Project 1 Decision

```
#Load all of the libraries
library(tidyr)
library(rpart)
```

```
Warning: package 'rpart' was built under R version 4.3.1
```

```
library(rpart.plot)
library(forecast)
```

```
Registered S3 method overwritten by 'quantmod':
  method            from
  as.zoo.data.frame zoo
```

```
library(caret)
```

```
Loading required package: ggplot2
```

```
Loading required package: lattice
```

```
library(ROSE)
```

```
Loaded ROSE 0.0-4
```

kNN Model

```r
# Load data
data <- read.csv("credit_fa2023_23.csv", header = TRUE)

# Add new fields into data frame to improve model accuracy

for (i in 1:nrow(data)) {
  data$Income_Credit_Ratio[i] <- data$AMT_INCOME_TOTAL[i] / data$AMT_CREDIT[i]
}

for (i in 1:nrow(data)) {
  data$Annuity_Income_Ratio[i] <- data$AMT_ANNUITY[i] / data$AMT_INCOME_TOTAL[i]
}

for (i in 1:nrow(data)) {
  data$Credit_As_Percentage[i] <- data$AMT_CREDIT[i] / data$AMT_INCOME_TOTAL[i]
}

for (i in 1:nrow(data)) {
  data$Percent_Days_Employed[i] <- data$DAYS_EMPLOYED[i] / data$DAYS_BIRTH[i]
}

for (i in 1:nrow(data)) {
  data$Income_Per_Person[i] <- data$AMT_INCOME_TOTAL[i] / data$CNT_FAM_MEMBERS[i]
}

# Remove XNA from CODE_GENDER variable and convert to factor
data <- data[data$CODE_GENDER != "XNA", ]

data$CODE_GENDER <- factor(data$CODE_GENDER)

# Explore data
names(data)
```

```
 [1] "X"                           "SK_ID_CURR"
 [3] "TARGET"                      "NAME_CONTRACT_TYPE"
 [5] "CODE_GENDER"                 "FLAG_OWN_CAR"
 [7] "FLAG_OWN_REALTY"             "CNT_CHILDREN"
 [9] "AMT_INCOME_TOTAL"            "AMT_CREDIT"
[11] "AMT_ANNUITY"                 "AMT_GOODS_PRICE"
[13] "NAME_TYPE_SUITE"             "NAME_INCOME_TYPE"
[15] "NAME_EDUCATION_TYPE"         "NAME_FAMILY_STATUS"
[17] "NAME_HOUSING_TYPE"           "DAYS_BIRTH"
```

```
[19] "DAYS_EMPLOYED"               "DAYS_REGISTRATION"
[21] "DAYS_ID_PUBLISH"             "OWN_CAR_AGE"
[23] "FLAG_MOBIL"                  "FLAG_EMP_PHONE"
[25] "FLAG_WORK_PHONE"             "FLAG_CONT_MOBILE"
[27] "FLAG_PHONE"                  "FLAG_EMAIL"
[29] "OCCUPATION_TYPE"             "CNT_FAM_MEMBERS"
[31] "REGION_RATING_CLIENT"        "REGION_RATING_CLIENT_W_CITY"
[33] "WEEKDAY_APPR_PROCESS_START"  "HOUR_APPR_PROCESS_START"
[35] "REG_REGION_NOT_LIVE_REGION"  "REG_REGION_NOT_WORK_REGION"
[37] "LIVE_REGION_NOT_WORK_REGION" "REG_CITY_NOT_LIVE_CITY"
[39] "REG_CITY_NOT_WORK_CITY"      "LIVE_CITY_NOT_WORK_CITY"
[41] "ORGANIZATION_TYPE"           "DAYS_LAST_PHONE_CHANGE"
[43] "FLAG_DOCUMENT_2"             "FLAG_DOCUMENT_3"
[45] "FLAG_DOCUMENT_4"             "FLAG_DOCUMENT_5"
[47] "FLAG_DOCUMENT_6"             "FLAG_DOCUMENT_7"
[49] "FLAG_DOCUMENT_8"             "FLAG_DOCUMENT_9"
[51] "FLAG_DOCUMENT_10"            "FLAG_DOCUMENT_11"
[53] "FLAG_DOCUMENT_12"            "FLAG_DOCUMENT_13"
[55] "FLAG_DOCUMENT_14"            "FLAG_DOCUMENT_15"
[57] "FLAG_DOCUMENT_16"            "FLAG_DOCUMENT_17"
[59] "FLAG_DOCUMENT_18"            "FLAG_DOCUMENT_19"
[61] "FLAG_DOCUMENT_20"            "FLAG_DOCUMENT_21"
[63] "AMT_REQ_CREDIT_BUREAU_HOUR"  "AMT_REQ_CREDIT_BUREAU_DAY"
[65] "AMT_REQ_CREDIT_BUREAU_WEEK"  "AMT_REQ_CREDIT_BUREAU_MON"
[67] "AMT_REQ_CREDIT_BUREAU_QRT"   "AMT_REQ_CREDIT_BUREAU_YEAR"
[69] "Income_Credit_Ratio"         "Annuity_Income_Ratio"
[71] "Credit_As_Percentage"        "Percent_Days_Employed"
[73] "Income_Per_Person"
```

```
str(data)
```

```
'data.frame':   29999 obs. of  73 variables:
 $ X                          : int  300440 217645 70440 300551 86881 146804 263212 70439 114
 $ SK_ID_CURR                 : int  448070 352176 181716 448195 200835 270206 404772 181715
 $ TARGET                     : int  0 0 1 0 1 0 0 1 1 0 ...
 $ NAME_CONTRACT_TYPE         : chr  "Cash loans" "Cash loans" "Cash loans" "Cash loans" ...
 $ CODE_GENDER                : Factor w/ 2 levels "F","M": 1 1 2 1 2 1 2 1 2 1 ...
 $ FLAG_OWN_CAR               : chr  "N" "N" "N" "N" ...
 $ FLAG_OWN_REALTY            : chr  "N" "Y" "Y" "Y" ...
 $ CNT_CHILDREN               : int  0 0 0 0 0 1 0 0 0 1 ...
 $ AMT_INCOME_TOTAL           : num  157500 90000 180000 171000 135000 ...
 $ AMT_CREDIT                 : num  640080 573628 292500 757598 381528 ...
 $ AMT_ANNUITY                : num  29970 22878 34844 40491 25628 ...
 $ AMT_GOODS_PRICE            : num  450000 463500 292500 702000 315000 ...
 $ NAME_TYPE_SUITE            : chr  "Unaccompanied" "Children" "Unaccompanied" "Unaccompanie
 $ NAME_INCOME_TYPE           : chr  "Commercial associate" "Working" "Working" "State serva
 $ NAME_EDUCATION_TYPE        : chr  "Secondary / secondary special" "Secondary / secondary 
 $ NAME_FAMILY_STATUS         : chr  "Separated" "Widow" "Civil marriage" "Married" ...
 $ NAME_HOUSING_TYPE          : chr  "With parents" "House / apartment" "House / apartment" 
 $ DAYS_BIRTH                 : int  -10953 -20075 -13898 -21445 -10240 -13857 -21167 -17146
 $ DAYS_EMPLOYED              : int  -3005 -1715 -539 -4657 -921 -3113 -3320 -1509 -295 3652
 $ DAYS_REGISTRATION          : int  -5485 -1409 -2070 -3980 -1113 -7952 -2434 -8268 -205 -4
 $ DAYS_ID_PUBLISH            : int  -1284 -3573 -258 -4154 -123 -4604 -3253 -695 -1315 -446
 $ OWN_CAR_AGE                : int  NA NA NA NA 2 NA NA NA NA NA ...
 $ FLAG_MOBIL                 : int  1 1 1 1 1 1 1 1 1 1 ...
 $ FLAG_EMP_PHONE             : int  1 1 1 1 1 1 1 1 1 0 ...
 $ FLAG_WORK_PHONE            : int  0 0 0 0 0 0 0 1 0 0 ...
 $ FLAG_CONT_MOBILE           : int  1 1 1 1 1 1 1 1 1 1 ...
 $ FLAG_PHONE                 : int  0 1 0 0 0 0 1 1 0 0 ...
 $ FLAG_EMAIL                 : int  0 0 0 0 0 0 0 0 0 0 ...
 $ OCCUPATION_TYPE            : chr  "" "Cooking staff" "Laborers" "Core staff" ...
 $ CNT_FAM_MEMBERS            : int  1 1 2 2 1 3 2 1 2 3 ...
 $ REGION_RATING_CLIENT       : int  2 2 2 2 3 1 2 3 3 2 ...
 $ REGION_RATING_CLIENT_W_CITY: int  2 2 2 2 3 1 2 3 3 2 ...
 $ WEEKDAY_APPR_PROCESS_START : chr  "SATURDAY" "FRIDAY" "TUESDAY" "SATURDAY" ...
 $ HOUR_APPR_PROCESS_START    : int  13 11 8 11 17 18 12 11 10 11 ...
 $ REG_REGION_NOT_LIVE_REGION : int  0 0 0 0 0 0 0 0 0 0 ...
 $ REG_REGION_NOT_WORK_REGION : int  0 0 1 0 0 0 0 0 0 0 ...
 $ LIVE_REGION_NOT_WORK_REGION: int  0 0 1 0 0 0 0 0 0 0 ...
 $ REG_CITY_NOT_LIVE_CITY     : int  1 0 0 0 1 0 0 0 0 0 ...
 $ REG_CITY_NOT_WORK_CITY     : int  1 0 0 0 1 0 0 0 1 0 ...
 $ LIVE_CITY_NOT_WORK_CITY    : int  0 0 0 0 0 0 0 0 1 0 ...
 $ ORGANIZATION_TYPE          : chr  "Self-employed" "Hotel" "Business Entity Type 1" "School
 $ DAYS_LAST_PHONE_CHANGE     : int  -2411 -1513 0 -2778 -20 -1827 -1471 -657 0 -796 ...
 $ FLAG_DOCUMENT_2            : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_3            : int  1 1 1 1 1 1 1 1 1 0 ...
 $ FLAG_DOCUMENT_4            : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_5            : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_6            : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_7            : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_8            : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_9            : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_10           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_11           : int  0 0 0 0 0 0 0 0 0 0 ...
```

```
$ FLAG_DOCUMENT_12          : int  0 0 0 0 0 0 0 0 0 0 ...
$ FLAG_DOCUMENT_13          : int  0 0 0 0 0 0 0 0 0 0 ...
$ FLAG_DOCUMENT_14          : int  0 0 0 0 0 0 0 0 0 0 ...
$ FLAG_DOCUMENT_15          : int  0 0 0 0 0 0 0 0 0 0 ...
$ FLAG_DOCUMENT_16          : int  0 0 0 0 0 0 0 0 0 0 ...
$ FLAG_DOCUMENT_17          : int  0 0 0 0 0 0 0 0 0 0 ...
$ FLAG_DOCUMENT_18          : int  0 0 0 0 0 0 0 0 0 0 ...
$ FLAG_DOCUMENT_19          : int  0 0 0 0 0 0 0 0 0 0 ...
$ FLAG_DOCUMENT_20          : int  0 0 0 0 0 0 0 0 0 0 ...
$ FLAG_DOCUMENT_21          : int  0 0 0 0 0 0 0 0 0 0 ...
$ AMT_REQ_CREDIT_BUREAU_HOUR : int  0 0 0 0 0 0 0 0 0 0 ...
$ AMT_REQ_CREDIT_BUREAU_DAY  : int  0 0 0 0 0 0 0 0 0 0 ...
$ AMT_REQ_CREDIT_BUREAU_WEEK : int  0 0 0 0 0 0 0 0 0 0 ...
$ AMT_REQ_CREDIT_BUREAU_MON  : int  0 0 1 0 0 0 0 0 0 0 ...
$ AMT_REQ_CREDIT_BUREAU_QRT  : int  0 0 0 0 0 0 0 0 0 0 ...
$ AMT_REQ_CREDIT_BUREAU_YEAR : int  0 4 2 3 1 1 5 2 3 6 ...
$ Income_Credit_Ratio        : num  0.246 0.157 0.615 0.226 0.354 ...
$ Annuity_Income_Ratio       : num  0.19 0.254 0.194 0.237 0.19 ...
$ Credit_As_Percentage       : num  4.06 6.37 1.62 4.43 2.83 ...
$ Percent_Days_Employed      : num  0.2744 0.0854 0.0388 0.2172 0.0899 ...
$ Income_Per_Person          : num  157500 90000 90000 85500 135000 ...
```

```
summary(data$TARGET)
```

```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.0000  0.0000  0.0000  0.1951  0.0000  1.0000
```

```
# Convert education type to factor with levels across education
data$NAME_EDUCATION_TYPE <- factor(data$NAME_EDUCATION_TYPE, levels = c(
  "Secondary / secondary special",
  "Higher education",
  "Lower secondary",
  "Incomplete higher",
  "Academic degree"))

# Set Target variable as factor
data$TARGET <- as.factor(data$TARGET)
```

```
# Variable list
# Percent_Days_Employed, NAME_EDUCATION_TYPE, REGION_RATING_CLIENT_W_CITY, AMT_GOODS_PRICE
```

```r
# Remove unused variables
data <- data[ , -c(1:2, 4, 6:9, 13:14, 16:17, 22:31, 33:69, 71, 73)]
names(data)
```

```
 [1] "TARGET"                      "CODE_GENDER"
 [3] "AMT_CREDIT"                  "AMT_ANNUITY"
 [5] "AMT_GOODS_PRICE"             "NAME_EDUCATION_TYPE"
 [7] "DAYS_BIRTH"                  "DAYS_EMPLOYED"
 [9] "DAYS_REGISTRATION"           "DAYS_ID_PUBLISH"
[11] "REGION_RATING_CLIENT_W_CITY" "Annuity_Income_Ratio"
[13] "Percent_Days_Employed"
```

```r
# Training - Validation split
set.seed(666)
train_index <- sample(1:nrow(data), 0.7 * nrow(data))
valid_index <- setdiff(1:nrow(data), train_index)
train_df <- data[train_index, ]
valid_df <- data[valid_index, ]

# Double check
nrow(train_df)
```

```
[1] 20999
```

```r
nrow(valid_df)
```

```
[1] 9000
```

```r
head(train_df)
```

```
      TARGET CODE_GENDER AMT_CREDIT AMT_ANNUITY AMT_GOODS_PRICE
17983      1           F   467257.5     17743.5          328500
12926      0           F   260640.0     31059.0          225000
13195      0           F   855882.0     36391.5          765000
23676      0           M   266832.0     24601.5          238500
15901      1           F   360000.0     10102.5          360000
873        0           F   157500.0     14575.5          157500
```

```
                NAME_EDUCATION_TYPE DAYS_BIRTH DAYS_EMPLOYED DAYS_REGISTRATION
17983              Higher education     -10631          -200             -4795
12926 Secondary / secondary special     -17424        -10763             -9653
13195 Secondary / secondary special     -13370         -3493             -6795
23676 Secondary / secondary special     -10606          -383             -5120
15901              Higher education     -14135         -1400             -7982
873                Higher education     -11931         -1154            -10268
      DAYS_ID_PUBLISH REGION_RATING_CLIENT_W_CITY Annuity_Income_Ratio
17983           -3303                           2           0.14082143
12926            -969                           2           0.16051163
13195           -4908                           2           0.10782667
23676           -3208                           2           0.14775676
15901           -3265                           2           0.10204545
873              -788                           2           0.09254286
      Percent_Days_Employed
17983            0.01881291
12926            0.61771120
13195            0.26125654
23676            0.03611163
15901            0.09904492
873              0.09672282
```

```
head(valid_df)
```

```
   TARGET CODE_GENDER AMT_CREDIT AMT_ANNUITY AMT_GOODS_PRICE
2       0           F   573628.5     22878.0          463500
3       1           M   292500.0     34843.5          292500
6       0           F  1198548.0     50913.0         1102500
9       1           M   545040.0     26640.0          450000
10      0           F   337500.0     16875.0          337500
12      1           F   312768.0     17095.5          270000
             NAME_EDUCATION_TYPE DAYS_BIRTH DAYS_EMPLOYED DAYS_REGISTRATION
2  Secondary / secondary special     -20075         -1715             -1409
3  Secondary / secondary special     -13898          -539             -2070
6               Higher education     -13857         -3113             -7952
9  Secondary / secondary special     -18234          -295              -205
10 Secondary / secondary special     -16897        365243             -4399
12 Secondary / secondary special     -12185         -2602             -4412
   DAYS_ID_PUBLISH REGION_RATING_CLIENT_W_CITY Annuity_Income_Ratio
2            -3573                           2            0.2542000
3             -258                           2            0.1935750
```

```
6           -4604                                   1            0.2057091
9           -1315                                   3            0.1315556
10           -446                                   2            0.1500000
12          -3509                                   2            0.2532667
   Percent_Days_Employed
2             0.08542964
3             0.03878256
6             0.22465180
9             0.01617857
10          -21.61584897
12            0.21354124
```

```r
str(train_df)
```

```
'data.frame':   20999 obs. of  13 variables:
 $ TARGET                     : Factor w/ 2 levels "0","1": 2 1 1 1 2 1 1 2 1 1 ...
 $ CODE_GENDER                : Factor w/ 2 levels "F","M": 1 1 1 2 1 1 2 2 2 1 ...
 $ AMT_CREDIT                 : num  467258 260640 855882 266832 360000 ...
 $ AMT_ANNUITY                : num  17744 31059 36392 24602 10102 ...
 $ AMT_GOODS_PRICE            : num  328500 225000 765000 238500 360000 ...
 $ NAME_EDUCATION_TYPE        : Factor w/ 5 levels "Secondary / secondary special",..: 2 1 1
 $ DAYS_BIRTH                 : int  -10631 -17424 -13370 -10606 -14135 -11931 -18336 -12296
 $ DAYS_EMPLOYED              : int  -200 -10763 -3493 -383 -1400 -1154 -6561 -1212 -2285 -6
 $ DAYS_REGISTRATION          : int  -4795 -9653 -6795 -5120 -7982 -10268 -1038 -2417 -3801 -
 $ DAYS_ID_PUBLISH            : int  -3303 -969 -4908 -3208 -3265 -788 -1898 -2422 -4156 -47
 $ REGION_RATING_CLIENT_W_CITY: int  2 2 2 2 2 2 2 2 2 2 ...
 $ Annuity_Income_Ratio       : num  0.141 0.161 0.108 0.148 0.102 ...
 $ Percent_Days_Employed      : num  0.0188 0.6177 0.2613 0.0361 0.099 ...
```

```r
str(valid_df)
```

```
'data.frame':   9000 obs. of  13 variables:
 $ TARGET                     : Factor w/ 2 levels "0","1": 1 2 1 2 1 2 1 2 1 1 ...
 $ CODE_GENDER                : Factor w/ 2 levels "F","M": 1 2 1 2 1 1 2 1 1 1 ...
 $ AMT_CREDIT                 : num  573628 292500 1198548 545040 337500 ...
 $ AMT_ANNUITY                : num  22878 34844 50913 26640 16875 ...
 $ AMT_GOODS_PRICE            : num  463500 292500 1102500 450000 337500 ...
 $ NAME_EDUCATION_TYPE        : Factor w/ 5 levels "Secondary / secondary special",..: 1 1 2
 $ DAYS_BIRTH                 : int  -20075 -13898 -13857 -18234 -16897 -12185 -10579 -11938
 $ DAYS_EMPLOYED              : int  -1715 -539 -3113 -295 365243 -2602 -246 -680 365243 -15
 $ DAYS_REGISTRATION          : int  -1409 -2070 -7952 -205 -4399 -4412 -4665 -5949 -4764 -47
 $ DAYS_ID_PUBLISH            : int  -3573 -258 -4604 -1315 -446 -3509 -3226 -4153 -82 -2312
 $ REGION_RATING_CLIENT_W_CITY: int  2 2 1 3 2 2 2 2 2 2 ...
 $ Annuity_Income_Ratio       : num  0.254 0.194 0.206 0.132 0.15 ...
 $ Percent_Days_Employed      : num  0.0854 0.0388 0.2247 0.0162 -21.6158 ...
```

```r
# Use ROSE to to balance model
train_df_rose <- ROSE(TARGET ~ Percent_Days_Employed + NAME_EDUCATION_TYPE + REGION_RATING
                      data = train_df, seed = 666)$data

table(train_df_rose$TARGET)
```

```

    0     1 
10337 10640 
```

```r
# Check variables
names(data)
```

```
 [1] "TARGET"                      "CODE_GENDER"                
 [3] "AMT_CREDIT"                  "AMT_ANNUITY"                
 [5] "AMT_GOODS_PRICE"             "NAME_EDUCATION_TYPE"        
 [7] "DAYS_BIRTH"                  "DAYS_EMPLOYED"              
 [9] "DAYS_REGISTRATION"           "DAYS_ID_PUBLISH"            
[11] "REGION_RATING_CLIENT_W_CITY" "Annuity_Income_Ratio"       
[13] "Percent_Days_Employed"      
```

```r
# Normalization algorithm
train_norm <- train_df_rose
valid_norm <- valid_df

names(train_df)
```

```
 [1] "TARGET"                      "CODE_GENDER"                
 [3] "AMT_CREDIT"                  "AMT_ANNUITY"                
 [5] "AMT_GOODS_PRICE"             "NAME_EDUCATION_TYPE"        
 [7] "DAYS_BIRTH"                  "DAYS_EMPLOYED"              
 [9] "DAYS_REGISTRATION"           "DAYS_ID_PUBLISH"            
[11] "REGION_RATING_CLIENT_W_CITY" "Annuity_Income_Ratio"       
[13] "Percent_Days_Employed"      
```

```
norm_values <- preProcess(train_df_rose[, -c(1)],
                          method = c("center",
                                     "scale"))
train_norm[, -c(1)] <- predict(norm_values,
                                train_df_rose[, -c(1)])

head(train_norm)
```

```
  TARGET CODE_GENDER  AMT_CREDIT AMT_ANNUITY AMT_GOODS_PRICE
1      0           F  0.99268089   1.5528871       1.0235252
2      0           F -0.55697501  -1.7304333      -0.8443363
3      0           F  1.72159500   0.8693993       1.2456352
4      0           M -1.62704136  -1.0453429      -0.6409312
5      0           F  0.02538308  -0.9634157      -0.7529268
6      0           M  0.79678441   0.2707055       0.3107206
            NAME_EDUCATION_TYPE  DAYS_BIRTH DAYS_EMPLOYED DAYS_REGISTRATION
1 Secondary / secondary special  0.07110103    -0.6251980         1.0871392
2              Incomplete higher  1.93744114    -1.4197362        -0.5301268
3                Lower secondary -1.64281333     1.8690287         0.7773543
4 Secondary / secondary special  1.00341927     0.6162391        -0.4667237
5 Secondary / secondary special  0.51491802    -0.2473539         0.8483135
6 Secondary / secondary special  1.67805126    -0.6616488         0.5031771
  DAYS_ID_PUBLISH REGION_RATING_CLIENT_W_CITY Annuity_Income_Ratio
1     -0.63250269                -1.919067483          -0.09484436
2      0.77819545                -0.330154347          -0.81335026
3     -1.80369065                 0.045849235           0.97555523
4     -0.04891247                -0.171642515          -1.10781415
5     -0.35311456                -0.009725689          -0.19377584
6      0.80158561                -1.541164503          -0.29677068
  Percent_Days_Employed
1            0.95414420
2           -0.13951258
3           -2.57417168
4            0.40890915
5            0.13264204
6            0.06014959
```

```
# Apply to validation set
valid_norm[, -c(1)] <- predict(norm_values,
                                valid_df[, -c(1)])
```

```r
head(valid_norm)
```

```
   TARGET CODE_GENDER  AMT_CREDIT AMT_ANNUITY AMT_GOODS_PRICE
2       0           F -0.02155441 -0.27565214      -0.1405352
3       1           M -0.69261860  0.52747947      -0.5902807
6       0           F  1.47015169  1.60607405       1.5400928
9       1           M -0.08979622 -0.02314442      -0.1760414
10      0           F -0.58520193 -0.67857714      -0.4719267
12      1           F -0.64423813 -0.66377704      -0.6494578
             NAME_EDUCATION_TYPE DAYS_BIRTH DAYS_EMPLOYED DAYS_REGISTRATION
2  Secondary / secondary special -0.9462974    -0.3838620        0.87385304
3  Secondary / secondary special  0.3392137    -0.3759077        0.69759733
6               Higher education  0.3477464    -0.3933178       -0.87083853
9  Secondary / secondary special -0.5631622    -0.3742573        1.19489975
10 Secondary / secondary special -0.2849158     2.0981828        0.07656926
12 Secondary / secondary special  0.6957105    -0.3898615        0.07310280
   DAYS_ID_PUBLISH REGION_RATING_CLIENT_W_CITY Annuity_Income_Ratio
2       -0.4122812                  -0.1450661           0.64784698
3        1.5503652                  -0.1450661           0.07502373
6       -1.0226850                  -1.9364656           0.18967428
9        0.9245682                   1.6463335          -0.51097511
10       1.4390598                  -0.1450661          -0.33670036
12      -0.3743900                  -0.1450661           0.63902826
   Percent_Days_Employed
2              0.3728429
3              0.3661824
6              0.3927217
9              0.3629550
10            -2.7257592
12             0.3911352
```

```r
# drop missing values
valid_norm <- drop_na(valid_norm)
```

```r
# Train kNN model using k = 5
knn_model <- caret::knn3(TARGET ~ ., data = train_norm, k = 5)
```

```r
# Prediction on training set
knn_pred_train <- predict(knn_model, newdata = train_norm[, -c(1)],
                          type = "class")
```

```r
head(knn_pred_train)
```

```
[1] 0 0 0 1 0 0
Levels: 0 1
```

```r
# Prediction on validation set
knn_pred_valid <- predict(knn_model, newdata = valid_norm[, -c(1)],
                          type = "class")
head(knn_pred_valid)
```

```
[1] 1 1 0 1 1 1
Levels: 0 1
```

```r
# Confusion matrix on training set
confusionMatrix(knn_pred_train, as.factor(train_norm[, 1]),
                positive = "1")
```

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 7120 2257
         1 3217 8383

               Accuracy : 0.739
                 95% CI : (0.733, 0.745)
    No Information Rate : 0.5072
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4773

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.7879
            Specificity : 0.6888
         Pos Pred Value : 0.7227
         Neg Pred Value : 0.7593
             Prevalence : 0.5072
         Detection Rate : 0.3996
```

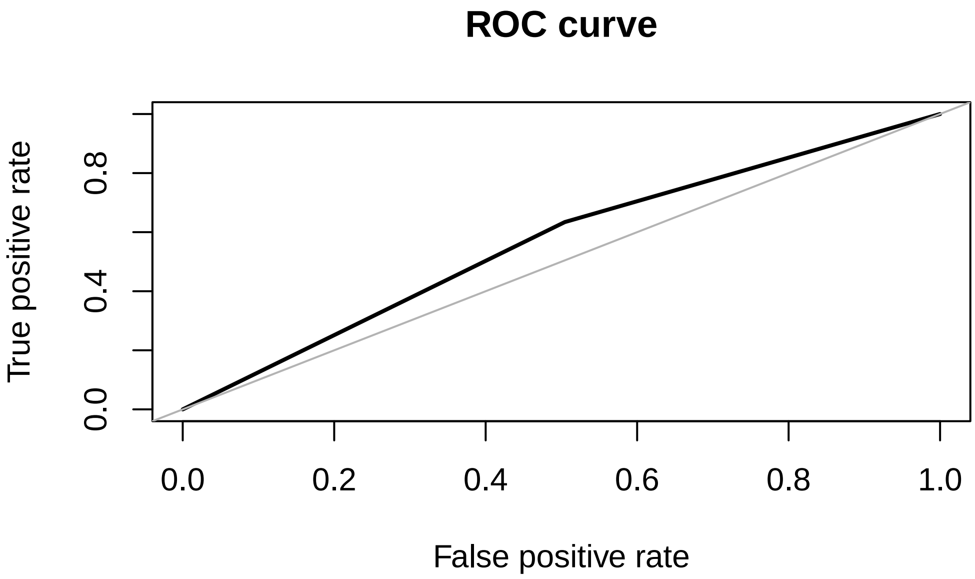
```
   Detection Prevalence : 0.5530
      Balanced Accuracy : 0.7383

       'Positive' Class : 1
```

```r
# Confusion matrix on validation set
confusionMatrix(knn_pred_valid, as.factor(valid_norm[, 1]),
                positive = "1")
```

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 3601  631
         1 3667 1094

               Accuracy : 0.5221
                 95% CI : (0.5117, 0.5324)
    No Information Rate : 0.8082
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0776

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

            Sensitivity : 0.6342
            Specificity : 0.4955
         Pos Pred Value : 0.2298
         Neg Pred Value : 0.8509
             Prevalence : 0.1918
         Detection Rate : 0.1217
   Detection Prevalence : 0.5294
      Balanced Accuracy : 0.5648

       'Positive' Class : 1
```

Model Evaluation

```r
library(ROSE)

ROSE::roc.curve(valid_norm$TARGET, knn_pred_valid)
```



```
Area under the curve (AUC): 0.565
```

Decision Tree Model

```r
# Load data for decision tree model and add new fields again
data1 <- read.csv("credit_fa2023_23.csv", header = TRUE)
data1$Income_Credit_Ratio <- NA

for (i in 1:nrow(data1)) {
  data1$Income_Credit_Ratio[i] <- data1$AMT_INCOME_TOTAL[i] / data1$AMT_CREDIT[i]
}

for (i in 1:nrow(data1)) {
  data1$Annuity_Income_Ratio[i] <- data1$AMT_ANNUITY[i] / data1$AMT_INCOME_TOTAL[i]
}

for (i in 1:nrow(data1)) {
  data1$Credit_As_Percentage[i] <- data1$AMT_CREDIT[i] / data1$AMT_INCOME_TOTAL[i]
}

for (i in 1:nrow(data1)) {
  data1$Percent_Days_Employed[i] <- data1$DAYS_EMPLOYED[i] / data1$DAYS_BIRTH[i]
}

for (i in 1:nrow(data1)) {
  data1$Income_Per_Person[i] <- data1$AMT_INCOME_TOTAL[i] / data1$CNT_FAM_MEMBERS[i]
}

#Remove XNA from CODE_GENDER variable
data1 <- data1[data1$CODE_GENDER != "XNA", ]
```

```r
#Save modified data into new data variable. Investigate data
data <- data1
names(data)
```

```
 [1] "X"                           "SK_ID_CURR"
 [3] "TARGET"                      "NAME_CONTRACT_TYPE"
 [5] "CODE_GENDER"                 "FLAG_OWN_CAR"
 [7] "FLAG_OWN_REALTY"             "CNT_CHILDREN"
 [9] "AMT_INCOME_TOTAL"            "AMT_CREDIT"
[11] "AMT_ANNUITY"                 "AMT_GOODS_PRICE"
[13] "NAME_TYPE_SUITE"             "NAME_INCOME_TYPE"
[15] "NAME_EDUCATION_TYPE"         "NAME_FAMILY_STATUS"
[17] "NAME_HOUSING_TYPE"           "DAYS_BIRTH"
[19] "DAYS_EMPLOYED"               "DAYS_REGISTRATION"
[21] "DAYS_ID_PUBLISH"             "OWN_CAR_AGE"
[23] "FLAG_MOBIL"                  "FLAG_EMP_PHONE"
[25] "FLAG_WORK_PHONE"             "FLAG_CONT_MOBILE"
[27] "FLAG_PHONE"                  "FLAG_EMAIL"
[29] "OCCUPATION_TYPE"             "CNT_FAM_MEMBERS"
[31] "REGION_RATING_CLIENT"        "REGION_RATING_CLIENT_W_CITY"
[33] "WEEKDAY_APPR_PROCESS_START"  "HOUR_APPR_PROCESS_START"
[35] "REG_REGION_NOT_LIVE_REGION"  "REG_REGION_NOT_WORK_REGION"
[37] "LIVE_REGION_NOT_WORK_REGION" "REG_CITY_NOT_LIVE_CITY"
[39] "REG_CITY_NOT_WORK_CITY"      "LIVE_CITY_NOT_WORK_CITY"
[41] "ORGANIZATION_TYPE"           "DAYS_LAST_PHONE_CHANGE"
[43] "FLAG_DOCUMENT_2"             "FLAG_DOCUMENT_3"
[45] "FLAG_DOCUMENT_4"             "FLAG_DOCUMENT_5"
```

```
[47] "FLAG_DOCUMENT_6"            "FLAG_DOCUMENT_7"
[49] "FLAG_DOCUMENT_8"            "FLAG_DOCUMENT_9"
[51] "FLAG_DOCUMENT_10"           "FLAG_DOCUMENT_11"
[53] "FLAG_DOCUMENT_12"           "FLAG_DOCUMENT_13"
[55] "FLAG_DOCUMENT_14"           "FLAG_DOCUMENT_15"
[57] "FLAG_DOCUMENT_16"           "FLAG_DOCUMENT_17"
[59] "FLAG_DOCUMENT_18"           "FLAG_DOCUMENT_19"
[61] "FLAG_DOCUMENT_20"           "FLAG_DOCUMENT_21"
[63] "AMT_REQ_CREDIT_BUREAU_HOUR" "AMT_REQ_CREDIT_BUREAU_DAY"
[65] "AMT_REQ_CREDIT_BUREAU_WEEK" "AMT_REQ_CREDIT_BUREAU_MON"
[67] "AMT_REQ_CREDIT_BUREAU_QRT"  "AMT_REQ_CREDIT_BUREAU_YEAR"
[69] "Income_Credit_Ratio"        "Annuity_Income_Ratio"
[71] "Credit_As_Percentage"       "Percent_Days_Employed"
[73] "Income_Per_Person"
```

```r
str(data)
```

```
'data.frame':   29999 obs. of  73 variables:
 $ X                         : int  300440 217645 70440 300551 86881 146804 263212 70439 11
 $ SK_ID_CURR                : int  448070 352176 181716 448195 200835 270206 404772 181715
 $ TARGET                    : int  0 0 1 0 1 0 0 1 1 0 ...
 $ NAME_CONTRACT_TYPE        : chr  "Cash loans" "Cash loans" "Cash loans" "Cash loans" ...
 $ CODE_GENDER               : chr  "F" "F" "M" "F" ...
 $ FLAG_OWN_CAR              : chr  "N" "N" "N" "N" ...
 $ FLAG_OWN_REALTY           : chr  "N" "Y" "Y" "Y" ...
 $ CNT_CHILDREN              : int  0 0 0 0 0 1 0 0 0 1 ...
 $ AMT_INCOME_TOTAL          : num  157500 90000 180000 171000 135000 ...
 $ AMT_CREDIT                : num  640080 573628 292500 757598 381528 ...
 $ AMT_ANNUITY               : num  29970 22878 34844 40491 25628 ...
 $ AMT_GOODS_PRICE           : num  450000 463500 292500 702000 315000 ...
 $ NAME_TYPE_SUITE           : chr  "Unaccompanied" "Children" "Unaccompanied" "Unaccompanie
 $ NAME_INCOME_TYPE          : chr  "Commercial associate" "Working" "Working" "State serva
 $ NAME_EDUCATION_TYPE       : chr  "Secondary / secondary special" "Secondary / secondary
 $ NAME_FAMILY_STATUS        : chr  "Separated" "Widow" "Civil marriage" "Married" ...
 $ NAME_HOUSING_TYPE         : chr  "With parents" "House / apartment" "House / apartment" "
 $ DAYS_BIRTH                : int  -10953 -20075 -13898 -21445 -10240 -13857 -21167 -17146
 $ DAYS_EMPLOYED             : int  -3005 -1715 -539 -4657 -921 -3113 -3320 -1509 -295 3652
 $ DAYS_REGISTRATION         : int  -5485 -1409 -2070 -3980 -1113 -7952 -2434 -8268 -205 -4
 $ DAYS_ID_PUBLISH           : int  -1284 -3573 -258 -4154 -123 -4604 -3253 -695 -1315 -446
 $ OWN_CAR_AGE               : int  NA NA NA NA 2 NA NA NA NA NA ...
 $ FLAG_MOBIL                : int  1 1 1 1 1 1 1 1 1 1 ...
 $ FLAG_EMP_PHONE             : int  1 1 1 1 1 1 1 1 1 0 ...
 $ FLAG_WORK_PHONE            : int  0 0 0 0 0 0 0 1 0 0 ...
 $ FLAG_CONT_MOBILE           : int  1 1 1 1 1 1 1 1 1 1 ...
 $ FLAG_PHONE                 : int  0 1 0 0 0 0 1 1 0 0 ...
 $ FLAG_EMAIL                 : int  0 0 0 0 0 0 0 0 0 0 ...
 $ OCCUPATION_TYPE            : chr  "" "Cooking staff" "Laborers" "Core staff" ...
 $ CNT_FAM_MEMBERS            : int  1 1 2 2 1 3 2 1 2 3 ...
 $ REGION_RATING_CLIENT       : int  2 2 2 2 3 1 2 3 3 2 ...
 $ REGION_RATING_CLIENT_W_CITY: int  2 2 2 2 3 1 2 3 3 2 ...
 $ WEEKDAY_APPR_PROCESS_START : chr  "SATURDAY" "FRIDAY" "TUESDAY" "SATURDAY" ...
 $ HOUR_APPR_PROCESS_START    : int  13 11 8 11 17 18 12 11 10 11 ...
 $ REG_REGION_NOT_LIVE_REGION : int  0 0 0 0 0 0 0 0 0 0 ...
 $ REG_REGION_NOT_WORK_REGION : int  0 0 1 0 0 0 0 0 0 0 ...
 $ LIVE_REGION_NOT_WORK_REGION: int  0 0 1 0 0 0 0 0 0 0 ...
 $ REG_CITY_NOT_LIVE_CITY     : int  1 0 0 0 1 0 0 0 0 0 ...
 $ REG_CITY_NOT_WORK_CITY     : int  1 0 0 0 1 0 0 0 1 0 ...
 $ LIVE_CITY_NOT_WORK_CITY    : int  0 0 0 0 0 0 0 0 1 0 ...
 $ ORGANIZATION_TYPE          : chr  "Self-employed" "Hotel" "Business Entity Type 1" "School
 $ DAYS_LAST_PHONE_CHANGE     : int  -2411 -1513 0 -2778 -20 -1827 -1471 -657 0 -796 ...
 $ FLAG_DOCUMENT_2            : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_3            : int  1 1 1 1 1 1 1 1 1 0 ...
 $ FLAG_DOCUMENT_4            : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_5            : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_6            : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_7            : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_8            : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_9            : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_10           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_11           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_12           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_13           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_14           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_15           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_16           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_17           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_18           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_19           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_20           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_21           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ AMT_REQ_CREDIT_BUREAU_HOUR : int  0 0 0 0 0 0 0 0 0 0 ...
 $ AMT_REQ_CREDIT_BUREAU_DAY  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ AMT_REQ_CREDIT_BUREAU_WEEK : int  0 0 0 0 0 0 0 0 0 0 ...
 $ AMT_REQ_CREDIT_BUREAU_MON  : int  0 0 1 0 0 0 0 0 0 0 ...
 $ AMT_REQ_CREDIT_BUREAU_QRT  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ AMT_REQ_CREDIT_BUREAU_YEAR : int  0 4 2 3 1 1 5 2 3 6 ...
 $ Income_Credit_Ratio        : num  0.246 0.157 0.615 0.226 0.354 ...
 $ Annuity_Income_Ratio       : num  0.19 0.254 0.194 0.237 0.19 ...
 $ Credit_As_Percentage       : num  4.06 6.37 1.62 4.43 2.83 ...
 $ Percent_Days_Employed      : num  0.2744 0.0854 0.0388 0.2172 0.0899 ...
 $ Income_Per_Person          : num  157500 90000 90000 85500 135000 ...
```

```
head(data)
```

```
       X SK_ID_CURR TARGET NAME_CONTRACT_TYPE CODE_GENDER FLAG_OWN_CAR
1 300440     448070      0         Cash loans           F            N
2 217645     352176      0         Cash loans           F            N
3  70440     181716      1         Cash loans           M            N
4 300551     448195      0         Cash loans           F            N
5  86881     200835      1         Cash loans           M            Y
6 146804     270206      0         Cash loans           F            N
  FLAG_OWN_REALTY CNT_CHILDREN AMT_INCOME_TOTAL AMT_CREDIT AMT_ANNUITY
1               N            0           157500   640080.0     29970.0
2               Y            0            90000   573628.5     22878.0
3               Y            0           180000   292500.0     34843.5
4               Y            0           171000   757597.5     40491.0
5               Y            0           135000   381528.0     25627.5
6               Y            1           247500  1198548.0     50913.0
  AMT_GOODS_PRICE NAME_TYPE_SUITE     NAME_INCOME_TYPE
1          450000   Unaccompanied Commercial associate
2          463500        Children              Working
3          292500   Unaccompanied              Working
4          702000   Unaccompanied        State servant
5          315000   Unaccompanied              Working
6         1102500   Unaccompanied Commercial associate
            NAME_EDUCATION_TYPE   NAME_FAMILY_STATUS NAME_HOUSING_TYPE
1 Secondary / secondary special            Separated      With parents
2 Secondary / secondary special                Widow House / apartment
3 Secondary / secondary special       Civil marriage House / apartment
4 Secondary / secondary special              Married House / apartment
5 Secondary / secondary special Single / not married House / apartment
6              Higher education       Civil marriage House / apartment
  DAYS_BIRTH DAYS_EMPLOYED DAYS_REGISTRATION DAYS_ID_PUBLISH OWN_CAR_AGE
1     -10953         -3005             -5485           -1284          NA
2     -20075         -1715             -1409           -3573          NA
```

```
3     -13898          -539             -2070            -258          NA
4     -21445         -4657             -3980           -4154          NA
5     -10240          -921             -1113            -123           2
6     -13857         -3113             -7952           -4604          NA
  FLAG_MOBIL FLAG_EMP_PHONE FLAG_WORK_PHONE FLAG_CONT_MOBILE FLAG_PHONE
1          1              1               0                1          0
2          1              1               0                1          1
3          1              1               0                1          0
4          1              1               0                1          0
5          1              1               0                1          0
6          1              1               0                1          0
  FLAG_EMAIL OCCUPATION_TYPE CNT_FAM_MEMBERS REGION_RATING_CLIENT
1          0                               1                    2
2          0   Cooking staff               1                    2
3          0        Laborers               2                    2
4          0      Core staff               2                    2
5          0                               1                    3
6          0                               3                    1
  REGION_RATING_CLIENT_W_CITY WEEKDAY_APPR_PROCESS_START
1                           2                   SATURDAY
2                           2                     FRIDAY
3                           2                    TUESDAY
4                           2                   SATURDAY
5                           3                  WEDNESDAY
6                           1                  WEDNESDAY
  HOUR_APPR_PROCESS_START REG_REGION_NOT_LIVE_REGION REG_REGION_NOT_WORK_REGION
1                      13                          0                          0
2                      11                          0                          0
3                       8                          0                          1
4                      11                          0                          0
5                      17                          0                          0
6                      18                          0                          0
  LIVE_REGION_NOT_WORK_REGION REG_CITY_NOT_LIVE_CITY REG_CITY_NOT_WORK_CITY
1                           0                      1                      1
2                           0                      0                      0
3                           1                      0                      0
4                           0                      0                      0
5                           0                      1                      1
6                           0                      0                      0
  LIVE_CITY_NOT_WORK_CITY      ORGANIZATION_TYPE DAYS_LAST_PHONE_CHANGE
1                       0          Self-employed                  -2411
2                       0                  Hotel                  -1513
3                       0 Business Entity Type 1                      0
```

```
4                       0                 School                  -2778
5                       0       Industry: type 4                    -20
6                       0 Business Entity Type 3                  -1827
  FLAG_DOCUMENT_2 FLAG_DOCUMENT_3 FLAG_DOCUMENT_4 FLAG_DOCUMENT_5
1               0               1               0               0
2               0               1               0               0
3               0               1               0               0
4               0               1               0               0
5               0               1               0               0
6               0               1               0               0
  FLAG_DOCUMENT_6 FLAG_DOCUMENT_7 FLAG_DOCUMENT_8 FLAG_DOCUMENT_9
1               0               0               0               0
2               0               0               0               0
3               0               0               0               0
4               0               0               0               0
5               0               0               0               0
6               0               0               0               0
  FLAG_DOCUMENT_10 FLAG_DOCUMENT_11 FLAG_DOCUMENT_12 FLAG_DOCUMENT_13
1                0                0                0                0
2                0                0                0                0
3                0                0                0                0
4                0                0                0                0
5                0                0                0                0
6                0                0                0                0
  FLAG_DOCUMENT_14 FLAG_DOCUMENT_15 FLAG_DOCUMENT_16 FLAG_DOCUMENT_17
1                0                0                0                0
2                0                0                0                0
3                0                0                0                0
4                0                0                0                0
5                0                0                0                0
6                0                0                0                0
  FLAG_DOCUMENT_18 FLAG_DOCUMENT_19 FLAG_DOCUMENT_20 FLAG_DOCUMENT_21
1                0                0                0                0
2                0                0                0                0
3                0                0                0                0
4                0                0                0                0
5                0                0                0                0
6                0                0                0                0
  AMT_REQ_CREDIT_BUREAU_HOUR AMT_REQ_CREDIT_BUREAU_DAY
1                          0                         0
2                          0                         0
3                          0                         0
4                          0                         0
```

```
5                          0                         0
6                          0                         0
  AMT_REQ_CREDIT_BUREAU_WEEK AMT_REQ_CREDIT_BUREAU_MON
1                          0                         0
2                          0                         0
3                          0                         1
4                          0                         0
5                          0                         0
6                          0                         0
  AMT_REQ_CREDIT_BUREAU_QRT AMT_REQ_CREDIT_BUREAU_YEAR Income_Credit_Ratio
1                         0                          0           0.2460630
2                         0                          4           0.1568960
3                         0                          2           0.6153846
4                         0                          3           0.2257135
5                         0                          1           0.3538403
6                         0                          1           0.2064999
  Annuity_Income_Ratio Credit_As_Percentage Percent_Days_Employed
1            0.1902857             4.064000            0.27435406
2            0.2542000             6.373650            0.08542964
3            0.1935750             1.625000            0.03878256
4            0.2367895             4.430395            0.21716018
5            0.1898333             2.826133            0.08994141
6            0.2057091             4.842618            0.22465180
  Income_Per_Person
1            157500
2             90000
3             90000
4             85500
5            135000
6             82500
```

```r
set.seed(666)
train_index <- sample(1:nrow(data), 0.7 * nrow(data))
valid_index <- setdiff(1:nrow(data), train_index)
train_df <- data[train_index, ]
valid_df <- data[valid_index, ]
nrow(train_df)
```

```
[1] 20999
```

```r
nrow(valid_df)
```

```
[1] 9000
```

```r
head(train_df)
```

```
           X SK_ID_CURR TARGET NAME_CONTRACT_TYPE CODE_GENDER FLAG_OWN_CAR
17983 235506     372786      1         Cash loans           F            N
12926  92468     207374      0         Cash loans           F            Y
13195 190897     321346      0         Cash loans           F            N
23676 103209     219794      0         Cash loans           M            N
15901 166080     292539      1         Cash loans           F            N
873   222231     357432      0         Cash loans           F            N
      FLAG_OWN_REALTY CNT_CHILDREN AMT_INCOME_TOTAL AMT_CREDIT AMT_ANNUITY
17983               Y            0           126000   467257.5     17743.5
12926               Y            0           193500   260640.0     31059.0
13195               N            2           337500   855882.0     36391.5
23676               Y            0           166500   266832.0     24601.5
15901               N            1            99000   360000.0     10102.5
873                 Y            1           157500   157500.0     14575.5
      AMT_GOODS_PRICE NAME_TYPE_SUITE     NAME_INCOME_TYPE
17983          328500   Unaccompanied Commercial associate
12926          225000          Family              Working
13195          765000   Unaccompanied Commercial associate
23676          238500          Family              Working
15901          360000   Unaccompanied              Working
873            157500   Unaccompanied              Working
                NAME_EDUCATION_TYPE   NAME_FAMILY_STATUS NAME_HOUSING_TYPE
17983              Higher education Single / not married House / apartment
12926 Secondary / secondary special Single / not married House / apartment
13195 Secondary / secondary special            Separated House / apartment
23676 Secondary / secondary special       Civil marriage House / apartment
15901              Higher education Single / not married House / apartment
873                Higher education              Married House / apartment
      DAYS_BIRTH DAYS_EMPLOYED DAYS_REGISTRATION DAYS_ID_PUBLISH OWN_CAR_AGE
17983     -10631          -200             -4795           -3303          NA
12926     -17424        -10763             -9653            -969          13
13195     -13370         -3493             -6795           -4908          NA
23676     -10606          -383             -5120           -3208          NA
15901     -14135         -1400             -7982           -3265          NA
873       -11931         -1154            -10268            -788          NA
      FLAG_MOBIL FLAG_EMP_PHONE FLAG_WORK_PHONE FLAG_CONT_MOBILE FLAG_PHONE
17983          1              1               0                1          0
```

```
12926          1              1               0                1          0
13195          1              1               0                1          0
23676          1              1               0                1          0
15901          1              1               0                1          1
873            1              1               0                1          1
      FLAG_EMAIL       OCCUPATION_TYPE CNT_FAM_MEMBERS REGION_RATING_CLIENT
17983          0           Sales staff               1                    2
12926          0              Laborers               1                    2
13195          0        Cleaning staff               3                    2
23676          1              Laborers               2                    2
15901          0 High skill tech staff               2                    2
873            0           Sales staff               3                    2
      REGION_RATING_CLIENT_W_CITY WEEKDAY_APPR_PROCESS_START
17983                           2                  WEDNESDAY
12926                           2                    TUESDAY
13195                           2                  WEDNESDAY
23676                           2                     SUNDAY
15901                           2                  WEDNESDAY
873                             2                    TUESDAY
      HOUR_APPR_PROCESS_START REG_REGION_NOT_LIVE_REGION
17983                      17                          0
12926                      16                          0
13195                      13                          0
23676                      12                          0
15901                      19                          0
873                        12                          0
      REG_REGION_NOT_WORK_REGION LIVE_REGION_NOT_WORK_REGION
17983                          0                           0
12926                          0                           0
13195                          0                           0
23676                          1                           1
15901                          0                           0
873                            0                           0
      REG_CITY_NOT_LIVE_CITY REG_CITY_NOT_WORK_CITY LIVE_CITY_NOT_WORK_CITY
17983                      1                      1                       0
12926                      0                      1                       1
13195                      0                      0                       0
23676                      0                      0                       0
15901                      0                      0                       0
873                        1                      1                       0
           ORGANIZATION_TYPE DAYS_LAST_PHONE_CHANGE FLAG_DOCUMENT_2
17983          Self-employed                      0               0
12926 Business Entity Type 2                  -1918               0
```

```
13195 Business Entity Type 3                  -1370               0
23676 Business Entity Type 3                   -282               0
15901       Industry: type 5                  -1498               0
873            Self-employed                  -1457               0
      FLAG_DOCUMENT_3 FLAG_DOCUMENT_4 FLAG_DOCUMENT_5 FLAG_DOCUMENT_6
17983               1               0               0               0
12926               1               0               0               0
13195               1               0               0               0
23676               1               0               0               0
15901               1               0               0               0
873                 1               0               0               0
      FLAG_DOCUMENT_7 FLAG_DOCUMENT_8 FLAG_DOCUMENT_9 FLAG_DOCUMENT_10
17983               0               0               0                0
12926               0               0               0                0
13195               0               0               0                0
23676               0               0               0                0
15901               0               0               0                0
873                 0               0               0                0
      FLAG_DOCUMENT_11 FLAG_DOCUMENT_12 FLAG_DOCUMENT_13 FLAG_DOCUMENT_14
17983                0                0                0                0
12926                0                0                0                0
13195                0                0                0                0
23676                0                0                0                0
15901                0                0                0                0
873                  0                0                0                0
      FLAG_DOCUMENT_15 FLAG_DOCUMENT_16 FLAG_DOCUMENT_17 FLAG_DOCUMENT_18
17983                0                0                0                0
12926                0                0                0                0
13195                0                0                0                0
23676                0                0                0                0
15901                0                0                0                0
873                  0                0                0                0
      FLAG_DOCUMENT_19 FLAG_DOCUMENT_20 FLAG_DOCUMENT_21
17983                0                0                0
12926                0                0                0
13195                0                0                0
23676                0                0                0
15901                0                0                0
873                  0                0                0
      AMT_REQ_CREDIT_BUREAU_HOUR AMT_REQ_CREDIT_BUREAU_DAY
17983                          0                         0
12926                          0                         0
13195                          0                         0
```

```
23676                          0                         0
15901                          0                         0
873                            0                         0
      AMT_REQ_CREDIT_BUREAU_WEEK AMT_REQ_CREDIT_BUREAU_MON
17983                          0                         0
12926                          0                         0
13195                          0                         0
23676                          0                         0
15901                          0                         0
873                            0                         1
      AMT_REQ_CREDIT_BUREAU_QRT AMT_REQ_CREDIT_BUREAU_YEAR Income_Credit_Ratio
17983                         0                          2           0.2696586
12926                         1                          4           0.7424033
13195                         0                          2           0.3943301
23676                         0                          6           0.6239881
15901                         2                          4           0.2750000
873                           0                          1           1.0000000
      Annuity_Income_Ratio Credit_As_Percentage Percent_Days_Employed
17983           0.14082143             3.708393            0.01881291
12926           0.16051163             1.346977            0.61771120
13195           0.10782667             2.535947            0.26125654
23676           0.14775676             1.602595            0.03611163
15901           0.10204545             3.636364            0.09904492
873             0.09254286             1.000000            0.09672282
      Income_Per_Person
17983            126000
12926            193500
13195            112500
23676             83250
15901             49500
873               52500
```

```
head(valid_df)
```

```
        X SK_ID_CURR TARGET NAME_CONTRACT_TYPE CODE_GENDER FLAG_OWN_CAR
2  217645     352176      0         Cash loans           F            N
3   70440     181716      1         Cash loans           M            N
6  146804     270206      0         Cash loans           F            N
9  114242     232484      1         Cash loans           M            N
10 251026     390460      0    Revolving loans           F            N
12 229721     366072      1         Cash loans           F            N
```

```
   FLAG_OWN_REALTY CNT_CHILDREN AMT_INCOME_TOTAL AMT_CREDIT AMT_ANNUITY
2                Y            0            90000   573628.5     22878.0
3                Y            0           180000   292500.0     34843.5
6                Y            1           247500  1198548.0     50913.0
9                N            0           202500   545040.0     26640.0
10               Y            1           112500   337500.0     16875.0
12               Y            3            67500   312768.0     17095.5
   AMT_GOODS_PRICE NAME_TYPE_SUITE     NAME_INCOME_TYPE
2           463500        Children              Working
3           292500   Unaccompanied              Working
6          1102500   Unaccompanied Commercial associate
9           450000   Unaccompanied              Working
10          337500   Unaccompanied            Pensioner
12          270000   Unaccompanied        State servant
             NAME_EDUCATION_TYPE NAME_FAMILY_STATUS NAME_HOUSING_TYPE
2  Secondary / secondary special              Widow House / apartment
3  Secondary / secondary special     Civil marriage House / apartment
6               Higher education     Civil marriage House / apartment
9  Secondary / secondary special     Civil marriage House / apartment
10 Secondary / secondary special     Civil marriage House / apartment
12 Secondary / secondary special     Civil marriage      With parents
   DAYS_BIRTH DAYS_EMPLOYED DAYS_REGISTRATION DAYS_ID_PUBLISH OWN_CAR_AGE
2      -20075         -1715             -1409           -3573          NA
3      -13898          -539             -2070            -258          NA
6      -13857         -3113             -7952           -4604          NA
9      -18234          -295              -205           -1315          NA
10     -16897        365243             -4399            -446          NA
12     -12185         -2602             -4412           -3509          NA
   FLAG_MOBIL FLAG_EMP_PHONE FLAG_WORK_PHONE FLAG_CONT_MOBILE FLAG_PHONE
2           1              1               0                1          1
3           1              1               0                1          0
6           1              1               0                1          0
9           1              1               0                1          0
10          1              0               0                1          0
12          1              1               1                1          0
   FLAG_EMAIL OCCUPATION_TYPE CNT_FAM_MEMBERS REGION_RATING_CLIENT
2           0   Cooking staff               1                    2
3           0        Laborers               2                    2
6           0                               3                    1
9           0        Laborers               2                    3
10          0                               3                    2
12          0      Core staff               5                    2
   REGION_RATING_CLIENT_W_CITY WEEKDAY_APPR_PROCESS_START
```

```
2                            2                     FRIDAY
3                            2                    TUESDAY
6                            1                  WEDNESDAY
9                            3                   THURSDAY
10                           2                   THURSDAY
12                           2                   SATURDAY
   HOUR_APPR_PROCESS_START REG_REGION_NOT_LIVE_REGION
2                       11                          0
3                        8                          0
6                       18                          0
9                       10                          0
10                      11                          0
12                       9                          0
   REG_REGION_NOT_WORK_REGION LIVE_REGION_NOT_WORK_REGION
2                           0                           0
3                           1                           1
6                           0                           0
9                           0                           0
10                          0                           0
12                          0                           0
   REG_CITY_NOT_LIVE_CITY REG_CITY_NOT_WORK_CITY LIVE_CITY_NOT_WORK_CITY
2                       0                      0                       0
3                       0                      0                       0
6                       0                      0                       0
9                       0                      1                       1
10                      0                      0                       0
12                      0                      0                       0
        ORGANIZATION_TYPE DAYS_LAST_PHONE_CHANGE FLAG_DOCUMENT_2
2                   Hotel                  -1513               0
3  Business Entity Type 1                      0               0
6  Business Entity Type 3                  -1827               0
9  Business Entity Type 3                      0               0
10                    XNA                   -796               0
12           Kindergarten                  -1079               0
   FLAG_DOCUMENT_3 FLAG_DOCUMENT_4 FLAG_DOCUMENT_5 FLAG_DOCUMENT_6
2                1               0               0               0
3                1               0               0               0
6                1               0               0               0
9                1               0               0               0
10               0               0               0               0
12               1               0               0               0
   FLAG_DOCUMENT_7 FLAG_DOCUMENT_8 FLAG_DOCUMENT_9 FLAG_DOCUMENT_10
2                0               0               0                0
```

```
3                0               0               0                0
6                0               0               0                0
9                0               0               0                0
10               0               0               0                0
12               0               0               0                0
   FLAG_DOCUMENT_11 FLAG_DOCUMENT_12 FLAG_DOCUMENT_13 FLAG_DOCUMENT_14
2                 0                0                0                0
3                 0                0                0                0
6                 0                0                0                0
9                 0                0                0                0
10                0                0                0                0
12                0                0                0                0
   FLAG_DOCUMENT_15 FLAG_DOCUMENT_16 FLAG_DOCUMENT_17 FLAG_DOCUMENT_18
2                 0                0                0                0
3                 0                0                0                0
6                 0                0                0                0
9                 0                0                0                0
10                0                0                0                0
12                0                0                0                0
   FLAG_DOCUMENT_19 FLAG_DOCUMENT_20 FLAG_DOCUMENT_21
2                 0                0                0
3                 0                0                0
6                 0                0                0
9                 0                0                0
10                0                0                0
12                0                0                0
   AMT_REQ_CREDIT_BUREAU_HOUR AMT_REQ_CREDIT_BUREAU_DAY
2                           0                         0
3                           0                         0
6                           0                         0
9                           0                         0
10                          0                         0
12                          0                         0
   AMT_REQ_CREDIT_BUREAU_WEEK AMT_REQ_CREDIT_BUREAU_MON
2                           0                         0
3                           0                         1
6                           0                         0
9                           0                         0
10                          0                         0
12                          0                         0
   AMT_REQ_CREDIT_BUREAU_QRT AMT_REQ_CREDIT_BUREAU_YEAR Income_Credit_Ratio
2                          0                          4           0.1568960
3                          0                          2           0.6153846
```

```
6                          0                          1           0.2064999
9                          0                          3           0.3715324
10                         0                          6           0.3333333
12                         0                          1           0.2158149
   Annuity_Income_Ratio Credit_As_Percentage Percent_Days_Employed
2             0.2542000             6.373650            0.08542964
3             0.1935750             1.625000            0.03878256
6             0.2057091             4.842618            0.22465180
9             0.1315556             2.691556            0.01617857
10            0.1500000             3.000000          -21.61584897
12            0.2532667             4.633600            0.21354124
   Income_Per_Person
2              90000
3              90000
6              82500
9             101250
10             37500
12             13500
```

```
str(train_df)
```

```
'data.frame':   20999 obs. of  73 variables:
 $ X                          : int  235506 92468 190897 103209 166080 222231 30005 194123 18
 $ SK_ID_CURR                 : int  372786 207374 321346 219794 292539 357432 134832 325093
 $ TARGET                     : int  1 0 0 0 1 0 0 1 0 0 ...
 $ NAME_CONTRACT_TYPE         : chr  "Cash loans" "Cash loans" "Cash loans" "Cash loans" ...
 $ CODE_GENDER                : chr  "F" "F" "F" "M" ...
 $ FLAG_OWN_CAR               : chr  "N" "Y" "N" "N" ...
 $ FLAG_OWN_REALTY            : chr  "Y" "Y" "N" "Y" ...
 $ CNT_CHILDREN               : int  0 0 2 0 1 1 0 0 0 0 ...
 $ AMT_INCOME_TOTAL           : num  126000 193500 337500 166500 99000 ...
 $ AMT_CREDIT                 : num  467258 260640 855882 266832 360000 ...
 $ AMT_ANNUITY                : num  17744 31059 36392 24602 10102 ...
 $ AMT_GOODS_PRICE            : num  328500 225000 765000 238500 360000 ...
 $ NAME_TYPE_SUITE            : chr  "Unaccompanied" "Family" "Unaccompanied" "Family" ...
 $ NAME_INCOME_TYPE           : chr  "Commercial associate" "Working" "Commercial associate"
 $ NAME_EDUCATION_TYPE        : chr  "Higher education" "Secondary / secondary special" "Seco
 $ NAME_FAMILY_STATUS         : chr  "Single / not married" "Single / not married" "Separated
 $ NAME_HOUSING_TYPE          : chr  "House / apartment" "House / apartment" "House / apartme
 $ DAYS_BIRTH                 : int  -10631 -17424 -13370 -10606 -14135 -11931 -18336 -12296
 $ DAYS_EMPLOYED              : int  -200 -10763 -3493 -383 -1400 -1154 -6561 -1212 -2285 -66
 $ DAYS_REGISTRATION          : int  -4795 -9653 -6795 -5120 -7982 -10268 -1038 -2417 -3801 -
 $ DAYS_ID_PUBLISH            : int  -3303 -969 -4908 -3208 -3265 -788 -1898 -2422 -4156 -47
 $ OWN_CAR_AGE                : int  NA 13 NA NA NA NA NA NA 15 NA ...
 $ FLAG_MOBIL                 : int  1 1 1 1 1 1 1 1 1 1 ...
 $ FLAG_EMP_PHONE             : int  1 1 1 1 1 1 1 1 1 1 ...
 $ FLAG_WORK_PHONE            : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_CONT_MOBILE           : int  1 1 1 1 1 1 1 1 1 1 ...
 $ FLAG_PHONE                 : int  0 0 0 0 1 1 0 0 0 0 ...
 $ FLAG_EMAIL                 : int  0 0 0 1 0 0 0 0 0 0 ...
 $ OCCUPATION_TYPE            : chr  "Sales staff" "Laborers" "Cleaning staff" "Laborers" ..
 $ CNT_FAM_MEMBERS            : int  1 1 3 2 2 3 2 1 1 2 ...
 $ REGION_RATING_CLIENT       : int  2 2 2 2 2 2 2 2 2 2 ...
 $ REGION_RATING_CLIENT_W_CITY: int  2 2 2 2 2 2 2 2 2 2 ...
 $ WEEKDAY_APPR_PROCESS_START : chr  "WEDNESDAY" "TUESDAY" "WEDNESDAY" "SUNDAY" ...
 $ HOUR_APPR_PROCESS_START    : int  17 16 13 12 19 12 14 19 10 12 ...
 $ REG_REGION_NOT_LIVE_REGION : int  0 0 0 0 0 0 0 0 0 0 ...
 $ REG_REGION_NOT_WORK_REGION : int  0 0 0 1 0 0 0 0 0 0 ...
 $ LIVE_REGION_NOT_WORK_REGION: int  0 0 0 1 0 0 0 0 0 0 ...
 $ REG_CITY_NOT_LIVE_CITY     : int  1 0 0 0 0 1 0 0 0 0 ...
 $ REG_CITY_NOT_WORK_CITY     : int  1 1 0 0 0 1 0 1 0 1 ...
 $ LIVE_CITY_NOT_WORK_CITY    : int  0 1 0 0 0 0 0 1 0 1 ...
 $ ORGANIZATION_TYPE          : chr  "Self-employed" "Business Entity Type 2" "Business Entit
 $ DAYS_LAST_PHONE_CHANGE     : int  0 -1918 -1370 -282 -1498 -1457 -108 -394 -740 -1555 ...
 $ FLAG_DOCUMENT_2            : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_3            : int  1 1 1 1 1 1 1 1 1 1 ...
 $ FLAG_DOCUMENT_4            : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_5            : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_6            : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_7            : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_8            : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_9            : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_10           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_11           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_12           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_13           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_14           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_15           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_16           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_17           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_18           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_19           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_20           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_21           : int  0 0 0 0 0 0 0 0 0 0 ...
```

```
 $ AMT_REQ_CREDIT_BUREAU_HOUR : int  0 0 0 0 0 0 0 0 0 0 ...
 $ AMT_REQ_CREDIT_BUREAU_DAY  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ AMT_REQ_CREDIT_BUREAU_WEEK : int  0 0 0 0 0 0 0 0 0 0 ...
 $ AMT_REQ_CREDIT_BUREAU_MON  : int  0 0 0 0 0 1 0 0 0 1 ...
 $ AMT_REQ_CREDIT_BUREAU_QRT  : int  0 1 0 0 2 0 0 0 0 0 ...
 $ AMT_REQ_CREDIT_BUREAU_YEAR : int  2 4 2 6 4 1 3 7 0 5 ...
 $ Income_Credit_Ratio        : num  0.27 0.742 0.394 0.624 0.275 ...
 $ Annuity_Income_Ratio       : num  0.141 0.161 0.108 0.148 0.102 ...
 $ Credit_As_Percentage       : num  3.71 1.35 2.54 1.6 3.64 ...
 $ Percent_Days_Employed      : num  0.0188 0.6177 0.2613 0.0361 0.099 ...
 $ Income_Per_Person          : num  126000 193500 112500 83250 49500 ...
```

```
str(valid_df)
```

```
'data.frame':   9000 obs. of  73 variables:
 $ X                          : int  217645 70440 146804 114242 251026 229721 85195 283368 14
 $ SK_ID_CURR                 : int  352176 181716 270206 232484 390460 366072 198841 428168
 $ TARGET                     : int  0 1 0 1 0 1 0 1 0 0 ...
 $ NAME_CONTRACT_TYPE         : chr  "Cash loans" "Cash loans" "Cash loans" "Cash loans" ...
 $ CODE_GENDER                : chr  "F" "M" "F" "M" ...
 $ FLAG_OWN_CAR               : chr  "N" "N" "N" "N" ...
 $ FLAG_OWN_REALTY            : chr  "Y" "Y" "Y" "N" ...
 $ CNT_CHILDREN               : int  0 0 1 0 1 3 0 0 0 1 ...
 $ AMT_INCOME_TOTAL           : num  90000 180000 247500 202500 112500 ...
 $ AMT_CREDIT                 : num  573628 292500 1198548 545040 337500 ...
 $ AMT_ANNUITY                : num  22878 34844 50913 26640 16875 ...
 $ AMT_GOODS_PRICE            : num  463500 292500 1102500 450000 337500 ...
 $ NAME_TYPE_SUITE            : chr  "Children" "Unaccompanied" "Unaccompanied" "Unaccompanie
 $ NAME_INCOME_TYPE           : chr  "Working" "Working" "Commercial associate" "Working" ..
 $ NAME_EDUCATION_TYPE        : chr  "Secondary / secondary special" "Secondary / secondary s
 $ NAME_FAMILY_STATUS         : chr  "Widow" "Civil marriage" "Civil marriage" "Civil marriag
 $ NAME_HOUSING_TYPE          : chr  "House / apartment" "House / apartment" "House / apartme
 $ DAYS_BIRTH                 : int  -20075 -13898 -13857 -18234 -16897 -12185 -10579 -11938
 $ DAYS_EMPLOYED              : int  -1715 -539 -3113 -295 365243 -2602 -246 -680 365243 -156
 $ DAYS_REGISTRATION          : int  -1409 -2070 -7952 -205 -4399 -4412 -4665 -5949 -4764 -47
 $ DAYS_ID_PUBLISH            : int  -3573 -258 -4604 -1315 -446 -3509 -3226 -4153 -82 -2312
 $ OWN_CAR_AGE                : int  NA NA NA NA NA NA NA NA NA 8 ...
 $ FLAG_MOBIL                 : int  1 1 1 1 1 1 1 1 1 1 ...
 $ FLAG_EMP_PHONE             : int  1 1 1 1 0 1 1 1 0 1 ...
 $ FLAG_WORK_PHONE            : int  0 0 0 0 0 1 0 0 0 1 ...
 $ FLAG_CONT_MOBILE           : int  1 1 1 1 1 1 1 1 1 1 ...
 $ FLAG_PHONE                 : int  1 0 0 0 0 0 1 0 0 1 ...
 $ FLAG_EMAIL                 : int  0 0 0 0 0 0 0 0 0 0 ...
 $ OCCUPATION_TYPE            : chr  "Cooking staff" "Laborers" "" "Laborers" ...
 $ CNT_FAM_MEMBERS            : int  1 2 3 2 3 5 1 1 2 3 ...
 $ REGION_RATING_CLIENT       : int  2 2 1 3 2 2 2 2 2 2 ...
 $ REGION_RATING_CLIENT_W_CITY: int  2 2 1 3 2 2 2 2 2 2 ...
 $ WEEKDAY_APPR_PROCESS_START : chr  "FRIDAY" "TUESDAY" "WEDNESDAY" "THURSDAY" ...
 $ HOUR_APPR_PROCESS_START    : int  11 8 18 10 11 9 14 8 9 11 ...
 $ REG_REGION_NOT_LIVE_REGION : int  0 0 0 0 0 0 0 0 0 0 ...
 $ REG_REGION_NOT_WORK_REGION : int  0 1 0 0 0 0 0 0 0 0 ...
 $ LIVE_REGION_NOT_WORK_REGION: int  0 1 0 0 0 0 0 0 0 0 ...
 $ REG_CITY_NOT_LIVE_CITY     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ REG_CITY_NOT_WORK_CITY     : int  0 0 0 1 0 0 1 0 0 1 ...
 $ LIVE_CITY_NOT_WORK_CITY    : int  0 0 0 1 0 0 1 0 0 1 ...
 $ ORGANIZATION_TYPE          : chr  "Hotel" "Business Entity Type 1" "Business Entity Type
 $ DAYS_LAST_PHONE_CHANGE     : int  -1513 0 -1827 0 -796 -1079 -558 -2209 0 -1105 ...
 $ FLAG_DOCUMENT_2            : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_3            : int  1 1 1 1 0 1 1 1 1 1 ...
 $ FLAG_DOCUMENT_4            : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_5            : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_6            : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_7            : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_8            : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_9            : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_10           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_11           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_12           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_13           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_14           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_15           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_16           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_17           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_18           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_19           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_20           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FLAG_DOCUMENT_21           : int  0 0 0 0 0 0 0 0 0 0 ...
 $ AMT_REQ_CREDIT_BUREAU_HOUR : int  0 0 0 0 0 0 0 0 0 NA ...
 $ AMT_REQ_CREDIT_BUREAU_DAY  : int  0 0 0 0 0 0 0 0 0 NA ...
 $ AMT_REQ_CREDIT_BUREAU_WEEK : int  0 0 0 0 0 0 0 0 1 NA ...
 $ AMT_REQ_CREDIT_BUREAU_MON  : int  0 1 0 0 0 0 0 0 0 NA ...
 $ AMT_REQ_CREDIT_BUREAU_QRT  : int  0 0 0 0 0 0 0 0 0 NA ...
 $ AMT_REQ_CREDIT_BUREAU_YEAR : int  4 2 1 3 6 1 0 4 0 NA ...
 $ Income_Credit_Ratio        : num  0.157 0.615 0.206 0.372 0.333 ...
```

```
 $ Annuity_Income_Ratio        : num  0.254 0.194 0.206 0.132 0.15 ...
 $ Credit_As_Percentage        : num  6.37 1.62 4.84 2.69 3 ...
 $ Percent_Days_Employed       : num  0.0854 0.0388 0.2247 0.0162 -21.6158 ...
 $ Income_Per_Person           : num  90000 90000 82500 101250 37500 ...
```

```r
table(train_df$CODE_GENDER)
```

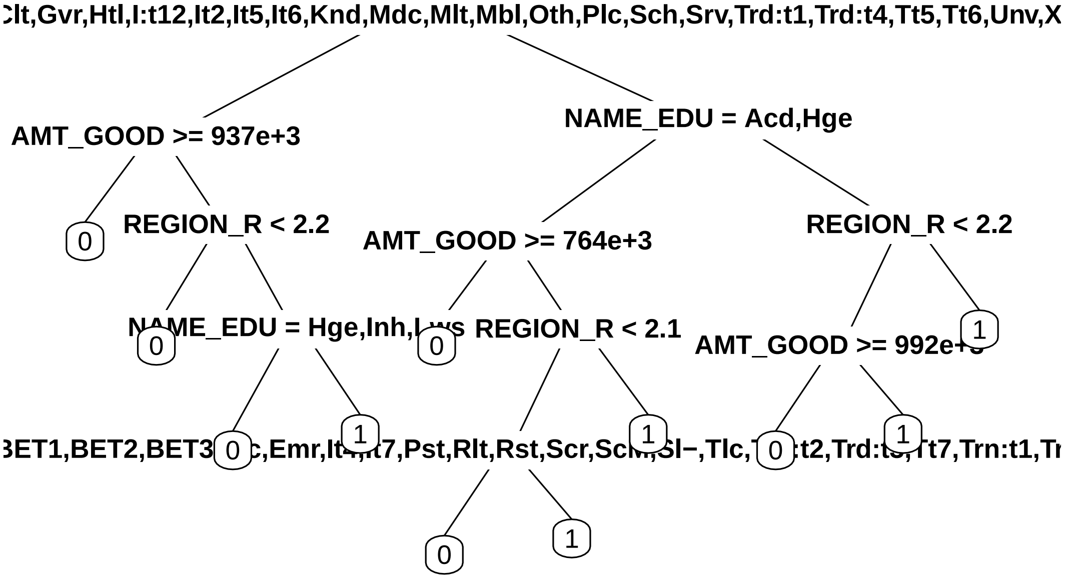
```
    F     M
13627  7372
```

```r
#Variable list
#Income_Credit_Ratio + Annuity_Income_Ratio + Credit_As_Percentage + Percent_Days_Employed
```

```r
#Convert all categorical variables into factors
train_df$OCCUPATION_TYPE <- as.factor(train_df$OCCUPATION_TYPE)
train_df$ORGANIZATION_TYPE <- as.factor(train_df$ORGANIZATION_TYPE)
train_df$NAME_EDUCATION_TYPE <- as.factor(train_df$NAME_EDUCATION_TYPE)
train_df <- train_df[train_df$CODE_GENDER != "XNA", ]


#Use ROSE to oversample target variable in order to balance model
train_df_rose <- ROSE(TARGET ~ Percent_Days_Employed + NAME_EDUCATION_TYPE + REGION_RATING
                      data = train_df, seed = 666)$data

table(train_df_rose$TARGET)
```

```
    0     1
10337 10640
```

```r
#Create classification decision tree with relevant fields
class_tr_cl <- rpart(TARGET ~ Percent_Days_Employed + NAME_EDUCATION_TYPE + REGION_RATING_
                    data = train_df_rose, method = "class", control = rpart.control(cp = 0

prp(class_tr_cl, cex = 0.8, tweak = 1)
```



```r
#Apply classificiation decision tree to training set and validation set
class_tr_train_predict <- predict(class_tr_cl, train_df_rose,
                                  type = "class")
class_tr_valid <- predict(class_tr_cl, valid_df,
                                  type = "class")
```

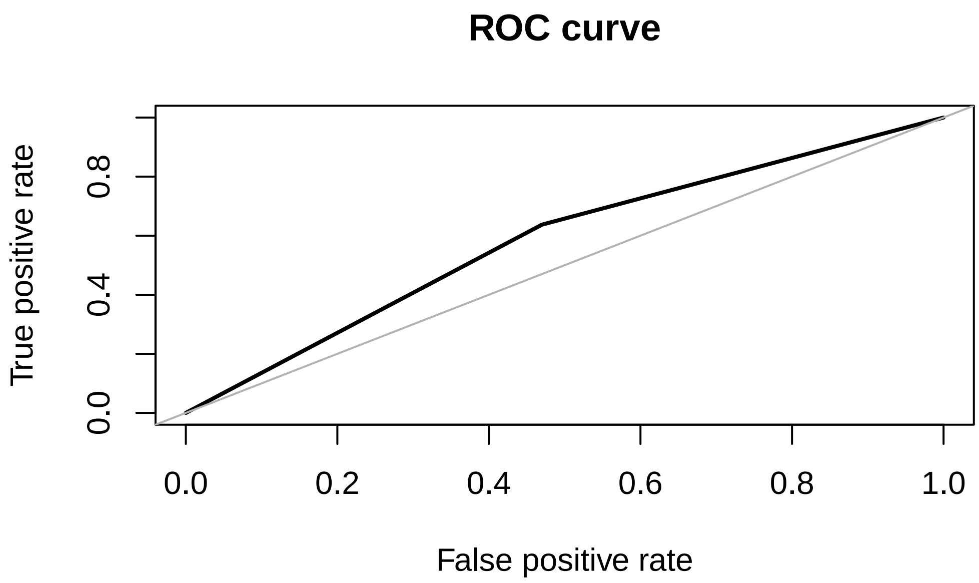
```r
#Check confusion matrix of training and validation model to determine model's accuracy and
train_df_rose$TARGET <- as.factor(train_df_rose$TARGET)
valid_df$TARGET <- as.factor(valid_df$TARGET)
confusionMatrix(class_tr_train_predict, train_df_rose$TARGET, positive = "1")
```

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 5443 3386
         1 4894 7254

               Accuracy : 0.6053
                 95% CI : (0.5986, 0.6119)
    No Information Rate : 0.5072
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.2088

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.6818
            Specificity : 0.5266
         Pos Pred Value : 0.5971
         Neg Pred Value : 0.6165
             Prevalence : 0.5072
         Detection Rate : 0.3458
   Detection Prevalence : 0.5791
      Balanced Accuracy : 0.6042

       'Positive' Class : 1
```

```r
confusionMatrix(class_tr_valid, valid_df$TARGET, positive = "1")
```

```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 3853  626
         1 3420 1101

               Accuracy : 0.5504
                 95% CI : (0.5401, 0.5608)
    No Information Rate : 0.8081
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1035

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

            Sensitivity : 0.6375
            Specificity : 0.5298
         Pos Pred Value : 0.2435
         Neg Pred Value : 0.8602
             Prevalence : 0.1919
         Detection Rate : 0.1223
   Detection Prevalence : 0.5023
      Balanced Accuracy : 0.5836

       'Positive' Class : 1
```

```
library(ROSE)
ROSE::roc.curve(valid_df$TARGET, class_tr_valid)
```



```
Area under the curve (AUC): 0.584
```

```
# Load new customer data
test <- read.csv("credit_test_fa2023_23.csv", header = TRUE)
#Add percent_days_employed field
for (i in 1:nrow(test)) {
  test$Percent_Days_Employed[i] <- test$DAYS_EMPLOYED[i] / test$DAYS_BIRTH[i]
}

# Predict risk of new customers
```

```
credit_prediction <- predict(class_tr_cl, newdata = test,
                               type = "class")
credit_prediction
```

```
1 2 3 4 5
0 0 1 1 0
Levels: 0 1
```

Write-up:

Problem Description:

Stark Enterprises decided it wanted to branch out into the financial industry. They want to create a model to assist them in this endeavor.

Objective: Create a model to predict which customers for a loan are likely to be high risk.

Data description: The data includes the characteristics and financial situation of our customers. This includes fields which are stricly personal, like their gender and education, and fields which are finance-related, like their income and loan annuity

Data Modifications: percent_days_employed was created by dividing Days employed by customer's age. The target variable for the training set was also over-sampled to raise sensitivity.

Both models were tailored to sensitivity. As a company which is branching into a new industry, it may be more important for them to catch all high risk customers to prevent losses. We chose to select the decision tree model because it was better at distinguishing risk (58.4% AUC to 56.6% AUC). Our chosen model correctly identifies high risk candidates 63.75% of the time. While this metric is high, the model pays for it with a lower accuracy of 55.04% and a low pos pred value of 24.35%. What this means is that this model has a tendency to incorrectly classify low risk candidates as high risk candidates. While this is regrettable, improving this metric would necessitate a reduction in sensitivity, which would be dangerous as more high risk customers would slip under the radar. Looking at it another way, 86.02% of low risk predictions are accurate, meaning that ~14% of low risk candidates may be unfairly rejected for a loan due to this model. The question of the efficacy of this model therefore lies in what the revenue loss will be from 14% of low risk candidates rejected vs. the potential revenue loss from giving loans to more high-risk candidates. Ultimately, as Stark Enterprises becomes more established, we believe the model could be tweaked to have less false positives, at the expense of false negatives, in order to give out more loans.

```
 Accuracy : 0.5504
                95% CI : (0.5401, 0.5608)
    No Information Rate : 0.8081
```

```
   P-Value [Acc > NIR] : 1

                 Kappa : 0.1035

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

           Sensitivity : 0.6375
           Specificity : 0.5298
        Pos Pred Value : 0.2435
        Neg Pred Value : 0.8602
            Prevalence : 0.1919
        Detection Rate : 0.1223
  Detection Prevalence : 0.5023
     Balanced Accuracy : 0.5836
```