

Modeling Churning of Customers in Bank using Supervised and Unsupervised Machine Learning

Contents

Modeling Churning of Customers in Bank using Supervised and Unsupervised Machine Learning	1
Abstract.....	2
<i>Gordon Cameron, Parul Chaudhary, Yohann Gazdar, , Paola Merrick, Melkamu Gishu.</i>	2
Introduction	3
Hypothesis:	4
Data description.....	4
Data Preprocessing	5
Feature Engineering.....	6
Analysing Exited Feature.....	6
Outlier detection and transformation	7
Feature Selection	7
Multicollinearity.....	8
Data Visualization	8
Modelling the Churn of the Bank Customers.....	10
Logistic Regression and Discriminant Analysis.....	10
Modelling target customers.....	13
Decision Tree.....	16
Model Evaluation	19
Non-Supervised - K-Mean clustering	20
Conclusions	25
References	26

Abstract

Gordon Cameron, Parul Chaudhary, Yohann Gazdar, , Paola Merrick, Melkamu Gishu.

Customer churn has become an increasing challenge for the banking industry, where available technology has multiplied the banking options easily accessible to all customers. This project will focus on identifying the characteristics of banking customers who are more likely to churn, i.e. close their accounts, for a particular bank. To find out the likely behavior or a combination of factors that potentially leads to churn, we applied both supervised and unsupervised algorithms to model the customer churning process and used the following methods of data analytics: Logistic Regression, Decision Tree, Random Forest, K-means and Clustering. By identifying the characteristics that can better explain customer churns, banks should be able to increase their Customer Lifetime Value and be more profitable. Our analysis indicates that age, gender, location, balance and number of products impact the likelihood of churn and make recommendations for each one.

Keywords: Churn, Clustering, Regression, Random Forest, Decision Trees

Introduction

Customers are now surrounded by a number of resources of information in today's digitized environment, and it's all available at the tip of their fingers. Smartphones, e.g., provide instant access to various products and services, mobile banking is one example. Customer perspectives are based on advanced technology, convenience, price sensitivity, service quality and social factors[1]. Due to the availability of different options as a consequence of advanced technologies, customers find it much easier changing from one service provider to another. That is why, for most companies, it is difficult to retain customers. The procedure of customers leaving their service providers is called churn[2].

For the banking sector, customer churn prediction has become a key part of their operations strategy and can have an enormous impact on the bottom line [3]. Thus, customer retention schemes can be targeted on high-risk customers who wish to discontinue their services and switch to another competitor. To minimize the cost of banks' marketing investment in customer retention schemes, an accurate and prior identification of these customers is hypercritical[4].

Customer churn is an estimate of the ratio of customers who leave their current bank by closing all accounts and services provided . It is one of the most common problems witnessed in various industries[5, 6]. Banking is one industry that focuses a lot on customer's behavior by tracking their level of engagement with the bank. Several studies have shown that it is more expensive to add a new customer than to retain it [7]. Companies can raise their profits by concentrating on these customers and avoiding churn. Hence, there is a need to retain the existing customers, which will be achieved only by understanding the customer behaviour, with the help of data analysis.

It is clear that customer retention is important for a company and for its business strategy. To reduce customer churn, companies started adapting machine learning techniques for prediction models. Among the available literature, Data Mining by author 'Sen K' aims to analyze large datasets by converting big data into useful insights. A customer churn prediction model is developed and is measured, using accuracy, sensitivity and specificity and Kappa's statistics[8]. The support vector machine (SVM) is a popular technique which provides a guide to the bank for customer strategy. SVM has a larger probability of customer churning in the samples. With a good number of vectors, SVM provides good precision in predicting technique models. SVM gives a high fitting accuracy rate of 0.59 by the author Zhao Jing[9]. Guoxun Wang' focuses on the comparison of all techniques used to build credit card holder churn models for the banks in China based on multi-criteria decision algorithms and constructing techniques using PROMETHEE and TOPSIS methods[10]. In the MCDM algorithm, decision tree methods are implemented. Shaoying Cui presents an improved FCM algorithm to facilitate the banks with a new idea for predicting customer churn[11]. It achieved an accuracy rate of 80% for high-value customers and 83% for low-value customers. Pradeep B proposed to construct a model for churn prediction for a company using logistic regression and decision trees techniques. In Pradeep's approach there is a trial to retrieve the important factors of the customer churn that provides additional and useful

knowledge which supports decision making[12]. Alisa Bilal Zorić applied a data mining technique neural network in the software package ANN to predict churn in bank customers. Using this model, the reason for customers leaving the bank can be easily predicted by entering the parameters[13]. Abinash Mishra proposed methodology of ensemble classifiers comprising bagging, boosting and random forest to predict customer churn for the telecom industry. Random forest achieves high accuracy of 96% with low specificity and high sensitivity and low error rate[14]. Ning Lu presented a paper in which an experimental evaluation proves that the boosting provides a good source of churn data, efficiently providing the customer churn model. The measures for churn prediction are calculated using a training set of customers over a period of six months[15]. Hend Sayed presented a methodology of decision tree in which two packages ML and MLib were conducted, to evaluate accuracy, model training and model evaluation. They got effective results with the ML package[16].

The analysis in this project will focus on the behavior of banking customers who are more likely to close their accounts. To find out the likely behavior or a combination of factors that potentially leads to churn we start with explanatory analysis. We also use predictive, classification and cluster models to determine the customers who are most likely to churn. The business objective is to help the banking industry with insights on customer behavior that would be helpful in developing targeted strategies for retention of customers. The overall goal will be to understand what correlations can be drawn between customers' information and their decision to leave the bank, and identify customers who have the highest likelihood of leaving. This goal will help the bank formulate effective retention strategies.

Hypothesis:

By gaining insights into which customers are more likely to churn, the bank can leverage this information to develop new and improved marketing, pricing, and/or service strategies to retain customers. The impact of lowering customer churn may result in an increase in profits / revenue in the long run.

Data description

Dataset used for this supervised prediction is acquired from an online source. The target dataset contains information on customers who have an existing relationship with the bank as well as those who have left. There are 10,000 rows of customer data with 14 features for each customer. The customers of the bank are identified as churn or loyal based on the potential features like credit score, age, gender, estimated salary, etc. Customers are classified as churn or “exited” if they left the bank by closing an account. The variable exited in the dataset gives the actual status of the customer whether stayed or left the bank.

- Customer ID: This attribute is unique and assume that primary key
- Surname: it belongs to surname of customer and string values
- Geography: it shows country of customer

- Gender: male/female
- Credit Score: it gives credit score of customers. That score calculates the interbank system. High score shows that the customer debt has a high repayment capacity.
- Age: age of customers
- Tenure: The number of ages the customer is in the bank.
- Balance: Customer's money in the bank.
- Number of Products: Number of products owned by the customer.
- Credit Card: Whether the customer has a credit card
- Active Status: Customer's presence in the bank
- Estimated Salary: Customer's estimated salary
- Exited: Churn or not

Data Preprocessing

Feature scaling or data normalization technique used to standardize the range of independent variables in the dataset.

To perform the analysis, certain R libraries were used. The code below was used to load and initialize the library.

```
**Perm link**
set.seed(12)
library(dplyr) # data manipulation
library(corrplot)
library(ranger) # required by Boruta
library(Boruta) # feature engineering selection
library(DescTools) # to use Winsorize
library(ggplot2) # basic and advanced graphs
library(reshape)

BankChurn<-read.csv("stable link for the dataset")

str(BankChurn)

'data.frame':      10000 obs. of  14 variables:
 $ RowNumber   : int  1 2 3 4 5 6 7 8 9 10 ...
 $ CustomerId  : int  15634602 15647311 15619304 15701354 15737888...
 $ Surname     : Factor w/ 2932 levels "Abazu","Abbie",..: 1116 1178 ...
 $ CreditScore : int  619 608 502 699 850 645 822 376 501 684 ...
 $ Geography   : Factor w/ 3 levels "France","Germany",..: 1 3 1 1 3 3...
 $ Gender      : Factor w/ 2 levels "Female","Male": 1 1 1 1 2 1 2 2 ...
 $ Age         : int  42 41 42 39 43 44 50 29 44 27 ...
```

```
$ Tenure      : int 2 1 8 1 2 8 7 4 4 2 ...
$ Balance     : num 0 83808 159661 0 125511 ...
$ NumOfProducts : int 1 1 3 2 1 2 2 4 2 1 ...
$ HasCrCard   : int 1 0 1 0 1 1 1 0 1 ...
$ IsActiveMember : int 1 1 0 0 1 0 1 0 1 ...
$ EstimatedSalary: num 101349 112543 113932 93827 79084 ...
$ Exited      : int 1 0 1 0 0 1 0 1 0 0 ...
```

Feature Engineering

It is desirable to reduce the number of input variables to both reduce the computational cost of modeling and, to improve the performance of the model. Statistical-based feature selection methods involve evaluating the relationship between each input variable and the target variable using statistics and selecting those input variables that have the strongest relationship with the target variable. These methods can be fast and effective, although the choice of statistical measures depends on the data type of both the input and output variables. As such, it can be challenging to choose an appropriate statistical measure for a dataset when performing feature selection.

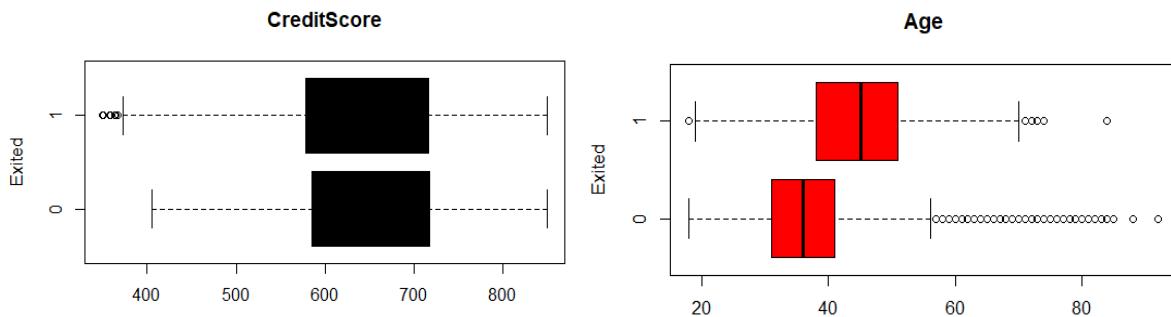
In our analysis we initially used filter methods, feature selection methods which use statistical measures to score the correlation or dependence between each input and output variables independently (univariate and bivariate) to aid us filter the most relevant variables. We then applied wrapper methods of feature selection to evaluate multiple models using procedures that add and/or remove predictors to find the optimal combination that maximizes model performance. Based on the type of dataset, method of analysis we are going to do and with consultation with the existing literature, we found the Boruta algorithm which is built around the random forest classification algorithm to be efficient for our final screening. The algorithm tries to capture all the important, interesting features in the dataset with respect to an outcome variable.

Analysing Exited Feature

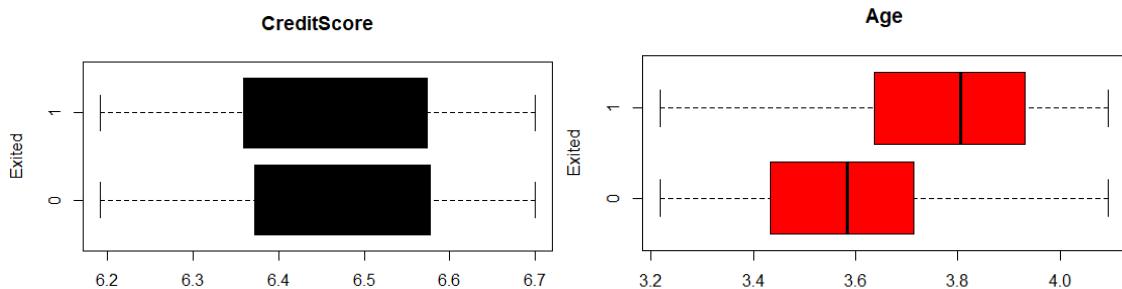
The pie chart below shows us that the data set has more records of customers who have remained loyal to the bank as opposed those that have left the bank.



Outlier detection and transformation



For unsupervised algorithms, since the objective is mainly classification, we transformed all variables including those who don't show severe outliers.

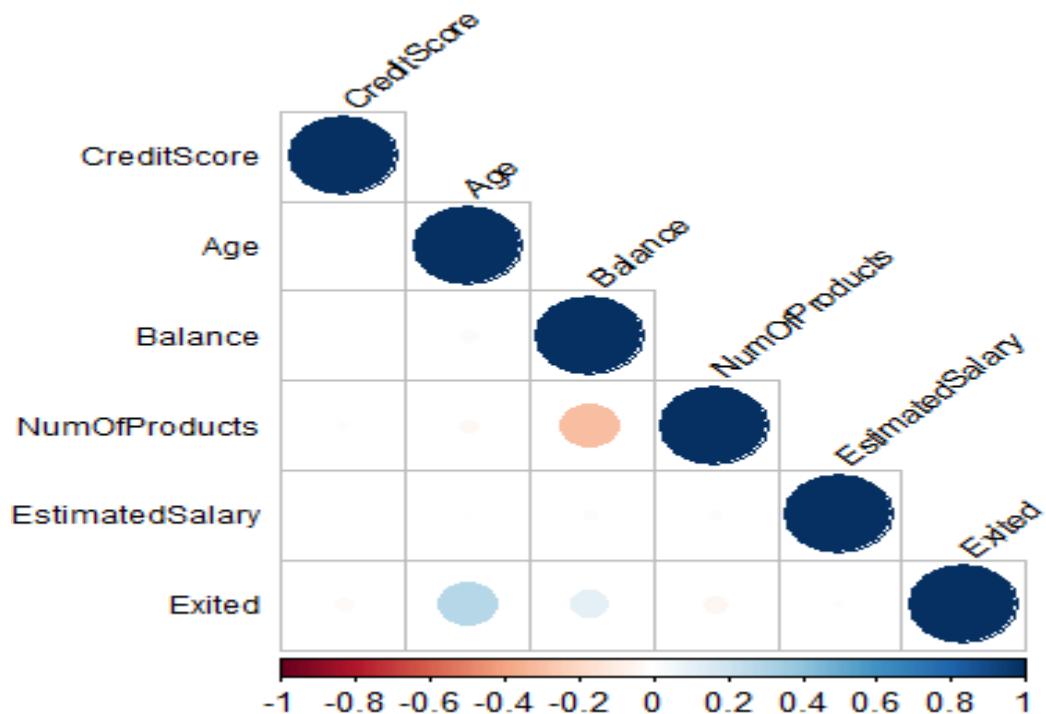


Feature Selection

For the final feature selection purposes, Boruta model was fitted. The model identified the significant variables which were used to further subset the original dataset. After 38 iterations, and +13.5 mins of computational time: "CreditScore", "Geography", "Gender", "Age", "Balance", "NumOfProducts" and "EstimatedSalary" found to be relevant candidates for the supervised machine learning analysis.

Multicollinearity

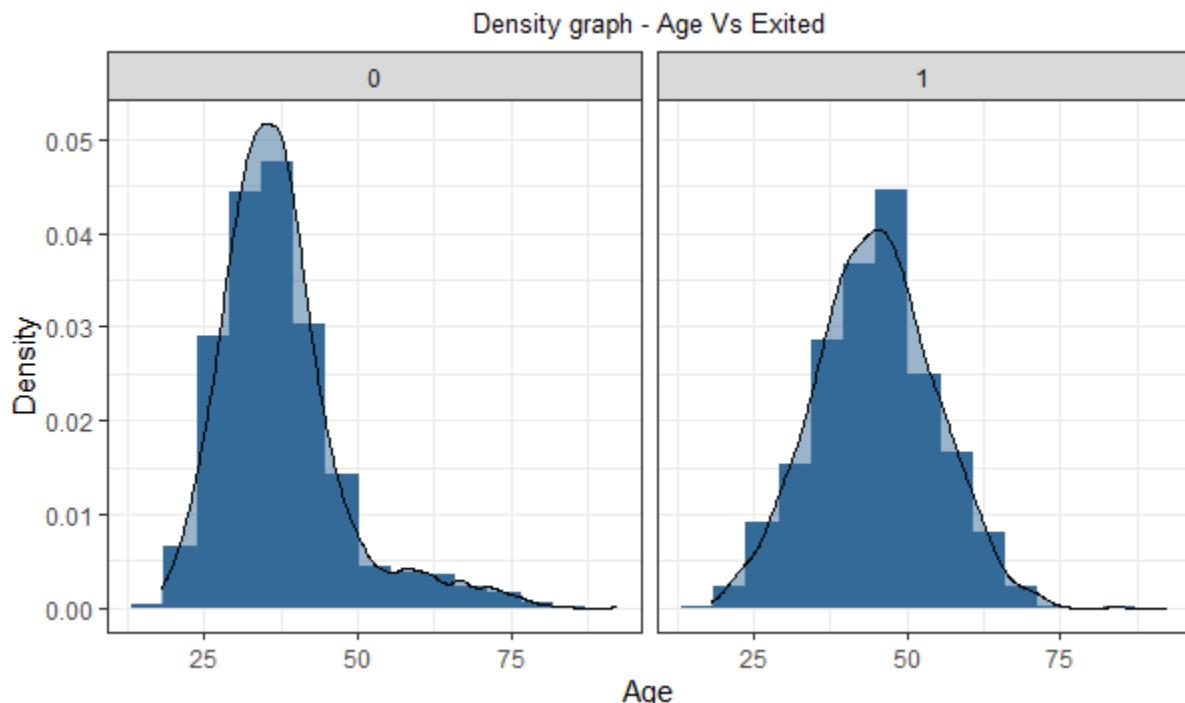
A CorroloPlot was used to highlight the correlated numeric variables which will potentially result in multicollinearity



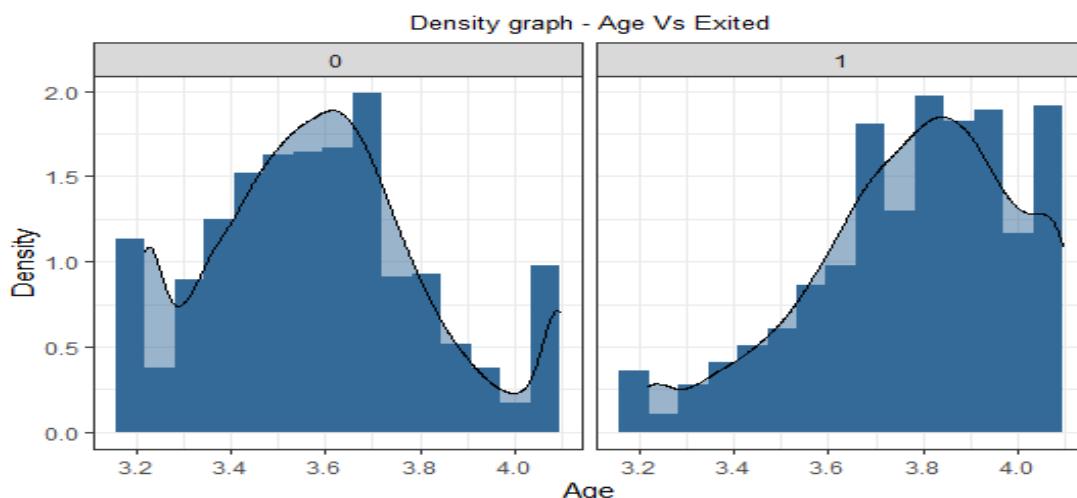
As it is shown in the graph no further reduction is needed as the correlation between the independent variables within acceptable range.

Data Visualization

Density plot shows some interesting relationships between Age vs. the Exited features.



```
s2=aggregate(BankChurn_sig$Age, by=list(BankChurn_sig$Exited), FUN = summary)
> print(s2)
Group.1 x.Min. x.1st Qu. x.Median x.Mean x.3rd Qu. x.Max.
1     0 18.00000 31.00000 36.00000 37.40839 41.00000 92.00000
2     1 18.00000 38.00000 45.00000 44.83800 51.00000 84.00000
```



```
s2=aggregate(BankChurn_sig2$Age, by=list(BankChurn_sig2$Exited), FUN = summary)
> print(s2)
Group.1 x.Min. x.1st Qu. x.Median x.Mean x.3rd Qu. x.Max.
1     0 3.218876 3.433987 3.583519 3.590174 3.713572 4.094345
2     1 3.218876 3.637586 3.806662 3.776004 3.931826 4.094345
```

Different trends for customers who leave and those who do not are shown in the transformed range 3.6 and 3.82. For those within that age group that may show us the presence of factor/s affecting the decision to stay or will leave.

Modelling the Churn of the Bank Customers

To predict the churn of customers, the dataset is split for training and testing. At this instance, splitting has 75% training rate and 25% testing rate. The value of the attribute ‘Exited’ will be 1 if the customer has left the bank and 0 if remained there. The dataset with the screened variables (significant variables to be included in the supervised models) has been split in such a way that train and test sets would have the same distribution of the EXITED attribute. The data does not have any missing values so we don’t need any imputations and the dataset is ready for the next analysis at this point.

```
spliter=createDataPartition(BankChurn_sig$Exited,times = 1,p=0.75,list = FALSE)
Train=BankChurn_sig[spliter,]
Test=BankChurn_sig[-spliter,]
prop.table(table(Train$Exited))
```

In this project, with focus in finding the best overall accuracy we applied both supervised and unsupervised algorithms to model the customer churning process. Logistic Regression, Decision Tree, Random Forest, K-means and Clustering.

Logistic Regression and Discriminant Analysis

Logistic regression model belongs to the class of generalized linear models. Logistic regression modeling is a commonly used strategy in analyzing data which have categorical dependent variables and the explanatory variables can be continuous and/or categorical. It can be classified as binary, ordinal and multinomial logistic regression based on the category of the dependent variable. In this project, binary logistic regression technique is used. Detailed discussion on logistic regression models can be found on Hosmer and Lemeshow (2000)[18]. Discrimination techniques concerned with separating distinct sets of objects or observations and allocating new objects (observations) to previously defined groups. There are two goals of discrimination, the first one is to describe graphically or algebraically the differential features of objects(observations) from several known collections (populations).The second goal is to sort objects (observations) into two or more labeled classes ,that is ,deriving a rule that can be used to optimally assign new objects to the labeled classes. Details are available in Johnson and Wichern (2007)[19].

```
LogReg2=glm(Exited~, data = Train, family = "binomial")
```

```

Coefficients: Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.412e+00 2.663e-01 -12.816 < 2e-16 ***
CreditScore -7.676e-04 3.163e-04 -2.427 0.015233 *
GeographyGermany 7.928e-01 7.688e-02 10.312 < 2e-16 ***
GeographySpain 6.048e-02 7.925e-02 0.763 0.445396
GenderMale -5.046e-01 6.149e-02 -8.207 2.26e-16 ***
Age 6.234e-02 2.777e-03 22.446 < 2e-16 ***
Balance 2.040e-06 5.829e-07 3.499 0.000467 ***
NumOfProducts -1.278e-01 5.419e-02 -2.359 0.018339 *
EstimatedSalary 2.926e-07 5.344e-07 0.548 0.583961
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Null deviance: 7552.9 on 7499 degrees of freedom
Residual deviance: 6710.6 on 7491 degrees of freedom
AIC: 6728.6

Once again, we rechecked multicollinearity from the model fit.

```

car::vif(LogReg2)
GVIF Df GVIF^(1/(2*Df))
CreditScore 1.000496 1 1.000248
Geography 1.225031 2 1.052051
Gender 1.002164 1 1.001082
Age 1.006103 1 1.003047
Balance 1.306089 1 1.142842
NumOfProducts 1.084458 1 1.041373
EstimatedSalary 1.001836 1 1.000918

```

The variance inflation factor for each of the variables is way below the recommended threshold value, and we are good to go to manipulate the variables.

Maximum likelihood - which tries to find the value of coefficients (β_0, β_1, \dots) such that the predicted probabilities are as close as possible to the observed probabilities - is used to determine the best coefficients and eventually a good model fit.

From the output of the logistic regression model, we can see that all the coefficients are significant and for ease of interpretation we can exponentiate them as follows:

```
data.frame(LogOfOdds=round(exp(coef(LogReg2)),3))
```

	Variable	LogOfOdds
1	(Intercept)	0.033
2	CreditScore	0.999
3	GeographyGermany	2.210
4	GeographySpain	1.062
5	GenderMale	0.604
6	Age	1.064

```

7 Balance 1.001
8 NumOfProducts 0.880
9 EstimatedSalary 1.000

```

Interpretation for continuous variables: Holding the other variables constant, for one unit increase in age will increase the chance of churn by 6%.

Interpretation for dichotomous variables: Holding the other variables constant, the odd of churn for males is 60% less likely compared to females (odd ratio less than one). Holding the other variables constant, the odds of churn being in Germany is 121% higher than the other geographic regions.

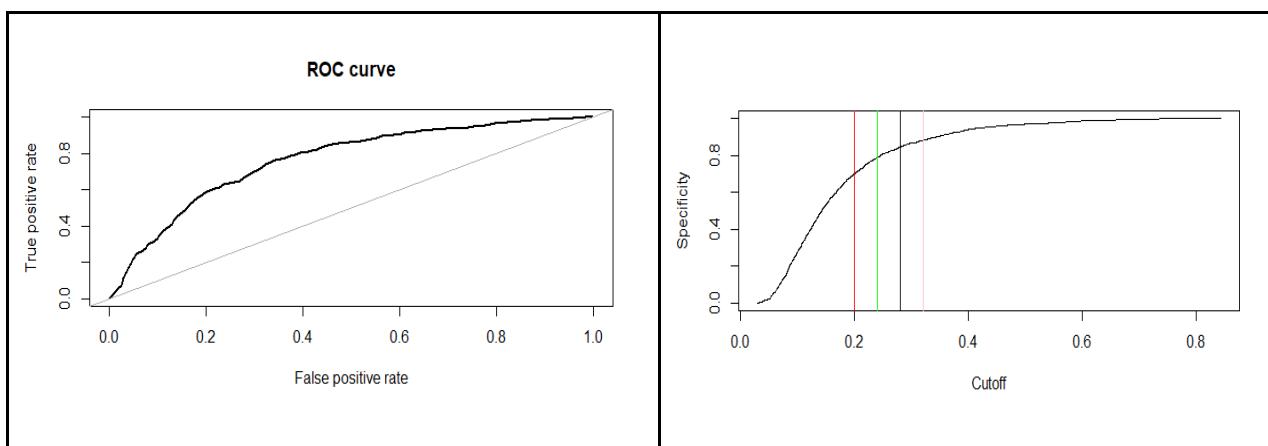
The regression coefficients explain the change in log(odds) in the response for a unit change in predictor. However, since the relationship here is not straight line, unlike linear regression, a unit change in input feature doesn't really affect the model output directly but it affects the odds ratio.

We also pulled of a list of variables with more than 50% probability of changing the decision of the customer for every 1 unit change in the respective independent variable by using the code below:

```
LOR %>% arrange(desc(LogOfOdds)) %>% filter(LogOfOdds >= 1) %>% mutate(Probability = round(LogOfOdds / (1 + LogOfOdds), 3))

Variable LogOfOdds Probability
1 GeographyGermany 2.210 0.688
2 Age 1.064 0.516
3 GeographySpain 1.062 0.515
4 Balance 1.000 0.500
5 EstimatedSalary 1.000 0.500
```

To find the best cutoff point, plotted the ROC graph and simulate different cutoff points based on the original fit value.



```
cut_offs %>% filter(fpr<=0.4, tpr>=0.6)
```

cut	fpr	tpr
-----	-----	-----

0.1683683	0.3979798	0.8057692
-----------	-----------	-----------

cutoff of 16 seems to give the highest tpr and relatively low fpr.

We will fit a model for the cutting points 16, 18, 22, 24, 32

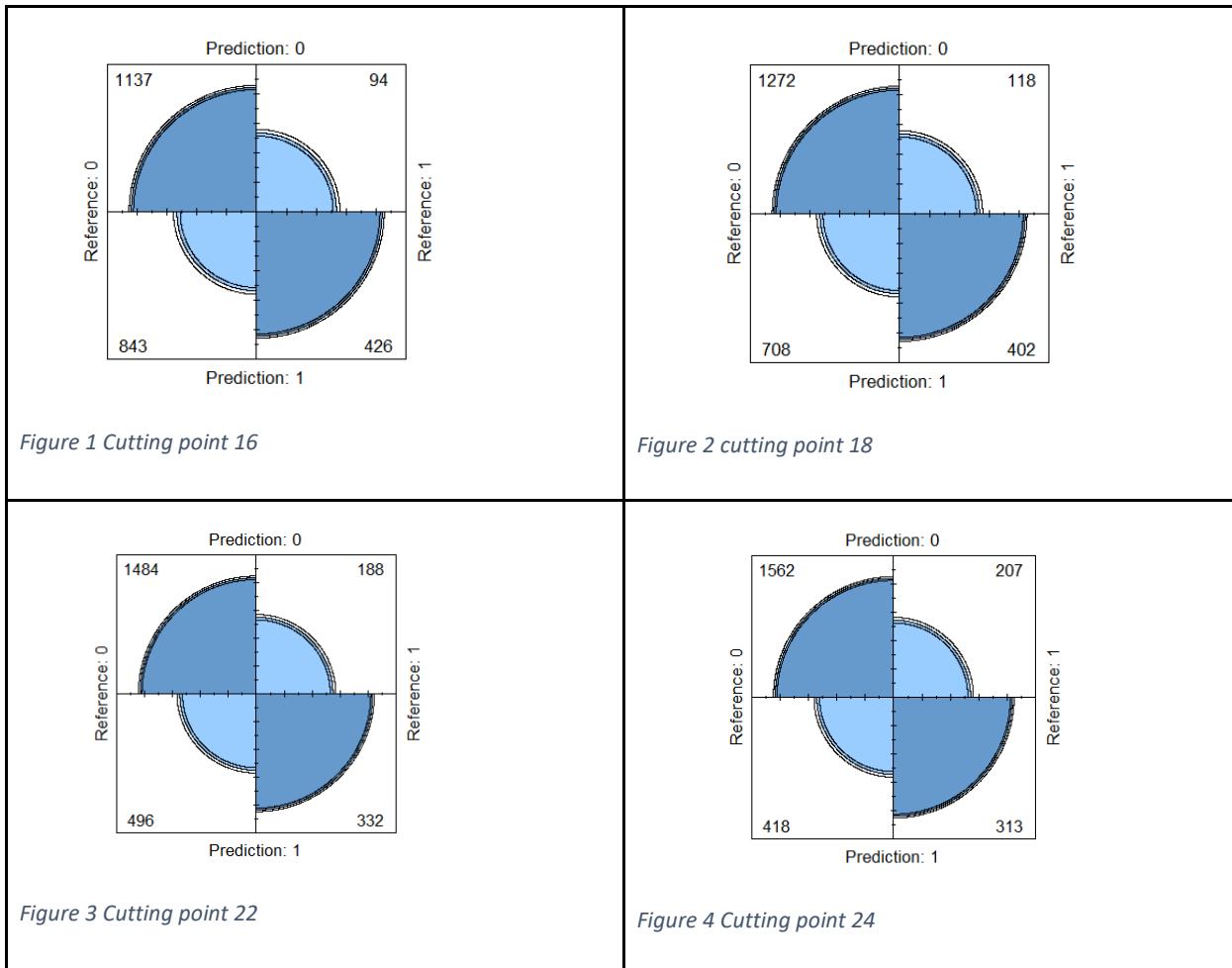


Figure 1 Cutting point 16

Figure 2 cutting point 18

Figure 3 Cutting point 22

Figure 4 Cutting point 24

Modelling target customers

We used a model for prioritization of customers for a proactive retention campaign

Gains – Lift Chart

```
gains(as.numeric(Target_customers$Exited), predict(Logtoselect, type="response",
, newdata=Target_customers), groups = 10)
```

Depth of File	N	Cume N	Mean Resp	Cume Mean Resp	Cume of Total Resp	Lift Index	Cume Lift	Mean Model Score
10	1000	1000	0.52	0.52	25.4%	254	254	0.54
20	1000	2000	0.44	0.48	46.8%	214	234	0.37
30	1000	3000	0.28	0.41	60.8%	140	203	0.29
40	1000	4000	0.23	0.37	72.1%	112	180	0.22
50	1000	5000	0.15	0.32	79.2%	72	158	0.18
60	1000	6000	0.14	0.29	86.4%	71	144	0.14
70	1000	7000	0.11	0.27	91.9%	55	131	0.11
80	1000	8000	0.07	0.24	95.5%	37	119	0.09
90	1000	9000	0.05	0.22	98.0%	25	109	0.06
100	1000	10000	0.04	0.20	100.0%	20	100	0.04

Theoretical Advertising Costs

<ul style="list-style-type: none"> Based on the gains/lift table in the previous slide, the bank can use this data to help optimize their advertising dollars 	Percentile of Population	% of Clients who Churn	Advertising Cost
	10	25.4%	\$1000
	20	46.8%	\$2000
<ul style="list-style-type: none"> Let's take a theoretical advertising budget of \$1000 per 1000 views/impressions. 	30	60.8%	\$3000
	40	72.1%	\$4000
	50	79.2%	\$5000
	60	86.4%	\$6000
	70	91.9%	\$7000
<ul style="list-style-type: none"> This model results in capturing over 60% of the target population with only \$3,000 advertising dollars spent. In a 	80	95.5%	\$8000

totally random advertising campaign, the same \$3,000 would target less than 1/3rd of the target population.	90	98.0%	\$9000
	100	100%	\$10000

The 20% customers who need to be proactively worked with to prevent churn were identified with.

```
quantile(Target_customers$prob,prob=c(0.10,0.20,0.30,0.40,0.50,0.60,0.70,0.80,0.90,1))
```

Quartile	Probability
10%	0.04971721
20%	0.07370845
30%	0.09775277
40%	0.12517418
50%	0.15645218
60%	0.19948253
70%	0.25341981
80%	0.32575619
90%	0.43197534

100%	0.70000452
------	------------

The customers whose probability of churn is greater than 32.57% and less than 70%. Generated using the code given below .

```
set.seed(121212)
Index=createDataPartition(Target_customers$Exited,times = 1,p=0.75,list = FALSE)
Train_target=Target_customers[Index,]
Test_target=Target_customers[-Index,]

Logtoselect=glm(formula = Exited ~. , family = "binomial",
                 data = Target_customers)

gains(as.numeric(Target_customers$Exited),predict(Logtoselect,type="response",
                                                 newdata=Target_customers),groups = 10)
Target_customers$Cust_ID<-original_data$CustomerId
Target_customers$prob<-predict(Logtoselect,type="response",newdata=Target_customers)
quantile(Target_customers$prob,prob=c(0.10,0.20,0.30,0.40,0.50,0.60,0.70,0.80,0.90,1))
targeted=Target_customers%>%filter(prob>0.32575619 & prob<=0.70000452)%>%dplyr::select(cust_ID)
head(targeted)
write.csv(targeted,"targetedCustomers.csv")
```

The above code helped us identify 1999 customers, who need to be pursued for further sales opportunities.

Decision Tree

A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules. In decision analysis, a decision tree and the closely related influence diagram are used as a visual and analytical decision support tool, where the expected values (or expected utility) of competing alternatives are calculated[20].

In this project Id3 and J4.8 decision tree algorithm with PART, is used. Id3 (basic divide-and conquer decision tree algorithm), J4.8 (implementation of C4.5 decision tree learner), PART (obtain rules from partial decision trees built using J4.8), J[1].

```
LogReg2DT=rpart(Exited ~., data = Train, method="class")
> LogReg2DT
n= 7500

node), split, n, loss, yval, (yprob)
  * denotes terminal node

1) root 7500 1517 0 (0.79773333 0.20226667)
  2) Age< 42.5 5317 614 0 (0.88452135 0.11547865)
    4) NumOfProducts< 2.5 5214 535 0 (0.89739164 0.10260836) *
      5) NumOfProducts>=2.5 103 24 1 (0.23300971 0.76699029) *
```

```

3) Age>=42.5 2183 903 0 (0.58634906 0.41365094)
6) NumOfProducts< 2.5 2059 783 0 (0.61971831 0.38028169)
12) NumOfProducts>=1.5 803 154 0 (0.80821918 0.19178082) *
13) NumOfProducts< 1.5 1256 627 1 (0.49920382 0.50079618)
26) Geography=France,Spain 877 367 0 (0.58152794 0.41847206)
52) Balance>=46504.34 612 204 0 (0.66666667 0.33333333) *
53) Balance< 46504.34 265 102 1 (0.38490566 0.61509434) *
27) Geography=Germany 379 117 1 (0.30870712 0.69129288) *
7) NumOfProducts>=2.5 124 4 1 (0.03225806 0.96774194) *
confusionMatrix(LogReg2DT.predicted , as.factor(Test$Exited))

```

Confusion Matrix and Statistics

	Reference	
Prediction	0	1
0	1897	295
1	83	225

Accuracy : 0.8488

95% CI : (0.8341, 0.8626)

No Information Rate : 0.792

P-Value [Acc > NIR] : 0.0000000000002297

Kappa : 0.4599

Mcnemar's Test P-Value : < 0.00000000000000022

Sensitivity : 0.9581

Specificity : 0.4327

Pos Pred Value : 0.8654

Neg Pred Value : 0.7305

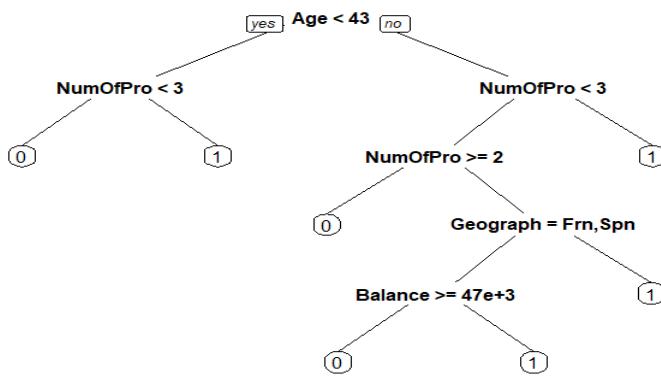
Prevalence : 0.7920

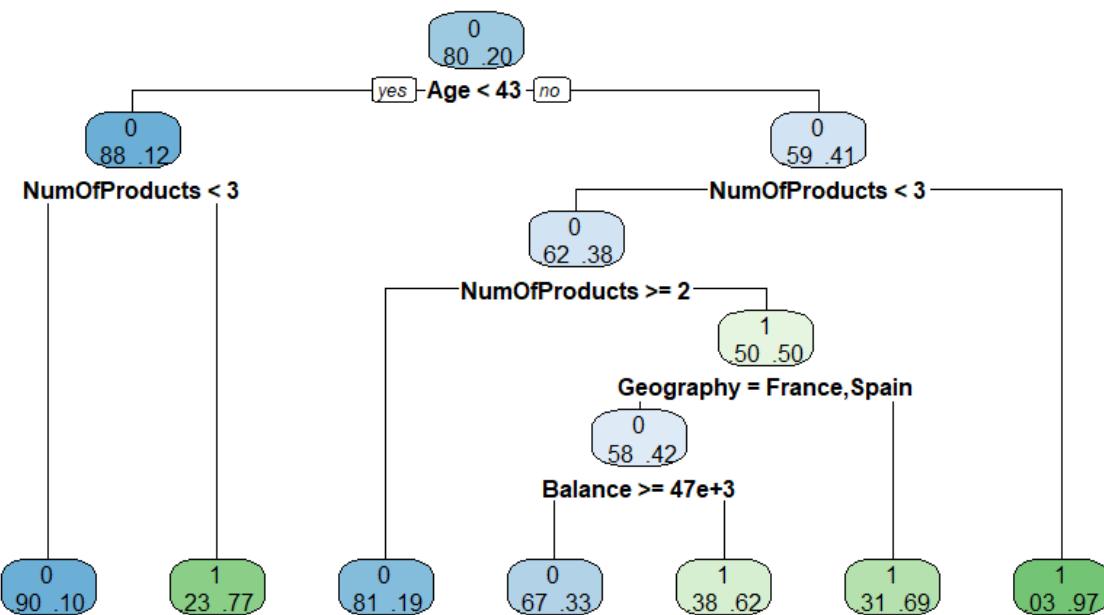
Detection Rate : 0.7588

Detection Prevalence : 0.8768

Balanced Accuracy : 0.6954

'Positive' Class : 0





Random Forest

A random forest can improve upon a decision tree by using multiple trees, and averaging the results to make predictions.

The target variable has only about 20% of the Exited customers and hence random forest model gets negatively affected by the imbalance. SMOTE technique is used to have more balanced examples of the less represented level of the target variable.

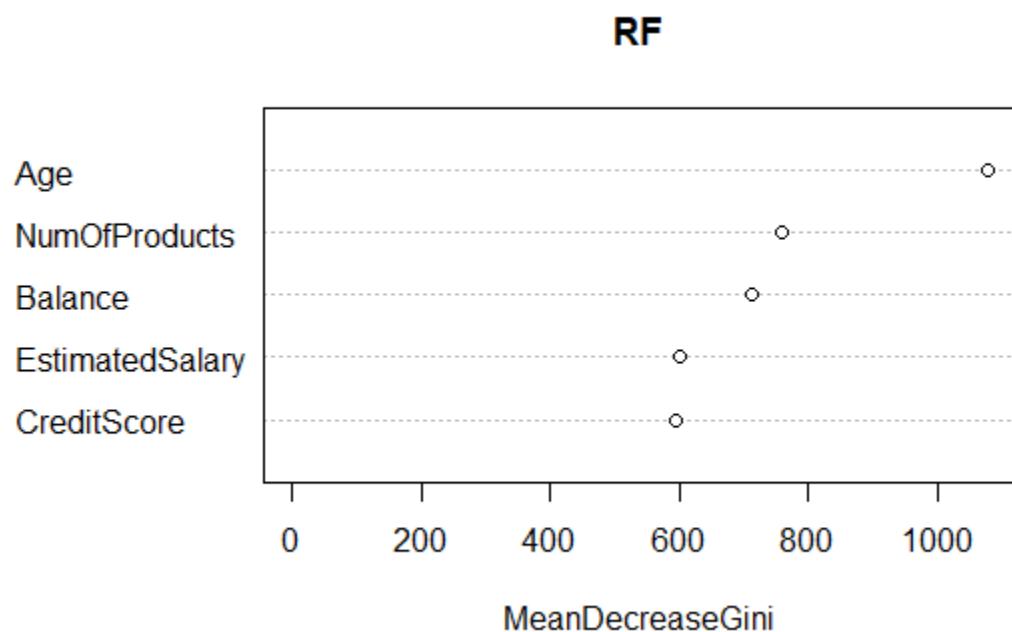
```

set.seed(12121212)
d.Rose=ROSE(Exited~, data = Winsorized, seed = 1)$data
prop.table(table(d.Rose$Exited))
dim(Winsorized)
Index=createDataPartition(d.Rose$Exited, p=0.75, list = FALSE, times = 1)
Train.R=d.Rose[Index,]
Test.R=d.Rose[-Index,]
RF=randomForest(as.factor(Exited)~, data = Train.R)
RF1=randomForest(as.factor(Exited)~, data = Train.R)
  
```

```
randomForest::varImpPlot(RF)
```

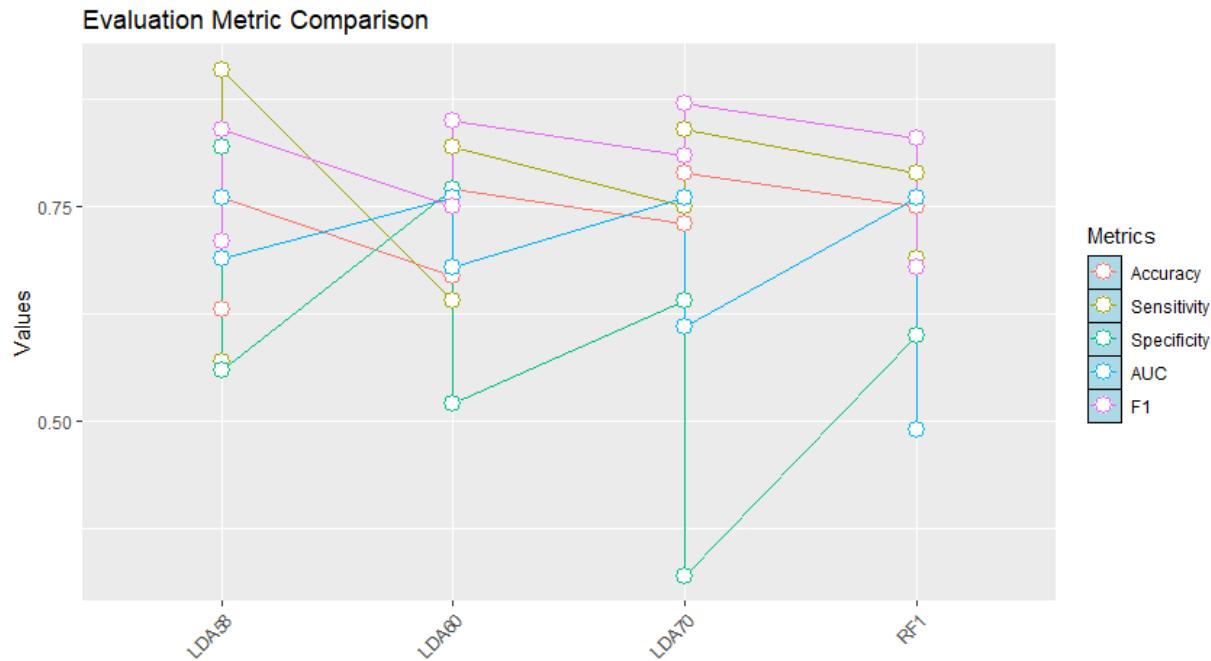
```
Pred_RF=predict(RF,Test.R)
```

Reference	
Prediction	0 1
0	845 413
1	378 864



Model Evaluation

Evaluation Metrics compared all the models on the following Metrics: Accuracy, Sensitivity, Specificity, AUC and F1 Score. The Evaluation Metric Comparison helps us identify the top four models.



Non-Supervised - K-Mean clustering

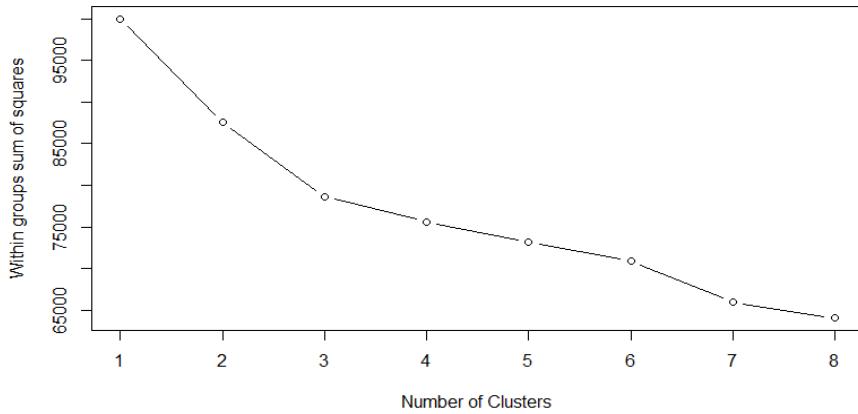
In K-Means clustering, dataset, D, is given from n data points, and k, the number of clusters. Partitioning algorithms organize data points into partition (cluster), where $k \leq n$. [21].

The steps of K-means clustering are [21]:

1. Determine the number of cluster (k).
2. Select initial centroids.
3. Map each data point into a nearest cluster (most similar to centroid).
4. Update mean value (centroid) of each cluster.
5. Repeat step 3-4 until all centroids are not changed

To initialize the number of clusterlets use the elbow method

K=3 or 6



K-means clustering with 3 clusters of sizes 3570, 3921, 2509

Cluster means:

```
CreditScore Geography.Germany Geography.Spain Gender.Male
1 0.007581993 -0.578707 0.1879821 0.05560871
2 -0.013025638 -0.578707 0.1960008 -0.02340192
3 0.009567880 1.727817 -0.5737805 -0.04255248

Age Tenure Balance NumOfProducts
1 -0.02141545 -0.0234086718 0.7434990 -0.54553850
2 -0.03235143 0.0219406055 -1.1204137 0.50822235
3 0.08102954 -0.0009805325 0.6930453 -0.01800214

HasCrCard EstimatedSalary
1 -0.0298860 0.003104425
2 0.0155171 -0.014211056
3 0.0182744 0.017791453
```

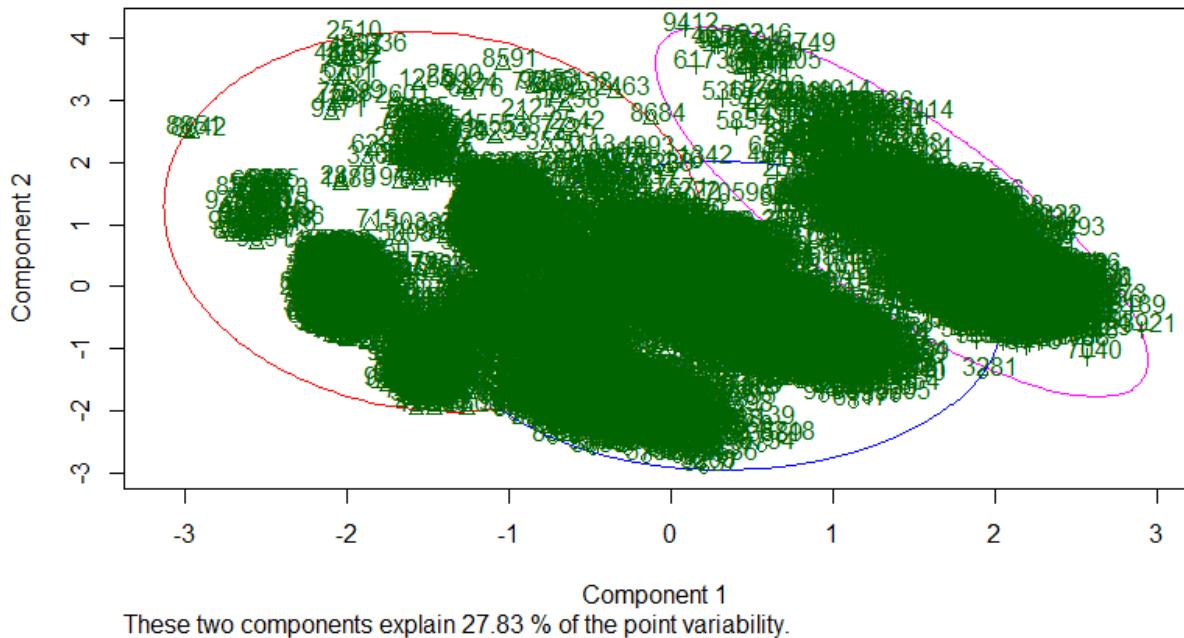
Within cluster sum of squares by cluster:

```
[1] 28194.43 31893.10 18572.43
(between_SS / total_SS = 21.3 %)
```

C	CreditScore	Geography.Germany	Geography.Spain	Gender.Male	Age	Tenure	Balance	NumOfProducts	HasCrCard	EstimatedSalary
1	0.007	-0.5787	0.1879	0.055	-0.021	-0.023	0.743	-0.545	-0.029	0.003
2	-0.013	-0.5787	0.1960	-0.023	0.032	0.021	-1.12	0.508	0.015	-0.014
3	0.009	1.7278	-0.5737	-0.042	0.082	0.001	0.693	-0.018	0.018	0.017

State of customers	1	2	3
Not Exited	2973	3295	1695
Exited	597	626	814

Three clusters were generated based on the churn status, most of the customers who Exited were found in cluster 3. We tried different combinations of clusters to look for any insight, but ended up without any meaningful insight.



The three clusters based on churn status

[centers](#)

Cluster	1	2	3
Not Exited (0)	2973	3295	1695
Exited (1)	597	626	814

Most of those who churned were in the cluster 3

Difference between cluster 1 and 2

[Difference](#)

Balance	NumOfProducts	Gender.Male
1.863913e+00	-1.053761e+00	7.901062e-02
HasCrCard	Tenure	CreditScore
-4.540310e-02	-4.534928e-02	2.060763e-02
EstimatedSalary	Age	Geography.Spain
1.731548e-02	1.093598e-02	-8.018668e-03

Geography.Germany

1.221245e-15

Difference between cluster 1 and 3

Difference

Geography.Germany	Geography.Spain	NumOfProducts
-2.306524409	0.761762573	-0.527536358
Age	Gender.Male	Balance
-0.102444995	0.098161185	0.050453726
HasCrCard	Tenure	EstimatedSalary
-0.048160403	-0.022428139	-0.014687028
CreditScore		
-0.001985887		

Difference between cluster 2 and 3

Difference

Geography.Germany	Balance	Geography.Spain
-2.306524409	-1.813458952	0.769781241
NumOfProducts	Age	EstimatedSalary
0.526224488	-0.113380973	-0.032002509
Tenure	CreditScore	Gender.Male
0.022921138	-0.022593518	0.019150564
HasCrCard		
-0.002757298		

We again fitted the model with k=6 to look for any possible discovery.

With K=6

K-means clustering with 6 clusters of sizes 1751, 2224, 873, 3095, 1343, 714

Cluster means:

CreditScore	Geography.Germany	Geography.Spain	Gender.Male
1 -0.016750946	1.727817	-0.5737805	-0.032711765
2 0.008415401	-0.578707	0.1730192	0.066248099
3 0.011061579	-0.459814	0.1001875	-0.097530759
4 -0.015389469	-0.578707	0.2046009	0.001984622
5 -0.001434289	-0.578707	0.2299856	0.028599148
6 0.070749233	1.724587	-0.5737805	-0.069277565
Age	Tenure	Balance	NumOfProducts
1 -0.003652323	0.02400917	0.7061529	0.0006085558
2 -0.288493778	-0.01826238	0.7473460	-0.5373880491
3 2.035666635	-0.04680419	-0.3786641	-0.1907593612
4 -0.332164214	0.04573469	-1.1551107	0.6677133023
5 -0.124455943	-0.04433109	0.3910833	-0.4993313787
6 0.092522385	-0.05963121	0.6748532	-0.0495129786
HasCrCard	EstimatedSalary		

```

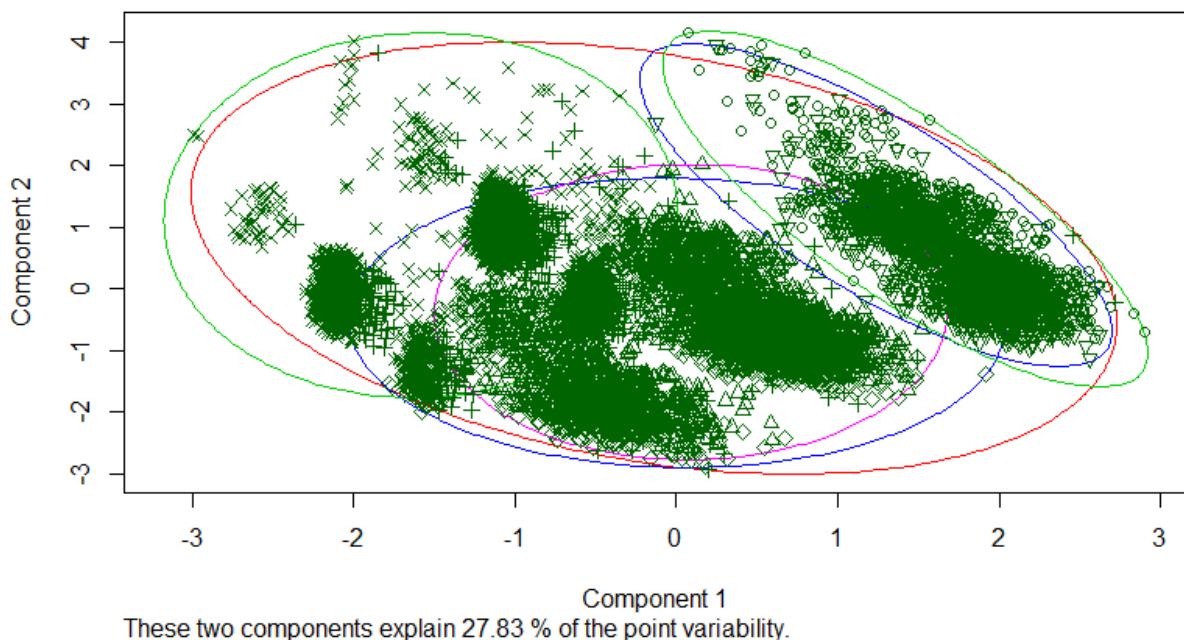
1 0.6460594 0.0334418357
2 0.6460594 0.0004311528
3 0.2666135 -0.0868249634
4 0.1236704 -0.0141955123
5 -1.5476906 0.0457737974
6 -1.5476906 -0.0017597263

```

C	CreditScore	Geography.Germany	Geography.Spain	Gender. Male	Age	Tenure	Balance	NumOfProducts	HasCrCard	EstimatedSalary
1	-0.0167	1.7278	-0.5737	-0.032	-0.003	0.024	0.706	0.0006	0.646	0.0334
2	0.0084	-0.5787	0.1730	0.066	-0.288	-0.018	0.747	-0.5373	0.646	0.0004
3	0.0110	-0.4598	0.1001	-0.097	0.035	-0.046	-0.378	-0.1907	0.266	-0.0868
4	-0.0153	-0.5787	0.2046	0.001	-0.332	0.045	-1.155	0.6677	0.123	-0.0141
5	-0.0014	-0.5787	0.2299	0.028	-0.124	-0.044	0.39	-0.4993	-1.547	0.0457
6	0.0707	1.7245	-0.5737	-0.069	0.092	-0.059	0.674	-0.0495	-1.547	-0.001

State of customers	1	2	3	4	5	6
Not Exited	1184	1908	544	2781	1070	476
Exited	567	316	329	314	273	238

The six clusters based on churn status



Difference between cluster 1 and 3	
CreditScore	0.06481899
Geography.Germany	2.18763140
Geography.Spain	-0.67396792
Gender.Male	0.06481899
Age	-2.03931896
Tenure	0.07081336
Balance	1.08481698
NumOfProducts	0.19136792
HasCrCard	0.37944587
EstimatedSalary	0.19136792

While we are looking for a undiscovered knowledge we ended up in an interesting insight that showed a distinction between cluster 1 and 3 centering the variable exit status, that is in line with our result from logistic regression that is, Being in Germany, higher age and higher balance associated with customers likely or potentially to close their account.

Conclusions

The focus of this project is to predict which banking customers are more likely to close their accounts. We used exploratory analysis as well as predictive, classification and cluster models to determine the customers who are most likely to churn. By identifying the customers with a higher likelihood to leave, our aim is to help the bank formulate effective and targeted retention strategies.

To reduce the number of input variables to improve the performance of our models, the use of feature selection methods helped us filter the most relevant variables. We found the Boruta algorithm, built around the random forest classification algorithm, to be the most efficient for the purpose of identifying which variables are better at predicting customer churn. Of initial 13 input variables, only 7 were found to be relevant candidates for supervised machine learning analysis ("CreditScore", "Geography", "Gender", "Age", "Balance", "NumOfProducts" and "EstimatedSalary"). From a Density Graph of "Age" and our "Exited" variables, we identified higher density in the age cohort of 40-50 years for "Exited" customers and a lower one for customers in the cohort of 30-40 years old.

To find the best accuracy we applied both supervised and unsupervised algorithms to model the customer churning process. Logistic Regression, Decision Tree, Random Forest, K-means and Clustering. From the logistic regression we identified that, holding the other variables constant, for one unit increase in age will increase the chance of churn by 6%, which is consistent with the findings from the density graph. Also the odds of churn for males is 60% less

likely compared to females (odd ratio less than one), and customers in Germany are more likely to churn than in other regions. The logistic regression model showed us that 20% of the customers who need to be proactively worked with to prevent from leaving were identified. In our Random Forest Model, we saw a higher Mean Decrease in Gini, indicating a higher variable importance for Age, number of products and balance.

From the K-means and Clustering we identified that the customers' location being in Germany, higher age and higher balance again increased the likelihood of churn.

By identifying the variables that have the most impact in customer churn, banks will be able to better target their customers, optimize their advertising dollars and protect their profits by increasing customer lifetime value.

You can find all project files at

<https://github.com/melkyed/Modeling-Churning-of-in-Bank-Customers-using-Supervised-and-Unsupervised-Machine-Learning>

References

1. Hashmi, N., N.A. Butt, and M. Iqbal, *Customer churn prediction in telecommunication a decade review and classification*. International Journal of Computer Science Issues (IJCSI), 2013. **10**(5): p. 271.

2. Mahajan, V., R. Misra, and R. Mahajan, *Review of data mining techniques for churn prediction in telecom*. Journal of Information and Organizational Sciences, 2015. **39**(2): p. 183-197.
3. Yan, L., R.H. Wolniewicz, and R. Dodier, *Predicting customer behavior in telecommunications*. IEEE Intelligent Systems, 2004. **19**(2): p. 50-58.
4. Kadeřábková, B. and P. Maleček, *Churning and Labour Market Flows in the New EU Member States*. Procedia Economics and Finance, 2015. **30**: p. 372-378.
5. Qureshi, S.A., et al. *Telecommunication subscribers' churn prediction model using machine learning*. in *Eighth International Conference on Digital Information Management (ICDIM 2013)*. 2013. IEEE.
6. Mishachandar, B. and K.A. Kumar, *Predicting customer churn using targeted proactive retention*. International Journal of Engineering & Technology, 2018. **7**(2.27): p. 69-76.
7. Mishra, K. and R. Rani. *Churn prediction in telecommunication using machine learning*. in *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*. 2017. IEEE.
8. Dalmia, H., C.V. Nikil, and S. Kumar, *Churning of Bank Customers Using Supervised Learning*, in *Innovations in Electronics and Communication Engineering*. 2020, Springer. p. 681-691.
9. Zhao, J. and X.-H. Dang. *Bank customer churn prediction based on support vector machine: Taking a commercial bank's VIP customer churn as the example*. in *2008 4th International Conference on Wireless Communications, Networking and Mobile Computing*. 2008. IEEE.
10. Wang, G., et al. *Predicting credit card holder churn in banks of China using data mining and MCDM*. in *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. 2010. IEEE.
11. Cui, S. and N. Ding. *Customer churn prediction using improved FCM algorithm*. in *2017 3rd International Conference on Information Management (ICIM)*. 2017. IEEE.
12. Pradeep, B., et al., *Analysis of Customer Churn prediction in Logistic Industry using Machine Learning*. Int. J. of Scientific and Res. Publications, 2017. **7**(11).
13. Bilal Zorić, A., *Predicting customer churn in banking industry using neural networks*. Interdisciplinary Description of Complex Systems: INDECS, 2016. **14**(2): p. 116-124.
14. Mishra, A. and U.S. Reddy. *A comparative study of customer churn prediction in telecom industry using ensemble based classifiers*. in *2017 International Conference on Inventive Computing and Informatics (ICICI)*. 2017. IEEE.
15. Lu, N., et al., *A customer churn prediction model in telecom industry using boosting*. IEEE Transactions on Industrial Informatics, 2012. **10**(2): p. 1659-1665.
16. Sayed, H., M.A. Abdel-Fattah, and S. Kholief, *Predicting Potential Banking Customer Churn using Apache Spark ML and MLlib Packages: A Comparative Study*. IJACSA) International Journal of Advanced Computer Science and Applications, 2018. **9**: p. 674-677.
17. Iyyer, S. *Churn Modelling: classification data set*. 2019 [cited 2020 15/04/2020]; Kaggle dataset]. Available from: <https://www.kaggle.com/shrutimechlearn/churn-modelling>.
18. Hosmer, D.W. and S. Lemeshow, *Applied Logistic Regression*. John Wiley & Sons. New York, 2000.

19. Johnson, R. and D. Wichern, *Applied multivariate statistical analysis.. PrenticeHall International*. INC., New Jersey, 2007.
20. Wikipedia, F. *Decision tree*. 2020 [cited 2020 06/05/2020]; Available from: https://en.wikipedia.org/wiki/Decision_tree.
211. Witten, I.H., E. Frank, and M.A. Hall, *Practical machine learning tools and techniques*. Morgan Kaufmann, 2005: p. 578.