz vorgenommen, indem 25% der Daten zufällig ausgewählt und als fehlend markiert werden. Ein Imputations-"Kernel" basierend auf dem Originaldatensatz und spezifischen Parametern wird erstellt. Der MICE-Algorithmus wird auf diesen Kernel angewandt, um die fehlenden Daten zu ersetzen. Nach Vervollständigung des Trainingsdatensatzes wird der gleiche Kernel verwendet, um fehlende Daten im Testdatensatz zu ersetzen, wodurch eine konsistente Imputation basierend auf den im Trainingsdatensatz gelernten Informationen sichergestellt wird.

3.4 Umgang mit Ausreißern

Ausreißer sind Datenpunkte, die signifikant von den restlichen Daten abweichen und das Gesamtbild eines Datensatzes verzerren können. Sie beeinträchtigen oft die Anpassung und die Vorhersagegenauigkeit von Modellen. Daher ist es von großer Bedeutung, Ausreißer zu identifizieren und angemessen zu behandeln, um die Leistungsfähigkeit von Klassifikations- und Regressionsalgorithmen zu verbessern. Die Entfernung von Ausreißern aus dem Trainingsdatensatz optimiert das Vorhersagemodell und steigert dessen Vorhersageleistung [vgl. DeSa2023, S. 2503].

Um die Übertragbarkeit des Vorhersagemodells auf unbekannte Daten und reale Situationen zu gewährleisten, ist es ratsam, Ausreißer nur aus dem Trainingsdatensatz zu entfernen. Der Testdatensatz sollte seine Ausreißer behalten, um eine realistische Bewertung der Modellrobustheit in verschiedenen Szenarien zu ermöglichen.

Der Interquartilbereich (IQR), dargestellt durch Boxplot-Diagramme, wird genutzt, um Ausreißer im Trainingsdatensatz zu identifizieren. Werte unterhalb oder oberhalb der IQR-Grenzen gelten als Ausreißer.

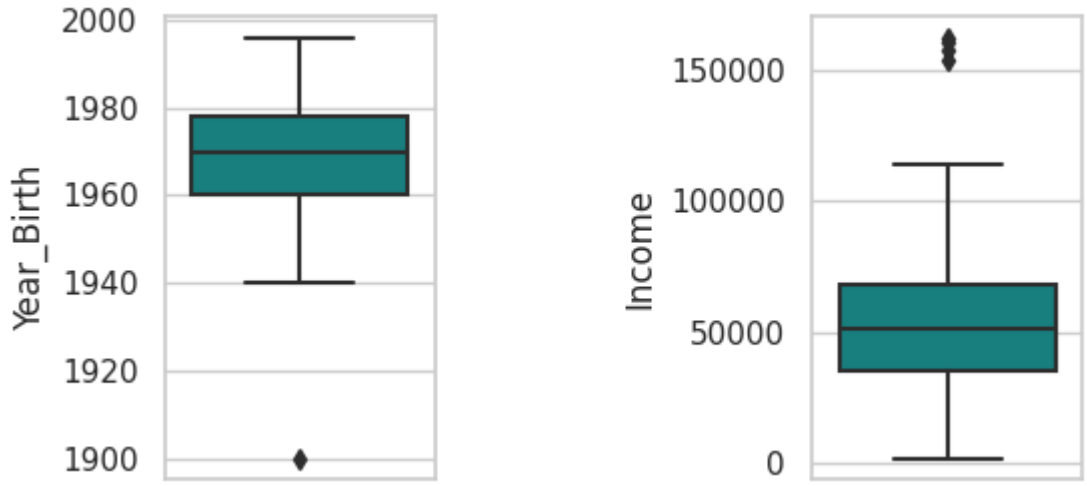

Abbildung 7: Boxplot der Merkmale „Year_Birth“ und „Income“ im Trainingsdatensatz

In Abbildung 7 wird ein Ausreißer sichtbar, der einem Kunden zuzuordnen ist, welcher vor dem Jahr 1900 geboren ist. Dieser Ausreißer wird aus dem Datensatz entfernt. Des Weiteren zeigt die Abbildung eine Gruppe von Einkommenswerten in der Nähe von 160.000. Trotz ihrer Abweichung bleiben diese Werte im Datensatz, da sie nicht als extreme Ausreißer gelten. Die Eliminierung solcher Werte könnte wertvolle Informationen im Lern- und Trainingsprozess der Klassifikatoren vermissen lassen.

In den weiteren Merkmalen des Trainingsdatensatzes werden keine Ausreißer zur Entfernung identifiziert. Die Anzahl der Datenpunkte im Trainingsdatensatz vor der Ausreißerbehandlung beträgt 1680, nach der Behandlung verringert sie sich auf 1679.

3.5 Feature Engineering

Im Kontext des Feature Engineerings werden neue Merkmale entwickelt, um die vorhandene Datenstruktur zu erweitern und bedeutungsvollere Informationen für die Modellierung zu liefern. Diese neu generierten Merkmale basieren auf existierenden Daten und der zugrunde liegenden Geschäftslogik. Nachfolgend werden die speziell erstellten Merkmale und deren Beschreibungen dargelegt.

Merkmal	Beschreibung
Age	Alter des Kunden, basierend auf dem Referenzjahr 2014.
Age_group	Kategorische Aufteilung des Alters in Gruppen: • Young Adult: Bis einschließlich 30 Jahre. • Adult: Zwischen 31 und 45 Jahre. • Senior Adult: Über 45 Jahre.
Has_child	Binäres Merkmal, das angibt, ob der Kunde mindestens ein Kind hat, basierend auf den Daten aus „Kidhome“ und „Teenhome“.
Dependents	Gesamtzahl der Abhängigen eines Kunden, abgeleitet aus der Summe von „Kidhome“ und „Teenhome“.
Lifetime	Monate seit dem ersten Kauf des Kunden.
Spending	Gesamtausgaben eines Kunden für alle Produkte.
Total_Purchases	Gesamtanzahl der Käufe eines Kunden, unabhängig vom Verkaufskanal.
Ever_Accept	Binäres Merkmal, das angibt, ob ein Kunde mindestens eine Kampagne akzeptiert hat.
Total_Cmp	Gesamtzahl der vom Kunden akzeptierten Marketingkampagnen.
Total_revenue	Gesamterlös, der durch die Akzeptanz von Marketingkampagnen generiert wurde. Es wird berechnet, indem die Anzahl der akzeptierten Kampagnen („Total_Cmp“) mit dem durchschnittlichen Erlös („Z_Revenue“) multipliziert wird.
Income_sgmt	Kategorisiert das Einkommen einer Person in Segmente basierend auf Quartilen des gesamten Einkommensdatensatzes: • High: Werte größer oder gleich dem dritten Quartil. • Medium: Werte zwischen dem ersten und dritten Quartil. • Low: Werte kleiner als das erste Quartil.
Conversion_rate_web	Verhältnis von Gesamtkäufen zur Anzahl der Website-Besucher.
Recency_sgmt	Bereichseinteilung der Aktualität des letzten Kaufs eines Kunden. • Score 4: Kunden, deren letzter Kauf vor weniger als oder genau 19 Tagen war. • Score 3: Kunden, deren letzter Kauf zwischen 20 und 39 Tagen zurückliegt. • Score 2: Kunden, deren letzter Kauf zwischen 40 und 59 Tagen zurückliegt. • Score 1: Kunden, deren letzter Kauf zwischen 60 und 79 Tagen zurückliegt. • Score 0: Kunden, die vor mehr als 79 Tagen gekauft haben.

Tabelle 6: Übersicht der neu generierten Merkmale und deren Beschreibungen

Im Rahmen des Feature Engineering werden die Trainings- und Testdatensätze temporär zu einem einzelnen Datenrahmen vereint, um den Prozess der Merkmalsentwicklung zu erleichtern. Diese Herangehensweise wird gewählt, da in diesem Prozess kein Potenzial für Datenleckagen besteht. Nach der vollständigen Entwicklung der Merkmale werden die Datensätze erneut in ihre ursprünglichen Trainings- und Testsegmente aufgeteilt.

3.6 Feature Transformation

In der Vorbereitungsphase für maschinelles Lernen ist die Transformation numerischer Variablen ein kritischer Schritt. Diese Notwendigkeit ergibt sich häufig aus der ausgeprägten Schiefe und der Abweichung von der Normalverteilung bei bestimmten Merkmalen, verursacht durch Ausreißer oder exponentielle Verteilungen. Die Schiefe ist ein Maß für die Asymmetrie einer Datendistribution [vgl. RaRo2021].

Die Analyse zeigt, dass mehrere Spalten des Datensatzes eine positive Schiefe aufweisen. Eine positive Schiefe impliziert, dass sich die Mehrzahl der Datenwerte links vom Mittelwert befindet, während der „Schwanz“ der Verteilung nach rechts verlängert ist.

Zur Korrektur dieser Schiefe und zur Normalisierung der Daten werden häufig Transformationstechniken wie die Log-, Box-Cox- und Yeo-Johnson-Transformation eingesetzt. Die Log-Transformation, die auf jedem Datenpunkt den natürlichen Logarithmus anwendet, ist besonders wirksam bei exponentiellen Datenverteilungen. Jedoch ist sie auf positive Daten beschränkt und kann bei Null- oder Negativwerten problematisch sein.

Eine vielseitigere Option ist die Box-Cox-Transformation, die durch einen variablen Parameter (λ) eine größere Anwendungsbreite ermöglicht. Sie strebt danach, die Schiefe der transformierten Daten zu minimieren, ist aber ebenfalls auf positive Daten beschränkt. Die Yeo-Johnson-Transformation, eine Weiterentwicklung der Box-Cox-Transformation, behebt die Einschränkungen der vorherigen Methoden, indem sie auch für negative Daten geeignet ist. Sie passt den Parameter λ so an, dass er für positive und negative Zahlen gleichermaßen funktioniert, was eine breitere Anwendbarkeit auf verschiedene Datensätze ermöglicht [vgl. OtAl2021, S. 2]. Nach dem Testen dieser

Methoden erwies sich die Yeo-Johnson-Transformation als effektivste Methode, um eine einer Normalverteilung ähnliche Verteilung zu erreichen. Diese Eigenschaft ist insbesondere für baumbasierte Algorithmen von Vorteil.

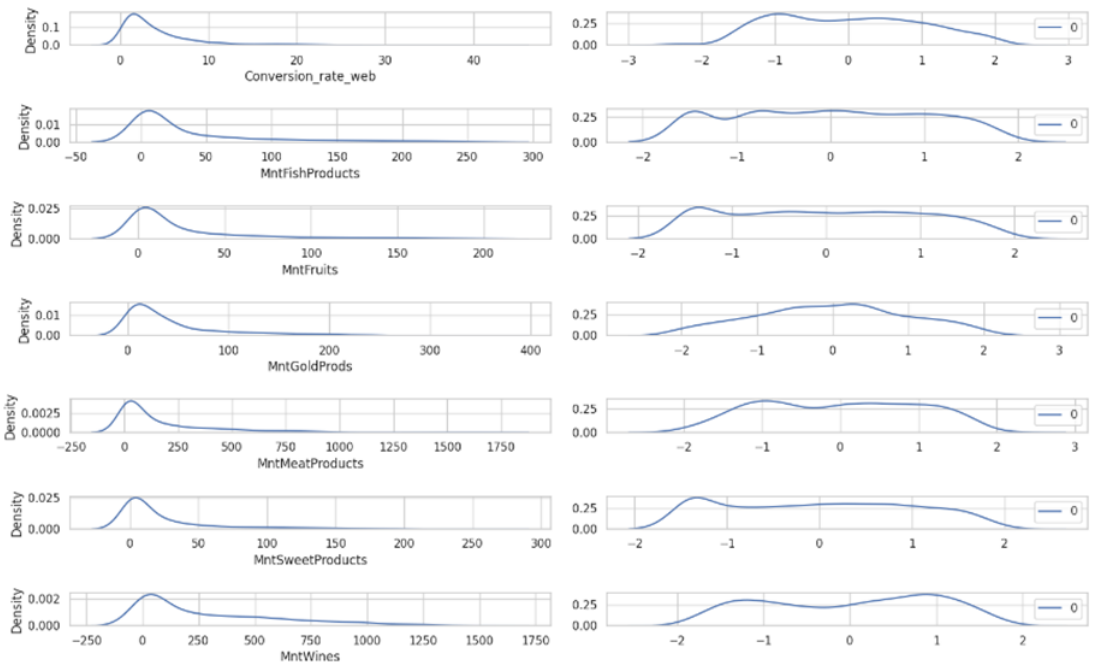

Abbildung 8: Verteilungskurven ausgewählter Merkmale vor und nach der Yeo-Johnson-Transformation

Die Anwendung der Yeo-Johnson-Transformation führte zu einer signifikanten Reduzierung der Schiefe, was auf eine erfolgreiche Normalisierung der Daten hindeutet. Dies verbessert die Qualität des Datensatzes für die nachfolgende Modellierung und trägt zur Genauigkeit und Effizienz der Vorhersagemodelle bei.

3.7 Kodierung kategorischer Merkmale

Kategoriale Daten sind in vielen Datenbanken präsent und erfordern oft eine besondere Behandlung, um in maschinellen Lernmodellen effektiv verwendet zu werden. Der Grund dafür ist, dass viele Algorithmen eine numerische Eingabe erfordern und daher kategoriale Daten, insbesondere solche, die als Text oder String-Typen vorliegen, kodiert werden müssen.

Der vorliegende Datensatz umfasst vier kategorische Merkmale: „Education“, „Marital_Status“, „Age_group“ und „Income_sgmt“. Bei der Kodierung dieser Daten ist zu entscheiden, ob es sich um ordinale oder nominale Kategorien handelt. Ordinale Daten besitzen eine eindeutige und sinnvolle Rangfolge, während nominale Daten keine solche Ordnung aufweisen.

Für das Merkmal „Education“, dass eine Hierarchie in der Bildungsabfolge aufzeigt, wird ebenso wie für „Age_group“ und „Income_sgmt“ das Label Encoding angewandt. Hierbei werden die Kategorien gemäß ihrer Reihenfolge in numerische Werte umgewandelt [vgl. DaJo2021].

Dagegen ist „Marital_Status“ ein nominelles Merkmal, da es keine inhärente Reihenfolge zwischen den Kategorien gibt. Daher wird hier das One-Hot Encoding verwendet. Dieses Verfahren wandelt jede Kategorie in eine neue Spalte um und kennzeichnet sie mit einer „1“ oder „0“, je nachdem, ob die Kategorie für eine bestimmte Beobachtung zutrifft oder nicht.

Die Art und Weise, wie kategoriale Merkmale kodiert werden, kann sich unmittelbar auf die Performance eines Modells auswirken. Während Label Encoding platzsparend ist, kann One-Hot Encoding die Dimensionalität der Daten erheblich erhöhen, was bei bestimmten Algorithmen Schwierigkeiten verursachen kann.

Schlussendlich ist die adäquate Kodierung kategorialer Merkmale ein zentraler Faktor für den Erfolg maschineller Lernmodelle. Dabei sollten die gewählten Methoden immer in Bezug auf die konkreten Daten und den angewendeten Algorithmus evaluiert werden.

3.8 Merkmalsselektion

Merkmalsselektion ist ein Prozess, bei dem aus einem umfangreichen Satz von Merkmalen eine relevante Teilmenge für den Einsatz im maschinellen Lernen ausgewählt wird. Ziel dieser Selektion ist es, die Datenqualität und -relevanz zu verbessern, indem irrelevante oder überflüssige Informationen entfernt werden.

Die Notwendigkeit für eine Merkmalsselektion entsteht aus verschiedenen Gründen: Erstens kann die Reduktion der Anzahl der Merkmale die Trainingszeiten verkürzen und Ressourcen einsparen. Zweitens hilft das Entfernen nicht relevanter Merkmale, die Überanpassung (Overfitting) zu verhindern, was wiederum die Generalisierungsfähigkeit des Modells verbessert. Drittens resultiert eine sorgfältige Merkmalsselektion oft in einfacheren und leichter interpretierbaren Modellen. Dies vereinfacht das Verständnis und die Erklärbarkeit der Modelle [vgl. LC++2017].

3.8.1 Eliminierung irrelevanter Merkmale

Im Rahmen der Merkmalsselektion ist die Identifikation und der Ausschluss irrelevanter Merkmale für die Modellierung von großer Bedeutung. Die Spalte „ID“, die lediglich eine Vielzahl von Kategorien ohne prädiktiven Wert enthält, trägt nicht zur Vorhersagekraft des Modells bei und wird daher entfernt. Ebenso wird die Spalte „Year_Birth“ nicht berücksichtigt, da das Alter der Kunden bereits durch die vorausgegangene Merkmalsextraktion für das Jahr 2014 erfasst wurde. Die Spalte „Dt_Customer“, obwohl sie Daten enthält, hat nur einen marginalen Einfluss auf das Vorhersagemodell, weshalb sie nach der Datenextraktion ausgeschlossen wird. Schließlich bieten die Spalten „Z_CostContact“ und „Z_Revenue“, die jeweils konstante Werte aufweisen, keine differenzierbaren Informationen und werden daher ebenfalls aus dem Datensatz entfernt, um die Effizienz der Modellierung zu optimieren.

3.8.2 Numerische Feature-Selektion

Bei numerischen Eingabedaten, bei denen die Zielvariable kategorial definiert ist, beispielsweise bei prädiktiven Klassifikationsmodellen, gelten die ANOVA F-Test-Statistik und die Mutual Information (MI) als die prominentesten Ansätze [vgl. Brow2020, S. 138].

Die ANOVA F-Test-Merkmalsauswahl ist eine statistische Methode, die dazu dient, die Bedeutung einzelner Merkmale in Bezug auf eine Zielvariable zu bewerten. Hierbei wird besonders die Varianzuntersuchung genutzt. Der ANOVA F-Test ist besonders für Klassifikationsprobleme geeignet, bei denen die Zielvariable kategorisch ist. Das Hauptprinzip des ANOVA F-Tests ist es, die Varianzen zwischen den verschiedenen Kategorien mit den Varianzen innerhalb dieser Kategorien zu vergleichen.

Ein hoher F-Wert zeigt an, dass das Merkmal signifikante Unterschiede zwischen den Kategorien aufweist und somit potenziell wichtig für die Klassifikation ist. Ein niedriger F-Wert hingegen deutet darauf hin, dass die Unterschiede zwischen den Kategorien nicht signifikant sind und das Merkmal möglicherweise für das Modell nicht relevant ist. In einem praktischen Kontext kann der ANOVA F-Test mit Werkzeugen wie der "Scikit-learn"-Bibliothek in Python durchgeführt werden. Nach Durchführung des Tests können Merkmale

mit niedrigen F-Werten, die unter einem bestimmten Schwellenwert liegen, aus dem Datensatz entfernt werden, um die Effizienz und Genauigkeit des Modells zu verbessern [vgl. ebd., S. 141].

Im Kontext der Merkmalsauswahl im maschinellen Lernen wird Mutual Information (MI) dazu verwendet, zu bewerten, wie sehr ein bestimmtes Merkmal die Vorhersage einer Zielvariablen beeinflusst. Ein hoher MI-Wert zwischen einem Merkmal und einer Zielvariablen zeigt an, dass das Wissen über das Merkmal wertvolle Informationen über die Zielvariable liefert. Ein MI-Wert von 0 deutet darauf hin, dass die beiden Variablen unabhängig sind und somit keine Informationen übereinander liefern. Je größer der MI-Wert, desto stärker ist die Abhängigkeit zwischen den beiden Variablen.

Ein wichtiger Vorteil der Mutual Information ist ihre Vielseitigkeit. Sie kann sowohl für kategorische als auch für kontinuierliche Daten verwendet werden, was sie besonders nützlich für verschiedene Arten von maschinellen Lernproblemen macht, sei es Klassifikation oder Regression. In der Praxis kann die Mutual Information mit Hilfe von Softwarebibliotheken wie „Scikit-learn“ in Python berechnet und zur Merkmalsauswahl genutzt werden [vgl. ebd., S. 143].

Die Anwendung von ANOVA F-Tests und Mutual Information (MI) auf den vorliegenden Datensatz hat ergeben, dass bestimmte Merkmale, wie zum Beispiel „Complain“ oder „Divorced“, nicht wesentlich zur Vorhersagegenauigkeit des Modells beitragen. Diese Erkenntnis führt zu der Überlegung, diese Merkmale aus dem Modell zu eliminieren. Jedoch ist es von entscheidender Bedeutung, statistische Methoden im Kontext des spezifischen Anwendungsfalls zu betrachten. Die Relevanz der Merkmale kann stark kontextabhängig sein.

Bei der Entscheidung, Merkmale zu entfernen, sollte stets der spezifische Kontext des Datensatzes und das Ziel des Modells berücksichtigt werden. Die Auswirkungen der Feature-Selektion auf das resultierende Modell müssen ebenfalls in Betracht gezogen werden. Einerseits kann die Reduktion der Merkmale die Effizienz und Performance des Modells verbessern, andererseits besteht das Risiko, dass durch die Entfernung wichtiger Informationen die Generalisierbarkeit und Validität des Modells beeinträchtigt werden könnten.

Zusammenfassend ist die Kombination aus statistischen Methoden und einem tiefgreifenden Verständnis des Datensatzes sowie seines Kontextes für eine effektive Feature-Selektion unabdingbar.

3.8.3 Multikollinearitätsprüfung

Multikollinearität, definiert als das Vorhandensein starker Korrelationen zwischen zwei oder mehr unabhängigen Merkmalen, kann die Interpretierbarkeit und Stabilität statistischer Modelle beeinträchtigen. Um dieses Problem anzugehen, werden zwei zentrale Methoden angewendet: die Erstellung einer Korrelationsmatrix und die Analyse des Varianzinflationsfaktors (VIF) [vgl. Shre2020].

Zunächst wird eine Korrelationsmatrix für die als wesentlich betrachteten Merkmale erstellt. Diese Matrix hilft dabei, Merkmalspaare zu identifizieren, die eine Korrelation von 0,7 oder höher aufweisen. Für jedes dieser Paare wird die Korrelation zum Zielmerkmal ermittelt. Das Merkmal eines Paares, das die geringere Korrelation zum Zielmerkmal aufweist, wird als redundant betrachtet und für die weitere Analyse ausgeschlossen. Hierbei ist zu beachten, dass das Merkmal „Income“ aufgrund seiner analytischen Relevanz beibehalten wird, selbst wenn es mit anderen Merkmalen stark korreliert.

Ergänzend zur Korrelationsmatrix wird der Varianzinflationsfaktor (VIF) herangezogen. Der VIF quantifiziert das Ausmaß der Multikollinearität in einem Regressionsmodell. Ein VIF-Wert über 5 deutet auf eine potenziell problematische Multikollinearität hin, was eine genauere Untersuchung der betroffenen Merkmale erfordert. Diese Methode ermöglicht eine differenzierte Bewertung der Multikollinearität, die über die einfache Korrelationsanalyse hinausgeht.

Es ist jedoch wichtig zu beachten, dass das Entfernen von Merkmalen aufgrund hoher Korrelation zwar zur Reduzierung von Multikollinearität beiträgt, aber auch zu einem Verlust wichtiger Informationen führen kann. Dieser Schritt sollte daher mit Vorsicht und unter Berücksichtigung der potenziellen Auswirkungen auf das Modell erfolgen.

Durch die Kombination dieser Methoden wird die potenzielle Beeinträchtigung durch Multikollinearität reduziert, was zu einem robusteren statistischen Modell

führt. Die sorgfältige Identifikation und der Ausschluss von korrelierten Merkmalen haben den Datensatz auf 17 wesentliche Merkmale reduziert, was das Risiko von Überanpassung (Overfitting) mindert und die Interpretierbarkeit des Modells erhöht. Diese konzentrierte Auswahl an Merkmalen fördert die Effizienz und Präzision des Klassifikationsmodells, was zu einer optimierten Vorhersagegenauigkeit beiträgt.

3.9 Behandlung unausgewogener Daten

Die Darstellung der Zielvariable „Response“ in Abbildung 9 zeigt ein deutliches Ungleichgewicht: 85,1% der Daten weisen keine Antwort („No Response“) auf, während nur 14,9% auf die Marketingaktionen reagieren („Response“). Dieses Ungleichgewicht ist auch im Balkendiagramm ersichtlich, wobei die Ereignisse „No Response“ weitaus häufiger vorkommen als „Response“.

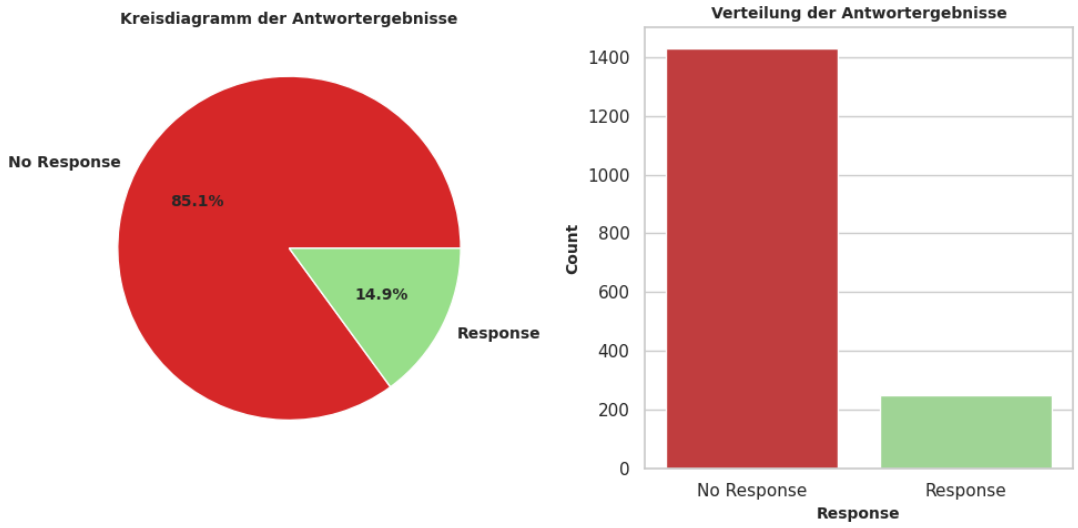

Abbildung 9: Verteilung der Antwortkategorien im Trainingsdatensatz

Ein solches Klassenungleichgewicht kann bei der Modellierung erhebliche Herausforderungen mit sich bringen. Modelle könnten sich beispielsweise dazu neigen, die dominierende Klasse übermäßig zu favorisieren. Dies führt dazu, dass die Modelle zwar eine hohe Genauigkeit aufweisen, jedoch auf Kosten der Erkennung der Minderheitsklasse. Zudem könnte das Modell in neuen, unbekannten Daten Schwierigkeiten haben, die Minderheitsklasse korrekt zu identifizieren, was die Generalisierbarkeit des Modells beeinträchtigt.

In der Datenanalyse gibt es zwei Hauptansätze zum Resampling: Oversampling und Undersampling. Beim Undersampling werden Beobachtungen der dominierenden Klasse reduziert, um ein ausgeglicheneres Verhältnis zu erreichen. Dieser Ansatz kann jedoch zu einem Verlust wichtiger Informationen führen, insbesondere in kleineren Datensätzen [vgl. KiJu2023].

Oversampling hingegen zielt darauf ab, die Anzahl der Beobachtungen in der Minderheitsklasse zu erhöhen. Ein prominentes Verfahren in diesem Bereich ist die „Synthetic Minority Oversampling Technique“ („SMOTE“). Anstatt nur existierende Datenpunkte zu duplizieren, generiert „SMOTE“ synthetische Beobachtungen. Vor der Anwendung von SMOTE zeigt der Training-Datensatz eine deutliche Diskrepanz in der Klassenverteilung. Durch die Anwendung von SMOTE mit einer „sampling_strategy“ von 0,5 wird die Anzahl der Beobachtungen in der Minderheitsklasse auf insgesamt 714 Proben erhöht, was einer Synthese von 463 neuen Proben entspricht.

Es ist essenziell zu betonen, dass „SMOTE“ ausschließlich auf den Training-Datensatz angewendet wird, um die Integrität des Testdatensatzes zu gewährleisten. Dies stellt sicher, dass die Bewertung des Modells authentisch und repräsentativ für den tatsächlichen Datenverlauf bleibt.

4 Modellentwicklung und -evaluation

In diesem Kapitel erfolgt die Fokussierung auf die Entwicklung und Bewertung diverser Klassifikationsmodelle, welche als zentrale Elemente zur Optimierung von Marketingkampagnen mittels künstlicher Intelligenz dienen. Basierend auf der detaillierten Datenaufbereitung und Analyse, die in den vorherigen Kapiteln umrissen wurde, schreitet diese Arbeit zur konkreten Anwendung und Evaluation maschineller Lernverfahren voran. Hierbei wird besonderes Augenmerk auf Modelle wie logistische Regression, Random Forest, XGBoost Classifier, Support Vector Machine und MLP Classifier (neuronales Netzwerk) gelegt. Jedes dieser Modelle unterzieht sich einer eingehenden Prüfung hinsichtlich seiner Fähigkeit, genaue Vorhersagen über Kundenreaktionen auf Marketingaktivitäten zu liefern. Ergänzend dazu wird eine Clustering-Analyse mittels „k-medoids“ auf RFM-Metriken angewendet, um eine zielgerichtete Kundensegmentierung zu erreichen.

Das primäre Ziel dieses Kapitels ist es, ein tiefgreifendes Verständnis für die Effektivität verschiedener Modelle zu erlangen, insbesondere in Bezug auf ihre Fähigkeit, die Reaktionsraten zu steigern und die Rentabilität von Marketingkampagnen zu verbessern.

4.1 Klassifikationsmodelle

Die Auswahl der Klassifikationsmodelle für diese Analyse umfasst ein breites Spektrum an maschinellen Lernmethoden, von grundlegenden bis zu fortgeschrittenen Techniken. Ziel ist es, die Effektivität der Marketingkampagnen durch die Minimierung falsch positiver und falsch negativer Vorhersagen zu steigern. Die Modelle werden anhand ihrer Genauigkeit (Präzision), Recall-Werte und F1-Scores bewertet, um die Antwortrate auf Kampagnen zu erhöhen und gleichzeitig die Marketingkosten zu senken.

Die Modelle sind aufgrund ihrer nachgewiesenen Wirksamkeit in ähnlichen Anwendungsfällen der binären Klassifikation ausgewählt. Jedes Modell verfügt über spezifische Stärken, die entscheidend für die Vorhersage der Kundenreaktionen auf Marketingaktivitäten sind. Beispielsweise bietet die logistische Regression ein robustes Verständnis linearer Zusammenhänge. Random Forest und XGBoost sind aufgrund ihrer Entscheidungsbaumstrukturen besonders effektiv im Umgang mit nichtlinearen Beziehungen und komplexen Interaktionen zwischen Merkmalen. Die Support Vector Machine wird wegen ihrer Effizienz in hochdimensionalen Räumen berücksichtigt, während der MLP Classifier (ein neuronales Netzwerk) aufgrund seiner Flexibilität und Fähigkeit, komplexe Muster in den Daten zu erkennen, verwendet wird.

4.1.1 Random Forest

Der Random Forest Classifier ist ein etabliertes Ensemble-Verfahren im maschinellen Lernen, das die Vorhersagen zahlreicher Entscheidungsbäume kombiniert, um die Genauigkeit und Stabilität des Modells zu verbessern. Dieses Verfahren trainiert jeden Baum auf einer separaten Teilstichprobe des Datensatzes und wählt die Merkmale für die Knotenpunkte zufällig aus, um Vielfalt zu schaffen und Überanpassung zu vermeiden [vgl. BeDr2016, S. 25].

Zur Maximierung der Leistung des Random Forest Classifiers wird eine sorgfältige Abstimmung der Hyperparameter durchgeführt. Hierzu wird das Tool „GridSearchCV“ aus der „sklearn.model_selection“-Bibliothek eingesetzt, welches verschiedene Kombinationen von Hyperparametern testet [vgl. MaPu2022, S. 429]. Dabei werden die Anzahl der Bäume (n_estimators), die maximale Tiefe der Bäume (max_depth), die minimale Anzahl an Stichproben für eine Aufteilung (min_samples_split) und die minimale Anzahl an Stichproben für Blattknoten (min_samples_leaf) berücksichtigt.

Die systematische Hyperparameter-Suche mit „GridSearchCV“ beinhaltet eine fünffache Kreuzvalidierung, um die optimale Kombination zu identifizieren. Nach der Bestimmung dieser Parameter wird das Modell auf den Trainingsdaten neu trainiert. Diese Phase dient der Bewertung der Leistungsfähigkeit des Modells. Daran anschließend erfolgt eine Evaluierung auf den Testdaten. Dieser Schritt ist entscheidend, um die Generalisierungsfähigkeit des Modells auf neue, unbekannte Daten zu überprüfen. Die Analyse der Testergebnisse liefert wichtige Erkenntnisse über die praktische Anwendbarkeit des Modells, ermöglicht ein tieferes Verständnis seiner Stärken und Grenzen und unterstützt die Beurteilung seiner Eignung im Kontext realer Marketingkampagnen.

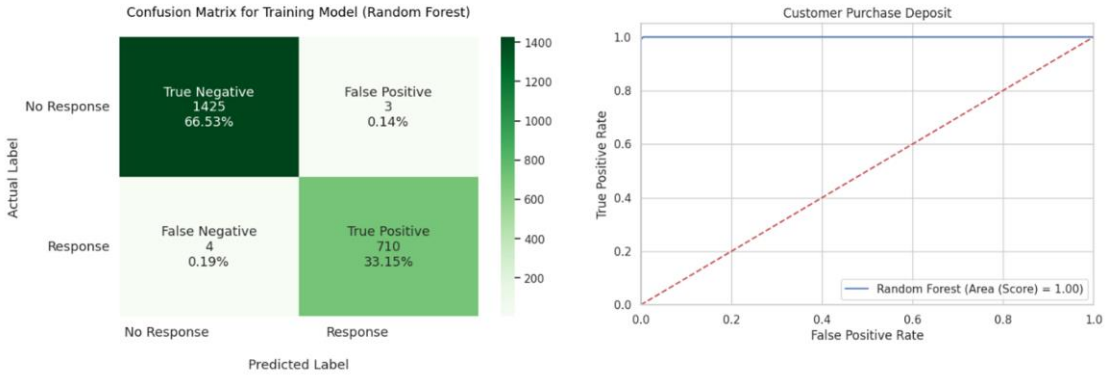

Abbildung 10: Leistung des Random Forest Classifiers auf den Trainingsdaten

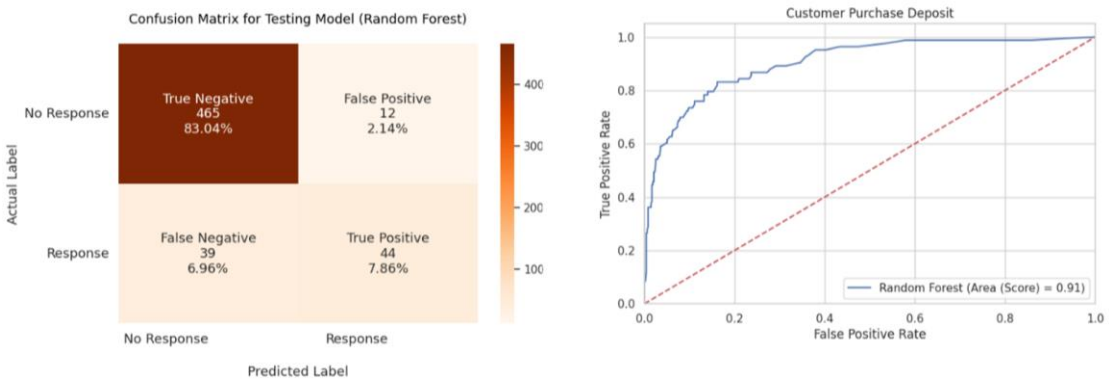

Abbildung 11: Leistung des Random Forest Classifiers auf den Testdaten

Abbildung 10 zeigt eine hohe Leistung des Random Forest Classifiers auf Trainingsdaten, was sich in einer präzisen Klassifikation in der Konfusionsmatrix und einer AUC von 1,00 auf der ROC-Kurve widerspiegelt. In Abbildung 11 wird die Testleistung des Modells dargestellt, mit einer AUC von 0,91 auf der ROC-Kurve. Diese Ergebnisse signalisieren eine starke Vorhersagefähigkeit, allerdings mit einer leichten Verringerung im Vergleich zur Trainingsleistung. Die Konfusionsmatrix im Test zeigt eine erhöhte Anzahl falsch positiver und falsch negativer Ergebnisse, was auf eine robuste, aber realistischere Leistung des Modells bei unbekannten Daten hinweist.

4.1.2 Logistische Regression

Die logistische Regression dient als Modellierungsansatz für die Wahrscheinlichkeit binärer Kundenreaktionen auf Marketingaktionen. Durch die Umwandlung unabhängiger Variablen in Wahrscheinlichkeitswerte ermöglicht sie eine detaillierte Bewertung der Zusammenhänge zwischen Prädiktoren und Zielvariablen [vgl. Stol2011, S.1099].

Für die technische Umsetzung wird „LogisticRegression“ aus „sklearn.linear_model“ mit einer Hyperparametersuche mittels „GridSearchCV“ verwendet. Dieses Verfahren berücksichtigt diverse Regularisierungsnormen (penalty), Regularisierungsstärken (C), Algorithmen (solver) und Iterationszahlen (max_iter) im Rahmen einer fünffachen Kreuzvalidierung, um die beste Parameterkombination zu ermitteln. Das daraufhin optimierte Modell wird zunächst auf den Trainingsdaten angewendet und evaluiert, um seine Leistungsfähigkeit zu beurteilen. Anschließend wird das Modell auch auf den Testdaten eingesetzt, um seine Generalisierbarkeit und Effektivität auf neuen,

unbekannten Daten zu überprüfen. Die Abbildungen 12 und 13 illustrieren die Leistung des logistischen Regressionsmodells sowohl auf dem Trainings- als auch auf dem Testdatensatz.

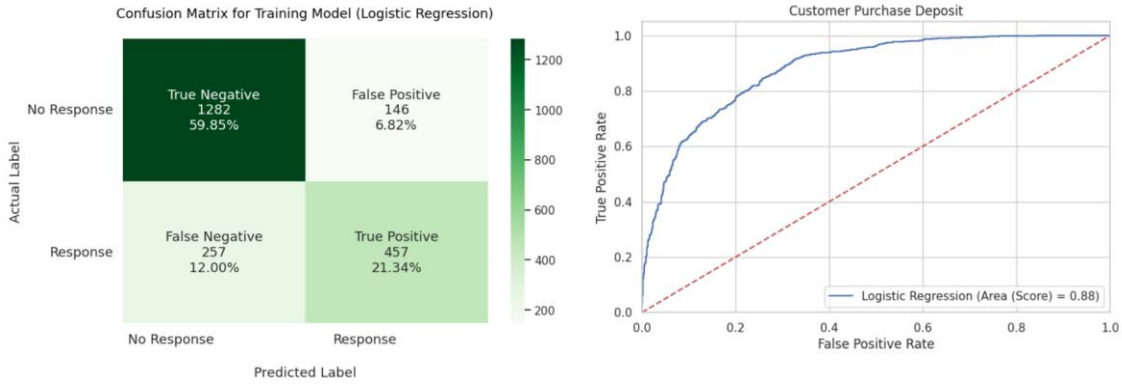

Abbildung 12: Konfusionsmatrix und ROC-Kurve des logistischen Regressionsmodells - Training

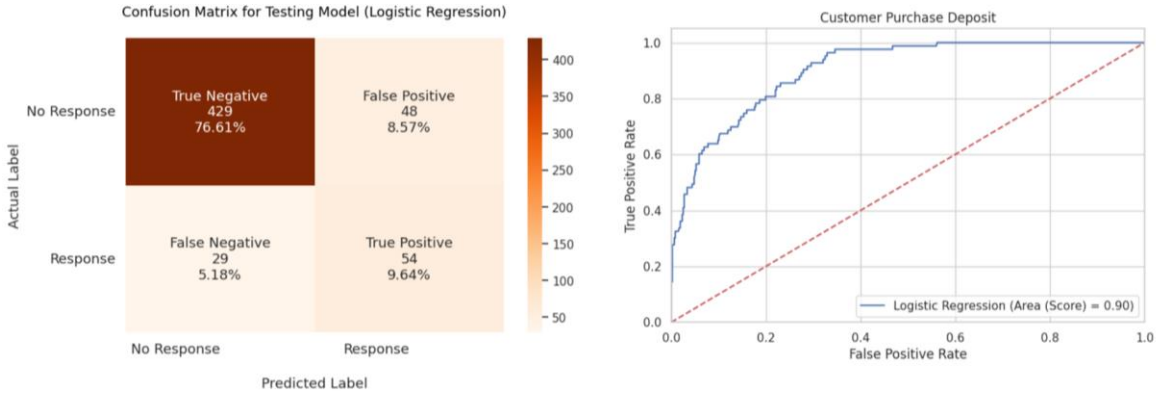

Abbildung 13: Konfusionsmatrix und ROC-Kurve des logistischen Regressionsmodells - Test

4.1.3 XGBoost Classifier

XGBoost, kurz für „Extreme Gradient Boosting“, ist ein fortschrittliches, verteiltes Gradient-Boosting-System, das für seine Effizienz und Leistungsfähigkeit bekannt ist. Dieses System zeichnet sich durch den Einsatz einer Reihe von Entscheidungsbäumen aus, die in einer sequenziellen Weise aufgebaut werden. Jeder nachfolgende Baum in dieser Sequenz zielt darauf ab, die Unzulänglichkeiten seiner Vorgänger zu korrigieren, indem er sich auf deren Fehler konzentriert. XGBoost verwendet den Gradientenabstieg als Kern-Optimierungsverfahren, um die Verlustfunktion zu minimieren, welche die Genauigkeit des Modells auf den Trainingsdaten bewertet. Zusätzlich integriert XGBoost Regularisierungstechniken in die Verlustfunktion, um die

Modellkomplexität zu kontrollieren und Überanpassung zu vermeiden [vgl. ZQ++2018, S. 21024].

Für die Feinabstimmung des XGBoost Classifiers werden Hyperparameter eingestellt, welche die Architektur des Modells und das Lernverhalten steuern. Dazu gehören die Anzahl der zu erstellenden Bäume (n_estimators), die maximale Tiefe dieser Bäume (max_depth), die Lernrate (learning_rate), die bestimmt, wie schnell das Modell lernt, und der Subsampling-Parameter (subsample), der angibt, welcher Anteil der Trainingsdaten für das Training eines Baumes verwendet wird. Die Optimierung dieser Hyperparameter erfolgt durch „GridSearchCV“. Die identifizierten optimalen Parameter werden verwendet, um den XGBoost Classifier auf den Trainingsdaten zu kalibrieren. Das finale Modell, resultierend aus diesem Prozess, wird anschließend mithilfe von Testdaten auf seine Vorhersagegenauigkeit und Generalisierungsfähigkeit hin untersucht.

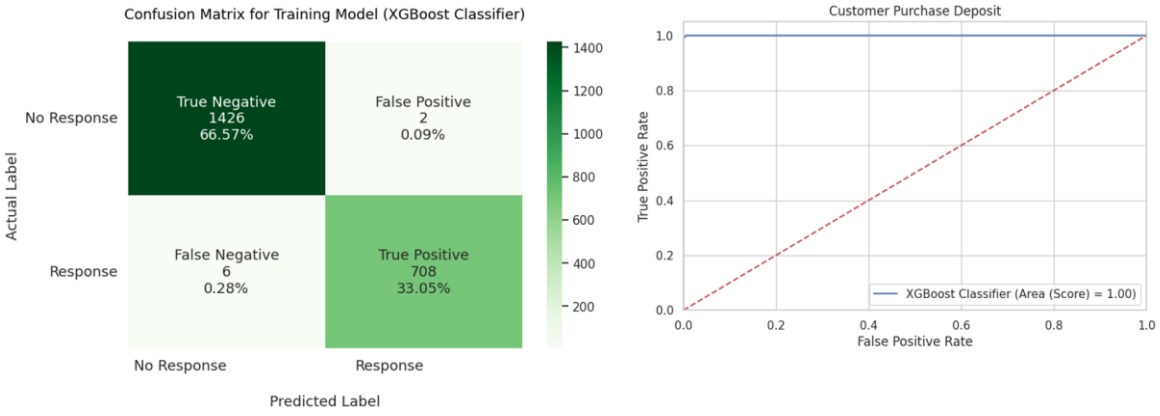

Abbildung 14: Trainingsleistung des XGBoost Classifiers: Konfusionsmatrix und ROC-Kurve

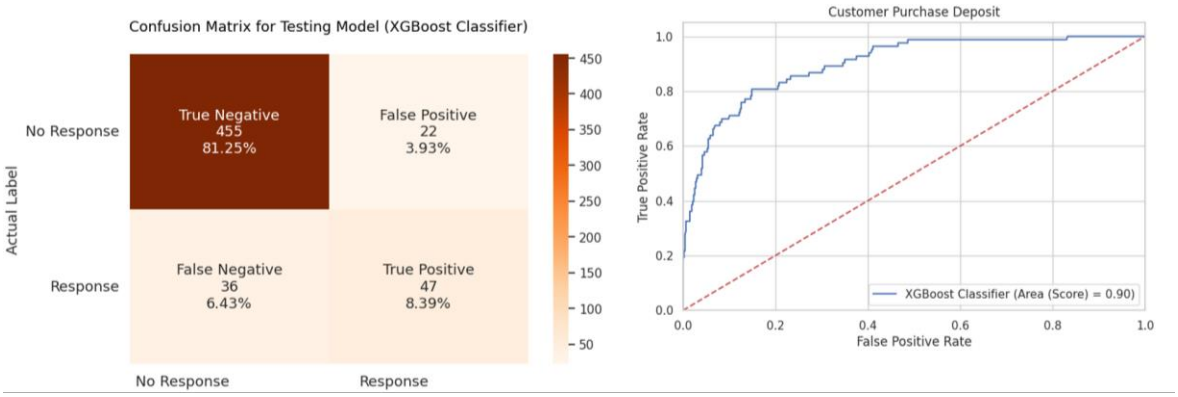

Abbildung 15: Testleistung des XGBoost Classifiers: Konfusionsmatrix und ROC-Kurve

4.1.4 Support Vector Machine

Support Vector Machines (SVM) sind zentrale Algorithmen im maschinellen Lernen für Klassifikationsaufgaben. Sie zeichnen sich durch ihre Robustheit und Effizienz bei der Trennung von Datensätzen in verschiedene Klassen aus. SVMs arbeiten, indem sie eine optimale Hyperplane im Feature-Raum finden, die den Abstand zwischen den Klassen maximiert. Bei linearen SVMs ist diese Hyperplane eine Linie oder Ebene, die Klassen auf Basis ihrer Merkmale trennt. Für nicht-lineare Datensätze verwendet der Kernel-Trick eine Transformation in höherdimensionale Räume, um eine lineare Trennung zu ermöglichen. Die mathematische Basis von SVMs beinhaltet die Maximierung der Marge zwischen den Klassen und die Minimierung einer Verlustfunktion, oft des Hinge-Loss. Die Lösung dieses Optimierungsproblems identifiziert die Support Vektoren, die kritisch für die Definition der Hyperplane sind [vgl. AwKh2015, S. 42]. Die Support Vector Machine wird in dieser Analyse durch den Einsatz des Standardkonstruktors „SVC“ aus dem sklearn.svm-Paket realisiert, wobei der RBF-Kernel zur Anwendung kommt. Erweiterte Verfahren zur Hyperparameter-Feinabstimmung wie „Grid Search“ und „Random Search“ führten nicht zu einer nennenswerten Performancesteigerung. Daher erfolgt das Modelltraining auf Basis der Standardparameter.

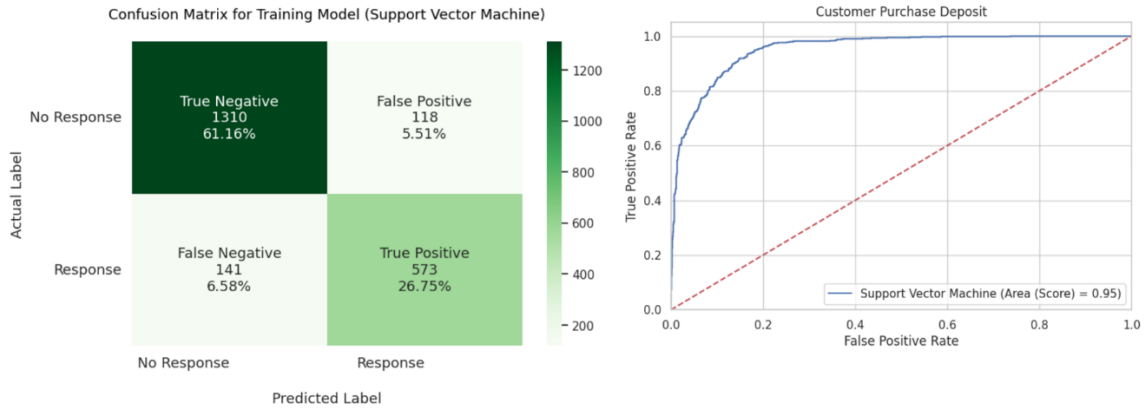

Abbildung 16: Konfusionsmatrix und ROC-Analyse der Support Vector Machine auf Trainingsdaten

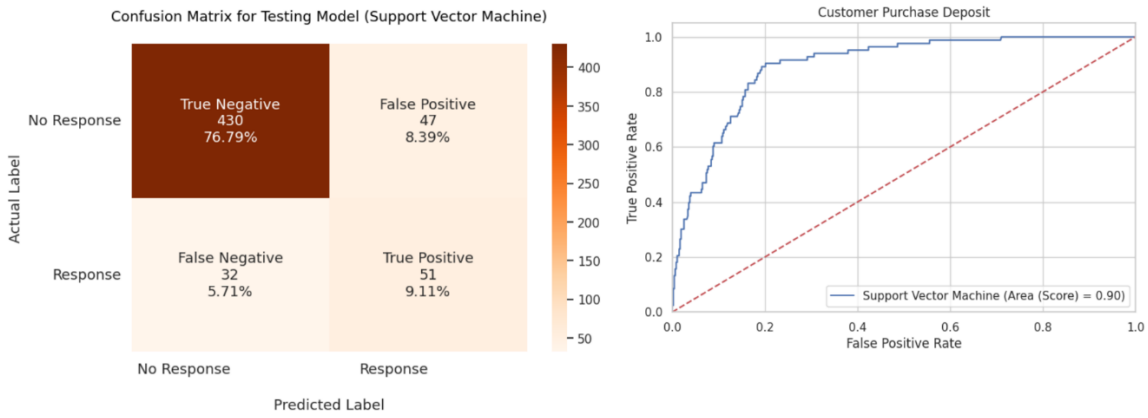

Abbildung 17: Evaluierung der Support Vector Machine auf Testdaten: Konfusionsmatrix und ROC-Kurve

4.1.5 Multi-Layer Perceptrons (MLP)

Das Multi-Layer Perceptron (MLP) ist ein fortgeschrittenes neuronales Netzwerkmodell, das sich besonders für die Modellierung komplexer, nichtlinearer Beziehungen eignet. Es besteht aus mehreren Schichten: Eingabe-, versteckten und Ausgabeschichten. In jedem Neuron werden Eingangssignale gewichtet und mittels Aktivierungsfunktionen verarbeitet, welche für die Einführung von Nichtlinearitäten entscheidend sind. Ein Schlüsselaspekt des MLPs ist das Backpropagation-Lernverfahren, bei dem Fehler vom Ausgang zurück durch das Netzwerk geführt werden, um die Gewichte effektiv anzupassen. Diese Anpassung erfolgt im Trainingsprozess, um die zugrundeliegenden Datenmuster optimal abzubilden [vgl. GB++2016, S. 196].

Im Rahmen dieser Untersuchung wird der MLPClassifier aus dem sklearn.neural_network-Modul genutzt, um ein Multi-Layer Perceptron zu entwickeln. Zur Feinjustierung des Modells dient der Einsatz von „RandomizedSearchCV“, welcher eine systematische Exploration des Hyperparameterraums ermöglicht. Konkret werden Parameter wie die Anzahl und Größe der verborgenen Schichten (hidden_layer_sizes), die Auswahl der Aktivierungsfunktion (activation), der Optimierungsalgorithmus (solver), der Regularisierungsparameter (alpha), die Anfangslernrate (learning_rate_init) und die Batch-Größe (batch_size) untersucht. Die Leistungsfähigkeit und Generalisierbarkeit des MLP wird mittels stratifizierter Kreuzvalidierung evaluiert, was sicherstellt, dass das Modell nicht nur auf die Trainingsdaten abgestimmt ist, sondern auch auf neue, unbekannte Datenmengen effektiv

angewendet werden kann.

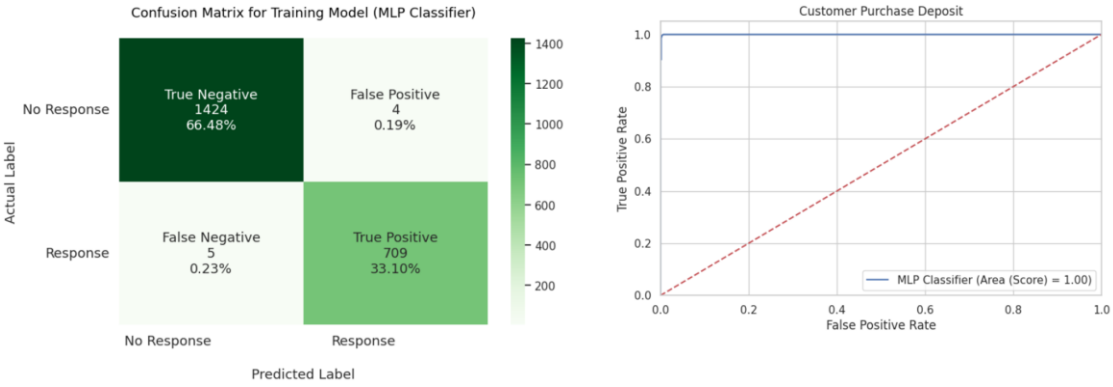

Abbildung 18: Konfusionsmatrix und ROC-Kurve des MLP-Classifiers - Training

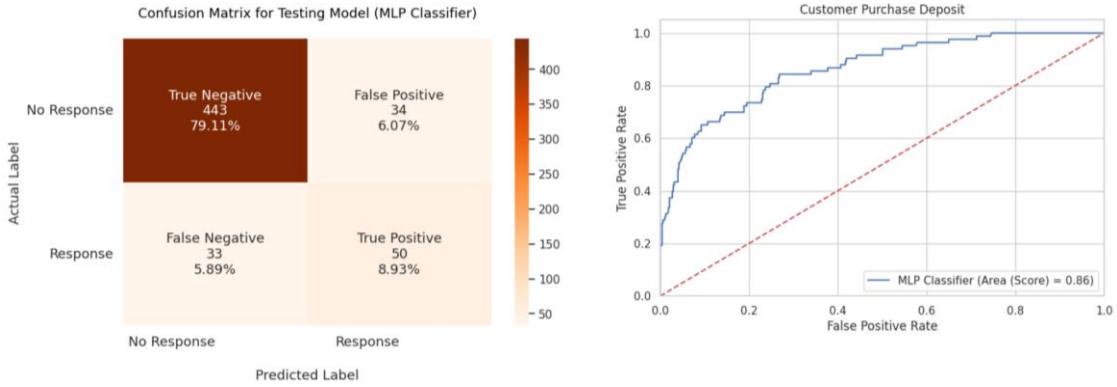

Abbildung 19: Konfusionsmatrix und ROC-Kurve des MLP-Classifiers – Test

4.1.6 Vergleich und Bewertung von Modellen

Im Kontext der Optimierung von Marketingkampagnen stützt sich die Modellbewertung auf drei Hauptparameter: Precision, Recall und F1 Score. Diese Metriken sind essenziell, um die Leistung von Klassifikationsmodellen präzise zu beurteilen und den Erfolg von Marketinginitiativen zu maximieren.

Die Präzision eines Modells gibt an, welcher Prozentsatz der als positiv klassifizierten Fälle tatsächlich positiv ist. Formal berechnet sich die Precision als $Precision = \frac{TP}{TP+FP}$, wobei TP die Anzahl der wahren positiven und FP die Anzahl der falschen positiven Ergebnisse darstellt [vgl. KuCB2021]. Eine hohe Precision reduziert die Zahl der Kunden, die irrtümlicherweise als reagierend klassifiziert werden, und trägt somit zur Effizienzsteigerung bei, indem Ressourcen gezielt für jene Kunden eingesetzt werden, die eine hohe Wahrscheinlichkeit zur Reaktion aufweisen.

Der Recall misst, welcher Anteil der tatsächlich positiven Fälle vom Modell erfasst wird. Die Formel lautet $Recall = \frac{TP}{TP+FN}$, wobei FN für die Anzahl der falsch negativen Ergebnisse steht. Ein hoher Recall sichert, dass möglichst wenige potenziell reagierende Kunden übersehen werden, was insbesondere für die Maximierung des Umsatzpotenzials von Kampagnen relevant ist.

Der F1 Score vereint Precision und Recall in einer Metrik und wird als $F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$ berechnet. Er ist besonders aussagekräftig in Situationen mit unausgeglichenen Daten, da er ein harmonisches Mittel darstellt, welches beide Aspekte berücksichtigt. Ein ausgewogener F1 Score signalisiert ein ausgewogenes Verhältnis zwischen der Identifizierung von Interessenten und der Vermeidung von Fehlansprachen.

Diese Parameter bilden die Grundlage für die Auswahl des optimalen Modells. Ziel ist es, ein Modell zu identifizieren, das die Reaktionsrate effektiv erhöht und gleichzeitig die Kosten durch die Minimierung von Falschpositiven senkt, ohne dabei Interessenten zu übersehen, was durch einen ausgewogenen F1 Score unterstützt wird.

Model (Test)	Accuracy	Precision	Recall	F1 Score	Cross Val F1 (k=5)	ROC AUC	Cross Val ROC AUC (k=5)
Random Forest	0.909	0.786	0.530	0.633	0.519	0.905	0.896
XGBoost Classifier	0.896	0.681	0.566	0.618	0.552	0.900	0.896
MLP Classifier	0.880	0.595	0.602	0.599	0.541	0.864	0.876
Logistic Regression	0.862	0.529	0.651	0.584	0.473	0.903	0.880
Support Vector Machine	0.859	0.520	0.614	0.564	0.448	0.895	0.884

Tabelle 7: Leistungsvergleich von Klassifikationsmodellen auf Testdaten

Im Vergleich zu anderen betrachteten Modellen, wie dem XGBoost Classifier, MLP Classifier, Logistic Regression und Support Vector Machine, zeichnet sich der Random Forest durch eine signifikant höhere Präzision aus. Dies ist ausschlaggebend, da die Präzision die zentrale Metrik für die Optimierung der Kampagnenkosten darstellt. Obwohl der XGBoost Classifier in den Metriken F1 Score und ROC AUC nahe an den Random Forest heranreicht, zeigt er eine geringere Präzision, was die Kostenwirksamkeit beeinträchtigen kann.

Die Modelle MLP Classifier und Support Vector Machine weisen zwar eine hohe Sensitivität (Recall) auf, ihre niedrigere Präzision könnte jedoch zu ineffizienten Marketingausgaben führen. Die Logistic Regression, obgleich sie

einen hohen Recall aufweist, bietet nicht die Präzision des Random Forest Classifiers, was die ökonomische Effizienz schmälert.

Der Random Forest Classifier überzeugt ferner durch einen ausgewogenen F1 Score, der eine harmonische Balance zwischen Präzision und Sensitivität symbolisiert, und somit die Effizienz in der Kundensegmentierung widerspiegelt. Ergänzend demonstrieren die hohen Werte der ROC AUC und der Cross-Validated ROC AUC eine hervorragende Diskriminierungsfähigkeit zwischen den Klassen, was die Zuverlässigkeit des Modells unterstreicht.

Zusammenfassend lässt sich festhalten, dass der Random-Forest-Classifier die Geschäftsziele effektiv unterstützt. Er trägt signifikant zur Steigerung der Antwortrate bei und senkt gleichzeitig die Marketingkosten, was wesentlich zur Gewinnmaximierung beiträgt. Diese Merkmale qualifizieren den Random-Forest-Classifier als bevorzugtes Modell für die Implementierung von Optimierungsmaßnahmen in Marketingkampagnen.

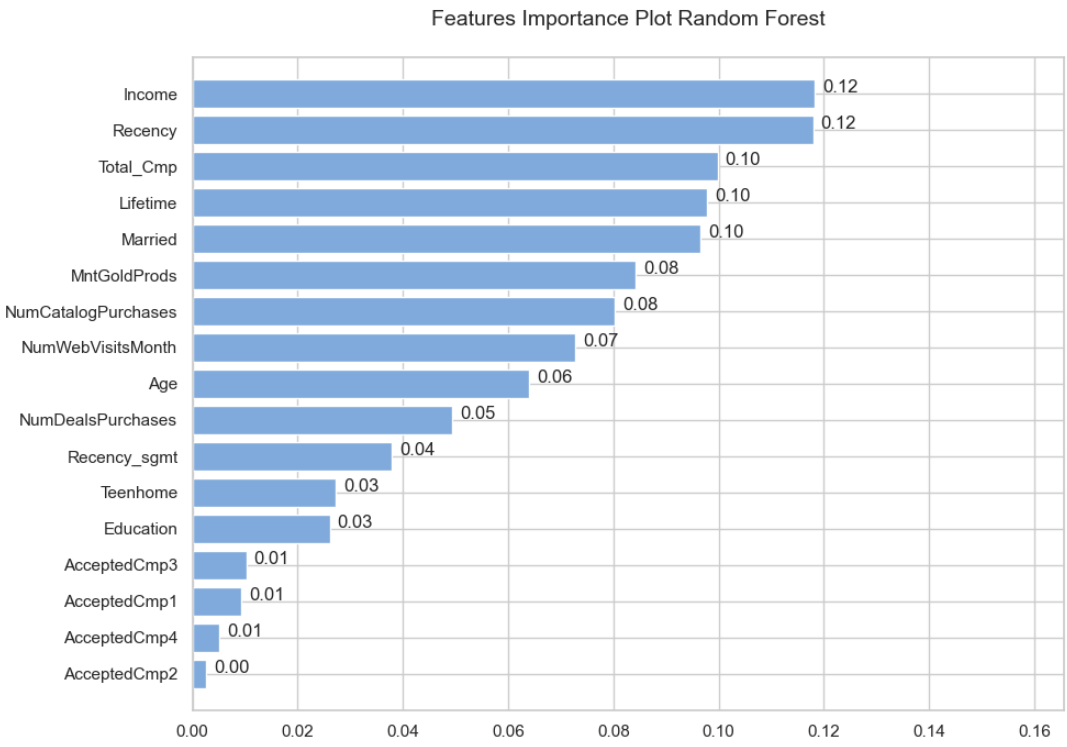

Abbildung 20: Feature-Importance-Analyse im Random Forest-Modell

Die Analyse der Feature-Importance im Random Forest-Modell identifiziert das Einkommen und die Recency als maßgebliche Prädiktoren für die Vorhersagekraft des Modells, was durch die höchsten Werte für die Merkmalswichtigkeit (jeweils 0,12) unterstrichen wird. Zusätzlich tragen die

Gesamtanzahl der Kampagnenteilnahmen (Total_Cmp), die Länge der Kundenbeziehung (Lifetime) und der Familienstand (Married) wesentlich zu den Vorhersagen bei. Andere Faktoren wie Ausgaben für Goldprodukte (MntGoldProds), die Anzahl der Katalogkäufe (NumCatalogPurchases), die Häufigkeit der Webseitenbesuche (NumWebVisitsMonth) und das Alter der Kunden spielen ebenfalls eine Rolle, allerdings mit einer geringeren Gewichtung. Diese Einsichten sind von großer Bedeutung für die strategische Ausrichtung der Marketingbemühungen, da sie eine fokussierte Kundenkommunikation unterstützen, bei der die relevanten Merkmale in den Mittelpunkt gestellt werden.

4.2 Clustering mit RFM-Analyse

4.2.1 Grundlagen und Anwendung

Die RFM-Analyse, eine Methode zur Kundensegmentierung, zielt darauf ab, Kunden basierend auf ihrem Kaufverhalten in Gruppen zu unterteilen. Die Abkürzung RFM steht für „Recency“ (Aktualität des letzten Kaufs), „Frequency“ (Kaufhäufigkeit) und „Monetary“ (monetärer Wert der Käufe). Diese drei Variablen bieten aussagekräftige Indikatoren für das Engagement und die Loyalität der Kunden.

Die Integration von Clustering-Verfahren, basierend auf RFM-Daten, in die Marketingstrategie stellt einen wesentlichen Schritt zur Effektivitätssteigerung von Kampagnen dar. Diese Technik ermöglicht es, Kundenbeziehungen detaillierter zu analysieren und trägt entscheidend zur Optimierung der Marketingaktivitäten bei. Durch die gezielte Segmentierung der Kundenbasis anhand von Kaufverhalten und Wertbeitrag können Marketingressourcen effizienter zugewiesen und maßgeschneiderte Kampagnen für spezifische Kundengruppen entwickelt werden. Dieser Ansatz erhöht nicht nur die Reaktionsraten, sondern verbessert auch das Kosten-Umsatz-Verhältnis signifikant, indem er die Marketingbemühungen auf die Segmente mit dem höchsten Potenzial konzentriert. Dadurch wird ein direkter Beitrag zur Steigerung der Rentabilität der Marketingkampagnen und folglich zum Unternehmenserfolg geleistet [vgl. ErBK2021, S. 1].

4.2.2 Modellierung und Bewertung

Im initialen Schritt dieser Analyse wird ein DataFrame für die RFM-Segmentierung erstellt, in dem zunächst die Gesamtausgaben („Spending“) jedes Kunden durch Addition der Ausgaben in verschiedenen Produktkategorien berechnet werden. Parallel dazu werden die Gesamtkäufe („Total_Purchases“) durch Summierung der Transaktionen über verschiedene Einkaufskanäle erfasst. Anschließend wird der DataFrame auf Schlüsselspalten wie „ID“, „Spending“, „Total_Purchases“ und „Recency“ reduziert. Für die RFM-Analyse werden die Schlüsselkomponenten Recency, Frequency und Monetary erarbeitet. Der abschließende DataFrame vereint diese Elemente, um eine fundierte Grundlage für die Segmentierung der Kunden nach ihren spezifischen RFM-Profilen zu bieten.

Um die optimale Anzahl von Clustern für jede Dimension der RFM-Analyse zu bestimmen, wird die Elbow-Methode angewandt. Dabei berechnet der KMedoids-Algorithmus, der auf euklidischen Distanzen basiert, die Summe der quadratischen Abstände der Datenpunkte zu den jeweiligen Clusterzentren für unterschiedliche Anzahlen von Clustern. Die graphische Darstellung dieser Summen ermöglicht die Identifikation des Punktes, an dem zusätzliche Cluster keine signifikante Verringerung der Varianz mehr bewirken. Dieser Punkt, häufig als „Elbow“ bezeichnet, markiert die Anzahl der Cluster, bei der ein effizienter Kompromiss zwischen der Anzahl der Cluster und der Homogenität innerhalb der Cluster erreicht wird. Aus dieser Analyse ergibt sich die Entscheidung für eine Unterteilung der Kundendaten in fünf Cluster pro RFM-Dimension, was in einer Gesamtheit von 5×5×5 unterschiedlichen Clustern resultiert.

RecencyCluster	count	mean	std	min	25%	50%	75%	max
0	392.000000	90.418367	5.137841	82.000000	86.000000	91.000000	95.000000	99.000000
1	430.000000	72.153488	5.593009	63.000000	67.000000	72.000000	77.000000	81.000000
2	472.000000	52.027542	5.706301	42.000000	47.750000	52.000000	56.000000	62.000000
3	490.000000	30.240816	6.173296	20.000000	25.000000	30.000000	36.000000	41.000000
4	456.000000	9.122807	5.849227	0.000000	4.000000	9.000000	14.000000	19.000000

Tabelle 8: Verteilung und deskriptive Statistiken der Recency-Cluster

FrequencyCluster	count	mean	std	min	25%	50%	75%	max
0	832.000000	6.608173	2.012077	0.000000	5.000000	7.000000	8.000000	10.000000
1	310.000000	12.996774	1.493239	11.000000	12.000000	13.000000	14.000000	15.000000
2	491.000000	17.940937	1.425201	16.000000	17.000000	18.000000	19.000000	20.000000
3	400.000000	22.770000	1.398925	21.000000	22.000000	23.000000	24.000000	25.000000
4	207.000000	28.246377	2.808132	26.000000	26.000000	27.000000	29.000000	44.000000

Tabelle 9: Verteilung und deskriptive Statistiken der Frequency-Cluster

MonetaryCluster	count	mean	std	min	25%	50%	75%	max
0	939.000000	74.423855	53.566384	5.000000	36.000000	57.000000	97.500000	231.000000
1	398.000000	407.756281	110.683876	232.000000	311.000000	406.000000	493.750000	615.000000
2	314.000000	828.961783	113.620066	622.000000	731.000000	832.500000	928.000000	1009.000000
3	315.000000	1208.228571	122.867096	1012.000000	1102.500000	1189.000000	1315.000000	1445.000000
4	274.000000	1766.171533	243.027162	1449.000000	1574.000000	1701.500000	1919.000000	2525.000000

Tabelle 10: Verteilung und deskriptive Statistiken der Monetary-Cluster

Die Tabellen 8, 9 und 10 verdeutlichen die Streuung der Kundencharakteristika innerhalb der definierten RFM-Cluster. In der Recency-Dimension illustriert Cluster 0 die Kunden mit dem längsten Zeitraum seit dem letzten Kauf, während Cluster 4 Kunden umfasst, die kürzlich Einkäufe getätigt haben. Die Frequency-Dimension zeigt in Cluster 0 Kunden mit der geringsten Kaufhäufigkeit; dem gegenüber steht Cluster 4, das Kunden mit der höchsten Transaktionsanzahl kennzeichnet. Im Bereich des Monetary-Werts weist Cluster 0 die Kundensegmente mit den niedrigsten Ausgaben auf, wohingegen Cluster 4 jene mit den höchsten Ausgaben auszeichnet. Diese differenzierte Darstellung ermöglicht eine tiefgehende Analyse des Kaufverhaltens und liefert wertvolle Ansatzpunkte für gezielte Marketingstrategien.

Die Erstellung von 5x5x5 Clustern resultiert in 125 unterschiedlichen Kundensegmenten. Diese Anzahl stellt in der Praxis eine Herausforderung dar, da die effektive Arbeit mit einer solchen Vielzahl an Segmenten kaum realisierbar ist. Zur Verbesserung der Handhabbarkeit ist es sinnvoll, die 125 initialen Segmente in 10 umfassendere Kundensegmente zu konsolidieren. Diese Konsolidierung ermöglicht eine praxisorientierte und effizientere Segmentierung, die sowohl die Vielfalt der Kundenbedürfnisse berücksichtigt als auch die Effektivität der marketingstrategischen Maßnahmen erhöht [vgl. Wutt2023].

Die Entscheidung, Kunden in zehn Segmente zu unterteilen, basiert auf einer eingehenden Analyse des Silhouette Scores. Dieser Score bietet ein Maß für die Güte der Clusterbildung, indem er die Kohäsion innerhalb der Cluster und die Separation zwischen ihnen bewertet. Ein hoher durchschnittlicher Silhouette Score, der sich dem Wert 1 annähert, deutet auf eine ausgeprägte Qualität der Clusterbildung hin.

Der Silhouettenplot in Abbildung 21 zeigt eine konsistente Clusterzuordnung für das KMedoids-Clustering von 2240 Proben, gekennzeichnet durch einen durchschnittlichen Silhouettenwert von 0,76. Dieser Wert unterstreicht eine allgemein gute Separierung der Proben innerhalb ihrer jeweiligen Cluster. Die breiten Bereiche der Plots über die verschiedenen Gruppen hinweg bestätigen eine stabile Clusterstruktur, obwohl einige Proben niedrigere Silhouettenwerte aufweisen, was eine geringere Passgenauigkeit innerhalb ihres Clusters anzeigt.

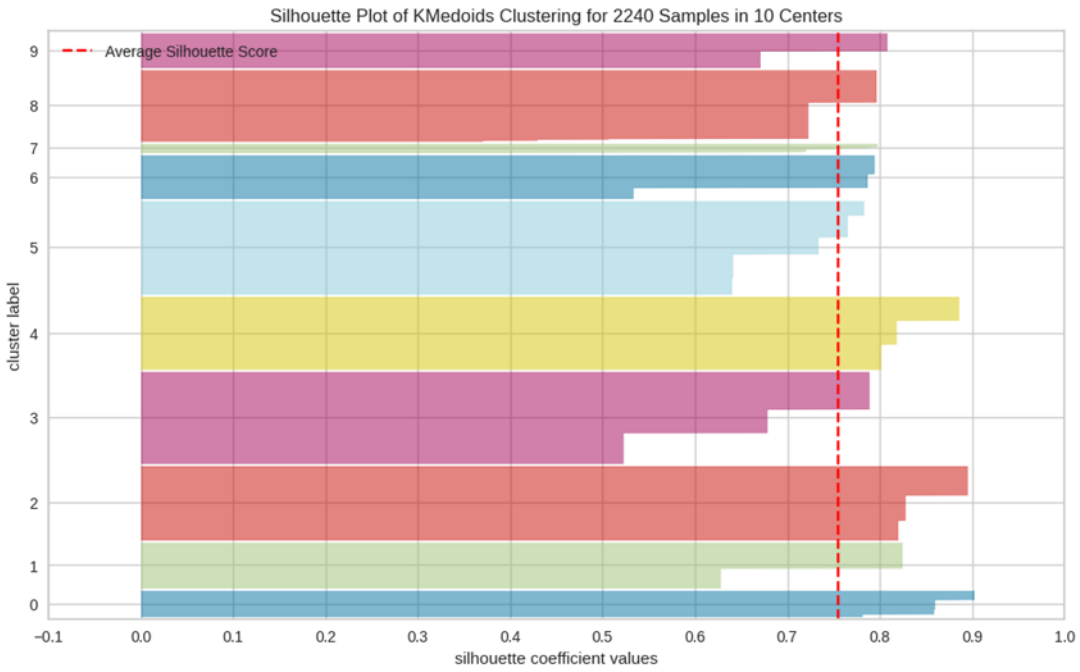

Abbildung 21: Silhouettenanalyse der KMedoids-Clusterung mit zehn Clustern

Die Segmentierung von Kunden erfolgt durch die Zuordnung zu bestimmten Gruppen basierend auf deren Kaufverhalten, welches durch die Metriken Recency und Frequency definiert wird. Jeder Kunde erhält eine numerische Repräsentation, die seiner Aktivität entspricht: die erste Ziffer steht für Recency, die zweite für Frequency. Die aggregierten Ziffern werden in vordefinierte Segmente übersetzt, die aussagekräftige Bezeichnungen tragen

und das Engagement der Kunden widerspiegeln:

Kundensegment	Beschreibung
Champions (44)	Kunden, die kürzlich eingekauft haben und dies häufig tun. Sie sind sehr wertvoll.
Treue Bestandskunden (33, 34, 43, 23, 24)	Kunden, die regelmäßig einkaufen. Sie haben eine starke Bindung zum Unternehmen.
Potenzielle treue Bestandskunden (32, 31, 42, 41)	Kunden, die neu sind, aber bereits ein Muster regelmäßiger Einkäufe zeigen. Sie haben das Potenzial, treue Kunden zu werden.
Kunden, die Aufmerksamkeit benötigen (21, 22, 12, 11)	Kunden, die weniger häufig einkaufen und deren letzter Kauf schon eine Weile her ist. Sie benötigen Anreize, um wieder aktiver zu werden.
Wichtig, aber inaktiv (14, 04)	Kunden, die früher regelmäßig eingekauft haben, aber schon seit längerem nicht mehr. Sie sind risikobehaftet, da sie früher wertvoll waren.
Gefährdete Kunden (02, 03, 13)	Kunden, die früher häufiger eingekauft haben, aber schon seit längerem nicht mehr. Sie sind in Gefahr, verloren zu gehen.
Gefahr zur Inaktivität (20)	Kunden, die selten einkaufen und deren letzter Kauf schon länger zurückliegt. Sie sind in Gefahr, inaktiv zu werden.
Neue Kunden (40)	Kunden, die vor kurzem ihren ersten Kauf getätigt haben. Ihre zukünftige Loyalität ist noch unklar.
Offenes Potential (30)	Kunden, die erst kürzlich ihren ersten Einkauf getätigt haben. Ihre zukünftige Bindung ist noch ungewiss.
Verloren (10, 00, 01)	Kunden, die selten einkaufen und deren letzter Einkauf sehr lange zurückliegt. Sie sind wahrscheinlich bereits verloren.

Tabelle 11: Übersicht der Kundensegmente und ihre Charakteristika

Durch die Ermittlung durchschnittlicher Werte für Recency, Frequency und Monetary wird die Größe und Beschaffenheit der definierten Kundensegmente klar ersichtlich. Dabei werden die Segmente nach der relativen Position ihrer durchschnittlichen Metrik-Werte in den Verteilungsquartilen des gesamten Datensatzes klassifiziert. Diese Einteilung in „niedrig“, „mittel“ und „hoch“ ermöglicht eine präzise Differenzierung der Kundenprofile nach ihrem Kaufverhalten.

Auf Grundlage dieser Methodik können Marketingstrategien präziser ausgerichtet werden. Die strukturierte Übersicht in der Tabelle 12 ermöglicht es Marketingspezialisten, maßgeschneiderte Strategien zu entwerfen, die auf die spezifischen Eigenschaften und das Potenzial jedes Kundensegments zugeschnitten sind. Dies trägt zu einer gesteigerten Ansprechrate bei und fördert die Rentabilität der Marketinginitiativen.

Segment	Recency	Frequency	Monetary
Champions	niedrig	hoch	mittel
Gefahr zur Inaktivität	mittel	niedrig	niedrig
Gefährdete Kunden	hoch	mittel	hoch
Kunden, die Aufmerksamkeit benötigen	mittel	mittel	mittel
Neue Kunden	niedrig	niedrig	niedrig
Offenes Potential	mittel	niedrig	niedrig
Potentielle treue Bestandskunden	niedrig	mittel	mittel
Treue Bestandskunden	mittel	hoch	hoch
Verloren	hoch	mittel	mittel
Wichtig, aber inaktiv	hoch	hoch	hoch

Tabelle 12: Kundensegmentierung nach RFM-Werten

5 Strategische Marketinganalyse und Ergebnisbewertung

Im abschließenden Kapitel dieser Arbeit erfolgt eine umfassende Auswertung der zuvor erarbeiteten Modelle zur Datenanalyse, um ihre Auswirkungen auf die Optimierung von Marketingstrategien zu beurteilen. Die Kernkompetenz dieser Modelle zeigt sich in ihrer Fähigkeit, fundierte Entscheidungen zu unterstützen und die Wirksamkeit von Marketinginitiativen zu erhöhen.

Das traditionelle Marketingvorgehen, dargestellt in Abbildung 22, offenbart die Herausforderungen einer breitgefächerten Kundenansprache ohne spezifische Zielgruppenfokussierung. Diese Methode resultierte oft in geringen Reaktionsraten und einer negativen Netto-Profit-Marge (NPM), was auf ein ineffizientes Verhältnis von Investition zu Ertrag hinweist.

Abbildung 22: Traditionelle Marketingstrategie vor Einsatz von Klassifizierungs- und Clusteringmodellen

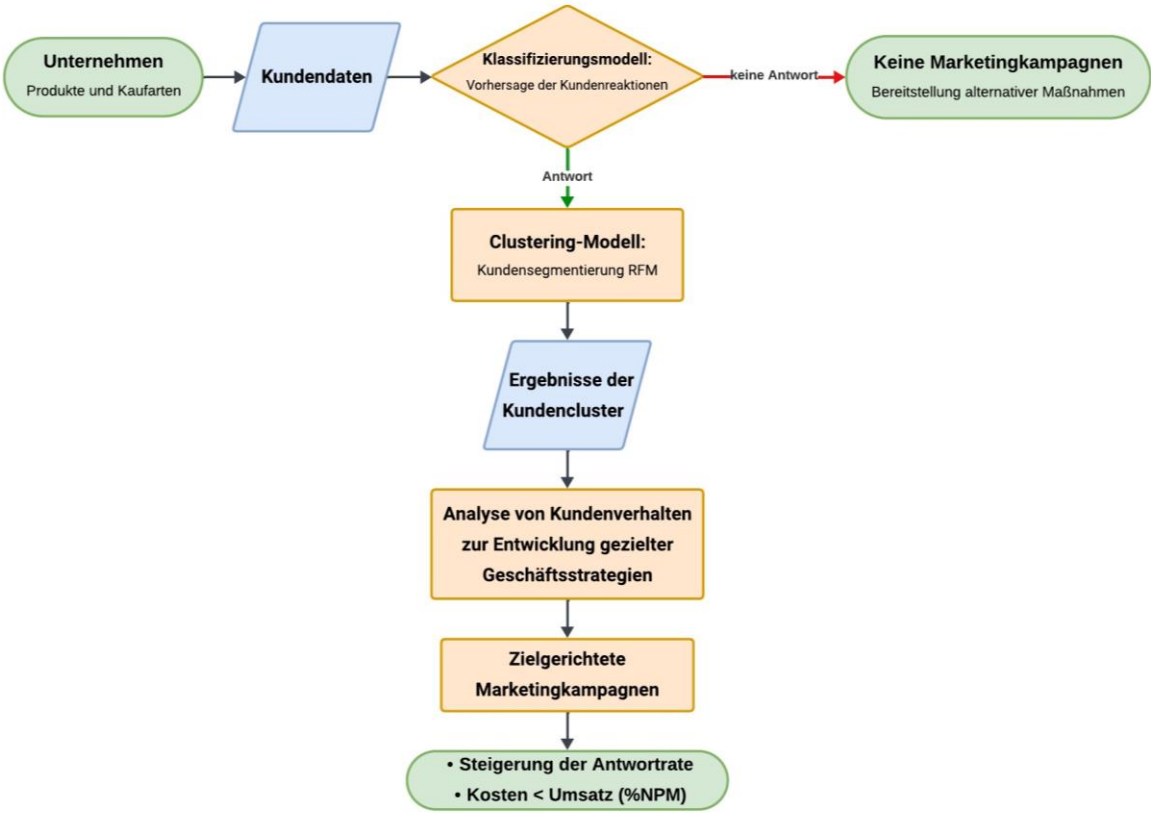

Abbildung 23: Datengesteuerte Marketingoptimierung durch Klassifizierung und RFM-Clustering

Demgegenüber illustriert Abbildung 23 einen fortschrittlichen, datenbasierten Ansatz. Durch den Einsatz von Klassifizierungsmodellen wird eine präzise Vorhersage der Kundenreaktionen ermöglicht. Nicht reagierende Kunden werden von Kampagnen ausgeschlossen, was zu einer effizienteren Ressourcennutzung führt. Zusätzlich wird das RFM-Modell angewandt, um Kunden anhand ihres Kaufverhaltens zu segmentieren. Dieses Vorgehen ermöglicht eine zielgerichtete Strategieentwicklung mit dem Fokus auf die Maximierung der Reaktionsrate und die Optimierung des Kosten-Umsatz-Verhältnisses.

Dieses Kapitel analysiert die Vorteile eines datengesteuerten Marketingansatzes und dessen Beitrag zur Erreichung der zentralen Ziele dieser Arbeit: Steigerung der Reaktionsrate und Erhöhung der Rentabilität. Insbesondere wird der Einfluss auf wichtige Geschäftsmetriken wie die Reaktionsrate und Nettomarge diskutiert. Es wird aufgezeigt, wie derartige strategische Neuausrichtungen durch datenbasierte Kundensegmentierung die betriebliche Effizienz steigern und die Profitabilität nachhaltig verbessern können.

5.1 Zielgerichtete Marketingkampagnen für Kundensegmente

Die Optimierung von Marketingkampagnen durch die gezielte Ausrichtung auf spezifische Kundensegmente ist ein wesentlicher Schritt zur Steigerung der Reaktionsraten und der Rentabilität von Marketingaktivitäten. Die Anwendung der RFM-Analyse ermöglicht es, differenzierte Kundeninformationen zu extrahieren und Marketingmaßnahmen zu entwickeln, die auf die individuellen Eigenschaften und Verhaltensweisen der Kundensegmente zugeschnitten sind. Ziel ist eine Personalisierung der Kommunikation, durch die Kunden eine höhere Wertschätzung erfahren und eine engere Bindung zum Unternehmen aufbauen. Die segmentspezifische Ansprache, basierend auf Engagement und Kaufverhalten der Kunden, erlaubt die Erstellung maßgeschneiderter Angebote, die sowohl den Kundenanforderungen als auch den Unternehmenszielen gerecht werden.

Die zielgerichtete Fokussierung auf Kundensegmente, die durch eine rezent bis mittlere Kaufrate (Recency), eine mittlere bis hohe Kaufhäufigkeit (Frequency) und eine beträchtliche Ausgabebereitschaft (Monetary) charakterisiert sind, ermöglicht eine optimale Nutzung der Ressourcen. Dieser Ansatz basiert auf der Hypothese, dass Kunden mit einer frischen Interaktion mit dem Unternehmen eine höhere Wahrscheinlichkeit für Folgekäufe aufweisen. Kunden mit mittlerer Recency sind potenziell wieder aktivierbar und bieten somit eine günstige Gelegenheit für erneute Marketinginitiativen. Eine ausgeprägte Kaufhäufigkeit indiziert eine feste Kundenbindung und ein tiefes Engagement, was diese Kundengruppe als Ziel für fortlaufende Marketingbemühungen attraktiv macht. Die hohe Kaufhäufigkeit signalisiert eine bereits bestehende positive Beziehung zum Unternehmen und erhöht die Wahrscheinlichkeit, dass diese Kunden auf weitere Marketingmaßnahmen positiv reagieren. Die Ausgabebereitschaft reflektiert den direkten finanziellen Beitrag der Kunden zum Unternehmensumsatz. Kunden mit einer hohen „Monetary“-Kennzahl sind daher für Investitionen in individuellere Marketingmaßnahmen prädestiniert, da sie einen bedeutenden Einfluss auf die Umsatzentwicklung haben.

Die Segmente „Champions“, „Treue Bestandskunden“, „Potenzielle treue Bestandskunden“ und „Kunden, die Aufmerksamkeit benötigen“ stehen im

Mittelpunkt, da sie aufgrund ihrer bisherigen Kauftätigkeiten das Potenzial für wiederkehrende und steigende Umsätze aufzeigen. „Champions“ und „Treue Bestandskunden“ repräsentieren die Kernbasis des Unternehmens mit regelmäßigen und hohen Umsätzen. „Potenzielle treue Bestandskunden“ bieten ein Wachstumspotenzial für zukünftige Umsätze, während das Segment „Kunden, die Aufmerksamkeit benötigen“ Möglichkeiten zur Intensivierung der Kundenbeziehung durch gezielte Marketingaktionen aufzeigt.

Die Priorisierung dieser Segmente über andere wie „Neue Kunden“ oder „Gefährdete Kunden“ ist eine strategische Entscheidung, die sich an den Zielsetzungen der Marketingstrategie orientiert. „Neue Kunden“ bedürfen einer andersartigen Herangehensweise, um ihr Loyalitätspotenzial zu entwickeln, und „Gefährdete Kunden“ erfordern Maßnahmen zur Vermeidung von Kundenabwanderungen, die in einer separaten strategischen Überlegung berücksichtigt werden. Die Konzentration auf die genannten vier Segmente ermöglicht es dem Unternehmen, die Marketingeffizienz zu erhöhen und die Rentabilität der Marketingaktivitäten zu maximieren.

Die Tabelle 13 illustriert die ausgewählten Kundensegmente und die entsprechenden strategischen Ansätze, die spezifisch auf die charakteristischen Bedürfnisse und Verhaltensweisen jedes Kundensegments zugeschnitten sind. Diese Strategien reichen von Loyalitäts- und Rabattprogrammen bis hin zu gezielten Verkaufsförderungsaktionen und Produktbündelungen, um die Interaktion und das Engagement der Kunden zu steigern. Die Umsetzung dieser Strategien erfolgt primär durch Direktmarketing-Maßnahmen wie personalisierte E-Mail- oder SMS-Kampagnen. Diese ermöglichen eine direkte Ansprache der Kunden mit maßgeschneiderten Angeboten, die auf deren Kaufhistorie und Präferenzen basieren. Ziel ist es, durch diese individualisierte Kommunikation nicht nur die Aufmerksamkeit der Kunden zu erhöhen, sondern auch eine stärkere Bindung zum Unternehmen zu fördern und letztendlich die Reaktionsraten auf Marketinginitiativen zu optimieren.

Kundentyp	Charakteristika	Strategien
Champions	Recency: Niedrig Frequency: Hoch Monetary: Mittel	Spezialbehandlung mit Belohnungen wie Rabattprogramme für spezifische Artikel mittels Bonuspunkten, Prämien für bestimmte Kaufsummen, etc.
Treue Bestandskunden	Recency: Mittel Frequency: Hoch Monetary: Hoch	Steigerung der Kundenaktivität durch ein Punktesammelprogramm, das gegen bestimmte Artikel/Promotions eingetauscht werden kann, Rabatte für Mitglieder, usw.
Potenzielle treue Bestandskunden	Recency: Niedrig Frequency: Mittel Monetary: Mittel	Steigerung der Einkaufsfrequenz durch Verkaufsförderungen/Rabatte auf bestimmte Artikel an bestimmten Tagen oder Daten.
Kunden, die Aufmerksamkeit benötigen	Recency: Mittel Frequency: Mittel Monetary: Mittel	Zusätzliche Aufmerksamkeit, um die Einkaufsaktivität und -intensität zu erhöhen, wie Bündelungsaktionen für beliebte Artikel. Beispiel: Bündelungspromo für Rindfleisch + Wein.

Tabelle 13: Strategische Marketingansätze für Kundensegmente

Die Konzentration auf diese Kundensegmente fußt auf fundierten Datenanalysen und einem vertieften Verständnis des Kundenverhaltens. Dies ermöglicht die Entwicklung präziser Marketingstrategien, die auf die spezifischen Bedürfnisse und das Potenzial der einzelnen Segmente abgestimmt sind. Eine solche Strategie fördert nicht nur eine höhere Reaktionsrate auf Marketingkampagnen, sondern trägt auch zur Reduktion der Kosten bei der Kundenakquise und zur Umsatzsteigerung bei.

5.2 Leistungsbewertung anhand Geschäftsmetrik

Die Nutzung von Geschäftsmetriken ist für die Bewertung der Effektivität von Unternehmensstrategien unerlässlich. Sie stellen quantitative Indikatoren dar, die den Erfolg verschiedener Geschäftsbereiche messen und sind daher entscheidend für fundierte strategische Entscheidungen. In dieser Bachelorarbeit dienen sie insbesondere dazu, die Auswirkungen der implementierten Klassifizierungs- und Clustering-Modelle auf die Marketingeffizienz und Unternehmensleistung zu analysieren. Durch den Vergleich von Metriken wie Reaktionsrate und Nettomarge vor und nach der Modellanwendung wird die Effektivität der angewandten Methoden beurteilt. Diese Herangehensweise ermöglicht es, den konkreten Nutzen der

datengestützten Kundenanalyse für die Verbesserung der Geschäftsergebnisse zu veranschaulichen.

5.2.1 Analyse der Reaktionsraten: Vor und nach der Modellimplementierung

Die Reaktionsrate ist ein wesentlicher Indikator für die Effektivität von Marketingkampagnen. Sie wird berechnet als das Verhältnis der Anzahl der positiven Kundenreaktionen zur Gesamtzahl der angesprochenen Kunden und gibt Aufschluss über die Wirksamkeit der Kampagnen.

Vor der Implementierung des Klassifizierungsmodells wurden Marketingkampagnen an die gesamte Kundenbasis von 2240 Kunden gerichtet. Davon reagierten 334 Kunden positiv, während 1906 keine Reaktion zeigten. Die Reaktionsrate wird mittels der Formel $\left(\frac{\text{Anzahl der Reaktionen}}{\text{Gesamtanzahl der Kunden}}\right) \times 100$ berechnet, was in diesem Fall $\left(\frac{334}{2240}\right) \times 100 \approx \mathbf{14.91}\%$ ergibt. Diese niedrige Reaktionsrate weist auf eine geringe Effizienz der damaligen Marketingansätze hin.

Nach der Einführung des Klassifizierungsmodells mittels Random Forest wurden folgende Ergebnisse erzielt: 44 True Positives (TP) und 12 False Positives (FP). Insgesamt wurden die Marketingkampagnen somit an 56 Kunden gesendet. Die neue Reaktionsrate berechnet sich nun zu $\left(\frac{44}{56}\right) \times 100 \approx \mathbf{78.57}\%$.

Die deutliche Verbesserung der Reaktionsrate nach der Modellimplementierung zeigt, wie effektiv die gezielte Kundenansprache durch datengestützte Modelle sein kann. Diese Methode ermöglicht eine präzisere Kundensegmentierung und effizientere Ressourcennutzung, was sich in einer wesentlich höheren Reaktionsrate und somit in einer effektiveren Marketingstrategie niederschlägt.

5.2.2 Vergleich der Nettomargen: Bewertung der finanziellen Effizienz

Die Nettomarge ist ein entscheidender Indikator für die finanzielle Effizienz von Marketingkampagnen. Sie gibt Aufschluss darüber, inwieweit die Einnahmen die Kosten der Kampagnen übersteigen und spiegelt somit die Rentabilität der

Marketinginvestitionen wider. Aus dem Datensatz geht hervor, dass die Kosten für den Kontakt mit einem Kunden 3 Dollar und die Einnahmen nach Annahme des Angebots durch den Kunden 11 Dollar betragen.

Vor der Implementierung des Klassifizierungsmodells resultierten die Kampagnen in Gesamtkosten von $3\,\$ \times 2240\,\text{Kunden} = 6720\,\$$. Bei einer Reaktionsrate von 334 Kunden beliefen sich die Gesamteinnahmen auf $11\,\$ \times 334\,Kunden = 3674\,\$$. Der Gesamtgewinn lag somit bei $3674\,\$ - 6720\,\$ = -3046\,\$$. Die Netto-Profit-Marge, berechnet als $\frac{\text{Gesamtgewinn}}{\text{Gesamteinnahmen}} \times 100 = \frac{-3046}{3674} \times 100 \approx \mathbf{-82.91}\%$, was auf eine ineffiziente Kostenstruktur hinweist.

Nach der Modellimplementierung veränderte sich das Bild deutlich. Die Gesamtkosten für die Marketingkampagne beliefen sich auf $3\,\$ \times 56\,Kunden = 168\,\$$, während die Gesamteinnahmen bei $11\,\$ \times 44\,\text{Kunden} = 484\,\$$ lagen. Dies führte zu einem Gesamtgewinn von $484\,\$ - 168\,\$ = 316\,\$$. Die daraus resultierende Netto-Profit-Marge betrug $\frac{316\,\$}{484\,\$} \times 100 \approx \mathbf{65.29}\%$, was eine erhebliche Verbesserung der Rentabilität bedeutet.

Diese Analyse unterstreicht die Effektivität des datengestützten Ansatzes zur Optimierung von Marketingkampagnen. Durch die gezielte Ansprache reaktionsfreudiger Kunden konnte eine Reduktion der Marketingkosten bei gleichzeitiger Erhöhung der Einnahmen erreicht werden. Somit wird die Bedeutung der Implementierung von datenbasierten Modellen für die Steigerung der finanziellen Effizienz und der Profitabilität von Marketingaktivitäten hervorgehoben.

6 Zusammenfassung und Ausblick

Diese Bachelorarbeit hat die transformative Kraft der Künstlichen Intelligenz (KI) im Bereich des Marketings untersucht, wobei ein besonderer Fokus auf die Verbesserung der Präzision und Effektivität von Marketingkampagnen gelegt wurde. Es wurde gezeigt, dass durch den Einsatz von KI-Techniken und Data-Mining-Methoden eine signifikante Verbesserung der Kundenansprache erreicht werden kann, was sich in einer effizienteren Segmentierung und einer

zielgerichteteren Kommunikation manifestiert. Diese Optimierung trägt dazu bei, die Herausforderungen traditioneller Marketingansätze zu überwinden, indem sie eine differenzierte und effektive Kundenansprache ermöglicht.

Im Rahmen der Arbeit wurde ein spezifischer Datensatz von der Plattform „Kaggle“ verwendet, um die praktische Anwendung und das Potenzial von KI im Marketing zu demonstrieren. Die Analyse dieses Datensatzes, die Datenaufbereitung sowie die Entwicklung und Bewertung verschiedener Klassifikationsmodelle bildeten den Kern der Forschung. Besonderes Augenmerk wurde dabei auf die Anwendung verschiedener Modelle gelegt. Zusätzlich wurde eine Clustering-Analyse mit RFM-Metriken durchgeführt, um eine effektive Segmentierung der Kunden zu ermöglichen.

Die kritische Reflexion der Ergebnisse zeigt, dass, obwohl wichtige Einsichten gewonnen wurden, die Grenzen der angewandten Methoden und Modelle berücksichtigt werden müssen. Zukünftige Forschungen könnten sich auf die Weiterentwicklung der Modelle und die Einbeziehung weiterer Datenquellen konzentrieren. Ebenso ist eine Auseinandersetzung mit ethischen Aspekten und Datenschutzfragen im Kontext der KI im Marketing von hoher Bedeutung.

Der Ausblick auf zukünftige Entwicklungen im Bereich des Marketings deutet darauf hin, dass die fortschreitende Evolution von KI-Technologien zu noch präziseren und personalisierten Marketingstrategien führen wird. Die Fähigkeit, auf Marktveränderungen in Echtzeit zu reagieren und Marketingkampagnen dynamisch anzupassen, wird zunehmend zu einem entscheidenden Faktor für den Unternehmenserfolg.

Abschließend lässt sich festhalten, dass die Implementierung von KI im Marketing erhebliche Vorteile mit sich bringt, aber auch Herausforderungen und Verantwortungen birgt. Die in dieser Arbeit entwickelten Ansätze und Modelle bieten wertvolle Ausgangspunkte für weiterführende Forschungen und praktische Anwendungen in diesem dynamischen und innovativen Feld.

Literaturverzeichnis

[AwKh2015] AWAD, Mariette; KHANNA, Rahul, 2015. Support Vector Machines for Classification [online]. In: M. AWAD und R. KHANNA, Hg. *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers.* Berkeley, CA: Apress, S. 39-66. ISBN 978-1-4302-5990-9. Verfügbar unter: https://link.springer.com/chapter/10.1007/978-1-4302-5990-9_3, zuletzt geprüft am: 11.11.2023.

[BeDr2016] BELGIU, Mariana; DRĂGUŢ, Lucian: Random forest in remote sensing: A review of applications and future directions [online]. *ISPRS Journal of Photogrammetry and Remote Sensing*, **114**, 24-31, 2016. Verfügbar unter: doi:10.1016/j.isprsjprs.2016.01.011

[Brow2020] BROWNLEE, J., 2020. *Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and Data Transforms in Python:* Machine Learning Mastery.

[DaJo2021] DAHOUDA, Mwamba K.; JOE, Inwhee: A Deep-Learned Embedding Technique for Categorical Features Encoding [online]. *IEEE Access*, **9**, 114381-114391, 2021. Verfügbar unter: doi:10.1109/ACCESS.2021.3104357

[DeSa2023] DEMIR, Selçuk; SAHIN, Emrehan K.: Application of state-of-the-art machine learning algorithms for slope stability prediction by handling outliers of the dataset [online]. *Earth Science Informatics*, **16**(3), 2497-2509, 2023. Verfügbar unter: doi:10.1007/s12145-023-01059-8

[ErBK2021] ERNAWATI E; BAHARIN S S K; KASMIN F: A review of data mining methods in RFM-based customer segmentation [online]. *Journal of Physics: Conference Series*, **1869**(1), 12085, 2021. Verfügbar unter: doi:10.1088/1742-6596/1869/1/012085

[GB++2016] GHOLAMI, Azadeh; BONAKDARI, Hossein; ZAJI, Amir H.; AJEEL FENJAN, Salma; AKHTARI, Ali A.: Design of modified structure multi-layer perceptron networks based on decision trees for the prediction of flow parameters in 90° open-channel bends [online]. *Engineering Applications of Computational Fluid Mechanics*, **10**(1), 193-208, 2016. Verfügbar unter: doi:10.1080/19942060.2015.1128358

[KiJu2023] KIM, Annie; JUNG, Inkyung: Optimal selection of resampling methods for imbalanced data with high complexity [online]. *PloS one*, **18**(7), 1-18, 2023. Verfügbar unter: doi:10.1371/journal.pone.0288540

[KuCB2021] KULKARNI, Ajay; CHONG, Deri; BATARSEH, Feras A.: Foundations of data imbalance and solutions for a data democracy [online]. *Data Democracy: 1st Edition At the Nexus of Artificial Intelligence, Software Development, and Knowledge Engineering,* 2021. Verfügbar unter: doi:10.48550/arXiv.2108.00071

[LC++2017] LI, Jundong; CHENG, Kewei; WANG, Suhang; MORSTATTER, Fred; TREVINO, Robert P.; TANG, Jiliang; LIU, Huan: Feature Selection: A Data Perspective [online]. *ACM Computing Surveys*, **50**(6), 1-45, 2017. Verfügbar unter: doi:10.1145/3136625

[MaPu2022] MANGKUNEGARA, Iis S.; PURWONO, Purwono: Analysis of DNA Sequence Classification Using SVM Model with Hyperparameter Tuning Grid Search CV [online]. *2022 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom)*, 427-432, 2022. Verfügbar unter: doi:10.1109/CyberneticsCom55287.2022.9865624

[Marr2022] MARR, Bernard, 9. September 2022: Artificial Intelligence And The Future Of Marketing [online]. *Forbes*, **2022.** Verfügbar unter: https://www.forbes.com/sites/bernardmarr/2022/09/09/artificial-intelligence-and-the-future-of-marketing, zuletzt geprüft am: 01.12.2023.

[Nuzz2016] NUZZO, Regina L.: The Box Plots Alternative for Visualizing Quantitative Data [online]. *PM & R: the journal of injury, function, and rehabilitation*, **8**(3), 268-272, 2016. Verfügbar unter: doi:10.1016/j.pmrj.2016.02.001

[OtAl2021] OTHMAN, Sameera A.; ALI, Haithem T.M.: Improvement of the Nonparametric Estimation of Functional Stationary Time Series Using Yeo-Johnson Transformation with Application to Temperature Curves [online]. *Advances in Mathematical Physics*, **2021**, 1-6, 2021. Verfügbar unter: doi:10.1155/2021/6676400

[Patr2002] PATRICIAN, Patricia A.: Multiple imputation for missing data [online]. *Research in nursing & health*, **25**(1), 76-84, 2002. Verfügbar unter: doi:10.1002/nur.10015

[Pric2017] PRICEWATERHOUSECOOPERS: *PwC's Global Artificial Intelligence Study: Sizing the prize* [online], 2017. Verfügbar unter: https://www.pwc.com/gx/en/issues/data-and-analytics/publications/artificial-intelligence-study.html, zuletzt geprüft am: 01.12.2023

[RaRo2021] RAYMAEKERS, Jakob; ROUSSEEUW, Peter J.: Transforming variables to central normality [online]. *Machine Learning,* 2021. Verfügbar unter: doi:10.1007/s10994-021-05960-5

[Sald2020] SALDANHA, Rodolfo: *Marketing Campaign* [online]. Verfügbar unter: https://www.kaggle.com/datasets/rodsaldanha/arketing-campaign/data, zuletzt geprüft am: 07.12.2023

[Shre2020] SHRESTHA, Noora: Detecting Multicollinearity in Regression Analysis [online]. *American Journal of Applied Mathematics and Statistics*, **8**(2), 39-42, 2020. Verfügbar unter: doi:10.12691/ajams-8-2-1

[SSPP2019] SAHOO, Kabita; SAMAL, Abhaya K.; PRAMANIK, Jitendra; PANI, Subhendu K.: Exploratory Data Analysis using Python [online]. *International Journal of Innovative Technology and Exploring Engineering*, **8**(12), 4727-4735, 2019. Verfügbar unter: doi:10.35940/ijitee.L3591.1081219

[Stol2011] STOLTZFUS, Jill C.: Logistic Regression: A Brief Primer [online]. *Academic Emergency Medicine*, **18**(10), 1099-1104, 2011. Verfügbar unter: doi:10.1111/j.1553-2712.2011.01185.x

[Wutt2023] WUTTKE, Laurenz, 2023: RFM-Analyse: Marketing optimieren durch intelligente Segmentierung [online]. *datasolut GmbH.* Verfügbar unter: https://datasolut.com/rfm-analyse, zuletzt geprüft am: 28.11.2023.

[ZQ++2018] ZHANG, Dahai; QIAN, Liyang; MAO, Baijin; HUANG, Can; HUANG, Bin; SI, Yulin: A Data-Driven Design for Fault Detection of Wind Turbines Using Random Forests and XGboost [online]. *IEEE Access*, **6**, 21020-21031, 2018. Verfügbar unter: doi:10.1109/ACCESS.2018.2818678

Eidesstattliche Versicherung:

Hiermit versichere ich, die vorliegende Arbeit selbstständig und unter ausschließlicher Verwendung der angegebenen Literatur und Hilfsmittel erstellt zu haben.

Die Arbeit wurde bisher in gleicher oder ähnlicher Form keiner anderen Prüfungsbehörde vorgelegt und auch nicht veröffentlicht.

Weiterhin bin ich damit einverstanden, dass meine Arbeit der THM-internen Plagiatsprüfung unterzogen wird.

Frankfurt am Main, 07.12.2023	
Ort, Datum	Unterschrift