STATS/CSE 780
Assignment 2

Pao Zhu Vivian Hsu (Student Number: 400547994)

2023-10-14

```
      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
```

```
          train_exit
bank_knn_1    0    1
         0 3170  804
         1  812  215
```

```
[1] 0.6768646
```

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

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -3.666e+00  3.807e-01  -9.627  < 2e-16 ***
CreditScore     -6.162e-04  3.983e-04  -1.547   0.1219
Age              7.915e-02  3.743e-03  21.145  < 2e-16 ***
```

```
Tenure          -6.519e-03  1.322e-02  -0.493   0.6219
Balance          5.319e-06  6.624e-07   8.029 9.82e-16 ***
NumOfProducts    2.984e-03  6.516e-02   0.046   0.9635
HasCrCard        6.539e-03  8.485e-02   0.077   0.9386
IsActiveMember  -1.056e+00  8.098e-02 -13.042  < 2e-16 ***
EstimatedSalary  1.027e-06  6.719e-07   1.529   0.1262
Geography        7.998e-02  4.748e-02   1.684   0.0921 .
Gender          -5.267e-01  7.691e-02  -6.849 7.44e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5056.7  on 5000  degrees of freedom
Residual deviance: 4295.4  on 4990  degrees of freedom
AIC: 4317.4

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)    -3.767e+00  2.014e-01 -18.703  < 2e-16 ***
Age             7.907e-02  3.738e-03  21.151  < 2e-16 ***
Balance         5.345e-06  6.410e-07   8.339  < 2e-16 ***
IsActiveMember -1.063e+00  8.083e-02 -13.147  < 2e-16 ***
Gender         -5.216e-01  7.672e-02  -6.798 1.06e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5056.7  on 5000  degrees of freedom
Residual deviance: 4303.1  on 4996  degrees of freedom
AIC: 4313.1

Number of Fisher Scoring iterations: 5
```

Introduction

This dataset was sourced from . KNN classification goal: Predict if the customer will churn (yes or no) KNN regression goal: Predict the customer's tenure - filter for out some columns because there is no description about what they mean

Methods

Results

Discussion

Supplementary material

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

library(tidyverse)
library(ggplot2)
library(class)

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
         )

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

# ----- DATA EXPLORATION ----- #
```

```r
nrow(bank)
ncol(bank)

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

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

set.seed(123456789)
train_index <- sample(1:nrow(bank), round(nrow(bank)/2, 0), replace = FALSE)
train_data <- bank[train_index, ]
train_exit <- pull(train_data, Exited)

test_data <- bank[-train_index, ]

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

set.seed(123456789)
bank_knn_1 <- knn(train=train_data, test=test_data, cl=train_exit, k=2)

table(bank_knn_1, train_exit)
mean(bank_knn_1 == train_exit) # percent of churn correctly predicted

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

set.seed(123456789)

# Include all variables as predictors in the regression
bank_reg_1 <- glm(Exited ~ ., family = binomial("logit"), data = train_data)
summary(bank_reg_1)

# Remove predictors with p-values that are not significant (i.e. > 0.05)
```

```
bank_reg_2 <- update(bank_reg_1, ~ . -CreditScore -Tenure -NumOfProducts
                     -HasCrCard -EstimatedSalary -Geography)
summary(bank_reg_2)
```

References

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