PERFORMANCE EVALUATION OF ML METHODS

Based on Credit Card Data Set

ABSTRACT

Hardly can we draw the conclusion that machine learning methods can do much better than classical ones. We implement several ML methods and traditional regression model as well, comparing between statistics in the generated confusion matrix, to test whether the ML can improve the performance of the model greatly.

Naixin Zhang, Mingxuan Wu

AAE 772: Economics of Machine Learning

Credit card project

Naixin Zhang, Mingxuan Wu

1. Introduction

The objective of this paper is to apply multiple popular machine learning methods to identify whether a credit card holder will default next month payment based on several attributes of the card holder such as payment history, education level, age and so on. By using the data provided by Professor Du and the knowledge we got from the course AAE 722: Machine Learning in Applied Economic Analysis. We will evaluate and compare the performance of traditional Econometrics model with different supervised machine learning methods in order to choose the most robust model according to the dataset background.

As a result, the lending institutions will be able to use our model to reduce the high delinquency rate by classifying the credit card defaulters and non-defaulters.

This paper mainly contains six parts:

- Introduction of this paper
- data preparation and introduction to the methods we will use in this paper
- Discussion among traditional Econometrics model and machine learning models using given dataset
- Make conclusion based on the results
- Reference
- Appendix (Separate R code)

2. Data and Methods

2.1 Introduction of the input and output variables

The dataset is related to customer's credit card payment default in Taiwan in October, 2005, there are 30,000 observations. This research employed a binary variable – default.payment.next.month: (Yes = 1, No = 0) whether the individual defaulted credit card payment in the next month, as the output variable. This study used the following 24 variables as input/feature variables:

- ID: ID of individual
- LIMIT_BAL: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.
- SEX: Gender (1 = male; 2 = female).
- EDUCATION: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).
- MARRIAGE: Marital status (1 = married; 2 = single; 3 = others).
- AGE: Age (year).
- PAY_0 - PAY_6: History of past payment. Past monthly payment records are tracked (from April to September, 2005) as follows:
    - PAY_0 = the repayment status in September, 2005; PAY_1 = the repayment status in August, 2005;

      . . .
    - PAY_6 = the repayment status in April, 2005.
- The measurement scale for the repayment status:
    - -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.

- BILL_AMT1 - BILL_AMT6: Amount of bill statement (NT dollar).
  - BILL_AMT1 = amount of bill statement in September, 2005;
  - BILL_AMT2 = amount of bill statement in August, 2005;

  . . .

  - BILL_AMT6 = amount of bill statement in April, 2005.
- PAY_AMT1 - PAY_AMT6: Amount of previous payment (NT dollar).
  - PAY_AMT1 = amount paid in September, 2005;
  - PAY_AMT2 = amount paid in August, 2005;

  . . .

  - PAY_AMT6 = amount paid in April, 2005.

2.2 Data and Model selection

2.2.1 Check missing values

First of all, we checked whether this dataset includes any missing values. The result indicates that there is no missing value in the entire dataset. After the cleaning process, we can investigate more details about the data.

Table 2.1 check missing values

	id	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	…	default.payment.next.month
# of missing values	0	0	0	0	0	…	0

2.2.2 Rename the dataset variable.

Notice that the first payment variable in the dataset is PAY_1 not PAY_0, which is a type error in the data set. For convenience, we can use the command names(dat)[names(dat)=="PAY_0"] <- "PAY_1" to rename the variable in the data set.

Table 2.2 List of Variable Names after revision of Type Error

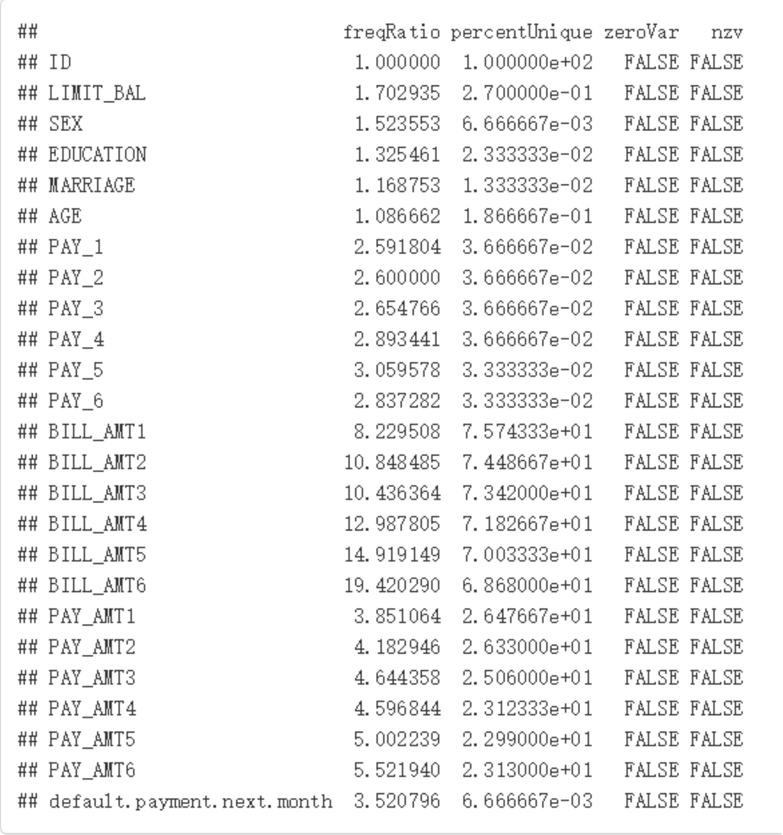
before	PAY_0	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6
after	PAY_1	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6

2.2.3 Zero- and Near Zero-Variance Predictors

In some situations, the data generating mechanism can create predictors that only have a single unique value (i.e. a “zero-variance predictor”). The concern here that these predictors may become zero-variance predictors when the data are split into cross-validation sub-samples or that a few samples may have an undue influence on the model. These “near-zero-variance” predictors need to be identified and eliminated prior to modeling.

we use the nearZeroVar function to detect if there are these predictors. The result above indicates that all variables in the column named zeroVar are all False. Thus, there is no such variables.

Table 2.3 check Zero- and Near Zero-Variance Predictors

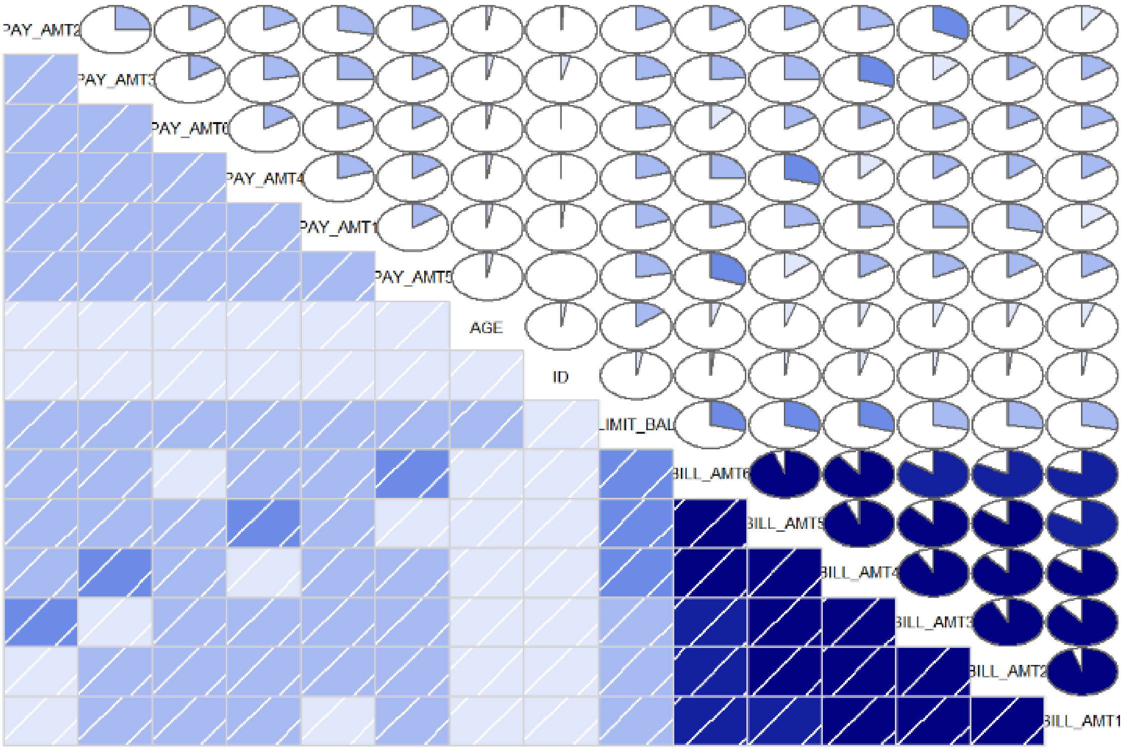


```
##                           freqRatio percentUnique zeroVar   nzv
## ID                         1.000000  1.000000e+02   FALSE FALSE
## LIMIT_BAL                  1.702935  2.700000e-01   FALSE FALSE
## SEX                        1.523553  6.666667e-03   FALSE FALSE
## EDUCATION                  1.325461  2.333333e-02   FALSE FALSE
## MARRIAGE                   1.168753  1.333333e-02   FALSE FALSE
## AGE                        1.086662  1.866667e-01   FALSE FALSE
## PAY_1                      2.591804  3.666667e-02   FALSE FALSE
## PAY_2                      2.600000  3.666667e-02   FALSE FALSE
## PAY_3                      2.654766  3.666667e-02   FALSE FALSE
## PAY_4                      2.893441  3.666667e-02   FALSE FALSE
## PAY_5                      3.059578  3.333333e-02   FALSE FALSE
## PAY_6                      2.837282  3.333333e-02   FALSE FALSE
## BILL_AMT1                  8.229508  7.574333e+01   FALSE FALSE
## BILL_AMT2                 10.848485  7.448667e+01   FALSE FALSE
## BILL_AMT3                 10.436364  7.342000e+01   FALSE FALSE
## BILL_AMT4                 12.987805  7.182667e+01   FALSE FALSE
## BILL_AMT5                 14.919149  7.003333e+01   FALSE FALSE
## BILL_AMT6                 19.420290  6.868000e+01   FALSE FALSE
## PAY_AMT1                   3.851064  2.647667e+01   FALSE FALSE
## PAY_AMT2                   4.182946  2.633000e+01   FALSE FALSE
## PAY_AMT3                   4.644358  2.506000e+01   FALSE FALSE
## PAY_AMT4                   4.596844  2.312333e+01   FALSE FALSE
## PAY_AMT5                   5.002239  2.299000e+01   FALSE FALSE
## PAY_AMT6                   5.521940  2.313000e+01   FALSE FALSE
## default.payment.next.month 3.520796  6.666667e-03   FALSE FALSE
```

2.2.4 Predictors' correlation detect

The package 'corrgram' enables us to visualize the correlation among variables in a direct way. We can see a high level of linear correlations between the amount of bill statements in different months. In case of the multicollinearity we need to use such techniques as Ridge and Lasso regression method.

Figure 2.1 check if there are correlated among numerical variables



2.2.5. resampling method

The data was randomly divided into two groups, we choose 70% of observations for model training and 30% of observations to validate the model. Because this paper aims to compare the performance of different models. To make the result more accurate, we can do standardization by doing ten-folder cross validation to resampling the training data for machine learning methods. For simplicity, we do not choose another method called repeated cross validation which may generate more accurate results.

2.3 Model selection

2.3.1. conventional method

- Logistic Regression

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is binary. Because the output variable in our dataset is binary, we use simple logistic regression to describe data and to explain the relationship between one dependent binary variable and independent variables.

2.3.2 Machine learning method

- K-Nearest Neighbors Classifiers

If we don't know the conditional probability distribution of Y given X, we always use this method. Given a test observation X_0, find the K training points X_r, r =1…, K closest in distance to X_0, represented by N_0.

- Lasso and Ridge regression

Lasso and ridge regression are two alternatives, or complements – to ordinary least squares (OLS). They both start with the standard OLS form and add a penalty for model complexity. The only difference between the two methods is the form of the penalty term.

Ridge regression uses the l_2-norm while lasso regression uses the l_1-norm. Specifically, the forms are shown below.

$$\text{Ridge Regression: } \hat{\beta}^{\mathbf{ridge}} = \mathbf{arg}\min_{\beta} \sum_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{y_i} - (\beta_{\mathbf{0}} + \beta^{\mathbf{T}}\mathbf{x_i}))^{\mathbf{2}} + \lambda\|\beta\|_{\mathbf{2}}^{\mathbf{2}}$$

$$\text{Lasso Regression: } \hat{\beta}^{\mathbf{lasso}} = \mathbf{arg}\min_{\beta} \sum_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{y_i} - (\beta_{\mathbf{0}} + \beta^{\mathbf{T}}\mathbf{x_i}))^{\mathbf{2}} + \lambda\|\beta\|_{\mathbf{1}}$$

- Decision Tree

Decision tree is a simple method for prediction, it's really popular because we it can easily be understand and explain. We can completely describe the algorithm and interpret the results.

- Random Forest

When building decision trees on bootstrapped training samples, each time a split in a tree is considered, a random sample of the predictors is chosen as split candidates from the full set of p predictors.

3. Discussion

The method we use in this paper for analyzing the performance of each model is confusion matrix, which is generally used as a performance measurement for machine learning classification problem. According to our dataset, a credit card company might particularly wish to avoid incorrectly classifying an individual who will default, whereas incorrectly classifying an individual who will not default, though still to be avoided, is less problematic. Since the matrix can tell us two types of errors, false positive (FP)and false negative (FN) rate. As we know low FN rate, low sensitivity. So, we will use both the accuracy and the sensitivity rate as the standard to compare the performance of different model. We can summarize the main output of our six models in one table.

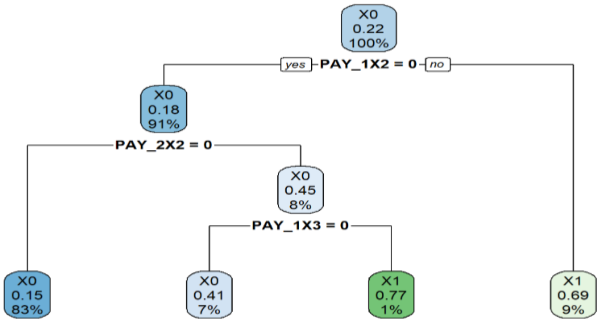
Table 3.1 Main output summary

Model name	Accuracy	Sensitivity	default rate next month
KNN model	76.7%	93.74%	8.17%
Random Forest	80.01%	98.11%	4.71%
Decision tree	81.89%	96.22%	9.54%
Ridge	81.93%	95.33%	10.99%
Lasso	82.03%	95.14%	11.37%
logistic model	82.03%	95.10%	11.44%

The table above indicates that KNN model has the lowest accuracy rate, lasso and logistic model has the most accuracy rate. However, the random forest model has the highest sensitivity and KNN has the lowest sensitivity. Balanced between accuracy rate and sensitivity rate, the random forest method has the best performance and KNN model has the lowest performance. And the random forest method predicts the default rate next month is 4.71%. That makes sense. The KNN method does not produce a simple classification probability formula and its predictive accuracy is highly affected by the measure of distance and the cardinality k of the neighborhood, after calculation, we choose k as 10 as the best tune for this model. Random forest aims to reduce the correlation issue we talked in data part before by choosing only a subsample of the feature space at each split. Essentially, it aims to make the trees de-correlated and prune the trees by setting a stopping criterion for node splits.

For logistic regression. The accuracy rate is 82.03%, which indicates that about 82.03% of test sets are predicted correctly It's a good performance. The prediction but we can't solve non-linear problems with logistic regression since its decision surface is linear.

Figure 3.1 Decision Tree



For decision tree model, the accuracy rate is 82.03% which indicates that about 81.89% of test sets are predicted correctly using only 3 predictors: PAY_0X2, PAY_2X2, PAY_0X3. Where PAY_1X2 means having payment delay for two months in September, 2005. PAY_2X2 means having payment delay for two months in August, 2005.PAY_1X3 means having payment delay for three months in September 2005.

The result shows that when a credit card company plans to issue the client a credit card, it's very important for the institution to check the payment history. The sensitivity is 96.22% which is a good performance for credit card company. However, decision trees have a lower accuracy compared with lasso method. It tends to have high variance when they utilize different training and test sets of the same data, since they tend to overfit on training data.

In ridge regression, we add a penalty by way of a tuning parameter called lambda which is chosen using cross validation. As lambda gets larger, the bias is unchanged, but the variance drops. After several times attempt, we choose alpha as 0 and lambda as 0.02. the drawback of ridge is that it doesn't select variables. It includes all of the variables in the final model.

In lasso, the penalty is the sum of the absolute values of the coefficients. Lasso shrinks the coefficient estimates towards zero and it has the effect of setting variables exactly equal to zero when lambda is large enough while ridge does not. Hence, lasso performs variable selection. The tuning parameter lambda is chosen by cross validation. When lambda is small, the result is essentially the least squares estimates. As lambda increases, shrinkage occurs so that variables that are at zero can be thrown away. After several attempts we choose alpha as1 and lambda as 0.001. A major advantage of lasso is that it is a combination of both shrinkage and selection of variables. In cases with very large number of features, lasso allow us to efficiently find the sparse model that involve a small subset of the features.

4. Conclusions

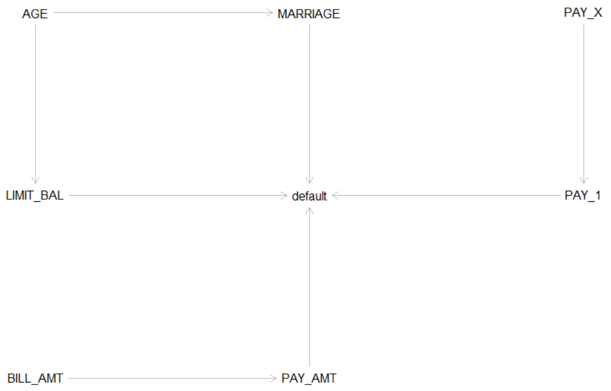
The objective of this paper is to train multiple supervised learning algorithms to predict customers' behavior on paying off credit card balance. We first investigated the data by using exploratory data analysis techniques including cleaning missing or invalid values and exploring the relationship between different features. We started with the logistic regression algorithm, then built KNN model, ridge model, lasso model. regression tree model and random forest model. By using accuracy and sensitivity rate to evaluate the model performance, we conclude that the random forest model has the best performance for customers' behavior.

We can also draw the causal diagram under the following rule,

1. Select all variables in the summary table of the Logit model which has significant coefficients.
2. Use the figure drawn by 'corrgram' package to pick variables with high correlations with the selected significant variables.
3. Put variables picked from step one at the vertical and horizontal position, and those picked up from step two at the corner position. Notice that AGE has correlation with two significant variables, so we cannot put LIMIT_BAL and MARRIAGE directly across the origin.
4. Put the outcome variable default.payment.next.month at the center of the figure, then draw the line indicating the relationship.

Then, we can get the figure which indicates the causal relationship in the credit card data set.

Figure 4.1 The causal relationship of each variable



We got some meaning points from the analyzing procedure above:

For model selection, we need to take background or feature of different datasets as consideration, for example, for credit card company they focus more on sensitivity rate rather than the traditional measurement way such as accuracy rate.

For binary dependent variable model, the logistic model gives us better accuracy performance among other machine learning methods. Moreover, we have run both regression methods and classification methods, the result shows that there is not too much difference in the level of performance between regression and classification given that the dependent variable is binary.

There are still a few possible improvements in the future. For example, because of the memory limits of our computer, we can't train the Neural nets model efficiently, especially training the Neural Net across different size of hidden layers, which may generate a better result as it is one of the most robust and popular algorithms.

5. References

[1] Yeh, I. and C. Lien. 2009. The Comparison of Data MiningTechniques for the Predictive Accuracy of Probability of Default ofCredit Card Clients. *Expert Systems with Applications* 36:2473-2480.

[2] James G, Witten D, Hastie T, et al. An Introduction to Statistical Learning: with Applications in R[M]// An Introduction to Statistical Learning. 2013.

[3] https://topepo.github.io/caret/index.html instruction for the package 'caret'.

[4] http://uc-r.github.io/dalex

[5]https://rawgit.com/pbiecek/DALEX_docs/master/vignettes/DALEX_caret.html#22_model_performance

[6] https://www.rdocumentation.org/packages/caret/versions/4.47/topics/train

[7]https://rstudio-pubs-static.s3.amazonaws.com/281390_8a4ea1f1d23043479814ec4a38dbbfd9.html

[8] https://jamesmccammon.com/2014/04/20/lasso-and-ridge-regression-in-r/

[9] https://www.datascience.com/resources/notebooks/random-forest-intro/