Capstone Project Submission

Instructions:
i) Please fill in all the required information.
ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:
Sanjay Yadav neer.ping@gmail.com **Contribution :** **1.Exploratory Data Analysis** **2.Handle class imbalance** **3. Feature Engineering** **4.Building Classification Model** **5. Model Evaluation**
Please paste the GitHub Repo link.
Github Link:- https://github.com/sanjay2097/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) **Problem Statement :** **Credit risk plays a major role in the banking industry business. Banks' main activities involve granting loan, credit card, investment, mortgage, and others. Credit cards have been one of the most booming financial services by banks over the past years. However, with the growing number of credit card users, banks have been facing an escalating credit card default rate. As such data analytics can provide solutions to tackle the current phenomenon and management credit risks.** **This dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005. (30,000rows and 25 columns)** **Approach :** **In this project we covered various aspects of classification using Machine learning models. During feature engineering we grouped certain misclassified features inside columns , we also had to perform outlier treatment for those**

records where the label was giving a wrong prediction when payment was fully made throughout the months .Further we did feature selection to filter and gather only the optimal features which are more significantly correlated to the target variable.The imbalanced training set was balanced using SMOTE. Then finally we trained the models on the optimum featureset to get the results and evaluated the models using various scores and confusion matrix to find the optimum classification model for our dataset.

Conclusion :

- **XGBoost and Random Forest classification models both have the best performance with a recall score of more than 80% and high AUC.**