STATISTICAL FOUNDATIONS OF DATA ANALYTICS
PROFESSOR: JUAN CAMILO SERPA

THE CHURNING CLIENTS CHALLENGE

A REPORT BY

ALISON LIU

The Churning Clients Challenge

PROJECT OVERVIEW

1 . Executive Summary

Since the COVID-19 pandemic began in January, the global banking industry has seen a 30% quarter on quarter decrease in market capitalization for the whole year of 2020. At the same time, the emerging entrance of Fin-Tech start-ups such as open banking and neo-banks are making the situation even worse. Banks are facing an unprecedented challenge of clients "churning up" their profitability by moving money between, switching or leaving to other competitors.

The motivation of this project is to develop a model that can identify which customers have a tendency to attrit based on their demographic and historical transaction data. This will enable banks to have a new level of personalization in service and marketing offers. Models in this project are based on the "Credit Card customers" dataset on Kaggle provided by Sakshi Goyal[1], explanations of dataset variables can be seen in Table 1.

Fostering customer loyalty is not only crucial but also truly beneficial for banks. According to Bain & Co., a 5% increase in retention rate can boost business profitability by 75%[2]. Especially for credit card services, the market expansion is limited, and the loss of one client due to the latest technology, unfriendly staff, high interest rates or location distance is hard to gain back. Hence, an accurate predictive and descriptive model for churning clients will be valuable and imperative for today's banks.

[1] https://www.statista.com/statistics/265135/market-capitalization-of-the-banking-sector-worldwide/#statisticContainer

[2] https://www.kaggle.com/sakshigoyal7/credit-card-customers/notebooks?datasetId=982921&sortBy=voteCount

In order to build a model with satisfactory performance, data describing the characteristics of a client were firstly explored by plotting and observing the distribution of the target variable along with numerical and categorical variables. Then, I explored the correlation between each variable and the target variable. Simple logistic regression was also utilized to evaluate the relationship between the target variable and each variable. After that, multiple models were developed with multiple algorithms to select the model with the best predicting power. With the final model defined, cross-validation was performed to prevent overfitting, and potential improvements were considered for the development of a future optimized model.

2 . Exploratory Data Analysis

The dataset provided contains 10,127 client information records with 23 dependent variables describing their characteristics such as age and income level. One target variable, ‘Attrition_Flag’, represents whether a customer has paused his or her service with the bank or not. It was transformed from string to binary number: '1' and '0', accordingly representing 'attrited' and 'not attrited'. Around 16% of the client observations provided have attrited. The two last columns contained the prediction results of a Naive Bayes classification model that will not contribute to the model; therefore, they were removed from the dataset. One label variable was also removed. The rest 19 independent variables were then separated into two categories: 14 numerical variables and 5 categorical variables.

2.1. EDA - Numercal Variables

Most of the numerical variables were discrete variables, therefore, I chose to visualize them by histograms. Considering that the target variable is binary, stacked histograms were used to better illustrate the proportion of observations with different target categories (see Exhibit 1). Six variables are positively skewed, meaning that most of the observations have a value smaller than the mean. Skewed data often contain outliers, which will impact the accuracy of linear regression algorithms such as GLM and LDA. But there are also other classification algorithms, such as tree-based models, that are not sensitive to variable skewness. Therefore, I decided to only include valuable skewed variables in tree-based models. If the performance of tree-based models is not ideal, I will then consider correcting skewness using techniques such as Logarithm transformation or outlier removal.

2.2. EDA - Categorical Variables

Categorical variables were visualized using stacked box plots (see Exhibit 2). For every variable, if one category has a distinct proportion of churned and unchurned clients compared to other categories of this variable, it represents an effect of this category on the target variable. In other words, if one bar has a distinct proportion of blue and yellow compared to other bars, the category represented by this bar has a strong impact, either positive or negative, on whether a client attrits or not. Visually, I cannot see one category that has a strong influence on the attrition ratio from the plots. Proportions were then calculated and compared relatively within categories (see Table 2). Three features were identified out of five variables and twenty-one categories. For example, clients who hold a doctorate degree have a relatively higher attrition ratio compared to other clients. Therefore, three binary variables were created for building models.

3 . Feature Selection

3.1. Numerical Variables Selection

The purpose of this process is to identify the most valuable variables for model development. First of all, the data was standardized, renamed, and mapped with PCA to get an overview of the variable relationships (see Exhibit 3). Standardization was performed with the same calculation method as the function StandardScaler() in Python[3]. Columns were then renamed to x_i and y for better visualization (see Table 3). From the PCA map, we can see that y, which is the target variable, is in the same direction as x1, 3, 5, 6, which means that they are highly positively correlated. Similarly, x10, 13 are in opposite directions as y, meaning that they are highly negatively correlated. While x7, 9, 14 are orthogonal to y, which means that they are rather unrelated to the target variable. However, it is noticeable that PC1 and PC2 only explain 17% and 15% variability across the dataset. So correlation map was drawn for a more accurate understanding of correlations between variables (see Exhibit 4).

We can see similar patterns in the correlation map. For example, x7 and x9 are highly positively correlated, while x9 and x14 are highly negatively correlated. On the other hand, some features displayed by the PCA map are not fully correct. For instance, x1 and x3 have little correlation with y.

[3] https://scikit-learn.org/stable/auto_examples/preprocessing/plot_scaling_importance.html

Then, simple logistic regression models were generated between every independent variable and y. P-values of different variables, together with their correlations with y were summarized and sorted from high to low to identify their impact on y (see Table 4). Finally, collinear variables were removed from the highest to the lowest. Seven significant variables remained and were therefore selected for future model development. Note that two of them are skewed, which will only be included in the development of tree-based models.

3.2. Categorical Variables Selection

Simple logistic regression was performed accordingly on the three binary variables created. All of them have significant p-values (less than 0.05), therefore, all three variables were remained. The ten selected variables were then gathered into one dataset along with the target variable, variables were renamed for better code layout (see Table 5). This dataset called 'bankChurners_selected' will be the dataset for model development.

4 . Model Selection

In this section, three types of classification techniques were utilized to build models and compared on performance. They are GLM (Generalized Linear Model), Discriminant Analysis including LDA (Linear Discriminant Analysis) and QDA (Quadratic Discriminant Analysis), and tree-based models including rpart (recursive partitioning), Random Forest and GBM (Gradient Boosting Machines). Model performances were evaluated by accuracy score.

Accuracy score is prioritized over MSE as the prior goal of this project is to correctly identify clients who have a high potential (over 50% of probability) to attribute, instead of correctly predicting a client's attrition probability. Out-of-bag error rate (error rate = 1 - accuracy) is prioritized over in-bag error rate as the relatively high number of predictors chosen in the model may result in overfitting.

Firstly, six models were generated based on the above six algorithms (see Table 6). A reminder that only 16% of the dataset provided are attrited clients (target variable = 1), so a model can easily gain a high accuracy of 84% by predicting all clients as existing clients (target variable = 0), but it is a useless model. Confusion Matrix was generated for applicable models to prevent this problem. Among the six algorithms, Random Forest and GBM showed better accuracy. From these two models, predictor importance rankings were generated (see Table 7), more models were built by removing predictors with low importance or correlation. Through trial and error, the final model was chosen based on the best performance score and predictor simplicity.

5 . Managerial Implications

The final model uses 8 variables as predictors, achieving an accuracy of 96.42% in prediction (see Table 8). The most important variables, in other words, the most influential factors on a client's attrition, are the number of products client holds from the bank, the number of months that client is inactive, client's credit limit, and client's number of contacts with the bank in the past 12 months. The number of transactions the client made in the past 12 months, and gender of the client are also relevant influencers.

Knowing the important factors, banks can work on improving retention rates more effectively. By observing the distributions of the four highly influential variables (see Exhibit 5), we can identify the range of variables that have a high proportion of attrition. First of all, clients that have less than two products with the bank have a higher tendency to leave the service, therefore, promoting suitable new products or services to clients with shallow engagement can foster the relationship. Also, clients who have been inactive for more than three months have a higher rate in attrition, which makes sense since credit card has become useless for them. Therefore, targeting these clients and get timely feedback from them is a key process. On the other hand, clients who have contacted the bank more than three times have a relatively high attrition ratio. When clients call for help or complaint, their feedback should be taken as seriously as those from inactive clients. If the problems remain unsolved, the relationship will be lost.

In addition, utilizing the predicting power of the model, clients with churning potential can be accurately identified, and banks can act on effective solutions to recover the most relationships. The solutions can be developed based on the clients' demographic, usage or psychological features. For example, if many 'churning' clients are in one city or neighbourhood, the bank can choose to do an advertising campaign, open more branches, or improve staff training based on local market research. If a large number of 'churning' clients are students, a cooperative offer with coffee stores like Tim Hortons or Starbucks can widely incentivize their credit card usage. If clients with a long relationship with the bank are 'churning' more and more, providing them with higher standing such as higher credit limit, bonus points for discounts, or card renewal with various benefits can increase their sense of belonging.

Overall, this project developed a model to analyze the influential factors in a credit card client's attrition, and to predict clients that have a high probability of attrition. Banks have one of the most large, accurate and detailed datasets among all businesses, covering their clients' demographic characteristics, financial condition, and spending habits. The proper exploration and utilization of the information provided by this data can bring remarkable value.

Appendix

Exhibit 1: Data Exploration Example - Numerical

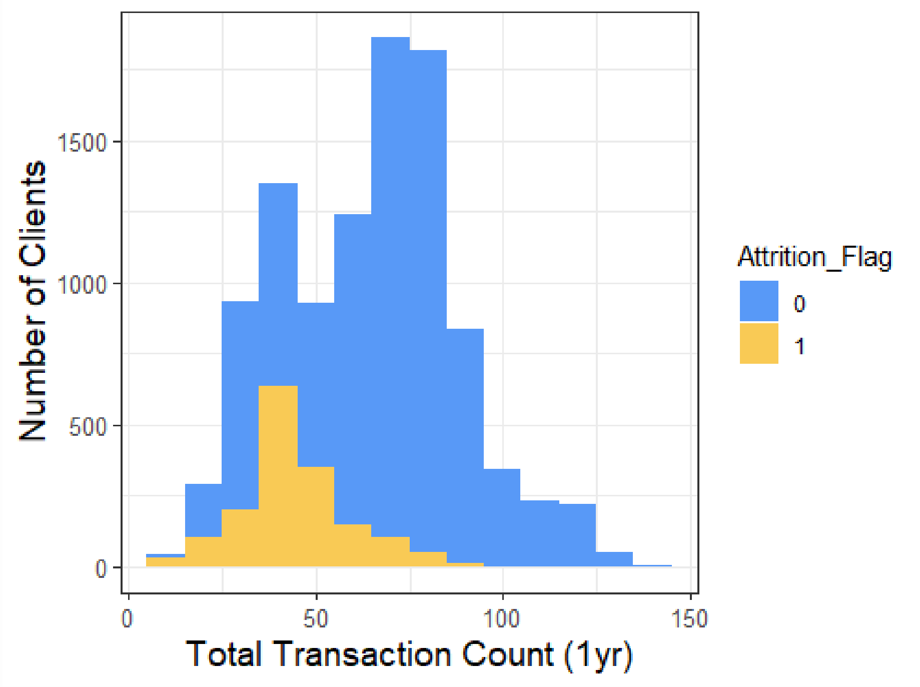
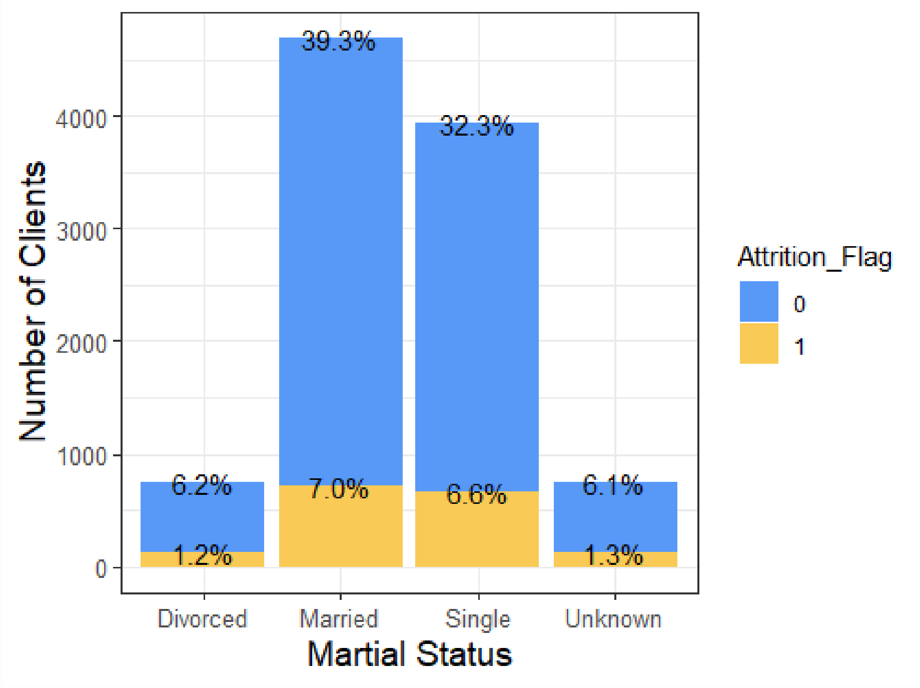


Exhibit 2: Data Exploration Example - Categorial



Appendix

Exhibit 3: PCA Map - Y and Numerical Variables

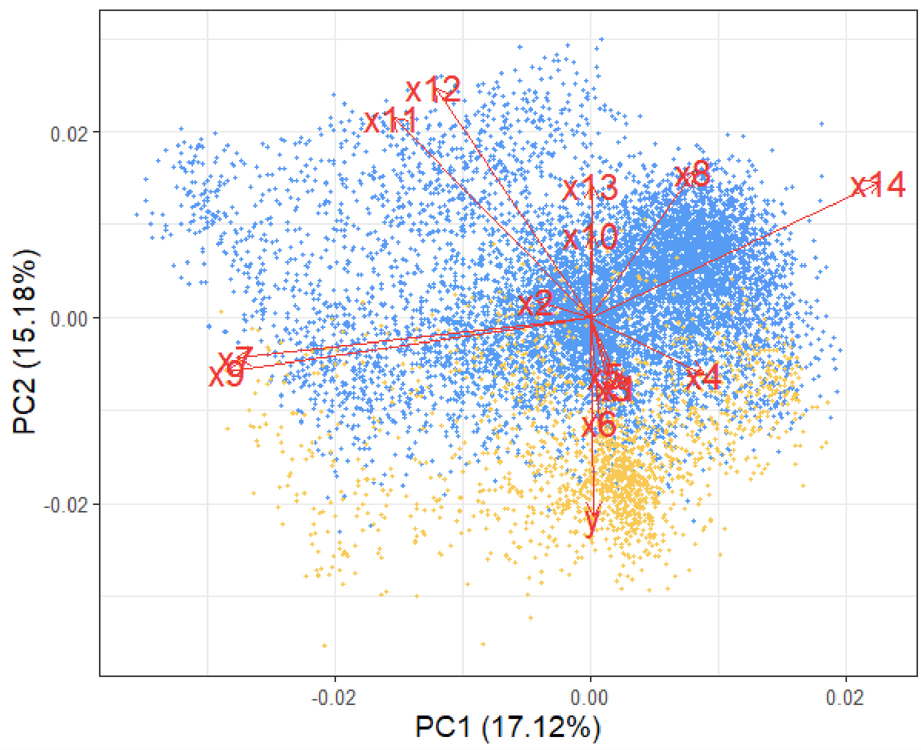
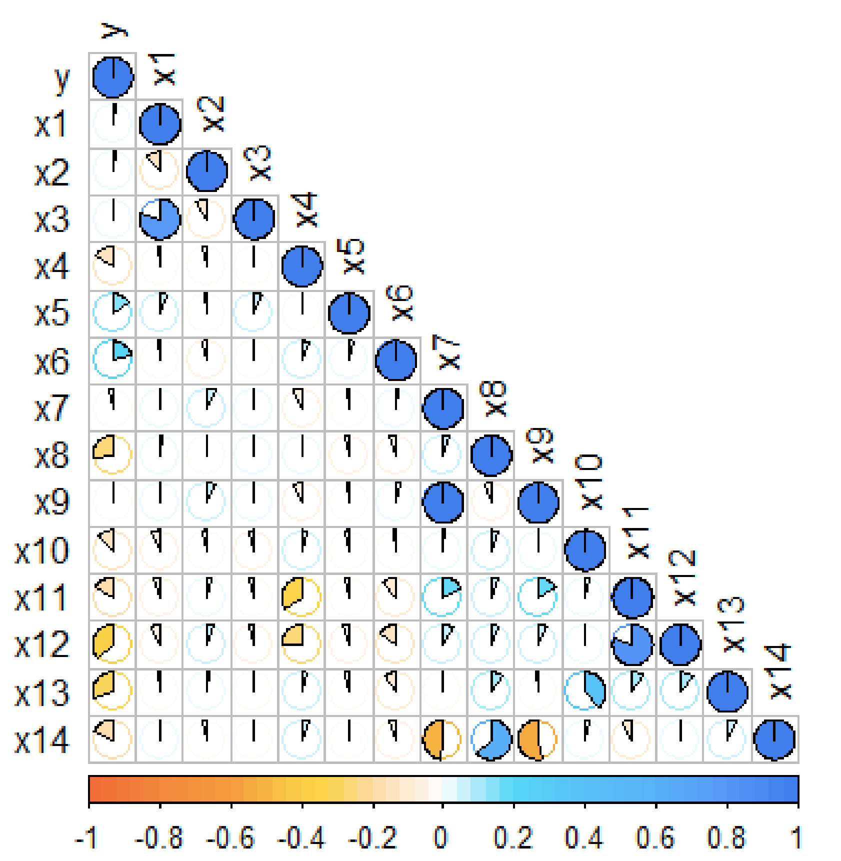
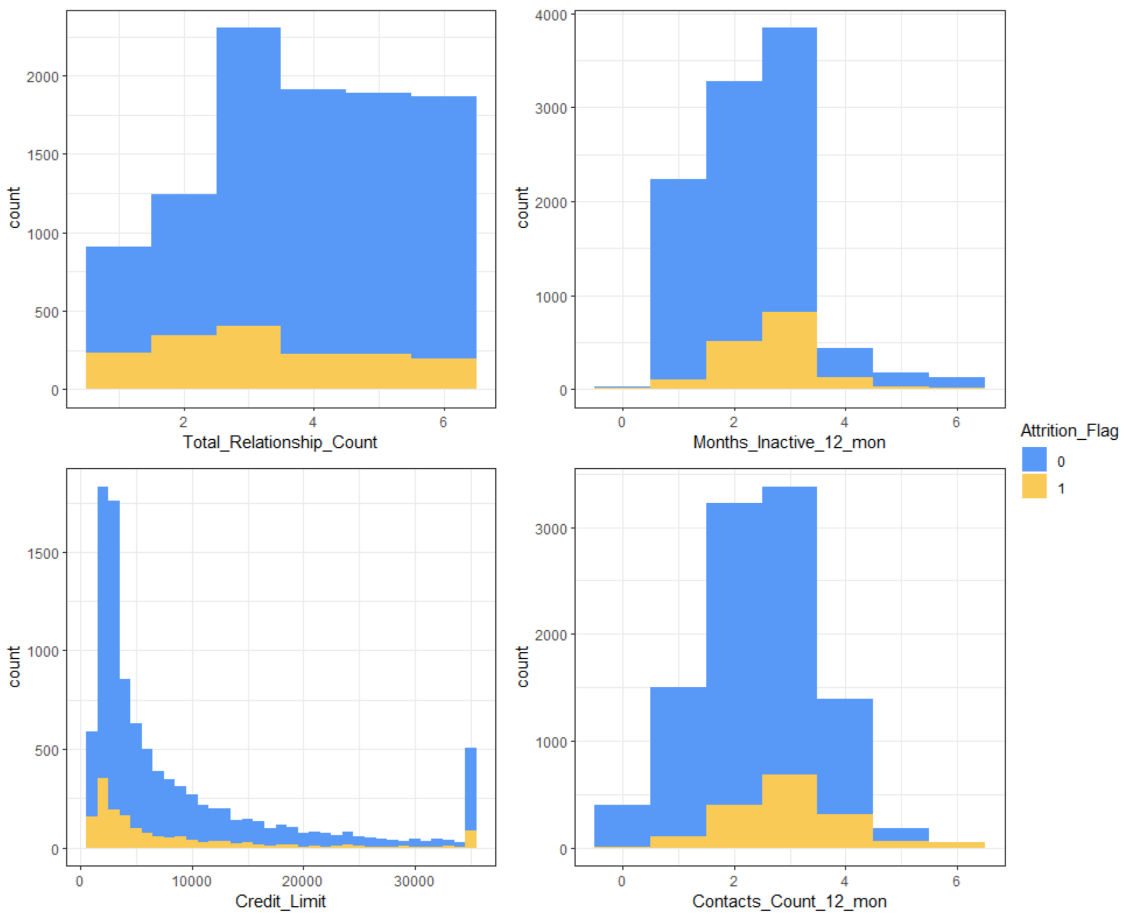


Exhibit 4: Correlation Matrix - Y and Numerical Variables



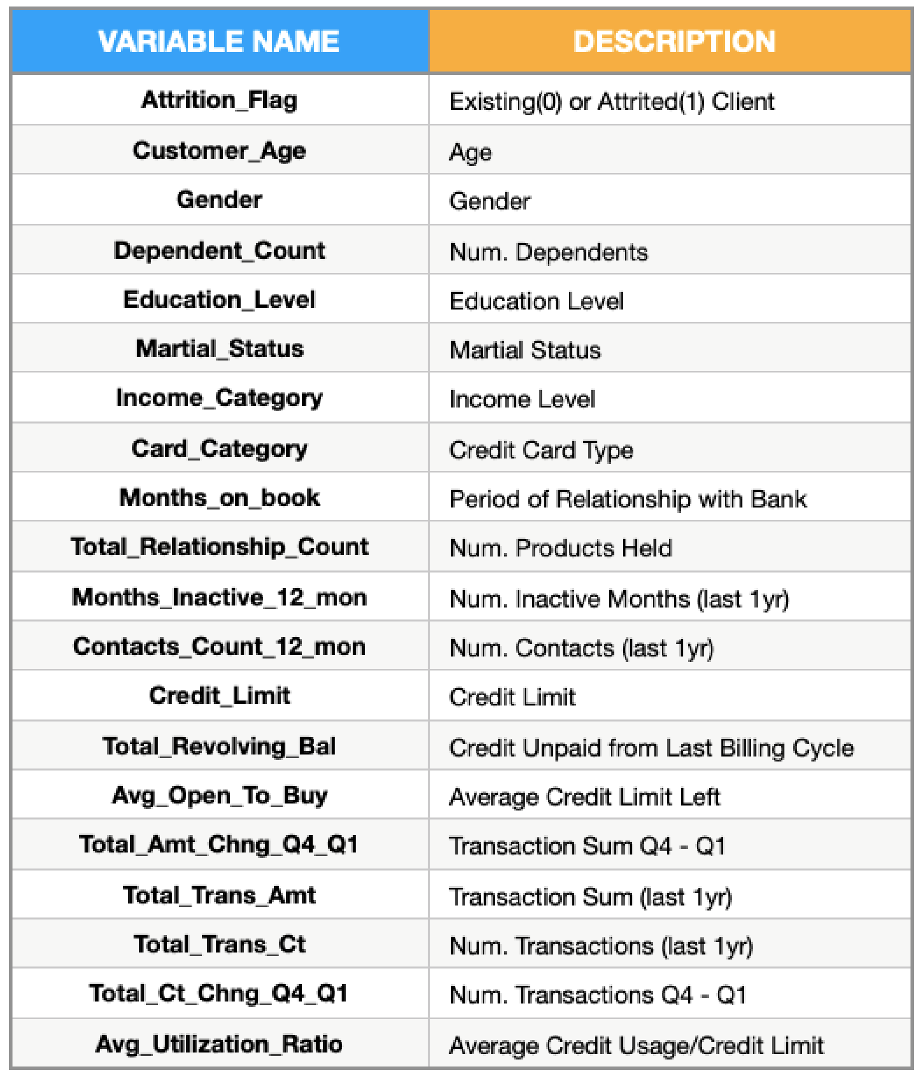
Appendix

Exhibit 5: EDA - Important Influencers Overview



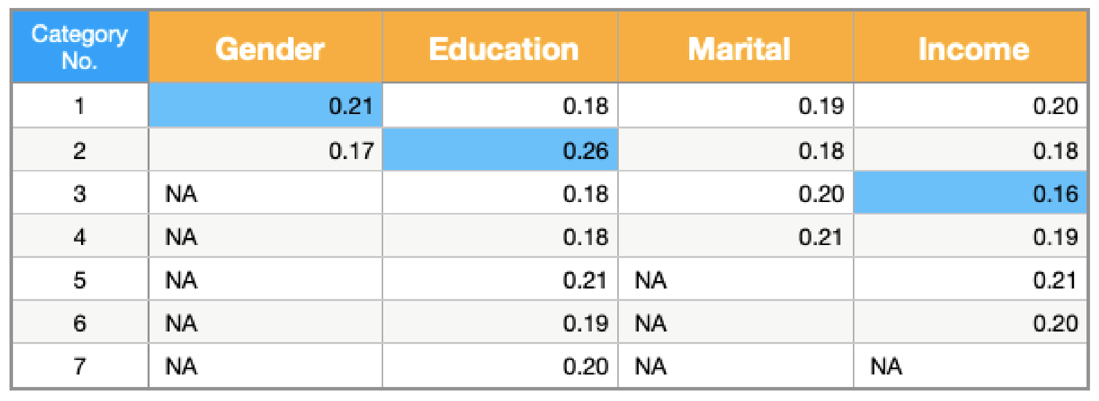
Appendix

Table 1: Variable Description



VARIABLE NAME	DESCRIPTION
Attrition_Flag	Existing(0) or Attrited(1) Client
Customer_Age	Age
Gender	Gender
Dependent_Count	Num. Dependents
Education_Level	Education Level
Martial_Status	Martial Status
Income_Category	Income Level
Card_Category	Credit Card Type
Months_on_book	Period of Relationship with Bank
Total_Relationship_Count	Num. Products Held
Months_Inactive_12_mon	Num. Inactive Months (last 1yr)
Contacts_Count_12_mon	Num. Contacts (last 1yr)
Credit_Limit	Credit Limit
Total_Revolving_Bal	Credit Unpaid from Last Billing Cycle
Avg_Open_To_Buy	Average Credit Limit Left
Total_Amt_Chng_Q4_Q1	Transaction Sum Q4 - Q1
Total_Trans_Amt	Transaction Sum (last 1yr)
Total_Trans_Ct	Num. Transactions (last 1yr)
Total_Ct_Chng_Q4_Q1	Num. Transactions Q4 - Q1
Avg_Utilization_Ratio	Average Credit Usage/Credit Limit

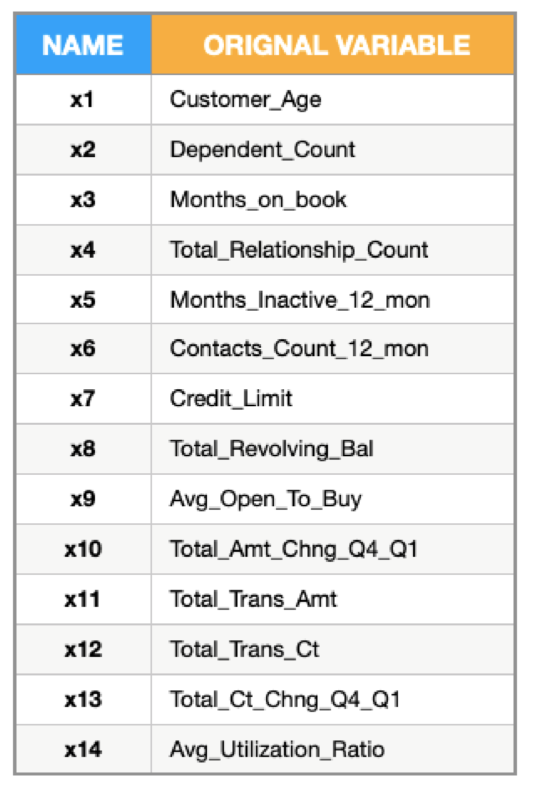
Table 2: Category Selection



Category No.	Gender	Education	Marital	Income
1	0.21	0.18	0.19	0.20
2	0.17	0.26	0.18	0.18
3	NA	0.18	0.20	0.16
4	NA	0.18	0.21	0.19
5	NA	0.21	NA	0.21
6	NA	0.19	NA	0.20
7	NA	0.20	NA	NA

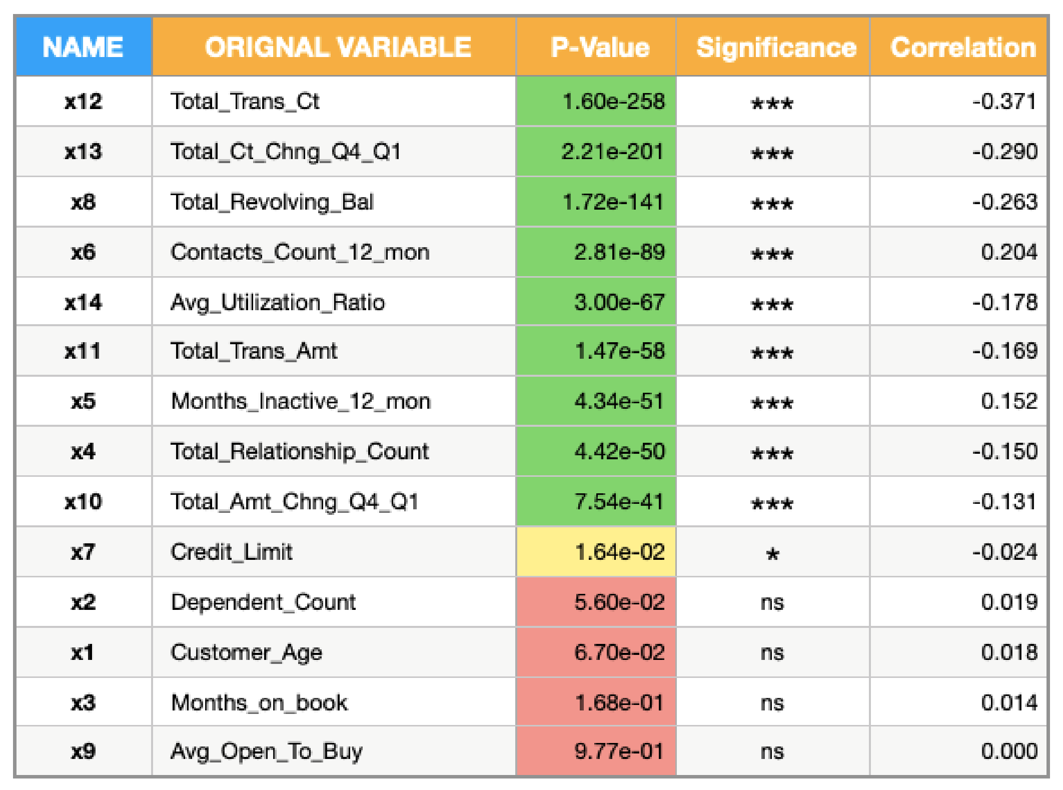
Appendix

Table 3: Illustration Variable Name



NAME	ORIGNAL VARIABLE
x1	Customer_Age
x2	Dependent_Count
x3	Months_on_book
x4	Total_Relationship_Count
x5	Months_Inactive_12_mon
x6	Contacts_Count_12_mon
x7	Credit_Limit
x8	Total_Revolving_Bal
x9	Avg_Open_To_Buy
x10	Total_Amt_Chng_Q4_Q1
x11	Total_Trans_Amt
x12	Total_Trans_Ct
x13	Total_Ct_Chng_Q4_Q1
x14	Avg_Utilization_Ratio

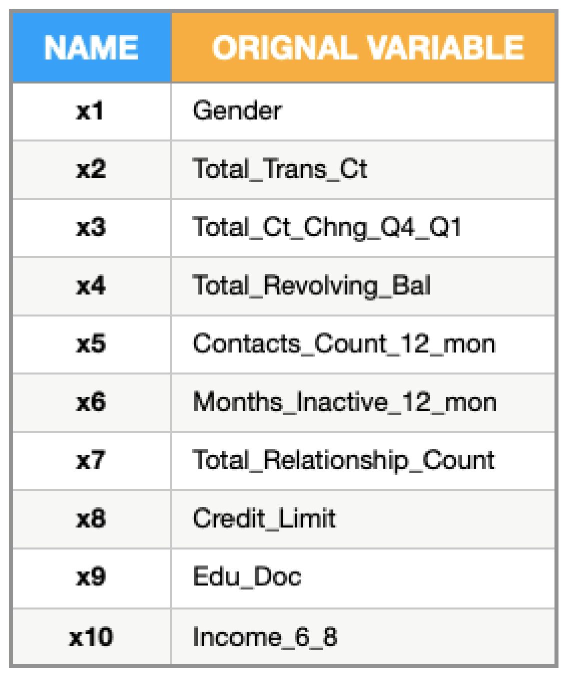
Table 4: Feature Selection - Numercial



NAME	ORIGNAL VARIABLE	P-Value	Significance	Correlation
x12	Total_Trans_Ct	1.60e-258	***	-0.371
x13	Total_Ct_Chng_Q4_Q1	2.21e-201	***	-0.290
x8	Total_Revolving_Bal	1.72e-141	***	-0.263
x6	Contacts_Count_12_mon	2.81e-89	***	0.204
x14	Avg_Utilization_Ratio	3.00e-67	***	-0.178
x11	Total_Trans_Amt	1.47e-58	***	-0.169
x5	Months_Inactive_12_mon	4.34e-51	***	0.152
x4	Total_Relationship_Count	4.42e-50	***	-0.150
x10	Total_Amt_Chng_Q4_Q1	7.54e-41	***	-0.131
x7	Credit_Limit	1.64e-02	*	-0.024
x2	Dependent_Count	5.60e-02	ns	0.019
x1	Customer_Age	6.70e-02	ns	0.018
x3	Months_on_book	1.68e-01	ns	0.014
x9	Avg_Open_To_Buy	9.77e-01	ns	0.000

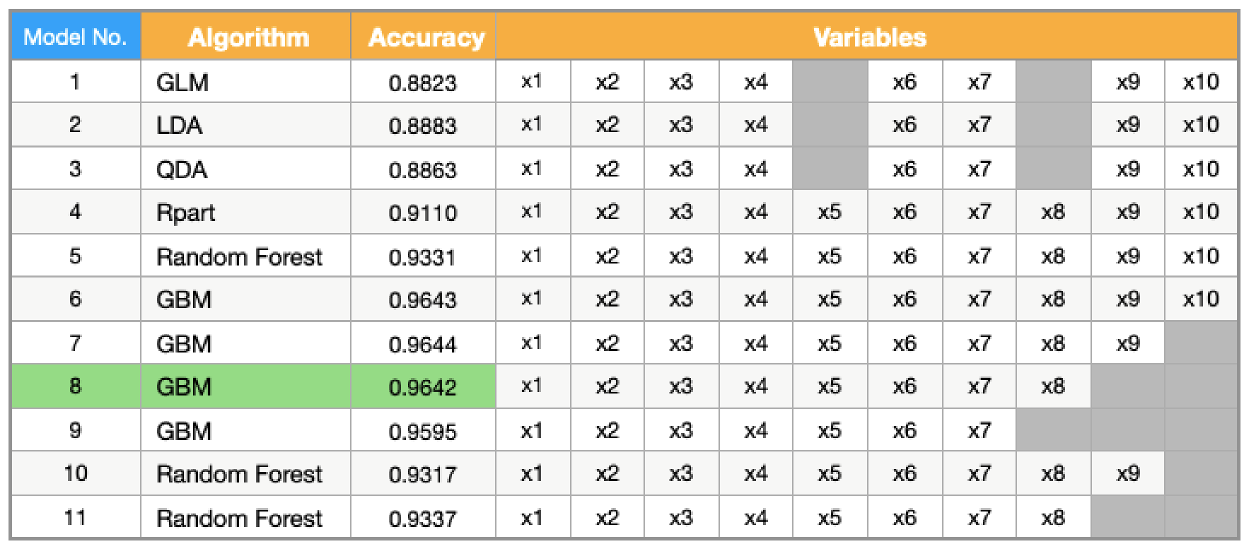
Appendix

Table 5: Model Development Variable Name



NAME	ORIGNAL VARIABLE
x1	Gender
x2	Total_Trans_Ct
x3	Total_Ct_Chng_Q4_Q1
x4	Total_Revolving_Bal
x5	Contacts_Count_12_mon
x6	Months_Inactive_12_mon
x7	Total_Relationship_Count
x8	Credit_Limit
x9	Edu_Doc
x10	Income_6_8

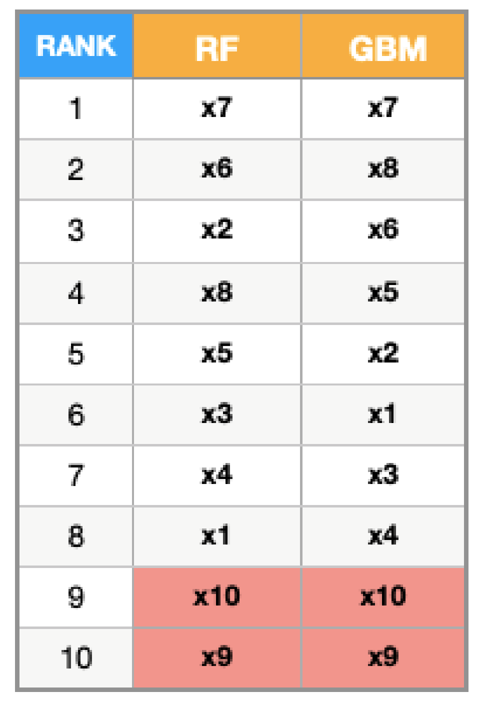
Table 6: Model Selection Summary



Model No.	Algorithm	Accuracy	Variables									
1	GLM	0.8823	x1	x2	x3	x4		x6	x7		x9	x10
2	LDA	0.8883	x1	x2	x3	x4		x6	x7		x9	x10
3	QDA	0.8863	x1	x2	x3	x4		x6	x7		x9	x10
4	Rpart	0.9110	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
5	Random Forest	0.9331	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
6	GBM	0.9643	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
7	GBM	0.9644	x1	x2	x3	x4	x5	x6	x7	x8	x9	
8	GBM	0.9642	x1	x2	x3	x4	x5	x6	x7	x8		
9	GBM	0.9595	x1	x2	x3	x4	x5	x6	x7			
10	Random Forest	0.9317	x1	x2	x3	x4	x5	x6	x7	x8	x9	
11	Random Forest	0.9337	x1	x2	x3	x4	x5	x6	x7	x8		

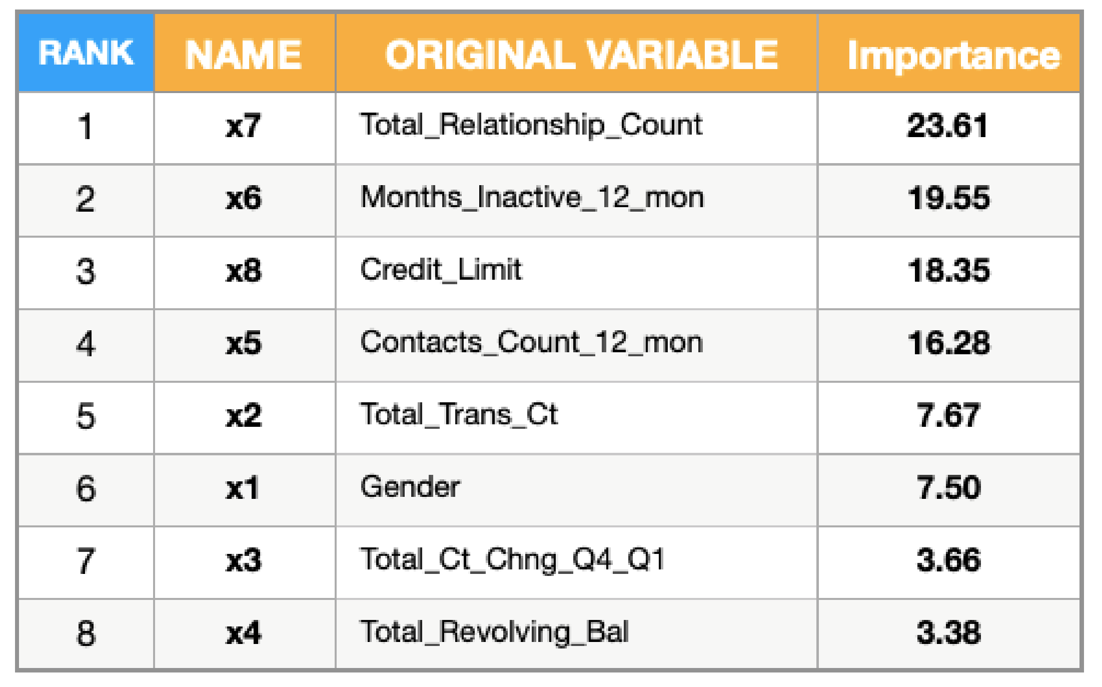
Appendix

Table 7: Variable Importance (Model No. 5 & 6)



RANK	RF	GBM
1	x7	x7
2	x6	x8
3	x2	x6
4	x8	x5
5	x5	x2
6	x3	x1
7	x4	x3
8	x1	x4
9	x10	x10
10	x9	x9

Table 8: Final Model Summary

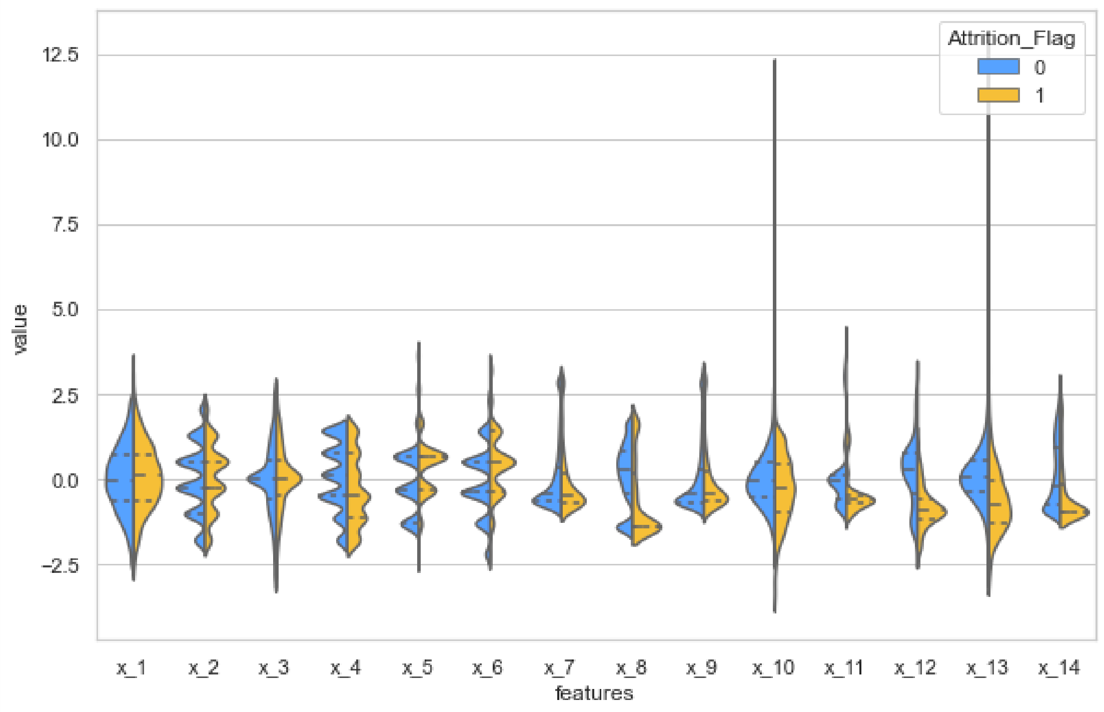


RANK	NAME	ORIGINAL VARIABLE	Importance
1	x7	Total_Relationship_Count	23.61
2	x6	Months_Inactive_12_mon	19.55
3	x8	Credit_Limit	18.35
4	x5	Contacts_Count_12_mon	16.28
5	x2	Total_Trans_Ct	7.67
6	x1	Gender	7.50
7	x3	Total_Ct_Chng_Q4_Q1	3.66
8	x4	Total_Revolving_Bal	3.38

Appendix

Exhibit Extra: EDA - Numerical

A violinplot was generated in Python to get an overview on all numerical variables


Attrition_Flag
value
x_1
x_2
x_3
x_4
x_5
x_6
x_7
x_8
x_9
x_10
x_11
x_12
x_13
x_14
features