EXPLORING AND ANALYZING THE FACTORS THAT CONTRIBUTE TO THE CUSTOMER CHURN IN BANK

Project By – M.R. USMAAN

BACHELOR OF SCIENCE (HONS) IN DATA SCIENCE

NATIONAL INSTITUE OF BUSINESS MANAGEMENT

COLOMBO, SRI LANKA

Acknowledgement

I would like to express my thankfulness to my lecturers, colleagues, and family who contributed to the success of this research project. First and foremost, I convey special gratitude to my project research supervisor Mr. Thurairasa Balakumar for his guidance and support and next to that I would like to thank my course director Mrs. W.M.S.G.D.C Wanigasekara for her support and guidance. I am thankful to my colleagues for their ideas and suggestions that they gave whenever I needed. Finally, I am grateful to my family for their support.

Executive Summary

Main scope of this research is to explore and analyze the factors that contribute to customer churn in bank organizations. To this research I got secondary dataset so, I have some limitations those are, dataset did not mention the bank properly and currency they have used in dataset, according to this dataset's bank has four types of services/products but they did not mention that what are the services are that and, according to this data set I assume there should be more factors because, there is no positive correlations between numerical variables. For my research I have used logistic regression, correlation and anova tests, with these tests I have used data visualization and descriptive statistics too. At the end of my research, I got to know that what are the factors contribute to the customer churn, there is no correlation between customer's bank balance and points that customer earned and sample mean of activeness and sample mean of customer bank balance are equal. I recommend future researchers to focus on my limitations and collect the data from more than one bank. And research and analyze on each factor not only to focus on churn it will help to know more things. Finally, I recommend that not to use or extend this data set use another or collect new dataset to future research.

Abstract

The main component of the bank is the customer. Therefor the knowledge of customer behavior and customer retention is very important to bank organizations. Some customers stopped the relationship with the bank for some reason. So, this study identify that which factors contribute to the customer churn and this research has been mainly focused on Europe banks. Additionally, this research has been checked whether there is an association between the bank balance of the customer and points that customer earned and significant mean difference between bank balance and activeness of customer. To predict the customer churn, I have used logistic regression model and additionally I have used ANOVA and correlation test for additional research.

Table of Contents

Chapter 1: Introduction 1
1.1 Background 1
1.2 Research Problem 1
1.3 Objectives 1
1.4 Research Questions 2
1.5 Scope of the Research 2
1.6 Justification of the Research 2
1.7 Expected Limitations 3
1.8 Proposed Work Schedule 3
Chapter 2: Literature Review 4
2.1 Introduction to the research theme 4
2.2 Theoretical Explanation About the Key Words in the Topic 4
2.3 Finding by Other Researchers 5
2.4 Research Gap 8
2.5 Table of Variables 8
2.6 Chapter Conclusion 10
Chapter 3: METHODOLOGY 11
3.1 Introduction 11
3.2 Population, Sample, and Sampling Technique 11
3.3 Type of Data to be Collected and Data Sources 11
3.4 Data Collection Tools and Plan 11
3.5 Conceptual Framework 11
3.6 Hypothesis 12
3.7 Operationalization Table 13
3.8 Methods of Data Analysis 14
Chapter 4: Data Pre-Processing, Data Analysis & Data Visualization 15
4.1 Introduction 15
4.2 Data Pre-Processing 15
4.3 Descriptive Statistics and Data Visualization 18
4.3.1 Data Visualization for Categorical Variables 18
4.3.2 Data Visualization and Distribution of Numerical Variables 24

4.3.3 Comparison of Categorical and Numerical Variables 29
4.4 Statistical Test to Research 32
4.4.1 Logistic Regression 32
4.4.2 Correlation Test between Point Earned and Bank Balance 34
4.4.3 One-Way ANOVA Test between Activeness and Bank Balance 34
Chapter 5: Discussion and Recommendation 36
5.1 Discussion 36
5.2 Recommendation 36
5.3 Conclusion 37
References 38

Table of Figures

Figure 1NULL values of dataset 15
Figure 2 Structure of the dataset before convert classes 16
Figure 3 Converting classes 16
Figure 4 Age and Credit Score 's boxplots before remove the outliers 16
Figure 5 Age and Credit Score 's boxplots after remove the outliers 17
Figure 6 Converting string values to numerical values 17
Figure 7 Pie chart for gender 18
Figure 8 Pie chart for churn 18
Figure 9 Pie chart for geography 19
Figure 10 Bar chart to number of products 19
Figure 11 Count of products 20
Figure 12 Bar graph and count of credit card customers 20
Figure 13 Bar graph and count of activeness of customers 21
Figure 14 Bar graph and count of customer complain 21
Figure 15 Bar graph of satisfaction score 22
Figure 16 Bar graph for Satisfaction Score with Churn 23
Figure 17 Summary of categorical variables 23
Figure 18 Summary of numerical variables 24
Figure 19 Histogram of Credit score 24
Figure 20 Histogram of Points that Customer Earned 25
Figure 21 Histogram of Age 26
Figure 22 Histogram of Tenure 27
Figure 23 Histogram of Balance 27
Figure 24 Histogram of Salary 28
Figure 25 Correlation plot to numerical variables 29
Figure 26 Boxplot Balance vs Churn 29
Figure 27 Boxplot Tenure/Salary/Credit Score/Points earned vs Churn 30
Figure 28 Boxplot Age vs Churn 31
Figure 29 Accuracy of model 32
Figure 30 Confusion matrix 32
Figure 31 Logistic regression result 33
Figure 32 correlation result 34
Figure 33 Scatter plot points earned vs balance 34
Figure 34 ANOVA result 34

Table of Table

Table 1 Work Schedule 3
Table 2 Key Word Explanation 5
Table 3 Table of Variables 9
Table 4 Table of Hypothesis Testing 12
Table 5 Operationalization Table 14
Table 6 Data Types 15
Table 7 Count of Products 20
Table 8 Table of Credit Card Customers with Churn 20
Table 9 Table of Activeness of Customers with Churn 21
Table 10 Table of Complain with Churn 22
Table 11 Count of Satisfaction Score 22
Table 12 Table of Score Range 25
Table 13 Table of Point Range 25
Table 14 Table of Age Range 26
Table 15 Table of Tenure Range 27
Table 16 Table of Balance Range 28
Table 17 Customer Salary Range 28
Table 18 Predicted Observations 33

Chapter 1: Introduction

1.1 Background

Currently the market is very highly competitive because of the availability of vast service providers, when we focus to bank organizations there are lot of competitors. The main component of the bank is the customers, and the bank depends on them. According to U.S 81.5 percent of U.S. households (approximately 107.9 million) were fully banked in 2021.Worldwide account ownerships have reached 76 percent in world population and 71 percent of people in developing countries. Change in customer behavior will highly effect the bank organizations. So, when customer end their relationship with a bank like closing their account and discontinuing specific bank products or services, it will have high negative effect to bank.

Bank organizations must understand the customer behavior and they need to know which factors effect to customer churn. In depth, it is important to analyze that which customer demographic group churns and whether satisfaction level effects on churn. So, focusing these the proposed research seeks to explore and analyze factors that contribute to the customer churn. So this study is going to predict whether customer will churn or not and going to evaluate the prediction results.

1.2 Research Problem

At present, customers are stopping and quitting their relationships with bank and they use alternative ways for their financial needs. For an example they invest their money in trading and real estates and they invest in cryptocurrency. So, this activity of consumers hast the biggest negative impact in bank organizations. The goal of this research study is to explore the customer groups and their behavior to analyze customer churn and identify the key factors that influence churn.

1.3 Objectives

The objectives for this project that are derived from research problem are list down below,

I. Investigate the factors that affect customer churn.

II. Investigate whether there is an association between bank balance and points that customer earned.
III. Investigate the means different between activeness of customer and bank balance.

1.4 Research Questions

I. What are the factors that contribute to the customer churn in bank?
II. Is there a significant difference in the mean of bank balance and activeness of customer?
III. Does there is an association between bank balance and points that customer earned?

These questions will be answered at the conclusion of this research.

1.5 Scope of the Research

The main scope of this study is to reduce customer churn in bank organizations by the help of exploring the factors that contribute to the churn. So, this study tells us how demographic factors and other factors affect churn. Even though research focus on Europe banks it can help to better understand the customer churn for worldwide banks and this study will show the correlation between bank balance and points earned by customer in specific bank.

1.6 Justification of the Research

This proposed research justified by falling of profit and loosing consumers in bank organizations. The main components of bank is customer. So, when they loose customer it will be huge effect to organization. Therefor every bank must know the consumer behavior and why they churn.

Bank should not be only profit making industry it should be peoples' well wisher. So, they need to know how to satisfy the customers. Bank accounts are safe and they follow highly secured protocols so nowadays it is one of the best option to handle money. So, this have benefits for account holders too. So, bank should focus the customer well also.

1.7 Expected Limitations

In this research study I get secondary data set from Kaggle. So, the details that are in dataset collected anonymously by a user in effective way and this data set has 10.00 usability in Kaggle but we cannot be sure about that every details are hundred percent honest. Specially this data has only Europe banks details and user did not mention the bank also.

1.8 Proposed Work Schedule

Date	**Work**
02nd of July 2023	Started to find a dataset
12th of July 2023	Finalize the data set and research topic
17th of July 2023	Proposal presentation submission
19th of July 2023	Introduction
20th – 21st of July 2023	Study articles and research papers and then Literature review
22nd of July 2023	Methodology and finalize the proposal
23rd of July 2023	Proposal submission
26th – 1st of August 2023	Data cleaning and Data preprocessing
20th - 25th of August 2023	Data Visualization and statistics tests.
12th of September 2023	Finalize the report
15th of September 2023	Report Submission
18th of September 2023	Viva

Table 1 Work Schedule

Chapter 2: Literature Review

2.1 Introduction to the research theme

One of the most important service providers are bank organizations. This organizations' most important component is customers. So, banks mostly depend on their customers. When banks well know the customer behaviors and the factors that have relationship and association with customer it is best strategy to achieve their goals. The theme of the research aims to focus on explore and analyze the factors that contribute to customer churn. So, to gain valuable insights about customer churn this research will explore the relationship between customers characteristics and churn.

2.2 Theoretical Explanation About the Key Words in the Topic

Key Words	**Explanation / Definition**
Customer Churn	This refers to customer who leaves one company to go to another company. (Sharma and Panigrahi (2011))
Electronic Banking	Initial alternative channels to the traditional bankbranches. (Liebana-Cabanillas et al. (2013))
Customer Loyalty	Deeply held commitment to re purchase or re patronize a preferred product /service consistently in the future.
Customer behavior	Study of how individuals, groups and organizations select, buy, use and dispose of goods, services, ideas or experiences to satisfy their needs and wants (Kotler and Keller,2006)

Customer Satisfaction	Extent to which customers' expectations and needs are met or exceeded by a product or service. (P. Kotler & G. Armstrong (2012))
Service quality	Refers to the extent to which a bank's services meet or exceed customers' expectation and needs (A. Parasuraman (1988))
Islamic banking	This is a financial systems that operates in accordance with the principles of Islamic law (M.N. Siddiqi (1983))
Multiple linear regression	This is a form of predictive modelling techniques that identify the relationships between dependent and independent variables. (Eric Goh Ming Hui (2019))
Six Sigma	This is a statistical measure of how far a process deviates from perfection.
Service Delivery Technology (STD)	SDT is the end to end process of providing a services to customers by using technology (ATM, Internet banking)

Table 2 Key Word Explanation

2.3 Finding by Other Researchers

1. How customer satisfaction changes behavior: A case study of banking industry

Bank industry have lot of competition so they need to focus on customer retention for that industry must learn about customer behavior. So, researchers implemented their research study in two Iranian banks. This research investigated the relationship between seven aspects of banking services and customers, those are how bank employees' treat customers, service prices, how to promote and market synergies, place and time to serve customers,

products, equipment and process. This study mainly focused in treating behavior of employees toward customers So if they do not treat properly customers will shift the bank.

2. Credit card churn forecasting by logistic regression and decision tree.

In this study credit card data collected from a real Chinese bank to predict credit card churn. They have designed 135 variables to summarize the behaviors and choices of the credit card users. After considering multicollinearity 95 variables are chosen to build the model. So finally this research have shown that the demographic information makes little contribution to the churn prediction.

3. Developing a prediction model for customer churn from electronic banking services using data mining

This research article also focuses on exploring and analyzing the factors that contribute to customer churn by using data mining techniques. This study emphasizes the significance of the customer retention as a fundamental requirement for bank. Researchers identify the characteristics of churned customers by extracting customer information, card information, risk information and transaction activity information from bank's database.

4. Predicting customer churn in electronic banking

The specific focus on this study is electronic churn. Electronic churn means the customers leaving electronic channels. So this behavior may or may not indicate they have churned from the bank. At present mostly banks focus to increase digital customers because its is a strategic goal. So, article highlights the importance of electronic churn prediction in the bank.

5. Machine Learning to Develop Credit Card Customer Churn Prediction.

In this study they have used feature selection method and five machine learning models to develop a churn prediction model. According to this study research most important variables for developing a credit card customer churn prediction model were the total transaction count, total revolving balance on the credit card, and the change in the transaction count. In this study researcher highlights the importance of feature selection to improve the churn prediction.

6. Analyzing the effect of service quality on customer satisfaction of Islamic Banking in Indonesia.

This article has investigated the factors that contribute to customer churn in Islamic banking. For this research researcher used multiple linear regression analysis for data analysis. The result of this research shows that service quality variables (tangible, reliability, assurance, empathy and responsive) influence the customer satisfaction in Islamic banking. The analysis reveal 70.1% customer satisfaction is explained by service quality while the remaining 29.9% influenced by other unexamined variables.

7. Impact of electronic banking on customer satisfaction in Ethiopian banking industry

This study specially focused in e-banking and it has revealed that majority of electronic bank users are young, educated, salaried individuals and students. According to this research e-banking positively effect customer satisfaction by reducing branch visits and waiting times. Research also highlight the demographic characteristics of customers influence the customer satisfaction.

8. Research on using Six Sigma management to improve bank customer satisfaction

This article has research to improve customer satisfaction level in bank using six sigma management. This article has missed to define the factors that affect the customer satisfaction score, but they have suggest some issues that customers face and the issues that affect customer satisfaction. Those issues are long waiting times for customers, lack of staffs in bank and other common issues that bank faces. In this research they have used six sigma infrastructure to analyze and improve the customer satisfaction level.

9. The 'Net Effect' on Bank Customer Complaining Behavior.

According to this research article, The arrival of online banking has saved the costs of banks and unexperienced convenience to consumers. There are lot of advantages in online banking but there are struggles for bank too. This research finding reveals that there are lot of usual problems in online banking and there is a high rate of customer complains but interestingly complains have made in traditional way like telephone calls to voice their complains rather than online complains. According to this study rate of complains in online

banking are greater than traditional banking. Technical issue is the root cause for this problem and bank is responsible to it according to consumers' view.

10. Importance-performance analysis of UK and US bank customer perceptions of service delivery technologies.

This research study address the question that how customers perceive and evaluate the self delivery technology in banking sector. Research have use important performance analysis to understand which factors that customer consider crucial when evaluating the performances of STD's. And according to this research organizational strategists are advised to continuously customers perceptions and adapt to changing expectations.

2.4 Research Gap

There are some places where above studies either did not research or get solution. In past research articles some of those focused on electronic banking and balance researches have been focused on credit card customers. When we look into demographic characteristics in past articles they have been missed to focus those factors properly. These researches have missed to research whether customer satisfaction level and relationship period between customer and bank contribute to churn or not. And they have missed to check whether bank balance and customer salary effect the customer churn.

2.5 Table of Variables

Variable Name	**Explanation**
CustomerID	Unique IDs for bank customers
Surname	Last name of customer
CreditScore	Score earned from customer's credit behavior

Geography	Country where customer belongs
Gender	Gender of customer
Age	Age of customer
Tenure	Number of years that customer have relationship with bank
Balance	Customer's bank balance
NumOfProducts	Number of products that customer purchased through a bank
HasCrCard	Whether customer is a credit card holder or no
IsActiveMember	Whether customer is an active member or no
EstimatedSalary	Estimated salary of the customer in euro
Exited	Whether customer churn or no
Complain	Whether customer have complain to bank or no
Satisfaction Score	Satisfaction level of the customer(scale – 1 to 5)
Point Earned	Points earned by customer

Table 3 Table of Variables

Note – Sources for these variables are get from Kaggle (https://www.kaggle.com/datasets/radheshyamkollipara/bank-customer-churn/discussion/422316)

2.6 Chapter Conclusion

In this chapter I studied previous research articles.SO, with the help of analyzing and machine learning techniques banks can predict the customer churn effectively. According to those articles the knowledge of customer behavior and customer retainment is most important to retain bank organizations. Some of these articles identified the some factors that contribute to customer churn but they missed some factors. So, those missing factors will be analyzed in research.

Chapter 3: METHODOLOGY

3.1 Introduction

This chapter aims to analyze the factors that contribute to customer churn in bank and focuses on the sampling technique, type of data to be analyze, data collection tools and plan, conceptual framework, hypothesis test for research and methods of data analysis. So, for research methodology is going to done with the huge help of statistical knowledge.

3.2 Population, Sample, and Sampling Technique

I have downloaded the dataset from Kaggle. This dataset was uploaded by Radeshyam Kollipara. The data has been obtained from Europe banks. So, population of this study may be Europe banks that data collected from. Uploader did not mention the bank name or names.

3.3 Type of Data to be Collected and Data Sources

In this dataset there are qualitative and quantitative data. Categorical variables are surname, geography, gender, number of products, has credit card, is active member, exited, complain and satisfaction score. Numerical variables are customer id, credit score, age, balance, estimated salary, and point earned. Data collection method did not mentioned by uploader but this dataset published on Kaggle which is world largest data science community and main platform to get secondary datasets. So, this dataset is secondary and it is gathered in csv format.

3.4 Data Collection Tools and Plan

This dataset obtained from Kaggle . If anyone interested can get the dataset from https://www.kaggle.com/datasets/radheshyamkollipara/bank-customer-churn/discussion/422316 this link. There is no sufficient information about the data collection tools and plan from the uploader.

3.5 Conceptual Framework

I. Identifying the relationship between customer churn and customers' characteristics.
    - Dependent variable - Exited.

- Independent variables – Geography, Gender, Age, Tenure, Balance, HasCrCard, IsActiveMember, Complain, and Satisfaction score.

II. Finding whether there is an association between bank balance and point earned.
    - Dependent – Point Earned
    - Independent – Balance

III. Checking whether there is a significance difference in the mean of balance and activeness of customer.
    - Dependent – Balance
    - Independent – IsActiveMember

3.6 Hypothesis

Test	Null Hypothesis	Alternative Hypothesis
Identifying the relationship between customers' characteristics and churn	There is no statistically significant relationship between X and Y	There is a statistically significant relationship between X and Y
Checking whether there is an association between bank balance and point earned	There is no association between X and Y	There is an association between X and Y
Checking whether there is a significance difference in the mean of balance and customer activeness	Means are equal $\overline{X} = \overline{Y}$	Means are not equal $\overline{X} \neq \overline{Y}$

Table 4 Table of Hypothesis Testing

Note - X = Independent Variable, Y = Dependent Variable

3.7 Operationalization Table

Variable	Indicators	Measures
CustomerID	Primary Key – Unique ID	-
Surname	Name	Categorical
CreditScore	Score that customer earn for credit loyality	Numerical
Geography	Region or country customer belongs to	Categorical
Gender	Male or Female	Categorical
Age	Age	Numerical
Tenure	Years	Numerical
Balance	Euro	Numerical
NumOfProducts	Count	Numerical
HasCrCard	Yes or no	Categorical
IsActiveMember	Yes or No	Categorical
EstimatedSalary	Euro	Numerical
Exited	Churn or Not	Categorical

Complain	Yes or No	Categorical
Satisfaction Score	Points	Numerical
Point Earned	Points	Numerical

Table 5 Operationalization Table

3.8 Methods of Data Analysis

So, qualitative, and quantitative data that are in my dataset will be analyzed using Python and R according to statistical knowledge. So graphical representation and statistical tools will be used for analyzing purposes. So, data will be analyzed by data visualization (pie charts, bar graphs and etc.), descriptive statistics (mean, median, maximum, minimum and etc.), ANOVA (Analysis of Variance) test, Correlation test and logistic regression model.

Therefor in my objectives I will use logistic regression model to predict whether customer will churn or not, use anova test for check whether there are significant different between bank balance and activeness of customer and chi-squared test for check whether there is an association between bank balance and credit score of customers.

Chapter 4: Data Pre-Processing, Data Analysis & Data Visualization

4.1 Introduction

This section provides a brief idea about variables that are in data set. So, data visualization helps to study about variables and descriptive statistics provides a summary of characteristics in data set. Correlation heatmap helps to know the relationship between numeric variables.

In this data set there are categorical and numerical variables.

Categorical Data	Numerical Data
Geography	CreditScore
Gender	Age
HasCrCard	Tenure
IsActiveMember	Balance
Exited	NumofProducts
Complain	EstimatedSalary
Satisfaction.Score	Point.Earned

Table 6 Data Types

4.2 Data Pre-Processing

First of all I checked for null values and I did not get any null values in my dataset.

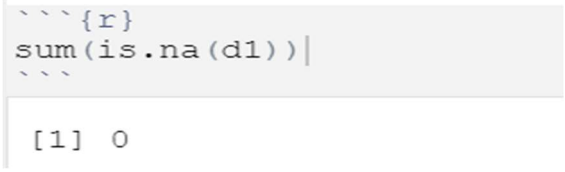

Figure 1NULL values of dataset

And after that I checked for structure,

```
$ Surname            : chr  "Hargrave" "Hill" "Onio"
$ CreditScore        : num  619 608 502 699 850 645
$ Geography          : Factor w/ 3 levels "France","
$ Gender             : Factor w/ 2 levels "Female","
$ Age                : num  42 41 42 39 43 44 50 44
$ Tenure             : num  2 1 8 1 2 8 7 4 2 6 ...
$ Balance            : num  0 83808 159661 0 125511
$ NumOfProducts      : Factor w/ 4 levels "1","2","3
$ HasCrCard          : Factor w/ 2 levels "0","1": 2
$ IsActiveMember     : Factor w/ 2 levels "0","1": 2
$ EstimatedSalary    : num  101349 112543 113932 938
$ Exited             : Factor w/ 2 levels "0","1": 2
$ Complain           : Factor w/ 2 levels "0","1": 2
$ Satisfaction.Score: Factor w/ 5 levels "1","2","3
$ Point.Earned       : num  464 456 377 350 425 484
```

Figure 2 Structure of the dataset before convert classes

Converting the Classes

First, I convert every variable to numerical and factor.

```
d1$Geography<- as.factor(d1$Geography)
d1$Gender <- as.factor(d1$Gender)
d1$Satisfaction.Score <- as.factor(d1$Satisfaction.Score)
d1$HasCrCard<- as.factor(d1$HasCrCard)
d1$IsActiveMember <- as.factor(d1$IsActiveMember)
d1$Exited <- as.factor(d1$Exited)
d1$Complain <- as.factor(d1$Complain)
d1$NumOfProducts<-as.factor(d1$NumOfProducts)
d1$CreditScore<-as.numeric(d1$CreditScore)
d1$Age<-as.numeric(d1$Age)
d1$Tenure<-as.numeric(d1$Tenure)
d1$Balance<-as.numeric(d1$Balance)
d1$EstimatedSalary<-as.numeric(d1$EstimatedSalary)
d1$Point.Earned<-as.numeric(d1$Point.Earned)
```

Figure 3 Converting classes

Checking for Outliers and Removing it.

So, when I was checking the outliers in numerical variables I got outliers in my two numerical variables that are age and credit score.

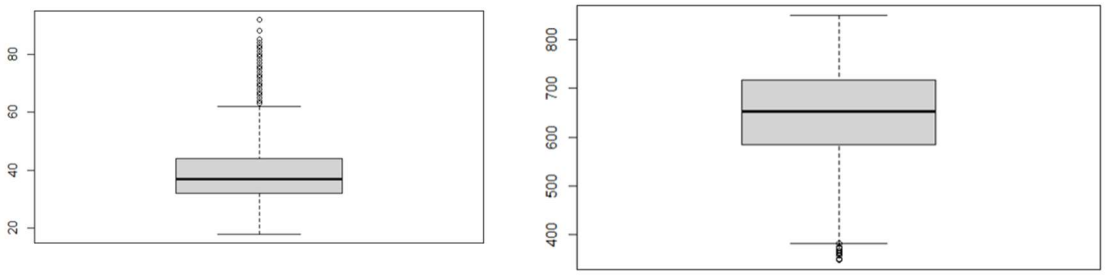

Figure 4 Age and Credit Score 's boxplots before remove the outliers

After that I removed the outliers of these two variables using IQR(Inter Quartile Range) method. So, at the beginning I had 10000 entries and after removed the outliers I got 9459 entries.

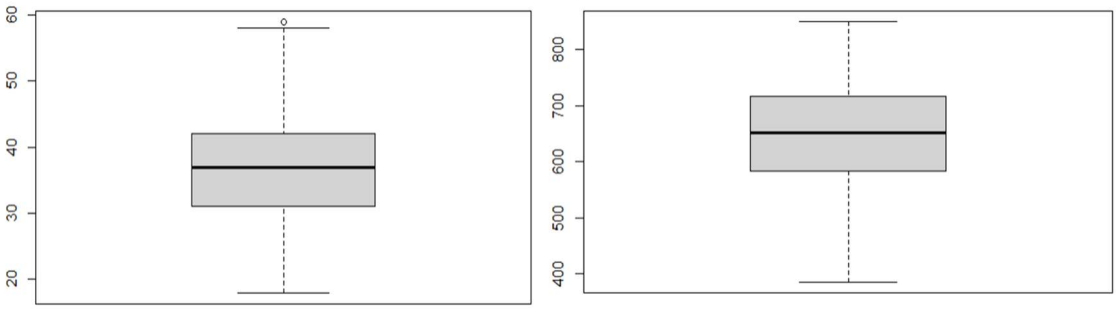

Figure 5 Age and Credit Score 's boxplots after remove the outliers

Converting String Values to Numerical Values

For logistic regression test we cannot do the test with string values, so we need to convert those string values to numerical values. In my dataset geography and gender variables in string value so I changed them to numerical value for test.

```
df['Geography'] = df['Geography'].astype('category')
df['Geography'] = df['Geography'].cat.codes
df
```

	CustomerId	Surname	CreditScore	Geography	Gender	A
0	15634602	Hargrave	619	0	Female	
1	15647311	Hill	608	2	Female	
2	15619304	Onio	502	0	Female	

```
df['Gender'] = df['Gender'].astype('category')
df['Gender'] = df['Gender'].cat.codes
df
```

	CustomerId	Surname	CreditScore	Geography	Gender
0	15634602	Hargrave	619	0	0
1	15647311	Hill	608	2	0
2	15619304	Onio	502	0	0

Figure 6 Converting string values to numerical values

4.3 Descriptive Statistics and Data Visualization

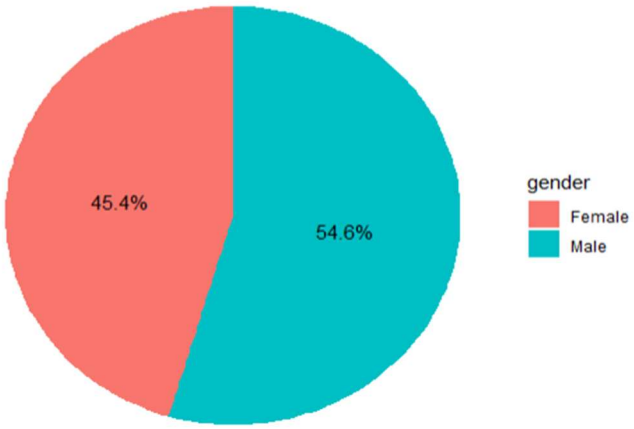

Figure 7 Pie chart for gender

In this dataset there are 9459 entries. So, here 5165 male and 4294 female.

4.3.1 Data Visualization for Categorical Variables

Churn vs Not Churn

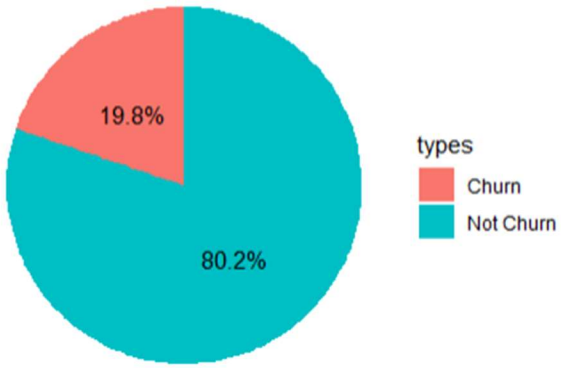

Figure 8 Pie chart for churn

Most of the customers did not churn. Here, 1876 customers churned, and 7583 did not churn.

Geography

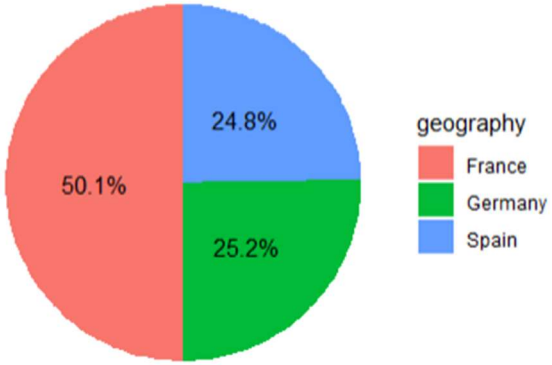

Figure 9 Pie chart for geography

According to my dataset,4735 from France,2382 from Germany and2342 from Spain.

Number of Products

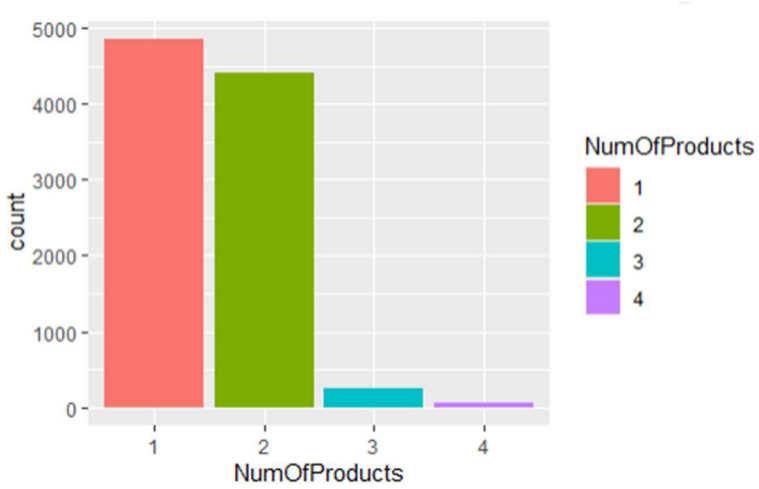

Figure 10 Bar chart to number of products

According to the uploader of dataset he mentioned the number of products so, there are four products in bank.4788 customers have purchased one product,4371 customers have purchased two products,245 customers have purchased three products and 55 customers have purchased four products.

NumOfProducts <fctr>	Frequency <int>
1	4788
2	4371
3	245
4	55

Table 7 Count of Products

Number of Credit Card Customers

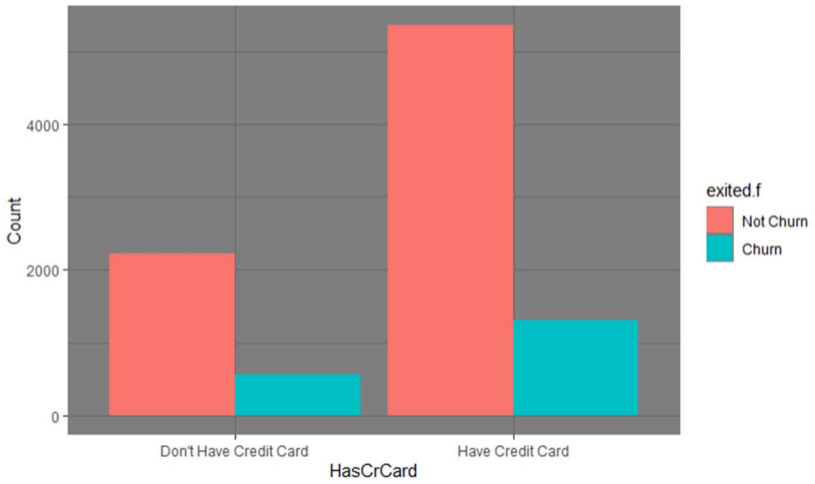

Figure 12 Bar graph and count of credit card customers

HasCrCard <fctr>	Churn <fctr>	Count <int>
0	0	2222
1	0	5361
0	1	568
1	1	1308

Table 8 Table of Credit Card Customers with Churn

According to this bar graph and data frame most of the customers have credit card and the customer who have churned are mostly credit card holders.

Active and Not-Active Customers

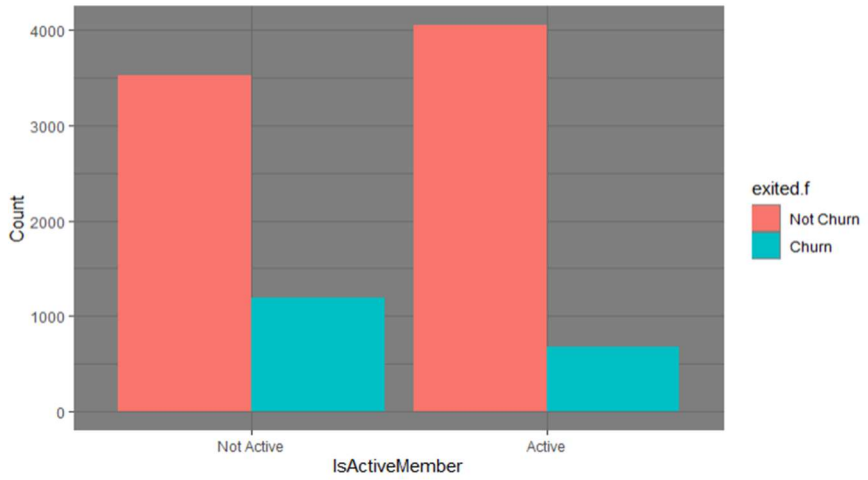

Figure 13 Bar graph and count of activeness of customers

IsActive <fctr>	Churn <fctr>	Count <int>
0	0	3530
1	0	4053
0	1	1197
1	1	679

Table 9 Table of Activeness of Customers with Churn

According to my dataset most of the customers are active customers. So according to above data frame we can recognize that customers who have churned are mostly in active members of bank.

Count of customers who Complained.

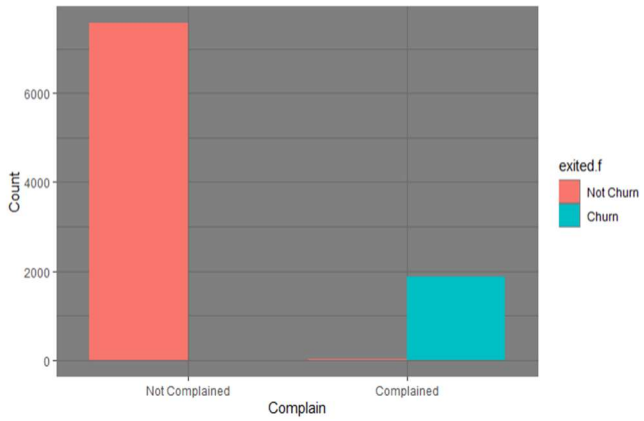

Figure 14 Bar graph and count of customer complain

Complain <fctr>	Churn <fctr>	Count <int>
0	0	7573
1	0	10
0	1	4
1	1	1872

Table 10 Table of Complain with Churn

According to my dataset most of the customers did not make complain to bank but when we analyze the churn customers most of the customers have make complains.

Count of Satisfaction Score

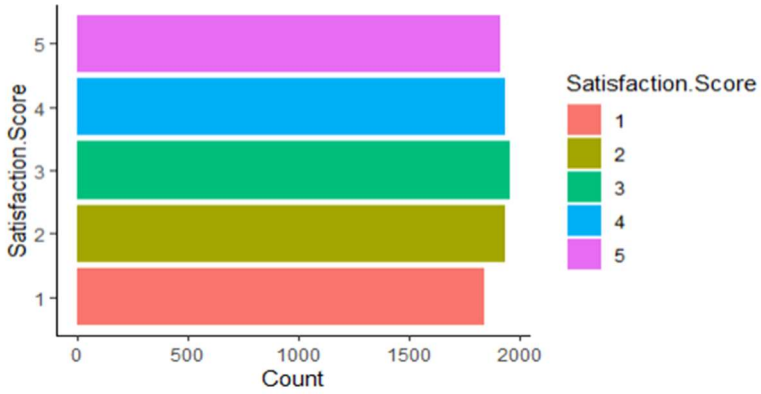

Figure 15 Bar graph of satisfaction score

According to my dataset there are five levels of satisfaction. So, below table will shows each satisfaction level with count of customers.

Satisfaction Level <fctr>	Frequency <int>
1	1808
2	1921
3	1933
4	1905
5	1892

Table 11 Count of Satisfaction Score

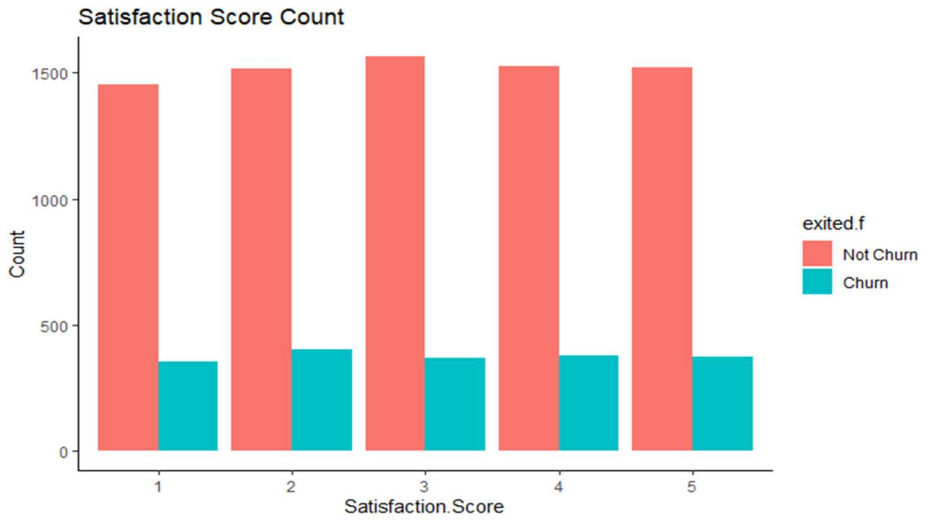

Figure 16 Bar graph for Satisfaction Score with Churn

According to this bar graph we can assume that there is no affect in churn by the customers satisfaction sore.

Summary of Categorical Variables

```
  Geography       Gender     NumOfProducts HasCrCard IsActiveMember Exited   Complain
France :4735   Female:4294   1:4788        0:2790    0:4727         0:7583   0:7577
Germany:2382   Male  :5165   2:4371        1:6669    1:4732         1:1876   1:1882
Spain  :2342                 3: 245
                             4:  55

Satisfaction.Score
1:1808
2:1921
3:1933
4:1905
5:1892
```

Figure 17 Summary of categorical variables

4.3.2 Data Visualization and Distribution of Numerical Variables

In my dataset there are four numerical variables. Such as credit score, age, tenure, balance, salary, and points.

```
  CreditScore          Age            Tenure          Balance
Min.   :386.0   Min.   :18.00   Min.   : 0.000   Min.   :     0
1st Qu.:584.0   1st Qu.:31.00   1st Qu.: 3.000   1st Qu.:     0
Median :652.0   Median :37.00   Median : 5.000   Median : 97325
Mean   :650.8   Mean   :37.38   Mean   : 5.017   Mean   : 76536
3rd Qu.:717.0   3rd Qu.:42.00   3rd Qu.: 7.500   3rd Qu.:127652
Max.   :850.0   Max.   :59.00   Max.   :10.000   Max.   :250898
             EstimatedSalary      Point.Earned
             Min.   :    11.58   Min.   : 119.0
             1st Qu.: 51227.75   1st Qu.: 411.0
             Median :100335.60   Median : 603.0
             Mean   :100168.07   Mean   : 606.1
             3rd Qu.:149401.95   3rd Qu.: 801.0
             Max.   :199992.50   Max.   :1000.0
```

Figure 18 Summary of numerical variables

Credit Card Score

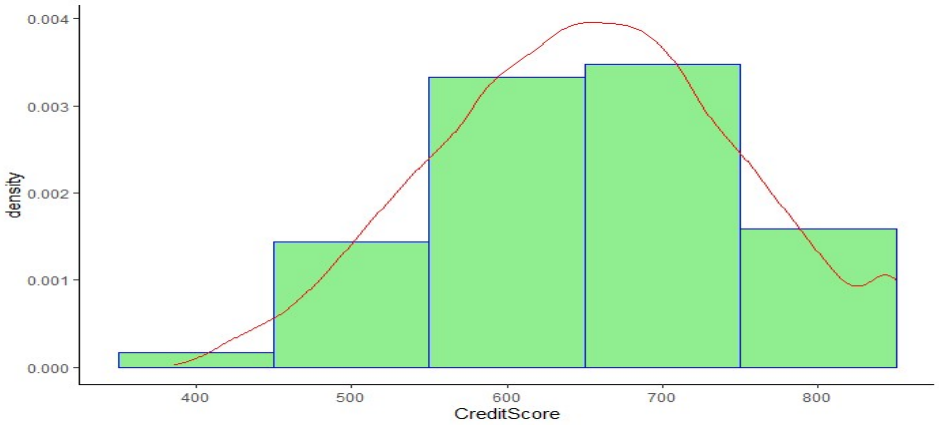

Figure 19 Histogram of Credit score

Score Range <fctr>	Frequency <int>
300-400	3
400-500	593
500-600	2300
600-700	3618
700-800	2337
800-900	608

Table 12 Table of Score Range

Credit card score is slightly negatively skewedl distributed. Most amount of score spread between the range 600 -700.

Point.Earned

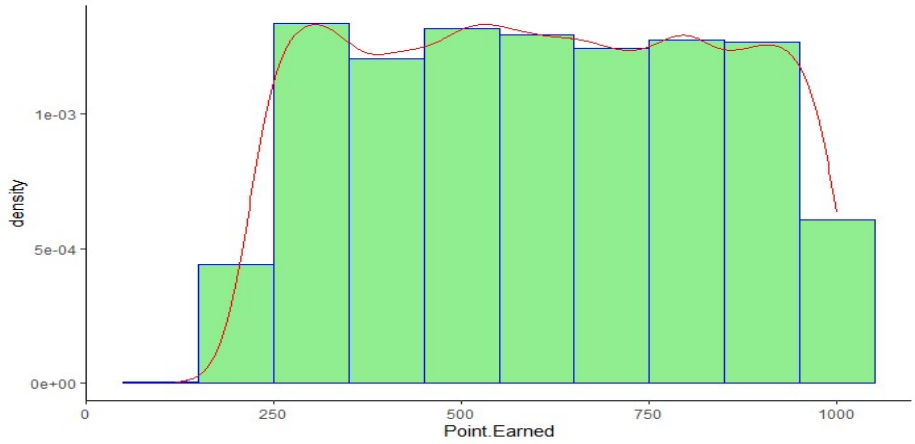

Figure 20 Histogram of Points that Customer Earned

Point Range <fctr>	Frequency <int>
100-200	2
200-300	1035
300-400	1190
400-500	1191
500-600	1273
600-700	1200
700-800	1197
800-900	1182
900-1000	1189

Table 13 Table of Point Range

Variable point earned is not a normal distributed. In range of 100-200 points there are only 2 customers and in other ranges approximately customers count is equal.

Age

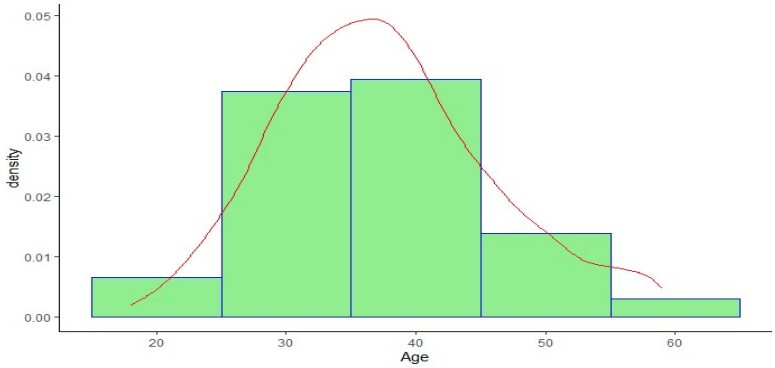

Figure 21 Histogram of Age

Age Range <fctr>	Frequency <int>
0-10	0
10-20	89
20-30	1876
30-40	4448
40-50	2315
50-60	731

Table 14 Table of Age Range

This variable is positive skewed unimodal distributed. I choose 10 as a bin width so, according to that 0 – 10 age there are no customers. Most of the customers from age range 30-40.

Tenure

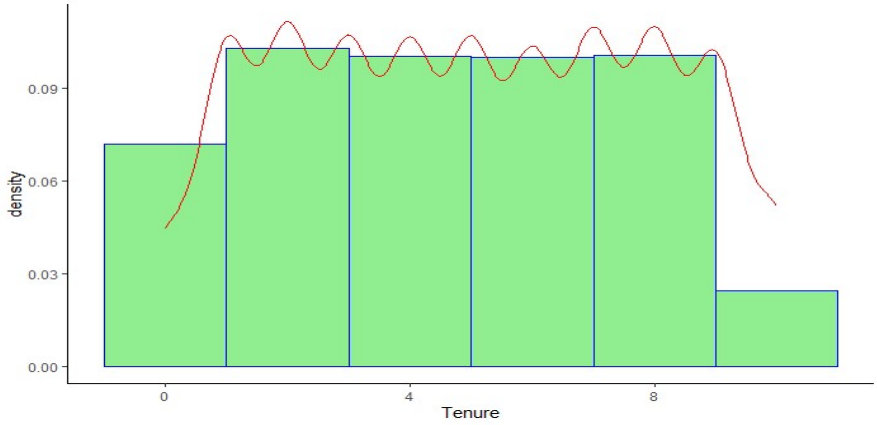

Figure 22 Histogram of Tenure

Tenure Range <fctr>	Frequency <int>
0-2	2353
2-4	1896
4-6	1867
6-8	1956
8-10	1387

Table 15 Table of Tenure Range

This variable is not normally distributed. So, according to this most of the customers are staying in bank for 0 – 2 years. This is very low.

Balance

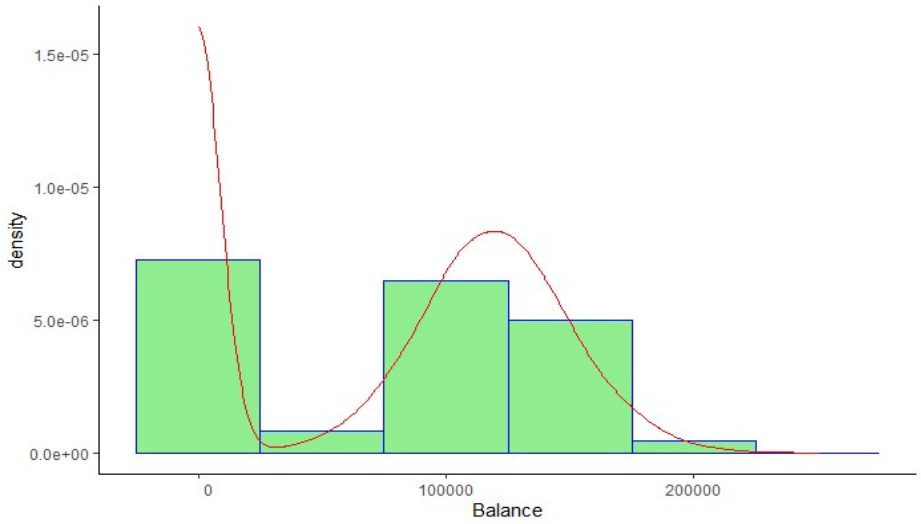

Figure 23 Histogram of Balance

Euro Range <fctr>	Frequency <int>
0-50K	3490
50K-100K	1425
100K-150K	3619
150K-200K	891
200K-250K	33
250K-300K	1

Table 16 Table of Balance Range

This is bimodal distribution. There is only one customer who has bank balance between 250K to 300K and most of the customers bank balance range is 100K-150K.

Estimated Salary

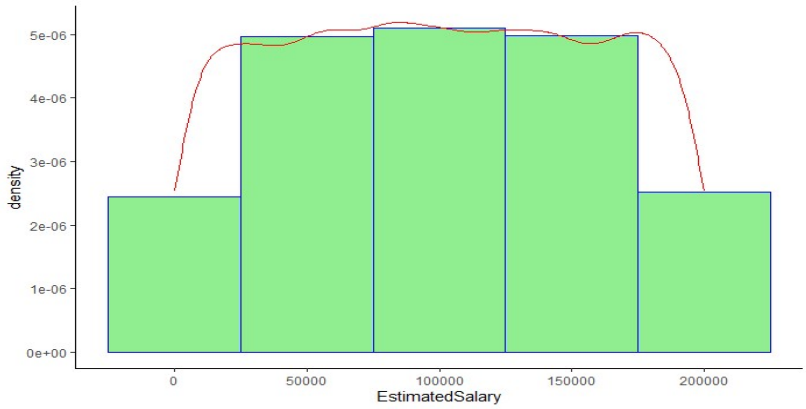

Figure 24 Histogram of Salary

Euro Range <fctr>	Frequency <int>
0-50K	2313
50K-100K	2401
100K-150K	2421
150K-200K	2324

Table 17 Customer Salary Range

This is not normally distributed. I choose 50,000 as a my bin width so according to this there are approximately equal number of customers in every range.

Correlation Plot

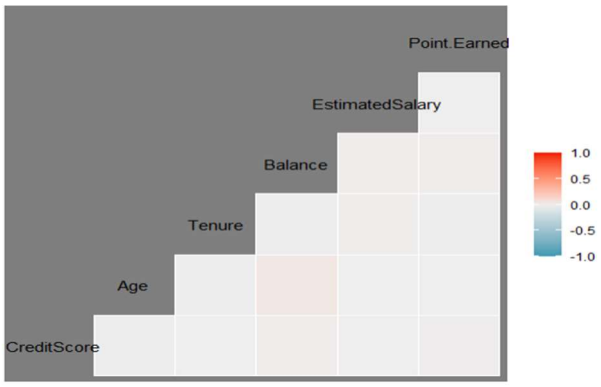

Figure 25 Correlation plot to numerical variables

According to this correlation heatmap we can assume that each and every numerical variable do not have correlation between them.

4.3.3 Comparison of Categorical and Numerical Variables

Balance vs Churn

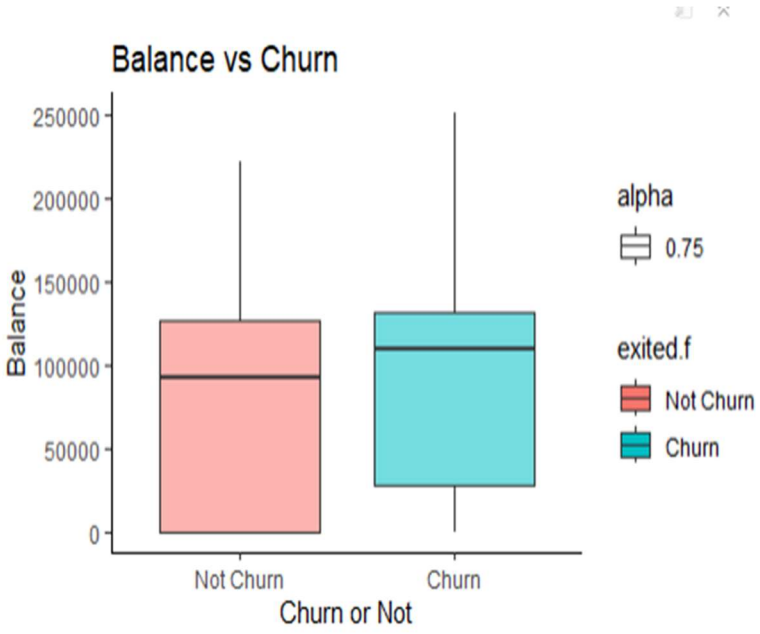

Figure 26 Boxplot Balance vs Churn

According to this boxplot we can assume that customers who have good bank balance have churned than not churned customers. So, balance have effect on customer churn.

Tenure / Salary / Points Earned / Credit Score vs Churn.

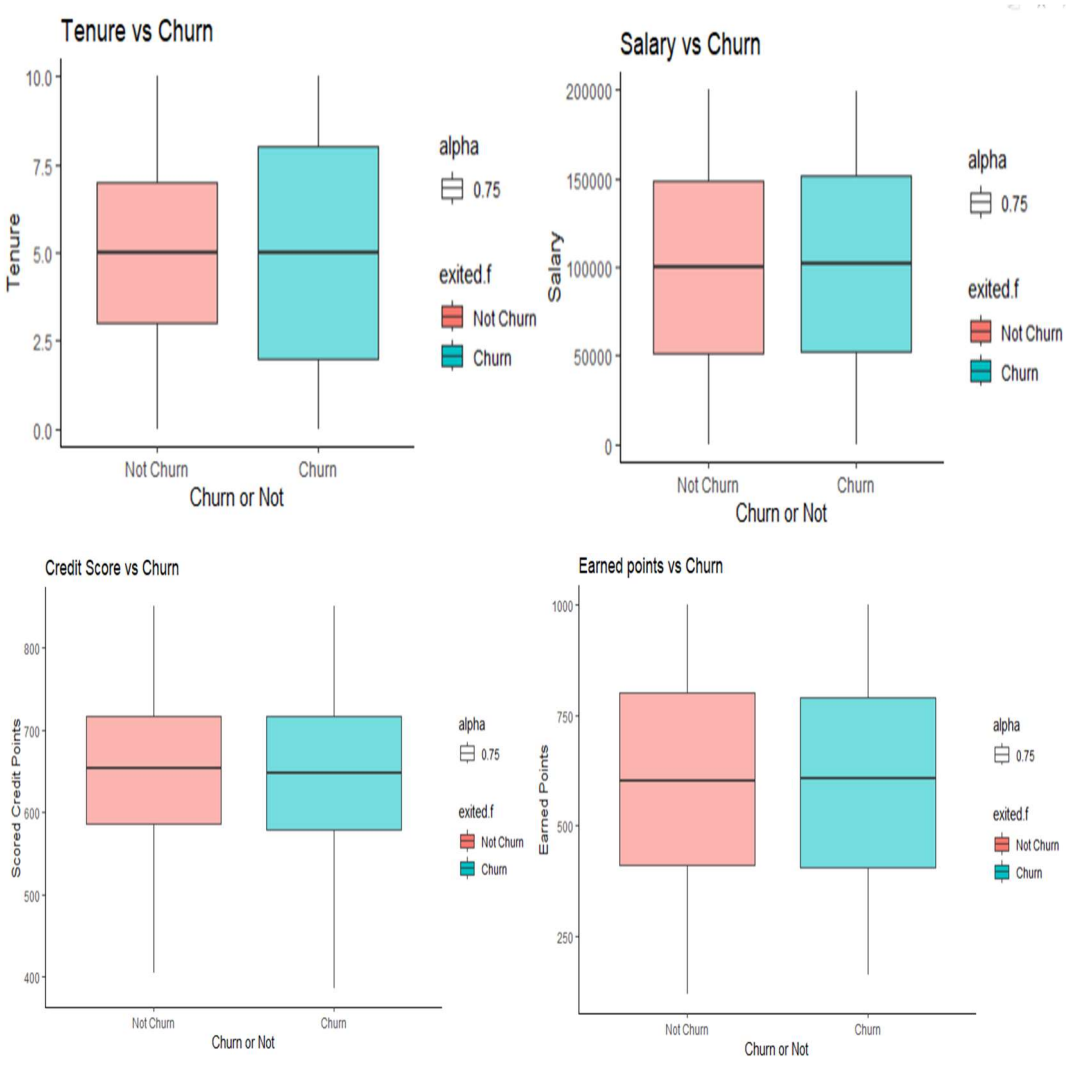

Figure 27 Boxplot Tenure/Salary/Credit Score/Points earned vs Churn

According to these boxplots we can assume salary, tenure, credit score, and points that customer earned wasn't affected by the churn but when we look into tenure vs churn boxplot churned customers' spread range is little wide than not churned customers.

Age vs Churn

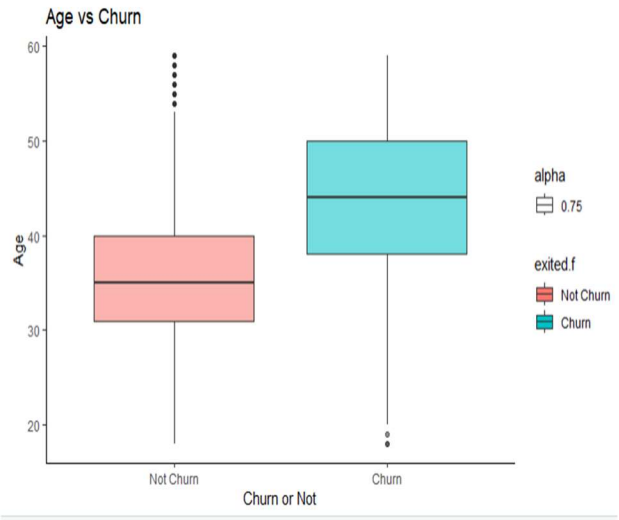

Figure 28 Boxplot Age vs Churn

According to this boxplot we can assume that customers who are old have churned. So, most of the churned customers who are in younger age not churned. So, age have effect on customer churn.

4.4 Statistical Test to Research

4.4.1 Logistic Regression

For this test I totally choose 14 variables and my dependent variable is churn (Exited) and balance 13 variables are independent.

Dependent Variable (Y) = Exited

Independent Variables (X) = CreditScore, Gender, Geography, Age, Tenure, Balance, NumOfProducts, HasCrCard, IsActiveMember, Complain, Satisfaction.Score, Point.Earned

After the preprocessing there are 9459 entries in my dataset. I split the data set into 30% and 70%. Because 30% for testing and 70% for training. I have used MinMaxScaling technique as my scaling technique.I trained and test my data set, I got the 0.99 accuracy

```
In [22]: log_reg.score(X_train_scaled,Y_train)
Out[22]: 0.9986406887177164

In [23]: log_reg.score(X_test_scaled,Y_test)
Out[23]: 0.9982381959126145
```

Figure 29 Accuracy of model

Confusion Matrix,

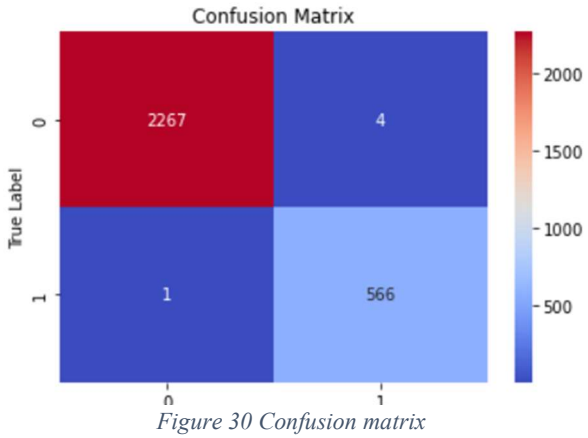

Figure 30 Confusion matrix

Label	No of Observations
True Positive	566
True Negative	2267
False Positive	4
False Negative	1

Table 18 Predicted Observations

I choose 0.05 as a my significant level there for only three variables have significant relationship between churn or only three variables' P value are less than 0.05. Those variables are Age, Complain, and IsActiveMember.

```
Optimization terminated successfully.
         Current function value: 0.007951
         Iterations 12
                           Logit Regression Results
==============================================================================
Dep. Variable:                 Exited   No. Observations:                 6621
Model:                          Logit   Df Residuals:                     6607
Method:                           MLE   Df Model:                           13
Date:                Thu, 07 Sep 2023   Pseudo R-squ.:                  0.9840
Time:                        11:14:39   Log-Likelihood:                -52.646
converged:                       True   LL-Null:                       -3292.0
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
                     coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const            -10.8384      3.571     -3.035      0.002     -17.837      -3.840
CreditScore        0.0006      0.004      0.169      0.866      -0.007       0.008
Geography         -0.0572      0.412     -0.139      0.889      -0.864       0.749
Gender            -0.1263      0.693     -0.182      0.855      -1.484       1.232
Age                0.1296      0.040      3.204      0.001       0.050       0.209
Tenure            -0.0797      0.120     -0.667      0.505      -0.314       0.155
Balance         7.192e-06   5.59e-06      1.287      0.198   -3.76e-06    1.81e-05
NumOfProducts     -0.2049      0.446     -0.460      0.646      -1.079       0.669
HasCrCard         -0.1201      0.737     -0.163      0.871      -1.565       1.324
IsActiveMember    -1.6744      0.815     -2.056      0.040      -3.271      -0.078
EstimatedSalary -1.556e-06   5.81e-06     -0.268      0.789   -1.29e-05    9.83e-06
Complain          13.7175      1.015     13.519      0.000      11.729      15.706
Satisfaction.Score -0.0141     0.241     -0.059      0.953      -0.486       0.458
Point.Earned      -0.0020      0.002     -1.234      0.217      -0.005       0.001
==============================================================================
```

Figure 31 Logistic regression result

4.4.2 Correlation Test between Point Earned and Bank Balance

```
        Pearson's product-moment correlation

data:  d1$Point.Earned and d1$Balance
t = 1.1809, df = 9457, p-value = 0.2377
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.008012044  0.032287631
sample estimates:
       cor
0.01214272
```

Figure 32 correlation result

Correlation between customer's bank balance and points that customer earned is 0.01. This prove us there is no association between two variables because 0.01 approximately equal to 0.

P value is greater than 0.05 so we need to accept null hypothesis test. So, there is no association between customer bank balance and points that customer earned.

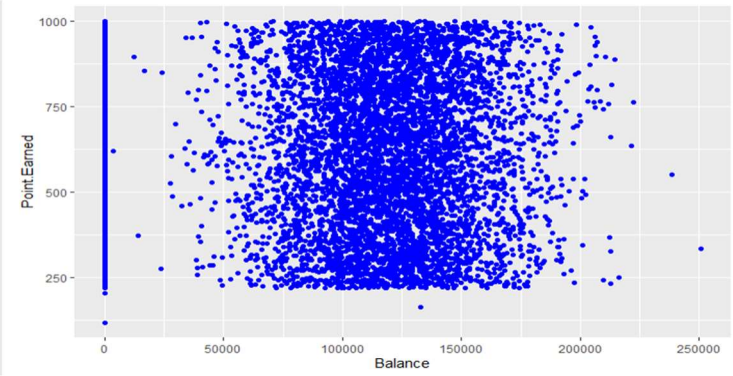

Figure 33 Scatter plot points earned vs balance

4.4.3 One-Way ANOVA Test between Activeness and Bank Balance

I have chosen 0.05 as my significant level for this test.

Dependent Variable (Y) = Balance

Independent Variable (X) = IsActiveMember

```
                 Df    Sum Sq   Mean Sq F value Pr(>F)
IsActiveMember    1 8.824e+08 8.824e+08   0.226  0.634
Residuals      9457 3.687e+13 3.899e+09
```

Figure 34 ANOVA result

So, P value is greater than 0.05 so we need to accept the null hypothesis so sample means are equal and do not have significant different between variables.

Chapter 5: Discussion and Recommendation

5.1 Discussion

Knowledge of statistics and machine learning algorithms are very helpful to analyze and predict the customer churn and factors that contribute to churn. From my dataset I test the logistic regression to every variables except row number, customer ID and card type. My logistic regression model has best accuracy to predict the churn with other characteristics that I had in dataset. According to my correlation test I got to know that there is no association between the bank balance and the points that customer earned. So, we can decide that points score by analyzing other factors except bank balance of the customer. And we can decide that the sample means of activeness of customer and bank balance of the customer is equal and there is no significant different between these two factors. Finally this study is essential to bank organizations to retent their customers.

5.2 Recommendation

According to my research I suggest some points,

Banks need to check the credit cards' rules and regulation and need to adjust it because, when we compare to churn customers that who have and haven't have credit card the customers who have credit cards are mostly churned.

When we look into tenure that we can assume that most of the customers have lowest tenure so we need to focus that.

And banks need to analyze the inactive customers and need to know what are the reasons for their inactiveness.

When we focus on complained customers in this data set most of the customers have churned that who are complained. So, we must look in to their complains and need to get solutions.

When we look into customers bank balance, the customers who have left the bank have good bank balance, so we need to check and analyze our balance points / interest rate that we offer to customer.

According to my visualization and my logistic model age have effect on churn. So, customers who are old are churning so we need to analyze the reason for that.

5.3 Conclusion

This research project's final aim is to explore and analyze the factors which contribute to the churn. So, according to my data visualization and analysis and logistic test proved that age, bank balance, complain of customer towards bank, activeness of the customer in bank, and credit card rules and regulation are contributing to the customer churn. Additionally, research proved that there is no association between the bank balance and the points that customer earned, and the sample means of bank balance and the activeness of the customers are equal.

This dataset is secondary data set so, according to that there are some areas should to investigate more,

- There is a variable call number of products but this dataset did not define what are the product or service is it.
- Dataset did not define the currency I assumed as a euro for this research.
- This dataset did not mention the period of churn and there is no clear information about data and tenure.
- The dataset did not define which banks detail is this.
- According to this dataset there is no positive relationship between numerical variables. Such as bank balance, customer salary, points that customer earned, and credit card score of the customer. So future researches should focus on what are the factors that effect these factors positively.

References

- Vazifedoost,H., Ansar,M., Yekezare,A.(2013)'How customer satisfaction changes behavior: A case study of banking industry'. *Management Science Letters* [online] 3 (6), 1587-1892. available from < http://growingscience.com/beta/msl/917-how-customer-satisfaction-changes-behavior-a-case-study-of-banking-industry.html > [21 July 2023]
- Guangli,N., Wei,R., Lingling,Z., Ying,T.,Yong,S. (2011)'Credit card churn forecasting by logistic regression and decision tree'. *Fictitious Economy and Data Science* [online] 38 (12), 15273 – 15285. available from <https://www.sciencedirect.com/science/article/abs/pii/S0957417411009237> [20 July 2023]
- Abbas,K., Hajar,G., Seyed,M. (2016)'Developing a prediction model for customer churn from electronic banking services using data mining'. [online] available from <https://www.academia.edu/49571730/Developing_a_prediction_model_for_customer_churn_from_electronic_banking_services_using_data_mining> [20 July 2023]
- Marcin,S.(2018) 'Predicting Customer Churn in Electronic Bank'. *Business Information Processing* [online] 339, 18-20 available from <https://annas-archive.org/md5/3eb1726d90687d373c9c7839848b1afe> [19 July 2023]
- Dana,A.,Al-Rousan,N.,Hazem,A. (2022) 'Machine Learning to Develop Credit Card Customer Churn Prediction' *Journal of Theoretical and Applied Electronic Commerce Research* [online] 17 (4), 1529-1549 available from <https://www.researchgate.net/publication/365485893_Machine_Learning_to_Develop_Credit_Card_Customer_Churn_Prediction> [19 July 2023]
- Deni Kamaludin,Y. (2019) 'Analyzing the effect of service quality on customer satisfaction of Islamic banking in west Java province, Indonesia' *Management Science Letters* [online] , 2347-2356 available form <https://annas-archive.org/md5/3b36457dfd7807eba27b5c17ebcb0f52> [20 July 2023]
- Millon,A. (2013) 'The Impacts of Electronic Banking on Customer Satisfaction in Ethiopian Banking Industry'. SSRN Electronic Journal [online]. available from <https://annas-archive.org/md5/34fbdf9da05059093a04da18abbef2cc>[21 July 2023]

- DataScientist(2023) *Kaggle : All you need to know about this platform* [online] available from <https://datascientest.com/en/kaggle-all-about-this-platform#:~:text=Kaggle%20is%20a%20web%20platform,achieve%20all%20data%20science%20advancements.> [20 July 2023]
- The Federal Deposit Insurance Corporation (2022*) 2021 FDIC National Survey of Unbanked and Underbanked Households* [online] available from <https://www.fdic.gov/analysis/household-survey/index.html#:~:text=An%20estimated%2081.5%20percent%20of,above%20nonbank%20transactions%20and%20credit.> [19 July 2023]
- The World Bank (2022) The Global Findex database 2021: Financial Inclusion, Digital Payments, and Resilience in the Age of COVID – 19 [online] available from < https://www.worldbank.org/en/publication/globalfindex#:~:text=Adults%20with%20an%20account%20(%25)%2C%202011%E2%80%932021&text=Worldwide%20account%20ownership%20has%20reached,it%20hovered%20for%20many%20years.> [20 July 2023]
- Stewart,K.,McCabe.(2006)'The 'Net Effect' on Bank Customer Complaining Behaviour'. *The Marketing Review* [online] 1 (6), 41-53. available from < https://www.ingentaconnect.com/content/westburn/tmr/2006/00000006/00000001/art00004> [12 September 2023]
- Zhiya Zhuo.(2019) 'Research on using Six Sigma management to improve bank customer satisfaction'. *International Journal of Quality Innovation* [online] 1 (3), .available form < https://jqualityinnovation.springeropen.com/articles/10.1186/s40887-019-0028-6 > [12 September 2023]
- Joseph,M.(2005)' Importance-performance analysis of UK and US bank customer perceptions of service delivery technologies'. *International Journal of Financial Services Management*[online], 66-88. available from < https://www.researchgate.net/publication/5171757_Importance-performance_analysis_of_UK_and_US_bank_customer_perceptions_of_service_delivery_technologies> [12 September 2023]
- Statistics (n.d) *Shapes of Distribution* [online] available from <https://www.studypug.com/statistics-help/shapes-of-distributions> [02 August 2023]