Data Mining of Franchise Failure by Brand

Tingwei Lin
Department of Computer Science and Engineering
University of Bridgeport, Bridgeport, CT

Abstract

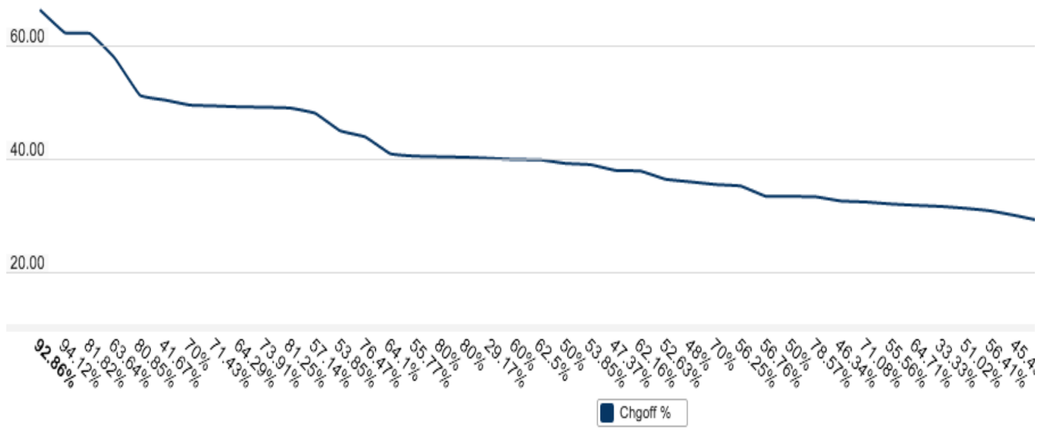
The SBA(Small Business Administration) offered a list of failure rates of small business loans sorted by franchise brands from 2001 to 2011. This is a collection of data of over 580 franchise brands who fail to pay back general SBA 7(a), real estate and equipment loans. Franchisee loan pay-back rates are tracked by brand for the ten-year period. Ideally, to be informed about where to put one's money, a franchise buyer should look at the profit of a brand's average store and compare the rate of return on the investment with all other brands' returns on investments. The purpose of this project is mining the relationship between the attributes in the dataset and offers a reference to the SBA about which kind of brand franchise worth a loan. And the failure percent of a brand can be accurately predicted with the program. Thus, the risk of loaning would be reduced and the percentage of success case could be guaranteed in a certain extent.

Problem Definition

Firstly, analyze the relationships between Brand name, failure percent, charge off percent, disbursements# and Disbursement $. Failure percent, which stands for the failure percentage to pay back the loan and indicates the success percent, can works as the target object. Chgoff percent means the charge off percent of bad loans incurred by the lenders after the small business borrowers' collateral assets have been collected. disbursements# means the number of case that the brands got loan.
Secondly, the problem could be concluded: (1) with the rising of Chgoff percent, the failure percentage also rise sharply; (2) low disbursements $ is always accompanied with high failure percentage;
Last, the main work of mining is determining the relationship and make a precise prediction on the target object taking advantages of the k-means and density-based clustering algorithm.

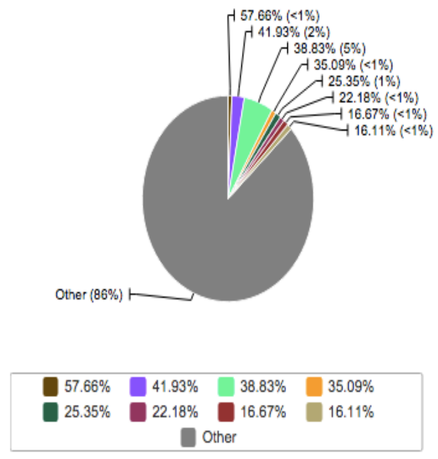
Design and Implementation

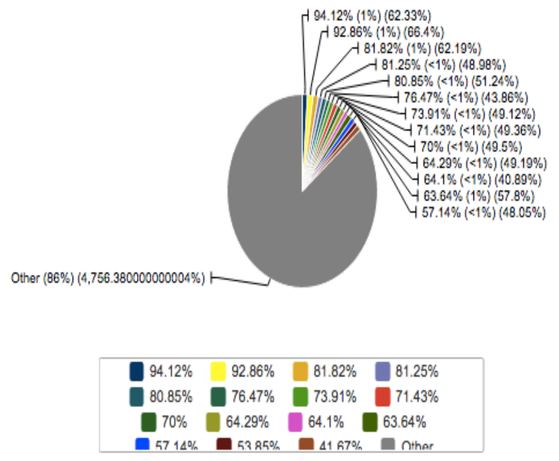
First, I smoothed the dataset and excluded some outliers. Then, I applied the K-means algorithm on the disbursement # and chgoff %, the cluster numbers vary from 3 to 6, then I found that the definition1 is correct and the chgoff% has the strongest connection with the target object.

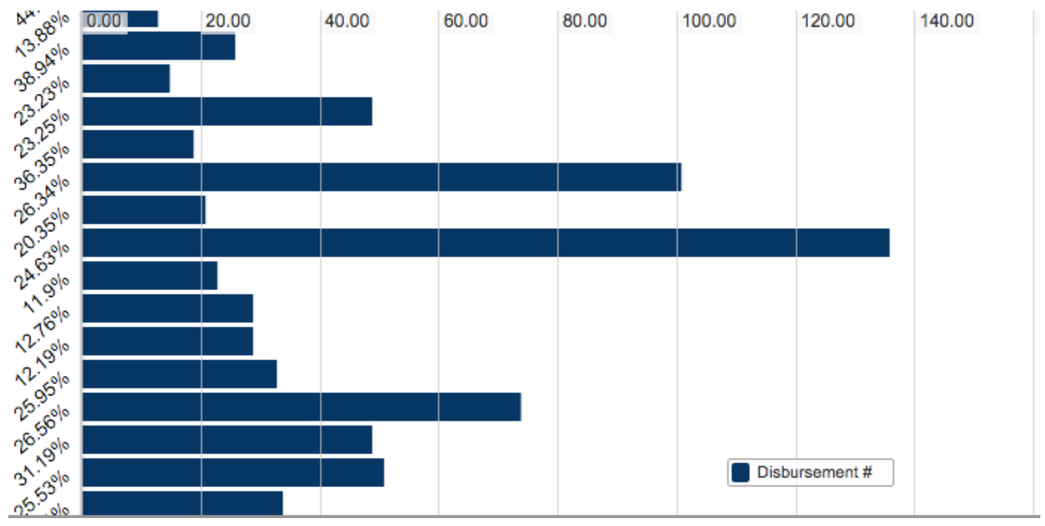


Result

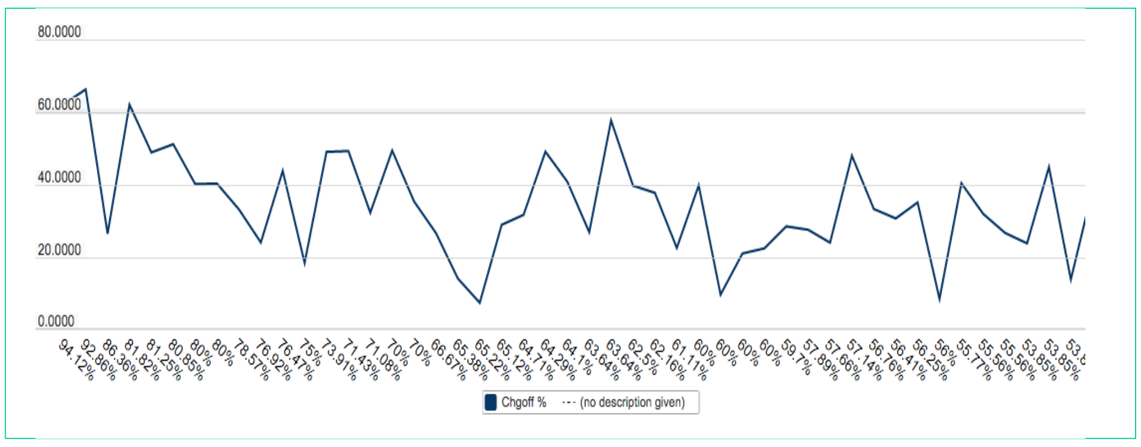
As shown in the graph, with the descent of the chgoff %, the failure percent decrease stably. It's obvious that according to the brands' loan charge condition, SBA could predict the failure percent. However the result of disbursement mining is undesirable and this ought to be the most essential part in prediction.







Relationship between Disbursement # and Failure Percent



Relationship between Charge off percent and Failure Percent

Performance Evaluation

The most important thing of K-means is to determine the number of clusters and this took a lot of time to test. In order to measure the clustering result precisely, I choose the K equals 5.Through contrast to the density-based clustering algorithm, K-means simplify the analysis and shows the result more visualized and straight. Maybe combining the Apriori algorithm, which focus on the frequent pattern, and K-means could lead to a better result. And because of the processing of data in advance, the result has obviously been better than before.

Conclusion and Future Work

The output result and evaluation are the most important. In pursuit of best consequence, the evaluation methods are still expected to be better and make the result more accurate. It is to design a program for mining data in such a economic field. But the purpose at the beginning is to find the significance of the disbursement#, which is the most suitable object for prediction. Thus, the mining method maybe can be improved in details. And if condition allows, 3-D print will be added to the project to give the result a strong sense of visualization.