Credit Card Customers Churn Prediction

Written by:

Binqian Chai

He Jiang

Ella Lin

Xuchen Liu

1. Introduction

The exploration of credit card customer churn in banking resonates deeply with our group for its direct relevance to our experiences as university students, a demographic often at the cusp of making significant financial decisions. Our interest in this topic stems from a desire to understand the underlying factors that influence customers to switch their banking services. A primary driver of our curiosity is the observation of our peers and ourselves being swayed by various bank offerings and benefits. This scenario presents a quintessential example of customer behavior in the banking sector, where decisions are frequently influenced by the perceived value and benefits offered by financial institutions. This commonality among students and other individuals choosing one bank over another due to better benefits underlines the importance of understanding customer preferences and expectations.

We aim to analyze and predict credit card customer churn using a dataset of 10,000 customers, which includes variables such as age, salary, marital status, credit card limit, and credit card category. This comprehensive dataset offers a window into the customer-bank relationship and allows us to explore the nuances of customer behavior in the banking sector. By analyzing these elements through a comprehensive dataset, we hope to uncover the key reasons behind customer churn and gain insights into what makes a bank's offering more attractive to certain demographics.

2. Literature Review

The topic of customer churn in the banking sector, especially concerning credit card users, has been the subject of considerable research. This literature review briefly summarizes key findings from recent studies, focusing on methods used to predict churn and the factors influencing it. A major area of research is the use of data analytics in predicting churn. Zhang

and Yang (2021) explored machine learning techniques like logistic regression and decision trees, highlighting their effectiveness in understanding customer behavior from large datasets. Patel and Kumar (2020) extended this by using deep learning models, which proved adept at identifying complex patterns in customer data.

Alongside technological methods, understanding why customers leave is also critical. Sharma and Kumar (2019) identified customer service quality, fee structures, and reward programs as key factors influencing customer decisions. Dissatisfaction in these areas often leads to increased churn. Similarly, Lopez and Lee (2022) emphasized the need for personalization in banking services. Their research suggests that banks failing to tailor services to individual needs see higher rates of customer departure. These studies collectively highlight the importance of using sophisticated data analysis techniques alongside a deep understanding of customer needs and preferences in tackling the issue of customer churn in banking.

3. Theoretical Analysis

3.1. Data

The dataset we used consists of 10,000 customers mentioning their age, salary, marital status, credit card limit, credit card category, etc. The data is from the Kaggle website. Since we have only 16.07% of customers who have churned, Thus, it's a bit difficult to train a model to predict churning customers and we should consider the problem of imbalance data. we leveraged all the variables and rename them as the following: customer_age, gender, dependent_count, education_level, marital_status, income_level, card_level, months_on_book and so on in total of 18 variables (See **Table1**)

Table 1. Variables

Variable	Type	Description	Measure Format
Attrition Flag	Dependent	Internal event (customer activity) variable - 1 if the account is closed then else 0	1: if the account is closed 0: if the account is not closed
Customer Age	Independent	Demographic variable - Customer's Age in Years	Measure in integer
Gender	Independent	Demographic variable - M=Male, F=Female	Measure in Nominal data
Dependent Count	Independent	Demographic variable - Number of dependents	Measure in Integer
Education Level	Independent	Demographic variable - Educational Qualification of the account holder	Measure in Nominal data
Marital Status	Independent	Demographic variable - Married, Single, Divorced, Unknown	Measure in Nominal data
Income Category	Independent	Demographic variable - Annual Income Category of the account holder	Measure in ordinary data
Card Category	Independent	Product Variable - Type of Card (Blue, Silver, Gold, Platinum)	Measure in Nominal data
Months on Book	Independent	Period of relationship with bank	Measure in integer
Total Relationship	Independent	Total no. of products held by the customer	Measure in integer
Month Inactive	Independent	No. of months inactive in the last 12 months	Measure in integer
Contacts Count	Independent	No. of Contacts in the last 12 months	Measure in integer
Credit Limits	Independent	Credit Limit on the Credit Card	Measure in integer
Total Revolving	Independent	Total Revolving Balance on the Credit Card	Measure in integer
Avg Open to Buy	Independent	Open to Buy Credit Line	Measure in integer
Total Amount Change	Independent	Change in Transaction Amount (Q4 over Q1)	Measure in percentage of change
Total Trans Count	Independent	Total Transaction Count (Last 12 months)	Measure in integer
Ave Utilization Ratio	Independent	Average Card Utilization Ratio	Measure in percentage

3.2 Preprocessing

We first have to guarantee that the dataset has every value for each variable in all columns and erase the outliers of each variable. Then, certain variables have to be changed from categorical variables into numerical variables. Lastly, we need to be able to handle the imbalance data problem in the dataset. We eliminated the missing values first, and then dropped all the rows that contained NaN values. We then tried to handle the outliers. We implemented detect_outliers, by applying the Interquartile Range (IQR) to identify outliers. Then, we dropped the outliers to make the data more precise since we need to avoid all the outliers that might cause the dataset to be unstable or skewed. The categorical data includes types like nominal or ordinal, and it needs to be converted into a numerical format. We chose to use the One-hot encoding to convert the ordinary data into more numbers to represent them but not multi-binaries because we want to avoid collinearity. For example, the card level is converted as 'Blue': 0, 'Silver': 1, 'Gold': 2, 'Platinum': 3. As we found in the dataset, there is only 16.07% of customers who have churned, there is a huge difference between the binary distribution, and such imbalanced data problem would cause skewed class proportions in directed variables, leading to a potentially problematic model performance. We have considered this potential problem existing, but since the data volume is big enough, we tried to just train the model first, and make this as a future refinement. Possibly in the future, we can refine the data set by using the over-sampling. Specifically, we might use SMOTE, which is an algorithm for oversampling data while the dataset faces the problem of imbalanced data, being used to over-sample the dataset to increase model performance.

We considered three classification algorithms: Logistic Regression, Decision Tree, and Random Forest. We split the data set into two parts, the training set and the test (holdout) set,

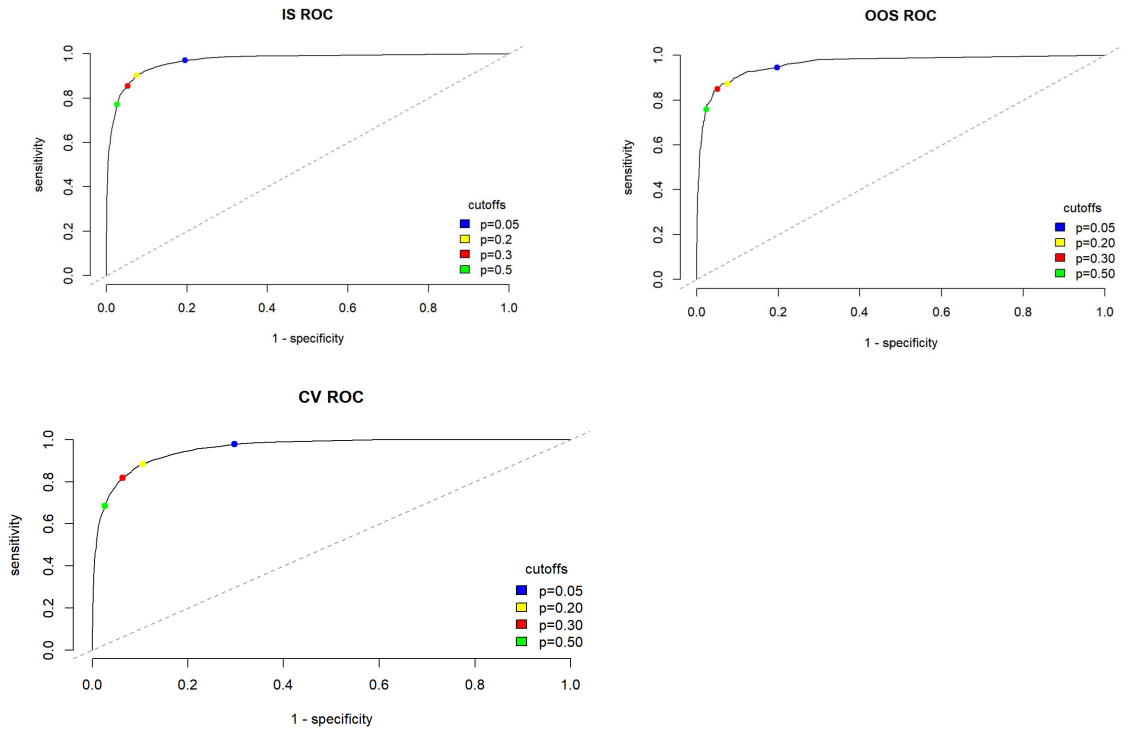
each with a percentage of 80% and 20% respectively. We trained all three models with the training set and measured the model performance measures and accuracy through the test set.

3.3 Methodology

We primarily used logistic regression to classify the data, with 100% training data, 80% training data and 20% testing data, and k-fold cross-validation (k=5). For each resulting model, we drew the ROC curve with different cutting-off p and found the optimal p that balances sensitivity and specificity as the final model. Also, we used the decision tree and random forest to classify the data, with 80% training data and 20% testing data. We finally compared the results and found the best model.

4. Result

Graph 1. ROC of in-sample, out-of-sample, and cross-validation



For the three models, we draw the ROC curve. Then, we used different values of cutting-off p values and drew the corresponding points on the graph. We observed the positions of points on the graph and found the optimal cutting-off value that balances sensitivity and specificity (sensitivity≈specificity) as the final value of the model. Then, we found the sensitivity and specificity of the models. The following graphs show the result.

Table 2. Result of logistic regression

Model	Sensitivity	Specificity
Logistic- 100% training	0.912	0.915
Logistic- 80% training, 20% testing	0.903	0.904
Logistic- k-fold cross-validation (k=5)	0.894	0.883

The three models generate similar results, but the 100% training shows the best-fitting result, while k-fold cross-validation shows the relatively worst-fitting result. The result matches the expectation. The 100% training set results in the highest level of fitting because the training set is the same as the testing set. But this overlap of sets could also have the overfitting issue. In comparison, the k-fold has relatively worse results because it uses different training sets and testing sets. This result should have higher generalizability and decrease the extent of overfitting. Finally, the second result, with 80% training and 20% testing, shows the middle level of fitting. We think the result could be dependent on the choice of seed. But in general, the result has the intermediary generalizability and overfitting problem.

Table 3. Result of Decision Tree

Model	Sensitivity	Specificity
Decision Tree- 80% training, 20% testing	0.969	0.779

We splitted the dataset to 80% training and 20% testing (with the same seed as the previous one). The decision tree generates high sensitivity but relatively low specificity. This could be caused by the imbalance ratio of data points in the two categories. We could possibly generate better results by further processing data to eliminate the imbalance.

Table 4. Result of Random Forest

Model	Sensitivity	Specificity
Random Forest- 80% training, 20% testing	0.989	0.828

We split the data set to 80% training and 20% testing (with the same seed as the previous). The random forest generates both high sensitivity and specificity, but the specificity is also lower than the sensitivity. Further refinement could be done by processing imbalanced data.

Compare the three methods: Comparing models generated by the three methods: logistic regression, decision tree, and random forest. Sensitivity: Random Forest> Decision tree> Logistic regression. Specificity: Logistic regression> Random forest> Decision tree. Through the analysis, we observed the problem of the imbalanced dataset, which impacts the result of random forest and decision tree. After re-processing the data and balancing the dataset, we predicted the performance of random forest and decision tree will be better. Also, another difference is that we included all interaction terms in the logistic regression, while we excluded these terms in the decision tree and random forest (because R cannot run the model with interaction terms). So, if we could further improve the decision tree model or random forest model to include these terms, we predict their performances will be better than the logistic regression model. So, after potential refinement, we predicted the performance ranking will be random forest> decision tree> logistic regression.

5. Conclusion

The three methods, logistic regression, decision tree, and random forest, show similarly good results of fitting. However, we observed the problem of data imbalance by analyzing the results. The problem does not manifest in the logistic regression by choosing the cutting-off p, but the problem presents in the result of the decision tree and random forest. Also, since R cannot run models with interaction terms in the random forest and decision tree model, the results of the two models are further underestimated. Further refinement to the model could include re-processing data to balance data and adding interaction terms to the decision tree and random forest model. We predict that the performance of methods after refinement should be random forest> decision tree> logistic regression.

Works Cited

Goyal, Sakshi. "Credit Card Customers." Online dataset. Kaggle, 2023. https://www.kaggle.com/datasets/sakshigoyal7/credit-card-customers.

Lopez, Maria, and David Lee. 2022. "The Importance of Personalization in Banking Services and Its Impact on Customer Churn." *Journal of Bank Marketing* 24, no. 4 (2022): 112-130.

Patel, Neha, and Amit Kumar. 2020. "Deep Learning Models in Banking Customer Churn Prediction." *Journal of Financial Technology* 21, no. 3 (2020): 58-75.

Sharma, Priya, and Anil Kumar. 2019. "Factors Influencing Credit Card Churn: Customer Service, Fee Structures, and Reward Programs." *Consumer Banking Quarterly* 17, no. 2: 200-215.

Zhang, Ming, and Yang Liu. 2021. "Machine Learning Techniques for Customer Churn Prediction: A Comparative Study." *Journal of Data Science and Analytics* 18, no. 6: 89-104.