HarvardX Data Science Program

Predict customer churn in a bank

Do Quang Anh

2023-03-22

Contents

1 Introduction

Customer churn is a significant issue for businesses across industries, and the banking industry is no exception. Customer churn refers to the phenomenon where customers stop doing business with a company, which can result in significant losses for the company. In the banking industry, churn can be caused by a variety of factors, such as poor customer service, high fees, or better offers from competitors.

In this project, we will explore a dataset of bank customers to predict customer churn using machine learning techniques. The dataset contains information about bank customers, including demographics, account information, and transaction history. Our goal is to use this data to build a model that can predict which customers are likely to churn in the future, allowing the bank to take proactive measures to retain these customers.

2 About Dataset

The dataset we will use in this project is available on Kaggle. The dataset contains information about bank customers, including demographics, account information, and transaction history. The dataset contains 10,000 observations and 14 variables, including:

RowNumber: The row number. CustomerId: The customer ID. Surname: The surname of the customer. CreditScore: The credit score of the customer. Geography: The country of the customer. Gender: The gender of the customer. Age: The age of the customer. Tenure: The number of years the customer has been with the bank. Balance: The balance of the customer. NumOfProducts: The number of bank products the customer has. HasCrCard: Whether the customer has a credit card (1 = yes, 0 = no). IsActiveMember: Whether the customer is an active member (1 = yes, 0 = no). EstimatedSalary: The estimated salary of the customer. Exited: Whether the customer has churned (1 = yes, 0 = no). The dataset also includes a data dictionary that provides more detailed information about each variable. In the next section, we will explore the dataset in more detail and perform data cleaning to prepare it for analysis.

Dataset	Number of Rows	Number of Columns
Customer churn	10,000	14

A preview of the data structure is shown below from the first few rows in data.

RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProdu
1	15,634,602	Hargrave	619	France	Female	42	2	0.00	
2	15,647,311	Hill	608	Spain	Female	41	1	83,807.86	
3	15,619,304	Onio	502	France	Female	42	8	159,660.80	
4	15,701,354	Boni	699	France	Female	39	1	0.00	
5	15,737,888	Mitchell	850	Spain	Female	43	2	125,510.82	
6	15,574,012	Chu	645	Spain	Male	44	8	113,755.78	

Show dimension, datatype, content of the data set

```
## '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 15574012 15592531 15656148 157
##  $ Surname        : chr  "Hargrave" "Hill" "Onio" "Boni" ...
##  $ CreditScore    : int  619 608 502 699 850 645 822 376 501 684 ...
##  $ Geography      : chr  "France" "Spain" "France" "France" ...
##  $ Gender         : chr  "Female" "Female" "Female" "Female" ...
##  $ 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 1 0 1 ...
##  $ IsActiveMember : int  1 1 0 0 1 0 1 0 1 1 ...
##  $ EstimatedSalary: num  101349 112543 113932 93827 79084 ...
##  $ Exited         : int  1 0 1 0 0 1 0 1 0 0 ...
```

3 Exploratory data analysis and data cleaning

Let's see whether there is any missing data.

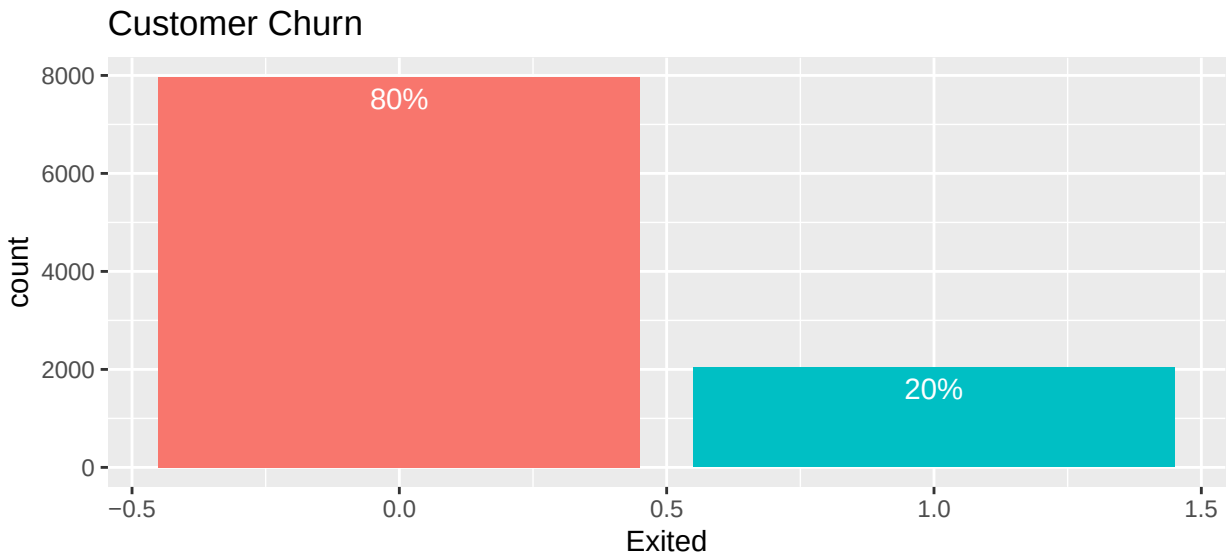
	x
RowNumber	0
CustomerId	0
Surname	0
CreditScore	0
Geography	0
Gender	0
Age	0
Tenure	0
Balance	0
NumOfProducts	0
HasCrCard	0
IsActiveMember	0
EstimatedSalary	0
Exited	0

There are no NA values in the data. Let see unique values for each attribute

RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts
10,000	10,000	2,932	460	3	2	70	11	6,382	4

3.1 Check for imbalance

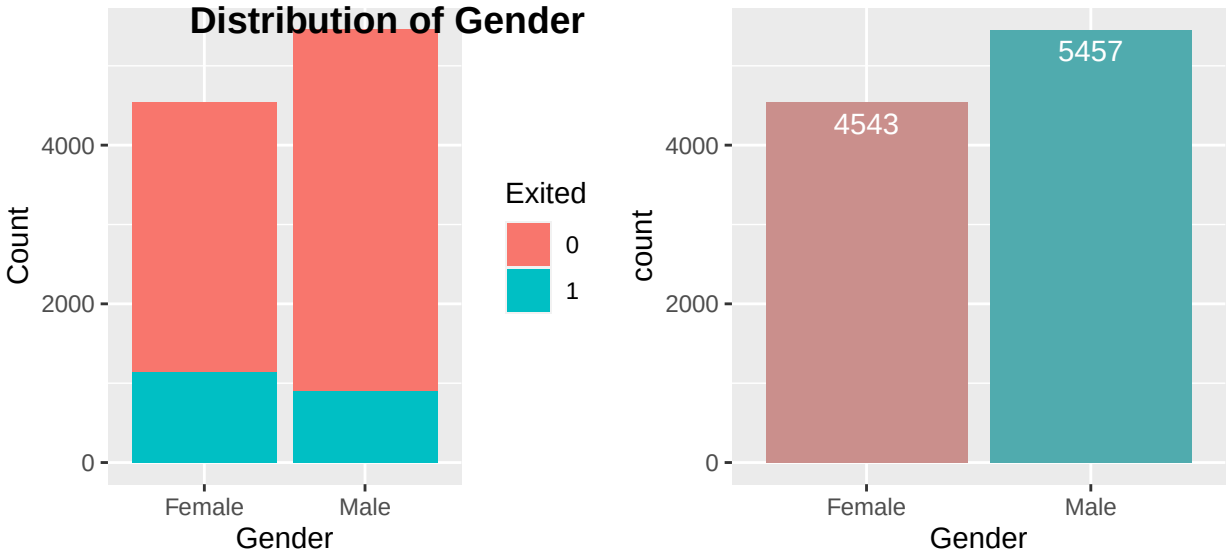
Exited	Count
0	7,963
1	2,037



About 20.4% ~ 2,037 customers left the bank and 79.6% ~ 7963 customers stayed at the bank

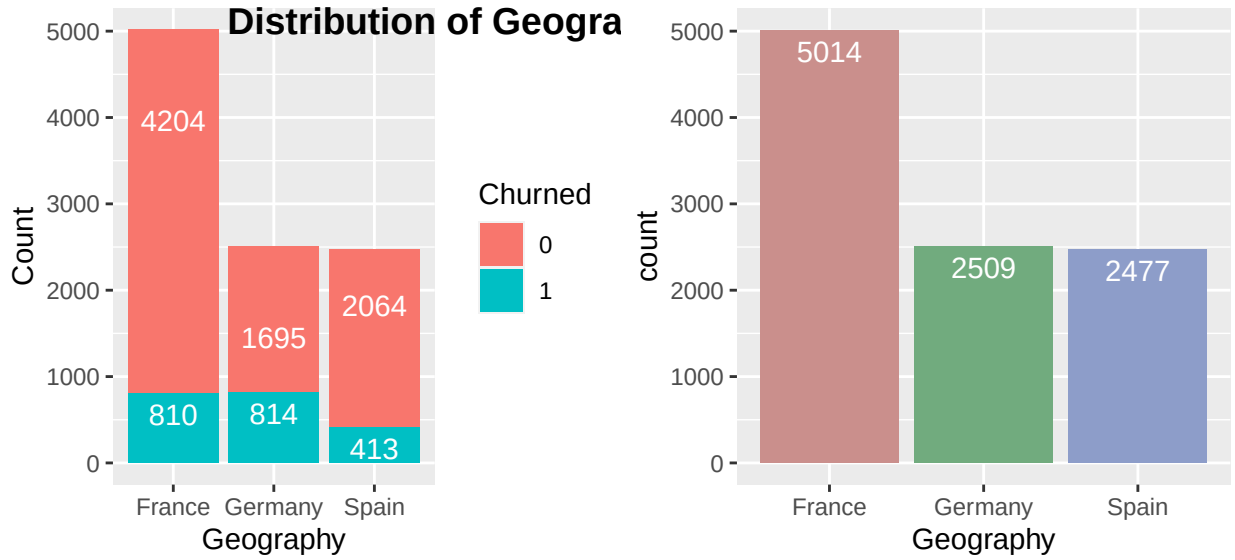
3.2 Distribution of class variables

3.2.1 Gender



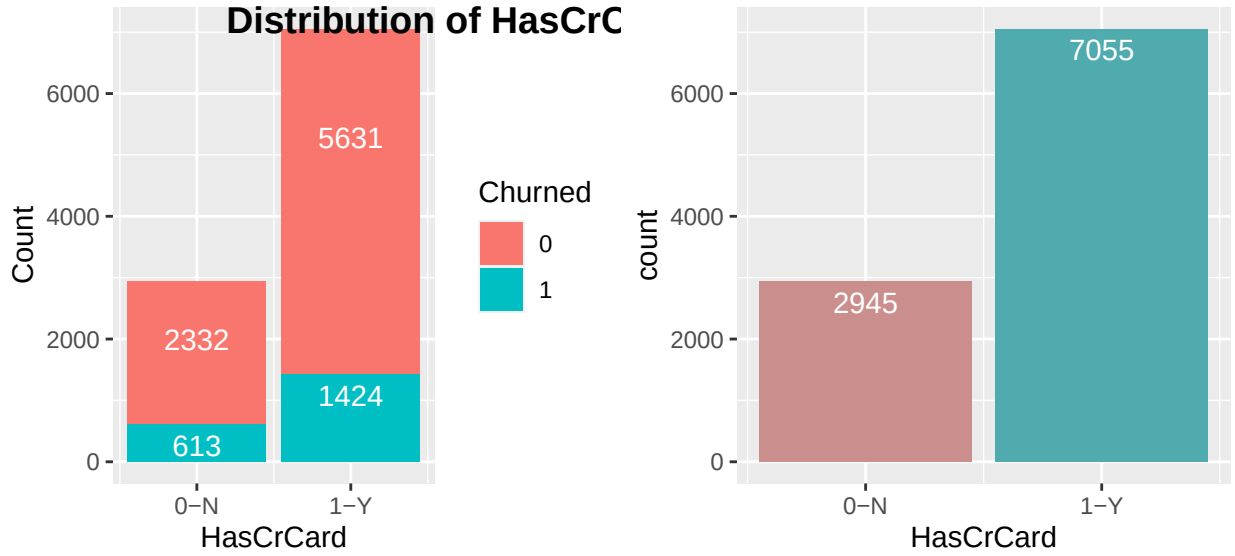
- Female customers are more likely to leave the bank than male customers - More male customers than females

3.2.2 Geography

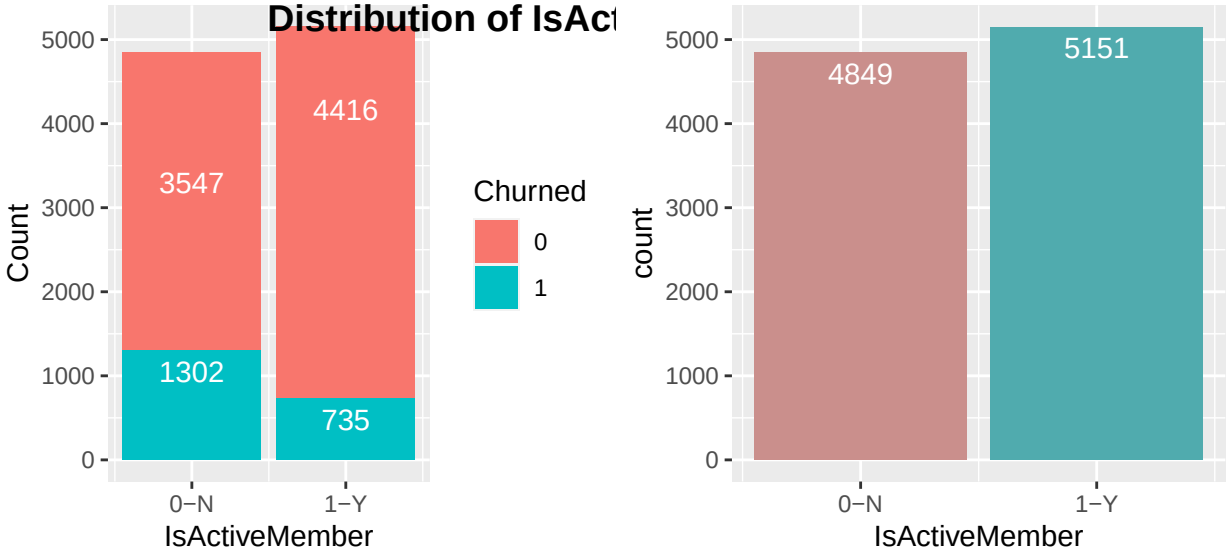


- Customers in France have the highest percentage compared to Germany and Spain, but German customers are more likely to leave the bank than French and Spanish customers

3.2.3 HasCrCard

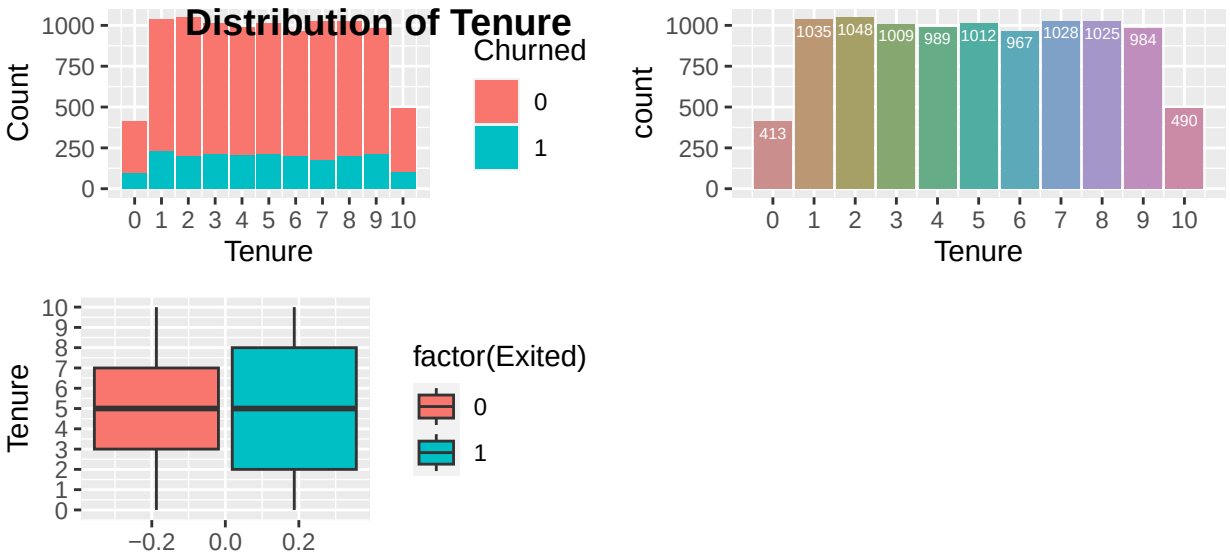


3.2.4 Is active member



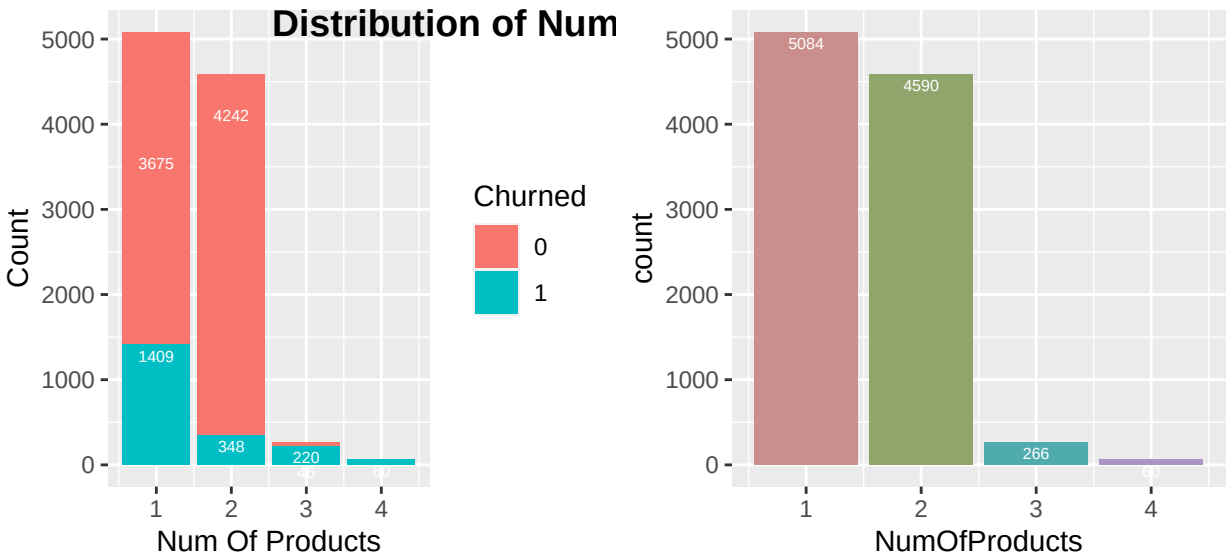
- Customers who are not active members have a higher rate of customer abandonment than customers who are members. Therefore, banks need to have a strategy to move customers from not active members to active members

3.2.5 Tenure



- Customers at both poles (spending less time with the bank or more time with the bank) are more likely to leave than customers with average tenure

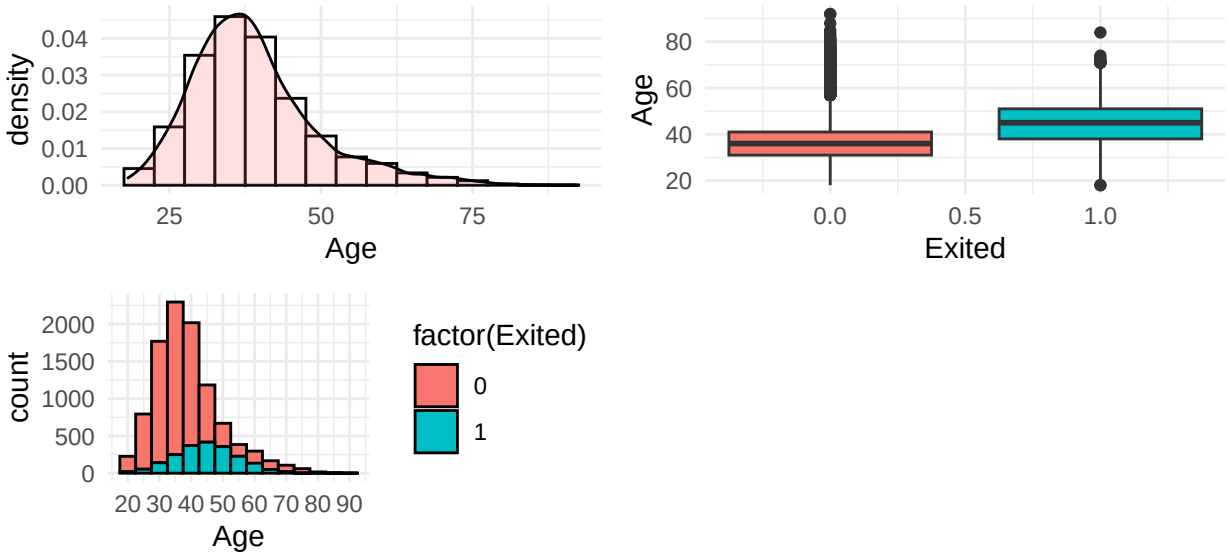
3.2.6 NumOfProducts



- Customers mainly use 1,2 products. customer using only one product are the most churned, however, customers using 3.4 products are more likely to leave. Pretty weird

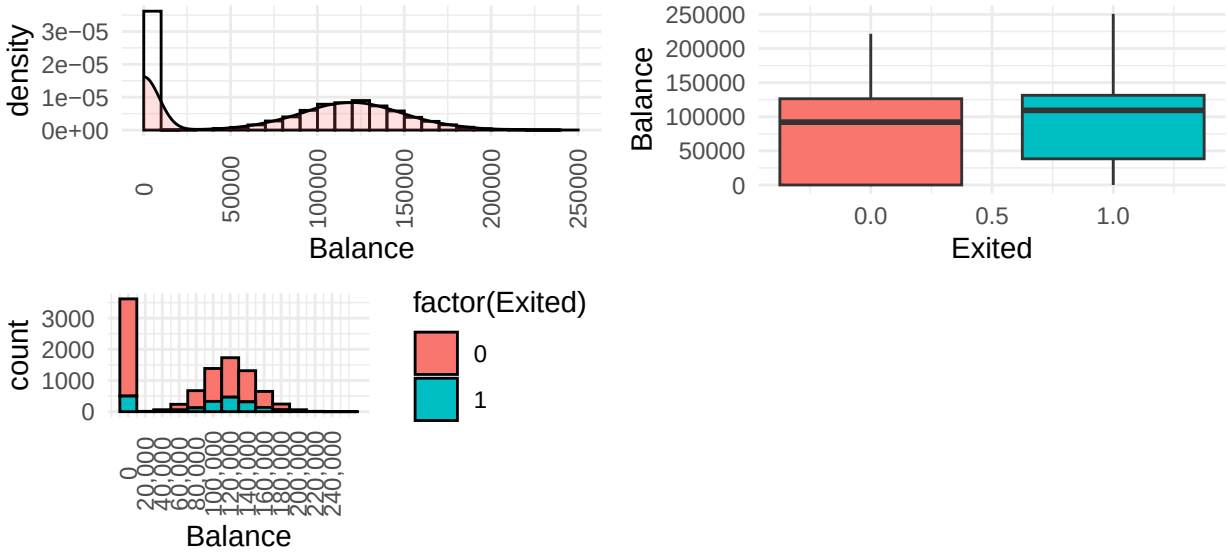
3.3 Distribution of Continuous Variables

3.3.1 Age



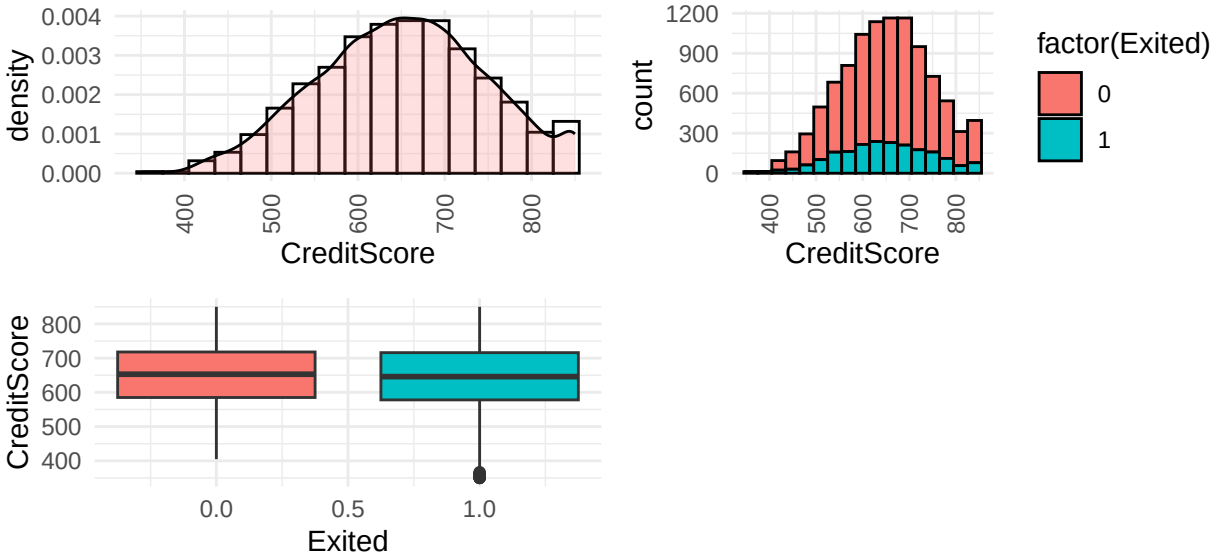
- Older customers are more likely to leave the bank

3.3.2 Balance



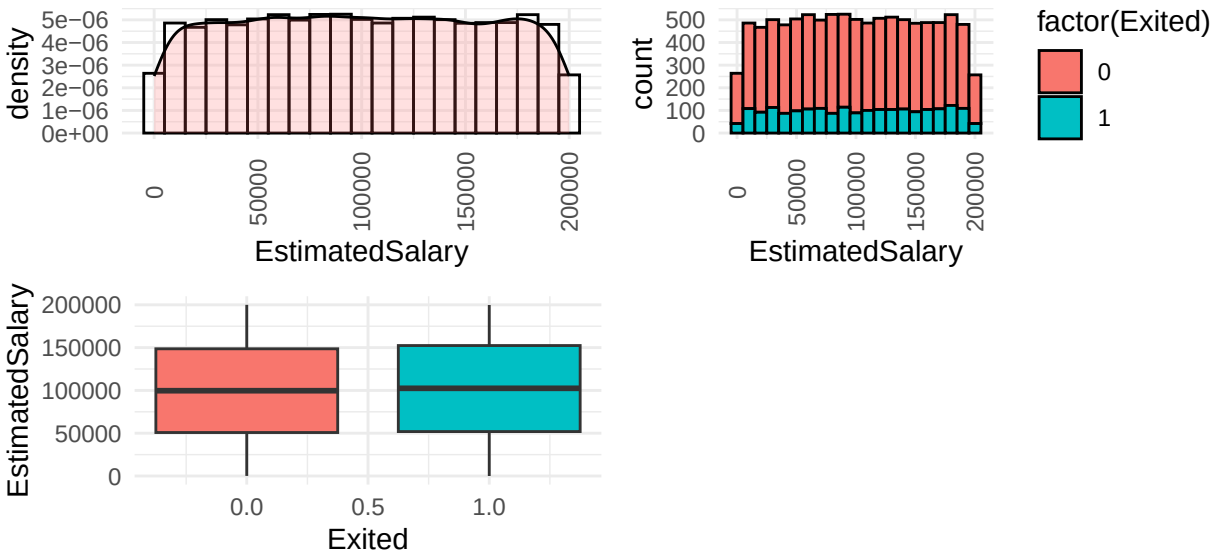
- Customers with large balances tend to leave the bank more than the remaining customers

3.3.3 CreditScore



- The majority of customers have credit scores above 600. We find that customers with low credit scores below 400 are customers who leave the bank

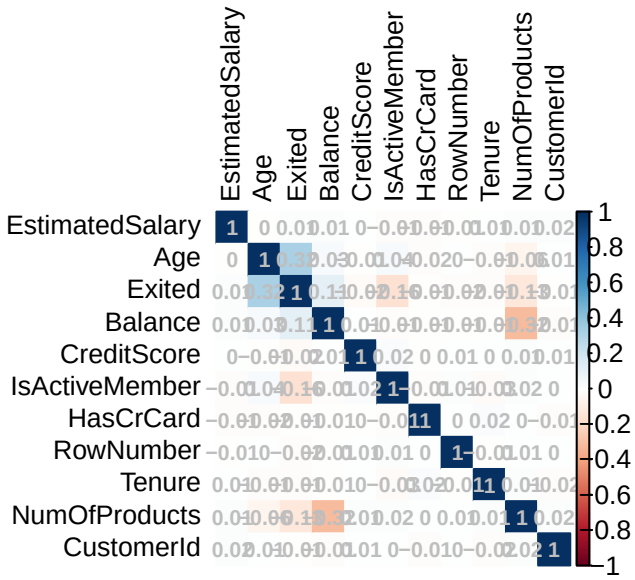
3.3.4 EstimatedSalary



- Looking at the chart, we see the same distribution of wages in both leaving and remaining customers. Therefore, this variable does not have much influence on whether customers leave or stay

3.4 Correlations between each variables

```
##                 RowNumber CustomerId CreditScore   Age Tenure Balance
## RowNumber            1.00       0.00        0.01  0.00  -0.01   -0.01
## CustomerId           0.00       1.00        0.01  0.01  -0.02   -0.01
## CreditScore          0.01       0.01        1.00 -0.01   0.00    0.01
## Age                  0.00       0.01       -0.01  1.00  -0.01    0.03
## Tenure              -0.01      -0.02        0.00 -0.01   1.00   -0.01
## Balance             -0.01      -0.01        0.01  0.03  -0.01    1.00
## NumOfProducts        0.01       0.02        0.01 -0.06   0.01   -0.32
## HasCrCard            0.00      -0.01        0.00 -0.02   0.02   -0.01
## IsActiveMember       0.01       0.00        0.02  0.04  -0.03   -0.01
## EstimatedSalary     -0.01       0.02        0.00  0.00   0.01    0.01
## Exited              -0.02      -0.01       -0.02  0.32  -0.01    0.11
##                 NumOfProducts HasCrCard IsActiveMember EstimatedSalary Exited
## RowNumber                0.01      0.00           0.01           -0.01  -0.02
## CustomerId               0.02     -0.01           0.00            0.02  -0.01
## CreditScore              0.01      0.00           0.02            0.00  -0.02
## Age                     -0.06     -0.02           0.04            0.00   0.32
## Tenure                   0.01      0.02          -0.03            0.01  -0.01
## Balance                 -0.32     -0.01          -0.01            0.01   0.11
## NumOfProducts            1.00      0.00           0.02            0.01  -0.13
## HasCrCard                0.00      1.00          -0.01           -0.01  -0.01
## IsActiveMember           0.02     -0.01           1.00           -0.01  -0.16
## EstimatedSalary          0.01     -0.01          -0.01            1.00   0.01
## Exited                  -0.13     -0.01          -0.16            0.01   1.00
```



- Churn has a positive correlation with age, balance. Generally the correlation coefficients are not so high.
- Balance attribute is negatively correlated with numberofproducts attribute

4 Data Preprocessing & Modeling

4.1 Data Cleansing & Engenerring

4.1.1 Remove unused rows with RowNumber, CustomerId, Surname

```
## 'data.frame':    10000 obs. of  11 variables:
##  $ 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 1 2 1 1 ...
##  $ Gender         : Factor w/ 2 levels "Female","Male": 1 1 1 1 1 2 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      : Factor w/ 2 levels "0","1": 2 1 2 1 2 2 2 2 1 2 ...
##  $ IsActiveMember : int  1 1 0 0 1 0 1 0 1 1 ...
##  $ EstimatedSalary: num  101349 112543 113932 93827 79084 ...
##  $ Exited         : Factor w/ 2 levels "0","1": 2 1 2 1 1 2 1 2 1 1 ...
```

4.1.2 One hot encoding Geography and Gender dataframe

We saw that Geography and Gender columns housed categorical values, so we need to change that as Machine Learning Models take only numerical values

```
## 'data.frame':    10000 obs. of  14 variables:
##  $ Geography.France : num  1 0 1 1 0 0 1 0 1 1 ...
##  $ Geography.Germany: num  0 0 0 0 0 0 0 1 0 0 ...
##  $ Geography.Spain  : num  0 1 0 0 1 1 0 0 0 0 ...
##  $ Gender.Female    : num  1 1 1 1 1 0 0 1 0 0 ...
##  $ Gender.Male      : num  0 0 0 0 0 1 1 0 1 1 ...
```

```
##  $ CreditScore      : int  619 608 502 699 850 645 822 376 501 684 ...
##  $ 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        : Factor w/ 2 levels "0","1": 2 1 2 1 2 2 2 2 1 2 ...
##  $ IsActiveMember   : int  1 1 0 0 1 0 1 0 1 1 ...
##  $ EstimatedSalary  : num  101349 112543 113932 93827 79084 ...
##  $ Exited           : Factor w/ 2 levels "0","1": 2 1 2 1 1 2 1 2 1 1 ...
```

4.1.3 Imbalance data handling

Table 1: Before imbalance data handling

Exited	Count
0	7963
1	2037

Table 2: After imbalance data handling

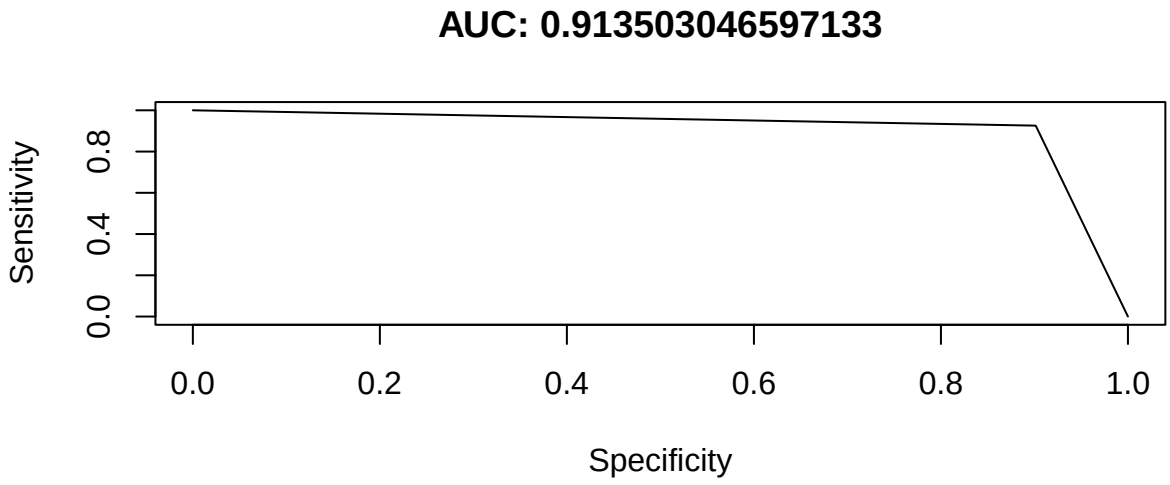
Exited	Count
0	5047
1	4953

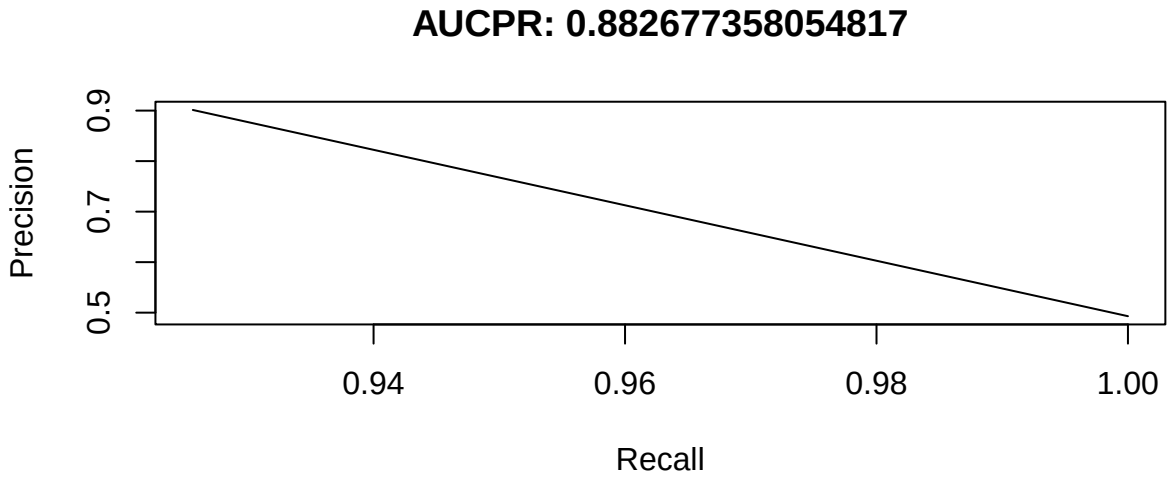
4.1.4 Training and Testing Split

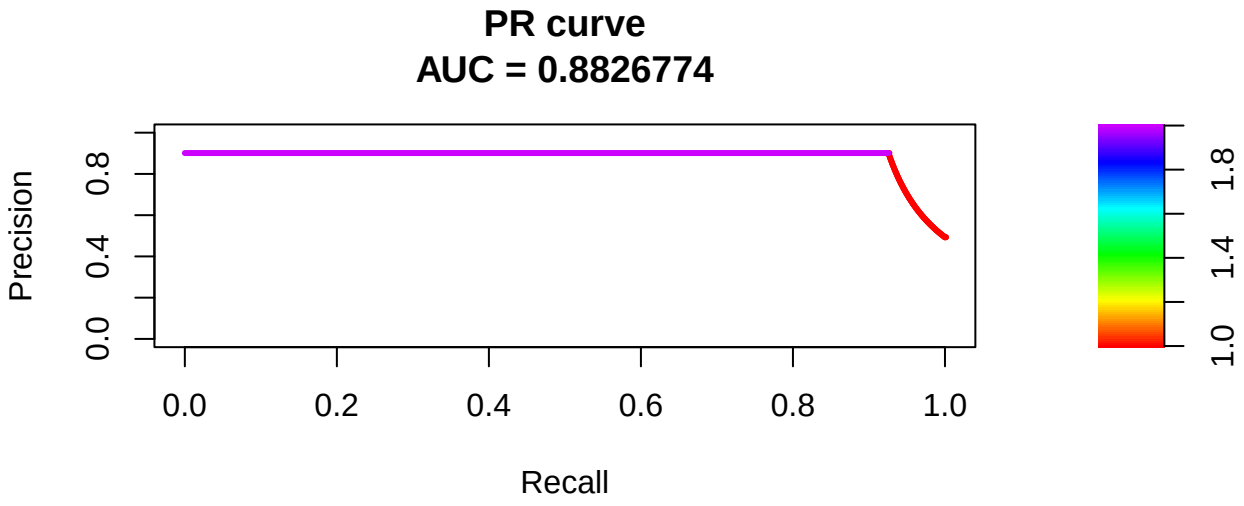
```
# calculates the number of rows in the data_balanced data frame and assigns the result to the row varia
row <- dim(data_balanced)[1]
# randomly samples 70% of the total number of rows in data_balanced and assigns the indices of the sele
train_idx <- sample(row, row * 0.7)
#selects the rows in data_balanced that correspond to the indices stored in train_idx
training_df <- data_balanced[train_idx,]
#Selects all rows in data_balanced except for those with the indices stored in train_idx
testing_df <- data_balanced[-train_idx,]

rm(train_idx)
```

4.2 Modeling

4.2.1 RandomForest

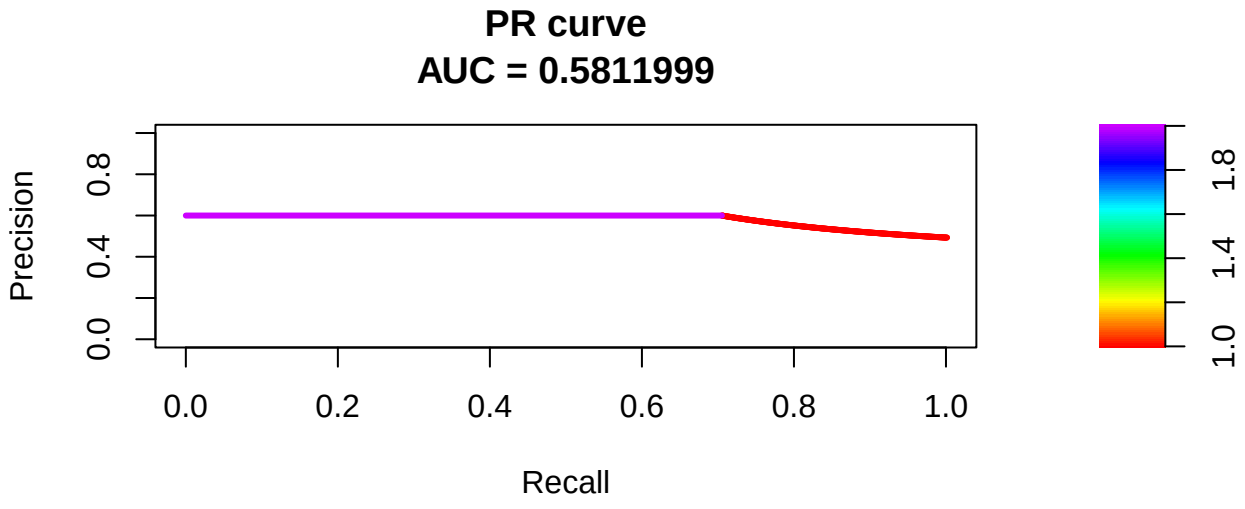


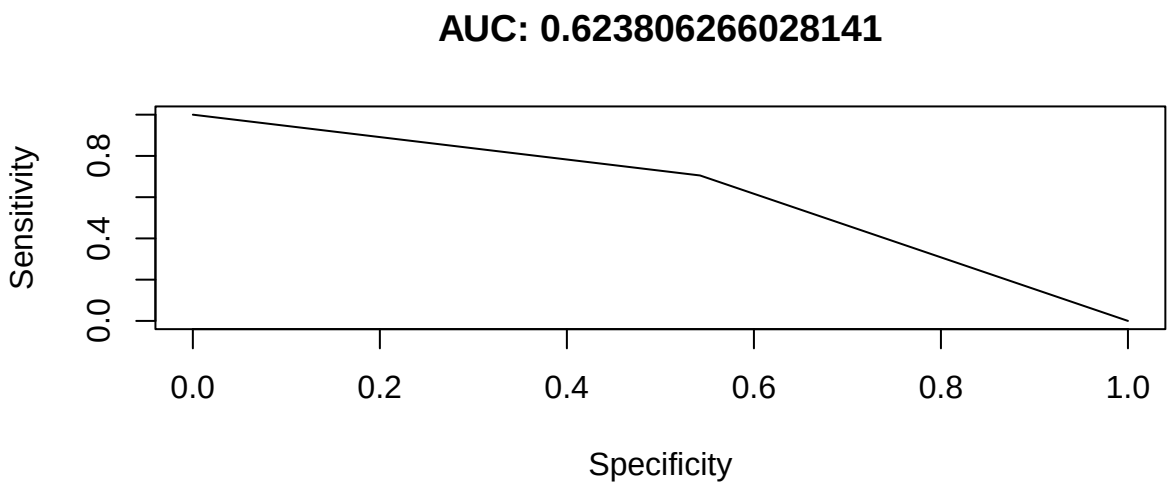


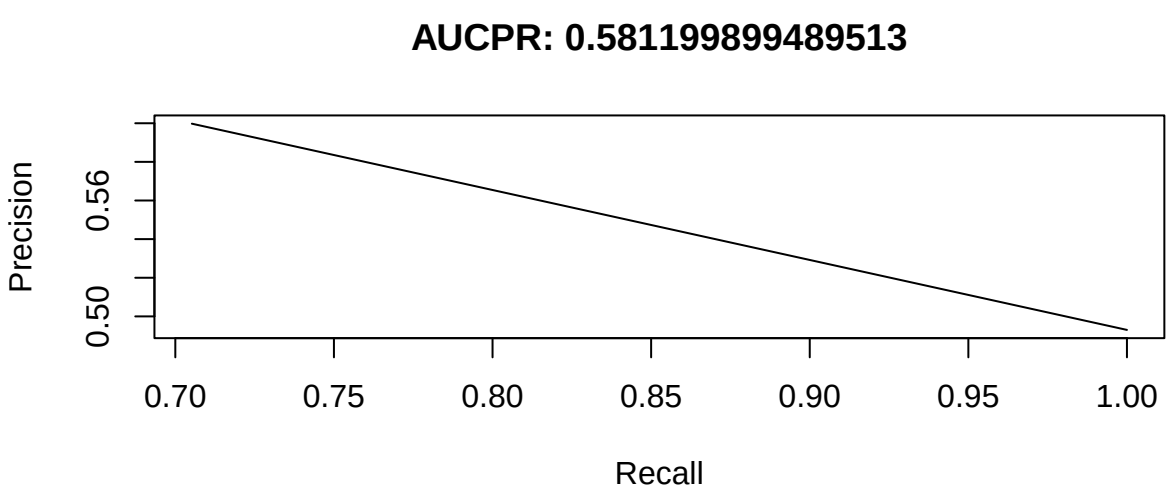


Model	Accuracy	AUC	AUCPR
Random Forest	0.9133333	0.913503	0.8826774

4.2.2 KNN

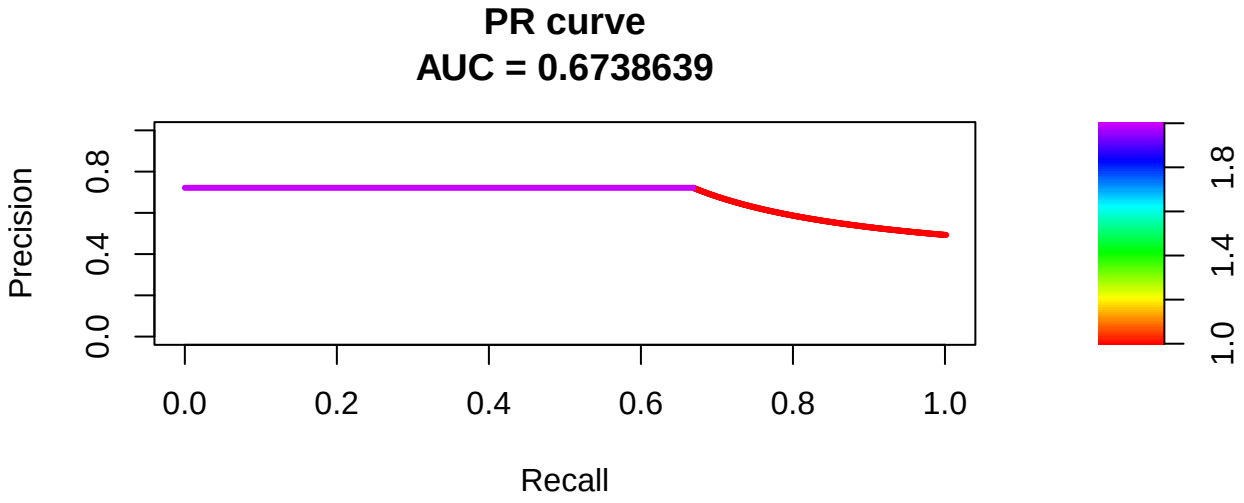

PR curve
AUC = 0.5811999
Precision
0.0
0.4
0.8
0.0
0.2
0.4
0.6
0.8
1.0
Recall
1.0
1.4
1.8

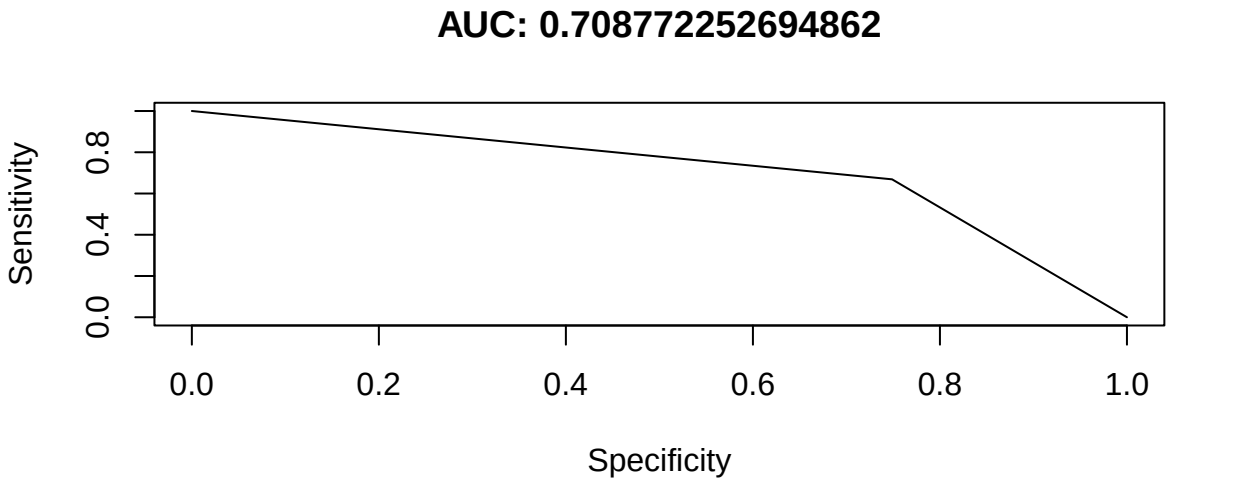

AUC: 0.623806266028141
Sensitivity
0.0
0.4
0.8
0.0
0.2
0.4
0.6
0.8
1.0
Specificity

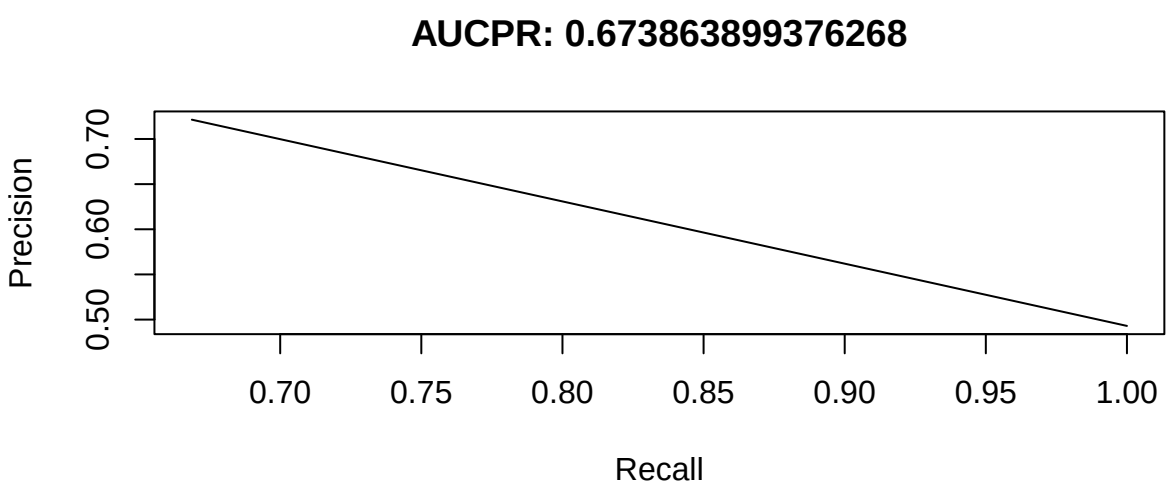


Model	Accuracy	AUC	AUCPR
Random Forest	0.9133333	0.9135030	0.8826774
K-Nearest Neighbors k=5	0.6226667	0.6238063	0.5811999

4.2.3 Naive Algorithm



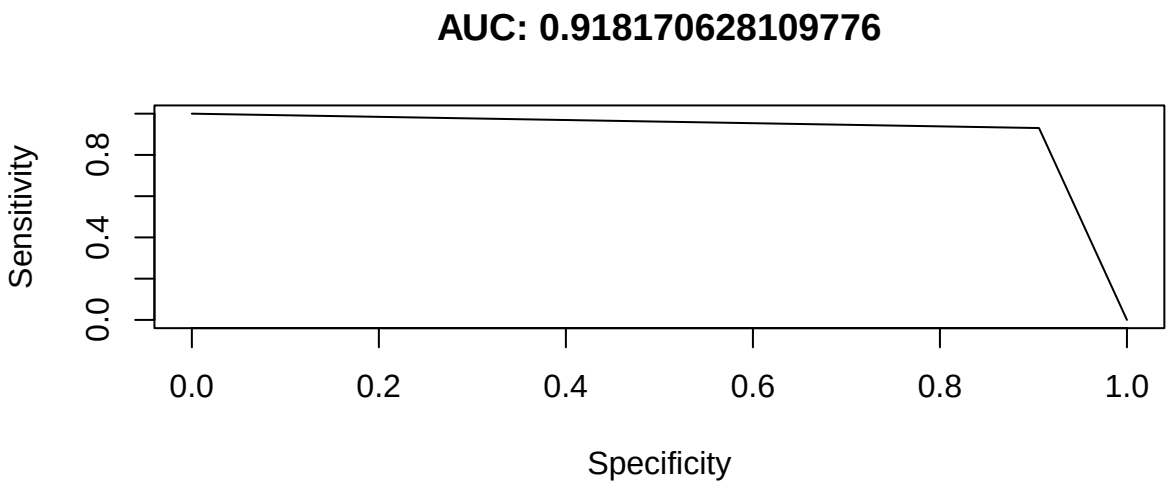


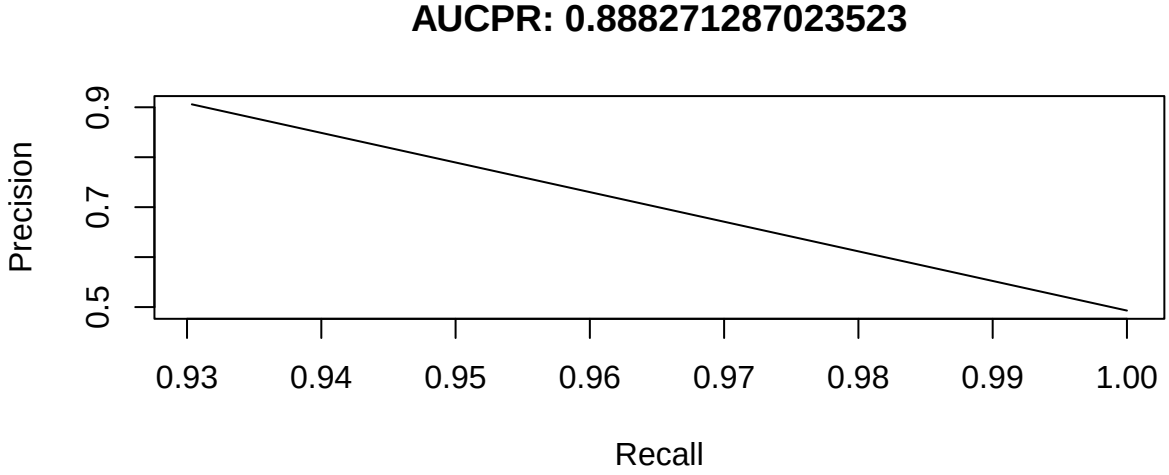


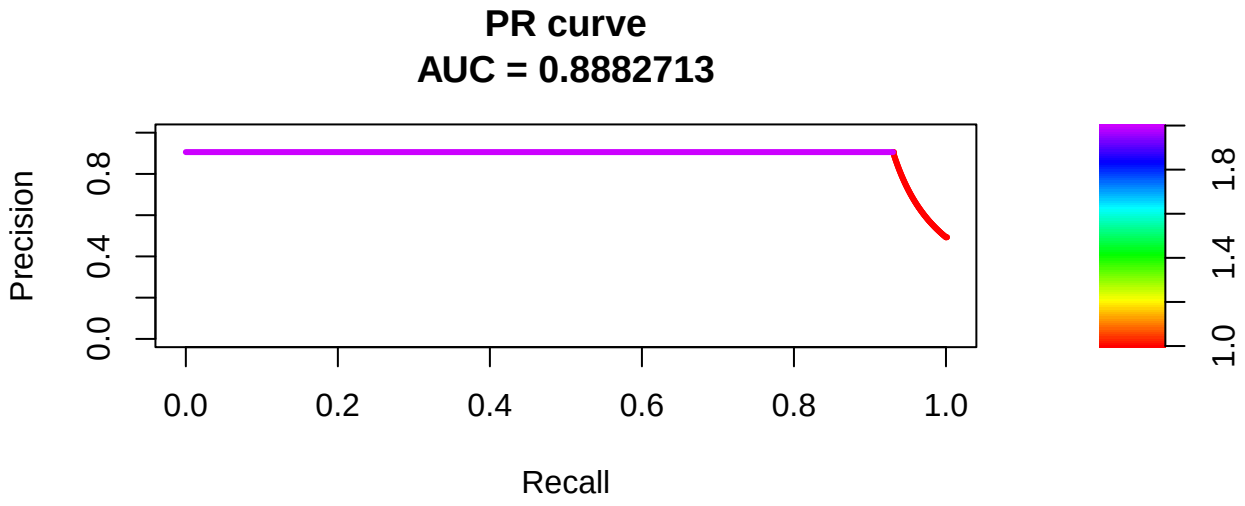
Model	Accuracy	AUC	AUCPR
Random Forest	0.9133333	0.9135030	0.8826774
K-Nearest Neighbors k=5	0.6226667	0.6238063	0.5811999
Naive Bayes	0.7093333	0.7087723	0.6738639

4.2.4 XGBoost

```
##     model        parameter                          label forReg forClass
## 1 xgbTree          nrounds          # Boosting Iterations   TRUE     TRUE
## 2 xgbTree        max_depth                 Max Tree Depth   TRUE     TRUE
## 3 xgbTree              eta                     Shrinkage   TRUE     TRUE
## 4 xgbTree            gamma         Minimum Loss Reduction   TRUE     TRUE
## 5 xgbTree colsample_bytree     Subsample Ratio of Columns   TRUE     TRUE
## 6 xgbTree min_child_weight Minimum Sum of Instance Weight   TRUE     TRUE
## 7 xgbTree        subsample           Subsample Percentage   TRUE     TRUE
##   probModel
## 1      TRUE
## 2      TRUE
## 3      TRUE
## 4      TRUE
## 5      TRUE
## 6      TRUE
## 7      TRUE
```


AUC: 0.918170628109776
Sensitivity
0.0
0.4
0.8
0.0
0.2
0.4
0.6
0.8
1.0
Specificity


AUCPR: 0.888271287023523
Precision
0.5
0.7
0.9
0.93
0.94
0.95
0.96
0.97
0.98
0.99
1.00
Recall



Model	Accuracy	AUC	AUCPR
Random Forest	0.9133333	0.9135030	0.8826774
K-Nearest Neighbors k=5	0.6226667	0.6238063	0.5811999
Naive Bayes	0.7093333	0.7087723	0.6738639
XGBoost	0.9180000	0.9181706	0.8882713

5 Result

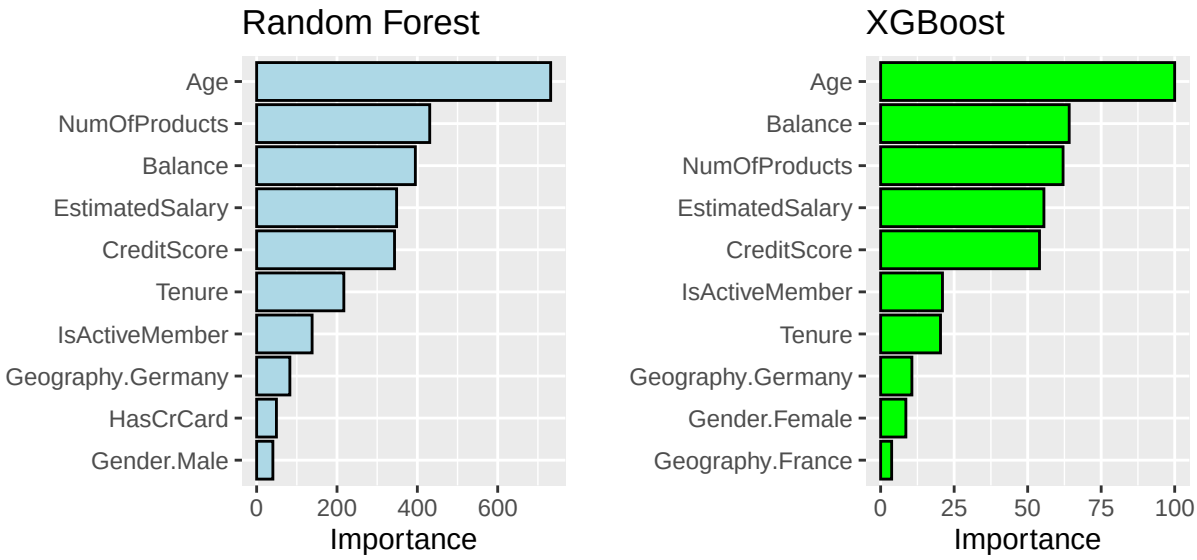
5.1 Compare model perfomance

This is a summary table of customer churn prediction results of different models. We see that the XGBoost model has the highest results with Accuracy = 0.918, AUC = 0.9182 , AUCPR= 0.8882

Table 3: Comparison Table

Model	Accuracy	AUC	AUCPR
Random Forest	0.9133333	0.9135030	0.8826774
K-Nearest Neighbors k=5	0.6226667	0.6238063	0.5811999
Naive Bayes	0.7093333	0.7087723	0.6738639
XGBoost	**0.9180000**	**0.9181706**	**0.8882713**

5.2 Feature Importance



- We only compare the feature important in the two models with the highest scores, Random Forest and XGBoost - The features Importance of the two models are quite similar - Although there are more customers from France but customers from Germany have more influence on the model

6 Conclusion

During our data analysis, we discovered that female customers are the most likely to leave, customers in Germany are most likely to leave, and customers who use only one product are also most likely to leave.After building some models, we find that the Random Forest and XGBoost models are more accurate than others. XGBoost is the model with the highest accuracy of ~92% and AUCPR~90%. However, the data from Kaggle is only a small set, so the model can achieve better performance when providing more historical data for the training phase.

7 Literature

1. Rafael A. Irizarry, Introduction to Data Science
2. HarvardX Data Science Program