Which customers are most likely to churn? A predictive banking analysis

Ishika Gupta

May 7, 2022

Abstract

Customer churn analysis is important for identifying old customers without loss and developing new products and making new strategic decisions for retaining customers. Identifying customer churn in banks will help the management to classify customers who are likely to churn early and target customers using promotions, as well as provide insight into which factors should be considered when retaining customers. This paper performs bank customer churn analysis with a predictive random forest model that predicts which customers are likely to churn. We found that the variables: total transaction amount in the last 12 months, total transaction count in the last 12 months, change in transaction amount from Q4 to Q1 and total revolving balance are the most important variables in predicting whether a customer will churn or not.

1. Introduction

Like any other industry, customer acquisition and retention is the primary way to stay competitive in the banking industry. With growing competition in the banking industry, customer churn is a serious issue and banks are required to retain their existing customers along with attracting new customers. Growth and profit for a firm come from customer retention. Churning refers to a customer who leaves one banking company to go to another banking company. The churn rate, also known as the rate of attrition or customer churn, is the rate at which customers opt out of business with a banking company which can be expressed as follows:

$$Churn\ Rate\ = \frac{Number\ of\ Lost\ or\ Canceled\ Customers}{Ending\ Total\ Customer\ Count}$$

Customer churn analysis is important as acquiring new customers is usually arduous and costlier than retaining old customers. According to most research, the cost of acquiring new

consumers is 1/5 the cost of keeping existing ones (Guliyev and Tatoğlu, 2021). If a banking company is able to retain their customers, their average customer lifetime value increases, making every future sale even more valuable and ultimately improving your unit margins. To solve this issue for banks and help them retain customers, we will build a predictive model to classify bank customers who are at the highest risk of churn. With this predictive model, banks can offer different types of incentives to those customers who are more likely to switch to different banks.

2. Literature Review

Growth in electronic services provided by banks increases the need for retaining customers. CRISP is a methodology used for predicting customer churn in electronic banking services. The study by Keramati et al., (2016) aims to identify the features of churners from electronic banking services. They extracted demographic variables, transaction data, the length of the customer association, and customer complaints from the bank's database to analyze the most important variables and features that impact a customer's decision to churn. They found that with a better understanding of the features of churners, banks can consider some strategies to prevent churn (Keramati et al., 2016).

Churn analysis is useful for the management of a bank to predict customers who are expected to churn early as well as impose techniques for customer retention. Guliyev and Tatoğlu (2021) focus on machine learning tools for churn analysis by using SHapely Additive exPlanations (SHAP) values to support the machine learning model evaluation. Their data was split into two parts in a 80:20 ratio to train and test their models. To deal with imbalance in their data, they used tree-based models such as Decision Tree, Random Forest, and XgBoost. Their results suggested that XgBoost outperformed all the other models (Guliyev and Tatoğlu, 2021).

Based on statistical & machine learning models, the study by Verma (2020) explores churn prediction for savings accounts for a customer. He uses Logistic Regression, C 5.0, CHAID, ANN, XG-BOOST, and Decision Tree techniques models for comparison and the components used for comparison of models are: Area under the curve (AUC), Model Accuracy, Gini coefficient, and Receiver Operating Characteristics (ROC) . The results show that Random Forest predicted the model with the highest accuracy. Customer vintage, customer's age, average balance, occupation code, population type, average debit amount, and an average number of transactions are found to be the variables with high predictive power for the churn prediction model. By giving better customer satisfaction and experience, the commercial banks can limit the customer churn and maintain their deposits (Verma, 2020).

To increase profits for continuing operations and enhance the core competitiveness, commercial banks must avoid the loss of customers while acquiring new customers. He et al., (2014) discuss commercial bank customer churn prediction based on SVM model, and use random sampling method to improve SVM model, considering the imbalance characteristics of customer data sets. Their results show that this method can effectively enhance the prediction accuracy of the selected model (He et al., 2014).

3. Data

3.1 Pre-processing

The dataset was chosen from Kaggle to predict the customer churn status as it has the most number of columns among the publicly available bank customer churn datasets. It has 10127 observations with 21 variables, and it includes various customer demographic data such as age, gender, geography, education level, and marital status. This dataset also provides various bank information and transaction data such as credit limit, balance, change in transaction amount, total transaction amount, total transaction count, and average card

utilization ratio. The outcome variable is “attrition_flag” which is the binary variable that reflects whether the customer left the bank.

Table 1: It shows all the available variables along with their interpretation and possible values.

Variable	**Description**
Attrition Flag	Binary variable. If the customer is churned then 1 else 0.
Customer Age	Customer’s age in years (26-73)
Gender	M=Male, F=Female
Dependent Count	Number of dependents (0-5)
Education Level	Education qualification of the account holder - Graduate - College - High School - Unknown - Uneducated
Marital Status	- Married - Single - Divorced - Unknown
Income Category	Annual income category of the account holder - < $40K - $40K - 60K - $60K - $80K - $80K-$120K - $120K <
Months on Book	Period of association with bank in months
Months Inactive 12 mon	Number of months inactive in the last 12 months
Total Relationship Count	Total number of products held by the customer (1-6)
Contacts Count 12 mon	Number of contacts in the last 12 months
Credit Limit	Credit limit on the credit card
Total Revolving Bal	Total Revolving Balance
Avg Open to Buy	Average open to buy credit limit for the past 12 months

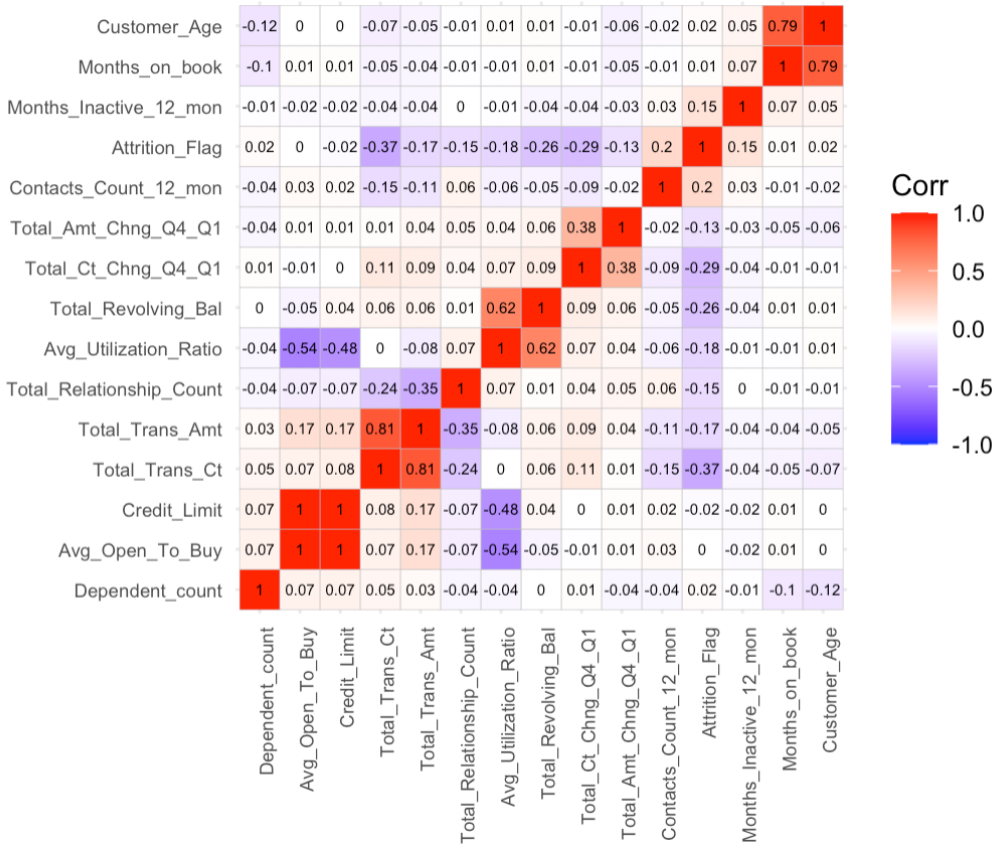
Total Amount Chng Q4-Q1	Change in transaction amount from Q4 to Q1
Total Trans Amt	Total transaction amount in the last 12 months
Total Trans Ct	Total transaction count in the last 12 months
Total Ct Chng Q4-Q1	Change in transaction count from Q4 to Q1
Avg Utilization Ratio	Average Card Utilization Ratio. Range: 0-1

It is worth mentioning that some variables such as client number and card category have been dropped from the dataset as they will not contribute to building a predictive model. The variable card category consists of “blue”, “silver”, “gold”, and “platinum” cards and using this variable will make the predictive model not generalizable to other bank customer data. On the other hand, other variables like credit limit or total transaction amount are data that can be found in customer databases of every other bank.

3.2 Exploratory Data Analysis

To study more about the relationship between these variables, this correlation matrix was created. From this correlation matrix, it can be concluded that variables such as total transaction count, total revolving balance, and change in transaction count have the highest negative correlation with the outcome variable “attrition_flag”, whereas variables like number of months inactive in the last 12 months, and number of contacts count in the last 12 months have the highest positive correlation with the dependent variable.

Fig 1: Confusion Matrix



4. Methodology

4.1 Machine Learning: Decision Trees

Decision tree learning is one of the predictive modeling techniques used in machine learning. The technique uses a decision tree as a predictive model to go from observations about an item represented in the branches, to conclusions about the item's target value represented in the leaves. Decision trees have the advantages of simple use, easy understanding, high accuracy, and high prediction ability. Tree models where the target variable can take a discrete set of values are called classification trees, where leaves represent class labels and branches represent conjunction of features that lead to those class labels. Decision trees where the target variable can take continuous values are called regression trees. Decision

trees are among the most popular machine learning algorithms given their intelligibility and simplicity.

4.2 Machine Learning: Random Forest

A group of trees underpins the Random Forest. It is complemented with an aggregate of the prediction's mean value, which is produced at the conclusion of each of the trees, reducing the lack of robustness of a single tree. One of the advantages of using a random forest model is that cross-validation or a different test set are not needed to estimate unbiased test set error, because each tree in a random forest model is built using a different bootstrap sample from the original data.

4.3 Modeling Process

Decision trees suffer from high variance. Bagging or bootstrap aggregation, is a general-purpose procedure for reducing the variance of a statistical learning method. It is the process of averaging a set of observations to reduce variance. In our modeling process, we did not use the split sampling method to assess the accuracy of our random forest model because the internal validation is implemented as the model is constructed. ⅔ of the training data is used to perform each tree and the remaining ⅓ of the training data is used to calculate out-of-bag error to evaluate the model. Therefore, out-of-bag error is a number of correctly predicted rows from the out-of-bag sample, and it can be interpreted as a validation score.

5. Results and Discussion

We randomly split the dataset to train (70%) and test (30%) datasets to measure the accuracy of the decision tree model. Figure 2 shows that the decision tree model has the accuracy of 91.4%. The depth of the tree was chosen based on the accuracy rate of models on the test dataset. Based on our exploratory data analysis and correlation matrix, we used these variables to run both the decision tree and random forest classifier model: “Total_Trans_Amt”,

“Total_Trans_Ct", “Total_Amt_Chng_Q4_Q1”, “Total_Revolving_Bal”, “Contacts_Count_12_mon”, “Months_Inactive_12_mon”, “Gender”, “Education”, “Marital_Status”, “Customer_Age”, and “Avg_Utilization_Ratio”.

Fig 2: Decision tree

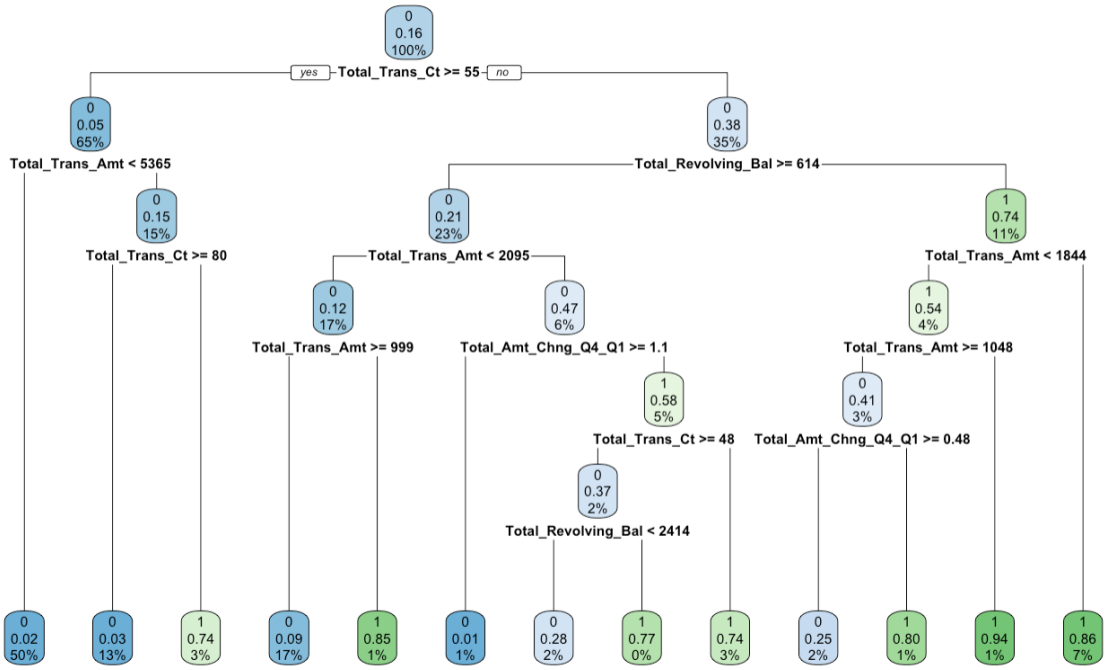


Table 2 shows the result of the random forest model. This result shows that the overall accuracy of the model is 95.9%, but the out-of-bag error is equal to 4.08%. Moreover, the false negative error of 18.3% is much higher than the false positive error of 1.3%. The false negative error implies wrongly identifying bank customers are not churned, and the false positive error implies wrongly identifying bank customers are churned.

Table 2: Confusion matrix and out-of-bag error of the model using imbalanced dataset

Table 2	Not churned (Predicted)	Churned (Predicted)	Class Error	Out-of-Bag Error
Not churned (Actual)	8380	120	1.3%	4.08%
Churned (Actual)	293	1334	18.3%	

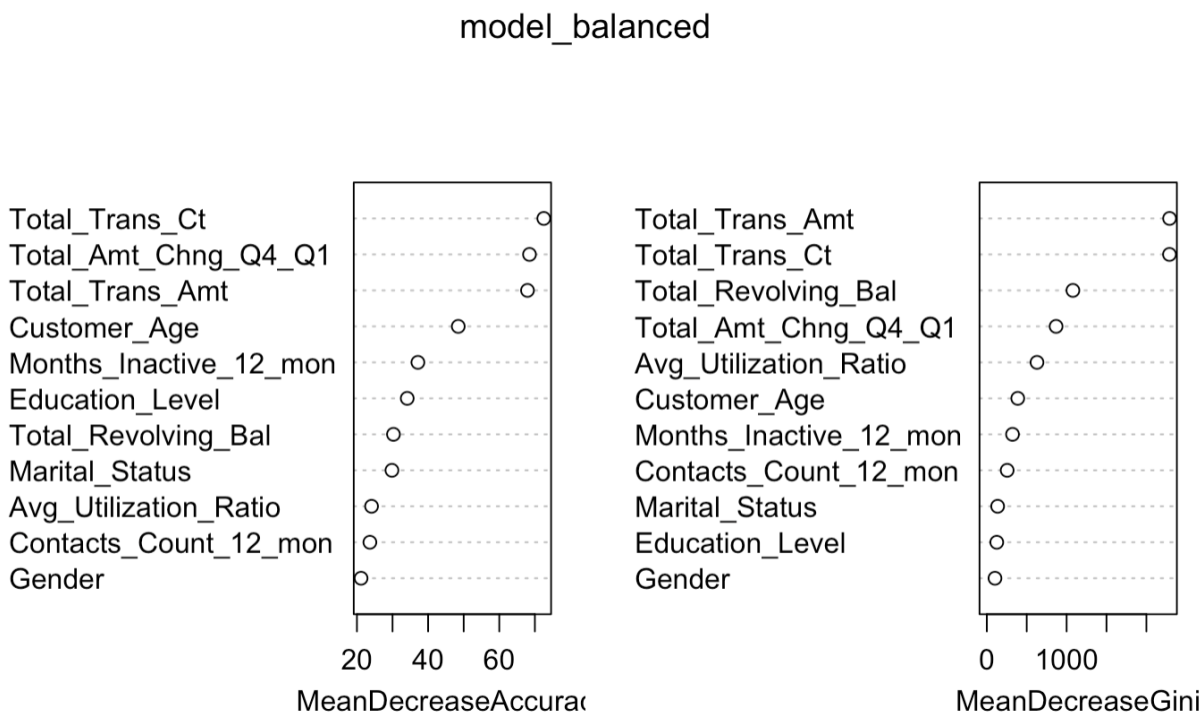
Of the 10127 bank customer records used in this predictive modeling, only 16.06% of customers are churned. Thus, an oversampling technique was implemented to mitigate this imbalance. Table 3 shows the result of the model after the oversampling technique is implemented.

Table 3: Confusion matrix and out-of-bag error of the model using balanced dataset

Table 3	Not churned (Predicted)	Churned (Predicted)	Class Error	Out-of-Bag Error
Not churned (Actual)	8314	192	2.3%	1.16%
Churned (Actual)	6	8494	0.1%	

Using an oversampling technique decreases the out-of-bag error to 1.16%, compared to a random forest model, where the out-of-bag error was found to be 4.08%. The oversampling technique uses a balanced dataset, whereas the random forest model uses an imbalanced dataset. Although the false positive error increased slightly to 2.3%, the false negative error decreased significantly to 0.1%.

Fig 3: Variable Importance Plots



The gini index measures the degree or probability of a particular variable being wrongly classified when it is randomly chosen. Therefore, the higher the mean decrease gini index, the higher the importance of the variable in our model. This figure shows that variables such as “Total_Trans_Amt”, “Total_Trans_Ct”, and “Total_Revolving_Bal” are important variables to predict whether a customer will churn or not. On the other hand, the mean decrease accuracy tells us how much accuracy the model loses by excluding each variable. Therefore, variables such as “Total_Trans_Ct”, “Total_Amt_Chng_Q4_Q1”, and “Total_Trans_Amt” are important variables to successfully classify the bank customers. A decrease in the total transaction amount in the last 12 months, total transaction count in the last 12 months, change in transaction amount from Q4 to Q1 or total revolving balance implies that the customers are more likely to churn.

6. Conclusion

In this paper, we proposed Machine Learning methods for predicting customer churn. Firstly, to test and train the model, the sample dataset is divided into 70% for training and 30% for testing. In addition, we contended with another problem: the data was not balanced and only about 16.06% of the dataset had included churn customers. To solve this problem, an oversampling technique was implemented to mitigate this imbalance. We applied the tree-based models such as Decision Tree and Random Forest. The Random Forest outperformed for each metric, with an accuracy of 98.8 percent for the balanced dataset as compared to the Decision Tree accuracy of 95.9%.

In the future, it might be important to gather data in the form of a survey from churned users to understand why customers decide to switch to other banks. This will allow banks to understand what services they lack and what they can improve on. In this way, banks do not have to spend money on offering incentives to these customers, rather they can improve services they lack.

Many of the predictive variables show that transaction and bank data is helpful to know which customer is likely to churn, but these customers withdraw all their money after making the decision to switch to another bank. So, the question then arises how do we predict the likelihood of churning before customers make their decision to churn and become inactive members of the bank. With lack of specificity about the source of the dataset, it is difficult to determine if this predictive model can be generalized to other banks in the United States, or if it is only usable for this specific bank where the data was collected.

References

Guliyev, H., & Tatoğlu, F. Y. (2021). Customer churn analysis in banking sector: Evidence from explainable machine learning models. *JOURNAL OF APPLIED MICROECONOMETRICS*, *1*(2), 85-99.

He, B., Shi, Y., Wan, Q., & Zhao, X. (2014). Prediction of customer attrition of commercial banks based on SVM model. *Procedia computer science*, *31*, 423-430.

Keramati, A., Ghaneei, H., & Mirmohammadi, S. M. (2016). Developing a prediction model for customer churn from electronic banking services using data mining. *Financial Innovation*, *2*(1), 1-13.

Verma, P. (2020). Churn Prediction for Savings Bank Customers: A Machine Learning Approach.