STATS/CSE 780
Assignment 2

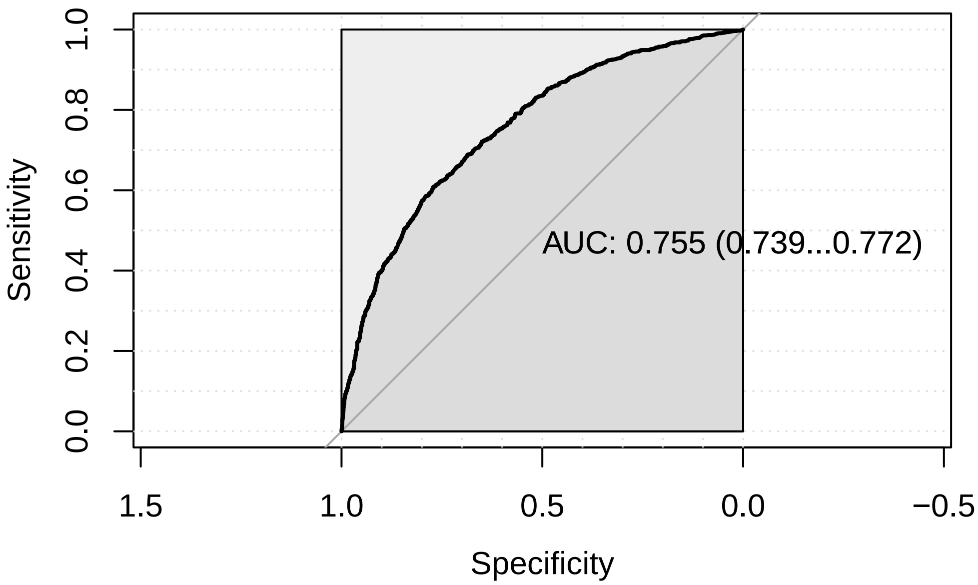
Pao Zhu Vivian Hsu (Student Number: 400547994)

2023-10-15

```
      RowNumber      CustomerId         Surname     CreditScore       Geography 
              0               0               0               0               0 
         Gender             Age          Tenure         Balance   NumOfProducts 
              0               1               0               0               0 
      HasCrCard  IsActiveMember EstimatedSalary          Exited 
              1               1               0               0 

      RowNumber      CustomerId         Surname     CreditScore       Geography 
              0               0               0               0               1 
         Gender             Age          Tenure         Balance   NumOfProducts 
              0              NA               0               0               0 
      HasCrCard  IsActiveMember EstimatedSalary          Exited 
             NA              NA               0               0 

[1] 10002

[1] 11

Call:
glm(formula = Exited ~ ., family = binomial("logit"), data = train_data)


Coefficients:
                  Estimate Std. Error z value Pr(>|z|)    
(Intercept)     -3.632e+00  3.923e-01  -9.258  < 2e-16 ***
CreditScore     -3.266e-04  4.023e-04  -0.812   0.4169    
Age              7.163e-02  3.691e-03  19.408  < 2e-16 ***
Tenure          -2.503e-02  1.352e-02  -1.851   0.0642 .  
Balance          5.531e-06  6.721e-07   8.229  < 2e-16 ***
NumOfProducts   -7.173e-02  6.880e-02  -1.043   0.2972    
HasCrCard        1.165e-01  8.611e-02   1.353   0.1761    
IsActiveMember  -1.125e+00  8.395e-02 -13.400  < 2e-16 ***
EstimatedSalary  9.086e-07  6.840e-07   1.328   0.1841    
Geography        7.272e-02  4.804e-02   1.514   0.1301    
```

```
Gender          -4.156e-01  7.816e-02  -5.317 1.06e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4836.4  on 5000  degrees of freedom
Residual deviance: 4190.2  on 4990  degrees of freedom
AIC: 4212.2

Number of Fisher Scoring iterations: 5

Call:
glm(formula = Exited ~ Age + Balance + IsActiveMember + Gender,
    family = binomial("logit"), data = train_data)

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)    -3.721e+00  2.007e-01 -18.538  < 2e-16 ***
Age             7.175e-02  3.684e-03  19.473  < 2e-16 ***
Balance         5.747e-06  6.475e-07   8.876  < 2e-16 ***
IsActiveMember -1.122e+00  8.359e-02 -13.418  < 2e-16 ***
Gender         -4.173e-01  7.795e-02  -5.353 8.64e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4836.4  on 5000  degrees of freedom
Residual deviance: 4201.0  on 4996  degrees of freedom
AIC: 4211
```

```
Number of Fisher Scoring iterations: 5
```




```
         test_y
log_mod1_y    0    1
         0 3009  429
         1  895  668

[1] 0.2647471

[1] 0.7352529

[1] 0.4273832

[1] 0.8752182

         test_y
knn_mod1_y    0    1
         0 3181  866
         1  723  231

[1] 0.3177365
```

```
[1] 0.6822635
```

```
[1] 0.2421384
```

```
[1] 0.7860143
```

Introduction

The goal of this study is to predict customer churn at a bank.

Methods

Data involving a bank's customers and churn was downloaded from Kaggle (Meshram, n.d.). The original data set consisted of 14 variables and about 10,000 rows of observations. This data was selected because it includes a binary variable indicating customer churn status that is suitable for the purpose of logistic regression and K-nearest neighbour classification. It also contained a variety of variables describing the customer such as estimated salary, age, bank balance, and more. Row numbers, customer ids, and surnames were removed from the data set because they are not important for the purpose of studying customer churn. Based on Harrell's 1:15 rule of choosing predictor variables with respect to sample size (2015), the remaining 10 variables were used as predictor variables for classification. A full description of each variable along with their data types can be found in the Supplementary Materials section (Meshram, n.d.).

Results

Discussion

Supplementary material

Data Description

Code

```
# ----- LOAD PACKAGES AND DATA ----- #

library(dplyr)
library(tidyverse)
library(ggplot2)
library(class)
library(pROC)

bankRaw <- read.csv("Churn_Modelling.csv")



# ----- DATA CLEANSING ----- #

# Check for missing values
sapply(bankRaw, function(x) sum(is.na(x))) # null values
sapply(bankRaw, function(x) sum(x == "")) # blank values

# Clean data
bankWithDef <- bankRaw %>%
  select(-c("RowNumber", "CustomerId", "Surname")) %>% # not needed for analysis
  mutate(Geography_Unclass = unclass(as.factor(Geography)),
         Gender_Unclass = unclass(as.factor(Gender)),
         Age = replace_na(Age, round(mean(Age,na.rm=TRUE),0)), # impute with mean
         HasCrCard = replace_na(HasCrCard, round(mean(HasCrCard,na.rm=TRUE),0)), # impute wi
         IsActiveMember = replace_na(IsActiveMember, round(mean(IsActiveMember,na.rm=TRUE),0
         )
```

```r
# Remove
bank <- bankWithDef %>%
  select(-c("Gender", "Geography")) %>%
  rename(Gender = Gender_Unclass, Geography = Geography_Unclass)



# ----- DATA EXPLORATION ----- #
nrow(bank)
ncol(bank)



# ----- DATA VISUALIZATION  ----- #



# ----- SPLIT INTO TRAIN & TEST DATA ----- #

set.seed(2023780)
train_index <- sample(1:nrow(bank), round(nrow(bank)/2, 0), replace = FALSE)

# Training set
train_data <- bank[train_index, ]
train_x <- dplyr::select(train_data, -Exited)
train_y <- dplyr::pull(train_data, Exited)

# Testing set
test_data <- bank[-train_index, ]
test_x <- dplyr::select(test_data, -Exited)
test_y <- dplyr::pull(test_data, Exited)



# ----- LOGISTIC REGRESSION ----- #
```

```r
set.seed(2023780)

# Include all variables as predictors in the regression
log_mod1 <- glm(Exited ~ ., family = binomial("logit"), data = train_data)
summary(log_mod1)

# Remove predictors with p-values that are not significant (i.e. > 0.05)
log_mod1 <- update(log_mod1, ~ . -CreditScore -Tenure -NumOfProducts
                   -HasCrCard -EstimatedSalary -Geography)
summary(log_mod1)

# Predict outcome using test set
log_mod1_y_prob <- predict(log_mod1, newdata = test_data, type = "response") # y probabiliti


# ----- K-NEAREST NEIGHBOUR CLASSIFICATION ----- #

set.seed(2023780)

# Develop KNN model and predict outcome using test set
knn_mod1_y <- knn(train=train_x, test=test_x, cl=train_y, k=2)


# ----- CLASSIFIER PERFORMANCE ----- #

# --- LOGISTIC REGRESSION MODEL PERFORMANCE --- #
# Find the optimal cut-off value using ROC curve
log_mod1_pROC <- roc(test_y, log_mod1_y_prob, smoothed = TRUE, ci=TRUE, ci.alpha=0.9,
                     plot=TRUE, auc.polygon=TRUE, max.auc.polygon=TRUE, grid=TRUE,
                     print.auc=TRUE, show.thres=TRUE)
```

```
cutoff <- coords(log_mod1_pROC, "best")$threshold

# Assign labels to prediction results using cut-off value
log_mod1_y <- ifelse(log_mod1_y_prob > cutoff, 1, 0)

# Stats on model performance
log_mod1_cmatrix <- table(log_mod1_y, test_y) # Confusion matrix
log_mod1_cmatrix
mean(log_mod1_y != test_y) # Miss-classification error rate (% of churn incorrectly predicte
mean(log_mod1_y == test_y) # Accuracy (% of churn correctly predicted)
log_mod1_cmatrix[2,2]/sum(log_mod1_cmatrix[2,]) # Sensitivity (% correctly predicted as chur
log_mod1_cmatrix[1,1]/sum(log_mod1_cmatrix[1,]) # Specificity (% correctly predicted as not

# --- KNN MODEL PERFORMANCE --- #
# Stats on model performance
knn_mod1_cmatrix <- table(knn_mod1_y, test_y) # Confusion matrix
knn_mod1_cmatrix
mean(knn_mod1_y != test_y) # Miss-classification error rate (% of churn incorrectly predicte
mean(knn_mod1_y == test_y) # Accuracy (% of churn correctly predicted)
knn_mod1_cmatrix[2,2]/sum(knn_mod1_cmatrix[2,]) # Sensitivity (% correctly predicted as chur
knn_mod1_cmatrix[1,1]/sum(knn_mod1_cmatrix[1,]) # Specificity (% correctly predicted as not


# ----- LOGISTIC REGRESSION WITH SHRINKAGE ----- #

set.seed(2023780)
```

References

Harrell, F. E. (2015). Multivariable modeling strategies. In *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis* (pp. 63–102). Springer International Publishing. https://doi.org/10.1007/978-3-319-19425-7_4

Meshram, S. (n.d.). *Bank Customer Churn Prediction*. Kaggle. https://www.kaggle.com/datasets/shubhammeshram579/bank-customer-churn-prediction/

Xie, Y., Dervieux, C., & Riederer, E. (2020). *R markdown cookbook*. CRC Press.