Credit Card Default Risk Analysis

- A Case Study of 3 Classification Algorithms

Project notebook: https://github.com/teresanan/capstone_project_1

Teresa Nan

June 2020

Table of Contents

Executive Summary **2**

1. Introduction **3**

1.1 Background 3

1.2 Method Overview 3

2. Data Exploration **4**

2.1 Gender Variable 4

2.2 Education Variable 5

2.3 Age Variable 7

2.4 Credit Limit Variable 7

2.5 Inactive Customers 8

3. Modeling **8**

3.1 Modeling Preparation 8

3.2 Predictive Modeling 9

3.3 Model Comparisons 11

4. Conclusions **12**

Appendix A **13**

Appendix B **13**

References **14**

Executive Summary

The purpose of this project is to conduct quantitative analysis on credit card default risk by applying 3 classification machine learning models. Despite machine learning and big data have been adopted by the banking industry, the current applications are mainly focused on credit score predicting. Heavily relying on credit scores could cause banks to miss valuable customers who are new immigrants with repaying power but little to no credit history. This analysis is a machine learning application on default risk itself and the predictor features do not include credit score or credit history. Due to the regulatory constraints that banks are facing, for example, The Fair Credit Reporting Act (FCRA), the algorithms used in this analysis are relatively simple and interpretable.

This analysis used a Kaggle public dataset that consists of 30,000 credit card usage records and 3 machine learning models - Logistic Regression, Random Forest and XGBoost.There might be other classification models that could yield better performances, due to the scope of the project, we did not cover other algorithms. Among the 3 models, Random Forest is the one with the best precision score as 0.512 and recall score as 0.515[1]. It may appear that these scores are not satisfactory, however, predicting default risk is an inherently challenging task and there is an inevitable trade-off between precision and recall. More importantly, this analysis is intended to be an aid to human decision by flagging high default risk customers, instead of automating the decision making.

From this study, we discovered a few interesting insights which may or may not hold for other datasets. We learned the most important predictors of default are not human characteristics, but the most recent 2 months' payment status and customers' credit limit.The conventional thinking of younger people tend to have higher default risk is proven to be only partially true in this dataset. Also, surprisingly, customers being inactive for months doesn't mean they have no default risk.

We understand creditors need to make decisions efficiently and in the meantime to abide by regulations, the machine learning models in this analysis can be served as an aid to credit card companies, loan lenders, and banks make informed decisions on creditworthiness based on accessible customer data. We suggest the model outputs probabilities rather than predictions, so that we can achieve higher accuracy and allow more control for human managers in decision making.

[1] Recall: The percentage of defaults are being successfully identified. Precision: Among all the flagged potential defaults, the percentage of truly defaults. Precision and recall trade-off: High recall will cause low precision and vice versa.

1. Introduction

1.1 Background

Credit risk has traditionally been the greatest risk among all the risks that the banking and credit card industry are facing, and it is usually the one requiring the most capital. This can be proven by industry business reports and statistical data. For example, "The Federal Reserve Bank of New York measures credit card delinquencies based on the percent of balances that are at least 90 days late. For the third quarter of 2019, that rate was about 8%, about the same level as in the previous quarter."[2] Thus, assessing, detecting and managing default risk is the key factor in generating revenue and reducing loss for the banking and credit card industry.

Despite machine learning and big data have been adopted by the banking industry, the current applications are mainly focused on credit score predicting. The disadvantage of heavily relying on credit score is banks would miss valuable customers who come from countries that are traditionally underbanked with no credit history or new immigrants who have repaying power but lack credit history. According to a literature review report on analyzing credit risk using machine and deep learning models, "credit risk management problems researched have been around credit scoring; it would go a long way to research how machine learning can be applied to quantitative areas for better computations of credit risk exposure by predicting probabilities of default." [3]

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.

1.2 Method Overview

- **Data**: We acquire the data from a public dataset from Kaggle. The original dataset can be found here at the UCI Machine Learning Repository. Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[2] https://www.creditcards.com/credit-card-news/credit-card-delinquency-statistics-1276.php

[3] Addo, P.M.; Guegan, D.; Hassani, B. Credit Risk Analysis Using Machine and Deep Learning Models. *Risks* **2018**, *6*, 38. https://www.mdpi.com/2227-9091/6/2/38

- **Models:** Due to the scope of the project and lack of computational resources, this analysis is not intended to be exhaustive, we only applied 3 classification machine learning models - Logistic Regression, Random Forest and XGBoost.

- **Tools:** The programming is done in Python. Scikit-learn Python libraries are utilized for machine learning modeling. For data analysis and visualization, we use Numpy, Pandas, imblean, matplotlib and seaborn.

2. Data Exploration

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. It includes 30,000 rows and 25 columns, and there is no credit score or credit history information. Data dictionary is available in Appendix A.

Overall, the dataset is very clean, but there are several undocumented column values. As a result, most of the data wrangling effort was spent on searching information and interpreting the columns. More details about the data cleaning can be found in this Jupyter Notebook.

The purpose of exploratory data analysis is to identify the variables that impact payment default likelihood and the correlations between them. We use graphical and statistical data exploratory analysis tools to check every categorical variable. Each starts with a visualization and is followed by a statistical test to verify the findings.

The main findings from exploratory analysis are as following:

- 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.
- Customers with higher education have less default payments and higher credit limits.
- 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.
- There appears to be no correlation between default payment and marital status.
- 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.

2.1 Gender Variable

Males have more delayed payments than females in this dataset.

Which gender group tends to have more delayed payment? Since there are more females than males in the dataset, we use percentage of default within each sex group. Figure 1 shows 30% males have default payment while only 26% females have default payment. The difference is not significant. To verify if this is due to chance, we use a permutation test and a t-test on each group's default proportions and mean respectively, and the results support the findings.

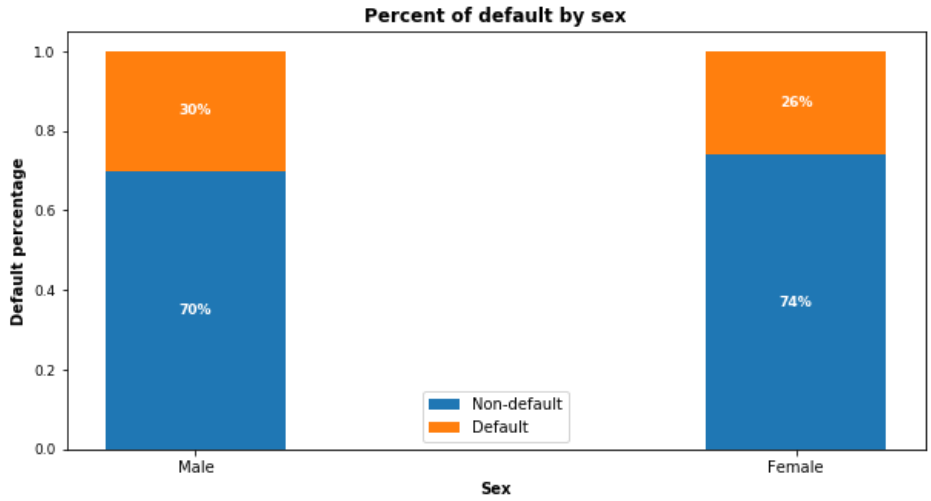

Figure 1: Percent of default by sex.

2.2 Education Variable

Customers with higher education have less delayed payment.

Figure 2 indicates customers with lower education levels default more. Customers with high school and university educational level have higher default percentages than customers with grad school education. Notice there is an education group “others” which appears to have the least default payment, but this group only has 468 (or 1.56%) customers, and we don’t know what consists of this group.

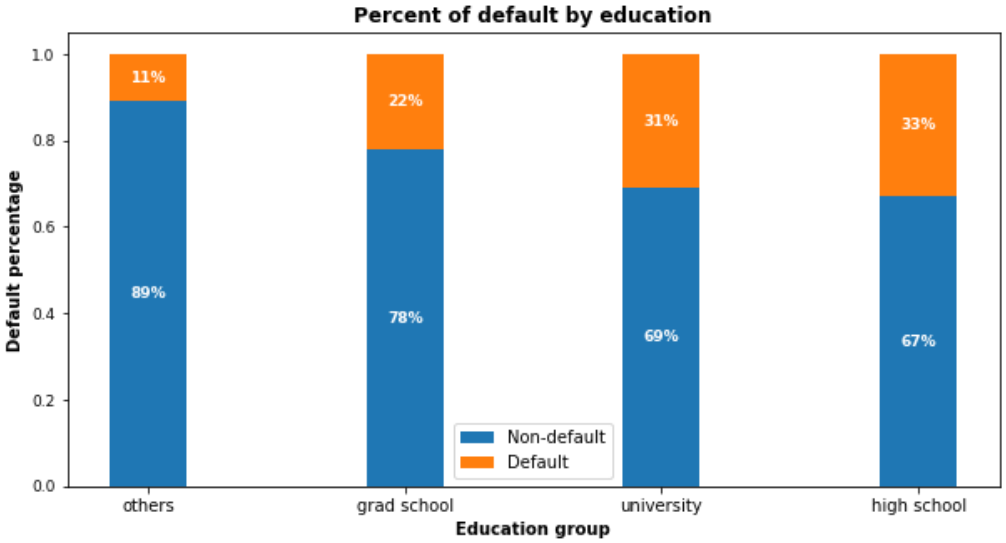

Figure 2: Percent of default by education level. The "others" group only consists of 1.56% of total customers.

Customers with a high education level get higher credit limits.

From the boxplot in figure 3, it is obvious that customers with grad school education have the highest 25% percentile, highest median, highest 75th percentile and highest maximum numbers, which suggests customers with higher education levels do get higher credit limits.

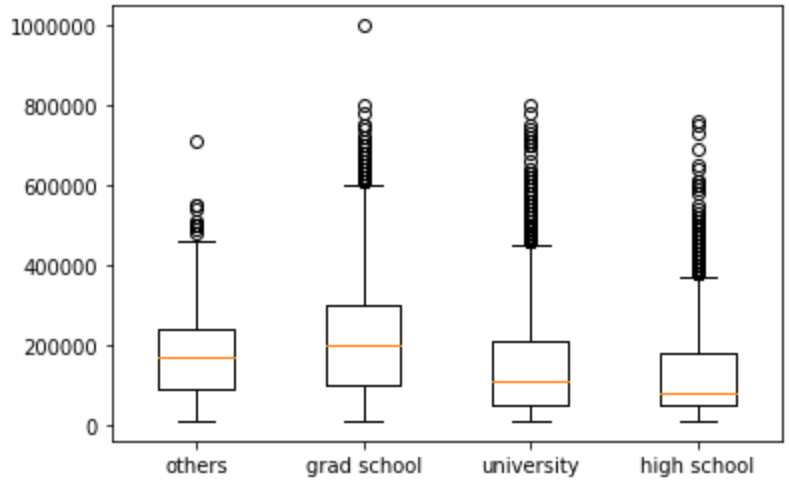

Figure 3: Education and credit limit boxplot

2.3 Age Variable

Middle-aged customers have the lowest default rate.

The bar chart in figure 4 shows the default probability increases for customers younger than 30 and older than 70. Customers aged between 30 and 50 have the lowest delayed payment rate, while younger groups (20-30) and older groups (50-70) all have higher delayed payment rates. This aligns with social reality that customers aged 30-50 typically have the strongest earning power. We also notice the delayed rate drops slightly again in customers older than 70. This is understandable because elder customers' consumption tends to decrease. Lastly, we use a Chi-squared test to verify this finding and the test statistics support it.

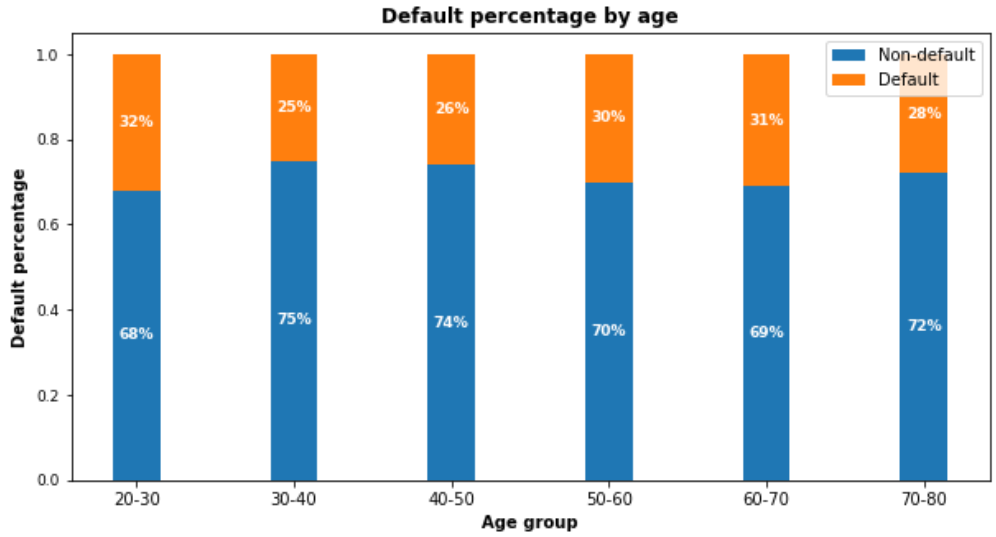

Figure 4: Percent of default by age group

2.4 Credit Limit Variable

Higher credit limit is associated with lower default risk.

Are there any correlations between credit limit and the default payment next month? Figure 5 gives us a clear answer. Unsurprisingly, customers with higher credit limits have lower delayed payment rates. We do a t-test to compare the mean of credit limits in these two groups and the test supports the hypothesis that higher credit limit is associated with lower default risk.

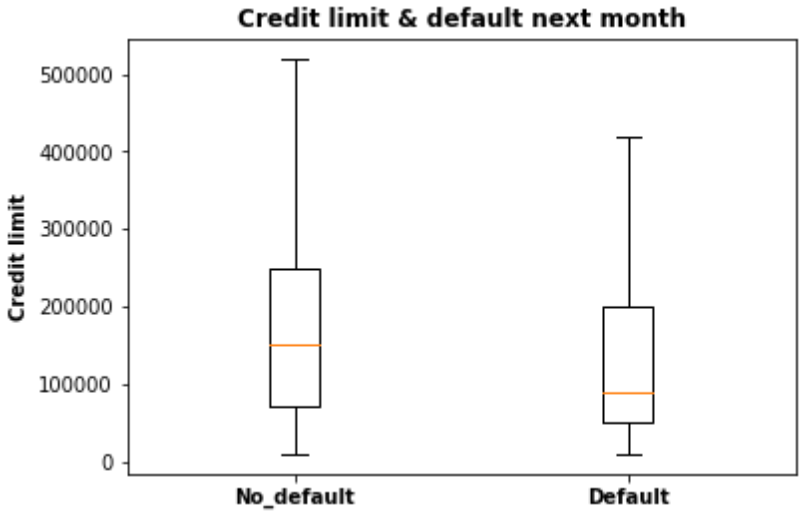

Figure 5: Credit limit and default next month

2.5 Inactive Customers

Customers who had no consumption in 6 months then default in the next month.

Do dormant customers have no default risk? We first detected 870 customers that were inactive for 6 months. Then we checked if these customers all had no default next month.To our surprise, 317 out of 870 inactive customers had default payment next month. We are not sure of the reasons behind this.

3. Modeling

3.1 Modeling Preparation

Since there are labeled data and the expected outcome is the probability of customer default, we define this as supervised machine learning and it is a binary classification problem. For better model performance, we first take a few preprocessing steps to prepare for modeling.

- **Feature Selection:**There are 25 columns in this dataset and the target variable is the column 'DEF_PAY_NMO' (means "default next month") . We drop the column 'ID' and 'DEF_PAY_NMO', save the rest 23 as predictor features. Those predictor variables include categorical variables such as sex, age, education level and marital stauts, along with numerical variables, such as payment status, credit limit, bill amount, etc. With this dataset, we don't need to do PCA or dimensionality reduction.

- **Check Class Imbalance**: It is common sense that most customers do not default. This dataset is likely to be dominated by 0s (non-default) with rare 1s (default). Imbalanced dataset will mislead machine learning algorithms and affect their performances. 'DEF_PAY_NMO' variable shows 22% of customers have default and 78% have no default. The class ratio is roughly 1:4. We consider this dataset is imbalanced and will use SMOTE oversampling technique after train-test data split to balance the data.

- **Transform Categorical Column**: In the dataset, 'AGE' column has continuous values which are the individual customer's age. In the business context, we are more concerned about the age groups than the specific age, so we bin the 'AGE' column to 6 bins - 21~29,30~39,40~49,50~59,60~69, and 70~79. Finally, we convert this column into numerical data type because sklearn does not accept categorical data type.

- **Data Rescaling:**The feature variables' value vary vastly. For example, the credit limit value is up to 100,000 NTD and the payment status only ranges from 0 to 8. In order to make all variables have similar ranges, so the Logistic Regression model can perform well in regularization, we rescale the training data. In this process, we make sure to only fit training data (X_train) and then transform training data and test data (X_train, X_test), instead of fit and transform the entire X (consists of X_train and X_test).

- **Split Training and Test Data**: For each model, we use the same ratio for training and test data split (70% for training, 30% for test) to ensure consistency. After splitting the data, we set the test data aside and leave it for the very end, which is the final testing after hyperparameter tuning.

3.2 Predictive Modeling

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.

- **SMOTE Oversampling:** In the initial model fitting, we start by using all models' default parameters. To compensate for the rare classes in the imbalance dataset, we use SMOTE(Synthetic Minority Over-Sampling Technique) method to over sample the minority class and ensure the sampling is not biased. What this technique does under the hood is simply duplicating examples from the minority class in the training dataset prior to fitting a mode. lAfter SMOTE sampling, the dataset has equal size of 0s and 1s. In order to verify if SMOTE improves models' performance, all 3 models are trained with SMOTE and without SMOTE. Below table shows the ROC_ AUC scores on training data

improved significantly with all models after over sampling with SMOTE. This proves SMOTE is an effective method in sampling imbalanced dataset.

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

Table 1: Model ROC_AUC score on training data with default parameters

- **Hyperparameters Tuning**: We utilize Scikit-Learn library's built in functions such as cross-validation, randomized search and grid search to make this process easier. In Logistic Regression, the only hyperparameter C penalizes a large number of features, reduces model complexity and prevents overfitting. We use randomized search to find the best C because C has a large search space and randomized search saves computing time. With Random Forest, there are many hyperparameters available for tuning, but we use most of the default settings in sklearn and only focus on a few. After creating a parameter grid, we use grid search to find the best parameters combinations. The third model XGBoost is known for its good performance on low-medium sized structured tabular data, but the downside is there are quite some hyperparameters to tune. We initially try grid search but this turns out to be not feasible because it requires substantial computational resources, then we switch to randomized search and find a suitable hyperparameters combination.

- **Performance Metrics**: Since this is a classification problem with imbalanced classes, accuracy is not the best metric because the data is dominated by non-default class, thus precision and recall is a better choice. In the credit card default risk business context, detecting as many defaults as possible is our ultimate goal because misclassifying a default as non-default is costly, therefore a high recall score is the best metric. However, there is a known trade-off between precision and recall. We can raise recall to abiturally high, but the precision will decrease. We use below metrics to measure model performances.
    a. Confusion matrix
    b. ROC_AUC curve
    c. Precision_recall curve

- **Feature Importance**: By plotting the feature importance on tuned Random Forest model, it is clear that 'PAY_1','PAY_2' (the most recent 2 months' payment status), along with credit limit(LIMIT_BAL) are the most important predictors. Since we don't have customer income data, generally speaking, higher credit limits are associated with lower default risk.

3.3 Model Comparisons

- **Compare to Sklearn.DummyClassifier**. In a real business context, we should have a benchmark to measure a predictive model's performance. For example, we could compare model performance to existing ways of making the same classification. Since we don't have any benchmark in this project, we compare the models to Scikit-Learn's dummy classifier. As shown in table 1, all 3 models have better classification performance than the dummy model, which suggests our modeling has some merit.

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	

Table 2: Model Performance Comparison

- **Compare within the 3 models**. Logistic Regression has the highest recall but also the lowest precision. Random Forest outperforms Logistic Regression and XGBoost if measured on their F1 scores, which is the balance between precision and recall. XGBoost has a decent performance but it takes the most time to tune the model.

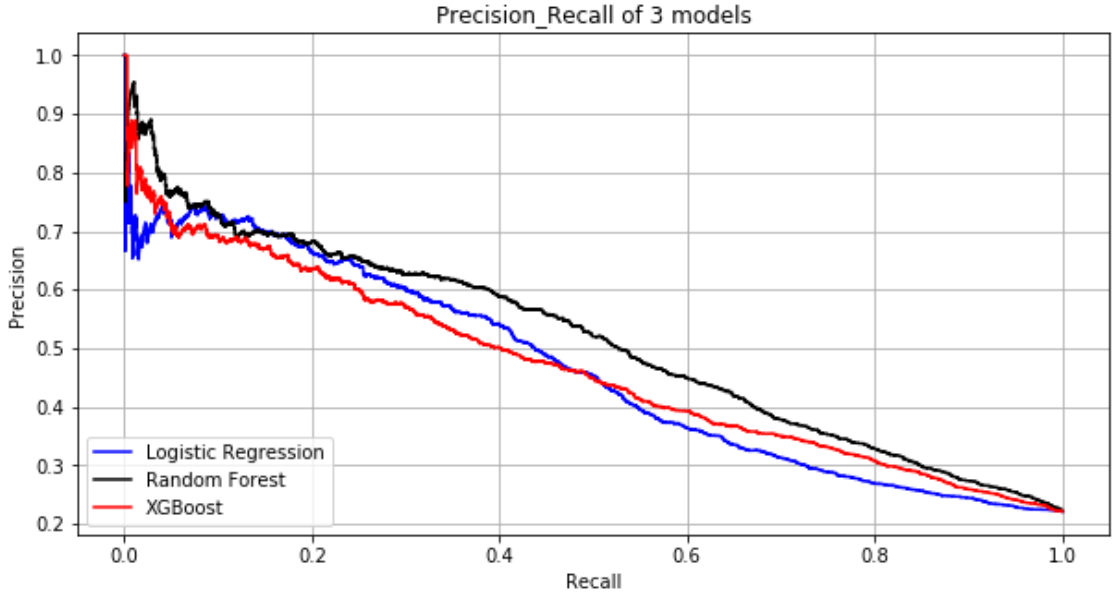

Figure 6: Precision and Recall curve of 3 models

4. Conclusions

Based on the exploratory data analysis, we discover that human characteristics are not the most important predictors of default, the payment status of the most 2 months and credit limit are. From the modeling, we are able to classify default risk with accessible customer data and find a decent model.

- **Suggestions to Clients:** With every classification model, there is a general trade-off between precision and recall. A model's recall can be adjusted to arbitrarily high at the cost of lower precision. 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.

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

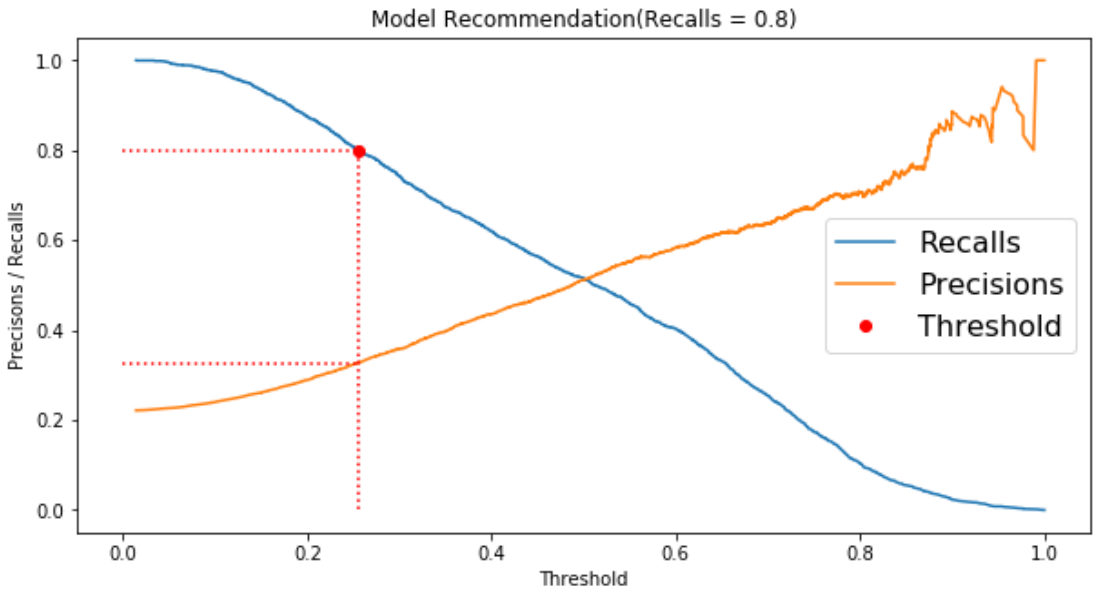

Figure 7: Suggested recalls and precisions

- **Limitations:** The dataset only includes 30,000 records of non-American consumers, there might be differences between US consumers and non-US consumers. Even with the best model Random Forest, we can only detect 51.5% of default customers, and among those that are being flagged as default, only 51.2% of them indeed have default.

Therefore, this model can only be served as an aid in decision making instead of replacing human decision. Lastly, we suggest the model output probabilities rather than predictions, so that we can achieve higher accuracy and allow more control for human managers to quantify default risk.

- **Future Work:** Due to the scope of this project and lack of computational resources, this study is far from exhaustive, there are other classification models that could perform better, we will leave it for future work.

Appendix A

Data Dictionary

There are 25 variables in this dataset:
• ID: ID of each client
• LIMIT_BAL: Amount of given credit in NT dollars (includes individual and family/supplementary credit
• SEX: Gender (1=male, 2=female)
• EDUCATION: (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
• MARRIAGE: Marital status (1=married, 2=single, 3=others)
• AGE: Age in years
• PAY_1: Repayment status in September, 2005 (-1=pay duly, 1=payment delay for one month, 2=payment delay for two months, … 8=payment delay for eight months, 9=payment delay for nine months and above)
• PAY_2: Repayment status in August, 2005 (scale same as above)
• PAY_3: Repayment status in July, 2005 (scale same as above)
• PAY_4: Repayment status in June, 2005 (scale same as above)
• PAY_5: Repayment status in May, 2005 (scale same as above)
• PAY_6: Repayment status in April, 2005 (scale same as above)
• BILL_AMT1: Amount of bill statement in September, 2005 (NT dollar)
• BILL_AMT2: Amount of bill statement in August, 2005 (NT dollar)
• BILL_AMT3: Amount of bill statement in July, 2005 (NT dollar)
• BILL_AMT4: Amount of bill statement in June, 2005 (NT dollar)
• BILL_AMT5: Amount of bill statement in May, 2005 (NT dollar)

• BILL_AMT6: Amount of bill statement in April, 2005 (NT dollar)
• PAY_AMT1: Amount of previous payment in September, 2005 (NT dollar)
• PAY_AMT2: Amount of previous payment in August, 2005 (NT dollar)
• PAY_AMT3: Amount of previous payment in July, 2005 (NT dollar)
• PAY_AMT4: Amount of previous payment in June, 2005 (NT dollar)
• PAY_AMT5: Amount of previous payment in May, 2005 (NT dollar)
• PAY_AMT6: Amount of previous payment in April, 2005 (NT dollar)
• DEF_PAY_NMO: Default payment (1=yes, 0=no)

Appendix B

Marital Status Variable

This marital status variable and default payment chart plot shows marital status has no impact on default risk.

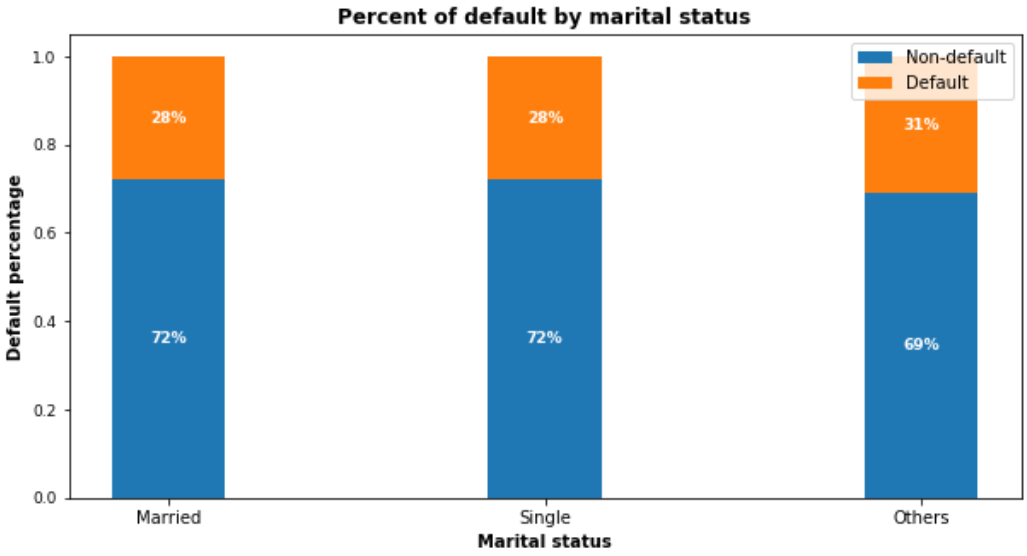

Figure 7: Percent of default by marital status. "other" group only consists of 0.18% of the dataset.

References

1. Addo, P.M.; Guegan, D.; Hassani, B. Credit Risk Analysis Using Machine and Deep Learning Models. Risks 2018, 6, 38. https://www.mdpi.com/2227-9091/6/2/38

2. Simeon Kostadinov: "My Analysis from 50+ papers on the Application of ML in Credit Lending": https://towardsdatascience.com/my-analysis-from-50-papers-on-the-application-of-ml-in-credit-lending-b9b810a3f38
3. Khyati Mahendru: "How to Deal with Imbalanced Data using SMOTE ":https://medium.com/analytics-vidhya/balance-your-data-using-smote-98e4d79fcddb