Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

Name: Sayali Kamalapurkar
Email: kumthekar.sayali19@gmail.com
Contribution:

- Data collection
- Understanding the data variables
- Importing python libraries
- Data Preprocessing
- Outlier Treatment
- Feature Engineering
- EDA on features
- RFM Modelling
- K-Means Clustering Algorithm
- Elbow and Silhouette score methods to find optimal number of clusters
- Cohort Analysis
- Conclusions

Please paste the GitHub Repo link.

GitHub Link: sskamalapurkar/Unsupervised_ML_Clustering (github.com)

Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches, and your conclusions. (200-400 words)

Companies aim to gain a deeper approach of the customer they are targeting. Therefore, their aim must be specific and should be tailored to address the requirements of every individual customer. Furthermore, through the data collected, companies can gain a deeper understanding of customer preferences as well as the requirements for discovering valuable segments that would reap them maximum profit. This way, they can strategize their marketing techniques more efficiently and minimize the possibility of risk to their investment.

PROBLEM STATEMENT: In this project, your task is to identify major customer segments on a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.

APPROACHES: This project started with importing important python libraries such as NumPy, pandas, visualization libraries such seaborn, matplotlib, clustering algorithm from sci-Kit learn libraries such as K-Means etc. Then, we imported dataset and did little bit of basic activity such as, checking out the dataset, data info, understanding columns using data description, etc. Further, looked for null and duplicate values in the dataset. We found that CustomerID column was having a lot of NaN values. we have to segment customers based on the different features, hence without CustomerID it is not possible. Thus, removed all NaN values. This all included under data preprocessing. Then found outliers in the columns Quantity and UnitPrice as these would be troublesome for the feature that we extracted from these two columns – Sales_Amount. Also, did some feature engineering by extracting new features from the InvoiceDate column such as – Hour, Month, Day, day_name, day_time.

Further, we did EDA on features to find out highest selling product, most spending customers, days, and months in which highest purchases are done, in which time of the day maximum purchases are done, highest revenue generated w.r.t. products, days, months, etc. Next, we did RFM Modeling (Recency, Frequency and Monetary), based on RFM and by using elbow method, silhouette score to find best number of clusters, we implemented K-Means clustering algorithm to segment customers. Then we did cohort analysis to segment customers using cohort month based on retention, acquisition, and average spending by the customers.

CONCLUSIONS:

1. Throughout the exercise, we went through various steps to perform customer segmentation. We started with importing data and important libraries. Then, did rigorous data wrangling.
2. We have performed RFM Analysis on the data, where we clustered customers based on Recency, Monetary and Frequency aspect. We used Elbow method, Silhouette score method to find appropriate number of clusters. We discovered 4 clusters based on RFM data.
3. Further, did cohort analysis to understand how retention and acquisition rate, average amount spends changes over the time.
4. However, there can be more modifications on this analysis. One may choose to cluster into more no. depending on company objectives and preferences. The labelled feature after clustering can be fed into classification supervised machine learning algorithms that could predict the classes for new set of observations. The clustering can also be performed on new set of features such as type of products each customer prefers to buy often, finding out customer lifetime value (clv) and much more.