Credit Card Default Prediction

Sourav Chowdhury(Data science trainee)
AlmaBetter, Bangalore

Abstract:

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 faced 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. The Data set used is of customers of Taiwan many of them have Defaulted. In classification problems, an imbalanced dataset is also crucial to improve the performance of the model because most of the cases lie in one class, and only a few examples are in other categories. Traditional statistical approaches are not suitable to deal with imbalanced data. Data level resampling techniques are employed to overcome the problem of the data imbalance.

1. 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.

Variable breakdown:

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.

Thought Process:

- Will do basic cleaning of Data(removing null values, Univariate and Multivariate Analysis)
- What makes a client Default on his/her credit card payment?
    1. History of past payment records and Transactions, Amount of bill payments, amount of previous debits(we all check through last 3 months bank statement)
    2. So how much important these variables are(eg. if last month's bill payments are good but the amount is very less or vice versa.)
    3. Do every variable is important or are these variables interrelated to each other?
    4. Once deciding the variables we will train our model on various Machine learning Algorithms.

2. INTRODUCTION :

Credit risk has traditionally been the greatest risk among all the risks that the banking and credit card companies are facing and usually one requiring more capital. This can be proven by industry business reports and statistical data. Despite machine Learning and big data have been adopted by the banking industry, the current applications are mainly focused on credit score prediction. Because of the risks inherent in such a large portion of the economy, building models for consumer spending behaviors to limit risk exposures in this sector is becoming more critical. For this to be a viable option, the predictions need to be reasonably accurate.

A robust model is not only a useful tool for the lending institutions to decide on credit applications, but it can also help the clients to be aware of the behaviors that may damage their credit scores. The primary motivation behind risk prediction is to utilize financial data.

3. Exploratory Data Analysis :

Data Preparation: At this stage one needs to analyze data step by step from starting by loading data into DataFrames. The relevance of the data to the data mining goals, the quality of the data, and technical constraints such as limits on data volume or data types. For data Preparation, number of outliers were removed

- Null Values Treatment

  Dataset do not contain any number of null values thus get an advantage of not removing any data from the dataset.

- Univariate Analysis / Multivariate Analysis

  Every column was Analyzed, starting from the box plot along with the distribution graph for different variables.

Variables Involved :

LIMIT_BAL

LIMIT_BAL states the amount of given credit. This is the maximum amount a customer can spend with their card in a single month. The amount of balance limit is dependent on the bank's own Screening processes and other unknown factors.

GENDER

GENDER states the gender of the person whose data is to be analyzed, it was given that value 1 was opted for male and value 2 for female, Analyzed it using Pie chart and Bar graphs.

EDUCATION

The Education level of a customer is represented as one of four values : 1 – Graduate School, 2 – University, 3 – High School, 4 – Other. For the purpose of Analyzing customer groups, this is assumed to indicate the highest level of education completed.

MARRIAGE

The dataset contains the data for married single and others, it is being represented with 1- married, 2 - Single, 3 - Others.

AGE

The dataset contains the age of customers which is given in years. This is a numerical data , their were outliers in age, Thus after making an analysis , all age groups were having Defaulters.

4. Handling Class Imbalance

SMOTE Oversampling : Since the data was highly imbalance between Defaulters and non defaulters

The class of Defaulters contained 6636 rows while non defaulters were 23364 . Thus Class needed to be balanced , we used SMOTE(Synthetic Minority Oversampling Technique) to over sample the minority class i.e Defaulters and ensure the sampling technique is not biased.

After application of SMOTE the dataset has equal no. of 0's and 1's i.e Non Defaulters and Defaulters.

```
Original unbalanced dataset shape 30000 rows

Resampled balanced dataset shape 46728 rows
```

Thus the size of the dataset increased.

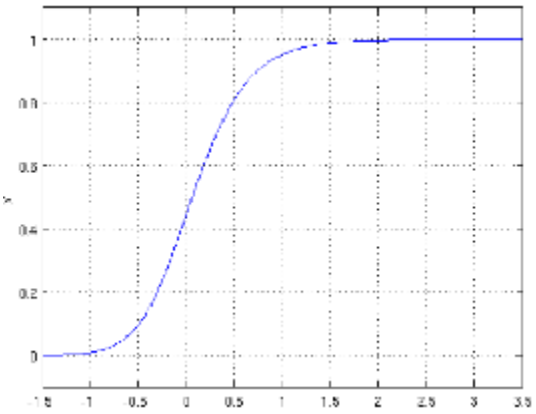
5. Fitting Different Models

For Modeling we tried on Various Classification Algorithms like:

1. **Logistic Regression**
2. **Decision Tree**
3. **Random Forest**
4. **Gradient Boost**
5. **XG Boost**
6. **Support Vector Machines**

5.1 Algorithms :

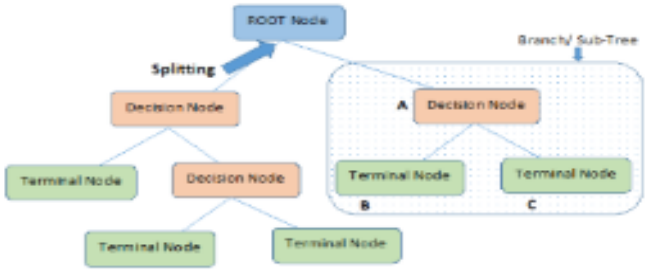
1. **Logistic Regression:** LR is actually a classification algorithm, The function used in LR is known as Sigmoid Function given by : f(x) = 1/1+e^(-x)



The cost function used calculated through **maximum likelihood estimation**:

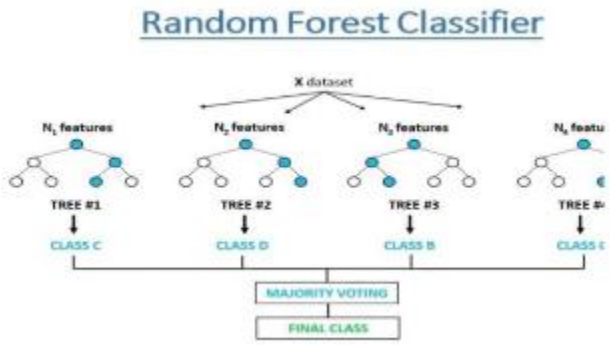
$$\ell(w) = \sum_{i=1}^{N} y_i log(h_w(x_i)) + (1 - y_i) log(1 - h_w(x_i))$$

2. **Decision Tree Classifier :** Decision Tree is a Tree based algorithm that splits the data based on some initial value with a node called **Root Node** (It represents entire population or sample and this further gets divided into two sets.)

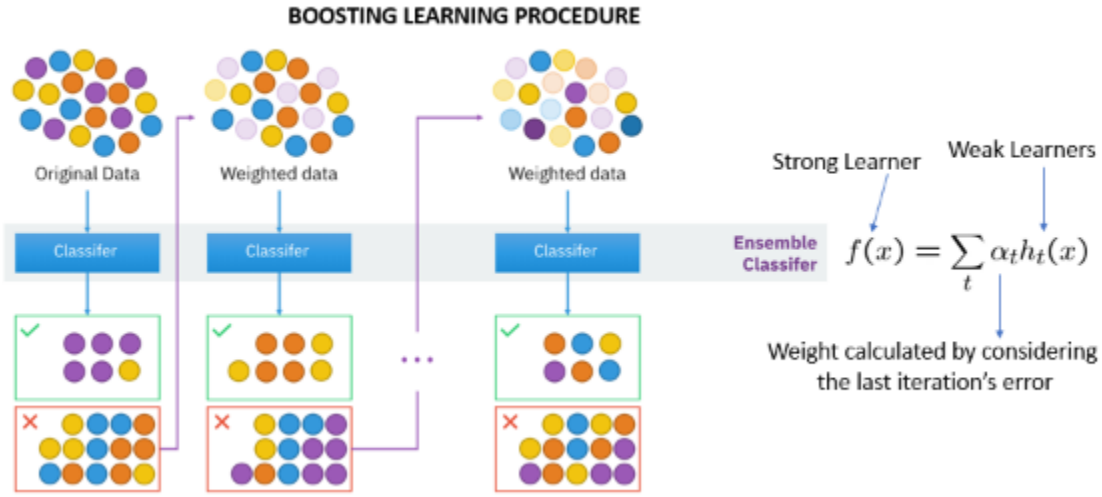


and then it further splits the data having methods like Information Gain, Gini , Chi Square, Reduction in Variance. It then collects the data based on the label of the final terminal node and generally takes the probabilities greater than 0.5.

3. **Random Forest Classifier:** Random Forest is a Special case of Bagging in Decision Tree, That creates trees based on random selection subset of training data, collects the labels from these subsets and then averages the final prediction depending on most no. of times a label has been predicted out of all.



4. **Gradient Boost Classifier:** It is the special case of Decision Trees In this case, there are 2 kinds of parameters P: the weight of each leaf, W : the no. of leaves in each tree



5. **XG Boosting Classifier :** It is one of the fastest implementations of gradient boosting trees. It does this by tackling one of the major inefficiencies of gradient boosted trees, concerning the potential loss for all possible splits to create a new branch (especially if we consider a case where

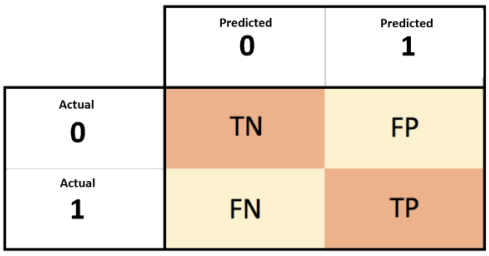
there are thousands of features). XG boost tackles this inefficiency by looking at the distribution of features across all data points in a leaf and using this information to reduce the search of possible feature splits.

6. **Support Vector Machines** : An SVM classifier builds a model that assigns new data points to one of the given categories. Thus, it can be viewed as a non-probabilistic binary linear classifier.Thus it is a kind of **Discriminative Classification.** SVMs can be used for linear classification purposes. In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using the kernel trick. It enables us to implicitly map the inputs into high dimensional feature spaces.

5.2 Model Performance :

Models evaluation is done based on below mentioned metrics:

1. **Confusion Matrix :** The confusion matrix is a table that summarizes how successful the classification models at predicting.



2. **Precision/Recall :** Precision is the ratio of correct positive predictions to the overall number of positive predictions : TP/TP+FP
   Recall is the ratio of correct positive predictions to the overall number of positive examples in the set: TP/FN+TP
3. **Accuracy :** Accuracy is given by the number of correctly classified examples divided by the total number of classified examples. In terms of the confusion matrix, it is given by:
   TP+TN/TP+TN+FP+FN
4. **Area Under ROC Curve :** ROC curve uses a combination of the true positive rate(the proportion of positive examples predicted correctly, defined exactly as recall) and false positive rate (the proportion of negative examples predicted incorrectly) to build up a summary picture of the classification performance.

5.3 Cross Validation and Hyper Parameter Tuning :

Hyperparameters 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 hyperparameters. This set of values affects performance, stability and interpretation of a model. Each algorithm requires a specific hyperparameter grid that can be adjusted according to the business problem. Hyperparameters alter the way a model learns to trigger this training algorithm after parameters to generate outputs.

We used Grid Search CV, Randomized Search CV Optimization for hyperparameter tuning. This also results in cross validation and in our case we have divided into different folds. The best performance among the two is shown by Grid Search CV.

1. **Grid Search CV :** Grid Search combines a selection of hyperparameters established by the user and runs through all of the combinations to evaluate the model's performance using accuracy metrics in our case. Its advantage is that it is a simple technique that will go through all the programmed combinations. The Biggest disadvantage is that it transverses a specific region of the parameter space and cannot understand which movement or which region of the space is important to optimize the model.
2. **Randomized Search CV :** In Random Search, the hyperparameters are chosen at random within a range of values that it can assume. The advantage of this method is that there is a greater chance of finding regions of the cost minimization space with more suitable hyperparameters, since the choice for each iteration is random. The disadvantage of this method is that the combination of hyperparameters is beyond the user's control.

6. Model Comparison

- **Baseline model of Random forest and decision tree shows huge difference in train and test accuracy which shows overfitting.**

- **After cross validation and hyperparameter tuning, XGBoost shows the highest test accuracy score of 88.10% and ROC Score is 0.884.**

Conclusion:

Based on the EDA, we discovered the human characteristics are not that much important in predicting the defaulter. From Modeling we are able to classify Default risk with accessible customer data and find a decent model.

References:

Geeks for Geeks

Kaggle

TowardsDataScience