

# Capstone Project Submission

## Instructions:

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

### **Team Member's Name, Email and Contribution:**

sibani choudhury

Mail id [choudhursibani120@gmail.com](mailto:choudhursibani120@gmail.com)

Contribution

- Preview Data
- Check total number of entries and column types
- Check the null values
- Plot distribution of numeric data
- Plot distribution of categorical data
- Remove the outliers
- PPT and Team Colab Building Contribution
- Confusion matrix, precision and recall
- Building the model
  - Logistic regression
  - XgBOOST
  - Random forest
- Model interpretation
- Conclusion

Rasik Jain

Mail id [rasik6627265@gmail.com](mailto:rasik6627265@gmail.com)

Contribution

- Data Cleaning
- EDA
- Feature engineering and Feature selection
- distribution check for dependent and independent features numeric data
- Outlier detection and elimination
- Confusion matrix
- PPT and Team Colab Building Contribution
- Building and evaluating the model
- Conclusion

**Please paste the GitHub Repo link.**

<https://github.com/choudhursibani120/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)**

This project is aimed at predicting the case of customers default payments in Taiwan. K-S chart can be used to evaluate which customers will default on their credit card payments.

The purpose of this project is to conduct quantitative analysis on credit card default risk by using interpretable machine learning models with accessible customer data, instead of credit score or credit history, with the goal of assisting and speeding up the human decision making process.

First of all we have done some exploratory data analysis to understand about the data. From the modeling, we are able to classify default risk with accessible customer data and find a decent model. using a Logistic Regression classifier, we can predict with 73% accuracy, whether a customer is likely to default next month. Using a Random Forest classifier, we can predict with 92% accuracy, whether a customer is likely to default next month. Using a XGBOOST classifier, we can predict with 88% accuracy, whether a customer is likely to default next month.

This analysis uses 3 classification models - Logistic Regression, Random Forest and XGBoost. Since Random Forest and XGBoost are tree based on algorithms, rescaling is only performed on Logistic Regression, not on these 2 models. For each model, we first try the model's default parameters, train each model without SMOTE and with SMOTE samplings. Then tune each model's hyperparameters to find the optimal performance. As mentioned earlier, this dataset has imbalanced classes, therefore we use precision and recall, instead of accuracy as the performance metrics.

If the balance of recall and precision is the most important metric, then Random Forest is the ideal model.