**Credit Card Default Risk Analysis**

**Akmal Jah Quamri**

**Data Science Trainees,**

**AlmaBetter**

**Abstract:**

The purpose of this project is to conduct quantitative analysis on credit card default risk by applying classification machine learning algorithms. Since 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 Report Act (FCRA), the algorithm used in this analysis are relatively simple and interpretable.

This dataset consists of 30000 credit card usage records and 4 machine learning models – Logistic Regression, Decision Tree Classifier, Random Forest Classifier, XGBoost Classifier. There might be other classification models that could yield better performances, due to the scope of the work, we did not cover other algorithms. Among the 4 models, XGBoost is the one with the best precision and recall scores. 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 tends 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. **Introduction**

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. 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.

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. **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.

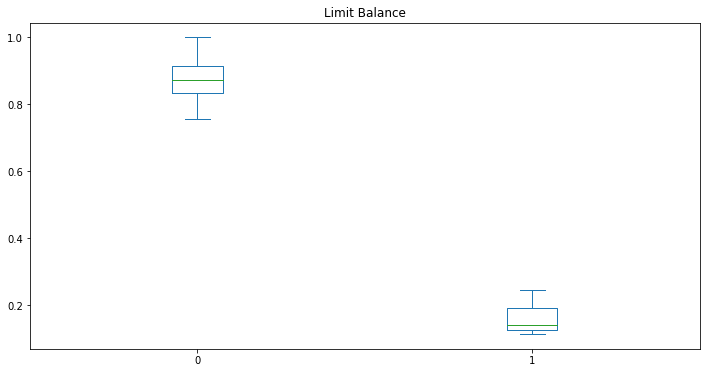
The purpose of exploratory data analysis is to identify the variables that impact payment default likelihood and the correlations between them. We use graphical and statistical data exploratory analysis tools to check every categorical variable. Each starts with a visualization and is followed by a statistical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

**The main findings from exploratory analysis are as following:**

* The lower the limit the higher the chances for being default
* With highest limit no default is recorded
* Older people will tend to fall more in default category
* Male has higher chances of being default compared to female.
* Higher educated persons have less rate to be default whereas lower educated will maximum chances to be default.

**2.1 The lower the limit the higher the chances for being default**

Are there any correlations between credit limit and the default payment next month? Figure below gives us a clear answer. Unsurprisingly, customers with higher credit limits have lower delayed payment rates.



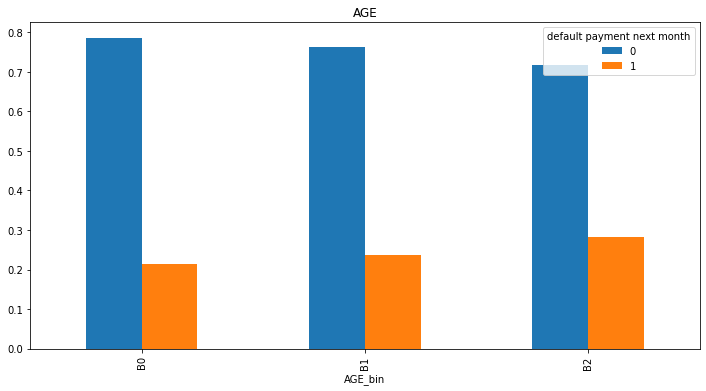
0 – No default

1 – Default

* 1. **Older people will tend to fall more in default category**

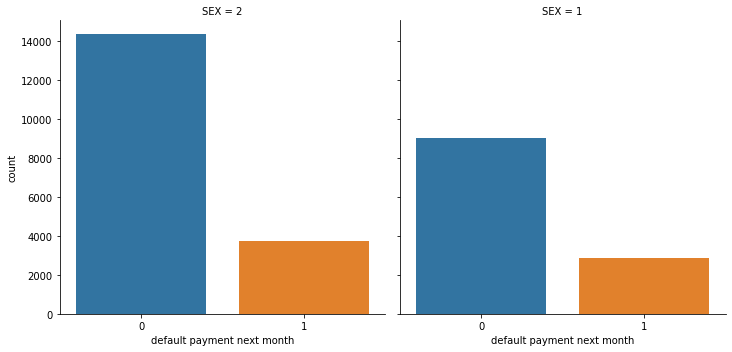
The bar chart in figure shows the default probability increases for customers younger than 30 and older than 70. Customers aged between 21 and 40 have the lowest delayed payment rate.

B0 – 21 – 40, B1 – 40 – 60, B2 – 60 – 79

****

* 1. **Male has higher chances of being default compared to female**

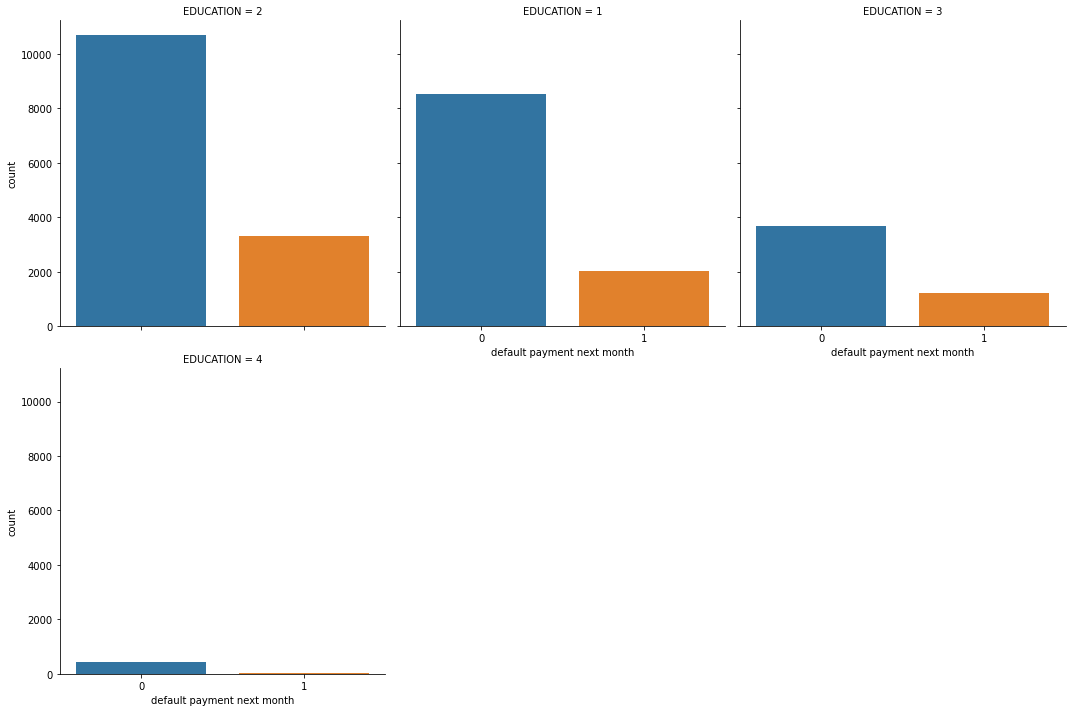
Which gender group tends to have more delayed payment? Since there are more females than males in the dataset, we use percentage of default within each sex group. Figure shows 30% males have default payment while only 26% females have default payment. The difference is not significant.

****

SEX – 1 – Male, SEX – 2 – Female

* 1. **Customer with higher education have less delay payment**

Figure 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. EDUCATION – 1 – Graduate School, EDUCATION – 2 – University, EDUCATION – 3 – High School, EDUCATION – 4 – Other.



1. **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.

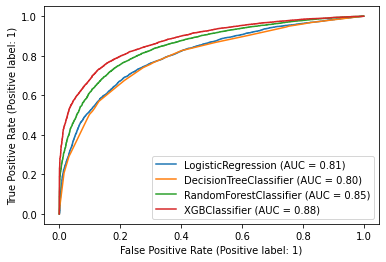
* **Feature Selection:** There are 25 columns in this dataset and the target variable is the column ‘DEF\_PAY\_NMO’ (means “default next month”) . We drop the column ‘ID’ and ‘DEF\_PAY\_NMO’, save the rest 23 as predictor features. Those predictor variables include categorical variables such as sex, age, education level and marital status, 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.
* **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.
* **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.
* **Data Rescaling:** The feature variables’ value varies 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).
* **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 ADASYN oversampling technique after train-test data split to balance the data.
  1. **Predictive Modeling**

This analysis uses 4 classification models - Logistic Regression, Random Forest, Decision Tree and XGBoost. Since Random Forest and XGBoost are tree based on algorithms, rescaling is only performed on Logistic Regression. For each model, we first try the model’s default parameters, train each model without handling imbalance and with imbalance handling 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.

* **ADASYN:** 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 ADASYN 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. After ADASYN sampling, the dataset has equal size of 0s and 1s. In order to verify if ADASYN improves models’ performance, all 4 models are trained with ADASYN and without ADASYN. Below table shows the ROC\_ AUC scores on training data improved significantly with all models after over sampling with ADASYN. This proves ADASYN is an effective method in sampling imbalanced dataset.

|  |  |  |
| --- | --- | --- |
| Models | AUC score without ADASYN | AUC with ADASYN |
| Logistic Regression | 0.73 | 0.81 |
| Decision Tree | 0.77 | 0.80 |
| Random Forest | 0.78 | 0.85 |
| XGBoost | - | 0.88 |

* **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 arbitrarily high but the precision will decrease. We use below metrics to measure model performances.
* Confusion matrix
* ROC\_AUC curve
* Classification Report



1. **Conclusions**

* XGBoost classifier is performing great with roc\_auc score 0.88, whereas randomforest is second highest with 0.85 score, and decision tree classifier is with least score of 0.80.
* From the above reports we can see that xgboost is giving good accuracy precision and recall as well. In all way xgboost is performing well. Since our aim highly interested in finding positive class so we will lean towards recall in this case and XGboost is giving high recall values.
* We will go ahead and deploy XGBoost model with handling class imbalance