**Capstone Project Submission**

**Instructions:**

if) Please fill in all the required information.

ii) Avoid grammatical errors.

|  |
| --- |
| **Team Member’s Name, Email and Contribution:** |
| **Name: Anish Johnson.**  **Email:** [**anishjohnson05@gmail.com**](mailto:anishjohnson05@gmail.com)**.**  **Contribution: Individual project.** |
| **Please paste the GitHub Repo link.** |
| GitHub Link: - <https://github.com/anishjohnson/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)** |
| As more and more consumers rely on credit cards to pay their everyday purchases in an online and physical retail store, the amount of issued credit cards and the overwhelming amount of credit card debt by the cardholders have rapidly increased.  Therefore, most financial institutions must deal with the issues of credit card default in addition to credit card fraud.    Our objective is to conduct quantitative analysis on credit card default risk by using machine learning models with accessible customer data to assist in predicting the case of customers' default payments in Taiwan.  After cleaning the data and renaming a few variables, in the next step, we began EDA (Exploratory Data Analysis) to understand the relationship between the dependent and the independent variables better and identify the necessary trends in them.  We then split the data into two sets, the train (70%) and the test (30%), and standardized it using Standard-Scaler.  Then we implemented five machine-learning approaches (***Logistic Regression, XGB Classifier, KNeighbors Classifier, Random Forest Classifier, ExtraTrees Classifier***) to predict the default cases on the provided data; however, most models encountered challenges to resolve the imbalance problem of default cases in data sets.  To avoid this, we oversampled the data using SMOTE Tomek and observed an increase in the overall performance of the models.  XG Boost Classifier with SMOTE Tomek gave the best accuracy of 82% but lacked Recall which is crucial in classifying the defaulters, whereas **XG Boost** with imbalanced data while applying the scale\_pos\_weight=3.521 provided the best Recall of **64%** approx.  In terms of overall balance, Random Forest Classifier provided the best results.  We also achieved a higher Recall of 80% (approx.) for XG Boost (on imbalanced data) by moving the threshold value (default threshold value=0.5) to 0.3605385.  From the feature importance plots for XG Boost created using SHAP, we observed that PAY\_SEPT\_0, PAY\_SEPT\_2, and *LIMIT*\_BAL are the most vital features in predicting the defaults.  From the above observations, we can say that XG Boost outperforms other models in terms of Recall and Accuracy; hence we can proceed with XG B­oost for future classifications. |