Final Report: Causes for Customer Attrition

MGT 4050-A: Data Analytics for Business

Professor Frederic Bien

Spring 2021

Group Members: Semeen Hajira, Eunice Kim, Victor Lai

Introduction & Dataset

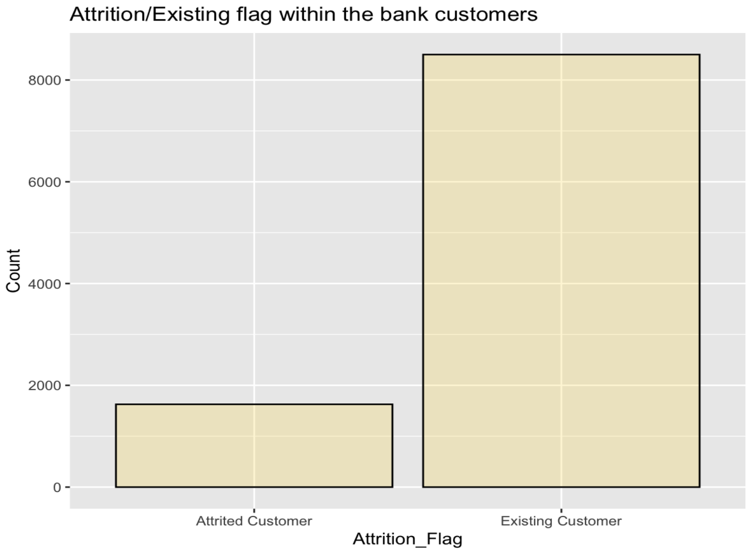
This report is centered around the analysis of a bank's dataset to determine which variables have the highest impact on customer attrition, also known as the customer turnover rate. Specifically, our research question is: "According to the bank's data, what variables are most influential in customer attrition?" The analysis took into account numerous socioeconomic factors, such as the number of dependents a customer had, a customer's marital status, and a customer's education level. The analysis also considered categorical factors such as gender and credit card type.

The dataset in use was developed by Sakshi Goyal and was found on kaggle.com under the name "Credit Card Customers Dataset". While the exact number of previous uses of this dataset cannot be determined, the dataset profile had more than 27,000 downloads and nearly 1,500 upvotes as of April 2021. Unfortunately, no prior uses of the dataset can be found, and therefore it is unknown what conclusions previous analyses of the dataset have been developed.

The dataset was developed for the specific purpose of predicting what kind of credit card customers were leaving the bank. The hope was once the variables were better understood relative to the customer attrition rate, the bank will be more able to adjust their services that improves customer retention.

Some key variables of the dataset were the average card utilization ratio and customer income category. However, the most important variable was the customer attrition flag which shows whether each customer was retained or lost through a binary variable. Some interesting characteristics we can immediately see from basic charts of the dataset include a slightly larger female population compared to male, the most common by far education level among customers being a graduate degree, the vast majority of customers having a blue card (the lowest tier of credit cards the bank offered), and about a 85:16 ratio for customers retained: customers lost, this is shown in figure 1.1.

Figure 1.1



Since there are over ten thousand individual customers logged into the dataset, there is no concern of there being too little data. Additionally, the dataset contains many different variables and the analysis utilized all data entries within the dataset, therefore reducing any bias concerns.

Both linear regression and logistic regression were used in the analysis. Linear regression was chosen to help identify the correlation strength between certain variables and average utilization ratio since it is assumed average utilization ratio can be used as an indicator of customer attrition. Logistic regression was later used to better determine conclusions due to the numerous categorical variables involved in the analysis.

Linear Regression

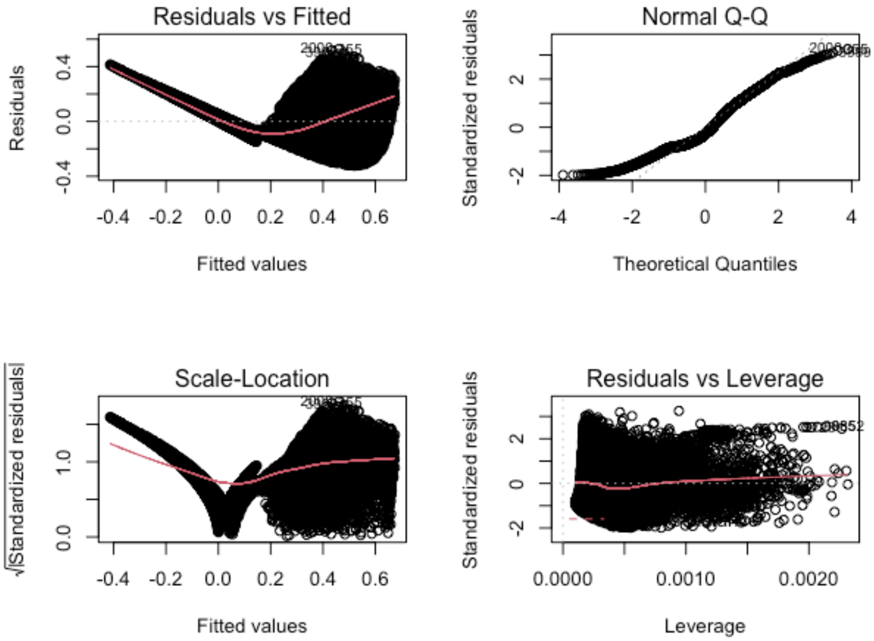
Linear regression was performed on the continuous variables within the dataset. The relationship between average utilization ratio and credit limit, months on book, total revolving balance, and total transaction was analyzed with the assumption that average utilization ratio could be used to indicate a customer's likelihood of attrition. Figure 2.1 details the results of the linear regression.

Figure 2.1: Output of Linear Regression Model

	Estimate	*Std. Error*	*t value*	*Pr(>|z|)*
(Intercept)	1.792e-01	8.172e-03	21.924	<2e-16
Credit_Limit	-1.528e-05	1.809e-07	-84.466	<2e-16
Months_On_Book	-3.741e-04	2.029e-04	-1.844	0.0652
Total_Revolving_Balance	2.192e-04	1.992e-06	110.062	<2e-16
Total_Transaction_Amount	-3.137e-06	4.850e-07	-6.468	1.04e-10

With an R-squared value of 0.651, the results show that months on book was the only independent variable that had a statistically insignificant p-value, and therefore was not included in any analysis. Total revolving balance was the only variable that had a non-negative coefficient, indicating that a one unit ($1.00) increase in total revolving balance would increase average utilization rate by 0.0002192. Since we were interested in which variables would cause customer attrition, and in turn a lower average utilization ratio, a deeper analysis of credit limit was necessary. Credit limit proved to be the variable that had the greatest negative effect on average utilization ratio, while remaining statistically significant.

Figure 2.2: Graph Output of Linear Regression Model



To look further into the normality of the data, the graphs corresponding to the linear regression model were outputted and analyzed. In Figure 2.2, the Residuals vs. fitted graph indicates heteroscedasticity, while the slightly tailed Q-Q plot suggests that there could be some non-linearity. Because these findings violate the assumptions of linear regression, non-linear transformations were utilized to achieve a more normal distribution and more constant variance.

Non-Linear Transformations

Model 1 consisted of a linear-linear model between credit limit and average utilization ratio. With an R-squared value of 0.2333, the results indicate that a one dollar increase in credit limit will decrease average utilization ratio by 0.00001465 units, holding all other factors constant. Model 2 consisted of a linear-logistic model, where the log of credit limit was evaluated. With an R-squared value of 0.2874, the results from Model 2 indicate that a 1% increase in credit limit will decrease average utilization ratio by 0.00258276 units, holding all other factors constant. Model 3 consisted of a logistic-linear model, where the log of average utilization ratio was evaluated. With an R-squared value of 0.2377, the results show that as credit limit increases by $1.00, average utilization ratio decreases by approximately 0.001107 units, holding all other factors constant. Lastly, Model 4 was a logistic-logistic model, in which the log of both average utilization ratio and credit were taken and utilized in the regression. With an R-squared value of 0.2803, the results indicate that as credit limit increases by 1%, average utilization ratio decreases by 0.116972%, holding all other factors constant.

When comparing all non-linear transformation models and their respective R-squared values (shown in Figure 2.3), Model 2 showed the largest R-squared value relative to the other models. However, the R-squared value of Model 2 was not high enough to provide a clear conclusion and did not suggest a strong correlation. Therefore, further tests will be necessary to be able to make concrete conclusions and suggestions to the bank in order to keep customers.

Figure 2.3: R-Squared & Adjusted R-Squared Comparison for Non-Linear Transformation Models

	R-Squared	***Adjusted R-Squared***
Model 1: Linear-Linear	0.2333	0.2332
Model 2: Linear-Log	0.2874	0.2873
Model 3: Log-Linear	0.2377	0.2376
Model 4: Log-Log	0.2803	0.2803

Logistic Regression

A logistic regression analysis was conducted in order to investigate if there is a relationship between the dependent binary variable named Attrited and the independent variables named AVG_Utilization_Ratio, Credit_Limit, and Months_Inactive_12_mon. Additionally, because the income category has 5 different ranges, we included four dummy variables in our logistic regression model. Out of all the predictor variables that were tested, 5 were found to contribute to the model: AVG_Utilization_Ratio, Credit_Limit, Months_Inactive_12_mon, and the 60K-80K income category. Needless to say, the other variables emerged to be statistically insignificant in the logistic regression model. Figure 3.1 summarizes the results.

Figure 3.1: Output of the Logistic Regression Model

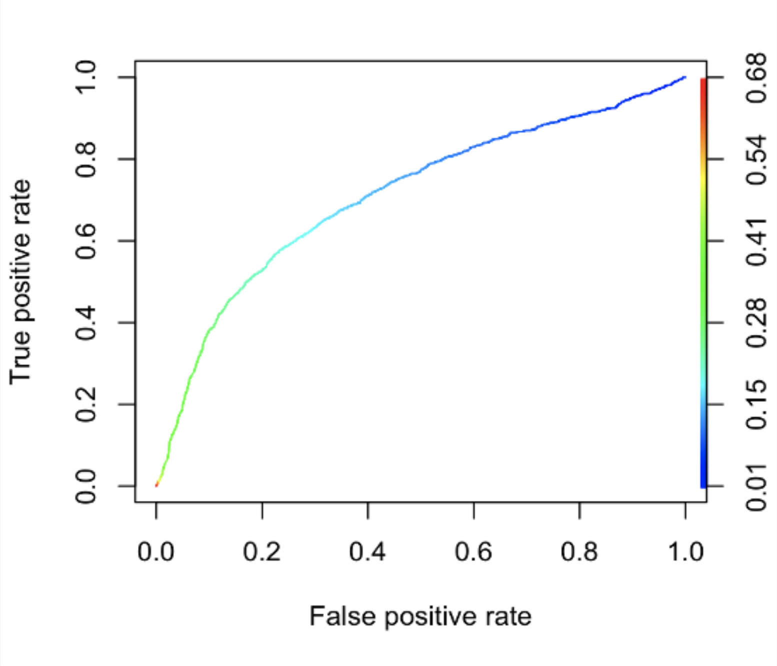
	Estimate	***Std. Error***	***z value***	***Pr(>|z|)***
(Intercept)	-1.657e+00	1.110e-01	-14.923	<2e-16
Avg_Utilization_Ratio	-2.759e+00	1.328e-01	-20.777	<2e-16
Credit_Limit	-3.648e-05	4.019e-06	-9.075	<2e-16
Months_Inactive_12_mon	3.950e-01	2.692e-02	14.670	<2e-16
Less_than_40K	7.998e-02	8.636e-02	0.926	0.35435
I40_60K	-1.184e-01	9.761e-02	-1.213	0.22530
I60_80K	-3.190e-01	1.030e-01	-3.098	0.00195
I80_120K	-4.128e-02	9.677e-02	-0.427	0.66968

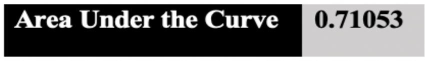
The estimated log odds favored a decrease of nearly 2.759 in the likeliness of customer attrition every one unit increase in AVG_Utilization_Ratio. Similarly, the estimated log odds also favored a 0.00003648 and 0.395 decrease in the likeliness of customer attrition for every one unit increase in a customer's credit limit and the number of months that customer was inactive respectively. Lastly, a customer being within the income category of $60K-$80K decreases the log odds and odds of the prospect of customer attrition by 38%.

ROC Curve, Sensitivity, & Specificity of the Logistic Regression Model

As a means of determining the ideal cutoff point, we analyzed the ROC Curve for our model. The ROC curve highlighted all possible true positive and false positive rates for our model. According to the high Area Under the Curve, it is fitting to say that the logistic regression model did well in terms of distinguishing between attritted and existing customers. This is illustrated by figure 3.2.

Figure 3.2: ROC Curve





Area Under the Curve	0.71053

After taking the ROC curve into consideration and some trial and error, we decided that the most ideal cutoff point would be 0.20. As the false positive rate (specificity) increases the true positive rate (sensitivity) also increases. In order to measure the validity of our logistic regression model, we wanted to calculate the sensitivity and specificity. Sensitivity, in regards to our test, would be the ability of the test to correctly identify the customers that have left the bank. On the contrary, specificity would be the ability of the test to correctly identify the bank customers as existing or non-attrited customers. The information needed to calculate these figures is demonstrated in figure 3.3.

Figure 3.3: Confusion Matrix

	0	***1***	***Total***
0	6579 (TN)	1921 (FP)	8500
1	703 (FN)	924 (TP)	1627
Total	7282	2845	2845

Calculated with cutoff = 0.20

Figure 3.3 shows that 1627 bank customers were flagged for attrition. Of those flagged for attrition, our test correctly identified 924 customers as such. Therefore, sensitivity of the test was calculated to 56.79% (924 (TP) / (924 (TP) + 703 (FN))). Because this is relatively high value, it was concluded that the test did well in correctly identifying customers who have left the bank. Only a few amount of attrited customers were missed by the test. Next, we determine that 8500 customers in the bank dataset were existing customers. Of those not flagged for attrition, our test correctly identified 6579 customers as such. Therefore, specificity of the test was calculated to be

77.4% (6579 (TN) / (6579 (TN) + 1921 (FP))). As a result, it was concluded that the test was strong at correctly identifying customers who have not left the bank. Additionally, it can also be implied that because our test was so highly specific, our test has a low false positive rate (22.6%) i.e. very few results of the test indicate that a customer has left the bank (attrited customer) when in reality the customer has not left the bank (existing customer).

Finally, we checked for the precision and accuracy of our model. The precision calculated to be approximately 32% (924/(924+1921)). This indicates that our model was weak when evaluating customers in the dataset that were predicted to be Attrited. However, the overall accuracy of the model was relatively high at approximately 74% ((924+6579)/10127). With this, we can conclude that the logistic regression test that was performed was considerably good at accurately predicting customer attrition at the bank from the data that was provided and used.

Conclusion

In summary, the analyses show the variables with the greatest impact on customer attrition are the average card utilization ratio and the number of months a customer is inactive. A notable trend of the bank's current customer base is customers within the income range of $60,000 to $80,000 are most likely to be long term loyal customers of the bank. This is presumably due to the bank's current services and offerings matching the needs of customers of that income bracket.

As for recommendations, the data supports targeted advertising towards individuals within the $60,000 to $80,000 income bracket. Specific actions that can be initiated would be to run small promotions to attract customers of that income range. Additionally, an introduction of a referral rewards system could be beneficial. The reasoning behind the referral reward system is the idea people are generally friends with others of a similar socioeconomic status. With people in the $60,000 to $80,000 income bracket referring others with financial similarities, the population associated with long-term loyalty will grow. These efforts will also enable a lower customer attrition rate.