# Analyzing Credit Card Attrition

MS in Business Analytics and Project Management, University of Connecticut

OPIM 5604 – Predictive Modeling

Instructor: Jose Cruz

December 1, 2021


**Team**

Adithya KR

Anisha Biswal

Yin Yue

Yun-Ting Yuan

## 1. Background and Problem Statement

### Background

Credit Card processing companies play a significant role in the finance and payment ecosystem. According to a report from Business Insider, the credit card is expected to remain the largest model of payment for an in-store purchase with an expected market cap of 1.82 trillion USD. Banks such as Bank of America, JP Morgan Chase, Wells Fargo, and/or financial institutions like Discover collaborate with payment processing networks like Visa, MasterCard to issue credit cards to their customers.

To obtain a personal credit card in the United States or most countries, the applicant must be 18 years or older. With the widespread use of mobile and internet banking, access to credit cards has rapidly grown. According to the ascent survey, there is a positive correlation between owning a credit card and education level. 47.3% of students get access to their first credit during or before the end of their undergraduate. Though anyone over the age of 18 can apply for a credit card, the approval and credit limit depends on various factors such as income, payment history, etc.

### Problem Statement

The large market of credit cards is predominantly dominated by millennials and Generation- Z. This poses the risk of increased credit card attrition that a lot of banks and financial institutions face today. In spite of an increase in the number of new credit cards issued, the high attrition rate ( around 15% in the United States) limits the revenue. Hence the case we propose is to analyze the credit card data to derive insights, influencing factors, and probability of a customer churn.

## 2. Data Source

The data source we will use is called BankChurners, which is obtained from the Kaggle website. It includes 23 columns in total. 17 of them are continuous, and 6 are nominal. Due to achieving the goal analysis, we decided only to use variables as follows

| ***Column Name*** | ***Descriptions*** |
|---|---|
| Attrition Flag | Customer activity variable--1 closed, 0 open |
| Customer Age | In years |
| Gender | M=male F=female |
| Dependent Count | The number of dependents |
| Education Level | Educational qualification of the account holder |
| Marital Status | Married, Single, Divorced, Unknown |
| Income Category | Annual income |
| Card Category | Type of card--blue, silver, gold, platinum |
| Month on book | Period of relationship with the bank |
| Total relationship count | Total no. of products held by the customer |
| Month inactive in 12mo | No. of months inactive in the last 12 months |
| Credit limit | Credit Limit on the Credit Card |
| Average open to buy | Open to Buy Credit Line (Average of last 12 months) |
| Total transaction amount | Last 12 months |
| Total_Amt_Chng_Q4_Q1 | Change in Transaction Amount (Q4 over Q1) |
| Total_Ct_Chng_Q4_Q1 | Change in Transaction Count (Q4 over Q1) |
| Average utilization rate | Average Card Utilization Ratio |

The data is clear enough and ready for use, so there is no more work to do after selecting the variables we will use.

**3. Goals of Our Analysis**

I. Look for influential factors that lead to customer churn.

II. Show behavior of credit limit with different categories of variables.

III. Predict the relationship of the customer with the bank based on gender, income in a year.

IV. Look for factors that influence transaction patterns.

**4. Data Analysis Tools We Plan to Use**

We will use JMP as our data analysis tool and Tableau for additional data visualization. First, we will get our sample through kaggle.com. Then use JMP to explore, modify, model, and assess our dataset.

**5. Data Products Our Project Will Produce**

**Performance Analysis**

We will create the five models we learned in class that are neutral model, logistic regression, decision tree, bootstrap forest, and boosted tree. And then, we will compare the performances of these models based on the validation data.

**Data Visualization**

In order to find some business insights, we will use distributions for model parameters, tabulations, charts, graphs to visualize the data.

**Business Insights**

We decided to figure out the affecting factors which result in the customers being churners. And we will give some recommendations and strategies to the bank to avoid customer churn.

**6. Analyzing variable type from the dataset**

The dataset has a total of 21 columns that are used to classify the likelihood of a customer leaving the credit card services of a bank. Out of the 21 columns, 2 columns are ordinal, 3 are nominal, and the remaining 16 are continuous variables. 16.7% of 10,127 customers discontinue their credit cards, causing a churn. Since the percentage of attrited customers is small, it is important to stratify the sample.

Though Income_category can be converted to a continuous variable by averaging the salary in the bracket or by generating a random number in the salary range, we decided to retain it as a nominal variable to bucket customers based on salary and determine the impact of salary class on credit card usage.

**7. Determining columns for building the model**

Upon careful analysis of the dataset, we realized that column Clientnum does not contribute to building an effective classification model and hence is excluded from the analysis. Furthermore, we had decided to add a column *start_year* that will capture the year when the account/ credit card was first activated. The start_year column is populated based on the assumption that the dataset was curated in the year December 2020. This will be used to determine the time period after which customers tend to churn for the given dataset.

| **Column** | **Reason for exclusion** |
|---|---|
| CLIENTNUM | Serves as a unique identifier of a customer record and does not impact the model used to classify. |

| Column | Reason for inclusion |
|---|---|
| customer_status | A nominal data type column is derived from the Attrition flag to determine if a customer is still existing or churned. |

**8. Outlier Analysis**

We conducted outliers analysis by exploring outliers through Huber, Cauchy, and quantile methods. In these methods, the outlier numbers are more than 1200, i.e., more than 10% of the data set. So it is not advisable to exclude a significant number of records from the data set to avoid any exclusion of useful records. So we decided to transform the outliers.

Through distribution, we checked that a few variables have less than 300 outliers, which is less than 3% of the data set. So we did not transform those outliers. In total, we transformed 8 variables with more than 350 outliers. Those variables are 'Total_Trans_Amt', 'Avg_Open_To_Buy' , 'Credit_Limit', 'Contacts_Count_12_mon' , 'Months_Inactive_12_mon' ,'Months_On_Book', 'Total_Amt_Chng_Q4_Q1', 'Total_Ct_Chng_Q4_Q1' .

**9. Missing Value Analysis**

We conducted the missing value pattern and found that the number of missing value columns is 0. Therefore, it means that there is no missing value in this dataset.

**10. Data Reduction**

We plan to use 20 columns to build our model, so the dataset is large. Thus, we decided to use multivariate methods to check the correlations between these columns. Because the result showed that most of them have low correlation, we decided not to combine any columns and not reduce data.

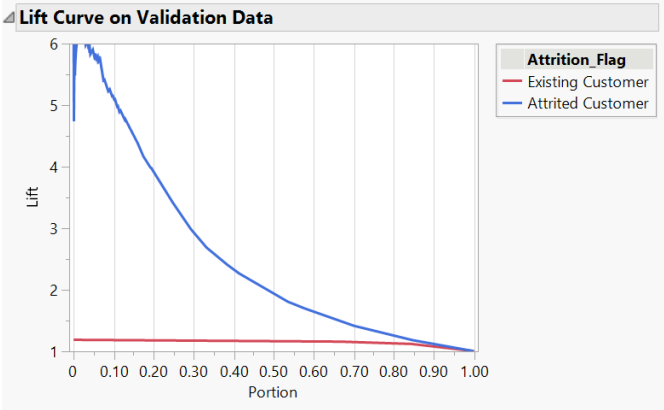
**11. Model Building**

We created five models in total. They are logistic regression, decision tree, booted tree, bootstrap forest, and neural network, respectively. Before we created models, we split 60% of data into a training set and 40% of data into a validation set. Our target variable is the attrition flag since we want to create a model to predict whether the customer will become a churning customer or not. And the other variables are the potential factors to affect the result. The first model we built is logistic regression. We used forward stepwise as direction, and the effect summary showed that only 10 columns contributed to this model. They are shown as follows

**Effect Summary**

| Source | LogWorth | PValue |
|---|---|---|
| Total_Trans_Ct | 197.320 | 0.00000 |
| Total_Revolving_Bal | 68.220 | 0.00000 |
| Total_Trans_Amt | 56.566 | 0.00000 |
| Total_Ct_Chng_Q4_Q1 | 48.989 | 0.00000 |
| Total_Relationship_Count | 41.411 | 0.00000 |
| Contacts_Count_12_mon | 25.111 | 0.00000 |
| Months_Inactive_12_mon | 22.526 | 0.00000 |
| Gender | 9.829 | 0.00000 |
| Marital_Status{Married-Divorced&Single&Unknown} | 9.238 | 0.00000 |
| Dependent_count | 3.487 | 0.00033 |

The lift curve for the attrited customers in the validation set doesn't perform well since the curve doesn't have a flat zone at the beginning of the curve at all, which means this model is not reliable enough. And it just reaches 6 at 0.1 portions of data. As for the misclassification rate in the validation set, because we want to avoid considering churning customers as existing customers, we pay attention to the false-negative rate of all models. The logistic regression has a high false-negative rate which is 0.416.




Validation

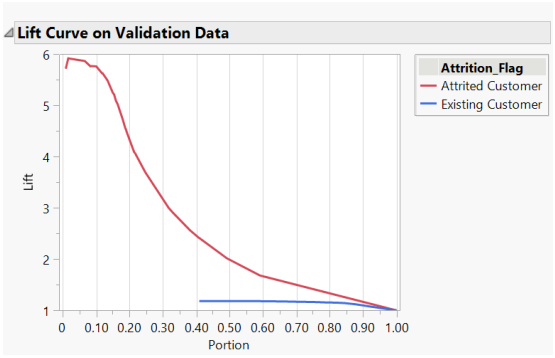
| Actual Attrition_Flag | Predicted Count Existing Customer | Predicted Count Attrited Customer |
|---|---|---|
| Existing Customer | 3300 | 117 |
| Attrited Customer | 264 | 370 |

| Actual Attrition_Flag | Predicted Rate Existing Customer | Predicted Rate Attrited Customer |
|---|---|---|
| Existing Customer | 0.966 | 0.034 |
| Attrited Customer | 0.416 | 0.584 |

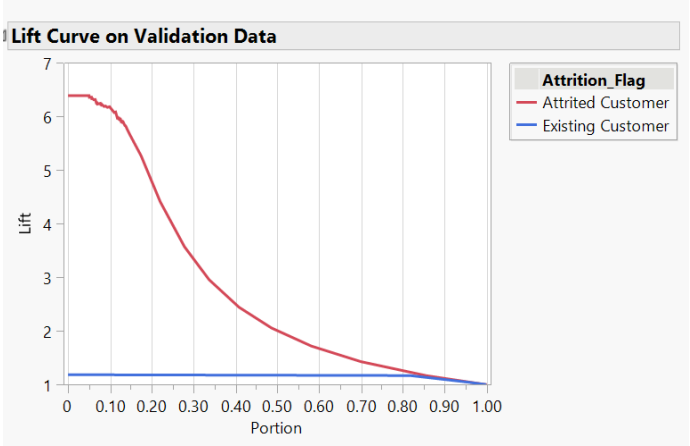
The second model we created is the decision tree. The lift curve of attrited customers in the validation set is smoother than that of the logistic regression model. But it doesn't reach 6 at 0.1 portions, either. The top five column contributions are all related to the total transaction amount and changing transaction.

**Column Contributions**

| Term | Number of Splits | G^2 | Portion |
|---|---|---|---|
| Total_Trans_Ct | 5 | 1664.44089 | 0.4464 |
| Total_Revolving_Bal | 3 | 661.82676 | 0.1775 |
| Total_Trans_Amt | 8 | 624.337324 | 0.1675 |
| Total_Relationship_Count | 2 | 303.472686 | 0.0814 |
| Total_Ct_Chng_Q4_Q1 | 5 | 290.425311 | 0.0779 |
| Customer_Age | 2 | 74.9398116 | 0.0201 |
| Total_Amt_Chng_Q4_Q1 | 1 | 33.4906849 | 0.0090 |
| Marital_Status | 1 | 26.2686926 | 0.0070 |
| Months_Inactive_12_mon | 1 | 25.8720357 | 0.0069 |
| Avg_Open_To_Buy | 1 | 23.2419078 | 0.0062 |
| Gender | 0 | 0 | 0.0000 |
| Dependent_count | 0 | 0 | 0.0000 |
| Education_Level | 0 | 0 | 0.0000 |
| Income_Category | 0 | 0 | 0.0000 |
| Card_Category | 0 | 0 | 0.0000 |
| Months_on_book | 0 | 0 | 0.0000 |
| Contacts_Count_12_mon | 0 | 0 | 0.0000 |
| Credit_Limit | 0 | 0 | 0.0000 |
| Avg_Utilization_Ratio | 0 | 0 | 0.0000 |




The third model is bootstrap forest. The lift curve of attrited customers in the validation set is very smooth and has a longer flat zone at the beginning of the curve, and it is at over 6 at 0.1 portions. The misclassification rate of the false-negative rate is 0.18.

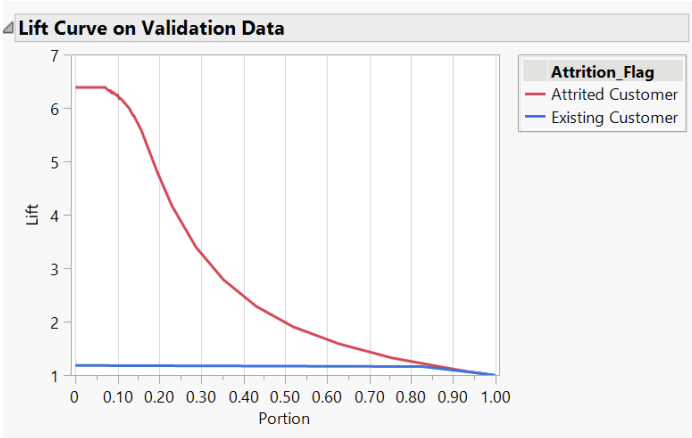



Validation

| Actual Attrition_Flag | Predicted Count Attrited Customer | Existing Customer |
|---|---|---|
| Attrited Customer | 520 | 114 |
| Existing Customer | 64 | 3353 |

| Actual Attrition_Flag | Predicted Rate Attrited Customer | Existing Customer |
|---|---|---|
| Attrited Customer | 0.820 | 0.180 |
| Existing Customer | 0.019 | 0.981 |

The fourth model is the boosted tree which has a similar lift curve to the bootstrap forest. But it has a higher false-negative rate which is 0.232.

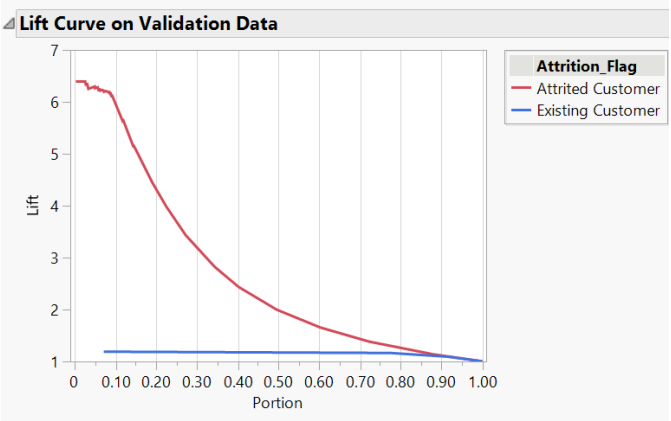



Validation

| Actual Attrition_Flag | Predicted Count Attrited Customer | Existing Customer |
|---|---|---|
| Attrited Customer | 487 | 147 |
| Existing Customer | 33 | 3384 |

| Actual Attrition_Flag | Predicted Rate Attrited Customer | Existing Customer |
|---|---|---|
| Attrited Customer | 0.768 | 0.232 |
| Existing Customer | 0.010 | 0.990 |

The last model we built is the neural network. The activation function we used is tanH since we want to create a classification model. And we set three nodes in the first layer and two nodes in the second layer. We kept default values for all other things. Its lift curve of the attrited customers reaches over 6 at 0.1 portions, but it isn't smooth and flat in the beginning. The false-negative rate is 0.292.




Confusion Rates

| Actual Attrition_Flag | Predicted Rate Attrited Customer | Predicted Rate Existing Customer |
|---|---|---|
| Attrited Customer | 0.708 | 0.292 |
| Existing Customer | 0.026 | 0.974 |

As so far, based on the performance of the lift curve and false negative rate, the best model is bootstrap forest since it has the lowest false-negative rate and well-performed lift curve. We decided to use model comparison to verify our speculation. Because we should compare the models in the validation set, the results shown below are only the model comparison in the validation set. To compare the performance of models, we should take into account generalized RSquare, RASE, and misclassification rate. The best model should have the highest RSquare, the lowest RASE, and the lowest misclassification rate. As we can see from the following table, the bootstrap forest has the highest RSquare, which is 0.8188, the lowest RASE, which is 0.1773, and the lowest misclassification rate, which is 0.0410. Therefore, we chose bootstrap forest as the best model to predict the possible churning customers for the bank.

**Model Comparison**

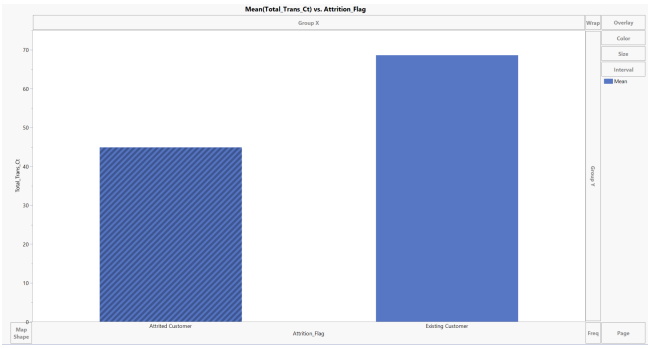
**Predictors**

**Measures of Fit for Attrition_Flag**

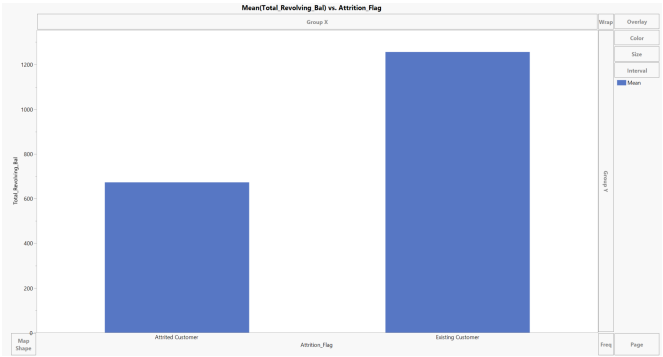
| Creator | .2 .4 .6 .8 | Entropy RSquare | Generalized RSquare | Mean -Log p | RASE | Mean Abs Dev | Misclassification Rate | N |
|---|---|---|---|---|---|---|---|---|
| Fit Nominal Logistic | | 0.4675 | 0.5748 | 0.231 | 0.2619 | 0.1393 | 0.0941 | 4051 |
| Partition | | 0.6122 | 0.7104 | 0.1682 | 0.2192 | 0.0880 | 0.0617 | 4051 |
| Bootstrap Forest | | 0.7425 | 0.8188 | 0.1117 | 0.1773 | 0.0819 | 0.0410 | 4051 |
| Boosted Tree | | 0.7239 | 0.8041 | 0.1198 | 0.1803 | 0.0817 | 0.0444 | 4051 |
| Neural Model NTanH(3)NTanH2(2) | | 0.6706 | 0.7605 | 0.1429 | 0.2033 | 0.0818 | 0.0575 | 4051 |

## 12. Visualization

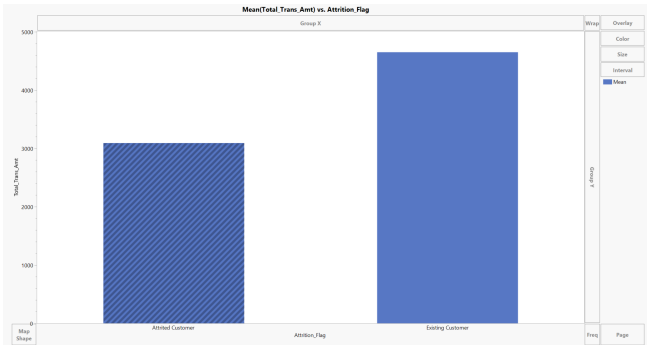
For the mean of total transaction count, the attrited customer group is lower than the existing group.



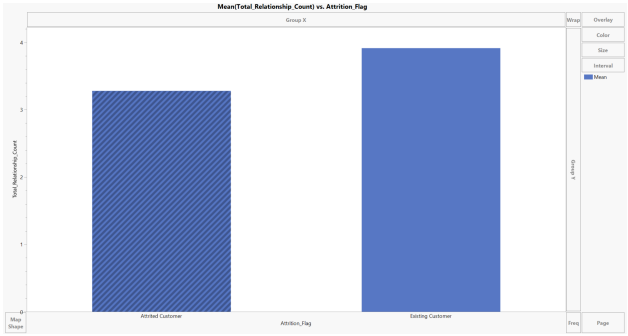

For the mean of total revolving balance on the credit card, the attrited customer group is lower than the existing group.



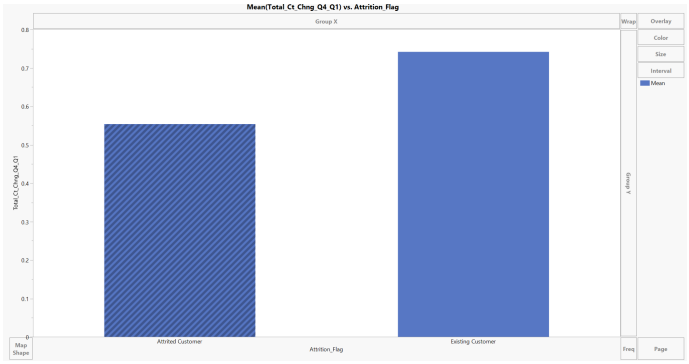

For the mean of the total transaction amount in the last 12 months, the attrited customer group is lower than the existing group.



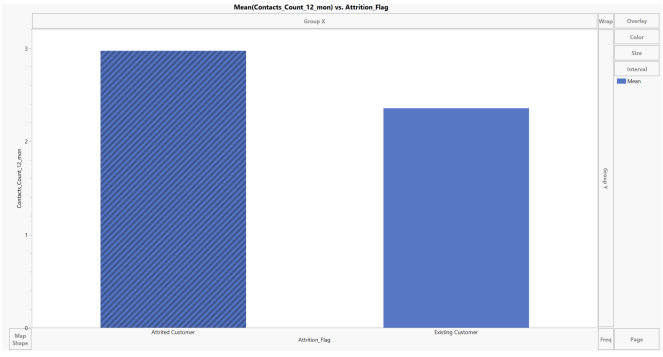
For the mean of the total number of products held by the customer, the attrited customer group is lower than the existing group.



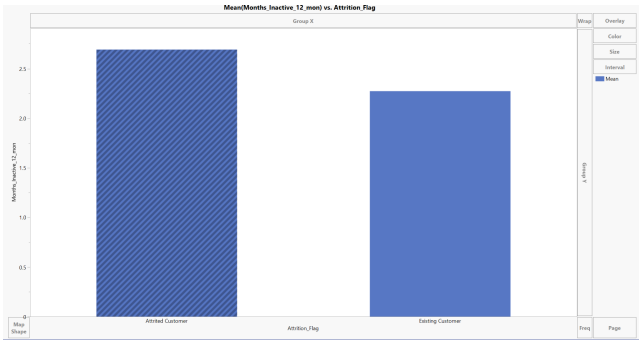

For the mean of change in transaction count (Q4 over Q1), the attrited customer group is lower than the existing group.



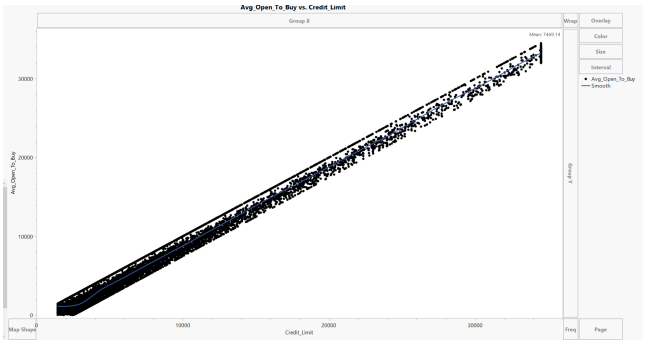

For the mean of the number of contacts in the last 12 months, the attrited customer group is higher than the existing group.



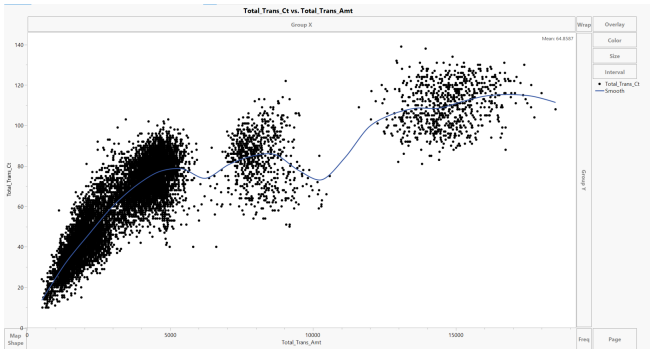
For the mean of the number of months inactive in the last 12 months, the attrited customer group is higher than the existing group.



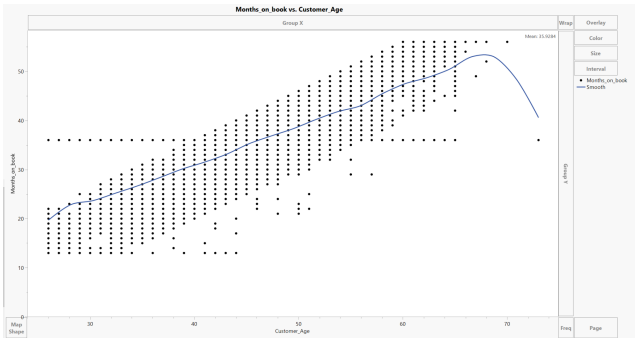

Credit Limit on the Credit Card and Open to Buy Credit Line (Average of last 12 months) have a strong relationship.




Total Transaction Amount (Last 12 months) and Total Transaction Count (Last 12 months) have a strong relationship.



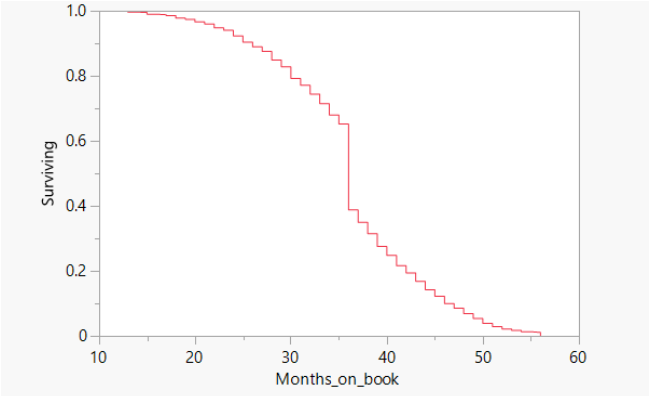

The customer's Age and Period of relationship with the bank have a strong relationship.




**13. Predicting the duration of the relationship with Survival Analytics**

For a financial institution, it is important to understand the period after which a customer holding a credit card is expected to churn ( stop using the credit card service). This will help them be proactive in their approach to strategize their marketing campaigns to retain their customers.

Analyzing the months on book vs. customer status indicates that the median duration for customers to use the credit card before churning is 36 months, and the 3rd quantile is 40 months. There is a sharp 30% decrease in survival when the number of months on book nears 36 months.




Time to event: Months_on_book
Censored by Customer Status
Censor Code 1

**Summary**

| Group | Number failed | Number censored | Mean | Std Error |
|---|---|---|---|---|
| Combined | 1627 | 0 | 36.1782 | 0.19329 |

**Quantiles**

| Group | Median Time | Lower 95% | Upper 95% | 25% Failures | 75% Failures |
|---|---|---|---|---|---|
| Combined | 36 | . | . | 32 | 40 |

Number of dependents, number of contact counts in the last 12 months from the customer, Credit Limit, Total Transaction Amount contribute to the risk. That is, a one-unit change in any of these predictor variables impacts the customer's association with the credit card.

**Risk Ratios**

**Unit Risk Ratios**

Per unit change in regressor

| Term | Risk Ratio | Lower 95% | Upper 95% | Reciprocal |
|---|---|---|---|---|
| Customer_Age | 0.874156 | 0.867248 | 0.881119 | 1.1439602 |
| Dependent_count | 1.149069 | 1.104316 | 1.195635 | 0.8702699 |
| Total_Relationship_Count | 0.823755 | 0.796186 | 0.852279 | 1.2139527 |
| Months_Inactive_12_mon | 1.1402 | 1.088411 | 1.194453 | 0.8770393 |
| Contacts_Count_12_mon | 1.242548 | 1.187159 | 1.300521 | 0.804798 |
| Credit_Limit | 1.000575 | 1.000566 | 1.000585 | 0.999425 |
| Total_Revolving_Bal | 0.998958 | 0.998855 | 0.999061 | 1.0010434 |
| Avg_Open_To_Buy | 0.999415 | 0 | 0 | 1.0005854 |
| Total_Amt_Chng_Q4_Q1 | 0.824744 | 0.657438 | 1.034625 | 1.2124981 |
| Total_Trans_Amt | 1.000275 | 1.000243 | 1.000307 | 0.9997251 |
| Total_Trans_Ct | 0.93766 | 0.933458 | 0.941882 | 1.0664841 |
| Total_Ct_Chng_Q4_Q1 | 0.361591 | 0.287714 | 0.454438 | 2.7655536 |
| Avg_Utilization_Ratio | 1.080638 | 0.758128 | 1.540344 | 0.9253795 |

Similarly, it is observed from the dataset that the risk of credit card churns increases by 1.66 times when the gender changes from F to M.

**14. Recommendations**

As the Bootstrap forest is the best model for predictions among others, we analyzed the model deeply and found that focusing on five factors will increase the customers and revenue for the bank, that is 'Total transaction amount.' 'Total Transaction Count', 'Total_Revolving_Bal', 'Total_ct_Chng_Q4_Q1', 'Total_Relationship_Count' . So based on these factors the decisions should be made to avoid customers getting churned.

After going through all the models and visualizations, a few suggestions can be made for the bank as below:

- The credit limit of the attrited customer is lower than the existing customer, even though the mean income of the attrited customer is higher than the existing customer. One of

the reasons for the customer to churn might be the inability to use cards after a certain limit, even if the customer is capable of spending that much amount. So the bank should increase the credit limit for them.

- The dependent count for the attrited customer is higher than the existing customer. The bank should come up with friendly policies for the dependent of the customers, which will help the customer to feel secure about their finances.
- Months of Inactiveness is more for attrited customers. So the bank should follow up frequently with them. Also, in months the attrited customers have contacted the bank more than the existing customers, which indicates that the customers are confused and have queries. So to mitigate this and to increase customer satisfaction, the bank should employ a few workers specifically to resolve customers' queries/ complaints.
- The total transaction count and the transaction amount is higher for Singles and very low for married people. To not lose the customers whose transactions are more, the bank should not be stringent with them regarding any kind of policies and to increase the transactions in married people, the bank should come up with friendly discounts on shopping.

**15. Reference**


https://www.kaggle.com/sakshigoyal7/credit-card-customers?select=BankChurners.csv

https://www.businessinsider.com/credit-card-industry

https://www.fool.com/the-ascent/research/study-when-does-average-american-get-their-first-credit-card/