PROJECT PROPOSAL

Analyze The Behaviour of Customers on E-commerce Platforms

UNIVERSITY OF CALGARY

Fall 2023: DATA 601- L01

Group #2:
Prashant Dhungana-30080130
Prateek Kaushik-30229287
Mariya Mathews-30192182
Arteen Rafiei-30043409
Nisha Pillai-30158934

Introduction

The retail landscape has undergone a profound transformation over the past decade. Traditional stores have been supplemented and in the majority of cases replaced by online storefronts also known as e-commerce platforms. E-commerce platforms have redefined the way people shop offering variety, convenience, and accessibility. In just a matter of the last two years e-commerce sales increased by 67.5% in the Canadian retail market alone[3]. As the customer demand for more convenient ways of shopping increases, understanding their behavior becomes an essential paramount for businesses to thrive. Thus, the aim of our project is to analyze customer behavior on e-commerce platforms for the sake of customer retention. As the use of e-commerce continues to evolve, it becomes important for businesses to conduct strategic methods to understand customer behavior to target and enhance customer experience ultimately correlating with higher sales.

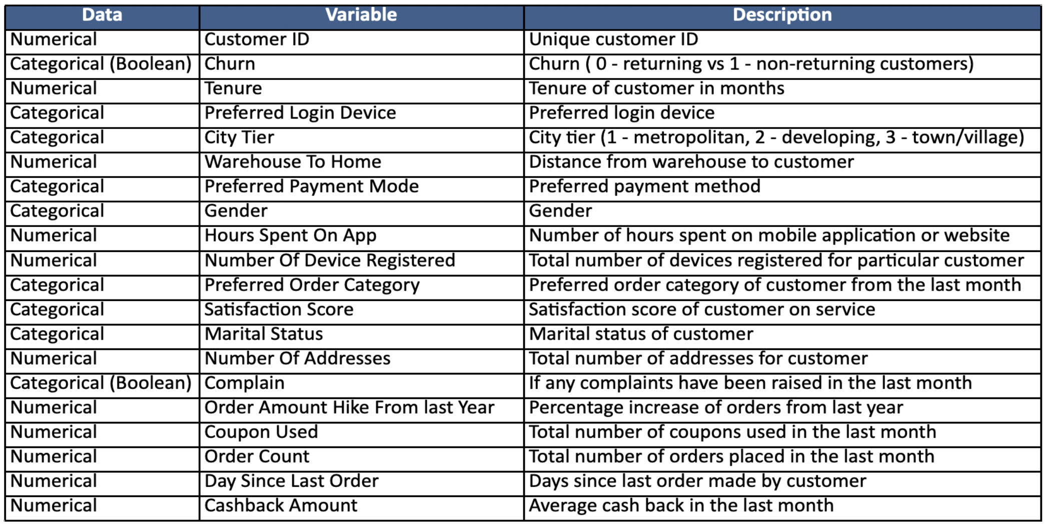
The primary objective of our project is to analyze and understand attributes and factors that influence customer behavior on e-commerce platforms. The analysis will provide valuable and explicit information regarding customer behavior on e-commerce which is an important insight that businesses can use to improve their services and enhance customer satisfaction.

Dataset

The dataset selected for undertaking our objective of diving deeper into online customer behavior is a comprehensive collection of customer interactions and purchasing patterns within the leading e-commerce ecosystem[6]. The initial purpose of the author was to find out which customers would churn. Customers who churn, in this context, refers to customers who do not return to the respective e-commerce platform. In essence, the author's intention with their dataset is similar to our objective.

This dataset offers a structured and tabular format. It is organized into rows and columns, making it suitable for traditional analysis techniques. Each row represents a unique customer interaction and activity with respect to the e-commerce platform, while each column represents a specific attribute or variable associated with those interactions. It is worth mentioning that there are some missing values which will be explained in more detail in the data cleaning section of the task section.

This dataset has 5360 unique values collected from 2019, that is 5360 unique customer interactions, and 20 columns. Each column provides crucial information that can be used for analysis. The columns, descriptions and potential use are listed below:



Data	Variable	Description
Numerical	Customer ID	Unique customer ID
Categorical (Boolean)	Churn	Churn (0 - returning vs 1 - non-returning customers)
Numerical	Tenure	Tenure of customer in months
Categorical	Preferred Login Device	Preferred login device
Categorical	City Tier	City tier (1 - metropolitan, 2 - developing, 3 - town/village)
Numerical	Warehouse To Home	Distance from warehouse to customer
Categorical	Preferred Payment Mode	Preferred payment method
Categorical	Gender	Gender
Numerical	Hours Spent On App	Number of hours spent on mobile application or website
Numerical	Number Of Device Registered	Total number of devices registered for particular customer
Categorical	Preferred Order Category	Preferred order category of customer from the last month
Categorical	Satisfaction Score	Satisfaction score of customer on service
Categorical	Marital Status	Marital status of customer
Numerical	Number Of Addresses	Total number of addresses for customer
Categorical (Boolean)	Complain	If any complaints have been raised in the last month
Numerical	Order Amount Hike From last Year	Percentage increase of orders from last year
Numerical	Coupon Used	Total number of coupons used in the last month
Numerical	Order Count	Total number of orders placed in the last month
Numerical	Day Since Last Order	Days since last order made by customer
Numerical	Cashback Amount	Average cash back in the last month

It is important to note that this dataset was initially found through Kaggle. However, this data was collected directly through creativecommon.org with proper permissions and attributions followed. Proper attribution to the original source and contributor is provided to adhere to ethical and legal considerations[6]. (CC BY-NC-SA 4.0)

Guiding questions

Customer Demographic and Preferences:

1. Does the hours spent on the app depend on the purchase frequency(Order count)?

- Is this correlation strong among men or women?
- What is this correlation with the preferred login device?

The importance of this question is to understand if the purchase frequency of customers increases as they spend more time on the app. Additionally, it aims to explore how different genders behave concerning purchase frequency (order count) and whether this behavior correlates with their preferred login device, which can either be a computer or a mobile phone. In summary, our aim is to find out if there will be any correlation between these features and analyze to carry out what insights it can unfold.

Customer Behavior and Engagement:

2. Are there any patterns forming between Preferred Order category and gender?

- Does the preferred order category have any correlation with marital status?

The significance of this question is to analyze and identify any potential relationship between preferred order category type and gender.By analyzing we can gain insights on the customer preferences on the type of order and group them by gender to uncover any distinct pattern that may exist between different gender groups and their preferred order category types. Additionally, comparing the order category to marital status, we hope to gain insights into customers preferences for their order type relation to their marital status.

3. Does the number of orders change based on the coupons and cashback that the customer receives?

The significance of this inquiry lies in comprehending how coupons and cashback impact the customer's order frequency. The insight we derive from this exploration can help determine whether e-commerce platforms should prioritize offering more coupons and cashback incentives to boost order volume.

Customer Churn and Retention:

4. What is the correlation of satisfaction score and complaints with the churn rate?

The importance of this question lies in understanding customer retention rates.The insight we gain from this is to understand how likely a customer is to stay with the e-commerce website for future purchases based on customer level of satisfaction with their purchase experience.

Customer Satisfaction and Feedback:

5. Does tenure have an impact on satisfaction scores and the number of complaints raised?

The importance of this question lies in determining if there is a relationship between the customer's number of complaints registered within a time period. The insight we gain from this analysis is to understand if customer tenure affects their satisfaction score.

Tasks

In order to conduct meaningful analyses of customer behavior on e-commerce platforms, we have identified and highlighted several tasks, including:

Data Exploration & Cleaning

Having sifted through kaggle we acquired our chosen dataset and went into detail about how the dataset is structured along with the different kinds of variables to carry out meaningful analyses according to the guiding questions.The data set consists of both categorical and numerical data. Although we are working on a clean dataset, a few variables need cleaning by removal of missing and null values.In this dataset the variables like 'Tenure', ‘WarehouseToHome’, ‘HourSpendOnApp’, ‘OrderAmountHikeFromlastYear’, ‘CouponUsed’, ‘OrderCount’ & ’DaySinceLastOrder’ consist of missing values which will be cleaned. Few variables in our data set consist of repetitive attributes. For instance, in the variables ‘PreferredLoginDevice’ & ‘PreferedOrderCat’ we have repetitive attribute values of ‘mobile phone’ and ‘phone’ that will also be cleaned to avoid any confusion . Panda’s will be used to perform this data cleaning task of removing missing values, replacing repetitive values, dealing with duplicate values, outlier detection, validation of data and removal of special characters etc in the dataset.

Data Visualization

To extract meaningful insights, it is crucial to utilize appropriate data visualization techniques, as they are a fundamental aspect of understanding customer behavior on e-commerce platforms. In Python, we have a variety of visualization tools at our disposal, such as Matplotlib, Seaborn and Plotly. All of which allow us to explore customer data and their actions influenced by various factors. For instance:

- Scatter plot to visualize the relationship between the hours spent on the app and purchase frequency (Order count) look for patterns or trends. If there is a positive correlation, we might notice that customers who make more frequent purchases tend to spend more hours on the app.
- Bar chart for relationship between preferred order category and count of each category. Using different colors or patterns within each bar to represent the gender of the customer enhances the clarity of the chart and allows for a more comprehensive understanding of how gender relates to the preferred order category.
- Box plots to show the median, quartiles, and outliers in order counts for customers with and without coupons, providing a summary of the data distribution
- Clustered bar graphs can provide a clear comparison of how churn rates vary across different levels of satisfaction score and complaint count.

We have listed a few data visualization techniques, and additional visualizations may be used based on the project's specific requirements and demands.

Customer Segmentation

To perform customer segmentation, we first need to have processed the data to ensure it is clean and appropriately formatted. We then select the features to group them according to the guiding questions that we aim to address, such as gender, purchase frequency, and shopping satisfaction. Once we have done that, we can ensure the features are standardized. Subsequently, we can perform segmentation via the popular method of KMeans clustering[1] by defining the number of clusters and creating distinct customer segments based on these criteria.

Project Timeline: We anticipate completing the project tasks outlined by the first week of October.

References:

[1]*Arvai, Kevincommerce (2023, August 4). “K-mean Clustering in Python: A Practical Guide.”*[*https://realpython.com/k-means-clustering-python/*](https://realpython.com/k-means-clustering-python/)

[2]*Countants. (2020, January 5). Why consumer behavior analysis is so relevant to the ecommerce business? Medium* [*https://medium.datadriveninvestor.com/why-consumer-behavior-analysis-is-so-relevant-to-the-ecommerce-business-8f49c250ca9c*](https://medium.datadriveninvestor.com/why-consumer-behavior-analysis-is-so-relevant-to-the-ecommerce-business-8f49c250ca9c)

[3]*“Ecommerce - Canada: Statista Market Forecast.” Statista. (n.d).* [*https://www.statista.com/outlook/dmo/ecommerce/canada*](https://www.statista.com/outlook/dmo/ecommerce/canada)*.*

[4]*Matplotlib. plot(x, y) - Matplotlib 3.8.0 documentation. (n.d.).* [*https://matplotlib.org/stable/plot_types/basic/plot.html*](https://matplotlib.org/stable/plot_types/basic/plot.html)

[5]*Predoiu, O. (2023, September 21). Customer behavior analysis. Omniconvert Ecommerce Growth Blog.* [*https://www.omniconvert.com/blog/customer-behavior-analysis/*](https://www.omniconvert.com/blog/customer-behavior-analysis/)

[6]*Verma, Ankit.(2023, July 6). “E-commerce Dataset.”* (CC BY-NC-SA 4.0) *creativecommon.org* [*https://creativecommons.org/licenses/by-nc-sa/4.0/*](https://creativecommons.org/licenses/by-nc-sa/4.0/)

[7]*Verma,(2021, January 26) A. Ecommerce customer churn analysis and prediction. Kaggle.* [*https://www.kaggle.com/datasets/ankitverma2010/ecommerce-customer-churn-analysis-and-prediction*](https://www.kaggle.com/datasets/ankitverma2010/ecommerce-customer-churn-analysis-and-prediction)

[8]*Zanzana, Salim, and Jessica Martin. (2023, February 21). Retail e-commerce and COVID-19: How online sales evolved as in-person shopping resumed.* [*https://www150.statcan.gc.ca/n1/pub/11-621-m/11-621-m2023002-eng.htm*](https://www150.statcan.gc.ca/n1/pub/11-621-m/11-621-m2023002-eng.htm)*.*