Capstone Project Submission

Individual Member's Name, Email and Contribution:

Individual member role:
Aamir Sohail

EMAIL: aamirsohail23081998@gmail.com

1. Data understanding
2. Feature analysis
3. Data visualization
4. Multivariate analysis
5. Exploratory data analysis
6. RFM model
7. K-means
8. Silhouette analysis
9. Elbow analysis
10. DBSCAN
11. Research analysis
12. Technical document

Please paste the GitHub Repo link.

Github Link :- https://github.com/Asohail115/Online-Retail-Customer-Segmentation

write a short summary of the Capstone project and its components. Describe the problem statement, approaches and conclusions.

Customer segmentation is the practice of dividing a company's customers into groups that reflect similarity among customers in each group. The goal of segmenting customers is to decide how to relate to customers in each segment in order to maximize the value of each customer to the business. The contents of the dataset had features such as invoiceno., stockcode, description, quantity, unitprice, customerID, and country. The problem statement was to build an unsupervised machine learning algorithm to perform customer segmentation. We started with data wrangling in which we tried to handle null values, duplicates and performed feature modifications. Next, we did some exploratory data analysis and tried to draw observations from the features we had in the dataset. Next, we formulated some quantitative factors such as recency, frequency and monetary known as rfm model for each of the customers. We implemented KMeans clustering algorithm on these features. We also performed silhouette and elbow method analysis to determine the optimal no. of clusters which was 2. We saw customers having high recency and low frequency and monetary values were part of one cluster and customers having low recency and high frequency, monetary values were part of another cluster. We also implemented shap techniques to understand what is going on inside our model. We saw higher values of frequency, monetary and low values of recency is deciding one class and low values of frequency, monetary and high values of recency is deciding other class. However, there can be more modifications on this analysis. One may choose to cluster into more numbers 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),

segmenting on the basis of time period they visit and much more. As machine learning has become more of an ART, there is nothing such as right or wrong. We only try to get the best outcomes that can suit our final objectives. There is, and always will be, a need to improve, going forward