Logistic Regression Model: Predict Customer Churn Rate for the Credit Card

Yingren Luo (1004194873)

2020/11/30

Abstract

Customers are quite significant, which has a significant impact on any business. Therefore, monitoring and forecasting the customer churn rate is a crucial point for any business. This study aims to investigate the suitable predictors for predicting customer churn rate using credit card customer data from *kaggle*. The data is separated into a training set and testing. The logistic regression model is used, and stepwise regression with BIC is used to selected specific predictors. The result shows that using data in the training set, there are 12 predictors useful in the model, which are **Gender**, **Marital Status**, **Total number of products**, **Average Open to Buy Credit Line in the Last 12 Months**, **Average Card Utilization Ratio**, **Number of Contacts in the Last 12 Months**, **Number of Dependents**, **Number of Months on Book**, **Number of Months Inactive in the Last 12 Months**, **Change in Transaction Count in the Last 12 Months**, **Total Transaction Amount in the Last 12 Months** and **Total Transaction Count in the Last 12 Months**. The estimated customer churn rate is 11.45%, and the accuracy is 89.26%. In conclusion, the logistic model needs to be improved, and further research needs to be done.

Key words: Credit Card, Customer Churn Prediction, Logistic Regression Model

Introduction

Customers are individuals or business that purchases another company's products or services (3), which significantly impact any business. As we can see from the adage, "The customer is always right.", the importance of the customer can be interpreted that happy customers are more likely to do business with those companies again. Therefore, many companies pay attention to keep the customer relationship, but the loss of the customer, customer churn still happens. Customer churn, also known as customer attrition, is a phenomenon that existing customers stop doing business with a company (1). It has a significant adverse effect on the business if the customer churn rate is high. Thus, having the ability to predict the customer churn rate is a crucial point for any business.

In this case, I would like to focus on the bank and find suitable predictors for predicting the customer churn rate of the credit card. Nowadays, the loss of customers of commercial banks is a quite serious problem. For instance, many retail banks' customer attrition rate is between 20% and 25% (2). Besides, attracting new customers has a higher cost comparing to maintain the existing customers (4), which indicates that long-term customers would produce more profits. For example, by reducing the customer churn rate by about 5%, a bank can increase its profits by up to 85% (5), which indicates the importance of reducing the customer churn rate. That is the reason why I want to demonstrate what might be related to customer churn rate.

The data set, "BankChurners.csv", will be used to predict the customer churn rate via a logistic model. In the *Methodology* section, I will describe the data, the way used to find a parsimonious model and the model used to predict the customer churn rate. In the *Result* section, the model validation analysis will be performed. In the *Discussion* section, the assumptions of logistic regression model will be checked and the results will be summarized and interpreted. The weakness and future direction will be mentioned.

Methodology

Data

The data used to do the analysis comes from *kaggle* website, which is provided by the bank manager who is facing a problem that more and more customers stop using their credit card services (10). There is not much information about how this data set is collected and how they deal with the non-response problem. It is probably collected through a survey via email or phone call. If customers do not answer the phone or reply to the email, they may try to call people or send them again. The target population is all the customers in this bank. The frame population is all the customers that they can reach out to by email or phone. The sampling population is all the customers who have or used to have credit card services in this bank. There are 10127 samples in the data set.

Table 1 – Characteristics of Data

Variable	Description
Clientnum	Client number, which is the unique identifier for customer holding the account
Attrition_Flag	Customer activity whether the account is closed or not (Attrited Customer, Existing Customer)
Customer_Age	The age of customer in years
Gender	The gender of customer (F indicates female, M indicates male)
Dependent_count	The number of dependents, people who get financial support from customer
Education_Level	The educational level of customer (Uneducated, High School, College, Graduate, Post-Graduate, Doctorate, Unknown)
Marital_Status	Whether customer get married or not (Married, Single, Divorced, Unknown)
Income_Category	Annual income of account holder (Less than $40K, $40K-$60K, $60K-$80K, $80K-$120K, $120K+, Unknown)
Card_Category	The type of card (Blue, Silver, Gold, Platinum)
Months_on_book	Number of months on book, which gives the period of relationship with bank
Total_Relationship_Count	Total number of products that customer has
Months_Inactive_12_mon	Number of months inactive in the last 12 months
Contacts_Count_12_mon	Number of contacts in the last 12 months
Credit_Limit	Credit limit to the credit card
Total_Revolving_Bal	Total revolving balance on the credit card
Avg_Open_To_Buy	open to buy credit line (Average of last 12 months)
Total_Amt_Chng_Q4_Q1	Change in transaction calculated by Quarter 4 Over Quarter 1 (Q/Q)
Total_Trans_Amt	Total transaction amount (Last 12 months)
Total_Trans_Ct	Total transaction count (Last 12 months)
Total_Ct_Chng_Q4_Q1	Change in transaction count calculated by Quarter 4 Over Quarter 1 (Q/Q)
Avg_Utilization_Ratio	Average card utilization ratio

As we can see from the **Table 1**, it shows the 21 variables in the original data set and also its descriptions.

For the drawbacks, as we mentioned above, there are some *Unknown* existing in the data set, which indicates the missing information and may affect the accuracy of the following analysis. The answer for variables, *Income_Category*, is separated into different categories instead of giving the exact numbers, which means that the data is not so accurate. The Uneducated in *Education_Level* should include a primary school, secondary school and people who do not go to school at all. Despite the problems in the survey, there is also a problem with the non-response, which indicates bias. What is more, customer may be dishonest, which would make the result inaccurate. For the advantages, the questionnaire has a similar structure for all the questions and

is easy to answer.

For the following analysis, variables *Clientnum*, *Credit_Limit*, and *Total_Revolving_Bal* are removed from the data set. The reason for removing *Clientnum* is that it is just an identifier, which would not help to build the model. For *Credit_Limit* and *Total_Revolving_Bal*, the reason is that variable *Avg_Open_To_Buy* is the difference between these two variables and variable *Avg_Utilization_Ratio* is calculated by *Total_Revolving_Bal* dividing *Credit_Limit* (8), that both have already included the information in *Credit_Limit* and *Total_Revolving_Bal*. All the *Unknown* are also removed. The cleaned data set is randomly divided into a training set used to build the model and a testing set to evaluate the model.

Figure 1 – Barplot for Customer Credit Card Activity

As we can see from *Figure 1*, there are many more existing customers than attrited customers.

Table 2 – Counts and Percentage for Credit Card Activity

	Attrited Customer	Existing Customer
Counts	891	4775
Percentage	0.1572538	0.8427462

Table 2 shows 891 attrited customers, which is 15.72538%, while the number of existing customers is 4775, which is 84.27462% of all the customers in the training data set.

Figure 2 – Histograms for all the numeric variables

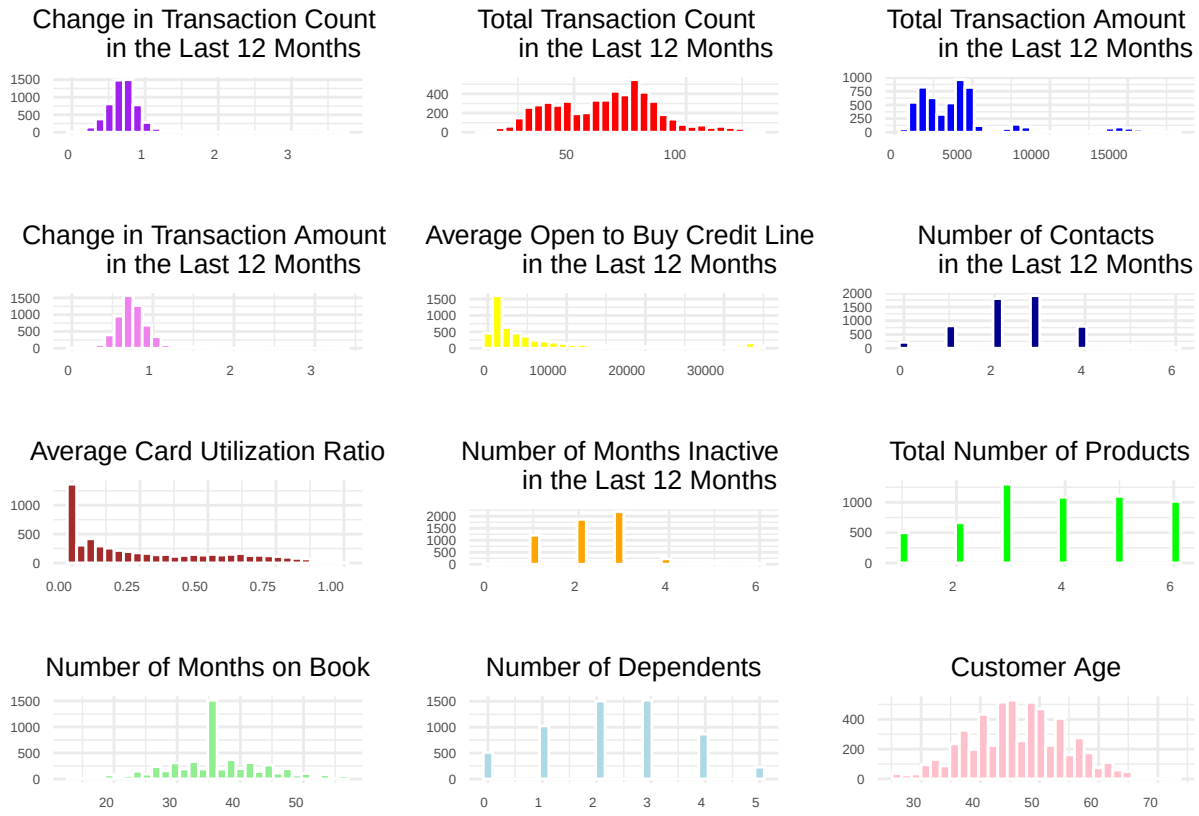


Figure 1 includes 12 small histograms for each numeric variable. For the **Average Open to Buy Credit Line in the Last 12 Months** and **Average Card Utilization Ratio**, the histogram is strongly right-skewed. For **Customer Age**, **Number of Contacts in the Last 12 Months**, **Number of Dependents** and **Number of Months on Book**, the histogram is symmetric and bell-curved. For **Number of Months Inactive in the Last 12 Months**, **Change in Transaction Amount in the Last 12 Months** and **Change in Transaction Count in the Last 12 Months**, the histogram is little right-skewed. For **Total Transaction Amount in the Last 12 Months** and **Total Transaction Count in the Last 12 Months**, the histogram is bi-model.

Figure 3 – Barplots for all the categorical variables

Figure 2 contains all the bar plots for the categorical variables. For **Gender**, the number of females is a little bit smaller than the male. For **Educational Level**, most customers are graduates. A small number of customers are doctorate or post-graduate, which are similar to each other. For **Marital Status**, most customers are married or single. The number of the married customer is a little bit more than single. There is only a small amount of customers divorced. For **Income Level**, most customers have income less than 40K. The number of customers whose income is between 40K and 60K, between 60K and 80K, and between 80K and 120K is quite similar. There is only a small amount of customers with an income of more than 120K. For **Card Category**, most customers blue card. There is a small number of customers choosing silver, gold or platinum cards.

Model

In this final project, I want to use a logistic regression model to model the probability of customer churn rate of credit cards. "Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary response variable." (6) The reason for using the logistic regression model, in this case, is that the response variable, *Attrition_Flag*, the customer activity, whether the account is active or not, is binary.

Here, the model's specific predictors are chosen by stepwise regression using the bayesian information criterion (BIC). The stepwise selection method is a step-by-step iterative construction of a regression model that combines forward selection and backward selection (16). In this case, it starts with a logistic regression model containing all the independent variables and adds or removes a potential predictor at a time based on the BIC. "BIC is an estimate of a function of the posterior probability of a model being true, under a certain Bayesian setup." (17) In this case, the model with lower BIC will be chosen. I use stepwise selection because it gives the removed predictor a chance to enter the model again, which checks more combinations of different predictors. The reason why I use BIC instead of AIC is that it has a larger penalty for adding a predictor, which would be less likely to get an over-fitting model.

The logistics regression model is:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{Gender\_M} + \beta_2 x_{Dependent\_count} + \beta_3 x_{Marital\_Status\_Married} + \beta_4 x_{Marital\_Status\_Single} +$$

$$\beta_5 x_{Months\_on\_book} + \beta_6 x_{Total\_Relationship\_Count} + \beta_7 x_{Months\_Inactive\_12\_mon} + \beta_8 x_{Contacts\_Count\_12\_mon} +$$

$$\beta_9 x_{Avg\_Open\_To\_Buy} + \beta_{10} x_{Total\_Trans\_Amt} + \beta_{11} x_{Total\_Trans\_Ct} + \beta_{12} x_{Total\_Ct\_Chng\_Q4\_Q1} +$$

$$\beta_{13} x_{Avg\_Utilization\_Ratio} + \epsilon$$

p represents the probability of customer stopping using credit card.

β_0 represents the intercept of the model, which is the log odds when all predictors is 0.

β_1 represents log of odds ratio between the female group and male group.

β_2 represents the slope of the model. The number of dependents increases by 1 unit, the log odds also increases by a β_1 unit when other predictors stay the same.

β_3 represents log of odds ratio between customer who is married and customer who is not.

β_4 represents log of odds ratio between customer who is single and customer who is not.

β_5 represents the slope of the model. The period of relationship with bank increases by 1 unit, the log odds also increases by β_5 unit when other predictors stay the same.

β_6 represents the slope of the model. The total number of products that customer has increases by 1 unit, the log odds also increased by β_6 unit when other predictors stay the same.

β_7 represents the slope of the model. The number of months inactive in the last 12 months increases by 1 unit, the log odds also increased by β_7 unit when other predictors stay the same.

β_8 represents the slope of the model. The number of contacts in the last 12 months increases by 1 unit, the log odds also increased by β_8 unit when other predictors stay the same.

β_9 represents the slope of the model. The average open to buy credit line in the last 12 months increases by 1 unit, the log odds also increased by β_9 unit when other predictors stay the same.

β_{10} represents the slope of the model. The total transaction amount increases by 1 unit, the log odds also increased by β_10 unit when other predictors stay the same.

β_{11} represents the slope of the model. The total transaction count increases by 1 unit, the log odds also increased by β_{11} unit when other predictors stay the same.

β_{12} represents the slope of the model. The change in transaction count increases by 1 unit, the log odds also increased by β_{12} unit when other predictors stay the same.

β_{13} represents the slope of the model. The average card utilization ration increases by 1 unit, the log odds also increased by β_{13} unit when other predictors stay the same.

Result

Table 3 – Summary table of logistics regression model

Coefficients	Estimate	Std. error	z value	Pr(>
(Intercept)	-5.482e+00	4.449e-01	-12.323	< 2e-16
GenderM	7.649e-01	1.121e-01	6.824	8.85e-12
Dependent_count	-1.302e-01	3.828e-02	-3.400	0.000674
Marital_StatusMarried	5.404e-01	1.856e-01	2.911	0.003604
Marital_StatusSingle	-8.836e-02	1.859e-01	-0.475	0.634543
Months_on_book	1.800e-02	6.073e-03	2.964	0.003035
Total_Relationship_Count	4.393e-01	3.559e-02	12.344	< 2e-16
Months_Inactive_12_mon	-5.729e-01	5.159e-02	-11.104	< 2e-16
Contacts_Count_12_mon	-4.552e-01	4.641e-02	-9.807	< 2e-16
Avg_Open_To_Buy	2.060e-05	6.781e-06	3.038	0.002380
Total_Trans_Amt	-4.224e-04	2.812e-05	-15.021	< 2e-16
Total_Trans_Ct	1.066e-01	4.627e-03	23.039	< 2e-16
Total_Ct_Chng_Q4_Q1	3.226e+00	2.382e-01	13.547	< 2e-16
Avg_Utilization_Ratio	2.609e+00	2.238e-01	11.656	< 2e-16

From the *Table 3*, we can see that the estimated logistics regression model is:

$$log(\frac{\hat{p}}{1-\hat{p}}) = \hat{\beta}_0 + \hat{\beta}_1 x_{Gender\_M} + \hat{\beta}_2 x_{Dependent\_count} + \hat{\beta}_3 x_{Marital\_Status\_Married} + \hat{\beta}_4 x_{Marital\_Status\_Single} +$$

$$\hat{\beta}_5 x_{Months\_on\_book} + \hat{\beta}_6 x_{Total\_Relationship\_Count} + \hat{\beta}_7 x_{Months\_Inactive\_12\_mon} + \hat{\beta}_8 x_{Contacts\_Count\_12\_mon} +$$

$$\hat{\beta}_9 x_{Avg\_Open\_To\_Buy} + \hat{\beta}_{10} x_{Total\_Trans\_Amt} + \hat{\beta}_{11} x_{Total\_Trans\_Ct} + \hat{\beta}_{12} x_{Total\_Ct\_Chng\_Q4\_Q1} +$$

$$\hat{\beta}_{13} x_{Avg\_Utilization\_Ratio}$$

$$= -5.482 + 0.7649 x_{Gender\_M} - 0.1302 x_{Dependent\_count} + 0.5404 x_{Marital\_Status\_Married} - 0.08836 x_{Marital\_Status\_Single}$$

$$+ 0.018 x_{Months\_on\_book} + 0.4393 x_{Total\_Relationship\_Count} - 0.5729 x_{Months\_Inactive\_12\_mon} - 0.4552 x_{Contacts\_Count\_12\_mon} +$$

$$+ 0.0000206 x_{Avg\_Open\_To\_Buy} - 0.0004224 x_{Total\_Trans\_Amt} + 0.1066 x_{Total\_Trans\_Ct} + 3.226 x_{Total\_Ct\_Chng\_Q4\_Q1}$$

$$+ 2.609 x_{Avg\_Utilization\_Ratio}$$

The $\hat{\beta}_0$ is the estimated intercept of the estimated logistic regression model, which is -5.482. The standard error of $\hat{\beta}_0$ is 0.4449. The null hypothesis, H_0 is $\beta_0 = 0$, while the alternative hypothesis, H_a is $\beta_0 \neq 0$. Since the p-value is < 2e-16, which is smaller than 0.05, there is very strong evidence against the null hypothesis, H_0.

The $\hat{\beta}_1$ is the estimated coefficient of gender, whether is male or not, of the estimated logistic regression model, which is 0.7649. The standard error of $\hat{\beta}_1$ is 0.1121. The null hypothesis, H_0 is $\beta_1 = 0$, while the alternative hypothesis, H_a is $\beta_1 \neq 0$. Since the p-value is 8.85e-12, which is smaller than 0.05, there is very strong evidence against the null hypothesis, H_0.

The $\hat{\beta}_2$ is the estimated coefficient of the number of dependents of the estimated logistic regression model, which is -0.1302.
The standard error of $\hat{\beta}_2$ is 0.03828. The null hypothesis, H_0 is $\beta_2 = 0$, while the alternative hypothesis, H_a is $\beta_2 \neq 0$. Since the p-value is 0.000674, which is smaller than 0.05, there is very strong evidence against the null hypothesis, H_0.

The $\hat{\beta}_3$ is the estimated coefficient of marital status, whether customer is married or not of the estimated logistic regression model, which is 0.5404. The standard error of $\hat{\beta}_3$ is 0.1856. The null hypothesis, H_0 is $\beta_3 = 0$, while the alternative hypothesis, H_a is $\beta_3 \neq 0$. Since the p-value is 0.003604, which is smaller than 0.05, there is very strong evidence against the null hypothesis, H_0.

The $\hat{\beta}_4$ is the estimated coefficient of marital status, whether customer is single or not of the estimated logistic regression model, which is -0.08836. The standard error of $\hat{\beta}_4$ is 0.1859. The null hypothesis, H_0 is $\beta_4 = 0$, while the alternative hypothesis, H_a is $\beta_4 \neq 0$. Since the p-value is 0.634543, which is larger than 0.05, there is no evidence against the null hypothesis, H_0.

The $\hat{\beta}_5$ is the estimated coefficient of the period of relationship with bank of the estimated logistic regression model, which is 0.018. The standard error of $\hat{\beta}_5$ is 0.006073. The null hypothesis, H_0 is $\beta_5 = 0$, while the alternative hypothesis, H_a is $\beta_5 \neq 0$. Since the p-value is 0.003035, which is smaller than 0.05, there is very strong evidence against the null hypothesis, H_0.

The $\hat{\beta}_6$ is the estimated coefficient of the total number of products that customer has of the estimated logistic regression model, which is 0.4393.
The standard error of $\hat{\beta}_6$ is 0.03559.
The null hypothesis, H_0 is $\beta_6 = 0$, while the alternative hypothesis, H_a is $\beta_6 \neq 0$. Since the p-value is $<$ 2e-16, which is smaller than 0.05, there is very strong evidence against the null hypothesis, H_0.

The $\hat{\beta}_7$ is the estimated coefficient of the number of month inactive in the last 12 months of the estimated logistic regression model, which is -0.5729.
The standard error of $\hat{\beta}_7$ is 0.05159.
The null hypothesis, H_0 is $\beta_7 = 0$, while the alternative hypothesis, H_a is $\beta_7 \neq 0$. Since the p-value is $<$ 2e-16, which is smaller than 0.05, there is very strong evidence against the null hypothesis, H_0.

The $\hat{\beta}_8$ is the estimated coefficient of the number of contacts in the last 12 months of the estimated logistic regression model, which is -0.4552
The standard error of $\hat{\beta}_8$ is 0.04641
The null hypothesis, H_0 is $\beta_8 = 0$, while the alternative hypothesis, H_a is $\beta_8 \neq 0$. Since the p-value is $<$ 2e-16, which is smaller than 0.05, there is very strong evidence against the null hypothesis, H_0.

The $\hat{\beta}_9$ is the estimated coefficient of the open to buy credit line of the estimated logistic regression model, which is 0.0000206. The standard error of $\hat{\beta}_9$ is 0.000006781.
The null hypothesis, H_0 is $\beta_9 = 0$, while the alternative hypothesis, H_a is $\beta_9 \neq 0$. Since the p-value is 0.00238, which is smaller than 0.05, there is very strong evidence against the null hypothesis, H_0.

The $\hat{\beta}_{10}$ is the estimated coefficient of the total transaction amount of the estimated logistic regression model, which is - 0.0004224.
The standard error of $\hat{\beta}_{10}$ is 0.00002812.
The null hypothesis, H_0 is $\beta_{10} = 0$, while the alternative hypothesis, H_a is $\beta_{10} \neq 0$. Since the p-value is $<$ 2e-16, which is smaller than 0.05, there is very strong evidence against the null hypothesis, H_0.

The $\hat{\beta}_{11}$ is the estimated coefficient of the total transaction count of the estimated logistic regression model, which is 0.1066
The standard error of $\hat{\beta}_{11}$ is 0.004627.

The null hypothesis, H_0 is $\beta_{11} = 0$, while the alternative hypothesis, H_a is $\beta_{11} \neq 0$. Since the p-value is < 2e-16, which is smaller than 0.05, there is very strong evidence against the null hypothesis, H_0.

The $\hat{\beta}_{12}$ is the estimated coefficient of the change in transaction count of the estimated logistic regression model, which is 3.226.
The standard error of $\hat{\beta}_{12}$ is 0.2382.
The null hypothesis, H_0 is $\beta_{12} = 0$, while the alternative hypothesis, H_a is $\beta_{12} \neq 0$. Since the p-value is < 2e-16, which is smaller than 0.05, there is very strong evidence against the null hypothesis, H_0.

The $\hat{\beta}_{13}$ is the estimated coefficient of the change in transaction count of the estimated logistic regression model, which is 2.609.
The standard error of $\hat{\beta}_{13}$ is 0.2238
The null hypothesis, H_0 is $\beta_{13} = 0$, while the alternative hypothesis, H_a is $\beta_{13} \neq 0$. Since the p-value is < 2e-16, which is smaller than 0.05, there is very strong evidence against the null hypothesis, H_0.

Table 4 – Confusion Matrix

n=1415	Reference:0	Reference:1
Prediction:0	116	46
Prediction:1	106	1147

Table 4 is the confusion matrix for the testing data set. The true positive is 1147, which means that there are 1147 customers in the testing set who are still existing and also predict existing customers. The true negative is 116, which means that customers predict as attrited customers and stop the business. The false positive is 106, and the false negative is 46, which are the wrong predictions. The accuracy of this logistic regression model is 0.8926. The customer churn rate predicted by the model is 11.45%, while the actual customer churn rate for the testing set is 15.69%.

Discussion

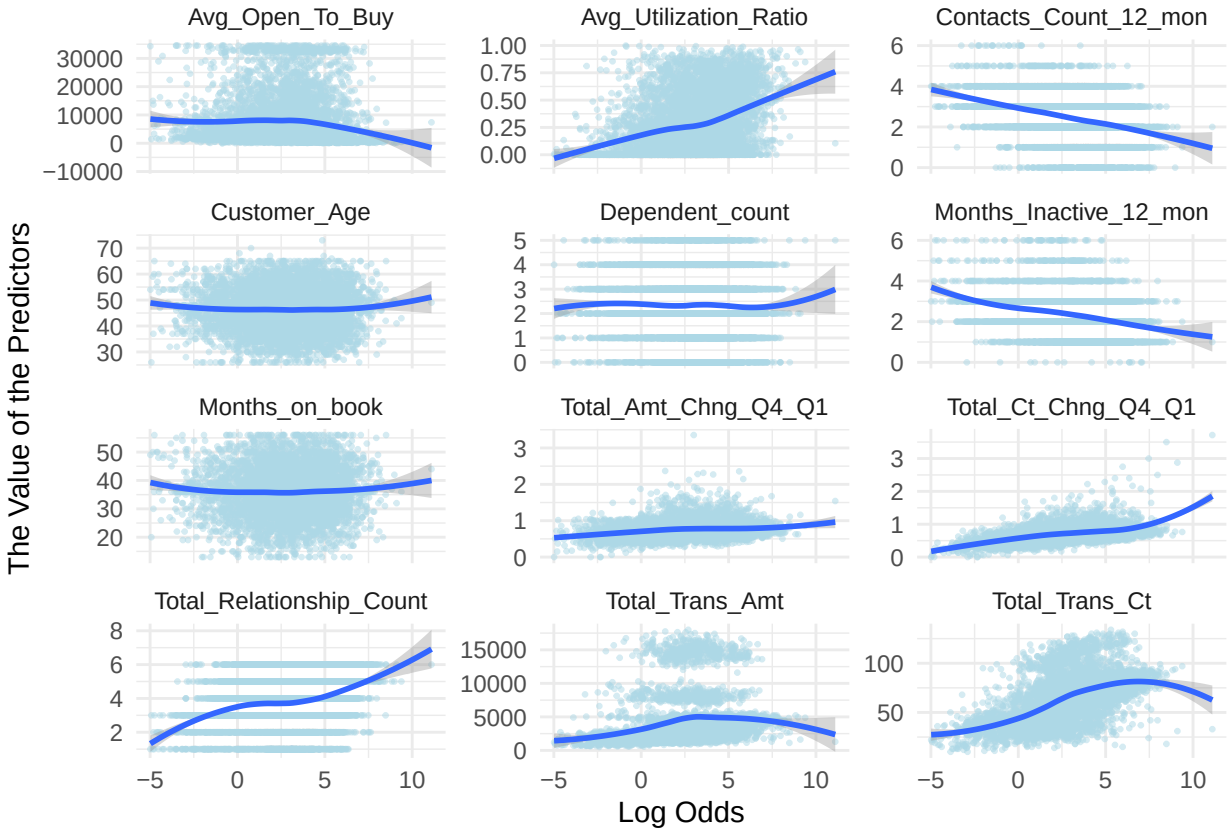
Summary

During the *Introduction*, the importance of customer and knowing the customer churn rate is stated. Using the training data set, a bar plot and a table of the summary are built for the response variables, *Attrition_Flag*. I also build 12 histograms and 5 bar plots for all other predictors to overview the data. Using the stepwise regression with BIC, the specific predictors of logistic regression model based on the training data set is selected, which are *Gender*, *Dependent_count*, *Marital_Status*, *Months_on_book*, *Total_Relationship_Count*, *Months_Inactive_12_mon*, *Contacts_Count_12_mon*, *Avg_Open_To_Buy*, *Total_Trans_Amt*, *Total_Trans_Ct*, *Total_Ct_Chng_Q4_Q1*, and *Avg_Utilization_Ratio*. The confusion matrix is created for the testing data set to predict whether customers will leave or not and evaluate the accuracy of the model.

Conclusions

The estimated proportion of customer churn rate is 11.45% for the testing data set, which means that only a few customers are likely to stop the business with the bank. The accuracy is 89.26%, which is relatively high. However, there is still some difference between the reference customer churn rate, 15.69% and the estimated churn rate, 11.45%. This shows that more customers tend to leave.

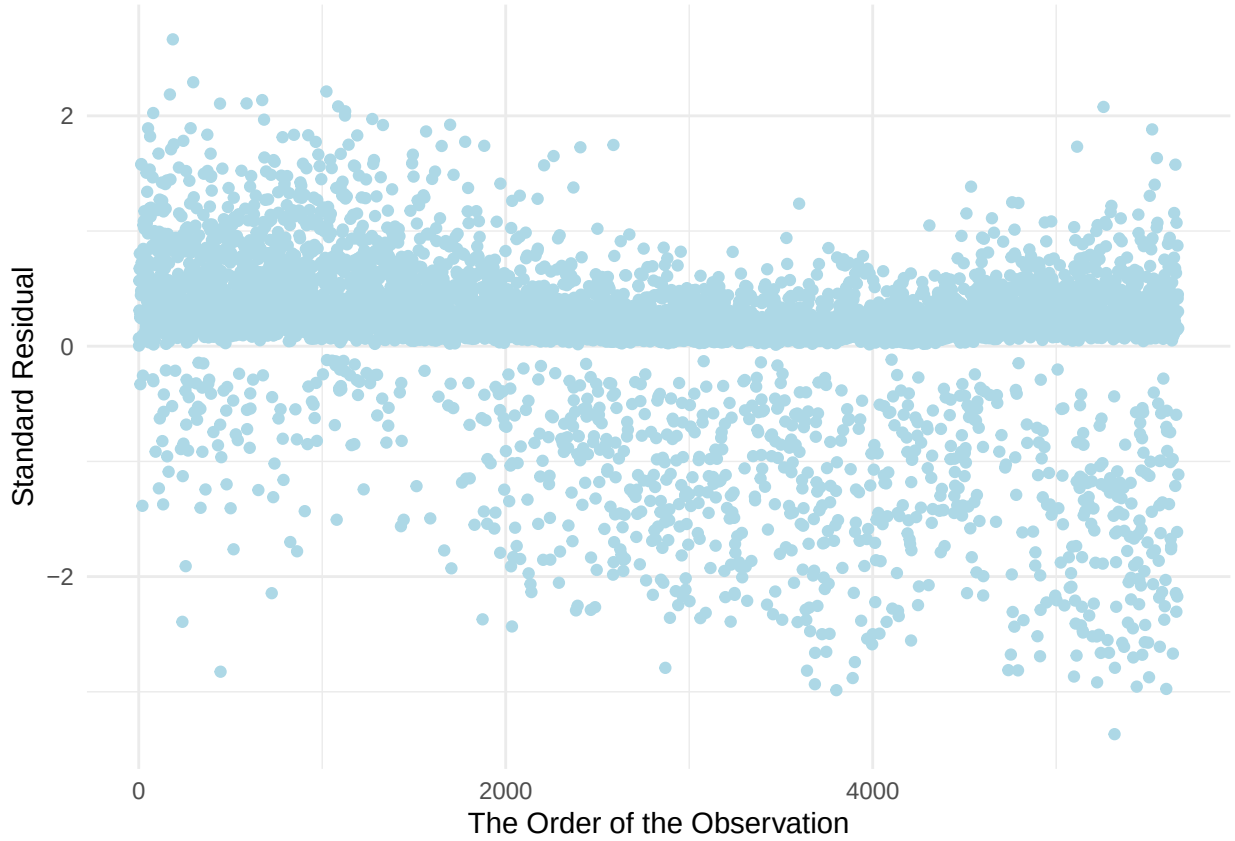
Figure 4 – Scatterplots for all numeric variable



The first assumption of the logistic model is linearity. As we can see from *Figure 4*, the linearity assumption for most of the numeric variables is valid, except for *Total_Trans_Ct* and *Total_Relationship_Count*. This indicates that some kinds of transformation may need for these two variables.

The second assumption that the response variable is binary is valid since the *Attrition_Flag* only contains outcomes, *Attrited Customer* and *Existing Customer*.

Figure 5 – Scatterplot between Standard Residual and Order of Observation



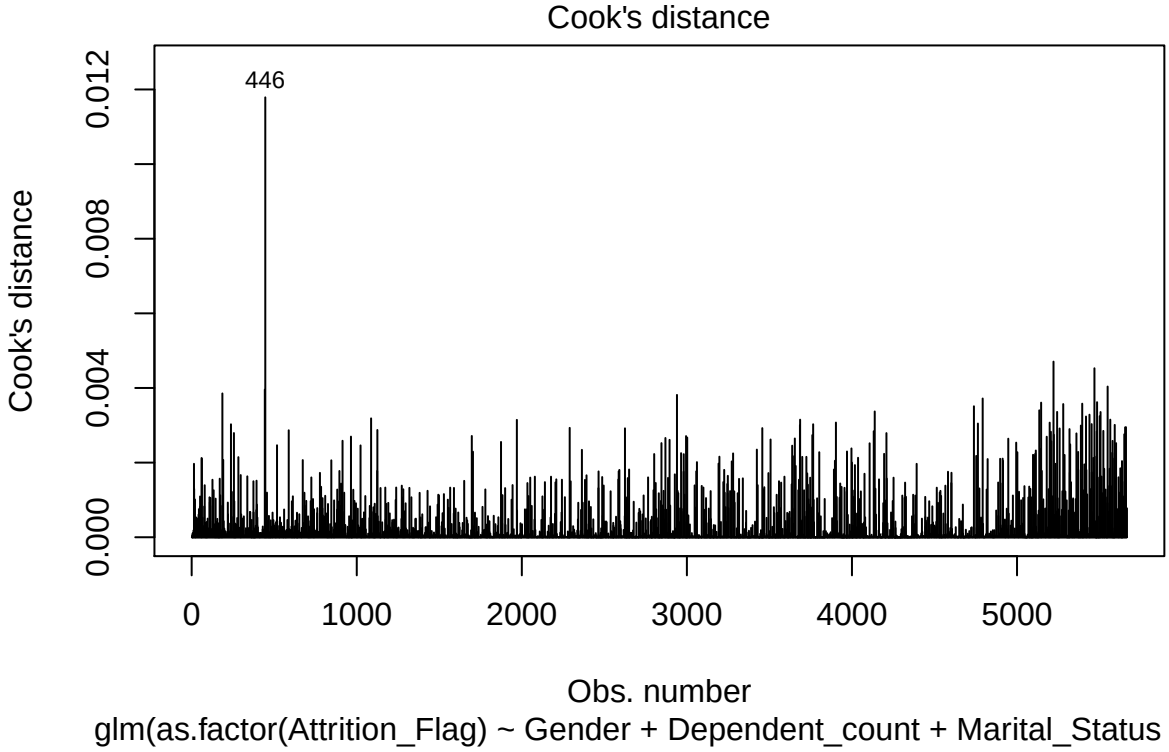
The third assumption that the observations are independent is also valid, since there is as random pattern showing in the *Figure 5*.

Table 5 – Variance Inflation Factor for all the predictors

	GVIF	Df	GVIF^(1/(2*Df))
Gender	1.322483	1	1.149993
Dependent_count	1.032288	1	1.016016
Marital_Status	1.080158	2	1.019464
Months_on_book	1.041413	1	1.020496
Total_Relationship_Count	1.184337	1	1.088272
Months_Inactive_12_mon	1.055146	1	1.027203
Contacts_Count_12_mon	1.025315	1	1.012579
Avg_Open_To_Buy	1.546600	1	1.243624
Total_Trans_Amt	4.250805	1	2.061748
Total_Trans_Ct	4.431203	1	2.105042
Total_Ct_Chng_Q4_Q1	1.062547	1	1.030799
Avg_Utilization_Ratio	1.274379	1	1.128884

The forth assumption that there is no multicollinearity among explanatory variables is also valid, since all the GVIF^(1/(2Df)), which is equivalent to VIF are smaller than 5 showing in the *Table 5*.

Figure 6 – Cook's distance for all the observations



The fifth assumption is that there are no extreme outliers. This assumption is invalid that there is an outlier, point 446 showing in the *Figure 6*.

The last assumption is that the sample size is sufficiently large. This assumption is valid, since the sample size of training data set is 5666.

Weakness and Next Step

For the weakness, when making the confusion matrix, the probability of customers who are likely to exist is first calculated for the testing set. If the probability is larger than 0.5, it would be recorded as 1 in the *pred_test*, leading to a difference in the customer churn rate between the reference and prediction. If calculating the average of the actual probability of customer churn rate, which is 15.61%, this is much closer to the reference customer churn rate, 15.69%. Also, the customer churn rate of this data set is relatively low, which indicates that this data set may not be that suitable for predicting the customer churn rate. What is more, some assumptions are violated.

For the next step, since there is an outlier showing in *Figure 6*, I would like to replace it with the mean value. I would also like to try other methods such as decision tree, "a non-parametric supervised learning method used for classification and regression" (19), random forest, "a classification algorithm consisting of many decisions trees" (20) and eXtreme Gradient Boost (XGBoost), "one of the implementations of Gradient Boosting concept" (21). Then, compare all the models using the ROC curve and confusion matrix to find a better model. Then, I would like to do a post-hoc analysis to determine how I could improve the model. For example, the questionnaire can be given for the customers to ask why they choose to leave or stay. Then, use this information to improve the model or reduce the customer churn rate.

Reference

1. Customer Churn: Definition, Rate, Calculation, Analysis and Prediction. ?QuestionPro. https://www.questionpro.com/blog/customer-churn/ (2020)
2. How to Calculate (And Lower!) Your Customer Churn Rate. WordStream. https://www.wordstream.com/blog/ws/2014/05/12/customer-churn#:~:text=Many%20retail%20banks%20have%20churn%20rates%20of%20between%2020%2D25%25 (2020)
3. Customer. Investopedia. https://www.investopedia.com/terms/c/customer.asp (2020)
4. Verbeke W, Martens D, Mues C, Baesens B (2011), Building comprehensible customer churn prediction models with advanced rule induction techniques. Expert Syst Appl. 38:2354-2364.
5. Nie G, Rowe W, Zhang L, Tian Y, Shi Y (2011), Credit card churn forecasting by logistic regression and decision tree. Expert Syst Appl. 38:15273-15285.
6. Caetano, S. (2020). Introduction to Logistic Regression. Lecture.
7. The 6 Assumptions of Logistic Regression (With Examples). Statology. https://www.statology.org/assumptions-of-logistic-regression/ (2020)
8. What Is the Credit Utilization Ration?. the balance. https://www.thebalance.com/what-is-a-good-credit-utilization-ratio-960548#:~:text=Generally%2C%20an%20ideal%20credit%20utilization,balances%20rise%20above%20that%20threshold. (2020)
9. Quarter Over Quarter (Q/Q). Investopedia. https://www.investopedia.com/terms/q/quarter-over-quarter.asp. (2020)
10. Credit Card customers. kaggle. https://www.kaggle.com/sakshigoyal7/credit-card-customers (2020)
11. Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
12. car. RDocumentation. https://www.rdocumentation.org/packages/car/versions/3.0-10 (2019)
13. caret. RDocumentation. https://www.rdocumentation.org/packages/caret/versions/6.0-86 (2020)
14. dplyr. RDocumentation. https://www.rdocumentation.org/packages/dplyr/versions/0.7.8 (2018)
15. ggpubr. RDocumentation. https://www.rdocumentation.org/packages/ggpubr/versions/0.4.0 (2020)
16. Stepwise Regression. Investopedia. https://www.investopedia.com/terms/s/stepwise-regression.asp (2020)
17. AIC VS. BIC. PennState. https://www.methodology.psu.edu/resources/AIC-vs-BIC/#:~:text=BIC%20is%20an%20estimate%20of,various%20assumptions%20and%20asymptotic%20approximations. (2020)
18. Logistic Regression Asuumptions and Diagnostics in R. STHDA. http://www.sthda.com/english/articles/36-classification-methods-essentials/148-logistic-regression-assumptions-and-diagnostics-in-r/ (2018)
19. Decision Trees. Scikit learn. https://scikit-learn.org/stable/modules/tree.html (2020)
20. Understanding Random Forest. towards data science. https://towardsdatascience.com/understanding-random-forest-58381e0602d2 (2019)
21. Introduction to Extreme Gradient Boosting. Medium. https://blog.exploratory.io/introduction-to-extreme-gradient-boosting-in-exploratory-7bbec554ac7 (2017)
22. broom. RDocumentation. https://www.rdocumentation.org/packages/broom/versions/0.7.3 (2020)

Appendix

GitHub link: https://github.com/Yingren-Luo/STA304-Final