Credit Card Default Risk Analysis

The purpose of this project is to conduct quantitative analysis on credit card default risk by using 3 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.

Dataset Source

This dataset contains information on default payments, demographic factors, credit limit, history of payments, and bill statements of credit card clients in Taiwan from April 2005 to September 2005.

Project Overview

The analysis consists of 2 Jupyter notebooks.

1. Exploratory Data Analysis.
2. Machine Learning Modeling.

Machine Learning Models Used:

1. Logistic Regression
2. Random Forest
3. XGBoost

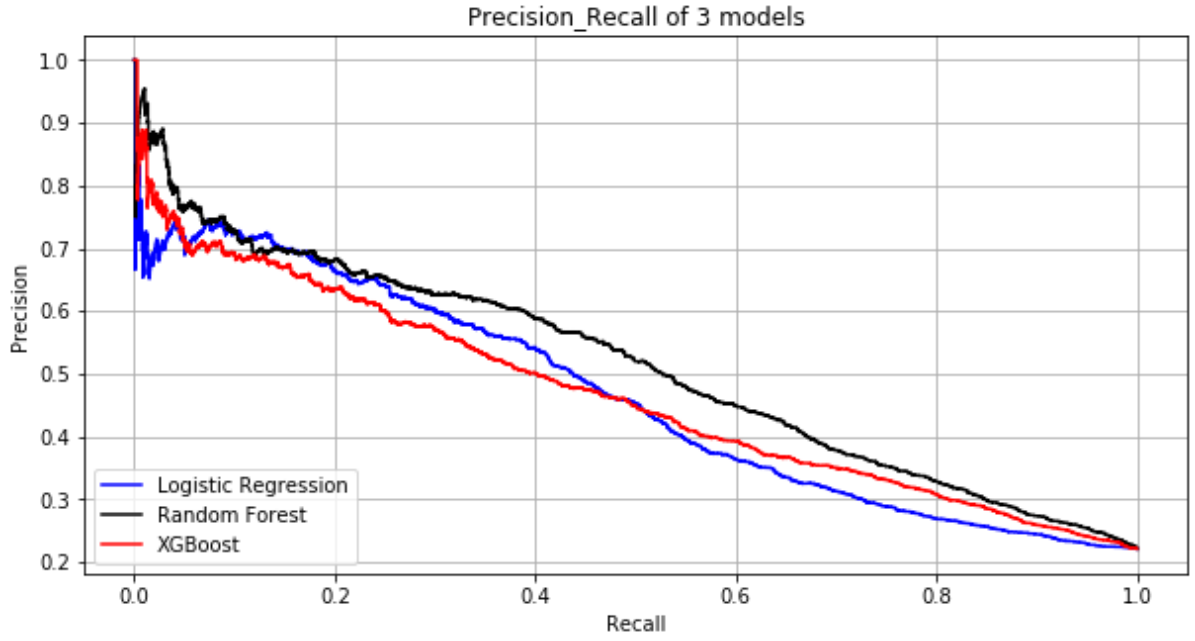
Key Findings from EDA

1. Males have more delayed payment than females in this dataset. Keep in mind that this finding only applies to this dataset, it does not imply this is true for other datasets.
2. Customers with higher education have less default payments and higher credit limits.
3. Customers aged between 30-50 have the lowest delayed payment rate, while younger groups (20-30) and older groups (50-70) all have higher delayed payment rates. However, the delayed rate drops slightly again in customers older than 70.
4. There appears to be no correlation between default payment and marital status.

5. Customers being inactive doesn't mean they have no default risk. We found 317 out of 870 inactive customers who had no consumption in 6 months then defaulted next month.

Model Comparison

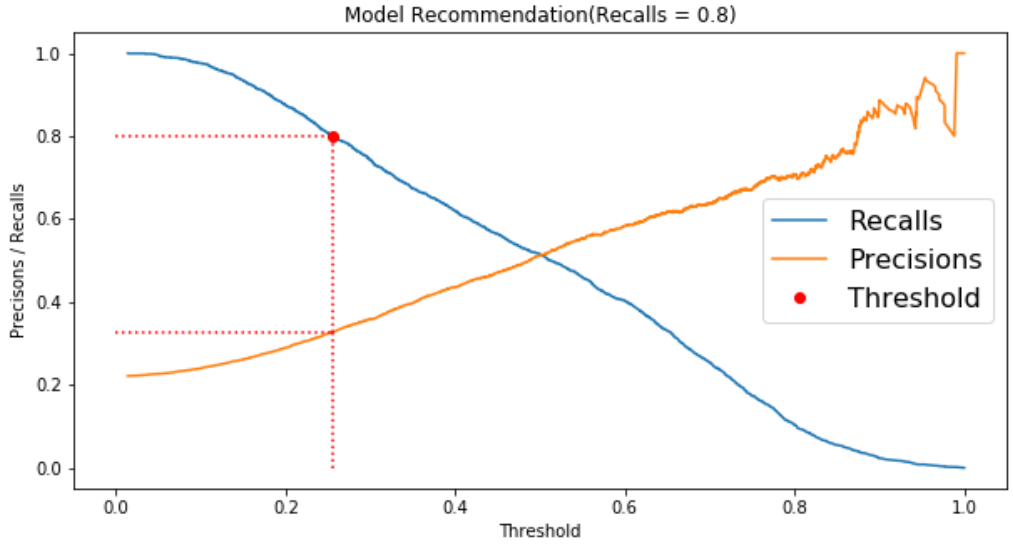
In these 3 models, Logistic Regression model has the highest recall but the lowest precision, if the firm expects high recall, then this model is the best candidate. If the balance of recall and precision is the most important metric, then Random Forest is the ideal model. Since Random Forest has slightly lower recall but much higher precision than Logistic Regression, we recommend the Random Forest model.



Recommendations Based on Modeling

Below is our suggested recall plot. Note the threshold can be adjusted to reach higher

recall.



Limitations

1. Best model Random Forest can only detect 51% of default.
2. Model can only be served as an aid in decision making instead of replacing human decision.

Future Work

1. Models are not exhaustive. Other models could perform better.
2. Get more computational resources to tune XGBoost parameters.
3. Acquire US customer data and more useful features.I.e.customer income.

Contributing Member:

Name	Email
Bharath P	Bharath0924.bp@gmail.com