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

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

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.

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