CHURN ANALYSIS

Introduction

Customer Churn occurs when a customer stops using a company's products or services. Customer churn affects profitability, especially in industries where revenues are heaveily dependent on subscriptions(e.g. banks, telephone and internet service providers, pay-TV companies, insurance firsms, etc.). It is estimated that acquiring a new customer can cost up to five times more than retaining an existing one. Therefore, customer churn analysis is eesential as it can help a business:

- Identify problems in its services(e.g. poor quality product/service, poor customer support, wrong target audience, etc.)
- Make correct strategic decisions that would lead to higher customer satisfaction and consequently higher customer retention.

Main objective of the analysis

The goal of this work is to understand and predict customer churn for a bank. Starting with the exploratory analysis of the data we try to identify and visualize the factors that most contribute to explain the probability for customer to churn bank services. This analysis will later help us to build Machine Learning models to predict whether a customer will churn or not. This is a typical classification problem which we will try to find an answer for through Logistic Regression and Tree based models.

Data

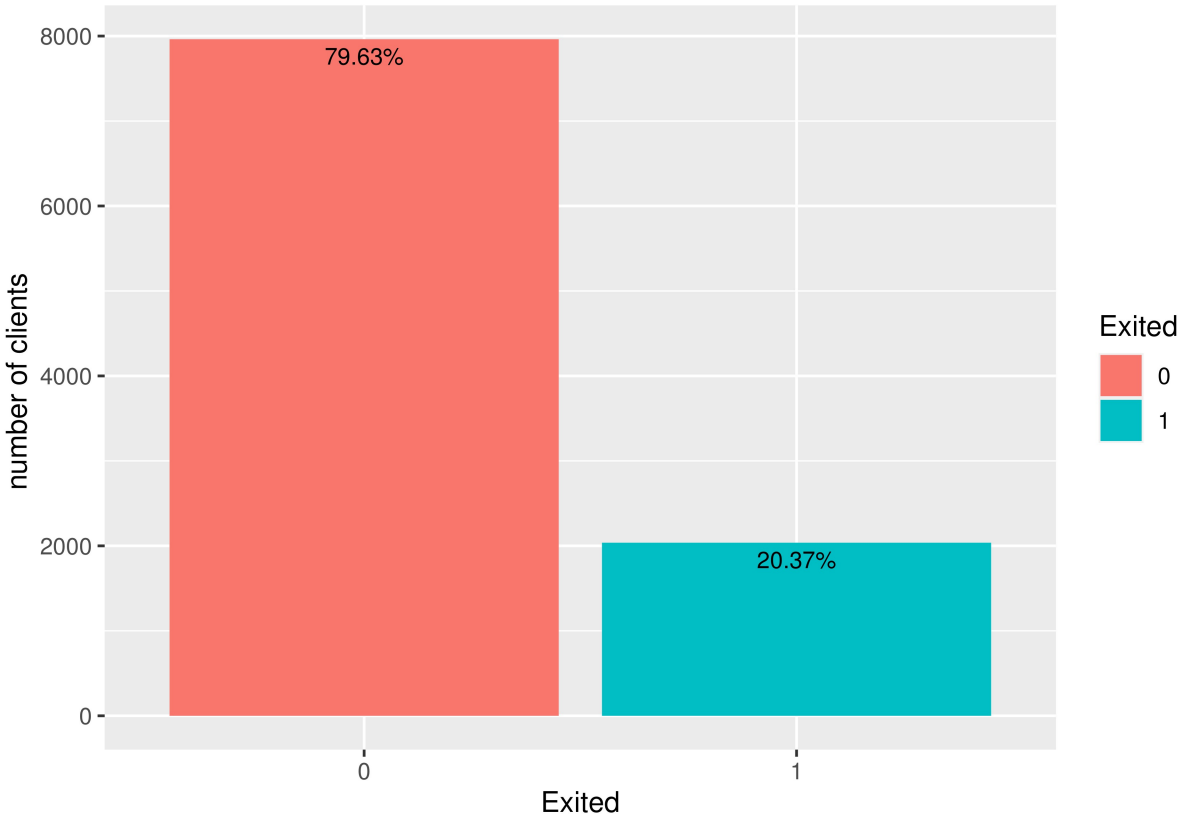
We start our analysis with a data frame of 10K observations on the following 14 variables.

- *Exited:* our target variable indicating whether a client exited or not the bank services (*Yes/No*)
- *RowNumver:* row indexing count
- *CustomerId:* the unique identifier for each bank client
- *Surname:* client's surname
- *Geography:* the country of origin of the client. A factor with 3 levels (France, Germany and Spain)
- *Gender:* a factor with 2 levels (*Male/Female*)
- *CreditScore:* a numerical variable ranging from 0 to 850
- *Age:* a descrete variable ranging from 18 to 92
- *Tenure:* a factor with 11 levels corresponding to the number of months of retention
- *Balance:* numerical variable reporting client's balance amount
- *NumOfProducts:* a factor with 4 levels (*1, 2, 3 and 4*)
- *HasCrCard:* a factor reporting *Yes* if client owns a credit card and *No* otherwise
- *IsActiveMember:* a factor with two levels
- *EstimatedSalary:* a numerical variable giving bank's estimation of their clients annual salary (in euro)

A few considerations on the data summary:

- The age of customers range from 18 to 92 with mean value almost 40.
- The mean and median tenure is 5 months, so the majority of customers is loyal (tenure > 3)
- Approximately 50% of the customers are not active.
- EstimatedSalary displays values below 1000 euros, with the lowest record of 11. Such low values for the annual salary are not realistic. However, given that these salary values cover the entire range of values below 1000 it is unlikely a matter of missrecording.

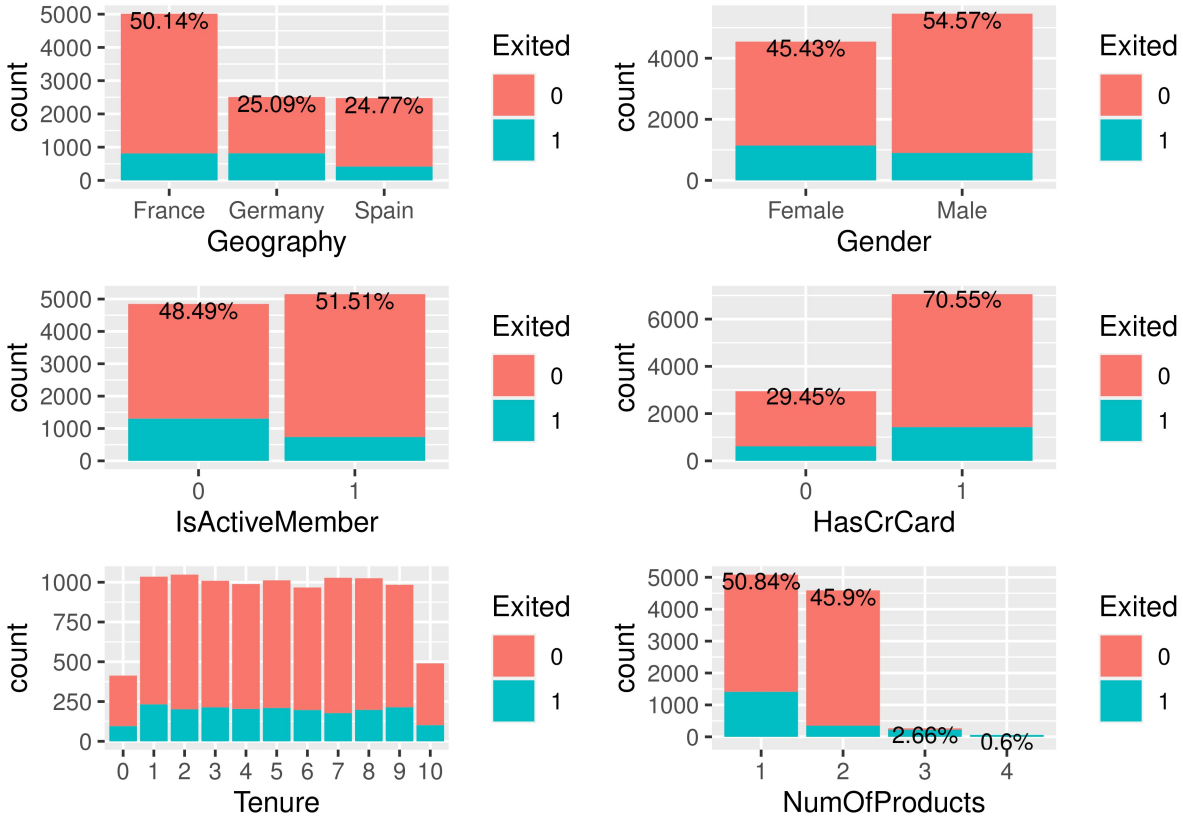
Data exploration



From the plot above of the target variable, we observe that out of total 10000 clients, almost 20% exited bank services. Out task is to understand what might have caused clients to churn the bank, using several observables logically related to such behaviour. In the following section we will first explore one-to-one relationship between each predictor and our target, to get better understanding of the explanatory power of each measured feature on the probability to churn the bank services. We also pause on the relationship among indipentent variables to become aware of any possible threat to the later modelling procedure.

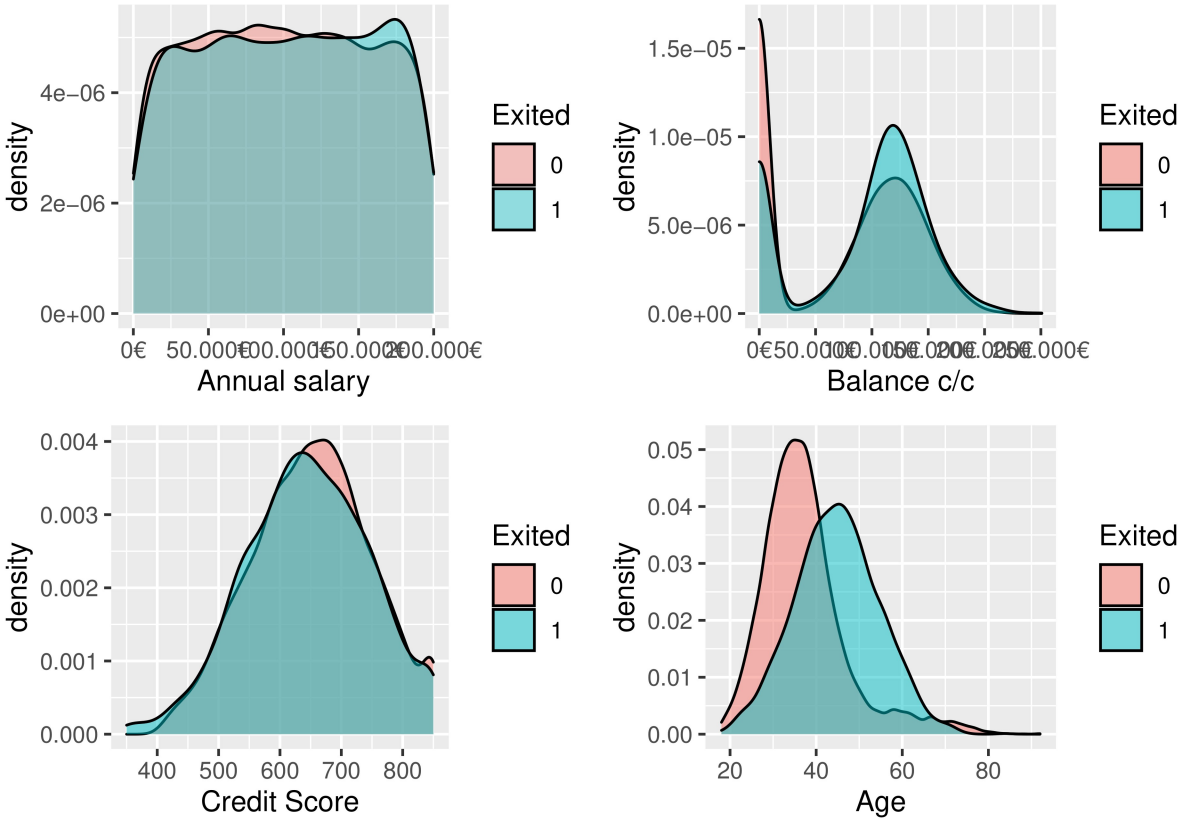
Descriptive statistics

Plotting frequency distribution for categorical predictors by Exited category



- The bank has customers in three different countries (France, Germany, Spain) and most of them are in France. Customers in Germany are more likely to churn, almost half of them are churned. There should be a particular situation regarding this but it is not provided in the data set.
- There are more males than females and females are more likely to churn.
- Only a small percentage leaves within the first year. The count of customers with respect to tenure years is almost the same. There seems to be no relationship with tenure.
- Most of the customers has purchased one or two products. The absolute majority of those who bought four products seems to have churned. It could potentially mean that the bank is unable to properly support customers with more products which in turn increases customer dissatisfaction. If that is the case, the bank might optimize it by not selling more than two products since its where there is the least churn.
- A significant majority of customers has credit card which does not seem to effect the churn rate.
- Almost 50% of customers are not active. Obviously, we would expect inactive members to churn more which is the case here. The bank should develop a strategy to engage their clients and avoid them to churn.

Plotting numerical predictors' distibution by Exired category



- There seems to be no significant difference between remained and churned customers in terms of their credit score, apart from slight prevalence of churn cases for very low scores.
- Both churned and remained customers don't differ in terms of their salary which seems to be uniformly distributed. On this basis the data does not provide evidence that salary could affect the likelihood to churn. We doubt this variable to contribute much to the predictive accuracy of our model.
- The age seems to play a role in explaining the exiting behavior. In fact, clients over 40 tend to be more likely to churn the bank rather than younger clients.
- The balance variable deserves a deeper studying. It is clear that the variable does impact the rate of churn. Lower values of calance account are less related to churn than the higher ones. Moreover, there is a considerable number of clients with null balance. For the modeling purposes we well regress the probability of churn to a piece-wise transformation of the balance variable. We transform it into a categorical variable with 5 categories.

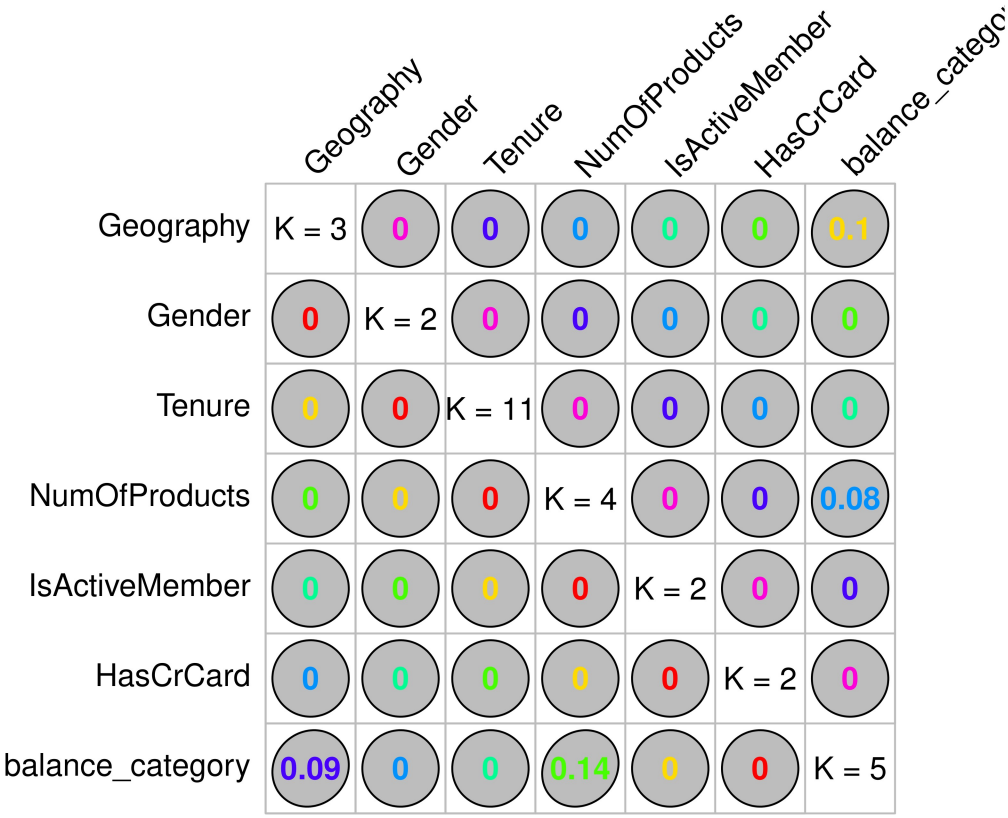
Chi-square test of independence

```
##                    statistic      p.value
## Geography        301.2553368 3.830318e-66
## Gender           112.9185706 2.248210e-26
## Tenure            13.9003726 1.775846e-01
## NumOfProducts   1503.6293615 0.000000e+00
## HasCrCard          0.4713378 4.923724e-01
```

```
## IsActiveMember  242.9853416 8.785858e-55
```

We run the Chi-square test to check whether our categorical predictors are independent from the response. Only Tenure and HasCrCard have high p-values, far above 0.05. Thus, we can't reject the null hypothesis of independence from the response variable for these two predictors. This result confirms our previous findings that these features are not informative to predict the exiting behaviour.

Association analysis



Goodman and Kruskal's gamma is an asymmetric statistics measuring the strength of association when both variables are measured at the ordinal level. Values near -1 indicate perfect negative association, while values close to $+1$ stand for positive association. A value of zero indicates the absence of association. There is a small positive association between balance and number of products hold by the client and the balance and geography.

Correlation analysis

```
##                  CreditScore          Age EstimatedSalary
## CreditScore      1.000000000 -0.003964906    -0.001384293
## Age             -0.003964906  1.000000000    -0.007201042
## EstimatedSalary -0.001384293 -0.007201042     1.000000000
```

From the matrix above we see that the correlation coefficients are vero close to zero for all 3 continuous variables. There is no significant correlation between continuous variables.

MODEL ESTIMATION and FEATURE SELECTION

Now, we will fit a logistic regression model in order to predict the churn event. Before starting to estimate the statistical model we perform a 70/30 split of our data into to training set and test set respectively. We will train our model on the training part and evaluate its performance on the validation set. Later on we will try to improve the efficiency of the model through a subset selection of the relevant predictors. Finally, we estimate the final model on the whole data set.

Full model incorporating all predictors:

```
##
## Call:
## glm(formula = Exited ~ ., family = binomial(link = "logit"),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6388  -0.5616  -0.3422  -0.1658   3.3329
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -2.949e+00  3.416e-01  -8.633  < 2e-16 ***
## CreditScore                -1.222e-04  3.733e-04  -0.327   0.7434
## GeographyGermany            1.114e+00  9.173e-02  12.149  < 2e-16 ***
## GeographySpain              7.598e-02  9.425e-02   0.806   0.4201
## GenderMale                 -5.556e-01  7.217e-02  -7.698 1.38e-14 ***
## Age                         7.313e-02  3.362e-03  21.749  < 2e-16 ***
## Tenure1                    -1.576e-01  1.912e-01  -0.825   0.4096
## Tenure2                    -3.224e-01  1.951e-01  -1.653   0.0984 .
## Tenure3                    -4.553e-01  1.961e-01  -2.322   0.0202 *
## Tenure4                    -1.801e-01  1.944e-01  -0.926   0.3543
## Tenure5                    -3.669e-01  1.959e-01  -1.873   0.0611 .
## Tenure6                    -1.941e-01  1.941e-01  -1.000   0.3171
## Tenure7                    -4.910e-01  1.985e-01  -2.474   0.0134 *
## Tenure8                    -4.774e-01  1.964e-01  -2.431   0.0150 *
## Tenure9                    -3.108e-01  1.937e-01  -1.604   0.1087
## Tenure10                   -4.212e-01  2.301e-01  -1.830   0.0672 .
## NumOfProducts2             -1.669e+00  8.803e-02 -18.956  < 2e-16 ***
## NumOfProducts3              2.826e+00  2.344e-01  12.057  < 2e-16 ***
## NumOfProducts4              1.628e+01  2.169e+02   0.075   0.9402
## HasCrCard1                  4.221e-02  7.917e-02   0.533   0.5940
## IsActiveMember1            -1.177e+00  7.645e-02 -15.390  < 2e-16 ***
## EstimatedSalary             2.548e-07  6.270e-07   0.406   0.6845
## balance_category50K-100K   -4.972e-01  1.228e-01  -4.050 5.11e-05 ***
## balance_category100K-150K  -2.590e-01  1.017e-01  -2.547   0.0109 *
## balance_category150K-200K  -2.720e-01  1.368e-01  -1.989   0.0467 *
## balance_category>200K       1.834e+00  4.518e-01   4.060 4.91e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 7019.5  on 6999  degrees of freedom
## Residual deviance: 5012.8  on 6974  degrees of freedom
## AIC: 5064.8
##
## Number of Fisher Scoring iterations: 14
```

Coefficients interpretation:

- Clients from Germany are more likely to churn bank services with respect to clients from France (baseline category), while the spanish are similar to the french in terms of exiting behaviour;
- Females display higher probability to churn;
- An additional year of age leads to an increase in logit of probability to churn equal to 0,07;
- Overall tenure seems to be weakly negatively associated with the probability to churn;
- Clients that have acquired 2 products from the bank have a lower probability than those with only one product. However holding three products generates an opposite effect, by increasing the probability to churn;
- Active members are less likely to churn;
- For balance level higher than 200K the probability to exit banking services is significantly larger than for lower balance categories;
- Neither the credit card ownership nor the estimated salary produce any statistically significant impact on the churn behavior.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1814  177
##          1  554  455
##
##                Accuracy : 0.7563
##                  95% CI : (0.7406, 0.7716)
##     No Information Rate : 0.7893
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.3988
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.7660
##             Specificity : 0.7199
##          Pos Pred Value : 0.9111
##          Neg Pred Value : 0.4509
##              Prevalence : 0.7893
##          Detection Rate : 0.6047
##    Detection Prevalence : 0.6637
##       Balanced Accuracy : 0.7430
##
##        'Positive' Class : 0
##
```

AUROC for full model:

```
## [1] 0.8140736
```

The misclassification rate is equal to 24.3%. Perhaps we could improve the performance of our model and reduce the variance by removing variables that appear not to be helpful to predict the churn event. We compute a stepwise logistic regression which performs a model selection by AIC. By using Akaike information criterion for the automatic feature selection, we retained 7 out of 10 initial predictors. In particular, confirming our previous expectation, the less informative variables removed from the model are the expected annual salary, the fact of owning a credit card and client's tenure. The error rate of the reduced model is 24.3%, showing a similar performance to the one of the full model. So, the stepwise selection reduced the complexity of the model without compromising its accuracy, which remains stable at 75,7%.

The one metric a bank could be much more interested in is the fraction of exited clients correctly predicted by the model. To compute such a metric we will set a threshold probability to 0.2 , since it seems the most optimal choice providing the best trade-off between model specificity and sensibility. The true positive rate of the reduced model is the same as in full model: We are still able to correctly identify 76% of churn events.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1816  176
##          1  552  456
##
##                Accuracy : 0.7573
##                  95% CI : (0.7416, 0.7726)
##     No Information Rate : 0.7893
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.401
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.7669
##             Specificity : 0.7215
##          Pos Pred Value : 0.9116
##          Neg Pred Value : 0.4524
##              Prevalence : 0.7893
##          Detection Rate : 0.6053
##    Detection Prevalence : 0.6640
##       Balanced Accuracy : 0.7442
##
##        'Positive' Class : 0
##
```

AUROC for reduced model:

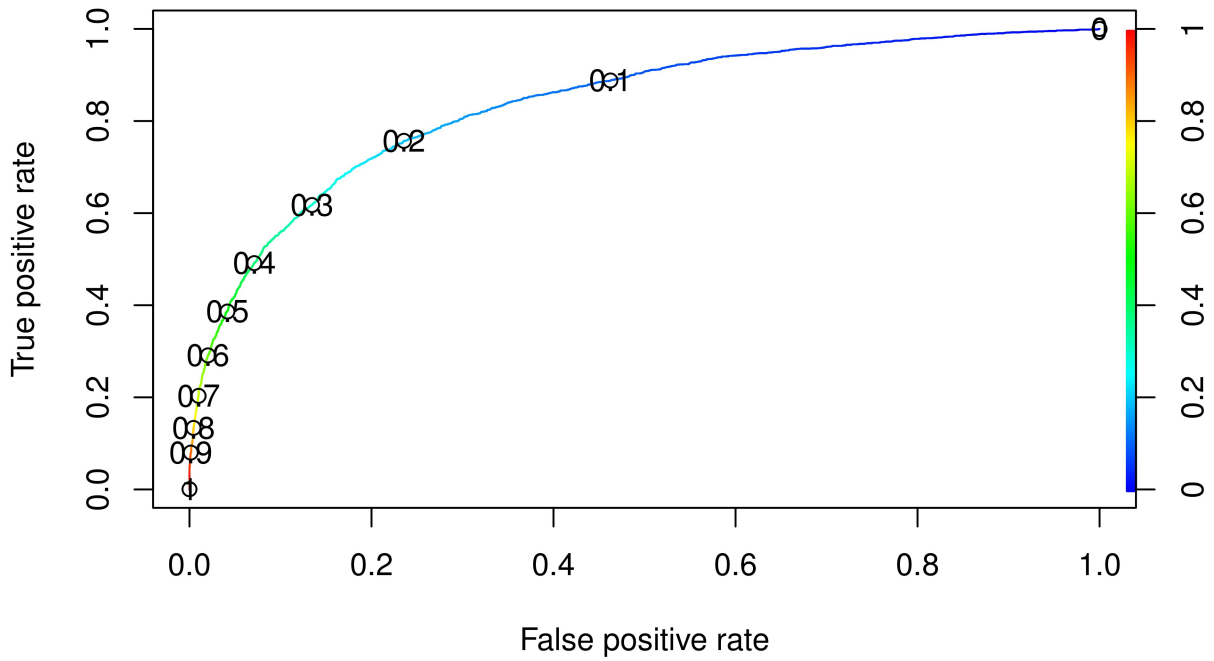
```
## [1] 0.8145417
```

Now, we retrain the final model on the whole dataset, opting for a simpler model returned by the stepwise regression.

```
##
## Call:
## glm(formula = Exited ~ CreditScore + Geography + Gender + Age +
##     NumOfProducts + IsActiveMember + balance_category, family = binomial,
##     data = clean_dta)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5215  -0.5803  -0.3603  -0.1793   3.2306
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                -2.835e+00  2.373e-01 -11.946  < 2e-16 ***
## CreditScore                -6.761e-04  3.040e-04  -2.224   0.0261 *
## GeographyGermany            1.002e+00  7.536e-02  13.292  < 2e-16 ***
## GeographySpain              6.090e-02  7.628e-02   0.798   0.4247
## GenderMale                 -5.238e-01  5.913e-02  -8.857  < 2e-16 ***
## Age                         7.136e-02  2.767e-03  25.794  < 2e-16 ***
## NumOfProducts2             -1.584e+00  7.226e-02 -21.924  < 2e-16 ***
## NumOfProducts3              2.533e+00  1.796e-01  14.103  < 2e-16 ***
## NumOfProducts4              1.631e+01  1.752e+02   0.093   0.9258
## IsActiveMember1            -1.099e+00  6.244e-02 -17.602  < 2e-16 ***
## balance_category50K-100K   -4.032e-01  1.005e-01  -4.012 6.03e-05 ***
## balance_category100K-150K  -1.581e-01  8.304e-02  -1.904   0.0569 .
## balance_category150K-200K  -2.320e-01  1.131e-01  -2.052   0.0401 *
## balance_category>200K       1.671e+00  3.911e-01   4.273 1.93e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 10109.8  on 9999  degrees of freedom
## Residual deviance:  7396.5  on 9986  degrees of freedom
## AIC: 7424.5
##
## Number of Fisher Scoring iterations: 14

## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 6087  495
##          1 1876 1542
##
##                Accuracy : 0.7629
##                  95% CI : (0.7544, 0.7712)
##     No Information Rate : 0.7963
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.4164
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.7644
```

```
##             Specificity : 0.7570
##          Pos Pred Value : 0.9248
##          Neg Pred Value : 0.4511
##              Prevalence : 0.7963
##          Detection Rate : 0.6087
##    Detection Prevalence : 0.6582
##       Balanced Accuracy : 0.7607
##
##        'Positive' Class : 0
##
```



AUROC for the final model:

```
## [1] 0.8355427
```

The misclassification rate of our final model using 20% probability thereshold is 23.7%