Proceedings of Business and Economic Studies

Research Article

BIO-BYWORD SCIENTIFIC PUBLISHING PTY LTD

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

Jianyao Liu

Greater Atlanta Christian School, Georgia State, United states

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 preprocess the data set, employing the training data set to construct our model, examining the model by testing data set, and analyzing the results.

Publication date: April, 2020

Publication online: 30 April, 2020

****Corresponding author:*** Jianyao Liu, guoliming@xyzrgroup.com

1 Introduction

The purpose of this research 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 other 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.

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 were 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.

2 Methods

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

Step A: Data collection

The data, originally come from bank users and consist 135 independent variables 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 indicates 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 people collect data, variables are holding distinctive units, which means they represent different proportions. 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 also be neglected. In order to avoid this problem, a mathematical method of Normalization is used for uniform data.

The process that 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 various features to the scale in the range of [0,1]. The formula of the Min-Max Normalization is given as :

In this formula, the x’ means 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 one of the results of classification. 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 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 are presented in the Fig.5, which can effectively and accurately evaluate our model and show how our model performs.

3 Conclusion

Developing these 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 acquire from this 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 a bank sector. 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.