Credit Card Default Prediction

Shubham Srivastava
Data science trainee,
AlmaBetter, Bangalore

Abstract:

Nowadays, the use of credit card becomes an integral part of modern economies. Still, predicting credit card defaulters is considered as the most important. So, its assessment becomes a crucial task. In this context, a few Data mining and intelligent artificial techniques were used for extracting meaningful patterns from a given dataset.

In this study, classification models based on decision trees, random forest, XGBoost , support vector machines (SVM) are developed and applied on credit card fraud detection problem.

Keywords: machine learning, credit card, defaulter prediction ,classifiers

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. We can use the K-S chart to evaluate which customers will default on their credit card payments.

The main objective is to build a predictive model, which could help them in predicting the defaulters proactively.

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.

2. Introduction

As an unsecured credit facility, credit cards have huge risks behind the high returns of banks. The ever-increasing number of credit card circulation cards has brought about an increase in the amount of credit card defaults, and the resulting large amount of bills and repayment information data have also brought certain difficulties to the risk controllers. Therefore, how to use the data generated by users, and extract useful information to control risks, reduce default rate, and control the growth of non-performing rate has become one of the key concerns of banks.

Credit card default prediction is based on the historical data of credit card customers. The use of corresponding methods to predict and analyze credit card customer default behavior is a typical classification problem. Data mining algorithms have long been applied to the study of credit card default prediction problems.

3. Steps involved:

- **Exploratory Data Analysis**
  We do exploratory analysis during which we apply label encoding on categorical columns such as sex, education, marriage. Then, apply SMOTE(Synthetic Minority Oversampling Technique) to remediate Imbalance in dependent column.
  After which we apply feature engineering on the data to acquire some meaningful features for better prediction accuracy. At last, we apply One Hot encoding on features such as sex(Male, Female).

- **Feature Engineering**

  We apply feature engineering on the data to acquire some meaningful features for better prediction accuracy such as payment value which is sum of all payment or dues which is difference between total bill amount and payment.

- **One Hot Encoding**

  We apply One Hot encoding on features such as sex(Male, Female), marriage(married, single).

- **Fitting different models**

  For modelling we tried various classification algorithms like:

1. **Logistic Regression**
2. **Decision Tree Classifier**
3. **Random Forest Classifier**
4. **XGBoost Classifier**
5. **SVM**

- **Tuning the hyperparameters for better accuracy**

  Tuning the hyperparameters of respective algorithms is necessary for getting better accuracy and to avoid overfitting in case of tree based models like Random Forest, decision tree and XGBoost.

4.1. Model performance:

Model can be evaluated by various metrics such as:

1. **Confusion Matrix-**

   The confusion matrix is a table that summarizes how successful the classification model is at predicting examples belonging to various classes. One axis of the confusion matrix is the label that the model predicted, and the other axis is the actual label.

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 (AUC)**-
   ROC curves use 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.

8. Conclusion:

That's it! We reached the end of our exercise.
Starting with loading the data so far we have done EDA , null values treatment, encoding of categorical columns, feature engineering, one hot encoding and then model building.
In all of these models our accuracy revolves in the range of 75 to 83%.
And there is improvement in accuracy score after hyperparameter tuning.
So the accuracy of our best model(random forest classifier) is 83% which can be said to be good for this large dataset.

References-

1. MachineLearningMastery
2. GeeksforGeeks
3. Analytics Vidhya