The Comparisons of Machine Learning Approaches -- An example of probability of default of credit card clients

Qinghui Yang, Xinyuan Yang, and Duoduo Yu

***Abstract*—This project focuses on customer's default payments situation in Taiwan and compare the predictive accuracy among five models. Risk prediction is an effective way to prevent crisis. Therefore, this research of predicting credit card default is meaningful. We use data from UCI machine learning data base and analyze default situation of 30,000 customers in Taiwan and give prediction of whether customers default to banks or not. In this study, we fit logistic regression model, Lasso model, Ridge model, Kernel SVM model and Random Forest model to the data. According to confusion matric, ROC, AUC graph and accuracy rate, random forest works best.**

***Index Terms*—Machine learning, Probability, Data mining**

I. Introduction

Nowadays, cashless payment is becoming more popular, especially in western countries. People usually pay by credit card without cash, and banks are encouraging this payment method through cash back on credit card consumption. However, there exists default payments. Some customers consume a lot but cannot pay back to banks on time, which could induce credit risk. Our group is interested in this topic and would like to conduct risk prediction through five models. Finally, we compare the five models through recall, precision, confusion matrix, ROC, AUC and accuracy rate. The results show that random forest model performs best.

Below is a brief introduction of the five models.

A. Basic logistic regression

Logistic regression is usually used for classification problems. In LR model, the explanatory variables do not need to be normally distributed and it does not require a linear relationship between explanatory variables and response variables. The binary response (0 or 1) stands for different class, which is easy to interpret.

B. Logistic regression with grid search (Lasso)

Lasso is an advanced version of LR, it adds a penalty term (absolute value of magnitude) to loss function. Shrinking and removing less important feature's coefficients can reduce variance without increasing bias. Lasso can be seen as a form of automatic feature selection and it could raise the accuracy and reduce over-fitting.

C. Stochastic gradient descent (Ridge)

Ridge regression is similar to Lasso regression. The difference is that the form of penalty term is a squared magnitude. If lambda is large in loss function, it will lead to under-fitting problem. This model works well on avoiding over-fitting.

D. Kernel SVM

SVM is used to solve classification problems. Kernel is better for making optimization when there are many features. Kernel SVM is a non-linear classification model to classify binary results by projecting data into a higher dimension without increasing too much complexity of loss function.

E. Random forest

Random forest can be used as an ensemble learning method for classification by establishing a multitude of decision trees at training time and outputting the class. It is efficient on large data bases.

II. Task description

We use default payment data from Taiwan to predict credit card default situation. First we preprocess data and then fit 5 models to the data. For every model, we calculate confusion matrix, recall and precision rate, ROC and AUC, and accuracy rate to measure the predictive performance.

III. Major challenges and solutions

At first, we are not sure about using which models to fit the data. Logistic regression is good but the accuracy rate is not high after our implementation. Then we try to fit Lasso and Ridge regression. Lasso's accuracy rate even lower than LR. Changing to kernel SVM and found it is also lower than LR. Eventually we tried random forest, it performs best.

In addition, we also consider about the reasons for the different performance of each model.

Qinghui Yang is with the Department of Economics, UCLA, CA 90025 USA (e-mail: qhyang83@ucla.edu).

Xinyuan Yang is with the Department of Economics, UCLA, CA 90025 USA (e-mail: xinyuanyoung@outlook.com).

Duoduo Yu is with the Department of Economics, UCLA, CA 90025 USA (e-mail: duoduoyu18@ucla.edu).

IV. Experiments

A. Dataset description

This study utilized 30,000 payment data from a significant bank in Taiwan in October, 2005, and the subjects investigated were credit card holders of the bank. We applied binary classification to the dependent variable, default payment:

1) Y: Default payment: when the credit card holder had a default payment in October, 2005, the dependent variable equals to 1; when the credit card holder did not have a default payment in October, 2005, the dependent variable equals to 0.

Among the total 30,000 data points, 6636 subjects investigated (22.12%) had default payment problem. We reviewed the previous studies (Yeh, I. C., & Lien, C. H., 2009; Lee, Yen, Lin, Tseng, & Ma, 2004; Steenackers & Goovaerts, 1989; Updegrave, 1987) and employed the following 23 variables as independent variables:

1) X1: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.
2) X2: Gender (1 = male; 2 = female).
3) X3: Education (1 = graduate school; 2 = university; 3= high school; 4 = others).
4) X4: Marital status (1 = married; 2 = single; 3 = others).
5) X5: Age (year).
6) 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.
7) 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.
8) 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. Tables

We randomly split the data into training dataset and testing dataset respectively. In this research, we train on 80% (24,000 data points) and test the rest 20%.

B. Data exploration

We first download the data from UCI website. To make it more clear, we did several steps to clean the data, including renaming the columns 'default payment next month' as 'def_pay' and 'PAY_0' as 'PAY_1'; dropping the ID column. The data shape is (30000,24) after checking the missing data.

Then we can describe some details of our data set. For instance, there are 30,000 credit card clients in total. The average value for the amount of credit card limit is 167,484. The standard deviation is unusually large, max value being 1 million. Education levels are mainly distributed on graduate school and university. Most of the clients are either married or single (less frequent the other status). Average age is 35.5 years, with a standard deviation of 9.2. Value 0 for default payment means 'not default' and value 1 means 'default', the mean of 0.221 indicates that 22.1% of credit card contracts may default next month.

As we can see from the figure1, the density of the predict value are skewed, showing that most amount of credit limit belongs to 0-200,000. Moreover, after combing age situation with marriage dummy, we can see there is a difference between men and women for all marriage levels.

Figure 3 is the distribution of delays in the past 6 months. Figure 4 is the delay dummy distribution. We can also check correlations in the figure 5 as below.

FIGURE 1

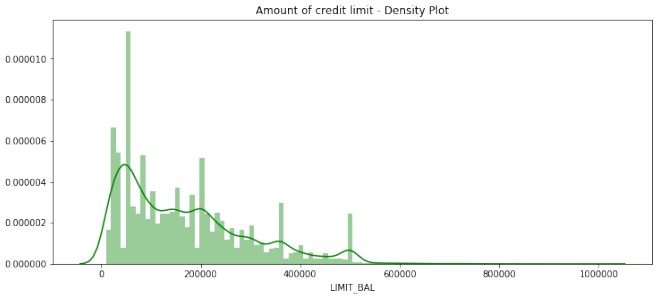


FIGURE 2

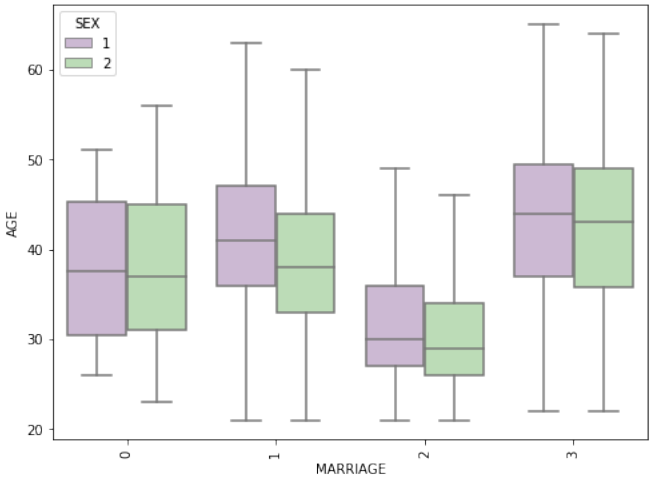


FIGURE 3

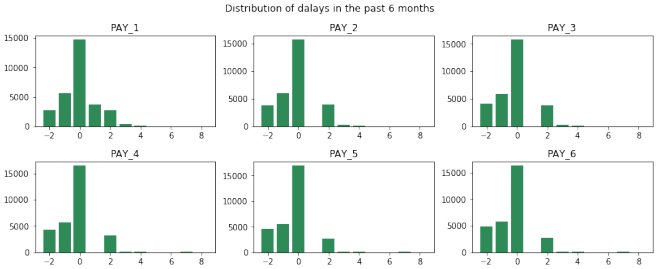


TABLE 3
FIGURE 4

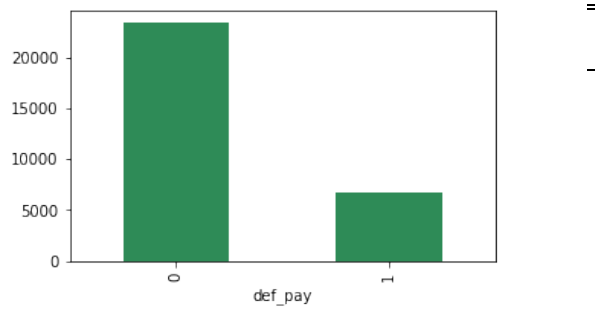
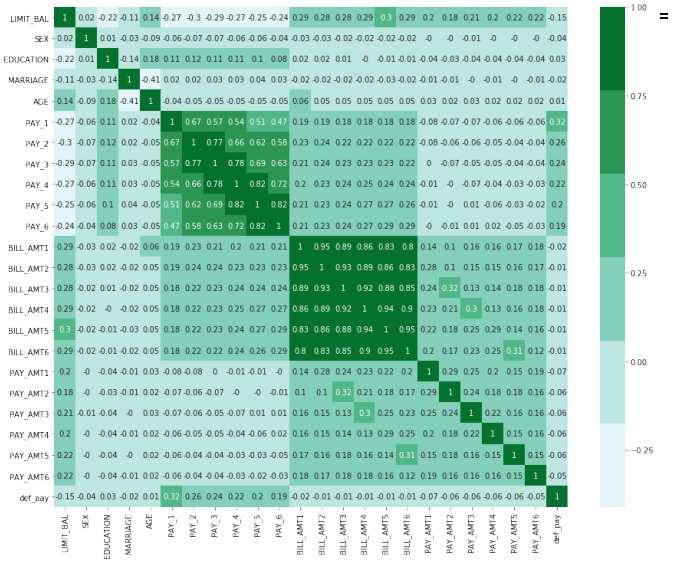


FIGURE 5



So it looks like the BILL_AMTX, PAY_X variables are the strongest predictors of default.

Then we will split the data and define the functions. After all these preparations, we will focus on prediction and evaluation.

C. *Evaluation metrics*

For we employed binary classification in this research, the discrimination evaluation of optimal methods of classification problems are based on confusion matrix. We apply the metrics of accuracy, precision, recall, F1-score, and ROC curve to value the quality of produced solution in this research (Table 1).

TABLE I
EVALUATION METRICS FOR CLASSIFICATION EVALUATIONS

Metrics	Equation	Evaluation Emphasis
Accuracy	$\frac{TP+TN}{TP+FP+FN+TN}$	Accuracy metric measures the ratio of correct predictions over the total number of instances evaluated.
Precision	$\frac{TP}{TP+FP}$	Precision metric measures the fraction of positive patterns that are correctly classified.
Recall	$\frac{TP}{TP+FN}$	Recall metric is used to measure the fraction of positive patterns that are correctly classified.
F1-score	$\frac{2*Recall*Precision}{Recall+Precision}$	F1-score metric represents the harmonic mean between recall and precision values.

Note: TP - true positive; FP - false positive; FN - false negative; TN - true negative

Additionally, we use ROC (Receiver Operating Characteristics) curve to visualize the performance of how well models are capable to distinguish between classes (D. J. Hand & R. J. Till., 2001). In the lift chart, the horizontal axis shows the number of total data. The vertical axis represents the cumulative number of target data. There are three curves– model curve and diagonal baseline curve, in the lift chart. Higher the ROC area, better the model is at predicting between classes.

D. *Results and analysis*

The evaluation metrics of the five machine learning methods are shown below (Table 2 and Table 3). From Table 2, we can see that based on accuracy evaluation metrics, in the training data, random forest model has the highest score (0.821), meanwhile basic logistic model and Stochastic Gradient Descent (Ridge) Model have relatively high score as well (0.8096 and 0.8023). In the test data, the situation is the same. Random forest has the highest score (0.814), Basic logistic model and Stochastic Gradient Descent (Ridge) Model have relatively high score (0.81 and 0.801). Because the test dataset is the effective dataset used to measure the generalization classification accuracy of models, so we can conclude that random forest model has best performance in the aspect of accuracy.

TABLE 2
CLASSIFICATION ACCURACY

Method	Model Score		Accuracy Score	
	Training set	Test set	Training set	Test set
Basic Logistic Regression	0.8096	0.81	0.8087	0.81
Logistic Regression with Grid Search (Lasso)	0.6944	0.6923	0.6924	0.6923
Stochastic Gradient Descent (Ridge)	0.8023	0.801	0.7983	0.801
Kernel SVN	0.7864	0.788	0.7832	0.788
Random Forest	0.821	0.814	0.8160	0.814

From Table 3, we can see that based on precision evaluation metrics, basic logistic regression model and Kernel SVN model have the highest score (0.70), Stochastic Gradient Descent (Ridge) Model and random forest model have relatively high value as well (0.69 and 0.66), but the precision is low of logistic regression model with Grid Search (Lasso). Based on recall evaluation metrics, logistic regression model with Grid Search (Lasso) has highest value (0.64), basic logistic regression model, Stochastic Gradient Descent (Ridge) Model and random forest model have medium value while Kernel SVN model has the lowest score (0.05). The models with higher recall metrics value are capable to catch more default, but will make more mistakes in identifying non-default at the same time. So if the bank cares more about identifying people who are going to default, that is, more conservative, the models may be a better choice.

To represent the harmonic mean between recall and precision values, we employed F1-Score. Logistic regression model with

Table 3
EVALUATION METRICS

Method	Precision	Recall	F1-Score	ROC Area
Basic Logistic Regression	0.70	0.23	0.34	0.72
Logistic Regression with Grid Search (Lasso)	0.38	0.64	0.48	0.73
Stochastic Gradient Descent (Ridge)	0.69	0.17	0.27	0.72
Kernel SVN	0.70	0.05	0.10	0.72
Random Forest	0.66	0.31	0.42	0.77

Grid Search (Lasso) has the highest score (0.48) and random forest model has the second highest score (0.42). The rest three are ranked as: basic logistic regression model (0.34), Stochastic Gradient Descent (Ridge) Model (0.27), and Kernel SVN model (0.10). So in the comprehensive evaluation of precision metrics and recall metrics, logistic regression model with Grid Search (Lasso) and random forest model are the best two models.

Additionally, we use ROC curve to visualize the performance of how well models are capable to distinguish between classes. The classification result shows the performance of the five data mining methods is ranked as: random forest model (0.77), Logistic Regression with Grid Search (Lasso) (0.73), Basic Logistic Regression (0.72), Stochastic Gradient Descent (Ridge) (0.72), Kernel SVN (0.72). To conclude, random forest model has the best capability to distinguish between classes.

V. CONCLUSION AND FUTURE WORK

In this thesis, we address the problem of customers' default payments in Taiwan and compare the classification performance of data points among five data mining techniques. We applied binary classification to the dependent variable, default payment, and employed the following 23 variables as independent variables. After splitting the data into training dataset and testing dataset respectively, we applied the metrics of accuracy, precision, recall, F1-score, and ROC curve to value the quality of produced solution in this research.

In the classification accuracy, random forest model has best performance in the aspect of accuracy. Basic logistic regression model and Kernel SVN model have the highest score in the precision perspective, but their recall metrics are relatively low. Meanwhile, based on recall evaluation metrics, logistic regression model with Grid Search (Lasso) has highest value (0.64), but its precision value is low. To represent the harmonic mean between recall and precision values, we employed F1-Score, logistic regression model with Grid Search (Lasso) and random forest model are the best two models in this case. Lastly, through ROC curve, we can conclude that random forest model has the best capability to distinguish between classes.

Several different adaptations, tests, and experiments have been left for the future. Future work concerns deeper analysis of particular mechanisms, new proposals to try different methods, or simply curiosity. For instance, we can apply the artificial neural network model, test its performance and compare it to the existing models. Moreover, it is necessary to update the data, because the data utilized in this study are not the newest.

REFERENCES

[1] Hand, D. J., & Till, R. J. (2001). A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine learning*, *45*(2), 171-186.

[2] Lee, Y. S., Yen, S. J., Lin, C. H., Tseng, Y. N., & Ma, L. Y. (2004). A data mining approach to constructing probability of default scoring model. In *Proceedings of 10th conference on information management and implementation* (pp. 1799-1813).

[3] Updegrave, W. L. (1987). How lenders size you up. *Money*, *16*(4), 145-151.

[4] 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.

FIGURE 6
BASIC LOGISTIC REGRESSION

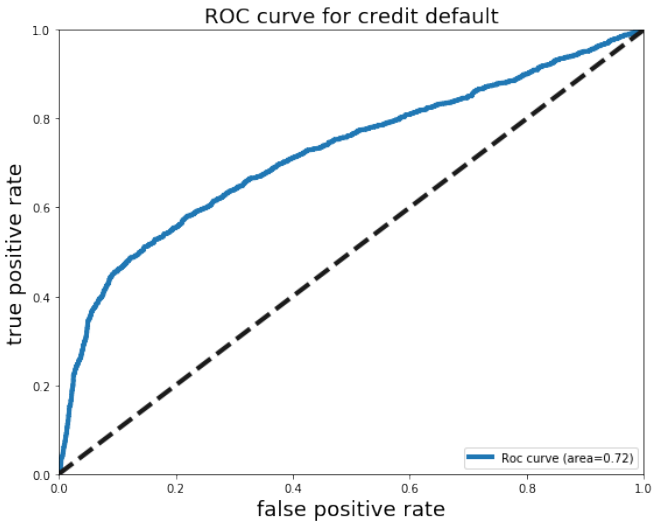


FIGURE 7
LOGISTIC REGRESSION WITH GRID SEARCH (LASSO)

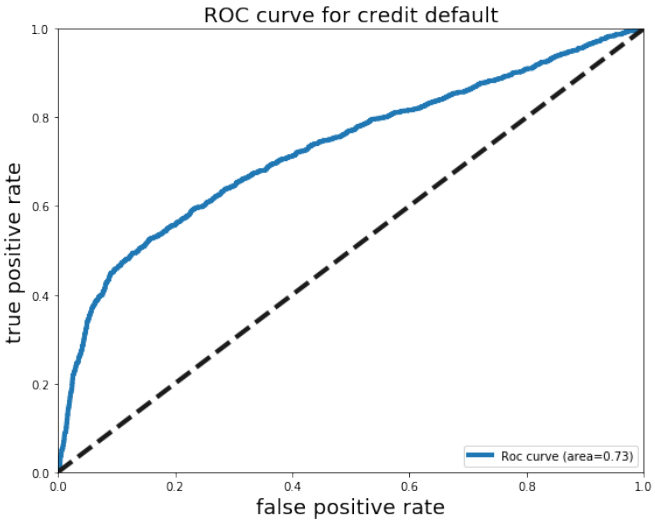


FIGURE 8
STOCHASTIC GRADIENT DESCENT (RIDGE)

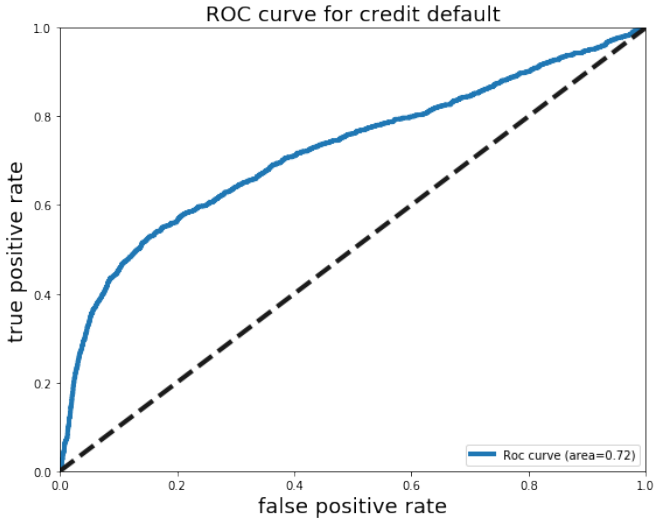


FIGURE 9
KERNEL SVN

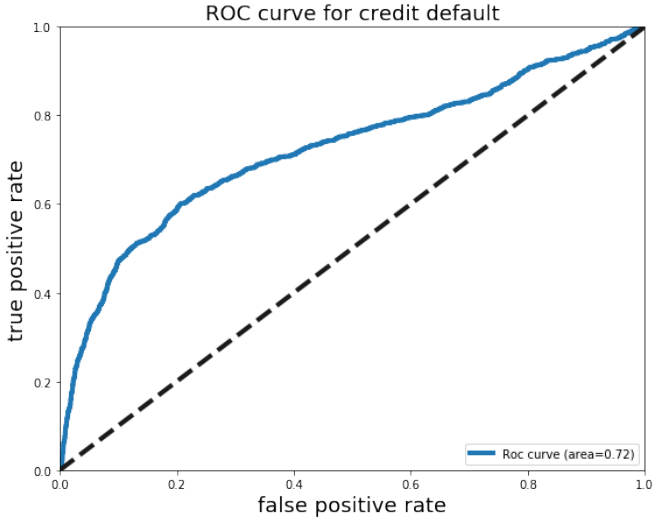


FIGURE 10
RANDOM FOREST