Credit Card Default Risk Analysis

A Case Study of 2 Classification Models

Sfiso Hlatshwayo

Why did you build a model?

Purpose of Project

- Conduct quantitative analysis on credit default risk by applying two interpretable machine learning models without utilizing credit score or credit history.
- To predict customers who would potentially default.

Who Should Care?

Credit Card Companies

Commercial Banks

* Image source:Google image

Approach Overview

Data Cleaning

Understand and Clean

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

Data Exploration

Graphical & Statistical

- Exam data with visualization
- Verify findings with statistical tests

Predictive Modeling

Machine Learning

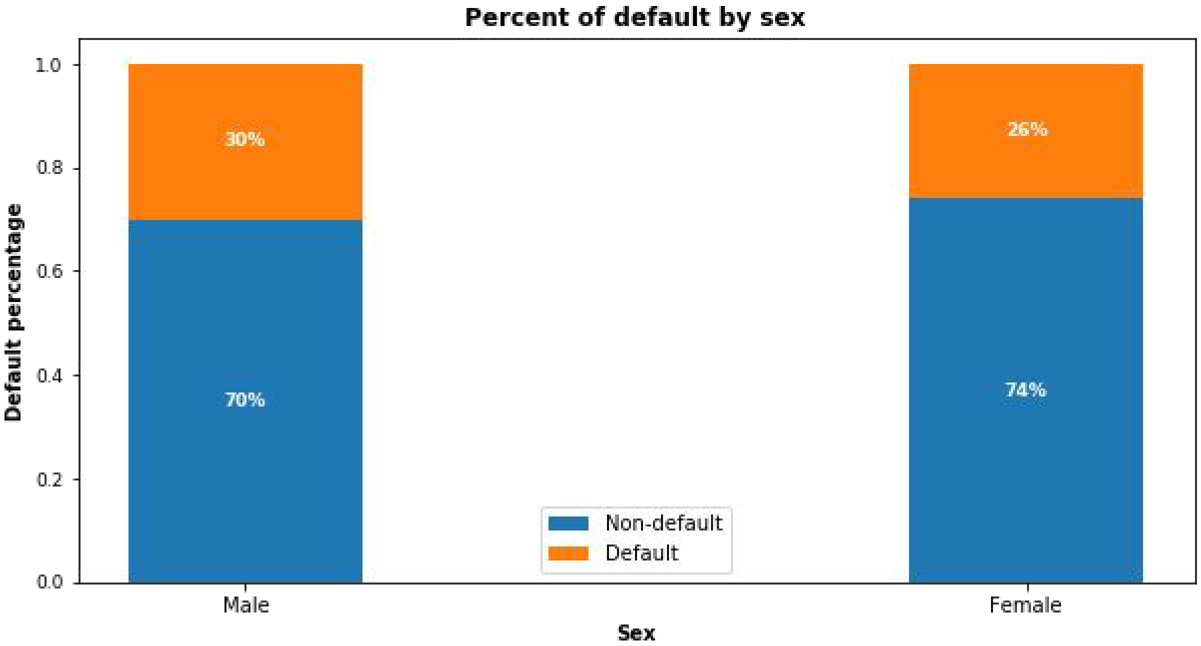
- Logistic Regression
- Random Forest
-

Part 1

Exploratory Data Analysis

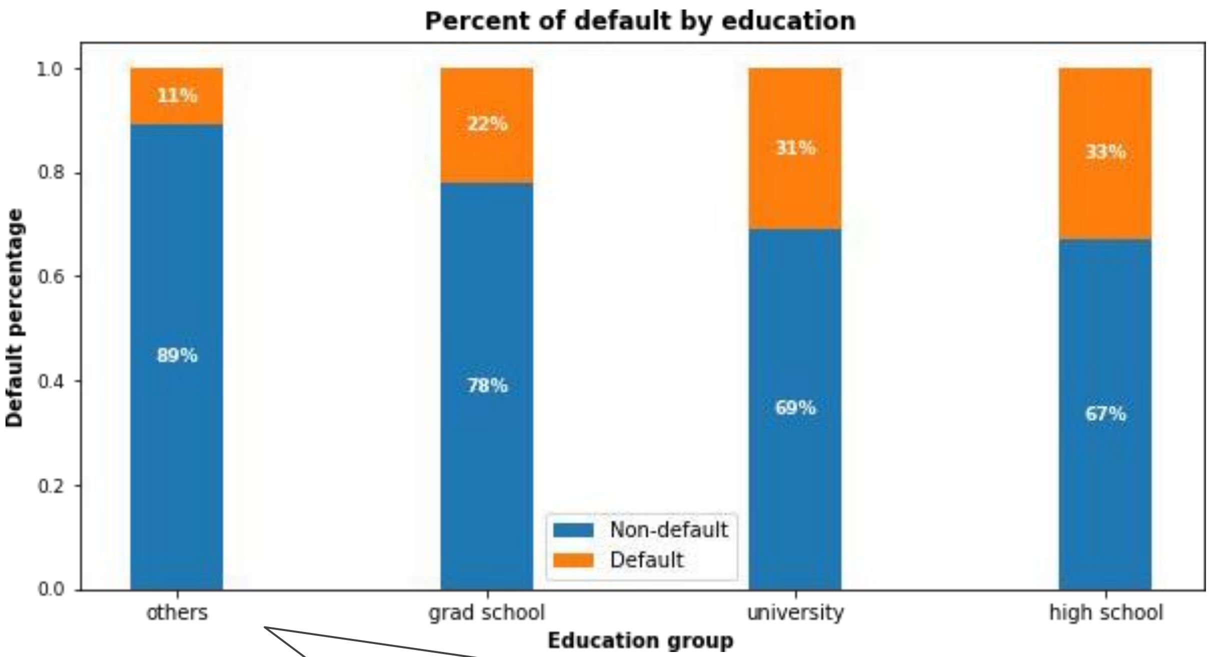
What demographics factors impact payment default risk?

Gender Variable



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

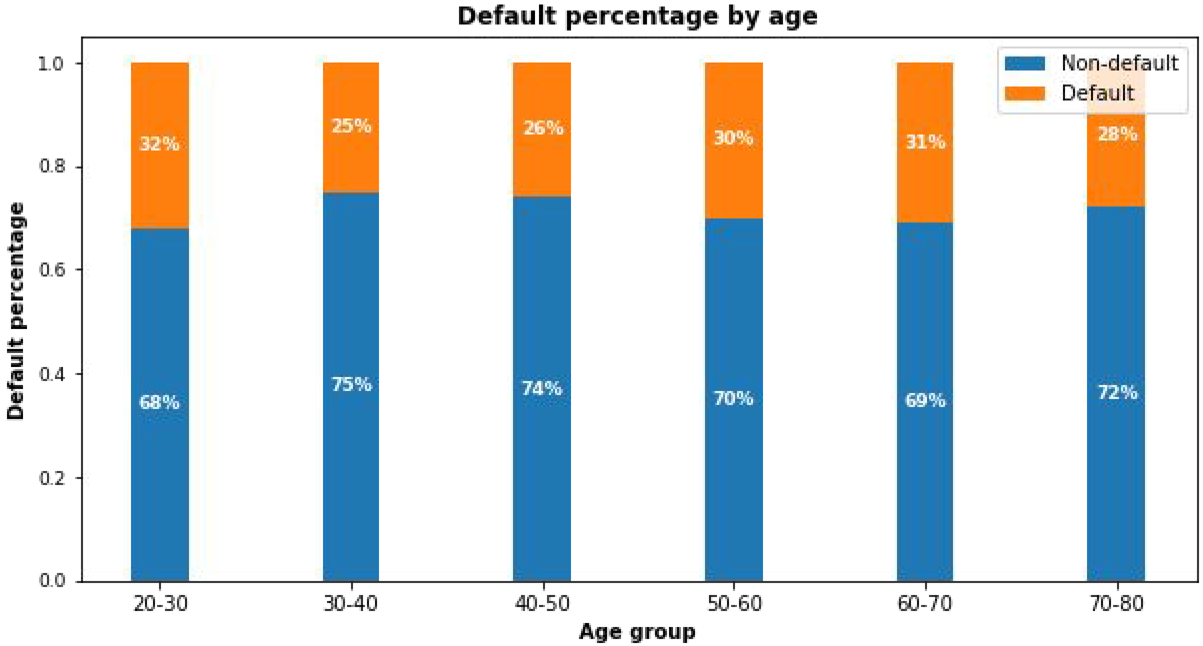
Education Variable



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

“Others” only consists 1.56% of total customers even if they appear to have the least default.

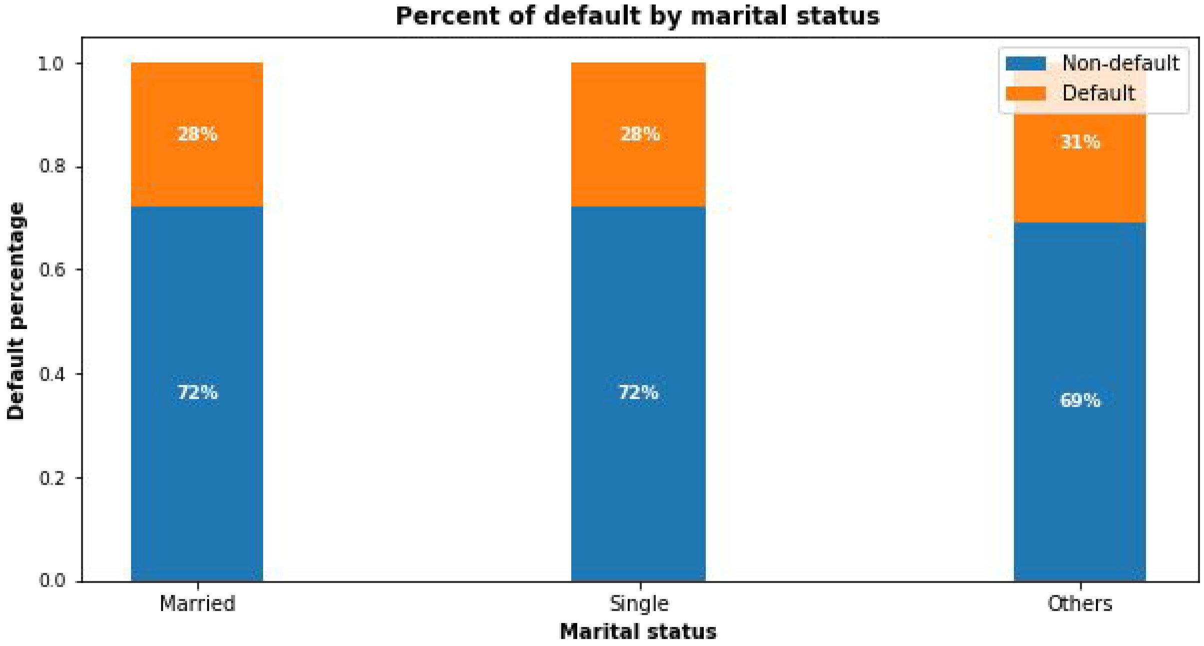
Age Variable



30-50:
Lowest risk

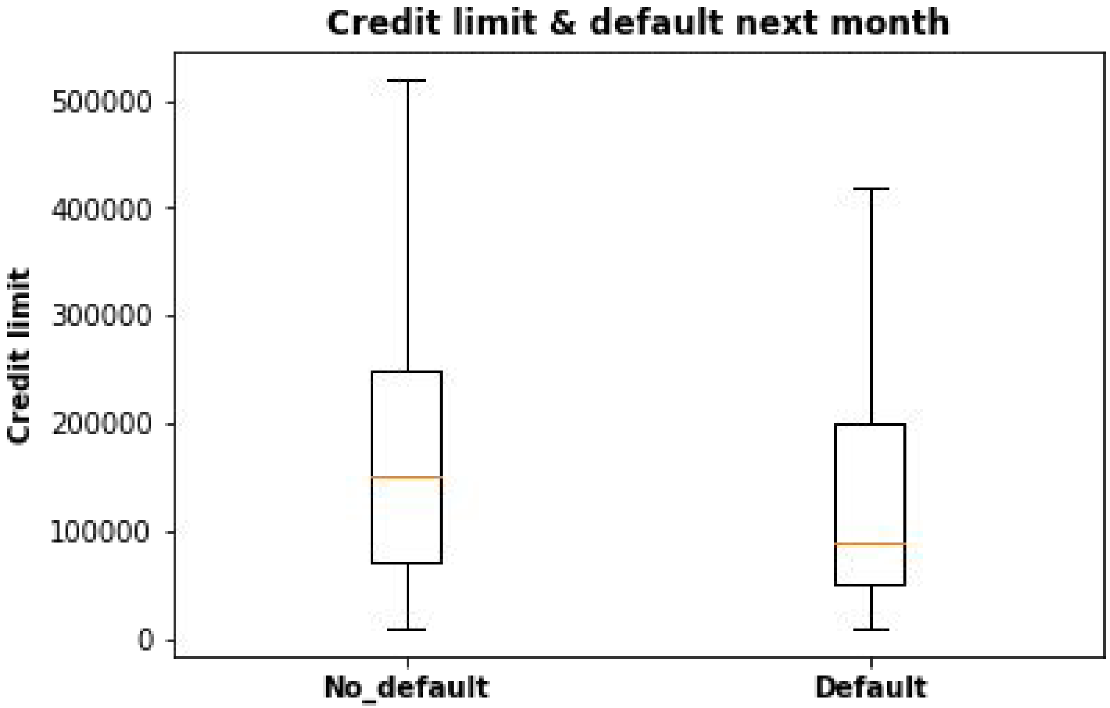
< 30 or>50:
Risk increases

Marital Status Variable



No significant correlations of default risk and marital status

Credit Limit Variable



Higher credit limits,

lower default risk

Part 2

Predictive Modeling

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

Modeling Steps

Data Preprocessing

- 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 2 models

Correct Imbalanced Classes

- Fit every model without and with SMOTE (synthetic minority oversampling technique) 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

Hyperparameters Tuning

- **Randomized Search** on Logistic Regression since C has large search space.
- **Grid Search** on Random Forest on limited parameters combinations.

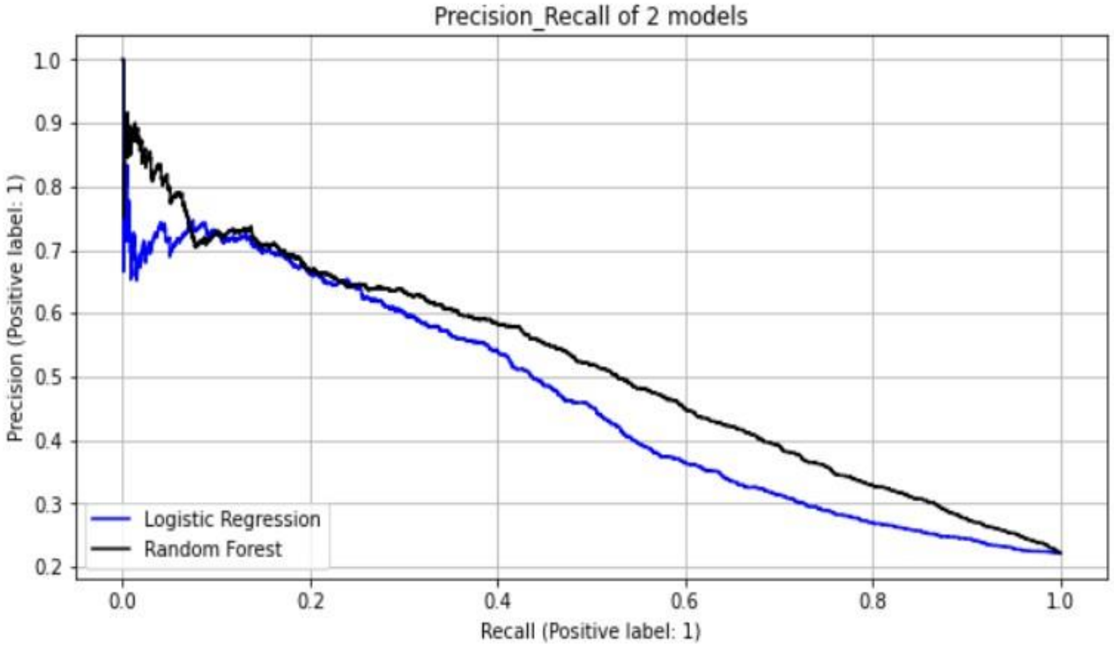
Model Comparisons

- 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**

Model Comparisons

- Compare within 2 models.
- Random Forest (black line) has the best precision_recall score.



Terminology/Layman terms:

★ Recall: Out of all the defaulters, how many did the model actually get correct?

Precision: How correct is the model based on

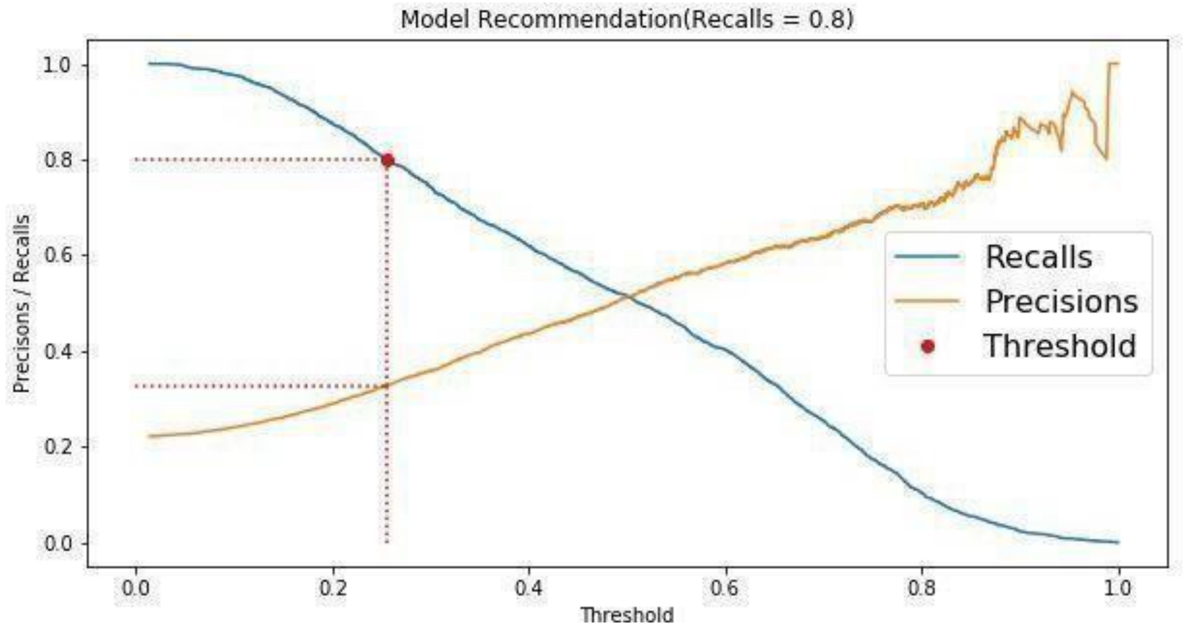
★ it's own predictions?

Precision and recall

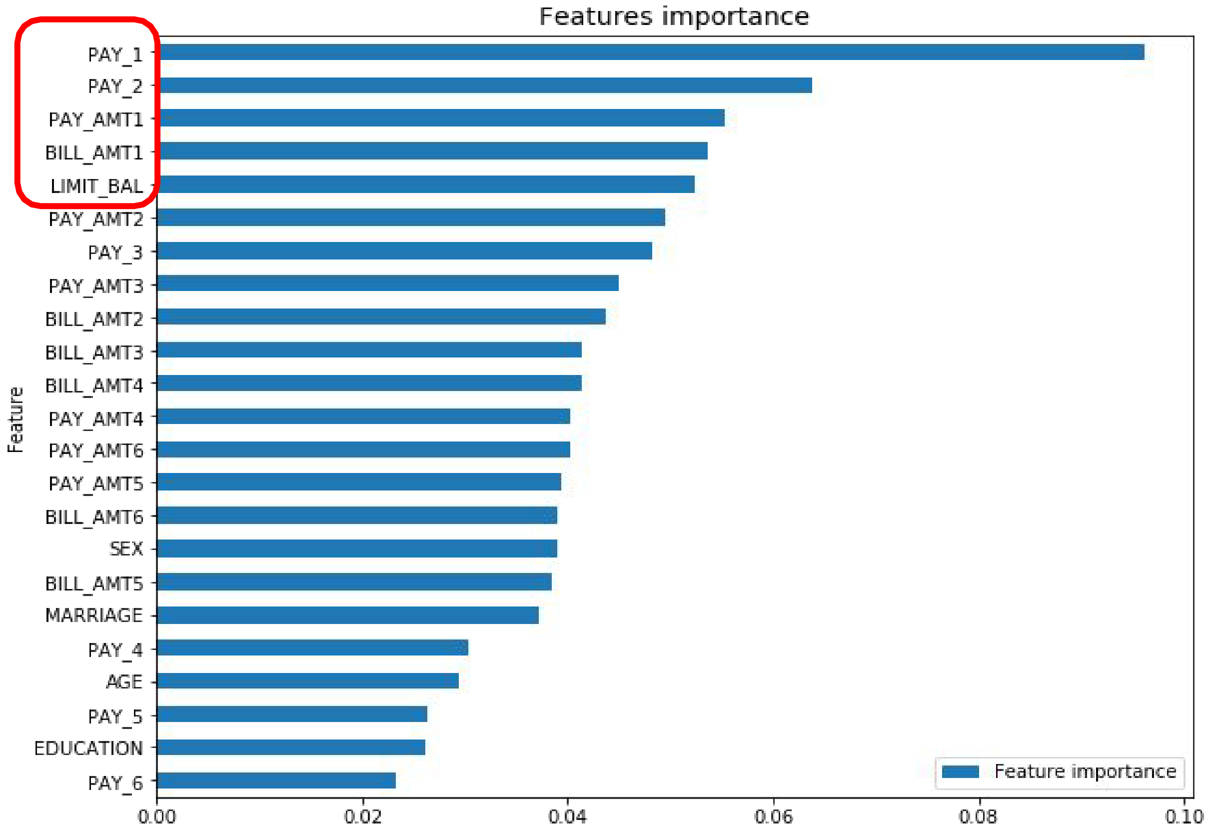
★ trade-off: high recall will cause low precision

Model Usage - Recommendation

- I.e. recall = 0.8. Threshold can be adjusted to reach higher recall.



Feature Importances



Best model Random Forest feature importances plot.

- ★ PAY_1: most recent month's payment status.
- ★ PAY_2: the month prior to current month's payment status.
- ★ BILL_AMT1: most recent month's bill amount.
- ★ LIMIT_BAL: credit limit

Limitations & Future

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.

Future Work

- Other models could perform better.
- Models such as Neural networks.
- More useful features.I.e.customer income.

Conclusions

- Recent 2 payment status and credit limit are the strongest default predictors.
- Random Forest has the best precision and recall balance.
- Higher recall can be achieved if low precision is acceptable.

Thank you