Capstone Project-3

Credit-Card-Default-Prediction

Ansh Bhatnagar
Sandeep Kumar Maurya

Problems to resolve

Problem Statement

- ML applications focused on credit score predicting.
- Relying on credit scores and credit history.
- Miss valuable customers with no credit history. I.e. immigrants.
- Regulatory constraints on banking industry forbids some ML algorithms.

Purpose of Project

- Conduct quantitative analysis on credit default risk by applying three interpretable machine learning models without utilizing credit score or credit history.

Who Should Care?

Credit Card Companies

Commercial Banks

Approach Overview

Data Cleaning

Data Exploration

Predictive Modeling

Understand and Clean

- Find information on undocumented columns values
- Clean data to get it ready for analysis

Graphical and Statistical

- Exam data with visualization
- Verify findings with statistical tests

Machine Learning

- Logistic Regression
- Random Forest
- XGBoost

Data Acquisition

Dataset

- Default Payments of Credit Card Clients in Taiwan from 2005
- Source: Public dataset from Almabetter.
- Original Source: UCI Machine Learning Repository*

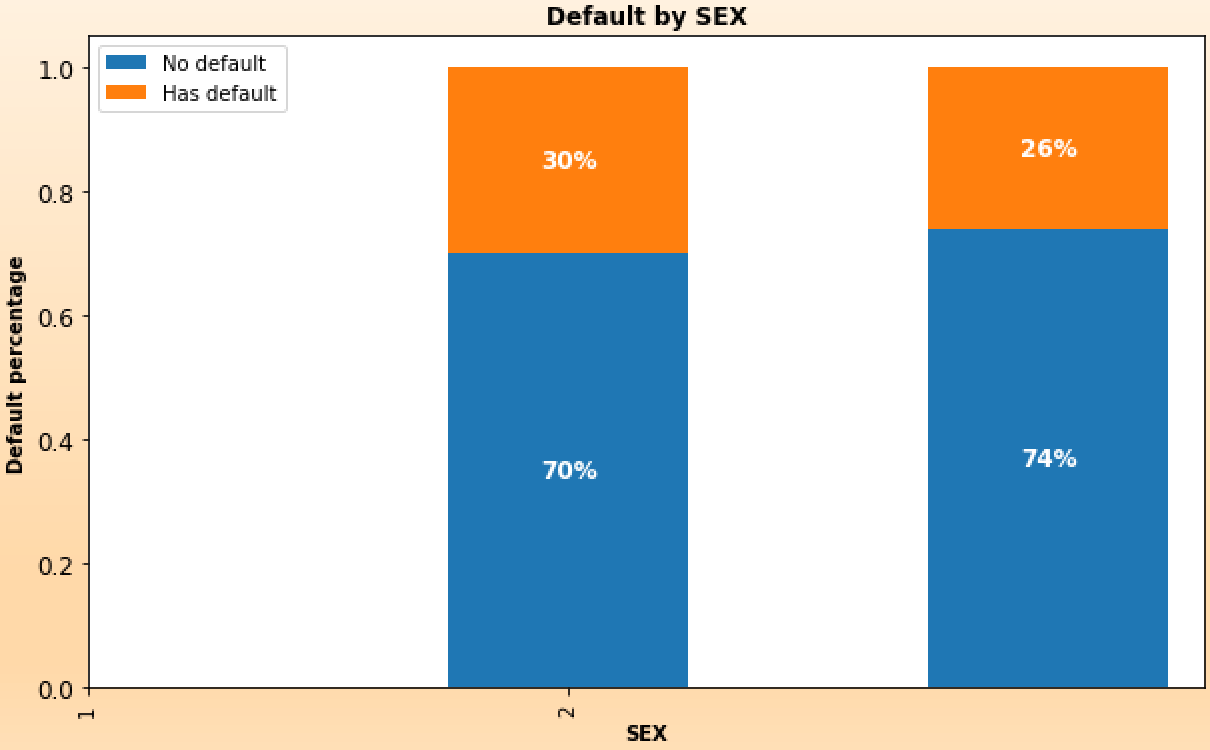
Why This Dataset?

- Real credit card data
- Comprehensive and complete
- 30,000 customers
- Usage of 6 months
- Age from 20-79
- Demographic factors
- No credit score or credit history

Part 1
Exploratory Data Analysis

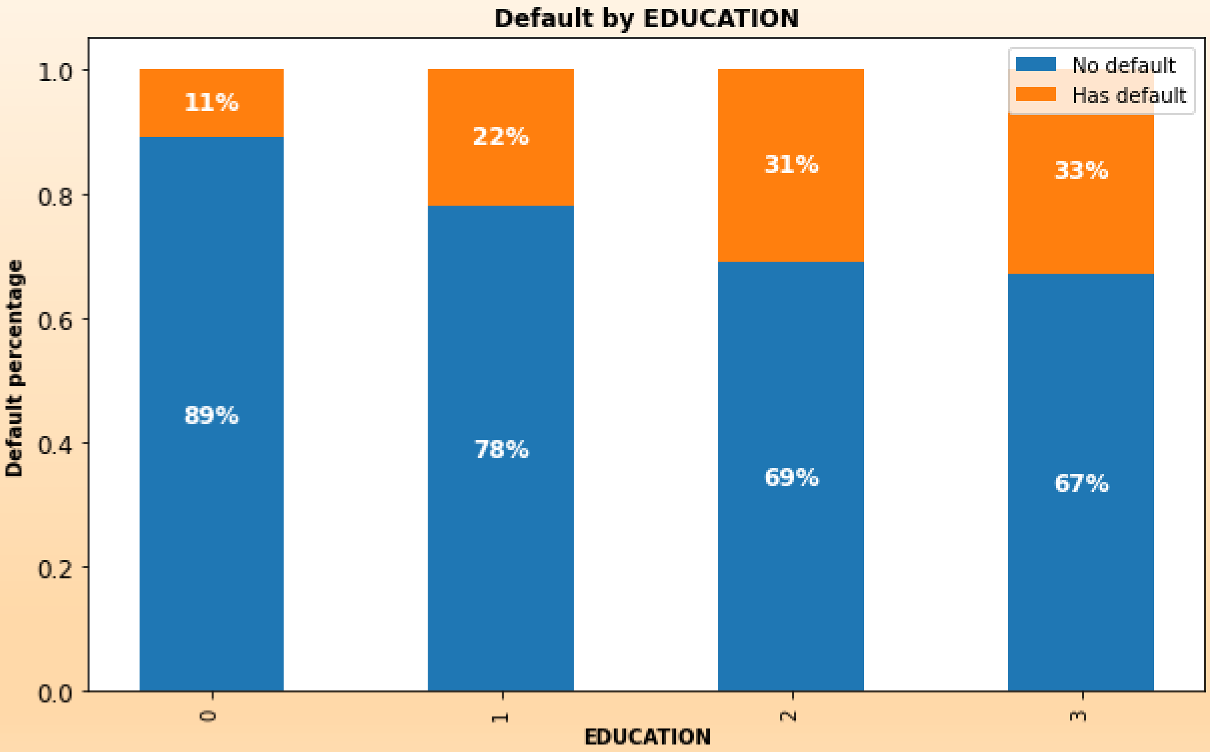
What demographic factors impact payment default risk?

Gender Variable



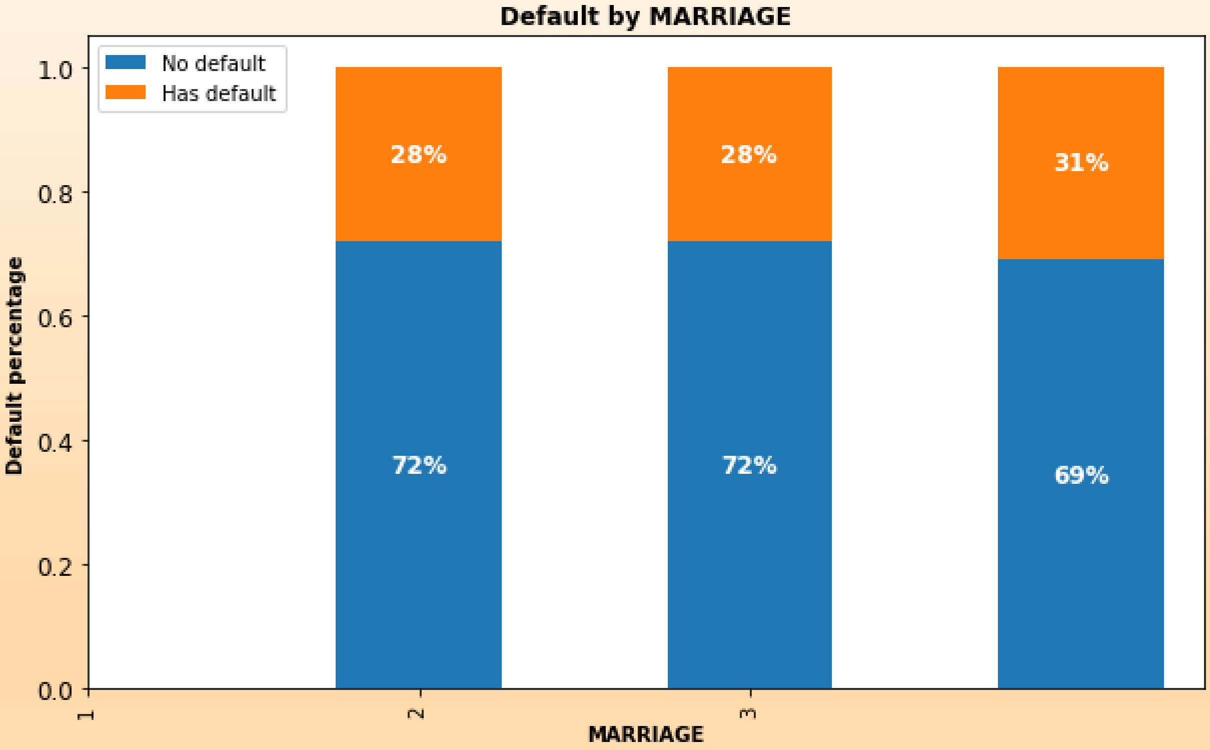
30% of males and **26%** of females have payment default.

Education Variable



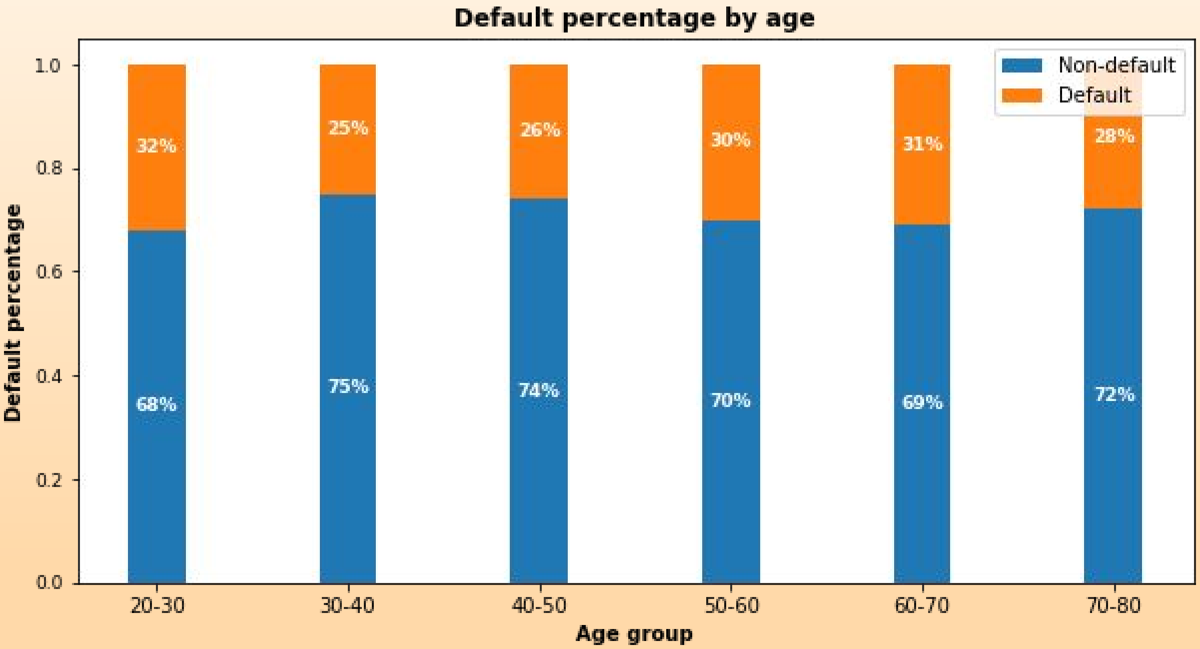
Higher education level, **lower** default risk.

Marital Status Variable



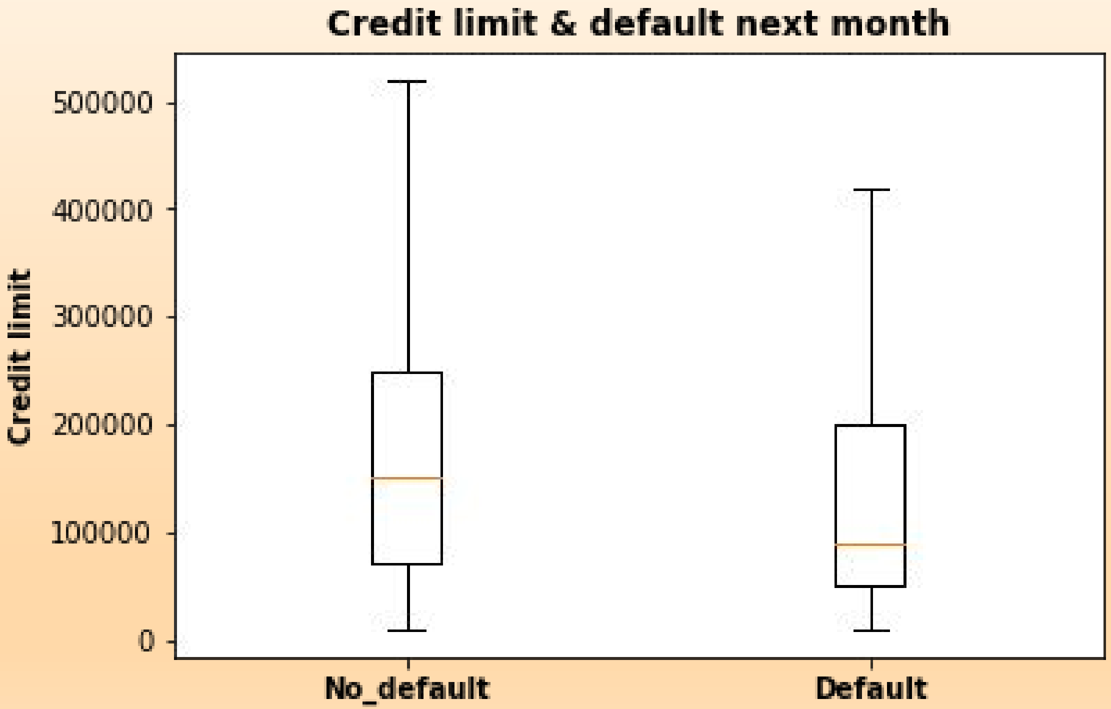
No significant correlations of default risk and marital status

Age Variable



30-50:
Lowest risk
< 30 or >50:
Risk increases

Credit Limit Variable



Higher credit limits,
lower default risk.

EDA Summary

- Demographic factors that impact default risk are:
  - Education: Higher education is associated with lower default risk.
  - Age: Customers aged 30-50 have the lowest default risk.
  - Sex: Females have lower default risk than males in this dataset.
  - Credit limit: Higher credit limit is associated with lower default risk.

Part 2
Predictive Modeling

What demographic factors impact payment default risk?

Modeling Overview

Define Problem: Supervised learning / binary classification

Imbalanced Classes: 78% non-default vs. 22% default

Tools Used: Scikit learn library and imblearn

Models Applied: Logistic Regression / Random Forest / XGBoost

Modeling Steps

Data Pre-processing

- Feature selection
- Feature engineering
- Train-test data splitting (70%/30%)
- Training data rescaling
- SMOTE oversampling

Fitting and Tuning

- Start with default model parameters
- Hyperparameters tuning
- Measure ROC_AUC on training data

Model Evaluation

- Models testing
- Precision_Recall score
- Compare with sklearn dummy classifier
- Compare within the 3 models

Correct Imbalanced Classes

- Fit every model without and with SMOTE oversampling for comparison.
- Training AUC scores improved significantly with SMOTE.

Models	AUC Without SMOTE	AUC With SMOTE
Logistic Regression	0.726	0.797
Random Forest	0.764	0.916
XGBoost	0.762	0.899

Hyperparameters Tuning

- **K-Fold Cross Validation** to get average performance on the folds.
- **Randomized Search** on Logistic Regression since C has large search space.
- **Grid Search** on Random Forest on limited parameters combinations.
- **Randomized Search** on XG Boost because multiple hyperparameters to tune.

Model Comparisons

- Compare the models to Scikit-learn's dummy classifier.
- All models performed better than dummy model.

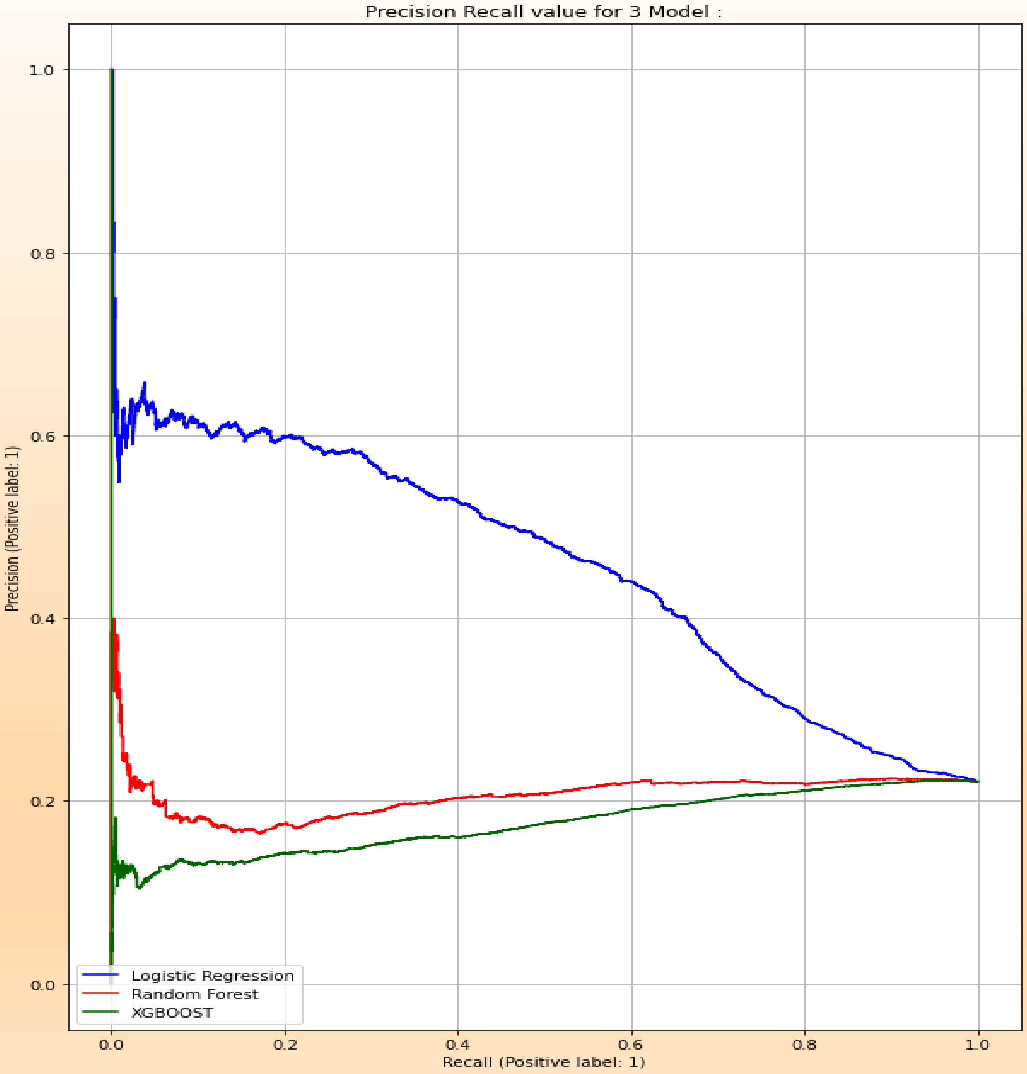
Models	Precision	Recall	F1 Score	Conclusion
Dummy Model	0.217	0.500	0.303	**Benchmark**
Logistic Regression	0.384	0.566	0.457	**Best recall**
Random Forest	0.513	0.514	0.514	**Best F1**
XGBoost	0.444	0.505	0.474	

Model Comparisons

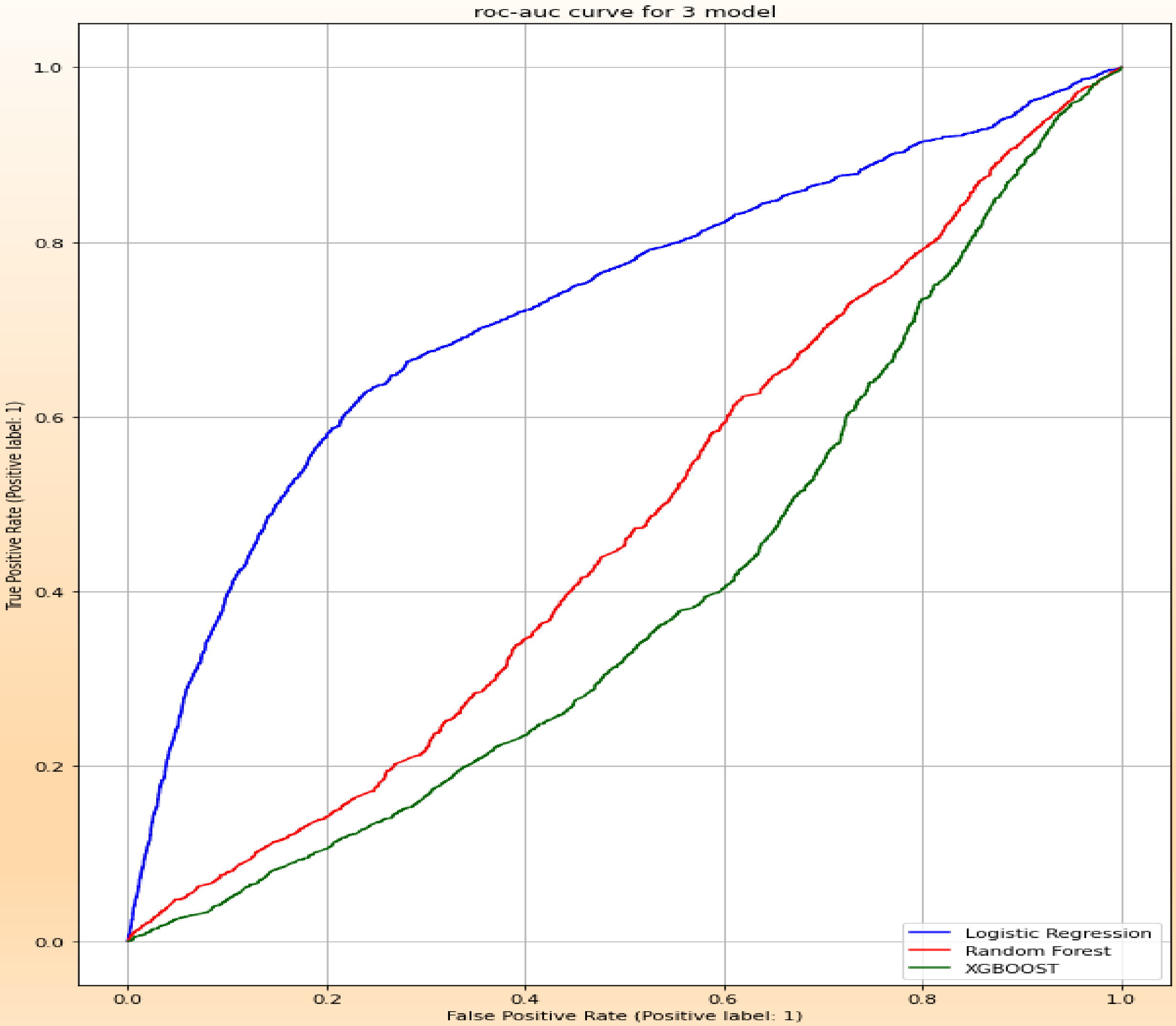
- Compare within 3 models.
- Random Forest (red line) has the best precision_recall score.

Terminology:

.
★ Recall: how many 1s are being identified?
★ Precision: Among all the 1s that are flagged, how many are truly 1s?
★ Precision and recall trade-off: high recall will cause low precision



ROC-AUC curve for 3-D Model



Limitations & Future Work

Limitations

- Best model Random Forest can only detect 51% of default.
- Model can only be served as an aid in decision making instead of replacing human decision.
- Used only 30,000 records and not from US consumers.

Future Work

- Models are not exhaustive. Other models could perform better.
- Get more computational resources to tune XG Boost parameters.
- Acquire US customer data and more useful features I.e. customer income.

Conclusions

- Recent 2 payment status and credit limit are the strongest default predictors.
- Dormant customers can also have default risk.
- Random Forest has the best precision and recall balance.
- Higher recall can be achieved if low precision is acceptable.
- Model can be served as an aid to human decision.
- Suggest output probabilities rather than predictions.
- Model can be improved with more data and computational resources.

Thank you!

Ansh Bhatnagar- 123anshbhatnagar@gmail.com
Sandeep Kumar Maurya- sandeepskm13@gmail.com

GitHub Link: -

Ansh Bhatnagar: - https://github.com/AnshRockstar/Credit-Card-Default-Analysis

Sandeep Kumar Maurya: - https://github.com/San13deep/Credit-Card-Default-Prediction