Credit Card Churning Classification using Decision Tree, Random Forest, and Logistic Regression

Xinyan Chen[ϒ], Steven Smith[ϕ], Susanna Wang[ϰ]

Zicklin School of Business, Baruch College, 55 Lexington Ave, New York, NY 10010

April 29, 2021

Abstract - Being able to predict potential credit card churners is helpful for banks to prevent potential churners from leaving their service. Banks will be able to proactively approach identified potential churners to provide them better services in the hopes of persuading them to stay. Managers at banks hope that the help of prediction models will provide early detection of churners. This paper applied three machine learning algorithms, Decision Tree, Random Forest, and Logistic Regression, in the credit card customers dataset to classify customers, and comparison results are presented.

I. Introduction

Many central bank managers are growing increasingly troubled by seeing more customers leaving their credit card services. So managers at these central banks want to know what are some of the reasons that cause these customers to go to their credit card services and potentially create a plan for the customers that are more likely to leave. So with the data we are being provided, we can create models to identify what kind of customers are more likely to go for another credit card service.

We performed classification modeling that takes in 19 predictor variables that describe the basic information of the customers' credit card services as the predictor variable. Our response variable will be the attrition flag column of the dataset with the value of "Existing Customers" and "Attrited Customers." We used Decision Tree, Random Forest, and Logistic Regression to assist us through this project better. This project aims to improve the model's accuracy with the limited data that we have for churning customers to provide the best result for central bank managers.

The rest of the paper is organized as follows: Related Work in Section II. Next, dataset description and methodology are described in Section III. Then, the Machine Learning algorithms used are explained in Section IV. Finally, the experiment results are discussed in Section V and our conclusion in Section VI.

II. Related Work

Analyzing potential churners using prediction models is common practice for many firms determining the best way to optimize revenue. Some models rely heavily on behavioral attributes instead of demographics by focusing on the

Spatio-temporal patterns that influence a client's financial decision [3]. Research students from China performed a similar study using Logistic Regression on a dataset ranging over 5000 clients. The data mart consisted primarily of "existing customers" with very few churners[4]. Another group performed the analysis using Decision Tree and concluded that demographics have very little with churn rate[5].

III. Methodology

A. Dataset

We use the Credit Card Customers dataset released on Kaggle by Sakshi Goyal, containing 10,127 customers' data. Only 16.07% of customers from the entire dataset have churned, and the remaining 83.93% of customers are still using their current credit card service. Each of the 10,127 customer records will contain 20 valuable attributes: client number, customer age, dependent count, education level, marital status, income category, card category, months on a book, total relationship count, months inactive, contacts count, credit limit, total revolving balance, average open to buy credit line, change in transaction amount, total transaction amount, total transaction count, change in transaction count, average card utilization ratio, and attrition flag which will be our response variable.

B. Data pre-process

The existing customer to attrited customer ratio is 5:1. Initial data analysis revealed that the total transaction amount in the last 12 months was the most important feature out of the rest. For the attrition flag column, we renamed the values "Existing Customer" and "Attrited Customer" as "Stay" and "Quit," respectively, for a more accessible analysis of model results. Features are either in a factor or numerical form, and there is no missing value.

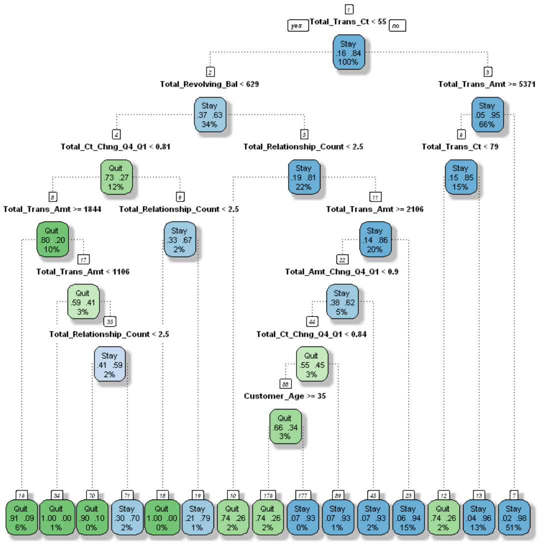
C. Classifier design

Three types of supervised classifiers were chosen: Decision Tree, Random Forest, and Logistic Regression. Baseline models of all three classifiers were trained and evaluated on the same dataset, and their accuracies were compared.

IV. Experiments

A. Decision Tree

The first model we utilized was Decision Tree; we used 19 predictors for our customers. We can see that customers who did more than 55 transactions and have a transaction amount of over $5,471 have a 52% chance of staying is also currently the highest percentage. Whereas, if a customer has less than 55 transactions, they're likely to be churners. The testing model accuracy is 92.89%. Five essential features: total revolving balance, total transaction amount, change in transaction count, total transaction count, and average card utilization ratio were selected based on variables importance. A separate model was trained with these selected features, and the test accuracy is 92.29.



B. Random Forest

The Random Forest model was performed using a cross-validation method with 10 folds and the number of variables available for splitting at each tree node of 4. The training model with all 19 predictors has an accuracy of 94.9%, and the testing model with 94.38% accuracy. Five important features: total transaction amount, total transaction count, change in transaction count, total revolving balance on the credit card, and average card utilization ratio were chosen based on variable importance. A separate model was trained with selected features with the same number of folds and the same number of variables splitting at each node. The training model has an accuracy of 93.51%, and the testing model has a 93.38% accuracy.

C. Logistic Regression

Logistic Regression models are dependent on binary variables that imply an event would take place or not. Using the given dataset, we decided "Existing" and "Attrited" would be leading candidates to determine a client would "Quit/Stay." The test accuracy of the model is 90.37%. Five features: the total transaction count, total transaction amount, total relationship count, change in transaction amount, and contacts count, were selected based on variable importance. A separate model was trained with the selected features using a cross-validation method of 10 folds. The accuracy for that model is 87.91%.$^{\phi}$

V. Results

We performed three types of supervised classifiers and the results that we have gotten is Random Forest has a higher percentage of accuracy with a 94.38%. With Decision Tree following behind with 92.89%.

Table 1: Base model accuracy

Model	Test Accuracy
Random Forest	**94.38%**
Decision Tree	92.89%
Logistic Regression	90.37%

Next, important feature selection was performed based on the importance of the variables of each model. Finally, four different models were trained with the selected features to improve the test accuracy. The results of the models are in table 2, with random forest having the highest test accuracy of 93.38%.

Table 2: Models performance on selected features

Model(tuned)	Test Accuracy
Random Forest	**93.38%**
Decision Tree	92.29%
Logistic Regression	87.91%

VI. Conclusion

Using the correct type of model to classify these dates is essential. The model with higher accuracy can help central banks predict the possible customers who would be more likely to leave their credit card services. With the data we have in hand, it is concluded that Random Forest performed better out of the three models. We can then better assist these central bank managers with the data we have found, and they can make a plan on how to keep the possible customers from leaving their services.

REFERENCES

[1] Goyal, Sakshi. “Credit Card Customers.” Accessed April 29, 2021. https://kaggle.com/sakshigoyal7/credit-card-customers.

[2] Analytics Vidhya. “Classification Models in Machine Learning | Classification Models,” November 30, 2020. https://www.analyticsvidhya.com/blog/2020/11/popular-classification-models-for-machine-learning/.

[3] Kaya, Erdem. “Behavioral attributes and financial churn prediction” 2018. https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-018-0165-5

[4] Nie, Guangli. “Finding the Hidden Pattern of Credit Card Holder’s Churn: A Case of China” 2009. (Attached PDF)

[5]Chiang, Ding-An. “Goal-oriented sequential pattern for network banking churn analysis” 2003. https://www.journals.elsevier.com/european-journal-of-operational-research