Chaitra Setty

Jacob Rhymes

Samuel Oh

Soumik Choudhuri

**1. Description**

The dataset our group decided to investigate regards credit card customer churn. The data includes details about the customer as well as their credit card activity. Based on the features provided in the dataset, we endeavor to understand the underlying factors surrounding credit card customer attrition (churn). If enabled to predict when a customer may leave, a bank could take action that may convince these churning customers to stay.

**2. Importance of the problem:**

We chose to analyze this question because of the large economic impact of churn on banks that offer credit cards. BCG research shows that American corporate banks annually lose 10% to 15% of gross revenue due to attrition alone.**1** Thus, reducing this attrition rate, even marginally, would greatly increase annual performance for many corporate banks.

By understanding the key indicators of attrition, the bank can implement targeted improvements in their business processes and therefore increase retention rate. Banks with this knowledge can make more informed business decisions by determining the focus areas to increase their investments, such as an increase in digital marketing, an improvement in customer service, or an expansion in incentives for the at-risk customers.

**3. Exploratory analysis**

We began by conducting preliminary exploratory analysis to better understand our dataset. First, we noticed that our dataset was imbalanced; only about 16% of our observations were attrited customers. We decided to use the SMOTE (synthetic minority oversampling technique) method to balance our data. After utilizing SMOTE, our dataset has an equivalent sample of existing customers and attrited customers. This enables meaningful analysis and more accurate modeling.

In order to further understand our dataset, we spent some time analyzing each feature through histograms, bar graphs, pie charts, and correlation matrices to identify patterns. We noticed that most of our customers are adults in their mid-40 with 2 to 3 dependents. The majority of the customers in our dataset have undergraduate degrees and almost 60% of the customers have an income of less than $60K.

There are 4 different types of cards: blue, silver, gold, and platinum. The most common credit card is blue, but the card category with the highest relative attrition rate was platinum. However, this may not be overly meaningful since we only have 11 customers in our dataset that own the platinum card. We need more data to validate any durable assumptions regarding this association.

The customers are also split between 2 main credit limits of $1,483 and $34,516. The majority of the customers with the lower credit limits are those with an income lower than $40k and most of the customers with the higher credit limit have an income greater than $120k. This seems logical. Attrition rate seems relatively stable across income categories.

We also found that the customers who churned had a lower average transaction amount. This means they were not using their cards as much as those who stayed.

Almost a quarter of the customers were on the books for 36 months. We assume that this is the default card expiration time period and these customers decided not to renew their card.

These were some of the significant insights we were able to derive from our preliminary analysis. Overall, the only large differentiator we noticed between the existing and attrited customers was the average transaction amount and transaction count.

**4. Solution and insights**

Our goal in this project was to predict customer attrition with high accuracy and recall so that the bank can efficiently target customers. This led us to a classification problem. So the first approach was to develop a logistic regression model setting 1 as the attrited customers and 0 as the existing customers for our target.

We obtained the correlation matrix for the numeric variables and found that 2 variables - “Credit\_limit” and “Avg\_open\_to\_buy” have a perfect correlation. So we kept “Credit\_limit” and dropped the other. We also initially converted the character variables into indicator variables to be used in the model.

Upon running the logistic regression model, we found that some variables like “Contact\_count\_12\_mon”, “months\_inactive\_12\_mon”, “Dependent\_count” have a positive coefficient when predicting attrition. This validated our intuition that more customer complaints (contacts) and higher inactivity lead to higher attrition rates.

On the other hand, variables like “Total relationship count” which indicates number of products a customer possesses or “Total\_Trans\_Ct” which indicates the total transactions related to a customer have a negative coefficient. This validated our intuition that higher customer usage leads to higher chances of retention.

We obtained an accuracy of around 83% and recall of around 77% for the test dataset in the logistic regression model. With the help of p-values, we were able to drop a few variables which had p-values > 0.05.

We then tried out k-nearest neighbors on our dataset along with cross validation. The graph of the number of nearest neighbors versus cross validation accuracy shows that we get the highest accuracy of classification at k = 10 and it decreases with further increase of neighbors as it starts capturing the noise. So, our classifier took the optimal value of 10 nearest training points and predicted with the highest accuracy. With the KNN model, our accuracy improved from 83% to 89.5%.

Next, we proceeded to non-linear models like decision trees.

In the decision tree algorithm, we used minimum entropy as the splitting factor. Upon performing 5-fold cross validation and varying the max depth of the tree from 1 to 5, we found that the average prediction accuracy in the training dataset varies from 78% to 92% for depth 1 to 5. So we chose to proceed with depth = 5 because that gave a significant improvement in accuracy of the test dataset as well.

A decision tree is easy to implement and interpret, but a single tree might not be efficient enough to provide good results because it might capture noise. With this firmly in mind, we applied the random forest technique which, as the name suggests, is a “Forest” of randomly created decision trees. It combines the output of individual trees to generate the final output. Through this process, the random forest approach filters out noise. In our random forest model we used n estimators=100 (i.e. the random forest would use the results of 100 decision trees and predict the class with majority votes). From this, our accuracy improved to 95% for training and 92% for test data while recall improved to 78%. This was a significant improvement from the logistic model with accuracy of 83%. Of our series of modeling techniques, random forest proves the most successful.

Based on the random forest, the most important variables are transaction amount, transaction counts, revolving balance in card, and utilization ratio.

Following this, we performed secondary exploratory analysis in an attempt to confirm or deny the variable importances generated by our models. At this point, our foundational supposition was that variables approximating usage, such as total transaction count, total revolving balance, utilization ratio and total transaction amount are the most important indicators of churn. We grouped customers by marital status, income category, gender, age, dependants and months on book in an attempt to find a deviation from this supposition. Each group confirms our conclusion that utilization is strongly associated with attrition.

Following our modeling endeavors and secondary exploratory analysis to confirm the conclusions of our models, we found that the customers who are likely to stay are those with higher transaction counts, higher balances in their cards and higher overall utilisation rates.

To apply this knowledge practically, we suggest that banks keep detailed data of and perform active analysis on customer usage patterns. For example, if there is a change in usage pattern of a customer resembling decreasing transaction counts, that might be an indication of a potential churn. Additionally, the bank should encourage routine usage of credit cards with reward incentives for such users like cash-backs or discount coupons upon reaching a certain level of transaction amount or count. These would help the bank retain its customers through predictive and efficient targeted marketing.

References

**1** Karthikeyan, S., Goyal, D., Khodabandeh, S., Dye, T. and Chhajer, S., 2021. *How Banks Can Close the Back Door on Attrition*. [online] BCG Global. Available at: <https://www.bcg.com/publications/2017/financial-institutions-marketing-sales-how-banks-close-back-door-attrition> [Accessed 5 July 2017].