Duale Hochschule Baden-Württemberg Mannheim

Kundensegmentierung

Wirtschaftsinformatik Schwerpunkt Data Science

Verfasser:	Hannah Laier, Mya-Melissa Jahic, Julia Pfützer
Matrikelnummern:	1010595, 2774085, 7411297
Kurs:	WWI 20 DSB
Modul:	Data Exploration
Studiengangsleiter:	Prof. Dr. Bernhard Drabant
Dozent:	Simon Poll
Bearbeitungszeitraum:	05.07.2022 – 29.07.2022

Inhaltsverzeichnis

Abbildungsverzeichnis

Quelltextverzeichnis

Der Quelltext ist in GitHub unter dem Link:

git@github.com:Julia-Pfuetzer/Data-Exploration_Kundensegmentierung.git

Motivation

Durchführung einer Kundensegmentierung von Supermarktdaten mit Orientierung an vier Businessfragen zur Verfeinerung der Zielgruppe zum exakteren Einsatz von Werbung.

Die Segmentierung der Kundendaten kann relevante Informationen über die Kunden liefern und die Zielgruppen für die jeweiligen Produktgruppen oder einzelnen Produkte genauer eingrenzen und beleuchten. Daraus folgt, dass Werbung genauer geschaltet und platziert werden kann. Es könnte Werbung für jeden geschaltet werden, ohne eine Kundensegmentierung durchzuführen, oder es kann Werbung nur für die relevanten potenziellen Käufer erscheinen und das Geld gespart werden, was die Schaltung der Werbung für jeden auch nicht-potentiellen Kunden gekostet hätte.

Die Segmentierung von Kundendaten eines Supermarktes liefert dem Supermarkt zusätzlich die Möglichkeit seinen Kundenstamm genauer einzuschätzen und so den Fokus des Sortiments anzupassen.

Bei der Kundensegmentierung wird zuerst festgelegt welche Merkmale der Kunden relevant sind und über die Einteilung in ein Kundensegment entscheiden. Danach werden die Kunden anhand der Merkmale in Cluster eingeteilt. Hier wurde dieser Prozess Produktbezogen durchgeführt und auf Basis der vorher ausgewählten Businessfragen und nicht primär Kundenbezogen. [1]

Die eingeteilten Cluster geben Hinweise auf die Zielgruppe der ausgewählten Produktgruppen oder einzelnen Produkte.

[1] Qualtrics (2022): Kundensegmentierung: Definition, Modelle & Ablauf | Qualtrics. Online verfügbar unter https://www.qualtrics.com/de/erlebnismanagement/kunden/kundensegmentierung/, zuletzt aktualisiert am 10.05.2022, zuletzt geprüft am 13.07.2022.

Related Work

Es gibt viele Publikationen zu dem Thema mit unterschiedlichen Verfahren. Der Fokus liegt auf dem Verfahren, dass hier auch verwendet wurde.

- In dem Buch[2] Nutzungsbasierte Kundensegmentierung wird die Clusteranalyse beleuchtet auch in Verbindung mit Marketing. Genauso wie in dieser Arbeit hier.
- Die Daten des Projektes stammen aus Kaggle Customer Segmentation: Clustering[3]
- Informationen zum Hierarchischen Clustern, was zur Cluster Anzahl Bestimmung verwendet wurde. [4]
- Eine weiter Publikation zu diesem Thema ist auf BigData-Insider zu finden[5]
- Codeinspiration [6]

[2] Gossens, Thomas (2000): Nutzungsbasierte Kundensegmentierung. In: Data Mining im praktischen Einsatz: Vieweg+Teubner Verlag, S. 143–180. Online verfügbar unter https://link.springer.com/chapter/10.1007/978-3-322-89950-7_7.

[3] karnikakapoor (2021): Customer Segmentation: Clustering 🛍️🛒🛒. In: *Kaggle*, 08.10.2021. Online verfügbar unter https://www.kaggle.com/code/karnikakapoor/customer-segmentation-clustering, zuletzt geprüft am 13.07.2022.

[4] Sascha (2021): Clusteranalyse zur Kundensegmentierung (mit Beispielen). Online verfügbar unter https://back2marketingschool.com/de/clusteranalyse-beispiele/, zuletzt aktualisiert am 31.12.2021, zuletzt geprüft am 13.07.2022.

[5] Matzer, Michael (2018): Optimale Clusteranalyse und Segmentierung mit dem k-Means-Algorithmus. In: *BigData-Insider*, 19.11.2018. Online verfügbar unter https://www.bigdata-insider.de/optimale-clusteranalyse-und-segmentierung-mit-dem-k-means-algorithmus-a-773713/, zuletzt geprüft am 13.07.2022.

[6] Wuttke, Laurenz (2022): Kundensegmentierung: Beispiel mit K-Means. In: *datasolut GmbH*, 27.04.2022. Online verfügbar unter https://datasolut.com/kundensegmentierung-beispiel-mit-k-means/, zuletzt geprüft am 13.07.2022.

Verwendete Technologien und Bibliotheken

Um eine Clusteranalyse durchzuführen, wurde Python gewählt, da Python anwenderfreundlich ist und auch Zugriffe auf viele Bibliotheken hat. Für dieses Programm wird die Bibliothek pandas, ein schnelles und effizientes DataFrame-Objekt zur Datenmanipulation mit integrierter Indexierung genutzt[7]. Mit numpy können einfach mathematische Rechnungen berechnet werden[8]. SciPy steht für Scientific Python und bietet nützliche Funktionen für Optimierung, Statistik und Signalverarbeitung[9]. Mit mathplotlib.pyplot[10] und seaborn[11] können die Daten grafich dargestellt werden. Wie man an den nachfolgenden Importen sehen kann, wurde nicht immer die komplette Bibliothek importiert, sondern nur einzelne Funktionen. Sklearn ist für supervised oder unsupervised learning zuständig und es kann die Güte eines Vorhersagemodell bewerten[12].

[7] pandas - Python Data Analysis Library (2.11.2021). url: https://pandas.pydata. org/about/index.html.
[8] What is NumPy? NumPy v1.21 Manual (23.06.2021). url: https://numpy.org/ doc/stable/user/whatisnumpy.html
[9] Introduction to SciPy (2022). Online verfügbar unter https://www.w3schools.com/python/scipy/scipy_intro.php, zuletzt aktualisiert am 12.07.2022, zuletzt geprüft am 12.07.2022.
[10] matplotlib.pyplot — Matplotlib 3.5.2 documentation (2022). Online verfügbar unter https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.html, zuletzt aktualisiert am 14.06.2022, zuletzt geprüft am 12.07.2022.
[11] seaborn: statistical data visualization — seaborn 0.11.2 documentation (2022). Online verfügbar unter https://seaborn.pydata.org/, zuletzt aktualisiert am 28.06.2022, zuletzt geprüft am 12.07.2022.
[12] Luber, Stefan (17.09.2018). Was ist Scikit-learn? In: BigData-Insider. url: https : //www.bigdata-insider.de/was-ist-scikit-learn-a-756150/.

Ergebnisse der Kundensegmentierung

Zielsetzung

Nachdem die Daten bereinigt wurden, wurden Zielgruppen definiert, um mit Werbung gezielter eine spezifische Zielgruppe zu erreichen. Dazu wurden die Kundendaten eines Supermarktes zum Einkaufsverhalten untersucht und vier konkrete Fragen zur Analyse aufgestellt:

1. Kaufen Kunden mit Kindern mehr Süßigkeiten?
2. Kaufen Kunden mit höherem Einkommen mehr Fleisch bzw. Fisch?
3. Kaufen Kunden mit einem höheren Bildungsgrad mehr Wein?
4. Präferieren Kunden bestimmte Werbekanäle?

Datenvorverarbeitung

Die verwendeten Daten stammen aus Kaggle. Der Datensatz bestand aus 2240 Zeilen und 29 Spalten. Bevor mit dem Datensatz gearbeitet werden konnte, musste noch Folgendes in Python angepasst werden. Zum einen wurden alle Zeilen mit fehlenden Werten und zum andern auch die für diesen Fall nicht relevante Spalten herausgelöscht. Auch wurde der Beziehungsstatus in 1, für Single, und 2, für in einer Beziehung, umgewandelt, da dies die Analyse später erleichtert. Zudem wurde auch der Bildungsgrad in Zahlen von 1-5 umgewandelt, da sonst der K-Means Algorithmus nicht damit arbeiten kann.

Mit dem angepassten Datensatz konnten nun Korrelationsmatrizen und verschiedene Grafiken zur Analyse erstellt werden.

Datenanalyse

Im Weiteren folgt die Analyse der Daten. Dafür wurden vier Fragen für den wirtschaftlichen Kontext aufgestellt, damit das Unternehmen die nächsten Schritte im Hinblick auf die Zielgruppe und den Werbekanal ableiten kann.

Die erste Frage lautet: Kaufen Kunden mit Kindern mehr Süßigkeiten?

Wenn man einen Blick auf die Korrelationsmatrix in Abbildung 1 wirft, kann man erkennen, dass eine negative Korrelation mit –0,378000 zwischen einem Haushalt mit Kindern und Süßigkeiten besteht. Das heißt Kinder haben keine wirkliche Beziehung zu Süßigkeiten. Um diese Erkenntnis zu Unterstützen wurde der K-Means Algorithmus eingesetzt, welcher Gruppierungen bzw. Cluster erstellt. In der Abbildung 2 kann man eindeutig erkennen, dass

	Kidhome	Teenhome	MntSweetProducts
Kidhome	1.000000	-0.039000	-0.378000
Teenhome	-0.039000	1.000000	-0.163000

Abbildung 1: Korrelationsmatrix - Kinder & Süßigkeiten

Personen bzw. Haushalte ohne oder mit einem Kind am meisten Süßigkeiten kaufen. Ebenso erstellt der Algorithmus zwei Gruppierungen. Die orangefarbige Gruppierung sind Personen mit einem geringeren Einkommen und die blaue Gruppierung andere sind Personen mit einem höheren Einkommen. Somit sind für diesen Fall Haushalte ohne oder mit einem Kind die Zielgruppe.

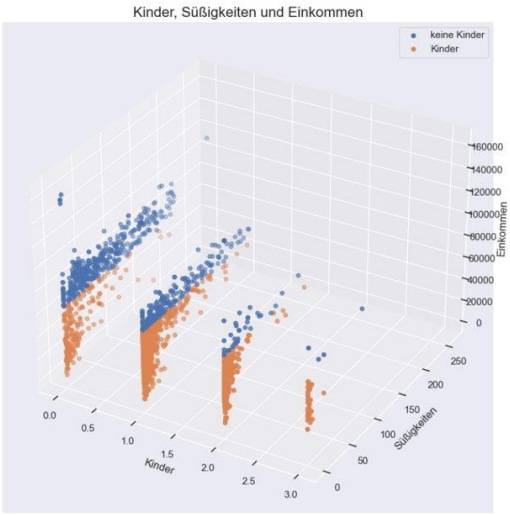

Abbildung 2: K-Means - Kinder & Süßgikeiten

Die nächste Frage, die aufgestellt wurde, lautet: Kaufen Kunden mit höherem Einkommen mehr Fleisch/Fisch?

Hier zeigt die Korrelationsmatrix in Abbildung 3 eine positive Korrelation zwischen dem Einkommen und Fleischprodukten mit 0,585000 sowie Fischprodukten mit 0,439000. Die Abbildung 4 mit Verwendung des Algorithmus zeigt zwei Gruppen, bei der die rote Gruppe eindeutig zeigt, dass Leute mit höherem Einkommen mehr Fleisch- und Fischprodukte kaufen. Leute mit niedrigerem Einkommen, hier die blaue Gruppe, hingegen kaufen geringere Mengen an Fleisch und Fisch ein. Somit zeigen sich hier Personen mit höherem Einkommen als Zielgruppe für Fleisch- und Fischprodukte.

	Income	MntMeatProducts	MntFishProducts
Income	1.000000	0.585000	0.439000
MntMeatProducts	0.585000	1.000000	0.568000
MntFishProducts	0.439000	0.568000	1.000000

Abbildung 3: Korrelationsmatrix - Einkommen und Fleisch-/Fischprodukte

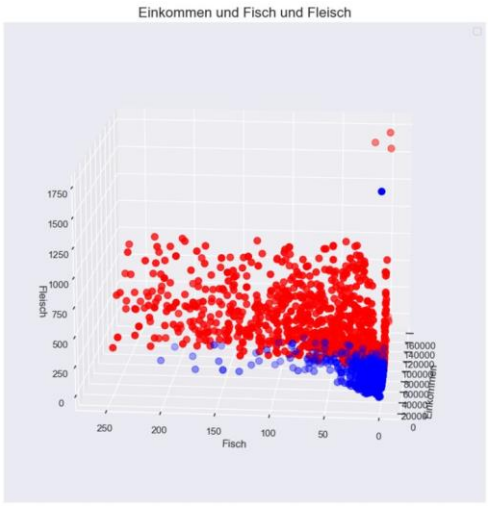

Abbildung 4: K-Means - Einkommen und Fleisch-/Fischprodukte

Eine weitere Frage lautet: Kaufen Kunden mit höherem Bildungsgrad mehr Wein?

Zu dieser Frage stellte die Korrelationsmatrix keine Ergebnisse und der Algorithmus keine zufriedenstellenden Ergebnisse dar, weshalb keine Zielgruppe identifiziert werden konnte.

Anschließend stellte sich die Frage: Präferieren Kunden bestimmte Werbekanäle?

Hier wurden einmal die Werbekanäle mit dem Einkommen und einmal mit dem Geburtsjahr gegenübergestellt. Die Korrelationsmatrix in Abbildung 5 zeigt die beste positive Korrelation zwischen Einkäufen über den Katalog und der gekauften Menge mit 0,780000 sowie dem Einkommen mit 0,697000. Die Abbildung 6 mit Bezug auf das Einkommen zeigt drei Gruppierungen. Zu erkennen ist, dass Personen mit wenigem bis mittleren Einkommen weniger im Katalog kaufen, hier die rote und grüne Gruppe. Personen mit mittlerem Einkommen, heißt die grüne Gruppe, kaufen ebenso häufiger im Laden ein. Bei Leuten mit höherem Einkommen erkennt man, dass doch mehr im Katalog als im Laden eingekauft wird, wobei es doch ausgeglichen ist. Die Abbildung 7 mit Bezug auf das Geburtsjahr zeigt zwei Gruppen, die sehr ähnlich sind und beide überwiegend im Katalog einkaufen. Es gibt ebenso einzeln verteilt Personen, die nur im Katalog oder im Laden einkaufen. Jedoch kann man feststellen, dass die rote Gruppe, also die ältesten Personen, häufiger über den Katalog einkaufen. Im Hinblick auf die Ergebnisse zeigt sich der Werbekanal Katalog als bestes Mittel für Leute mit hohem Einkommen und älteren Leuten.

	NumWebPurchases	NumCatalogPurchases	NumStorePurchases	Menge	Income
NumWebPurchases	1.000000	0.387000	0.516000	0.529000	0.459000
NumCatalogPurchases	0.387000	1.000000	0.518000	0.780000	0.697000
NumStorePurchases	0.516000	0.518000	1.000000	0.675000	0.630000
Menge	0.529000	0.780000	0.675000	1.000000	0.793000
Income	0.459000	0.697000	0.630000	0.793000	1.000000

Abbildung 5: Korrelationsmatrix - Präferenz von Werbekanälen

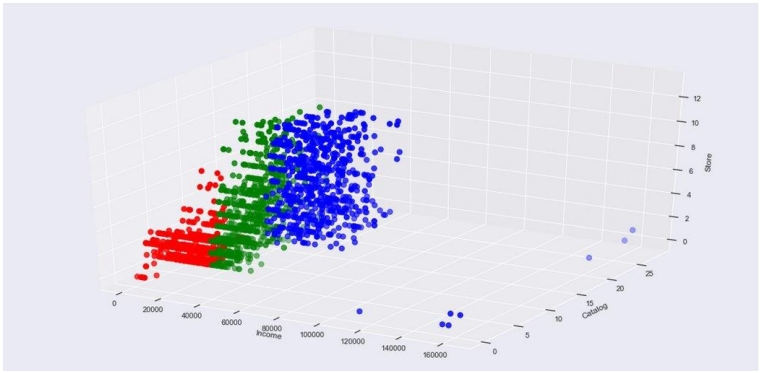

Abbildung 6: K-Means - Präferenz von Werbekanälen mit Einkommen

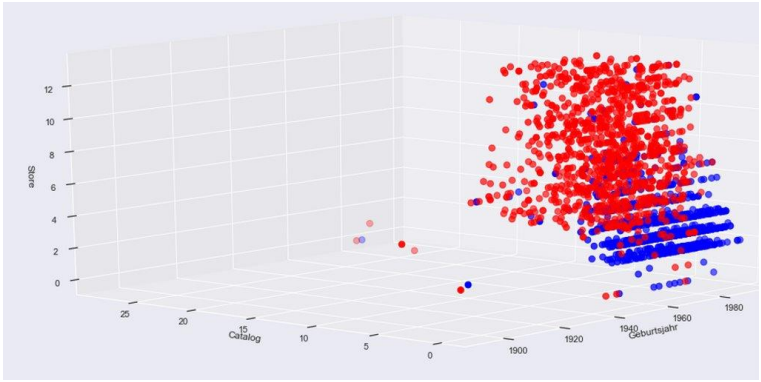

Abbildung 7: K-Means - Präferenz von Werbekanälen mit Geburtsjahr

Kritische Bewertung der Ergebnisse

Zusammenfassend kann man sagen, dass die Zielgruppe dieses Datensatzes für die Produktgruppen Personen mit höherem Einkommen und Personen ohne bzw. mit einem Kind sind. Je nach Businessfrage könnten ältere Personen auch dazu gehören. Dafür kann das Unternehmen Handlungen wie eine attraktivere Gestaltung des Katalogs ableiten. Jedoch ist zu beachten, dass die Daten vermutlich älter sind, da die meisten Personen zwischen 1950 und 1980 geboren wurden und der Werbekanal Katalog heutzutage nicht effektiv sein könnte. Des Weiteren muss für eine genauere Zielgruppenbestimmung ein größerer und der Realität entsprechender Datensatz verarbeitet werden, welcher möglicherweise eine granulare Produktebene hat, um den effektivsten Werbekanal zu bestimmen. Damit der Fokus für dieses Unternehmen nicht allein auf die Werbung im Katalog liegt und weitere Zielgruppe in Betracht gezogen werden, können weitere Analysen verknüpft werden, um beispielsweise die Ladenfläche attraktiver zu gestalten.

Literaturverzeichnis

- Gossens, Thomas (2000): Nutzungsbasierte Kundensegmentierung. In: Data Mining im praktischen Einsatz: Vieweg+Teubner Verlag, S. 143–180. Online verfügbar unter https://link.springer.com/chapter/10.1007/978-3-322-89950-7_7.
- Introduction to SciPy (2022). Online verfügbar unter https://www.w3schools.com/python/scipy/scipy_intro.php, zuletzt aktualisiert am 12.07.2022, zuletzt geprüft am 12.07.2022.
- karnikakapoor (2021): Customer Segmentation: Clustering. In: *Kaggle*, 08.10.2021. Online verfügbar unter https://www.kaggle.com/code/karnikakapoor/customer-segmentation-clustering, zuletzt geprüft am 13.07.2022.
- Luber, Stefan (17.09.2018). Was ist Scikit-learn? In: BigData-Insider. url: https : //www.bigdata-insider.de/was-ist-scikit-learn-a-756150/.
- matplotlib.pyplot — Matplotlib 3.5.2 documentation (2022). Online verfügbar unter https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.html, zuletzt aktualisiert am 14.06.2022, zuletzt geprüft am 12.07.2022.
- Matzer, Michael (2018): Optimale Clusteranalyse und Segmentierung mit dem k-Means-Algorithmus. In: *BigData-Insider*, 19.11.2018. Online verfügbar unter https://www.bigdata-insider.de/optimale-clusteranalyse-und-segmentierung-mit-dem-k-means-algorithmus-a-773713/, zuletzt geprüft am 13.07.2022.
- pandas - Python Data Analysis Library (2.11.2021). url: https://pandas.pydata. org/about/index.html.
- Qualtrics (2022): Kundensegmentierung: Definition, Modelle & Ablauf | Qualtrics. Online verfügbar unter https://www.qualtrics.com/de/erlebnismanagement/kunden/kundensegmentierung/, zuletzt aktualisiert am 10.05.2022, zuletzt geprüft am 13.07.2022.
- Sascha (2021): Clusteranalyse zur Kundensegmentierung (mit Beispielen). Online verfügbar unter https://back2marketingschool.com/de/clusteranalyse-beispiele/, zuletzt aktualisiert am 31.12.2021, zuletzt geprüft am 13.07.2022.
- seaborn: statistical data visualization — seaborn 0.11.2 documentation (2022). Online verfügbar unter https://seaborn.pydata.org/, zuletzt aktualisiert am 28.06.2022, zuletzt geprüft am 12.07.2022.
- What is NumPy? NumPy v1.21 Manual (23.06.2021). url: https://numpy.org/ doc/stable/user/whatisnumpy.html.
- Wuttke, Laurenz (2022): Kundensegmentierung: Beispiel mit K-Means. In: *datasolut GmbH*, 27.04.2022. Online verfügbar unter https://datasolut.com/kundensegmentierung-beispiel-mit-k-means/, zuletzt geprüft am 13.07.2022.