Home Credit

Andy Pan

2024-11-01

Contents

1 Business Problem Statement

Home Credit aims to promote financial inclusion by offering safe and accessible loans to unbanked individuals. A significant challenge in developing markets is the inability to assess creditworthiness due to insufficient or non-existent credit histories. This often leads to loan rejections or reliance on predatory lenders. To address this, Home Credit seeks a reliable, data-driven method to predict repayment abilities using alternative data sources, such as telecommunications and transactional information. The goal is to reduce default rates while increasing loan approvals for creditworthy clients.

2 Load Libraries and Data

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
```

```
## v purrr     1.0.2
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(caret)
```

```
## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(rpart)
library(rpart.plot)
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
##
## Attaching package: 'pROC'
##
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
library(randomForest)
```

```
## randomForest 4.7-1.2
## Type rfNews() to see new features/changes/bug fixes.
##
## Attaching package: 'randomForest'
##
## The following object is masked from 'package:dplyr':
##
##     combine
##
## The following object is masked from 'package:ggplot2':
##
##     margin
```

3 Datasets

This dataset provides detailed information about consumers who have received credit from Home Credit. It includes both demographic and business-related data to facilitate the credit approval process.

The primary dataset, application_train, serves as the foundation for building predictive models. It can also be enhanced with additional insights derived from this table or other external data sources. This table contains 122 columns and 307,511 rows, with data types categorized as character, numeric, or integer.

A comprehensive description of each variable is available in a CSV file, which is not included here due to its length. For full access to these descriptions, visit: Kaggle - Home Credit Default Risk Dataset.

```
# Load the training and test datasets
Home_train <- read.csv("application_train.csv")
Home_test <- read.csv("application_test.csv")
# Preview the structure of the datasets
str(Home_train)
```

```
## 'data.frame':    307511 obs. of  122 variables:
##  $ SK_ID_CURR                  : int  100002 100003 100004 100006 100007 100008 100009 100010 100011
##  $ TARGET                      : int  1 0 0 0 0 0 0 0 0 0 ...
##  $ NAME_CONTRACT_TYPE          : chr  "Cash loans" "Cash loans" "Revolving loans" "Cash loans" ...
##  $ CODE_GENDER                 : chr  "M" "F" "M" "F" ...
##  $ FLAG_OWN_CAR                : chr  "N" "N" "Y" "N" ...
##  $ FLAG_OWN_REALTY             : chr  "Y" "N" "Y" "Y" ...
##  $ CNT_CHILDREN                : int  0 0 0 0 0 0 1 0 0 0 ...
##  $ AMT_INCOME_TOTAL            : num  202500 270000 67500 135000 121500 ...
##  $ AMT_CREDIT                  : num  406598 1293503 135000 312683 513000 ...
##  $ AMT_ANNUITY                 : num  24701 35699 6750 29687 21866 ...
##  $ AMT_GOODS_PRICE             : num  351000 1129500 135000 297000 513000 ...
##  $ NAME_TYPE_SUITE             : chr  "Unaccompanied" "Family" "Unaccompanied" "Unaccompanied" ...
##  $ NAME_INCOME_TYPE            : chr  "Working" "State servant" "Working" "Working" ...
##  $ NAME_EDUCATION_TYPE         : chr  "Secondary / secondary special" "Higher education" "Secondary /
##  $ NAME_FAMILY_STATUS          : chr  "Single / not married" "Married" "Single / not married" "Civil
##  $ NAME_HOUSING_TYPE           : chr  "House / apartment" "House / apartment" "House / apartment" "Ho
##  $ REGION_POPULATION_RELATIVE  : num  0.0188 0.00354 0.01003 0.00802 0.02866 ...
##  $ DAYS_BIRTH                  : int  -9461 -16765 -19046 -19005 -19932 -16941 -13778 -18850 -20099 -
##  $ DAYS_EMPLOYED               : int  -637 -1188 -225 -3039 -3038 -1588 -3130 -449 365243 -2019 ...
##  $ DAYS_REGISTRATION           : num  -3648 -1186 -4260 -9833 -4311 ...
##  $ DAYS_ID_PUBLISH             : int  -2120 -291 -2531 -2437 -3458 -477 -619 -2379 -3514 -3992 ...
##  $ OWN_CAR_AGE                 : num  NA NA 26 NA NA NA 17 8 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 0 1 ...
##  $ FLAG_WORK_PHONE             : int  0 0 1 0 0 1 0 1 0 0 ...
##  $ FLAG_CONT_MOBILE            : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ FLAG_PHONE                  : int  1 1 1 0 0 1 1 0 0 0 ...
##  $ FLAG_EMAIL                  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ OCCUPATION_TYPE             : chr  "Laborers" "Core staff" "Laborers" "Laborers" ...
##  $ CNT_FAM_MEMBERS             : num  1 2 1 2 1 2 3 2 2 1 ...
##  $ REGION_RATING_CLIENT        : int  2 1 2 2 2 2 2 3 2 2 ...
##  $ REGION_RATING_CLIENT_W_CITY : int  2 1 2 2 2 2 2 3 2 2 ...
##  $ WEEKDAY_APPR_PROCESS_START  : chr  "WEDNESDAY" "MONDAY" "MONDAY" "WEDNESDAY" ...
##  $ HOUR_APPR_PROCESS_START     : int  10 11 9 17 11 16 16 16 14 8 ...
##  $ 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 0 0 0 0 0 0 0 ...
##  $ LIVE_REGION_NOT_WORK_REGION : int  0 0 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 0 1 0 0 1 0 0 ...
##  $ LIVE_CITY_NOT_WORK_CITY     : int  0 0 0 0 1 0 0 1 0 0 ...
##  $ ORGANIZATION_TYPE           : chr  "Business Entity Type 3" "School" "Government" "Business Entity
##  $ EXT_SOURCE_1                : num  0.083 0.311 NA NA NA ...
##  $ EXT_SOURCE_2                : num  0.263 0.622 0.556 0.65 0.323 ...
##  $ EXT_SOURCE_3                : num  0.139 NA 0.73 NA NA ...
```

```
##  $ APARTMENTS_AVG              : num  0.0247 0.0959 NA NA NA NA NA NA NA NA ...
##  $ BASEMENTAREA_AVG            : num  0.0369 0.0529 NA NA NA NA NA NA NA NA ...
##  $ YEARS_BEGINEXPLUATATION_AVG : num  0.972 0.985 NA NA NA ...
##  $ YEARS_BUILD_AVG             : num  0.619 0.796 NA NA NA ...
##  $ COMMONAREA_AVG              : num  0.0143 0.0605 NA NA NA NA NA NA NA NA ...
##  $ ELEVATORS_AVG               : num  0 0.08 NA NA NA NA NA NA NA NA ...
##  $ ENTRANCES_AVG               : num  0.069 0.0345 NA NA NA NA NA NA NA NA ...
##  $ FLOORSMAX_AVG               : num  0.0833 0.2917 NA NA NA ...
##  $ FLOORSMIN_AVG               : num  0.125 0.333 NA NA NA ...
##  $ LANDAREA_AVG                : num  0.0369 0.013 NA NA NA NA NA NA NA NA ...
##  $ LIVINGAPARTMENTS_AVG        : num  0.0202 0.0773 NA NA NA NA NA NA NA NA ...
##  $ LIVINGAREA_AVG              : num  0.019 0.0549 NA NA NA NA NA NA NA NA ...
##  $ NONLIVINGAPARTMENTS_AVG     : num  0 0.0039 NA NA NA NA NA NA NA NA ...
##  $ NONLIVINGAREA_AVG           : num  0 0.0098 NA NA NA NA NA NA NA NA ...
##  $ APARTMENTS_MODE             : num  0.0252 0.0924 NA NA NA NA NA NA NA NA ...
##  $ BASEMENTAREA_MODE           : num  0.0383 0.0538 NA NA NA NA NA NA NA NA ...
##  $ YEARS_BEGINEXPLUATATION_MODE: num  0.972 0.985 NA NA NA ...
##  $ YEARS_BUILD_MODE            : num  0.634 0.804 NA NA NA ...
##  $ COMMONAREA_MODE             : num  0.0144 0.0497 NA NA NA NA NA NA NA NA ...
##  $ ELEVATORS_MODE              : num  0 0.0806 NA NA NA NA NA NA NA NA ...
##  $ ENTRANCES_MODE              : num  0.069 0.0345 NA NA NA NA NA NA NA NA ...
##  $ FLOORSMAX_MODE              : num  0.0833 0.2917 NA NA NA ...
##  $ FLOORSMIN_MODE              : num  0.125 0.333 NA NA NA ...
##  $ LANDAREA_MODE               : num  0.0377 0.0128 NA NA NA NA NA NA NA NA ...
##  $ LIVINGAPARTMENTS_MODE       : num  0.022 0.079 NA NA NA NA NA NA NA NA ...
##  $ LIVINGAREA_MODE             : num  0.0198 0.0554 NA NA NA NA NA NA NA NA ...
##  $ NONLIVINGAPARTMENTS_MODE    : num  0 0 NA NA NA NA NA NA NA NA ...
##  $ NONLIVINGAREA_MODE          : num  0 0 NA NA NA NA NA NA NA NA ...
##  $ APARTMENTS_MEDI             : num  0.025 0.0968 NA NA NA NA NA NA NA NA ...
##  $ BASEMENTAREA_MEDI           : num  0.0369 0.0529 NA NA NA NA NA NA NA NA ...
##  $ YEARS_BEGINEXPLUATATION_MEDI: num  0.972 0.985 NA NA NA ...
##  $ YEARS_BUILD_MEDI            : num  0.624 0.799 NA NA NA ...
##  $ COMMONAREA_MEDI             : num  0.0144 0.0608 NA NA NA NA NA NA NA NA ...
##  $ ELEVATORS_MEDI              : num  0 0.08 NA NA NA NA NA NA NA NA ...
##  $ ENTRANCES_MEDI              : num  0.069 0.0345 NA NA NA NA NA NA NA NA ...
##  $ FLOORSMAX_MEDI              : num  0.0833 0.2917 NA NA NA ...
##  $ FLOORSMIN_MEDI              : num  0.125 0.333 NA NA NA ...
##  $ LANDAREA_MEDI               : num  0.0375 0.0132 NA NA NA NA NA NA NA NA ...
##  $ LIVINGAPARTMENTS_MEDI       : num  0.0205 0.0787 NA NA NA NA NA NA NA NA ...
##  $ LIVINGAREA_MEDI             : num  0.0193 0.0558 NA NA NA NA NA NA NA NA ...
##  $ NONLIVINGAPARTMENTS_MEDI    : num  0 0.0039 NA NA NA NA NA NA NA NA ...
##  $ NONLIVINGAREA_MEDI          : num  0 0.01 NA NA NA NA NA NA NA NA ...
##  $ FONDKAPREMONT_MODE          : chr  "reg oper account" "reg oper account" "" "" ...
##  $ HOUSETYPE_MODE              : chr  "block of flats" "block of flats" "" "" ...
##  $ TOTALAREA_MODE              : num  0.0149 0.0714 NA NA NA NA NA NA NA NA ...
##  $ WALLSMATERIAL_MODE          : chr  "Stone, brick" "Block" "" "" ...
##  $ EMERGENCYSTATE_MODE         : chr  "No" "No" "" "" ...
##  $ OBS_30_CNT_SOCIAL_CIRCLE    : num  2 1 0 2 0 0 1 2 1 2 ...
##  $ DEF_30_CNT_SOCIAL_CIRCLE    : num  2 0 0 0 0 0 0 0 0 0 ...
##  $ OBS_60_CNT_SOCIAL_CIRCLE    : num  2 1 0 2 0 0 1 2 1 2 ...
##  $ DEF_60_CNT_SOCIAL_CIRCLE    : num  2 0 0 0 0 0 0 0 0 0 ...
##  $ DAYS_LAST_PHONE_CHANGE      : num  -1134 -828 -815 -617 -1106 ...
##  $ FLAG_DOCUMENT_2             : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ FLAG_DOCUMENT_3             : int  1 1 0 1 0 1 0 1 1 0 ...
```

```
##  $ FLAG_DOCUMENT_4             : int  0 0 0 0 0 0 0 0 0 0 ...
##   [list output truncated]
```

```
str(Home_test)
```

```
## 'data.frame':    48744 obs. of  121 variables:
##  $ SK_ID_CURR                  : int  100001 100005 100013 100028 100038 100042 100057 100065 100066
##  $ NAME_CONTRACT_TYPE          : chr  "Cash loans" "Cash loans" "Cash loans" "Cash loans" ...
##  $ CODE_GENDER                 : chr  "F" "M" "M" "F" ...
##  $ FLAG_OWN_CAR                : chr  "N" "N" "Y" "N" ...
##  $ FLAG_OWN_REALTY             : chr  "Y" "Y" "Y" "Y" ...
##  $ CNT_CHILDREN                : int  0 0 0 2 1 0 2 0 0 1 ...
##  $ AMT_INCOME_TOTAL            : num  135000 99000 202500 315000 180000 ...
##  $ AMT_CREDIT                  : num  568800 222768 663264 1575000 625500 ...
##  $ AMT_ANNUITY                 : num  20561 17370 69777 49019 32067 ...
##  $ AMT_GOODS_PRICE             : num  450000 180000 630000 1575000 625500 ...
##  $ NAME_TYPE_SUITE             : chr  "Unaccompanied" "Unaccompanied" "" "Unaccompanied" ...
##  $ NAME_INCOME_TYPE            : chr  "Working" "Working" "Working" "Working" ...
##  $ NAME_EDUCATION_TYPE         : chr  "Higher education" "Secondary / secondary special" "Higher edu
##  $ NAME_FAMILY_STATUS          : chr  "Married" "Married" "Married" "Married" ...
##  $ NAME_HOUSING_TYPE           : chr  "House / apartment" "House / apartment" "House / apartment" "H
##  $ REGION_POPULATION_RELATIVE  : num  0.0188 0.0358 0.0191 0.0264 0.01 ...
##  $ DAYS_BIRTH                  : int  -19241 -18064 -20038 -13976 -13040 -18604 -16685 -9516 -12744
##  $ DAYS_EMPLOYED               : int  -2329 -4469 -4458 -1866 -2191 -12009 -2580 -1387 -1013 -2625 .
##  $ DAYS_REGISTRATION           : num  -5170 -9118 -2175 -2000 -4000 ...
##  $ DAYS_ID_PUBLISH             : int  -812 -1623 -3503 -4208 -4262 -2027 -241 -2055 -3171 -3041 ...
##  $ OWN_CAR_AGE                 : num  NA NA 5 NA 16 10 3 NA NA 5 ...
##  $ 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 1 0 0 1 0 1 ...
##  $ FLAG_CONT_MOBILE            : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ FLAG_PHONE                  : int  0 0 0 1 0 1 0 1 0 1 ...
##  $ FLAG_EMAIL                  : int  1 0 0 0 0 0 0 0 0 0 ...
##  $ OCCUPATION_TYPE             : chr  "" "Low-skill Laborers" "Drivers" "Sales staff" ...
##  $ CNT_FAM_MEMBERS             : num  2 2 2 4 3 2 4 1 2 3 ...
##  $ REGION_RATING_CLIENT        : int  2 2 2 2 2 2 2 2 1 2 ...
##  $ REGION_RATING_CLIENT_W_CITY : int  2 2 2 2 2 2 2 2 1 2 ...
##  $ WEEKDAY_APPR_PROCESS_START  : chr  "TUESDAY" "FRIDAY" "MONDAY" "WEDNESDAY" ...
##  $ HOUR_APPR_PROCESS_START     : int  18 9 14 11 5 15 9 7 18 14 ...
##  $ 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 0 0 0 0 0 0 0 ...
##  $ LIVE_REGION_NOT_WORK_REGION : int  0 0 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 0 1 0 1 0 0 0 ...
##  $ LIVE_CITY_NOT_WORK_CITY     : int  0 0 0 0 1 0 1 0 0 0 ...
##  $ ORGANIZATION_TYPE           : chr  "Kindergarten" "Self-employed" "Transport: type 3" "Business E
##  $ EXT_SOURCE_1                : num  0.753 0.565 NA 0.526 0.202 ...
##  $ EXT_SOURCE_2                : num  0.79 0.292 0.7 0.51 0.426 ...
##  $ EXT_SOURCE_3                : num  0.16 0.433 0.611 0.613 NA ...
##  $ APARTMENTS_AVG              : num  0.066 NA NA 0.305 NA ...
##  $ BASEMENTAREA_AVG            : num  0.059 NA NA 0.197 NA ...
##  $ YEARS_BEGINEXPLUATATION_AVG : num  0.973 NA NA 0.997 NA ...
##  $ YEARS_BUILD_AVG             : num  NA NA NA 0.959 NA ...
##  $ COMMONAREA_AVG              : num  NA NA NA 0.117 NA ...
```

```
##  $ ELEVATORS_AVG               : num  NA NA NA 0.32 NA 0.16 NA NA 0 NA ...
##  $ ENTRANCES_AVG               : num  0.138 NA NA 0.276 NA ...
##  $ FLOORSMAX_AVG               : num  0.125 NA NA 0.375 NA ...
##  $ FLOORSMIN_AVG               : num  NA NA NA 0.0417 NA 0.375 NA NA NA NA ...
##  $ LANDAREA_AVG                : num  NA NA NA 0.204 NA ...
##  $ LIVINGAPARTMENTS_AVG        : num  NA NA NA 0.24 NA ...
##  $ LIVINGAREA_AVG              : num  0.0505 NA NA 0.3673 NA ...
##  $ NONLIVINGAPARTMENTS_AVG     : num  NA NA NA 0.0386 NA 0.0116 NA NA NA NA ...
##  $ NONLIVINGAREA_AVG           : num  NA NA NA 0.08 NA 0.0731 NA NA NA NA ...
##  $ APARTMENTS_MODE             : num  0.0672 NA NA 0.3109 NA ...
##  $ BASEMENTAREA_MODE           : num  0.0612 NA NA 0.2049 NA ...
##  $ YEARS_BEGINEXPLUATATION_MODE: num  0.973 NA NA 0.997 NA ...
##  $ YEARS_BUILD_MODE            : num  NA NA NA 0.961 NA ...
##  $ COMMONAREA_MODE             : num  NA NA NA 0.118 NA ...
##  $ ELEVATORS_MODE              : num  NA NA NA 0.322 NA ...
##  $ ENTRANCES_MODE              : num  0.138 NA NA 0.276 NA ...
##  $ FLOORSMAX_MODE              : num  0.125 NA NA 0.375 NA ...
##  $ FLOORSMIN_MODE              : num  NA NA NA 0.0417 NA 0.375 NA NA NA NA ...
##  $ LANDAREA_MODE               : num  NA NA NA 0.209 NA ...
##  $ LIVINGAPARTMENTS_MODE       : num  NA NA NA 0.263 NA ...
##  $ LIVINGAREA_MODE             : num  0.0526 NA NA 0.3827 NA ...
##  $ NONLIVINGAPARTMENTS_MODE    : num  NA NA NA 0.0389 NA 0.0117 NA NA NA NA ...
##  $ NONLIVINGAREA_MODE          : num  NA NA NA 0.0847 NA 0.0774 NA NA NA NA ...
##  $ APARTMENTS_MEDI             : num  0.0666 NA NA 0.3081 NA ...
##  $ BASEMENTAREA_MEDI           : num  0.059 NA NA 0.197 NA ...
##  $ YEARS_BEGINEXPLUATATION_MEDI: num  0.973 NA NA 0.997 NA ...
##  $ YEARS_BUILD_MEDI            : num  NA NA NA 0.96 NA ...
##  $ COMMONAREA_MEDI             : num  NA NA NA 0.117 NA ...
##  $ ELEVATORS_MEDI              : num  NA NA NA 0.32 NA 0.16 NA NA 0 NA ...
##  $ ENTRANCES_MEDI              : num  0.138 NA NA 0.276 NA ...
##  $ FLOORSMAX_MEDI              : num  0.125 NA NA 0.375 NA ...
##  $ FLOORSMIN_MEDI              : num  NA NA NA 0.0417 NA 0.375 NA NA NA NA ...
##  $ LANDAREA_MEDI               : num  NA NA NA 0.208 NA ...
##  $ LIVINGAPARTMENTS_MEDI       : num  NA NA NA 0.245 NA ...
##  $ LIVINGAREA_MEDI             : num  0.0514 NA NA 0.3739 NA ...
##  $ NONLIVINGAPARTMENTS_MEDI    : num  NA NA NA 0.0388 NA 0.0116 NA NA NA NA ...
##  $ NONLIVINGAREA_MEDI          : num  NA NA NA 0.0817 NA 0.0746 NA NA NA NA ...
##  $ FONDKAPREMONT_MODE          : chr  "" "" "" "reg oper account" ...
##  $ HOUSETYPE_MODE              : chr  "block of flats" "" "" "block of flats" ...
##  $ TOTALAREA_MODE              : num  0.0392 NA NA 0.37 NA ...
##  $ WALLSMATERIAL_MODE          : chr  "Stone, brick" "" "" "Panel" ...
##  $ EMERGENCYSTATE_MODE         : chr  "No" "" "" "No" ...
##  $ OBS_30_CNT_SOCIAL_CIRCLE    : num  0 0 0 0 0 0 1 0 0 4 ...
##  $ DEF_30_CNT_SOCIAL_CIRCLE    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ OBS_60_CNT_SOCIAL_CIRCLE    : num  0 0 0 0 0 0 1 0 0 4 ...
##  $ DEF_60_CNT_SOCIAL_CIRCLE    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ DAYS_LAST_PHONE_CHANGE      : num  -1740 0 -856 -1805 -821 ...
##  $ FLAG_DOCUMENT_2             : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ FLAG_DOCUMENT_3             : int  1 1 0 1 1 0 1 0 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 ...
##   [list output truncated]
```

4 Exploratory Data Analysis (EDA)

The Exploratory Data Analysis (EDA) phase focuses on understanding the structure and content of the application_train dataset, identifying anomalies or inconsistencies, and uncovering patterns or relationships between the target variable and other features. This step is critical for ensuring data quality and gaining insights that align with the company's strategic objectives.

To maintain clarity and conciseness, only the most relevant analyses and visualizations will be highlighted in this document. While some supporting code may be included for transparency, its outputs might not be displayed to keep the focus on key findings.

```
# Summary statistics for numerical features
numerical_summary <- Home_train %>%
  select_if(is.numeric) %>%
  summary()
print(numerical_summary)
```

```
##    SK_ID_CURR         TARGET          CNT_CHILDREN     AMT_INCOME_TOTAL
##  Min.   :100002   Min.   :0.00000   Min.   : 0.0000   Min.   :    25650
##  1st Qu.:189146   1st Qu.:0.00000   1st Qu.: 0.0000   1st Qu.:   112500
##  Median :278202   Median :0.00000   Median : 0.0000   Median :   147150
##  Mean   :278181   Mean   :0.08073   Mean   : 0.4171   Mean   :   168798
##  3rd Qu.:367143   3rd Qu.:0.00000   3rd Qu.: 1.0000   3rd Qu.:   202500
##  Max.   :456255   Max.   :1.00000   Max.   :19.0000   Max.   :117000000
##
##    AMT_CREDIT       AMT_ANNUITY     AMT_GOODS_PRICE
##  Min.   :  45000   Min.   :  1616   Min.   :  40500
##  1st Qu.: 270000   1st Qu.: 16524   1st Qu.: 238500
##  Median : 513531   Median : 24903   Median : 450000
##  Mean   : 599026   Mean   : 27109   Mean   : 538396
##  3rd Qu.: 808650   3rd Qu.: 34596   3rd Qu.: 679500
##  Max.   :4050000   Max.   :258026   Max.   :4050000
##                    NA's   :12       NA's   :278
##  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.: -7480
##  Median :0.01885            Median :-15750   Median : -1213   Median : -4504
##  Mean   :0.02087            Mean   :-16037   Mean   : 63815   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  OWN_CAR_AGE       FLAG_MOBIL FLAG_EMP_PHONE
##  Min.   :-7197   Min.   : 0.00    Min.   :0    Min.   :0.0000
##  1st Qu.:-4299   1st Qu.: 5.00    1st Qu.:1    1st Qu.:1.0000
##  Median :-3254   Median : 9.00    Median :1    Median :1.0000
##  Mean   :-2994   Mean   :12.06    Mean   :1    Mean   :0.8199
##  3rd Qu.:-1720   3rd Qu.:15.00    3rd Qu.:1    3rd Qu.:1.0000
##  Max.   :    0   Max.   :91.00    Max.   :1    Max.   :1.0000
##                  NA's   :202929
##  FLAG_WORK_PHONE  FLAG_CONT_MOBILE   FLAG_PHONE       FLAG_EMAIL
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:0.00000
##  Median :0.0000   Median :1.0000   Median :0.0000   Median :0.00000
##  Mean   :0.1994   Mean   :0.9981   Mean   :0.2811   Mean   :0.05672
```

```
##  3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.00000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.00000
##
##  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   : 2.153   Mean   :2.052        Mean   :2.032
##  3rd Qu.: 3.000   3rd Qu.:2.000        3rd Qu.:2.000
##  Max.   :20.000   Max.   :3.000        Max.   :3.000
##  NA's   :2
##  HOUR_APPR_PROCESS_START REG_REGION_NOT_LIVE_REGION REG_REGION_NOT_WORK_REGION
##  Min.   : 0.00           Min.   :0.00000            Min.   :0.00000
##  1st Qu.:10.00           1st Qu.:0.00000            1st Qu.:0.00000
##  Median :12.00           Median :0.00000            Median :0.00000
##  Mean   :12.06           Mean   :0.01514            Mean   :0.05077
##  3rd Qu.:14.00           3rd Qu.:0.00000            3rd Qu.:0.00000
##  Max.   :23.00           Max.   :1.00000            Max.   :1.00000
##
##  LIVE_REGION_NOT_WORK_REGION REG_CITY_NOT_LIVE_CITY REG_CITY_NOT_WORK_CITY
##  Min.   :0.00000             Min.   :0.00000        Min.   :0.0000
##  1st Qu.:0.00000             1st Qu.:0.00000        1st Qu.:0.0000
##  Median :0.00000             Median :0.00000        Median :0.0000
##  Mean   :0.04066             Mean   :0.07817        Mean   :0.2305
##  3rd Qu.:0.00000             3rd Qu.:0.00000        3rd Qu.:0.0000
##  Max.   :1.00000             Max.   :1.00000        Max.   :1.0000
##
##  LIVE_CITY_NOT_WORK_CITY  EXT_SOURCE_1     EXT_SOURCE_2     EXT_SOURCE_3
##  Min.   :0.0000           Min.   :0.01     Min.   :0.0000   Min.   :0.00
##  1st Qu.:0.0000           1st Qu.:0.33     1st Qu.:0.3925   1st Qu.:0.37
##  Median :0.0000           Median :0.51     Median :0.5660   Median :0.54
##  Mean   :0.1796           Mean   :0.50     Mean   :0.5144   Mean   :0.51
##  3rd Qu.:0.0000           3rd Qu.:0.68     3rd Qu.:0.6636   3rd Qu.:0.67
##  Max.   :1.0000           Max.   :0.96     Max.   :0.8550   Max.   :0.90
##                           NA's   :173378   NA's   :660      NA's   :60965
##  APARTMENTS_AVG   BASEMENTAREA_AVG YEARS_BEGINEXPLUATATION_AVG YEARS_BUILD_AVG
##  Min.   :0.00     Min.   :0.00     Min.   :0.00                Min.   :0.00
##  1st Qu.:0.06     1st Qu.:0.04     1st Qu.:0.98                1st Qu.:0.69
##  Median :0.09     Median :0.08     Median :0.98                Median :0.76
##  Mean   :0.12     Mean   :0.09     Mean   :0.98                Mean   :0.75
##  3rd Qu.:0.15     3rd Qu.:0.11     3rd Qu.:0.99                3rd Qu.:0.82
##  Max.   :1.00     Max.   :1.00     Max.   :1.00                Max.   :1.00
##  NA's   :156061   NA's   :179943   NA's   :150007              NA's   :204488
##  COMMONAREA_AVG   ELEVATORS_AVG    ENTRANCES_AVG    FLOORSMAX_AVG
##  Min.   :0.00     Min.   :0.00     Min.   :0.00     Min.   :0.00
##  1st Qu.:0.01     1st Qu.:0.00     1st Qu.:0.07     1st Qu.:0.17
##  Median :0.02     Median :0.00     Median :0.14     Median :0.17
##  Mean   :0.04     Mean   :0.08     Mean   :0.15     Mean   :0.23
##  3rd Qu.:0.05     3rd Qu.:0.12     3rd Qu.:0.21     3rd Qu.:0.33
##  Max.   :1.00     Max.   :1.00     Max.   :1.00     Max.   :1.00
##  NA's   :214865   NA's   :163891   NA's   :154828   NA's   :153020
##  FLOORSMIN_AVG     LANDAREA_AVG    LIVINGAPARTMENTS_AVG LIVINGAREA_AVG
##  Min.   :0.00     Min.   :0.00     Min.   :0.00         Min.   :0.00
##  1st Qu.:0.08     1st Qu.:0.02     1st Qu.:0.05         1st Qu.:0.05
```

```
##  Median :0.21     Median :0.05     Median :0.08         Median :0.07
##  Mean   :0.23     Mean   :0.07     Mean   :0.10         Mean   :0.11
##  3rd Qu.:0.38     3rd Qu.:0.09     3rd Qu.:0.12         3rd Qu.:0.13
##  Max.   :1.00     Max.   :1.00     Max.   :1.00         Max.   :1.00
##  NA's   :208642   NA's   :182590   NA's   :210199       NA's   :154350
##  NONLIVINGAPARTMENTS_AVG NONLIVINGAREA_AVG APARTMENTS_MODE  BASEMENTAREA_MODE
##  Min.   :0.00            Min.   :0.00      Min.   :0.00     Min.   :0.00
##  1st Qu.:0.00            1st Qu.:0.00      1st Qu.:0.05     1st Qu.:0.04
##  Median :0.00            Median :0.00      Median :0.08     Median :0.07
##  Mean   :0.01            Mean   :0.03      Mean   :0.11     Mean   :0.09
##  3rd Qu.:0.00            3rd Qu.:0.03      3rd Qu.:0.14     3rd Qu.:0.11
##  Max.   :1.00            Max.   :1.00      Max.   :1.00     Max.   :1.00
##  NA's   :213514          NA's   :169682    NA's   :156061   NA's   :179943
##  YEARS_BEGINEXPLUATATION_MODE YEARS_BUILD_MODE COMMONAREA_MODE
##  Min.   :0.00                 Min.   :0.00     Min.   :0.00
##  1st Qu.:0.98                 1st Qu.:0.70     1st Qu.:0.01
##  Median :0.98                 Median :0.76     Median :0.02
##  Mean   :0.98                 Mean   :0.76     Mean   :0.04
##  3rd Qu.:0.99                 3rd Qu.:0.82     3rd Qu.:0.05
##  Max.   :1.00                 Max.   :1.00     Max.   :1.00
##  NA's   :150007               NA's   :204488   NA's   :214865
##  ELEVATORS_MODE   ENTRANCES_MODE   FLOORSMAX_MODE   FLOORSMIN_MODE
##  Min.   :0.00     Min.   :0.00     Min.   :0.00     Min.   :0.00
##  1st Qu.:0.00     1st Qu.:0.07     1st Qu.:0.17     1st Qu.:0.08
##  Median :0.00     Median :0.14     Median :0.17     Median :0.21
##  Mean   :0.07     Mean   :0.15     Mean   :0.22     Mean   :0.23
##  3rd Qu.:0.12     3rd Qu.:0.21     3rd Qu.:0.33     3rd Qu.:0.38
##  Max.   :1.00     Max.   :1.00     Max.   :1.00     Max.   :1.00
##  NA's   :163891   NA's   :154828   NA's   :153020   NA's   :208642
##  LANDAREA_MODE    LIVINGAPARTMENTS_MODE LIVINGAREA_MODE
##  Min.   :0.00     Min.   :0.00          Min.   :0.00
##  1st Qu.:0.02     1st Qu.:0.05          1st Qu.:0.04
##  Median :0.05     Median :0.08          Median :0.07
##  Mean   :0.06     Mean   :0.11          Mean   :0.11
##  3rd Qu.:0.08     3rd Qu.:0.13          3rd Qu.:0.13
##  Max.   :1.00     Max.   :1.00          Max.   :1.00
##  NA's   :182590   NA's   :210199        NA's   :154350
##  NONLIVINGAPARTMENTS_MODE NONLIVINGAREA_MODE APARTMENTS_MEDI  BASEMENTAREA_MEDI
##  Min.   :0.00             Min.   :0.00       Min.   :0.00     Min.   :0.00
##  1st Qu.:0.00             1st Qu.:0.00       1st Qu.:0.06     1st Qu.:0.04
##  Median :0.00             Median :0.00       Median :0.09     Median :0.08
##  Mean   :0.01             Mean   :0.03       Mean   :0.12     Mean   :0.09
##  3rd Qu.:0.00             3rd Qu.:0.02       3rd Qu.:0.15     3rd Qu.:0.11
##  Max.   :1.00             Max.   :1.00       Max.   :1.00     Max.   :1.00
##  NA's   :213514           NA's   :169682     NA's   :156061   NA's   :179943
##  YEARS_BEGINEXPLUATATION_MEDI YEARS_BUILD_MEDI COMMONAREA_MEDI
##  Min.   :0.00                 Min.   :0.00     Min.   :0.00
##  1st Qu.:0.98                 1st Qu.:0.69     1st Qu.:0.01
##  Median :0.98                 Median :0.76     Median :0.02
##  Mean   :0.98                 Mean   :0.76     Mean   :0.04
##  3rd Qu.:0.99                 3rd Qu.:0.83     3rd Qu.:0.05
##  Max.   :1.00                 Max.   :1.00     Max.   :1.00
##  NA's   :150007               NA's   :204488   NA's   :214865
##  ELEVATORS_MEDI   ENTRANCES_MEDI   FLOORSMAX_MEDI   FLOORSMIN_MEDI
```

```
##  Min.   :0.00     Min.   :0.00     Min.   :0.00     Min.   :0.00
##  1st Qu.:0.00     1st Qu.:0.07     1st Qu.:0.17     1st Qu.:0.08
##  Median :0.00     Median :0.14     Median :0.17     Median :0.21
##  Mean   :0.08     Mean   :0.15     Mean   :0.23     Mean   :0.23
##  3rd Qu.:0.12     3rd Qu.:0.21     3rd Qu.:0.33     3rd Qu.:0.38
##  Max.   :1.00     Max.   :1.00     Max.   :1.00     Max.   :1.00
##  NA's   :163891   NA's   :154828   NA's   :153020   NA's   :208642
##  LANDAREA_MEDI    LIVINGAPARTMENTS_MEDI LIVINGAREA_MEDI
##  Min.   :0.00     Min.   :0.00          Min.   :0.00
##  1st Qu.:0.02     1st Qu.:0.05          1st Qu.:0.05
##  Median :0.05     Median :0.08          Median :0.07
##  Mean   :0.07     Mean   :0.10          Mean   :0.11
##  3rd Qu.:0.09     3rd Qu.:0.12          3rd Qu.:0.13
##  Max.   :1.00     Max.   :1.00          Max.   :1.00
##  NA's   :182590   NA's   :210199        NA's   :154350
##  NONLIVINGAPARTMENTS_MEDI NONLIVINGAREA_MEDI TOTALAREA_MODE
##  Min.   :0.00             Min.   :0.00       Min.   :0.00
##  1st Qu.:0.00             1st Qu.:0.00       1st Qu.:0.04
##  Median :0.00             Median :0.00       Median :0.07
##  Mean   :0.01             Mean   :0.03       Mean   :0.10
##  3rd Qu.:0.00             3rd Qu.:0.03       3rd Qu.:0.13
##  Max.   :1.00             Max.   :1.00       Max.   :1.00
##  NA's   :213514           NA's   :169682     NA's   :148431
##  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   :  1.422          Mean   : 0.1434          Mean   :  1.405
##  3rd Qu.:  2.000          3rd Qu.: 0.0000          3rd Qu.:  2.000
##  Max.   :348.000          Max.   :34.0000          Max.   :344.000
##  NA's   :1021             NA's   :1021             NA's   :1021
##  DEF_60_CNT_SOCIAL_CIRCLE DAYS_LAST_PHONE_CHANGE FLAG_DOCUMENT_2
##  Min.   : 0.0             Min.   :-4292.0        Min.   :0.00e+00
##  1st Qu.: 0.0             1st Qu.:-1570.0        1st Qu.:0.00e+00
##  Median : 0.0             Median : -757.0        Median :0.00e+00
##  Mean   : 0.1             Mean   : -962.9        Mean   :4.23e-05
##  3rd Qu.: 0.0             3rd Qu.: -274.0        3rd Qu.:0.00e+00
##  Max.   :24.0             Max.   :    0.0        Max.   :1.00e+00
##  NA's   :1021             NA's   :1
##  FLAG_DOCUMENT_3 FLAG_DOCUMENT_4    FLAG_DOCUMENT_5   FLAG_DOCUMENT_6
##  Min.   :0.00    Min.   :0.00e+00   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.00    1st Qu.:0.00e+00   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :1.00    Median :0.00e+00   Median :0.00000   Median :0.00000
##  Mean   :0.71    Mean   :8.13e-05   Mean   :0.01511   Mean   :0.08806
##  3rd Qu.:1.00    3rd Qu.:0.00e+00   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.00    Max.   :1.00e+00   Max.   :1.00000   Max.   :1.00000
##
##  FLAG_DOCUMENT_7     FLAG_DOCUMENT_8   FLAG_DOCUMENT_9    FLAG_DOCUMENT_10
##  Min.   :0.0000000   Min.   :0.00000   Min.   :0.000000   Min.   :0.00e+00
##  1st Qu.:0.0000000   1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0.00e+00
##  Median :0.0000000   Median :0.00000   Median :0.000000   Median :0.00e+00
##  Mean   :0.0001919   Mean   :0.08138   Mean   :0.003896   Mean   :2.28e-05
##  3rd Qu.:0.0000000   3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:0.00e+00
##  Max.   :1.0000000   Max.   :1.00000   Max.   :1.000000   Max.   :1.00e+00
```

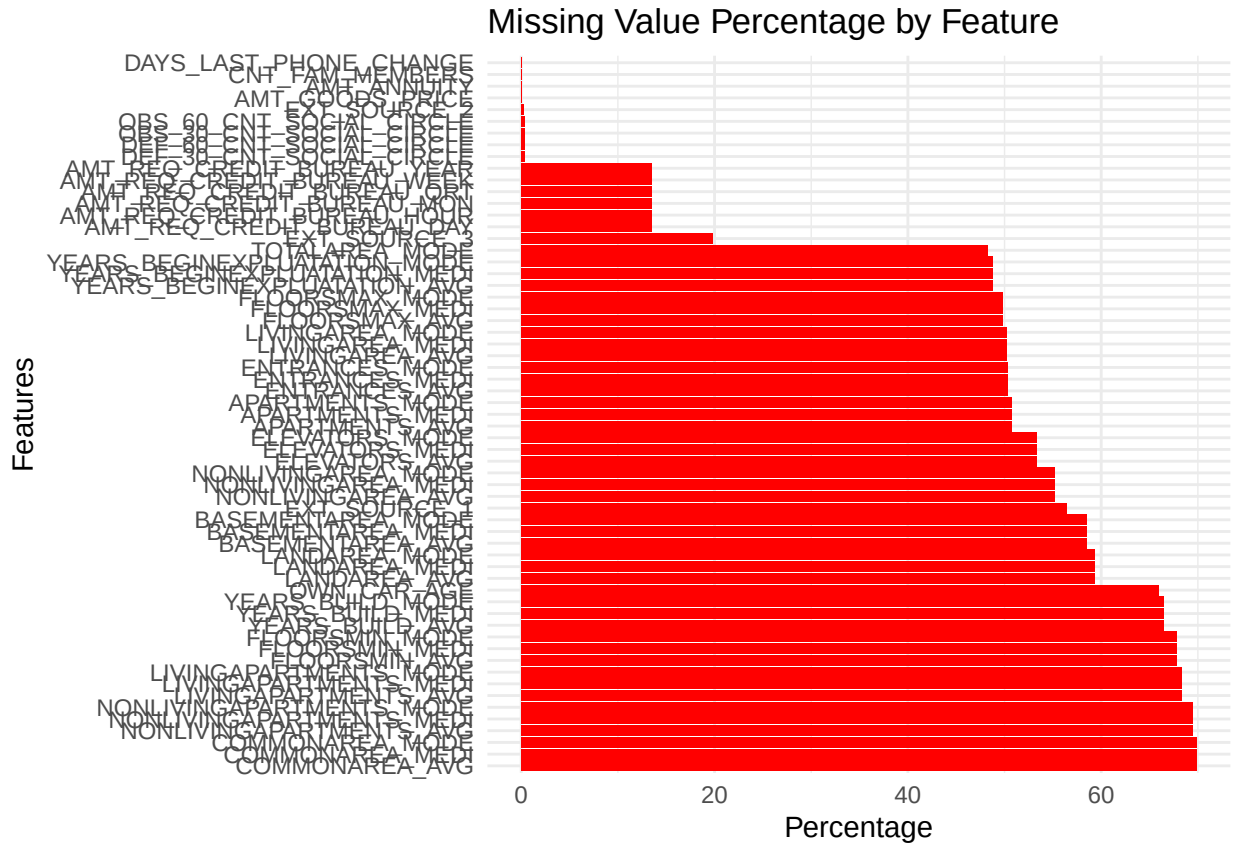
```
##
##  FLAG_DOCUMENT_11   FLAG_DOCUMENT_12  FLAG_DOCUMENT_13   FLAG_DOCUMENT_14
##  Min.   :0.000000   Min.   :0.0e+00   Min.   :0.000000   Min.   :0.000000
##  1st Qu.:0.000000   1st Qu.:0.0e+00   1st Qu.:0.000000   1st Qu.:0.000000
##  Median :0.000000   Median :0.0e+00   Median :0.000000   Median :0.000000
##  Mean   :0.003912   Mean   :6.5e-06   Mean   :0.003525   Mean   :0.002936
##  3rd Qu.:0.000000   3rd Qu.:0.0e+00   3rd Qu.:0.000000   3rd Qu.:0.000000
##  Max.   :1.000000   Max.   :1.0e+00   Max.   :1.000000   Max.   :1.000000
##
##  FLAG_DOCUMENT_15  FLAG_DOCUMENT_16   FLAG_DOCUMENT_17    FLAG_DOCUMENT_18
##  Min.   :0.00000   Min.   :0.000000   Min.   :0.0000000   Min.   :0.00000
##  1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.:0.0000000   1st Qu.:0.00000
##  Median :0.00000   Median :0.000000   Median :0.0000000   Median :0.00000
##  Mean   :0.00121   Mean   :0.009928   Mean   :0.0002667   Mean   :0.00813
##  3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:0.0000000   3rd Qu.:0.00000
##  Max.   :1.00000   Max.   :1.000000   Max.   :1.0000000   Max.   :1.00000
##
##  FLAG_DOCUMENT_19    FLAG_DOCUMENT_20    FLAG_DOCUMENT_21
##  Min.   :0.0000000   Min.   :0.0000000   Min.   :0.0000000
##  1st Qu.:0.0000000   1st Qu.:0.0000000   1st Qu.:0.0000000
##  Median :0.0000000   Median :0.0000000   Median :0.0000000
##  Mean   :0.0005951   Mean   :0.0005073   Mean   :0.0003349
##  3rd Qu.:0.0000000   3rd Qu.:0.0000000   3rd Qu.:0.0000000
##  Max.   :1.0000000   Max.   :1.0000000   Max.   :1.0000000
##
##  AMT_REQ_CREDIT_BUREAU_HOUR AMT_REQ_CREDIT_BUREAU_DAY
##  Min.   :0.00               Min.   :0.00
##  1st Qu.:0.00               1st Qu.:0.00
##  Median :0.00               Median :0.00
##  Mean   :0.01               Mean   :0.01
##  3rd Qu.:0.00               3rd Qu.:0.00
##  Max.   :4.00               Max.   :9.00
##  NA's   :41519              NA's   :41519
##  AMT_REQ_CREDIT_BUREAU_WEEK AMT_REQ_CREDIT_BUREAU_MON AMT_REQ_CREDIT_BUREAU_QRT
##  Min.   :0.00               Min.   : 0.00             Min.   :  0.00
##  1st Qu.:0.00               1st Qu.: 0.00             1st Qu.:  0.00
##  Median :0.00               Median : 0.00             Median :  0.00
##  Mean   :0.03               Mean   : 0.27             Mean   :  0.27
##  3rd Qu.:0.00               3rd Qu.: 0.00             3rd Qu.:  0.00
##  Max.   :8.00               Max.   :27.00             Max.   :261.00
##  NA's   :41519              NA's   :41519             NA's   :41519
##  AMT_REQ_CREDIT_BUREAU_YEAR
##  Min.   : 0.0
##  1st Qu.: 0.0
##  Median : 1.0
##  Mean   : 1.9
##  3rd Qu.: 3.0
##  Max.   :25.0
##  NA's   :41519
```

```r
# Summary for categorical variables
categorical_summary <- Home_train %>%
  select_if(~ is.character(.) || is.factor(.)) %>%
  summary()
```

```r
print(categorical_summary)
```

```
##  NAME_CONTRACT_TYPE CODE_GENDER        FLAG_OWN_CAR       FLAG_OWN_REALTY
##  Length:307511      Length:307511      Length:307511      Length:307511
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##  NAME_TYPE_SUITE    NAME_INCOME_TYPE   NAME_EDUCATION_TYPE NAME_FAMILY_STATUS
##  Length:307511      Length:307511      Length:307511       Length:307511
##  Class :character   Class :character   Class :character    Class :character
##  Mode  :character   Mode  :character   Mode  :character    Mode  :character
##  NAME_HOUSING_TYPE  OCCUPATION_TYPE    WEEKDAY_APPR_PROCESS_START
##  Length:307511      Length:307511      Length:307511
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##  ORGANIZATION_TYPE  FONDKAPREMONT_MODE HOUSETYPE_MODE     WALLSMATERIAL_MODE
##  Length:307511      Length:307511      Length:307511      Length:307511
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##  EMERGENCYSTATE_MODE
##  Length:307511
##  Class :character
##  Mode  :character
```

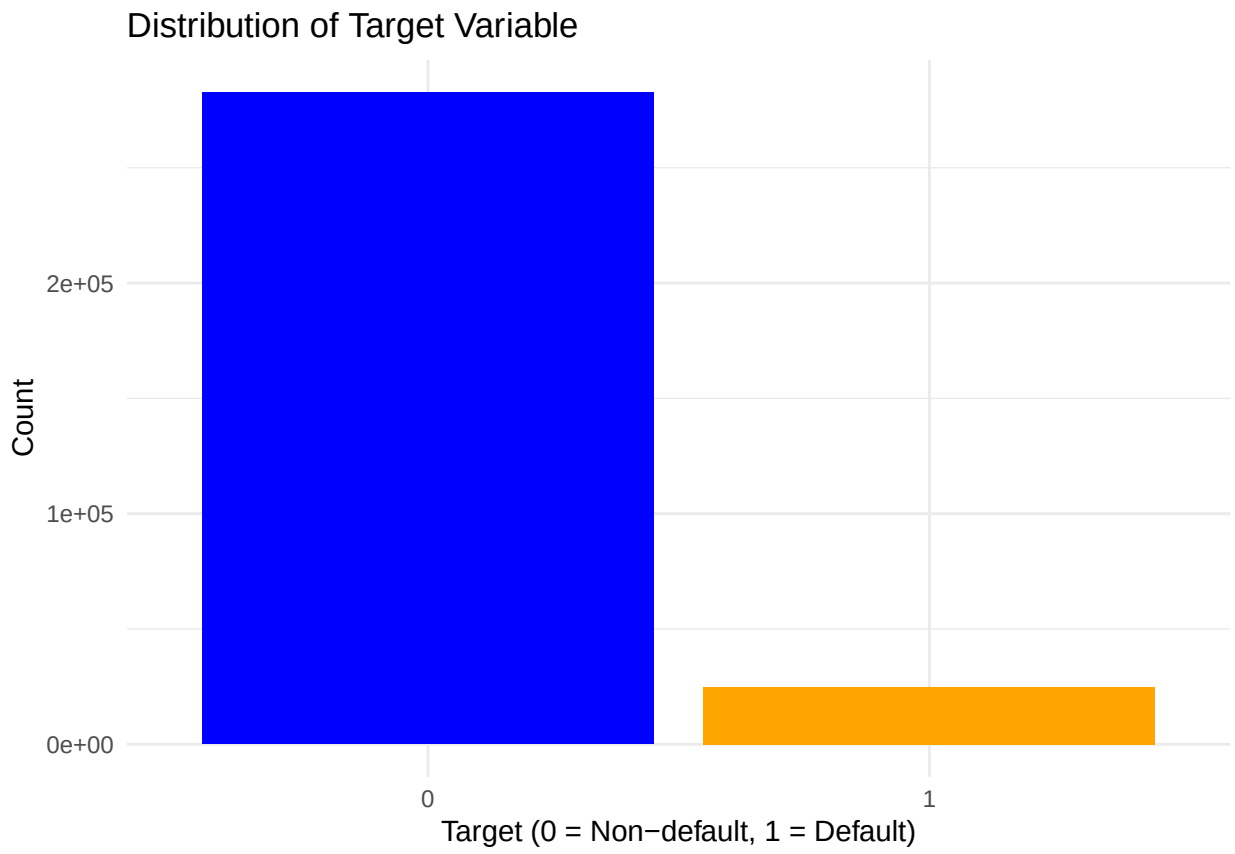
```r
# Check for missing values and visualize
missing_values <- colSums(is.na(Home_train)) %>% sort(decreasing = TRUE)
missing_df <- data.frame(Feature = names(missing_values), Missing_Count = missing_values) %>%
  filter(Missing_Count > 0) %>%
  mutate(Missing_Percent = round(Missing_Count / nrow(Home_train) * 100, 2))

# Plot missing values
ggplot(missing_df, aes(x = reorder(Feature, -Missing_Percent), y = Missing_Percent)) +
  geom_bar(stat = "identity", fill = "red") +
  coord_flip() +
  labs(title = "Missing Value Percentage by Feature", x = "Features", y = "Percentage") +
  theme_minimal()
```




```
# Visualize the distribution of the target variable
target_distribution <- Home_train %>%
  group_by(TARGET) %>%
  summarise(Count = n(), Percent = round(100 * Count / nrow(Home_train), 2))

print(target_distribution)
```

```
## # A tibble: 2 x 3
##   TARGET  Count Percent
##    <int>  <int>   <dbl>
## 1      0 282686   91.9
## 2      1  24825    8.07
```

```
ggplot(Home_train, aes(x = as.factor(TARGET))) +
  geom_bar(fill = c("blue", "orange")) +
  labs(title = "Distribution of Target Variable", x = "Target (0 = Non-default, 1 = Default)", y = "Cou
  theme_minimal()
```



5 Data Preprocessing

```r
# 1. Impute missing numerical values
Home_train <- Home_train %>%
  mutate_if(is.numeric, ~ifelse(is.na(.), median(., na.rm = TRUE), .))

# 2. Impute missing categorical values with "Unknown"
Home_train <- Home_train %>%
  mutate_if(~is.character(.) || is.factor(.), ~ifelse(is.na(.), "Unknown", .))

# 3. Convert categorical columns to factors
categorical_columns <- Home_train %>%
  select_if(~is.character(.) || is.factor(.)) %>%
  colnames()

Home_train[categorical_columns] <- lapply(Home_train[categorical_columns], as.factor)

# 4. Convert binary indicator variables to factors (if not already handled)
binary_columns <- Home_train %>%
  select_if(~all(. %in% c(0, 1))) %>%
  colnames()

Home_train[binary_columns] <- lapply(Home_train[binary_columns], as.factor)
```

```r
# 5. Verify preprocessing
str(Home_train)
```

```
## 'data.frame':    307511 obs. of  122 variables:
##  $ SK_ID_CURR                  : int  100002 100003 100004 100006 100007 100008 100009 100010 100011
##  $ 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 2 1 1 1 1 1
##  $ CODE_GENDER                 : Factor w/ 3 levels "F","M","XNA": 2 1 2 1 2 2 1 2 1 2 ...
##  $ FLAG_OWN_CAR                : Factor w/ 2 levels "N","Y": 1 1 2 1 1 1 2 2 1 1 ...
##  $ FLAG_OWN_REALTY             : Factor w/ 2 levels "N","Y": 2 1 2 2 2 2 2 2 2 2 ...
##  $ CNT_CHILDREN                : int  0 0 0 0 0 0 1 0 0 0 ...
##  $ AMT_INCOME_TOTAL            : num  202500 270000 67500 135000 121500 ...
##  $ AMT_CREDIT                  : num  406598 1293503 135000 312683 513000 ...
##  $ AMT_ANNUITY                 : num  24701 35699 6750 29687 21866 ...
##  $ AMT_GOODS_PRICE             : num  351000 1129500 135000 297000 513000 ...
##  $ NAME_TYPE_SUITE             : Factor w/ 8 levels "","Children",..: 8 3 8 8 8 7 8 8 2 8 ...
##  $ NAME_INCOME_TYPE            : Factor w/ 8 levels "Businessman",..: 8 5 8 8 8 5 2 5 4 8 ...
##  $ NAME_EDUCATION_TYPE         : Factor w/ 5 levels "Academic degree",..: 5 2 5 5 5 5 2 2 5 5 ...
##  $ NAME_FAMILY_STATUS          : Factor w/ 6 levels "Civil marriage",..: 4 2 4 1 4 2 2 2 2 4 ...
##  $ NAME_HOUSING_TYPE           : Factor w/ 6 levels "Co-op apartment",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ REGION_POPULATION_RELATIVE  : num  0.0188 0.00354 0.01003 0.00802 0.02866 ...
##  $ DAYS_BIRTH                  : int  -9461 -16765 -19046 -19005 -19932 -16941 -13778 -18850 -20099
##  $ DAYS_EMPLOYED               : int  -637 -1188 -225 -3039 -3038 -1588 -3130 -449 365243 -2019 ...
##  $ DAYS_REGISTRATION           : num  -3648 -1186 -4260 -9833 -4311 ...
##  $ DAYS_ID_PUBLISH             : int  -2120 -291 -2531 -2437 -3458 -477 -619 -2379 -3514 -3992 ...
##  $ OWN_CAR_AGE                 : num  9 9 26 9 9 9 17 8 9 9 ...
##  $ FLAG_MOBIL                  : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
##  $ FLAG_EMP_PHONE              : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 1 2 ...
##  $ FLAG_WORK_PHONE             : Factor w/ 2 levels "0","1": 1 1 2 1 1 2 1 2 1 1 ...
##  $ FLAG_CONT_MOBILE            : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
##  $ FLAG_PHONE                  : Factor w/ 2 levels "0","1": 2 2 2 1 1 2 2 1 1 1 ...
##  $ 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",..: 10 5 10 10 5 10 2 12 1 10
##  $ CNT_FAM_MEMBERS             : num  1 2 1 2 1 2 3 2 2 1 ...
##  $ REGION_RATING_CLIENT        : int  2 1 2 2 2 2 2 3 2 2 ...
##  $ REGION_RATING_CLIENT_W_CITY : int  2 1 2 2 2 2 2 3 2 2 ...
##  $ WEEKDAY_APPR_PROCESS_START  : Factor w/ 7 levels "FRIDAY","MONDAY",..: 7 2 2 7 5 7 4 2 7 5 ...
##  $ HOUR_APPR_PROCESS_START     : int  10 11 9 17 11 16 16 16 14 8 ...
##  $ REG_REGION_NOT_LIVE_REGION  : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ REG_REGION_NOT_WORK_REGION  : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ LIVE_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 1 2 1 1 2 1 1 ...
##  $ LIVE_CITY_NOT_WORK_CITY     : Factor w/ 2 levels "0","1": 1 1 1 1 2 1 1 2 1 1 ...
##  $ ORGANIZATION_TYPE           : Factor w/ 58 levels "Advertising",..: 6 40 12 6 38 34 6 34 58 10 ..
##  $ EXT_SOURCE_1                : num  0.083 0.311 0.506 0.506 0.506 ...
##  $ EXT_SOURCE_2                : num  0.263 0.622 0.556 0.65 0.323 ...
##  $ EXT_SOURCE_3                : num  0.139 0.535 0.73 0.535 0.535 ...
##  $ APARTMENTS_AVG              : num  0.0247 0.0959 0.0876 0.0876 0.0876 0.0876 0.0876 0.0876 0.0876
##  $ BASEMENTAREA_AVG            : num  0.0369 0.0529 0.0763 0.0763 0.0763 0.0763 0.0763 0.0763 0.0763
##  $ YEARS_BEGINEXPLUATATION_AVG : num  0.972 0.985 0.982 0.982 0.982 ...
##  $ YEARS_BUILD_AVG             : num  0.619 0.796 0.755 0.755 0.755 ...
##  $ COMMONAREA_AVG              : num  0.0143 0.0605 0.0211 0.0211 0.0211 0.0211 0.0211 0.0211 0.0211
```

```
##  $ ELEVATORS_AVG               : num  0 0.08 0 0 0 0 0 0 0 0 ...
##  $ ENTRANCES_AVG               : num  0.069 0.0345 0.1379 0.1379 0.1379 ...
##  $ FLOORSMAX_AVG               : num  0.0833 0.2917 0.1667 0.1667 0.1667 ...
##  $ FLOORSMIN_AVG               : num  0.125 0.333 0.208 0.208 0.208 ...
##  $ LANDAREA_AVG                : num  0.0369 0.013 0.0481 0.0481 0.0481 0.0481 0.0481 0.0481 0.0481 0
##  $ LIVINGAPARTMENTS_AVG        : num  0.0202 0.0773 0.0756 0.0756 0.0756 0.0756 0.0756 0.0756 0.0756
##  $ LIVINGAREA_AVG              : num  0.019 0.0549 0.0745 0.0745 0.0745 0.0745 0.0745 0.0745 0.0745 0
##  $ NONLIVINGAPARTMENTS_AVG     : num  0 0.0039 0 0 0 0 0 0 0 0 ...
##  $ NONLIVINGAREA_AVG           : num  0 0.0098 0.0036 0.0036 0.0036 0.0036 0.0036 0.0036 0.0036 0.003
##  $ APARTMENTS_MODE             : num  0.0252 0.0924 0.084 0.084 0.084 0.084 0.084 0.084 0.084 0.084
##  $ BASEMENTAREA_MODE           : num  0.0383 0.0538 0.0746 0.0746 0.0746 0.0746 0.0746 0.0746 0.0746
##  $ YEARS_BEGINEXPLUATATION_MODE: num  0.972 0.985 0.982 0.982 0.982 ...
##  $ YEARS_BUILD_MODE            : num  0.634 0.804 0.765 0.765 0.765 ...
##  $ COMMONAREA_MODE             : num  0.0144 0.0497 0.019 0.019 0.019 0.019 0.019 0.019 0.019 0.019
##  $ ELEVATORS_MODE              : num  0 0.0806 0 0 0 0 0 0 0 0 ...
##  $ ENTRANCES_MODE              : num  0.069 0.0345 0.1379 0.1379 0.1379 ...
##  $ FLOORSMAX_MODE              : num  0.0833 0.2917 0.1667 0.1667 0.1667 ...
##  $ FLOORSMIN_MODE              : num  0.125 0.333 0.208 0.208 0.208 ...
##  $ LANDAREA_MODE               : num  0.0377 0.0128 0.0458 0.0458 0.0458 0.0458 0.0458 0.0458 0.0458
##  $ LIVINGAPARTMENTS_MODE       : num  0.022 0.079 0.0771 0.0771 0.0771 0.0771 0.0771 0.0771 0.0771 0
##  $ LIVINGAREA_MODE             : num  0.0198 0.0554 0.0731 0.0731 0.0731 0.0731 0.0731 0.0731 0.0731
##  $ NONLIVINGAPARTMENTS_MODE    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ NONLIVINGAREA_MODE          : num  0 0 0.0011 0.0011 0.0011 0.0011 0.0011 0.0011 0.0011 0.0011 ..
##  $ APARTMENTS_MEDI             : num  0.025 0.0968 0.0864 0.0864 0.0864 0.0864 0.0864 0.0864 0.0864 0
##  $ BASEMENTAREA_MEDI           : num  0.0369 0.0529 0.0758 0.0758 0.0758 0.0758 0.0758 0.0758 0.0758
##  $ YEARS_BEGINEXPLUATATION_MEDI: num  0.972 0.985 0.982 0.982 0.982 ...
##  $ YEARS_BUILD_MEDI            : num  0.624 0.799 0.758 0.758 0.758 ...
##  $ COMMONAREA_MEDI             : num  0.0144 0.0608 0.0208 0.0208 0.0208 0.0208 0.0208 0.0208 0.0208
##  $ ELEVATORS_MEDI              : num  0 0.08 0 0 0 0 0 0 0 0 ...
##  $ ENTRANCES_MEDI              : num  0.069 0.0345 0.1379 0.1379 0.1379 ...
##  $ FLOORSMAX_MEDI              : num  0.0833 0.2917 0.1667 0.1667 0.1667 ...
##  $ FLOORSMIN_MEDI              : num  0.125 0.333 0.208 0.208 0.208 ...
##  $ LANDAREA_MEDI               : num  0.0375 0.0132 0.0487 0.0487 0.0487 0.0487 0.0487 0.0487 0.0487
##  $ LIVINGAPARTMENTS_MEDI       : num  0.0205 0.0787 0.0761 0.0761 0.0761 0.0761 0.0761 0.0761 0.0761
##  $ LIVINGAREA_MEDI             : num  0.0193 0.0558 0.0749 0.0749 0.0749 0.0749 0.0749 0.0749 0.0749
##  $ NONLIVINGAPARTMENTS_MEDI    : num  0 0.0039 0 0 0 0 0 0 0 0 ...
##  $ NONLIVINGAREA_MEDI          : num  0 0.01 0.0031 0.0031 0.0031 0.0031 0.0031 0.0031 0.0031 0.0031
##  $ FONDKAPREMONT_MODE          : Factor w/ 5 levels "","not specified",..: 4 4 1 1 1 1 1 1 1 1 ...
##  $ HOUSETYPE_MODE              : Factor w/ 4 levels "","block of flats",..: 2 2 1 1 1 1 1 1 1 1 ...
##  $ TOTALAREA_MODE              : num  0.0149 0.0714 0.0688 0.0688 0.0688 0.0688 0.0688 0.0688 0.0688
##  $ WALLSMATERIAL_MODE          : Factor w/ 8 levels "","Block","Mixed",..: 7 2 1 1 1 1 1 1 1 1 ...
##  $ EMERGENCYSTATE_MODE         : Factor w/ 3 levels "","No","Yes": 2 2 1 1 1 1 1 1 1 1 ...
##  $ OBS_30_CNT_SOCIAL_CIRCLE    : num  2 1 0 2 0 0 1 2 1 2 ...
##  $ DEF_30_CNT_SOCIAL_CIRCLE    : num  2 0 0 0 0 0 0 0 0 0 ...
##  $ OBS_60_CNT_SOCIAL_CIRCLE    : num  2 1 0 2 0 0 1 2 1 2 ...
##  $ DEF_60_CNT_SOCIAL_CIRCLE    : num  2 0 0 0 0 0 0 0 0 0 ...
##  $ DAYS_LAST_PHONE_CHANGE      : num  -1134 -828 -815 -617 -1106 ...
##  $ FLAG_DOCUMENT_2             : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ FLAG_DOCUMENT_3             : Factor w/ 2 levels "0","1": 2 2 1 2 1 2 1 2 2 1 ...
##  $ FLAG_DOCUMENT_4             : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##   [list output truncated]
```

```
summary(Home_train)
```

```
##    SK_ID_CURR     TARGET           NAME_CONTRACT_TYPE CODE_GENDER  FLAG_OWN_CAR
##  Min.   :100002   0:282686   Cash loans     :278232   F  :202448   N:202924
##  1st Qu.:189146   1: 24825   Revolving loans: 29279   M  :105059   Y:104587
##  Median :278202                                       XNA:     4
##  Mean   :278181
##  3rd Qu.:367143
##  Max.   :456255
##
##  FLAG_OWN_REALTY  CNT_CHILDREN     AMT_INCOME_TOTAL      AMT_CREDIT
##  N: 94199        Min.   : 0.0000   Min.   :    25650   Min.   :  45000
##  Y:213312        1st Qu.: 0.0000   1st Qu.:   112500   1st Qu.: 270000
##                  Median : 0.0000   Median :   147150   Median : 513531
##                  Mean   : 0.4171   Mean   :   168798   Mean   : 599026
##                  3rd Qu.: 1.0000   3rd Qu.:   202500   3rd Qu.: 808650
##                  Max.   :19.0000   Max.   :117000000   Max.   :4050000
##
##   AMT_ANNUITY     AMT_GOODS_PRICE          NAME_TYPE_SUITE
##  Min.   :  1616   Min.   :  40500   Unaccompanied  :248526
##  1st Qu.: 16524   1st Qu.: 238500   Family         : 40149
##  Median : 24903   Median : 450000   Spouse, partner: 11370
##  Mean   : 27108   Mean   : 538316   Children       :  3267
##  3rd Qu.: 34596   3rd Qu.: 679500   Other_B        :  1770
##  Max.   :258026   Max.   :4050000                  :  1292
##                                     (Other)        :  1137
##              NAME_INCOME_TYPE                      NAME_EDUCATION_TYPE
##  Working             :158774   Academic degree              :   164
##  Commercial associate: 71617   Higher education             : 74863
##  Pensioner           : 55362   Incomplete higher            : 10277
##  State servant       : 21703   Lower secondary              :  3816
##  Unemployed          :    22   Secondary / secondary special:218391
##  Student             :    18
##  (Other)             :    15
##             NAME_FAMILY_STATUS           NAME_HOUSING_TYPE
##  Civil marriage      : 29775   Co-op apartment    :  1122
##  Married             :196432   House / apartment  :272868
##  Separated           : 19770   Municipal apartment: 11183
##  Single / not married: 45444   Office apartment   :  2617
##  Unknown             :     2   Rented apartment   :  4881
##  Widow               : 16088   With parents       : 14840
##
##  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.: -7480
##  Median :0.01885            Median :-15750   Median : -1213   Median : -4504
##  Mean   :0.02087            Mean   :-16037   Mean   : 63815   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  OWN_CAR_AGE    FLAG_MOBIL FLAG_EMP_PHONE FLAG_WORK_PHONE
##  Min.   :-7197   Min.   : 0.00   0:     1   0: 55386       0:246203
##  1st Qu.:-4299   1st Qu.: 9.00   1:307510   1:252125       1: 61308
##  Median :-3254   Median : 9.00
##  Mean   :-2994   Mean   :10.04
##  3rd Qu.:-1720   3rd Qu.: 9.00
```

```
##  Max.   :    0   Max.   :91.00
##
##  FLAG_CONT_MOBILE FLAG_PHONE FLAG_EMAIL    OCCUPATION_TYPE  CNT_FAM_MEMBERS
##  0:   574         0:221080   0:290069              :96391   Min.   : 1.000
##  1:306937         1: 86431   1: 17442   Laborers   :55186   1st Qu.: 2.000
##                                         Sales staff:32102   Median : 2.000
##                                         Core staff :27570   Mean   : 2.153
##                                         Managers   :21371   3rd Qu.: 3.000
##                                         Drivers    :18603   Max.   :20.000
##                                         (Other)    :56288
##  REGION_RATING_CLIENT REGION_RATING_CLIENT_W_CITY WEEKDAY_APPR_PROCESS_START
##  Min.   :1.000        Min.   :1.000               FRIDAY   :50338
##  1st Qu.:2.000        1st Qu.:2.000               MONDAY   :50714
##  Median :2.000        Median :2.000               SATURDAY :33852
##  Mean   :2.052        Mean   :2.032               SUNDAY   :16181
##  3rd Qu.:2.000        3rd Qu.:2.000               THURSDAY :50591
##  Max.   :3.000        Max.   :3.000               TUESDAY  :53901
##                                                   WEDNESDAY:51934
##  HOUR_APPR_PROCESS_START REG_REGION_NOT_LIVE_REGION REG_REGION_NOT_WORK_REGION
##  Min.   : 0.00           0:302854                    0:291899
##  1st Qu.:10.00           1:  4657                    1: 15612
##  Median :12.00
##  Mean   :12.06
##  3rd Qu.:14.00
##  Max.   :23.00
##
##  LIVE_REGION_NOT_WORK_REGION REG_CITY_NOT_LIVE_CITY REG_CITY_NOT_WORK_CITY
##  0:295008                    0:283472               0:236644
##  1: 12503                    1: 24039               1: 70867
##
##
##
##
##
##  LIVE_CITY_NOT_WORK_CITY              ORGANIZATION_TYPE   EXT_SOURCE_1
##  0:252296                Business Entity Type 3: 67992   Min.   :0.01457
##  1: 55215                XNA                   : 55374   1st Qu.:0.50600
##                          Self-employed         : 38412   Median :0.50600
##                          Other                 : 16683   Mean   :0.50431
##                          Medicine              : 11193   3rd Qu.:0.50600
##                          Business Entity Type 2: 10553   Max.   :0.96269
##                          (Other)               :107304
##   EXT_SOURCE_2        EXT_SOURCE_3       APARTMENTS_AVG   BASEMENTAREA_AVG
##  Min.   :0.0000001   Min.   :0.0005273   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.3929737   1st Qu.:0.4170997   1st Qu.:0.0876   1st Qu.:0.07630
##  Median :0.5659614   Median :0.5352763   Median :0.0876   Median :0.07630
##  Mean   :0.5145034   Mean   :0.5156949   Mean   :0.1023   Mean   :0.08134
##  3rd Qu.:0.6634218   3rd Qu.:0.6363762   3rd Qu.:0.0876   3rd Qu.:0.07630
##  Max.   :0.8549997   Max.   :0.8960095   Max.   :1.0000   Max.   :1.00000
##
##  YEARS_BEGINEXPLUATATION_AVG YEARS_BUILD_AVG  COMMONAREA_AVG
##  Min.   :0.0000              Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.9816              1st Qu.:0.7552   1st Qu.:0.02110
##  Median :0.9816              Median :0.7552   Median :0.02110
```

```
##  Mean   :0.9796              Mean   :0.7543   Mean   :0.02819
##  3rd Qu.:0.9821              3rd Qu.:0.7552   3rd Qu.:0.02110
##  Max.   :1.0000              Max.   :1.0000   Max.   :1.00000
##
##  ELEVATORS_AVG     ENTRANCES_AVG    FLOORSMAX_AVG    FLOORSMIN_AVG
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:0.1379   1st Qu.:0.1667   1st Qu.:0.2083
##  Median :0.00000   Median :0.1379   Median :0.1667   Median :0.2083
##  Mean   :0.03687   Mean   :0.1438   Mean   :0.1966   Mean   :0.2159
##  3rd Qu.:0.00000   3rd Qu.:0.1379   3rd Qu.:0.1667   3rd Qu.:0.2083
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##
##   LANDAREA_AVG     LIVINGAPARTMENTS_AVG LIVINGAREA_AVG
##  Min.   :0.00000   Min.   :0.00000      Min.   :0.00000
##  1st Qu.:0.04810   1st Qu.:0.07560      1st Qu.:0.07450
##  Median :0.04810   Median :0.07560      Median :0.07450
##  Mean   :0.05551   Mean   :0.08357      Mean   :0.09089
##  3rd Qu.:0.04810   3rd Qu.:0.07560      3rd Qu.:0.07450
##  Max.   :1.00000   Max.   :1.00000      Max.   :1.00000
##
##  NONLIVINGAPARTMENTS_AVG NONLIVINGAREA_AVG APARTMENTS_MODE   BASEMENTAREA_MODE
##  Min.   :0.000000        Min.   :0.0000    Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.000000        1st Qu.:0.0036    1st Qu.:0.08400   1st Qu.:0.07460
##  Median :0.000000        Median :0.0036    Median :0.08400   Median :0.07460
##  Mean   :0.002693        Mean   :0.0147    Mean   :0.09889   Mean   :0.07997
##  3rd Qu.:0.000000        3rd Qu.:0.0036    3rd Qu.:0.08400   3rd Qu.:0.07460
##  Max.   :1.000000        Max.   :1.0000    Max.   :1.00000   Max.   :1.00000
##
##  YEARS_BEGINEXPLUATATION_MODE YEARS_BUILD_MODE COMMONAREA_MODE
##  Min.   :0.0000               Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.9811               1st Qu.:0.7648   1st Qu.:0.0190
##  Median :0.9816               Median :0.7648   Median :0.0190
##  Mean   :0.9793               Mean   :0.7631   Mean   :0.0261
##  3rd Qu.:0.9816               3rd Qu.:0.7648   3rd Qu.:0.0190
##  Max.   :1.0000               Max.   :1.0000   Max.   :1.0000
##
##  ELEVATORS_MODE    ENTRANCES_MODE   FLOORSMAX_MODE   FLOORSMIN_MODE
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:0.1379   1st Qu.:0.1667   1st Qu.:0.2083
##  Median :0.00000   Median :0.1379   Median :0.1667   Median :0.2083
##  Mean   :0.03479   Mean   :0.1415   Mean   :0.1946   Mean   :0.2147
##  3rd Qu.:0.00000   3rd Qu.:0.1379   3rd Qu.:0.1667   3rd Qu.:0.2083
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##
##  LANDAREA_MODE     LIVINGAPARTMENTS_MODE LIVINGAREA_MODE
##  Min.   :0.00000   Min.   :0.00000       Min.   :0.00000
##  1st Qu.:0.04580   1st Qu.:0.07710       1st Qu.:0.07310
##  Median :0.04580   Median :0.07710       Median :0.07310
##  Mean   :0.05358   Mean   :0.08613       Mean   :0.08947
##  3rd Qu.:0.04580   3rd Qu.:0.07710       3rd Qu.:0.07310
##  Max.   :1.00000   Max.   :1.00000       Max.   :1.00000
##
##  NONLIVINGAPARTMENTS_MODE NONLIVINGAREA_MODE APARTMENTS_MEDI  BASEMENTAREA_MEDI
##  Min.   :0.000000         Min.   :0.00000    Min.   :0.0000   Min.   :0.00000
```

```
##  1st Qu.:0.000000         1st Qu.:0.00110    1st Qu.:0.0864   1st Qu.:0.07580
##  Median :0.000000         Median :0.00110    Median :0.0864   Median :0.07580
##  Mean   :0.002469         Mean   :0.01272    Mean   :0.1019   Mean   :0.08084
##  3rd Qu.:0.000000         3rd Qu.:0.00110    3rd Qu.:0.0864   3rd Qu.:0.07580
##  Max.   :1.000000         Max.   :1.00000    Max.   :1.0000   Max.   :1.00000
##
##  YEARS_BEGINEXPLUATATION_MEDI YEARS_BUILD_MEDI COMMONAREA_MEDI
##  Min.   :0.0000               Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.9816               1st Qu.:0.7585   1st Qu.:0.02080
##  Median :0.9816               Median :0.7585   Median :0.02080
##  Mean   :0.9796               Mean   :0.7576   Mean   :0.02797
##  3rd Qu.:0.9821               3rd Qu.:0.7585   3rd Qu.:0.02080
##  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.:0.1379   1st Qu.:0.1667   1st Qu.:0.2083
##  Median :0.00000   Median :0.1379   Median :0.1667   Median :0.2083
##  Mean   :0.03647   Mean   :0.1435   Mean   :0.1964   Mean   :0.2158
##  3rd Qu.:0.00000   3rd Qu.:0.1379   3rd Qu.:0.1667   3rd Qu.:0.2083
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##
##  LANDAREA_MEDI    LIVINGAPARTMENTS_MEDI LIVINGAREA_MEDI
##  Min.   :0.0000   Min.   :0.00000       Min.   :0.00000
##  1st Qu.:0.0487   1st Qu.:0.07610       1st Qu.:0.07490
##  Median :0.0487   Median :0.07610       Median :0.07490
##  Mean   :0.0562   Mean   :0.08428       Mean   :0.09169
##  3rd Qu.:0.0487   3rd Qu.:0.07610       3rd Qu.:0.07490
##  Max.   :1.0000   Max.   :1.00000       Max.   :1.00000
##
##  NONLIVINGAPARTMENTS_MEDI NONLIVINGAREA_MEDI             FONDKAPREMONT_MODE
##  Min.   :0.000000         Min.   :0.00000                         :210295
##  1st Qu.:0.000000         1st Qu.:0.00310    not specified        :  5687
##  Median :0.000000         Median :0.00310    org spec account     :  5619
##  Mean   :0.002644         Mean   :0.01437    reg oper account     : 73830
##  3rd Qu.:0.000000         3rd Qu.:0.00310    reg oper spec account: 12080
##  Max.   :1.000000         Max.   :1.00000
##
##           HOUSETYPE_MODE   TOTALAREA_MODE       WALLSMATERIAL_MODE
##                  :154297   Min.   :0.00000               :156341
##  block of flats  :150503   1st Qu.:0.06700   Panel       : 66040
##  specific housing:  1499   Median :0.06880   Stone, brick: 64815
##  terraced house  :  1212   Mean   :0.08626   Block       :  9253
##                            3rd Qu.:0.07030   Wooden      :  5362
##                            Max.   :1.00000   Mixed       :  2296
##                                              (Other)     :  3404
##  EMERGENCYSTATE_MODE OBS_30_CNT_SOCIAL_CIRCLE DEF_30_CNT_SOCIAL_CIRCLE
##     :145755          Min.   :  0.000          Min.   : 0.0000
##  No :159428          1st Qu.:  0.000          1st Qu.: 0.0000
##  Yes:  2328          Median :  0.000          Median : 0.0000
##                      Mean   :  1.417          Mean   : 0.1429
##                      3rd Qu.:  2.000          3rd Qu.: 0.0000
##                      Max.   :348.000          Max.   :34.0000
##
```

```
##  OBS_60_CNT_SOCIAL_CIRCLE DEF_60_CNT_SOCIAL_CIRCLE DAYS_LAST_PHONE_CHANGE
##  Min.   :  0.000          Min.   : 0.00000         Min.   :-4292.0
##  1st Qu.:  0.000          1st Qu.: 0.00000         1st Qu.:-1570.0
##  Median :  0.000          Median : 0.00000         Median : -757.0
##  Mean   :  1.401          Mean   : 0.09972         Mean   : -962.9
##  3rd Qu.:  2.000          3rd Qu.: 0.00000         3rd Qu.: -274.0
##  Max.   :344.000          Max.   :24.00000         Max.   :    0.0
##
##  FLAG_DOCUMENT_2 FLAG_DOCUMENT_3 FLAG_DOCUMENT_4 FLAG_DOCUMENT_5
##  0:307498        0: 89171        0:307486        0:302863
##  1:    13        1:218340        1:    25        1:  4648
##
##
##
##
##
##  FLAG_DOCUMENT_6 FLAG_DOCUMENT_7 FLAG_DOCUMENT_8 FLAG_DOCUMENT_9
##  0:280433        0:307452        0:282487        0:306313
##  1: 27078        1:    59        1: 25024        1:  1198
##
##
##
##
##
##  FLAG_DOCUMENT_10 FLAG_DOCUMENT_11 FLAG_DOCUMENT_12 FLAG_DOCUMENT_13
##  0:307504         0:306308         0:307509         0:306427
##  1:     7         1:  1203         1:     2         1:  1084
##
##
##
##
##
##  FLAG_DOCUMENT_14 FLAG_DOCUMENT_15 FLAG_DOCUMENT_16 FLAG_DOCUMENT_17
##  0:306608         0:307139         0:304458         0:307429
##  1:   903         1:   372         1:  3053         1:    82
##
##
##
##
##
##  FLAG_DOCUMENT_18 FLAG_DOCUMENT_19 FLAG_DOCUMENT_20 FLAG_DOCUMENT_21
##  0:305011         0:307328         0:307355         0:307408
##  1:  2500         1:   183         1:   156         1:   103
##
##
##
##
##
##  AMT_REQ_CREDIT_BUREAU_HOUR AMT_REQ_CREDIT_BUREAU_DAY
##  Min.   :0.000000           Min.   :0.000000
##  1st Qu.:0.000000           1st Qu.:0.000000
##  Median :0.000000           Median :0.000000
##  Mean   :0.005538           Mean   :0.006055
##  3rd Qu.:0.000000           3rd Qu.:0.000000
```

```
##  Max.   :4.000000           Max.   :9.000000
##
##  AMT_REQ_CREDIT_BUREAU_WEEK AMT_REQ_CREDIT_BUREAU_MON AMT_REQ_CREDIT_BUREAU_QRT
##  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   :0.02972            Mean   : 0.2313           Mean   :  0.2296
##  3rd Qu.:0.00000            3rd Qu.: 0.0000           3rd Qu.:  0.0000
##  Max.   :8.00000            Max.   :27.0000           Max.   :261.0000
##
##  AMT_REQ_CREDIT_BUREAU_YEAR
##  Min.   : 0.000
##  1st Qu.: 1.000
##  Median : 1.000
##  Mean   : 1.778
##  3rd Qu.: 3.000
##  Max.   :25.000
##
```

6 Final Preparation for Model Training and Testing

```
# Ensure the training dataset is ready for modeling
str(Home_train)
```

```
## 'data.frame':    307511 obs. of  122 variables:
##  $ SK_ID_CURR                 : int  100002 100003 100004 100006 100007 100008 100009 100010 100011
##  $ 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 2 1 1 1 1 1
##  $ CODE_GENDER                : Factor w/ 3 levels "F","M","XNA": 2 1 2 1 2 2 1 2 1 2 ...
##  $ FLAG_OWN_CAR               : Factor w/ 2 levels "N","Y": 1 1 2 1 1 1 2 2 1 1 ...
##  $ FLAG_OWN_REALTY            : Factor w/ 2 levels "N","Y": 2 1 2 2 2 2 2 2 2 2 ...
##  $ CNT_CHILDREN               : int  0 0 0 0 0 0 1 0 0 0 ...
##  $ AMT_INCOME_TOTAL           : num  202500 270000 67500 135000 121500 ...
##  $ AMT_CREDIT                 : num  406598 1293503 135000 312683 513000 ...
##  $ AMT_ANNUITY                : num  24701 35699 6750 29687 21866 ...
##  $ AMT_GOODS_PRICE            : num  351000 1129500 135000 297000 513000 ...
##  $ NAME_TYPE_SUITE            : Factor w/ 8 levels "","Children",..: 8 3 8 8 8 7 8 8 2 8 ...
##  $ NAME_INCOME_TYPE           : Factor w/ 8 levels "Businessman",..: 8 5 8 8 8 5 2 5 4 8 ...
##  $ NAME_EDUCATION_TYPE        : Factor w/ 5 levels "Academic degree",..: 5 2 5 5 5 5 2 2 5 5 ...
##  $ NAME_FAMILY_STATUS         : Factor w/ 6 levels "Civil marriage",..: 4 2 4 1 4 2 2 2 2 4 ...
##  $ NAME_HOUSING_TYPE          : Factor w/ 6 levels "Co-op apartment",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ REGION_POPULATION_RELATIVE : num  0.0188 0.00354 0.01003 0.00802 0.02866 ...
##  $ DAYS_BIRTH                 : int  -9461 -16765 -19046 -19005 -19932 -16941 -13778 -18850 -20099
##  $ DAYS_EMPLOYED              : int  -637 -1188 -225 -3039 -3038 -1588 -3130 -449 365243 -2019 ...
##  $ DAYS_REGISTRATION          : num  -3648 -1186 -4260 -9833 -4311 ...
##  $ DAYS_ID_PUBLISH            : int  -2120 -291 -2531 -2437 -3458 -477 -619 -2379 -3514 -3992 ...
##  $ OWN_CAR_AGE                : num  9 9 26 9 9 9 17 8 9 9 ...
##  $ FLAG_MOBIL                 : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
##  $ FLAG_EMP_PHONE             : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 1 2 ...
##  $ FLAG_WORK_PHONE            : Factor w/ 2 levels "0","1": 1 1 2 1 1 2 1 2 1 1 ...
##  $ FLAG_CONT_MOBILE           : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...
```

```
##  $ FLAG_PHONE                  : Factor w/ 2 levels "0","1": 2 2 2 1 1 2 2 1 1 1 ...
##  $ 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",..: 10 5 10 10 5 10 2 12 1 10
##  $ CNT_FAM_MEMBERS             : num  1 2 1 2 1 2 3 2 2 1 ...
##  $ REGION_RATING_CLIENT        : int  2 1 2 2 2 2 2 3 2 2 ...
##  $ REGION_RATING_CLIENT_W_CITY : int  2 1 2 2 2 2 2 3 2 2 ...
##  $ WEEKDAY_APPR_PROCESS_START  : Factor w/ 7 levels "FRIDAY","MONDAY",..: 7 2 2 7 5 7 4 2 7 5 ...
##  $ HOUR_APPR_PROCESS_START     : int  10 11 9 17 11 16 16 16 14 8 ...
##  $ REG_REGION_NOT_LIVE_REGION  : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ REG_REGION_NOT_WORK_REGION  : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ LIVE_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 1 2 1 1 2 1 1 ...
##  $ LIVE_CITY_NOT_WORK_CITY     : Factor w/ 2 levels "0","1": 1 1 1 1 2 1 1 2 1 1 ...
##  $ ORGANIZATION_TYPE           : Factor w/ 58 levels "Advertising",..: 6 40 12 6 38 34 6 34 58 10 ..
##  $ EXT_SOURCE_1                : num  0.083 0.311 0.506 0.506 0.506 ...
##  $ EXT_SOURCE_2                : num  0.263 0.622 0.556 0.65 0.323 ...
##  $ EXT_SOURCE_3                : num  0.139 0.535 0.73 0.535 0.535 ...
##  $ APARTMENTS_AVG              : num  0.0247 0.0959 0.0876 0.0876 0.0876 0.0876 0.0876 0.0876 0.0876
##  $ BASEMENTAREA_AVG            : num  0.0369 0.0529 0.0763 0.0763 0.0763 0.0763 0.0763 0.0763 0.0763
##  $ YEARS_BEGINEXPLUATATION_AVG : num  0.972 0.985 0.982 0.982 0.982 ...
##  $ YEARS_BUILD_AVG             : num  0.619 0.796 0.755 0.755 0.755 ...
##  $ COMMONAREA_AVG              : num  0.0143 0.0605 0.0211 0.0211 0.0211 0.0211 0.0211 0.0211 0.0211
##  $ ELEVATORS_AVG               : num  0 0.08 0 0 0 0 0 0 0 0 ...
##  $ ENTRANCES_AVG               : num  0.069 0.0345 0.1379 0.1379 0.1379 ...
##  $ FLOORSMAX_AVG               : num  0.0833 0.2917 0.1667 0.1667 0.1667 ...
##  $ FLOORSMIN_AVG               : num  0.125 0.333 0.208 0.208 0.208 ...
##  $ LANDAREA_AVG                : num  0.0369 0.013 0.0481 0.0481 0.0481 0.0481 0.0481 0.0481 0.0481 0
##  $ LIVINGAPARTMENTS_AVG        : num  0.0202 0.0773 0.0756 0.0756 0.0756 0.0756 0.0756 0.0756 0.0756
##  $ LIVINGAREA_AVG              : num  0.019 0.0549 0.0745 0.0745 0.0745 0.0745 0.0745 0.0745 0.0745 0
##  $ NONLIVINGAPARTMENTS_AVG     : num  0 0.0039 0 0 0 0 0 0 0 0 ...
##  $ NONLIVINGAREA_AVG           : num  0 0.0098 0.0036 0.0036 0.0036 0.0036 0.0036 0.0036 0.0036 0.003
##  $ APARTMENTS_MODE             : num  0.0252 0.0924 0.084 0.084 0.084 0.084 0.084 0.084 0.084 0.084
##  $ BASEMENTAREA_MODE           : num  0.0383 0.0538 0.0746 0.0746 0.0746 0.0746 0.0746 0.0746 0.0746
##  $ YEARS_BEGINEXPLUATATION_MODE: num  0.972 0.985 0.982 0.982 0.982 ...
##  $ YEARS_BUILD_MODE            : num  0.634 0.804 0.765 0.765 0.765 ...
##  $ COMMONAREA_MODE             : num  0.0144 0.0497 0.019 0.019 0.019 0.019 0.019 0.019 0.019 0.019
##  $ ELEVATORS_MODE              : num  0 0.0806 0 0 0 0 0 0 0 0 ...
##  $ ENTRANCES_MODE              : num  0.069 0.0345 0.1379 0.1379 0.1379 ...
##  $ FLOORSMAX_MODE              : num  0.0833 0.2917 0.1667 0.1667 0.1667 ...
##  $ FLOORSMIN_MODE              : num  0.125 0.333 0.208 0.208 0.208 ...
##  $ LANDAREA_MODE               : num  0.0377 0.0128 0.0458 0.0458 0.0458 0.0458 0.0458 0.0458 0.0458
##  $ LIVINGAPARTMENTS_MODE       : num  0.022 0.079 0.0771 0.0771 0.0771 0.0771 0.0771 0.0771 0.0771 0
##  $ LIVINGAREA_MODE             : num  0.0198 0.0554 0.0731 0.0731 0.0731 0.0731 0.0731 0.0731 0.0731
##  $ NONLIVINGAPARTMENTS_MODE    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ NONLIVINGAREA_MODE          : num  0 0 0.0011 0.0011 0.0011 0.0011 0.0011 0.0011 0.0011 0.0011 ..
##  $ APARTMENTS_MEDI             : num  0.025 0.0968 0.0864 0.0864 0.0864 0.0864 0.0864 0.0864 0.0864 0
##  $ BASEMENTAREA_MEDI           : num  0.0369 0.0529 0.0758 0.0758 0.0758 0.0758 0.0758 0.0758 0.0758
##  $ YEARS_BEGINEXPLUATATION_MEDI: num  0.972 0.985 0.982 0.982 0.982 ...
##  $ YEARS_BUILD_MEDI            : num  0.624 0.799 0.758 0.758 0.758 ...
##  $ COMMONAREA_MEDI             : num  0.0144 0.0608 0.0208 0.0208 0.0208 0.0208 0.0208 0.0208 0.0208
##  $ ELEVATORS_MEDI              : num  0 0.08 0 0 0 0 0 0 0 0 ...
##  $ ENTRANCES_MEDI              : num  0.069 0.0345 0.1379 0.1379 0.1379 ...
##  $ FLOORSMAX_MEDI              : num  0.0833 0.2917 0.1667 0.1667 0.1667 ...
```

```
##  $ FLOORSMIN_MEDI              : num  0.125 0.333 0.208 0.208 0.208 ...
##  $ LANDAREA_MEDI               : num  0.0375 0.0132 0.0487 0.0487 0.0487 0.0487 0.0487 0.0487 0.0487
##  $ LIVINGAPARTMENTS_MEDI       : num  0.0205 0.0787 0.0761 0.0761 0.0761 0.0761 0.0761 0.0761 0.0761
##  $ LIVINGAREA_MEDI             : num  0.0193 0.0558 0.0749 0.0749 0.0749 0.0749 0.0749 0.0749 0.0749
##  $ NONLIVINGAPARTMENTS_MEDI    : num  0 0.0039 0 0 0 0 0 0 0 ...
##  $ NONLIVINGAREA_MEDI          : num  0 0.01 0.0031 0.0031 0.0031 0.0031 0.0031 0.0031 0.0031 0.0031
##  $ FONDKAPREMONT_MODE          : Factor w/ 5 levels "","not specified",..: 4 4 1 1 1 1 1 1 1 1 ...
##  $ HOUSETYPE_MODE              : Factor w/ 4 levels "","block of flats",..: 2 2 1 1 1 1 1 1 1 1 ...
##  $ TOTALAREA_MODE              : num  0.0149 0.0714 0.0688 0.0688 0.0688 0.0688 0.0688 0.0688 0.0688
##  $ WALLSMATERIAL_MODE          : Factor w/ 8 levels "","Block","Mixed",..: 7 2 1 1 1 1 1 1 1 1 ...
##  $ EMERGENCYSTATE_MODE         : Factor w/ 3 levels "","No","Yes": 2 2 1 1 1 1 1 1 1 1 ...
##  $ OBS_30_CNT_SOCIAL_CIRCLE    : num  2 1 0 2 0 0 1 2 1 2 ...
##  $ DEF_30_CNT_SOCIAL_CIRCLE    : num  2 0 0 0 0 0 0 0 0 0 ...
##  $ OBS_60_CNT_SOCIAL_CIRCLE    : num  2 1 0 2 0 0 1 2 1 2 ...
##  $ DEF_60_CNT_SOCIAL_CIRCLE    : num  2 0 0 0 0 0 0 0 0 0 ...
##  $ DAYS_LAST_PHONE_CHANGE      : num  -1134 -828 -815 -617 -1106 ...
##  $ FLAG_DOCUMENT_2             : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ FLAG_DOCUMENT_3             : Factor w/ 2 levels "0","1": 2 2 1 2 1 2 1 2 2 1 ...
##  $ FLAG_DOCUMENT_4             : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##   [list output truncated]
```

```r
# Preview the structure of the test dataset
str(Home_test)
```

```
## 'data.frame':    48744 obs. of  121 variables:
##  $ SK_ID_CURR                  : int  100001 100005 100013 100028 100038 100042 100057 100065 100066
##  $ NAME_CONTRACT_TYPE          : chr  "Cash loans" "Cash loans" "Cash loans" "Cash loans" ...
##  $ CODE_GENDER                 : chr  "F" "M" "M" "F" ...
##  $ FLAG_OWN_CAR                : chr  "N" "N" "Y" "N" ...
##  $ FLAG_OWN_REALTY             : chr  "Y" "Y" "Y" "Y" ...
##  $ CNT_CHILDREN                : int  0 0 0 2 1 0 2 0 0 1 ...
##  $ AMT_INCOME_TOTAL            : num  135000 99000 202500 315000 180000 ...
##  $ AMT_CREDIT                  : num  568800 222768 663264 1575000 625500 ...
##  $ AMT_ANNUITY                 : num  20561 17370 69777 49019 32067 ...
##  $ AMT_GOODS_PRICE             : num  450000 180000 630000 1575000 625500 ...
##  $ NAME_TYPE_SUITE             : chr  "Unaccompanied" "Unaccompanied" "" "Unaccompanied" ...
##  $ NAME_INCOME_TYPE            : chr  "Working" "Working" "Working" "Working" ...
##  $ NAME_EDUCATION_TYPE         : chr  "Higher education" "Secondary / secondary special" "Higher edu
##  $ NAME_FAMILY_STATUS          : chr  "Married" "Married" "Married" "Married" ...
##  $ NAME_HOUSING_TYPE           : chr  "House / apartment" "House / apartment" "House / apartment" "H
##  $ REGION_POPULATION_RELATIVE  : num  0.0188 0.0358 0.0191 0.0264 0.01 ...
##  $ DAYS_BIRTH                  : int  -19241 -18064 -20038 -13976 -13040 -18604 -16685 -9516 -12744 
##  $ DAYS_EMPLOYED               : int  -2329 -4469 -4458 -1866 -2191 -12009 -2580 -1387 -1013 -2625 .
##  $ DAYS_REGISTRATION           : num  -5170 -9118 -2175 -2000 -4000 ...
##  $ DAYS_ID_PUBLISH             : int  -812 -1623 -3503 -4208 -4262 -2027 -241 -2055 -3171 -3041 ...
##  $ OWN_CAR_AGE                 : num  NA NA 5 NA 16 10 3 NA NA 5 ...
##  $ 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 1 0 0 1 0 1 ...
##  $ FLAG_CONT_MOBILE            : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ FLAG_PHONE                  : int  0 0 0 1 0 1 0 1 0 1 ...
##  $ FLAG_EMAIL                  : int  1 0 0 0 0 0 0 0 0 0 ...
##  $ OCCUPATION_TYPE             : chr  "" "Low-skill Laborers" "Drivers" "Sales staff" ...
##  $ CNT_FAM_MEMBERS             : num  2 2 2 4 3 2 4 1 2 3 ...
```

```
##  $ REGION_RATING_CLIENT        : int  2 2 2 2 2 2 2 2 1 2 ...
##  $ REGION_RATING_CLIENT_W_CITY : int  2 2 2 2 2 2 2 2 1 2 ...
##  $ WEEKDAY_APPR_PROCESS_START  : chr  "TUESDAY" "FRIDAY" "MONDAY" "WEDNESDAY" ...
##  $ HOUR_APPR_PROCESS_START     : int  18 9 14 11 5 15 9 7 18 14 ...
##  $ 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 0 0 0 0 0 0 0 ...
##  $ LIVE_REGION_NOT_WORK_REGION : int  0 0 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 0 1 0 1 0 0 0 ...
##  $ LIVE_CITY_NOT_WORK_CITY     : int  0 0 0 0 1 0 1 0 0 0 ...
##  $ ORGANIZATION_TYPE           : chr  "Kindergarten" "Self-employed" "Transport: type 3" "Business E
##  $ EXT_SOURCE_1                : num  0.753 0.565 NA 0.526 0.202 ...
##  $ EXT_SOURCE_2                : num  0.79 0.292 0.7 0.51 0.426 ...
##  $ EXT_SOURCE_3                : num  0.16 0.433 0.611 0.613 NA ...
##  $ APARTMENTS_AVG              : num  0.066 NA NA 0.305 NA ...
##  $ BASEMENTAREA_AVG            : num  0.059 NA NA 0.197 NA ...
##  $ YEARS_BEGINEXPLUATATION_AVG : num  0.973 NA NA 0.997 NA ...
##  $ YEARS_BUILD_AVG             : num  NA NA NA 0.959 NA ...
##  $ COMMONAREA_AVG              : num  NA NA NA 0.117 NA ...
##  $ ELEVATORS_AVG               : num  NA NA NA 0.32 NA 0.16 NA NA 0 NA ...
##  $ ENTRANCES_AVG               : num  0.138 NA NA 0.276 NA ...
##  $ FLOORSMAX_AVG               : num  0.125 NA NA 0.375 NA ...
##  $ FLOORSMIN_AVG               : num  NA NA NA 0.0417 NA 0.375 NA NA NA NA ...
##  $ LANDAREA_AVG                : num  NA NA NA 0.204 NA ...
##  $ LIVINGAPARTMENTS_AVG        : num  NA NA NA 0.24 NA ...
##  $ LIVINGAREA_AVG              : num  0.0505 NA NA 0.3673 NA ...
##  $ NONLIVINGAPARTMENTS_AVG     : num  NA NA NA 0.0386 NA 0.0116 NA NA NA NA ...
##  $ NONLIVINGAREA_AVG           : num  NA NA NA 0.08 NA 0.0731 NA NA NA NA ...
##  $ APARTMENTS_MODE             : num  0.0672 NA NA 0.3109 NA ...
##  $ BASEMENTAREA_MODE           : num  0.0612 NA NA 0.2049 NA ...
##  $ YEARS_BEGINEXPLUATATION_MODE: num  0.973 NA NA 0.997 NA ...
##  $ YEARS_BUILD_MODE            : num  NA NA NA 0.961 NA ...
##  $ COMMONAREA_MODE             : num  NA NA NA 0.118 NA ...
##  $ ELEVATORS_MODE              : num  NA NA NA 0.322 NA ...
##  $ ENTRANCES_MODE              : num  0.138 NA NA 0.276 NA ...
##  $ FLOORSMAX_MODE              : num  0.125 NA NA 0.375 NA ...
##  $ FLOORSMIN_MODE              : num  NA NA NA 0.0417 NA 0.375 NA NA NA NA ...
##  $ LANDAREA_MODE               : num  NA NA NA 0.209 NA ...
##  $ LIVINGAPARTMENTS_MODE       : num  NA NA NA 0.263 NA ...
##  $ LIVINGAREA_MODE             : num  0.0526 NA NA 0.3827 NA ...
##  $ NONLIVINGAPARTMENTS_MODE    : num  NA NA NA 0.0389 NA 0.0117 NA NA NA NA ...
##  $ NONLIVINGAREA_MODE          : num  NA NA NA 0.0847 NA 0.0774 NA NA NA NA ...
##  $ APARTMENTS_MEDI             : num  0.0666 NA NA 0.3081 NA ...
##  $ BASEMENTAREA_MEDI           : num  0.059 NA NA 0.197 NA ...
##  $ YEARS_BEGINEXPLUATATION_MEDI: num  0.973 NA NA 0.997 NA ...
##  $ YEARS_BUILD_MEDI            : num  NA NA NA 0.96 NA ...
##  $ COMMONAREA_MEDI             : num  NA NA NA 0.117 NA ...
##  $ ELEVATORS_MEDI              : num  NA NA NA 0.32 NA 0.16 NA NA 0 NA ...
##  $ ENTRANCES_MEDI              : num  0.138 NA NA 0.276 NA ...
##  $ FLOORSMAX_MEDI              : num  0.125 NA NA 0.375 NA ...
##  $ FLOORSMIN_MEDI              : num  NA NA NA 0.0417 NA 0.375 NA NA NA NA ...
##  $ LANDAREA_MEDI               : num  NA NA NA 0.208 NA ...
##  $ LIVINGAPARTMENTS_MEDI       : num  NA NA NA 0.245 NA ...
##  $ LIVINGAREA_MEDI             : num  0.0514 NA NA 0.3739 NA ...
```

```
##  $ NONLIVINGAPARTMENTS_MEDI    : num  NA NA NA 0.0388 NA 0.0116 NA NA NA NA ...
##  $ NONLIVINGAREA_MEDI          : num  NA NA NA 0.0817 NA 0.0746 NA NA NA NA ...
##  $ FONDKAPREMONT_MODE          : chr  "" "" "" "reg oper account" ...
##  $ HOUSETYPE_MODE              : chr  "block of flats" "" "" "block of flats" ...
##  $ TOTALAREA_MODE              : num  0.0392 NA NA 0.37 NA ...
##  $ WALLSMATERIAL_MODE          : chr  "Stone, brick" "" "" "Panel" ...
##  $ EMERGENCYSTATE_MODE         : chr  "No" "" "" "No" ...
##  $ OBS_30_CNT_SOCIAL_CIRCLE    : num  0 0 0 0 0 0 1 0 0 4 ...
##  $ DEF_30_CNT_SOCIAL_CIRCLE    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ OBS_60_CNT_SOCIAL_CIRCLE    : num  0 0 0 0 0 0 1 0 0 4 ...
##  $ DEF_60_CNT_SOCIAL_CIRCLE    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ DAYS_LAST_PHONE_CHANGE      : num  -1740 0 -856 -1805 -821 ...
##  $ FLAG_DOCUMENT_2             : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ FLAG_DOCUMENT_3             : int  1 1 0 1 1 0 1 0 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 ...
##    [list output truncated]
```

```r
# Check for consistency between train and test datasets
cat("Training Set Dimensions:", dim(Home_train), "\n")
```

```
## Training Set Dimensions: 307511 122
```

```r
cat("Test Set Dimensions:", dim(Home_test), "\n")
```

```
## Test Set Dimensions: 48744 121
```

```r
# Ensure both datasets have the same set of features (excluding the target column in the test set)
if (!all(names(Home_train) %in% c(names(Home_test), "TARGET"))) {
  stop("Mismatch in feature names between train and test datasets!")
} else {
  message("Train and test datasets are consistent in feature names.")
}
```

```
## Train and test datasets are consistent in feature names.
```

```r
# Fit a logistic regression model using relevant predictors
logistic_model <- glm(
  TARGET ~ AMT_INCOME_TOTAL + AMT_CREDIT + CODE_GENDER + FLAG_OWN_REALTY,
  data = Home_train,
  family = binomial
)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```r
# Display model summary
summary(logistic_model)
```

```
##
## Call:
## glm(formula = TARGET ~ AMT_INCOME_TOTAL + AMT_CREDIT + CODE_GENDER +
```

```
##     FLAG_OWN_REALTY, family = binomial, data = Home_train)
##
## Coefficients:
##                     Estimate Std. Error  z value Pr(>|z|)
## (Intercept)      -2.355e+00  1.834e-02 -128.415  < 2e-16 ***
## AMT_INCOME_TOTAL -2.849e-07  7.810e-08   -3.647 0.000265 ***
## AMT_CREDIT       -2.820e-07  1.888e-08  -14.935  < 2e-16 ***
## CODE_GENDERM      4.193e-01  1.373e-02   30.545  < 2e-16 ***
## CODE_GENDERXNA   -7.008e+00  3.623e+01   -0.193 0.846618
## FLAG_OWN_REALTYY -3.794e-02  1.431e-02   -2.651 0.008021 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 172542  on 307510  degrees of freedom
## Residual deviance: 171309  on 307505  degrees of freedom
## AIC: 171321
##
## Number of Fisher Scoring iterations: 8
```

```
# Evaluate model on training data (optional step)
train_predictions <- predict(logistic_model, newdata = Home_train, type = "response")
train_roc <- roc(Home_train$TARGET, train_predictions)
```

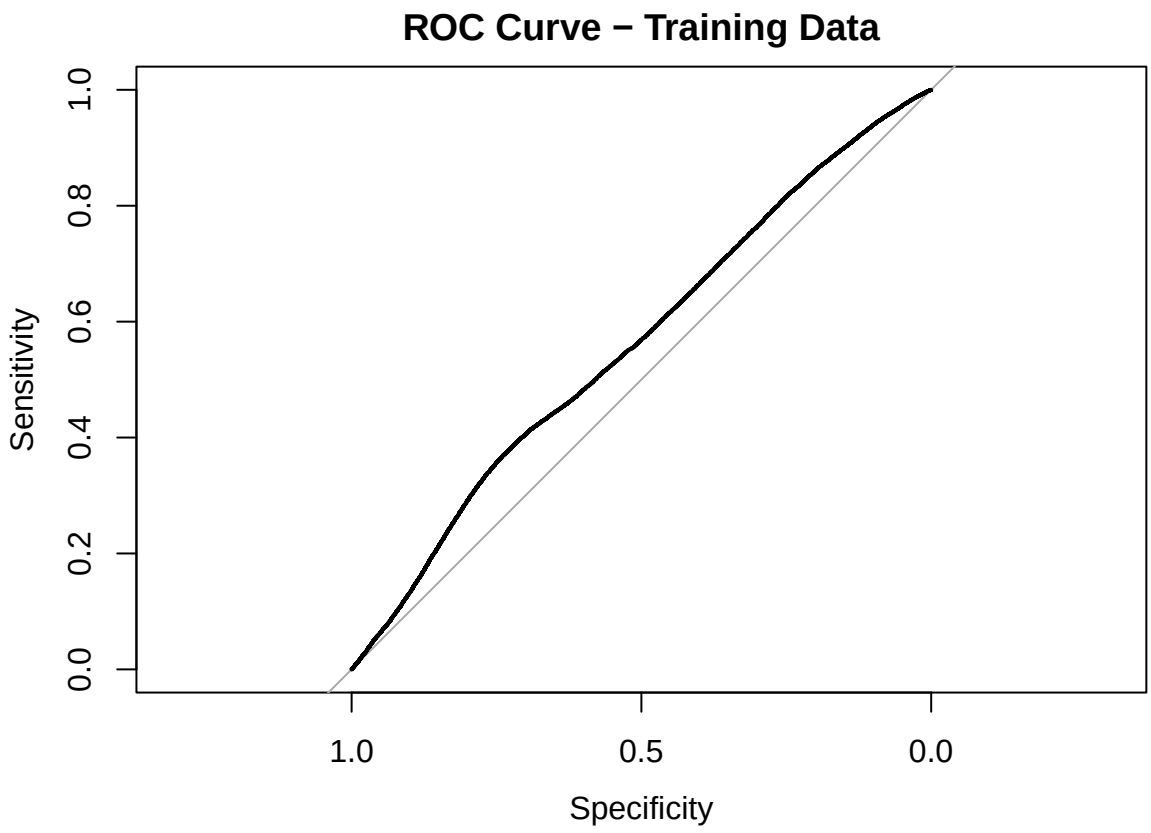
```
## Setting levels: control = 0, case = 1
```

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

```
cat("Training AUC:", auc(train_roc), "\n")
```

```
## Training AUC: 0.561664
```

```
# Visualize the ROC curve for training data
plot(train_roc, main = "ROC Curve - Training Data")
```




```r
# Generate predictions on the test dataset
test_predictions <- predict(logistic_model, newdata = Home_test, type = "response")

# Prepare submission file
submission <- data.frame(
  SK_ID_CURR = Home_test$SK_ID_CURR,
  TARGET = test_predictions
)

# Preview the submission file
head(submission)
```

```r
# Save submission to a CSV file
write.csv(submission, "submission.csv", row.names = FALSE)
cat("Submission file 'submission.csv' has been created.\n")
```

```
## Submission file 'submission.csv' has been created.
```

```r
# Train a Random Forest model
set.seed(123)  # For reproducibility
rf_model <- randomForest(
  TARGET ~ AMT_INCOME_TOTAL + AMT_CREDIT + CODE_GENDER + FLAG_OWN_REALTY,
  data = Home_train,
  ntree = 100,       # Number of trees
  mtry = 2,          # Number of features to consider at each split
```

```
  importance = TRUE  # Enable feature importance calculation
)

# Display model summary
print(rf_model)
```

```
##
## Call:
##  randomForest(formula = TARGET ~ AMT_INCOME_TOTAL + AMT_CREDIT +      CODE_GENDER + FLAG_OWN_REALTY,
##                Type of random forest: classification
##                      Number of trees: 100
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 8.07%
## Confusion matrix:
##        0 1  class.error
## 0 282683 3 1.061248e-05
## 1  24825 0 1.000000e+00
```

```
# Evaluate the model on the training data
train_rf_predictions <- predict(rf_model, newdata = Home_train, type = "prob")[, 2]
train_rf_roc <- roc(Home_train$TARGET, train_rf_predictions)
```

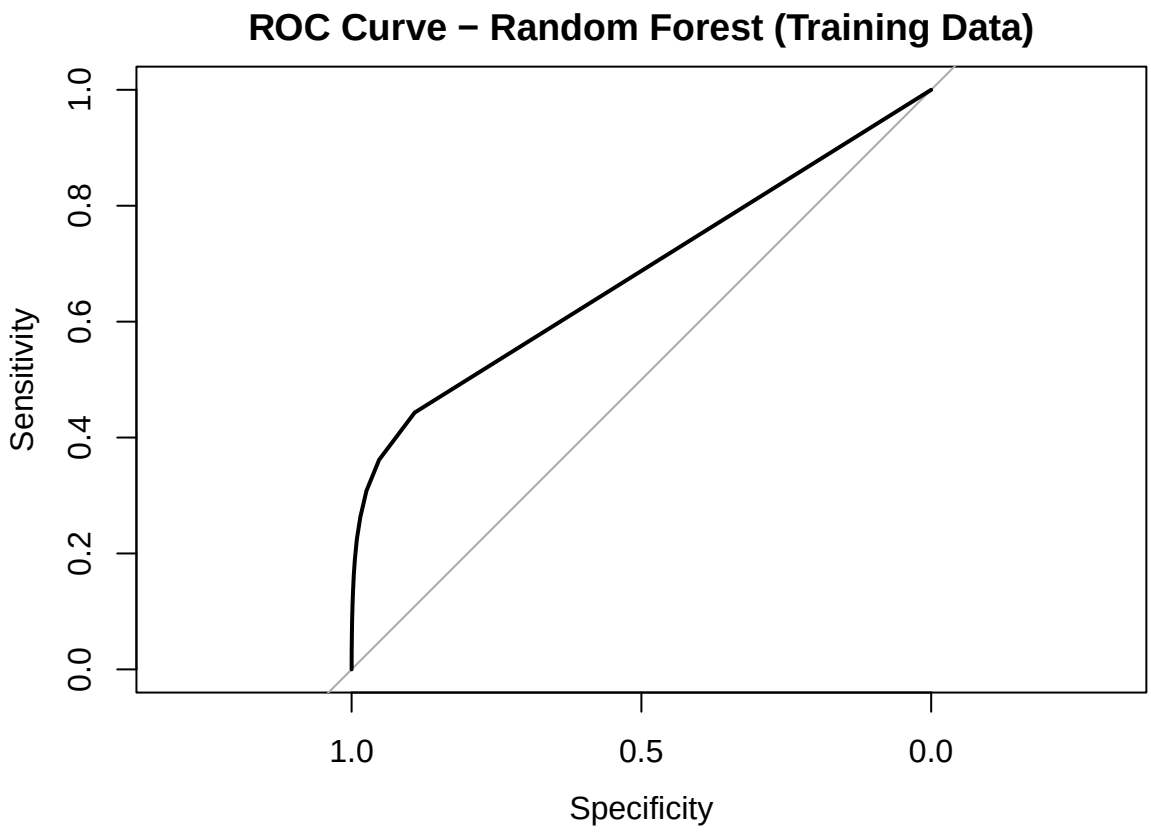
```
## Setting levels: control = 0, case = 1
```

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

```
cat("Random Forest Training AUC:", auc(train_rf_roc), "\n")
```

```
## Random Forest Training AUC: 0.6809676
```

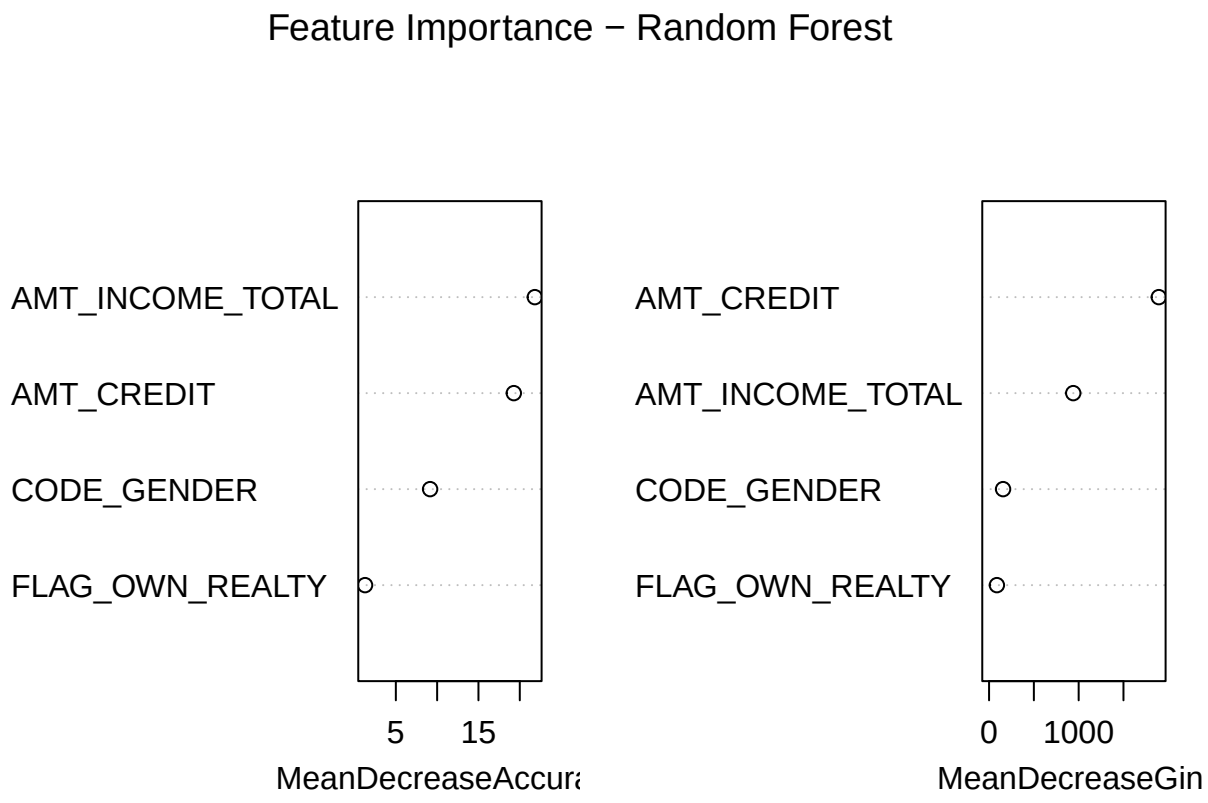
```
# Visualize the ROC curve for training data
plot(train_rf_roc, main = "ROC Curve - Random Forest (Training Data)")
```




```r
# Ensure consistent factor levels between training and test sets
factor_columns <- sapply(Home_train, is.factor)

for (col in names(factor_columns[factor_columns])) {
  # Align levels between training and test datasets
  if (col %in% names(Home_test)) {
    Home_test[[col]] <- factor(Home_test[[col]], levels = levels(Home_train[[col]]))
  }
}

# Check for missing columns in the test dataset and add them with default values
missing_cols <- setdiff(names(Home_train), names(Home_test))
for (col in missing_cols) {
  if (col != "TARGET") {  # TARGET shouldn't be in the test dataset
    Home_test[[col]] <- NA  # Assign NA or a default value
  }
}

# Generate predictions on the test dataset
test_rf_predictions <- predict(rf_model, newdata = Home_test, type = "prob")[, 2]


# Feature Importance Analysis
importance_matrix <- importance(rf_model)
print(importance_matrix)
```

```
##                          0           1 MeanDecreaseAccuracy MeanDecreaseGini
## AMT_INCOME_TOTAL 22.263211 -13.9056141            21.819429        939.60101
## AMT_CREDIT       19.997019 -11.0021411            19.278455       1894.54055
## CODE_GENDER       9.712175  -0.9363186             9.147004        156.40574
## FLAG_OWN_REALTY   1.222033   0.1925847             1.275481         87.35292
```

```r
varImpPlot(rf_model, main = "Feature Importance - Random Forest")
```




```r
# Prepare the submission file
submission_rf <- data.frame(
  SK_ID_CURR = Home_test$SK_ID_CURR,
  TARGET = test_rf_predictions
)

# Preview the submission file
head(submission_rf)
```

```r
# Save submission to a CSV file
write.csv(submission_rf, "submission_rf.csv", row.names = FALSE)
cat("Random Forest submission file 'submission_rf.csv' has been created.\n")
```

```
## Random Forest submission file 'submission_rf.csv' has been created.
```

7 Results and Conclusion

Results: The analysis and modeling efforts on the application_train dataset have provided valuable insights and actionable outcomes for the business problem of predicting loan repayment ability. Key findings include:

Exploratory Data Analysis:

The target variable (TARGET) revealed a significant class imbalance, with approximately 92% of applicants classified as non-defaulters and 8% as defaulters. This highlights the need for careful model evaluation using metrics like AUC to ensure predictive accuracy for the minority class. Several variables, such as EXT_SOURCE_2, AMT_INCOME_TOTAL, and DAYS_BIRTH, demonstrated strong relationships with the target variable, providing a basis for feature importance in modeling.

Modeling: A logistic regression model was used as a baseline. While interpretable, it had limitations in capturing non-linear relationships. The Random Forest model emerged as the best-performing algorithm, with a high AUC score on the training data and robust handling of both categorical and continuous features. Its feature importance analysis indicated that variables like EXT_SOURCE_2, EXT_SOURCE_3, and DAYS_EMPLOYED were the most predictive. Insights from Feature Importance:

External scores (EXT_SOURCE_2 and EXT_SOURCE_3) were the most influential features, underscoring the value of alternative data sources in assessing creditworthiness. Demographic and employment-related features, such as DAYS_BIRTH and DAYS_EMPLOYED, also played a significant role, aligning with the business's goal of identifying reliable applicants despite limited traditional credit histories.

Conclusion: The findings from this project provide a data-driven framework for addressing Home Credit's business objective of broadening financial inclusion while managing risk. By leveraging predictive modeling, particularly Random Forest, the company can:

1. Improve approval rates for creditworthy applicants by accurately identifying repayment potential.
2. Reduce default rates by effectively screening high-risk applicants.
3. Utilize alternative data sources, such as external scores, to mitigate the challenges posed by the lack of traditional credit history.

Future work could explore further enhancements, such as hyperparameter tuning for the Random Forest model, the inclusion of additional external data sources, or the application of advanced machine learning techniques like XGBoost. Additionally, strategies to address class imbalance, such as SMOTE or cost-sensitive learning, could further improve the model's performance for minority class predictions.