Credit Card Default Prediction

Vishal Raul
Data science trainees,
AlmaBetter, Bangalore

Abstract:

Credit risk plays a major role in the banking industry business. Banks' main activities involve granting loan, credit card, investment, mortgage, and others. Credit card has been one of the most booming financial services by banks over the past years. However, with the growing number of credit card users, banks have been facing an escalating credit card default rate. As such data analytics can provide solutions to tackle the current phenomenon and management credit risks. This paper provides a performance evaluation of credit card default prediction. Thus, logistic regression, decision tree, and random forest are used to test the variable in predicting credit default and random forest proved to have the higher accuracy and area under the curve. This result shows that random forest best describes which factors should be considered with an accuracy of 82 % and an Area under Curve of 77 % when assessing the credit risk of credit card customers.

Keywords: Credit Card, Credit Card Defaults, Machine Learning Techniques

1. Introduction

Recently, the state vigorously promotes the economic construction of large- and medium-sized cities, which not only improves people's living standards but also changes people's consumption concept and Consumption mode. People are more and more inclined to spend ahead of time and Mortgage their "credit" to the bank to enjoy certain things in advance. However, when consuming, people often lack rational thinking and overestimate their ability to repay loans to banks in time. On the one hand, it increases the loan risk of banks; on the other hand, it increases the credit crisis of consumers themselves. With a large number of banks selling credit cards, the phenomenon of credit card default emerges one after another. It is very important for banks to effectively identify high-risk credit card default users. Generally speaking, compared with the credit card customers who have not paid their loans overdue, there are fewer overdue repayments. The variable feature of overdue and overdue loan repayment is called "two classifications" in machine learning prediction. In the prediction of "two classifications," a few categories are called positive examples (default), and most categories are called counter examples (non-default). However, most of the credit card loan data are unbalanced. In view of this situation, domestic and overseas scholars have taken up a large amount of research. Proposed an evolutionary sampling method for unbalanced data, which uses genetic algorithms to selectively delete most types of samples and retain samples with a lot of feature information. Compared with other existing data sampling technologies, evolutionary sampling technology has better performance and is more conducive to empirical replication.

2. 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.

3. Data

Data Description:

It consists of **30000 observations** that represent distinct credit card clients. Each observation has **24 attributes** that contain information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan.

Data fields:

The first group of variables contains information about the **client personal information**:

1. ID: ID of each client, categorical variable
2. LIMIT_BAL: Amount of given credit in NT dollars (includes individual and family/supplementary credit)
3. SEX: Gender, categorical variable (1=male, 2=female)
4. EDUCATION: level of education, categorical variable (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
5. MARRIAGE: Marital status, categorical variable (1=married, 2=single, 3=others)
6. AGE: Age in years, numerical variable

The following attributes contains information about the **delay of the past payment** referred to a specific month:

1. PAY_0: 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)
2. PAY_2: Repayment status in August 2005 (same scale as before)
3. PAY_3: Repayment status in July
4. PAY_4: Repayment status in June
5. PAY_5: Repayment status in May
6. PAY_6: Repayment status in April 2005

Other variables instead consider the information related to the **amount of bill statement** (i.e. a monthly report that credit card companies issue to credit card holders in a specific month):

1. BILL_AMT1: Amount of bill statement in September
2. BILL_AMT2: Amount of bill statement in August
3. BILL_AMT3: Amount of bill statement in July
4. BILL_AMT4: Amount of bill statement in June
5. BILL_AMT5: Amount of bill statement in May
6. BILL_AMT6: Amount of bill statement in April

The following variables instead consider the **amount of previous payment** in a specific month:

1. PAY_AMT1: Amount of previous payment in September
2. PAY_AMT2: Amount of previous payment in August
3. PAY_AMT3: Amount of previous payment in July
4. PAY_AMT4: Amount of previous payment in June

5. PAY_AMT5: Amount of previous payment in May
6. PAY_AMT6: Amount of previous payment in April

The last variable is the one to be predicted:

1. default.payment.next.month: indicate whether the credit card holders are defaulters or non-defaulters (1=yes, 0=no)

4. Steps involved

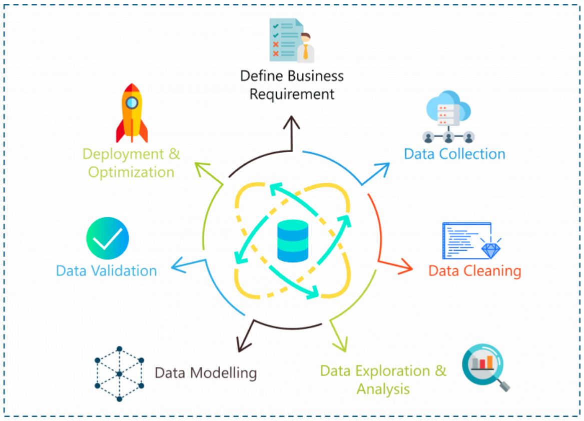

Define Business Requirement
Data Collection
Data Cleaning
Data Exploration & Analysis
Data Modelling
Data Validation
Deployment & Optimization

Image credit: edureka.com

- **Define Problem Statement**

  Before you even begin a Data Science project, you must define the problem you're trying to solve. At this stage, you should be clear with the objectives of your project.

- **Data Collection**

  Our dataset contains a large number of null values which might tend to disturb our accuracy hence we dropped them at the beginning of our project in order to get a better result.

- **Data Cleaning**

  Data cleaning is the process of removing redundant, missing, duplicate and unnecessary data. This stage is considered to be one of the most time-consuming stages in Data Science. However, in order to prevent wrongful predictions, it is important to get rid of any inconsistencies in the data.

- **Data Analysis and Exploration**

  Once you're done cleaning the data, it is time to get the inner Sherlock Holmes out. At this stage in a Data Science life-cycle, you must detect patterns and trends in the data. This is where you retrieve useful insights and study the behavior of the data. At the end of this stage, you must start to form hypotheses about your data and the problem you are tackling.

- **Data Modelling**

  This stage is all about building a model that best solves your problem. A model can be a Machine Learning Algorithm that is trained and tested using the data. This stage always begins with a process called Data Splicing, where you split your entire data set into two proportions. One for training the model (training data set) and the other for testing the efficiency of the model (testing data set).

- **Optimization and Deployment**

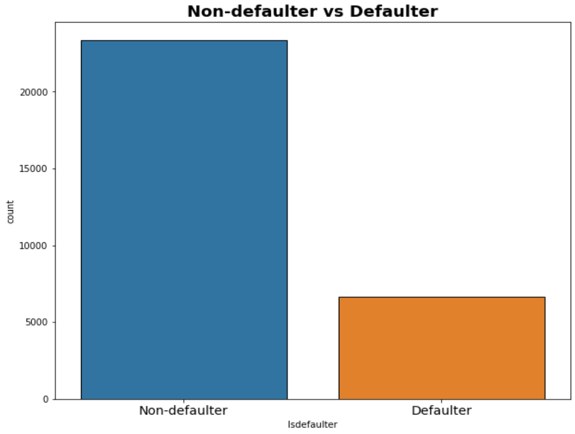
  This is the last stage of the Data Science life-cycle. At this stage, you must try to improve the efficiency of the data model, so that it can make more accurate predictions. The end goal is to deploy the model into production or production-like environments for final user acceptance. The users must validate the performance of the models and if there are any issues with the model then they must be fixed in this stage.

5. Exploratory Data Analysis:

5.1. Isdefaulter

The main aim of the data is to discriminate clients that are predicted to credit card default the next month, according to the default.payment.next.month column which is set to "0" for non-defaulters and "1" for defaulters. Thus, it is a ***binary classification problem*** on a relatively *unbalanced dataset*, as shown in the following figure.

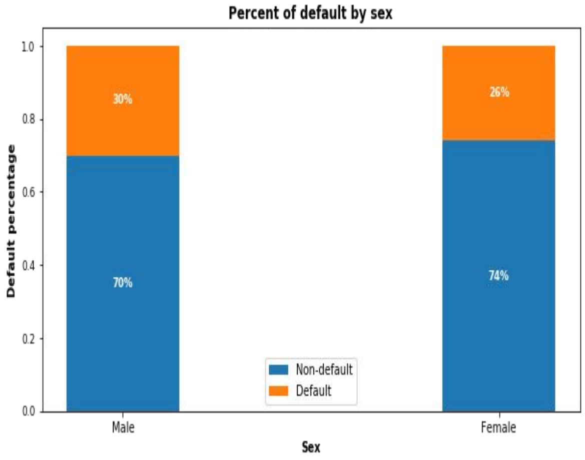
.**Graph 1. Distribution of defaulter**



5.2. Gender Variable

Whether it is male or female, the proportion of default consumers is still relatively low, which is in line with the general situation. Conventionally, most of the default data such as credit card fraud are uneven, and we need to make some adjustments to the model based on the actual situation. For the feature sex, we draw a stacked histogram according to the target variable, as shown in Graph 2.

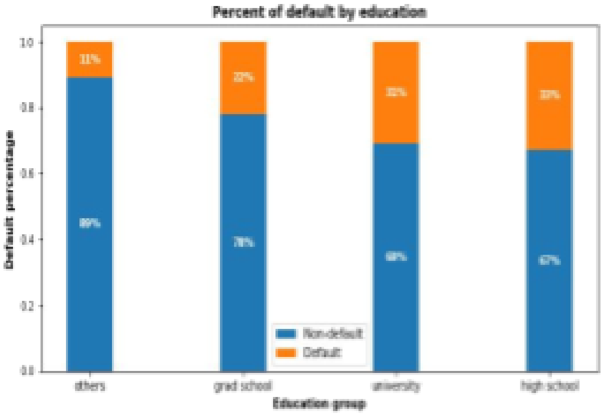
Graph 2. Distribution of Gender variable



5.3. Education Variable

The sample set is unbalanced in the corresponding attribute values of the three characteristics of gender, education, and marriage. For the feature series payment status, we draw different stacked histograms according to different months, and the results are shown in Graph 2. Maximum number of defaulter are educated people (University and Graduate school).

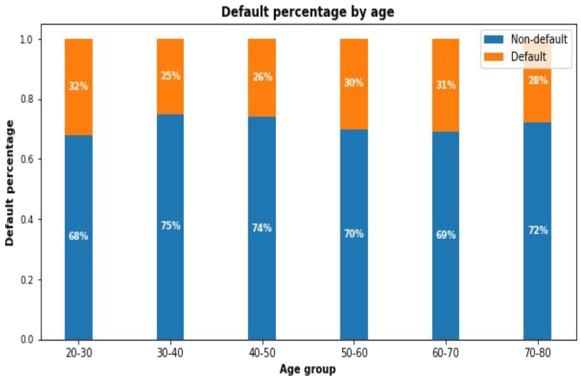
Graph 3. Geography of pickup points



5.4. Age Variable

For the feature age, we also performed a visual analysis, as shown in graph 3. Graph shows that the probability of non-default age between approximately 25 and 40 is higher, which indicates that consumers in this age group are more capable of repaying credit card loans. 0is may be because their work and family tend to be stable without too much pressure.

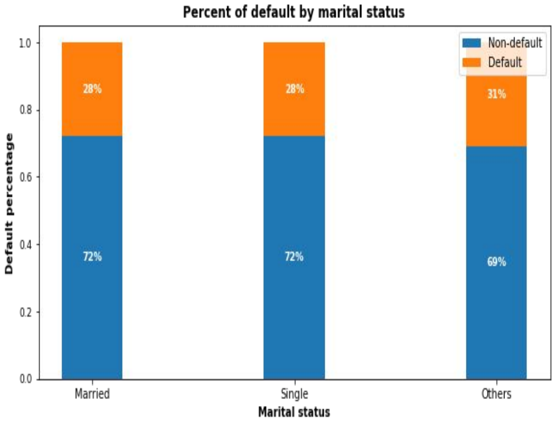
Graph 5. Age Distribution



5.5. Marital Status Variable

For the feature marriage, we draw the same graph as the feature sex and education. The default and non-default conditions of this feature are shown in Figure.

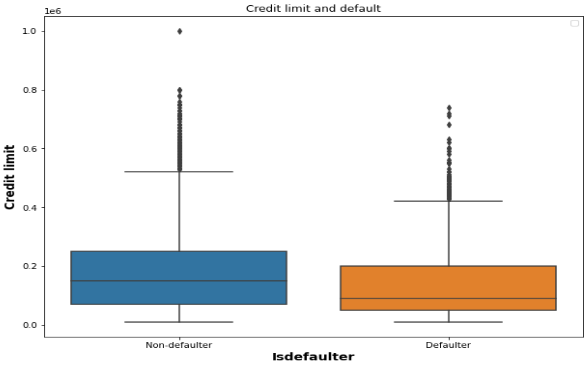
Graph 6. Marital Status



5.6. Credit limit

Higher credit limit is associated with lower default risk. Payment amount has also increased the difficulty for banks to adjust the credit card loan limit.

Graph 6. Credit limit



6. 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, and 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

$$Xnorms = \frac{X - Xmin}{Xmax - Xmin}$$

All the features are scaled except categorical features.

7. RESAMPLING METHODS

Any dataset can be considered as imbalanced if the number of instances between classes is not equal. Resampling methods for imbalanced learning applications typically means to add a bias to balance the dataset.

Credit Card Default Prediction in the Imbalanced Datasets of balancing the dataset to achieve more robust results. All the credit-related datasets employed in this study leads to the data imbalance problem. Besides, all of the resampling techniques allow resampling until the desired ratio of balance dataset, allowing us to directly compare different resampling methods for a given proportion of minority and majority class data points in the final training set. Resampling techniques have been implemented on the full datasets. Data level resampling approaches have most commonly been used to deal with class imbalance, so various undersampling and oversampling based approaches have been used in this study.

1) Random Undersampling

Random undersampling is a simple undersampling based approach. Majority class instances in the training set are randomly eliminated until the ratio between the minority, and the majority class is at the desired level. Theoretically, one of the problems with random undersampling is that one cannot control what information about the majority class is thrown away. In particular, crucial details on the decision boundary between the minority and majority class may be eliminated. Despite its simplicity, random undersampling has empirically been shown to be one of the most effective resampling methods. In particular, few of the more sophisticated undersampling methods have outperformed random undersampling in empirical studies. In random undersampling, examples have been randomly removed from the majority class to balance the class instances, which results in the removal of vital information from the majority class. This approach also results in a downsizing of the training data considerably. Therefore it is the most naive approach in data undersampling

2) Random Oversampling

Like random undersampling, random oversampling is a simple yet effective approach to resampling. Random Oversampling is a very naive approach to data oversampling. It merely replicates the minority class examples and adds them to the training data. By using this technique, new examples come from the existing minority class examples in the training set that results in the problem of overfitting. Overfitting is a problem that occurs when all the training examples are very similar to each other, and the classifier correctly classifies these examples. In such a scenario, if a test example is slightly different from the training examples, then the classifier is not able to classify it correctly and results in poor classification for the new examples. In other words, the classifier is trained to classify only a very narrow set of examples correctly. The random oversampling method operates by replicating the randomly selected set of examples from the minority class so that the majority class does not have an overbearing presence during the training process. Since the resampling process is random, it becomes difficult for the decision function to find a clear borderline between the two classes. Therefore, although it is widely used, Random oversampling might be ineffective at improving recognition of the minority class by a large margin. Some potential drawbacks of random oversampling include an increase in training time for the classifier and overfitting on account of duplication of examples of the minority class. However, other oversampling methods have been built based on this method.

8. Evaluation Metrics

Model evaluation is of paramount importance in any predictive modeling task. It becomes even more critical in ensemble predictive modeling, where the relative performance and diversity of models must be thoroughly evaluated. All the evaluation metrics are built on four types of classifications:

A. Accuracy

Typically, accuracy is used to assess the effectiveness of a model with the help of the confusion matrix. The accuracy of the model has been computed through

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

B. Precision

Precision compares the number of true positives to the number of true positives and the number of false positives. That is, of all the instances the classifier said were positive, precision measures how many of them were positive. The Precision of the model has been computed through

$$Precision = \frac{TP}{TP + FP}$$

C. Recall

Recall compares the number of true positives to the number of true positives and false negatives. The Recall of the model has been computed through

$$Recall = \frac{TP}{TP + FN}$$

D. ROC

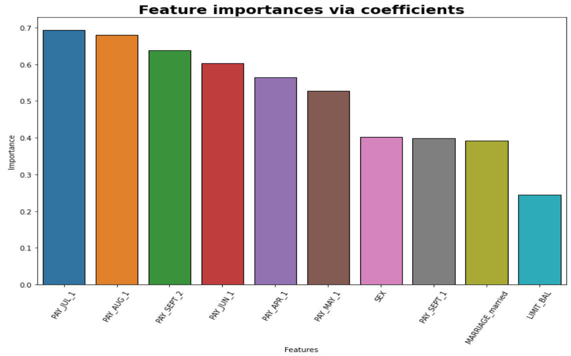
A receiver operating characteristic (ROC) curve plot is also a widely used measure to evaluate the performance of classifiers. Specifically, the plot is created by plotting the true positive rate (recall) against the false positive rate at various threshold levels.

9. ML Model Building

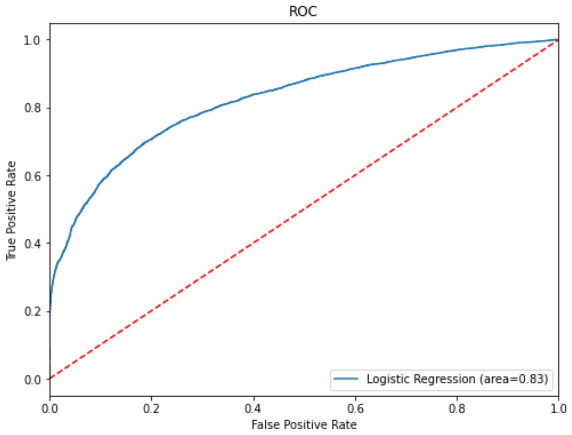
9.1. Logistic Regression

Logistic Regression is one of the simplest algorithms which estimates the relationship between one dependent binary variable and independent variables, computing the probability of occurrence of an event. The regulation parameter C controls the trade-off between increasing complexity (overfitting) and keeping the model simple (underfitting). For large values of C, the power of regulation is reduced and the model increases its complexity, thus overfitting the data.

* Features Importance

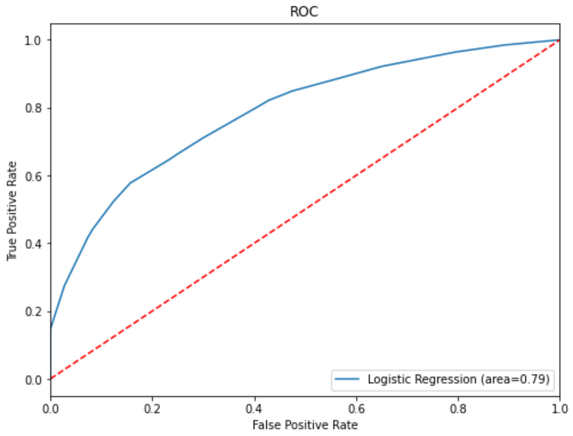


By implementing logistic regression we get f1-sore approx. 73%. As we have an imbalanced dataset, F1- score is a better parameter. Let's go ahead with other models and see if they can yield better results.



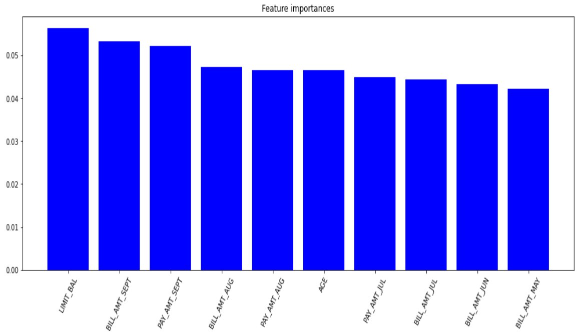
9.2. Decision Tree Classifier

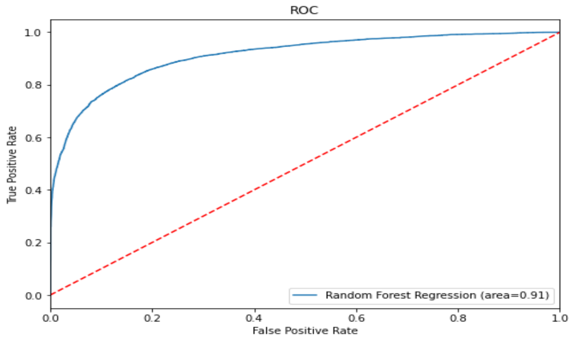
Decision Tree is another very popular algorithm for classification problems because it is easy to interpret and understand. An internal node represents a feature, the branch represents a decision rule, and each leaf node represents the outcome. Some advantages of decision trees are that they require less data preprocessing, i.e., no need to normalize features. However, noisy data can be easily overfitted and results in biased results when the data set is imbalanced.



9.3. Random Forest Classifier

Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

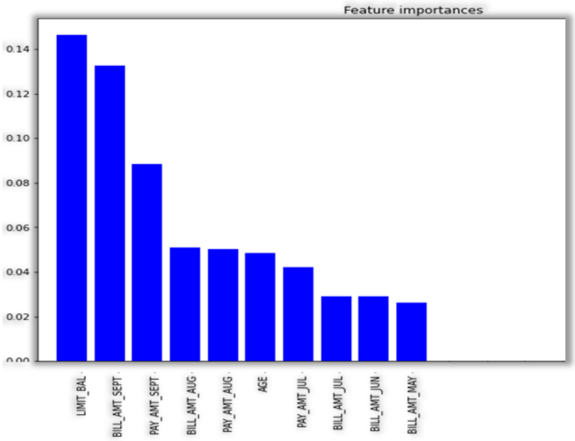


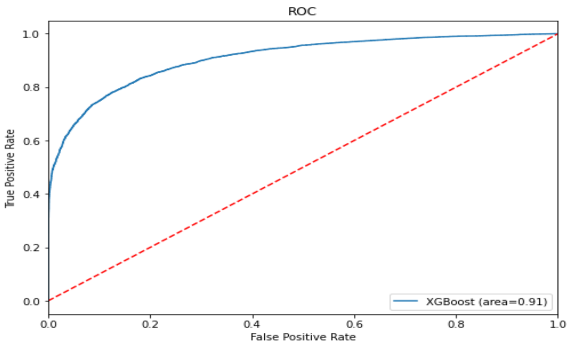


9.4. XGBoost Classifier

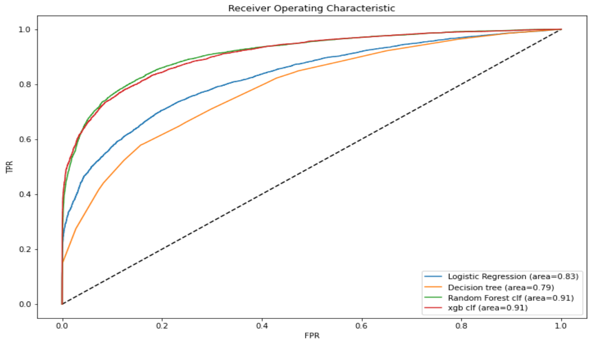
XGBoost is an implementation of Gradient Boosted decision trees. XGBoost models majorly dominate in many Kaggle Competitions.

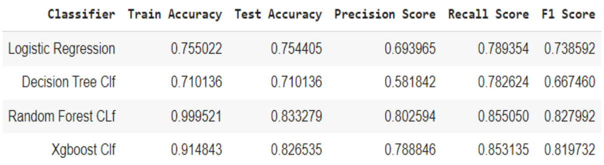
In this algorithm, decision trees are created in sequential form. Weights play an important role in XGBoost. Weights are assigned to all the independent variables which are then fed into the decision tree which predicts results. The weight of variables predicted wrong by the tree is increased and these variables are then fed to the second decision tree. These individual classifiers/predictors then ensemble to give a strong and more precise model. It can work on regression, classification, ranking, and user-defined prediction problems.





Model Evaluation Result





Classifier	Train Accuracy	Test Accuracy	Precision Score	Recall Score	F1 Score
Logistic Regression	0.755022	0.754405	0.693965	0.789354	0.738592
Decision Tree Clf	0.710136	0.710136	0.581842	0.782624	0.667460
Random Forest CLf	0.999521	0.833279	0.802594	0.855050	0.827992
Xgboost Clf	0.914843	0.826535	0.788846	0.853135	0.819732

10. Conclusion:

Machine learning methods, in conjunction with the use of imbalanced methods, have been utilized in various domains. The objective of this paper is to train various supervised learning algorithms to predict the client’s behavior in paying off the credit card balance. In classification problems, an imbalanced dataset is also crucial to enhance the performance of the model, so different resampling techniques were also used to balance the dataset. We first investigated the datasets by using exploratory data analysis techniques, including data normalization.
However, all the models implemented achieved comparable results in terms of accuracy.

- 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 serve as an aid to human decision.
- Suggest output probabilities rather than predictions.
- Model can be improved with more data and computational resources.

11. References:

1. https://www.researchgate.net/publication/344914401
2. https://www.kaggle.com/
3. https://www.analyticsvidhya.com/
4. https://learn.almabetter.com/