College of Professional Studies

Northeastern University San Jose

MPS Analytics

Course: ALY6015: Intermediate Analytics

Assignment: Final Project: Analysis Report

Bank Customer Churn

Submitted on: February 12, 2023

Submitted to:

Professor: PAROMITA GUHA

Submitted by:

ARCHIT BARUA

HEEJAE ROH

NIKSHITA RANGANATHAN

YUQING CHEN

INTRODUCTION

The goal of this project is to analyze customer behavior/pattern and determine the prediction of the exited target variable. The dataset used for this analysis is the "Churn Modeling" dataset which contains information about customers who have either exited or not exited from a particular credit card company. The target variable in this dataset is "Exited" and the project aims to predict this variable using various statistical methods and machine learning algorithms. The methods applied in this project include descriptive statistics, hypothesis testing, linear and logistic regression, decision tree and regularization techniques.

In addition to the mentioned techniques, we have also incorporated Lasso and Ridge regularization analysis to address the issue of overfitting and improve the performance of the models. Furthermore, decision trees will be used to gain a deeper understanding of the relationships between the various variables and the target variable "Exited".

The insights from this project can be used to develop targeted marketing campaigns, improve customer service, and create personalized retention strategies to reduce customer churn. Additionally, the results can be used to identify the most important drivers of customer attrition and prioritize initiatives aimed at reducing churn.

ANALYSIS

PART 1. PRIMARY QUESTIONS

1. What are the main factors that are driving customers to churn?
2. What are the customer segments that are most likely to churn?
3. What are the most effective strategies to reduce customer churn?
4. What are the most effective methods to increase customer loyalty/engagement?

We first looked at the EDA to see which segment the customers who exited. Based on this, we planned various tests. In the test, we will look at the real reasons why customers leave. We will consider strategies to prevent customers from churn.

Furthermore, beyond preventing churn, we would like to generate ideas by synthesizing the results of what strategies can be used to increase customer loyalty.

PART 2. DATA CLEANING

Data cleaning is an important step in data analytics projects because it helps to improve the quality and reliability of the data, making it easier to use and analyze. The steps we used to prepare the data for analysis are:

- Checking for NA Values - The 'churn' dataset does not have any missing or null values

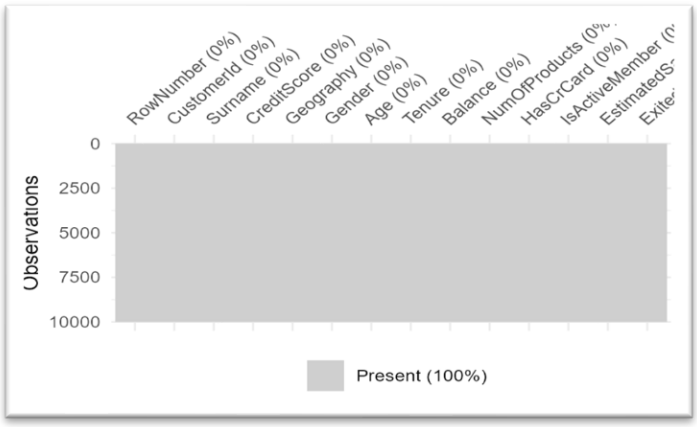

Figure 1 – Visualization of NA values

- Checking for duplicate rows - There are no duplicate rows in the dataset.

- Data Manipulation - The main objective of data manipulation is to prepare data for analysis by transforming it into a format that is easier to work with.
  - Adding a new column “Status” - We created a new variable Status based on the value of the Exited variable: if Exited is equal to 1, Status is assigned the value "Churned"; otherwise, Status is assigned the value "Retained".
  - Dropping column - We decided to remove column “RowNumber” because this column does not contribute to the analysis or the insights.
  - Manipulating data - The columns "HasCrCard" and "IsActiveMember" are modified by recoding the original variables from 0/1 to "No"/"Yes" and "Inactive"/"Active" respectively.
  - Updating the column names - Column names of two columns "HasCrCard" and "IsActiveMember" are revised in the "churn" dataset to make the names more descriptive and easier to understand.
  - Changing the datatypes - The data type of the columns is changed from one datatype to a factor by using the as.factor() function.

PART 3. DESCRIPTIVE STATISTICS & EDA

- Understanding the bank churn dataset by Headtail

RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure
1	15634602	Hargrave	619	France	Female	42	2
2	15647311	Hill	608	Spain	Female	41	1
3	15619304	Onio	502	France	Female	42	8
…				…			
9998	15584532	Liu	709	France	Female	36	7
9999	15682355	Sabbatini	772	Germany	Male	42	3
10000	15628319	Walker	792	France	Female	28	4

Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
0	1	1	1	101348.88	1
83807.86	1	1	1	112542.58	0
159660.8	3	3	0	113931.57	1
…				…	
0	1	0	1	42085.58	1
75075.31	2	1	0	92888.52	1
130142.79	1	1	0	38190.78	0

Figure 2 – headtail()

- The "churn" dataset contains information about 10,000 customers of a bank. It has 14 variables:

1. RowNumber: Row number specifying the number of customers
2. CustomerId: Unique Identifier for each customer
3. Surname: Surname of each customer
4. CreditScore: Credit score of each customer
5. Geography: The country where each customer is located
6. Gender: The gender of each customer
7. Age: The age of each customer
8. Tenure: Number of years each customer has been with the bank
9. Balance: The current balance of each customer
10. NumOfProducts: The number of bank products each customer is using
11. HasCrCard: Whether each customer has a credit card or not (recoded as "Yes" or "No")
12. IsActiveMember: Whether each customer is an active member or not (recoded as "Active" or "Inactive")
13. EstimatedSalary: The estimated salary of each customer
14. Exited: Whether each customer has left the bank or not (recoded as "1" or "0")

- The dataset contains 7 categorical variables, 1 character variable (Surname), and 6 numerical variables.

	Mean	Sd	Median	Trimmed	Mad	Min	Max	Range	Skew	Kurtosis	se
Surname*	1508.78	846.20	1543.00	1512.94	1085	1.00	2932.0	2931.0	-0.02	-1.20	8.46
CreditScore	650.53	96.65	652.00	651.01	99.33	350.00	850.0	500.0	-0.07	-0.43	0.97
Geography*	1.75	0.83	1.00	1.68	0.00	1.00	3.0	2.0	0.50	-1.36	0.01
Gender*	1.55	0.50	2.00	1.56	0.00	1.00	2.0	1.0	-0.18	-1.97	0.00
Age	38.92	10.49	37.00	37.91	8.90	18.00	92.0	74.0	1.01	1.39	0.10
Tenure	5.01	2.89	5.00	5.01	2.97	0.00	10.0	10.0	0.01	-1.17	0.03
Balance	76486	62397	97199	74828	69336	0.00	250898	250898	-0.14	-1.49	623.97
NumOfProducts	1.53	0.58	1.00	1.49	0.00	1.00	4.0	3.0	0.75	0.58	0.01
HasCrCard	0.71	0.46	1.00	0.76	0.00	0.00	1.0	1.0	-0.90	-1.19	0.00
IsActiveMember	0.52	0.50	1.00	0.52	0.00	0.00	1.0	1.0	-0.90	-1.19	0.00
EstimatedSalary	100090	57510	100194	100115	72941	11.58	199993	199980	0.00	-1.18	575.10
Exited	0.20	0.40	0.00	0.13	0.00	0.00	1.0	1.0	1.47	0.16	0.00

Figure 3 – describe()

- The mean value of the variable "Balance" is 76485.89, the median value is 97198.5, and the minimum and maximum values are 0 and 250898 respectively.
- The summary statistics of the "CreditScore" variable show a mean of 650.53, with a median of 652, and a range of 500 (from a minimum of 350 to a maximum of 850).
- On the other side, "EstimatedSalary" has a mean value of 100090.24 and a median value of 100193.5. Min and max values are 12 and 199992.
- Most of the variables have a small skew with the exception of the "Exited" and "Status" variables, which have a skewness of 1.47, indicating a positive skewness and a shift to the right.
- The kurtosis of most of the variables is close to 0, with the exception of the variable Age, which has a kurtosis value of 1.39, indicating a peaked distribution.
- The mean of Exited is 0.20, with a standard deviation of 0.40 and the descriptive statistics show that 20% of the customers in the dataset have exited.

1. Barplots

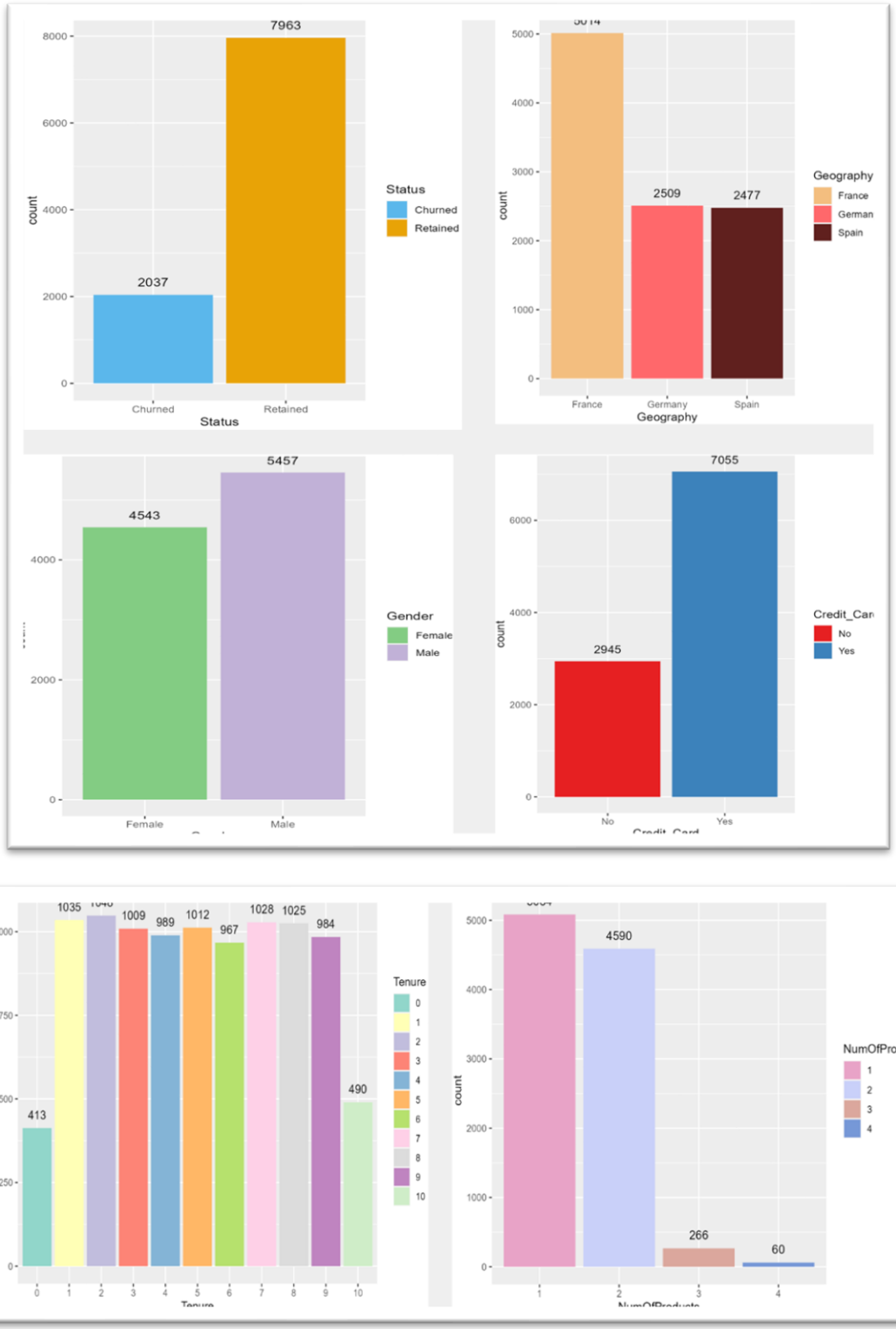

Figure 4 – Barplots

- Approximately 20% of the bank's customers left, resulting in a 20% churn rate, while the bank was able to retain 80% of its customers.
- A significant portion of the bank's customer base is comprised of individuals who reside in France.
- Male customers constitute the majority of the bank's customer base, making up 5457 of the customers, while female customers make up the remaining 4543.
- 71% of the customer base utilize credit cards, while the remaining 29% do not make use of them. This could be because of various factors like personal financial preference, income level, debt level, or banking behavior.
- The bank experiences a relatively consistent rate of customer churn over the years 1 to 9. A smaller proportion of customers leave the bank within the first year. This means the bank is doing a good job of retaining new customers in the early stages but may have difficulty maintaining customer loyalty over the long term.
- A majority of the customers have purchased either one or two products, while only a limited number have purchased three or four.

2. Histogram

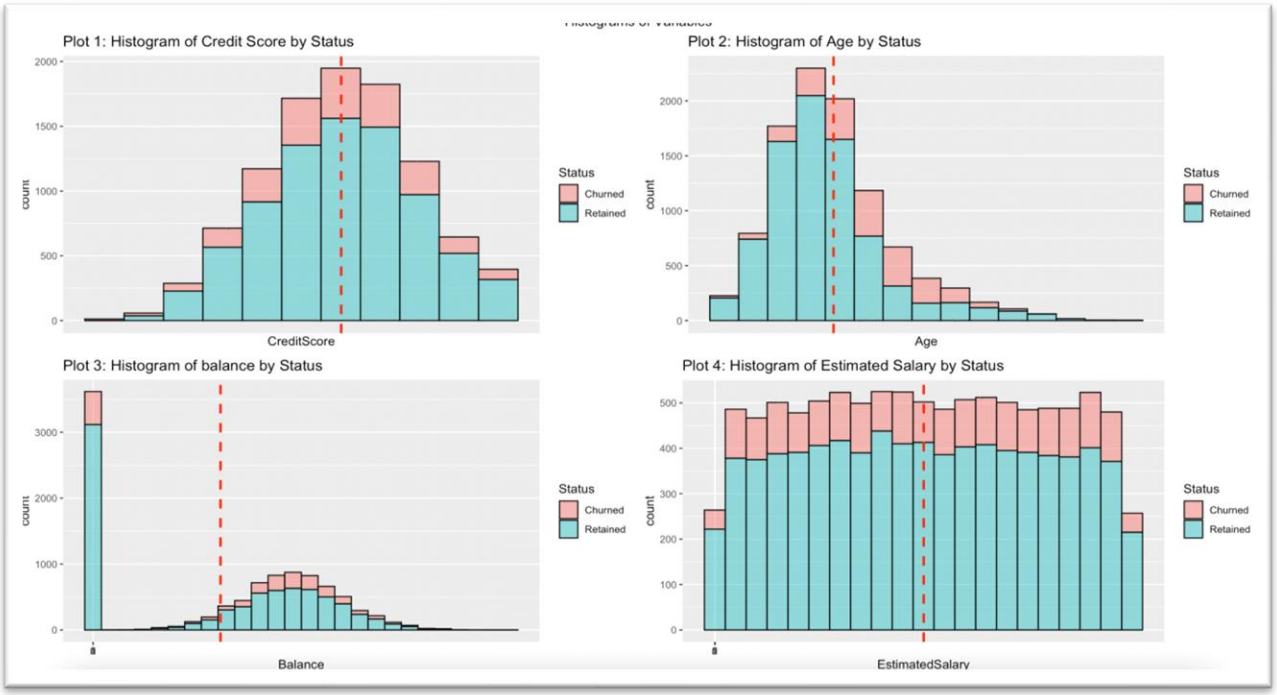

Figure 5 – Histogram

The histograms show the frequency of the data distributed by each variable. In each plot, the pink bar represents the customers who exited the bank, and the blue bar represents the customers who are retaining. And the red line represents the mean credit score of all observations.

- Plot 1 represents the distribution of Credit Scores by Status. We can see it is roughly normally distributed, and the average credit score is around 650.
- Plot 2 represents the distribution of Age by Status. The distribution of Age is right-skewed, The peak of the age occurs to the left of the median age, so the mean age comes to the right of the center.
- Plot 3 represents the distribution of Balance by Status. Except for the balance of zero, the histogram has a normal distribution with a mean of around 125k. We noticed that some customers with zero deposits didn't churn because they bought the products.
- Plot 4 represents the distribution of Estimated Salary by Status. From the graph, we can see the histogram is uniform distribution ranging between 0–200k, and the mean estimated salary is around 100k. This is the same for both churned and retained customers.

3. Churned rate

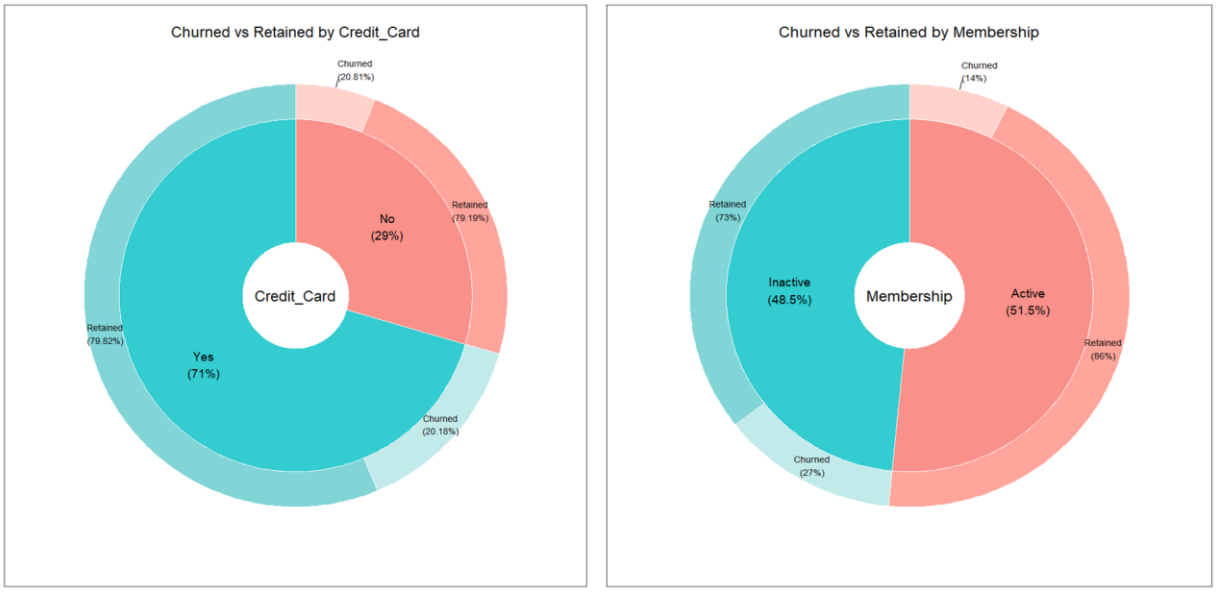

Figure 6 – pie chart

- Approximately 29% of clients do not possess a credit card. The retention rate is equal for both groups (Credit card Yes and Credit Card No), at around 80%.
- From the second pie donut chart, we can observe that about 50% of customers are considered active, while the remainder are inactive. The percentage of inactive customers who leave the bank is approximately 27%, whereas only 14% of active customers choose to leave.
- This shows the churned rate is different in Membership Status which is active and inactive. This can be used for Logistics regression and hypothesis testing

4. Histogram and Boxplot

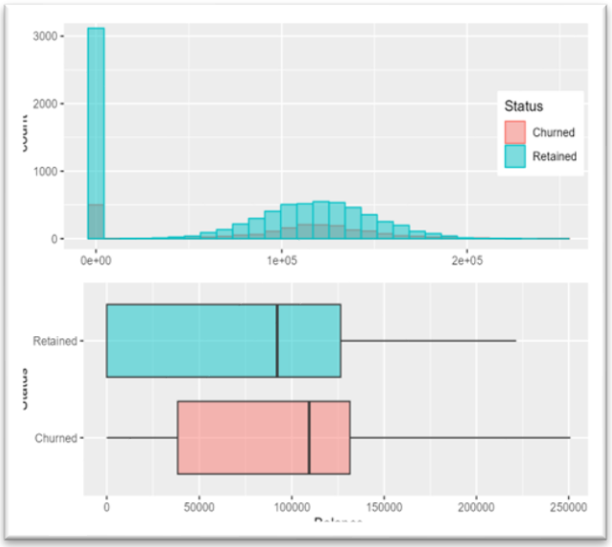

Figure 7a – Histogram+Boxplot A

The distributions of the two groups appear to be quite similar. A significant portion of customers who retained have low balance in their accounts.

Figure 7b – Histogram+Boxplot B

Salaries of both churned and retained customers have uniform distribution and salary does not have a major impact on the probability of churn.

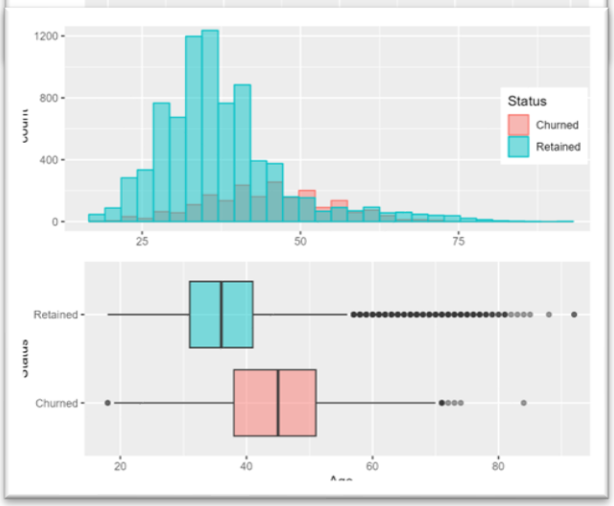

Figure 7c – Histogram+Boxplot C

Churned and Retained customers do not have a similar distribution for age variable. Retained customers have a higher peak compared to churned.

5. Correlation Matrix

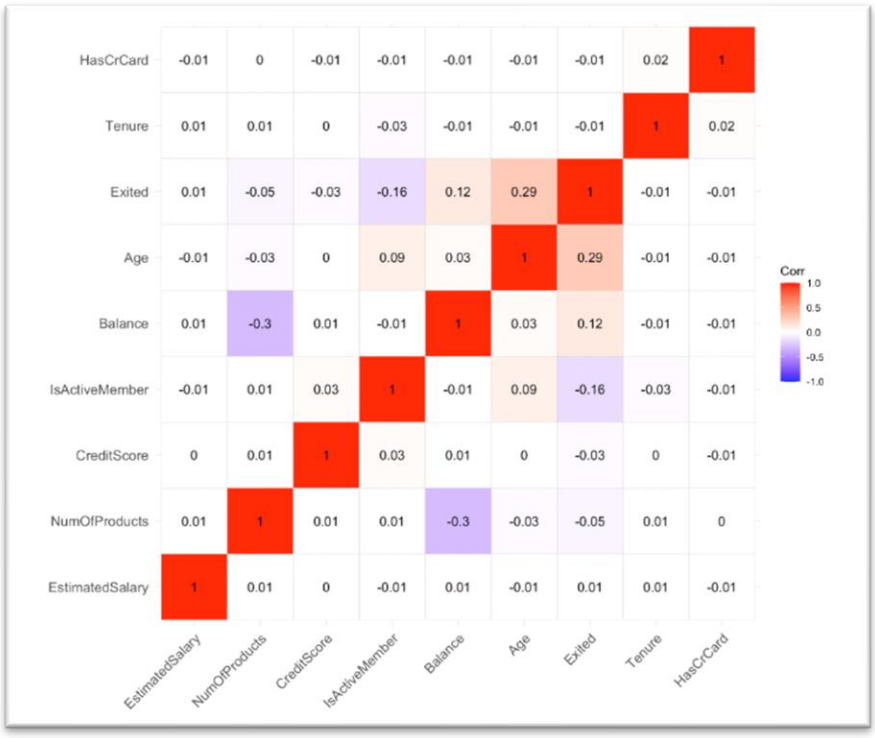

Figure 8 – Correlation Matrix

- From the correlation matrix, we could figure out that there are no high correlation variables with "Exited". "Balance", "Age" and "EstimatedSalary" have a weak positive correlation with " Exited", and the correlation coefficients are 0.12, 0.29, and 0.01. "IsActiveMember" , "CreditScore", “NumOfProducts”, “Tenure” and “HasCrCard” has a weak negative correlation with "Exited", and the correlation coefficients are -0.16,-0.03, -0.05, -0.01 and -0.01.

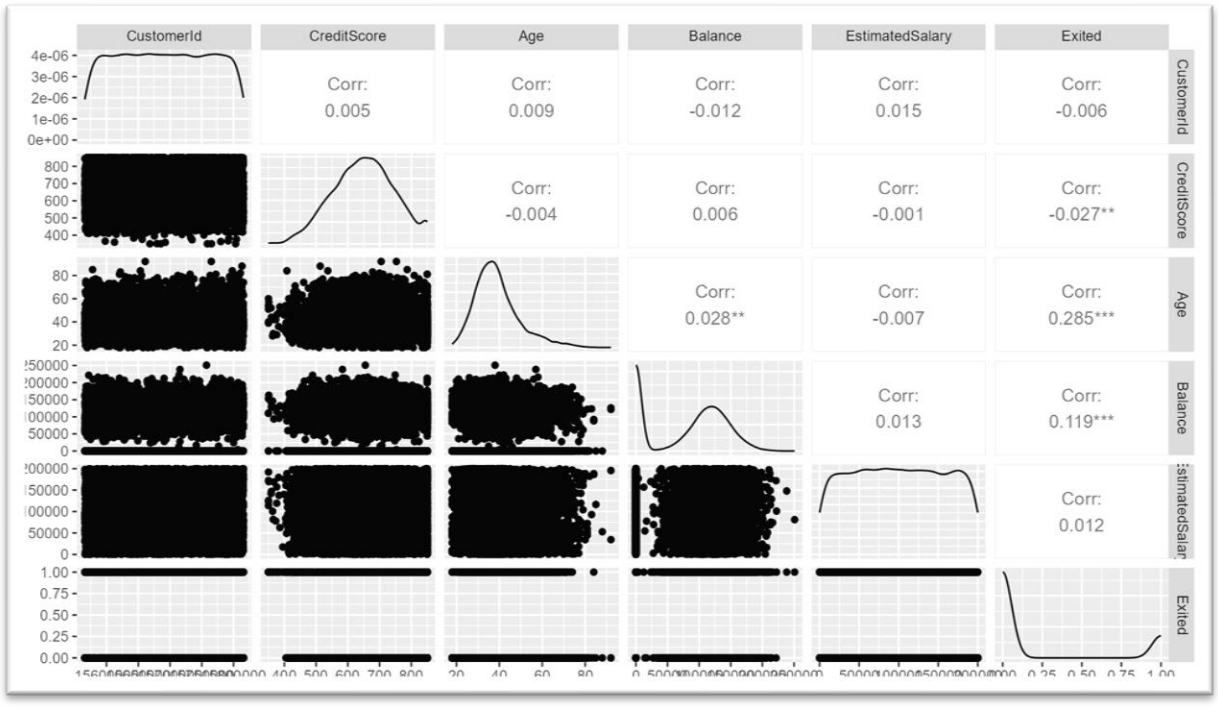

Figure 9 – ggpairs()

- ggpairs is a function in the R programming language that creates a matrix of scatter plots to visualize the relationship between multiple variables in a dataset. It provides a quick and easy way to view the distribution and relationship between variables in a data frame.
- From the above graph, we can see that Exited column is most correlated with CreditScore, Age, and Balance.

5. 3D plot

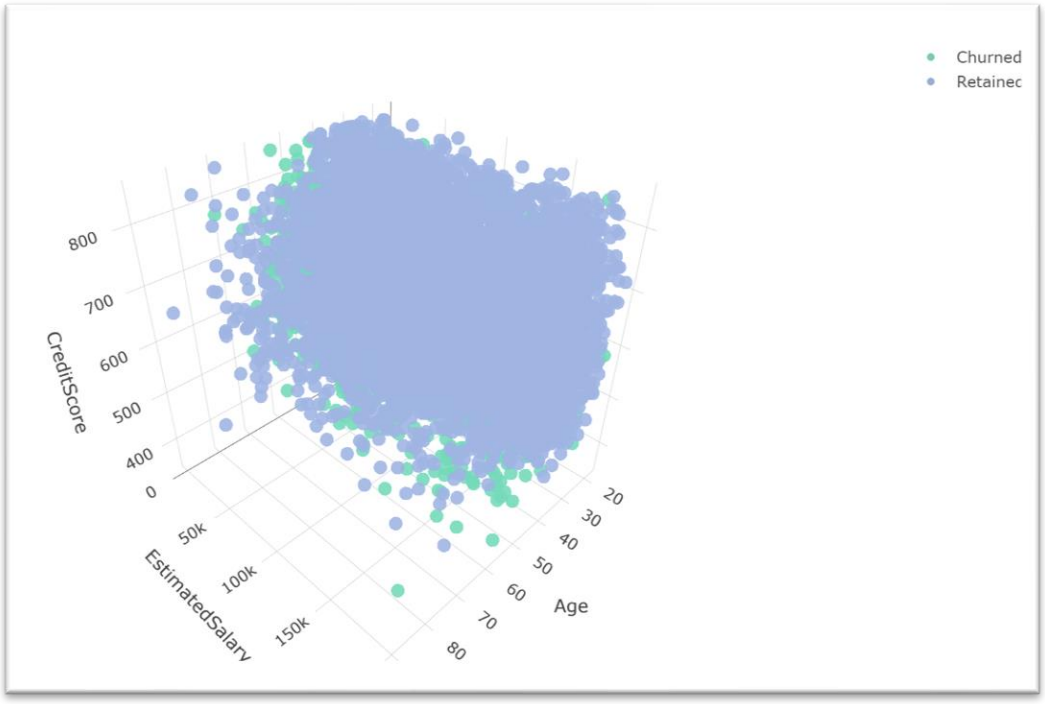

Figure 10 – 3D plot

- A 3D plot is a graphical representation of three variables in a three-dimensional space, where each variable is represented by a different axis.
- In the 3D plot of Credit score, Estimated Salary, and Age, we can see the relationship between the three variables. Through this plot we can get insights into patterns or trends in the data.

PART 4. RESULT OF METHODS & INTERPRETATION

1. Hypothesis Testing (Simplified)

a. One Sample t-test

Null: The mean of CreditScore is greater or equal to 600 (H0: μ1 ≥ 600)

Alternative: The population mean of CreditScore is less than 600 (H1: μ1 < 600, claim)

churn$CreditScore, mu=600, alternative = "less"		
t = 52.278	df = 9999	p-value = 1
95 percent confidence interval:	-Inf 652.1188	
sample estimates: mean of x	650.5288	

Result of One Sample t-test: The P-value (=1) is greater than 0.05, there is not enough evidence to reject H0.

b. Two Sample t-test

Null: The mean of CreditScore is equal between Males and Females (H0: μ1=μ2, claim)

Alternative: The mean of CreditScore differs between Male and Female (H1: μ1≠μ2)

Data: Male$CreditScore and Female$CreditScore		
t = -0.28563	df = 9998	p-value = 0.7752
95 percent confidence interval:	-4.359797	3.250804
sample estimates:	mean of x 650.2769	mean of y 650.8314

Result of Two Sample t-test: The P-value (=0.7752) is greater than 0.05, there is not enough evidence to reject H0.

c. F-test

Null: No difference in the variance of Salary between Male and Female (H0: σ2m = σ2f)

Alternative: Difference in the variance of Salary between M and F (H1:σ2m ≠ σ2f)

F test to compare two variances: Male$EstimatedSalary and Female$EstimatedSalary			
F = 1.009	num df = 5456	denom df = 4542	p-value = 0.7536
95 percent confidence interval:	0.954268 1.066681		
sample estimates: ratio of variances	1.008983		

Result of F-test: The P-value (=0.7536) is greater than 0.05,

The P-value is greater than 0.05, there is not enough evidence to reject H0.

Interpretation

In one-sample t-test, we checked the Creditscore and confirmed that There is not enough evidence to reject the claim that the CreditScore is greater than or equal to 600.

Two-sample t-test, we checked whether there is a difference in CreditScore by gender. There is not enough evidence to reject the claim that there is no difference in CreditScore by gender.

In the F-test, we checked if there was a difference in salary by Gender. There is not enough evidence to reject the claim that there was no difference in Salary by Gender.

d. one-way ANOVA

Null: There is no difference in mean of Balance according to Geography

H0: μ1 = μ2 = μ3

Alternative: At least one mean is different from the others (claim).

Balance ~ Geography, data=churn					
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Geography	2	6.264e+12	3.132e+12	958.4	<2e-16 ***
Residuals	9997	3.267e+13	3.268e+09		

Result of one-way ANOVA: The P-value (<2e-16 ***) is smaller than 0.05, there is enough evidence to reject H0. The mean Balance of Germany is 119,730, France is 62,093, and Spain is 61,818. There is enough evidence that the Balance of Spain & France differs from Germany's.

TukeyHSD(fit)

aov(formula = Balance ~ Geography, data = churn)				
	diff	lwr	upr	p adj
Germany-France	57637.4804	54360.765	60914.196	0.0000000
Spain-France	-274.4898	-3565.282	3016.302	0.9791459
Spain-Germany	-57911.9702	-61707.291	-54116.649	0.0000000

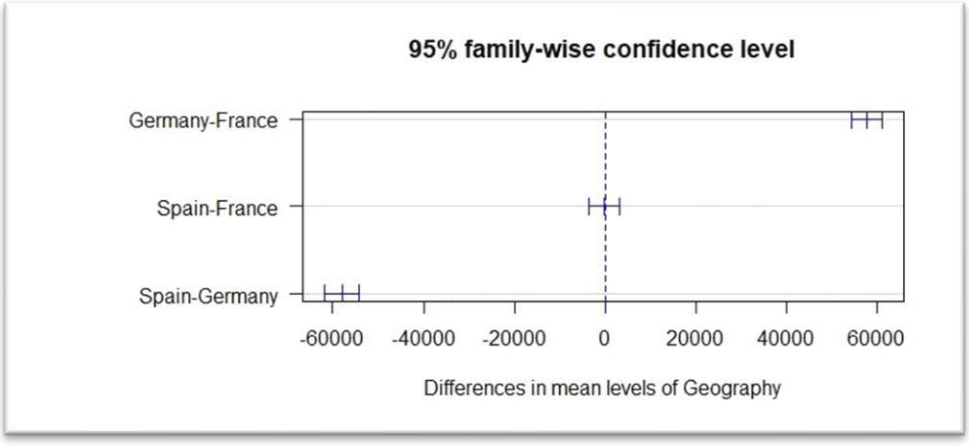

Figure 11 – Tukey plot

e. two-way ANOVA

1. The hypotheses for interaction are stated as follows.

Null: There is no interaction effect between Exited and Gender on Balance.

Alternative: There is an interaction effect between Exited and Gender on Balance.

2. The hypotheses regarding churn are stated as follows.

Null: There is no difference in means of Balance by Exited or not.

Alternative Hypothesis: There is a difference in means of Balance by Exited or not.

3. The hypotheses regarding sex are stated as follows:

Null: There is no difference in means of Balance by sex.

Alternative: There is a difference in means of Balance by sex.

aov(Balance ~ Exited*Gender, data=churn)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Exited	1	5.470e+11	5.470e+11	142.540	<2e-16 ***
Gender	1	2.405e+10	2.405e+10	6.266	0.0123 *
Exited:Gender	1	1.552e+09	1.552e+09	0.404	0.5248
Residuals	9996	3.836e+13	3.837e+09		

Result of two-way ANOVA:

1. Interaction, there is no interaction effect between Exited and Gender on Balance. since 0.5248 (p-value) is greater than 0.5248
2. Exited, there is a difference in means of balance by Exited or not. since 0.000186 (p-value) < 0.05
3. Gender, there is a difference in means of balance by Gender. since 0.0123 (p-value) < 0.05

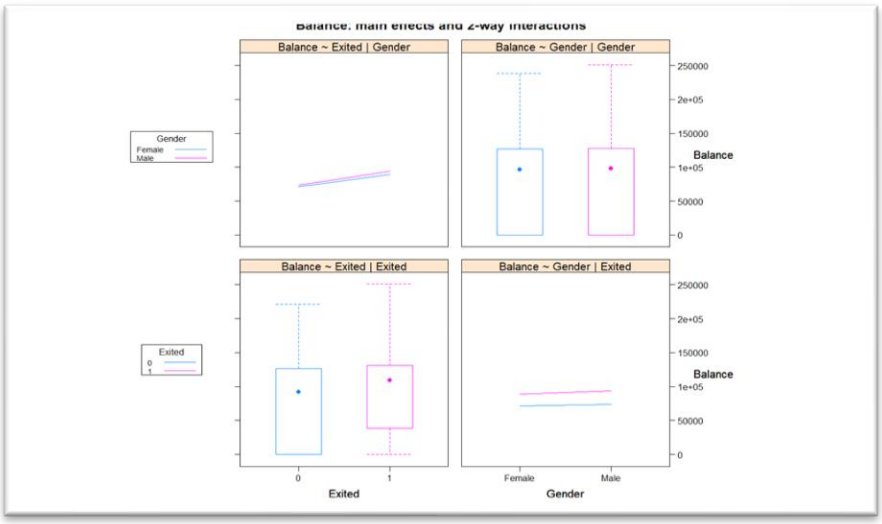

Figure 12 – ANOVA interaction

Interpretation

In one-way ANOVA, there is enough evidence to reject the claim that there is no difference in mean of Balance among Spain, France, and Germany. With Tukey test, we confirmed that there is enough evidence that Germany has difference in Balance comparing with other countries.

In two-way ANOVA, although the interaction hypothesis is not rejected, we can concluded that there is enough evidence that there is difference in Balance by Gender & Exited.

2. Linear Regression

Call: lm(formula = Exited ~ Balance + Age + NumOfProducts, data = churn2)

Residuals

Min	1Q	Median	3Q	Max
-0.81248	-0.22060	-0.13807	-0.02975	1.07857

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.651e-01	1.975e-02	-13.422	<2e-16 ***
Balance	7.016e-07	6.453e-08	10.872	<2e-16 ***
Age	1.083e-02	3.659e-04	29.602	<2e-16 ***
NumOfProducts	-4.227e-03	6.923e-03	-0.611	0.542

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3835 on 9996 degrees of freedom
Multiple R-squared: 0.09365, Adjusted R-squared: 0.09338
F-statistic: 344.3 on 3 and 9996 DF p-value: < 2.2e-16

Based on the Correlation coefficients, we formed a simple linear regression model: lm (Exited~Balance+Age+NumOfProducts, data = chrun2)
From the summary, we found the p-value of independent variables "Balance", "Age" and NumOfProducts" are all less than 0.05, so we can say these 3 variables have a significant influence on Exited. From the multiple r-square of this model, we could observe that our multiple r-square is around 0.09365 or 9.365%.

Interpretation
Although we could use lm() with dependent variables in R, We need to think about the linear regression assumption. Limitations of OLS include Linearity. In addition, normality is included, but when using binary variables, these two limitations are naturally not satisfied. The two most important assumptions of linear regression are not satisfied. Therefore, we need to apply the linear model interpreted above to logistic regression.

3. Logistic Regression, Confusion Matrix

Model 1

```
Call: glm(formula = Exited ~ CreditScore + Geography + Gender + Age + Tenure
                    +Balance + NumOfProducts + Credit_Card + Membership
                    +EstimatedSalary, family = "binomial", data = data_train)
```

Deviance Residuals

Min	1Q	Median	3Q	Max
-2.2809	-0.6601	-0.4621	-0.2763	2.8899

Coefficients:

	Estimate	Std. Error	z value	Pr(>|t|)
(Intercept)	-3.386e+00	2.929e-01	-11.559	< 2e-16 ***
CreditScore	-5.343e-04	3.332e-04	-1.604	0.10876
GeographyGermany	7.708e-01	8.105e-02	9.511	< 2e-16 ***
GeographySpain	6.620e-02	8.352e-02	0.793	0.42803
GenderMale	-5.188e-01	6.490e-02	-7.993	1.31e-15 ***
Age	7.044e-02	3.071e-03	22.941	< 2e-16 ***
Tenure	-1.129e-02	1.115e-02	-1.012	0.31149
Balance	2.370e-06	6.162e-07	3.846	0.00012 ***
NumOfProducts	-1.273e-01	5.685e-02	-2.238	0.02519 *
Credit_Card1	-1.557e-02	7.093e-02	-0.219	0.82627
Membership1	-1.070e+00	6.876e-02	-15.569	< 2e-16 ***
EstimatedSalary	5.209e-07	5.623e-07	0.926	0.35423

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Null deviance: 7063.5 on 6999 degrees of freedom
Residual deviance: 6032.4 on 6988 degrees of freedom
AIC: 6056.4

Number of Fisher Scoring iterations: 5

The output is the results of a logistic regression model using the "glm" function in R. The dependent variable, "Exited", is binary (0 or 1). The independent variables are "CreditScore", "Geography", "Gender", "Age", "Tenure", "Balance", "NumOfProducts", "Credit_Card", "Membership", and "EstimatedSalary".
The "Deviance Residuals" section shows the distribution of residuals from the model, with the minimum, first quartile (1Q), median, third quartile (3Q), and maximum values. The "Coefficients" section provides the estimated coefficients for each independent variable and their standard errors, the z-values and corresponding p-values (Pr(>|z|)) that test the null hypothesis that the true coefficient is equal to zero, and the significance codes.
The "Null deviance" measures the error of the intercept-only model, while the "Residual deviance" measures the error of the full model.

The "AIC" is the Akaike Information Criterion, which is a measure of the model's goodness-of-fit that balances the model's complexity and its ability to fit the data. The model had 5 iterations of Fisher Scoring.

Model 2

```
Call: glm(formula = Exited ~ Balance + Age + Membership + Gender,
                    family = "binomial", data = data_train)
```

Deviance Residuals

Min	1Q	Median	3Q	Max
-2.1415	-0.6766	-0.4737	-0.2862	2.9134

Coefficients:

	Estimate	Std. Error	z value	Pr(>|t|)
(Intercept)	3.886e+00	1.396e-01	-27.842	<2e-16 ***
Balance	4.870e-06	5.324e-07	9.148	<2e-16 ***
Age	7.031e-02	3.037e-03	23.153	<2e-16 ***
Membership1	-1.079e+00	6.814e-02	-15.833	<2e-16 ***
GenderMale	-5.306e-01	6.423e-02	-8.261	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Null deviance: 7063.5 on 6999 degrees of freedom
Residual deviance: 6135.0 on 6995 degrees of freedom
AIC: 6145

Number of Fisher Scoring iterations: 5

Comparing the models

Both the models are logistic regression models that predict the likelihood of a customer leaving the bank (Exited). The first model (logistic.m1) uses 10 predictor variables: CreditScore, Geography, Gender, Age, Tenure, Balance, NumOfProducts, Credit_Card, Membership, and EstimatedSalary. The second model (logistic.m2) uses only 4 predictor variables: Balance, Age, Membership, and Gender.

Comparing the two models, the first model has a lower residual deviance and a lower AIC, indicating that it fits the data better than the second model. However, the second model may be preferred if the goal is to reduce the number of predictors, as it has fewer predictor variables. Additionally, the second model may be easier to interpret as it has fewer variables to consider.

Confusion Matrix

Training data

		Actual Values	
		No	Yes
Predicted Values	No	5398	1343
	Yes	188	71

Accuracy	0.7813	95% CI	(0.7714, 0.7909)
No Information Rate	0.798	P-Value [Acc > NIR]	0.9997
Kappa	0.0238	Mcnemar's Test P-Value	<2e-16
Sensitivity	0.05021	Specificity	0.96634
Pos Pred Value	0.27413	Neg Pred Value	0.80077
Prevalence	0.20200	Detection Rate	0.01014
Detection Prevalence	0.03700	Balanced Accuracy	0.50828
'Positive' Class	Yes		

The confusion matrix shows the results of a binary classification problem. The classifier is trying to predict if a customer will leave the bank or not (Yes or No).

5398 customers were correctly classified as "No".
1343 customers were incorrectly classified as "No", when they actually left the bank.
188 customers were incorrectly classified as "Yes", when they actually did not leave the bank.
71 customers were correctly classified as "Yes".
The accuracy of the classifier is 0.7813, which means that 78.13% of the time, the classifier correctly predicts if a customer will leave the bank or not.

The sensitivity (recall) of the classifier is 0.05021, meaning that only 5.021% of the customers that actually left the bank were correctly identified as "Yes".

The specificity of the classifier is 0.96634, meaning that 96.634% of the customers that did not leave the bank were correctly identified as "No".

The prevalence of the positive class (customers that left the bank) is 0.202, meaning that 20.2% of the customers in the train data left the bank. The balanced accuracy is 0.50828, meaning that the classifier performs equally well on both classes.
False negatives would be more damaging in this case, as they represent customers who have actually left the bank, but the model predicted they would not. This would mean that the bank would not take any action to retain these customers, causing a loss in revenue. On the other hand, false positives represent customers who have not left the bank, but the model predicted they would. This would lead to the bank wasting resources on trying to retain these customers, but not actually retaining them, as they did not leave in the first place.

Testing Data

		Actual Values	
		No	Yes
Predicted Values	No	2298	591
	Yes	79	32

Accuracy	0.7767	95% CI	(0.7613, 0.7915)
No Information Rate	0.7923	P-Value [Acc > NIR]	0.9831
Kappa	0.026	Mcnemar's Test P-Value	<2e-16
Sensitivity	0.05136	Specificity	0.96676
Pos Pred Value	0.28829	Neg Pred Value	0.79543
Prevalence	0.20767	Detection Rate	0.01067
Detection Prevalence	0.3700	Balanced Accuracy	0.50906
'Positive' Class	Yes		

ROC curve

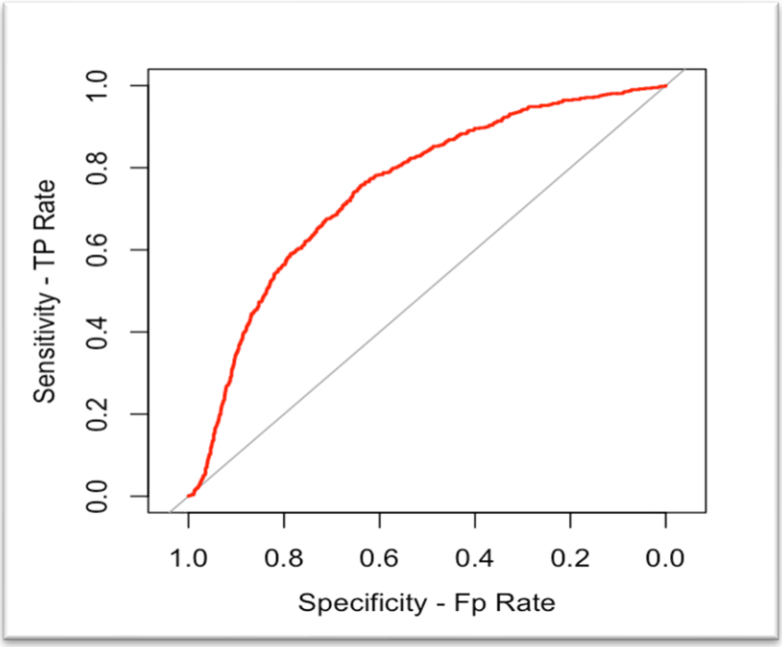

Figure 13 – ROC curve

The ROC (Receiver Operating Characteristic) curve is a tool used to evaluate the performance of a binary classification model. It plots the true positive rate (sensitivity) against the false positive rate (1-specificity) at various threshold settings. The closer the curve is to the top-left corner of the ROC space, the higher the model's accuracy in distinguishing between the two classes. A perfect classifier would have an ROC curve that hugs the top-left border, indicating 100% sensitivity and 100% specificity.

AUC
Area under the curve: 0.7485

The area under the curve (AUC) is a metric used to evaluate the performance of a binary classifier. In this case, the AUC value of 0.7485 indicates that the classifier has an average performance in distinguishing between the positive and negative classes. A value of 1 would indicate perfect performance and a value of 0.5 indicates a random classifier.

4. Regularization

Split the data into a train and test set

We randomly divided the data set into two parts, with 70% of the random data going into the training set and the remaining 30% going into the testing set. There are 7000 observations in the train data frame and 3000 observations in the test data frame.
To determine the churn rate, we set “Exited” as the dependent variable and we use all other variables as predictor variables. The x and y variables in the training and testing data sets correspond to the predictor and responding variables.

LASSO Logistic Regression

Lasso (Least Absolute Shrinkage and Selection Operator) regression method employs L1 regularization, which imposes a penalty equal to the absolute magnitude of the coefficients. It has the potential to reduce coefficients to zero, making it suitable for models with high levels of multicollinearity and eliminating selected features.

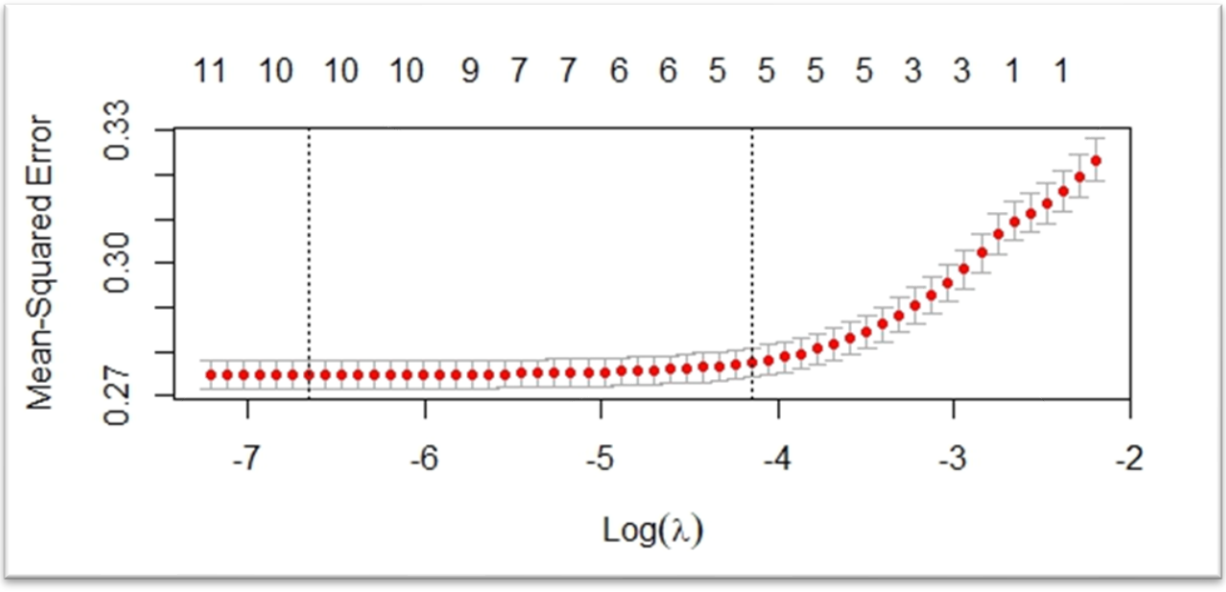

Figure 14 – LASSO diagram

The Best Lambda

log(λ$min) ①	-6.66026412	λ$min	0.00128080
log(λ$1se) ②	-4.14835311	λ$1se	0.01579039

Find best value of lambda using cross-validation

In order to fit the lasso regression, we need to find the best value of lambda first. We use the cross-validation method to find the best value of lambda by using the cv.glmnet( ) function and the training set of x and y. The non-zero coefficients on the top means the non-zero coefficients in the model for the particular value of lambda. The x-axis represents the value of the logarithmic of lambda and the y-axis represents the Mean-Squared Error. Each straight line with the red dot shows the confidence interval for the error estimate. The red dots are the error estimate. Red dots represent the loss metric which is computed through the cross-validation process. The left vertical dot lines represent the minimum value of the lambda, and the right vertical dot line is known as Lambda.1se, which represents the maximum value within 1 standard error of the minimum.

Interpretation:

- The left dot line has coefficient of 10, which means there are 10 non-zero coefficients in the model with minimum lambda.The minimum mean-square error can be considered at 10 features.
- The right dot line has coefficient of 5, which means there are 5 non-zero coefficients in the model with 1SE of lambda.The minimum logarithm value of lambda is -6.6603
- The logarithm of one standard error of lambda is -4.1484
- The minimum value of lambda is 0.0013
- The value of lambda at one standard error is 0.0158

Fit a model with the best Lambda in LASSO

We use glmnet ( ) function to fit the model on the training set using lambda.min and lambda.1se to get the coefficients tables. We can see the coefficients are different for lambda min model and lambda 1se model. By using lambda.1se, we eliminated more variables, and get a fitter model than using lambda min.

1）Model with minimum lambda

glmnet(x = train_x, y = train_y, alpha = 1, lambda = cv.lasso$lambda.min)			
	Df	%Dev	Lambda
1	10	14.38	0.001281

coef(model with min Lambda)			
13 x 1 sparse Matrix of class "dgCMatrix"	s0	λ$min	0.00128080 s0
(Intercept)	-8.837045e-02	CreditScore	-6.587130e-05
GeographyGermany	1.213995e-01	GeographySpain	3.299456e-03
GenderMale	-7.108533e-02	Age	1.071278e-02
Tenure	-7.654091e-04	Balance	2.711338e-07
NumOfProducts	-1.765865e-02	HasCrCard	.
IsActiveMember	-1.387397e-01	EstimateSalary	6.167052e-08

There are 10 variables with this minimum lambda model, and the minimum lambda is 0.00128. The table shows the coefficients on training test using lambda min in LASSO regression. The variables with coefficients equals to zero means there is high levels of multicollinearity, and we need to eliminate these variables. In this case, we eliminate “HasCrCard” from the model.

2) Model with one standard error of the minimum lambda

glmnet(x = train_x, y = train_y, alpha = 1, lambda = cv.lasso$lambda.1se)			
	Df	%Dev	Lambda
1	5	13.59	0.01579

coef(model with 1se Lambda)			
13 x 1 sparse Matrix of class "dgCMatrix"	s0	λ$min	0.01579 s0
(Intercept)	-1.177223e-01	CreditScore	.
GeographyGermany	9.513705e-02	GeographySpain	.
GenderMale	-4.309673e-02	Age	9.344845e-03
Tenure	.	Balance	1.661237e-07
NumOfProducts	.	HasCrCard	.
IsActiveMember	-1.088924e-01	EstimateSalary	.

There are 5 variables with this one standard error of the minimum lambda model, and the minimum lambda is 0.01579. The table shows the coefficients on training test using lambda 1se in LASSO regression. The variables with coefficients equals to zero means there is high levels of multicollinearity, and we need to eliminate these variables. In this case, we eliminate “CreditScore”, “GeographySpain”, “Tenure”, “NumOfProducts”, “HasCrCard”, “EstimatedSalary” from the model to get a more accurate model to predict the churn rate.r.

Ridge Logistic Regression

Ridge regression method employs L2 regularization, which adds a penalty equal to the square of the magnitude of the coefficients. Unlike Lasso, the coefficient of ridge regression will approach to zero but will not equal to zero, all coefficients are shrunk by the same factor (none are eliminated). Ridge regression is fitting for multicollinear models or when the number of predictors exceeds the number of observations.

Find best value of lambda using cross-validation

In order to fit the Ridge regression, we need to find the best value of lambda first. We use the cross-validation method to find the best value of lambda by using the cv.glmnet( ) function and the training set of x and y. The non-zero coefficients on the top means the non-zero coefficients in the model for the particular value of lambda. The x-axis represents the value of logarithmic of lambda and y-axis represents the Mean-Squared Error. Each straight line with the red dot shows the confidence interval for the error estimate. The red dots are the error estimate. Red dots represent the loss metric which is computed through the cross-validation process. The left vertical dot lines represent the minimum value of the lambda, and the right vertical dot line is known as Lambda.1se, which represents the maximum value within 1 standard error of the minimum.

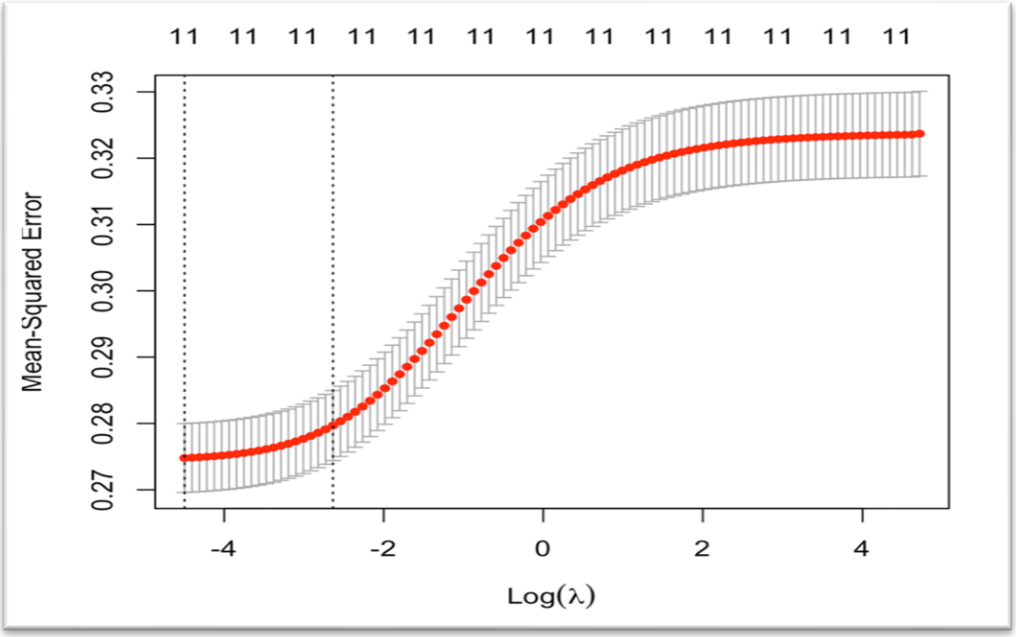

Figure 15 – Ridge diagram

The Best Lambda			
log(λ$min) ①	-4.49722296	λ$min	0.01113981
log(λ$1se) ②	-2.63655482	λ$1se	0.07160754

We found that:

- The left dot line has coefficient of 11, which means there are 11 non-zero coefficients in the model with minimum lambda.The minimum mean-square error can be considered at 11 features.
- The right dot line has coefficient of 11, which means there are 11 non-zero coefficients in the model with 1SE of lambda.
- The minimum logarithm value of lambda is -4.4972
- The logarithm of one standard error of lambda is -2.6366
- The minimum value of lambda is 0.0111
- The value of lambda at one standard error is 0.0716

Fit a model with the best Lambda in Ridge

We use glmnet ( ) function to fit the model on training set using lambda.min, and lambda.1se to get the coefficients tables. We can see the coefficients are different. The coefficients are smaller when we use the lambda 1.se than lambda.min.

The coefficients closer to zero, means the features need to be eliminated in order to regularize the model. From the coefficients table of Ridge regression on the training set for regularization, we found that only few variables cause the insignificant of the model. The higher the coefficients value the more significant to the model.

1) Model with minimum lambda

glmnet(x = train_x, y = train_y, alpha = 0, lambda = cv.ridge$lambda.min)			
	Df	%Dev	Lambda
1	11	14.39	0.001281

coef(model with min Lambda)			
13 x 1 sparse Matrix of class "dgCMatrix"		λ$min	0.001281
	s0		s0
(Intercept)	-8.010077e-02	CreditScore	-7.857682e-05
GeographyGermany	1.248050e-01	GeographySpain	7.619477e-03
GenderMale	-7.339789e-02	Age	1.079422e-02
Tenure	-1.206412e-03	Balance	2.795545e-07
NumOfProducts	-1.949586e-02	HasCrCard	-7.655719e-04
IsActiveMember	-1.409609e-01	EstimateSalary	8.373582e-08

From the table we found:

- This is coefficient of the model with minimum Lambda. There are 17 variables that contain every variable of the data set without deleting the variables.
- The higher the coefficients value the more significant to the model. From the model, we can see “EstimatedSalary” has been penalized the most.
- The value of minimum lambda in the minimum lambda model is 0.00128

2）Model with one standard error of the minimum lambda

glmnet(x = train_x, y = train_y, alpha = 0, lambda = cv.ridge$lambda.1se)			
	Df	%Dev	Lambda
1	11	14.39	0.01579

coef(model with 1se Lambda)			
13 x 1 sparse Matrix of class "dgCMatrix"		λ$min	0.01579
	s0		s0
(Intercept)	-7.024125e-02	CreditScore	-7.629348e-05
GeographyGermany	1.201053e-01	GeographySpain	5.791745e-03
GenderMale	-.7122500e-02	Age	1.040919e-02
Tenure	-1.151311e-03	Balance	2.850921e-07
NumOfProducts	-1.888965e-02	HasCrCard	-7.520278e-04
IsActiveMember	-1.355095e-01	EstimateSalary	8.028177e-08

There are 11 variables with this one standard error of the minimum lambda model, and the minimum lambda is 0.01579. The table shows the coefficients on training test using lambda 1se in Ridge regression. The variables with coefficients will not equals to zero, but will close to zero when there is high penalty, and it means there is high levels of multicollinearity. In this case, we eliminate “CreditScore” “Balance”, “EstimatedSalary” from the model are very close to zero.

5. Decision tree

A decision tree is a graphical representation of possible solutions to a decision based on certain conditions. It is a tree-like model of decisions and their possible consequences, including chance-event outcomes, resource costs, and utility.

Each internal node in the tree represents a "test" on an attribute (e.g., whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The topmost node in the tree is the root node. The decision tree is used in various fields such as machine learning, operations research, and decision analysis.

In the context of machine learning, decision trees are used as a predictive model for both classification and regression problems. The goal is to create a model that predicts the value of a target variable based on several input variables. The algorithm builds the tree by recursively splitting the data based on the feature that provides the most information gain (i.e., the feature that results in the most homogeneous groups of data).

Classification tree:

```
rpart(formula = Exited ~ Age + Balance + Geography + IsActiveMember
                         + NumOfProduct, data = data_train, method = “class”)
```

Variables actually used in tree construction:

[1] Age Balance Geography IsActiveMember NumOfProducts

Root node error: 1421/7000 = 0.203

n= 7000

	CP	nsplit	rel error	xerror	xstd
1	0.058761	0	1.0000	1.0000	0.0
2	0.033779	2	0.88248	0.88248	0.022578
3	0.029791	3	0.84870	0.85714	0.022321
4	0.027445	6	0.75932	0.79592	0.021671
5	0.012315	7	0.73188	0.73188	0.020941
6	0.010000	9	0.70725	0.72132	0.020816

A classification tree is a machine learning algorithm used to predict a categorical response variable based on one or more predictor variables. In the case of the output you provided, the categorical response variable is "Exited", and the predictor variables are "Age", "Balance", "Geography", "IsActiveMember", and "NumOfProducts". The decision tree algorithm works by recursively dividing the data into smaller and smaller subsets based on the values of the predictor variables, until the tree reaches a stopping criterion.

The "Variables actually used in tree construction" section lists the five predictor variables that are actually used to construct the tree. The "Root node error" indicates that the misclassification error rate at the root node is 1421/7000, or about 20.3%.

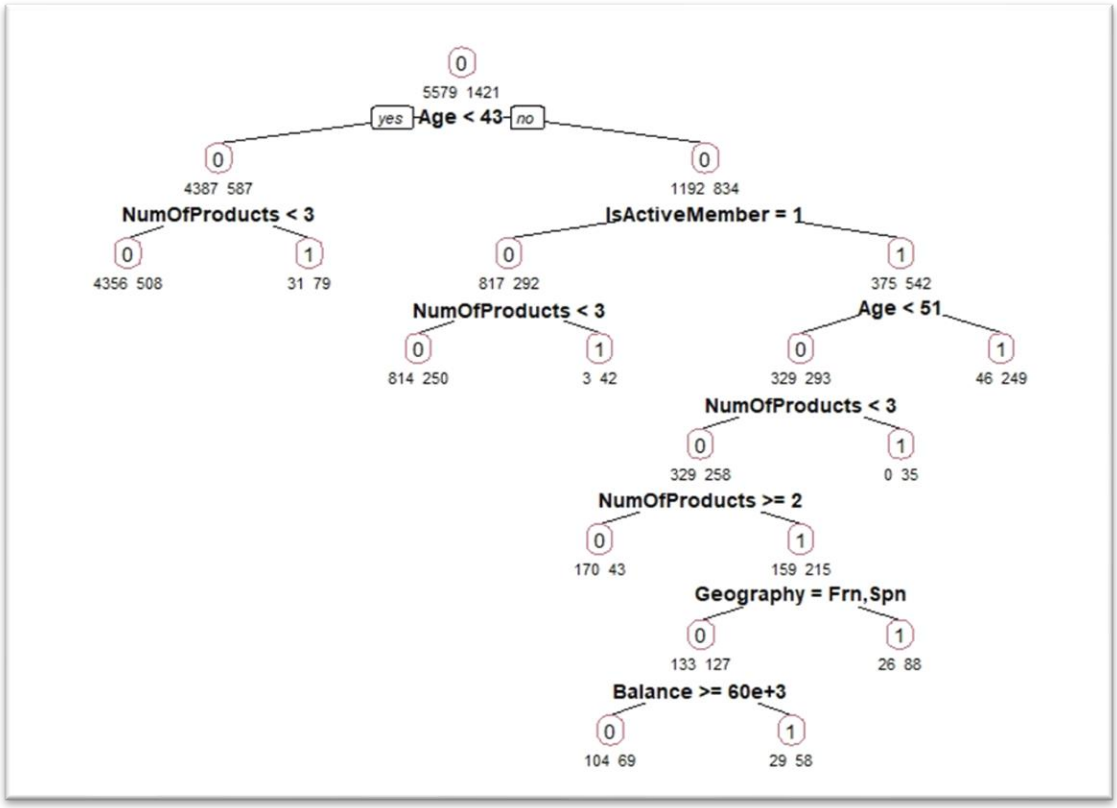

Figure 16 – Decision Tree

PART 5. ANSWERS for PRIMARY QUESTIONS

1. What are the main factors that are driving customers to churn?
    - Geography (German), Gender (Male), Age (Old), Membership (Not active)
2. What are the customer segments that are most likely to churn?
    - Most churn: German, female, older, not active
    - Barely churn: Not German, male, young, active
3. What are the most effective strategies to reduce customer churn?
    - Making membership status active
    - Other elements cannot be changed, but this element can be changed
4. What are the most effective methods to increase customer loyalty/engagement?
    - Using personalization to make customers feel special can help to increase their loyalty and engagement. This could involve personalizing emails, offering special discounts, or providing tailored recommendations
    - Offering rewards and loyalty programs to customers for their transactions and activities can encourage them to keep using the bank and its services.
    - Improving customer service can go a long way in increasing customer

loyalty and engagement. Making sure customers have all their questions answered quickly, have access to helpful customer service representatives, and generally have a pleasant experience when they interact with the bank can help to build trust and loyalty

- Offering new products and services that meet customers' needs and wants can help to keep them interested in the bank and its offerings

CONCLUSION

In conclusion, the "Churn Modeling" project aimed to predict the target variable "Exited" in a customer dataset, based on various predictor variables. Through the application of descriptive statistics, hypothesis testing, linear and logistic regression, decision tree analysis, and regularization techniques such as Lasso and Ridge, the project aimed to identify the most important drivers of customer attrition and prioritize initiatives aimed at reducing churn. The results of the analysis showed that the customer segments most likely to churn are German, female, older, and not active members. The most effective strategy to reduce customer churn was identified as making membership status active. Additionally, the project found that organizing events that may interest members and introducing benefits provided to active members could be effective methods to increase customer loyalty and engagement. The insights from this project can be used to inform targeted marketing campaigns, improve customer service, and create personalized retention strategies to reduce customer churn. Ultimately, the results of this project will provide valuable information to the credit card company and help them understand customer behavior and improve customer satisfaction, resulting in a lower rate of customer attrition.

REFERENCE

Northeastern. (2023, Jan). ALY 6015 - lesson 4-1,4-2, 4-3, 4-4, 4.5 — regularization . Retrieved from https://northeastern.instructure.com/courses/131212/pages/lesson-4-1-regularizati on?module_item_id=8227336

Rosane Rech. (2021, June 28). Pie-donut chart in R. Retrieved from https://statdoe.com/pie-donut-chart-in-r/

STHDA. (n.d.). ggplot2 legend : easy steps to change the position and the appearance of a graph legend in R software. Retrieved from http://www.sthda.com/english/wiki/ggplot2-legend-easy-steps-to-change-the-position-and-the-appearance-of-a-graph-legend-in-r-software

Vedda. (2018, May 31). Remove all of x axis labels in ggplot [duplicate]. Retrieved from https://stackoverflow.com/questions/35090883/remove-all-of-x-axis-labels-in-ggplot

STHDA. (n.d.). ggplot2 histogram plot : Quick start guide - R software and data visualization. Retrieved from http://www.sthda.com/english/wiki/ggplot2-histogram-plot-quick-start-guide-r-software-and-data-visualization

jmp. (n.d.). The one-sample t-Test. Retrieved from https://www.jmp.com/en_sg/statistics-knowledge-portal/t-test/one-sample-t-test.html

K. (2017, July 11). A gentle introduction to logistic regression and lasso regularisation using R. wordpress.com. Retrieved from https://eight2late.wordpress.com/2017/07/11/a-gentle-introduction-to-logistic-regression-and-lasso-regularisation-using-r/

finnstats. (2021, April 19). Decision trees in R. R-bloggers. Retrieved from https://www.r-bloggers.com/2021/04/decision-trees-in-r/

R-Code

```
# Installing and Loading the packages
library(dplyr)
library(corrplot)
library(GGally)
library(DAAG)
library(party)
library(rpart)
library(rpart.plot)
library(mlbench)
library(tree)
library(plotly)
library(ggcorrplot)
library(glmnet)
library(Metrics)
library(utils)
library(psych)
library(skimr)
library(wesanderson)
library(visdat)
library(grid)
library(gridExtra)
library(webr)
library(HH)
library(caret)
library(pROC)

# Importing the dataset
churn<-read.csv("churn_Modelling.csv")

# Checking for NA values
complete.cases(churn)
which(!complete.cases(churn))
vis_miss(churn)
sum(is.na(churn))
sum(is.null(churn))

# Checking for Duplication
duplicated(churn)
anyDuplicated(churn)
```

```
# Rounding the values for Balance and Estimated Salary
churn$Balance<-round(churn$Balance,0)
churn$EstimatedSalary<-round(churn$EstimatedSalary,0)

# Data Manipulation
churn$Status<-ifelse(churn$Exited=="1","Churned","Retained")
churn<-churn %>% mutate(HasCrCard = recode(HasCrCard, '0'='No', '1'='Yes'))
churn<-churn %>% mutate(IsActiveMember = recode(IsActiveMember, '0'='Inactive',
'1'='Active'))

# Changing the column names
colnames(churn)[11]<-"Credit_Card"
colnames(churn)[12]<-"Membership"

# Changing datatypes to factors
churn$Geography<-as.factor(churn$Geography)
churn$Gender<-as.factor(churn$Gender)
churn$Credit_Card<-as.factor(churn$Credit_Card)
churn$Membership<-as.factor(churn$Membership)
churn$Status<-as.factor(churn$Status)
churn$Tenure<-as.factor(churn$Tenure)
churn$NumOfProducts<-as.factor(churn$NumOfProducts)

# Dropping column Rownumber
drop<-c("RowNumber")
churn<-churn[,!(names(churn) %in% drop)]

# Understanding the dataset
str(churn)
View(churn)
glimpse(churn)
skim(churn)
headTail(churn,5)
dim(churn)
summary(churn)
describe(churn)
describeBy(churn,group="Geography")
describeBy(churn,group="Gender")

# EDA
# Histograms
hist5<-ggplot(churn, aes(x=CreditScore, fill=Status),labels=TRUE) +
```

```
geom_histogram(binwidth=50, alpha=.5,colour="black") +
scale_x_continuous(breaks=0:5)+ggtitle("Plot 1: Histogram of Credit Score by Status")+
geom_vline(aes(xintercept=mean(CreditScore, na.rm=T)),color="red",
linetype="dashed", size=1)
hist6<-ggplot(churn, aes(x=Age, fill=Status),labels=TRUE) +
geom_histogram(binwidth=5, alpha=.5,colour="black") +
scale_x_continuous(breaks=0:5)+ggtitle("Plot 2: Histogram of Age by Status")+
geom_vline(aes(xintercept=mean(Age, na.rm=T)),color="red", linetype="dashed",
size=1)
hist7<-ggplot(churn, aes(x=Balance, fill=Status),labels=TRUE) +
geom_histogram(binwidth=10000, alpha=.5,colour="black") +
scale_x_continuous(breaks=0:5)+ggtitle("Plot 3: Histogram of balance by Status")+
geom_vline(aes(xintercept=mean(Balance, na.rm=T)),color="red", linetype="dashed",
size=1)
hist8<-ggplot(churn, aes(x=EstimatedSalary, fill=Status),labels=TRUE)
+geom_histogram(binwidth=10000, alpha=.5,colour="black")
+scale_x_continuous(breaks=0:5)+ggtitle("Plot 4: Histogram of Estimated Salary by
Status")+ geom_vline(aes(xintercept=mean(EstimatedSalary, na.rm=T)),color="red",
linetype="dashed", size=1)
grid.arrange(hist5,hist6,hist7,hist8,top=textGrob("Histograms of Variables"))

# QQplots
qqnorm(churn$CreditScore, pch = 1, frame = FALSE,main="Q-Q Plot (Credit score)")
qqline(churn$CreditScore, col = "yellowgreen", lwd = 2)
qqnorm(churn$Age, pch = 1, frame = FALSE,main="Q-Q Plot (Age)")
qqline(churn$Age, col = "cornflowerblue", lwd = 2)
qqnorm(churn$Balance, pch = 1, frame = FALSE,main="Q-Q Plot (Balance)")
qqline(churn$Balance, col = "tomato2", lwd = 2)
qqnorm(churn$EstimatedSalary, pch = 1, frame = FALSE,main="Q-Q Plot (Estimated
Salary)")
qqline(churn$EstimatedSalary, col = "mediumorchid1", lwd = 2)

# Barplots
# Barplot 0 for Status
ggplot(churn, aes(x=Status, fill=Status)) + geom_bar() + geom_text(stat='count',
aes(label=..count..), vjust=-1)+scale_fill_manual(values=c("#56B4E9", "#E69F00"))
# Barplot 1 for Geography
ggplot(churn, aes(x=Geography, fill=Geography)) + geom_bar() +
geom_text(stat='count', aes(label=..count..), vjust=-
1)+scale_fill_manual(values=wes_palette(n=3, name="GrandBudapest1"))
# Barplot 2 for gender
ggplot(churn, aes(x=Gender, fill=Gender)) + geom_bar() + geom_text(stat='count',
```

```
aes(label=..count..), vjust=-1)+scale_fill_brewer(palette="Accent")
# Barplot 3 for HasCrCard
ggplot(churn, aes(x=Credit_Card, fill=Credit_Card)) +
geom_bar()+geom_text(stat='count', aes(label=..count..), vjust=-
1)+scale_fill_brewer(palette="Set1")
# Barplot 4 for Tenure
ggplot(churn, aes(x=Tenure, fill=Tenure)) + geom_bar()+geom_text(stat='count',
aes(label=..count..), vjust=-1)+scale_fill_brewer(palette="Set3")
# Barplot 5 for NumofProducts
ggplot(churn, aes(x=NumOfProducts, fill=NumOfProducts)) +
geom_bar()+geom_text(stat='count', aes(label=..count..), vjust=-
1)+scale_fill_manual(values=wes_palette(n=4, name="GrandBudapest2"))

# Piecharts
## categorical pie chart = HasCrcard
pie1 <- churn %>% group_by(Status, Credit_Card) %>% summarize(Freq=n())
PieDonut(pie1, aes(Credit_Card, Status, count=Freq), title = "Churned by Credit_Card")

## categorical pie chart = IsActiveMember
pie2 <- churn %>% group_by(Status, Membership) %>% summarize(Freq=n())
PieDonut(pie2, aes(Membership, Status, count=Freq), title = "Churned by
IsActiveMember")

# Age with his & box
Age.hist <- ggplot(churn, aes(x=Age, fill=Status, color=Status)) +
geom_histogram(position="identity", alpha=0.5)+
  theme(axis.title.x=element_blank())+ theme(legend.position = c(0.9, 0.5))
Age.box <- ggplot(churn, aes(x=Age, y=Status, fill=Status)) +
geom_boxplot(alpha=0.5)+theme(legend.position = "none")
grid.arrange(Age.hist,Age.box,nrow=2)

## Balance with his & box
B.hist <- ggplot(churn, aes(x=Balance, fill=Status, color=Status)) +
  geom_histogram(position="identity", alpha=0.5)+
  theme(axis.title.x=element_blank())+ theme(legend.position = c(0.9, 0.5))
B.box <- ggplot(churn, aes(x=Balance, y=Status, fill=Status)) +
  geom_boxplot(alpha=0.5)+theme(legend.position = "none")
grid.arrange(B.hist,B.box,nrow=2)

## Estimated Salary with his & box
S.hist <- ggplot(churn, aes(x=EstimatedSalary, fill=Status, color=Status)) +
  geom_histogram(position="identity", alpha=0.5)+
```

```
  theme(axis.title.x=element_blank())+theme(legend.position = c(0.9, 0.8))
S.box <- ggplot(churn, aes(x=EstimatedSalary, y=Status, fill=Status)) +
  geom_boxplot(alpha=0.5)+ theme(legend.position = "none")
grid.arrange(S.hist,S.box,nrow=2)

# 3d plot
plot_ly(churn, x = ~Age, y = ~EstimatedSalary, z = ~CreditScore,mode =
'markers',marker = list(size = 6),color = ~Status,alpha=0.9)

# Stacked bargraph
df1<-churn %>% group_by(Geography,Status) %>% summarise(count = n())
ggplot(df1, aes(y=count, x=Geography,fill=Status)) +
geom_bar(position="stack",stat="identity") + scale_fill_manual(values=wes_palette(n=2,
name="FantasticFox1"))
df2<-churn %>% group_by(Gender,Status) %>% summarise(count = n())
ggplot(df2, aes(y=count, x=Gender,fill=Status)) +
geom_bar(position="stack",stat="identity") +
scale_fill_manual(values=wes_palette(n=2, name="Cavalcanti1"))

# Correlation Matrix
corr <- select_if(churn, is.numeric)
cormatrix<-round(cor(corr,method = "pearson"),digits=2)
ggcorrplot(cor(corr,use = "complete.obs"),lab = TRUE,hc.order = TRUE)
ggpairs(corr)

# Hypothesis Testing
# One sample test
# null hypothesis: mean credit score is less than 600
# alternate hypothesis: mean credit score is <= 600
t.test(churn$CreditScore, mu=600, alternative = "less")

#null hypothesis: mean age is less than 50
#alternate hypothesis: mean age is >= 50
t.test(churn$Age, mu=50, alternative = "greater")

# Two Sample t test
# null hypothesis: both male and female have same mean credit score.
# alternate hypothesis: both male and female do not have same mean credit score.
Male<-subset(churn,Gender=="Male")
Female<-subset(churn,Gender=="Female")
t.test(Male$CreditScore, Female$CreditScore, var.equal=TRUE)
```

```
# F test
# null hypothesis: both male and female have same mean salary.
# alternate hypothesis: both male and female do not have same mean salary.
var.test(Male$EstimatedSalary, Female$EstimatedSalary, alternative = "two.sided")

# One-way Anova
# Null: There is no difference in mean of CreditScore according to Geography
# Alternative: At least one mean is different from the others (claim).
fit <- aov(Balance ~ Geography, data=churn)
summary(fit)
fit.tukey <- TukeyHSD(fit)

opar <- par(no.readonly = TRUE)
par(fig=c(0.2, 1, 0, 1))
plot(fit.tukey, col="blue", las=1)
par(opar)

balance.geo <- churn %>% group_by(Geography) %>%
  summarise(mean_Balance=mean(Balance),
            .groups = 'drop')
balance.geo

# Two-way Anova
# Null: There is no difference in mean of CreditScore according to Geography
# Alternative: At least one mean is different from the others (claim).
fit2 <- aov(Balance ~ Exited*Gender, data=churn)
summary(fit2)
attach(churn)

#plots
interaction2wt(Balance ~ Exited*Gender, data=churn)
detach(churn)

# Linear regression
cor(corr,use = "complete.obs")
linear_model <- lm (Exited~Balance+Age+NumOfProducts, data = churn)
summary(linear_model)

#Logistic Regression
set.seed(3456)
split.data1 <- createDataPartition(churn$Exited, p = 0.7, list = FALSE, times = 1)
data_train <- churn[ split.data1,]
```

```
data_test <- churn[-split.data1,]

logistic.m1 <- glm(Exited~CreditScore + Geography + Gender + Age + Tenure +
                   Balance + NumOfProducts + Credit_Card + Membership +
EstimatedSalary, data=data_train, family = "binomial")
summary(logistic.m1)

logistic.m2 <- glm(Exited~Balance + Age + Membership + Gender , data=data_train,
family = "binomial")
summary(logistic.m2)

#Confusion matrix
#train data
prob.train <- predict(logistic.m2, newdata=data_train, type="response")
predicted <- as.factor(ifelse(prob.train>=0.5, "Yes", "No"))
data_train$Exited <- as.factor(data_train$Exited)
data_train$Exited <- factor(ifelse(data_train$Exited==1, "Yes", "No"))
confusionMatrix(predicted, data_train$Exited, positive = "Yes")

#test data
prob.test <- predict(logistic.m2, newdata = data_test, type="response")
predicted1<- as.factor(ifelse(prob.test>=0.5, "Yes", "No"))
data_test$Exited <- as.factor(data_test$Exited)
data_test$Exited <- factor(ifelse(data_test$Exited==1, "Yes", "No"))
confusionMatrix(predicted1, data_test$Exited, positive = "Yes")

ROC <- roc(data_train$Exited, prob.train)
plot(ROC, col="red", ylab="Sensitivity - TP Rate", xlab= "Specificity - Fp Rate")

#AUC
AUC1 <- auc(ROC)

# Importing the dataset again
churn2<-read.csv("churn_Modelling.csv")

# LASSO Logistic Regression
set.seed(3456)
str(churn2)
drop<-c("CustomerId", "Surname","RowNumber")
churn2<-churn2[,!(names(churn2) %in% drop)]

split.data1 <- createDataPartition(churn2$Exited, p = 0.7, list = FALSE, times = 1)
```

```
data_train2 <- churn2[ split.data1,]
data_test2 <- churn2[-split.data1,]
train_x <- model.matrix(Exited~.,data_train2)
train_y <- data_train2$Exited

test_x <- model.matrix(Exited~.,data_test2)
test_y <- data_test2$Exited

cv.out <- cv.glmnet(train_x,train_y,alpha=1, family="binomial",type.measure = "mse")
summary(cv.out)
plot(cv.out)

# optimal value of lambda; minimizes the prediction error
# lambda min - minimizes out of sample loss
# labmda 1se - largest value of lambda within 1 standard error of lambda min
log(cv.out$lambda.min)
log(cv.out$lambda.1se)

cv.out$lambda.min
cv.out$lambda.1se

##########
# Fit models based on lambda
##########

# Fit the finla model on the training data using lambda.min
# alpha = 1 for Lasso (L1)
# alpha = 0 for Ridge (L2)
lasso.model.min <- glmnet(train_x, train_y, alpha = 1, lambda = cv.out$lambda.min)
lasso.model.min

# Display regression coefficients
coef(lasso.model.min)

# Fit the final model on the training data using lambda.1se
lasso.model.1se <- glmnet(train_x, train_y, alpha =1, lambda = cv.out$lambda.1se)
lasso.model.1se

# Display regression coefficients
coef(lasso.model.1se)

# Ridge Logistic Regression
```

```
set.seed(3456)
str(churn2)

cv.out2 <- cv.glmnet(train_x,train_y,alpha=0, family="binomial",type.measure = "mse")
summary(cv.out2)
plot(cv.out2)

# optimal value of lambda; minimizes the prediction error
# lambda min - minimizes out of sample loss
# labmda 1se - largest value of lambda within 1 standard error of lambda min
log(cv.out2$lambda.min)
log(cv.out2$lambda.1se)
cv.out2$lambda.min
cv.out2$lambda.1se

##########
# Fit models based on lambda
##########

# Fit the finla model on the training data using lambda.min
# alpha = 1 for Lasso (L2)
# alpha = 0 for Ridge (L1)
ridge.model.min <- glmnet(train_x, train_y, alpha = 0, lambda = cv.out$lambda.min)
ridge.model.min

# Display regression coefficients
coef(ridge.model.min)

# Fit the final model on the training data using lambda.1se
ridge.model.1se <- glmnet(train_x, train_y, alpha =0, lambda = cv.out$lambda.1se)
ridge.model.1se

# Display regression coefficients
coef(ridge.model.1se)

install.packages("rpart.plot")
library(rpart)
library(rpart.plot)
D.tree = rpart(Exited ~Age+Balance+Geography+IsActiveMember+NumOfProducts,
data = data_train, method = "class")
printcp(D.tree)
prp(D.tree, type = 2, extra = 1, under = TRUE, split.font = 2,border.col = 2, varlen = 0)
```