Dataset Information

Table1: Description of Variables in Dataset

S/N	Variables	Description of Variables	Class of Variable	Data Type
1	RowNumber	RowNumber	Dependent	Not Applicable
2	CustomerId	Bank assigned unique ID	Dependent	Not Applicable
3	Surname	Surname	Dependent	Not Applicable
4	Credit Score	Credit score between 600-800	Dependent	Numerical
5	Geography	The country the customer is from	Dependent	Categorical
6	Gender	Male or Female	Dependent	Categorical
7	Age	age	Dependent	Numerical
8	Tenure	How many years he/she is a customer of the bank	Dependent	Categorical
9	Balance	Amount in the account	Dependent	Continuous
10	NumberOfProducts	How many products he/she bought from the bank	Dependent	Numerical
11	HasCrCard	Has CreditCard (1) or not (0)	Dependent	Numerical
12	IsActiveMember	Whether he/she is an active member of the bank (1) or not (0)	Dependent	Categorical
13	EstimatedSalary	Salary of the person estimated by the bank	Dependent	Continuous
14	Exited	A customer leaving (1) or not leaving (0)	Independent	Categorical

Research Questions

To determine and predict the categories and age spectrum of customers that exited the bank.

Exploratory Data Analysis and Results

Data Cleansing

Checking for missing data

```
> #checking missing data
> sum(is.na(churn))
[1] 0
```

Figure1: Checking Missing Values

There was no missing value found

Importing the data file to R platform

Datasets in R platform

	RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
1	1	15634602	Hargrave	619	France	Female	42	2	0.00	1	1	1	101348.88	1
2	2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
3	3	15619304	Onio	502	France	Female	42	8	159660.80	3	1	0	113931.57	1
4	4	15701354	Boni	699	France	Female	39	1	0.00	2	0	0	93826.63	0
5	5	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1	79084.10	0
6	6	15574012	Chu	645	Spain	Male	44	8	113755.78	2	1	0	149756.71	1
7	7	15592531	Bartlett	822	France	Male	50	7	0.00	2	1	1	10062.80	0
8	8	15656148	Obinna	376	Germany	Female	29	4	115046.74	4	1	0	119346.88	1
9	9	15792365	He	501	France	Male	44	4	142051.07	2	0	1	74940.50	0
10	10	15592389	H?	684	France	Male	27	2	134603.88	1	1	1	71725.73	0
11	11	15767821	Bearce	528	France	Male	31	6	102016.72	2	0	0	80181.12	0
12	12	15737173	Andrews	497	Spain	Male	24	3	0.00	2	1	0	76390.01	0
13	13	15632264	Kay	476	France	Female	34	10	0.00	2	1	0	26260.98	0
14	14	15691483	Chin	549	France	Female	25	5	0.00	2	0	0	190857.79	0
15	15	15600882	Scott	635	Spain	Female	35	7	0.00	2	1	1	65951.65	0
16	16	15643966	Goforth	616	Germany	Male	45	3	143129.41	2	0	1	64327.26	0
17	17	15737452	Romeo	653	Germany	Male	58	1	132602.88	1	1	0	5097.67	1

Showing 1 to 18 of 10,000 entries, 14 total columns

Removing Irrelevant variables-Data Cleansing

RowNumber, CustomerId and Surname were removed because they do not contribute to the classification.

```
> churn <- data_churn %>%
+   select(-c(RowNumber,CustomerId, Surname,))%>%
+   mutate(HasCrCard = as.factor(HasCrCard),
+          IsActiveMember = as.factor(IsActiveMember),
+          Exited = as.factor(Exited))
> glimpse(churn)
Rows: 10,000
Columns: 11
$ CreditScore     <int> 619, 608, 502, 699, 850, 645, 822, 376, 501, 684, 528, 497, 476, 549, 635, 616, 653, 549, 587,…
$ Geography       <chr> "France", "Spain", "France", "France", "Spain", "Spain", "France", "Germany", "France", "Franc…
$ Gender          <chr> "Female", "Female", "Female", "Female", "Female", "Male", "Male", "Female", "Male", "Male", "M…
$ Age             <int> 42, 41, 42, 39, 43, 44, 50, 29, 44, 27, 31, 24, 34, 25, 35, 45, 58, 24, 45, 24, 41, 32, 38, 46…
$ Tenure          <int> 2, 1, 8, 1, 2, 8, 7, 4, 4, 2, 6, 3, 10, 5, 7, 3, 1, 9, 6, 6, 8, 8, 4, 3, 5, 3, 2, 9, 3, 0, 3, …
$ Balance         <dbl> 0.00, 83807.86, 159660.80, 0.00, 125510.82, 113755.78, 0.00, 115046.74, 142051.07, 134603.88, …
$ NumOfProducts   <int> 1, 1, 3, 2, 1, 2, 2, 4, 2, 1, 2, 2, 2, 2, 2, 2, 1, 2, 1, 2, 2, 2, 1, 2, 1, 2, 1, 2, 1, 2, 3, 1…
$ HasCrCard       <fct> 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 0, 1, 1, 0, 1, 0, 1, 1, 0, 1, 1, 1, 1, 0, 1, 0, 1, 0, 1, 1, 1, 0…
$ IsActiveMember  <fct> 1, 1, 0, 0, 1, 0, 1, 0, 1, 1, 0, 0, 0, 0, 1, 1, 0, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 0, 1, 1, 0, 1…
$ EstimatedSalary <dbl> 101348.88, 112542.58, 113931.57, 93826.63, 79084.10, 149756.71, 10062.80, 119346.88, 74940.50,…
$ Exited          <fct> 1, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0…
>
```

Figure 2: Removing Irrelevant Variables

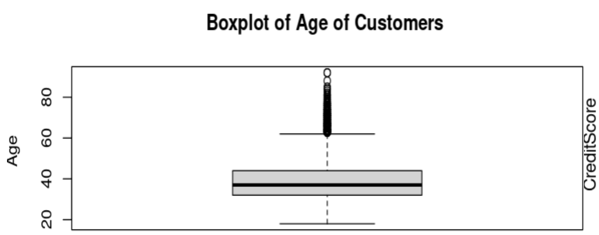
Summary Statistics

```
> summary(churn)
  CreditScore      Geography            Gender               Age            Tenure      
 Min.   :350.0   Length:10000       Length:10000       Min.   :18.00   Min.   : 0.000  
 1st Qu.:584.0   Class :character   Class :character   1st Qu.:32.00   1st Qu.: 3.000  
 Median :652.0   Mode  :character   Mode  :character   Median :37.00   Median : 5.000  
 Mean   :650.5                                         Mean   :38.92   Mean   : 5.013  
 3rd Qu.:718.0                                         3rd Qu.:44.00   3rd Qu.: 7.000  
 Max.   :850.0                                         Max.   :92.00   Max.   :10.000  
    Balance       NumOfProducts   HasCrCard IsActiveMember EstimatedSalary     Exited  
 Min.   :     0   Min.   :1.00   0:2945    0:4849         Min.   :    11.58   0:7963  
 1st Qu.:     0   1st Qu.:1.00   1:7055    1:5151         1st Qu.: 51002.11   1:2037  
 Median : 97199   Median :1.00                            Median :100193.91           
 Mean   : 76486   Mean   :1.53                            Mean   :100090.24           
 3rd Qu.:127644   3rd Qu.:2.00                            3rd Qu.:149388.25           
 Max.   :250898   Max.   :4.00                            Max.   :199992.48           
>
```

Figure3: Summary Statistics

Visualizing Numeric Data; Age, Tenure and Credit Score

Numeric data visualization was done using the box plot



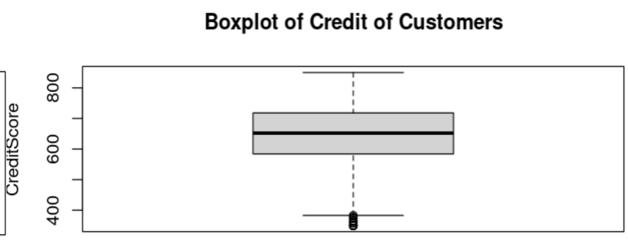


Figure 4: Boxplot of Age

Figure 5:Boxplot of Credit Score

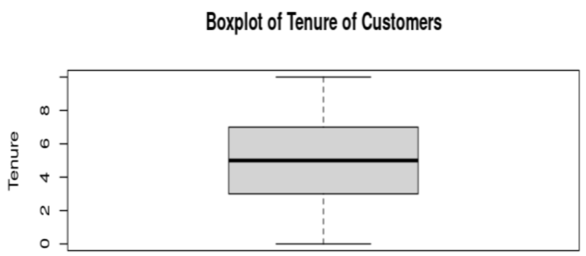
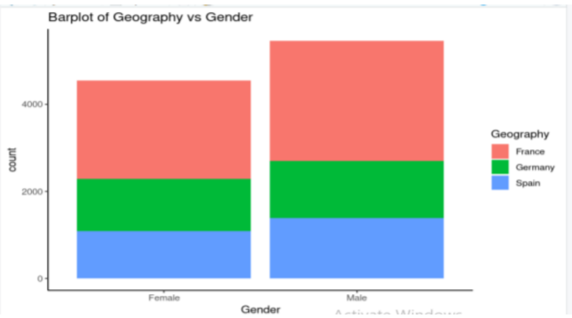


Figure 6: Boxplot of Tenure

The average age of customers is 38years, the average credit score of customers is 652 and the average tenure of customers is 5years.

Visualizing Categorical Variables; Geography, Gender



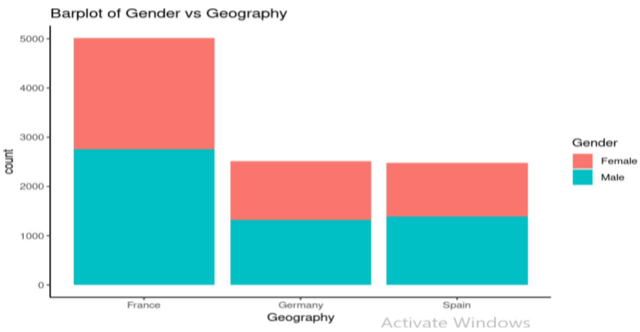
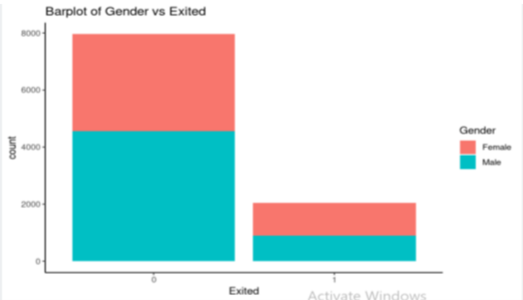


Figure 6: Geography Versus Gender *Figure 6: Boxplot of Gender Versus Geography*

In figure (6), There are more male customers than female customers in France.

In figure (7), There are more customers in France than Germany and Spain.



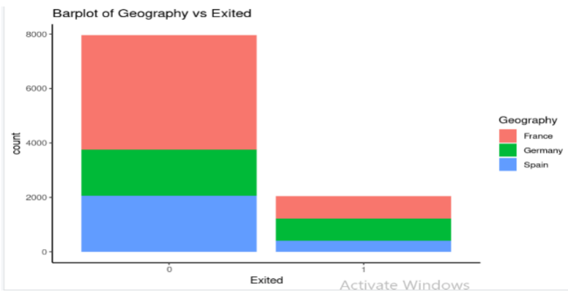
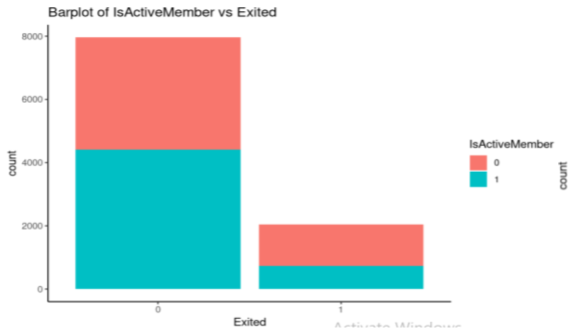


Figure 8: Exited Versus Gender *Figure 9: Exited versus Geography*

Females exited the bank more than males as shown in figure (8) and in figure (9) Spain customers exited more than Germany and France with France having the highest number of customers.



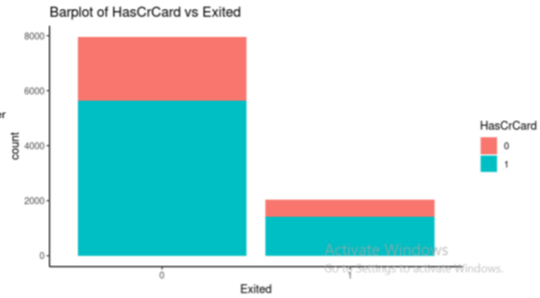
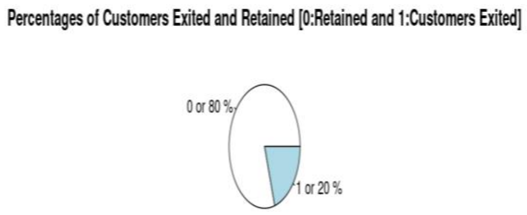


Figure 10: Active Member versus Exited *Figure 11: HasCrCard versus Exited*

From figure (10), there are more active customers in the bank than inactive customers and more inactive customers exited than active members. Also, in figure (11) Customers that have credit cards exited the bank more than those who do not have.



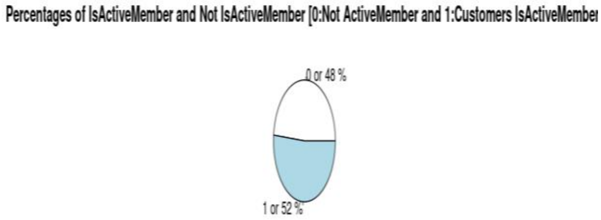


Figure12: Proportion of customers Exited *Figure13: Proportion of active/ inactive customers*

Figure (12) shows that 20% of customers exited and figure (13) shows 48% of customers are inactive while 52% are active.

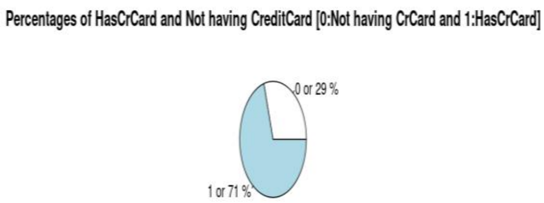
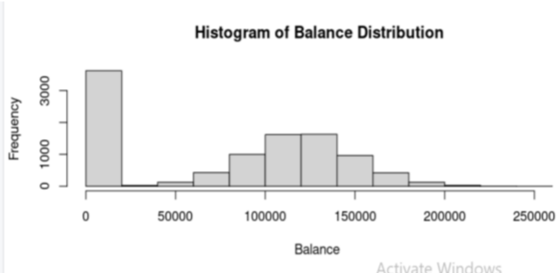


Figure13: Proportion of customers that have credit card

Figure 1 shows that 71% of customers have credit cards while 29% of customers do not have.

3.3.6 Visualizing continuous Variables



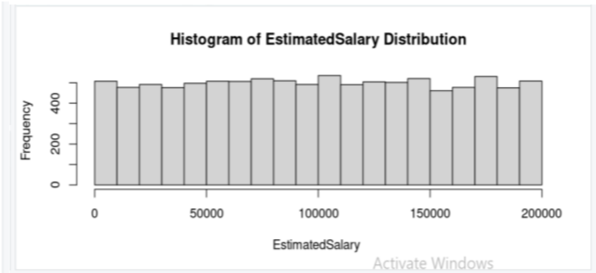
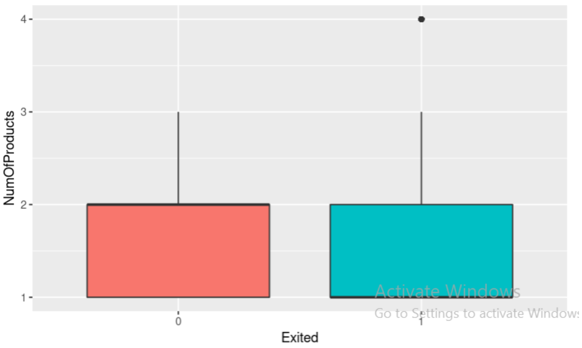


Figure16: Balance Distribution *Figure 17: Estimated Salary Distribution*

Figure (16) shows that a lot of customer has low balance and there is no negative balance, Figure 17) shows distribution of customers estimated salary.

Visualizing Numeric Variables and Exited Variable



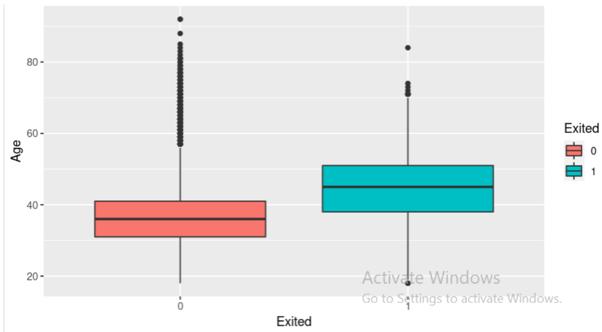
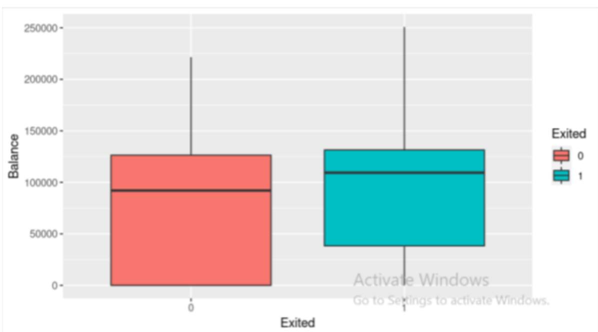


Figure18: Number of Products versus Exited *Figure 19: Age versus Exited*

The number of products has no effect on customer churn as shown in figure (18), older people exited compared to younger people.

Visualizing continuous Variable Versus Exited Variable



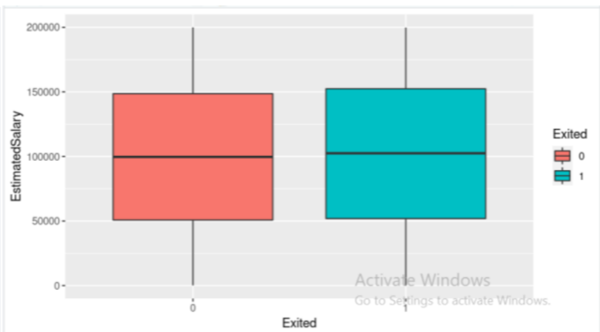


Figure20: Balance Distribution *Figure 21: Estimated Salary Distribution*

Customers with high balance churn more than customers with low balance as shown in figure (20) Estimated Salary has a low impact on customer churn.

Building Decision Tree Model

The decision tree was used to create a train and test model that was used to make predictions of customer churn (Exited) using the set of data given Figure (23) shows the train data.

Procedure in building the Decision Tree Model

-Set the sample using a set.seed (456), this is to enable the randomization process to follow a sequence.

-Carryout out random sampling and split the data into train and test samples.

```
>     # create a random sample for training and test data
>     RNGversion("3.5.2")
Warning message:
In RNGkind("Mersenne-Twister", "Inversion", "Rounding") :
  non-uniform 'Rounding' sampler used
>     set.seed(456)
>     train_sample <- sample(10000, 9000)
>     #see structure of train sample data
>     str(train_sample)
 int [1:9000] 896 2105 7329 8519 7881 3318 824 2854 2374 3849 ...
```

Figure22: Performing Random Sampling

```
>     # split the data frames
>     churn_train <- churn[train_sample, ] #training data
>     churn_test <- churn[-train_sample, ] #test data
>     # check the proportion of class variable
>     prop.table(summary(churn_train$Exited))
        0         1
0.7957778 0.2042222
>     prop.table(summary(churn_test$Exited))
    0     1
0.801 0.199
```

Figure23: Proportion of Train and Test Data

Trained the model

```
>     # display simple facts about the tree
>     churn_model

Call:
C5.0.default(x = churn_train[-11], y = churn_train$Exited, control = C5.0Control(minCases = 400))

Classification Tree
Number of samples: 9000
Number of predictors: 10

Tree size: 3

Non-standard options: attempt to group attributes, minimum number of cases: 400
```

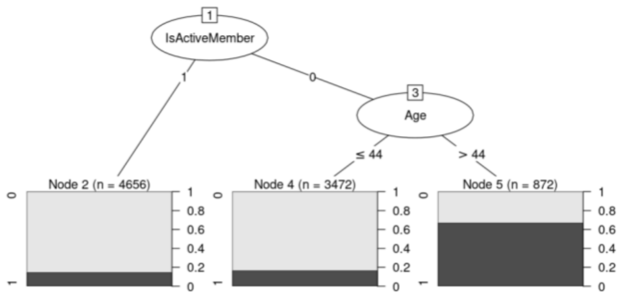
Figure24: Displaying facts about the Tree

```
Call:
C5.0.default(x = churn_train[-11], y = churn_train$Exited, control = C5.0Control(minCases = 400))


C5.0 [Release 2.07 GPL Edition]         Sun May  8 18:14:02 2022
-------------------------------

Class specified by attribute `outcome'

Read 9000 cases (11 attributes) from undefined.data

Decision tree:

IsActiveMember = 1: 0 (4656/674)
IsActiveMember = 0:
:...Age <= 44: 0 (3472/578)
    Age > 44: 1 (872/286)
```

```
Evaluation on training data (9000 cases):

	    Decision Tree
	  ----------------
	  Size      Errors

	     3 1538(17.1%)   <<


	   (a)   (b)    <-classified as
	  ----  ----
	  6876   286    (a): class 0
	  1252   586    (b): class 1


	Attribute usage:

	100.00% IsActiveMember
	 48.27% Age


Time: 0.0 secs

> |
```

Figure25: Details of the Tree /Evaluation of Train Model

Plot the decision tree



G

Figure26: Decision Tree Plot

```
   Cell Contents
|-------------------------|
|                       N |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  1000


             | predicted Exited
actual Exited |         0 |         1 | Row Total |
--------------|-----------|-----------|-----------|
            0 |       758 |        43 |       801 |
              |     0.758 |     0.043 |           |
--------------|-----------|-----------|-----------|
            1 |       112 |        87 |       199 |
              |     0.112 |     0.087 |           |
--------------|-----------|-----------|-----------|
 Column Total |       870 |       130 |      1000 |
--------------|-----------|-----------|-----------|
```

Figure 27: Evaluating Test Model

In the train sample of 9000 customers,10 predictors and tree size of 3 as shown in figure (25). The train model shows 17.1% error rate ,286 of customers who stayed were incorrectly predicted as exited and 1252 of customer who exited were incorrectly predicted as stayed.

The test sample of 1000 customers as shown in figure (27) gave an error rate of 15.5%, 43 customers who stayed were incorrectly predicted as exited and 112 of customers who exited were incorrectly predicted as stayed. The test sample had a higher accuracy than the train sample.

Interpretation of the decision tree model plot

Table 2: Interpretation of decision tree model plot as shown in figure (25)

Variables	Classification
IsActiveMember = 1: 0 (4656/674)	4656 customers classified as Exited,674 out of these customers were incorrectly classified as Exited
IsActiveMember = 0: ...Age <= 44: 0 (3472/578)	3472 customers classified as stayed,578 out of these customers were incorrectly classified as stayed
Age > 44: 1 (872/286)	872 customers classified as Exited,286 out of these customers were incorrectly classified as Exited

Recommendation

The rate of customer churn in this research is 20%, this can be reduced by identifying the variables that causes customer attrition and identifying customers that are likely to churn. Retaining existing customers is cheaper than obtaining new ones. In maintaining customers, the first stage is to keep an eye on the churn as this evaluates how good you are at keeping customers. A product that suits the female gender should be developed such that it creates awareness of breast cancer and gives an opportunity for female entrepreneurs to network, this will reduce the rate at which female customers exit the bank. More awareness through the use of advertisement and product development to suit Germany and Spain customers because these areas with few customers have higher churn rate than France which has the highest customers. Products such as the Internet and mobile banking, prompt response services with the use of chatbots in resolving customers' complaints on a 24/7 basis, encouraging the use of these products, and promoting self-service and personalized banking to reduce churn rate in Germany and Spain. Strategies should be put in place to reactivate inactive accounts. Review interest rates on credit cards and compare them with competitors in the banking industry, and maintain customer loyalty incentives on these cards to discourage credit card customers from exiting the bank. The Bank should introduce products that suit the older generation, for example liaising with health care services providers to provide special care services for older customers, this will reduce the rate of churn for older customers

Introduction of products that suit high net worth customers for example, VIP services with the use of ATM cards in lounges, will decrease churn in the number of customers with high account balances. Conduct market research and request customer feedback on all services and use data obtained from the feedback to formulate competitive products to increase its market share as well as reduce the customer churn rate.

The training model predicted accurately as its outcome aligns with the data visualization outcome that is more active customers stayed than inactive customers as shown in figure (10) and (26) and customers older than 44 years old exited while those less than 44 years old stayed as shown in figure (19) and (26).

Conclusion

AI/DS techniques were efficient in identifying areas of improvement of the banks products and services, making the right decisions, improving customer retention and satisfaction, reducing customer churn, and increasing profit. Further research should be carried out using with the given data using the Random Forest model which is a more advanced machine learning algorithm and results obtained compared with same as obtained in this research.