Capstone Project

Project Name: Online Retail Store Analysis

By

Vijayeshwaran T

Email: vijayeshwaran2k@gmail.com

Table of contents

Problem Statement

Despite the increasing prominence of online retail, understanding customer purchase patterns remains a challenge for many businesses. The lack of insight into customer behavior hinders the ability of online retail stores to effectively tailor their strategies and offerings. This project seeks to analyze and provide evidence-based insights into the diverse customer purchase patterns exhibited within an online retail store. By identifying key trends, preferences, and factors influencing purchasing decisions, the goal is to empower the store with actionable intelligence for enhancing customer engagement, optimizing product offerings, and ultimately, improving business performance.

Project Objectives

- **Insights:** Providing insights about the customer purchase pattern by identifying key trends, correlations within the customer purchase data. This will be the key to provide a ground breaking strategies among the customers and improving the business performance.
- **Segmenting the Customers:** Segment the customers into groups based on their purchase patterns. This cluster will enhance the understanding of customer needs for the retailers. And helps in providing offers for the right people for the business purpose.

Data Description

The dataset “online_retail.csv” contains 541909 rows and 8 columns. Which has the valuable information about Invoice number, Product ID, Product Description, Quantity of the product, Date of the invoice, Price of the product per unit, Customer ID, Region of Purchase.

Data Preprocessing steps and Inspiration

- Data Cleaning
    - **Handling Null values:** In “online_retail.csv” contains Null values in description and Customer ID features. Since, our aim is to analyze the customer purchase pattern based on the previous purchases, we drop the Null values. After removing the Null values, we have 406829 records in the dataset.
    - **Handling Duplicates:** There are 5225 duplicated values are in the dataset. After removing the duplicates, we have 401604 records in the dataset.
- Data Transformation
    - **Feature Engineering:** Created a new feature TotalSales for the better understanding and analysis of the customer’s purchase.
- Data Formatting
    - Convert the datatype of the feature “InvoiceDate” to datetime and set it as the index for better analysis

- Exploratory Data Analysis
    - **Sales over Time:** Calculated and visualized the total sales over the period of time. This gives the clear view of trends and peaks of the sales of a particular time which is a valuable insight.
    - **Yearly Sales:** Calculated and visualized the total sales for 2010 and 2011. Clearly as per the given data the sales in 2011 is much higher than 2010 (since we have only December data for 2010).
    - **Top selling products based on quantity:** Calculated and visualized top 10 selling products based on the sum of quantities. It provides insight about the product and how frequently people buy a particular product.
    - **Top profit winning product:** Calculated and visualized top 10 products which has the highest sales profit to get the insight about the product sales.
    - **Top Customers Who Buy more Products:** Calculated top customers who bought products more frequently.
    - **Sales over Country:** Calculated and visualized sales over different countries.

Choosing the algorithm for the project

K-Means clustering is well-suited for segmentation tasks, helping identify natural groupings within the data.

Since, this project aims to cluster the customers based on the customer's purchase pattern. K-Means clustering algorithm was applied.

- Implementation
    - Extracted the appropriate features from the data such as Frequency, Recency and Monetary to categorize customers into different segments.
    - Since, K-Means clustering is the distance-based model. The relevant features are standardized.
    - Found the optimal number of clusters using elbow plot and silhouette score.
    - Applied K-Means clustering algorithm on the scaled data.

Inferences from the same

The applied K-Means clustering model has successfully segmented the customers based on the purchase behaviour.

Three main clusters were identified.

- Cluster 0 - Active purchasers.
- Cluster 1 - Active visitors.
- Cluster 2 - High – value customers.

Cluster characteristics

- **Cluster 0 customers** - These are the customers who actively purchase products and visits the shop quite often. These are the customers who visits the shop with a decent frequency and purchase a decent number of products every time. About 74% of customers lies in this cluster.
  Recommendation: Provide personal promotions and offer to enhance customer retention.
- **Cluster 1 customers** - These are the customers who visits shop very frequently but purchase occasionally. They will buy a low amount of purchase and visit the shop with very high frequency compare to other two clusters. This cluster has about 25% of the total customers.
  Recommendation: Provide promotions and offer prices for active visitors as they actively visit the shop to increase the purchase.
- **Cluster 2 customers** - These are the customers who purchase in large quantity and visits the store frequently. This is the cluster which has low number of elements (0.27 percent) means that we can expect this kind of customers rarely.
  Recommendation: Provide Exclusive offers, Bumper offers, Premium services to encourage additional spending.

Future possibilities of the project

- Predictive Analytics:
  - Develop predictive models for each customer segment to forecast future trends, enabling the business to proactively address changing market dynamics.
- Feature Engineering:
  - Explore additional features that could enhance customer segmentation, such as customer demographics, purchase history, or seasonality factors.
- Advanced Modeling:
  - Improve the accuracy of the segmentation with ensemble methods

Conclusion

The analysis of customer purchase patterns in the online retail store has provided valuable insights that can significantly enhance business strategies. By applying the K-Means clustering algorithm, we successfully segmented customers into three main clusters: active purchasers, frequent visitors, and high-value customers. These segments allow for targeted marketing strategies, improving customer engagement, and maximizing sales.

The project's insights into sales trends, top-selling products, and customer behaviors offer a solid foundation for making informed business decisions. Implementing these findings can lead to optimized product offerings, better inventory management, and enhanced overall business performance.

Looking forward, there are several avenues for further development. Predictive analytics can be used to forecast future trends for each customer segment, enabling proactive adjustments to market strategies. Additional features, such as customer demographics and purchase history, can improve segmentation accuracy. Advanced modeling techniques can also be explored to refine customer segmentation and prediction models.

In summary, this project provides actionable intelligence that empowers the online retail store to tailor its strategies effectively, ultimately leading to improved customer satisfaction and business growth.