Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:
1. Shubham Bhadouria (shubhambhadouria@gmail.com) 1.1. Descriptive Analysis 1.1.1. Data frame description 1.1.2. Data frame shape 1.2. Data cleaning, transformation, and Analysis: 1.2.1. Extracting the information from the categorical variable 1.2.2. Rename the column name for better understanding of data set 1.3. Data Wrangling 1.3.1 Extracting the information from Education, Sex, Marriage, columns for Problem reading 1.3.2 Finding the correlation between the data set and drop the useless columns 1.4 Data Visualization 1.4.1 Relationship between 'Education' and 'Defaulter/Non-defaulter' 1.4.2 Relationship between 'AGE' and 'Defaulter/Non-defaulter' 1.4.3 Relationship between 'Married' and 'Defaulter/Non-defaulter' 1.4.4 Relationship between 'PAY_MONTHS' and 'Defaulter/Non-defaulter' 1.4.5 Relationship between 'Bill_amount' and 'Defaulter/Non-defaulter' 1.4.6 Relationship between 'AGE', 'LIMIT_BAL' and 'Defaulter/Non-defaulter' 1.5 Feature Engineering and Dummy variable 1.5.1 Create a column age group 1.5.2 After the data reading, we did one hot coding. 1.6 Machine learning Model Analysis 1.6.1 Fitting data on KNN,Random Forest, SVM, XGBoost, LGBM. 1.7 Model Explainability 1.7.1 SHAP features
Please paste the GitHub Repo link.
Github Link:- https://github.com/shubham-bhadouria/Credit-Card-Default-Prediction
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 Card Fraud Detection with Machine Learning models is a process of data investigation and the development of a model that will provide the best results in revealing and preventing fraudulent transactions. In this project, we analyze the data of the credit card users of Taiwan deeply and make the prediction whether the customer will be a defaulter or non-defaulter.

Initially we load the data and search for null values and duplicate rows in the dataset. No duplicate and null values are found in our dataset.

After that we did EDA on the dataset. Thereafter we extract all possible information about the 'default payment next month' (dependent variable) with the help of all independent variables like relationship between 'Education' and 'Defaulter/Non-defaulter', Relationship between 'AGE' and 'Defaulter/Non-defaulter', relationship between 'Married' and 'Defaulter/Non-defaulter', relationship between 'Pay_month' and 'Defaulter/Non-defaulter', relationship between 'Bill_amount' and 'Defaulter/Non-defaulter', relationship between 'AGE', 'LIMIT_BAL' and 'Defaulter/Non-defaulter'. After that, we introduce a feature 'age group' and find the relation between 'age group' and target variable. After that we plot correlation heatmap and found that we found that LIMIT_BAL, SEX, EDU, AGE and MARRIAGE are not highly correlated to each other and people who pay the bill on time has the high probability that they can pay the next bill on time.

After extracting useful data from the dataset, we did one hot encoding and dropped all the unnecessary columns and made the dataset ready for fitting in various machine learning models.

We fit our data in 5 models namely KNN model, Random Forest model, XGBboost model, SVM model and LGBT model. After fitting data in all the models, we found that LGBT model has highest test accuracy (0.861682), highest precision score (0.805032), highest F1 score (0.853379) and highest ROC score (0.866385) among all the models and hence outperforms all the other models.

We can conclude that Most of the defaulters are university pass-outs and have the balance limit in the range 50000 to 200000. Credit cardholders who have no consumption or paid in full every month or delay 1 month, the number of no-defaults is more than that of default. For those who use revolving credit, which means people who only pay the minimum amount every month, the non-default far exceeds the default. However, for those who delay the payment for more than one month, it turns out that the likelihood of default would then surpass the non-default, which also means the longer the payment delay, the higher risk for that person to default.