Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

Suvendu Nayak (kanhasuvendu@gmail.com)

Contribution :

- Defining the problem statement
- Data Cleaning
- EDA and data visualization
- Data pre-processing
- Feature selection
- Preparing Dataset for model
- Applying model
- Model validation and selection

Please paste the GitHub Repo link.

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

Drive Link-
https://drive.google.com/drive/folders/1B9dg7gpokmExsGpUnuqXNAlzQDqf9yUX?usp=sharing

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

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.

Approach:

This study started with importing dataset, analyzing dataset after this I have done preprocessing, I have checked for the null values as our dataset contains many null values in Customer id feature and we have to segment the customers, without customer id we are unable to segment customers therefore I have removed all the rows without Customer id.

After that I have done some exploratory data analysis (EDA) I came to know about top customers, Worst customers, periodical purchasing stats, most revenue generated weekdays, purchase stats of country, top and lease purchasing country, top sold product, most revenue generated product, Customer stats, etc.

After that I have done some feature engineering to build RFM model (recency, frequency and monetary value) . I have extracted and analyzed RFM score then I have created customer segments in 3 categories: bronze, silver and gold.

After that I have done data preprocessing for clustering with the help of log transformation, I have reduced Skewness of data then I have scaled data, after scaling I have extracted Silhouette Score Based on the inertia and silhouette score, I came to know that optimal number of clusters is 3.

Then I have implemented Kmeans clustering and plotted different graphs to visualize clusters. After that I merged the cluster column to data and used a classification model for prediction. I have used Logistic Regression, Random Forest Classifier and XGBoost and done evaluation of it.

Conclusion:

We used 3 Models i.e. Logistic Regression, Random Forest Classifier and XGBoost.
Out of this we came to the conclusion that , Optimal model was Random Forest as I got train Accuracy of 1.00 and test accuracy of 0.98 with it.