



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

Gendererklärung	i
Abstract	ii
Inhaltsverzeichnis	iii
Abbildungsverzeichnis	iv
Tabellenverzeichnis	v
1 Einleitung	1
1.1 Problemstellung	1
1.2 Zielsetzung	1
1.3 Vorgehensweise	1
2 Grundlagen	2
2.1 Kundensegmentierung im CRM Kontext	2
2.2 Zwei unabhängige Datensichten und deren Harmonisierung im Lichte von CRISP-DM	3
2.3 Datenquellen und Datenschutz RetailRocket und Online Retail II	3
2.4 Verfahren und Qualitätsmaße im Überblick K Means, HDBSCAN und pas- sende Indizes	4
2.5 Qualitäts- und Stabilitätsmaße im Überblick	5
2.6 Übergang zum Praxisteil im Lichte von CRISP-DM	7
3 Methodik und Umsetzung	9
3.1 Datenbasis, Harmonisierung und Bereinigung	9
3.2 Gemeinsames Merkmalset RFM, Konversionsproxys, Zeitmuster, Intervalle, Diversität	9
3.3 Clustertendenz und Modelle Hopkins, K-Means, HDBSCAN	9
3.4 Evaluationsdesign Silhouette, CH, DB, DBCV, Bootstrapping, ARI	9

4	Evaluation und Ergebnisse	10
4.1	Qualitätsübersicht RetailRocket und Online Retail II	10
4.2	Stabilität und Robustheit ARI und Rauschanteil	10
4.3	Vergleich K-Means vs. HDBSCAN	10
4.4	Segmentprofile und CRM-Interpretation	10
4.5	Limitationen und Übertragbarkeit	10
5	Fazit und Ausblick	11
5.1	Beantwortung der Forschungsfrage	11
5.2	Nutzen für CRM Kampagnensteuerung und nächste Schritte	11
5.3	Ausblick weitere Text Mining Merkmale und Uplift Experimente	11
	Literaturverzeichnis	13
	Anhang	14
	Anhangsverzeichnis	14
A:	CRISP-DM	15
A.1:	Das CRISP-DM-Modell	15
A.2:	Das CRISP-DM-Modell angewendet auf diese Studienarbeit	16
B:	Kundensegmentierung	18
B.1:	Einordnung der Kundensegmentierung unter Unsupervised Learning	18
C:	Erste Analyse der Daten	20
C.1:	Lorem	20
	Glossar	21

Abbildungsverzeichnis

A.1	Das CRISP-DM-Modell nach Wirth und Hipp (2000), S. 5	15
A.2	Eigene Darstellung des angewendeten CRISP-DM-Frameworks	16
A.3	Ansatzpunkte zur Lösung nach Chakraborty et al. (2022), S. 28	18

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, Merkmalerstellung, 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 und ist in Anhang A: CRISP-DM, Abschnitt A1: Das CRISP-DM-Modell grafisch dargestellt.

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 B: Kundensegmentierung, Abschnitt Anhang B.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; Tsipis 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 dichte-basiertes 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“ (Tsipis 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“ (Tsipis 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“ (Tsipis 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 Qualitäts- und Stabilitätsmaße im Überblick

Zur Bewertung der Clusterlösungen werden für jedes Verfahren passende Gütemaße verwendet. Für K-Means liegt der Schwerpunkt auf dem Silhouettenkoeffizienten. Dieses Maß erfasst, wie gut ein Objekt zu seinem eigenen Cluster passt, verglichen mit der Nähe zu anderen Clustern. Der Wert reicht von -1 bis $+1$, wobei höhere Werte auf klar abgegrenzte

Clusterstrukturen hinweisen. Tsipis und Chorianopoulos nennen die Silhouette als ein zentrales Kriterium für die technische Evaluation von Clustern (Tsipis und Chorianopoulos (2010), S. 209). Lai et al. betonen, dass der Silhouettenkoeffizient sowohl die Kohäsion innerhalb der Cluster als auch die Trennung zwischen Clustern berücksichtigt und damit ein objektives Maß für die Güte der Gruppierung liefert (Lai et al. (2025), S. 3063).

Für dichte-basierte Verfahren wie HDBSCAN eignet sich der DBCV-Index. Moulavi et al. betonen, dass „DBCV employs the concept of Hartigan’s model of density-contour trees to compute the least dense region inside a cluster and the most dense region between the clusters, which are used to measure the within and between-cluster density connectedness of clusters“ (Moulavi et al. (2014), S. 839). Damit wird die Qualität einer Clustering-Lösung nicht mehr über Abstände zwischen Zentren gemessen, sondern über dichte-basierte Eigenschaften. Ein guter Cluster hat dabei eine höhere Minstdichte im Inneren als die maximale Dichte, die ihn von anderen Clustern trennt. Zudem stellen die Autoren klar: „Unlike other relative validity indices, our method not only directly takes into account density and shape properties of clusters but also properly deals with noise objects, which are intrinsic to the definition of the density-based clustering“ (Moulavi et al. (2014), S. 845). Das bedeutet, dass das Verfahren auch Ausreißer und Rauschen korrekt berücksichtigt, die bei dichte-basierten Algorithmen wie HDBSCAN typischerweise auftreten. So lässt sich die Qualität der Lösung auch dann zuverlässig beurteilen, wenn die Cluster keine einfachen, kugelförmigen Strukturen bilden.

Ergänzend zur reinen Gütebewertung wird die Stabilität geprüft. Hierzu werden wiederholt Stichproben aus den Daten gezogen, Modelle neu geschätzt und die Übereinstimmung der Clusterzuordnungen mit dem Adjusted Rand Index gemessen. Werte nahe bei 1 deuten auf eine stabile Lösung hin. Vor der eigentlichen Modellierung wird zudem mit der Hopkins-Statistik getestet, ob der Datensatz überhaupt eine ausgeprägte Clustertendenz enthält. Werte nahe 0,5 sprechen für Zufälligkeit, während Werte über 0,75 auf deutliche Clusterstrukturen hindeuten.

Ergänzend zur reinen Gütebewertung wird die Stabilität der Clusterlösungen untersucht. Dazu werden wiederholt Stichproben aus den Daten gezogen und die Modelle neu geschätzt. Hennig beschreibt dieses Vorgehen als naheliegend: „The Jaccard coefficient, a similarity measure between sets, is used as a cluster-wise measure of cluster stability, which is assessed

by the bootstrap distribution of the Jaccard coefficient for every single cluster of a clustering compared to the most similar cluster in the bootstrapped data sets“ (Hennig (2007), S. 258). Auf Ebene der gesamten Clusterlösung wird häufig der Adjusted Rand Index eingesetzt. Hennig weist darauf hin, dass „the adjusted Rand index (Hubert and Arabie, 1985) has been used often to measure the similarity between two complete clusterings“ (Hennig (2007), S. 260). Auch Moulavi et al. nennen den Adjusted Rand Index als gängiges Maß, um Clustering-Ergebnisse mit einer bestehenden Struktur zu vergleichen: „External clustering validity approaches such as the Adjusted Rand Index compare clustering results with a pre-existing clustering (or class) structure, i.e., a ground truth solution“ (Moulavi et al. (2014), S. 840). Werte nahe bei 1 deuten in beiden Fällen auf eine stabile Lösung hin.

Vor der eigentlichen Modellierung wird zudem geprüft, ob die Daten überhaupt eine ausgeprägte Clustertendenz enthalten. Ein etabliertes Maß hierfür ist die Hopkins-Statistik. Sie „can be used to assess the clustering tendency of a data set by measuring the probability that a uniform data distribution generates a given data set. In other words, it considers the spatial randomness of the data“ (Acito (2023), S. 271). Werte nahe 0,5 sprechen für eine zufällige Verteilung, während Werte über 0,75 auf deutliche Clusterstrukturen schließen lassen.

2.6 Übergang zum Praxisteil im Lichte von CRISP-DM

Die folgenden Schritte orientieren sich eng am CRISP-DM-Referenzmodell und werden iterativ durchlaufen. Wirth und Hipp betonen: „The sequence of the phases is not strict. The arrows indicate only the most important and frequent dependencies between phases, but in a particular project, it depends on the outcome of each phase which phase, or which particular task of a phase, has to be performed next“ (Wirth und Hipp (2000), S. 4).

Genau diese flexible und zyklische Struktur nutzen wir, um Datenaufbereitung, Merkmalsbildung und Modellwahl eng mit den Evaluationsbefunden zu koppeln.

Konkret bedeutet das für den Praxisteil: Zunächst vertiefen wir Data Understanding für beide Quellen und legen Datenqualität, Beobachtungsfenster und Clustertendenz fest. In Data Preparation harmonisieren wir RetailRocket und Online Retail II in ein gemeinsames Ereignisschema, sodass RFM-, Zeit- und Diversitätsmerkmale messäquivalent berechnet

werden. In Modeling setzen wir zwei komplementäre Verfahren ein: K-Means als zentroidbasierte Baseline und HDBSCAN als hierarchisch dichtebasiertes Gegenstück. Die Evaluation folgt dem in Abschnitt Qualitäts- und Stabilitätsmaße im Überblick beschriebenen Set aus Silhouette, Calinski-Harabasz, Davies-Bouldin bzw. DBCV, ergänzt um Bootstrapping und ARI zur Stabilität sowie den Rauschanteil bei HDBSCAN.

Zur Orientierung, wie diese Schritte in dieser Arbeit konkret den CRISP-DM-Phasen zugeordnet sind, verweist Anhang A: CRISP-DM, Abschnitt A.2: Das CRISP-DM-Modell angewendet auf diese Studienarbeit.

3 Methodik und Umsetzung

3.1 Datenbasis, Harmonisierung und Bereinigung

3.2 Gemeinsames Merkmalset RFM, Konversionsproxys, Zeitmuster, Intervalle, Diversität

3.3 Clustertendenz und Modelle Hopkins, K-Means, HDBSCAN

3.4 Evaluationsdesign Silhouette, CH, DB, DBCV, Bootstrapping, ARI

4 Evaluation und Ergebnisse

4.1 Qualitätsübersicht RetailRocket und Online Retail II

4.2 Stabilität und Robustheit ARI und Rauschanteil

4.3 Vergleich K-Means vs. HDBSCAN

4.4 Segmentprofile und CRM-Interpretation

4.5 Limitationen und Übertragbarkeit

5 Fazit und Ausblick

5.1 Beantwortung der Forschungsfrage

Die Untersuchung zeigt, dass sich auf dem gemeinsamen Merkmalset in beiden Datensichten robuste und interpretierbare Kundensegmente ableiten lassen. K-Means liefert bei annähernd kugelförmigen Strukturen die höheren Silhouettenwerte und klarere Zentroidprofile, HDBSCAN identifiziert zusätzlich dichtebasierte Cluster und weist Rauschen explizit aus. Insgesamt bestätigt sich die Übertragbarkeit zentraler Muster zwischen RetailRocket und Online Retail II, wenn Merkmaldefinitionen harmonisiert und Skalierungen konsistent sind.

5.2 Nutzen für CRM Kampagnensteuerung und nächste Schritte

Die resultierenden Segmente sind operabel: Sie unterscheiden sich in RFM-Profilen, Kaufintervallen, Zeitmustern und Top-Kategorien. Daraus lassen sich Zielgruppenregeln für Kampagnen ableiten etwa Reaktivierung bei langen Intervallen, Cross-Sell auf Basis von TF-IDF-Kategorien, sowie Taktung nach bevorzugten Zeitfenstern. Nächste Schritte sind die Anbindung an Kampagnenlogik und ein A/B-Rahmen zur Wirkungsmessung.

5.3 Ausblick weitere Text Mining Merkmale und Uplift Experimente

Zur Verfeinerung eignen sich zusätzliche Text-Merkmale aus Titel, Beschreibung und Kategoriepfaden etwa n-Gramme, Embeddings oder Item2Vec. Für die Wirksamkeitseinordnung empfehlen sich Uplift-Experimente, um Segmente mit kausalem Mehrwert zu identifizieren

und Ressourcen dorthin zu priorisieren.

Was eine [Multi-Faktor-Authentifizierung \(MFA\)](#) ist, wird im Glossar beschrieben. Auch glspl und gslink sind möglich.

Literaturverzeichnis

- Acito, F. (2023). *Predictive Analytics with KNIME: Analytics for Citizen Data Scientists*. Springer Nature Switzerland. <https://doi.org/10.1007/978-3-031-45630-5>
- 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
- Hennig, C. (2007). Cluster-wise assessment of cluster stability. *Computational Statistics & Data Analysis*, 52(1), 258–271. <https://doi.org/10.1016/j.csda.2006.11.025>
- Lai, H., Huang, T., Lu, B., Zhang, S., und Xiaog, R. (2025). Silhouette coefficient-based weighting k-means algorithm. *Neural Computing and Applications*, 37(5), 3061–3075. <https://doi.org/10.1007/s00521-024-10706-0>
- Miyabon. (2009–2011, August 31–Dezember 30). *Online Retail II UCI*. <https://www.kaggle.com/datasets/mashlyn/online-retail-ii-uci>
- Moulavi, D., Jaskowiak, P. A., Campello, R. J. G. B., Zimek, A., und Sander, J. (2014). Density-Based Clustering Validation. *Proceedings of the 2014 SIAM International Conference on Data Mining*, 839–847. <https://doi.org/10.1137/1.9781611973440.96>
- Tsplitsis, 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 CRISP-DM	15
A.1 Das CRISP-DM-Modell	15
A.2 Das CRISP-DM-Modell angewendet auf diese Studienarbeit	16
B Kundensegmentierung	17
B.1 Einordnung der Kundensegmentierung unter Unsupervised Learning	18
C Erste Analyse der Daten	20
C.1 Ergebnisse der ersten Analyse	20

A: CRISP-DM

A.1: Das CRISP-DM-Modell

Die Abbildung A.1 zeigt das CRISP-DM-Referenzmodell mit seinen sechs Phasen Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation und Deployment. Pfeile markieren die zentralen Abhängigkeiten, während der äußere Kreis den iterativen, zyklischen Charakter des Vorgehens unterstreicht. Das Modell ist als allgemeiner Rahmen entworfen und unabhängig von Anwendungsdomäne oder eingesetzter Software.

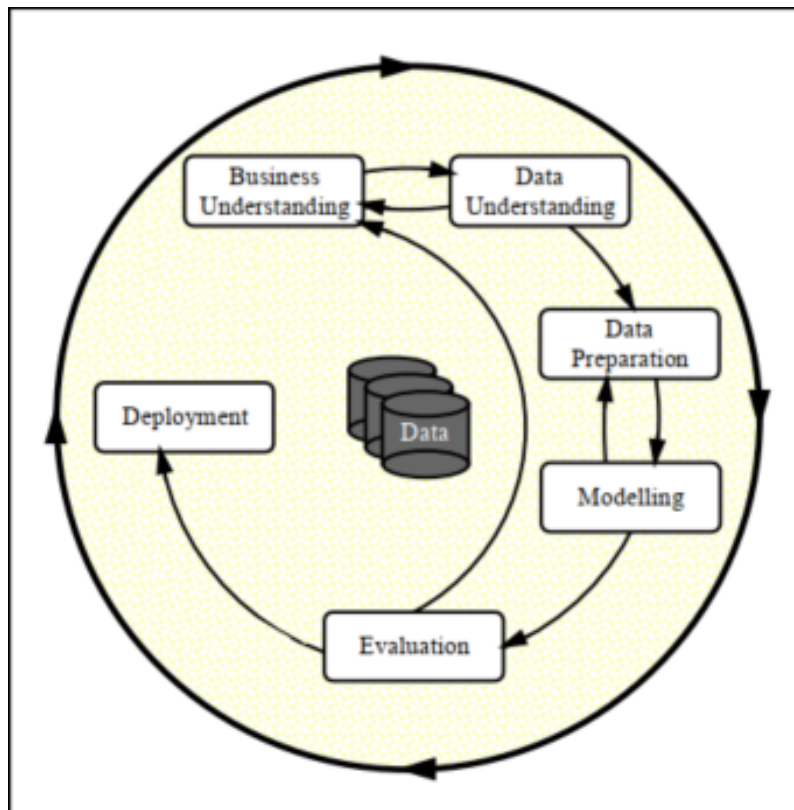


Abbildung A.1: Das CRISP-DM-Modell nach Wirth und Hipp (2000), S. 5

Jede der sechs Phasen ist in spezifische Aufgaben untergliedert, die den Ablauf systematisieren und nachvollziehbar machen. So umfasst Data Understanding das Beschreiben, Erkunden und Überprüfen der Datenqualität, während Data Preparation Aufgaben wie Auswählen, Bereinigen, Konstruieren und Integrieren von Daten vorsieht. Modeling beinhaltet die Wahl geeigneter Verfahren, die Erstellung eines Testdesigns sowie den Aufbau und die Bewertung von Modellen. Evaluation konzentriert sich auf die Überprüfung der Ergebnisse im Hinblick auf die Projektziele, und Deployment behandelt die Bereitstellung und Nutzung der Modelle im Anwendungskontext. Wirth und Hipp betonen außerdem den stark iterativen Charakter des Modells: Ergebnisse einer Phase können jederzeit zu Rücksprüngen führen, wodurch die Arbeitsschritte flexibel angepasst werden können (Wirth und Hipp (2000), S. 6–9).

A.2: Das CRISP-DM-Modell angewendet auf diese Studienarbeit

Die Abbildung A.2 zeigt, wie die einzelnen Phasen des CRISP-DM-Modells konkret auf die Schritte und Kapitel dieser Studienarbeit abgebildet wurden. Die Zuordnung verdeutlicht, welche Arbeitsschritte in welcher Phase umgesetzt und wie sie im Dokument strukturiert sind.

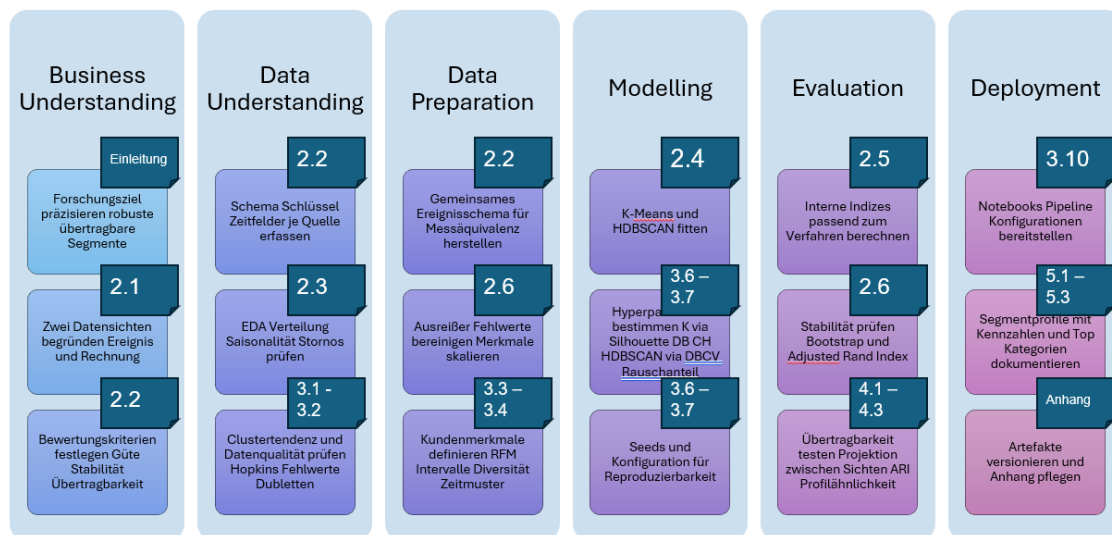


Abbildung A.2: Eigene Darstellung des angewendeten CRISP-DM-Frameworks

Die Grafik ordnet die Kapitel und Arbeitsschritte dieser Studienarbeit den einzelnen Phasen des CRISP-DM-Prozesses zu. So wird sichtbar, wie die theoretischen und praktischen Teile entlang des etablierten Data-Mining-Standards strukturiert und umgesetzt wurden.

B: Kundensegmentierung

B.1: 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.

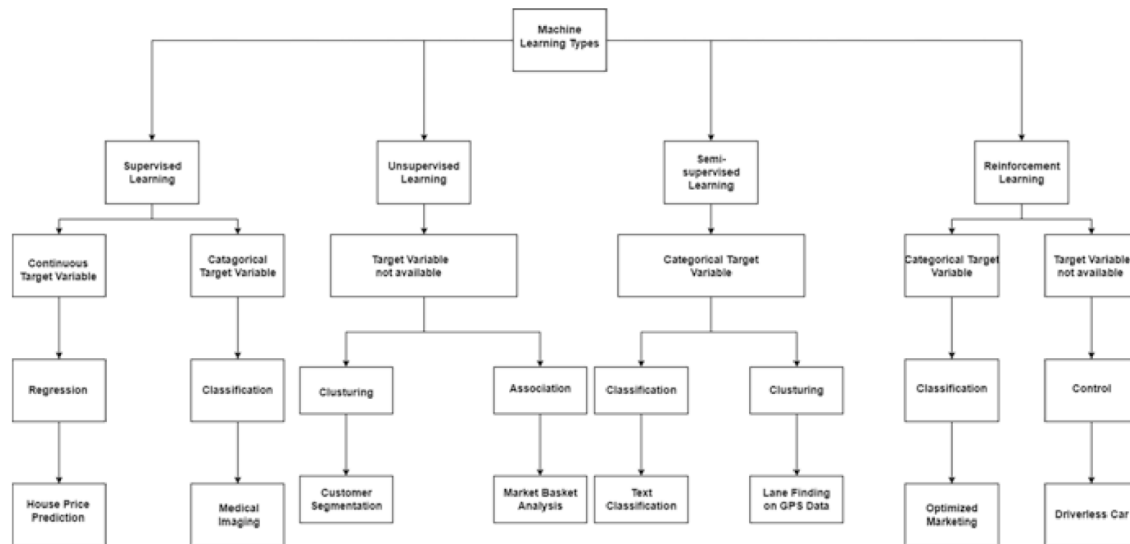


Abbildung A.3: Ansatzpunkte zur Lösung nach Chakraborty et al. (2022), S. 28

Wie in der Abbildung A.3 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.

C: 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.

C.1: 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