Credit Card Default Prediction

Abhay Deep Singh

Data Science Trainees

AlmaBetter, Bangalore

Abstract:

This project is aimed at predicting the case of customers default payments in Taiwan. From the perspective of risk management, the result of predictive accuracy of the estimated probability of default will be more valuable than the binary result of classification - credible or not credible clients. We can use the K-S Chart to evaluate which customers will default on their credit card payments.

Financial threats are displaying a trend about the credit risk of commercial banks as the incredible improvement in the financial industry has arisen. In this way, one of the biggest threats faces by commercial banks is the risk prediction of credit clients. Recent studies mostly focus on enhancing the classifier performance for credit card default prediction rather than an interpretable model. In classification problems, an imbalanced dataset is also crucial to improve the performance of the model because most of the cases lied in one class, and only a few examples are in other categories. Traditional statistical approaches are not suitable to deal with imbalanced data. In this study, a model is developed for credit default prediction by employing various credit-related datasets. There is often a significant difference between the minimum and maximum values in different features, so Min-Max normalization is used to scale the features within one range.

Problem Statement:

This project is aimed at predicting the case of customers default payments in Taiwan. From the perspective of risk management, the result of predictive accuracy of the estimated probability of default will be more valuable than the binary result of classification - credible or not credible clients. We can use the K-S Chart to evaluate which customers will default on their credit card payments.

Introduction:

Before your account goes into default, it will become delinquent. This happens after missing a payment for 30 days. Default usually happens after six months in a row of not making at least the minimum payment due, which means your credit card is seriously delinquent.

Delinquency means that you are behind on payments. Once you are delinquent for a certain period of time (usually nine months for federal loans), your lender will declare the loan to be in default. The entire loan balance will become due at that time.

Before your account goes into default, it will become delinquent. This happens after missing a payment for 30 days. Default usually happens after six months in a row of not making at least the minimum payment due, which means your credit card is seriously delinquent. During that time you will be contacted by your creditor and they will want to know what you're going to do about it. If they are not satisfied with your response (or the lack thereof), the account will be closed and after 180 days with no payment, reported as charged off to the credit bureaus.

Once your account is delinquent, your credit score is going to be negatively impacted. When your account is charged off, the damage becomes more serious. Late payments are not immediately reported—for instance, if you miss your due date by a few days. It typically happens when you're at least 30 days late. But once that happens, your score is in jeopardy of starting on a downward spiral. Every month of non-payment will result in another hit to your score, as the delinquency ages and balances owed increase due to interest and fees. And an actual default will likely result in the card's line of credit being closed, which will cause your credit utilization rate to soar to 100 percent on the account. It may be turned over to a collection agency, putting you in another category of bad for your score. (One notable exception to this rule is medical bills. They are not reported as past due until they're six months old.

Data Description:

Attribute Information:

This research employed a binary variable, default payment (Yes = 1, No = 0), as the response variable. This study reviewed the literature and used the following 23 variables as explanatory variables.

- **X1:** Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.
- **X2:** Gender (1 = male; 2 = female).
- **X3:** Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
- **X4:** Marital status (1 = married; 2 = single; 3 = others).
- **X5:** Age (year).
- **X6 - X11:** History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows:
- **X6** = the repayment status in September, 2005;
- **X7** = the repayment status in August, 2005;
- **X11** = the repayment status in April, 2005. The measurement scale for the repayment status is:
- **-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.
- **X12-X17:** Amount of bill statement (NT dollar).
- **X12** = amount of bill statement in September, 2005;
- **X13** = amount of bill statement in August, 2005; . . .;
- **X17** = amount of bill statement in April, 2005.
- **X18-X23:** Amount of previous payment (NT dollar).
- **X18** = amount paid in September, 2005;
- **X19** = amount paid in August, 2005; . . .;
- **X23** = amount paid in April, 2005.

Step involved:

- Exploratory Data Analysis

  Exploratory data analysis (EDA) is a term for certain kinds of initial analysis and findings done with data sets, usually early on in an analytical process. Some experts describe it as “taking a peek” at the data to understand more about what it represents and how to apply it. Exploratory data analysis is often a precursor to other kinds of work with statistics and data.

- Encoding of Categorical columns

We used One Hot Encoding to produce binary integers of 0 and 1 to encode our categorical features because categorical features that are in string format cannot be understood by the machine and needs to be converted to numerical format.

- Standardization of features

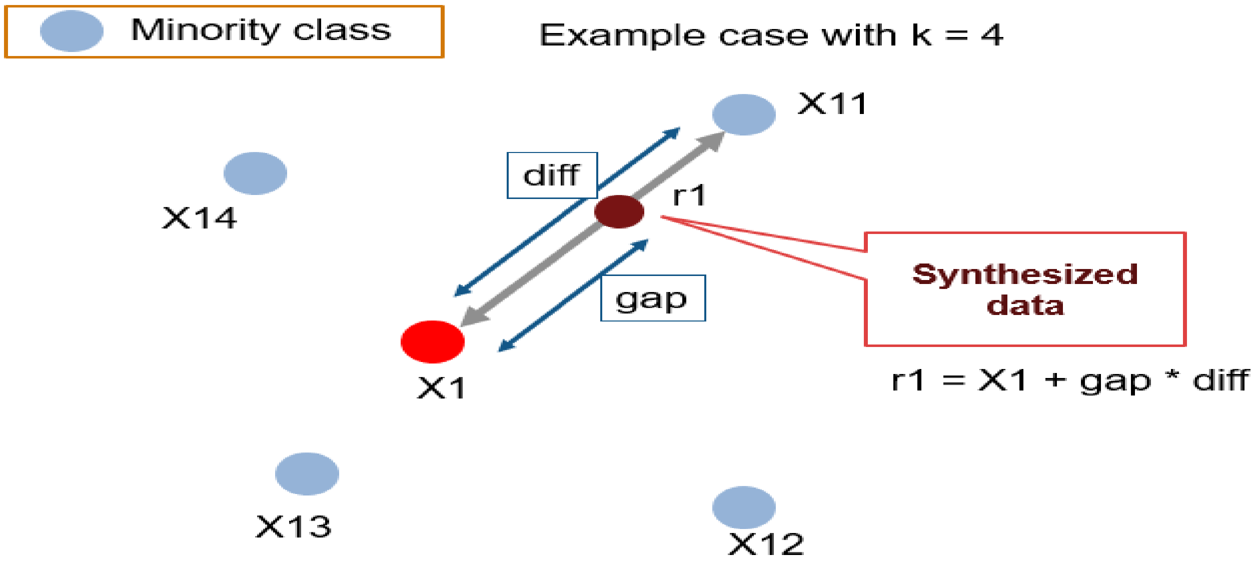
Feature standardization makes the values of each feature in the data have zero-mean (when subtracting the mean in the numerator) and unit-variance. This method is widely used for normalization in many machine learning.

Data Normalization:

The major problem in the various datasets is that numerical features are all measured in different units. Therefore, data normalization is a useful data preparation scheme for tabular data, should be considered so that the comparison between measurements can be more accessible when building a model. Data normalization is a process of re-scaling the feature values to make the new inputs follow the standard normal distribution. Within the different features, there is often a significant difference between the minimum and maximum value. The most common normalization method is the Min-Max normalization. This technique scaled all the numerical values of a numerical feature to a specified range and computed through.

SMOTE (Synthetic Minority Oversampling Technique):

A problem with imbalanced classification is that there are too few examples of the minority class for a model to effectively learn the decision boundary. One way to solve this problem is to oversample the examples in the minority class. This can be achieved by simply duplicating examples from the minority class in the training dataset prior to fitting a model. This can balance the class distribution but does not provide any additional information to the model. An improvement on duplicating examples from the minority class is to synthesize new examples from the minority class. This is a type of data augmentation for tabular data and can be very effective.

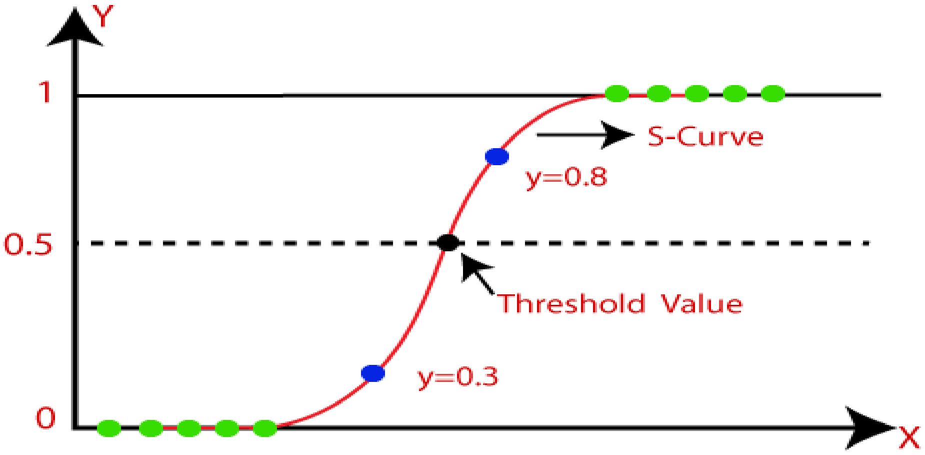


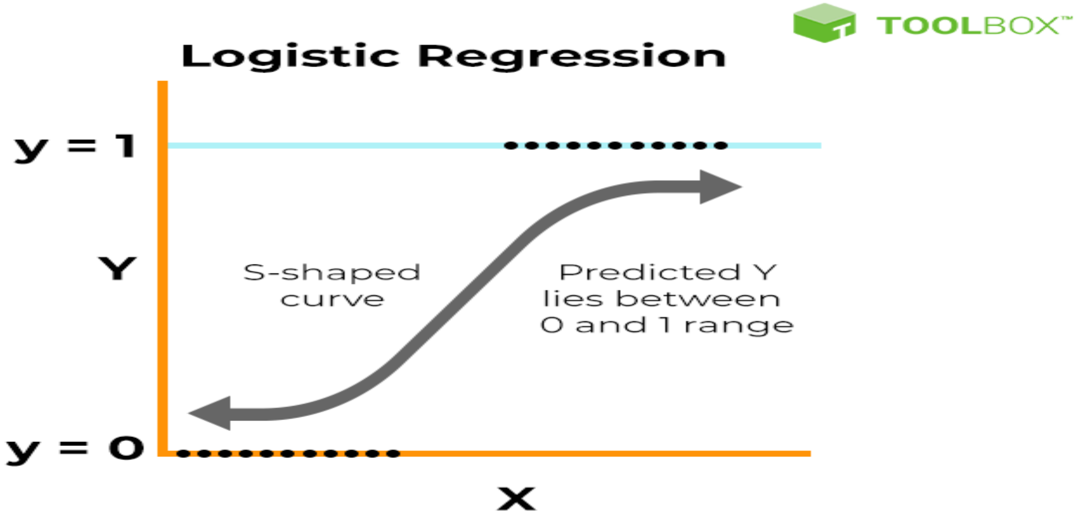
Logistic Regression Implementation:

Logistic regression, despite its name, is a classification algorithm rather than regression algorithm. Based on a given set of independent variables, it is used to estimate discrete value (0 or 1, yes/no, true/false). It is also called log it or Max Ent Classifier.

Below are the steps:

- Data Pre-processing step.
- Fitting Logistic Regression to the Training set.
- Predicting the test result.
- Test accuracy of the result(Creation of Confusion matrix)
- Visualizing the test set result.

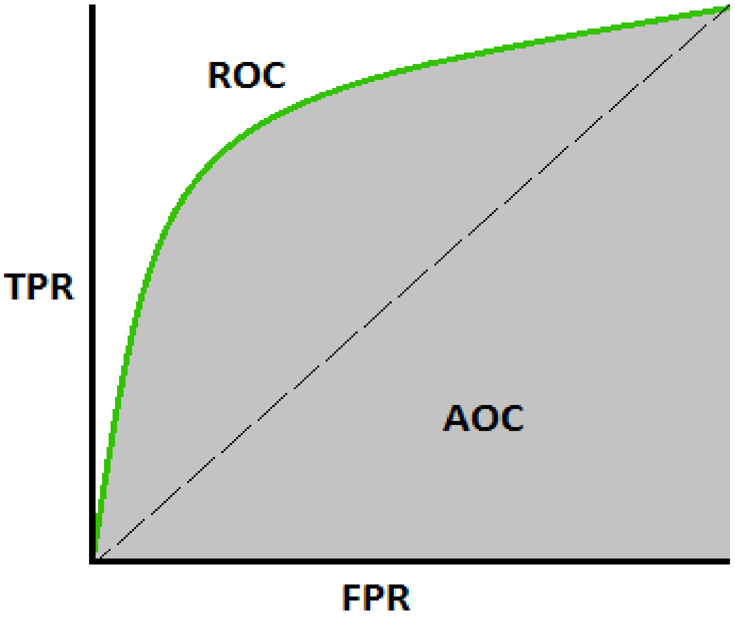




ROC AUC Curve:

- In Machine Learning, performance measurement is an essential task. So when it comes to a classification problem, we can count on an AUC - ROC Curve.
- An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters.

- True Positive Rate
- False Positive Rate



$$\text{TPR /Recall / Sensitivity} = \frac{TP}{TP + FN}$$

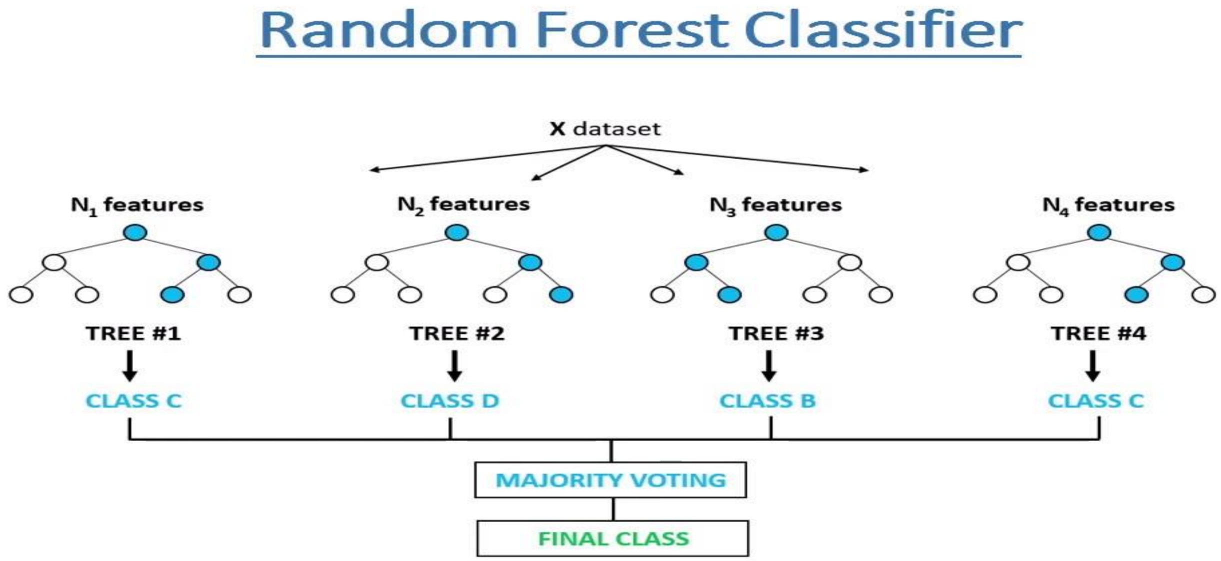
$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{FPR} = 1 - \text{Specificity}$$

$$= \frac{FP}{TN + FP}$$

Random Forest :

Random Forest is a bagging type of Decision Tree Algorithm that creates a number of decision trees from a randomly selected subset of the training set, collects the labels from these subsets and then averages the final prediction depending on the most number of times a label has been predicated out of all.



Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

Hyper parameter Tunning:

Hyper parameters are sets of information that are used to control the way of learning an algorithm. Their definitions impact parameters of the models, seen as a way of learning, change from the new hyper parameters. This set of values affects performance, stability and interpretation of a model. Each algorithm requires a specific hyper parameters grid that can be adjusted according to the business problem. Hyper parameters alter the way a model learns to trigger this training algorithm after parameters to generate outputs.

We used Grid Search CV, Randomized Search CV and Bayesian Optimization for hyper parameter tuning. This also results in cross validation and in our case we divided the dataset into different folds. The best performance improvement among the three was by Bayesian Optimization.

Grid Search CV

Grid Search combines a selection of hyper parameters established by the scientist and runs through all of them to evaluate the model's performance. Its advantage is that it is a simple technique that will go through all the programmed combinations. It helps to loop through predefined hyper parameters and fit your estimator (model) on your training set. So, in the end, you can select the best parameters from the listed hyper parameters.

Randomized Search CV

Random search is a technique where random combinations of the hyper parameters are used to find the best solution for the built model. It is similar to grid search, and yet it has proven to yield better results comparatively.

Bayesian Optimization

Bayesian optimization is a global optimization method for noisy black-box functions. Applied to hyper parameter optimization, Bayesian optimization builds a probabilistic model of the function mapping from hyper parameter values to the objective evaluated on a validation set.

Overall Conclusion

- Data categorical variables had minority classes
- List item which were added to their closest majority class.
- There were not huge gap but female clients tended to default the most.
- Labels of the data were imbalanced and had a significant difference.
- Gradient boost gave the highest accuracy of 82% on test dataset.
- Repayment in the month of September tended to be the most important feature for our machine learning model.
- The best accuracy is obtained for the Random forest and XG Boost classifier.
- From above table we can see that XG Boost Classifier having Recall = 86%, F1-score = 82%, and ROC Score = 83% and Random forest Classifier having Recall =86%, F1-score = 83% and ROC Score = 84%.
- XG Boost Classifier and Random Forest Classifier are giving us the best Recall, F1-score, and ROC Score among other algorithms. We can conclude that these two algorithms are the best to predict whether the credit card is default or not default according to our analysis on this dataset.