Capstone Project Submission

Team Member's Name, Email and Contribution:

Team Member's Name	Email id
Abhishek Kirar	abhishekkirar27@gmail.com

Contribution:

Abhishek Kirar :

- Data Wrangling
  - Loading and Pre-processing
  - Structuring data
  - Enriching data
- Data Mining
- Data Analysis
- Statistical Analysis
- Visualizations
  - Bar graphs and Distribution Graph
- Machine Learning -- Modelling and Predicting using Algorithms
  - Logistic Regression with Hypertuning, Decision Tree Classifier with Hypertuning, Random Forest Classifier with Hypertuning, KNN Classifier
- Observation
- Summarization
- Conclusions
- Technical Document
- PowerPoint Presentation

Please paste the GitHub Repo link.

Github Link:-

https://github.com/abkirar27/Credit-Card-Default-Prediction-Capstone-Project-03

Short summary of Capstone project and its components. With the problem statement, approaches and conclusions. (200-400 words)

Project Name: Credit Card Default Prediction

This project is aimed at predicting the case of customers' default payments in Taiwan. From the perspective of risk management, the result of predictive accuracy of the estimated probability of default will be more valuable than the binary result of classification - credible or not credible clients.

Approach:

As an individual I read the data present in the file and went through the details in each and every column. The data set was huge in which some of the data was not required for the analysis so the data was cleaned by dropping some unwanted columns and creating a new data frame, with the columns we required for the analysis. The first problem we faced was the name of the columns which was not proper and the nan values present in the data. We renamed the columns by using dictionary format and replaced all the nan values to zero in int dtype and unknown in object dtype by using replace syntax. Each and every column was compared to gain the insights about the data. I worked individually to gain some insights by doing the exploratory data analysis using python. Cleaning the dataset, analysing the data and visualizing the data by plotting the data into different graphs and charts so that the trend and relationship between the various indicators can be understood easily, Modelling and Predicting the model using Machine learning algorithms which model is best to predict .

Conclusion:

- XGBoost model has the highest recall, if the business cares about recall the most, then this model is the best candidate. If the balance of recall and precision is the most important metric, then Random Forest is the ideal model. Since Random Forest has slightly lower recall but much higher precision than Logistic Regression, I would recommend Random Forest.
- Data categorical variables had minority classes which were added to their closest majority class.
- There were not huge gaps but female clients tended to default the most. Labels of the data were imbalanced and had a significant difference.
- Gradient boost gave the highest accuracy of 82% on the test dataset. Repayment in the month of September tended to be the most important feature for our machine learning model.
- The best accuracy is obtained for the Random forest and XGBoost classifier.
- In general, all models have comparable accuracy. Nevertheless, because the classes are imbalanced (the proportion of non-default credit cards is higher than default) this metric is misleading.
- Furthermore, accuracy does not consider the rate of false positives (non-default credit cards that were predicted as default) and false negatives (default credit cards that were

incorrectly predicted as non-default).

- Both cases have a negative impact on the bank, since false positives lead to unsatisfied customers and false negatives lead to financial loss.
- From above table we can see that XGBoost Classifier having Recall, F1-score, and ROC Score values equals 86%, 82%, and 86% and Random forest Classifier having Recall, F1-score, and ROC Score values equals 86%, 83%, and 84%.
- XGBoost Classifier and Decision Tree Classifier are giving us the best Recall, F1-score, and ROC Score among other algorithms. We can conclude that these two algorithms are the best to predict whether the credit card is default or not according to our analysis.