SPOT A CREDIT CARD DEFAULTER

Authors

Aditya Patnaik
School of Information Studies,
Syracuse University, Syracuse
USA

Rohit Jain
School of Information Studies,
Syracuse University, Syracuse
USA

Siddharth Kumar
School of Information Studies,
Syracuse University, Syracuse
USA

Abstract

Credit card default is an all-time high in any county. While it is possible that such defaults are deliberate often, they can be completely unintentional. Such defaults can lead to increased interest rate, loss of credit card limit, and reduction in interest free period. We apply various Machine Learning algorithms on the dataset and tried to understand factors which contributes to default payment. We expect to gain a deeper understanding regarding factors which influences weather a person pays credit card due in time. From this project banks can get an insight about why certain customers tend to default on credit card payment. Such analysis will be beneficial not only for identifying defaulters but also for targeting customers for other marketing campaigns of the bank.

Introduction

The health of the credit card industry is best measured not by the number of people with cards, but rather the number who pay their bills. Bad payment habits begin by nicking you with more fees and lower credit scores, and, in advanced cases, can lead to the loss of a vehicle or home, garnishment and bankruptcy. According to TransUnion's Industry Insights Report, it was found that the credit card delinquency rate reached 1.79 percent in Q4 2016, an increase of 12.6 percent from 1.59 percent in Q4 2015. The credit card delinquency rate remains more than a full point below its peak in Q4 2009 (2.97 percent) (2)(3).

So how does a credit card default happen? Credit card default happens when you've become

severely delinquent on your credit card payment. For example, you agree to make your minimum payment by the due date listed on your credit card statement. If you miss the minimum credit card payment, your credit card will be in default. It's a serious credit card status that not only affects your standing with that credit card issuer, but also your credit standing in general and your ability to get approved for credit cards, loans, and other credit-based services(1).

In this study, we will try to predict if a person would default in his next credit card payment, using multiple features such as age, gender, education, payment habit etc. We will use UCI credit card data set pulled from Kaggle.com. The data set contain 30,000 observations with 25 features. We will try to find out most prominent features using different data modeling algorithms. Also, we will have to perform feature engineering on some of the columns in the data.

Material & Methods

Data

The data was downloaded from Kaggle.com. The main source of the data is UC Irvine Machine Learning Repository. There are 2 institutions involved in collection of the data set, that is, (1) Department of Information Management, Chung Hua University, Taiwan and (2) Department of Civil Engineering, Tamkang University, Taiwan. Our study took payment data in October 2005, from an important bank (a cash and credit card issuer) in Taiwan and the targets were credit card holders of the bank. Among the total 25,000 observations, 5529 observations (22.12%) are the cardholders with default payment. This research employed a binary variable – default payment (Yes = 1, No = 0), as the response variable and used the following 23 variables as explanatory variables. Here are the features after cleaning the data set:

- LIMIT_BAL: Amount of the given credit.
- SEX: Gender (0 = male; 1 = female).
- EDUCATION: Education (4 = graduate school; 3 = university; 2 = high school; 1 = others).
- MARRIAGE: Marital status (1 = married; 0=unmarried).
- AGE: Age (year).
- PAY_0–PAY_6: History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows: PAY_0 = the repayment status in September, 2005; PAY_1 = the repayment status in August, 2005. The measurement scale for the repayment status is: -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months and so on;
- BILL_AMT1–BILL_AMT6: Amount of bill statement. BILL_AMT1 = amount of bill statement in September, 2005; BILL_AMT2 = amount of bill statement in August, 2005 and so on.
- PAY_AMT1– PAY_AMT6: Amount of previous payment (NT dollar). PAY_AMT1 = amount paid in September, 2005; PAY_AMT2 = amount paid in August, 2005 and so on.
- Default.payment.next.month: Default payment next month (1=yes, 0=no);

Method

In our study, we decided to compare different model and their accuracy by checking area under curve score. The models that we tested were, Logistic regression, Logistic Regression with Elastic Net Regression, Decision Tree Model, Random Forest and Gradient Boosting. Since we are working with all the numerical fields, they are required to be regularized. To do this, the data was randomly divided into three groups, one for model training, other to validate the model and third one to test the model on the data set. Splitting of the data is done by 6:3:1 method.

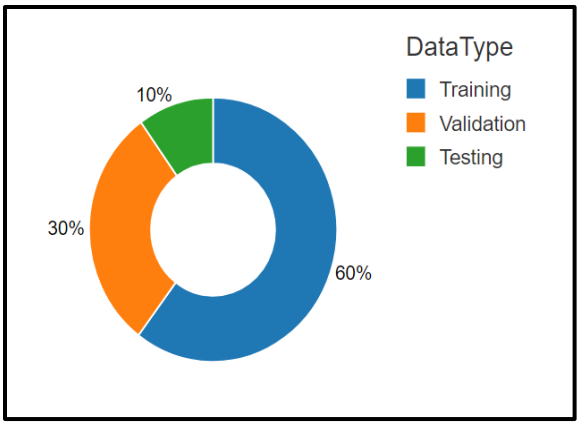

Fig 1

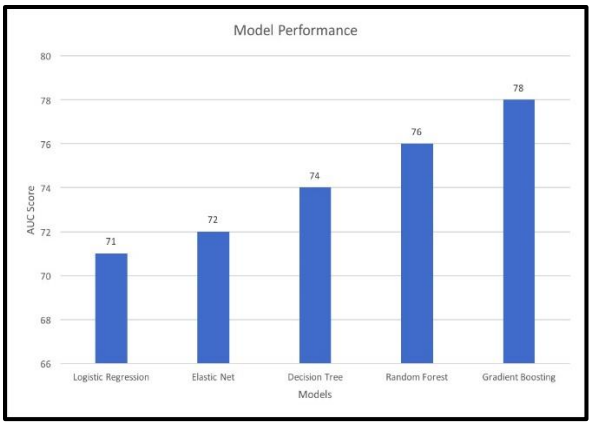

Fig 2

Once the split is done, we built a pipeline which would scale all the columns except, education, sex and marriage. This is done to improve the performance of the model and to bring homogeneity in magnitude of the data. The scaled results were passed through a vector assembler. The output/features were passed through different machine learning algorithm.

After fitting different models on training data, we obtained the AUC (Area Under Curve) score for each model and multiplied the result by 100 for ease of interpretation. Next, we used this pipeline on testing data and visualized the results.

Results

We ran our pipelines on the validation set and the following results were obtained:

Model	AUC Score
Logistic Regression	0.71
LR with Elastic Net Regularization	0.72
Decision Tree	0.74
Random Forest	0.76
Gradient Boost	0.78

From here we can clearly conclude that Gradient Boosting model is giving the best performance for validation data set.

We analyzed our outputs from LR with Elastic Net Regularization as the model provides with automatic feature selection for the given data set. In other words, features which are least important are not considered while building the model. The coefficients here represent the weight of the feature on the overall result of the model. We can clearly see that the feature 'Pay_0', that is, repayment for the month of September 2005, has the highest positive impact on the probability of the customer being a defaulter. Whereas, 'Bill_Amt1' has the maximum negative impact on the customer not being a defaulter. From here we can conclude that 'Pay_0', 'Pay 2', 'Pay_3', 'Pay_4', 'Pay_5', 'Pay_6', 'Age' are positively related to the probability of repayment of the credit card bill. Whereas, 'Sex', 'Marriage', 'Limit_Bal', 'Bill_Amt1' and 'Pay_Amt1' are negatively related to the repayment.

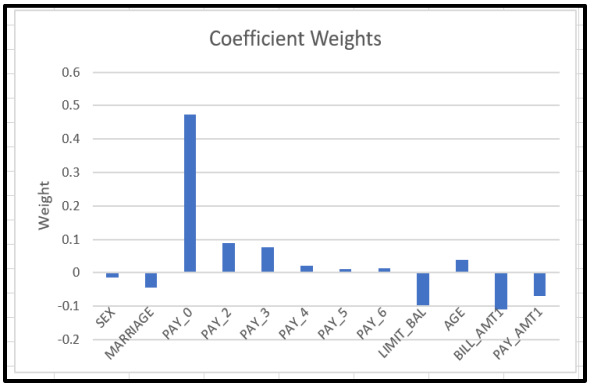

Fig 3

In a Receiver Operating Characteristic (ROC) curve the true positive rate (Sensitivity) is plotted in function of the false positive rate (100-Specificity) for different cut-off points. Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold.

In the figure below, we have plotted the ROC curve for random forest on testing data. The AUC is 0.77 which means that the model is giving good performance as compared to other models on testing data.

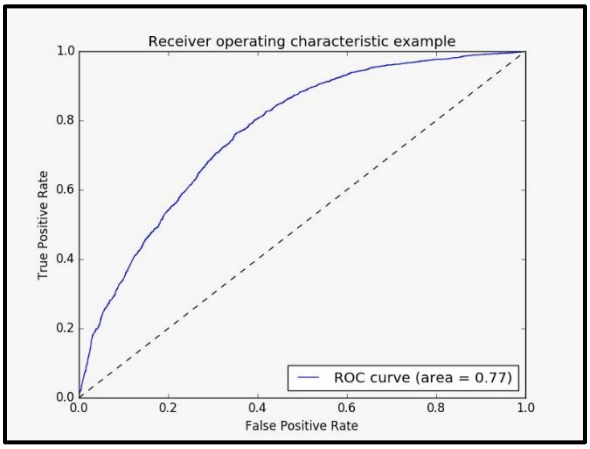

Fig 4

Conclusion

This paper examines the five major classification techniques in data mining and compares the performance of classification and predictive accuracy among them. The Random Forest Method, is presented to estimate the real probability of default with highest accuracy.

In the classification accuracy among the five data mining techniques, the results show that there are little differences in error rates among the five methods. However, there are relatively big differences in area ratio among the five techniques. Obviously, area ratio is more sensitive and is an appropriate criterion to measure the classification accuracy of models. In our case the area under the cure for Random forest is 0.77 which gives the accuracy of approximately 80%.

References

1. Retrieved May 1, 2018, from https://www.thebalance.com/what-is-credit-card-default-960209
2. Retrieved May 2, 2018, from https://www.creditcards.com/credit-card-news/credit-card-delinquency-statistics-1276.php
3. Retrieved May 2, 2018, from https://www.transunioninsights.com/IIR-2016Q4/