JAHNAVI RAMBHA

AWDCRJY

2ND YEAR,DATA SCIENCE

PROJECT NAME:

CHURN MODELLING

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S.NO** | **CONTENTS** | | **PAGE-NO** | |
| **1.** | **OVERVIEW AND PROBLEM STATEMENT** | | **3** | |
| **2.** | **ABOUT THE DATA SET** | | **4-5** | |
| **3.** | **EDA STEP BY STEP PROCESS** | | **6-28** | |
| **4.** | | **CONCLUSION** | | **29** |

**INDEX**

**OVERVIEW**

You have to use EDA to analyze the customer who are churned We will Perform this Exploratory Data Analysis in order to get actionable insights, and convert them into meaningful stories and present it so that the companies can take necessary actions in order to retain their customers

**PROBLEM STATEMENT**

Conduct a thorough Exploratory Data Analysis on the bank customer dataset to reveal key characteristics, trends, and potential factors influencing customer churn. Identify and visualize patterns in customer behavior, demographics, and interactions with the bank's services. Additionally, highlight any anomalies, missing values, or outliers that may impact the quality of the dataset. The insights derived from EDA will guide feature selection, preprocessing steps, and model development for predicting customer churn effectively.

**ABOUT THE DATA SET**

* **The data set is all about the churning customers in the last year**
* **Our data set consists of in total 14 columns which are:**

1. **Customer\_ID-It give the unique id of the customer**
2. **RouNumber:It gives the Row number of entry**
3. **Gender : It gives whether the customer is male or female male or female**
4. **Geography : It gives to what place does the customer belongs to**
5. **HasCrcard- It gives the information whether the person is having credit card or not**
6. **IsActiveMember: It gives information whether the customer is active or not**
7. **EstimateSalary: It gives estimated salary of a customer**
8. **Balance:It gives the information about the balance of customer**
9. **Tenure:It shows the months in data**
10. **Age:Is shows the age of a customer**
11. **NumOfProducts:it shows the number of products**
12. **Exited : it shows whether the customer is churned or not**
13. **CreditScore:It shows the credit score of a customer**
14. **Surname:It gives the surname of the customer**

**EXPLORATORY DATA ANALYSIS STEP BY STEP PROCESS**

* **STEP-1**:Installing and loading Required libraries for the data

library(ggplot2)

library(dplyr)

library(plotly)

library(plotrix)

library(cowplot)

library(Hmisc)

library(tidyr)

library(scales)

library(MASS)

library(psych)

library(naniar)

library(grid)

library(corrplot)

* **STEP-2**: Loading Our csv file into our R studio

churnData=read.csv(file.choose(),header = TRUE)

* **STEP-3:**Viewing the first and last elements for a overview on data

head(churnData)

tail(churnData)

* **STEP-4:** Knowing the dimensions,structure , and statistical information about data

##dimensions of data

dim(churnData)

##structure of data

str(churnData)

## statistical info

describe(churnData)

summary(churnData)

* **STEP-5**:Checking whether we are having null values are not

colSums(is.na(churnData))

* **STEP-6:**Since we are not having null values in our data.Now let us add few null values in our data

#Not to effect the original data let us make a copy of the churnData

churn2Data=churnData

##Now let us insert null values in our copy of data

set.seed(123)

churn2Data[sample(seq(NROW(churn2Data)),round(3 / 100 \* nrow(churn2Data))),"Age"]<-NA

set.seed(123)

churn2Data[sample(seq(NROW(churn2Data)),round(5 / 100 \* nrow(churn2Data))),"Gender"]<-NA

* **STEP-7:** Plotting graph for missing values

gg\_miss\_var(churn2Data)+ggtitle("Missing Values in a data")

* **STEP-8:**Starting the Data cleaning process

#filling age col with their mean

churn2Data <- churn2Data %>%

mutate(Age = ifelse(is.na(Age), median(churn2Data$Age,na.rm=TRUE), Age))

##Here gender is categorical so we should fill it with mode but in r

#We dont have mode function so let let us create mode function

Mode <- function(x) {

ux <- unique(x)

ux[which.max(tabulate(match(x, ux)))]

}

churn2Data<-churn2Data %>%

mutate(Gender=ifelse(is.na(Gender),Mode(Gender),Gender))

colSums(is.na(churn2Data))##We are having 0 null values

#2.Handling duplicated values

##Identifying whether we have duplicate values or not

churn2Data[duplicated(churn2Data),]#No duplicate values

#3.Outliers

# Selecting numerical columns

numericalColumns <- churn2Data[, c("CreditScore", "Age", "Tenure", "Balance", "NumOfProducts", "EstimatedSalary")]

# Creating boxplots for each numerical column

par(mfrow = c(2, 3)) # 2 rows, 3 columns #par-parameter #mfrow-multiple figure row

for (col in colnames(numericalColumns)) {

boxplot(numericalColumns[[col]], main = col)

}

for (col in colnames(churn2Data)[sapply(churn2Data, is.numeric)]) {

outliers <- boxplot(churn2Data[[col]], plot = FALSE)$out

cat("Column:", col, "\n")

cat("Number of outliers:", length(outliers), "\n")

cat("Outliers:", outliers, "\n\n")

}

#We are having extreme values but it is not effecting our result

* **STEP-9:** Now the further step is Data transformation

#5.Transforming Data for univariate analysis

churn2Data <- churn2Data %>%

mutate(Exited = ifelse(Exited == 1, "Yes", "No"))

churn2Data <- churn2Data %>%

mutate(IsActiveMember = ifelse(IsActiveMember == 1, "Yes", "No"))

churn2Data <- churn2Data %>%

mutate(HasCrCard = ifelse(HasCrCard== 1, "Yes", "No"))

##for tenure data

labels <- sprintf("%d - %d", seq(0, 11, 2), seq(1, 11, 2))

# Cut the 'tenure' column into bins

churn2Data$Tenure\_group <- cut(

churn2Data$Tenure,

breaks = c(seq(0, 11, 2), Inf),

right = FALSE,

labels = labels

)

#6.dropping the columns that are not required

churn2Data<-subset(churn2Data,select = -c(RowNumber,CustomerId,Surname))

head(churn2Data)

* **STEP-10:**Now its time for Data exploration

#1.Univariate analysis

##As our Problem statement is analyse the churn customers let us have a lookon that

custom\_colors <- c('#FF5733', '#33FF57', '#5733FF', '#FF5733', '#33FF57', '#5733FF')

#exited

valueCount=table(churn2Data$Exited)

df\_value\_counts=as.data.frame(valueCount)

colnames(df\_value\_counts) <- c("Variable", "Count")

df\_value\_counts

percentageCount=prop.table(valueCount)\*100

percentageCount

plot\_ly(df\_value\_counts, x = ~Variable, y = ~Count) %>%

layout(title = 'No.of customers churned', xaxis = list(title = 'Target variable'), yaxis = list(title = 'count')) %>%

add\_trace(marker = list(color = custom\_colors))

#card

valueCountcard=table(churn2Data$HasCrCard)

df\_value\_countcard=as.data.frame(valueCountcard)

colnames(df\_value\_countcard) <- c("Variable", "Count")

plot\_ly(df\_value\_countcard, x = ~Variable, y = ~Count, name = 'Bar Plot') %>%

layout(title = 'No.of Customers based on card', xaxis = list(title = 'Target variable'), yaxis = list(title = 'count'))%>%

add\_trace(marker = list(color = c('yellow','purple')))

#gender

valueCountGender=table(churn2Data$Gender)

df\_value\_countGender=as.data.frame(valueCountGender)

colnames(df\_value\_countGender) <- c("Variable", "Count")

plot\_ly(df\_value\_countGender, x = ~Variable, y = ~Count, name = 'Bar Plot') %>%

layout(title = 'No.of Customers based on card', xaxis = list(title = 'Target variable'), yaxis = list(title = 'count'))%>%

add\_trace(marker = list(color = c('blue','pink')))

#active

valueCountactive=table(churn2Data$IsActiveMember)

df\_value\_countactive=as.data.frame(valueCountactive)

colnames(df\_value\_countactive) <- c("Variable", "Count")

plot\_ly(df\_value\_countactive, x = ~Variable, y = ~Count, name = 'Bar Plot') %>%

layout(title = 'No.of Customers based on active', xaxis = list(title = 'Target variable'), yaxis = list(title = 'count'))%>%

add\_trace(marker = list(color = c('black','grey')))

#products

valueCountproducts=table(churn2Data$NumOfProducts)

df\_value\_countproducts=as.data.frame(valueCountproducts)

colnames(df\_value\_countproducts) <- c("Variable", "Count")

plot\_ly(df\_value\_countproducts, x = ~Variable, y = ~Count, name = 'Bar Plot') %>%

layout(title = 'No.of Customers based on products', xaxis = list(title = 'Target variable'), yaxis = list(title = 'count'))%>%

add\_trace(marker = list(color = c('brown','darkgreen','darkblue','red')))

#age

valueCountage=table(churn2Data$Age)

df\_value\_countage=as.data.frame(valueCountage)

colnames(df\_value\_countage) <- c("Variable", "Count")

plot\_ly(df\_value\_countage, x = ~Variable, y = ~Count, name = 'Bar Plot') %>%

layout(title = 'No.of Customers based on age', xaxis = list(title = 'Target variable'), yaxis = list(title = 'count'))%>%

add\_trace(marker = list(color = 'violet'))

#tenure

valueCounttenure=table(churn2Data$Tenure\_group)

df\_value\_counttenure=as.data.frame(valueCounttenure)

colnames(df\_value\_counttenure) <- c("Variable", "Count")

plot\_ly(df\_value\_counttenure, x = ~Variable, y = ~Count, name = 'Bar Plot') %>%

layout(

title = 'No.of Customers based on age',

xaxis = list(title = 'Target variable'),

yaxis = list(title = 'count'),

showlegend = TRUE,

bargap = 0.05 # Adjust the gap between bars if needed

) %>%

add\_trace(marker = list(color = c('green','skyblue','gold','blue','brown')))

##From the graph we can undersatand that exited ratio is almost 20:30

#Let us analyse this by other categories further

#Univariate analysis

#For gender

plotforgender <- ggplot(churn2Data, aes(x = Gender, fill = Exited)) +

geom\_bar(position = "dodge") +

geom\_text(

aes(label = stat(count)),

stat = "count",

position = position\_dodge(width = 0.9),

vjust = -0.5

) +

labs(title = "Countplot by Churn based on gender")

print(plotforgender)

#for card holders

plotforcard <- ggplot(churn2Data, aes(x = HasCrCard, fill = Exited)) +

geom\_bar(position = "dodge") +

geom\_text(

aes(label = stat(count)),

stat = "count",

position = position\_dodge(width = 0.9),

vjust = -0.5

) +

labs(title = "Countplot by Churn based on card holders")

print(plotforcard)

#For active members

plotforactivemembers <- ggplot(churn2Data, aes(x = IsActiveMember, fill = Exited)) +

geom\_bar(position = "dodge") +

geom\_text(

aes(label = stat(count)),

stat = "count",

position = position\_dodge(width = 0.9),

vjust = -0.5

) +

labs(title = "Countplot by Churn based on Active members")

print(plotforactivemembers)

#Geography

plotforgeography <- ggplot(churn2Data, aes(x = Geography, fill = Exited)) +

geom\_bar(position = "dodge") +

geom\_text(

aes(label = stat(count)),

stat = "count",

position = position\_dodge(width = 0.9),

vjust = -0.5

) +

labs(title = "Countplot by Churn based on Geography")

print(plotforgeography)

#For tenure

plotfortenure <- ggplot(churn2Data, aes(x = Tenure\_group, fill = Exited)) +

geom\_bar(position = "dodge") +

geom\_text(

aes(label = stat(count)),

stat = "count",

position = position\_dodge(width = 0.9),

vjust = -0.5

) +

labs(title = "Countplot by Churn based on Tenure")

print(plotfortenure)

#for products

plotforproducts <- ggplot(churn2Data, aes(x = NumOfProducts, fill = Exited)) +

geom\_bar(position = "dodge") +

geom\_text(

aes(label = stat(count)),

stat = "count",

position = position\_dodge(width = 0.9),

vjust = -0.5

) +

labs(title = "Countplot by Churn based on No.of Products")

print(plotforproducts)

#For credit score

ggplot(churnData, aes(x = CreditScore, fill = factor(Exited))) +

geom\_density(alpha = 0.8) +

labs(title = "Density Plot of CreditScore by Exited Status",

x = "Credit Score",

y = "Density")+

scale\_fill\_manual(values = c("green", "purple"), name = "Exited") +theme\_minimal()

#For Estimated Salary

ggplot(churnData, aes(x = EstimatedSalary, fill = factor(Exited))) +

geom\_density(alpha = 0.8) +

labs(title = "Density Plot of EstimatedSalary by Exited Status",

x = "Estimated Salary",

y = "Density") +

scale\_fill\_manual(values = c("red", "blue"), name = "Exited") +

theme\_minimal()

#For Balance

ggplot(churnData, aes(x = Balance, fill = factor(Exited))) +

geom\_density(alpha = 0.9) +

labs(title = "Density Plot of Balance by Exited Status",

x = "Balance",

y = "Density") +

scale\_fill\_manual(values = c("yellow", "purple"), name = "Exited") +

theme\_minimal()

#for age

ggplot(churn2Data, aes(x = Age,fill=factor(Exited))) +

geom\_density(alpha = 0.5) +

scale\_fill\_manual(values = c("brown", "skyblue")) +

labs(fill = "Churn") +

labs(y = "Density", x = "Age") +

ggtitle("Age by Churn") +

theme\_minimal()

* **STEP-11:** As we explored our data by univariate analysis still we didn’t got a correct insights about geography for that let us make bivariate analysis for them with respect to gender.
* **STEP-12:**To do the Bivariate analysis.Let us divide our data into exited and not exited groups as our problem statement lies on the churning which means exited and not exited

#for bivariate analysis dividing it into exited non exited

exited=churn2Data[churn2Data$Exited=='Yes',]

notexited=churn2Data[churn2Data$Exited=='No',]

* **STEP:13:**Plotting graphs for bivariate analysis

#Age and Gender

ggplot(exited, aes(x=Gender, fill=Geography)) +

geom\_bar(position="dodge", stat="count") +

labs(title="Exited customers based Gender and Geography",

x="Gender",

y="Count") +

geom\_text(

aes(label = stat(count)),

stat = "count",

position = position\_dodge(width = 0.9),

vjust = -0.5

)+

theme\_minimal()

#card and gender

ggplot(exited, aes(x=Gender, fill=HasCrCard)) +

geom\_bar(position="dodge", stat="count") +

labs(title="Exited customers based Gender and card holders",

x="Gender",

y="Count") +

geom\_text(

aes(label = stat(count)),

stat = "count",

position = position\_dodge(width = 1),

vjust = -0.5

)+

theme\_minimal()

#geography and gender

ggplot(exited, aes(x=Geography, fill=HasCrCard)) +

geom\_bar(position="dodge", stat="count") +

labs(title="Exited customers based Gender and card holders",

x="Gender",

y="Count") +

geom\_text(

aes(label = stat(count)),

stat = "count",

position = position\_dodge(width = 1),

vjust = -0.5

)+

theme\_minimal()

* **STEP:14:**Now let us have a look on not exited customers by plotting the graphs

#piechart for geography

chart1= ggplot(notexited, aes(x = "", fill = Geography)) +

geom\_bar(width = 1, stat = "count") +

coord\_polar("y") +

ggtitle("Distribution of Geography")

# Pie chart for Gender

chart2=ggplot(notexited, aes(x = "", fill = Gender)) +

geom\_bar(width = 1, stat = "count") +

coord\_polar("y") +

ggtitle("Distribution of Gender")

# Pie chart for HasCrcard

chart3=ggplot(notexited, aes(x = "", fill = as.factor(HasCrCard))) +

geom\_bar(width = 1, stat = "count") +

coord\_polar("y") +

ggtitle("Distribution of HasCrcard")

# Pie chart for Isactive

chart4=ggplot(notexited, aes(x = "", fill = as.factor(IsActiveMember))) +

geom\_bar(width = 1, stat = "count") +

coord\_polar("y") +

ggtitle("Distribution of Isactive")

grid.arrange(

top = textGrob("Combined Distribution of Charts for non-exited customers", gp = gpar(fontsize = 16, fontface = "bold")),

arrangeGrob(chart1, chart2, chart3, chart4, ncol = 2)

)

* **STEP-15:**Now let us make numerical analysis..by plotiing the heat map which gives correlation of variables

#Numerical Analysis

numeric\_data <- select\_if(churn2Data, is.numeric)

correlation\_matrix <- cor(numeric\_data)

print(as.data.frame(correlation\_matrix))

correlation\_with\_churn <- correlation\_matrix["Exited",]

# Create a plot of the correlation with 'Churn'

corrplot(correlation\_matrix, method = "color", col = c("darkblue", "green", "darkred",'purple'), scale = "none")

**CONCLUSION:**

* **According to our Findings Females are more likely to churn**
* **Tenure of six and seven month has high churning rate , this may be because of seasonal variations.**
* **Germany is having high churning rate compared to France and Spain**
* **Products are directly proportional to churning rate. Three and Four Products have high churning rate**
* **Age between 40-50 are more likely to churn**
* **Customers with credit score between 600-700 are more likely to churn**
* **As From our Analysis Having card is not playing a crucial role because cards alone doesn’t have anY insight and cards along with gender doesn’t have any insight**