Home Credit

Mckay Flake

April 26th 2024

Contents

Introduction

The problem faced by the Home Credit Group centers around loan default. As a business, they extend loans to a higher risk segment of the population, therefore limiting loan defaults is crucial to the business operations. It is clear that the target variable for this problem is binary, Yes or No found in the target column in the train set. A couple of questions arise initially:

- It appears that there's a lot of missing data, how should this be handled? Are there any variables that cannot be used because of missing data?
- Are there significant outliers or mistakes within the data and how should they be handled?
- Which of the predictors are the most correlated to a change in the target variable? What variables can be removed?

Data Processing

```
# Package loading
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.2.3
```

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.2.3
```

```
## Loading required package: lattice
```

```r
library(rmarkdown)
library(tictoc)
```

```
## Warning: package 'tictoc' was built under R version 4.2.3
```

```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --

## v tibble  3.2.1     v dplyr   1.1.3
## v tidyr   1.3.0     v stringr 1.5.0
## v readr   2.1.3     v forcats 0.5.2
## v purrr   1.0.1

## Warning: package 'tibble' was built under R version 4.2.3

## Warning: package 'tidyr' was built under R version 4.2.3

## Warning: package 'dplyr' was built under R version 4.2.3

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x purrr::lift()   masks caret::lift()
```

```r
library(caret)
library(knitr)
library(skimr)
```

```
## Warning: package 'skimr' was built under R version 4.2.3
```

```r
library(janitor)
```

```
##
## Attaching package: 'janitor'
##
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```r
library(dplyr)

tic()

cloud_wd <- getwd()
setwd(cloud_wd)

# Data importing
df <- read.csv(file = "application_train.csv", stringsAsFactors = FALSE)  # Load the dataset
```

```
echo = FALSE

# Remove columns with missing data
threshold <- nrow(df) * 0.5 # Define threshold
df <- df[, colSums(is.na(df)) < threshold]

columns_to_remove <- names(df)[colSums(is.na(df)) > 100] # Identify columns with over 100 missing value
df <- df[, !names(df) %in% columns_to_remove] # Remove columns with over 100 missing values

columns_to_remove <- c("FLAG_DOCUMENT_2", "FLAG_DOCUMENT_3", "FLAG_DOCUMENT_4",
                       "FLAG_DOCUMENT_5", "FLAG_DOCUMENT_6", "FLAG_DOCUMENT_7",
                       "FLAG_DOCUMENT_8", "FLAG_DOCUMENT_9", "FLAG_DOCUMENT_10",
                       "FLAG_DOCUMENT_11", "FLAG_DOCUMENT_12", "FLAG_DOCUMENT_13",
                       "FLAG_DOCUMENT_14", "FLAG_DOCUMENT_15", "FLAG_DOCUMENT_16",
                       "FLAG_DOCUMENT_17", "FLAG_DOCUMENT_18", "FLAG_DOCUMENT_19",
                       "FLAG_DOCUMENT_20", "FLAG_DOCUMENT_21", "SK_ID_CURR", "FLAG_MOBIL")

df <- df[, !names(df) %in% columns_to_remove] # Remove specific columns

# Remove rows with null values
df <- na.omit(df)

# Factorize variables
vars_to_factor <- c("NAME_CONTRACT_TYPE", "CODE_GENDER", "FLAG_OWN_CAR", "FLAG_OWN_REALTY",
     "NAME_TYPE_SUITE", "NAME_INCOME_TYPE", "NAME_EDUCATION_TYPE",
     "NAME_FAMILY_STATUS", "NAME_HOUSING_TYPE", "FLAG_EMP_PHONE",
     "FLAG_WORK_PHONE", "FLAG_CONT_MOBILE", "FLAG_PHONE", "FLAG_EMAIL",
     "OCCUPATION_TYPE", "REGION_RATING_CLIENT", "REGION_RATING_CLIENT_W_CITY",
     "WEEKDAY_APPR_PROCESS_START", "REG_REGION_NOT_LIVE_REGION",
     "REG_REGION_NOT_WORK_REGION", "LIVE_REGION_NOT_WORK_REGION",
     "REG_CITY_NOT_LIVE_CITY", "REG_CITY_NOT_WORK_CITY",
     "LIVE_CITY_NOT_WORK_CITY", "ORGANIZATION_TYPE", "FONDKAPREMONT_MODE",
     "HOUSETYPE_MODE", "WALLSMATERIAL_MODE", "EMERGENCYSTATE_MODE", "TARGET")

df[vars_to_factor] <- lapply(df[vars_to_factor], factor)

head(df)
```

```
##   TARGET NAME_CONTRACT_TYPE CODE_GENDER FLAG_OWN_CAR FLAG_OWN_REALTY
## 1      1         Cash loans           M            N               Y
## 2      0         Cash loans           F            N               N
## 3      0    Revolving loans           M            Y               Y
## 4      0         Cash loans           F            N               Y
## 5      0         Cash loans           M            N               Y
## 6      0         Cash loans           M            N               Y
##   CNT_CHILDREN AMT_INCOME_TOTAL AMT_CREDIT AMT_ANNUITY NAME_TYPE_SUITE
## 1            0           202500   406597.5     24700.5   Unaccompanied
## 2            0           270000  1293502.5     35698.5          Family
## 3            0            67500   135000.0      6750.0   Unaccompanied
## 4            0           135000   312682.5     29686.5   Unaccompanied
## 5            0           121500   513000.0     21865.5   Unaccompanied
## 6            0            99000   490495.5     27517.5 Spouse, partner
##   NAME_INCOME_TYPE           NAME_EDUCATION_TYPE   NAME_FAMILY_STATUS
```

```
## 1          Working Secondary / secondary special Single / not married
## 2    State servant              Higher education              Married
## 3          Working Secondary / secondary special Single / not married
## 4          Working Secondary / secondary special       Civil marriage
## 5          Working Secondary / secondary special Single / not married
## 6    State servant Secondary / secondary special              Married
##   NAME_HOUSING_TYPE REGION_POPULATION_RELATIVE DAYS_BIRTH DAYS_EMPLOYED
## 1 House / apartment                   0.018801      -9461          -637
## 2 House / apartment                   0.003541     -16765         -1188
## 3 House / apartment                   0.010032     -19046          -225
## 4 House / apartment                   0.008019     -19005         -3039
## 5 House / apartment                   0.028663     -19932         -3038
## 6 House / apartment                   0.035792     -16941         -1588
##   DAYS_REGISTRATION DAYS_ID_PUBLISH FLAG_EMP_PHONE FLAG_WORK_PHONE
## 1             -3648           -2120              1               0
## 2             -1186            -291              1               0
## 3             -4260           -2531              1               1
## 4             -9833           -2437              1               0
## 5             -4311           -3458              1               0
## 6             -4970            -477              1               1
##   FLAG_CONT_MOBILE FLAG_PHONE FLAG_EMAIL OCCUPATION_TYPE CNT_FAM_MEMBERS
## 1                1          1          0       Laborers               1
## 2                1          1          0     Core staff               2
## 3                1          1          0       Laborers               1
## 4                1          0          0       Laborers               2
## 5                1          0          0     Core staff               1
## 6                1          1          0       Laborers               2
##   REGION_RATING_CLIENT REGION_RATING_CLIENT_W_CITY WEEKDAY_APPR_PROCESS_START
## 1                    2                           2                  WEDNESDAY
## 2                    1                           1                     MONDAY
## 3                    2                           2                     MONDAY
## 4                    2                           2                  WEDNESDAY
## 5                    2                           2                   THURSDAY
## 6                    2                           2                  WEDNESDAY
##   HOUR_APPR_PROCESS_START REG_REGION_NOT_LIVE_REGION REG_REGION_NOT_WORK_REGION
## 1                      10                          0                          0
## 2                      11                          0                          0
## 3                       9                          0                          0
## 4                      17                          0                          0
## 5                      11                          0                          0
## 6                      16                          0                          0
##   LIVE_REGION_NOT_WORK_REGION REG_CITY_NOT_LIVE_CITY REG_CITY_NOT_WORK_CITY
## 1                           0                      0                      0
## 2                           0                      0                      0
## 3                           0                      0                      0
## 4                           0                      0                      0
## 5                           0                      0                      1
## 6                           0                      0                      0
##   LIVE_CITY_NOT_WORK_CITY      ORGANIZATION_TYPE FONDKAPREMONT_MODE
## 1                       0 Business Entity Type 3   reg oper account
## 2                       0                 School   reg oper account
## 3                       0             Government
## 4                       0 Business Entity Type 3
## 5                       1               Religion
```

```
## 6                       0                  Other
##   HOUSETYPE_MODE WALLSMATERIAL_MODE EMERGENCYSTATE_MODE DAYS_LAST_PHONE_CHANGE
## 1 block of flats       Stone, brick                  No                  -1134
## 2 block of flats              Block                  No                   -828
## 3                                                                         -815
## 4                                                                         -617
## 5                                                                        -1106
## 6                                                                        -2536
```

#Discussion of Missing Data

```
# Identify variables missing at least half of their data
missing_prop <- colSums(is.na(df)) / nrow(df)
threshold <- 0.5
cols_with_high_missing <- names(missing_prop[missing_prop > threshold])
print(cols_with_high_missing)
```

```
## character(0)
```

```
# Remove variables missing half of their data
threshold <- nrow(df) * 0.5 # Define threshold
df <- df[, colSums(is.na(df)) < threshold]

# Check for remaining missing data
missing_data <- colSums(is.na(df))
cols_with_missing <- names(missing_data[missing_data > 0])
print(missing_data)
```

```
##                      TARGET          NAME_CONTRACT_TYPE
##                           0                           0
##                 CODE_GENDER                FLAG_OWN_CAR
##                           0                           0
##             FLAG_OWN_REALTY                CNT_CHILDREN
##                           0                           0
##            AMT_INCOME_TOTAL                  AMT_CREDIT
##                           0                           0
##                  AMT_ANNUITY             NAME_TYPE_SUITE
##                           0                           0
##            NAME_INCOME_TYPE         NAME_EDUCATION_TYPE
##                           0                           0
##          NAME_FAMILY_STATUS           NAME_HOUSING_TYPE
##                           0                           0
##  REGION_POPULATION_RELATIVE                 DAYS_BIRTH
##                           0                           0
##               DAYS_EMPLOYED           DAYS_REGISTRATION
##                           0                           0
##             DAYS_ID_PUBLISH              FLAG_EMP_PHONE
##                           0                           0
##             FLAG_WORK_PHONE            FLAG_CONT_MOBILE
##                           0                           0
##                  FLAG_PHONE                 FLAG_EMAIL
##                           0                           0
##             OCCUPATION_TYPE             CNT_FAM_MEMBERS
```

```
##                            0                            0
##        REGION_RATING_CLIENT REGION_RATING_CLIENT_W_CITY
##                            0                            0
##  WEEKDAY_APPR_PROCESS_START     HOUR_APPR_PROCESS_START
##                            0                            0
##  REG_REGION_NOT_LIVE_REGION  REG_REGION_NOT_WORK_REGION
##                            0                            0
## LIVE_REGION_NOT_WORK_REGION      REG_CITY_NOT_LIVE_CITY
##                            0                            0
##      REG_CITY_NOT_WORK_CITY     LIVE_CITY_NOT_WORK_CITY
##                            0                            0
##           ORGANIZATION_TYPE          FONDKAPREMONT_MODE
##                            0                            0
##              HOUSETYPE_MODE          WALLSMATERIAL_MODE
##                            0                            0
##         EMERGENCYSTATE_MODE      DAYS_LAST_PHONE_CHANGE
##                            0                            0
```

```r
# Remove variables with over 100 missing values
columns_to_remove <- names(df)[colSums(is.na(df)) > 100]
df <- df[, !names(df) %in% columns_to_remove]

# Check remaining nulls
missing_data <- colSums(is.na(df))
cols_with_missing <- names(missing_data[missing_data > 0])
print(missing_data)
```

```
##                      TARGET          NAME_CONTRACT_TYPE
##                           0                           0
##                 CODE_GENDER               FLAG_OWN_CAR
##                           0                           0
##             FLAG_OWN_REALTY               CNT_CHILDREN
##                           0                           0
##            AMT_INCOME_TOTAL                 AMT_CREDIT
##                           0                           0
##                  AMT_ANNUITY             NAME_TYPE_SUITE
##                           0                           0
##            NAME_INCOME_TYPE         NAME_EDUCATION_TYPE
##                           0                           0
##          NAME_FAMILY_STATUS           NAME_HOUSING_TYPE
##                           0                           0
##  REGION_POPULATION_RELATIVE                 DAYS_BIRTH
##                           0                           0
##               DAYS_EMPLOYED           DAYS_REGISTRATION
##                           0                           0
##             DAYS_ID_PUBLISH              FLAG_EMP_PHONE
##                           0                           0
##             FLAG_WORK_PHONE            FLAG_CONT_MOBILE
##                           0                           0
##                  FLAG_PHONE                 FLAG_EMAIL
##                           0                           0
##             OCCUPATION_TYPE             CNT_FAM_MEMBERS
##                           0                           0
##        REGION_RATING_CLIENT REGION_RATING_CLIENT_W_CITY
```

```
##                            0                            0
##  WEEKDAY_APPR_PROCESS_START     HOUR_APPR_PROCESS_START
##                            0                            0
##  REG_REGION_NOT_LIVE_REGION  REG_REGION_NOT_WORK_REGION
##                            0                            0
## LIVE_REGION_NOT_WORK_REGION      REG_CITY_NOT_LIVE_CITY
##                            0                            0
##      REG_CITY_NOT_WORK_CITY     LIVE_CITY_NOT_WORK_CITY
##                            0                            0
##           ORGANIZATION_TYPE          FONDKAPREMONT_MODE
##                            0                            0
##              HOUSETYPE_MODE          WALLSMATERIAL_MODE
##                            0                            0
##         EMERGENCYSTATE_MODE      DAYS_LAST_PHONE_CHANGE
##                            0                            0
```

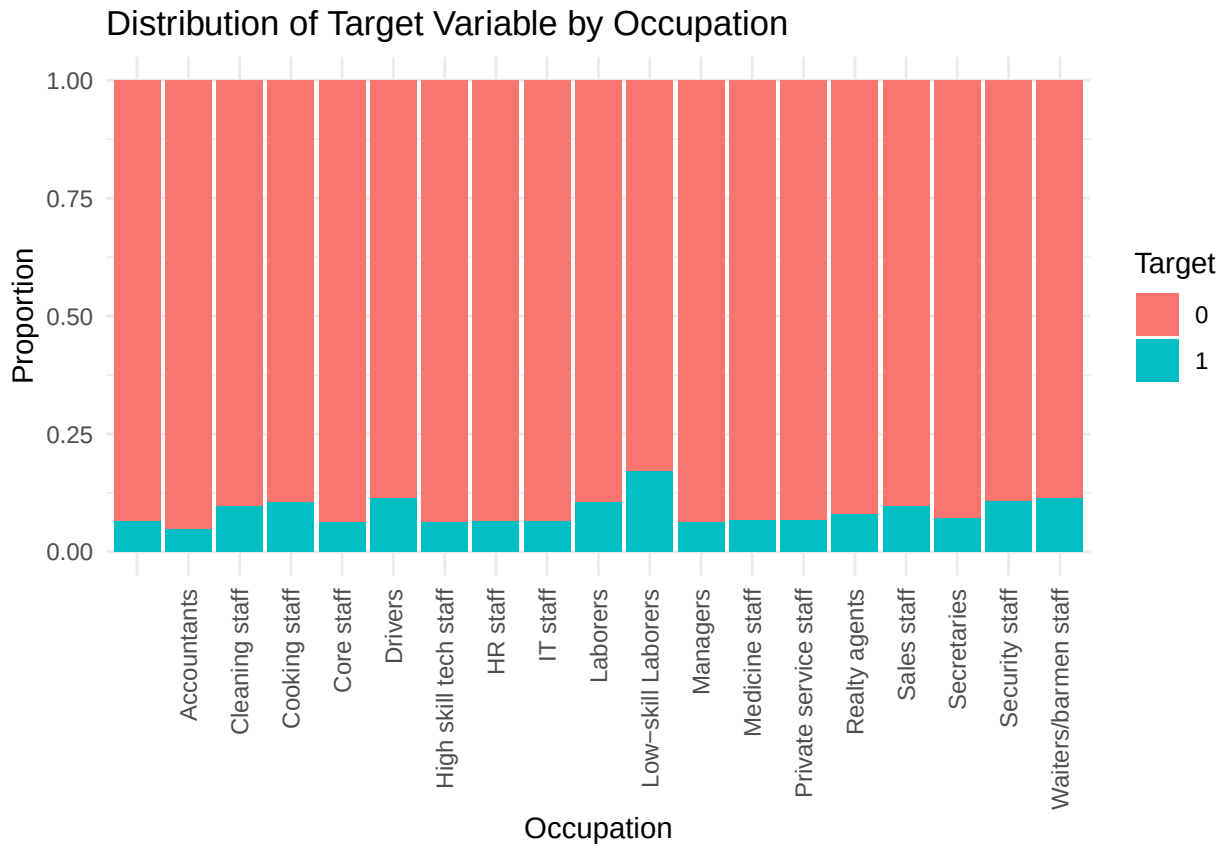
#Exploratory Visualizations

```
# Numeric Variables Summary
numeric_df <- df[sapply(df, is.numeric)]
summary(numeric_df)
```

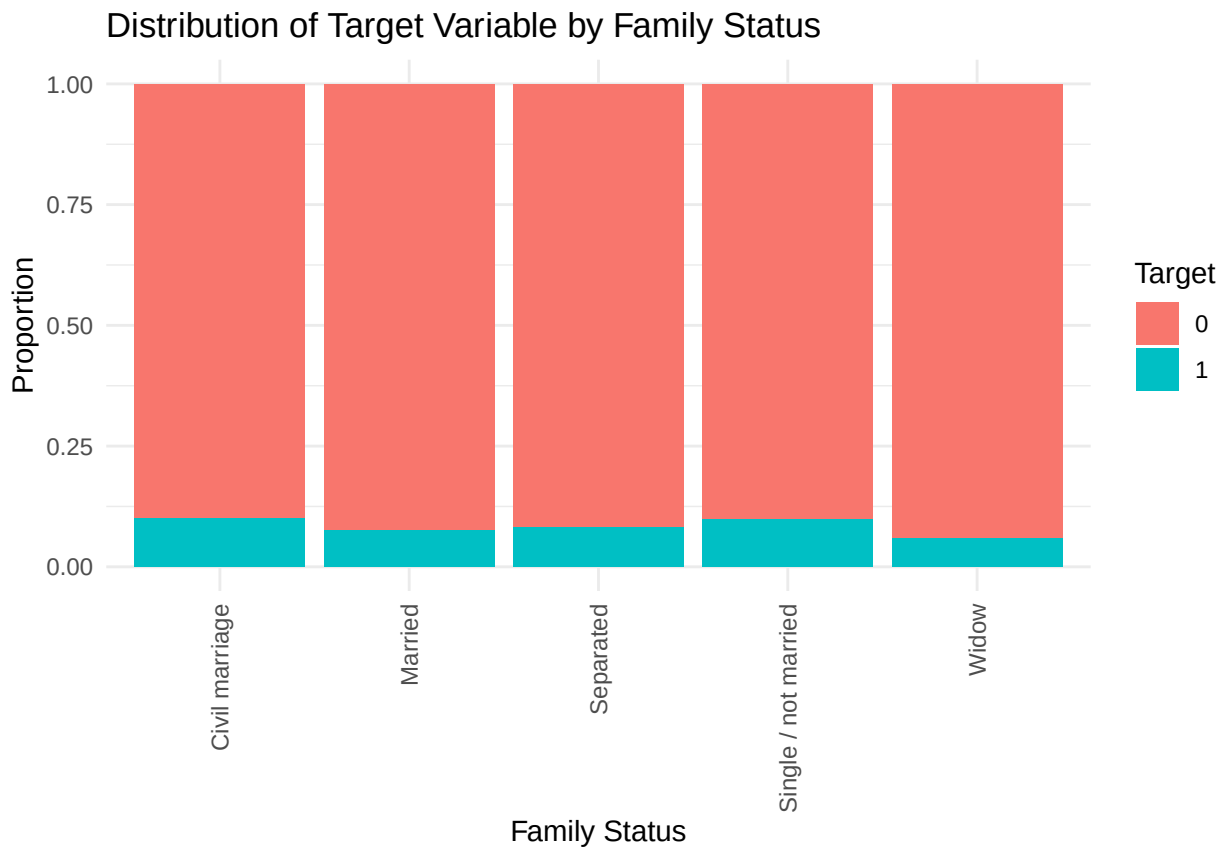
```
##   CNT_CHILDREN     AMT_INCOME_TOTAL      AMT_CREDIT       AMT_ANNUITY
##  Min.   : 0.0000   Min.   :    25650   Min.   :  45000   Min.   :  1616
##  1st Qu.: 0.0000   1st Qu.:   112500   1st Qu.: 270000   1st Qu.: 16524
##  Median : 0.0000   Median :   146905   Median : 513531   Median : 24903
##  Mean   : 0.4171   Mean   :   168796   Mean   : 599028   Mean   : 27109
##  3rd Qu.: 1.0000   3rd Qu.:   202500   3rd Qu.: 808650   3rd Qu.: 34596
##  Max.   :19.0000   Max.   :117000000   Max.   :4050000   Max.   :258026
##  REGION_POPULATION_RELATIVE   DAYS_BIRTH     DAYS_EMPLOYED    DAYS_REGISTRATION
##  Min.   :0.00029            Min.   :-25229   Min.   :-17912   Min.   :-24672
##  1st Qu.:0.01001            1st Qu.:-19682   1st Qu.: -2760   1st Qu.: -7479
##  Median :0.01885            Median :-15750   Median : -1213   Median : -4504
##  Mean   :0.02087            Mean   :-16037   Mean   : 63818   Mean   : -4986
##  3rd Qu.:0.02866            3rd Qu.:-12413   3rd Qu.:  -289   3rd Qu.: -2010
##  Max.   :0.07251            Max.   : -7489   Max.   :365243   Max.   :     0
##  DAYS_ID_PUBLISH CNT_FAM_MEMBERS  HOUR_APPR_PROCESS_START
##  Min.   :-7197   Min.   : 1.000   Min.   : 0.00
##  1st Qu.:-4299   1st Qu.: 2.000   1st Qu.:10.00
##  Median :-3254   Median : 2.000   Median :12.00
##  Mean   :-2994   Mean   : 2.153   Mean   :12.06
##  3rd Qu.:-1720   3rd Qu.: 3.000   3rd Qu.:14.00
##  Max.   :    0   Max.   :20.000   Max.   :23.00
##  DAYS_LAST_PHONE_CHANGE
##  Min.   :-4292.0
##  1st Qu.:-1570.0
##  Median : -757.0
##  Mean   : -962.9
##  3rd Qu.: -274.0
##  Max.   :    0.0
```

```
# Plot Distribution of Target Variable by Occupation
ggplot(df, aes(x = OCCUPATION_TYPE, fill = factor(TARGET))) +
```

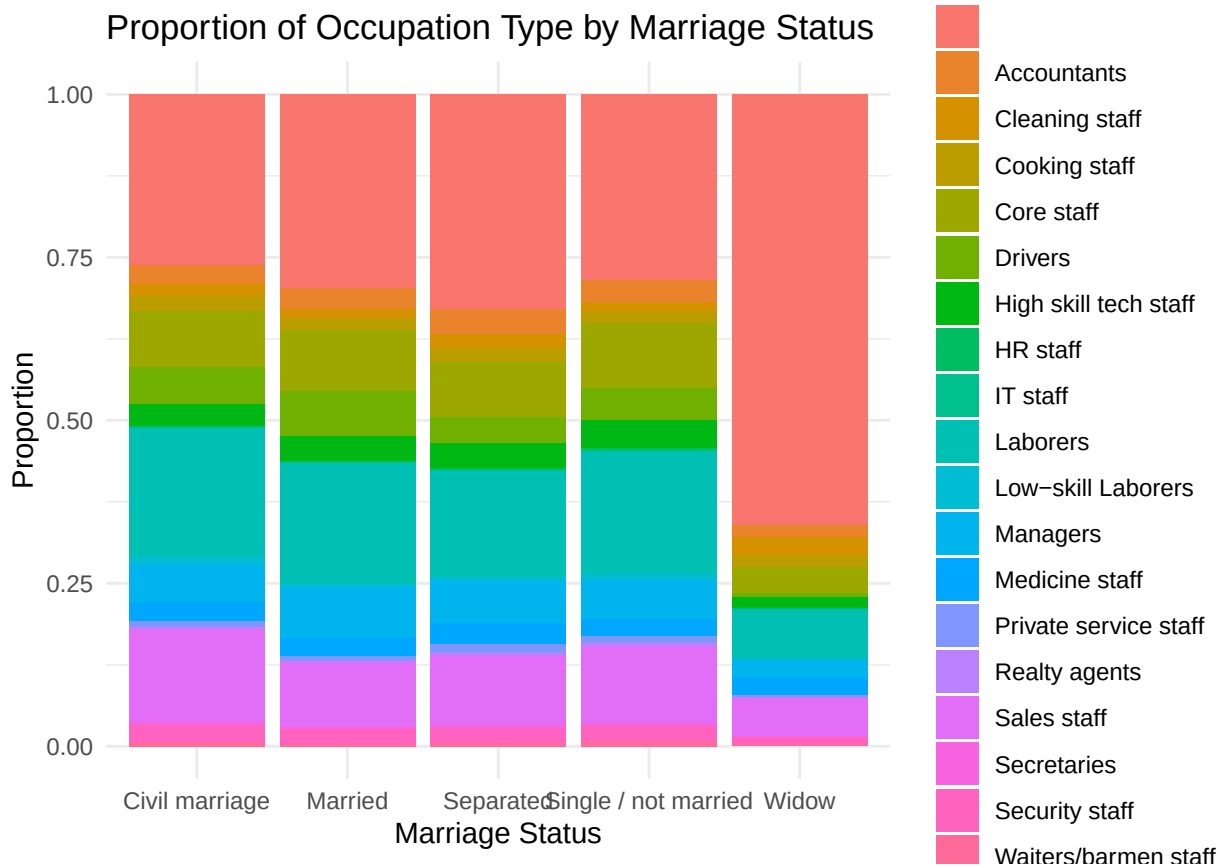
```
  geom_bar(position = "fill") +
  labs(title = "Distribution of Target Variable by Occupation",
       x = "Occupation",
       y = "Proportion",
       fill = "Target") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```




```
# Plot Distribution of Target Variable by Family Status
ggplot(df, aes(x = NAME_FAMILY_STATUS, fill = factor(TARGET))) +
  geom_bar(position = "fill") +
  labs(title = "Distribution of Target Variable by Family Status",
       x = "Family Status",
       y = "Proportion",
       fill = "Target") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```




```r
# Plot Proportion of Occupation Type by Marriage Status
ggplot(df, aes(x = NAME_FAMILY_STATUS, fill = OCCUPATION_TYPE)) +
  geom_bar(position = "fill") +
  labs(title = "Proportion of Occupation Type by Marriage Status",
       x = "Marriage Status",
       y = "Proportion",
       fill = "Occupation Type") +
  theme_minimal()
```



Modeling

```
library(pROC)
```

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

##
## Attaching package: 'pROC'

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

```
# Split data into training and validation sets
set.seed(123)
trainIndex <- createDataPartition(df$TARGET, p = .8,
                                  list = FALSE,
                                  times = 1)
train <- df[trainIndex,]
validation <- df[-trainIndex,]

# Fit logistic regression model
```

```r
model <- glm(TARGET ~ . + DAYS_BIRTH * DAYS_EMPLOYED, data = train, family = "binomial")

# Predict on validation set
pred <- predict(model, newdata = validation, type = "response")
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading
```

```r
# Calculate AUC
auc <- roc(validation$TARGET, pred)
```

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

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

```r
print(paste("Model AUC:", auc(auc)))
```

```
## [1] "Model AUC: 0.678602187577521"
```