Design and Implementation of Analytics System

Customer Retention Modeling for International Banking

IST 350-01: Business Analytics for Decision Making

Abhishek Tripathi

The College of New Jersey

December 15, 2018

Document Control

Work carried out by:

Name	**Email Address**	**Task description**	Total number of hours
Neel Kumtakar	kumtakn1@tcnj.edu	Team Leader	6
Michael Bifulco	bifulcm1@tcnj.edu	Team Member	2
Connor Introna	intronc@tcnj.edu	Team Member	2
Alice Li	lia2@tcnj.edu	Team Member	6
Anthony Sofia	sofiaa2@tcnj.edu	Team Member	1

Revision Sheet

Release No.	**Date**	**Revision Description**
1		Assigned to group
2	11.26.2018	Formed GroupMe for communication purposes
3	12.1.2018	Started working on project
4	12.7.2018	Created Google Doc to work collaboratively on project
5	12.15.2018	Finished the project
6		
7		
8		

TABLE OF CONTENTS

Project Description and Objectives

Problem Definition

Consumer banking is an expanding theme in the broader American economy. As a result, financial institutions must attempt to consistently meet the expectations of the consumer. It is of great value for banks to make a concerted effort to keep the existing client base rather than attract new customers. According to Anunay Gupta of consulting and business solutions company Brillio "acquiring new customer costs six to seven times more than keeping an existing customer." [4]

Per a recent infographic from Invesp focusing on general industry's approach to customer retention, more than half of companies put more resources into customer acquisition rather than retention, which is doing them a large financial disservice. The blog post infographic also shows, not only is retaining customers more cost effective than acquiring new customers, but existing clients are more likely to try new product offerings and spend more money than new clients.

"Looking at churn rates by customer segment illuminates which types of customers are at risk and which may require an intervention. It's a nice simple metric that tells us a lot about when and how to interact with customers," says Jill Avery of Harvard Business School[3]. By understanding and analyzing customer churn patterns, it can help banks refocus resources in meeting and exceeding customer expectations in an effort to increase retention ratings and improve profits.

Objectives

Our challenge is to develop a predictive analysis system to predict who may leave the bank. Any classification system providing a best fit for the dataset may be used to create a churn model to predict customer retention rates for future fiscal years. Additionally, group may want to explore research questions such as what are the key predictors the bank should be aware of with their customers? Examples are included, but not limited to: Are female customers leaving more than males? What age groups show an inclination to leave? Is there indication of customer departures associated with specific countries? Is there evidence of salary associated departures?

Data Collection

Data Sources

The dataset is located on SuperDataScience[1] (a data repository for data science training). The data was provided by Kirill Eremenko, a Data Science management consultant and was extracted from an international bank. The bank is concerned with the churn rate of their customer base and would like to mine their data for what may be causing customer departures. The bank does business in at least the following three countries: France, Spain, and Germany. A random sample of 10,000 customers was selected over a six month period. The data provides information on their current status with the bank as well as 12 additional attributes describing their demographic and banking information.

Data Collection Plan

The research question can be framed as "can we use data about this bank's customers to help the bank predict which customers will stay or exit?" This information will be valuable in helping the bank determine appropriate marketing strategies to attract and retain long-term customers.

Data Understanding and Descriptive Analysis

Data Dictionary

Attribute	Description	Input	Role	Data Values	Values Count
CUSTOMER_ID	Represents the customer's unique identifier as assigned by the bank.	Integer	Input	Min=15565701 Max=15815690	10,000
CUSTOMER_CREDIT_SCORE	Represents the customer's credit score based on FICO	Integer	Input	Min=15565701 Max=15815960	10,000
CUSTOMER_COUNTRY	Represents customer's country of residence.	Categoric	Input	"France", "Germany", "Spain"	France = 5014 Germany = 2509 Spain = 2477
CUSTOMER_GENDER	Represents gender of the customer	Categoric	Input	"Female", "Male"	Female = 4543 Male = 5457
CUSTOMER_AGE	Represents the age of the customer in years	Integer	Input	Min=18 Max=92	70
CUSTOMER_TENURE	Represents the number of years the customer has been a member of the bank	Integer	Input	Min=0 Max=10	11
CUSTOMER_ACCT_BALANCE	Indicates the current account balance for the customer in USD. Balance does not include negative assets.	Numeric	Input	Min=0 Max=250898	6382
CUSTOMER_PRODUCTS	Indicates the number of bank products assigned to a customer. Products include savings, loans, checking account, credit card. The value does not correspond to a product type but rather to a total # of products used.	Integer	Input	Min=1 Max=4	4
CUSTOMER_CREDIT_CARD	Indicates whether the customer has a credit card with the bank	Categoric	Input	Min=0 Max=1	0 = 2945 1 = 7055
CUSTOMER_ACTIVITY	Indicates whether a customer's account is active	Categoric	Input	Min=0 Max=1	0 = 4849 1 = 5151
CUSTOMER_SALARY	Indicates the customer's estimated annual salary in USD	Numeric	Input	Min=11.58 Max=199992.48	9999
CUSTOMER_STATUS	Response variable whether the customer left or stayed with the bank after 6 month observational period for the sample data	Categoric	Target	Min=0 Max=1	0 = 7983 1 = 2037

Descriptive Statistics with R and Rattle

Loading and preparing dataset

Using RStudio, we load and prepare the dataset. We see that the dataset has 10,000 observations and 14 variables. Interestingly, there seems to be no missing values, which is great as imputation is not needed. However, some of the names in the dataset do not match the one in our data description, hence we must rename the variables in the data cleaning. That way, anyone who uses our cleaned dataset will have the variables with the correct names. Additionally, referring to the data description table, we needed to delete some variables, i.e. surname and Row Number.

Data Cleaning and Pre-Processing

Data Cleaning

The Surname, RowNumber, and ID columns were all taken out of the because they do not relate to the target variable. The row number is already listed on the side of the data chart and a whole column dedicated to it is a waste of data use. The Surname also does not provide any valuable information or assist towards the target variable. Similarly, the customer ID also provides little useful information.

Also, for coding purposes and to match the data dictionary, we have renamed the variables in our R File adding the “Cust_” prefix to each them. “NumofProducts” was changed to “Products” in our R file.

Data Transformation

Input variables – categorical

Status, Credit_Card, and Activity are set as factor variables because they are considered a categorical variable. Categorical variables take on a limited number of different values, in the case of this data these variables are either a “1” or a “0”. Gender and Geography are also categorical variables because of the limited number of different values. In this data set there are only two genders (Male and Female) and 3 geographies (France, Spain, and Germany).

Input variables – numeric

Credit Score,Age, Tenure, Balance, NumOfProducts, EstimatedSalary, are all numeric variables. Each of these columns contains a number that is not limited to a certain few numbers.

Final Transformed Dataset

The final transformed dataset contained 10,000 observations of 11 variables. The factor variables were: Geography, Gender, HasCrCard, IsActiveMember, and Exited. Integer variables were CreditScore, Age, Tenure, and NumOfProducts. Numeric variables were Balance and EstimatedSalary.

Variable Selection

Multicollinearity Review

Before we do any kind of modeling, we have to check for multicollinearity. Multicollinearity is a property where 2 independent variables, (variables other than the target variable), have some sort of relationship with one another. This relationship could be a strong positive or a strong negative correlation.

We use a correlation matrix to identify all the possible correlations (in the form of correlation coefficients) in the data. If a correlation coefficient is strongly positive (coefficient near 1) or strongly negative (coefficient near -1), then we could say that multicollinearity has occurred between two variables that produced that coefficient.

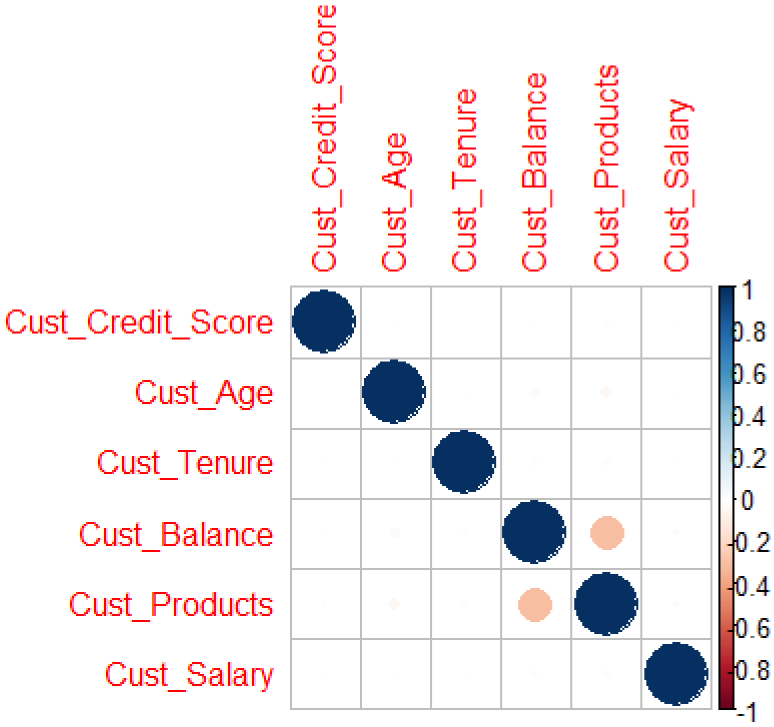
Ignoring multicollinearity can be a problem, as it could have drastic effects on the efficiency of the model and the results. Therefore, if multicollinearity occurs between two variables, we must remove one variable but keep the other. Recall in a model, all variables must be independent. Multicollinearity implies dependencies between variables, which we do not desire.

Correlation Matrix

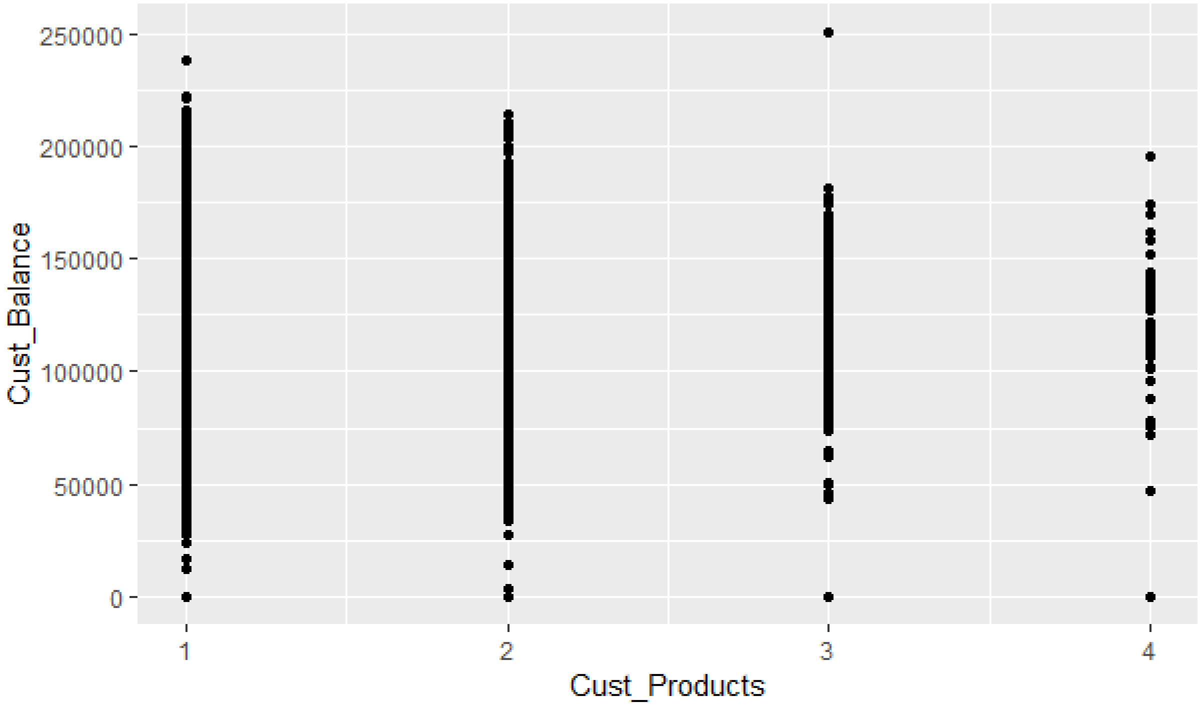
Factor variables cannot be used in correlation as they are categorical variables. Categories or “levels” in a categorical variable cannot be used to evaluate correlation. Thus, types of variables that are used in a correlation matrix are numeric. Using R we run a correlation matrix and we get the following matrix:

Variable Name	Credit Score	Age	Tenure	Balance	Num of Products	Salary
Credit Score	1.0000	-0.003	0.0008	0.0063	0.012	-0.0014
Age	-0.003	1.0000	-0.0099	0.0283	-0.0306	-0.0072
Tenure	0.0008	-0.0099	1.0000	-0.01	0.013	0.007
Balance	0.0063	0.0283	-0.01	1.0000	**-0.304**	0.012
Num of Products	0.012	-0.0306	0.013	**-0.304**	1.0000	0.014
Salary	-0.0014	-0.0072	0.007	0.012	0.014	1.0000

Based on the results of the matrix, there appears to be no multicollinearity occurring between the variables. There is a weak negative correlation between Product and Balance but we can ignore that, as the correlation coefficient (-0.3) is close to 0, which signifies no correlation. Below is the visual representation of the correlation matrix with respect to 6 numeric variables, Credit Score, Age, Tenure, Balance, NumOfProducts, and Salary.



From the data visualization plot below, we can see as products increases, the difference between the maximum and minimum balance shrinks, implying a negative correlation. However, it is not a negative linear correlation that we are used to seeing.



Variable Significance

All the variables are significant in the model. No variables need to be removed according to the correlation matrix.

Predictive Modeling Approach

Evaluation Criteria

To evaluate the model, we will use the following criteria:

1. Accuracy
2. Sensitivity
3. Specificity

Accuracy is the amount of correct classifications the model makes out of the total number of classifications. That is the number of total True Positive and True Negatives over N, the total number of observations/classifications.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

Sensitivity is the number of correct positive classifications out of the total number of positive classifications. Positive classifications are classifications where the condition in question exists (i.e., condition = 1).

$$Sensitivity = \frac{TP}{(TP + FN)}$$

Specificity is the number of correct negative classifications out of the total number of negative classifications. Negative Classifications Are classifications are classifications where the condition does not exist (i.e., condition = 0).

$$Specificity = \frac{TN}{(TN + FP)}$$

Modeling, Evaluation, and Validation

1. Naive Bayes

Model Description and Assumptions

Naive Bayes model is a probabilistic data model that calculates the posterior probabilities of the data. Posterior probabilities are conditional probabilities in the form P(C| X) where C is the Class/Dependent Variable, and X is the independent variable/ Predictor. The Naive Bayes model assumes that the effect that a predictor has on the dependent variable is independent of the other predictors in the model. In other words, Naive Bayes assumes that each predictor/feature makes an equal and independent effect on the dependent variable. This idea is based off the Bayes’s probability principle.

Model Analysis

In the Naïve Bayes model, a series of conditional probabilities are calculated. Below is a table of all the conditional probabilities. For the analysis of the conditional probabilities, we first look at the condition of each probability. If two rules have the same condition, we examine those probabilities and choose the highest one for the analysis of that condition. For the convenience of the writer, in every variable, the conditional probabilities colored in red have the highest probability for that condition and are reflective of their respective scenarios. Also, one should realize that the Naïve Bayes calculates conditional probabilities for qualitative and quantitative variables differently. For quantitative variables, a cutoff value is used and the frequency of the number of observations below/above the cutoff have been computed. For reader, we have manually calculated those probabilities.

Credit Score:

P(Status=0 | Credit Score <Cutoff value)= 652.96 / (652.96+ 95.67) = 0.872

P(Status=0 | Credit Score > Cutoff value)= 95.67 / (652.96+ 95.67) = 0.128

P(Status=1 | Credit Score <Cutoff value)= 644.70 / (644.70+ 101.63) = 0.864

P(Status=1 | Credit Score > Cutoff value)= 101.63 / (644.70+ 101.63) = 0.136

The conditional probabilities suggest, if a customer's credit score is low, then he or she has a high probability of leaving the bank after 6 months. On the other hand, if a customer has a high credit score, then they are more likely to stay at the bank.

Country:

P(Status=0 | Country= France) = 0.531

P(Status=0 | Country= Germany) = 0.212

P(Status=0 | Country= Spain) = 0.257

P(Status=1 | Country= France) = 0.383

P(Status=1 | Country= Germany) = 0.409

P(Status=1 | Country= Spain) = 0.208

Based on this conditional probability, customers from France are more likely to leave the bank while customers from Germany are more likely to stay in the bank after 6 months.

Gender

P(Status=0 | Gender = Male) = 0.431

P(Status=0 | Gender = Female) = 0.569

P(Status=1 | Gender = Male) = 0.562

P(Status=1 | Gender = Female) = 0.438

Based on the conditional probabilities, if a customer is male, he will most likely stay at the bank after 6 months. If the customer was a female, she is more likely to leave after 6 months. This could be based on the fact that females are more impulsive, while males are more patient.

Age

P(Status=0 | Age <Cutoff Age)= 37.44 / (37.44 + 10.17) = 0.786

P(Status=0 | Age > Cutoff Age)= 10.17 / (37.44 + 10.17) = 0.214

P(Status=1 |Age <Cutoff Age)= 45.18 / (45.18+ 9.71) = 0.823

P(Status=1 | Age > Cutoff Age)= 9.71 / (45.18+ 9.71) = 0.177

Based on the conditional probability, if a customer is young, then he/she is more likely to stay at the bank after 6 months. However, if a customer is old, then he/she is more likely to leave the bank after 6 months.

Tenure

P(Status=0 | Tenure <Cutoff Tenure)= 4.99 / (4.99 +2.89) = 0.633

P(Status=0 |Tenure > Cutoff Tenure)= 2.89 / (4.99 +2.89) = 0.367

P(Status=1 | Tenure <Cutoff Tenure)= 4.94 / (4.94+2.96) = 0.625

P(Status=1 | Tenure > Cutoff Tenure)= 2.96 / (4.94+2.96) = 0.375

Tenure is how many years the customer has been with the bank. If a customer has a low tenure, he or she has a high chance of leaving the bank. However, if a customer has a high tenure, he or she is more likely to stay at the bank. This makes sense as the “surviving” customers are the ones that have had a long history with their bank.

Balance

P(Status=0 | Balance <Cutoff Balance value) = 72799.43 / (72799.43+62822.11) = 0.537

P(Status=0 | Balance > Cutoff Balance value) = 62822.11/ (72799.43+62822.11) = 0.463

P(Status=1 | Balance <Cutoff Balance value) = 91796.71 / (57215.35+91796.71) = 0.616

P(Status=1 | Balance > Cutoff Balance value) = 57215.35 / (57215.35+91796.71) = 0.384

If customers have a low balance, then they are more likely to stay at the bank. However, if customers have a high balance, they are most likely to leave. It is possible that when customers achieve a high balance, they don’t feel the need to be associated with the bank, as they noticed minute increases in their money. They just don’t see the value in the bank other than a storage unit for their money. However, those with low balance, see value in the bank, a potential for higher increases in money. Thus, they will be more inclined to stay at the bank in the long term.

Products

P(Status=0 | Products <Cutoff Product value) = 1.55 / (1.55+0.51) = 0.752

P(Status=0 | Products> Cutoff Product value) = 0.51 / (1.55+0.51) = 0.248

P(Status=1 |Products <Cutoff Product value) = 1.47 / (1.47 +0.80) = 0.648

P(Status=1 |Products > Cutoff Product value) = 0. 80 / (0.80 +1.47) = 0.342

If a customer has a smaller number of bank products, then they are more likely to leave the bank. However, if they have a high number of products then they are more likely to stay. Recall that products include savings account, loans, checking account and credit card. To some people, such products represent a connection/association between customer and bank. The weaker the connection, the more likely it is that the customer will break that connection.

Credit Card

P(Status=0 | Credit Card=0) = 0.293

P(Status=0 |Credit Card=1) = 0.707

P(Status=1 | Credit Card=0) = 0.310

P(Status=1 | Credit Card=1) = 0.690

Customers that have a credit card with the bank are more likely to leave. On the other hand, customers that do not have a credit card with the bank are more likely to stay. It is possible that when customers have a credit card with the bank, they are more likely to overspend and thus have to pay off a huge amount of credit in the long run.

Activity

P(Status=0 | Activity=0) = 0.445

P(Status=0 | Activity=1) = 0.555

P(Status=1 | Activity=0) = 0.645

P(Status=1 | Activity=1) = 0.356

If a customer account is inactive then they are more likely to stay at the bank. However, if a user account is active, then they are more likely to leave the bank. This is a most unusual observation. One would think that active accounts are customers more likely to stay, while inactive accounts are accounts more likely to leave. But that's not the case. Suppose you open an account, and you don't touch the account for a long period of time. The money stored accumulates interest, so you are more likely to stay with the bank.

Salary

P(Status=0 | Salary <Cutoff Salary)= 99820.09 / (99820.09+57605.01) = **0.634**

P(Status=0 | Salary> Cutoff Salary)=57605.01 / (99820.09+57605.01) = 0.366

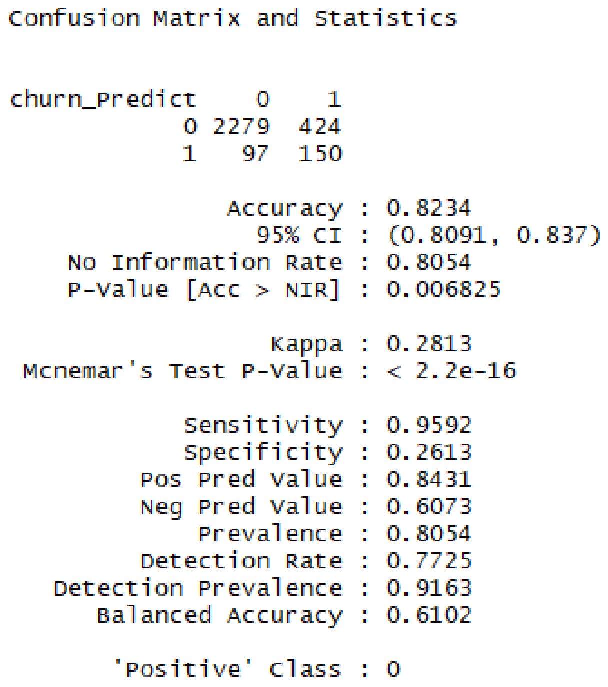
P(Status=1 |Salary <Cutoff Salary)= 101868.07 / (101868.07+57510.13) = 0.639

P(Status=1 | Salary > Cutoff Salary)=57510.13 / (57510.13 +101868.07) = 0.361

If a customer has a low salary, then they are more likely to stay with the bank. If a customer has a high salary then they are more likely to leave the bank.

Model Evaluation/Performance

The model’s accuracy is quite good, correctly classifying about 80% of the model’s observations/cases. Sensitivity is pretty good correctly classifying about 95% positive cases out of the total number of positive cases (True Positive and False Negative). Specificity is terrible, correctly classifying 26% negative cases out of the total number of negative cases (True Negative and False Positive).



```
Confusion Matrix and Statistics


churn_Predict    0    1
            0 2279  424
            1   97  150

               Accuracy : 0.8234
                 95% CI : (0.8091, 0.837)
    No Information Rate : 0.8054
    P-Value [Acc > NIR] : 0.006825

                  Kappa : 0.2813
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9592
            Specificity : 0.2613
         Pos Pred Value : 0.8431
         Neg Pred Value : 0.6073
             Prevalence : 0.8054
         Detection Rate : 0.7725
   Detection Prevalence : 0.9163
      Balanced Accuracy : 0.6102

       'Positive' Class : 0
```

Naive Bayes Conclusion

Overall, the Naive Bayes model is good in performance, considering accuracy and sensitivity. However, it has misclassified 20% of the observations. Naive in general is easy to implement, can take categorical as well as numeric variables. Most of the time however the independent assumption is wrong.

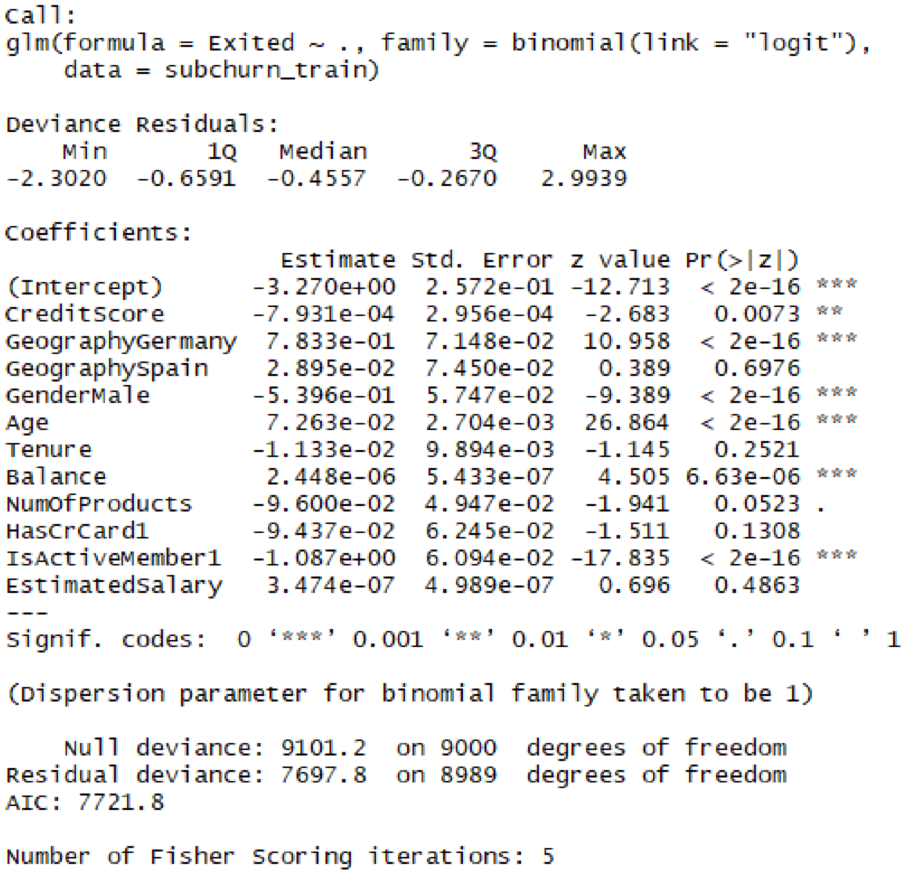
2. Regression

Model Description

Regression is a mathematical technique that finds a relationship between a dependent variable to one or more independent variables in the form of an equation. All regression models are predictive analyses. There are various regression methods for different types of data. For this dataset, a logistic regression approach is the most appropriate. Logistic regression is used to describe data and explain the relationship between a dependent binary variable and one or more independent variables. In this case, the dependent variable that we want to predict is Exited.

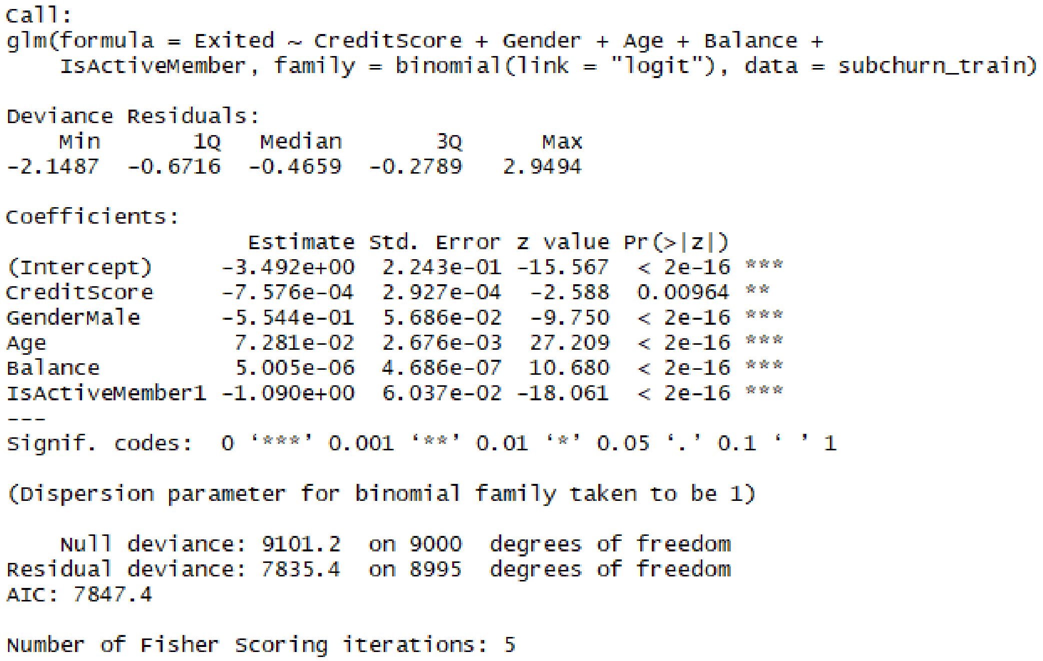
Data cleaning was first done to change the datatype of various variables. The categorical variables Exited, HasCrCard, and IsActiveMember were all changed to factor datatype. Then, a subset of the original dataset was created which removed the variables Rowname, Customer_id, and Surname. These variables would not be useful in regression since they are identifier variables. From this data subset, a testing and training dataset were created to validate the regression model.

A preliminary logistic model was fitted with all independent variables using the glm function in RStudio. Variables with insignificant p-values were then removed from the model. A 95% confidence level was chosen. Thus, the variables Geography, Tenure, NumOfProducts, HasCrCard, and EstimatedSalary were removed since they were insignificant at the 95% confidence level. The figure below shows the preliminary logistic regression model done in RStudio.



```
Call:
glm(formula = Exited ~ ., family = binomial(link = "logit"),
    data = subchurn_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3020  -0.6591  -0.4557  -0.2670   2.9939

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)      -3.270e+00  2.572e-01 -12.713  < 2e-16 ***
CreditScore      -7.931e-04  2.956e-04  -2.683   0.0073 **
GeographyGermany  7.833e-01  7.148e-02  10.958  < 2e-16 ***
GeographySpain    2.895e-02  7.450e-02   0.389   0.6976
GenderMale       -5.396e-01  5.747e-02  -9.389  < 2e-16 ***
Age               7.263e-02  2.704e-03  26.864  < 2e-16 ***
Tenure           -1.133e-02  9.894e-03  -1.145   0.2521
Balance           2.448e-06  5.433e-07   4.505 6.63e-06 ***
NumOfProducts    -9.600e-02  4.947e-02  -1.941   0.0523 .
HasCrCard1       -9.437e-02  6.245e-02  -1.511   0.1308
IsActiveMember1  -1.087e+00  6.094e-02 -17.835  < 2e-16 ***
EstimatedSalary   3.474e-07  4.989e-07   0.696   0.4863
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9101.2  on 9000  degrees of freedom
Residual deviance: 7697.8  on 8989  degrees of freedom
AIC: 7721.8

Number of Fisher Scoring iterations: 5
```

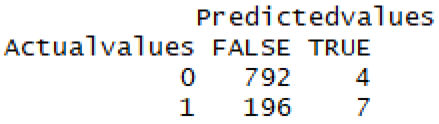
The final logistic regression model only included the following variables: CreditScore, Gender, Age, Balance, and IsActiveMember. The figure below displays the coefficients and residuals of the final logistic regression model.



```
Call:
glm(formula = Exited ~ CreditScore + Gender + Age + Balance +
    IsActiveMember, family = binomial(link = "logit"), data = subchurn_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1487  -0.6716  -0.4659  -0.2789   2.9494

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -3.492e+00  2.243e-01 -15.567  < 2e-16 ***
CreditScore     -7.576e-04  2.927e-04  -2.588  0.00964 **
GenderMale      -5.544e-01  5.686e-02  -9.750  < 2e-16 ***
Age              7.281e-02  2.676e-03  27.209  < 2e-16 ***
Balance          5.005e-06  4.686e-07  10.680  < 2e-16 ***
IsActiveMember1 -1.090e+00  6.037e-02 -18.061  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9101.2  on 9000  degrees of freedom
Residual deviance: 7835.4  on 8995  degrees of freedom
AIC: 7847.4

Number of Fisher Scoring iterations: 5
```

Model Analysis

The testing dataset was used for prediction and validation of the final logistic regression model. A confusion matrix was then generated based on the prediction results to assess the regression model. A cutoff value of 0.5 was used for the confusion matrix because that value yielded the best accuracy. The table below shows the Actual vs. Predicted values from the confusion matrix. The model had an accuracy of 0.808 or 80.8%. The sensitivity and specificity were calculated based on the formulas. Sensitivity was 0.153 and specificity was 0.976.



```
             Predictedvalues
Actualvalues FALSE TRUE
           0   792    4
           1   196    7
```

Looking at the confusion matrix, the model does a good job at predicting when customers will leave. However, there is not many true positive values (only 7). Furthermore, there are also 196 false negative values. This means that the customer who would have stayed would have been predicted to have left.

Regression Conclusion

$$\widehat{Exited} = \frac{e^{u}}{1+e^{u}} \ ,$$

$$where\ u = -3.492 - 0.0007576C - 0.5544G + 0.07281A + 0.000005B - 1.090I \ ,$$

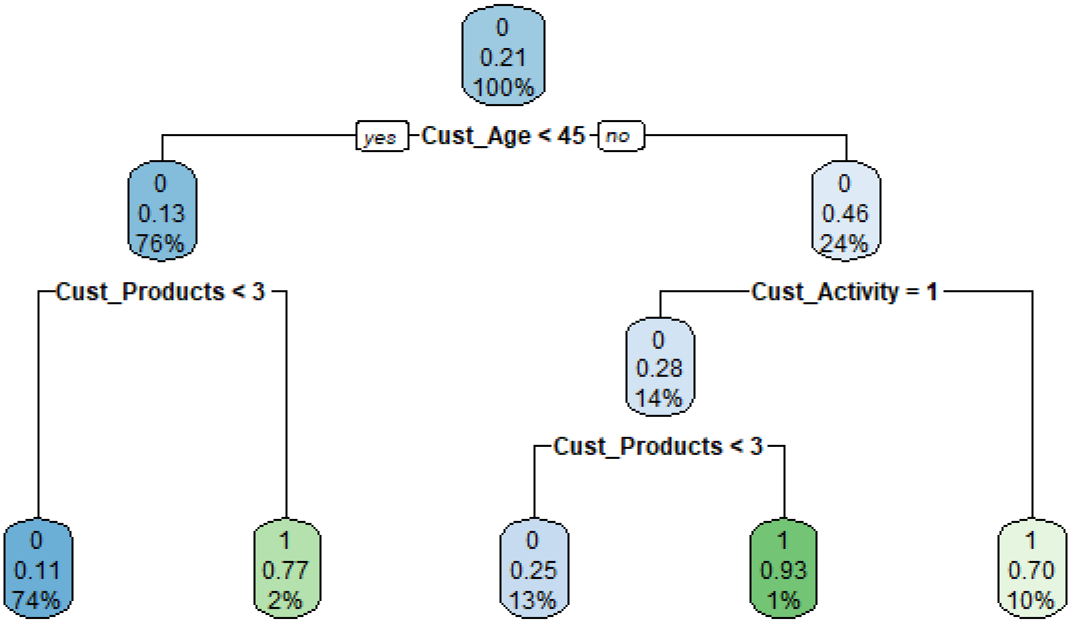
$$and\ C = CreditScore, G = Gender, A = Age, B = Balance, and\ I = IsActiveMember$$

The final logistic regression equation is detailed above. Overall, the final logistic regression model's accuracy is relatively good. It was surprising that many of the independent variables were insignificant and could not be used in predicting if customers would leave the bank or not.

3. Any Other Algorithms

Model Description

The decision tree is a tree-like model whose main functionality is to categorize the data based on certain rules. The "leaves" of the decision tree are the data partitions/groups. The "root" is the parent of the leaf, as in the group that the leaf in a part of. The path from root to leaf is the decision rules.



Model Analysis

Our decision tree training model for determined 3 rules that have categorized the bank customers

I. Customer Age < 45
II. Number of Customer Products < 3
III. Customer Activity = 1

From the graph, the decision tree has created a total of 5 groups of customers.

Group I: This group of customers was determined by the decision tree as a group of customers under the age of 45 and where the number of products purchased is under 3. In general, this was a young group and slightly uninvolved. This group accounts for **74%** of the training data. **According to the decision tree, this young group is deemed the customers most likely to leave after 6 months.**

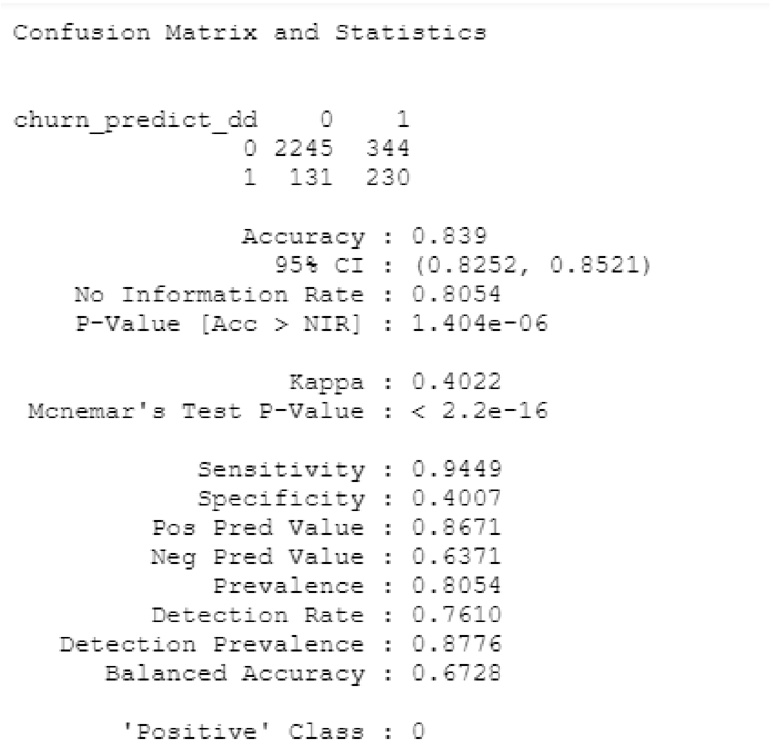
Group II: This group accounts for **2%** of the training data and is a young group like Group I. However, unlike the latter, this group bought more than 3 products. This young group is more involved compared to its counterpart. **According to the decision tree, this young group is deemed a group of customers likely to stay after 6 months.**

Group III: This group represents an older group of customers, aged 45 or greater, accounting for **13%** of the training data. This group is an active group of customers that have less than 3 customer products. **According to the decision tree, this old group is a group of customers more likely to leave after 6 months.**

Group IV: This is a group that accounts for **1%** of the training data and like Group III represents an older crowd of customers. Like Group III, this group is an active group, but this group has bought more than 3 products. **According to the decision tree, this old group is a group of customers more likely to stay after 6 months.**

Group V: This group accounts for **10%** of the training data and like the previous groups represents an older crowd of customers. However, unlike the previous two groups, this group is an inactive group. **According to the decision tree, this old group is a group of customers more likely to stay after 6 months.**

Model Evaluation/Performance



```
Confusion Matrix and Statistics


churn_predict_dd    0    1
               0 2245  344
               1  131  230

               Accuracy : 0.839
                 95% CI : (0.8252, 0.8521)
    No Information Rate : 0.8054
    P-Value [Acc > NIR] : 1.404e-06

                  Kappa : 0.4022
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9449
            Specificity : 0.4007
         Pos Pred Value : 0.8671
         Neg Pred Value : 0.6371
             Prevalence : 0.8054
         Detection Rate : 0.7610
   Detection Prevalence : 0.8776
      Balanced Accuracy : 0.6728

       'Positive' Class : 0
```

The decision tree accuracy is quite good, correctly classifying about 84% of the observations. Sensitivity is great, correctly classifying 94% of positive tests, but specificity could be better.

Decision Tree Conclusion

Instead of evaluating conditional probabilities, the Decision Tree model allows us to see the bigger picture of the classification of our data. Unfortunately, according to the decision tree model, most of the customers in the training data set are ones that leave after 6 months. The ones that do stay, in shorter amount, tend to have more products. Inactive accounts surprisingly are reflective of customers that do stay.

Final Model Selection

Model	Accuracy	Sensitivity	Specificity
Naive Bayes	0.823	**0.959**	0.261
Logistic Regression	0.809	0.153	**0.976**
Decision Tree	**0.839**	0.945	0.401

Conclusion

Predictive Model Summary

Analyzing all the models, it would seem that each one is top in one category. Decision Tree fairs the best in terms of accuracy. Naive Bayes is top for its sensitivity analysis, and Logistic Regression reigns supreme in terms of Specificity. We do feel that accuracy weights higher than sensitivity or specificity as it takes into account the all the correct positive and negative tests, rather than focusing on one element. Therefore, since the Decision Tree has a high accuracy, it is the best model for the classifying the churn data.

Customer Recommendations

Based on the decision tree training model, we feel that most of the bank's customers are young. However, a huge percentage of this young population is leaving the bank due to having less products. Therefore, the bank should try to cater to its fleeting youngsters by encouraging them to buy more bank products, such as checking accounts, savings accounts, and credit cards. By doing so, the bank could sway them into staying in the long term. However, the bank also should not completely ignore its older customers.

References

1. Eremenko, K. (2015, Aug 11). Churn Modelling. Retrieved from http://www.superdatascience.com/training
2. Saleh, K. (n.d). Customer Acquisition Vs.Retention Costs – Statistics And Trends [Web Log Post]. Retrieved August 29, 2017 from http://www.invespcro.com/
3. Naïve Bayesian (n.d). Retrieved from https://www.saedsayad.com/naive_bayesian.htm
4. Naïve Bayes Classifiers (n.d). Retrieved from https://www.geeksforgeeks.org/naive-bayes-classifiers/
5. Lander, Jared. "R for Everyone: Advanced Analytics and Graphics, 2nd edition". Pearson Education, 2017

Appendix 1.0 Project Plan

Tasks:	Notes/Status and Author
Data Reading	Completed by everyone
Data Cleaning	Completed by everyone
Data Dictionary table	Completed by Neel and Alice
Naïve Bayes Model	Completed by Neel
Regression Model	Completed by Alice
Model Comparison	Assigned to Michael Completed by Neel
Decision Tree Analysis	Assigned to Anthony Completed by Neel
Transformation Analysis	Completed by Michael and Neel
Conclusion	Completed by Neel and Connor
Documentation	Completed by Connor and Alice
Professional Appearance	Completed by Alice
References	Completed by Alice
Code Organization	Completed by Alice

Appendix 2.0 Various R Code

```
## Churn Data
## IST 301
## Neel Kumtakar, Anthony Sofia, Alice Li, Connor Introna, and Micheal Bifulco
## 12/15/18
## Grp2

churn<-read.csv(file.choose())  ## we read in the data
str(churn) ## figure out the structure of the data
summary(churn)  ##get the summary statistics, i.e min and max of each variable
View(churn)  ## View the dataset instead of opening it

 ##get the number of values for numeric variables or the distribution of values
for   categorical variables
length(churn$RowNumber)
length(churn$CreditScore)
table(churn$Geography)
table(churn$Gender)
length(churn$Age)
length(churn$Tenure)
length(churn$Balance)
length(churn$NumOfProducts)
table(churn$HasCrCard)
table(churn$IsActiveMember)
length(churn$EstimatedSalary)
table(churn$Exited)

## Data cleaning Step
library(dplyr)
## Dplyr's rename function comes in handy when renaming columns.
## We rename the column names to match the data dictionary names and for coding purposes.

churn= rename(churn, Cust_ID= CustomerId, Cust_Country= Geography, Cust_Credit_Score=
CreditScore, Cust_Gender= Gender, Cust_Age=Age, Cust_Tenure=Tenure, Cust_Balance=
Balance, Cust_Products=NumOfProducts, Cust_Credit_Card=HasCrCard,
Cust_Activity=IsActiveMember, Cust_Salary=EstimatedSalary, Cust_Status=Exited )
names(churn)

## Verify the names have been changed
## Data Transformation- We remove Row Number, Surname and ID as they  are variables that
do not relate to our target variable or are useful in our model. Also because
Cust_Status, Cust_Credit_Card, and Cust_activity are categorical variables, we set them
as factor variables.

churn$RowNumber=NULL
churn$Surname= NULL
churn$Cust_ID= NULL
churn$Cust_Status <- as.factor(churn$Cust_Status )
churn$Cust_Credit_Card<-as.factor(churn$Cust_Credit_Card)
churn$Cust_Activity<-as.factor(churn$Cust_Activity)
churn$Cust_Age <- as.numeric(churn$Cust_Age )
churn$Cust_Products<-as.numeric(churn$Cust_Products)
churn$Cust_Tenure<-as.numeric(churn$Cust_Tenure)
churn$Cust_Credit_Score<-as.numeric(churn$Cust_Credit_Score)

## To get the correltion matrix, you need to filter out the categorical variables.
## Correlation matrix only meant for numeric variables.

library(dplyr)
churn_subset=select(churn,-c(Cust_Country,Cust_Gender,Cust_Status,Cust_Credit_Card,
Cust_Activity)
## Setting up the Correlation Matrix
library(corrplot)
CHRN_corr<-cor(churn_subset)
corrplot(CHRN_corr)
## model the correlation between Balance and Products
library(ggplot2)
```

```
a<-ggplot(churn_subset, aes(y=Cust_Balance, x=Cust_Products)) + geom_point()


## Naive Bayes Procedure
set.seed(3)
## generate testing and training data- 70% testing, 30% train
c <- sample(2, nrow(churn), prob=c(0.70,0.30), replace=TRUE)
churn_Train= churn[c==1,]
churn_Test= churn[c==2,]
library(e1071)
churn_naive= naiveBayes(Cust_Status~., data=churn_Train)
churn_Predict=predict(churn_naive, churn_Test)
churn_Predict
## Generate Naive Bayesian classification statistics
library(caret)
confusionMatrix(table(churn_Predict, churn_Test$Cust_Status))


## Decision Tree Procedure
 library(rpart)
library(rpart.plot)
library(e1071)
 dt_churn <- rpart(Cust_Status~.,churn_Train, method= "class")
summary(dt_churn)
rpart.plot(dt_churn)
 churn_predict_dd<-predict(dt_churn,churn_Test, type = "class")
churn_predict_dd
 library(caret)
confusionMatrix(table(churn_predict_dd,churn_Test$Cust_Status))


## Regression Procedure
# Data cleaning
churn$Exited <- as.factor(churn$Exited)
churn$HasCrCard <- as.factor(churn$HasCrCard)
churn$IsActiveMember <- as.factor(churn$IsActiveMember)
colnames(churn)

# Create a subset that removes variables rowname, customer_id, and surname
# These variables will not be useful in regression model b/c they are identifiers

subchurn <- churn[,c("CreditScore", "Geography", "Gender", "Age", "Tenure", "Balance",
                        "NumOfProducts", "HasCrCard", "IsActiveMember",
                        "EstimatedSalary", "Exited")]
head(subchurn)
str(subchurn)

# Splitting the data
install.packages("caret")
library(caret)

set.seed(99)
id <- createDataPartition(subchurn$Exited, p = 0.90, list = F)
id
subchurn_train <- subchurn[id,]
subchurn_test <- subchurn[-id,]
str(subchurn_train)
str(subchurn_test)

# Logistic Regression model
logreg <- glm(Exited~., data = subchurn_train, family = "binomial"(link = "logit"))
summary(logreg)

final_logreg <- glm(Exited ~ CreditScore + Gender +
                      Age + Balance + IsActiveMember, data = subchurn_train,
                      family = "binomial"(link = "logit"))
summary(final_logreg)
```

```
# Coefficients of final logistic regression model
invlogit <- function(x)
   {     1 / (1 + exp(-x))     }
invlogit(final_logreg$coefficients)

# Use test dataset for prediction
logreg_pred <- predict(final_logreg, newdata = subchurn_test, type = "response")
head(logreg_pred)

# Confusion Matrix
# test different cutoff values to find one with best accuracy
# cutoff value of 0.3
table(Actualvalues = subchurn_test$Exited, Predictedvalues = logreg_pred > 0.3)

# cutoff value of 0.4
table(Actualvalues = subchurn_test$Exited, Predictedvalues = logreg_pred > 0.4)

# cutoff value of 0.5
table(Actualvalues = subchurn_test$Exited, Predictedvalues = logreg_pred > 0.5)

# cutoff value of 0.6
table(Actualvalues = subchurn_test$Exited, Predictedvalues = logreg_pred > 0.6)

# cutoff value of 0.7
table(Actualvalues = subchurn_test$Exited, Predictedvalues = logreg_pred > 0.7)
```