University of Central Florida

2018

Predicting Default Payments of Credit Card Clients

Chengle Huang

clhuang@Knights.ucf,edu

Predicting Default Payments of Credit Card Clients

Chengle Huang

University of Central Florida

4000 Central Florida Blvd, Orlando, FL 32816

August 2, 2018

Abstract

As an important business of the bank, credit card also brings high income and high risk to the bank. As the credit card business flourishes, banks also need to set up databases to rate customers. Based on the information provided by the credit customer, the bank can effectively evaluate the customer's credit rating to determine whether the customer will bring a default risk. According to the credit card client data of 2005 in Taiwan, this paper built Random Forest model, Lasso-Logistic model and Support Vector Machine model to explore the key factors which effect customer credit, including icredit data, sex, education, marriage, age, history of payment, and bill statements. Through comparing the AUC value of the models, we found the model of better prediction effect to forecast the bank credit card defaults. The fitted credit card default prediction model not only helps banks to choose safe clients, but also can provide certain theoretical support for the credit decisions. In addition, it has a strong theoretical and practical significance.

Keywords: Credit Cards, Default Payment, Sampling Methods, Feature Selection.

1. Introduction

A credit card is a payment card issued by a bank to a client, enabling the cardholder to pay the merchant for the merchandise according to the their commitment to the card issuer. For banks, risk management and default detection has been a crucial challenge in issuing credit cards. The traditional credit evaluation method is a manual credit risk assessment, which is performed by credit analysts through the review of the information submitted by credit card applicants, generally including customer personal information, income, assets, stable repayment ability, etc. . Nowadays, the number of credit card applications is increasing, and the bank's credit card circulation is gradually increasing. This makes us realize that traditional manual estimation is becoming more and more incompetent. Therefore, with the advent of data mining technology, we can try to find an accurate model for the credit risk assessment of credit cards.

A credit card issuer wants to better predict the likelihood of a customer defaulting, which will affect the issuer's decision whether to provide the customer with a credit card and what credit line to offer. In order to reduce the financial loss of clients' default, the credit card issuer has the gathered information on 30000 clients, including their credit data, sex, education, marriage, age, and history of payment.

In this paper, we will build Lasso-Logistic Regression, Random Forest, and Support Vector Machine models to predict the likelihood of a customer defaulting. The research aimed at the case of customers' default payments and compares the predictive accuracy of probability of default among three data mining methods.

Due to the dataset is an imbalanced dataset, we used sampling methods to generate balanced training dataset. Three methods are considered: Oversampling, Both Sampling, and SMOTE. While there are three balanced dataset, we choose to training Random Forest models using each dataset and evaluate each models' performance.

Next, we choose to select relevant variables to fit the models. We choose to use Lasso working for Logistic Regression method, and Best-first search working for SVM model. In addition, For SVM model, we have tuned the kernels and two parameters to optimize the SVM model.

Finally, we could evaluate the performance for all models. By comparing the accuracy, sensitivity, specificity, and AUC value, we could find the best performance model.

2. Data preparation

The original dataset comes from UCI Machine Learning Repository (Data source: https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients.) The research aimed at the case of customers' default payments in Taiwan and compares the predictive accuracy of probability of default among data mining methods. The credit card issuer has the gathered information on 30000 clients. The dataset contains information on 24 variables, including credit data, sex, education, marriage, age, history of payment, and bill statements of credit card customers from April 2005 to September 2005, as well as information on the outcome: did the customer default payment next month or not? I set default as the response variable Y, and other variables as predictors X's. Table 1 shows the description for those variables.

Name	Description
LIMIT_BAL	Amount of given credit in NT dollars (includes individual and family/supplementary credit)
SEX	Gender (1=male, 2=female)
EDUCATION	(1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
MARRIAGE	Marital status (1=married, 2=single, 3=others)
AGE	Age of the client
PAY_0	Repayment status in September, 2005 (-2=no consumption, -1=pay duly, 0=the use of revolving credit, 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)
PAY_2	Repayment status in August, 2005 (scale same as above)
PAY_3	Repayment status in July, 2005 (scale same as above)
PAY_4	Repayment status in June, 2005 (scale same as above)
PAY_5	Repayment status in May, 2005 (scale same as above)
PAY_6	Repayment status in April, 2005 (scale same as above)
BILL_AMT1	Amount of bill statement in September, 2005 (NT dollar)
BILL_AMT2	Amount of bill statement in August, 2005 (NT dollar)
BILL_AMT3	Amount of bill statement in July, 2005 (NT dollar)
BILL_AMT4	Amount of bill statement in June, 2005 (NT dollar)
BILL_AMT5	Amount of bill statement in May, 2005 (NT dollar)
BILL_AMT6	Amount of bill statement in April, 2005 (NT dollar)
PAY_AMT1	Amount of previous payment in September, 2005 (NT dollar)

PAY_AMT2	Amount of previous payment in August, 2005 (NT dollar)
PAY_AMT3	Amount of previous payment in July, 2005 (NT dollar)
PAY_AMT4	Amount of previous payment in June, 2005 (NT dollar)
PAY_AMT5	Amount of previous payment in May, 2005 (NT dollar)
PAY_AMT6	Amount of previous payment in April, 2005 (NT dollar)
default	Default payment next month(1=yes, 0=no)

Table 1

The dataset contains some categorical variables like SEX, EDUCATION, MARRIAGE that have numeric values, so we have converted these numeric variables into dummy variables to avoid losing important information.

After converting, there were 5607 clients' information is not completed, then we removed them and finally got 24411 complete client data.

Next, in order to measure the performance of the models. It is very important that we need to prepare test data. The idea is that in practice we want our models to be used for predicting the class of data we have not seen yet: although the performance of a classification method may be high in the data used to estimate the model parameters, it may be significantly poorer on data not used for parameter estimation, such as the future data. In this paper, we split the original dataset into a training dataset and a test dataset. We randomly choose 50% of the original observations as training samples, and other observation are treated as test dataset.

3. Handling Imbalanced Dataset

In the training dataset, there are 12205 clients' data, and the dependent variable "default" represents if the customer would default payment next month (1) or not (0)? After we analyzed the training dataset, we found that 22.6% clients would default and 77.4% would not. This is not a well balanced dataset.

With imbalanced dataset, the algorithms can not get the necessary information about the minority class to make an accurate prediction. Hence, it is desirable to use ML algorithms with balanced data sets. In this paper, we have used three sampling methods: Oversampling, Both Sampling, and SMOTE.

Oversampling works with minority class. It replicates the observations from minority class to balance the data. An advantage of using this method is that it leads to no

information loss. The disadvantage of using this method is that, since oversampling simply adds replicated observations in original data set, it ends up adding multiple observations of several types, thus leading to overfitting.

Both Sampling contains both Oversampling and Undersampling, Undersampling works with majority class. It reduces the number of observations from majority class to make the data set balanced.

SMOTE algorithm creates artificial data based on feature space similarities from minority samples. This method mitigates the problem of overfitting caused by random oversampling as synthetic examples are generated rather than replication of instances; however, while generating synthetic examples SMOTE does not take into consideration neighboring examples from other classes. This can result in increase in overlapping of classes and can introduce additional noise [1].

Table 2 shows the information of dependent variable in original training datasets and the datasets after sampling.

	default	not default
Original	2754	9451
Oversampling	9451	9451
Both Sampling	4935	5065
SMOTE	6149	6059

Table 2

Now, we already got three balanced datasets using three techniques, then we choose to training Random Forest models using each dataset and evaluate each models' performance. Finally, we could choose the best dataset to training other models.

4. Random Forest

Random Forest is a bagging technique which divide the data into several portions, use a relatively weak classification tree to process, and then combine them. Due to the datasets contains 23 predictors, we choose to use 5 variables when building a random forest of regression trees.

We evaluate the model performance by using AUC value. Figure 1 shows the ROC curves of each fitted models with each balanced dataset.

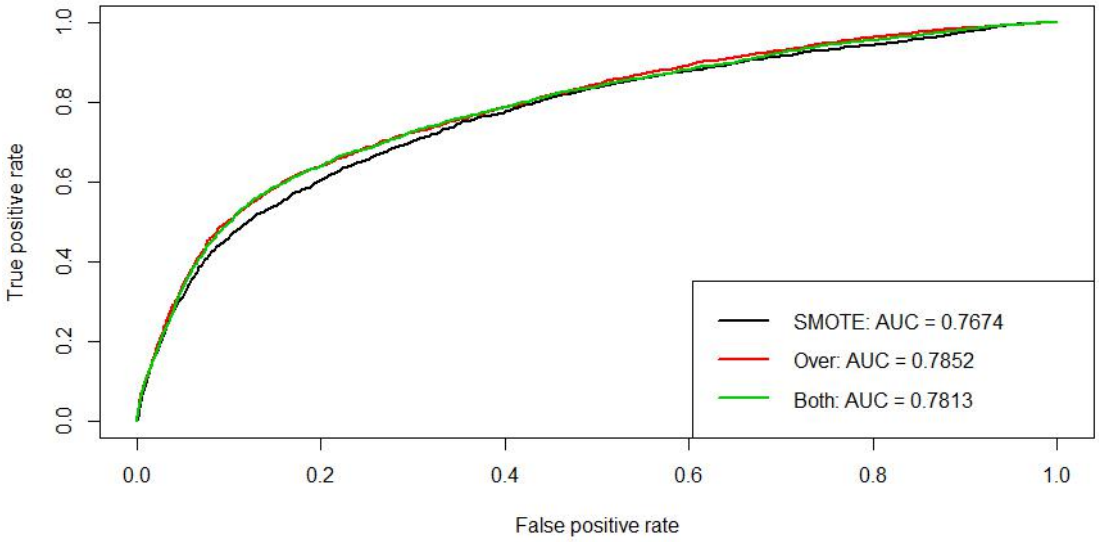

Figure 1

Due to the model using Oversampling dataset reach the greatest AUC value, we choose this dataset as our new training dataset.

Finally, we used the test dataset to estimate the performance of the selected Random Forest model. Table 3 shows the result.

		True Class	
		0	1
Predicted Class	0	8328	1497
	1	898	1483

(a)

Accuracy	Sensitivity	Specificity
0.8038	0.4977	0.9027

(b)

Table 3

5. Feature Selection

Feature selection is the process of selecting a subset of relevant features for use in model training. The feature selection process removes redundant and irrelevant features to improve model performance, which reduces the probability of overfitting. It can also simplify the models to make them easier to interpret and reduce the training times.

In this project, the clients' information contains 23 predictors, generally, some of them may be redundant or irrelevant, so feature selection is a necessary task for this project. In this project, we have tried two methods to do this: Lasso and Best-first search algorithm.

Lasso was introduced in order to improve the prediction accuracy and interpretability of regression models by altering the model fitting process to select only a subset of the provided covariates for use in the final model rather than using all of them. It also reveals that the coefficient estimates need not be unique if covariates are collinear. According to Figure 2, we choose to use lambda of 0.00734 to do feature selection for logistic regression model.

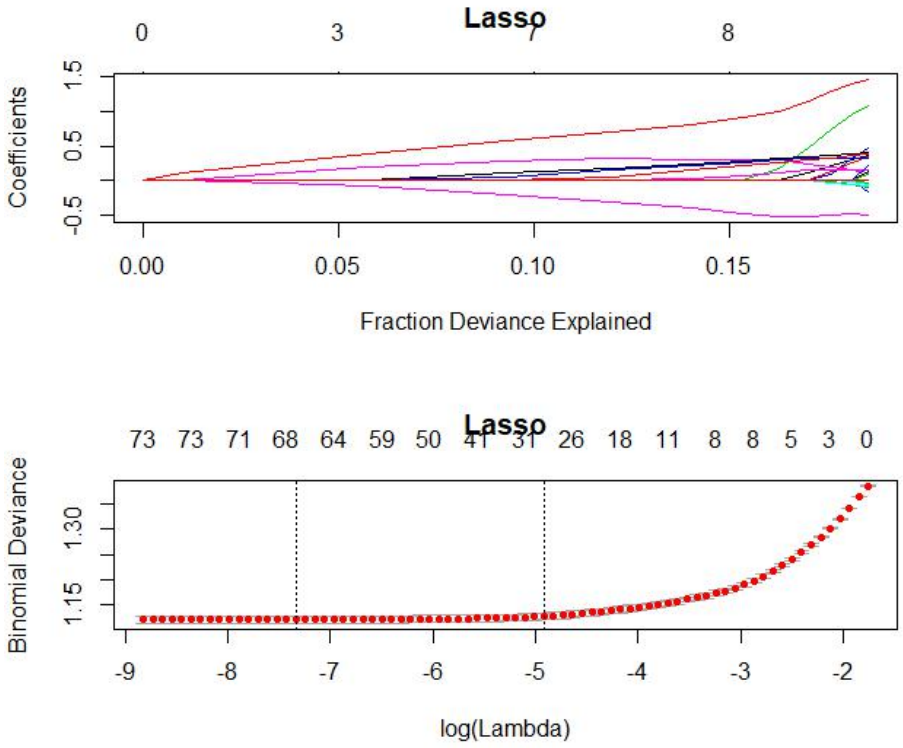

Figure 2

Best-first search is an algorithm that traverses a graph in search of one or more goal nodes. It uses a greedy algorithm, expand the first successor of the parent. After a successor is generated, if the successor's heuristic is better than its parent, the successor is set at the front of the queue, and the loop restarts. Else, the successor is inserted into the queue. The procedure will evaluate the remaining successors of the parent [2].

6. Logistic Regression

Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable. It is used to describe data and to explain the relationship between one dependent binary variable and one or more independent variables.

By using Lasso, we got the feature subset: LIMIT_BAL, SEX, EDUCATION, MARRIAGE, AGE, PAY_0, PAY_2, PAY_3, PAY_4, PAY_5, PAY_6, BILL_AMT1, BILL_AMT2, BILL_AMT3, BILL_AMT4.

Finally, we used the test dataset to estimate the performance of the Lasso-Logistic regression model. Table 4 shows the result.

		True Class	
		0	1
Predicted Class	0	7818	1213
	1	1408	1767

Accuracy	Sensitivity	Specificity
0.7853	0.593	0.8474

Table 4

7. Support Vector Machine

Support Vector Machine draws a boundary to separate different classes of data and tries to maximize the margin between each class and the boundary. In this project, it is trying to maximize the distance between margins for a decision boundary separating the default and non-default instances in hyper-dimensional space.

By using Best-first search algorithm, we got the feature subset: LIMIT_BAL, SEX, PAY_0, PAY_4.

Then, we have tuned three parameter for SVM models: Kernel, Cost, and Gama.

Kernel: Linear, Radial

Cost: 0.1, 1, 10, 100, 1000

Gama: 0.5, 1, 2, 3

The parameter tuning result shows best parameters are Linear kernel with a cost of 1000.

Finally, we used the test dataset to estimate the performance of the Linear kernel SVM model. Table 5 shows the result.

		True Class	
		0	1
Predicted Class	0	7921	1286
	1	1305	1694

Accuracy	Sensitivity	Specificity
0.7877	0.5685	0.8586

Table 5

8. Result

With the balanced training dataset, we have trained three classification models. In here, we combine the results of the models' performances.

	Accuracy	Sensitivity	Specificity
RF	0.8038	0.4977	0.9027
Logistic	0.7853	0.593	0.8474
SVM	0.7877	0.5685	0.8586

As we can see, Random Forest model has the highest Accuracy and Specificity, but its sensitivity is the lowest one. Logistic Regression looks similar with SVM model. The accuracy for the models are 0.8038, 0.7853, and 0.7877, and if a model blindly predicts all clients would not default, the accuracy can reach 0.7759. So, depend on the accuracy, these models have a few improvement than blindly guess. However, we

still can not decide which model is the best one. Therefore, I choose to use AUC value as the model performance measure.

Finally, I choose to use ROC curve to show the performance of models. In Figure 3, the true positive rate represents sensitivity and false positive rate means 1 - specificity. Therefore, we can adjust the threshold value of classification probability to change the Sensitivity and Specificity. In here, we use AUC value as evaluation measure, Higher AUC value means greater performance for models, as we can see, the AUC for all models are much better than random guess which is only 0.5, and random forest model has the best performance.

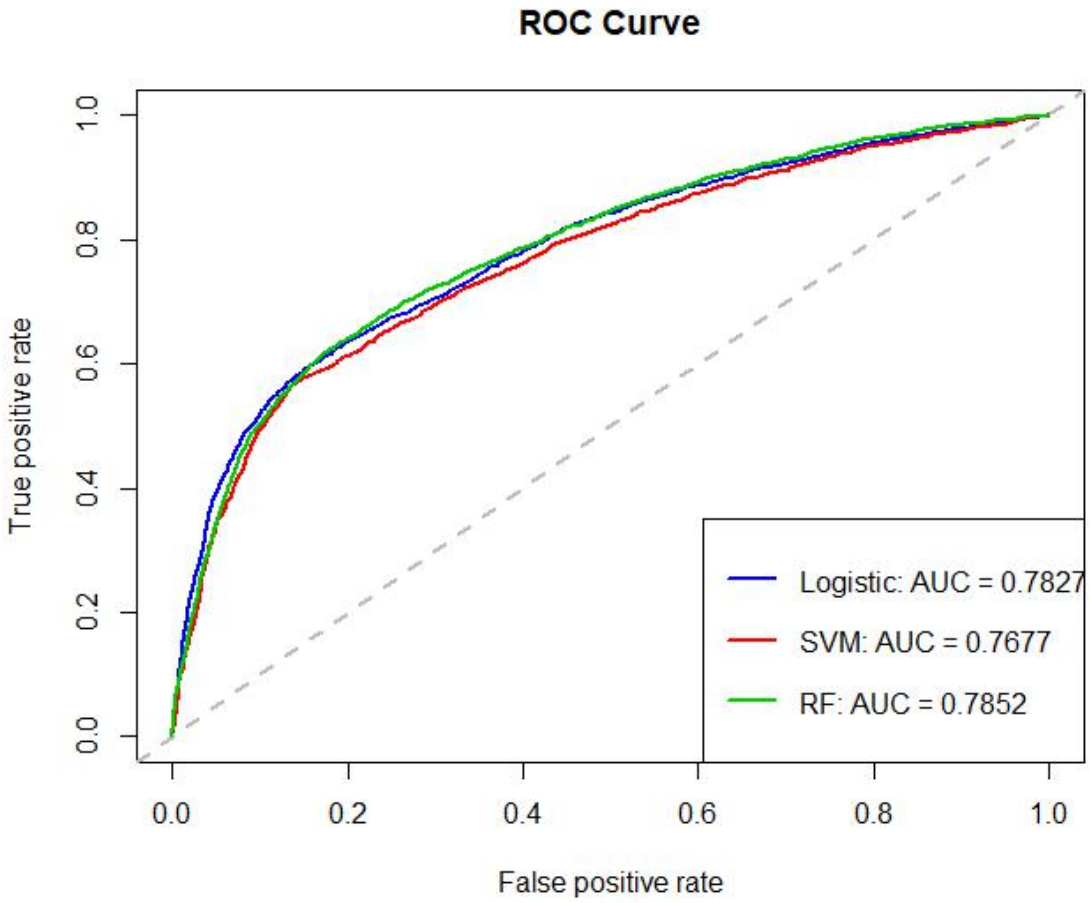

Figure 3

9. Discussion

This paper mainly establishes three kinds of customer default prediction models for the basic data of credit card customers, and compares the model prediction effects. It is found that the Random Forest model is better than the Lasso-Logistic model and SVM model. The Random Forest model has an accuracy of 0.8 that means it can effectively predict most of clients' credit rank. Its specificity is 0.9 that means this model would predict most safe clients would not default. Moreover, its sensitivity is 0.5 that means half of potential default clients would be predicted. Overall, this model is powerful for default payments prediction. For imbalanced dataset, sampling process is significant before training the model. Feature selection could help us removes redundant and irrelevant features to improve model performance, which reduces the probability of overfitting. The fitted credit card default prediction model not only helps banks to choose safe clients, but also can provide certain theoretical support for the credit decisions.

Reference

1. "How to handle Imbalanced Classification Problems in machine learning?" MARCH 17, 2017.
https://www.analyticsvidhya.com/blog/2017/03/imbalanced-classification-problem/

2. Carnegie Mellon. "Greedy Best-First Search when EHC Fails."
http://www.cs.cmu.edu/afs/cs/project/jair/pub/volume28/coles07a-html/node11.html#modifiedbestfs

3. Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications, 36(2), 2473-2480.