Predicting Churning Customers Lab Report for Business

Tristin Burdick

BIA 650 D1 Data Mining for Decision Making

Professor Awe

St. Mary’s University

**Abstract**

Customer churn can directly impact a firm’s bottom line, especially in the financial industry. The management team at a nationally recognized bank is seeking help from an analyst to help solve their customer churn problem. They want the analyst to develop a predictive model that will identify existing customers that have the highest probability of churning. This will give the management team and support staff the ability to preventively intervene before a customer churns, by offering incentives like discounts, reduced fees, and other offerings tailored to the individual.

**Introduction**

Predicting the churn of customers is important because it significantly impacts a firm’s bottom line. Losing customers is costly for any business, as it can lead to a decrease in revenue and an increase in marketing and acquisition costs to replace lost customers (Qualitrics, 2023). In the credit card industry, it is particularly important to retain customers, as credit card companies earn significant revenue through interest and fees. Therefore, predicting which customers are at risk of churning can help credit card companies take proactive steps to retain customers and reduce the overall impact of customer churn on their business.

Furthermore, predicting churning customers allows credit card companies to offer targeted retention incentives to customers who are at risk of churning. By understanding the factors that lead to customer churn, credit card companies can offer incentives such as lower interest rates, bonus rewards points, or other perks to incentivize customers to stay with the company (Gerdeman, 2013). This approach is often more cost-effective than trying to acquire new customers, as retaining existing customers is typically less expensive than acquiring new ones.

A manager from a nationally recognized bank is frustrated that more customers are leaving their credit card services. They contacted an analyst to help them with this problem and see if there was a better way to identify customers that are likely to churn. The manager requests that a predictive model be developed so they can identify existing customers that are the most likely to churn. The model will flag existing customers that are the most likely to churn so the manager or support staff can proactively intervene and prevent them from churning.

The purpose of this study is to help the banking service provider identify existing customers who are at risk of churning. The analyst will use a predictive machine-learning model on the bank’s historical data set. The model will analyze a large number of variables and factors that can influence the behavior of churning customers. By identifying customers who are at risk of churning, the banking firm can take proactive steps to reduce churn rates and improve customer retention.

**Methods**

Predictive modeling is a statistical process that involves developing a model to predict a future outcome based on historical data. The methods used in this report are commonly used in a wide range of industries and applications. This section details how the research was performed and the process that the analyst followed to solve the business problem.

**2.1 Data Collection**

The first step involves collecting and obtaining the data necessary to complete the project. Data was provided to the analyst by the bank manager. Alternatively, the data set is also available on a website called [Kaggle.com](https://www.kaggle.com/datasets/whenamancodes/credit-card-customers-prediction) – an online community of data enthusiasts that collaborate with each other and publish datasets. The data was downloaded from the website and saved on the analyst’s computer. The data set is relatively new and was made available in October 2022.

**2.2 Data Cleaning and Preparation**

Data cleaning and preparation is a critical step in predictive modeling. This step involved cleaning, transforming, and pre-processing the raw data for analysis and modeling. First, the analyst identified irrelevant information in the data set. For example, the data contained classifier metrics that needed to be removed because it wasn’t necessary for modeling or pertaining to the customers’ credit card accounts.

Then, the analyst performed data cleaning procedures by removing inaccuracies and duplicated data. This data set it was already cleaned and in a tabular form and didn’t contain any duplicate values. However, the analyst needed to transform the data by encoding the categorical variables in each column field. This made sure the dataset was in a proper format for predictive modeling and made it easier to analyze the data and spot outliers. The analyst was able to deal with outliers by observing each data field with a box and whisker plot.

The analyst will also perform descriptive statistics on each variable in the data. This will provide them with the necessary information they need to develop an initial hypothesis and determine the context of the data. Data types will be noted by the analyst (integer, character, string, etc.) and will be assessed in the next step. The dependent variable and independent variables will also be examined by the analyst.

**2.3 Feature Selection**

Feature selection was another important step in the predictive modeling process. This step involves selecting the most relevant and important features to include in the model. The purpose of feature selection is to improve the accuracy and performance of the model by reducing the noise, increasing interpretability, and reducing the risk of overfitting (Khandelwal, 2019). Feature selection can help prevent overfitting of the model and improve the model’s ability to make accurate predictions of new data.

The analyst will use two common methods for selecting features – which include univariate techniques and recursive feature elimination. The first thing the analyst will do is conduct a correlation analysis by examining the correlation matrix. They will identify which predictor variables are highly correlated to the target variable - these will be included in the model. The analyst must also be aware of the problems associated with multicollinearity as it can cause unreliable estimates in the model. They will have to exclude predictor variables that are correlated with each other from the model to prevent multicollinearity. The analyst will also perform a stepwise selection method when they are running the model. This is a recursive feature elimination technique that will help identify the most important predictor variables. This prevents the model from overfitting the data by incrementally adding or removing predictor variables.

**2.4 Model Selection**

The analyst had to determine which predictive modeling technique was appropriate for solving the business problem. The outcome was to predict the probability that an existing customer is likely to churn. Therefore, a logistic regression model would be the best model to use in this scenario. The primary reason for this is that logistic regression is designed specifically for predicting binary outcomes, whereas linear regression is designed for continuous outcomes.

Linear regression assumes that the outcome variable is continuous, thus providing a value within a continuous range. Since the business problem is not determining a value within a range of continuous values, but rather a probability between 0 and 1, the logistic model would be the right approach. The analyst will be using SAS Studio to develop the logistic regression model with the data provided. The analyst will first need to transform the data into binary format for the logistic regression model. Then the analyst will prepare the model for training by splitting the data into training and test sets.

**2.5 Model Training**

Model training is an essential part of developing a predictive model and plays a critical role in achieving accurate predictions. The process of training a model involves adjusting the model’s parameters on a training data set so it can learn patterns and relationships on unseen data – in this case accurately predicting the test data set (Roman, 2018).

The analyst will partition the original data set into two sets, one being the training set (80%) and the other being the test set (20%). The analyst will first run the logistic regression model with the training data set and produce a predictive outcome. The model will be evaluated and assessed by the analyst to determine its initial accuracy. The analyst will perform optimization techniques (like stepwise selection, and reevaluating feature selection) to improve the accuracy of the trained model. When the trained model has been successfully optimized, the test set will be provided to determine the overall accuracy of the model with unseen data.

**2.6 Model Evaluation**

The model needs to be evaluated to determine how well it accurately predicts unseen data. The goal of model evaluation is to tell the analyst how well the model performed and identify the inadequacies of the model. The model will be evaluated using common statistical methods.

First, the model will be evaluated using a confusion matrix. The confusion matrix, also known as the classification table, is used to compare the predicted values to the actual values in the data set. The f1 score is commonly used to assess the results of a confusion matrix. In general, an f1 score of 0.5 or 0.6 will be considered good depending on the context.

Next, the model will be evaluated on the ROC curve. The ROC curve is a graphical representation of the probability of distinguishing between classes (Narkhede, 2018). Its commonly used in statistics to evaluate the performance of a binary classifier. Typically, a higher value for the ROC curve indicates better performance.

Finally, the model will be evaluated using the model fit statistics. The model fit statistics, such as AIC (Akaike Information Criterion), SC (Schwarz Bayesian Criterion), and -2 Log L values, are important for evaluating the quality of the model. These statistics provide a quantitative measure of the goodness of fit of the model, which is the extent that the model provides an accurate representation of the data (Penn State, 2023). A model with a better fit will generally have a lower value of AIC, SC, or -2 Log L.

**2.7 Model Deployment**

The importance of model development lies in the ability of the model to improve decision-making. The last step in this project is to implement the model and provide access to stakeholders. This will involve integrating the model into a preexisting system or software at the banking firm. Upon integration, the model will need to continually be monitored, assessed, and evolve as the model is put into practice. Business stakeholders must consider factors of scalability, robustness, usability, and the continued collaboration between developers and data professionals (Shin, 2020).

**Results**

**3.1 Descriptive Statistics and Initial Findings**

The final data set was in binary form (containing 1’s and 0’s) and was fairly large for this project, boasting over 10,000 rows of data and a total of 87 columns. Categorical variables were transformed into their own column field. Continuous fields were split into four or five equally segmented fields. For example, the age range was broken down into five column fields that included customers under the age of 35, 35-44, 45-54, 55-64, and 65+. The data set contained outliers in nearly every column field. However, the values were not significant enough to exclude from the data set and were kept.

The statements below provide further insights from the analysis of credit card cancellation behavior among customers. The data analyst will need to continue to explore these trends to fully understand the factors contributing to credit card cancellations.

**Female customers are more likely to cancel their credit card services than male customers**

This could be due to many reasons, such as differences in spending habits, financial goals, or life events. For example, women may have more debt, lower income, or more caregiving responsibilities, making it harder for them to pay off their balances or manage their finances. The analyst could further investigate the reasons behind this gender gap by looking at demographic, transactional, and behavioral data. Results can be found in figure 1.0.

**The age of the customer had little to do with the cancelation of services**

This suggests that age may not be a significant predictor of credit card churn, contrary to the analyst's initial hypothesis. The reasons for this finding could be impeding factors, such as sample size, data quality, or other confounding variables. The analyst could validate these results by comparing them with other datasets provided by the credit card provider or conducting a sensitivity analysis. Results can be found in figure 1.0.

**Customers with more dependents in their household are more likely to switch credit card providers**

This could be because customers with larger families have more expenses and may be more sensitive to interest rates or fees. For instance, they may need to use credit cards more often to cover household expenses or emergencies, leading to higher balances or missed payments. The analyst could explore this hypothesis by looking at the transactional or demographic data that explain the number of dependents and their financial needs. Results can be found in figure 1.0.

**Customers that aren't married will have a higher likelihood of switching providers**

The possible explanation given is that older married customers are less likely to churn because they have little to no independents living at home. However, this may not apply to all age groups or household types. For example, younger unmarried customers may switch providers more often due to life transitions or job changes. The analyst could investigate this trend by comparing different segments of the population or examining the interaction effects between marital status and other variables. Results can be found in figure 1.0.

**Higher net-worth individuals are less likely to cancel their credit card services**

The reasoning for this could be the fact that they have more financial, better credit scores, or more stable income sources than the average customer. The analyst could confirm this hypothesis by looking at the financial data or conducting a regression analysis that controls for other factors. Results can be found in figure 1.0.

**3.2 Correlation Analysis**

The analyst conducted a correlation analysis to determine the relationships between the variables within the data set. This helped identify independent variables that exhibited characteristics of multicollinearity. These were noted in the data file and had to be tested within the model. The analyst used correlation analysis primarily to understand the relationships among variables. An example of the correlation output can be found in figure 1.1 in the appendix section. The statements below are the most interesting insights that were uncovered during the correlation analysis.

Males had a negative correlation with a credit limit of $5000, meaning that male customers were less likely to have a credit limit of $5000 compared to female customers. However, males had a positive correlation with all other credit limit amounts of higher denominations, indicating that male customers were more likely to have higher credit limits compared to female customers.

There is a strong negative correlation between silver card users and blue card users, indicating that as the number of silver card users in the dataset increases, the number of blue card users decreases. This could be due to a variety of factors, such as differences in credit score, income, or other demographic factors that may influence the type of credit card a person is approved for. It’s also important to note that blue card users tend to have a lower credit limit amount compared to other card types, which could also be a factor in their negative correlation with silver card users. Additionally, the blue card user demographic slightly favors females, which could suggest that the features or benefits of the blue card are more attractive to this particular group.

There is a positive relationship between credit card open rate and credit limit. Specifically, as the credit limit amount increases, the average credit card open rate also increases. This could indicate that customers with higher credit limits are more likely to apply for credit cards and are approved for more credit cards than customers with lower credit limits. It is important to note that correlation does not imply causation, so further analysis would be needed to determine the reasons for this relationship.

There is a positive correlation between the average total revolving balance of $600 and the average utilization rate of 20%, indicating that customers with a higher total revolving balance tend to utilize a higher percentage of their credit limit. A correlation of 54% suggests a moderate to strong positive relationship between these two variables. It is important to note that a high utilization rate may indicate a higher risk of default for the credit card company, and may impact the customer's credit score negatively (Experian, 2023). Therefore, credit card companies often use utilization rates as a key factor in assessing credit risk and determining credit limits for customers.

Finally, there is a moderately strong positive correlation between the total number of transactions made per month and the total transaction amount. In this case, a correlation of 59% indicates that as the total number of transactions made per month increases, the total transaction amount also tends to increase. Additionally, it is important to consider other factors that may influence the correlation between the number of transactions and the transaction amount, such as individual spending habits, demographics, and economic factors.

**3.3 Feature Selection**

Logistic regression models should not have multicollinearity among predictor variables in the model (Allison, 2012). This is a common problem with logistic regression models and classification problems. The analyst runs the risk of developing an unreliable or unstable model that produces inaccuracies. Therefore, this step will focus on ways in which the analyst has reduced the risk of multicollinearity and has produced the best feature set for the model.

Firstly, the analyst viewed the correlation analysis and made a note of which independent variables had exceptionally positive and negatively correlated values. These would be reviewed in the final output, but the analyst wanted to run the model first before excluding unwanted variables.

Next, the analyst ran the model through the logistic regression model in SAS Studio. The initial output of the model was given. The analyst examined the features and the model accuracy and determined that the model needed to be re-run with a new feature set. A stepwise model selection method was used to select a subset of available features that were repeated multiple times until the optimal set of features was determined.

The training set contained 19 features and can be viewed in figure 1.2 The training set shows that all the values selected were significant with a p-value < 0.5. The test set also used a stepwise model selection and excluded 10 features from the training set, having a total of 9 statistically significant features. The feature for the test set can be viewed in figure 1.3.

**3.4 Results from Logistic Model**

The model was first evaluated using the model fit statistics found in figure 1.4. AIC and SC are both measures of the relative quality of a statistical model. The lower the value of AIC and SC, the better the model. In this case, we can see that the AIC and SC for the model with intercept and covariates are much lower than the intercept-only model, indicating that the model with covariates is a better fit.

The -2 Log L statistic is a measure of the overall fit of the model. The lower the value of -2 Log L, the better the model. Again, we can see that the -2 Log L for the model with intercept and covariates is lower than the intercept-only model, indicating a better fit with the model covariates.

The results of the model fit statistics suggest that the model that was produced was *fairly accurate* at predicting the outcome variable. Both the intercept and covariates are a better fit than the intercept-only model. Therefore, the addition of covariates improved the model's ability to predict the outcome variable.

The model was next evaluated on performance by observing the ROC curve found in figure 1.5. The ROC (Receiver Operating Characteristics) curve is a graphical representation of the performance in the binary classifier. The ROC curve value of 0.9489 indicates that the classifier is skewing towards sensitivity, which means that the classifier has a higher true-positive rate than a false-positive rate (Steen, 2020). The model is more conservative in the way it makes predictions and is willing to accept a higher false positive rate to achieve a higher true positive rate. Overall, an ROC curve of 0.9489 is considered *excellent* by statistical standards.

Lastly, the model was evaluated using the confusion matrix, also known as the classification table, shown in figure 1.6. The classification table shows that the model is correctly predicting the outcome (either Event or Non-Event) for a large majority of the observations, even at low probability thresholds. For example, at a probability threshold of 0.02, the model correctly predicted 6800 out of 6824 observations, resulting in a sensitivity of 100% and a specificity of 84.2%. As the probability threshold increases, the model's sensitivity and positive predictive value also increase, indicating that the model is becoming more confident in its predictions and is better able to identify true positives. Overall, the high accuracy rates and increased sensitivity suggest that the logistic regression model is performing well and is a good fit for the data.

**Discussion**

Predicting the churn of customers is important because it significantly benefits the bank branch. By using a logistic regression model to predict customer churn, the management team can identify customers who are at high risk of churning. For example, figure 1.7 shows a table of 143 existing customers that the model identified has the highest likelihood of churn, at a 70% minimum probability rate. This gives the management team and support staff actionable steps to contact this list of customers to proactively intervene and prevent them from churning.

In the future, they can also develop offers and programs to prevent the most likely consumers from churning. The statements below offer some options in which the bank branch can provide value to consumers likely to churn (these are from the insights that were identified in section 3.1 above).

To prevent female customers from churning, the bank could offer tailored credit card services that address their unique needs and financial challenges (Luzio, 2018). For example, the bank could provide credit counseling or financial education programs to help female customers manage their debts or improve their credit scores. The bank could also offer rewards or discounts on expenses that are more common among women, such as childcare or healthcare expenses.

To prevent customers with dependents from churning, the bank could offer credit card services that cater to the specific needs of families, such as cash-back rewards on grocery or gas purchases or low-interest rates on home or car loans (Luzio, 2018). The bank could also provide financial planning or budgeting tools that help customers manage their household expenses more effectively.

To prevent unmarried customers from churning, the bank could provide credit card services that cater to the specific needs of singles, such as cash-back rewards on dining or entertainment expenses or low-interest rates on personal loans (Caporal, 2023).

To prevent high-net-worth customers from churning, the bank could offer exclusive perks or rewards that reflect their higher spending power or investment portfolios (Caporal, 2023). For example, the bank could provide access to premium credit card services that offer travel rewards or provide personalized investment advice and portfolio management services. The bank could also offer special rates or discounts on premium financial products, such as high-yield savings accounts or investment funds.

# References

Allison, P. (2012, September 10). *When Can You Safely Ignore Multicollinearity?* Retrieved from Statistical Horizons: https://statisticalhorizons.com/multicollinearity/#:~:text=Multicollinearity%20is%20a%20common%20problem,unstable%20estimates%20of%20regression%20coefficients.

Caporal, J. (2023, February 15). *Study: Banking Needs, Digital Banking Trends, and Consumer Priorities*. Retrieved from Fool: https://www.fool.com/the-ascent/research/digital-banking-trends/

Experian. (2023). *What is a Credit Utilization Rate?* Retrieved from Experian: https://www.experian.com/blogs/ask-experian/credit-education/score-basics/credit-utilization-rate/#:~:text=Your%20credit%20utilization%20rate%2C%20sometimes,generally%20expressed%20as%20a%20percent.

Gerdeman, D. (2013, November 11). *Forbes*. Retrieved from A Smarter Way To Reduce Customer Churn: https://www.forbes.com/sites/hbsworkingknowledge/2013/11/11/a-smarter-way-to-reduce-customer-churn/?sh=a22a7162c0a4

Khandelwal, R. (2019, October 24). *Feature Selection : Identifying the best input features*. Retrieved from Towards Data Science: https://towardsdatascience.com/feature-selection-identifying-the-best-input-features-2ba9c95b5cab

Luzio, C. (2018, November 1). *BankThink Why women keep leaving banking*. Retrieved from American Banker: https://www.americanbanker.com/opinion/why-women-keep-leaving-banking

Narkhede, S. (2018, June 26). *Understanding AUC - ROC Curve*. Retrieved from Towards Data Science: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

Penn State. (2023). *6.2.3 - More on Model-fitting*. Retrieved from Penn State: https://online.stat.psu.edu/stat504/lesson/6/6.2/6.2.3

Qualitrics. (2023). *Customer churn: how to calculate, measure, and stop losing customers*. Retrieved from qualtrics: https://www.qualtrics.com/experience-management/customer/customer-churn/

Roman, V. (2018, December 23). *How To Develop a Machine Learning Model From Scratch*. Retrieved from Towards Data Science: https://towardsdatascience.com/machine-learning-general-process-8f1b510bd8af

Shin, T. (2020, April 23). *What Does it Mean to Deploy A Machine Learning Model?* Retrieved from Towards Data Science: https://towardsdatascience.com/what-does-it-mean-to-deploy-a-machine-learning-model-dddb983ac416

Steen, D. (2020, September 12). *Understanding the ROC Curve and AUC*. Retrieved from Towards Data Science: https://towardsdatascience.com/understanding-the-roc-curve-and-auc-dd4f9a192ecb

**Appendix**

**Figure 1.0 Descriptive Statistics for Initial Insights**

Chart

Description automatically generated

**Figure 1.1 Correlation Matrix**

Table

Description automatically generated

**Figure 1.2 Training Set Feature Selection**

Table

Description automatically generated

**Figure 1.3 Test Set Feature Selection**

Table

Description automatically generated

**Figure 1.4 Model Fit Statistics**

Chart, line chart

Description automatically generatedChart

Description automatically generated

**Figure 1.5 ROC Curve**

Table

Description automatically generated

Table

Description automatically generated

**Figure 1.6 Confusion Matrix**

**Table

Description automatically generated**

**Figure 1.7 Flagged Customers**