Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:
Name: Rudrajit Bhattacharyya Email: rudrajitb24@gmail.com Contribution: 1. Importing the Libraries and Loading the Data 2. Data Summary 3. Data Cleaning/Pre-processing 4. Feature Engineering 5. Exploratory Data Analysis a. Which products are the most and least sold ones? b. Which countries had the most and least number of customers? c. Distribution of the numerical features d. Which day of the week had the most number of purchases? e. Which month of the year had the most number of purchases? f. Which time of the day had the most number of purchases? 6. RFM Segmentation and Analysis a. Filter UK data only b. Calculating RFM Scores 7. Building Clustering Models a. K-Means clustering on RM data b. K-Means clustering on FM data c. K-Means clustering on RFM data d. Hierarchical clustering on RFM data 8. Evaluation of all the Models 9. Comparing all the Models 10. Conclusion
Please paste the GitHub Repo link.
Github Link:- https://github.com/Rudrajit12/UnsupervisedML_Capstone_Project
Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)

In our online retail customer segmentation project, the task was to find out or identify major customer segments on a transactional dataset. Customer segmentation is a way to split customers into groups based on certain characteristics that those customers share. Customer segmentation will allow marketers to better tailor their marketing efforts to various audience subsets.

The dataset contains 541909 records of transactional data with 8 features. There were a good number of duplicated records and missing values present which were dropped before proceeding further. Few new features were created with the help of the data available in order to simplify and speed up data transformations while also enhancing model accuracy.

We discovered a few insights from EDA like

- which products are the most sold ones and the least sold ones,
- UK was the country with the most number of customers and Saudi Arabia with the least number of customers,
- most of the customers made their purchases on Thursday and the least number of purchases on Friday,
- most of the purchases were made during the festive period between Oct to Dec and least number of purchases during Jan and Feb,
- most of the purchases were made during the afternoon time and least purchases during evening,
- there are 4338 unique customers but only 10 customers had an order share of approx 9% which implies that they might be wholesalers.

There are 6 types of customer segmentation models present but we have decided to use the RFM method for our transactional data. RFM is a method often used to identify customers based on the recency of their last purchase, total number of purchases they have made and the amount they have spent. RFM analysis gives organizations a sense of how much revenue comes from repeat customers vs new customers. After the RFM analysis, K-Means clustering was implemented to identify different customer segments in our data. We have used the silhouette score method and the elbow method to help us determine the best possible ‘k’ value i.e, the optimal number of clusters. We found out that the optimal number of clusters for the K-Means model implemented on RFM data is 3.

The 3 clusters formed are as follows:

- **Loyal Customers**: The most frequent buyers and contribute more monetarily. These are the high valued customers for the retail store.
- **Casual Customers**: These buyers are not so frequent or recent but contribute moderately in terms of money.
- **New Customers**: These are the most recent buyers with very less frequency history.

We have additionally used K-Means clustering separately for Recency, Monetary and Frequency, Monetary along with Hierarchical clustering on RFM data to compare the clustering performances.

Please paste the drive link to your deliverables folder. Ensure that this folder consists of the project Colab notebook, project presentation and video.
Drive Link: https://drive.google.com/drive/folders/1GjovD4mwfHvWTgGOcSTvj4oUsU-1taoD?usp=sharing