Analysis for Clients Churn of Credit Cards in Model Construction in Banking Industry

Abstract:

Data mining technology has been more and more important in the economics and financial market. Helping the banks to predict a customers' behavior, which is that whether the existing customers will continue use their credit cards or not, we utilize the data mining technology to construct a convenient and effective model, Decision Tree. By using our Decision Tree model, which can classify the customers according to different features step by step, the banks are able to predict the customers' behavior well. The main steps of our experiment includes collecting statistics from the bank, utilizing Min-Max normalization to pre-process the data set,employing the training data set to construct our model, examining the model by testing data set, and analyzing the results.

Claims

The method we used to forecast the credit card churning contains the following steps: attaining data from one of the most authoritative banks. Then, through utilizing csv-reader reading the data sets, we employ the Min-Max method to normalize the data and partition data sets.

And we select the Decision Tree(DT) as our experiment's model. After model training and model test, we conclude our experiment through our DT models.

INTRODUCTION AND BACKGROUND

As a result of the advance of the banking system, banking services become an indispensable part of people's daily life. Some banking sectors and other such organizations furnish clients with a variety of financial products and services while managing different risks, such as credit and finance risks. These loans provided by banks involve a lot of money, and failure to collect them will result in significant losses for financial institutions. In which case, it is greatly crucial for these banking sectors to get an accurate assessment of clients before these banking organizations decide whether to lend money to some clients. Besides to minimizing the risks of granting credits, it is also of importance for banking sectors and such leading organization to reduce the errors in declining the number of the valid clients, with the fierce competition among banking organizations, which is extremely important for banking sectors. To achieve above purpose, an increasing number of banking sectors start to utilize the credit scoring and data mining technologies to

identify clients' credibility or make some decisions in order to get access to greater yields in recent years.

Considering the banking theory as a part of economic theory, the reasons for the development of Banks as a mechanism are obvious. Tracing back to the early civilization of Mesopotamia tribes in 6000BC, bartering system firstly was regard as an exchanging way. People of that areas, such as Phoenicians and Babylonians, bartered goods to local people or those located in various other cities across oceans. This period was followed by another, in which the exchange of money replaced barter to a large extent. The first paper money that can be traced appeared in China during Song dynasty, which played a significant role on the development of the paper money in the Europe. Within comparatively modern time, credit system has succeeded the systems of ancient and mediaeval times. Most of countries had their own comprehensive public credit system before entering 21st century, even though wars happened frequently in the 20th century. Entering 21st contrary, credit system become much more popular, and people rely on these credit products like credit cards to a considerable extent. For example, In India, there are an increasing number of people opting for loans for houses and cars and there is also a large demand for credit cards, after the prosperity of the credit system in 21st century.

Now, as the fierce competition among banking sectors, it is of significance for some banking sectors to maintain the number of clients in their own bank without losing. In which case, having an accurate prediction of the loss of clients is necessary. Thus, more and more scientists and bankers gather some models to analyze data and try to reach the conclusion, so that they could take some reactions to solve the problem of loss clients. With advancing technology, these scientists and bankers are more likely to achieve this goals.

The purpose of this invention is to analyze the loss of clients in one bank and try to find the solution to this problem. Maintaining valid clients is greatly crucial for a bank, under the condition of fierce competition with other banks sectors. To take an important place in the banking industry, some bank sectors or such leading organizations have to take some reactions. In the process of experiment, we develop some classification models by using the decision trees to test the accuracy of our data and analyze the loss of clients. This system provides us with a straightforward way to understand which kind of clients are much more likely to be lost.

SUMMARY OF THE INVENTION

The purpose of our experiment is to forecast whether the credit card churning will happen for a individual customer or not, a prediction

which

is beneficial for the bank to retain their existing customers instead of pursuing the new customers.

Firstly, we attain the statistics of nearly 3500 customers from one of the most authoritative banks in China. In order to construct a effective and efficient model, we, at first, import a great number of modules from “sklearn” and utilize the “csv-reader” read the data from our file. After finishing these steps, we need to do the data preprocessing to ensure that our data set is balanced. To do the data preprocessing, we employ the Min-Max normalization methods, a key process of data preprocessing, to make our statistics in the same order of magnitude so that we are able to eliminate the effect caused by different dimensions for different indicators and improve our accuracy. Then, we partition our data as two different parts: a training data set and a testing data set.

In modeling stage, because of its advantage of easy comprehension and the discontinuity of our statistics, we select the Decision Tree(DT)as our model to forecast the results. After constructing the model of Decision Tree, we give the training data set to the model. During the training process, the model can achieve information gain to select some important and influential features. Then, it can utilize these significant features to construct the best Decision Tree. In order to evaluate the model constructed during the training process, we employ the testing data

set to exam the AUC value in the ROC curve, confusion matrix, precision rate, recall rate, and accuracy rate.

The final result of model is recorded in detail and showed intuitively in the Fig.3, Fig.4, and Fig.5. Through our Decision Tree model, the banks is able to effectively and sufficiently forecast and predict the behaviors of their customers, which is whether they will continue to use their credit cards in the future.

DESCRIPTION OF DRAWINGS

The following drawings are only for the purpose of description and demonstration of the results, where in:

Fig.1 is the flow diagram of the core processes of our experiment, which shows the specific step in our work step by step.

Fig.2 is the description of the principles of our method, the decision tree, of our experiment, which includes the root nodes, directed edge, internal nodes, and leaf nodes.

Fig.3 is a matrix evaluating of our modeling based on the method of confusion matrix, the leading diagonal of which shows the number of TP and the number of NF.

Fig.4 presents a great number of statistics which are able to well evaluate our model, including training time, AUC, prediction rate, recall rate, and accuracy rate.

Fig.5 is the ROC curve of our model.

Fig.6 is the complete visualized process of our Decision Tree model, which includes the root node, x78, other internal nodes, and the final leaf nodes.

DESCRIPTION OF THE INVENTION

All the details of the experiment will be elaborated below in order to make the experiment comprehensive.

Step A: Data collection

The data, originally from bank users, consists 135 independent variable and 1 dependent variable are collected from a bank. The examiner will process multiple information of users, whether it be income, age, times of overdue loans, and etc. The results will be presented as “exist” or “expired” accounts based on the data above. These data are cleaned in a broad picture, which indicated that they are well qualified for input.

Step B: Data Normalization

To generate comprehensive data, disequilibrium of data and different

scale of values are the main two problems before applying models. At first, it is common that when collecting data, variables are holding distinctive units, which means they represent different proportion. Thus, data with greater proportion can influence the overall results more than the ones are not, and because of this, relationships between independent variables might be ignored and some important indexes could be neglected. In order to avoid the problem, mathematical method of Normalization is employed to uniform data.

The process changing sizes of multiple types of data into a fixed range is called “Normalization”. Usually, Z-score Normalization and Min-Max Scaling are the two main ways for Normalization. When comparing the experiment results, Min-Max Scaling was found to be more effective, since the Z-score Normalization was unable to delete abnormal values in the dataset.

The principle of the Min-Max normalization is that it can rescale the ranges of different and various features to the scale in the range of [0,1]. The formula of the Min-Max Normalization is given as :

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

In this formula, the x’ meas the final rescaling value of the feature and the x means the original value of the feature. Min(x) means the minimum

value in this specific feature and Max(x) means the maximum value in it. Therefore, through this formula, we can rescale the whole statistics in different features to the range in [0,1] so that we can completely the disadvantageous impact resulted from the different orders of magnitudes among various features.

Step C: Model Construction

We develop classification models by using decision tree. Firstly, dividing these data into two groups— trained group and tested group, according to the proportion of 3:7. After that, we could construct figure shown in figure 4. A intact figure of decision tree mainly composes of one root node, several internal nodes, and several leaf nodes. To be more specific, root node represents the beginning of the decision tree, each internal node represents one of the classifications, and each leaf node represents represent one of the results of classification. The rules for selecting attributes in the decision tree are based on information gain. To calculate the information gain, it is necessary for us to understand and calculate the the information entropy firstly. According to the formula below, getting the results of information entropy is straightforward. Since information gain is the change in information entropy, we could get access to the result easily. And then, we need to rank these different classifications from the

$$I(a_1, a_2, ..., a_n) = \sum_{i=1}^{n} I(a_i) = \sum_{i=1}^{n} p(a_i) \log_2 \frac{1}{p(a_i)}$$

largest to the smallest according to their entropy of information. Finally, putting corresponding classification on one of the internal nodes of the decision tree. The same processes repeat to finish constructing the decision tree.

Step D: Results Visualization

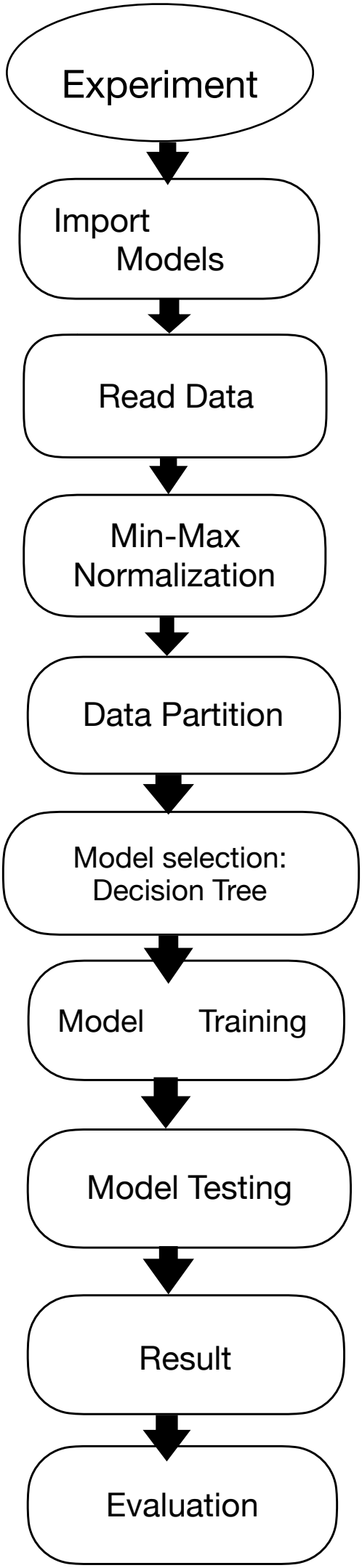
In order to demonstrate our experimental result visually, firstly, we utilized confusion matrix to evaluate the performance of our models during the test. According to the result in the Fig.3, the correct results, including TP and TN, are 825 and 83 but the incorrect results, including FN and FP, are 49 and 45. In order to present the evaluation of our model in details, we also employs the precision rate, recall rate and accuracy rate. According with the statistics result in the Fig.4, the precision rate of our model is 62.88%, the recall rate of our model is 64.84%, and the accuracy rate is 90.62%. However, because, in order to calculate these rates and confusion matrix, we need to set a threshold by ourselves which might has strongly disadvantageous impact and influence on our final results and the degree of accuracy of our results, we employ other evaluation indicators, ROC curve and AUC, to assess our models more precisely. The ROC curve and AUC is presented in the Fig.5, which can effectively and accurately evaluate our model and show how our model

performs.

CONCLUSION

Developing optimal classification models by utilizing decision trees is an effective way to analyze the data of the loss of the clients in a bank sector. The adoption of the decision trees can help to improve the accuracy of model prediction and find the solution of the problem. It is rational to state that these data that we get from the experiment is meaningful and can be seen as some references in real application. After analyzing these data, it is straightforward to find which kind of people are more likely to be lost. In which case, some banking sectors can take some reactions to avoid the loss of clients. However, for the future development of the data mining technologies in banking industry, it still exists space to improve the accuracy of the model prediction by using less information related to clients.

Fig.1


Experiment
Import Models
Read Data
Min-Max Normalization
Data Partition
Model selection: Decision Tree
Model Training
Model Testing
Result
Evaluation

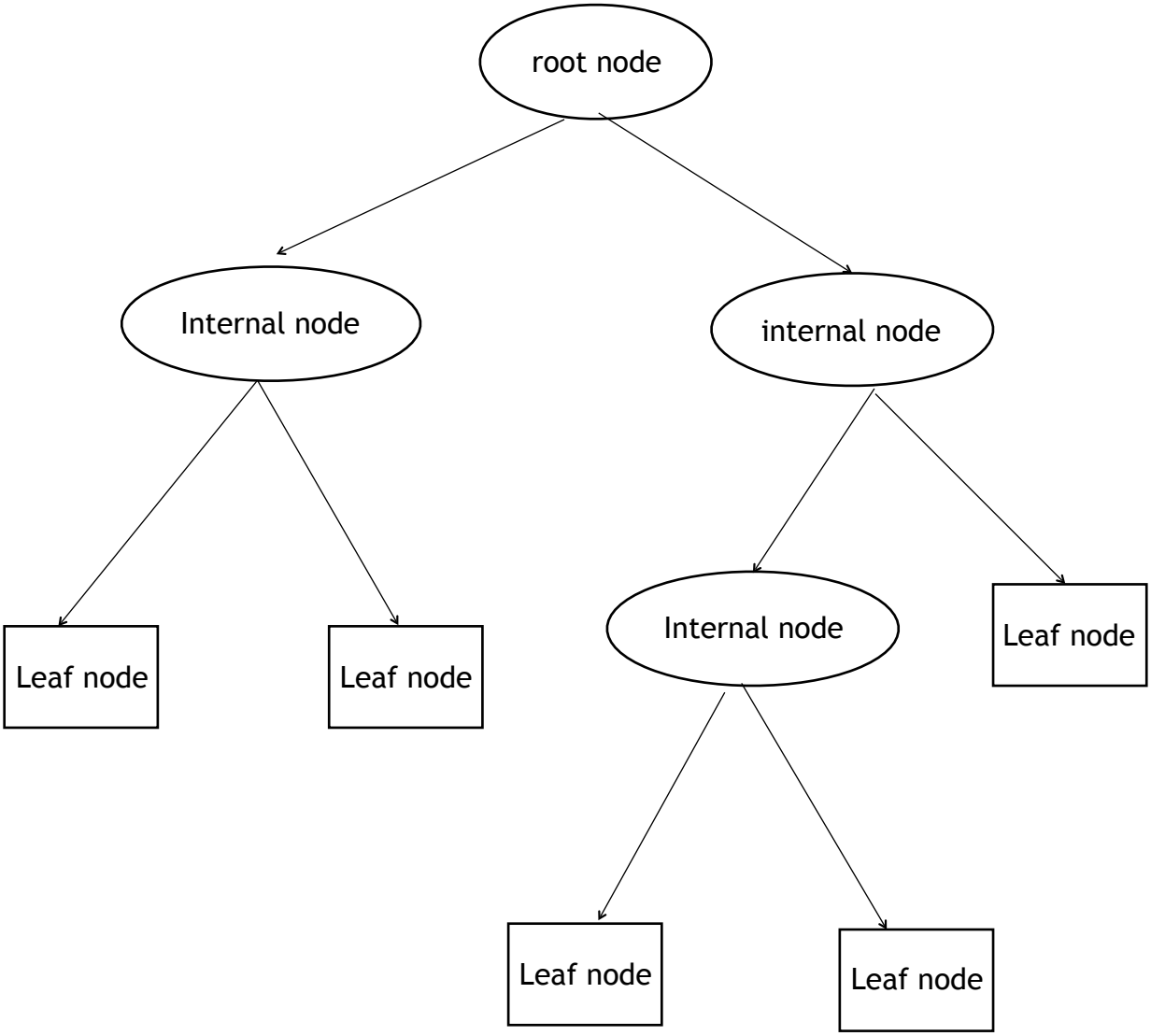

root node
Internal node
internal node
Leaf node
Leaf node
Internal node
Leaf node
Leaf node
Leaf node

Fig.2

$$\begin{bmatrix} 825 & 49 \\ 45 & 83 \end{bmatrix}$$

Fig.3

Training time:	0.358093s
AUC:	0.80
Precision rate:	62.88%
Recall rate:	64.84%
Accuracy rate:	90.62%

Fig.4

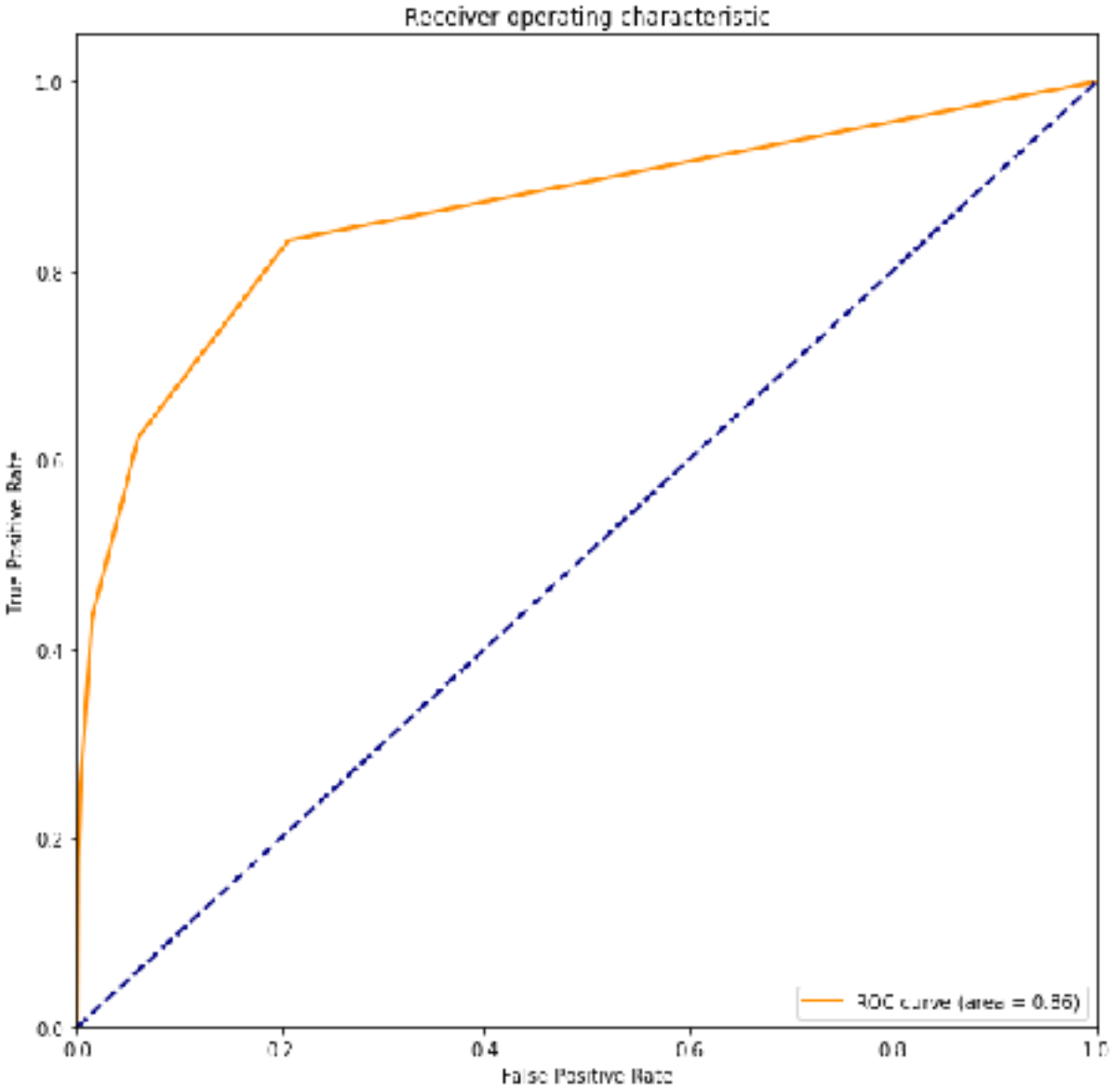

Receiver operating characteristic
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.86)

Fig.5

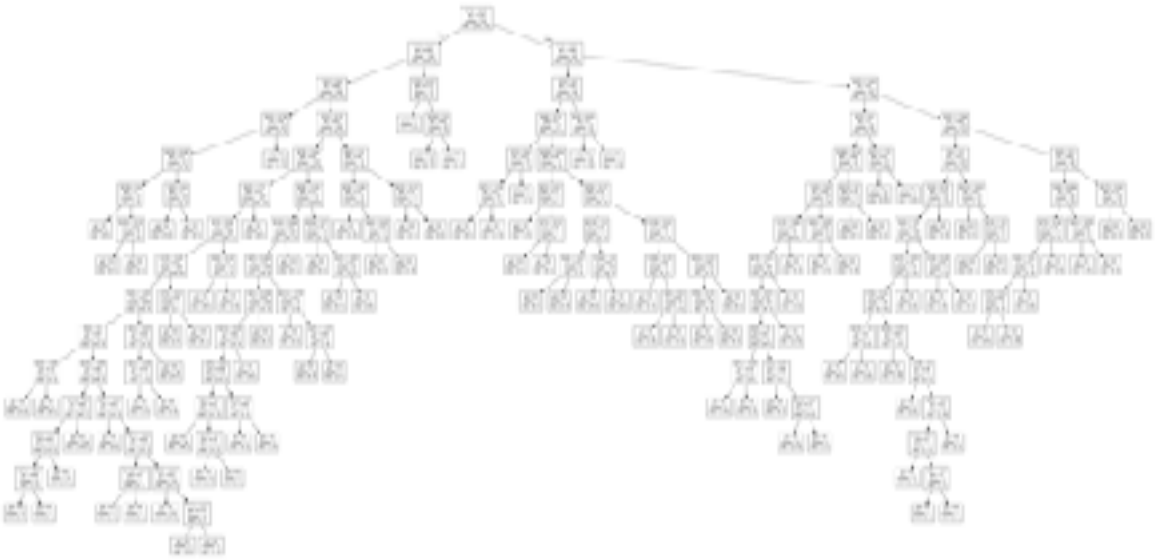


Fig.6