Data Analysis on Classification and Prediction

Li Sun

INTRODUCTION

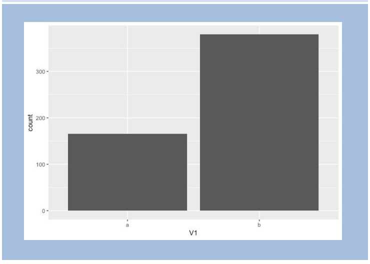
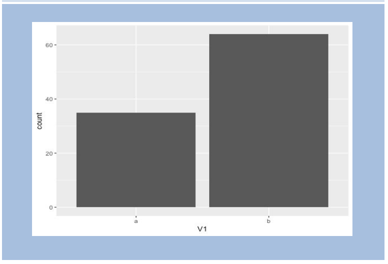
In this project, data set, *project.1.data.2.train.txt* and *project.1.data.2.test.txt* was analyzed. The data sets were associated with credit card applications. The original data was downloaded from *https://archive.ics.uci.edu/ml/datasets/Credit+Approval*, and had been splitted into a training data and a test data. The data analysis was focused on classification and prediction. A logistic regression model was built and it was used to obtain a classification rule which was used to predict whether a credit card application could be approved.

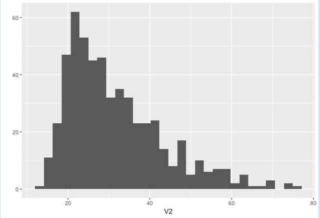
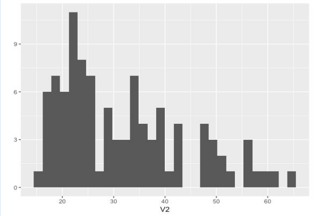
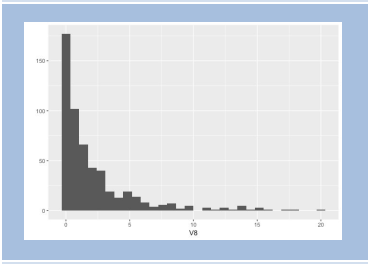
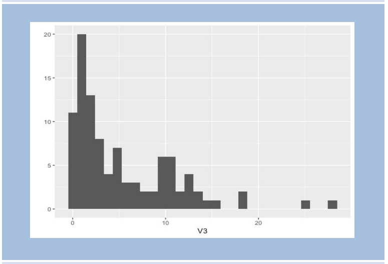
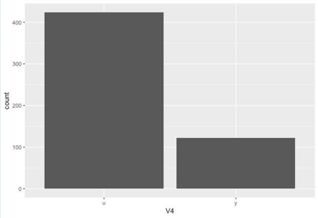
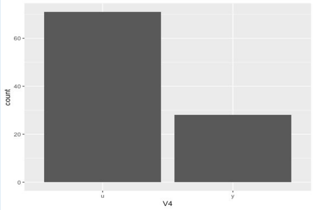
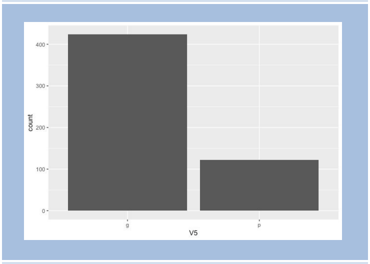
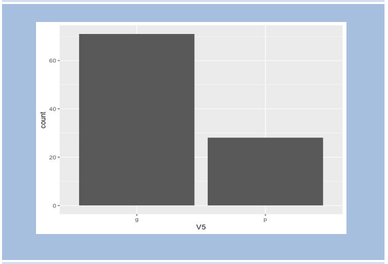
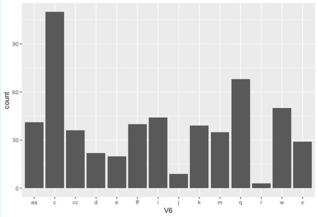
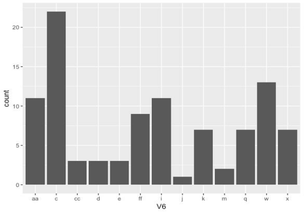
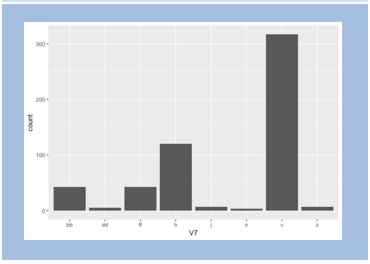
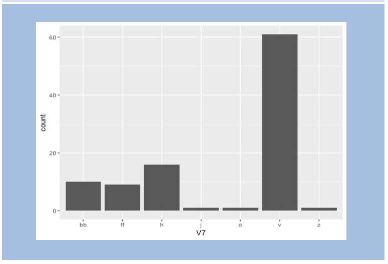
DATA DESCRIPTION

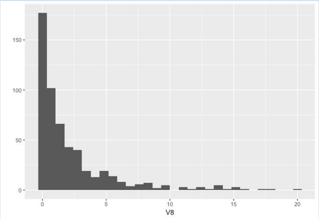
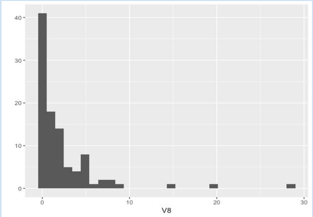
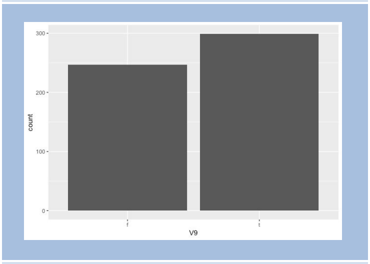
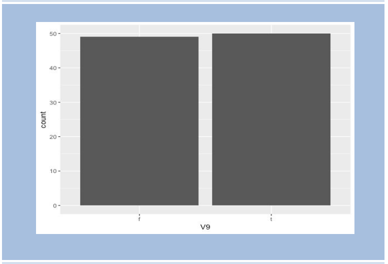
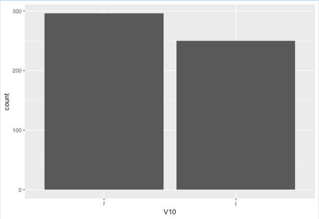
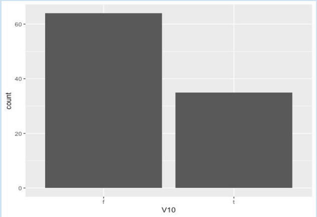
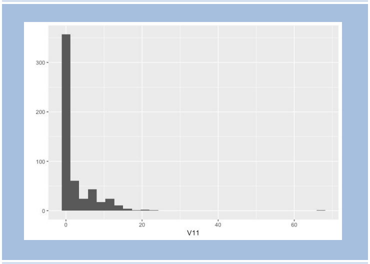
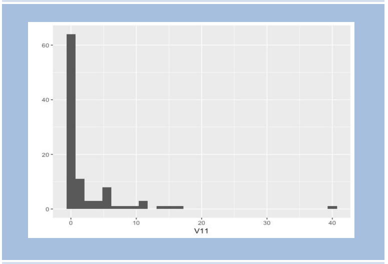
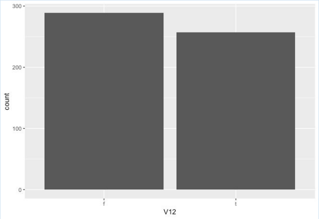
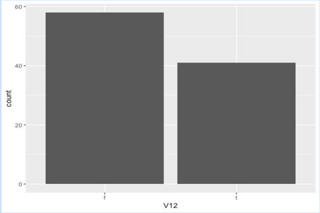
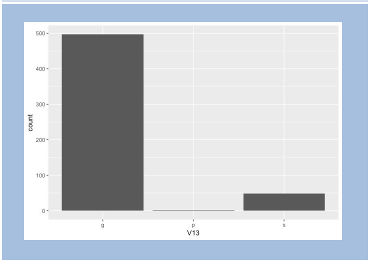
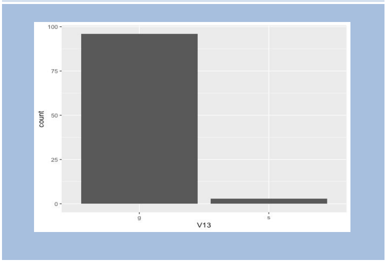
There are two datasets. One is training data, and the other is test data. The training data used in the research contains 583 observations with a total number of 16 attributes. The test data used in the research contains 100 observations with a total number of 16 attributes. Since the data involves private information, all attribute names and values have been changed to meaningless symbols to protect the confidentiality of the data. In both training and test data, the last attribute V16 is the response variable and is the one to be predicted. There are 6 continuous attributes, which are V2,V3,V8,V11,V14 and V15. There are 9 categorical attributes, which are V1, V4, V5, V6, V7, V9, V10, V12 and V13. There are also a few missing values. After removing the missing values and one observation in V4 with level 'l', there were 546 observations left in the final training data, and 99 observations left in the final test data.

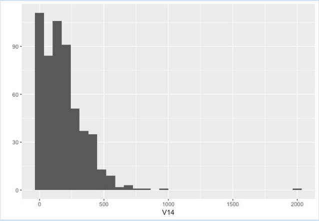
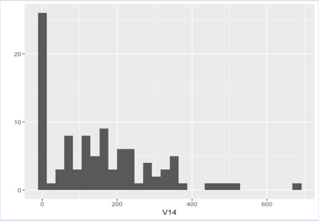
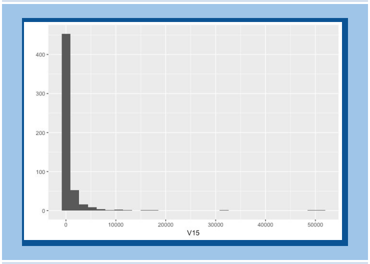
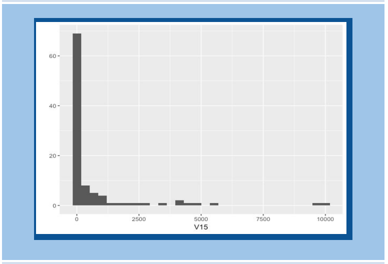
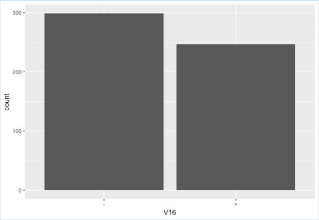
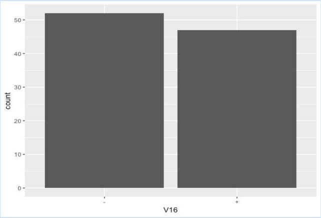
The 16 attributes were described in table 1 below. The listed information included attribute, type and histogram plot of training data and test data.

Table 1: Attribute Information

ATTRIBUTE	TYPE	HISTOGRAM PLOT (Training Data)	HISTOGRAM PLOT (Test Data)
V1	CATEGORICAL a b		

V2	NUMERICAL		
V3	NUMERICAL		
V4	CATEGORICAL u y		
V5	CATEGORICAL g gg p		
V6	CATEGORICAL aa, c, cc, d e, ff, i , j k, m, q r, w, x		
V7	CATEGORICAL bb, dd, ff, h, j, n, o, v, z		

V8	NUMERICAL		
V9	CATEGORICAL f t		
V10	CATEGORICAL f t		
V11	INTEGER		
V12	CATEGORICAL f t		
V13	CATEGORICAL g p s		

V14	INTEGER		
V15	INTEGER		
V16	CATEGORICAL + -		

PLOT INTERPRETATION

In order to obtain inspection about distribution of the attributes in the training data over the corresponding subsamples (247 V16 with value ‘+’, and 299 V16 with value ‘-’), histogram for continuous variables and bar charts for categorical variables of 16 attributes were plotted in Figure 1. In the figures, the frequency of V16 with level ‘-’ was shown in blue, and the frequency of V16 with level ‘+’ was shown in red. The histograms of continuous variables were shown in figure 1.1, and bar charts of categorical variables were shown in figure 1.2.

In this research, the level ‘+’ in V16 was defined as success, and level ‘-’ was defined as failure. The plots showed that, for variables V8, V11 and V15, the probability of success increases while their values increase. On the contrary, the probability of failure increases while value of V14 increases. Between the two levels of V4, the probability of success for level ‘y’ is lower than that for level ‘u’. Among the 14 levels of V6, level ‘cc’, ‘r’, ‘q’ and ‘x’ have more chance to be successful, and level ‘ff’ and ‘i’ have high probability to be failure. Between the levels of V9, level ‘t’ has a much higher chance to be successful than level ‘f’. Between the levels of V10, level ‘t’ also has a higher chance to be successful than level ‘f’ Therefore, from the exploring the plot information, we concluded that the variables V4, V6, V8, V9, V10, V11, V14 and V15 might be the factors that have important effects on credit card approval.

Figure 1: Plots of subsamples in the training data

Figure 1.1 histogram of subsamples for continuous variables

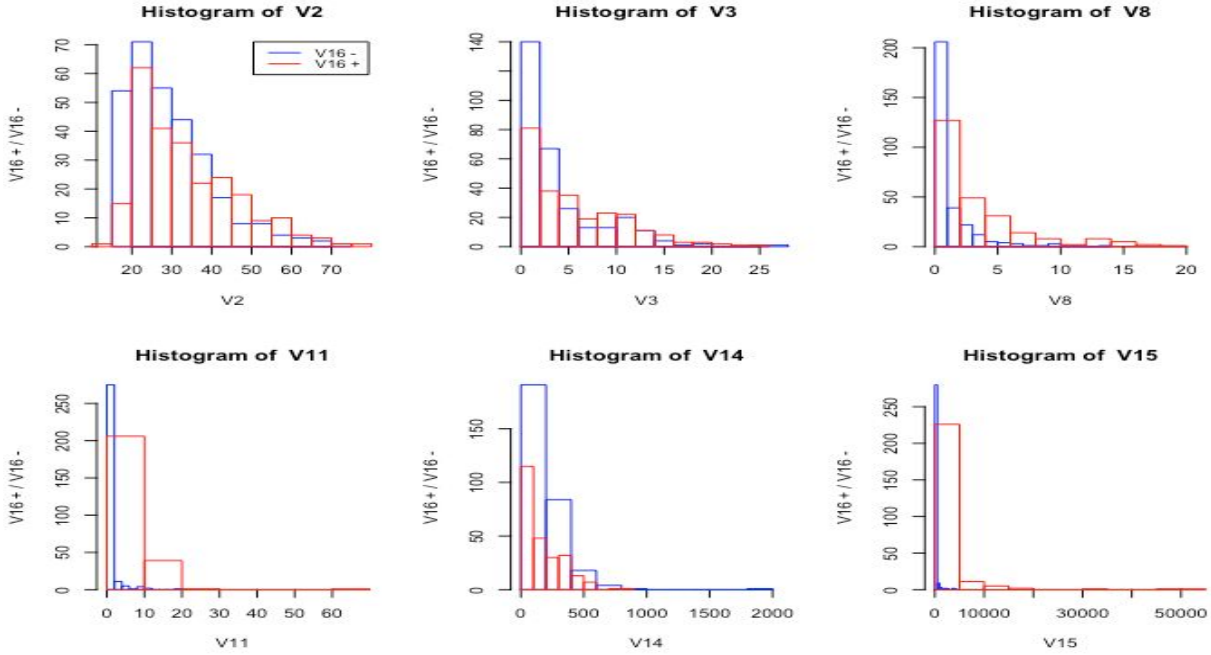
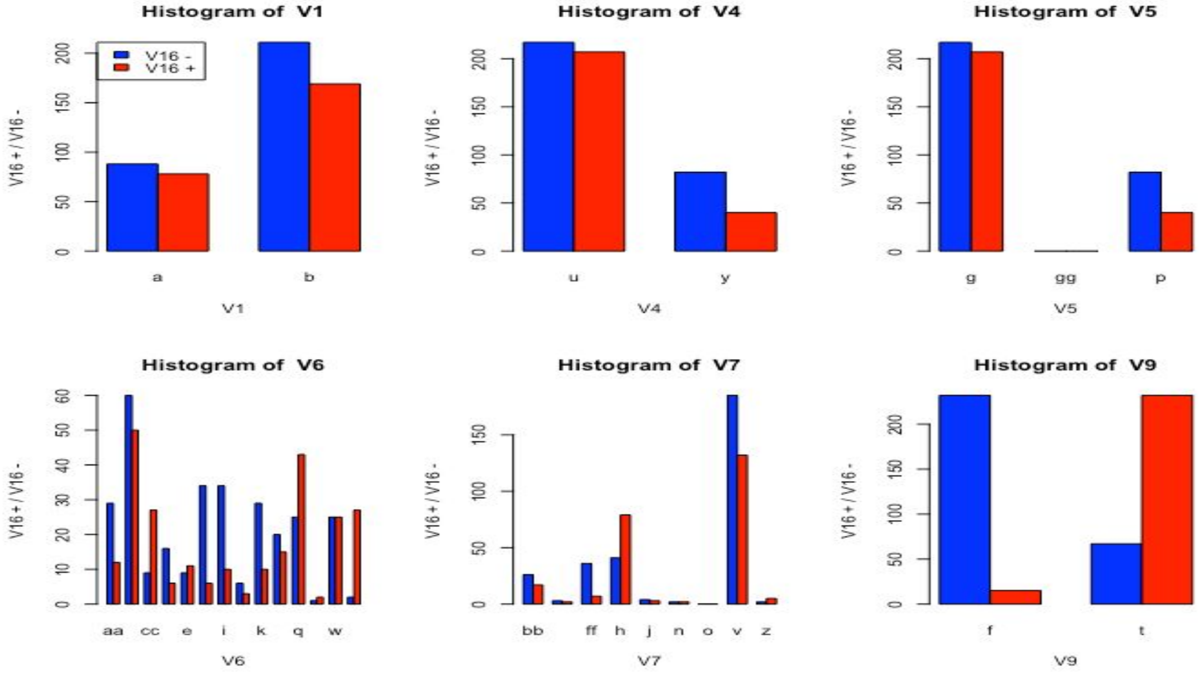
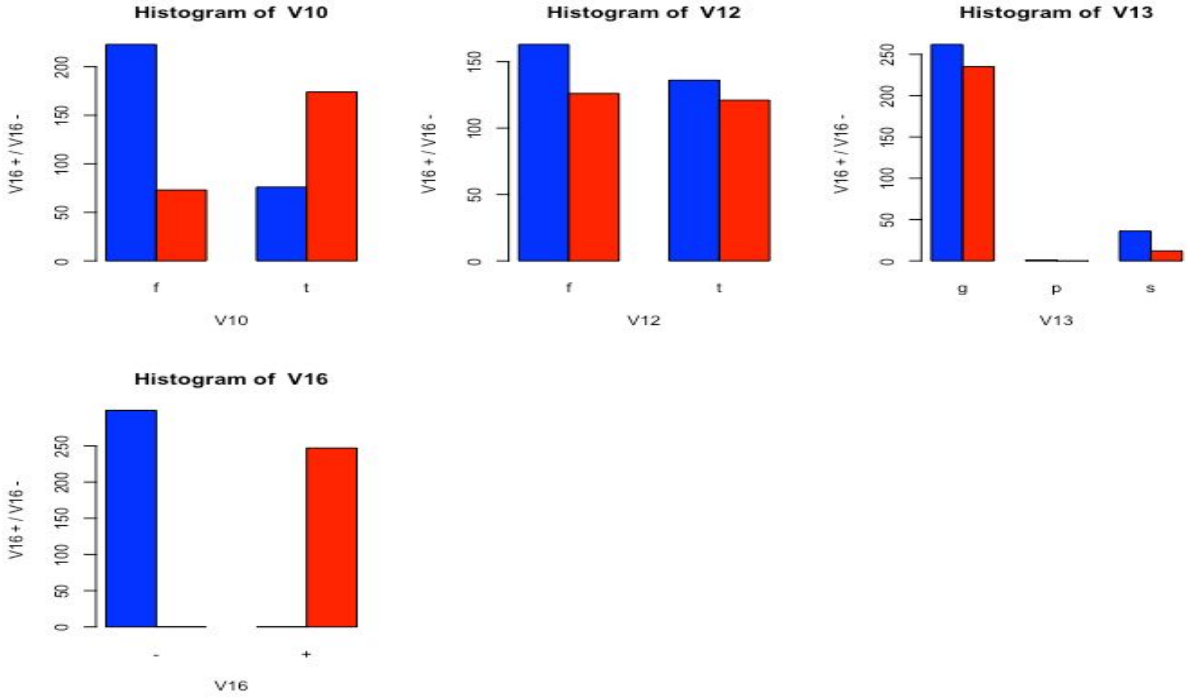


Figure 1.2 bar charts of subsamples for continuous variables



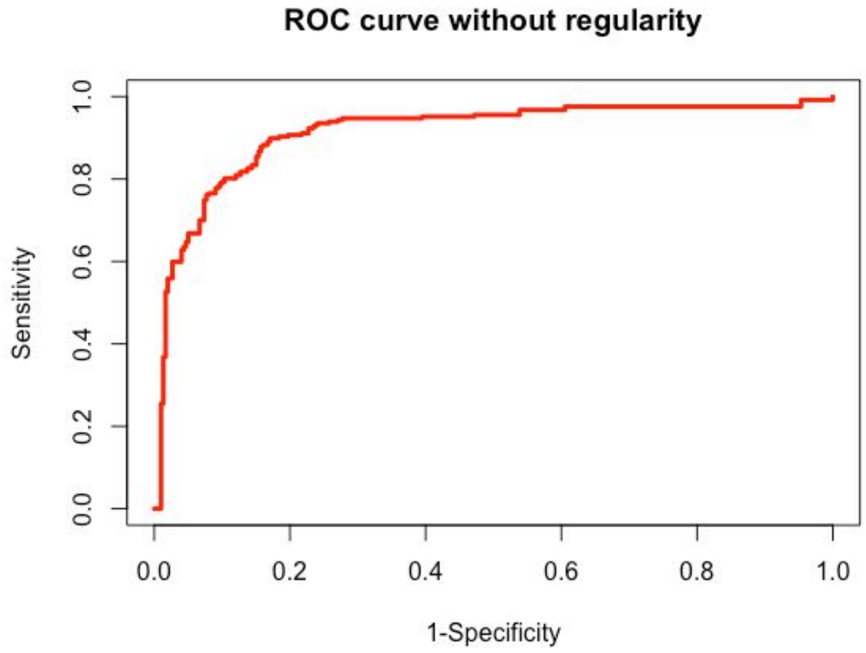


DATA ANALYSIS

In this project, two methods were used to build a logistic regression model and to determine the classification rule. The first method was implemented without considering regularity, and the second method was implemented with regularity. The details of these two methods were explained as follows.

In the first method, forward selection was applied to do the logistic regression selection. As a result, variables V4,V6, V8, V9, V11, V14 and V15 were selected and used in leave-one-out procedure to the training data. ROC curves were plotted (shown in figure 2) and AUC was calculated, which is 0.9137. The best cut-off point was π_0 identified as 0.46465. The prediction accuracy to the training data is 86.08%, which is high enough to prove that the model selected is acceptable. Then the selected model and the established classification rule were applied to the test data, and the classification accuracy for prediction whether a credit card application could be approved was 0.91919.

Figure 2: Plot of ROC curve without considering regularity



The second method was implemented with regularity. Both ridge penalty and lasso penalty were utilized to build their own logistic regression model and to choose their corresponding tuning parameter and cut-off point π_0. The result of the tuning parameter and optimal cut-off point for ridge penalty were 0.1309 and 0.5051 respectively, and the result of the tuning parameter and optimal cut-off point for lasso penalty were 0.0054 and 0.5456 respectively. After generating the final classification rules, the rules were applied to the test data. Ridge penalty method had a 0.8889 classification accuracy, and Lasso penalty method generated a 0.8990 classification accuracy.

COMPARISON AND COMMENT

The data in table 2 included the procedure time and classification accuracy for each method. It clearly demonstrated that the method without regularity, which was forward model selection with leave-one-out procedure, took the longest procedure time, but presented the highest classification accuracy. So if the size of a dataset is not too big, the combination of model selection and leave-one-out procedure might be a good choice for logistic regression model selection and classification rule determination for prediction purpose. But if a dataset is big and consuming time is a big concern, the method with regularity might be useful. The method with regularity can find a reduced set of variables resulting in an optimal performing model. In this scenario, if all variables need to be incorporated in the model according to domain knowledge, ridge penalty regression should be applied. If the number of variables is very large, lasso penalty regression is a suitable method, which only keeps the most significant variables in the final model.

Table 2: Result comparison

Method	Procedure	Running time	Classification Accuracy
Without regularity (100 iterations)	Forward selection Leave-one-out	394.116	0.91919
With regularity (100 iterations)	Ridge penalty	195.053	0.88889
	Lasso penalty	148.296	0.89899

CONCLUSION

In the research, data sets, which were associated with credit card applications, were analyzed. Two methods, one without regularity and the other one with regularity, were performed to generate classification rules. The rules were also used to test data to yield the classification accuracy for prediction purpose. In this work, the combination of forward model selection and leave-one-out procedure, which is one of the methods without regularity, produced the highest accuracy, but it also took the longest time. Compared with the method without regularity, the ridge and lasso penalty regressions provided reasonable accuracy and efficiency.