Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:
Team Members: 1. Kunal Mahadik Email id : kunalmahadik0811@gmail.com Contribution: 1. Data Wrangling 2. Data Preparation 3. Data Cleaning 4. Data Preprocessing 5. Implementation of Decision Trees and Random Forest. 1. Aashruti Agarwal Email id : aaashruti@gmail.com Contribution: 1. Data Wrangling 2. Data Preparation 3. Data Cleaning 4. Data Preprocessing 5. Implementation of KNN algorithm. 1. Raneev K Email id : raneevk36@gmail.com Contribution: 1. Data Wrangling 2. Data Preparation 3. Data Cleaning 4. Data Preprocessing 5. Implementation of Logistic Regression.
Please paste the GitHub Repo link. https://github.com/AashrutiA/ML_classification
Github Link :- https://github.com/AashrutiA/ML_classification Drive link : https://drive.google.com/drive/folders/1pIoEE9h5Q0pNCZgVKgMvAHb_f9nw4vi4?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)

Credit cards are one of the key business areas within the banking industry, and have gained huge success in the past few years.

Problem Statement:

The goal is to build an automated model for both identifying the key factors, and predicting a credit card default based on the information about the client and historical transactions.

Approach:

1. **Data Cleaning**: The dataset has 25 features and 30000 records without any missing or duplicated value. But, it has many outliers which was treated using floor capping IQR method.

2. **EDA:** we answered some hypothesized questions like defaulter as per different categories, total bills, reason of negative bill etc to find relation of each feature amongst themselves. We found that the rate of being a defaulter is comparatively higher in males than females irrespective of the fact that females forms higher customer proportion. Marital status does not affect the target feature significantly. Customers having higher credit limit tends to be non-defaulter. We found the credit limit is highly correlated with the target variable followed by the payment done through each month. The extra trees classifier was used for feature selection. We found that out data is imbalanced so we applied SMOTE (Synthetic Minority Oversampling Technique) resampling to balance the data.

3. **Feature selection:** we applied ExtraTree classifier to check the results of each feature i.e which feature is more important compared to our model and which is of less importance. We have also implemented ANOVA to confirm the best feature which we will be using further in our model.

4. **Model building and Evaluation:** we implemented various models like. Logistic Regression, Decision Trees, Random Forest, KNN to classify our target variable and evaluated them using recall score evaluation metric.

Challenges:

Major challenge was understanding the data, perform EDA and feature engineering as it is a tricky part with lot of insights. Also, handling label encoded and imbalanced dataset was a tricky part.

Conclusion:

The dataset was pretty clean, **imbalance**d data was balanced using **SMOTE** resampling. EDA was

performed for better understanding the data, extraclassifier and ANOVA test were used for feature selection.

Logistic Regression, Decision Trees, Random Forest algorithms were implemented were evaluated using '**Recall'** score. Even after applying SMOTE, there was imbalance in score as well. Logistic Regression had performed well comparatively with recall score of about 83% for class 0 and 56% for class 1.