Using DBSCAN to Identify Customer Segments with High Churn Risk on Amazon Consumer Behavior Data

Govind A¹ and Rohith Syam¹

¹Affiliation not available

January 25, 2024

Abstract

In modern business, understanding customer behavior is of paramount importance. It allows for the delivery of personalized experiences, which can profoundly influence engagement and revenue. Behavioral segmentation is a potent approach known to enhance customer engagement by up to 73 percent and boost revenue by 15 percent. This study employs DBSCAN (Density-Based Spatial Clustering of Applications with Noise) to address the research problem of segmenting Amazon customers based on a wide range of features. This methodology enables a comprehensive analysis of customer behavior. The key findings of this study reveal customer segments at high risk of churn, a critical concern in the e-commerce industry. This insight is invaluable for businesses looking to tailor retention strategies, ultimately enhancing customer satisfaction and fostering sustainable growth within the fiercely competitive e-commerce landscape. The research showcases the practical application of DBSCAN for customer segmentation and churn risk analysis in the context of Amazon. Academics, managers, and decision-makers can leverage these findings to refine their strategies and better serve their customers, contributing to increased customer satisfaction and long-term business success.
Using DBSCAN to Identify Customer Segments with High Churn Risk on Amazon Consumer Behavior Data

Govind A1* and Rohith Syam2†

1*Department of Computer Science and Engineering, Sree Chitra Thirunal College of Engineering, Trivandrum, 695018, Kerala, India.
2Department of Computer Science and Engineering, Sree Chitra Thirunal College of Engineering, Trivandrum, 695018, Kerala, India.

*Corresponding author(s). E-mail(s): govind123.ga@gmail.com; Contributing authors: rohithlayana@gmail.com; †These authors contributed equally to this work.

Abstract
In modern business, understanding customer behavior is of paramount importance. It allows for the delivery of personalized experiences, which can profoundly influence engagement and revenue. Behavioral segmentation is a potent approach known to enhance customer engagement by up to 73 percent and boost revenue by 15 percent. This study employs DBSCAN (Density-Based Spatial Clustering of Applications with Noise) to address the research problem of segmenting Amazon customers based on a wide range of features. This methodology enables a comprehensive analysis of customer behavior.

The key findings of this study reveal customer segments at high risk of churn, a critical concern in the e-commerce industry. This insight is invaluable for businesses looking to tailor retention strategies, ultimately enhancing customer satisfaction and fostering sustainable growth within the fiercely competitive e-commerce landscape.

The research showcases the practical application of DBSCAN for customer segmentation and churn risk analysis in the context of Amazon. Academics, managers, and decision-makers can leverage these findings to refine their strategies and better serve their customers, contributing to increased customer satisfaction and long-term business success.

Keywords: Customer segmentation, Churn risk, DBSCAN (Density-based spatial clustering of applications with noise), Clickstream data, Machine learning algorithms,
1 Introduction

Customer segmentation is the method of grouping customers based on their shared characteristics and behaviours. This approach allows businesses to identify customer segments with high churn risk, enabling them to tailor retention strategies and improve customer satisfaction. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a powerful clustering algorithm that can be used for customer segmentation, by grouping customers based on their density in the feature space, rather than relying on predefined parameters such as the number of clusters or the distance between customers. This makes DBSCAN an ideal algorithm for customer segmentation, as it can identify clusters of customers with similar behaviour, even if the clusters are of different sizes or shapes.

According to reference [1], the rapid expansion of online shopping activities has been a consequence of the Internet’s ever-increasing reach, with consumers transitioning from offline to online platforms. In this context, the analysis of data generated by online shopping activities plays a pivotal role in crafting effective sales strategies. Specifically, a critical type of data to be scrutinised is market consumer behaviour data, predominantly available in the form of click stream data. Notably, until now, there has been a dearth of research focusing on the comprehensive examination of click stream data components, the mechanisms employed for click stream data recording, and the methodologies to systematically analyse these components in the domain of online retail. This paper elucidates the architecture of a novel online shop application and introduces a dedicated module for recording click stream data, shedding light on the methodologies employed for its analysis.

In this study, we apply DBSCAN to cluster Amazon customers based on a range of features, including purchase history, demographics, and browsing behaviour. Our goal is to identify customer segments with high churn risk. By understanding the behaviours of these segments, businesses can develop targeted retention strategies to reduce churn and improve customer lifetime value (CLV). Here we demonstrate the practical application of DBSCAN for customer segmentation and churn risk analysis in an Amazon context. We identify several customer segments with high churn risk, and we provide insights into the behaviours of these segments. These insights can be used by businesses to develop targeted retention strategies and improve customer satisfaction. Using the data science algorithms we used we can employ targeted marketing.

2 Literature Review

Reference [2] emphasises box plots for outlier detection in data-sets with diverse distributions, highlighting their robustness using quartiles and medians, rather than
sensitive means. However, caution is advised against hasty outlier removal, emphasising meticulous root cause analysis to avoid premature data point dismissal. This approach aids in outlier identification and data quality assessment, enabling analytical comparisons with and without outliers to consider data point importance.

Reference [3] proposes a novel approach to collect and analyse market customer behaviour data on online shops. The approach involves collecting data from multiple sources, cleaning and preparing the data, and applying machine learning algorithms to identify patterns and trends in customer behaviour. The authors evaluate their approach using a case study of an online fashion retailer, and they find that it can be used to make valuable recommendations for improving marketing and sales strategies.

Reference [4] investigates how consumer perceptions of online shopping change with experience. It distinguishes between first-time and experienced online shoppers and explores the role of cultural factors in e-commerce. The study extends the Technology Acceptance Model (TAM) and has implications for e-commerce providers and academic research. The research was supported financially.

Reference [5] explores the impact of customer behaviour on e-commerce and modern market visit intentions. It emphasises the convenience and efficiency of e-commerce, while modern markets rely on physical infrastructure and strategic locations. Customer behaviour significantly influences both e-commerce and modern market visit intentions, fostering trust and convenience. Multiple related studies are referenced.

3 DATA SOURCE & OUTLIERS

The Amazon Consumer Behavior Dataset is a comprehensive dataset of customer interactions and browsing patterns within the Amazon ecosystem. The dataset was obtained from Kaggle. It includes a wide range of variables such as customer demographics, user interaction, and reviews. The dataset is designed to provide insights into customer preferences, shopping habits, and decision-making processes on the Amazon platform. Different utilities of the dataset include:

1. Customer demographics: The dataset includes information on customer demographics such as age and gender. This information can be used to understand the demographics of Amazon’s customers and target marketing campaigns accordingly.

2. Purchase behavior: The dataset also includes information on customer purchase behavior, such as purchase frequency and purchase categories. This information can be used to understand how often customers purchase on Amazon and what products they are buying.

3. Browsing behavior: The dataset also includes information on customer browsing behavior, such as browsing frequency, product search method, and search result exploration. This information can be used to understand how customers browse Amazon’s website or app and how they find the products they are looking for.

4. Review behavior: The dataset also includes information on customer review behavior, such as whether or not a customer has left a product review and how much they rely on product reviews when making a purchase. This information
can be used to understand how important customer reviews are to Amazon customers and how to improve the reliability and helpfulness of customer reviews.

5. Recommendation behavior: The dataset also includes information on customer recommendation behavior, such as how often a customer receives personalized product recommendations from Amazon and how helpful they find these recommendations. This information can be used to understand how effective Amazon’s personalized recommendation system is and how to improve it.

6. Overall shopping experience: The dataset also includes information on the overall customer experience on Amazon, such as customer satisfaction and areas for improvement. This information can be used to identify areas where Amazon can improve its customer experience.

Reference [8] suggests using box plots for outlier detection. Box plots are a reliable and visually intuitive method for identifying data points that deviate significantly from a dataset’s central tendency. They graphically represent data distribution by highlighting quartiles and the interquartile range (IQR), which makes them resilient to the influence of outliers. The method defines upper and lower fences, typically set at 1.5 times the IQR above Q3 and below Q1, as thresholds for identifying outliers. Data points outside these fences are considered potential outliers.

![Box Plot](image_url)

**Fig. 1** Box Plot displays outliers using colour-coding: Age in blue, Customer review importance in orange, Recommendation Frequency in green, Accuracy in red, and Customer Satisfaction in violet. Age exhibits multiple outliers, whereas Recommendation Frequency, Accuracy, and Customer Satisfaction each have one outlier. Customer Review Importance, however, shows no outliers.

### 4 IMPLEMENTING DBSCAN

An improved version of the DBSCAN clustering algorithm can resolve issues with border objects in closely adjacent clusters. This enhancement maintains DBSCAN’s core features while increasing its robustness, particularly in scenarios with dense and connected clusters, expanding its applicability in clustering diverse datasets [7].
Reference [8] discusses the extensive literature on the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm. It explores various aspects, including parameter selection, performance optimization, application domains, comparative evaluations, and contributions by researchers like Dingsheng Deng. This body of work emphasises DBSCAN’s enduring relevance and widespread use across diverse fields, underscoring its significance for researchers and data scientists.

Reference [9] puts forward the various DBSCAN variations that have been proposed in the literature to address the limitations of the original algorithm. The authors critically analyse the strengths and weaknesses of each variation, and they compare and contrast the different approaches.

The DBSCAN clustering algorithm was employed to analyze the customer dataset extracted from "amazondata.csv," focusing on two key attributes, 'Age' and 'Shopping Satisfaction.' First, the data was standardized using the StandardScaler to ensure consistent scales for accurate clustering. DBSCAN was then applied with specific configuration parameters, including an epsilon (eps) value of 0.3 and a minimum number of samples (min samples) set to 5. These parameters determined the clustering density and the point count required to consider a core point. The resulting clusters were graphically visualized in a line chart, showcasing the distribution of customers across various shopping satisfaction levels. Notably, the graph exhibited eight distinct clusters, ranging from highly dissatisfied (Cluster -1) to extremely satisfied (Cluster 7) customers. The largest cluster, Cluster 4, encompassed more than 2.5 customers, indicating moderate shopping satisfaction, while the second-largest cluster, Cluster 7, consisted of over 2 customers representing highly satisfied customers. This implementation of DBSCAN provided valuable insights into customer segmentation and preferences, facilitating targeted marketing strategies and service enhancements based on customer characteristics.
**Algorithm 1** DBSCAN

**Require:** Data, eps, MinPts

**Ensure:** List of clusters

1: Initialize an empty list of clusters
2: Initialize a list to keep track of visited data points
3: for each unvisited data point P in Data do
   4:   Mark P as visited
   5:   Find all data points within a distance of eps from P
   6:   if the number of points found ≥ MinPts then
       7:     Create a new cluster C
       8:     Add P to C
       9:     for each data point Q in the set of points found do
          10:       if Q is not visited then
             11:         Mark Q as visited
             12:         Find all data points within a distance of eps from Q
             13:         if the number of points found ≥ MinPts then
               14:             Add those points to the set of points found
             15:         end if
          16:       end if
          17:       if Q does not belong to any cluster then
             18:         Add Q to C
          19:       end if
       20:     end for
   21:   end if
22: end for
23: Return the list of clusters

**Fig. 2** DBSCAN clustering of shopping satisfaction. The largest cluster is cluster 4, representing customers who are somewhat satisfied. The next largest cluster is cluster 7, representing customers who are very satisfied. There are also small clusters of customers who are very dissatisfied (cluster -1).
According to Figure 2, Cluster -1, despite being the smallest, is the most distinctive, indicating that the customers in this cluster have unique needs and preferences compared to those in other clusters. Hence, the company should prioritize enhancing the customer experience for Cluster -1 by pinpointing their specific areas of dissatisfaction and implementing targeted solutions. Moreover, since Cluster 4 represents the largest customer segment, the company could launch marketing campaigns specifically tailored to boost their satisfaction levels.

5 RESULTS OBSERVATIONS

A range of data science analysis tools such as matplotlib, seaborn, word cloud, pandas, etc., have been employed to examine patterns and analyze data according to the dataset’s different use cases. The results of our analysis are detailed below:

5.1 CUSTOMER DEMOGRAPHICS

![Bar plot visualizing the distribution of gender percentages across different age groups](image)

Fig. 3 The Bar plot visualizes the distribution of gender percentages across different age groups, with age groups on the X-axis and percentage on the Y-axis.

From Figure 3 we can conclude that the percentage sales distribution is more among females in the below 60 age category while it increases exponentially with men count after this age. Non binary genders prefer the use of this shopping platform less than the regulars. This could be fixed by targeted market campaigns for them.
5.2 PURCHASE BEHAVIOUR

Fig. 4 It’s a double Pie Chart, where the inner segment depicts purchase frequency and outer segment illustrates purchase categories.

The insights drawn from Figure 4 indicate that the majority of customers make a limited number of monthly purchases, while only a small percentage make multiple weekly purchases. The bulk of sales are concentrated in specific categories such as clothing, fashion, and beauty and personal care. To enhance sales, it’s imperative to direct more attention towards other categories, such as through increased advertising efforts.

5.3 BROWSING BEHAVIOUR

Fig. 5 The figure represents a heatmap matrix showing the relationship between browsing frequency (X-axis) and product search methods (Y-axis) in percentage form.
The data in Figure 5 suggests that customers who utilize filters make more frequent weekly purchases, whereas customers who do not use filters make purchases less often. This observation implies that, in many instances, individuals may struggle to identify their true needs due to their failure to use filters, resulting in reduced purchasing activity.

5.4 REVIEW BEHAVIOUR

![Heatmap Matrix on Customer Behaviour](image)

**Fig. 6** The figure represents a heatmap matrix showing the relationship between reviews left (X-axis) and customer reviews importance (Y-axis) in percentage form.

Figure 6 demonstrates that 45 percentage of them leaves a review and also highly depends on someone else’s review to make decision while a majority of 55 percentage doesn’t leave behind a review while making a decision that depends on another review. Most of them give 1 or 5 as important since they either favour the product or dislike it. Meanwhile the rest that preferred other responses had a tendency to leave behind a review.

5.5 RECOMMENDATION BEHAVIOUR

We can draw the inference from this figure[7] that when personalized recommendations are scarce, cart completion occurs less frequently, whereas a higher incidence of cart completion is associated with an increased number of personalized recommendations.
Fig. 7 The figure represents a multi-layered bar graph, where X-axis represents cart completion frequency, Y-axis shows percentages, and the bars indicate personalized recommendation frequency using blue for ‘No’, green for ‘Sometimes’, and red for ‘Yes’.

Fig. 8 It's a bar graph which shows the distribution of cart abandonment factors, with the number of abandoned carts on the x-axis and the cart abandonment factors on the y-axis.

According to Figure 8, the primary cause of cart abandonment is customers finding a more appealing option elsewhere. This necessitates the development of improved customer retention strategies by the company.

Fig. 9 It’s a word cloud which shows the top reasons why customers abandon their shopping carts: better price elsewhere, higher price, no longer need the item, and high shipping costs.

In Figure 9, the predominant causes of cart abandonment, such as higher prices and better deals elsewhere, are clearly highlighted by the larger words in the word cloud.
cloud. In contrast, the less commonly cited reasons, like cost changes and shipping expenses, are depicted with smaller words.

6 Conclusion and Future work

This project’s application of DBSCAN for customer segmentation and churn risk analysis in the Amazon context, utilising a comprehensive dataset covering demographics, purchase behaviour, browsing behaviour, review behaviour, recommendation behaviour, and overall shopping experience, provides invaluable insights. These insights can inform highly targeted marketing campaigns, product development, and website improvements, while also enhancing the reliability and relevance of customer reviews and personalised recommendations. This data-driven approach not only enables Amazon to mitigate churn and develop tailored retention strategies but also positions the company for sustained growth in the dynamic and competitive landscape of e-commerce.

We conclude that, based on Figure[2], most customers are somewhat satisfied with their shopping experience. However, there is also a significant number of customers who are very satisfied, and a small number of customers who are very dissatisfied or extremely satisfied. This suggests that businesses should focus on improving the shopping experience for customers who are somewhat satisfied, as this group represents a potential churn risk. Additionally, businesses should ensure that they continue to provide a high-quality shopping experience for customers who are very satisfied to retain them. Furthermore, attention should be given to the needs of customers who are very dissatisfied or extremely satisfied, as these customers may also represent churn risks or opportunities for upselling and loyalty-building. By addressing the concerns and requests of these customer segments, businesses can proactively reduce churn risk and foster long-term customer relationships.

In future the scope of this project can be extended to customer review analysis using various NLP toolkits analysing customer mood and sentiment analysis. Another effective addition is performing Market Basket Analysis on shopping carts to identify customer motive and provide personalised services.

References


**Declarations**

- Funding-Not Applicable
- Conflict of interest-Not Applicable
- Ethics approval-Not Applicable
- Consent to participate-Not Applicable
- Consent for publication-Not Applicable
- Availability of data and materials-Not Applicable
- Code availability-Not Applicable
- Authors' contributions-Not Applicable