Homework #2: Classification task

1- The problem of this task is ABC Multistate Banks loss of customer. In the dataset we have 10000 customers datas(customer id, credit score,country,gender,age,tenure,balance,products number,credit card,active member,estimated salary and churn). The target is churn and the others our features.

```
library(DALEX)
```

```
Welcome to DALEX (version: 2.4.3).
Find examples and detailed introduction at: http://ema.drwhy.ai/
Additional features will be available after installation of: ggpubr.
Use 'install_dependencies()' to get all suggested dependencies
```

```
library(caret)
```

```
Loading required package: ggplot2
```

```
Loading required package: lattice
```

```
library(ROCR)
```

```
bank <- read.csv("bank.csv")
```

```
str(bank)
```

```
'data.frame':	10000 obs. of  12 variables:
 $ customer_id     : int  15634602 15647311 15619304 15701354 15737888 15574012 15592531 1565
 $ credit_score    : int  619 608 502 699 850 645 822 376 501 684 ...
 $ country         : 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 ...
 $ products_number : int  1 1 3 2 1 2 2 4 2 1 ...
 $ credit_card     : int  1 0 1 0 1 1 1 1 0 1 ...
 $ active_member   : int  1 1 0 0 1 0 1 0 1 1 ...
 $ estimated_salary: num  101349 112543 113932 93827 79084 ...
 $ churn           : int  1 0 1 0 0 1 0 1 0 0 ...
```

2- The data have 10000 observation and 12 features. The observation of Country and Gender are Character Customer id, credit score, age tenure, product number, credit card, active member and churn are İnteger Balance and estimated salary are Numeric. For the churn, 1 if the client has left the bank during some period or 0 if he/she has not.

Before starting, we need to check is there any missing value. Because missing values can be trouble for us.

```r
sum(is.na(bank))
```

```
[1] 0
```

There is no missing value at the data set. We can start.

3- Splitting The Dataset

```r
set.seed(1)
index <- sample(1 : nrow(bank), round(nrow(bank) * 0.80))
train <- bank[index, ]
test  <- bank[-index, ]
```

Train The Logistic Regression Model

```r
lr_model <- glm(churn ~ ., data = train, family = "binomial")
```

```r
lr_model
```

```
Call:  glm(formula = churn ~ ., family = "binomial", data = train)

Coefficients:
     (Intercept)       customer_id      credit_score    countryGermany  
      -2.120e+00        -8.026e-08        -7.489e-04         8.138e-01  
    countrySpain        genderMale               age            tenure  
       1.925e-02        -4.672e-01         7.279e-02        -1.949e-02  
         balance   products_number       credit_card     active_member  
       2.408e-06        -1.214e-01        -4.469e-02        -1.057e+00  
estimated_salary  
       9.932e-07  

Degrees of Freedom: 7999 Total (i.e. Null);  7987 Residual
Null Deviance:	    8045 
Residual Deviance: 6819 	AIC: 6845
```

```
summary(lr_model)
```

```
Call:
glm(formula = churn ~ ., family = "binomial", data = train)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.2852  -0.6542  -0.4555  -0.2727   2.9946  

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)    
(Intercept)      -2.120e+00  6.641e+00  -0.319   0.7495    
customer_id      -8.026e-08  4.229e-07  -0.190   0.8495    
credit_score     -7.489e-04  3.138e-04  -2.387   0.0170 *  
countryGermany    8.138e-01  7.570e-02  10.750  < 2e-16 ***
countrySpain      1.925e-02  7.992e-02   0.241   0.8097    
genderMale       -4.672e-01  6.107e-02  -7.651 1.99e-14 ***
age               7.279e-02  2.926e-03  24.881  < 2e-16 ***
tenure           -1.949e-02  1.051e-02  -1.854   0.0637 .  
balance           2.408e-06  5.799e-07   4.152 3.30e-05 ***
products_number  -1.214e-01  5.328e-02  -2.279   0.0227 *  
credit_card      -4.469e-02  6.657e-02  -0.671   0.5020    
active_member    -1.057e+00  6.454e-02 -16.379  < 2e-16 ***
```

```
estimated_salary  9.932e-07  5.304e-07   1.873   0.0611 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8045.1  on 7999  degrees of freedom
Residual deviance: 6818.9  on 7987  degrees of freedom
AIC: 6844.9

Number of Fisher Scoring iterations: 5
```

4-Performance of The Trained Model

```
predicted_probs <- predict(lr_model, test[,-12], type = "response")
head(predicted_probs)
```

```
         1          11          12          13          17          20
0.12530897 0.11645374 0.06141933 0.15279730 0.69735049 0.02728314
```

```
predicted_classes <- ifelse(predicted_probs > 0.5, 1, 0)
head(predicted_classes)
```

```
 1 11 12 13 17 20
 0  0  0  0  1  0
```

```
TP <- sum(predicted_classes[which(test$churn == "1")] == 1)
FP <- sum(predicted_classes[which(test$churn == "1")] == 0)
TN <- sum(predicted_classes[which(test$churn == "0")] == 0)
FN <- sum(predicted_classes[which(test$churn == "0")] == 1)

recall      <- TP / (TP + FN)
specificity <- TN / (TN + FP)
precision   <- TP / (TP + FP)
accuracy    <- (TN + TP) / (TP + FP + TN + FN)

recall
```

```
[1] 0.5986842
```

```r
specificity
```

```
[1] 0.8203463
```

```r
precision
```

```
[1] 0.21513
```

```r
accuracy
```

```
[1] 0.8035
```

The model classifies the observations with 0.80 accuracy. For the precision value shows that only 20% of customers who still using the bank classified correctly.

```r
table(train$churn) / dim(train)[1]
```

```
      0       1 
0.79825 0.20175 
```

It's shows that in the train set 79% observation is belonging to customers who leave the bank and 20% observation is still using the bank. This mean there is a imbalancedness problem.

```r
confusionMatrix(table(ifelse(test$churn == "1", "1", "0"), predicted_classes), positive =
```

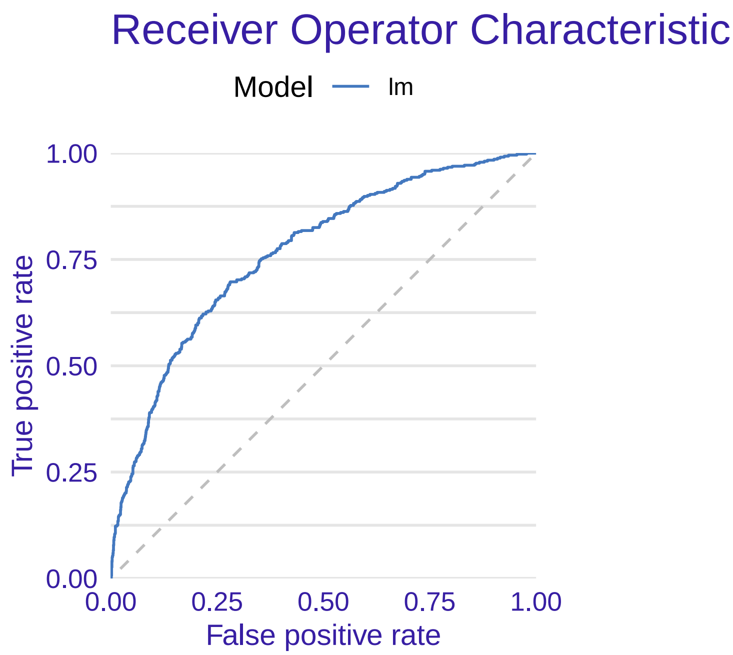
```
Confusion Matrix and Statistics

   predicted_classes
       0    1
  0 1516   61
  1  332   91
                                          
               Accuracy : 0.8035          
                 95% CI : (0.7854, 0.8207)
```

```
    No Information Rate : 0.924
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2305

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.5987
            Specificity : 0.8203
         Pos Pred Value : 0.2151
         Neg Pred Value : 0.9613
             Prevalence : 0.0760
         Detection Rate : 0.0455
   Detection Prevalence : 0.2115
      Balanced Accuracy : 0.7095

       'Positive' Class : 1
```

ROC CURVE

```
explain_lr <- explain(model   = lr_model,
                      data    = test[, -12],
                      y       = test$churn == "1",

                      type    = "classification",
                      verbose = FALSE)
```

```
performance_lr <- model_performance(explain_lr)
plot(performance_lr, geom = "roc")
```




```
performance_lr
```

```
Measures for:  classification
recall      : 0.21513
precision   : 0.5986842
f1          : 0.3165217
accuracy    : 0.8035
auc         : 0.7661074

Residuals:
         0%         10%         20%         30%         40%         50% 
-0.83202612 -0.32303050 -0.22698140 -0.17298866 -0.13301333 -0.09730289 
        60%         70%         80%         90%        100% 
-0.06955816 -0.04627318  0.31218357  0.70077084  0.97373542 
```

Area under curve is 76% not bad and the accuracy is 80%. Thats why there is no problem about the model performance.