Untitled2

Meenakshi Hariharan

2024-12-11

```
pacman::p_load(tidyverse,skimr,janitor,knitr,caret,rminer,mice,dbscan,tictoc,dplyr,ranger,psych, tidymod

library(xgboost)
```

```
##
## Attaching package: 'xgboost'

## The following object is masked from 'package:dplyr':
##
##     slice
```

```
library(Matrix)
```

```
##
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
library(pROC)   # to plot ROC-AUC
```

```
## Type 'citation("pROC")' for a citation.

##
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
library(e1071)  # for sum
```

```
##
## Attaching package: 'e1071'

## The following object is masked from 'package:tune':
##
##     tune
```

```
## The following object is masked from 'package:rsample':
##
##     permutations

## The following object is masked from 'package:parsnip':
##
##     tune
```

```
library(GGally) # to plot with ggpairs()
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(doParallel)  # for training xgboost using parallel processing
```

```
## Warning: package 'doParallel' was built under R version 4.4.2

## Loading required package: foreach

##
## Attaching package: 'foreach'

## The following objects are masked from 'package:purrr':
##
##     accumulate, when

## Loading required package: iterators

## Loading required package: parallel
```

```
library(MLmetrics)
```

```
## Warning: package 'MLmetrics' was built under R version 4.4.2

##
## Attaching package: 'MLmetrics'

## The following object is masked from 'package:psych':
##
##     AUC

## The following objects are masked from 'package:caret':
##
##     MAE, RMSE

## The following object is masked from 'package:base':
##
##     Recall
```

```r
library(readxl)
library(themis)
```

```
## Warning: package 'themis' was built under R version 4.4.2
```

```r
library(rpart)
```

```
##
## Attaching package: 'rpart'

## The following object is masked from 'package:dials':
##
##     prune
```

```r
library(rpart.plot)
```

```
## Warning: package 'rpart.plot' was built under R version 4.4.2
```

Loading Data and Clean up

```r
# load cleaned application_train and application_test data from EDA HW (with removed columns, NA's, out

train_clean <- read_csv("train_clean.csv")
```

```
## Rows: 237776 Columns: 64
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (12): NAME_CONTRACT_TYPE, CODE_GENDER, FLAG_OWN_CAR, FLAG_OWN_REALTY, NA...
## dbl (52): SK_ID_CURR, TARGET, CNT_CHILDREN, AMT_INCOME_TOTAL, AMT_CREDIT, AM...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
test_clean <- read_csv("test_clean.csv")
```

```
## Rows: 37740 Columns: 63
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (12): NAME_CONTRACT_TYPE, CODE_GENDER, FLAG_OWN_CAR, FLAG_OWN_REALTY, NA...
## dbl (51): SK_ID_CURR, CNT_CHILDREN, AMT_INCOME_TOTAL, AMT_CREDIT, AMT_ANNUIT...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
head(train_clean)  #first 6 rows
```

```
## # A tibble: 6 x 64
##   SK_ID_CURR TARGET NAME_CONTRACT_TYPE CODE_GENDER FLAG_OWN_CAR FLAG_OWN_REALTY
##        <dbl>  <dbl> <chr>              <chr>       <chr>        <chr>
## 1     100002      1 Cash loans         M           N            Y
## 2     100003      0 Cash loans         F           N            N
## 3     100006      0 Cash loans         F           N            Y
## 4     100007      0 Cash loans         M           N            Y
## 5     100008      0 Cash loans         M           N            Y
## 6     100011      0 Cash loans         F           N            Y
## # i 58 more variables: CNT_CHILDREN <dbl>, AMT_INCOME_TOTAL <dbl>,
## #   AMT_CREDIT <dbl>, AMT_ANNUITY <dbl>, AMT_GOODS_PRICE <dbl>,
## #   NAME_TYPE_SUITE <chr>, NAME_INCOME_TYPE <chr>, NAME_EDUCATION_TYPE <chr>,
## #   NAME_FAMILY_STATUS <chr>, NAME_HOUSING_TYPE <chr>,
## #   REGION_POPULATION_RELATIVE <dbl>, DAYS_BIRTH <dbl>,
## #   DAYS_REGISTRATION <dbl>, DAYS_ID_PUBLISH <dbl>, OWN_CAR_AGE <dbl>,
## #   FLAG_EMP_PHONE <dbl>, FLAG_WORK_PHONE <dbl>, FLAG_PHONE <dbl>, ...
```

```
dim(train_clean)  #shape of data
```

```
## [1] 237776     64
```

```
colnames(train_clean)  #column names
```

```
##  [1] "SK_ID_CURR"                  "TARGET"
##  [3] "NAME_CONTRACT_TYPE"          "CODE_GENDER"
##  [5] "FLAG_OWN_CAR"                "FLAG_OWN_REALTY"
##  [7] "CNT_CHILDREN"                "AMT_INCOME_TOTAL"
##  [9] "AMT_CREDIT"                  "AMT_ANNUITY"
## [11] "AMT_GOODS_PRICE"             "NAME_TYPE_SUITE"
## [13] "NAME_INCOME_TYPE"            "NAME_EDUCATION_TYPE"
## [15] "NAME_FAMILY_STATUS"          "NAME_HOUSING_TYPE"
## [17] "REGION_POPULATION_RELATIVE"  "DAYS_BIRTH"
## [19] "DAYS_REGISTRATION"           "DAYS_ID_PUBLISH"
## [21] "OWN_CAR_AGE"                 "FLAG_EMP_PHONE"
## [23] "FLAG_WORK_PHONE"             "FLAG_PHONE"
## [25] "FLAG_EMAIL"                  "OCCUPATION_TYPE"
## [27] "CNT_FAM_MEMBERS"             "REGION_RATING_CLIENT"
## [29] "REGION_RATING_CLIENT_W_CITY" "WEEKDAY_APPR_PROCESS_START"
## [31] "HOUR_APPR_PROCESS_START"     "REG_REGION_NOT_WORK_REGION"
## [33] "REG_CITY_NOT_LIVE_CITY"      "REG_CITY_NOT_WORK_CITY"
## [35] "LIVE_CITY_NOT_WORK_CITY"     "ORGANIZATION_TYPE"
## [37] "EXT_SOURCE_1"                "EXT_SOURCE_2"
## [39] "EXT_SOURCE_3"                "APARTMENTS_MEDI"
## [41] "YEARS_BUILD_MEDI"            "COMMONAREA_MEDI"
## [43] "ELEVATORS_MEDI"              "ENTRANCES_MEDI"
## [45] "FLOORSMAX_MEDI"              "FLOORSMIN_MEDI"
## [47] "LIVINGAPARTMENTS_MEDI"       "LIVINGAREA_MEDI"
## [49] "NONLIVINGAPARTMENTS_MEDI"    "OBS_30_CNT_SOCIAL_CIRCLE"
## [51] "DEF_30_CNT_SOCIAL_CIRCLE"    "OBS_60_CNT_SOCIAL_CIRCLE"
## [53] "DEF_60_CNT_SOCIAL_CIRCLE"    "DAYS_LAST_PHONE_CHANGE"
## [55] "FLAG_DOCUMENT_3"             "FLAG_DOCUMENT_6"
## [57] "FLAG_DOCUMENT_8"             "AMT_REQ_CREDIT_BUREAU_HOUR"
## [59] "AMT_REQ_CREDIT_BUREAU_DAY"   "AMT_REQ_CREDIT_BUREAU_WEEK"
```

```
## [61] "AMT_REQ_CREDIT_BUREAU_MON"    "AMT_REQ_CREDIT_BUREAU_QRT"
## [63] "AMT_REQ_CREDIT_BUREAU_YEAR"   "House_Attribute_Low_Variance"
```

```r
# Clean Train Data
train_clean <- train_clean%>% select(-SK_ID_CURR)  # Remove identifier

# Convert character variables to factors
train_clean <- train_clean %>% mutate_if(is.character, as.factor)

# Convert specific numeric variables to factors
num_cat_values <- c(
  "TARGET", "FLAG_EMP_PHONE", "FLAG_WORK_PHONE", "FLAG_EMAIL", "FLAG_PHONE",
  "REG_REGION_NOT_WORK_REGION", "LIVE_CITY_NOT_WORK_CITY",
  "REG_CITY_NOT_LIVE_CITY", "REG_CITY_NOT_WORK_CITY",
  "REGION_RATING_CLIENT_W_CITY", "FLAG_DOCUMENT_3",
  "FLAG_DOCUMENT_6", "FLAG_DOCUMENT_8", "REGION_RATING_CLIENT"
)
train_clean <- train_clean %>% mutate(across(all_of(num_cat_values), as.factor))

# Summary of Cleaned Train Data
str(train_clean)
```

```
## tibble [237,776 x 63] (S3: tbl_df/tbl/data.frame)
##  $ TARGET                      : Factor w/ 2 levels "0","1": 2 1 1 1 1 1 1 1 1 1 ...
##  $ NAME_CONTRACT_TYPE          : Factor w/ 2 levels "Cash loans","Revolving loans": 1 1 1 1 1 1 2 1 1
##  $ CODE_GENDER                 : Factor w/ 2 levels "F","M": 2 1 1 2 2 1 2 1 1 1 ...
##  $ FLAG_OWN_CAR                : Factor w/ 2 levels "N","Y": 1 1 1 1 1 1 1 1 1 1 ...
##  $ FLAG_OWN_REALTY             : Factor w/ 2 levels "N","Y": 2 1 2 2 2 2 2 2 2 2 ...
##  $ CNT_CHILDREN                : num [1:237776] 0 0 0 0 0 0 0 1 0 0 ...
##  $ AMT_INCOME_TOTAL            : num [1:237776] 202500 270000 135000 121500 99000 ...
##  $ AMT_CREDIT                  : num [1:237776] 406598 1293503 312683 513000 490496 ...
##  $ AMT_ANNUITY                 : num [1:237776] 24701 35699 29687 21866 27518 ...
##  $ AMT_GOODS_PRICE             : num [1:237776] 351000 1129500 297000 513000 454500 ...
##  $ NAME_TYPE_SUITE             : Factor w/ 7 levels "Children","Family",..: 7 2 7 7 6 1 7 7 1 7 ...
##  $ NAME_INCOME_TYPE            : Factor w/ 8 levels "Businessman",..: 8 5 8 8 5 4 8 8 4 8 ...
##  $ NAME_EDUCATION_TYPE         : Factor w/ 5 levels "Academic degree",..: 5 2 5 5 5 5 5 2 5 5 ...
##  $ NAME_FAMILY_STATUS          : Factor w/ 6 levels "Civil marriage",..: 4 2 1 4 2 2 4 2 2 2 ...
##  $ NAME_HOUSING_TYPE           : Factor w/ 6 levels "Co-op apartment",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ REGION_POPULATION_RELATIVE  : num [1:237776] 0.0188 0.00354 0.00802 0.02866 0.03579 ...
##  $ DAYS_BIRTH                  : num [1:237776] 9461 16765 19005 19932 16941 ...
##  $ DAYS_REGISTRATION           : num [1:237776] 3648 1186 9833 4311 4970 ...
##  $ DAYS_ID_PUBLISH             : num [1:237776] 2120 291 2437 3458 477 ...
##  $ OWN_CAR_AGE                 : num [1:237776] 0 0 0 0 0 0 0 0 0 0 ...
##  $ FLAG_EMP_PHONE              : Factor w/ 2 levels "0","1": 2 2 2 2 2 1 2 2 1 2 ...
##  $ FLAG_WORK_PHONE             : Factor w/ 2 levels "0","1": 1 1 1 1 2 1 1 1 1 2 ...
##  $ FLAG_PHONE                  : Factor w/ 2 levels "0","1": 2 2 1 1 2 1 1 1 2 2 ...
##  $ FLAG_EMAIL                  : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ OCCUPATION_TYPE             : Factor w/ 19 levels "Accountants",..: 9 4 9 4 9 18 9 4 18 9 ...
##  $ CNT_FAM_MEMBERS             : num [1:237776] 1 2 2 1 2 2 1 3 2 2 ...
##  $ REGION_RATING_CLIENT        : Factor w/ 3 levels "1","2","3": 2 1 2 2 2 2 2 2 2 2 ...
##  $ REGION_RATING_CLIENT_W_CITY : Factor w/ 3 levels "1","2","3": 2 1 2 2 2 2 2 2 2 2 ...
##  $ WEEKDAY_APPR_PROCESS_START  : Factor w/ 7 levels "FRIDAY","MONDAY",..: 7 2 7 5 7 7 5 3 1 1 ...
##  $ HOUR_APPR_PROCESS_START     : num [1:237776] 10 11 17 11 16 14 8 15 7 10 ...
##  $ REG_REGION_NOT_WORK_REGION  : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
```

```
##  $ REG_CITY_NOT_LIVE_CITY      : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ REG_CITY_NOT_WORK_CITY      : Factor w/ 2 levels "0","1": 1 1 1 2 1 1 1 1 1 1 ...
##  $ LIVE_CITY_NOT_WORK_CITY     : Factor w/ 2 levels "0","1": 1 1 1 2 1 1 1 1 1 1 ...
##  $ ORGANIZATION_TYPE           : Factor w/ 58 levels "Advertising",..: 6 40 6 38 34 57 10 31 57 5 ..
##  $ EXT_SOURCE_1                : num [1:237776] 0.083 0.311 0 0 0 ...
##  $ EXT_SOURCE_2                : num [1:237776] 0.263 0.622 0.65 0.323 0.354 ...
##  $ EXT_SOURCE_3                : num [1:237776] 0.139 0.64 0.555 0.458 0.621 ...
##  $ APARTMENTS_MEDI             : num [1:237776] 0.025 0.0968 0.0323 0.0749 0.1135 ...
##  $ YEARS_BUILD_MEDI            : num [1:237776] 0.624 0.799 0.591 0.644 0.698 ...
##  $ COMMONAREA_MEDI             : num [1:237776] 0.0144 0.0608 0 0.243 0.0076 0.0109 0.0338 0.0276 0.0
##  $ ELEVATORS_MEDI              : num [1:237776] 0 0.08 0.08 0.04 0.16 0.16 0.24 0.04 0 0 ...
##  $ ENTRANCES_MEDI              : num [1:237776] 0.069 0.0345 0.9655 0.3103 0.0345 ...
##  $ FLOORSMAX_MEDI              : num [1:237776] 0.0833 0.2917 0.3333 0.3333 0.0833 ...
##  $ FLOORSMIN_MEDI              : num [1:237776] 0.125 0.333 0.375 0.708 0.208 ...
##  $ LIVINGAPARTMENTS_MEDI       : num [1:237776] 0.0205 0.0787 0.0513 0.041 0.053 ...
##  $ LIVINGAREA_MEDI             : num [1:237776] 0.0193 0.0558 0.0897 0.0618 0.0096 ...
##  $ NONLIVINGAPARTMENTS_MEDI    : num [1:237776] 0 0.0039 0 0 0.0155 0.0155 0 0 0.0155 0 ...
##  $ OBS_30_CNT_SOCIAL_CIRCLE    : num [1:237776] 2 1 2 0 0 1 2 0 0 0 ...
##  $ DEF_30_CNT_SOCIAL_CIRCLE    : num [1:237776] 2 0 0 0 0 0 0 0 0 0 ...
##  $ OBS_60_CNT_SOCIAL_CIRCLE    : num [1:237776] 2 1 2 0 0 1 2 0 0 0 ...
##  $ DEF_60_CNT_SOCIAL_CIRCLE    : num [1:237776] 2 0 0 0 0 0 0 0 0 0 ...
##  $ DAYS_LAST_PHONE_CHANGE      : num [1:237776] 1134 828 617 1106 2536 ...
##  $ FLAG_DOCUMENT_3             : Factor w/ 2 levels "0","1": 2 2 2 1 2 2 1 2 1 2 ...
##  $ FLAG_DOCUMENT_6             : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 2 1 ...
##  $ FLAG_DOCUMENT_8             : Factor w/ 2 levels "0","1": 1 1 1 2 1 1 1 1 1 1 ...
##  $ AMT_REQ_CREDIT_BUREAU_HOUR  : num [1:237776] 0 0 0 0 0 0 0 0 0 0 ...
##  $ AMT_REQ_CREDIT_BUREAU_DAY   : num [1:237776] 0 0 0 0 0 0 0 0 0 0 ...
##  $ AMT_REQ_CREDIT_BUREAU_WEEK  : num [1:237776] 0 0 0 0 0 0 0 0 0 0 ...
##  $ AMT_REQ_CREDIT_BUREAU_MON   : num [1:237776] 0 0 0 0 0 0 0 1 0 1 ...
##  $ AMT_REQ_CREDIT_BUREAU_QRT   : num [1:237776] 0 0 0 0 1 0 0 0 0 0 ...
##  $ AMT_REQ_CREDIT_BUREAU_YEAR  : num [1:237776] 1 0 1 0 1 1 1 0 2 0 ...
##  $ House_Attribute_Low_Variance: num [1:237776] -1.463 -1.389 -0.522 -0.522 -0.522 ...
```

```
summary(train_clean)
```

```
##  TARGET           NAME_CONTRACT_TYPE CODE_GENDER FLAG_OWN_CAR FLAG_OWN_REALTY
##  0:218531   Cash loans     :215096   F:168782    N:184043     N: 72789
##  1: 19245   Revolving loans: 22680   M: 68994    Y: 53733     Y:164987
##
##
##
##
##
##   CNT_CHILDREN    AMT_INCOME_TOTAL   AMT_CREDIT       AMT_ANNUITY
##  Min.   :0.0000   Min.   : 25650   Min.   :  45000   Min.   :  1616
##  1st Qu.:0.0000   1st Qu.:108000   1st Qu.: 270000   1st Qu.: 16006
##  Median :0.0000   Median :135000   Median : 491031   Median : 23837
##  Mean   :0.3625   Mean   :148333   Mean   : 567430   Mean   : 25678
##  3rd Qu.:1.0000   3rd Qu.:180000   3rd Qu.: 781695   3rd Qu.: 32603
##  Max.   :3.0000   Max.   :299700   Max.   :3860019   Max.   :225000
##
##  AMT_GOODS_PRICE          NAME_TYPE_SUITE               NAME_INCOME_TYPE
##  Min.   :  40500   Children       :  2713   Working             :121870
##  1st Qu.: 229500   Family         : 31084   Commercial associate: 51983
```

```
##  Median : 450000   Group of people:   203   Pensioner           : 47417
##  Mean   : 509126   Other_A        :   682   State servant       : 16464
##  3rd Qu.: 675000   Other_B        :  1408   Unemployed          :    18
##  Max.   :3555000   Spouse, partner:  8462   Student             :    16
##                    Unaccompanied  :193224   (Other)             :     8
##                     NAME_EDUCATION_TYPE            NAME_FAMILY_STATUS
##  Academic degree              :    99   Civil marriage      : 23341
##  Higher education             : 53967   Married             :147215
##  Incomplete higher            :  7885   Separated           : 16124
##  Lower secondary              :  3062   Single / not married: 36790
##  Secondary / secondary special:172763   Unknown             :     1
##                                         Widow               : 14305
##
##            NAME_HOUSING_TYPE  REGION_POPULATION_RELATIVE   DAYS_BIRTH
##  Co-op apartment    :   745   Min.   :0.00029            Min.   : 7673
##  House / apartment  :210886   1st Qu.:0.01001            1st Qu.:12439
##  Municipal apartment:  8903   Median :0.01885            Median :15968
##  Office apartment   :  1884   Mean   :0.02044            Mean   :16187
##  Rented apartment   :  3778   3rd Qu.:0.02639            3rd Qu.:19961
##  With parents       : 11580   Max.   :0.07251            Max.   :25201
##
##  DAYS_REGISTRATION DAYS_ID_PUBLISH  OWN_CAR_AGE   FLAG_EMP_PHONE
##  Min.   :    0     Min.   :   0    Min.   : 0.0   0: 47432
##  1st Qu.: 2108     1st Qu.:1721    1st Qu.: 0.0   1:190344
##  Median : 4601     Median :3262    Median : 0.0
##  Mean   : 5097     Mean   :2996    Mean   : 1.4
##  3rd Qu.: 7649     3rd Qu.:4299    3rd Qu.: 0.0
##  Max.   :24672     Max.   :7197    Max.   :12.0
##
##  FLAG_WORK_PHONE FLAG_PHONE FLAG_EMAIL    OCCUPATION_TYPE  CNT_FAM_MEMBERS
##  0:190061        0:170246   0:225462   Unemployed :78776   Min.   :1.00
##  1: 47715        1: 67530   1: 12314   Laborers   :41576   1st Qu.:2.00
##                                        Sales staff:26320   Median :2.00
##                                        Core staff :21645   Mean   :2.08
##                                        Managers   :12915   3rd Qu.:2.00
##                                        Drivers    :11595   Max.   :4.00
##                                        (Other)    :44949
##  REGION_RATING_CLIENT REGION_RATING_CLIENT_W_CITY WEEKDAY_APPR_PROCESS_START
##  1: 21787             1: 23152                    FRIDAY   :38758
##  2:179447             2:181331                    MONDAY   :39153
##  3: 36542             3: 33293                    SATURDAY :26248
##                                                   SUNDAY   :12519
##                                                   THURSDAY :39084
##                                                   TUESDAY  :41640
##                                                   WEDNESDAY:40374
##  HOUR_APPR_PROCESS_START REG_REGION_NOT_WORK_REGION REG_CITY_NOT_LIVE_CITY
##  Min.   : 0.00           0:227280                     0:219288
##  1st Qu.:10.00           1: 10496                     1: 18488
##  Median :12.00
##  Mean   :12.09
##  3rd Qu.:14.00
##  Max.   :23.00
##
##  REG_CITY_NOT_WORK_CITY LIVE_CITY_NOT_WORK_CITY              ORGANIZATION_TYPE
```

```
##  0:185117               0:197449               Business Entity Type 3:49932
##  1: 52659               1: 40327               Unemployed             :47426
##                                                Self-employed          :29110
##                                                Other                  :12695
##                                                Medicine               : 9179
##                                                Government             : 8079
##                                                (Other)                :81355
##   EXT_SOURCE_1     EXT_SOURCE_2        EXT_SOURCE_3       APARTMENTS_MEDI
##  Min.   :0.0000   Min.   :0.0000001   Min.   :0.0005273   Min.   :0.00000
##  1st Qu.:0.0000   1st Qu.:0.3859750   1st Qu.:0.3706496   1st Qu.:0.03440
##  Median :0.0000   Median :0.5627356   Median :0.5388627   Median :0.07290
##  Mean   :0.2163   Mean   :0.5111018   Mean   :0.5118501   Mean   :0.09919
##  3rd Qu.:0.4528   3rd Qu.:0.6612173   3rd Qu.:0.6690567   3rd Qu.:0.12280
##  Max.   :0.9516   Max.   :0.8549997   Max.   :0.8960095   Max.   :1.00000
##
##  YEARS_BUILD_MEDI COMMONAREA_MEDI   ELEVATORS_MEDI    ENTRANCES_MEDI
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.00000   Min.   :0.0000
##  1st Qu.:0.6578   1st Qu.:0.00590   1st Qu.:0.00000   1st Qu.:0.0345
##  Median :0.7048   Median :0.01620   Median :0.00000   Median :0.1034
##  Mean   :0.7177   Mean   :0.03946   Mean   :0.07234   Mean   :0.1213
##  3rd Qu.:0.7920   3rd Qu.:0.04580   3rd Qu.:0.12000   3rd Qu.:0.1379
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.00000   Max.   :1.0000
##
##  FLOORSMAX_MEDI   FLOORSMIN_MEDI   LIVINGAPARTMENTS_MEDI LIVINGAREA_MEDI
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.00000       Min.   :0.0000
##  1st Qu.:0.1667   1st Qu.:0.0833   1st Qu.:0.02740       1st Qu.:0.0388
##  Median :0.1667   Median :0.2083   Median :0.06070       Median :0.0677
##  Mean   :0.2130   Mean   :0.2207   Mean   :0.08435       Mean   :0.1003
##  3rd Qu.:0.3333   3rd Qu.:0.3750   3rd Qu.:0.10180       3rd Qu.:0.1209
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.00000       Max.   :1.0000
##
##  NONLIVINGAPARTMENTS_MEDI OBS_30_CNT_SOCIAL_CIRCLE DEF_30_CNT_SOCIAL_CIRCLE
##  Min.   :0.00000          Min.   : 0.000           Min.   :0.0000
##  1st Qu.:0.00000          1st Qu.: 0.000           1st Qu.:0.0000
##  Median :0.00000          Median : 0.000           Median :0.0000
##  Mean   :0.00639          Mean   : 1.322           Mean   :0.1432
##  3rd Qu.:0.00000          3rd Qu.: 2.000           3rd Qu.:0.0000
##  Max.   :1.00000          Max.   :10.000           Max.   :6.0000
##
##  OBS_60_CNT_SOCIAL_CIRCLE DEF_60_CNT_SOCIAL_CIRCLE DAYS_LAST_PHONE_CHANGE
##  Min.   : 0.000           Min.   :0.0000           Min.   :   0.0
##  1st Qu.: 0.000           1st Qu.:0.0000           1st Qu.: 266.0
##  Median : 0.000           Median :0.0000           Median : 741.0
##  Mean   : 1.306           Mean   :0.1005           Mean   : 949.7
##  3rd Qu.: 2.000           3rd Qu.:0.0000           3rd Qu.:1554.0
##  Max.   :10.000           Max.   :6.0000           Max.   :4292.0
##
##  FLAG_DOCUMENT_3 FLAG_DOCUMENT_6 FLAG_DOCUMENT_8 AMT_REQ_CREDIT_BUREAU_HOUR
##  0: 66846        0:214506        0:222941        Min.   :0.000000
##  1:170930        1: 23270        1: 14835        1st Qu.:0.000000
##                                                  Median :0.000000
##                                                  Mean   :0.005451
##                                                  3rd Qu.:0.000000
##                                                  Max.   :4.000000
```

```
##
##  AMT_REQ_CREDIT_BUREAU_DAY AMT_REQ_CREDIT_BUREAU_WEEK AMT_REQ_CREDIT_BUREAU_MON
##  Min.   :0.000000          Min.   :0.00000            Min.   : 0.0000
##  1st Qu.:0.000000          1st Qu.:0.00000            1st Qu.: 0.0000
##  Median :0.000000          Median :0.00000            Median : 0.0000
##  Mean   :0.006157          Mean   :0.02947            Mean   : 0.2221
##  3rd Qu.:0.000000          3rd Qu.:0.00000            3rd Qu.: 0.0000
##  Max.   :9.000000          Max.   :8.00000            Max.   :24.0000
##
##  AMT_REQ_CREDIT_BUREAU_QRT AMT_REQ_CREDIT_BUREAU_YEAR
##  Min.   : 0.00             Min.   : 0.00
##  1st Qu.: 0.00             1st Qu.: 1.00
##  Median : 0.00             Median : 1.00
##  Mean   : 0.23             Mean   : 1.78
##  3rd Qu.: 0.00             3rd Qu.: 3.00
##  Max.   :19.00             Max.   :25.00
##
##  House_Attribute_Low_Variance
##  Min.   :-2.77162
##  1st Qu.:-0.52169
##  Median :-0.52169
##  Mean   :-0.00629
##  3rd Qu.:-0.34387
##  Max.   :53.15290
##
```

```r
# Clean Test Data
test_clean <- test_clean %>% select(-SK_ID_CURR)  # Remove identifier
test_clean <- test_clean %>% mutate_if(is.character, as.factor)


# Summary of Cleaned Test Data
str(test_clean)
```

```
## tibble [37,740 x 62] (S3: tbl_df/tbl/data.frame)
##  $ NAME_CONTRACT_TYPE          : Factor w/ 2 levels "Cash loans","Revolving loans": 1 1 1 1 1 1 1 1
##  $ CODE_GENDER                 : Factor w/ 2 levels "F","M": 1 2 2 1 2 2 1 1 1 1 ...
##  $ FLAG_OWN_CAR                : Factor w/ 2 levels "N","Y": 1 1 2 2 2 1 2 1 1 1 ...
##  $ FLAG_OWN_REALTY             : Factor w/ 2 levels "N","Y": 2 2 2 2 2 2 2 2 2 2 ...
##  $ CNT_CHILDREN                : num [1:37740] 0 0 0 0 2 0 1 0 0 0 ...
##  $ AMT_INCOME_TOTAL            : num [1:37740] 135000 99000 202500 270000 180000 ...
##  $ AMT_CREDIT                  : num [1:37740] 568800 222768 663264 959688 499221 ...
##  $ AMT_ANNUITY                 : num [1:37740] 20561 17370 69777 34601 22118 ...
##  $ AMT_GOODS_PRICE             : num [1:37740] 450000 180000 630000 810000 373500 ...
##  $ NAME_TYPE_SUITE             : Factor w/ 7 levels "Children","Family",..: 7 7 7 7 7 7 2 7 7 7 ...
##  $ NAME_INCOME_TYPE            : Factor w/ 6 levels "Commercial associate",..: 6 6 6 3 6 6 6 2 6 6 .
##  $ NAME_EDUCATION_TYPE         : Factor w/ 5 levels "Academic degree",..: 2 5 2 5 2 2 2 5 5 5 ...
##  $ NAME_FAMILY_STATUS          : Factor w/ 5 levels "Civil marriage",..: 2 2 2 2 2 4 1 2 5 1 ...
##  $ NAME_HOUSING_TYPE           : Factor w/ 6 levels "Co-op apartment",..: 2 2 2 2 2 6 2 2 2 2 ...
##  $ REGION_POPULATION_RELATIVE  : num [1:37740] 0.0188 0.0358 0.0191 0.0252 0.0228 ...
##  $ DAYS_BIRTH                  : num [1:37740] 19241 18064 20038 18604 16685 ...
##  $ DAYS_REGISTRATION           : num [1:37740] 5170 9118 2175 6116 10125 ...
##  $ DAYS_ID_PUBLISH             : num [1:37740] 812 1623 3503 2027 241 ...
##  $ OWN_CAR_AGE                 : num [1:37740] 0 0 5 10 3 0 5 0 0 0 ...
```

```
##  $ FLAG_EMP_PHONE              : num [1:37740] 1 1 1 1 1 1 1 0 1 1 ...
##  $ FLAG_WORK_PHONE             : num [1:37740] 0 0 0 0 0 1 1 0 0 0 ...
##  $ FLAG_PHONE                  : num [1:37740] 0 0 0 1 0 1 1 1 0 0 ...
##  $ FLAG_EMAIL                  : num [1:37740] 1 0 0 0 0 0 0 0 0 0 ...
##  $ OCCUPATION_TYPE             : Factor w/ 19 levels "Accountants",..: 18 10 5 5 6 4 15 18 9 18 ...
##  $ CNT_FAM_MEMBERS             : num [1:37740] 2 2 2 2 4 1 3 2 1 2 ...
##  $ REGION_RATING_CLIENT        : num [1:37740] 2 2 2 2 2 2 2 3 2 2 ...
##  $ REGION_RATING_CLIENT_W_CITY : num [1:37740] 2 2 2 2 2 2 2 3 2 2 ...
##  $ WEEKDAY_APPR_PROCESS_START  : Factor w/ 7 levels "FRIDAY","MONDAY",..: 6 1 2 2 5 1 6 6 1 5 ...
##  $ HOUR_APPR_PROCESS_START     : num [1:37740] 18 9 14 15 9 7 14 11 17 17 ...
##  $ REG_REGION_NOT_WORK_REGION  : num [1:37740] 0 0 0 0 0 0 0 0 0 0 ...
##  $ REG_CITY_NOT_LIVE_CITY      : num [1:37740] 0 0 0 0 0 0 0 0 0 0 ...
##  $ REG_CITY_NOT_WORK_CITY      : num [1:37740] 0 0 0 0 1 0 0 0 1 0 ...
##  $ LIVE_CITY_NOT_WORK_CITY     : num [1:37740] 0 0 0 0 1 0 0 0 1 0 ...
##  $ ORGANIZATION_TYPE           : Factor w/ 58 levels "Advertising",..: 29 43 55 12 27 43 47 57 43 43
##  $ EXT_SOURCE_1                : num [1:37740] 0.753 0.565 0 0 0.761 ...
##  $ EXT_SOURCE_2                : num [1:37740] 0.79 0.292 0.7 0.629 0.571 ...
##  $ EXT_SOURCE_3                : num [1:37740] 0.16 0.433 0.611 0.393 0.651 ...
##  $ APARTMENTS_MEDI             : num [1:37740] 0.0666 0.2946 0.0625 0.2436 0.0625 ...
##  $ YEARS_BUILD_MEDI            : num [1:37740] 0.698 0.685 0.685 0.758 0.685 ...
##  $ COMMONAREA_MEDI             : num [1:37740] 0.0883 0.0267 0.0062 0.0455 0.1237 ...
##  $ ELEVATORS_MEDI              : num [1:37740] 0 0 0.08 0.16 0 0.08 0 0 0 0.08 ...
##  $ ENTRANCES_MEDI              : num [1:37740] 0.1379 0.1379 0.0345 0.1379 0.1034 ...
##  $ FLOORSMAX_MEDI              : num [1:37740] 0.125 0.1875 0.1667 0.3333 0.0417 ...
##  $ FLOORSMIN_MEDI              : num [1:37740] 0.1667 0.2083 0.0417 0.375 0.125 ...
##  $ LIVINGAPARTMENTS_MEDI       : num [1:37740] 1 0.1509 0.0428 0.1975 0.1505 ...
##  $ LIVINGAREA_MEDI             : num [1:37740] 0.0514 0.1351 0.0388 0.2258 0.1457 ...
##  $ NONLIVINGAPARTMENTS_MEDI    : num [1:37740] 0.0155 0 0 0.0116 0 0 0.0311 0.0039 0 0 ...
##  $ OBS_30_CNT_SOCIAL_CIRCLE    : num [1:37740] 0 0 0 0 1 0 4 0 0 1 ...
##  $ DEF_30_CNT_SOCIAL_CIRCLE    : num [1:37740] 0 0 0 0 0 0 0 0 0 1 ...
##  $ OBS_60_CNT_SOCIAL_CIRCLE    : num [1:37740] 0 0 0 0 1 0 4 0 0 1 ...
##  $ DEF_60_CNT_SOCIAL_CIRCLE    : num [1:37740] 0 0 0 0 0 0 0 0 0 0 ...
##  $ DAYS_LAST_PHONE_CHANGE      : num [1:37740] 1740 0 856 1705 1182 ...
##  $ FLAG_DOCUMENT_3             : num [1:37740] 1 1 0 0 1 0 1 1 1 1 ...
##  $ FLAG_DOCUMENT_6             : num [1:37740] 0 0 0 0 0 0 0 0 0 0 ...
##  $ FLAG_DOCUMENT_8             : num [1:37740] 0 0 1 1 0 1 0 0 0 0 ...
##  $ AMT_REQ_CREDIT_BUREAU_HOUR  : num [1:37740] 0 0 0 0 0 0 0 0 0 0 ...
##  $ AMT_REQ_CREDIT_BUREAU_DAY   : num [1:37740] 0 0 0 0 0 0 0 0 0 0 ...
##  $ AMT_REQ_CREDIT_BUREAU_WEEK  : num [1:37740] 0 0 0 0 0 0 0 0 0 0 ...
##  $ AMT_REQ_CREDIT_BUREAU_MON   : num [1:37740] 0 0 0 0 0 0 0 0 0 0 ...
##  $ AMT_REQ_CREDIT_BUREAU_QRT   : num [1:37740] 0 0 1 1 0 0 0 0 0 1 ...
##  $ AMT_REQ_CREDIT_BUREAU_YEAR  : num [1:37740] 0 3 4 2 1 2 2 0 5 3 ...
##  $ House_Attribute_Low_Variance: num [1:37740] -0.853 -0.532 -0.532 2.071 -0.532 ...
```

```
summary(test_clean)
```

```
##        NAME_CONTRACT_TYPE CODE_GENDER FLAG_OWN_CAR FLAG_OWN_REALTY
##  Cash loans     :37391    F:27114     N:29341      N:11626
##  Revolving loans:  349    M:10626     Y: 8399      Y:26114
##
##
##
##
##
```

```
##   CNT_CHILDREN    AMT_INCOME_TOTAL   AMT_CREDIT       AMT_ANNUITY
##  Min.   :0.0000   Min.   : 26942   Min.   :  45000   Min.   :  2295
##  1st Qu.:0.0000   1st Qu.:112500   1st Qu.: 248760   1st Qu.: 17262
##  Median :0.0000   Median :153000   Median : 414612   Median : 24926
##  Mean   :0.3456   Mean   :155613   Mean   : 480048   Mean   : 27570
##  3rd Qu.:1.0000   3rd Qu.:202500   3rd Qu.: 610484   3rd Qu.: 34826
##  Max.   :3.0000   Max.   :299250   Max.   :2245500   Max.   :177827
##
##  AMT_GOODS_PRICE          NAME_TYPE_SUITE              NAME_INCOME_TYPE
##  Min.   :  45000   Children       :  328   Commercial associate: 8298
##  1st Qu.: 225000   Family         : 4619   Pensioner           : 7995
##  Median : 360000   Group of people:   38   State servant       : 2677
##  Mean   : 428636   Other_A        :   80   Student             :    1
##  3rd Qu.: 540000   Other_B        :  174   Unemployed          :    1
##  Max.   :2245500   Spouse, partner: 1056   Working             :18768
##                    Unaccompanied  :31445
##                     NAME_EDUCATION_TYPE            NAME_FAMILY_STATUS
##  Academic degree              :   31    Civil marriage      : 3367
##  Higher education             : 8954    Married             :24252
##  Incomplete higher            : 1319    Separated           : 2426
##  Lower secondary              :  390    Single / not married: 5722
##  Secondary / secondary special:27046    Widow               : 1973
##
##
##            NAME_HOUSING_TYPE REGION_POPULATION_RELATIVE   DAYS_BIRTH
##  Co-op apartment    :   77   Min.   :0.000253           Min.   : 7338
##  House / apartment  :33821   1st Qu.:0.010006           1st Qu.:12503
##  Municipal apartment: 1260   Median :0.018850           Median :16004
##  Office apartment   :  300   Mean   :0.020619           Mean   :16211
##  Rented apartment   :  553   3rd Qu.:0.028663           3rd Qu.:19942
##  With parents       : 1729   Max.   :0.072508           Max.   :25195
##
##  DAYS_REGISTRATION DAYS_ID_PUBLISH  OWN_CAR_AGE     FLAG_EMP_PHONE
##  Min.   :    0     Min.   :   0    Min.   : 0.000   Min.   :0.0000
##  1st Qu.: 2004     1st Qu.:1711    1st Qu.: 0.000   1st Qu.:1.0000
##  Median : 4622     Median :3249    Median : 0.000   Median :1.0000
##  Mean   : 5092     Mean   :3058    Mean   : 1.403   Mean   :0.7881
##  3rd Qu.: 7659     3rd Qu.:4448    3rd Qu.: 0.000   3rd Qu.:1.0000
##  Max.   :23722     Max.   :6348    Max.   :12.000   Max.   :1.0000
##
##  FLAG_WORK_PHONE    FLAG_PHONE       FLAG_EMAIL        OCCUPATION_TYPE
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Unemployed :12824
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   Laborers   : 6632
##  Median :0.0000   Median :0.0000   Median :0.0000   Sales staff: 4176
##  Mean   :0.2041   Mean   :0.2656   Mean   :0.1538   Core staff : 3469
##  3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:0.0000   Managers   : 2028
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Drivers    : 1718
##                                                     (Other)    : 6893
##  CNT_FAM_MEMBERS REGION_RATING_CLIENT REGION_RATING_CLIENT_W_CITY
##  Min.   :1.000   Min.   :1.000        Min.   :-1.000
##  1st Qu.:2.000   1st Qu.:2.000        1st Qu.: 2.000
##  Median :2.000   Median :2.000        Median : 2.000
##  Mean   :2.077   Mean   :2.053        Mean   : 2.031
##  3rd Qu.:2.000   3rd Qu.:2.000        3rd Qu.: 2.000
```

```
##  Max.   :4.000   Max.   :3.000        Max.   : 3.000
##
##  WEEKDAY_APPR_PROCESS_START HOUR_APPR_PROCESS_START REG_REGION_NOT_WORK_REGION
##  FRIDAY   :5670             Min.   : 0.00           Min.   :0.00000
##  MONDAY   :6456             1st Qu.:10.00           1st Qu.:0.00000
##  SATURDAY :3554             Median :12.00           Median :0.00000
##  SUNDAY   :1437             Mean   :12.04           Mean   :0.04642
##  THURSDAY :6495             3rd Qu.:14.00           3rd Qu.:0.00000
##  TUESDAY  :7562             Max.   :23.00           Max.   :1.00000
##  WEDNESDAY:6566
##  REG_CITY_NOT_LIVE_CITY REG_CITY_NOT_WORK_CITY LIVE_CITY_NOT_WORK_CITY
##  Min.   :0.00000        Min.   :0.0000         Min.   :0.0000
##  1st Qu.:0.00000        1st Qu.:0.0000         1st Qu.:0.0000
##  Median :0.00000        Median :0.0000         Median :0.0000
##  Mean   :0.07713        Mean   :0.2141         Mean   :0.1631
##  3rd Qu.:0.00000        3rd Qu.:0.0000         3rd Qu.:0.0000
##  Max.   :1.00000        Max.   :1.0000         Max.   :1.0000
##
##               ORGANIZATION_TYPE  EXT_SOURCE_1     EXT_SOURCE_2
##  Unemployed            : 7996   Min.   :0.0000   Min.   :0.0000081
##  Business Entity Type 3: 7900   1st Qu.:0.0000   1st Qu.:0.4039882
##  Self-employed         : 4477   Median :0.2366   Median :0.5556799
##  Other                 : 2004   Mean   :0.2853   Mean   :0.5149308
##  Medicine              : 1403   3rd Qu.:0.5464   3rd Qu.:0.6555860
##  Government            : 1180   Max.   :0.9391   Max.   :0.8549997
##  (Other)               :12780
##   EXT_SOURCE_3       APARTMENTS_MEDI  YEARS_BUILD_MEDI COMMONAREA_MEDI
##  Min.   :0.0005273   Min.   :0.0000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.3656165   1st Qu.:0.0416   1st Qu.:0.6847   1st Qu.:0.00760
##  Median :0.5208976   Median :0.0833   Median :0.7048   Median :0.02200
##  Mean   :0.5020387   Mean   :0.1109   Mean   :0.7486   Mean   :0.05325
##  3rd Qu.:0.6545293   3rd Qu.:0.1374   3rd Qu.:0.8390   3rd Qu.:0.05750
##  Max.   :0.8825303   Max.   :1.0000   Max.   :1.0000   Max.   :1.00000
##
##  ELEVATORS_MEDI    ENTRANCES_MEDI   FLOORSMAX_MEDI   FLOORSMIN_MEDI
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:0.0690   1st Qu.:0.1667   1st Qu.:0.0833
##  Median :0.00000   Median :0.1379   Median :0.1667   Median :0.2083
##  Mean   :0.08353   Mean   :0.1438   Mean   :0.2270   Mean   :0.2255
##  3rd Qu.:0.12000   3rd Qu.:0.2069   3rd Qu.:0.3333   3rd Qu.:0.3750
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##
##  LIVINGAPARTMENTS_MEDI LIVINGAREA_MEDI  NONLIVINGAPARTMENTS_MEDI
##  Min.   :0.00000       Min.   :0.0000   Min.   :0.000000
##  1st Qu.:0.03850       1st Qu.:0.0488   1st Qu.:0.000000
##  Median :0.06670       Median :0.0771   Median :0.000000
##  Mean   :0.09857       Mean   :0.1115   Mean   :0.008813
##  3rd Qu.:0.12310       3rd Qu.:0.1359   3rd Qu.:0.003900
##  Max.   :1.00000       Max.   :1.0000   Max.   :1.000000
##
##  OBS_30_CNT_SOCIAL_CIRCLE DEF_30_CNT_SOCIAL_CIRCLE OBS_60_CNT_SOCIAL_CIRCLE
##  Min.   : 0.000           Min.   :0.0000           Min.   : 0.000
##  1st Qu.: 0.000           1st Qu.:0.0000           1st Qu.: 0.000
##  Median : 0.000           Median :0.0000           Median : 0.000
```

```
##  Mean   : 1.332           Mean   :0.1427           Mean   : 1.321
##  3rd Qu.: 2.000           3rd Qu.:0.0000           3rd Qu.: 2.000
##  Max.   :10.000           Max.   :6.0000           Max.   :10.000
##
##  DEF_60_CNT_SOCIAL_CIRCLE DAYS_LAST_PHONE_CHANGE FLAG_DOCUMENT_3
##  Min.   :0.0000           Min.   :   0           Min.   :0.0000
##  1st Qu.:0.0000           1st Qu.: 354           1st Qu.:1.0000
##  Median :0.0000           Median : 839           Median :1.0000
##  Mean   :0.1015           Mean   :1062           Mean   :0.7964
##  3rd Qu.:0.0000           3rd Qu.:1755           3rd Qu.:1.0000
##  Max.   :5.0000           Max.   :4361           Max.   :1.0000
##
##  FLAG_DOCUMENT_6   FLAG_DOCUMENT_8   AMT_REQ_CREDIT_BUREAU_HOUR
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.000000
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.000000
##  Median :0.00000   Median :0.00000   Median :0.000000
##  Mean   :0.09764   Mean   :0.06778   Mean   :0.001696
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.000000
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.000000
##
##  AMT_REQ_CREDIT_BUREAU_DAY AMT_REQ_CREDIT_BUREAU_WEEK AMT_REQ_CREDIT_BUREAU_MON
##  Min.   :0.000000          Min.   :0.000000           Min.   :0.000000
##  1st Qu.:0.000000          1st Qu.:0.000000           1st Qu.:0.000000
##  Median :0.000000          Median :0.000000           Median :0.000000
##  Mean   :0.001431          Mean   :0.002544           Mean   :0.008161
##  3rd Qu.:0.000000          3rd Qu.:0.000000           3rd Qu.:0.000000
##  Max.   :2.000000          Max.   :2.000000           Max.   :6.000000
##
##  AMT_REQ_CREDIT_BUREAU_QRT AMT_REQ_CREDIT_BUREAU_YEAR
##  Min.   :0.0000            Min.   : 0.000
##  1st Qu.:0.0000            1st Qu.: 1.000
##  Median :0.0000            Median : 2.000
##  Mean   :0.4734            Mean   : 1.999
##  3rd Qu.:1.0000            3rd Qu.: 3.000
##  Max.   :7.0000            Max.   :17.000
##
##  House_Attribute_Low_Variance
##  Min.   :-2.78776
##  1st Qu.:-0.53249
##  Median :-0.53249
##  Mean   :-0.00739
##  3rd Qu.:-0.29941
##  Max.   :48.18664
##
```

#Class Imbalance

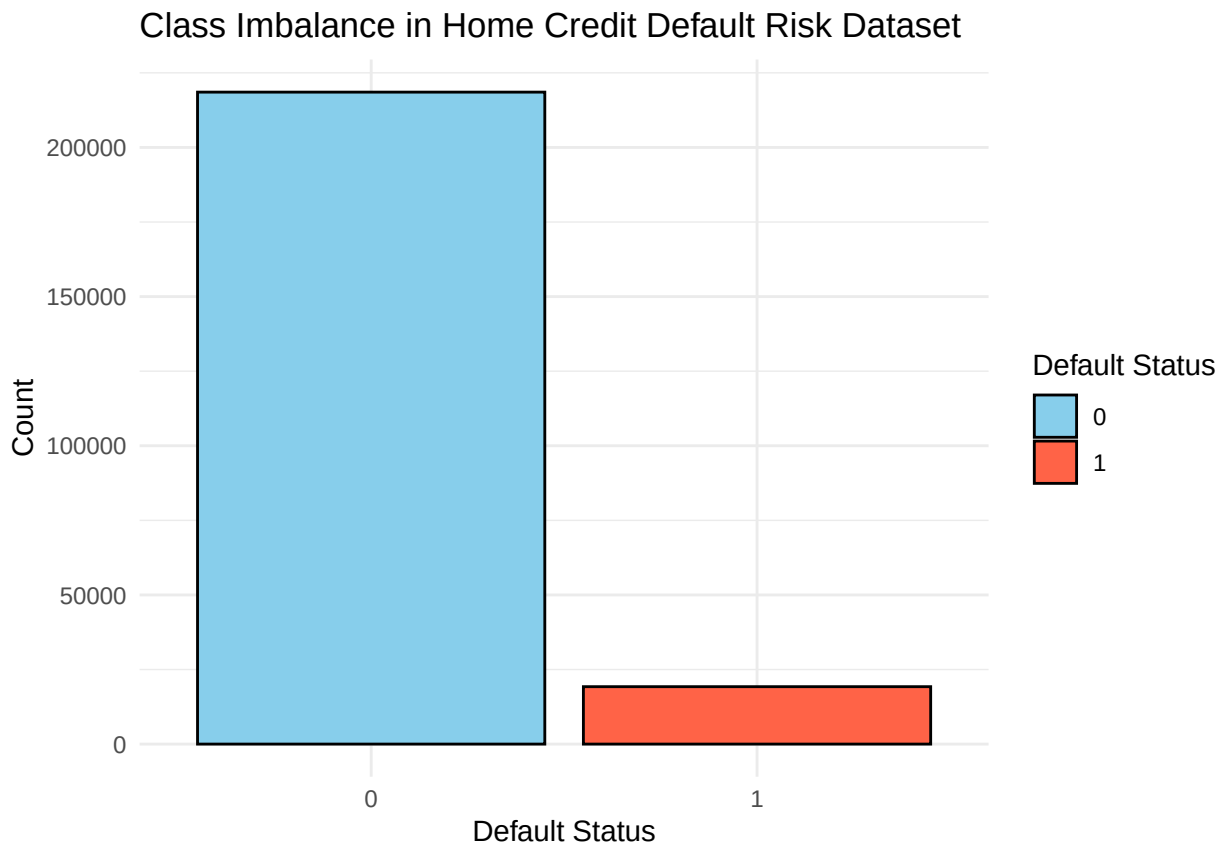
```
# Class Imbalance Analysis
table(train_clean$TARGET)
```

```
##
##      0      1
## 218531  19245
```

```r
round(prop.table(table(train_clean$TARGET)), 4) * 100
```

```
##
##     0     1
## 91.91  8.09
```

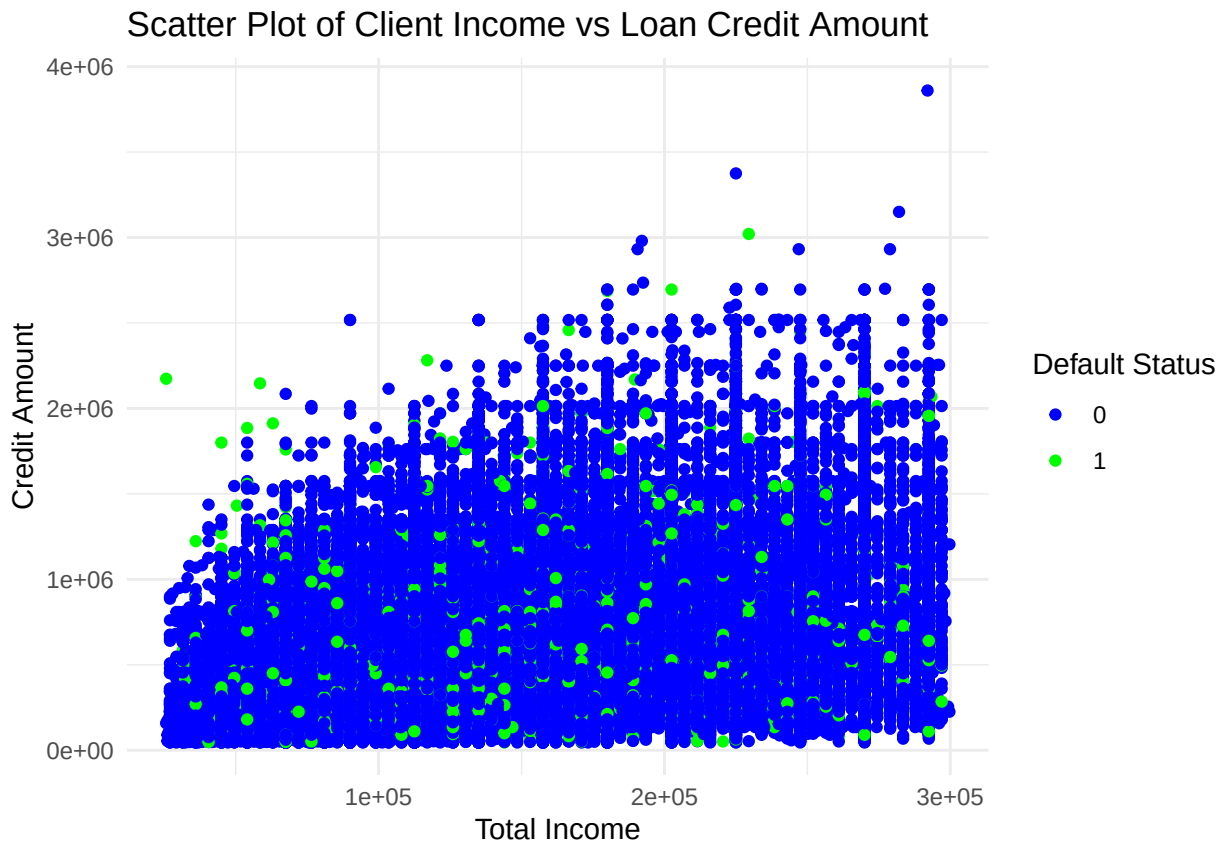
```r
ggplot(train_clean, aes(x = TARGET, fill = TARGET)) +
  geom_bar(color = "black") +
  scale_fill_manual(values = c("0" = "skyblue", "1" = "tomato")) +
  theme_minimal() +
  labs(
    title = "Class Imbalance in Home Credit Default Risk Dataset",
    x = "Default Status",
    y = "Count",
    fill = "Default Status"
  )
```



Non-Linear Separable Check

```r
ggplot(train_clean, aes(x = AMT_INCOME_TOTAL, y = AMT_CREDIT, color = as.factor(TARGET))) +
  geom_point() +
  labs(
    title = "Scatter Plot of Client Income vs Loan Credit Amount",
    x = "Total Income",
    y = "Credit Amount",
```

```
    color = "Default Status"
  ) +
  scale_color_manual(values = c("0" = "blue", "1" = "green")) +
  theme_minimal()
```




```
# Data Splitting
set.seed(123)
train_index <- createDataPartition(train_clean$TARGET, p = 0.7, list = FALSE)
train_data <- train_clean[train_index, ]
test_data <- train_clean[-train_index, ]
```

#Logistic Regression

```
# Logistic Regression
logistic_model <- glm(TARGET ~ ., data = train_data, family = binomial())
summary(logistic_model)
```

```
##
## Call:
## glm(formula = TARGET ~ ., family = binomial(), data = train_data)
##
## Coefficients: (1 not defined because of singularities)
##                                                 Estimate Std. Error z value
## (Intercept)                                   -3.213e+01  3.969e+02  -0.081
```

```
## NAME_CONTRACT_TYPERevolving loans                -7.497e-02  6.326e-02  -1.185
## CODE_GENDERM                                      3.375e-01  2.495e-02  13.529
## FLAG_OWN_CARY                                    -4.633e-01  4.829e-02  -9.594
## FLAG_OWN_REALTYY                                  1.954e-02  2.135e-02   0.915
## CNT_CHILDREN                                      2.960e-02  1.569e-02   1.887
## AMT_INCOME_TOTAL                                 -4.992e-08  2.036e-07  -0.245
## AMT_CREDIT                                        2.141e-06  1.568e-07  13.657
## AMT_ANNUITY                                       1.203e-05  1.239e-06   9.707
## AMT_GOODS_PRICE                                  -2.665e-06  1.787e-07 -14.912
## NAME_TYPE_SUITEFamily                            -4.353e-03  9.629e-02  -0.045
## NAME_TYPE_SUITEGroup of people                    2.024e-01  3.085e-01   0.656
## NAME_TYPE_SUITEOther_A                           -6.387e-02  1.926e-01  -0.332
## NAME_TYPE_SUITEOther_B                            5.848e-02  1.461e-01   0.400
## NAME_TYPE_SUITESpouse, partner                   -1.064e-01  1.056e-01  -1.007
## NAME_TYPE_SUITEUnaccompanied                     -4.463e-03  9.322e-02  -0.048
## NAME_INCOME_TYPECommercial associate              1.048e+01  3.017e+02   0.035
## NAME_INCOME_TYPEMaternity leave                   1.388e+01  3.017e+02   0.046
## NAME_INCOME_TYPEPensioner                        -2.327e-01  3.692e+02  -0.001
## NAME_INCOME_TYPEState servant                     1.049e+01  3.017e+02   0.035
## NAME_INCOME_TYPEStudent                          -5.952e-01  3.327e+02  -0.002
## NAME_INCOME_TYPEUnemployed                        2.973e+00  3.692e+02   0.008
## NAME_INCOME_TYPEWorking                           1.059e+01  3.017e+02   0.035
## NAME_EDUCATION_TYPEHigher education               1.070e+01  6.165e+01   0.174
## NAME_EDUCATION_TYPEIncomplete higher              1.077e+01  6.165e+01   0.175
## NAME_EDUCATION_TYPELower secondary                1.102e+01  6.165e+01   0.179
## NAME_EDUCATION_TYPESecondary / secondary special  1.094e+01  6.165e+01   0.177
## NAME_FAMILY_STATUSMarried                        -1.322e-01  3.046e-02  -4.338
## NAME_FAMILY_STATUSSeparated                       4.853e-02  4.526e-02   1.072
## NAME_FAMILY_STATUSSingle / not married           -5.443e-02  3.566e-02  -1.526
## NAME_FAMILY_STATUSUnknown                        -1.013e+01  5.354e+02  -0.019
## NAME_FAMILY_STATUSWidow                          -8.499e-02  5.350e-02  -1.588
## NAME_HOUSING_TYPEHouse / apartment                1.033e-01  1.711e-01   0.604
## NAME_HOUSING_TYPEMunicipal apartment              1.752e-01  1.777e-01   0.986
## NAME_HOUSING_TYPEOffice apartment                -8.875e-02  2.041e-01  -0.435
## NAME_HOUSING_TYPERented apartment                 2.583e-01  1.812e-01   1.426
## NAME_HOUSING_TYPEWith parents                     1.321e-01  1.745e-01   0.757
## REGION_POPULATION_RELATIVE                        2.760e+00  9.695e-01   2.847
## DAYS_BIRTH                                       -2.790e-05  3.250e-06  -8.586
## DAYS_REGISTRATION                                -1.109e-05  2.966e-06  -3.739
## DAYS_ID_PUBLISH                                  -5.185e-05  6.677e-06  -7.765
## OWN_CAR_AGE                                       2.103e-02  6.190e-03   3.397
## FLAG_EMP_PHONE1                                   1.052e+01  2.505e+02   0.042
## FLAG_WORK_PHONE1                                  1.934e-01  2.486e-02   7.780
## FLAG_PHONE1                                      -9.585e-02  2.316e-02  -4.138
## FLAG_EMAIL1                                       1.567e-02  4.307e-02   0.364
## OCCUPATION_TYPECleaning staff                     2.609e-01  9.389e-02   2.779
## OCCUPATION_TYPECooking staff                      1.601e-01  8.858e-02   1.808
## OCCUPATION_TYPECore staff                         1.505e-02  7.541e-02   0.200
## OCCUPATION_TYPEDrivers                            2.396e-01  7.884e-02   3.039
## OCCUPATION_TYPEHigh skill tech staff              7.016e-02  8.442e-02   0.831
## OCCUPATION_TYPEHR staff                           4.174e-01  2.266e-01   1.842
## OCCUPATION_TYPEIT staff                          -4.723e-02  2.921e-01  -0.162
## OCCUPATION_TYPELaborers                           2.054e-01  6.998e-02   2.935
## OCCUPATION_TYPELow-skill Laborers                 3.284e-01  1.074e-01   3.057
```

```
## OCCUPATION_TYPEManagers                          3.775e-02  7.961e-02   0.474
## OCCUPATION_TYPEMedicine staff                    3.623e-02  9.891e-02   0.366
## OCCUPATION_TYPEPrivate service staff             1.441e-02  1.282e-01   0.112
## OCCUPATION_TYPERealty agents                    -9.831e-02  2.163e-01  -0.454
## OCCUPATION_TYPESales staff                       1.299e-01  7.143e-02   1.819
## OCCUPATION_TYPESecretaries                       1.644e-01  1.620e-01   1.015
## OCCUPATION_TYPESecurity staff                    2.441e-01  9.652e-02   2.529
## OCCUPATION_TYPEUnemployed                        1.297e-01  7.021e-02   1.847
## OCCUPATION_TYPEWaiters/barmen staff              2.721e-01  1.328e-01   2.048
## CNT_FAM_MEMBERS                                         NA         NA      NA
## REGION_RATING_CLIENT2                           -3.482e-01  1.622e-01  -2.147
## REGION_RATING_CLIENT3                           -2.490e-01  1.638e-01  -1.520
## REGION_RATING_CLIENT_W_CITY2                     4.999e-01  1.548e-01   3.229
## REGION_RATING_CLIENT_W_CITY3                     5.553e-01  1.580e-01   3.514
## WEEKDAY_APPR_PROCESS_STARTMONDAY                -8.573e-02  3.285e-02  -2.610
## WEEKDAY_APPR_PROCESS_STARTSATURDAY              -6.681e-02  3.652e-02  -1.830
## WEEKDAY_APPR_PROCESS_STARTSUNDAY                -1.110e-01  4.741e-02  -2.341
## WEEKDAY_APPR_PROCESS_STARTTHURSDAY              -3.796e-02  3.257e-02  -1.166
## WEEKDAY_APPR_PROCESS_STARTTUESDAY                2.076e-02  3.171e-02   0.655
## WEEKDAY_APPR_PROCESS_STARTWEDNESDAY             -9.614e-03  3.221e-02  -0.298
## HOUR_APPR_PROCESS_START                          6.945e-04  3.068e-03   0.226
## REG_REGION_NOT_WORK_REGION1                     -7.786e-02  4.614e-02  -1.688
## REG_CITY_NOT_LIVE_CITY1                          1.861e-01  4.670e-02   3.986
## REG_CITY_NOT_WORK_CITY1                         -2.391e-03  5.219e-02  -0.046
## LIVE_CITY_NOT_WORK_CITY1                         3.323e-02  5.054e-02   0.658
## ORGANIZATION_TYPEAgriculture                    -2.935e-01  2.628e-01  -1.117
## ORGANIZATION_TYPEBank                           -4.943e-01  2.730e-01  -1.811
## ORGANIZATION_TYPEBusiness Entity Type 1         -3.538e-01  2.532e-01  -1.397
## ORGANIZATION_TYPEBusiness Entity Type 2         -3.117e-01  2.489e-01  -1.252
## ORGANIZATION_TYPEBusiness Entity Type 3         -1.771e-01  2.445e-01  -0.724
## ORGANIZATION_TYPECleaning                        1.500e-02  3.579e-01   0.042
## ORGANIZATION_TYPEConstruction                   -5.584e-02  2.508e-01  -0.223
## ORGANIZATION_TYPECulture                        -7.263e-02  3.625e-01  -0.200
## ORGANIZATION_TYPEElectricity                    -3.759e-01  3.034e-01  -1.239
## ORGANIZATION_TYPEEmergency                      -3.099e-01  3.398e-01  -0.912
## ORGANIZATION_TYPEGovernment                     -3.314e-01  2.498e-01  -1.326
## ORGANIZATION_TYPEHotel                          -4.523e-01  2.984e-01  -1.516
## ORGANIZATION_TYPEHousing                        -4.194e-01  2.639e-01  -1.589
## ORGANIZATION_TYPEIndustry: type 1               -1.198e-01  2.810e-01  -0.426
## ORGANIZATION_TYPEIndustry: type 10              -2.031e-01  5.569e-01  -0.365
## ORGANIZATION_TYPEIndustry: type 11              -2.816e-01  2.621e-01  -1.074
## ORGANIZATION_TYPEIndustry: type 12              -8.838e-01  4.611e-01  -1.917
## ORGANIZATION_TYPEIndustry: type 13              -2.529e-01  5.698e-01  -0.444
## ORGANIZATION_TYPEIndustry: type 2               -5.986e-01  3.482e-01  -1.719
## ORGANIZATION_TYPEIndustry: type 3               -1.463e-01  2.570e-01  -0.570
## ORGANIZATION_TYPEIndustry: type 4               -1.903e-01  2.867e-01  -0.664
## ORGANIZATION_TYPEIndustry: type 5               -4.307e-01  3.197e-01  -1.347
## ORGANIZATION_TYPEIndustry: type 6               -1.044e+00  7.728e-01  -1.351
## ORGANIZATION_TYPEIndustry: type 7               -3.597e-01  2.816e-01  -1.277
## ORGANIZATION_TYPEIndustry: type 8               -4.583e-01  1.077e+00  -0.425
## ORGANIZATION_TYPEIndustry: type 9               -6.141e-01  2.666e-01  -2.303
## ORGANIZATION_TYPEInsurance                      -4.141e-01  3.641e-01  -1.137
## ORGANIZATION_TYPEKindergarten                   -2.604e-01  2.525e-01  -1.031
## ORGANIZATION_TYPELegal Services                  2.940e-01  3.974e-01   0.740
```

```
## ORGANIZATION_TYPEMedicine                          -2.872e-01  2.525e-01  -1.137
## ORGANIZATION_TYPEMilitary                          -7.424e-01  2.776e-01  -2.675
## ORGANIZATION_TYPEMobile                            -1.855e-01  3.784e-01  -0.490
## ORGANIZATION_TYPEOther                             -3.066e-01  2.474e-01  -1.239
## ORGANIZATION_TYPEPolice                            -6.397e-01  2.813e-01  -2.274
## ORGANIZATION_TYPEPostal                            -1.588e-01  2.658e-01  -0.597
## ORGANIZATION_TYPERealtor                            5.412e-01  3.431e-01   1.577
## ORGANIZATION_TYPEReligion                           2.584e-01  5.467e-01   0.473
## ORGANIZATION_TYPERestaurant                        -6.858e-02  2.644e-01  -0.259
## ORGANIZATION_TYPESchool                            -4.187e-01  2.517e-01  -1.663
## ORGANIZATION_TYPESecurity                          -3.232e-01  2.646e-01  -1.222
## ORGANIZATION_TYPESecurity Ministries               -5.352e-01  2.844e-01  -1.882
## ORGANIZATION_TYPESelf-employed                     -9.029e-02  2.451e-01  -0.368
## ORGANIZATION_TYPEServices                          -1.618e-01  2.822e-01  -0.573
## ORGANIZATION_TYPETelecom                           -2.906e-01  3.303e-01  -0.880
## ORGANIZATION_TYPETrade: type 1                     -7.032e-02  3.425e-01  -0.205
## ORGANIZATION_TYPETrade: type 2                     -6.417e-01  2.758e-01  -2.327
## ORGANIZATION_TYPETrade: type 3                     -1.041e-01  2.561e-01  -0.407
## ORGANIZATION_TYPETrade: type 4                     -1.505e+00  1.072e+00  -1.404
## ORGANIZATION_TYPETrade: type 5                     -1.123e+01  9.562e+01  -0.117
## ORGANIZATION_TYPETrade: type 6                     -5.597e-01  3.575e-01  -1.565
## ORGANIZATION_TYPETrade: type 7                     -1.348e-01  2.500e-01  -0.539
## ORGANIZATION_TYPETransport: type 1                 -1.631e+00  7.618e-01  -2.141
## ORGANIZATION_TYPETransport: type 2                 -3.182e-01  2.689e-01  -1.183
## ORGANIZATION_TYPETransport: type 3                  5.851e-01  2.706e-01   2.162
## ORGANIZATION_TYPETransport: type 4                 -2.007e-01  2.535e-01  -0.792
## ORGANIZATION_TYPEUnemployed                         2.101e+01  3.287e+02   0.064
## ORGANIZATION_TYPEUniversity                        -4.568e-01  3.006e-01  -1.519
## EXT_SOURCE_1                                       -6.169e-01  4.050e-02 -15.232
## EXT_SOURCE_2                                       -2.015e+00  4.851e-02 -41.526
## EXT_SOURCE_3                                       -2.149e+00  4.754e-02 -45.197
## APARTMENTS_MEDI                                    -4.082e-03  1.154e-01  -0.035
## YEARS_BUILD_MEDI                                   -1.507e-01  9.518e-02  -1.584
## COMMONAREA_MEDI                                     1.065e-01  1.393e-01   0.765
## ELEVATORS_MEDI                                     -6.328e-02  9.660e-02  -0.655
## ENTRANCES_MEDI                                     -1.401e-01  1.055e-01  -1.328
## FLOORSMAX_MEDI                                     -1.572e-01  1.021e-01  -1.539
## FLOORSMIN_MEDI                                      8.244e-02  6.845e-02   1.204
## LIVINGAPARTMENTS_MEDI                               3.676e-02  1.100e-01   0.334
## LIVINGAREA_MEDI                                     8.516e-02  1.124e-01   0.758
## NONLIVINGAPARTMENTS_MEDI                           -3.982e-01  2.281e-01  -1.746
## OBS_30_CNT_SOCIAL_CIRCLE                            5.179e-02  7.085e-02   0.731
## DEF_30_CNT_SOCIAL_CIRCLE                            1.572e-01  3.875e-02   4.057
## OBS_60_CNT_SOCIAL_CIRCLE                           -5.786e-02  7.158e-02  -0.808
## DEF_60_CNT_SOCIAL_CIRCLE                            5.402e-02  4.556e-02   1.186
## DAYS_LAST_PHONE_CHANGE                             -5.456e-05  1.283e-05  -4.254
## FLAG_DOCUMENT_31                                    2.466e-01  5.526e-02   4.462
## FLAG_DOCUMENT_61                                    1.624e-01  7.099e-02   2.288
## FLAG_DOCUMENT_81                                   -3.124e-02  6.795e-02  -0.460
## AMT_REQ_CREDIT_BUREAU_HOUR                         -6.843e-02  1.260e-01  -0.543
## AMT_REQ_CREDIT_BUREAU_DAY                           1.206e-01  8.787e-02   1.373
## AMT_REQ_CREDIT_BUREAU_WEEK                         -5.026e-02  5.126e-02  -0.980
## AMT_REQ_CREDIT_BUREAU_MON                          -1.705e-02  1.372e-02  -1.242
## AMT_REQ_CREDIT_BUREAU_QRT                          -2.840e-02  1.626e-02  -1.747
```

```
## AMT_REQ_CREDIT_BUREAU_YEAR                          1.196e-02  5.361e-03   2.230
## House_Attribute_Low_Variance                       -1.686e-02  6.295e-03  -2.678
##                                                    Pr(>|z|)
## (Intercept)                                        0.935477
## NAME_CONTRACT_TYPERevolving loans                  0.235919
## CODE_GENDERM                                        < 2e-16 ***
## FLAG_OWN_CARY                                       < 2e-16 ***
## FLAG_OWN_REALTYY                                   0.360203
## CNT_CHILDREN                                       0.059103 .
## AMT_INCOME_TOTAL                                   0.806348
## AMT_CREDIT                                          < 2e-16 ***
## AMT_ANNUITY                                         < 2e-16 ***
## AMT_GOODS_PRICE                                     < 2e-16 ***
## NAME_TYPE_SUITEFamily                              0.963938
## NAME_TYPE_SUITEGroup of people                     0.511867
## NAME_TYPE_SUITEOther_A                             0.740244
## NAME_TYPE_SUITEOther_B                             0.688993
## NAME_TYPE_SUITESpouse, partner                     0.313820
## NAME_TYPE_SUITEUnaccompanied                       0.961813
## NAME_INCOME_TYPECommercial associate               0.972286
## NAME_INCOME_TYPEMaternity leave                    0.963305
## NAME_INCOME_TYPEPensioner                          0.999497
## NAME_INCOME_TYPEState servant                      0.972258
## NAME_INCOME_TYPEStudent                            0.998572
## NAME_INCOME_TYPEUnemployed                         0.993574
## NAME_INCOME_TYPEWorking                            0.971994
## NAME_EDUCATION_TYPEHigher education                0.862192
## NAME_EDUCATION_TYPEIncomplete higher               0.861346
## NAME_EDUCATION_TYPELower secondary                 0.858148
## NAME_EDUCATION_TYPESecondary / secondary special   0.859204
## NAME_FAMILY_STATUSMarried                          1.44e-05 ***
## NAME_FAMILY_STATUSSeparated                        0.283623
## NAME_FAMILY_STATUSSingle / not married             0.126921
## NAME_FAMILY_STATUSUnknown                          0.984904
## NAME_FAMILY_STATUSWidow                            0.112199
## NAME_HOUSING_TYPEHouse / apartment                 0.546017
## NAME_HOUSING_TYPEMunicipal apartment               0.323923
## NAME_HOUSING_TYPEOffice apartment                  0.663664
## NAME_HOUSING_TYPERented apartment                  0.153906
## NAME_HOUSING_TYPEWith parents                      0.449076
## REGION_POPULATION_RELATIVE                         0.004417 **
## DAYS_BIRTH                                          < 2e-16 ***
## DAYS_REGISTRATION                                  0.000184 ***
## DAYS_ID_PUBLISH                                    8.16e-15 ***
## OWN_CAR_AGE                                        0.000681 ***
## FLAG_EMP_PHONE1                                    0.966487
## FLAG_WORK_PHONE1                                   7.25e-15 ***
## FLAG_PHONE1                                        3.50e-05 ***
## FLAG_EMAIL1                                        0.715954
## OCCUPATION_TYPECleaning staff                      0.005453 **
## OCCUPATION_TYPECooking staff                       0.070658 .
## OCCUPATION_TYPECore staff                          0.841827
## OCCUPATION_TYPEDrivers                             0.002376 **
## OCCUPATION_TYPEHigh skill tech staff               0.405917
```

```
## OCCUPATION_TYPEHR staff                          0.065507 .
## OCCUPATION_TYPEIT staff                          0.871555
## OCCUPATION_TYPELaborers                          0.003337 **
## OCCUPATION_TYPELow-skill Laborers                0.002234 **
## OCCUPATION_TYPEManagers                          0.635358
## OCCUPATION_TYPEMedicine staff                    0.714150
## OCCUPATION_TYPEPrivate service staff             0.910532
## OCCUPATION_TYPERealty agents                     0.649500
## OCCUPATION_TYPESales staff                       0.068905 .
## OCCUPATION_TYPESecretaries                       0.309986
## OCCUPATION_TYPESecurity staff                    0.011424 *
## OCCUPATION_TYPEUnemployed                        0.064803 .
## OCCUPATION_TYPEWaiters/barmen staff              0.040528 *
## CNT_FAM_MEMBERS                                        NA
## REGION_RATING_CLIENT2                            0.031799 *
## REGION_RATING_CLIENT3                            0.128516
## REGION_RATING_CLIENT_W_CITY2                     0.001241 **
## REGION_RATING_CLIENT_W_CITY3                     0.000442 ***
## WEEKDAY_APPR_PROCESS_STARTMONDAY                 0.009064 **
## WEEKDAY_APPR_PROCESS_STARTSATURDAY               0.067308 .
## WEEKDAY_APPR_PROCESS_STARTSUNDAY                 0.019231 *
## WEEKDAY_APPR_PROCESS_STARTTHURSDAY               0.243762
## WEEKDAY_APPR_PROCESS_STARTTUESDAY                0.512630
## WEEKDAY_APPR_PROCESS_STARTWEDNESDAY              0.765336
## HOUR_APPR_PROCESS_START                          0.820895
## REG_REGION_NOT_WORK_REGION1                      0.091505 .
## REG_CITY_NOT_LIVE_CITY1                          6.73e-05 ***
## REG_CITY_NOT_WORK_CITY1                          0.963456
## LIVE_CITY_NOT_WORK_CITY1                         0.510802
## ORGANIZATION_TYPEAgriculture                     0.264040
## ORGANIZATION_TYPEBank                            0.070149 .
## ORGANIZATION_TYPEBusiness Entity Type 1          0.162385
## ORGANIZATION_TYPEBusiness Entity Type 2          0.210576
## ORGANIZATION_TYPEBusiness Entity Type 3          0.468992
## ORGANIZATION_TYPECleaning                        0.966556
## ORGANIZATION_TYPEConstruction                    0.823794
## ORGANIZATION_TYPECulture                         0.841194
## ORGANIZATION_TYPEElectricity                     0.215395
## ORGANIZATION_TYPEEmergency                       0.361776
## ORGANIZATION_TYPEGovernment                      0.184684
## ORGANIZATION_TYPEHotel                           0.129554
## ORGANIZATION_TYPEHousing                         0.111960
## ORGANIZATION_TYPEIndustry: type 1                0.669789
## ORGANIZATION_TYPEIndustry: type 10               0.715333
## ORGANIZATION_TYPEIndustry: type 11               0.282710
## ORGANIZATION_TYPEIndustry: type 12               0.055278 .
## ORGANIZATION_TYPEIndustry: type 13               0.657130
## ORGANIZATION_TYPEIndustry: type 2                0.085575 .
## ORGANIZATION_TYPEIndustry: type 3                0.568993
## ORGANIZATION_TYPEIndustry: type 4                0.506831
## ORGANIZATION_TYPEIndustry: type 5                0.177871
## ORGANIZATION_TYPEIndustry: type 6                0.176597
## ORGANIZATION_TYPEIndustry: type 7                0.201478
## ORGANIZATION_TYPEIndustry: type 8                0.670608
```

```
## ORGANIZATION_TYPEIndustry: type 9               0.021263 *
## ORGANIZATION_TYPEInsurance                      0.255464
## ORGANIZATION_TYPEKindergarten                   0.302459
## ORGANIZATION_TYPELegal Services                 0.459509
## ORGANIZATION_TYPEMedicine                       0.255454
## ORGANIZATION_TYPEMilitary                       0.007478 **
## ORGANIZATION_TYPEMobile                         0.623972
## ORGANIZATION_TYPEOther                          0.215188
## ORGANIZATION_TYPEPolice                         0.022973 *
## ORGANIZATION_TYPEPostal                         0.550271
## ORGANIZATION_TYPERealtor                        0.114768
## ORGANIZATION_TYPEReligion                       0.636550
## ORGANIZATION_TYPERestaurant                     0.795332
## ORGANIZATION_TYPESchool                         0.096217 .
## ORGANIZATION_TYPESecurity                       0.221891
## ORGANIZATION_TYPESecurity Ministries            0.059868 .
## ORGANIZATION_TYPESelf-employed                  0.712593
## ORGANIZATION_TYPEServices                       0.566395
## ORGANIZATION_TYPETelecom                        0.379055
## ORGANIZATION_TYPETrade: type 1                  0.837323
## ORGANIZATION_TYPETrade: type 2                  0.019972 *
## ORGANIZATION_TYPETrade: type 3                  0.684347
## ORGANIZATION_TYPETrade: type 4                  0.160273
## ORGANIZATION_TYPETrade: type 5                  0.906499
## ORGANIZATION_TYPETrade: type 6                  0.117467
## ORGANIZATION_TYPETrade: type 7                  0.589750
## ORGANIZATION_TYPETransport: type 1              0.032257 *
## ORGANIZATION_TYPETransport: type 2              0.236671
## ORGANIZATION_TYPETransport: type 3              0.030586 *
## ORGANIZATION_TYPETransport: type 4              0.428435
## ORGANIZATION_TYPEUnemployed                     0.949029
## ORGANIZATION_TYPEUniversity                     0.128648
## EXT_SOURCE_1                                     < 2e-16 ***
## EXT_SOURCE_2                                     < 2e-16 ***
## EXT_SOURCE_3                                     < 2e-16 ***
## APARTMENTS_MEDI                                 0.971790
## YEARS_BUILD_MEDI                                0.113291
## COMMONAREA_MEDI                                 0.444559
## ELEVATORS_MEDI                                  0.512398
## ENTRANCES_MEDI                                  0.184335
## FLOORSMAX_MEDI                                  0.123833
## FLOORSMIN_MEDI                                  0.228457
## LIVINGAPARTMENTS_MEDI                           0.738227
## LIVINGAREA_MEDI                                 0.448476
## NONLIVINGAPARTMENTS_MEDI                        0.080860 .
## OBS_30_CNT_SOCIAL_CIRCLE                        0.464746
## DEF_30_CNT_SOCIAL_CIRCLE                        4.97e-05 ***
## OBS_60_CNT_SOCIAL_CIRCLE                        0.418945
## DEF_60_CNT_SOCIAL_CIRCLE                        0.235814
## DAYS_LAST_PHONE_CHANGE                          2.10e-05 ***
## FLAG_DOCUMENT_31                                8.10e-06 ***
## FLAG_DOCUMENT_61                                0.022158 *
## FLAG_DOCUMENT_81                                0.645695
## AMT_REQ_CREDIT_BUREAU_HOUR                      0.587073
```

```
## AMT_REQ_CREDIT_BUREAU_DAY                          0.169888
## AMT_REQ_CREDIT_BUREAU_WEEK                         0.326845
## AMT_REQ_CREDIT_BUREAU_MON                          0.214147
## AMT_REQ_CREDIT_BUREAU_QRT                          0.080677 .
## AMT_REQ_CREDIT_BUREAU_YEAR                         0.025743 *
## House_Attribute_Low_Variance                       0.007411 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 93561  on 166443  degrees of freedom
## Residual deviance: 83686  on 166280  degrees of freedom
## AIC: 84014
##
## Number of Fisher Scoring iterations: 12
```

```
# Predictions and Evaluation
test_pred_probs <- predict(logistic_model, newdata = test_data, type = "response")
test_predictions <- ifelse(test_pred_probs > 0.5, 1, 0)


# ROC-AUC
roc_obj <- roc(test_data$TARGET, test_pred_probs)
```

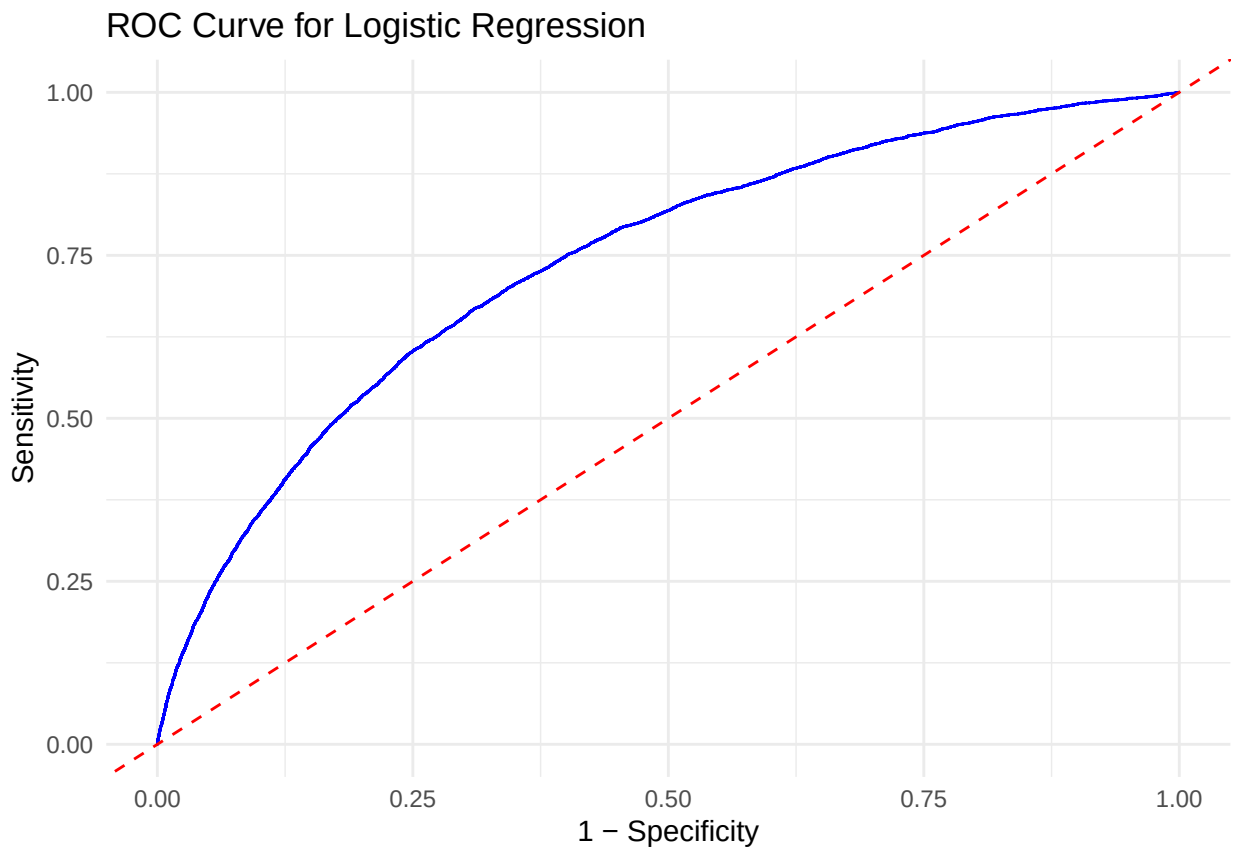
```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
roc_auc <- auc(roc_obj)
cat("ROC-AUC:", round(roc_auc, 4), "\n")
```

```
## ROC-AUC: 0.7391
```

```
ggplot(data.frame(FPR = 1 - roc_obj$specificities, TPR = roc_obj$sensitivities), aes(x = FPR, y = TPR)) +
  geom_line(color = "blue") +
  geom_abline(linetype = "dashed", color = "red") +
  labs(title = "ROC Curve for Logistic Regression", x = "1 - Specificity", y = "Sensitivity") +
  theme_minimal()
```




```
# Pseudo R²
null_model <- glm(TARGET ~ 1, data = train_data, family = binomial())
pseudo_r2 <- 1 - (logLik(logistic_model) / logLik(null_model))
cat("Pseudo R²:", round(pseudo_r2, 4), "\n")
```

```
## Pseudo R²: 0.1055
```

```
# Accuracy
accuracy <- mean(test_predictions == test_data$TARGET)
cat("Accuracy:", round(accuracy, 4), "\n")
```

```
## Accuracy: 0.919
```

```
table(train_data$TARGET)
```

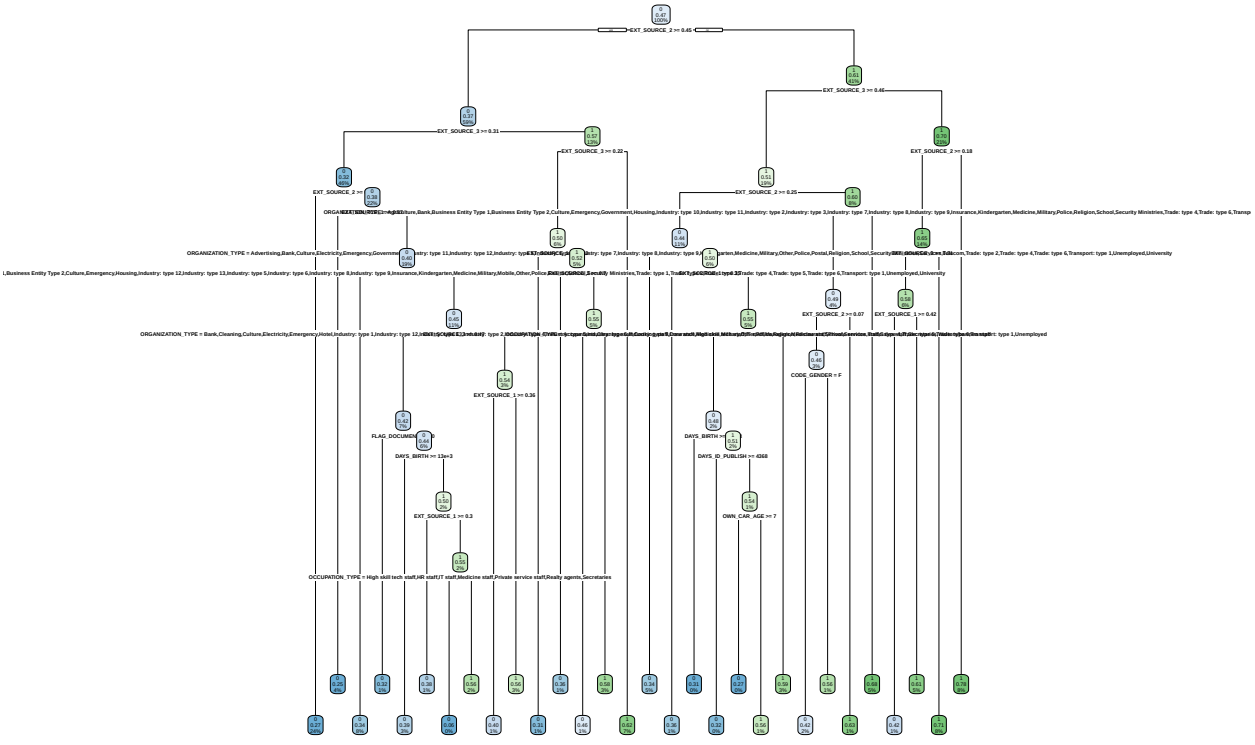
```
##
##      0      1
## 152972  13472
```

```
prop.table(table(train_data$TARGET))
```

```
##
##          0          1
## 0.91905986 0.08094014
```

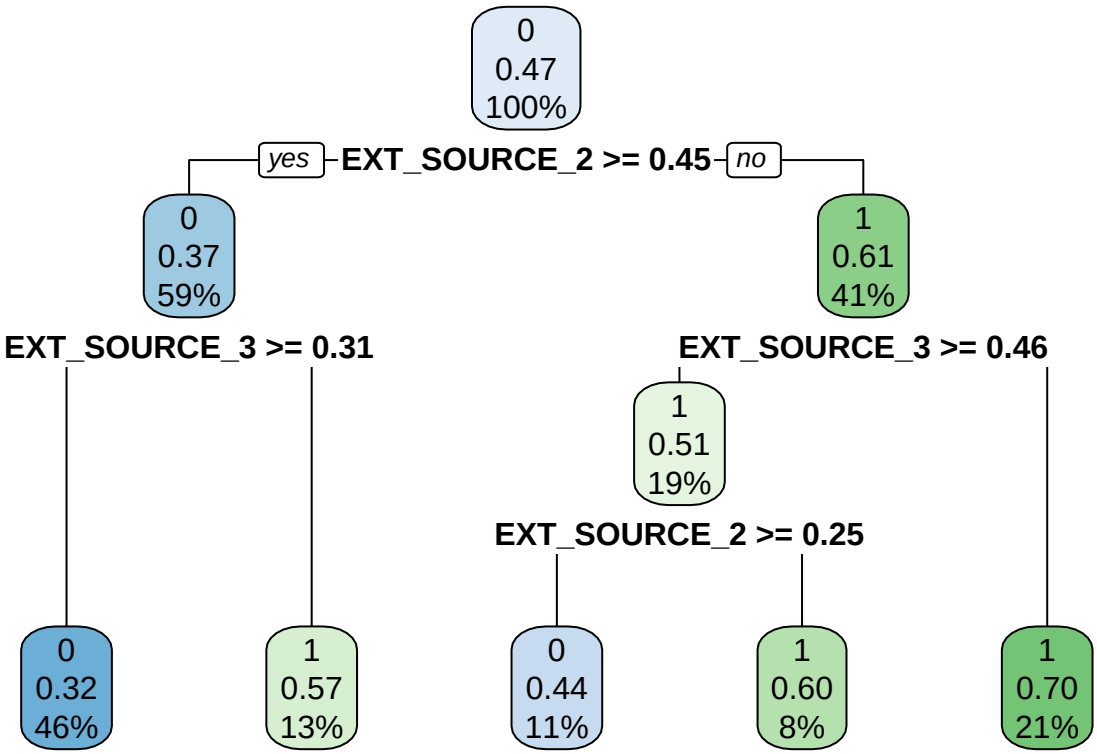
```
#Adjust the rpart.control parameters to encourage splitting.
tree_model <- rpart(
  TARGET ~ .,
  data = train_data,
  method = "class",
  control = rpart.control(minsplit = 2, cp = 0.001, maxdepth = 5)
)


#Handling Class Imbalance
class_weights <- ifelse(train_data$TARGET == 1, 10, 1)
tree_model <- rpart(
  TARGET ~ .,
  data = train_data,
  weights = class_weights,
  method = "class",
  control = rpart.control(cp = 0.001)
)
rpart.plot(tree_model)
```

```
## Warning: labs do not fit even at cex 0.15, there may be some overplotting
```




```
# Weighted Decision Tree for Class Imbalance
class_weights <- ifelse(train_data$TARGET == 1, 10, 1)
```

```r
tree_model_weighted <- rpart(TARGET ~ ., data = train_data, weights = class_weights, method = "class")
rpart.plot(tree_model_weighted)
```




```r
# Decision Tree Evaluation
tree_train_pred <- predict(tree_model, newdata = train_data, type = "class")
tree_test_pred <- predict(tree_model, newdata = test_data, type = "class")

weighted_train_pred <- predict(tree_model_weighted, newdata = train_data, type = "class")
weighted_test_pred <- predict(tree_model_weighted, newdata = test_data, type = "class")


cat("Decision Tree Accuracy - Train:", mean(tree_train_pred == train_data$TARGET), "\n")
```

```
## Decision Tree Accuracy - Train: 0.6969191
```

```r
cat("Decision Tree Accuracy - Test:", mean(tree_test_pred == test_data$TARGET), "\n")
```

```
## Decision Tree Accuracy - Test: 0.6909381
```

```r
cat("Weighted Decision Tree Accuracy - Train:", mean(weighted_train_pred == train_data$TARGET), "\n")
```

```
## Weighted Decision Tree Accuracy - Train: 0.7002595
```

```r
cat("Weighted Decision Tree Accuracy - Test:", mean(weighted_test_pred == test_data$TARGET), "\n")
```

```
## Weighted Decision Tree Accuracy - Test: 0.6946672
```