Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:

1) **Anas Mustafa** **Email: Mustafaanas84464@gmail.com**

Contribution:

1) Feature Engineering:
   - Data preprocessing
   - One-Hot encoding
2) EDA(Exploratory Data Analysis):
   - Countplots
   - Boxplot
   - Pair plot
3) Imbalanced dataset
   - SMOTE
4) Classification Analysis:
   - Logistic Regression
   - Decision Tree Classifier
   - Random Forest Classifier
   - XG-Boost Classifier
5) Model Explain ability
   - Shapley
   - Summary Plot
6) ROC-AUC curve
7) Presentation

2) **Chetan Rajput:** **Email: Chetan.rajput91@yahoo.com**

Contribution:

1) Feature Engineering:
   - Data Preprocessing
   - One-Hot encoding
2) EDA(Exploratory Data Analysis):
   - Countplots
   - Pairplot
   - Heat-Map
3) Hyperparameter tuning and Cross-validation
4) Classification Analysis:
   i. Logistic Regression
   ii. Decision Tree Classifier
   iii. XG-Boost Classifier
5) Group colab

3) **Sarthak Rastogi:** **Email: sartakrastogi1@gmail.com**

Contribution:

1) Data Mugging
2) Introducing New variables
3) Data Visualization:

- Countplot
- Boxplot
- Confusion Matrix(Heat-Map)

4) Classification Analysis:
   - Logistic Regression
   - Random Forest Classifier
   - XG-Boost Classifier

5) Feature Importance:

Github link:-

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 and their facilities are so lucrative nowadays among people, everything has its consequences and responsibilities. Banks issue credit cards to students, employees, and businessmen for their ease and of course to grow their business but if the institute will not take care of time to time payment of issued credit cards this can lead to huge business loss or even bankruptcy. Customers who don't pay their payment bill on credit cards at a time are known as defaulters and the institute call it 'Default Payment'. To overcome this problem(who would default or not) we have been provided with the dataset of credit cards (having 30000 instances and25 columns).

Our first step was to import the dataset through pandas 'read_csv' then data wrangling and featureengineering in our dataset. We did not get into the situation to remove NA values because there are 0 null values in the Credit Card dataset.

Next, EDA(exploratory data analysis) in which visualization of default payments are examined where we get to know that our dataset suffers imbalance in the dataset. Plotted count plot for thevariables Education, Age, Limit balance, Sex and all repayment columns. Boxplot and pair plot also plotted for the rest of bill amount of different months and pay amount of different months.

Our dataset has lots of categorical features having string labels and ordinal labels. So, we applieda one-hot encoding technique is applied on the features(Age, Marriage, Education and Repay Scale) to deal with it.

After splitting the dataset into training and testing, SMOTE has been applied to the training dataset to balance our dataset, this technique creates artificial instances based on the minoritydataset.

Applied Logistic Regression, Decision tree classifier, Random Forest and XGBoost Classifier, after fitting the model, training accuracy, testing accuracy, and confusion matrix is called for all the models. Among all these models Random Forest and XGBoost Classifier are performing in abetter way. It can be seen that after applying Hyperparameter Tuning and cross-validation performance of the RF classifier has improved.

Conclusion:

- Using a Logistic Regression classifier, we can predict with 66.37% accuracy, whether acustomer is likely to default next month.
- Using the Decision Tree classifier, we can predict with 71.83% accuracy whether acustomer is likely to default next month or not.
- Using Random Forest, we can predict with 81.38% accuracy whether a customer will be adefaulter in the next month or not.

By applying XGBoost Classifier with a recall of 60.60%, we can predict with 81.60% accuracy whether a customer is likely to default next month.