FORECASTING AND PREDICTIVE ANALYTICS OF CUSTOMER CHURN IN BANKING SECTOR

Contents

1. Abstract: 3
2. Introduction: 3
   1.1 Objectives of the Study: 4
   1.2 About the data Set: 4
   1.3 Case Profile: 5
3. Literature Review: 5
   3.1 Customer churn and retention: 5
   3.2 Predictive Analytics: 6
   3.3 Decision Tree: 6
   3.4 Logistic Regression: 7
   3.5 R Studio: 8
   3.6 SPSS: 8
4. Data Exploration and Preparation: 8
   4.1 Understanding the data set: 8
   4.2 Data cleaning and Transformation: 9
   4.3 Visualizing customer demographics: 12
   4.4 Visualizing account information: 14
5. Forecasting and Predictive modelling: 18
   5.2 Logistic Regression: 20
6. Results: 22
7. Conclusion: 22
8. Future Work: 22

References: 24

1. Abstract:

This report aims to illustrate an in-depth understanding of forecasting and predictive analytics using the case of customer churn in the banking sector. The report uses statistical and visualization tools such as R and SPSS for data cleaning, exploration, analysis, visualization, and prediction models of customer churn. Data visualization is used to show the trends and relationships among various attributes present in the data. It is a graphical way of presenting data and is widely used in analyzing complex data. In data visualization, data is visually summarized and the spatial variables like size, position, and shape form the key factors of the data (Sadiku, Shadare, Musa, Akujuobi & Perry, 2016). In this study data visualization will be used to understand the various attributes of the case study and their effect on customer churn in the banking sector. Predictive analytics, on the other hand, is used to predict future events based on existing and historical data (Kumar & Garg, 2018). It is used in machine learning, statistics, and optimization techniques. These techniques help in understanding the risk of customers individually (Kumar & Garg, 2018). Predictive analytics is not only limited to the banking sector but is widely used in retail, finance, insurance, and sales industries (Kumar & Garg, 2018). In this study predictive model is used to predict the customer churn of the bank and help the bank take the required measures to reduce customer churn. In this study, a decision tree is used as a predictive model. Decision trees can be easily interpreted and resilient in the existence of noise. The decision tree has advantages over other classification models and is a dominantly used model to predict customer churn (Bin, Peiji & Juan, 2007). On the other hand, logistic regression is used the understand the dependency of a variable on other attributes of the data. Logistic regression helps in understanding the attributes that affect a customer's decision to churn (Dalvi, Khandge, Deomore, Bankar & Kanade, 2016).

2. Introduction:

In a fast-changing world, the number of providers is increasing very swiftly in every industry. In the banking sector, there is no scarcity of options for a customer to choose where to place their money (Rahman & Kumar, 2020). In banking, customer

churn can be explained as a customer closing his accounts and withdrawing himself from all the products offered by the bank (Karvana, Yazid, Syalim & Mursanto, 2019). It not only reduces the profits made by the bank but also harms its operations (Karvana et al., 2019). With increasing customer churn, the pressure of customer retention is also increasing in the competitive setup. Customer retention is a very effective approach for banks to grow. Various relationship management tools for customer retention are formulated to reduce customer churn and boost customer retention (Chitra & Subashini, 2011). To solve the problem of customer churn and understand which group of customers are likely to churn, classifying customers is crucial. Churn prediction, therefore, is important to identify customers who are likely to churn in the future. This prediction is based on historical data which contains information on customers who have previously churned (Chitra & Subashini, 2011). In this research, a real-world data set is used to explore and visualize data to derive meaningful insights and predict customer churn (Chitra & Subashini, 2011). This report examines the customer churn of the bank SVND Pvt Ltd. and investigates the measure that needs to be taken to decrease customer churn.

1.1 Objectives of the Study:

The primary objective of this study is to estimate whether the customer of the bank, SVND Pvt Ltd. will churn or not. Using the data provided by the bank, perform exploratory data analysis to understand and visualize the variables affecting customer churn. The analysis will help in building a prediction model using Machine Learning to predict customer churn. Examine if customer demographics and attributes related to customer account information influences customer churn.

1.2 About the Data Set:

This report utilizes the data set titled “Predicting Churn for Bank Customers” and is sourced from Kaggle which is a prominent open-source platform and has numerous data sets. This data set has 6,383 rows and 14 columns. Each column represents an attribute of the data which are related to customer demographics, account information such as tenure, balance, number of products, and the churn variable.

1.3 Case Profile:

SVND Pvt Ltd is a prominent European bank in the banking sector based in the UK known for its authenticity and reliability. SVND Pvt Ltd has customers from different countries like France, Germany, and Spain. The customer-centric approach and quality of the services offered by the bank have contributed to its reputation and market position. The bank has reached out to help in understanding the attributes influencing its customers to switch to a different bank so that effective measures can be taken for customer retention and also predict customer churn.

3. Literature Review:

3.1 Customer churn and retention:

The most valued assets in any industry are customers and are considered the main source of profit for any organization (Saran & Chandrakala, 2016). Organizations have noticed that conscious efforts should be made not only to persuade customers to purchase their products and services but also to retain customers (Saran & Chandrakala, 2016). The concept of customer retention has been scrutinized in many sectors. A bank is represented in its industry by the customers who stay with the bank for a very long tenure (Mutanen, 2006). Customer churn happens when a customer decides to leave a bank and switch to another bank for their financial transactions and services. Customer churn not only harms the profits but also negatively affects the other operations of the bank (Keramati, Ghaneei & Mirmohammadi, 2016). The two main approaches to handling customer churn are proactive and reactive. In the reactive framework, the bank waits for the customer to request the bank to cancel their products and services relationship. In this case, the customer is offered attractive offers and benefits to retain their relationship with the bank (Saran & Chandrakala, 2016). On the other hand, in the proactive framework, the bank tries to identify the customers who are likely to leave and then provides attractive benefits to retain them (Saran & Chandrakala, 2016). The focus of customer churn analysis is to predict the customers who are most likely to churn and find measures to retain them (Mutanen, 2006). The discussion about customer retention in the banking sector is very straightforward. Retaining existing customers is more economical than acquiring new customers. The cost of gaining new

customers to replace the ones who have churned is very high (Cohen, Gan, Au Yong & Chong, 2007). Retaining customers is one of the most important management functions. Compared to new customers, existing customers will be more profitable if they continue their business relationship with the bank. Customer retention is not only a strategy to reduce customer churn but is a crucial approach to building customer loyalty and promoting sales of additional products and services offered by the bank (Darzi & Bhat, 2018).

3.2 Predictive Analytics:

Predictive analytics is a branch of advanced analytics used in statistics and analytical techniques (Kumar & Garg, 2018). Predictive analytics helps in predicting future events based on previous historical data with the help of machine learning techniques. The historical data which is collected is transformed and explored by techniques like filtering and correlation of the data (Mishra & Silakari, 2012). Predictive analytics is a four-step process that includes collecting data, processing, transforming, employing a machine learning model, and finally predicting the outcome (Mishra & Silakari, 2012). Various predictive analytics models provide a score. A higher score indicates the occurrence of the event. These predictive models help in recognizing the risk and opportunities of every customer (Kumar & Garg, 2018). It can be said that predictive analytics is the future of data mining. With the help of business knowledge, and statistical and analytical tools, predictive analytics can be used to produce business insights (Mishra & Silakari, 2012). The key aspect of predictive analytics is the predictor attribute. The predictor attribute is evaluated for an individual entity (Mishra & Silakari, 2012).

3.3 Decision Tree:

The decision tree is a classification algorithm that is used to separate data into categories with the help of certain conditions used in the decision-making process (Lee, Cheang & Moslehpour, 2022). The decision tree is an easy algorithm when compared to other classification algorithms and is popular due to the ease of use and reading of the outcome (Rahman & Kumar, 2020). The decision tree is used in multiple modeling techniques such as image processing, data mining, and machine

learning (Lee, Cheang & Moslephour, 2022). The decision tree is a tree model which relates the decision to its potential outcomes (Kumar & Garg, 2018). The decision tree structure comprises branches and nodes (Lee, Cheang & Moslephour, 2022). Each branch of the tree symbolizes the number of choices and each leaf symbolizes a decision (Rahman & Kumar, 2020). Each node of the decision tree consists of a specific number of values. The internal node extends from a parent node and their internal nodes are connected with the help of branches. Internal nodes will further have branches that are joined to other leaf nodes. The leaf nodes will have branches connected to them but branches do not extend out of leaf nodes. (Lee, Cheang & Moslephour, 2022). A decision tree has a lot of potential to be the most prominent predictive analytics tool as it is a very user-friendly method to be used without an in-depth understanding of statistics (Lee, Cheang & Moslephour 2022).

3.4 Logistic Regression:

Regression is a highly preferred statistical technique that is used to understand the relationship between various attributes. It is used to structure the relationship between a dependent variable and various independent attributes (Kumar & Garh, 2018). Regression helps in analyzing how the values of the dependent variable change when the values of the independent attributes are changed in the model. The regression function is crucial as it helps in mapping the dependent variable with the independent attributes (Kumar & Garh, 2018). The two types of regression models used in predictive analytics are linear regression and logistic regression. Binomial logistic regression is a regression in which the dependent variable is not continuous. Logistic regression is used to predict variables with discrete responses based on continuous attributes (Mutanen, 2006). Logistic regression is used to predict a binary variable based on categorical variables which are called predictors (Larose, 2015). The logistic regression the odds of a response is the fraction of the likelihood of a response to be the outcome divided by the probability of the response not being the outcome. Contrary to linear regression, logistic regression does not have a formula to estimate the value of β (LaValley, 2008).

3.5 R Studio:

R Studio is an open-source programming software that is used for statistical analytics and data visualizations. R can be run on various operating systems such as Windows, Mac OS, Unix, and Linux (Verzani, 2011). R Studio not only facilitates statistical analytics but also data analytics and predictive analytics which are used in problem-solving and decision-making (Shinde, Oza & Kamat, 2017). R is a binary format software that makes it easy to install. This language is influenced by the S language. R has a core scripting language but can be integrated with codes written in other languages like C, C++, and Java (Verzani, 2011). This programming language has CRAN which provides access to various functions and packages and the users can also develop their functions. R is used in various industries for business problem-solving and decision-making (Shinde et al., 2017).

3.6 SPSS:

IBM SPSS is a statistical software that performs operations related to data. SPSS performs a wide range of analytics from simple computations such as mean, median, and standard deviation to fundamental analysis such as logistic regression, correlation, and multivariate techniques such as scaling and factor analysis (Meyers, Gamst & Guarino, 2013). The acronym for SPSS is Statistical Package for Social Sciences. This software has a very user-friendly user interface. SPSS works with various types of files which include data files, syntax files, and output files. The output file displays the outcome of the analysis using tables and visualization plots (Wagner III, 2019). In this study, SPSS software is used to perform logistic regression.

4. Data Exploration and Preparation:

4.1 Understanding the data set:

The data set contains 6383 entries and 14 columns. Each row is the information of a customer and each column is related to customer demographics, account information, and churn information. Fig 1 shows the key variables of the data set and Fig 2 is a sample of the data set.

Key Variables	Description
Geography	The geographical location of the customer. Spain, France, and Germany.
Gender	The Gender of the Customer.
Age	Age of the customer.
Tenure	The tenure of the customer with the bank.
Balance	The bank balance of the customer.
NumOfProducts	The number of products and services the customer has subscribed to.
HasCrCard	The customer has a credit card with the bank or not.
IsActiveMember	Is the customer active with the bank or not?
EstimatedSalary	The estimated salary of the customer.
Exited	The customer has churned or not.
CreditScore	The credit score of the customer.

Fig 1: Key Variables of the Data Set

RowNumb	CustomerId	Surname	CreditScore	Geograp	Gender	Age	Tenure	Balance	NumOfPrc	HasCrCarc	IsActiveM	EstimatedSalary	Exited
2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	No	Yes	112542.58	No
3	15619304	Onio	502	France	Female	42	8	159660.8	3	Yes	No	113931.57	Yes
5	15737888	Mitchell	850	Spain	Female	43	2	125510.8	1	Yes	Yes	79084.1	No
6	15574012	Chu	645	Spain	Male	44	8	113755.8	2	Yes	No	149756.71	Yes
8	15656148	Obinna	376	Germany	Female	29	4	115046.7	4	Yes	No	119346.88	Yes
9	15792365	He	501	France	Male	44	4	142051.1	2	No	Yes	74940.5	No
10	15592389	H?	684	France	Male	27	2	134603.9	1	Yes	Yes	71725.73	No
11	15767821	Bearce	528	France	Male	31	6	102016.7	2	No	No	80181.12	No
16	15643966	Goforth	616	Germany	Male	45	3	143129.4	2	No	Yes	64327.26	No
17	15737452	Romeo	653	Germany	Male	58	1	132602.9	1	Yes	No	5097.67	Yes
27	15736816	Young	756	Germany	Male	36	2	136815.6	1	Yes	Yes	170041.95	No
29	15728693	McWillian	574	Germany	Female	43	3	141349.4	1	Yes	Yes	100187.43	No
30	15656300	Lucciano	411	France	Male	29	0	59697.17	2	Yes	Yes	53483.21	No
32	15706552	Odinakach	533	France	Male	36	7	85311.7	1	No	Yes	156731.91	No
33	15750181	Sandersor	553	Germany	Male	41	9	110112.5	2	No	No	81898.81	No
36	15794171	Lombardo	475	France	Female	45	0	134264	1	Yes	No	27822.99	Yes
37	15788448	Watson	490	Spain	Male	31	3	145260.2	1	No	Yes	114066.77	No
38	15729599	Lorenzo	804	Spain	Male	33	7	76548.6	1	No	Yes	98453.45	No
40	15585768	Cameron	582	Germany	Male	41	6	70349.48	2	No	Yes	178074.04	No
42	15738148	Clarke	465	France	Female	51	8	122522.3	1	No	No	181297.65	Yes

Fig 2: Sample of the data set

4.2 Data cleaning and Transformation:

The variable “Exited” is the target variable of the data set and shows if a customer has churned or not. The variables “CustomerId”, “Surname” and “RowNumber” are

customer specific and do not play an important role in the analysis and hence can be dropped. There are no missing values in this data set.

```
bank_churn$Surname <-  NULL
bank_churn$RowNumber <- NULL
bank_churn$CustomerId <- NULL
```

```
> any(is.na(bank_churn))
[1] FALSE
```

The age of the customers is from 18 to 92 and the mean value is approximately 40. The mean of the variable “tenure” is 5 years which means that the majority of the customers are tenured for more than 3 years and approximately 50% of the customers are active members. Further exploratory data analysis will be performed to understand the variables affecting customer churn.

```
age_max <- max(bank_churn$Age)
age_min <- min(bank_churn$Age)

mean_age <- mean(bank_churn$Age)

cat("Upper Limit:", age_max, "\n")
cat("Lower Limit:", age_min, "\n")
cat("Mean:", mean_age)
```

```
> cat("Upper Limit:", age_max, "\n")
Upper Limit: 92
> cat("Lower Limit:", age_min, "\n")
Lower Limit: 18
> cat("Mean:", mean_age)
Mean: 39.19771
```

Fig 3: R code with results showing the minimum, maximum, and mean values of Age

```
tenure_max <- max(bank_churn$Tenure, na.rm = TRUE)
tenure_min <- min(bank_churn$Tenure, na.rm = TRUE)

mean_tenure <- mean(bank_churn$Tenure, na.rm = TRUE)

print(paste("Upper Limit:", tenure_max))
print(paste("Lower Limit:", tenure_min))
print(paste("Mean:", mean_tenure))
```

```
> print(paste("Upper Limit:", tenure_max))
[1] "Upper Limit: 10"
> print(paste("Lower Limit:", tenure_min))
[1] "Lower Limit: 0"
> print(paste("Mean:", mean_tenure))
[1] "Mean: 4.979633401222"
```

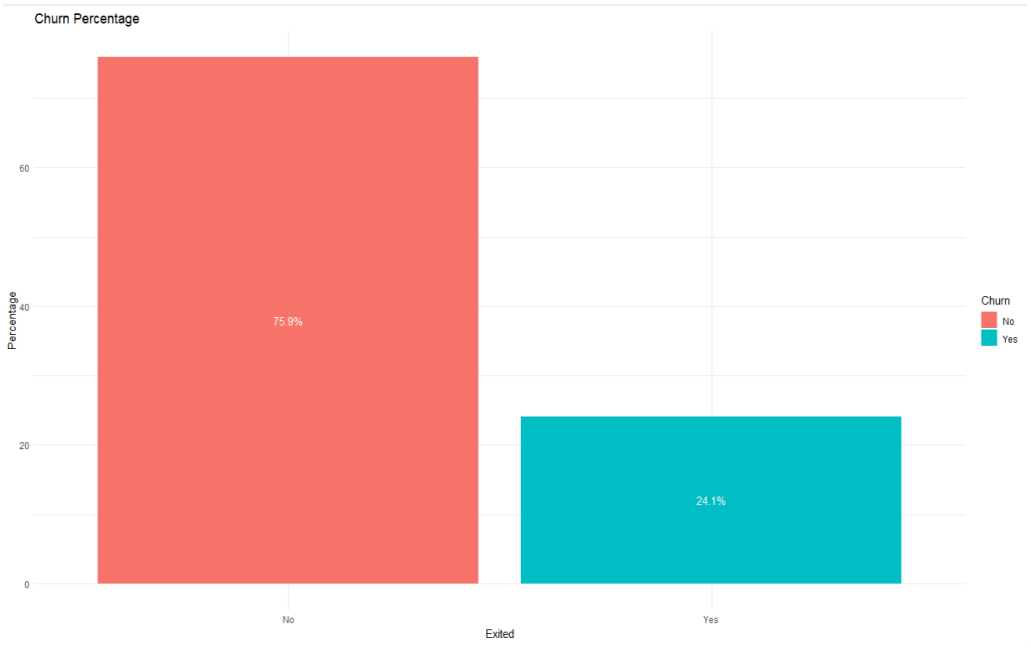
Fig 4: R code with results showing minimum, maximum, and mean values of Tenure

```
> count <- table(bank_churn$IsActiveMember)
> print(count)

  No  Yes
3105 3278
```

Fig 5: R code with results showing the count of active customers

1. Churn distribution:

Fig 6 shows the exit distribution of the customers. The bank is successful in retaining 75.9% of customers.



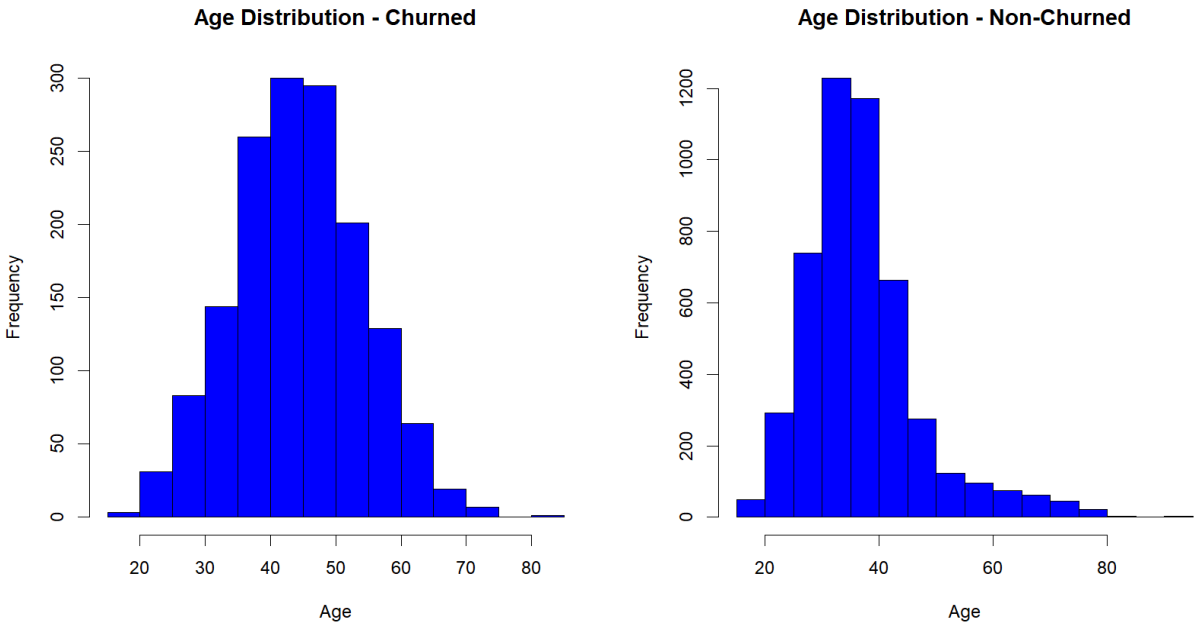
```r
exit_percent <- bank_churn %>%
  group_by(Exited) %>%
  summarise(count = n()) %>%
  mutate(percentage = count / sum(count) * 100)

ggplot(exit_percent, aes(x = Exited, y = percentage, fill = Exited)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = paste0(round(percentage, 1), "%")),
            position = position_stack(vjust = 0.5),
            color = "white",
            size = 4) +
  labs(x = "Exited", y = "Percentage", fill = "Churn") +
  ggtitle("Churn Percentage") +
  theme_minimal()
```

Fig 6: Churn distribution

4.3 Visualizing customer demographics:

1. Age distribution:

From Fig 7, it is visible that customers who churn are mostly between the age group of 40 to 50.



```r
churned <- bank_churn[bank_churn$Exited == "Yes", ]
non_churned <- bank_churn[bank_churn$Exited == "No", ]

par(mfrow = c(1, 2))  # Set the plotting layout to 1 row and 2 columns

hist(churned$Age, col  = "blue" , main = "Age Distribution - Churned", xlab = "Age", ylab = "Frequency")
hist(non_churned$Age,col  = "blue" , main = "Age Distribution - Non-Churned", xlab = "Age", ylab = "Frequency")
```

Fig 7: Age distribution vs churn

2. Geography distribution:

Customers who are from Germany are most likely churn among the three countries. The exit percentage of Germany is 32.4% followed by Spain at 19.6% and France is the least at 18.2%.

	Geography	Exited_No	Exited_Yes
	<chr>	<int>	<int>
1	France	2123	473
2	Germany	1695	814
3	Spain	1028	250

```r
bank_churn %>%
  group_by(Geography, Exited) %>%
  summarize(count = n()) %>%
  mutate(percentage = count / sum(count) * 100) %>%
  ggplot(aes(x = Geography, y = percentage, fill = factor(Exited))) +
  geom_bar(stat = "identity", position = "stack") +
  geom_text(aes(label = paste0(round(percentage, 1), "%")),
            position = position_stack(vjust = 0.5), color = "white") +
  labs(x = "Geography", y = "Percentage", fill = "Exited") +
  ggtitle("Exited vs. Geography with Percentages") +
  theme_minimal()
```

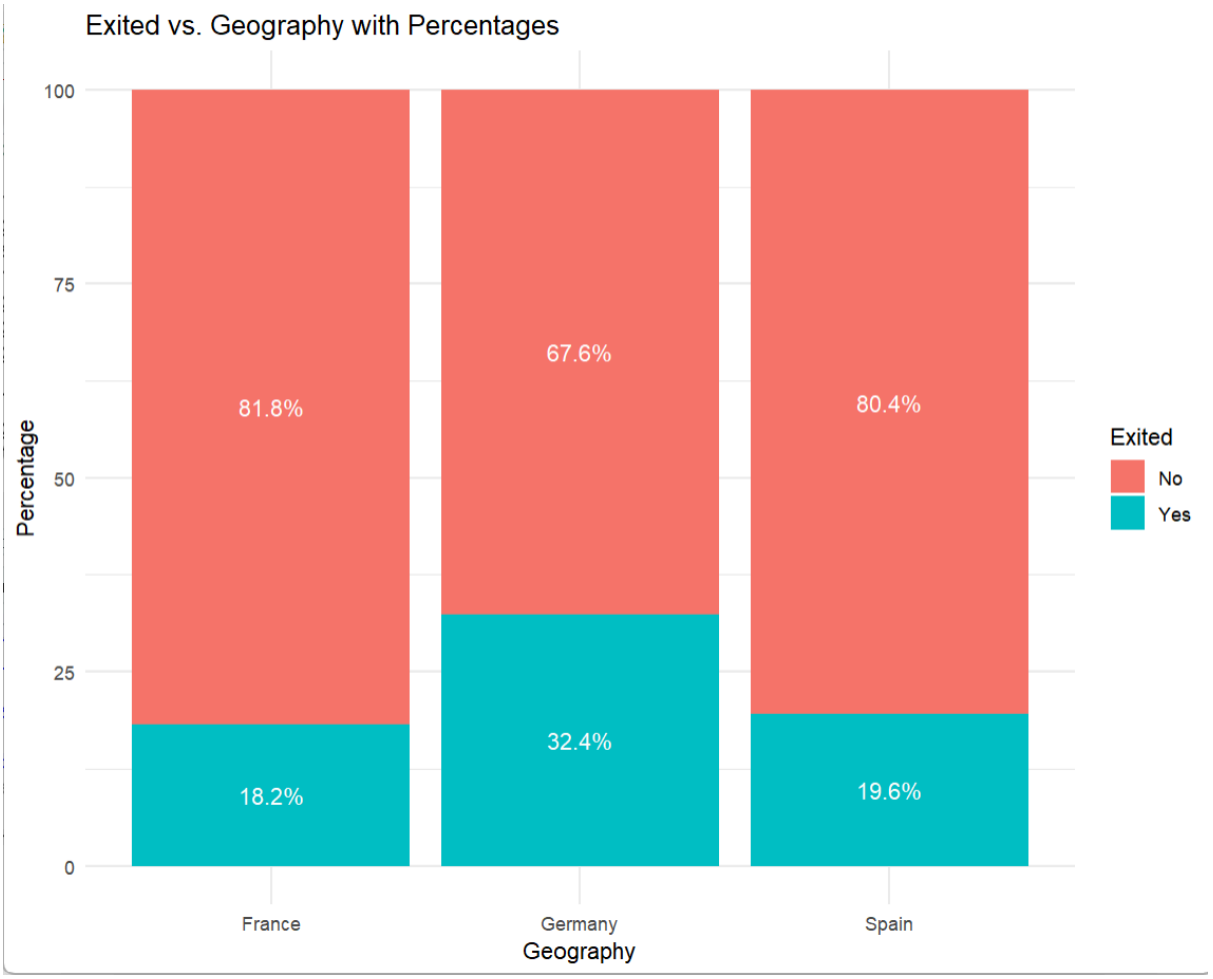
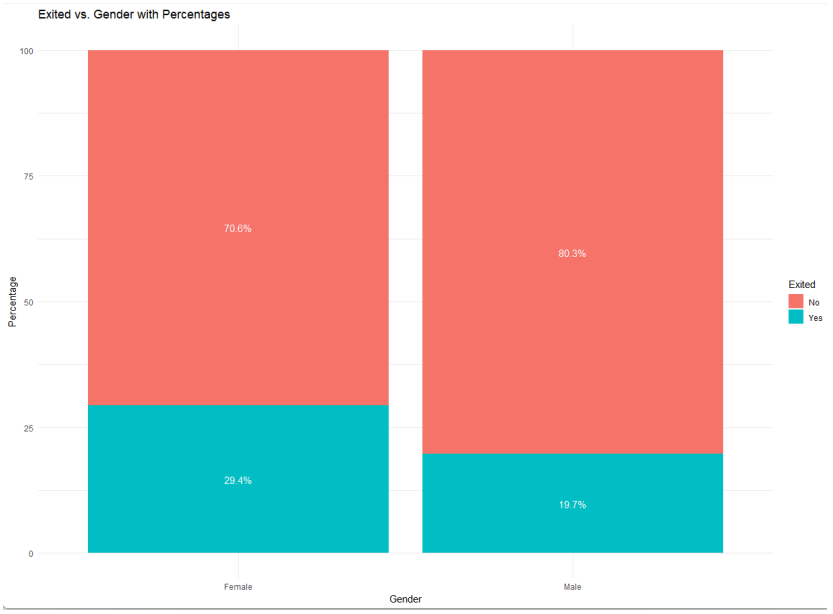


Fig 8: Churn distribution concerning Geography

3. Gender distribution:

From Fig 9, female customers are more likely to churn when compared to males.




```
gender_exit <- bank_churn %>%
  group_by(Gender, Exited) %>%
  summarise(count = n()) %>%
  mutate(percentage = count / sum(count) * 100)

ggplot(gender_exit, aes(x = Gender, y = percentage, fill = factor(Exited))) +
  geom_bar(stat = "identity", position = "stack") +
  geom_text(aes(label = paste0(round(percentage, 1), "%")),
            position = position_stack(vjust = 0.5), color = "white") +
  labs(x = "Gender", y = "Percentage", fill = "Exited") +
  ggtitle("Exited vs. Gender with Percentages") +
  theme_minimal()
```

Fig 9: Churn distribution concerning Gender

4.4 Visualizing account information:

1. Credit Score:

The customers who have churned have a credit score between 600 and 700.

```
ggplot(bank_churn, aes(x = CreditScore, fill = Exited)) +
  geom_bar() +
  labs(x = "Credit Score", y = "Count", fill = "Exited") +
  ggtitle("Exited vs Credit Score") +
  theme_minimal()
```

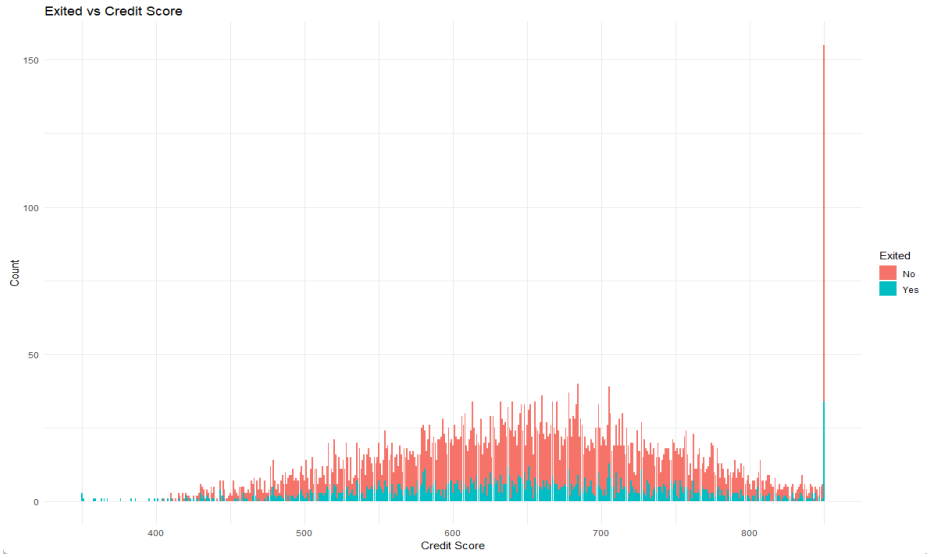
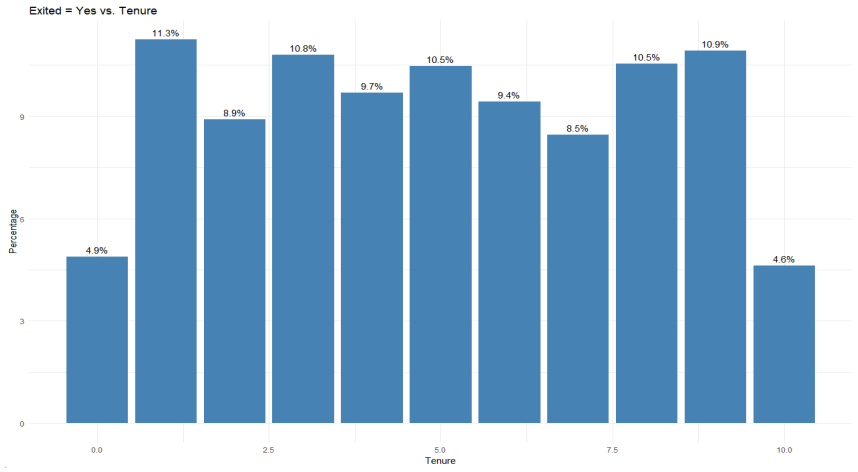


Fig 10: Churn concerning Credit Score

2. Visualizing tenure vs churn:

Tenure does not seem to have an effect on customer churn. However, it looks like new customers tend to leave more when compared to tenured customers.

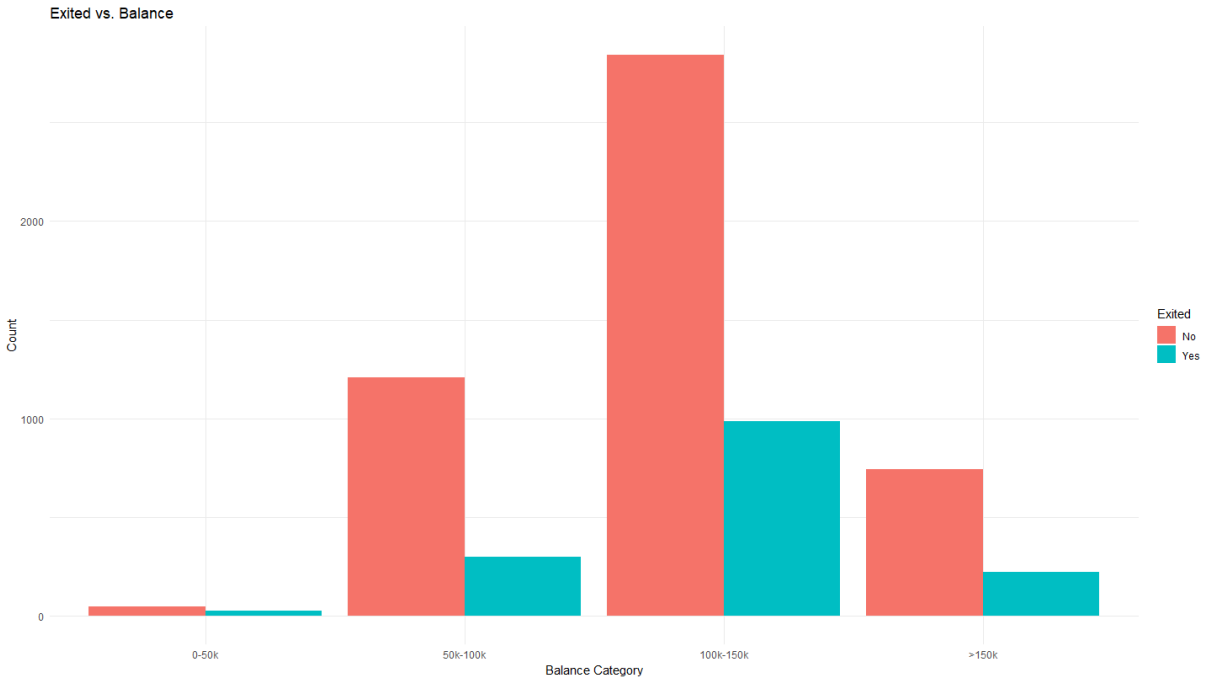



```
exited_yes <- bank_churn %>%
  filter(Exited == "Yes")

tenure_exit <- exited_yes %>%
  group_by(Tenure) %>%
  summarise(count = n()) %>%
  mutate(percentage = count / sum(count) * 100)

ggplot(tenure_exit, aes(x = Tenure, y = percentage)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  geom_text(aes(label = paste0(round(percentage, 1), "%")),
            vjust = -0.5) +
  labs(x = "Tenure", y = "Percentage", fill = "Exited") +
  ggtitle("Exited = Yes vs. Tenure") +
  theme_minimal()
```

Fig 11: Churn concerning Tenure

3. Distribution of Balance:

Fig 12 shows that customers with the balance between 100-150K are most likely to churn.

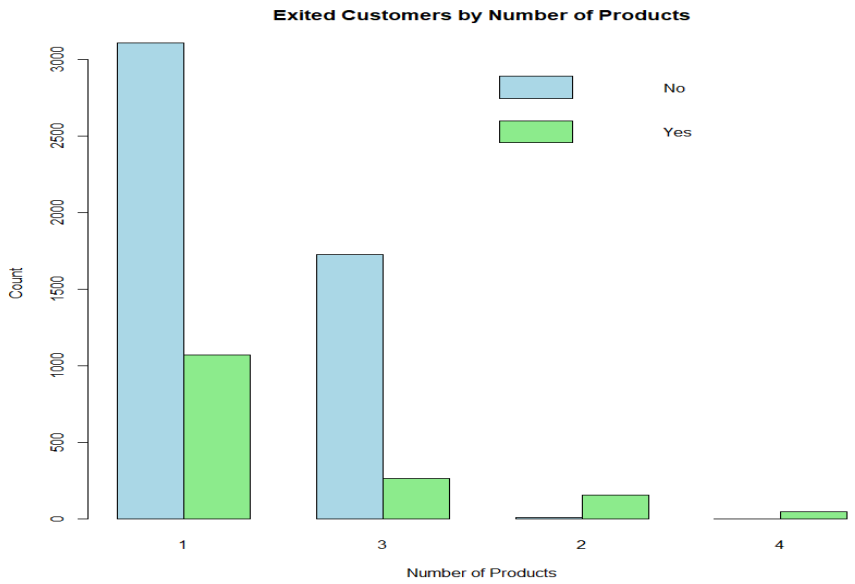



```
exit_bal <- bank_churn %>%
  group_by(Exited) %>%
  mutate(BalanceCategory = cut(Balance, breaks = c(0, 50000, 100000, 150000, Inf),
                               labels = c("0-50k", "50k-100k", "100k-150k", ">150k"))) %>%
  group_by(BalanceCategory, Exited) %>%
  summarise(count = n())

ggplot(exit_bal, aes(x = BalanceCategory, y = count, fill = Exited)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(x = "Balance Category", y = "Count", fill = "Exited") +
  ggtitle("Exited vs. Balance") +
  theme_minimal()
```

Fig 12: Churn concerning Balance

4. Number of Products vs Churn:

Most of the customer purchase 1 or 2 products from the bank and customers with only 1 product are most likely to churn.




```r
unique_products <- unique(bank_churn$NumOfProducts)

exit_by_products <- table(bank_churn$Exited, bank_churn$NumOfProducts)

barplot(exit_by_products, beside = TRUE, legend = TRUE,
        main = "Exited Customers by Number of Products",
        xlab = "Number of Products", ylab = "Count",
        col = c("lightblue", "lightgreen"),
        names.arg = unique_products,
        args.legend = list(x = "topright", bty = "n"))
```

Fig 13: Churn concerning the Number of Products purchased

5. Active members vs Customer churn:

Customers who are not active are most likely to churn. A significant number of customers are inactive and therefore the bank can change its policies to make sure that customers become active.

```r
activemember <- bank_churn %>%
  group_by(IsActiveMember, Exited) %>%
  summarise(count = n()) %>%
  mutate(percentage = count / sum(count) * 100)

ggplot(activemember, aes(x = IsActiveMember, y = percentage, fill = factor(Exited))) +
  geom_bar(stat = "identity", position = "dodge") +
  geom_text(aes(label = paste0(round(percentage, 1), "%")),
            position = position_dodge(width = 0.9), vjust = -0.5) +
  labs(x = "Active Member", y = "Percentage", fill = "Exited") +
  ggtitle("Exited vs. Active Members") +
  theme_minimal()
```

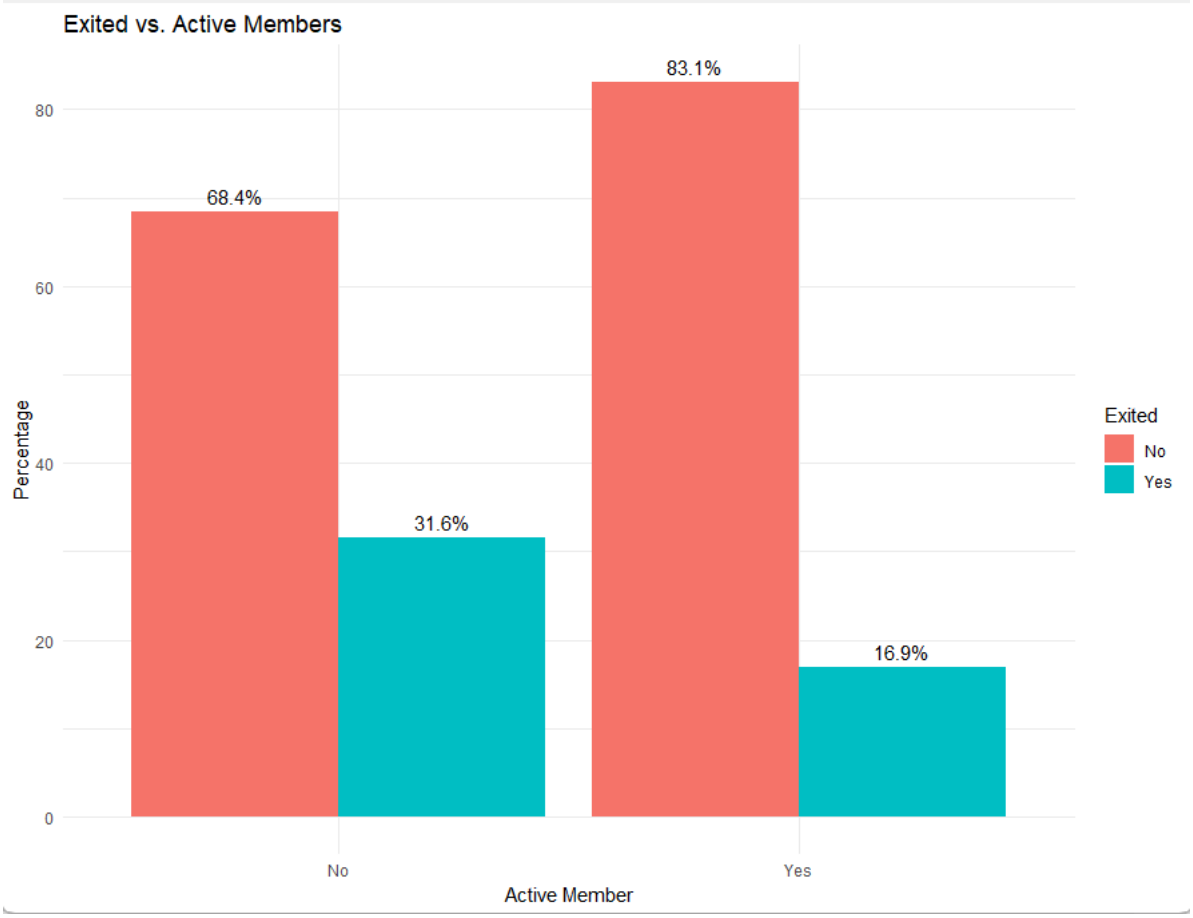


Fig 14: Customer churn concerning Active Members

5. Forecasting and Predictive Modeling:

5.1 Decision Tree:

The dataset is split into train and test data where 70% is assigned to train data and 30% is assigned to test data and then fit the decision tree model.

```
set.seed(123)  # for reproducibility
train_indices <- sample(nrow(bank_churn), nrow(bank_churn) * 0.7)
train_data <- bank_churn[train_indices, ]
test_data <- bank_churn[-train_indices, ]
```

```
data_model <- rpart(Exited ~ ., data=train_data, method="class")
data_pred <- predict(data_model, test_data, type="class")

data_predf = factor(data_pred)
test_data$Exited = factor(test_data$Exited)

rpart.plot(data_model)
```

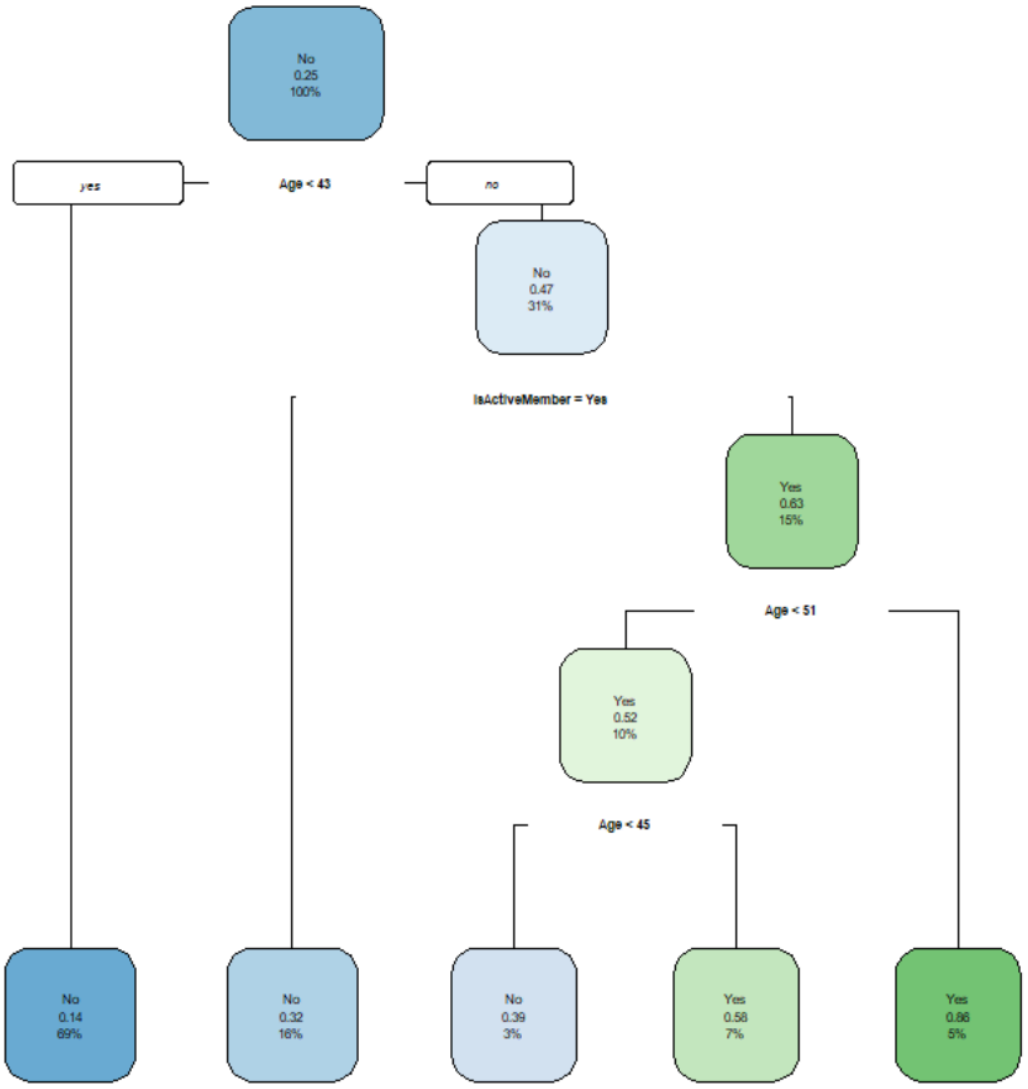


Fig 15: Decision Tree

According to the decision tree plot, depending on customer demographics and account information, the model predicted that 25% of customers are likely to churn.

```
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  1427  306
       Yes   47  135

               Accuracy : 0.8157
                 95% CI : (0.7976, 0.8328)
    No Information Rate : 0.7697
    P-Value [Acc > NIR] : 5.681e-07

                  Kappa : 0.3453

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

            Sensitivity : 0.9681
            Specificity : 0.3061
         Pos Pred Value : 0.8234
         Neg Pred Value : 0.7418
             Prevalence : 0.7697
         Detection Rate : 0.7452
   Detection Prevalence : 0.9050
      Balanced Accuracy : 0.6371
```

```
dt_cm <- confusionMatrix(data_predf, test_data$Exited)
print(dt_cm)
```

Fig 16: Confusion Matrix

The model predicted 1427 no cases and 135 yes cases correctly. The confusion matrix shows 81.5% accuracy. The sensitivity is 0.96 and the specificity is 0.306. The prevalence is 0.7697 which means that 76.97% of the instances of the test set are the response “no” which means that 76.97% of the customers would stay with the bank.

5.2 Logistic Regression:

Logistic regression is performed using SPSS to check the dependency of churn variables on other attributes. The variables used to check dependency are Age, Geography, Number of Products, and Gender. The below shows a significance <0.001 which means that the null hypothesis can be rejected.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	789.231	5	<.001
	Block	789.231	5	<.001
	Model	789.231	5	<.001

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	6257.544[a]	.116	.174

Variance

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

The model summary shows the variance and -2 log likelihood which represents minimization. The Nagelkerke R square which denotes variance is 0.174 and the Cox & Snell R square is 0.116 which means that there is an 11.6% probability of customer churn is explained by logistic regression.

Classification Table[a]

			Predicted		
			Exited		Percentage Correct
	Observed		0	1	
Step 1	Exited	0	4609	237	95.1
		1	1288	249	16.2
	Overall Percentage				76.1

a. The cut value is .500

The above table shows the results of classification with 76.1% accuracy. The below table is a crucial part of logistic regression as it contains the significance of the coefficients of each variable.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Geography			119.413	2	<.001	
	Geography(1)	.715	.071	101.500	1	<.001	2.044
	Geography(2)	.059	.092	.420	1	.517	1.061
	Gender(1)	-.526	.063	69.910	1	<.001	.591
	NumOfProducts	.208	.052	16.137	1	<.001	1.231
	Age	.066	.003	494.710	1	<.001	1.068
	Constant	-4.195	.158	708.622	1	<.001	.015

a. Variable(s) entered on step 1: Geography, Gender, NumOfProducts, Age.

6. Results:

From data exploration and visualization, the attributes contributing to customer churn were identified. It can be seen from exploration and visualization that the most significant attribute is “Age” where older customers are most likely to churn. The number of products is also an important variable that affects customer churn. Customers with a greater number of products stay with the bank. Tenure does not have to seem a huge impact on customer churn. However, the bank can run marketing campaigns to understand the needs of the tenured customers and also provide attractive offers to new customers which would help the bank retain them. The predictive analytics techniques like a decision tree and logistic regression showed an accuracy of 79.7% and 76.1% which means that the analytics could meaningfully detect approximately 80% of customers who are likely to churn. Adding more attributes to the dataset could improve the performance of predictive analytics.

7. Conclusion:

The study aims to predict the customers of a bank who are most likely to churn. The decision tree and classification have good accuracy and prediction. These models have helped in a good understanding of the data set. In conclusion, the report shows a good understanding of predictive and forecasting analytics besides data visualizations using tools like R and SPSS. By exploring the SVND Pvt Ltd business case and applying predictive modeling techniques, the report showed the effectiveness of the analytical tools and machine learning modeling to extract meaningful insights from the data.

8. Future Work:

The churn analysis is limited to the attributes present in the data set. Including more variables like marketing campaigns and products offered will further help predict customer churn accurately and effectively suggest customer retention measures. The study of churn analysis in this report was based on the current account balance

and if the study was based on loyalty program or active usage of internet banking, it could possibly provide a greater lifetime customer value (Mutanen, 2006).

References:

Bin, L., Peiji, S., & Juan, L. (2007, June). Customer churn prediction based on the decision tree in personal handyphone system service. In *2007 International Conference on Service Systems and Service Management* (pp. 1-5). IEEE.

Chitra, K., & Subashini, B. (2011). Customer retention in banking sector using predictive data mining technique. In *ICIT 2011 The 5th International Conference on Information Technology*.

Cohen, D., Gan, C., Au Yong, H. H., & Chong, E. (2007). *Customer retention by banks in New Zealand* (Doctoral dissertation, Українська академія банківської справи Національного банку України).

Dalvi, P. K., Khandge, S. K., Deomore, A., Bankar, A., & Kanade, V. A. (2016, March). Analysis of customer churn prediction in telecom industry using decision trees and logistic regression. In *2016 symposium on colossal data analysis and networking (CDAN)* (pp. 1-4). IEEE.

Darzi, M. A., & Bhat, S. A. (2018). Personnel capability and customer satisfaction as predictors of customer retention in the banking sector: A mediated-moderation study. *International journal of bank marketing*, *36*(4), 663-679.

Karvana, K. G. M., Yazid, S., Syalim, A., & Mursanto, P. (2019, October). Customer churn analysis and prediction using data mining models in banking industry. In *2019 International Workshop on Big Data and Information Security (IWBIS)* (pp. 33-38). IEEE.

Keramati, A., Ghaneei, H., & Mirmohammadi, S. M. (2016). Developing a prediction model for customer churn from electronic banking services using data mining. *Financial Innovation*, *2*, 1-13.

Kumar, V., & Garg, M. L. (2018). Predictive analytics: a review of trends and techniques. *International Journal of Computer Applications*, *182*(1), 31-37.

Larose, D. T. (2015). *Data mining and predictive analytics*. John Wiley & Sons.

LaValley, M. P. (2008). Logistic regression. *Circulation*, *117*(18), 2395-2399.

Lee, C. S., Cheang, P. Y. S., & Moslehpour, M. (2022). Predictive analytics in business analytics: decision tree. *Advances in Decision Sciences*, *26*(1), 1-29.

Meyers, L. S., Gamst, G. C., & Guarino, A. J. (2013). Performing data analysis using IBM SPSS. John Wiley & Sons.

Mishra, N., & Silakari, S. (2012). Predictive analytics: a survey, trends, applications, oppurtunities & challenges. *International Journal of Computer Science and Information Technologies*, *3*(3), 4434-4438.

Mutanen, T. (2006). Customer churn analysis–a case study. *Journal of Product and Brand Management*, *14*(1), 4-13.

Rahman, M., & Kumar, V. (2020, November). Machine learning based customer churn prediction in banking. In *2020 4th international conference on electronics, communication and aerospace technology (ICECA)* (pp. 1196-1201). IEEE.

Sadiku, M., Shadare, A. E., Musa, S. M., Akujuobi, C. M., & Perry, R. (2016). Data visualization. *International Journal of Engineering Research And Advanced Technology (IJERAT)*, *2*(12), 11-16.

Saran Kumar, A., & Chandrakala, D. (2016). A survey on customer churn prediction using machine learning techniques. *International Journal of Computer Applications*, *975*, 8887.

Shinde, P. P., Oza, K. S., & Kamat, R. K. (2017, February). Big data predictive analysis: Using R analytical tool. In *2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)* (pp. 839-842). IEEE.

Verzani, J. (2011). *Getting started with RStudio*. " O'Reilly Media, Inc.".

Wagner III, W. E. (2019). *Using IBM® SPSS® statistics for research methods and social science statistics*. Sage Publications.