Fachhochschule der Wirtschaft
FHDW Bielefeld

Studienarbeit im Modul Classification and Clustering

Kundensegmentierung mit RetailRocket und Online Retail II

Vorgestellt von:

Christian Roth
Senner Straße 68
33647 Bielefeld
christian.roth@edu.fhdw.de

Studiengang:

Wirtschaftsinformatik mit Schwerpunkt Data Science (*M.Sc.*)

Prüferin:

Prof. Dr. Yvonne Gorniak

Eingereicht am:

08. September 2025 in Bielefeld

Gendererklärung

Aus Gründen der besseren Lesbarkeit wird auf die gleichzeitige Verwendung der Sprachformen männlich, weiblich und divers (m/w/d) verzichtet. Sämtliche Personenbezeichnungen gelten gleichermaßen für alle Geschlechter.

Abstract

Diese Arbeit untersucht die Kundensegmentierung auf Basis zweier offener, datenschutzkonformer E-Commerce-Datensätze: RetailRocket mit Clickstream-Ereignissen und Online Retail II mit Transaktionen. Ziel ist es, robuste und geschäftlich interpretierbare Segmente zu gewinnen, die sowohl Nutzungsverhalten als auch Kaufmuster abbilden. Methodisch orientiert sich die Studie am CRISP-DM-Zyklus. Beide Quellen werden in ein einheitliches Ereignisschema überführt und durch ein gemeinsames Merkmalset beschrieben, das unter anderem RFM-Metriken, Konversionsraten, Zeitmuster, Diversität und Zwischenkaufintervalle umfasst. Ergänzend wird Text Mining auf Kategoriepfaden eingesetzt.

Zur Bildung der Cluster werden zwei kontrastierende Verfahren angewandt: K-Means als zentroidbasiertes und HDBSCAN als dichtebasiertes Verfahren. Die Qualität der Lösungen wird mit Silhouette- und Davies–Bouldin-Index sowie mit dem DBCV-Index bewertet, die Stabilität über Bootstrapping und den Adjusted Rand Index geprüft und die Clustertendenz mit der Hopkins-Statistik abgesichert. Darüber hinaus werden die Verfahren durch Ablationsanalysen und systematische Parameterabstimmungen optimiert.

Die Ergebnisse zeigen, dass sich konsistente Kundensegmente aus unterschiedlichen Datenwelten ableiten lassen und dass bestimmte Merkmalsgruppen die Clusterbildung besonders stark beeinflussen. Die Arbeit liefert eine reproduzierbare Referenzpipeline für Kundensegmentierung im CRM-Kontext und macht transparent, welche Zielkonflikte zwischen K-Means und HDBSCAN in Praxisdaten auftreten.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

1.1 Problemstellung

Unternehmen segmentieren Kunden oft je Datenquelle, etwa nur im Clickstream oder nur in Transaktionen. So entstehen Cluster mit begrenzter Übertragbarkeit. CRM benötigt jedoch stabile, interpretierbare Segmente, die Nutzungs- und Kaufverhalten zusammenführen und ohne personenbezogene Daten auskommen.

1.2 Zielsetzung

Die Arbeit vereinheitlicht RetailRocket und Online Retail II zu einem Ereignisschema und beschreibt Kunden mit einem gemeinsamen Merkmalset aus RFM, Konversionsraten, Zeitmustern, Diversität, Zwischenkaufintervallen sowie Text-Mining auf Kategoriepfaden. Darauf aufbauend werden K-Means und HDBSCAN angewandt und verglichen. Qualität, Stabilität und Clustertendenz werden mit Silhouette-Index, Davies-Bouldin-Index, DBCV, Bootstrapping, Adjusted Rand Index und der Hopkins-Statistik bewertet.

1.3 Vorgehensweise

Die Untersuchung folgt CRISP-DM mit Datenverständnis und EDA, Vereinheitlichung, Merkmalserstellung, Modellierung, Evaluation und Profiling zu CRM-Segmenten. Leitend ist die Frage, wie robust und übertragbar Kundencluster aus RetailRocket und Online Retail II sind, die mit K-Means und HDBSCAN gebildet werden, gemessen mit Silhouette-Index, Davies-Bouldin-Index, DBCV und Adjusted Rand Index, und wie sie zu geschäftlich interpretierbaren Segmenten profiliert werden können.

2 Grundlagen

CRISP-DM bildet den Prozessrahmen dieser Arbeit. Wirth und Hipp schreiben: „The life cycle of a data mining project is broken down in six phases“ (Wirth und Hipp (2000), S. 4). Die Phasen reichen von Business Understanding über Data Understanding, Data Preparation, Modeling und Evaluation bis zu Deployment; sie werden iterativ durchlaufen (Wirth und Hipp (2000), S. 4–7). Wirth und Hipp sprechen von „this highly iterative, creative process with many parallel activities“ und halten fest: „it is never the case that a phase is completely done before the subsequent phase starts“ (Wirth und Hipp (2000), S. 9). Dieser Rahmen stellt sicher, dass Ziele, Daten, Modelle und Bewertung systematisch aufeinander aufbauen.

2.1 Kundensegmentierung im CRM Kontext

Clustering wird im CRM als Verfahren des unüberwachten Lernens eingesetzt, das natürliche Gruppierungen erkennt und Kunden in intern kohäsive Gruppen einordnet. „Clustering techniques identify meaningful natural groupings of records and group customers into distinct segments with internal cohesion“ (Tsiptsis und Chorianopoulos (2010), S. 39). Ziel ist die Bildung unterscheidbarer Kundentypologien, „so that they can be marketed more effectively“ (Tsiptsis und Chorianopoulos (2010), S. 40). Der Nutzen einer Segmentierung misst sich daran, ob die resultierenden Typologien transparent, aussagekräftig und handlungsleitend sind. „The value of each solution depends on its ability to represent transparent, meaningful, and actionable customer typologies“ (Tsiptsis und Chorianopoulos (2010), S. 129). Als visuelle Einordnung dient im Anhang A - Kundensegmentierung, Abschnitt Anhang A.1 - Einordnung der Kundensegmentierung unter Unsupervised Learning die Lernarten-Übersicht, in der Kundensegmentierung unter Unsupervised Learning verortet ist.

2.2 Zwei unabhängige Datensichten und deren Harmonisierung im Lichte von CRISP-DM

Business Understanding konkretisiert das Ziel auf robuste und übertragbare Segmente. Data Understanding führt zwei unabhängige, aber komplementäre Datensichten ein: eine verhaltensnahe Ereignissicht und eine wertbasierte Rechnungssicht. Erst die Kombination erlaubt zu prüfen, ob Clusterstrukturen auf gleich definierten Merkmalen in unterschiedlichen Datenwelten wiederkehren und damit eher domänenweit gültige Muster darstellen als datensatzspezifische Effekte (Wirth und Hipp (2000), S. 5; Tsiptsis und Chorianopoulos (2010), S. 39–40).

In Data Preparation werden beide Quellen in ein gemeinsames Ereignisschema überführt, um Messäquivalenz herzustellen. Praktisch umfasst dies konsistente Felder wie customer_id, item_id, timestamp und event_type sowie bei Bedarf quantity, price, revenue und eine Kategoriendarstellung. Damit lassen sich zentrale Konstrukte wie Recency, Frequency, Interkaufintervalle, Diversität und Zeitmuster in beiden Datensichten identisch operationalisieren (Wirth und Hipp (2000), S. 5–6).

In Modeling werden komplementäre Verfahren auf denselben Kernmerkmalen je Datensicht kalibriert, etwa ein zentroidbasiertes und ein dichtebasiertes Verfahren, um unterschiedliche Strukturannahmen abzudecken (Wirth und Hipp (2000), S. 6). Die Evaluation kombiniert interne Qualitätsmaße, die zur jeweiligen Methode passen, mit Stabilitäts- und Übertragbarkeitstests zwischen den Datensichten, bevor Ergebnisse reproduzierbar dokumentiert werden (Wirth und Hipp (2000), S. 6–7, S. 9).

2.3 Datenquellen und Datenschutz RetailRocket und Online Retail II

Diese Studie nutzt zwei offen zugängliche, anonymisierte E-Commerce-Datensätze mit komplementärer Sicht. Der RetailRocket-Datensatz wurde vom Anbieter auf Kaggle veröffentlicht und enthält roh erfasste Shop-Ereignisse sowie artikelbezogene Eigenschaften und den Kategoriebaum. Zentrale Dateien sind events.csv mit den Ereignistypen view,

addtocart und transaction, item_properties_part1.csv, item_properties_part2.csv sowie category_tree.csv. Identifikatoren für Nutzer und Artikel sind gehasht; personenbezogene Daten liegen nicht vor. Die Kaggle-Seite gilt als Primärreferenz des Datensatzes (Zykov et al. (2022)). Online Retail II ist eine von UCI kuratierte Transaktionshistorie eines britischen Händlers mit Produktcodes, Mengen, Preisen, Zeitstempeln und einer Kundenkennung. Inhaltlich wird die UCI-Landingpage zitiert; der praktische Download erfolgte über den Kaggle-Mirror (*Datasets - UCI Machine Learning Repository* (2019); Miyabon (2009–2011)). Für diese Arbeit genügt ein knapper Datenschutzhinweis, da ausschließlich anonymisierte Forschungsdaten verwendet werden.

2.4 Verfahren und Qualitätsmaße im Überblick K Means, HDBSCAN und passende Indizes

K-Means dient in dieser Arbeit als zentroidbasierte Baseline. Die folgenden Zitate belegen genau die Punkte, die für unser Vorgehen entscheidend sind Iterationsprinzip und Zielgröße, Vorabwahl der Clusterzahl und Zentroidrepräsentation sowie Distanzmaß und Konvergenzkriterium. Erstens zum Iterationsprinzip und Ziel K-Means „starts with an initial cluster solution which is updated and adjusted until no further refinement is possible … Each iteration refines the solution by reducing the within-cluster variation“ (Tsiptsis und Chorianopoulos (2010), S. 85). Zweitens zur Modellannahme über die Segmentzahl „The “K” in the algorithm’s name comes from the fact that users should specify in advance the number of k clusters to be formed. The “means” part of the name refers to the fact that each cluster is represented by the means of its records … the cluster central point or centroid“ (Tsiptsis und Chorianopoulos (2010), S. 85). Drittens zum Ablauf und zur Konvergenz „K-means uses the Euclidean distance measure … The procedure starts by selecting k well-spaced initial records as cluster centers … This iterative procedure is repeated until it converges and the migration of records between clusters no longer refines the solution“ (Tsiptsis und Chorianopoulos (2010), S. 86). Die herangezogenen Quellen belegen präzise die methodischen Voraussetzungen dieser Studie: ein standardisierter numerischer Merkmalsraum mit euklidischer Distanz sowie die transparente Bestimmung der Clusterzahl auf Basis geeigneter Indizes. Zugleich verdeutlichen sie, warum K-Means klar interpretierbare Zentroidprofile

erzeugt, die sich unmittelbar für das anschließende Segmentprofiling nutzen lassen.

HDBSCAN steht für eine hierarchische Weiterentwicklung dichtebasierter Clusterverfahren. Dichtebasierte Ansätze identifizieren Gruppen dort, wo Beobachtungen in Regionen hoher Punktdichte liegen, und behandeln spärlich besetzte Bereiche als Rauschen. Das zugrunde liegende Prinzip wird bei DBSCAN anhand der Parameter ε und MinPts eingeführt; dadurch lassen sich nicht-konvexe Formen erkennen und Ausreißer explizit als noise ausweisen (Chakraborty et al. (2022), S. 87–89). HDBSCAN knüpft an dieses Prinzip an, setzt jedoch auf eine hierarchische Betrachtung der Dichte, sodass Cluster über ein Spektrum von Dichteschwellen hinweg ermittelt und stabilere Gruppierungen extrahiert werden können. Damit entfällt die feste Vorgabe einer Clusterzahl und es entstehen Lösungen, die mit lokal unterschiedlichen Dichten und Rauschen umgehen können.

Der methodische Kontrast zu zentroidbasierten Verfahren ist dabei grundlegend. Während K-Means vor allem bei annähernd kugelförmigen Strukturen überzeugt, eignen sich hierarchische Verfahren für arbiträr geformte Cluster (Chakraborty et al. (2022), S. 96). In dieser Arbeit wird HDBSCAN deshalb als dichtebasiertes Gegenstück zu K-Means eingesetzt; die konkrete Parametrisierung und der algorithmische Ablauf folgen im Methodenteil.

Abschließend zum Methodenvergleich ist der Umgang mit Ausreißern wichtig. In E-Commerce-Daten treten extreme Preise, Mengen oder Event-Bursts auf. K-Means reagiert darauf empfindlich, weil einzelne Extreme Zentren und Distanzen verschieben, während dichtebasierte Verfahren dünn besetzte Bereiche als Rauschen behandeln. Die zugrunde liegende Idee wird beim dichtebasierten Clustering mit ε und MinPts erläutert, die Outlier-Analyse ordnet Begriffe und Vorgehen ein (Chakraborty et al. (2022), S. 87–89). Darauf baut die regelbasierte Ausreißerbehandlung in Kapitel 3 auf.

2.5 Übergang zum Praxisteil Messäquivalenz und minimaler Merkmalssatz

Für einen fairen Vergleich werden beide Quellen in ein gemeinsames Ereignisschema überführt und ein identischer Kern von Merkmalen definiert. Dieser minimale, quellenübergreifende Merkmalssatz trägt die Modellierung und die Stabilitätsanalysen, während zusätzliche merkmals- oder datensatzspezifische Details erst in Kapitel 3 genutzt und begründet werden.

So bleibt der Grundlagenteil schlank und führt direkt in die praktische Umsetzung.

3 Praxisteil

3.1 Datenverständnis RetailRocket Ereignisschema und Eigenschaften

3.2 Datenverständnis Online Retail II Transaktionsschema und Eigenschaften

3.3 Vereinheitlichung beider Datensätze gemeinsames Ereignisschema

3.4 Datenaufbereitung und Feature Engineering RFM Konversionsproxys Zeitmuster Diversität Interkaufintervalle

3.5 Text Mining auf Kategoriepfaden TF IDF optional Item2Vec

3.6 Clustering Ansatz 1 K-Means Auswahl der Clusterzahl über Silhouette und Davies Bouldin

3.7 Clustering Ansatz 2 HDBSCAN Parameterwahl und DBCV

3.8 Stabilität und Robustheit Bootstrapping und Adjusted Rand Index

3.9 Segmentprofiling KPI Profile und Top Kategorien je Cluster

4 Evaluation und Ergebnisse

4.1 Clusterqualität je Datensatz RetailRocket und Online Retail II

4.2 Vergleich der Lösungen K-Means gegenüber HDBSCAN

4.3 Sensitivitätsanalyse Feature Varianten und Hyperparameter

4.4 Grenzen Datenqualität Sparsity Cold Start

5 Fazit und Ausblick

Was eine Multi-Faktor-Authentifizierung (MFA) ist, wird im Glossar beschrieben. Auch glspl und glslink sind möglich.

5.1 Beantwortung der Forschungsfrage

5.2 Nutzen für CRM, Kampagnensteuerung und nächste Schritte

5.3 Ausblick weitere Text Mining Merkmale und Uplift Experimente

Literaturverzeichnis

Chakraborty, S., Islam, S. H., und Samanta, D. (2022). *Data Classification and Incremental Clustering in Data Mining and Machine Learning*. Springer International Publishing. https://doi.org/10.1007/978-3-030-93088-2

Datasets - UCI Machine Learning Repository. (2019, September 21). https://archive.ics.uci.edu/datasets?search=Online+Retail&utm_source=chatgpt.com

Miyabon. (2009–2011, August 31–Dezember 30). *Online Retail II UCI*. https://www.kaggle.com/datasets/mashlyn/online-retail-ii-uci

Tsiptsis, K., und Chorianopoulos, A. (2010). *Data Mining Techniques in CRM: Inside Customer Segmentation* (1. Aufl.). Wiley. https://doi.org/10.1002/9780470685815

Wirth, R., und Hipp, J. (2000). *CRISP-DM: Towards a Standard Process Model for Data Mining*. 11. http://cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf

Zykov, R., Noskov, A., und Anokhin, A. (2022). *Retailrocket recommender system dataset*. Kaggle.com. https://www.kaggle.com/datasets/retailrocket/ecommerce-dataset

Anhang

Anhangsverzeichnis

A: Kundensegmentierung

A1: Einordnung der Kundensegmentierung unter Unsupervised Learning

Die Kundensegmentierung ist ein klassisches Beispiel für Clustering und damit ein zentrales Verfahren des unüberwachten Lernens. Im Machine-Learning-Kontext werden Kunden anhand ihrer Merkmale in Gruppen eingeteilt, ohne dass eine Zielvariable vorgegeben ist. Die folgende Abbildung zeigt die Einordnung von Clustering und Kundensegmentierung innerhalb der verschiedenen Typen von Machine Learning.

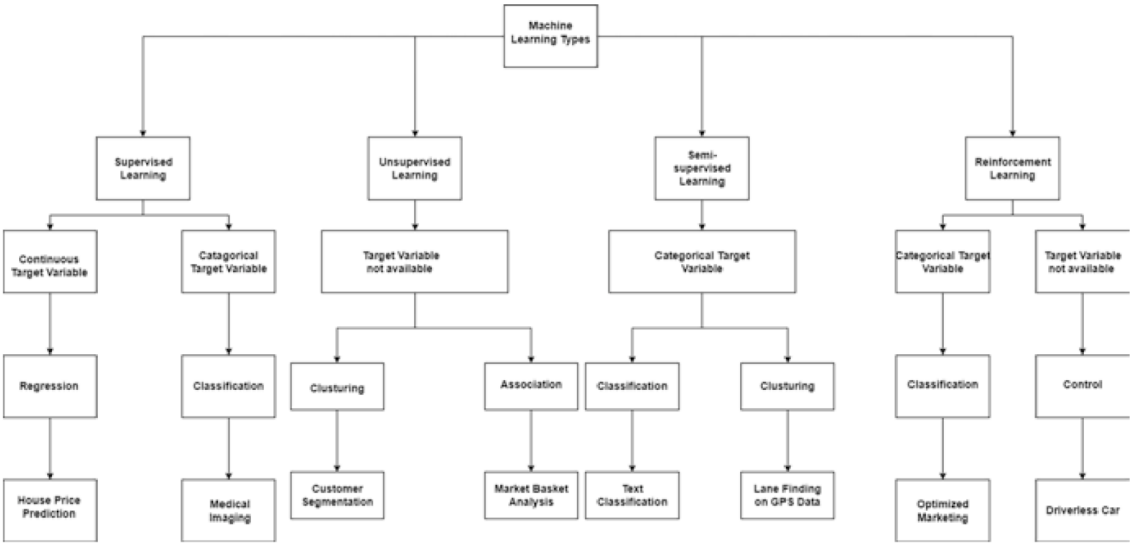

Abbildung A.1: Ansatzpunkte zur Lösung nach Chakraborty et al. (2022), S. 28

Wie in der Abbildung A.1 dargestellt, ist Clustering ein Teilbereich des Unsupervised Learning. Kundensegmentierung nutzt Clustering-Algorithmen, um natürliche Gruppen von Kunden zu identifizieren. Diese Segmentierung bildet die Grundlage für gezielte Marke-

tingmaßnahmen und eine personalisierte Kundenansprache, ohne dass vorher festgelegte Zielvariablen benötigt werden. Im CRM-Kontext ermöglicht dies eine datengetriebene und flexible Einteilung der Kundenbasis.

B: Erste Analyse der Daten

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

B1: Lorem

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Glossar

Multi-Faktor-Authentifizierung (MFA) eine Sicherheitsmethode, die eine zusätzliche Verifikationsebene erfordert, beispielsweise durch eine Kombination aus Passwort und biometrischer Authentifizierung.

Ehrenwörtliche Erklärung

Hiermit erkläre ich, dass ich die vorliegende Studienarbeitselbstständig angefertigt habe. Es wurden nur die in der Arbeit ausdrücklich benannten Quellen und Hilfsmittel benutzt. Wörtlich oder sinngemäß übernommenes Gedankengut habe ich als solches kenntlich gemacht. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

Bielefeld, den 08. September 2025

Christian Roth