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 mittels Clustering auf Basis zweier offener, datenschutzkonformer E-Commerce-Datensätze: RetailRocket mit Clickstream-Ereignissen und Online Retail II mit Transaktionen. Ziel ist es, robuste und übertragbare Kundencluster zu bilden, die sich geschäftlich interpretieren lassen. Methodisch folgt die Studie dem CRISP-DM-Zyklus. Beide Quellen werden in ein einheitliches Ereignisschema überführt und mit datenquellenübergreifenden Merkmalen beschrieben, darunter RFM Recency, Frequency, Monetary Value, Konversionsraten, Zeitmuster, Diversität sowie Zwischenkaufintervalle; ergänzend kommt Text-Mining über Kategoriepfade zum Einsatz. Als zwei kontrastierende Clusteransätze werden K-Means als zentroidbasiertes Verfahren und HDBSCAN als dichtebasiertes Verfahren eingesetzt. Die Güte wird mit Silhouette und Davies–Bouldin sowie dem DBCV-Index bewertet; die Stabilität wird über Bootstrapping und den Adjusted-Rand-Index geprüft, die Clustertendenz mit der Hopkins-Statistik abgesichert. Zur Aufwertung werden Parameter systematisch abgestimmt und Merkmalsvarianten in einer Ablationsanalyse verglichen. Die Ergebnisse zeigen, wie sich konsistente Segmente aus unterschiedlichen Datenwelten ableiten lassen und welche Merkmalsgruppen die Clusterbildung dominieren. 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

Laut einer Untersuchung von Mustermann (Mustermann, 2024) ist das Problem bekannt. Andere Autoren sind anderer Meinung (Musterfrau, 2023).

Was eine Multi-Faktor-Authentifizierung (MFA) ist, wird im Glossar beschrieben.

1.1 Problemstellung

Viele Unternehmen segmentieren Kunden ausschließlich innerhalb einer einzelnen Datenquelle, etwa nur im Shop-Clickstream oder nur in Bestelldaten. Dadurch entstehen Cluster, die schwer übertragbar sind und deren Aussagekraft stark von der jeweiligen Datenwelt abhängt. Gleichzeitig verlangt das CRM nach stabilen, interpretierbaren Segmenten, die sowohl Nutzungsverhalten als auch Kaufmuster abbilden und ohne personenbezogene Daten auskommen. Es fehlt eine kompakte, belastbare Vorgehensweise, die zwei Datensätze zusammenführt, zwei unterschiedliche Clusterverfahren systematisch vergleicht und die Qualität der resultierenden Segmente mit über die Vorlesung hinausgehenden Maßen belegt.

1.2 Zielsetzung

Diese Arbeit verfolgt drei Ziele. Erstens werden RetailRocket und Online Retail II in ein gemeinsames Ereignisschema überführt und mit datenquellenübergreifenden Merkmalen beschrieben RFM, Konversionsraten, Zeitmuster, Diversität, Zwischenkaufintervalle sowie Text-Mining über Kategoriepfade. Zweitens werden Kundencluster mit zwei kontrastierenden Verfahren gebildet K-Means und HDBSCAN und systematisch miteinander verglichen. Drittens wird die Clusterqualität umfassend bewertet mit Silhouette und Davies–Bouldin, zusätzlich DBCV für HDBSCAN, sowie die Stabilität über Bootstrapping und

den Adjusted-Rand-Index; die Clustertendenz wird mit der Hopkins-Statistik geprüft. Eine Ablationsanalyse quantifiziert den Beitrag der Merkmalsgruppen.

1.3 Vorgehensweise

Die Studie orientiert sich strikt an CRISP-DM. Im Datenverständnis werden Struktur und Eigenschaften beider Datensätze erfasst sowie Datenschutz und Zweckbindung erläutert. In der Datenaufbereitung werden beide Quellen auf ein einheitliches Schema mit customer_id, ts, event_type, product_id, category_path, qty, revenue, channel abgebildet und ein datenquellenübergreifendes Feature-Set konstruiert. In der Modellierung werden K-Means mit Auswahl der Clusterzahl über Silhouette und Davies–Bouldin sowie HDBSCAN mit Parameterwahl über den DBCV-Index angewendet. Die Stabilität wird über wiederholte Stichproben Bootstrapping mit dem Adjusted-Rand-Index gemessen; die Clustertendenz wird mit der Hopkins-Statistik geprüft. Die Evaluation vergleicht die Verfahren je Datensatz und diskutiert Clusterprofile anhand von Kennzahlen und dominanten Kategorien. Reproduzierbarkeit wird durch eine klar strukturierte Notebook-Pipeline, feste Seeds, dokumentierte Hyperparameter und exportierte Metriktabellen sichergestellt; personenbezogene Daten werden nicht verwendet.

2 Grundlagen

2.1 Kundensegmentierung im CRM Kontext

2.2 Datenquellen und Datenschutz RetailRocket und Online Retail II

2.3 Clusterverfahren im Überblick K-Means und HDBSCAN

2.4 Qualitätsmaße Silhouette Davies Bouldin DBCV Adjusted Rand Index Hopkins Calinski Harabasz

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

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

Musterfrau, E. (2023). Ein weiterer Testartikel. In *Test-Journal*.

Mustermann, M. (2024). Ein einfacher Testartikel. In *Beispiel-Zeitschrift*.

Anhang

Anhangsverzeichnis

A: Zero-Trust-Modell

A1: Zero-Trust-Säulen

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.

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