Klassifizierung für E-Commerce Nutzergruppen auf Basis von Data Mining unter Verwendung von künstlicher Intelligenz.

Untersuchung der Nutzerklassifizierung durch Datenanalyse

ENTWURF

vorgelegte Masterarbeit
zum Erlangen des akademischen Grades
Master of Science (M.Sc.)

eingereicht von: Wilfried Pahl
Matrikelnummer: 901932
Studiengang: Online Medieninformatik
Berliner Hochschule für Technik

Betreuender Prüfer Prof. Dr. S. Edlich Berliner Hochschule für Technik
Zweitgutachter noch nicht bekannt auch von einer Hochschule

Temmen-Ringenwalde, der 9. Januar 2023

Stichworte

Data-Mining, Big Data, künstliche Intelligenz, Clusterung, Nutzergruppen, k-Means-Algorithm.

Kurzzusammenfassung

Hier kommt das Abstract auf Deutsch.

ENTWURF

Keywords

Data-Mining, Big Data, künstliche Intelligenz, Clusterung, Nutzergruppen, k-Means-Algorithm.

Abstract

Here comes later the abstract on english.

ENTWURF

Danksagung

Hier sage ich auch mal zu irgendjemand Danke.

ENTWURF

Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig verfasst habe, dass ich sie zuvor an keiner anderen Hochschule und in keinem anderen Studiengang als Prüfungsleistung eingereicht habe und dass ich keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Alle Stellen der Arbeit, die wörtlich oder sinngemäß aus Veröffentlichungen oder aus anderweitigen fremden Äußerungen entnommen wurden, sind als solche kenntlich gemacht.

Datum Unterschrift

ENTWURF

ENTWURF

Inhaltsverzeichnis

ENTWURF

ENTWURF

Abbildungsverzeichnis

ENTWURF

ENTWURF

Tabellenverzeichnis

ENTWURF

ENTWURF

Kapitel 1

Einführung

Das Volumen der bestellten Waren im E-Commerce Bereich wächst stetig und durch die Corona-Krise erfuhr dieser Prozess eine Beschleunigung. Die meisten Umsätze verzeichnen virtuelle Marktplätze, wie Amazon und Ebay, aber auch kleine Onlineshops können durchaus bestehen.

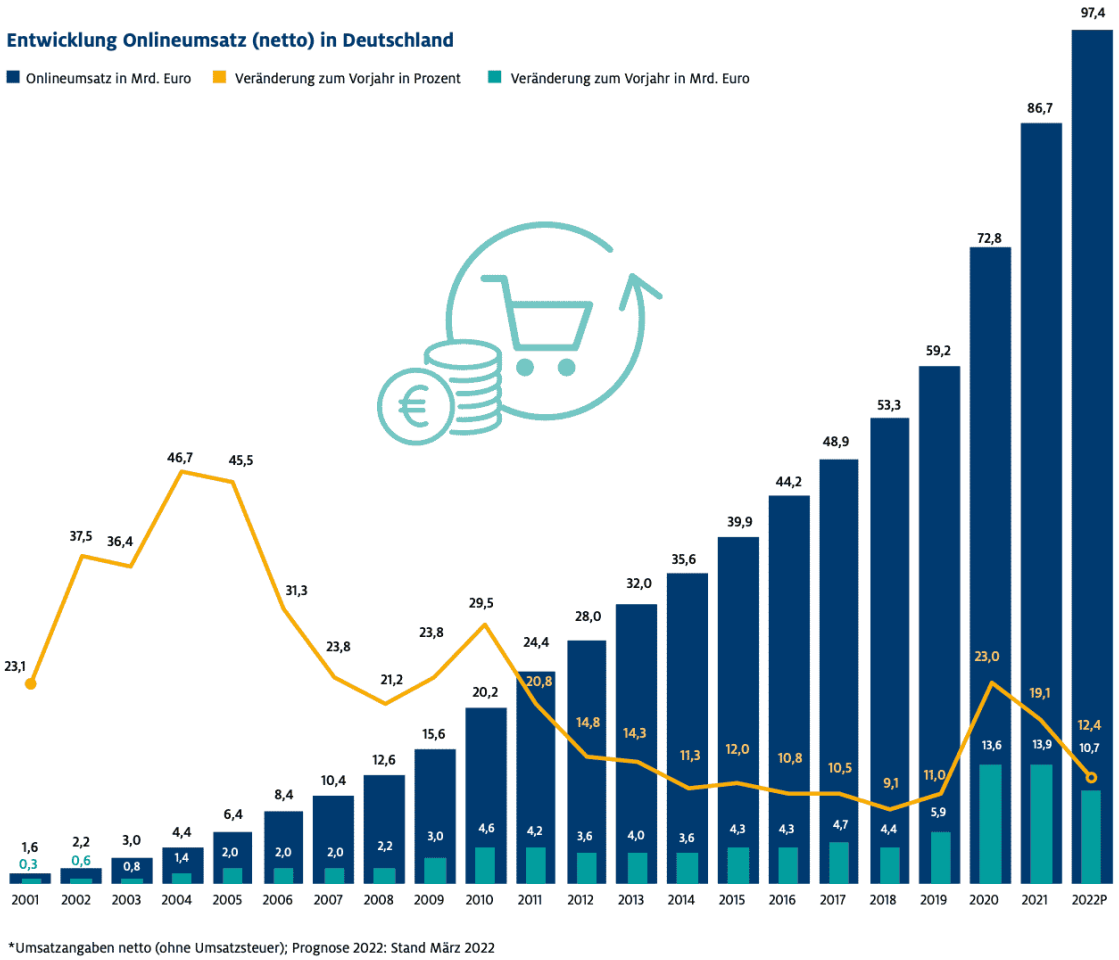

Abbildung 1.1: Umsatzentwicklung E-Commerce von 2001-2022

Mithilfe von Machine Learning können aus vorhandenen Daten, Nutzergruppe bestimmt werden, mit deren Hilfe besser Angebote für die Nutzergruppen erstellt und ausgespielt werden können.

Allgemein benennt [talend.com] folgende drei Vorteile für Unternehmen der durch den Einsatz von maschinellen Lernen entstehen.

1. Future-Proof: es ist eine Technologie der Zukunft, die in den kommenden Jahren weiter entwickelt wird.
2. Effizienz: durch die Automatisierung und Klassifizierung lassen sich große Datenmengen effizienter analysieren. Dadurch verbessern sich interne und externe Unternehmensprozesse.
3. Fördert Konzentration: ML übernimmt repetitive Aufgaben, so bleibt Mitarbeitern für anspruchsvollere Arbeiten.

1.1 Motivation

Im E-Commerce nutzen Onlineshop Betreiber Kundenklassifizierungen, um Kunden einzuteilen und ihnen besser Angebote unterbreiten zu können. Oft sind diese Klassifizierungen in den ersten Schritten bei der Shop-Planung oder bei einem Relaunch erstellt wurden, basierend auf einer erwarteten und anvisierten Kundengruppe. In vielen Fällen werden keine Kundenklassifizierungen erstellt. Es wird unterstellt das alle Menschen zur Zielgruppe gehören.

1.1.1 Ausgangslage

Im Rahmen dieser Arbeit soll ein kleiner Onlineshop eines Verlages aus dem Nordosten Deutschlands untersucht werden. Über den Shop werden hauptsächlich Print-Produkte wie Bücher und Zeitschriften. Eine vorherige Planung und Analyse der Nutzergruppen und ihrer Bedürfnisse sind einigen Mitarbeitern, die den Shop betreuen, nicht wichtig.

Folgende Aussagen trafen Mitarbeitern, die unmittelbar mit den Onlineshop Arbeiten und diesen betreuen.„Für den Onlineshop benötigen wir keine Nutzergruppen, dafür haben wir die verschiedene Menüpunkte. Nutzergruppen werden erst interessant, wenn wir auf verschiedenen Kanäle Werbung schalten.“ Aussage einer Mitarbeiterin für Kundenmanagement. „Zielgruppendefinition ist wichtig, aber für den Shop sollte keine angefertigt werden.“ Aussage einer Mitarbeiterin des Webdesigner Teams.

1.1.2 Vorteile durch die Vorhersage

Einer der wichtigsten Vorteile einer durch künstlicher Intelligenz vorhergesagten Nutzergruppenklassifizierung wäre das Ausschließen des menschlichen Einwirkens. Das Erhöht die Genauigkeit der Entscheidungen und es passieren weniger Fehler. Des Weiteren erkennt künstliche Intelligenz versteckte Muster aus tausenden und mehr Datensätzen. Ein weiterer Vorteil ist die Möglichkeit einer zeitnahen Anpassung der Verkaufsprozesse auf der Webseite und eine kontinuierliche Optimierung des Onlineauftritts.

1.2 Ziel der Arbeit

In dieser Arbeit ist die Prüfung, ob eine mögliche Vorhersage der Nutzergruppenklassifizierung durch künstliche Intelligenz möglich ist und ob diese Vorhersage Vorteile

ENTWURF

gegenüber einer klassischen Klassifizierung, beispielsweise durch Personas ist. Für die Vorhersage werden vorhandene Daten aus dem Shop und verschiedene externe Quellen verwendet.

Durch die Zielsetzung lässt sich folgende Forschungsfrage und dazugehörige Unterfragen formulieren:

„Ist die Erstellung eine Nutzergruppenklassifizierung mit Machine-Learning besser als eine klassische Analyse der Nutzergruppen?“

- Welche Systeme und Algorithmen eignen sich am besten für die Klassifizierung?
- Sind die klassischen Nutzergruppen (z.B. mit soziodemografischen Merkmalen) aus E-Commerce noch relevant oder stehen andere Merkmale im Vordergrund?

1.3 Inhaltlicher Aufbau der Arbeit

Diesen Teil schreibe ich während der Fertigstellung der einzelnen Kapitel.

ENTWURF

ENTWURF

Kapitel 2

Grundlagen

In diesem Kapitel wird auf die wichtigsten grundlegenden Begriffe, Methodiken und Verfahren eingegangen.

2.1 Knowledge Discovery in Data (-base)

Dies hat das Ziel aus vorhandenen meist großen Datenbeständen, fachliche Zusammenhänge zu erkennen. Zu den Teilschritten des KDD Prozesses gehören 1. Bereitstellung von Hintergrundwissen, 2. Definition der Ziele, 3. Datenauswahl, 4. Datenbereinigung, 5. Datenreduktion, 6. Auswahl eines Modells, 7. Data-Mining, die eigentliche Datenanalyse, 8. Interpretation der gewonnenen Erkenntnisse.

2.2 Künstliche Intelligenz

Eine einheitliche gemeingültige Definition von künstlicher Intelligenz zu geben ist nicht einfach. Zuvor muss Intelligenz definiert werden. Aber was ist Intelligenz? In der Literatur werden kognitive Fähigkeiten oft mit Intelligenz in Verbindung gebracht. In der Definition von künstlicher Intelligenz gibt es schwache oder enge künstliche Intelligenz, die auf die Lösung bestimmter Aufgaben beschränkt ist und menschliche Intelligenz nicht imitieren kann. Starke oder allgemeine künstliche Intelligenz hingegen ist in der Lage die kognitiven Fähigkeiten des Menschen zu erzielen.

2.3 Machine Learning

Maschinelles Lernen (eng. Machine Learning) ist nach [https://datasolut.com/was-ist-machine-learning] ein Teilbereich der künstlichen Intelligenz, der System in die Lage versetzt, automatisch aus Erfahrungen (Daten) zu lernen und sich zu verbessern. Aufgaben die Machine Learning erledigen kann, ist Berechnung von Wahrscheinlichkeiten für bestimmte Ereignisse, Erkennen von Gruppen und Clustern in Datensätzen, Erkennen von Zusammenhängen in Sequenzen, Reduktion von Dimensionen ohne großen Informationsverlust und Optimierung von Geschäftsprozessen.

ENTWURF

2.4 Big Data

Eine Abgrenzung des Begriff Big Data ist nicht eindeutig, da der Begriff sehr heterogen verwendet wird. Somit gibt es viele Definitionen vom Begriff Big Data.

Big Data beschäftigt sich nach [DataSolut3] und [Oracle] mit dem Sammeln, Verarbeiten und Zusammenführen von großen Datenmengen. Um diese Daten für die Entscheidungsfindung und Prozessautomatisierung zu verwenden. Dabei stammen die Daten aus den unterschiedlichsten Quellen, aus verschiedenen Datenbanken oder auch direkt aus Programmen. Als Datenquellen können folgende infrage kommen:

- Internetnutzung
- Social Media
- Geo-Tracking
- Cloud Computing
- Vitaldaten-Messung
- Media-Streaming

Diese Daten können strukturiert, aber auch unstrukturiert vorliegen. [Gratner] beschrieb Big Data anhand von den ,,4 V's". Mit der Zeit wurde es um ein ,,V" erweitert. Diese Beschreibung wird in unterschiedlichen Publikationen aufgegriffen, unter anderen auch in [Oracle].

Volume (Volumen)

Immer größere Datenmengen müssen Verarbeitet werden. Durch die stetig zunehmende Digitalisierung in immer mehr Lebensbereichen wächst die erzeugte Datenmenge pro Zeiteinheit immer mehr an. So werden großen Datenmengen nicht nur durch die oben genannten Quellen erzeugt, sondern auch z. B. durch Gerätesensoren. Hierbei können etliche Terabytes oder hunderte Petabytes an Daten anfallen. Wie die Abbildung 2.1 [Statistika] zeigt, wird das Datenvolumen im Jahr 2025 auf 181 Zettabyte vorhergesagt.

Variety (Vielfalt)

Durch die unterschiedlichen Bereiche, in denen die Datenmengen entstehen sind, diese sehr unterschiedlich und zu meist unstrukturiert. Oft liegen diese in relationalen Datenbanken und können dort nicht ausgewertet werden. Neben Texten liegen die Daten in Bildern und Videos vor die Analyse erfolgt durch Machine Learning Algorithmen .

Velocity (Geschwindigkeit)

Mit der Entwicklung der Technik produzieren Softwaresysteme mit einer höheren Geschwindigkeit mehr Daten. Bei vielen Produkten fließen die Daten nicht auf eine Festplatte, sondern werden direkt im Speicher verarbeitet. Solche Produkte arbeiten in Echzeit oder beinahe in Echtzeit. Deren Verarbeitung in immer kürzerer Zeit erfolgt. Für Unternehmen und verschiedenen Use Cases kann die Verarbeitung in Echtzeit einen erheblichen Wettbewerbsvorteil bedeuten.

Veracity (Wahrhaftigkeit)

ENTWURF

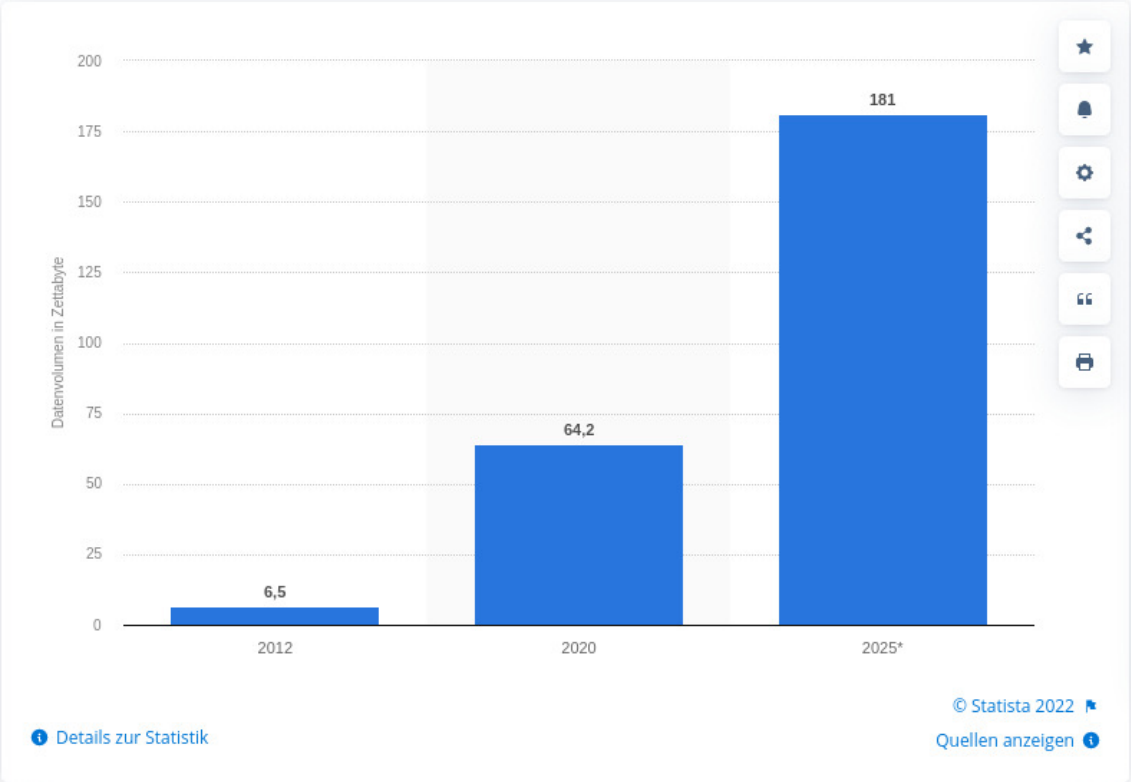

Abbildung 2.1: Umsatzprognose E-Commerce bis 2025

Da die Daten oft aus Quellen kommen, deren Wahrheitsgehalt nicht sicher ist und die Daten in nicht geeigneter Qualität vorliegen, können diese nicht ohne eine aufwendige Nachbearbeitung eingesetzt werden.

Value (Mehrwert)

Durch die Verknüpfung der Daten, die beim Einsatz der Techniken des Machine Learning entstehen, ist dieser Mehrwert eines der wichtigsten ,,V" bei Big Data. Ohne diesen Mehrwert würde Big Data keinen Sinn ergeben.

2.4.1 Technologien

Rund um das Thema Big Data, haben sich verschiedene Technologien entwickelt, die Ansätze für die Verarbeitung von großen Datenmengen liefern. Nachfolgende sind Open Source Produkte stellvertretend einige genannt.

Apache Spark

Es ist nach eigenen Angaben [Spark] eine mehrsprachige Engine zur Ausführung von Data Engineering, Data Science und maschinelles Lernen auf Single-Node-Maschinen oder Clustern.

Apache Hadoop

Hadoop ist nach [Hadoop] ein Framework das mit einfachen Programmiermodellen, welches eine verteilte Verarbeitung von großen Datenmengen anbietet, das von einzelnen Server auf mehrere tausend skaliert werden kann.

ENTWURF

Apache Cassandra

Nach eigener Beschreibung [Cassandra], ist Cassandra ist ein skalierbares und hochverfügbares verteiltes Datenbanksystem. Es basiert auf NoSQL, ist Open-Source und kann ebenfalls auf einzelnen Servern oder in der Cloud eingesetzt werden.

2.5 Data-Mining

Data-Mining die systematische Anwendung von statischen Methoden auf große Datenbestände, um neue Querverbindungen zu erkennen. Data-Mining ist ein Teilprozess KDD Prozesses. Mit Data-Mining findet ein Informationsgewinn oder -erweiterung, aus den Big Data statt. Hier werden Algorithmen aus dem Bereich ML angewandt.

2.6 Clustering

ENTWURF

Kapitel 3

Related Work

3.1 Verwandte Arbeiten

Mit weiteren Anwendungsfällen beschäftigt sich das ,,Fraunhofer Institut BIG DATA" in ihrem Paper [Big Data Frauenhofer].

Um den Einsatz künstlicher Intelligenz im werte-orientierten Marketing zu bewerten, befassten sich in [EConster] einige Professoren des ,,Leibniz-Informationszentrum Wirtschaft" mit diesem Thema.

Die Masterarbeit [Bitstream] von Eduard Weigandt befasst sich mit der Personalisierung im E-Commerce basierend auf Data-Mining. Interessante Grundlagen zum Einstieg in E-Commerce und künstlicher Intelligenz sind auf [Eqop] zu finden. Die Firma Kobold AI befasst sich in ihrem Artikel [Kobold.ai] ,,Optimale Segmentierung von Bestandskunden durch KI" und erläutert Methoden zur Clustering von Bestandskunden. ,,Datasolut" ist ein weiteres Unternehmen das sich in [Datasolut1] mit Kundenklassifizierung, Clusteranalyse und maschinellem Lernen befasst. Ebenfalls von ,,Datasolut" ist der Artikel [Datasolut2] in dem erfolgreiche Anwendungen und Beispiele zum Thema künstlicher Intelligenz im E-Commerce aufgezeigt werden.

ENTWURF

ENTWURF

Kapitel 4

Explorative Datenanalyse

Das Wichtigste für die künstliche Intelligenz sind die Daten. Beim Erheben der Daten für die künstliche Intelligenz sollten die Datensätze nicht aus kontrollierten Umgebungen stammen die mit festgelegten Rahmenbedingungen arbeiten. Sonst können die Algorithmen unter realen Bedingungen schlecht abschneiden, da hier andere Daten von den Trainingsdaten zu stark abweichen. Daher sollten bereist vorhandenen realen Daten verwendet werden. Daher sind die Modelle mit bereits vorhandenen realen Daten zu trainieren.

Die explorative Datenanalyse (EDA) wird verwendet, um Daten zu analysieren und ihre Merkmale zusammenzufassen und um besseres Verständnis für die Datensätze zu bekommen. Dies hilft dabei, um herauszufinden wie die Daten am besten verarbeitet werden können. Dabei hilft die EDA Muster oder Fehler und Anomalien zu finden, sowie Hypothesen testen und Annahmen zu überprüfen.

Sind die Daten analysiert können sie mithilfe, beispielsweise von maschinelles Lernen verarbeitet werden.

4.1 Arten der EDA

Primär unterschiedet [ibm.com] vier Arten der EDA.

Univariat, nicht-grafisch
Untersucht nur eine Variable und ist somit die einfachste Form der Datenanalyse. Hierbei geht es nicht um Ursachen Forschung oder Finden von Beziehungen, sondern das Beschreiben der Daten und Muster zu finden.

Univariat, grafisch
In dieser Art gibt es beispielsweise die Möglichkeit eine Variable beispielsweise in einem

- Stamm-Blatt-Kurvendiagramm[1],
- Histogramm[2] oder
- Box-Diagramm[3] dargestellt.

[1]Stamm-Blatt-Kurvendiagramm zeigt die alle Datenwerte und die Form der Verteilung.
[2]Histogramm mit Balken wird die Häufigkeit oder Anteil der Fälle für einen Wertebereich angezeigt.
[3]Box-Diagramm fünfstellige Zusammenfassung von 1. Minimum, 2. erstes Quartils, 3. Median, 4. dritten Quartils, 5. Maximum.

ENTWURF

Multivariat, nicht-grafisch
Diese Daten bestehen aus mehreren Variablen. Sie zeigen allgemeine Beziehungen zwischen zwei oder mehreren Variablen durch Kreuztabellen oder Statistiken.

Multivariat, grafisch
Zeigen grafisch die Beziehungen zwischen ein oder mehreren Variablen. Häufig werden zur Darstellung Streu-[4], Lauf-[5], Blasendiagramme[6], Multivariate Diagramme[7] und Heat-Maps[8] verwendet.

4.2 Kundengruppen

Die Kundengruppe muss aus einer homogenen Menge an Kunden bestehen, die gleich Arttribute besitzen. Neben der Unterteilung der Kundengruppen nach dem DiSG Modells werden die Kunden nach ihrer Vorlieben an Artikeln eingeteilt.

4.2.1 DiSG Modell

Das DiSG Modell beschreibt die vier Grundtypen von Persönlichkeiten der Nutzer. Diese sind **D**ominant, **i**nitiativ, **S**tetig und **G**ewissenhaft. Jeder der vier Grundtypen weist beim Besuch eines Onlineshops andere Verhaltensmuster auf.

Die Grundlagen für dieses Modell beruhen auf der Arbeit von William Moulton Marston aus dem Jahr 1928.
Der *Dominante*, hat konkrete Erwartungen an ein Produkt. Daher sollte eine Produktbeschreibung nicht nur klar und deutlich sein, sonder ebenso dessen Nutzen, welche Funktionen es hat, warum es so gut funktioniert und die wie es die Probleme des Kunden löst.

Der *Initiative* Kundentyp ist begeisterungsfähig, extrovertiert und optimistisch. Da dieser Typ eher die positiven Eigenschaften eines Produktes sieht, lohnt es sich diesen Kundentyp Rezensionen schreiben zu lassen. Er legt Wert auf eine Wertbeschreibung des Produktes. Dies kann mit Storytelling erfolgen.

Der *Stetige* kann eine harte Nuss für Onlinehändler sein, da dieser Kundentyp sehr viel Wert auf Produktbeschreibung legt, die dessen Bedürfnisse und Ziele erfasst. Andererseits ändern sich die Bedürfnisse dieses Typs nicht über Nacht. Somit kann er zu einem regelmäßigen kaufenden Stammkunden werden. Die Produktbeschreibung sollte die Unique Selling Point enthalten und herausstellen welche Ziele das Produkt, wie unterstützt.

Der *Gewissenhafte* geht systematisch vor und analysiert seine Erkenntnisse zum Produkt. Ist eher reserviert und zurückhaltend und schreibt somit weniger Rezensionen. Dieser Typ kann mit Daten und Fakten überzeugt werden. Dies sollte mit Anwendungsbeispielen und Studienergebnissen untermauert werden.

ENTWURF

[4] Streudiagramm, stellen Datenpunkte auf einer horizontalen und einer vertikalen Achse, die Abhängigkeit einer Variablen zu einer anderen zu zeigen.
[5] Laufdiagramm ist ein Liniendiagramm von Daten die über die Zeit aufgetragen werden.
[6] Blasendiagramm ist eine Datenvisualisierung mittels Kreisen.
[7] Multivariate Diagramm ist eine grafische Darstellung zwischen Faktoren und einer Antwort.
[8] Heat-Map hierbei werden die Daten durch Farben dargestellt.

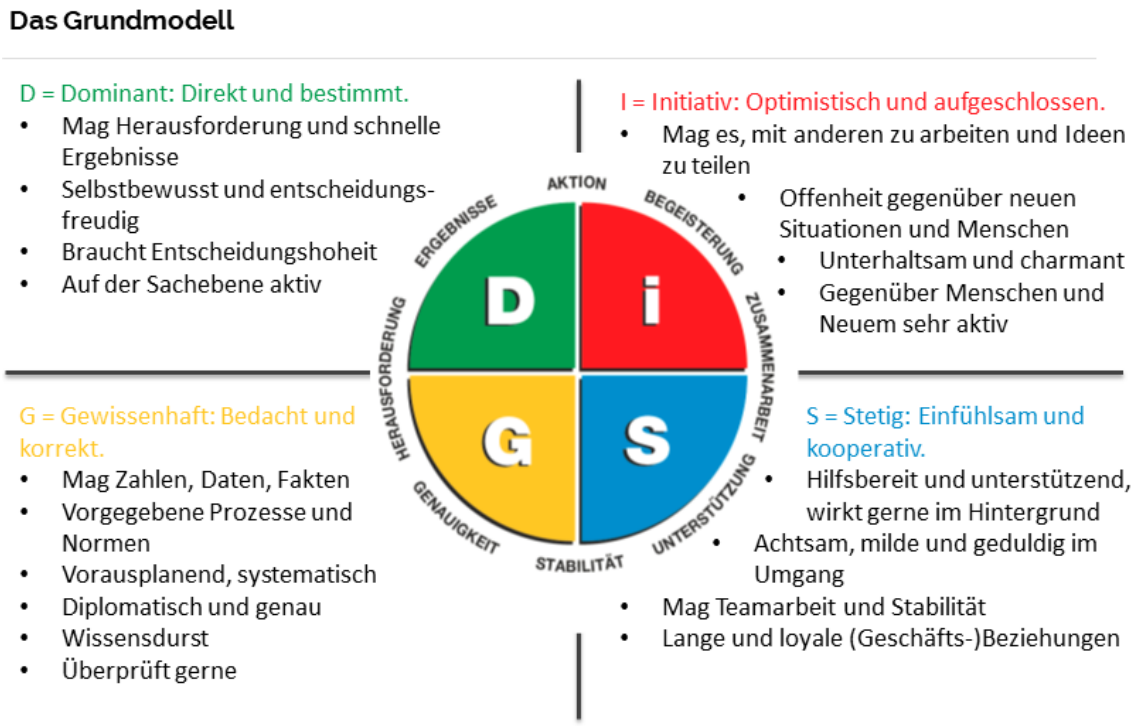

Abbildung 4.1: DiSG Übersicht https://www.disg-modell.de

4.3 Daten für die Onlineshop-Nutzeranalyse

Um das Modell zu trainieren kommen bereits erhobene, interne und offene Daten zu Anwendung. Dabei werden die Daten konsolidiert und deren Integrität geprüft.

Daten aus dem Shop

Diese Daten sind die ausführlichsten Daten. Sie enthalten unter anderem die Nutzerdaten, Daten zu den ab- und nicht abgeschlossenen Käufen. Dazu gehören u.a. die Höhe der Verkäufe, welche und wie viele Produkte und Zeitpunkt des Kaufabschlusses.

Daten von Google Analytics

Hier handelt es sich um Nutzerdaten, die deren Verhalten und Bewegung innerhalb der Webseite getrackt wurden.

interne Daten

Diese Daten werden verwendet, um Zeitpunkte berücksichtigt, bei denen Kunden aktiv mit bestimmten Produkten beworben wurden. Dies soll verhindert das in den Ergebnissen Verzerrungen zu bestimmten Produkten entstehen.

Offene Daten

Für die Auswertung werden sogenannte ,,Open Data" verwendet, um saisonale und Feiertage zu berücksichtigen.

Aus den gesammelten Daten, die das Nutzer- und Kaufverhalten repräsentieren, soll die Klassifizierung der Nutzer erfolgen. Für die Ermittlung der Klassifizierung wird bei der Berechnung die Daten verwendet, bei denen ein Nutzer einen Kauf abgeschlossen hat. Die folgenden Daten finden Berücksichtigung,

- Warenkorbhöhe

ENTWU

- Gekauften Artikel pro Warenkorb
- Anzahl der zuvor besuchten Seiten
- Einstiegsseite
- Datum, aus den Open Data

Nun erfolgt die Vergleichsberechnung mittel FP-Growth-Algorithmus und eines neuronalen Netzwerks.

ENTWURF

Kapitel 5

Maschinelles Lernen

Das maschinelle Lernen ist ein Teilgebiet der künstlichen Intelligenz und übernimmt Aufgaben die typischerweise menschliche Intelligenz erfordern. Sie soll dabei helfen Muster und Gesetzmäßigkeiten in Datensätzen zu erkennen. Aus vorhandenen Daten wird durch Algorithmen künstliches Wissen generiert.

5.1 Methoden maschinellen Lernens

Maschinelles Lernen lässt sich nach [talend.com] in vier Methoden unterteilen. Die in den folgenden Kapiteln näher betrachtet werden.

5.1.1 Überwachtes Lernen

Bei überwachtem Lernen erhält ein Computer strukturierte Inputs und gewünschte Ergebnisse. Nun muss der Computer Wege finden mit Inputs, um diese Ergebnisse zu erreichen, d.h der Algorithmus versucht eine vorhersage Funktion zu entwickeln. Die Vorhersagen über die unbekannten oder künftigen Daten wird als prädiktive Modellierung bezeichnet.

Das überwachte Lernen lässt sich in zwei Arten unterteilen.

- Klassifizierung: das Ergebnis ist eine Kategorie, z. B. Gruppenzugehörigkeit
- Regression: hier ist das Ergebnis ein realer Wert, z. B. Produktpreis

Mittels verschiedener Methoden lassen sich die Ergebnisse vorhersagen. Diese können Entscheidungsbäume, Random-Forest-Algorithmus, lineare Regression, Naive-Bayes-Verfahren, usw. Eignet sich für Probleme der Klassifizierung und Regression.

5.1.2 Unüberwachtes Lernen

Bei diesem Lernen sind keine strukturieren Daten vorhanden, eher legen sie unstrukturiert und unbeschriftet vor. Der Algorithmus muss die Strukturen selbst erkennen. Aus erkannten Mustern und Merkmalen, lassen sich weitere Muster und Korrelationen vorhersagen. Ist, gibt zwei Arten von unüberwachten Lernen.

- Clustering: Gruppierung von Daten, weitere lassen sich in bestehende Cluster zuordnen
- Assoziation: Regeln in Daten finden, so werden Daten durch Erfahrung definiert

ENTWURF

Zu diesem Lernen gehören u. a. K-Means, hierarchische Clusteranalyse und Dimensionsreduktion. Mit dieser Methode lassen sich Probleme des Clustering, Dimensionsreduktion und Lernen von Assoziationsregeln.

5.1.3 Teilweise überwachtes Lernen

Es ist ein Hybridverfahren zwischen unüberwachten und überwachten Lernen. Die Rohdaten sind nur teilweise strukturiert und beschriftet. Durch die strukturierten Daten werden die unstrukturierten aufgewertet.

Die strukturierten Daten finden anfangs Verwendung, um diese auf Muster und Korrelationen zu untersuchen. Im Anschluss können diese auf die unstrukturierten angewandt werden. Mit dem teilweise überwachten Lernen können Probleme der Klassifizierung und Regression gelöst werden.

5.1.4 Bestärktes Lernen

Ein Computerprogramm interagiert mit einer dynamischen Umgebung. Beim Ausführen bestimmter Aufgaben erhält das Programm gutes oder schlechtes Feedback für die Aktion. Durch die Belohnung und Bestrafung lernt das Programm die richtigen Verhaltensweisen. Belohnungen werden auf zwei verschiedene Arten vergeben.

- Monte Carlo: Vergabe erfolgt am Ende
- Temporal-Difference-Learning (TD-Learning): Vergabe der Belohnung erfolgt nach jedem Schritt

Als Algorithmen sind hier beispielsweise Q-Learning, Deep Q Network (DQN) und State-Action-Reward-State-Action (SARSA) zu nennen.

5.2 Algorithmen zum Analysieren von Onlineshop Daten

Für die Ermittlung der Klassifizierung werden Daten verwendet, bei denen ein Nutzer einen Kauf abgeschlossen hat. Die folgenden Daten werden berücksichtigt,

-
- Warenkorbhöhe
- Gekauften Artikel pro Warenkorb
- Anzahl der zuvor besuchten Seiten
- Einstiegsseite
- Datum

Diese Daten werden im FP-Growth-Algorithmus und mittels neuronalen Netzwerk ausgewertet und im Anschluss verglichen.

ENTWURF

5.2.1 FP-Growth-Algorithmus

FP-Growth steht für Frequent Pattern (deutsch: häufige Muster). Er ist eine Verbesserung des Apriori Algorithmus da und stellt die Daten in einer Baumstruktur dar, mit den häufigsten Muster. Dabei repräsentiert jedes Blatt des Baumes ein Item eines Itemsets.

Der FP-Growth ist ein schnell arbeitender Algorithmus, der gut skalieren ist und auch versteckte Muster in Datensätzen findet. Der Algorithmus arbeitet effizienter als beispielsweise der Apriori Algorithmus oder die TreeProjection.

Der Nachteil dieses Algorithmus ist, dass er sehr komplex ist und für kleine Datenmengen Algorithmen wie beispielsweise der bereits genannte Apriori Algorithmus ausreicht.

Dieser Algorithmus wird unter anderem von Apache Spark [1] eingesetzt.

5.2.2 Neuronales Netz

Bei neuronalen Netzen handelt es sich um dem Nervensystem nachempfundenes Netz. Die Idee stammt aus der Neurobiologie und soll die Signale weiterleiten. Dadurch soll ein neuronales Netz in die Lage versetzt werden abstrakte Konzepte zu erlernen. Die Übermittlung der Signale von einem Neuron zum nächsten kann bei künstlicher Intelligenz nur bei trainierten Netzen erfolgen.

Einsatzzwecke für neuronale Netze

Prinzipiell gibt es drei Hauptanwendungen für neuronale Netze. Die Abbildung 5.1 zeigt deren Verwendungszwecke.

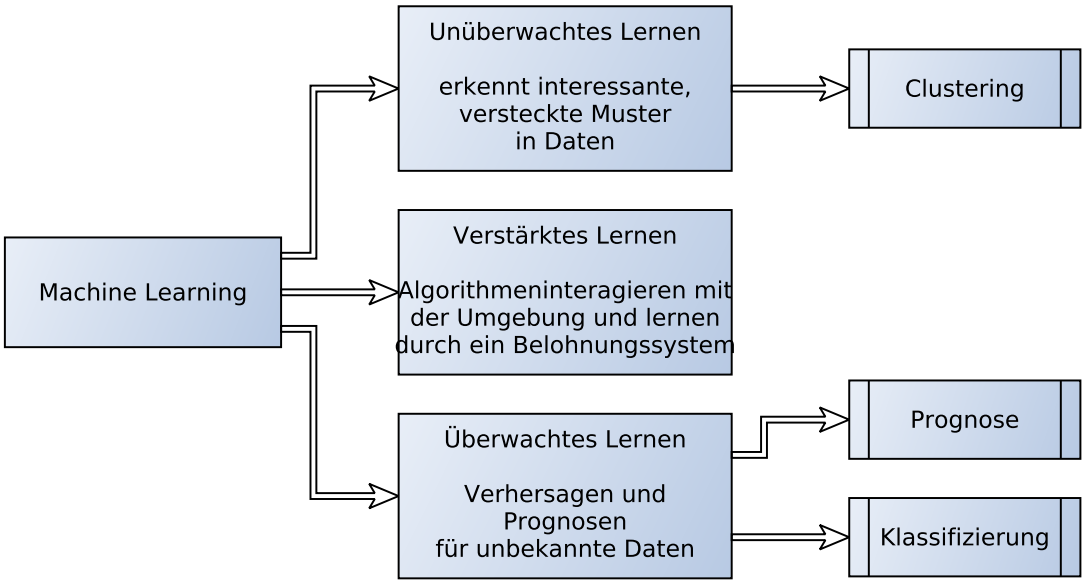

Abbildung 5.1: Übersicht zu neuronalen Netzen und ihrer Verwendung

Classification (Supervised Machine Learning) das Netz wird mit Trainingsdaten trainiert und berechnet für neue Datensätze die Kategorie, in der die Informationen eines Datensatzes passen.

ENTWURF

Regiession (Supervised Machine Learning) ähnlich wie bei der Classification werden diese Netze trainiert, um Vorhersagen zu treffen. Beispielsweise für optimale Preise, Retouren und Umsatzzahlen.

Clustering (Unsupervised Machine Learning) erkennt Gruppen anhand ähnlicher Merkmale. Diese Netze benötigen kein Training, da nur die reinen Informationen innerhalb der Datensätze nutzen, um Muster zuerkennen.

Arten von neuronalen Netzen

Perception sind die einfachsten Netze. Sie bestehen nur aus einem Input-Layer und einer Aktivierungsfunktion im Output-Layer. Hidden-Layer gibt es hierbei nicht.

Feedforward artificial neural network (FFNN) hierbei bewegen sich die Signale in eine Richtung. Diese Netze besitzen aber auch Hidden-Layer.

Deep Learning (DL) diese Netze besitzen mehrere Hidden-Layer Schichten.

Recurrent neural networks (RNN) betrachten, im Vergleich zu den FFNN Netzen auch temporäre Sequenzen, indem sie Informationen zwischenspeichern. Sie gelten deshalb als einer der stärksten neuronalen Netze.

Symmetrically connected neural network sind aufgebaut wie RNN's haben aber eine symmetrische Topologie in der Anordnung und Zuweisung der Gewichte.

Convoluted neural networks (CNN) diese Netze betrachten die Signale nicht linear, sondern betrachte regionale Signale. Dadurch werden Ergebnisse genauer und das Training wird schneller und effizienter.

Radial basis function network (RBFN) sind einfache FFNN Netzwerke die aber die Distanz zwischen den Knoten miteinbeziehen.

Self organizing neural network finden Verwendung, um Strukturen innerhalb von Datensätzen zu finden. Diese Netze arbeiten unsupervised. Moduar neural network kombinieren mehrere neurale Netzwerke, um verschiedene Teilaufgaben zu lösen.

Generative adversarial networks (GAN) generieren aus den zu lernenden Daten neue Daten. Hier werden zwei Netze kombiniert. Ein Netz generiert Daten das andere Netz überprüft diese.

Aufbau des neuronalen Netzes

In dieser Arbeit wird das Feedforward artificial neural network (FFNN) verwendet. Die einzelnen Nervenzellen werden hierbei als Neuronen bezeichnet und sind Funktionen die auf Basis des Inputs und der Gewichtung den Output berechnen. Die Schichten auf denen die Neuronen liegen als Layer. Bei den Layern gibt die Input-Layer, den oder die Hidden-Layer und den Output-Layer. Die Neuronen der einzelnen Layer sind mit Kanten untereinander verbunden, welche mit Gewichtungen, die Ausgaben für das nächste Neuron umrechnet.

Training neuronaler Netze

Zuerst erfolgt die Übergabe der Kundenattribute an die Input-Neuronen. Die Attribute enthalten die Bestelldaten, Daten zum Nutzerverhalten und zusätzlichen externen Daten.

ENTWURF

Zunächst wird das Netz trainiert. Die Initialisierung der Gewichtig muss gut überlegt sein, da es sonst dazu führen kann, das das Netz langsam lernt oder alle Hidden-Layer das gleiche lernen. Nur wird das Netz mit dem Input geladen und iterativ werden die Gewichte der Kanten definiert. Die Signale werden durch die Layer zum Output-Layer geleitet, also zum Ende hin. Dieses Verfahren wird auch als Forward Propagation bezeichnet. Hier sei noch einmal auf die Datenqualität hingewiesen. Ist die Datenqualität schlecht oder stehen nicht genügend Daten zur Verfügung kann das nicht oder schlecht trainiert werden. Sind die Signale am Output-Layer angekommen, definieren Aktivierungsfunktionen (eng. Activation Function), ob das Netz feuert oder nicht.

Nun wird der Output des neuronalen Netzes mit den vorher definierten Ausgaben. Bei einer Abweichung wird dann rückwärts durch das Netz propagiert (eng. Back Propagation) und das Netz lernt daraus. D. h. die Gewichte werden angepasst. Der Vorgang des Inputs laden, Fehler berechnen und Gewichte korrigieren, wird so lange wiederholt, bis ein festgelegtes Abbruchkriterium greift. Nun ist das Netz trainiert.

5.3 Umsetzung der Datenverarbeitung

Erste Tests in Sachen Python und Datenanalyse. Kommt ein Kapitel früher.

5.4 Umsetzung ds FP-Grwoth-Algorithmus

Zur Analyse der zusammen verkauften Artikel und Assoziationsregeln zum Nutzerverhalten aufstellen zu können und einen Merkmals-Vektor zu erstellen, um auf dessen Grundlage die Benutzerklassifizierung zu ermitteln.

Im ersten Schritt werden alle vorbereiteten Datensätze durchlaufen, um die Anzahl der Items zu finden, die für die Klassifizierung der Nutzer erforderlich ist. Im Anschluss werden alle Items erfasst, die den minimalen Support erfüllen. Anschließend werden die Datensätze wiederholt durchsucht und nach den minimalen Support-Items gesucht.

Daraus wird der Häufigkeitsbaum erstellt, in dem alle Items und Datensätze abgebildet werden.

5.5 Umsetzung mittels neuronalen Netz

ENTWURF

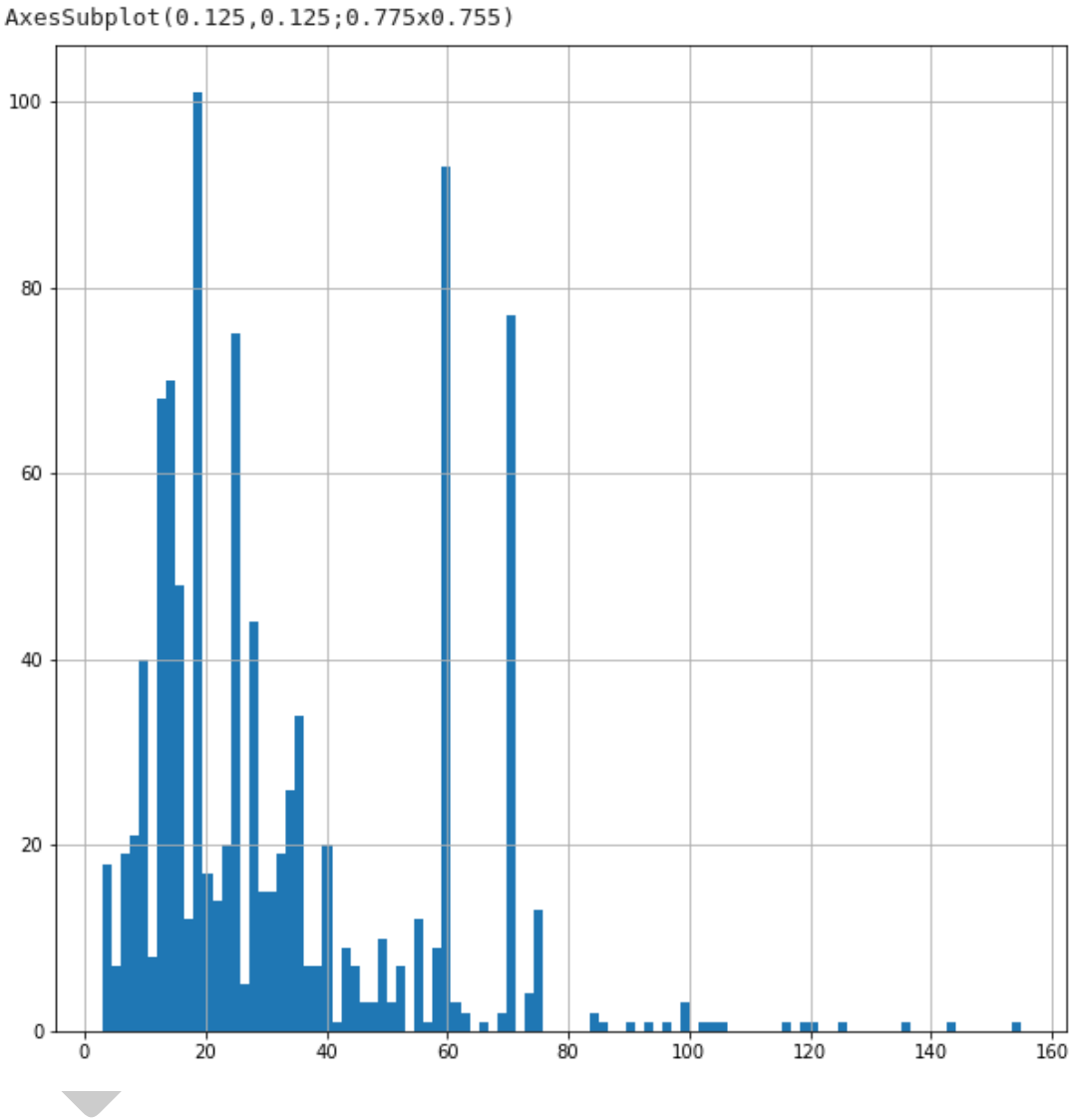

Abbildung 5.2: Anzahl der Bestellung in Abhängigkeit der Warenkorbhöhe

Kapitel 6

Evaluation

In diesem Kapitel werden die Ergebnisse der eigenen Arbeit bewertet und besprochen.

6.1 Aufbau der Umgebung

Hier könnte das CMS erwähnt werden.

6.2 Ergebnisse

6.3 Bewertung und Diskussion

ENTWURF

ENTWURF

Kapitel 7

Conclusion / Lessons Learned

In diesem Kapitel wird beschrieben, wie die Ergebnisse der Arbeit umgesetzt und beispielsweise bestehende Workflows angepasst werden können.

Es wird an die anfangs, im Kapitel 1 gestellten Zielsetzungen angeschlossen und diese beantwortet.

ENTWURF

ENTWURF

Kapitel 8

Future Work

Dieses Kapitel befasst sich mit die Arbeiten, die auf dieser aufbauen könnten.

ENTWURF

ENTWURF

Anhang

ENTWURF

ENTWURF

Literaturverzeichnis

[1] P. Versteegen, "Anlagetrend E-Commerce Aktien: Die besten E-Commerce-Wertpapiere." Finanzwissen, Aug. 2022. [Online]. https://finanzwissen.de/aktien/e-commerce/. [abgerufen am 21.11.2022].

ENTWURF

ENTWURF

Glossar

CMS Nutzerfreudliche Bedienungsoberfläche einer Software. 9

ENTWURF