Credit Rating Analysis with Decision Tree Method in Housing Loans through the United States

Yuechen Mu

muyueche@msu.edu

Department of Mathematics, Natural Science College, Michigan State University, East Lansing, Mi, USA

Abstract----With the development of society, the problem of buying houses has become one of the main debt pressures of contemporary people, and housing loans can effectively alleviate this situation. This article tries to use data statistics to discover the preferences of loan groups and find the best classification method for predicting the housing loan people. In order to analyze more accurately, all personal data are sampled from the United States, and their influence on housing loans is judged by comparing individuals' gender, marriage, graduation, employment relationship, and income. This article will use three statistical algorithms for data testing and training, include Decision Tree, Artificial Neural Network, and Support Vector Machines. Through sampling training of seventy percent of the total data, a statistical model is built based on the above three methods. Then the remaining data is used for testing respectively to verify or modify the training model to improve the accuracy of prediction. The final conclusion is that the sampled data cannot be one hundred percent through the evaluation parameters to get the result of the housing loan, the main reason is that the sources of parameters are limited. We can only assess the basic conditions of people, but there are no more sources of information to fully label individuals. In addition, the differences in loan conditions within each state have caused statistical difficulties in national data. However, the Decision Tree is the way to do the best prediction, it can still be used to verify and analyze certain filtered data.

Keywords: Decision Tree; Neural Network; Support Vector Machines

1 INTRODUCTION

The housing purchase problem is regarded as one of the main causes of income pressure, and housing loans provided by society can effectively alleviate this problem. This article aims to collect the characteristics and preferences of the loan crowd and use the mathematical classification method to study the big data, in order to evaluate the law of whether an individual is a housing loan and find the best prediction method. This article plans to use three statistical tools to classify, then compare and analyze which method can better predict the housing loan crowd. In order to prepare for my research, I first learned about classification methods and the programming process of classification software, which I think can effectively help me solve practical problems.

ICIDC 2022, June 17-19, Qingdao, People's Republic of China

DOI 10.4108/eai.17-6-2022.2322597

Some scholars have used similar classification methods to study different issues. According to S. R. Safavian and D. Landgrebe (1991) [1], they defined the decision tree classifier and the relationship between neutral networks and decision tree. And for Song, Yan-Yan, and Ying Lu (2015) [2], their mathematic group argue that decision tree structure and three methods to build a decision tree model. The following academic articles are researches on neural networks. In M. Feindt and U. Kerzel's (2006) [3] article, they discuss how to use Bayesian statistics in the neural network package and noted that it can be used to predict the probability density distribution. For H. A. Rowley, S. Baluja, and T. Kanade (1998) [4], they used the neural network to solve the applied problem and show how it works for face detection. Then, in the support vector part, In S. V. M. Vishwanathan and M. Narasimha Murty's (2002) [5] articles, they mentioned the fast iterative algorithm to support the support vector machine and compare the new approach with the traditional iterative algorithms. Moreover, in Vladimir Cherkassky and Yunqian Ma (2004) [6] supported the regression of the support vector machine, and further discuss the least-modulus.

This article intends to use classification methods to predict the situation of housing loans. First, the data will be processed through the clustering method, and then the model will be built through decision trees, neural networks, and support vector machines. Finally, the prediction conclusion of the three classification methods will be compared. And use the accurate rate of each results to choose the best method for predicting housing loans.

2 MATERIALS AND METHODS

We first searched for random data of 700 residents from the United States, and their housing loans were random. The personal information of these residents includes: Gender, Married Status, Education Status, Employed Status, Applicant Income, Co-applicant Income, Loan Amount and Loan Status. When we removed the missing error data of those individual parameters, we finally left a complete set of 500 random data.

Because there are more references for finance, in order to facilitate subsequent classification, we first group all income data into new clusters. The best way to do the cluster is the K-means process, we divide everyone's finance into group 1 and group 2. And use the finance group to parallel the other parameters. The following figure 1 and 2 show the result of the clustering. They are a total of 497 persons in the finance group1 and only 7 persons in the finance group 2. This clustering result is based on the assessment of income based on three finance items which included Applicant Income, Co-applicant Income, Loan Amount. The clustering result based on only one item will be more uniform, but because of the participation of other parameters, the clustering process becomes more demanding. Therefore, the size of people in the second group is very small.

1	Male	No	Graduate	No	5849	0	66	Y	1
2	Male	Yes	Graduate	No	4583	1508	128	N	1
165	Male	Yes	Graduate	No	33846	0	260	N	2
166	Female	Yes	Graduate	No	3625	0	108	Y	1
167	Male	Yes	Graduate	Yes	39147	4750	120	Y	2

Figure 1 The example data after clustering

Group	1	493.000
	2	7.000
valid		500.000
Missing		.000

Figure 2 The result of number of clustering

2.1 Decision Tree

Decision tree is one of our considered models. It can build a model based on known information, and then use the testing data to recheck the accuracy of the model. We extract 300 data from the cleaned 500 data to construct the training set, and determine the choice of the root node through the calculation of entropy.

In 300 people, there are 93 people did not have a housing loan, and 207 people chose a housing loan. Then without any conditions, the original entropy is:

E(original)= $\frac{-93}{300}\log_2\frac{93}{300}-\frac{207}{300}\log_2\frac{207}{300}$ =0.8932

In order to test the root node, we separately calculate the entropy of Gender, Married Status, Education Status, Employed Status and Finance Group.

In the gender part, there is a total of 242 males and 73 of them do not have a housing loan. And there is a total of 58 female, 20 of them do not have the housing loan.

E(male)= $\frac{-73}{242}\log_2\frac{73}{242}-\frac{169}{242}\log_2\frac{169}{242}$ =0.8833

E(female)=$\frac{-20}{58}\log_2\frac{20}{58}-\frac{38}{58}\log_2\frac{38}{58}$=0.9294

E(total)=$\frac{242}{300}\times 0.8833+\frac{58}{300}\times 0.9294$ =0.8922

Gain(gender)=0.8932-0.8922 =1×10^{-3}

Operate the other groups in the same way, we can get the answer as below:

Gain(married)= 4.77×10^{-3}

Gain(Education)= 7.51×10^{-3}

Gain(employed)=0.2651

Gain(finance)= 0.4696

0.4686>0.2651>7.51×10^{-3}>4.77×10^{-3}>1×10^{-3}

The gain is the original entropy minus the group entropy. When the entropy decrease more, it will be the effective way. Thus the largest number in the gain group is from finance, so the root node we decided as the Finance group.

They are two groups in the finance part, take finance group 1 as the example. They are 295 people include in it. We have the remain four parameters, they are Gender, Married Status, Education Status, Employed Status.

Now in 295 persons, there are 203 have the housing loan and 92 persons not have it.

E(original with finance group)=

$$\frac{-92}{295}\log_2\frac{92}{295}-\frac{203}{295}\log_2\frac{203}{295}=0.8953$$

In a group of Married Status, they are 193 married and 55 of them do not have housing loan. And there are 102 of them not married and 37 of them do not have housing loans.

E(married)= $\frac{-55}{193}\log_2\frac{55}{193}-\frac{138}{193}\log_2\frac{138}{193}$ = 0.8621

E(not married)= $\frac{-37}{102}\log_2\frac{37}{102}-\frac{-65}{102}\log_2\frac{65}{102}$ =0.9449

E(total)= $\frac{193}{295}\times 0.8621+\frac{102}{295}\times 0.9449=0.8907$

Gain(married)=0.8953-0.8907= 4.61×10^{-3}

Next, we use the same way to recalculate the remain part, the largest gain is Married Status, so the next node below the Finance group will be married and not married.

Then, we repeat this process again, until all the nodes are calculated, the final order is, Finance Group, Married Status, Education Status, Gender and Employed Status. In Figure 3, it shows the final conclusion of decision tree.

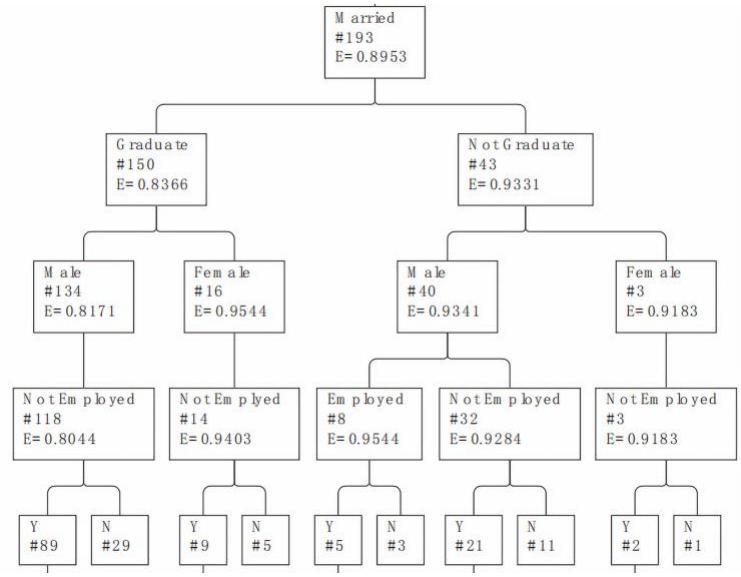

Figure 3 examples of the conclusion of decision tree

According to the entropy reduction principle of the decision tree, we have deleted some unreasonable nodes. Although after all classifications, we cannot get a clear result, then under the final branch, the number of housing loans is greater than the number of people who do not have housing loans, then it is judged that people in this branch will choose loans.

We use the remaining 200 data to test the model, and finally, make statistics on the error value of each branch. They are total of 41 persons in the testing data we can not determine them by the training model.

Accuracy= $1 - \frac{41}{200} = 0.795 = 79.5\% \approx 80\%$

This final prediction conclusion did not reach our ideal state, so we decided to model again through neural networks and support vector machines, and finally compare and analyze the accuracy results to find the best methods.

2.2 Neural Network

Neural network is used as the second method to evaluate the data and classify the results through the operation of the program. We convert Gender, Married Status, Education Status, Employed Status and Finance Group into digital codes and use them as input, and use House Loans as output to extract 350 from 500 data for model construction. Then use the remaining data to do the test. The process of program shows in Figure 4.

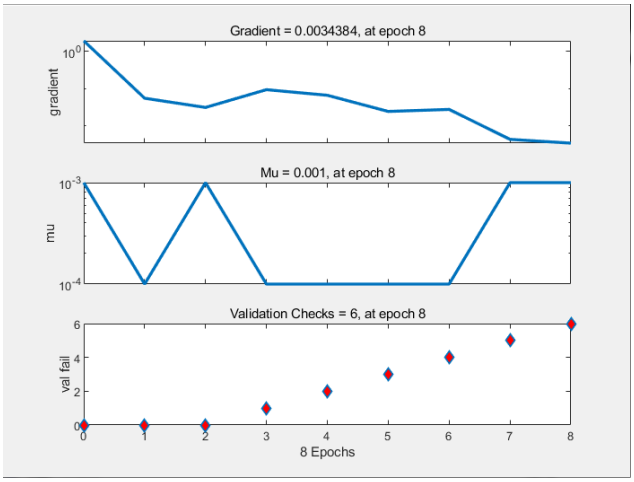

Figure 4 the Training State of Neural Network

When we determine the number of hidden neurons, when the hidden neurons is five, the conclusion shows clearer. We decided to use Levenberg-Maquardt to set up the model, the performance shows in Figure 5, 6 and 7.

Results

	Samples	MSE	R
Training:	350	2.18679e-1	1.19910e-1
Validation:	75	2.16679e-1	2.42611e-2
Testing:	75	1.70810e-1	2.40723e-1

Figure 5 the performance of Neural Network

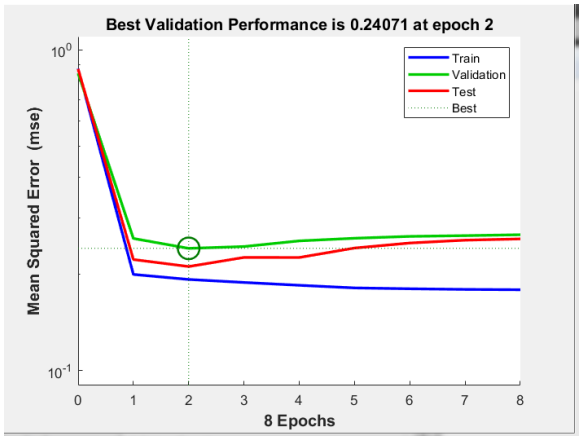

Figure 6 the performance of Neural Network

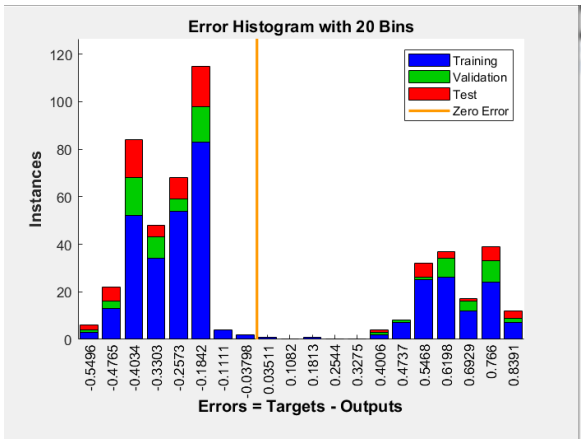

Figure 7 the error histogram of Neural Network

The final evaluation of the data is about regression which shows in Figure 8. However, the effect is not very satisfactory, the value of regression is a bit too low.

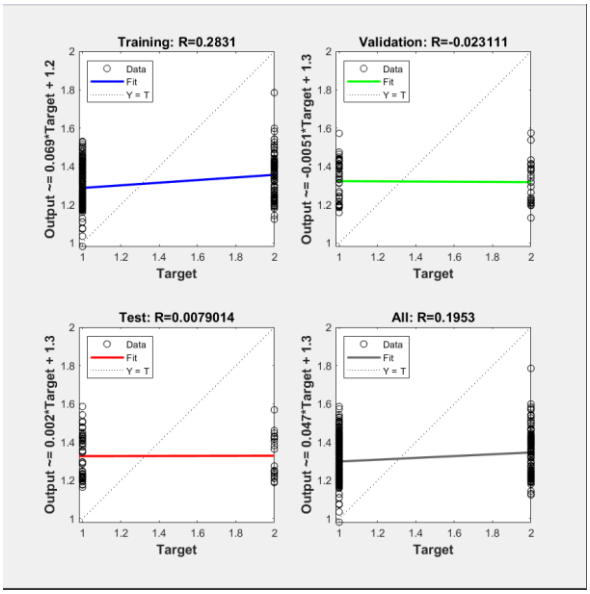

Figure 8 the regression of Neural Network

Accuracy= 1- 0.24=0.76=76%

2.3 Support Vector Machines

The final method is the support vector machine. It used to find the linear classifier with the largest spacing in a special space. Because of the characteristics of the model, we no longer cluster the interior of the financial group, but directly observe the linear relationship of all parameters.

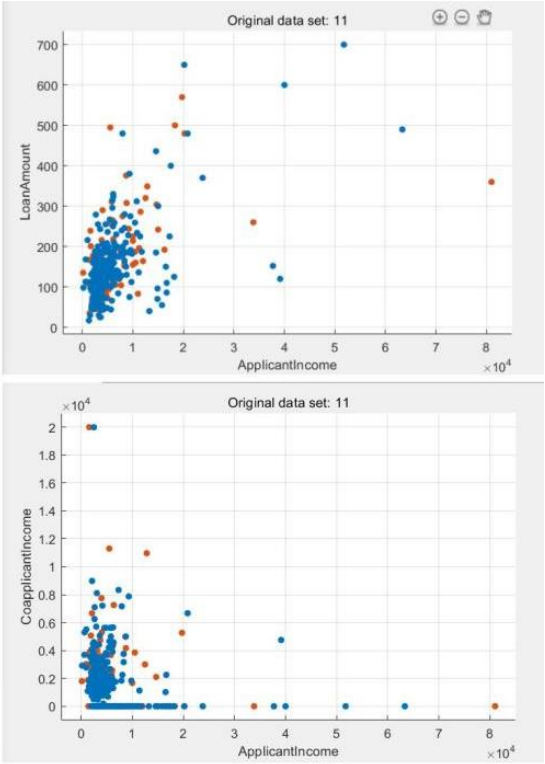

Figure 9 the relationships in Finance group

Then we estimate the average of the two linear relationships. Accuracy=0.57=57%

3 COMPARE AND CONTRASTS

Compare the accuracy of the three methods 80%>76%>57.2%, so the decision tree is still a more suitable model for this situation.

The amount of parameters of known complete data is limited, and we cannot fully evaluate an individual through the current database, such as the number of their family members, insurance, and other information. The least parameters make it impossible for us to continue to classify the housing loans crowd in detail, so in the end, the effects of all models did not have a result of more than 95%.

The possible social reasons maybe because each state has different policies on housing loans. Some states will reduce the procedures and requirements for housing loans on credit and marital status. Therefore, there are still local differences in the data sampled from the United States, and the current parameters cannot meet the situation base on the large population base.

4 CONCLUSION

Based on the comparative analysis of decision tree, neural network and support vector machine, we finally concluded that the decision tree is the statistically optimal result for the preference of the housing loan crowd. Although the prediction model of the decision tree cannot be one hundred percent same as the actual data, it still has an important reference value for predicting housing loan problems. The ultimate reason for this result is the limitation of the number of parameters. Based on the known information, we can only get the individual's gender, marital status, graduation status, employment and income for analysis. If there are more detailed background parameters, then the classification result might be more specific. At the same time, because it is an assessment of house loan in United States, we cannot ignore the influence of the different policies of house loan in different states.

Acknowledgment: Yuechen Mu thanks all the data from Kaggle.

REFERENCE

[1] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," in IEEE Transactions on Systems, Man, and Cybernetics, vol. 21, no. 3, pp. 660-674, May-June 1991, doi: 10.1109/21.97458.

[2] Song, Yan-Yan, and Ying Lu. “Decision tree methods: applications for classification and prediction.” Shanghai archives of psychiatry vol. 27,2 (2015): 130-5. doi:10.11919/j.issn.1002-0829.215044

[3] M. Feindt, U. Kerzel, “The NeuroBayes neural network package”, Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, Volume 559, Issue 1, 2006, Pages 190-194, ISSN 0168-9002,
https://doi.org/10.1016/j.nima.2005.11.166.

[4] H. A. Rowley, S. Baluja and T. Kanade, "Neural network-based face detection," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 1, pp. 23-38, Jan. 1998, doi: 10.1109/34.655647.

[5] S. V. M. Vishwanathan and M. Narasimha Murty, "SSVM: a simple SVM algorithm," Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No.02CH37290), 2002, pp. 2393-2398 vol.3, doi: 10.1109/IJCNN.2002.1007516.

[6] Vladimir Cherkassky, Yunqian Ma, “Practical selection of SVM parameters and noise estimation for SVM regression”, Neural Networks, Volume 17, Issue 1, 2004, Pages 113-126, SSN 0893-6080, https://doi.org/10.1016/S0893-6080(03)00169-2.