# HOME CREDIT: ACCURATE LOAN REPAYMENT PREDICTIONS WITH MACHINE LEARNING MODEL

Final Project Report

Group members' name: Yayuan Zhang, Peiling Zou

ISM 6353 Programming for Data Analytics: R & Machine Learning

Professor: Dr. Andy Chen

Seattle Pacific University

3/21/2023

Contents

## Abstract

Loan default is an inevitable risk in the loan industry and a severe challenge faced by many lending institutions. In this project, we used R programming to build a predictive model that can predict whether a loan applicant will default on their loan.

The article would display an outline of each phase of the data-mining process, including the business insight and findings, data pre-processing, feature engineering, exploratory data analysis, and the impact of variables on repayment, data analysis, train-test split approaches, the tweaking of parameters, modeling, and evaluation. We hope that our experience can inspire and inform others working in the loan industry or interested in machine learning.

**Keywords**: R programming, Prediction model, Loan default, Data pre-processing, Exploratory data analysis, Machine learning.

**R Notebooks**: Home_Credit_Final_Project_EDA_Data_Preprocessing, Home_Credit_Final_Project_Modeling

## 1. Business insight

### 1.1 Business Problems and Motivations

This article aims to address a significant problem in the lending industry – predicting the likelihood of loan default by borrowers. The inability to repay loans is a widespread issue that can lead to negative consequences for both lenders and borrowers. Lending institutions need to accurately assess the creditworthiness of their potential borrowers to mitigate the risk of loan default and minimize financial losses by using machine learning prediction model. The successful development of such models would significantly improve the efficiency and accuracy of credit risk management in the lending industry, benefiting both lenders and borrowers.

In this article, we explored the data of Home Credit Group, which is an international consumer finance provider with operations in eight countries. They focus on responsible lending primarily to people with little or no credit history, such as self-employed and low-income individuals and organizations. The main challenge faced by Home Credit is how to accurately assess the default risk of their clients before issuing loans. to avoid facing financial losses or negative impacts if their customer is unable to repay on time. Therefore, Home Credit need machine learning and data analysis techniques to develop an accurate predictive model to better understand their client's financial status and take appropriate measures.

### 1.2 Stakeholders

In this project, there are three primary stakeholders involved in the loan default prediction model. Firstly, the management team and risk management team at Home Credit have a vested interest in ensuring the financial stability of the company and reducing the risk of loan defaults. By utilizing an

accurate and efficient predictive model, Home Credit can better evaluate the creditworthiness of their borrowers and minimize potential losses resulting from loan defaults.

Secondly, qualified loan applicants seeking financing from Home Credit also benefit from the implementation of this model. By accurately assessing the borrower's creditworthiness, qualified applicants have a higher chance of being approved for a loan and obtaining the necessary financing to fulfill their needs.

Finally, the wider lending industry also has an interest in this project. The successful development and implementation of a predictive model could revolutionize the way credit risk management is performed across the industry. By utilizing machine learning techniques and data analysis, lending institutions could better evaluate the creditworthiness of their borrowers, ultimately leading to more efficient and accurate loan decisions.

### 1.3 Source

Data comes from a competition on Kaggle for Home Credit Group (https://www.kaggle.com/competitions/home-credit-default-risk).

### 1.4 Data

The dataset used in this project contains 307,511 samples, each with 122 attributes. Each row represents basic information about the loan applicant, such as their gender, education level, family status, employment status, and flag owning a car or a house. It is worth noting that 92% of the loan applicants in the dataset repay their loans on time, while 8% default on their loans. This results in an imbalanced data problem that needs to be addressed in the modeling process.

Several techniques were implemented to handle this issue, including threshold adjustment and Stochastic Gradient Boosting. These techniques aimed to improve the model's performance by addressing the class imbalance problem and allowing for a more accurate prediction of loan default risk. The process of identifying and addressing these challenges is crucial to the success of the model and ultimately the effectiveness of the credit risk management process for Home Credit. We will talk about details of the modeling part.

## 2. Data Analysis

### 2.1 Workflow Analysis process

Throughout the workflow analysis process, firstly we import libraries and datasets to understand the data, including but not limited to understanding the meaning of features, defining target variables, selecting features, removing irrelevant features, checking data types, checking for missing values, checking for outliers, checking statistical values of features, and checking relationships between features. Exploratory data analysis was performed after we had processed the necessary data preparation for the model, such as data cleaning and feature engineering.

In the predictive modeling process, we divided the data into training and testing sets and run into 3 different models and determine the performance of model with the relevant important metrics, such as highest accuracy, AUC values, and p-values etc. Finally, we tweaked the parameters to improve the performance of the three models.

## 2.2 Data Understanding & Data Pre-processing

### 2.2.1 Target variable

Firstly, we identify TARGET as our target variable. The "TARGET" indicates whether the client has loan default or not, if it is "1", the client has loan default which means this client might have some difficulties paying the loan on time; if it is "0", the client pays the loan on time. As "TARGET" is our predictor variable in our models.

Then, we analyzed the target distribution. It can be seen that the number of people who repay the loan on time is 282686, and the number of people who do not repay the loan on time is 24825.

| TARGET <int> | n <int> |
|---|---|
| 0 | 282686 |
| 1 | 24825 |

### 2.2.2 Removing and selecting features

In the data pre-processing, we performed four key steps to prepare the data for the model.

One of step is that we dropped all outsource features and selected relevant features according to the heatmap. For the part of features selection, we filter out the correlation values greater than 80% to avoid overfitting. However, highly correlated variables may suffer from multicollinearity issues, and it might lead to overfitting and inaccurate results.

For the part of removing all outsource features, we examined the missing values in each column and the percentage of missing values exceeding 45% in that column. From the following table, it can be seen that these columns are all from external resources.

| variables <chr> | Missing_values <dbl> | % of Total Values <dbl> |
|---|---|---|
| COMMONAREA_AVG | 214865 | 69.9 |
| COMMONAREA_MODE | 214865 | 69.9 |
| COMMONAREA_MEDI | 214865 | 69.9 |
| NONLIVINGAPARTMENTS_AVG | 213514 | 69.4 |
| NONLIVINGAPARTMENTS_MODE | 213514 | 69.4 |
| NONLIVINGAPARTMENTS_MEDI | 213514 | 69.4 |
| LIVINGAPARTMENTS_AVG | 210199 | 68.4 |
| LIVINGAPARTMENTS_MODE | 210199 | 68.4 |
| LIVINGAPARTMENTS_MEDI | 210199 | 68.4 |
| FLOORSMIN_AVG | 208642 | 67.8 |

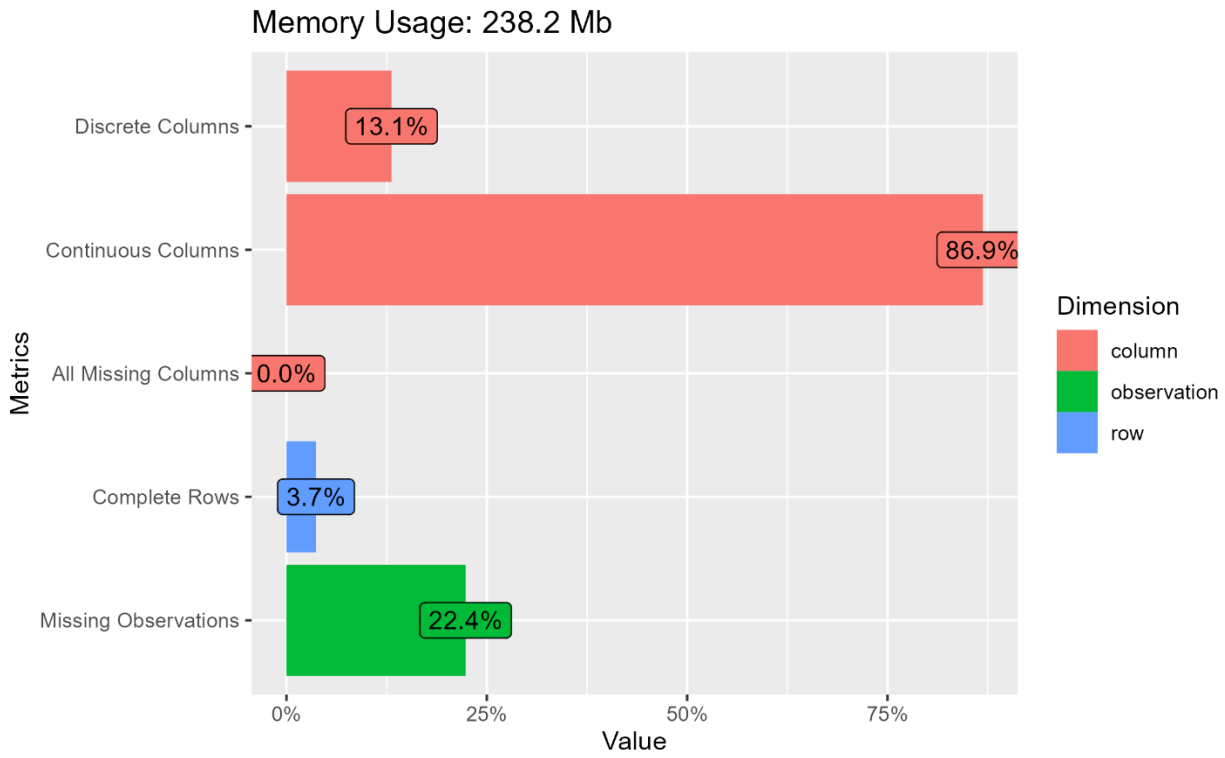
1-10 of 45 rows

After the above data processing, our dataset has 307511 rows and 29 columns.

| SK_ID_CURR <int> | TARG... <int> | NAME_CONTRACT_TY... <chr> |
|---|---|---|
| 100002 | 1 | Cash loans |
| 100003 | 0 | Cash loans |
| 100004 | 0 | Revolving loans |
| 100006 | 0 | Cash loans |
| 100007 | 0 | Cash loans |
| 100008 | 0 | Cash loans |
| 100009 | 0 | Cash loans |
| 100010 | 0 | Cash loans |
| 100011 | 0 | Cash loans |
| 100012 | 0 | Revolving loans |

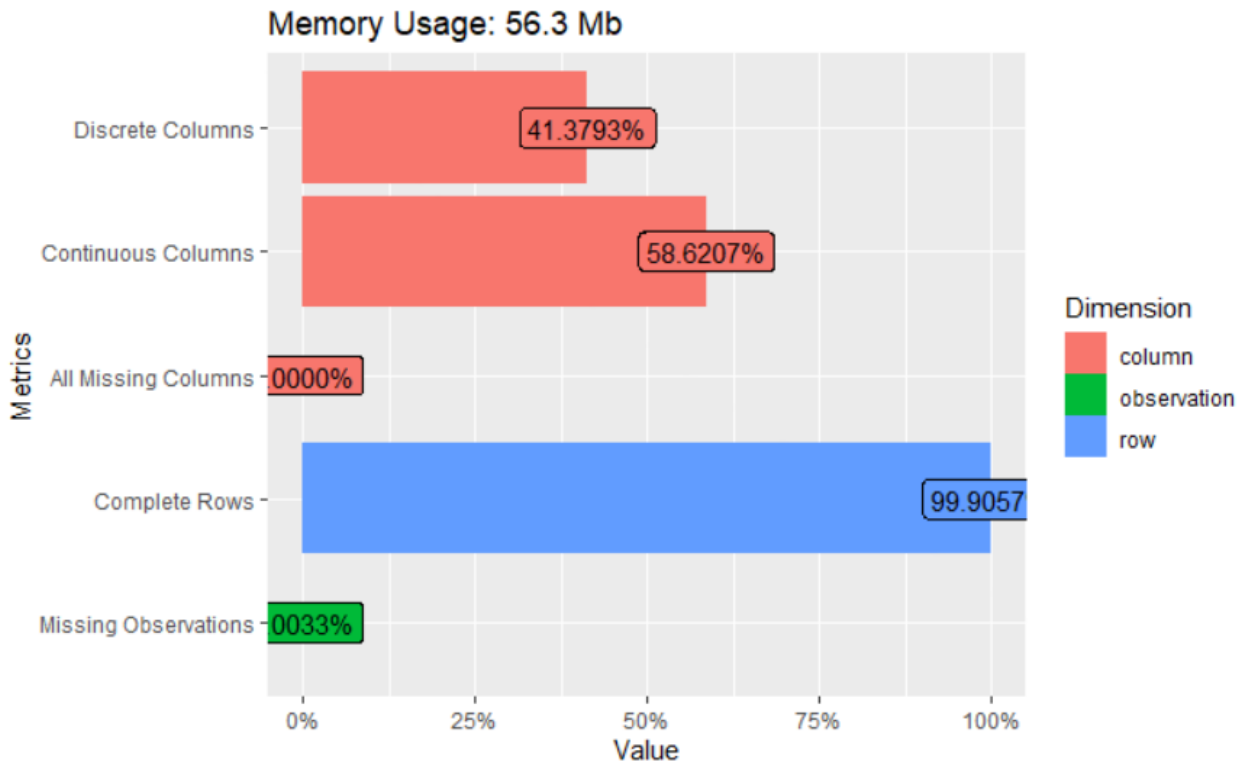
1-10 of 307,511 rows | 1-5 of 29 colu... Previous

### 2.2.3 Imputing missing values with median values




The original dataset had 61 columns missing values and 22.4% missing observations, so we imputed the missing values with the median values of their feature. A completeness rate of 3.7% means that only a small percentage of your data is fully complete, while a missing observations rate of 22.4% suggests that there might be a significant amount of missing information in the dataset.

It could have several implications for the data analysis. The missing data can result in biased estimates and reduced statistical power, as well as limit the generalizability of your findings. Next, we used multiple imputation techniques to address these issues.




After the above the step of remove and select features, we checked the complete rows is 99.91% and missing observations is almost zero. We checked missing values in each column by using coding - data %>% skim() %>% kable().

| skim_type | skim_variable | n_missing | complete_rate | character.min | character.max | chara |
|---|---|---|---|---|---|---|
| character | NAME_CONTRACT_TYPE | 0 | 1.0000000 | 10 | 15 | |
| character | CODE_GENDER | 0 | 1.0000000 | 1 | 3 | |
| character | FLAG_OWN_CAR | 0 | 1.0000000 | 1 | 1 | |
| character | FLAG_OWN_REALTY | 0 | 1.0000000 | 1 | 1 | |
| character | NAME_TYPE_SUITE | 0 | 1.0000000 | 0 | 15 | |
| character | NAME_INCOME_TYPE | 0 | 1.0000000 | 7 | 20 | |
| character | NAME_EDUCATION_TYPE | 0 | 1.0000000 | 15 | 29 | |
| character | NAME_FAMILY_STATUS | 0 | 1.0000000 | 5 | 20 | |
| character | NAME_HOUSING_TYPE | 0 | 1.0000000 | 12 | 19 | |
| character | OCCUPATION_TYPE | 0 | 1.0000000 | 0 | 21 | |
| character | WEEKDAY_APPR_PROCESS_START | 0 | 1.0000000 | 6 | 9 | |
| character | ORGANIZATION_TYPE | 0 | 1.0000000 | 3 | 22 | |
| numeric | SK_ID_CURR | 0 | 1.0000000 | NA | NA | |
| numeric | TARGET | 0 | 1.0000000 | NA | NA | |
| numeric | CNT_CHILDREN | 0 | 1.0000000 | NA | NA | |
| numeric | AMT_INCOME_TOTAL | 0 | 1.0000000 | NA | NA | |
| numeric | AMT_CREDIT | 0 | 1.0000000 | NA | NA | |
| numeric | AMT_ANNUITY | 12 | 0.9999610 | NA | NA | |
| numeric | AMT_GOODS_PRICE | 278 | 0.9990960 | NA | NA | |

According to the above table, it can be seen that the column for AMT_ANNUITY, AMT_GOODS_PRICE, and CNT_FAM_MEMBERS. Therefore, we imputing these missing values using median values of its features until the complete rows is 100 percent and missing observations is zero percent.

### 2.2.4 Handling outlier values

| DAYS_BIRTH <int> | DAYS_EMPLOYED <int> | DAYS_REGISTRATION <dbl> | DAYS_ID_PUBLISH <int> | AGE <dbl> |
|---|---|---|---|---|
| -9461 | -637 | -3648 | -2120 | 26 |
| -16765 | -1188 | -1186 | -291 | 46 |
| -19046 | -225 | -4260 | -2531 | 52 |
| -19005 | -3039 | -9833 | -2437 | 52 |
| -19932 | -3038 | -4311 | -3458 | 55 |
| -16941 | -1588 | -4970 | -477 | 46 |
| -13778 | -3130 | -1213 | -619 | 38 |
| -18850 | -449 | -4597 | -2379 | 52 |

We also noticed four features displaying negative values, indicating the number of days before the application submission date, such as negative values on the Birth variable indicating that the applicant was born before how many days of application. We convert the negative values as days to positive values as years and make them more meaningful.

| AGE <dbl> | DAYS_EMPLOYED <dbl> | DAYS_REGISTRATION <dbl> | DAYS_ID_PUBLISH <dbl> |
|---|---|---|---|
| 26 | 2 | 10 | 6 |
| 46 | 3 | 3 | 1 |
| 52 | 1 | 12 | 7 |
| 52 | 8 | 27 | 7 |
| 55 | 8 | 12 | 9 |
| 46 | 4 | 14 | 1 |
| 38 | 9 | 3 | 2 |
| 52 | 1 | 13 | 7 |

### 2.2.5 Adding nine new features

Additionally, we added nine new features based on the original dataset's features to improve the model's accuracy and performance using the following coding, as the original features may not capture the complex relationships between the target variables. For instance, The DAYS_EMPLOYED_PERC feature (percentage of client's lifespan that they have been employed) could be a useful measure of employment stability, which is an important factor in assessing a client's ability to repay a loan.

```
DAYS_EMPLOYED_PERC = sqrt(DAYS_EMPLOYED / DAYS_BIRTH),
INCOME_CREDIT_PERC = AMT_INCOME_TOTAL / AMT_CREDIT,
INCOME_PER_PERSON = log1p(AMT_INCOME_TOTAL / CNT_FAM_MEMBERS),
ANNUITY_INCOME_PERC = sqrt(AMT_ANNUITY / (1 + AMT_INCOME_TOTAL)),
LOAN_INCOME_RATIO = AMT_CREDIT / AMT_INCOME_TOTAL,
ANNUITY_LENGTH = AMT_CREDIT / AMT_ANNUITY,
CHILDREN_RATIO = CNT_CHILDREN / CNT_FAM_MEMBERS,
CREDIT_TO_GOODS_RATIO = AMT_CREDIT / AMT_GOODS_PRICE,
INC_PER_CHLD = AMT_INCOME_TOTAL / (1 + CNT_CHILDREN),)
```

- DAYS_EMPLOYED_PERC: The percentage of the client's lifespan (measured in days) that they have been employed for.
- INCOME_CREDIT_PERC: The ratio of the client's income to the amount of credit they are applying for.
- INCOME_PER_PERSON: The natural logarithm of the client's income divided by the number of family members.
- ANNUITY_INCOME_PERC: The ratio of the client's income to their monthly annuity payment.
- LOAN_INCOME_RATIO: The ratio of the amount of credit the client is applying for to their income.
- ANNUITY_LENGTH: The length of the client's annuity payment, measured in months.
- CHILDREN_RATIO: The ratio of the number of children the client has to the number of family members.
- CREDIT_TO_GOODS_RATIO: The ratio of the amount of credit the client is applying for to the price of the goods they are purchasing with the credit.
- INC_PER_CHLD: The client's income divided by the number of children they have plus 1.

So, we added nine new features into dataset and obtain the final dataset have 38 features in total.

## 2.3 Reading data again

Before we delve into the modeling part, we examined the processed dataset that was fed into the model. After data preprocessing, our dataset was reduced from 122 features to 38, including 9 new features.

| SK_ID_CURR <int> | TARGET <int> | NAME_CONTRACT_TYPE <chr> |
|---|---|---|
| 100002 | 1 | Cash loans |
| 100003 | 0 | Cash loans |
| 100004 | 0 | Revolving loans |
| 100006 | 0 | Cash loans |
| 100007 | 0 | Cash loans |
| 100008 | 0 | Cash loans |
| 100009 | 0 | Cash loans |
| 100010 | 0 | Cash loans |
| 100011 | 0 | Cash loans |
| 100012 | 0 | Revolving loans |

1-10 of 307,511 rows | 1-10 of 122 columns

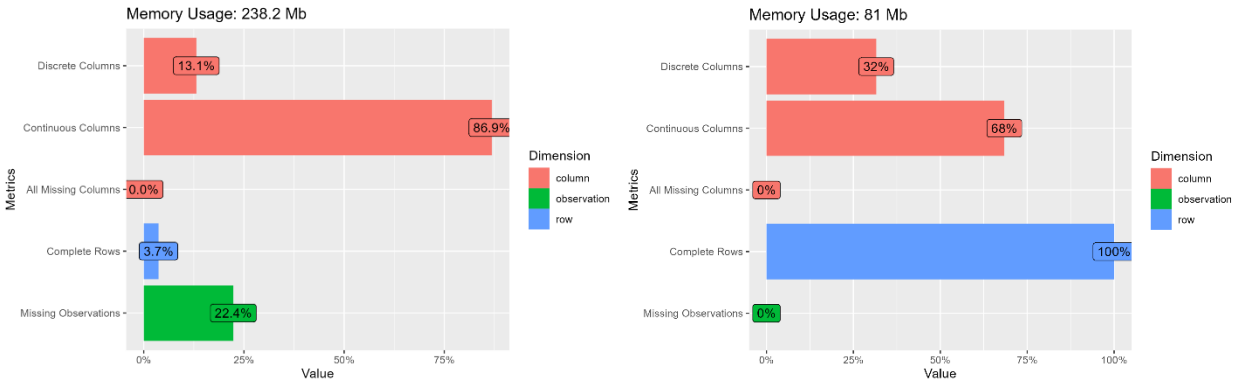
| SK_ID_CURR <int> | TARGET <int> | NAME_CONTRACT_TYPE <chr> |
|---|---|---|
| 100002 | 1 | Cash loans |
| 100003 | 0 | Cash loans |
| 100004 | 0 | Revolving loans |
| 100006 | 0 | Cash loans |
| 100007 | 0 | Cash loans |
| 100008 | 0 | Cash loans |
| 100009 | 0 | Cash loans |
| 100010 | 0 | Cash loans |
| 100011 | 0 | Cash loans |
| 100012 | 0 | Revolving loans |

1-10 of 307,511 rows | 1-10 of 38 columns

We converted negative values in "days" to positive values and displayed them in years for better understanding.

| DAYS_BIRTH <int> | DAYS_EMPLOYED <int> | DAYS_REGISTRATION <dbl> | DAYS_ID_PUBLISH <int> | AGE <dbl> | DAYS_EMPLOYED <dbl> | DAYS_REGISTRATION <dbl> | DAYS_ID_PUBLISH <dbl> |
|---|---|---|---|---|---|---|---|
| -9461 | -637 | -3648 | -2120 | 26 | 2 | 10 | 6 |
| -16765 | -1188 | -1186 | -291 | 46 | 3 | 3 | 1 |
| -19046 | -225 | -4260 | -2531 | 52 | 1 | 12 | 7 |
| -19005 | -3039 | -9833 | -2437 | 52 | 8 | 27 | 7 |
| -19932 | -3038 | -4311 | -3458 | 55 | 8 | 12 | 9 |
| -16941 | -1588 | -4970 | -477 | 46 | 4 | 14 | 1 |
| -13778 | -3130 | -1213 | -619 | 38 | 9 | 3 | 2 |
| -18850 | -449 | -4597 | -2379 | 52 | 1 | 13 | 7 |

We filled in 22.4% of missing observations, improving the completeness of each row to 100%.




From these two tables, it is clear that the complete rate of each row is 1, indicating that there are no missing values in each row anymore.

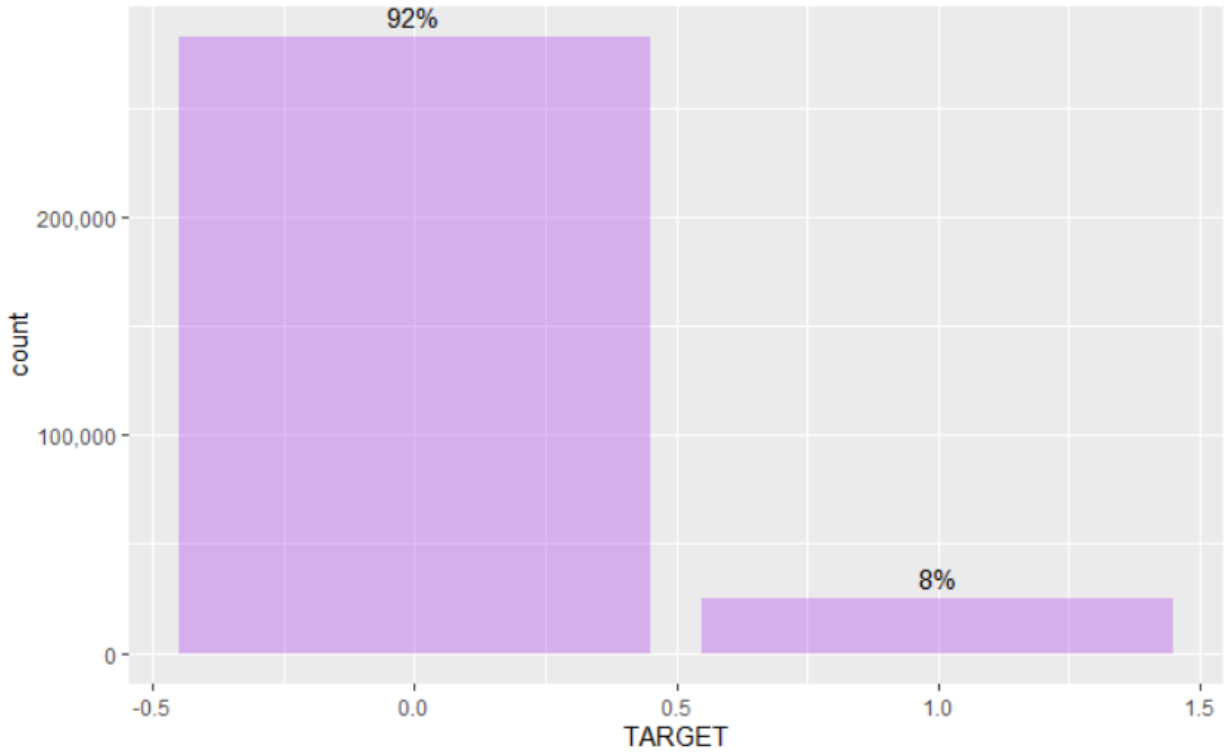
| skim_type | skim_variable | n_missing | complete_rate |
|---|---|---|---|
| numeric | APARTMENTS_AVG | 156061 | 0.4925027 |
| numeric | BASEMENTAREA_AVG | 179943 | 0.4148404 |
| numeric | YEARS_BEGINEXPLUATATION_AVG | 150007 | 0.5121898 |
| numeric | YEARS_BUILD_AVG | 204488 | 0.3350222 |
| numeric | COMMONAREA_AVG | 214865 | 0.3012770 |
| numeric | ELEVATORS_AVG | 163891 | 0.4670402 |
| numeric | ENTRANCES_AVG | 154828 | 0.4965123 |
| numeric | FLOORSMAX_AVG | 153020 | 0.5023918 |
| numeric | FLOORSMIN_AVG | 208642 | 0.3215137 |
| numeric | LANDAREA_AVG | 182590 | 0.4062326 |
| numeric | LIVINGAPARTMENTS_AVG | 210199 | 0.3164505 |
| numeric | LIVINGAREA_AVG | 154350 | 0.4980667 |
| numeric | NONLIVINGAPARTMENTS_AVG | 213514 | 0.3056704 |
| numeric | NONLIVINGAREA_AVG | 169682 | 0.4482084 |
| numeric | APARTMENTS_MODE | 156061 | 0.4925027 |
| numeric | BASEMENTAREA_MODE | 179943 | 0.4148404 |
| numeric | YEARS_BEGINEXPLUATATION_MODE | 150007 | 0.5121898 |
| numeric | YEARS_BUILD_MODE | 204488 | 0.3350222 |

| skim_type | skim_variable | n_missing | complete_rate |
|---|---|---|---|
| character | NAME_CONTRACT_TYPE | 0 | 1 |
| character | CODE_GENDER | 0 | 1 |
| character | FLAG_OWN_CAR | 0 | 1 |
| character | FLAG_OWN_REALTY | 0 | 1 |
| character | NAME_TYPE_SUITE | 0 | 1 |
| character | NAME_INCOME_TYPE | 0 | 1 |
| character | NAME_EDUCATION_TYPE | 0 | 1 |
| character | NAME_FAMILY_STATUS | 0 | 1 |
| character | NAME_HOUSING_TYPE | 0 | 1 |
| character | OCCUPATION_TYPE | 0 | 1 |
| character | WEEKDAY_APPR_PROCESS_START | 0 | 1 |
| character | ORGANIZATION_TYPE | 0 | 1 |
| numeric | SK_ID_CURR | 0 | 1 |
| numeric | TARGET | 0 | 1 |
| numeric | CNT_CHILDREN | 0 | 1 |
| numeric | AMT_INCOME_TOTAL | 0 | 1 |
| numeric | AMT_CREDIT | 0 | 1 |
| numeric | AMT_ANNUITY | 0 | 1 |
| numeric | AMT_GOODS_PRICE | 0 | 1 |

We cleaned up the data before moving on to EDA and predictive modeling. In the process of data cleaning, imputed missing values, incorrect, poorly formatted, or otherwise confusing data are sorted and corrected. Now our dataset is ready to model.

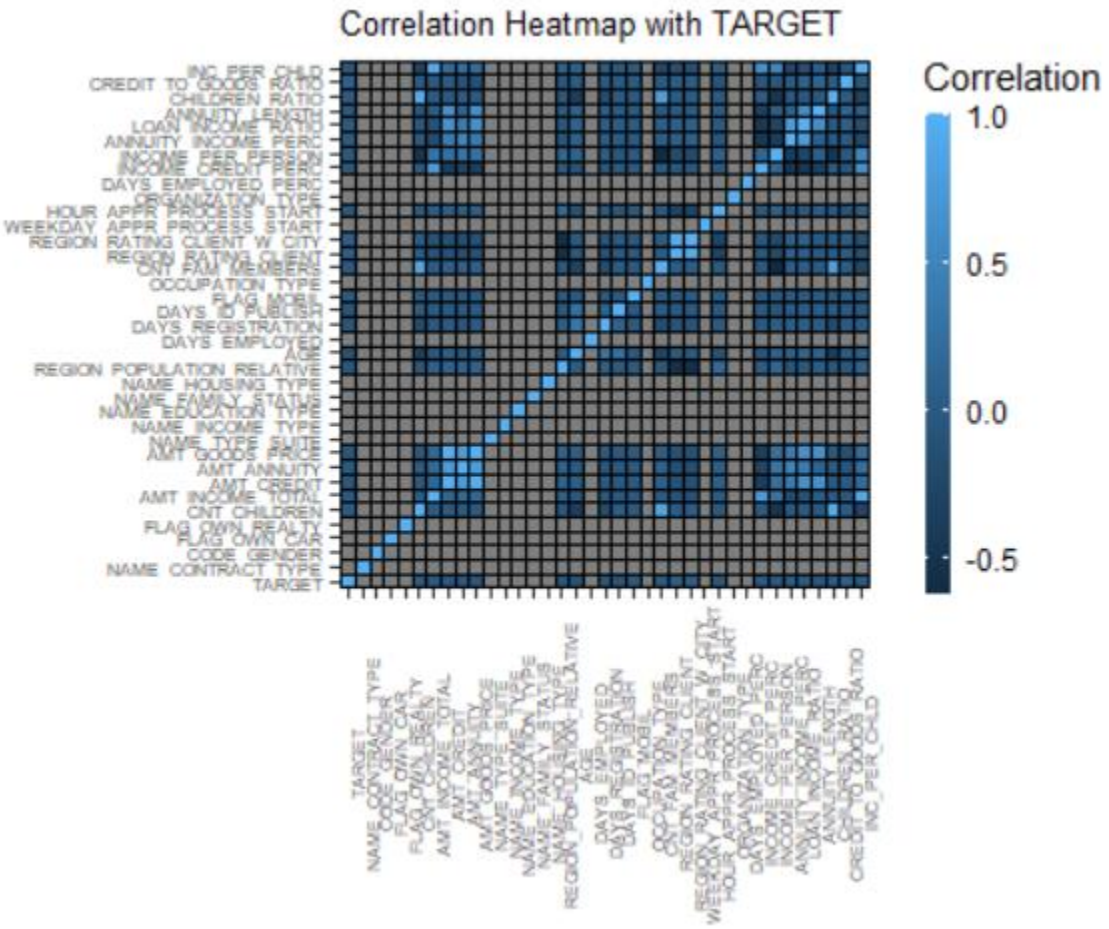
## 2.4 Explanatory Data Analysis (EDA)

### 2.4.1 Target variable



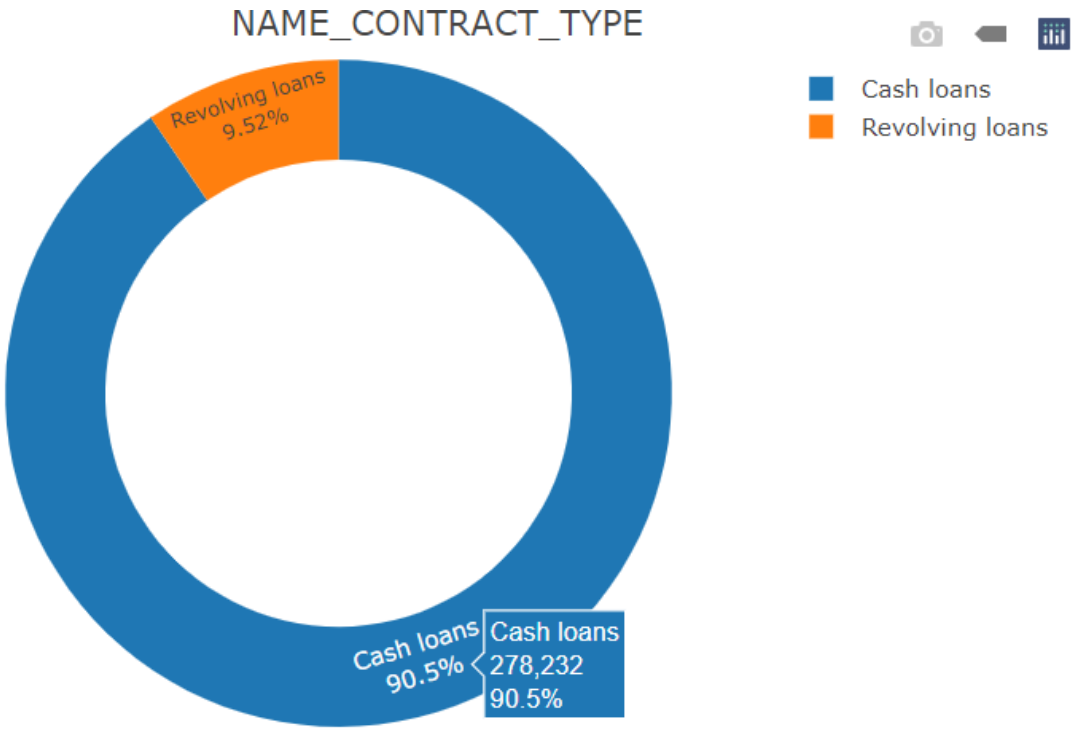

In this dataset, 92% of applicants repay the loan on time, and 8% of applicants don't. We see that this is an imbalanced data problem. There are far more loans repaid on time than outstanding loans. Once we get into more complex machine learning models, we can threshold adjust and XGBoosting through their representation in the data to reflect this imbalance.

### 2.4.2 Heatmap chart



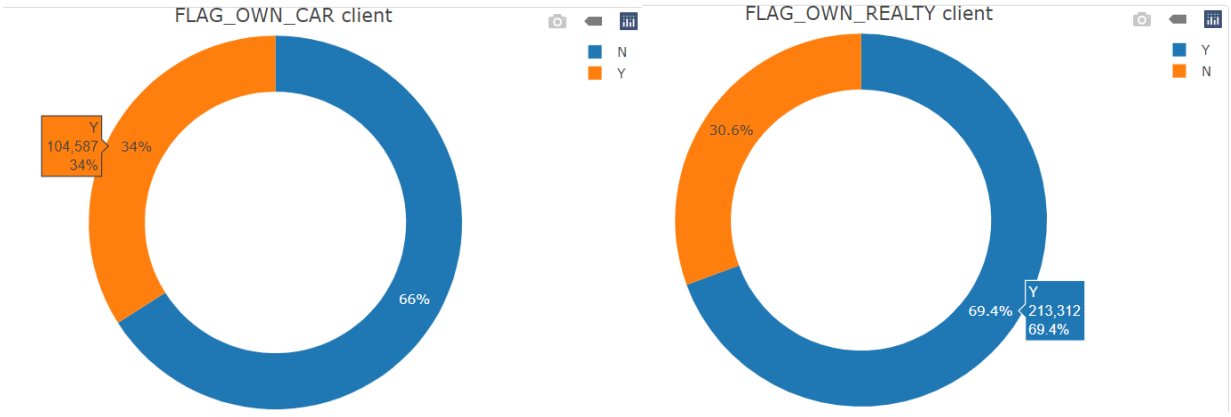

This heatmap shows the relationship between a lender's applicant's ability to repay their loan on time and their personal information. Darker colors indicate a high correlation between them. There is a high correlation among AMT_Goods_Price, AMT_Income_Total, AMT_credit. This heatmap helped us remove some highly correlated variables in the previous data cleaning step, and we removed variables with a correlation above 0.8 to ensure overfitting occurred.

### 2.4.3 Distribution for types of loan



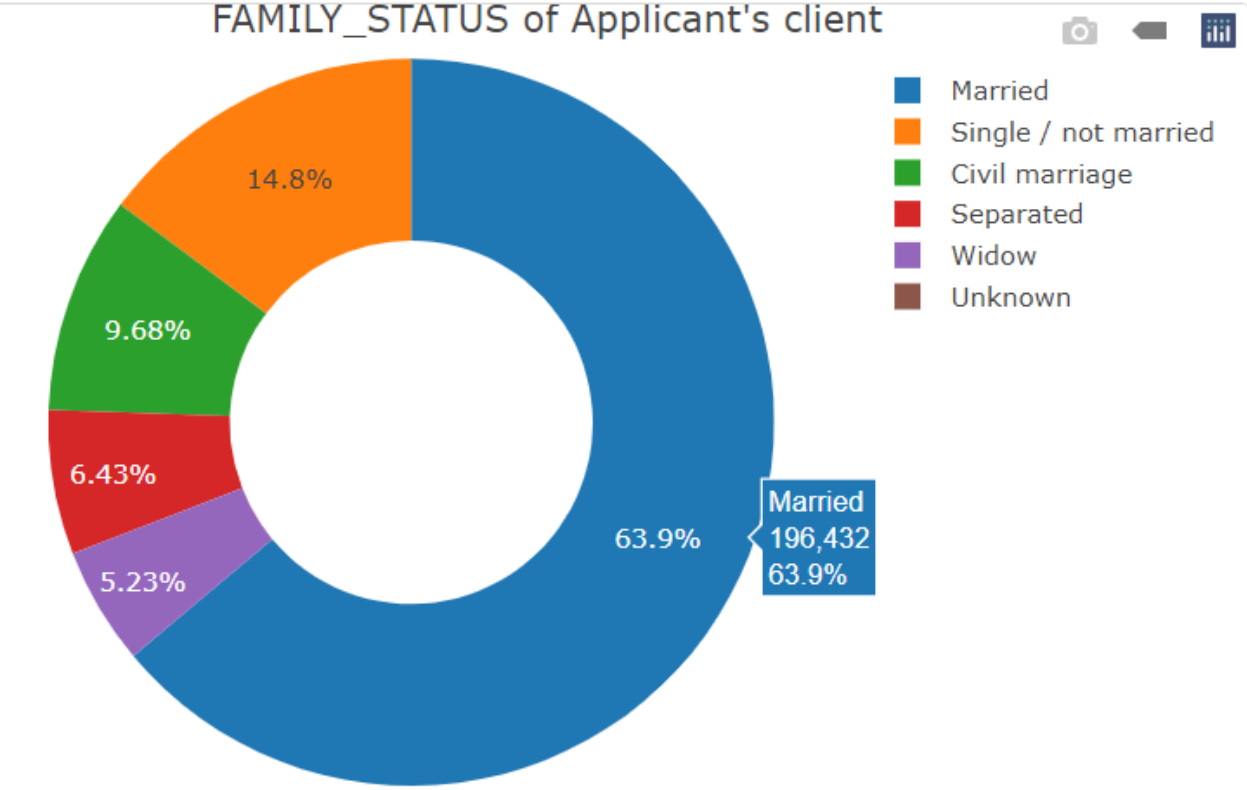

Most loans are cash loans obtained by applicants. 90.6% of loans are cash loans.

### 2.4.4 Distribution for purpose of loan



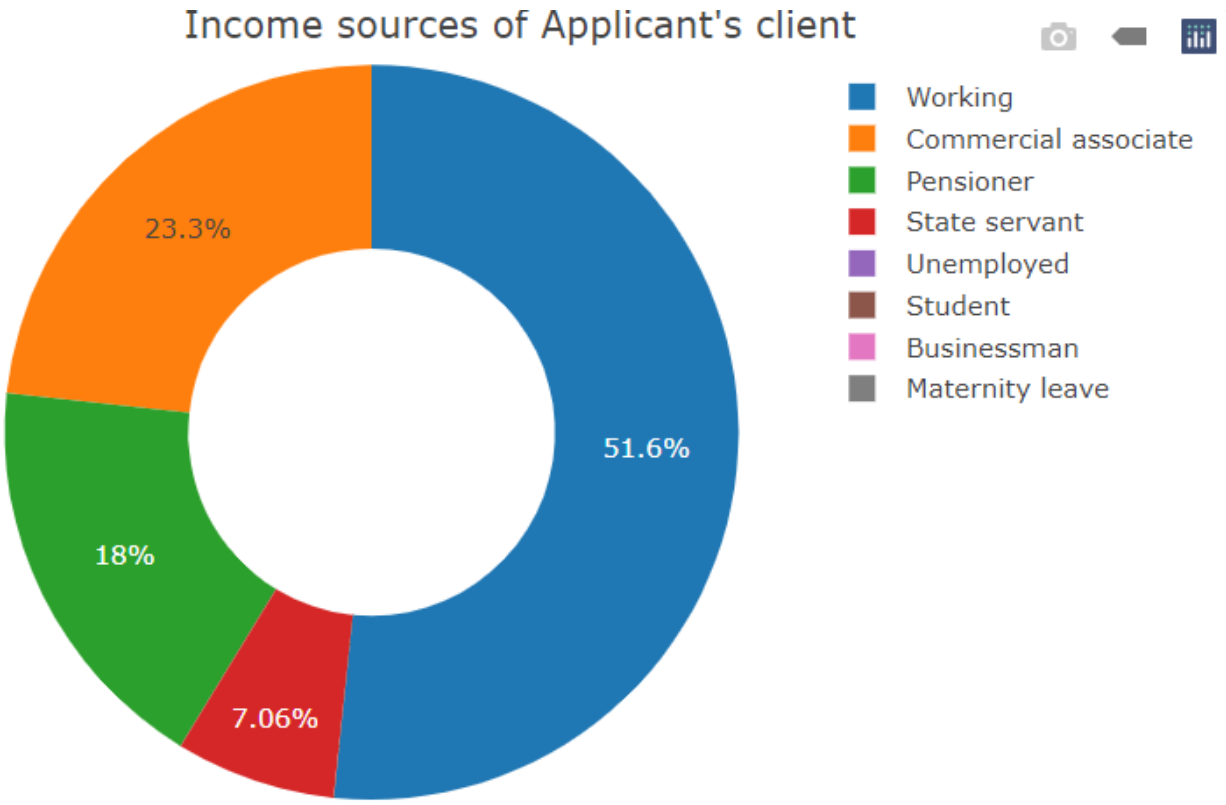

Most of the loans are mainly for the applicant's real estate.

### 2.4.5 Family Status of Applicant's who applied for loan



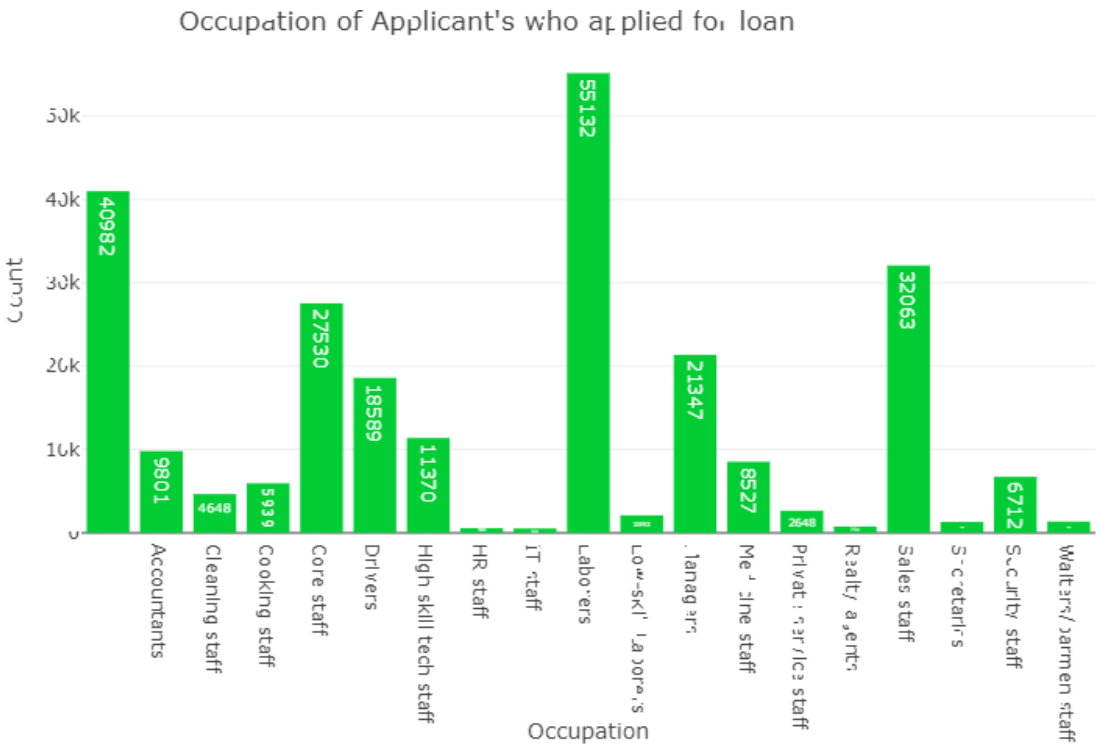

In this dataset, 63.9% of the applicants were married. 14.8% Single, etc.

2.4.6 Income sources of Applicants' who applied for loan




In this dataset, 51.6% of Applicants mentioned that they are working. 23.3% are Commercial Associate and 18 % are Pensioner etc.

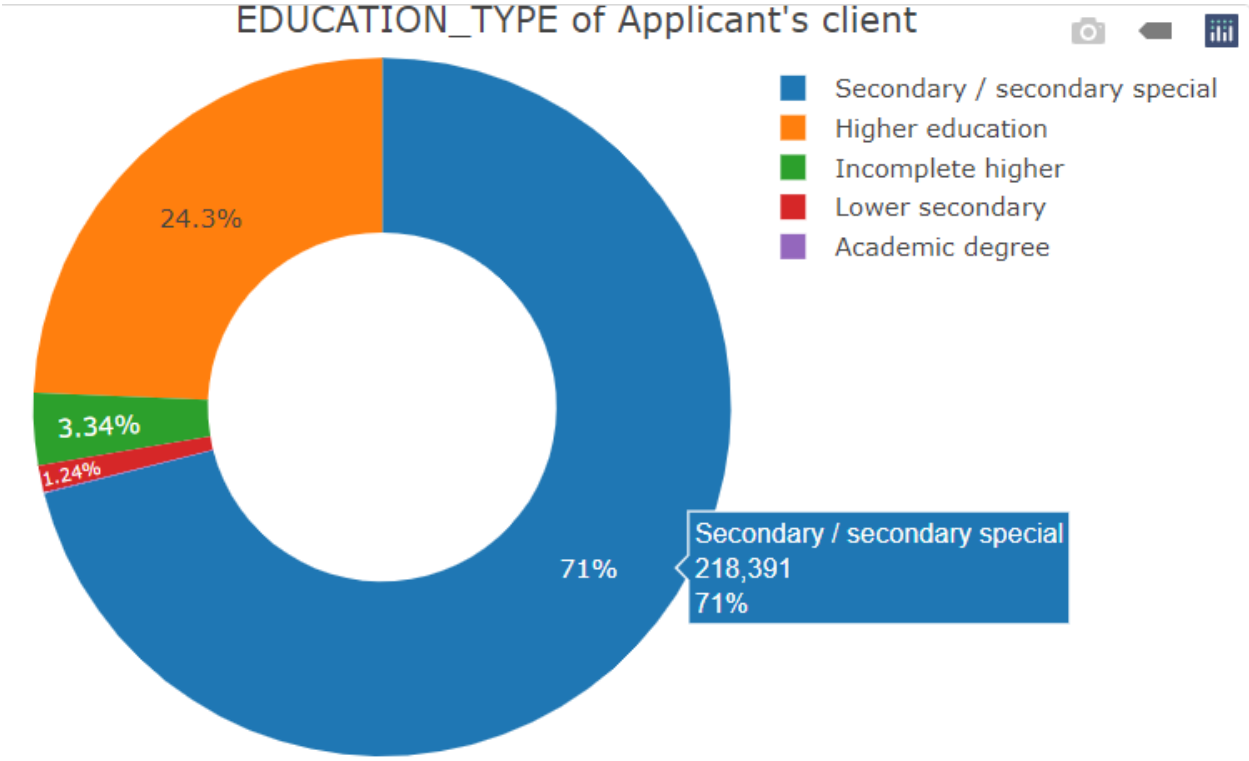
### 2.4.7 Occupation of Applicant's who applied for loan




In this bar chart, we can see that the top five applicant occupation who applied for a loan are:

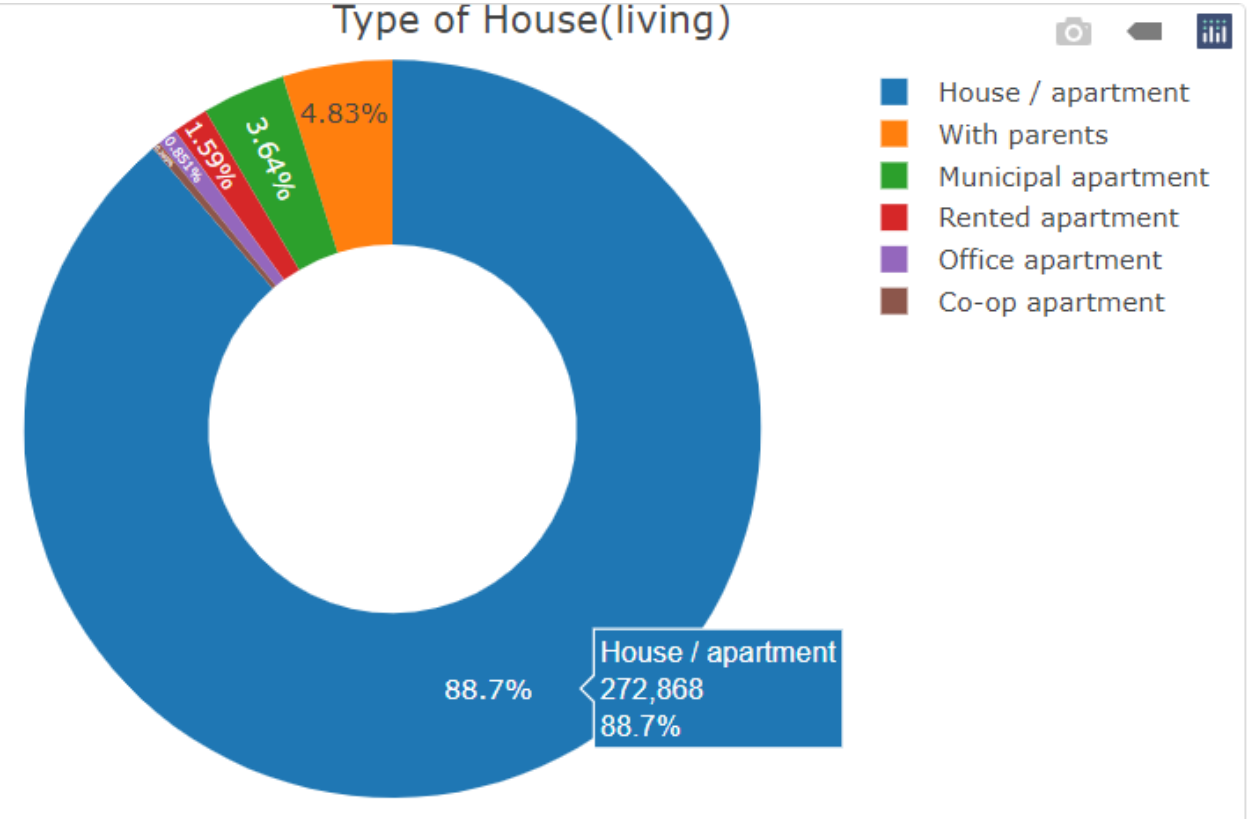
- Laborers with approximately 55,000 applications
- Sales Staff with approximately 32,000 applications
- Core staff with approximately 28,000 applications
- Managers with approximately 21,000 applications
- Drivers with approximately 19,000 applications

### 2.4.8 Education of Applicant's who applied for loan



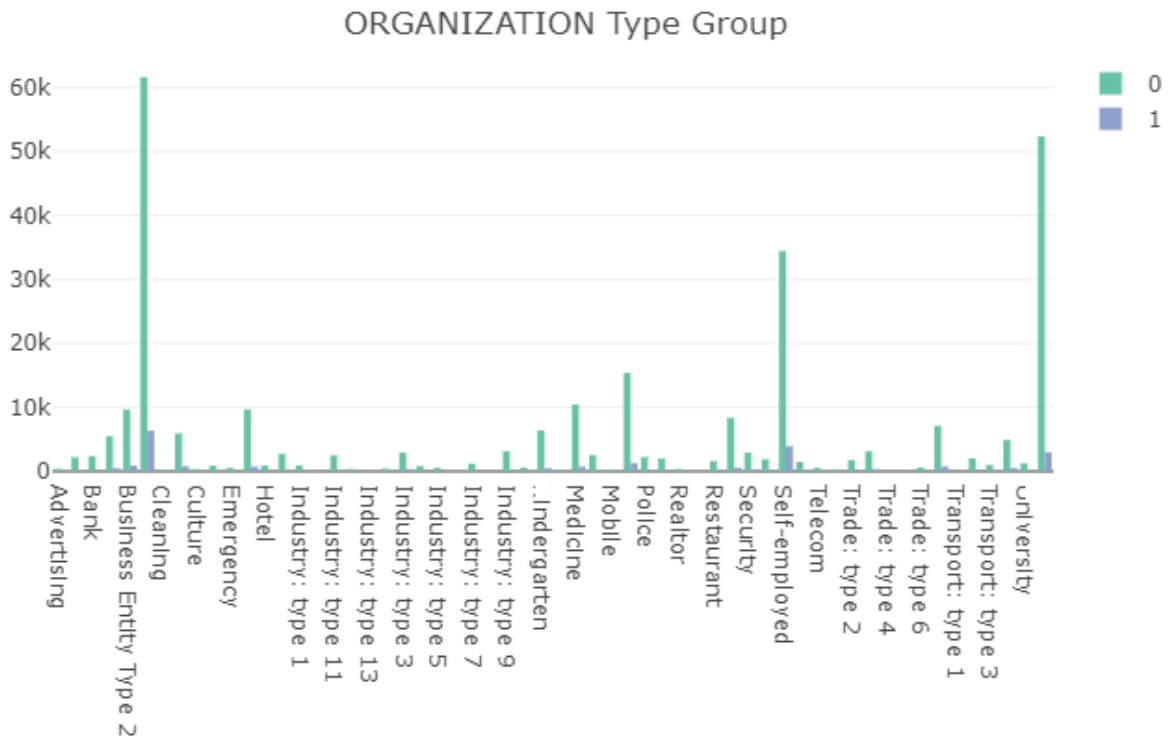

In this dataset, 71 % of applicants have secondary and 24.3 % having higher education.

2.4.9 which types of houses higher applicants applied for loan?




In this dataset, approximately 89% of the people who applied for a loan indicated that they live in a house or apartment.

### 2.4.10 Types of Organizations who applied for loan

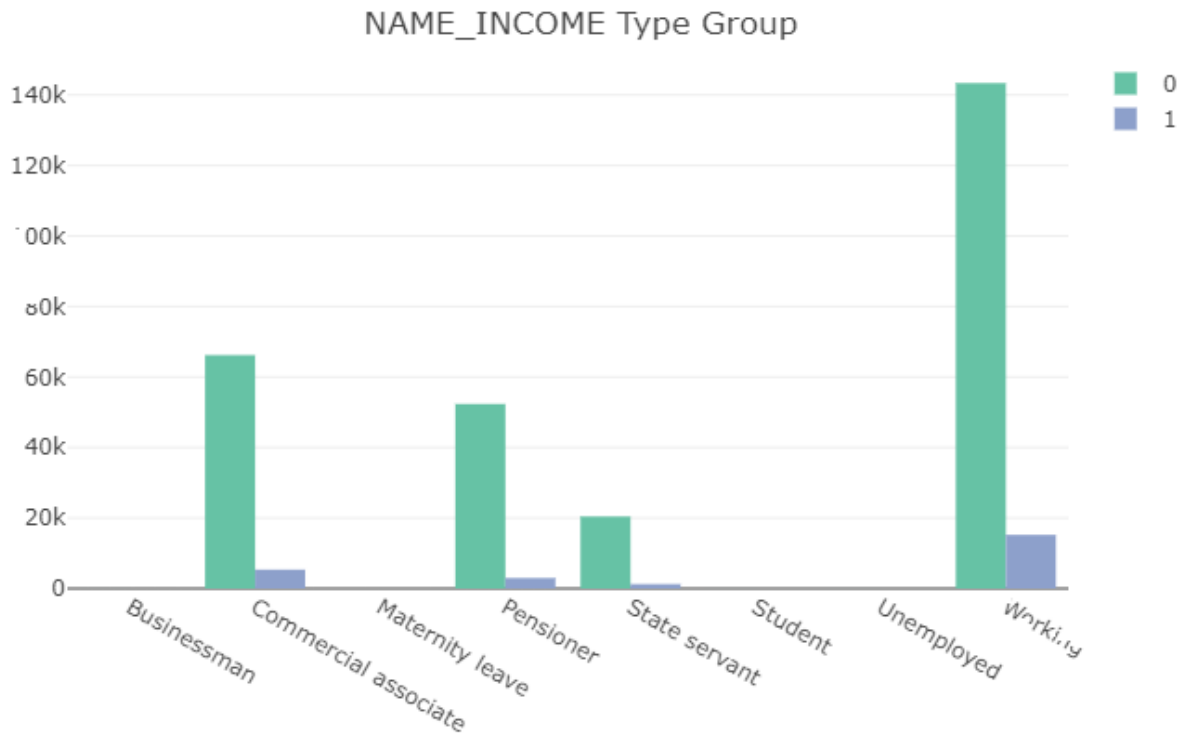



In this dataset, we can see that the types of organizations that applied for a loan are:

- Business Entity Type 3 with approximately 68,000 applications
- XNA with approximately 55,000 applications
- Self-employed with approximately 38,000 applications
- Others with approximately 17,000 applications
- Medicine with approximately 11,000 applications.

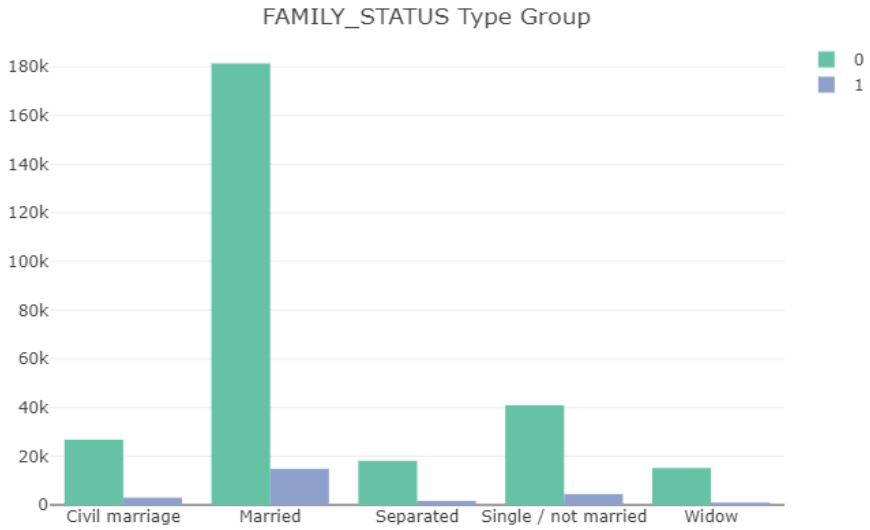
Next, we continue to explore whether the loan is repaid or not and other variables.

### 2.4.11 Effect of the Income sources on repayment



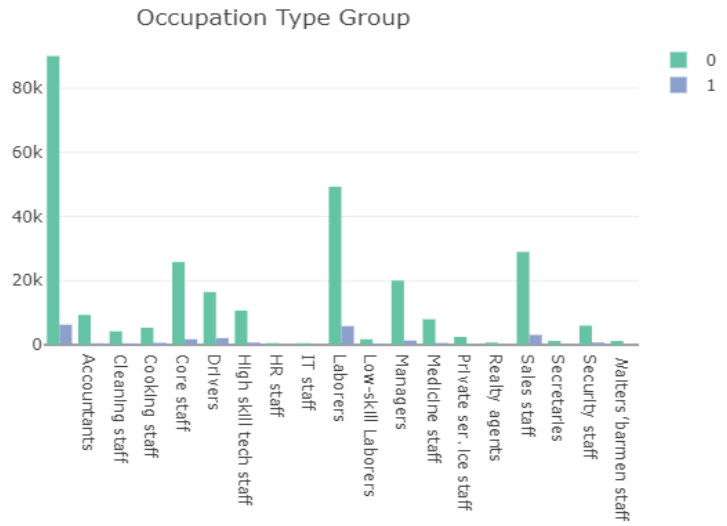

In this bar graph, we can see that the income source group from state civil servants has a higher share of repayments.

### 2.4.12 Effect of Family Status on repayment

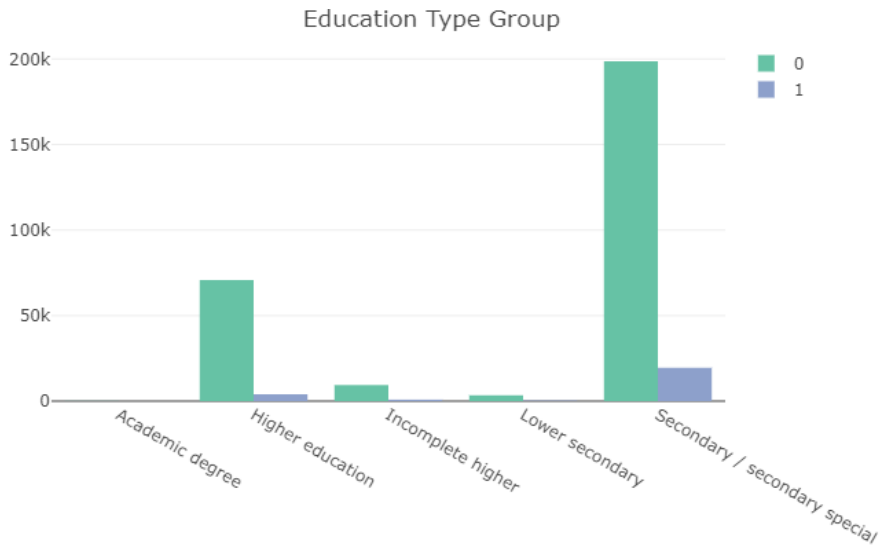


In this bar graph, we can see that the family status group from married has a higher share of repayments.

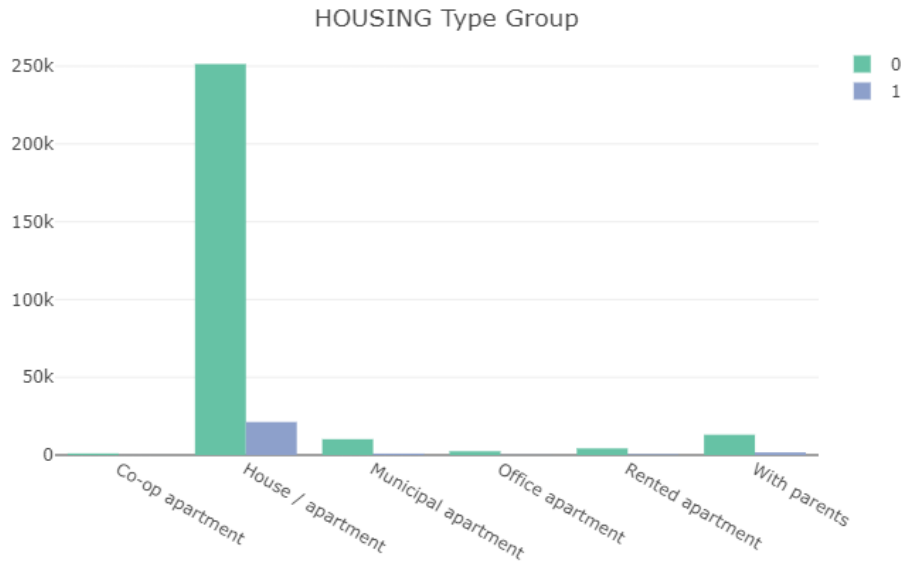
### 2.4.13 Effect of the Occupation on repayment



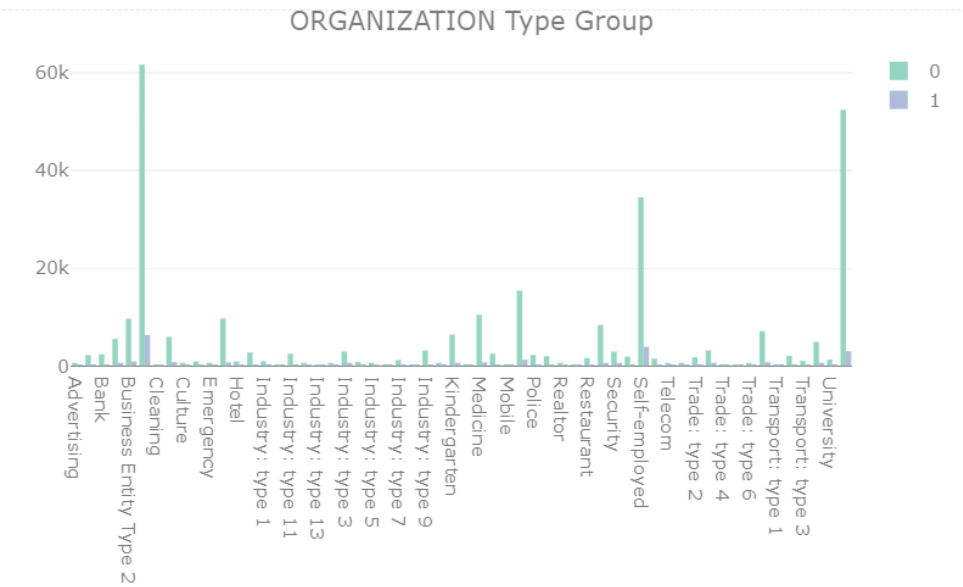

### 2.4.14 Effect of the Education on repayment



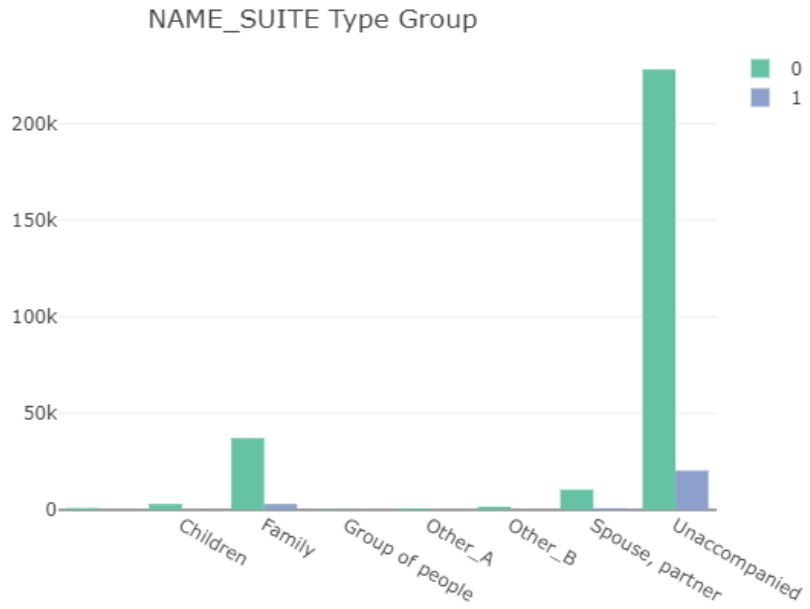
### 2.4.15 Effect of the housing types on repayment




### 2.4.16 Effect of the Organization types on repayment

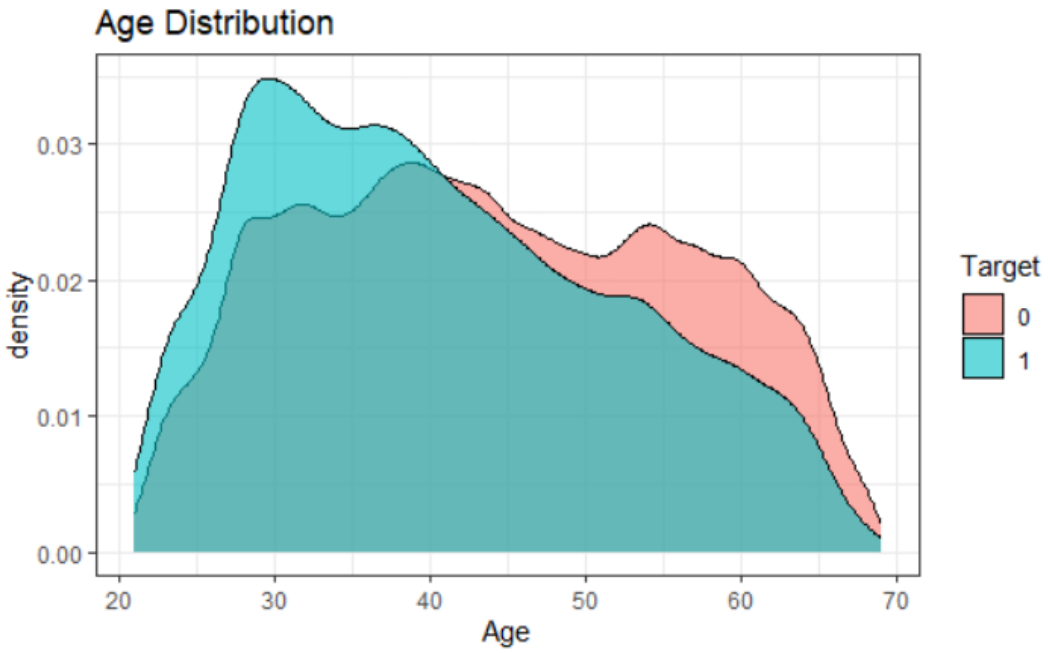


### 2.4.17 Distribution of Name of type of the Suite in terms of loan is repaid or not



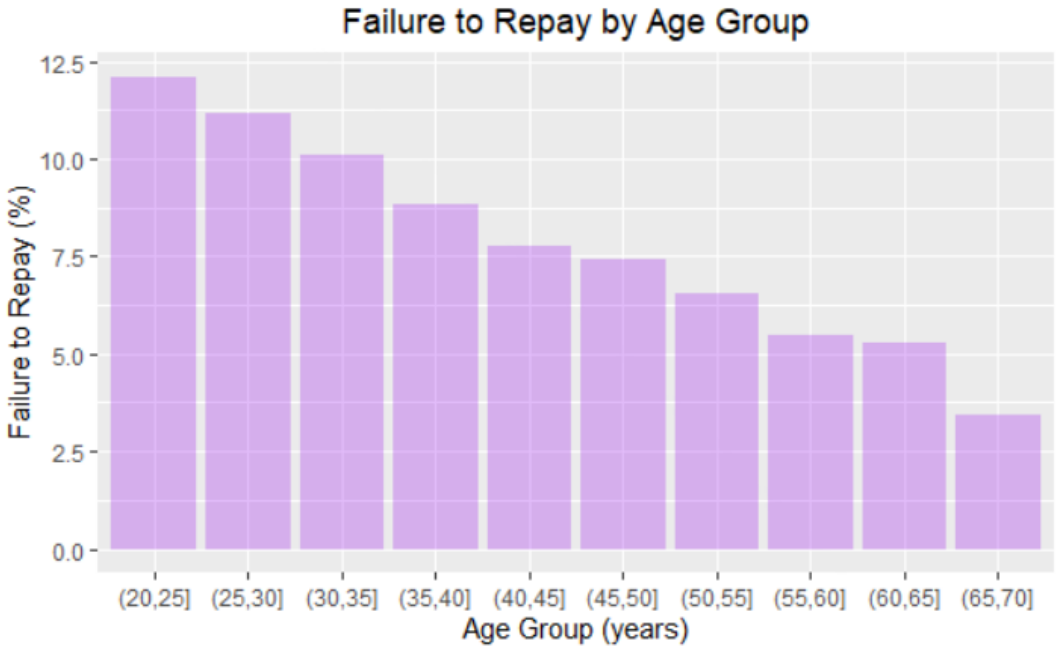

### 2.4.18 Effect of the age on repayment

**2.4.18.1 kernel density estimation plot (KDE) colored by the value of the target**



The result of our analysis suggests that there may be a correlation between age and loan default, with the curve of the target variable skewed towards the younger end of the range.

#### 2.4.18.2 Average failure to repay loans by age bracket




However, the above graph curve does not look obvious. In order to create this graph, we first divided the age categories into bins of 5 years each. Then, for each bin, we calculated the average value of the target variable, which represents the ratio of loans that were not repaid in each age category.

The results of our analysis reveal that the rate of failure to repay loans is above 10% for the youngest three age groups, and below 5% for the oldest age group. In addition, it is important to note that this does not mean that all young borrowers are more likely to default on their loans, nor does it suggest that all older borrowers are less likely to default. Other factors, such as income stability, debt-to-income ratio, and credit history, may also play important roles in determining loan default risk.

Therefore, it is important for lenders to consider a range of factors beyond age when assessing loan applications and determining lending policies.

## 3. Main Analysis

In the main analysis, we will mainly discuss a logistic regression model and Stochastic gradient boosting used to predict the probability of loan default based on various client attributes. We will explain the steps involved in building these models and analyzing the results.

### 3.1 Feature Engineering

In this project, we attempted to use Principal Component Analysis (PCA) as a feature engineering approach to decrease the dimensionality of the Home Credit Default Risk dataset. This was done in order to better understand the relationship between the variables. On the other hand, we did have some technical difficulties, which brought in R session crashes. As a consequence of this, we came to the conclusion that an alternate approach to the problem of overfitting would be to make use of a heatmap consisting of variables that are not significantly associated with one another. Our goal was to enhance the performance of our machine learning models by lowering the dimensionality of the dataset and narrowing our focus to include those variables that were of importance. With the aid of this strategy, we were able to find a subset of variables that are the most predictive of loan defaults while simultaneously reducing the danger of overfitting. The incorporation of heatmap variables not only assisted in the resolution of problems associated with overfitting but also enhanced the computational efficacy of our models. As a result, using this method was a workable alternative to use PCA within the scope of this investigation.

### 3.2 Heatmap variables

```r
variables = c('CNT_CHILDREN','AMT_INCOME_TOTAL','AMT_CREDIT','AMT_ANNUITY','AMT_GOODS_PRICE','REGION_POPULATION_RELATIVE','AGE','DAYS_REGISTRATION','DAYS_ID_PUBLISH' , 'FLAG_MOBIL', 'CNT_FAM_MEMBERS' ,  'REGION_RATING_CLIENT', 'HOUR_APPR_PROCESS_START', 'REGION_RATING_CLIENT_W_CITY', 'INCOME_CREDIT_PERC' ,'INCOME_PER_PERSON' ,'ANNUITY_INCOME_PERC','LOAN_INCOME_RATIO','ANNUITY_LENGTH','CHILDREN_RATIO', 'CREDIT_TO_GOODS_RATIO', 'INC_PER_CHLD')
```

### 3.3 Logistic Regression

Logistic regression is a statistical method used to model the probability of a binary outcome based on one or more predictor variables. In this case, we will be using logistic regression to model the probability of loan default based on various client attributes. The concept's objective is to support financial institutions in making informed decisions on the granting of loans to prospective customers and minimizing the risk of financial losses.

#### 3.3.1 Data Collection

For the model, we used customer demographic information such as age, gender, and marital status, financial information such as yearly income, and credit-related data such as credit ratings. To prevent overfitting, we utilized a heatmap to determine which variables had a strong correlation and omitted them from the model.

#### 3.3.2 Modeling

The data was split into a training set and a testing set. In this case, there are too many rows in the majority class (on-time loan repayment) as opposed to insufficient rows in the minority class (not paying the loan on time). This indicates that the dataset is very imbalanced, with the minority class

comprising just 8.1% of the majority class. To address with this imbalanced dataset, we may try applying a threshold below 0.5, which we set to 0.3, which increases the sample's likelihood of being classified as a minority class. Adjustments were made to the threshold for managing unbalanced data. The model of logistic regression was trained using the training set and evaluated with the testing set.

```
set.seed(100)

# Split the data into training and testing sets
train_idx <- createDataPartition(data$TARGET, p = 0.7, list = FALSE)
train <- data[train_idx, ]
test <- data[-train_idx, ]

# Fit the logistic regression model on the training data
logit_model <- glm(TARGET~ CNT_CHILDREN+AMT_INCOME_TOTAL+AMT_CREDIT+AMT_ANNUITY+AMT_GOODS_PRICE+REGION_POPULATION_RELATIVE+AGE+DAYS_R
EGISTRATION+DAYS_ID_PUBLISH + FLAG_MOBIL+ CNT_FAM_MEMBERS +  REGION_RATING_CLIENT+ HOUR_APPR_PROCESS_START+
REGION_RATING_CLIENT_W_CITY+ INCOME_CREDIT_PERC +INCOME_PER_PERSON
+ANNUITY_INCOME_PERC+LOAN_INCOME_RATIO+ANNUITY_LENGTH+CHILDREN_RATIO+ CREDIT_TO_GOODS_RATIO+ INC_PER_CHLD, family = binomial(link =
'logit'), data = train)

# Make predictions on the test data
prob <- predict(logit_model, newdata = test, type = 'response')
threshold <- 0.3 # Adjust the threshold for handling imbalanced data
pred <- ifelse(prob > threshold, 1, 0)

summary(logit_model)
acc <- mean(pred == test$TARGET)
auc <- roc(test$TARGET, prob)$auc

conf_matrix <- confusionMatrix(table(pred, test$TARGET))

# Print confusion matrix
conf_matrix

# Print F1 and recall
f1 <- 2 * conf_matrix$byClass['Pos Pred Value'] * conf_matrix$byClass['Sensitivity'] /
  (conf_matrix$byClass['Pos Pred Value'] + conf_matrix$byClass['Sensitivity'])
recall <- conf_matrix$byClass['Sensitivity']
cat(sprintf("F1 score: %0.3f, Recall: %0.3f\n", f1, recall))

cat("Accuracy: ", acc, "\n")
cat("AUC: ", auc, "\n")
```

### 3.3.3 Results

Predicting the risk that a customer would default on a loan was facilitated by the logistic regression model. For instance, the coefficient analysis found that income and age are one of the most important predictors of loan default which P value is less than 0.05. With each unit rise in a client's income, the log-odds of timely loan repayment increase by a coefficient value of 9.574e-07. This shows that customers with a greater income are more likely to repay their loans on time. With each one-unit rise in a client's age, the log-odds of late loan repayment fall by 0.021. This shows that older consumers

are more likely to make timely loan payments.

```
Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -1.640e+01  4.396e+01  -0.373  0.70913
CNT_CHILDREN                -1.871e-01  5.303e-02  -3.527  0.00042 ***
AMT_INCOME_TOTAL             9.575e-07  1.618e-07   5.917 3.28e-09 ***
AMT_CREDIT                   6.642e-07  2.350e-07   2.826  0.00472 **
AMT_ANNUITY                 -1.737e-05  2.618e-06  -6.632 3.31e-11 ***
AMT_GOODS_PRICE             -1.149e-06  2.491e-07  -4.613 3.97e-06 ***
REGION_POPULATION_RELATIVE  -1.626e+00  7.294e-01  -2.229  0.02583 *
AGE                         -2.052e-02  7.863e-04 -26.098  < 2e-16 ***
DAYS_REGISTRATION           -6.475e-03  9.336e-04  -6.935 4.06e-12 ***
DAYS_ID_PUBLISH             -2.456e-02  1.996e-03 -12.306  < 2e-16 ***
FLAG_MOBIL                   6.799e+00  4.395e+01   0.155  0.87708
CNT_FAM_MEMBERS              2.339e-01  3.914e-02   5.975 2.30e-09 ***
REGION_RATING_CLIENT        -6.401e-03  5.401e-02  -0.119  0.90566
HOUR_APPR_PROCESS_START     -1.337e-02  2.548e-03  -5.248 1.54e-07 ***
REGION_RATING_CLIENT_W_CITY  3.478e-01  5.440e-02   6.393 1.62e-10 ***
INCOME_CREDIT_PERC          -4.372e-01  5.456e-02  -8.012 1.12e-15 ***
INCOME_PER_PERSON            5.231e-01  5.281e-02   9.906  < 2e-16 ***
ANNUITY_INCOME_PERC          2.357e+00  3.728e-01   6.323 2.56e-10 ***
LOAN_INCOME_RATIO            2.703e-02  1.152e-02   2.346  0.01896 *
ANNUITY_LENGTH              -1.917e-02  3.625e-03  -5.290 1.22e-07 ***
CHILDREN_RATIO               4.232e-01  1.437e-01   2.946  0.00322 **
CREDIT_TO_GOODS_RATIO        1.335e+00  1.084e-01  12.314  < 2e-16 ***
INC_PER_CHLD                -3.553e-07  1.981e-07  -1.793  0.07290 .
```

The performance of the model may be assessed using the F1 score, Recall, Accuracy, and AUC. The F1 score of 0.958 and the Recall value of 1.000 suggest that the model is effective at detecting positive instances, with a strong balance between precision and recall, and that the model correctly recognized all positive cases in the dataset. The Accuracy reflects the overall predictive precision of the model for both positive and negative scenarios. It indicates that 92% of instances were properly categorized by the model. Nonetheless, the AUC of 0.6 indicates that the model's ability to differentiate between positive and negative situations is better than random guessing, but inadequate and might be termed overfitting to positive cases.

```
               Accuracy : 0.9196
                 95% CI : (0.9179, 0.9214)
    No Information Rate : 0.9197
    P-Value [Acc > NIR] : 0.5561

                  Kappa : 9e-04

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

            Sensitivity : 0.9998114
            Specificity : 0.0006753
         Pos Pred Value : 0.9197784
         Neg Pred Value : 0.2380952
             Prevalence : 0.9197424
         Detection Rate : 0.9195690
   Detection Prevalence : 0.9997724
      Balanced Accuracy : 0.5002434

       'Positive' Class : 0

F1 score: 0.958, Recall: 1.000
Accuracy:  0.9196232
AUC:  0.6440295
```

### 3.4 Stochastic Gradient Boosting (SGB)

The Stochastic Gradient Boosting (SGB) model is a machine learning algorithm that uses an ensemble of decision trees to make predictions. This model is commonly used in prediction tasks where there are non-linear relationships between the predictors and the outcome. Similar to Logistic Regression, the SGB model aims to analyze various factors that contribute to the likelihood of loan default. However, one of the advantages of using the SGB model is it can handle the highly imbalanced data by assigning different weights to the different classes. Handling imbalance data is essential because it can lead to bias in the model and inaccurate predictions.

#### 3.4.1 Setting parameters

The first step in using the SGB model is to define the model's parameters. This involves selecting the number of trees, the learning rate, and the depth of each tree. The choice of these parameters can impact the model's accuracy and performance.

### 3.4.2 Data Collection

The data used in the SGB model includes clients' demographic, financial, and credit-related information, such as age, gender, family status, annually income, credit scores, and loan amount. Similar to Logistic Regression, we used heatmap to check which variables have high correlation and avoid using them to reduce overfitting.

### 3.4.3 Modeling

After collecting the data, we split it into training and testing sets. Then, we adjusted the threshold for handling imbalanced data by assigning different weights to the different classes. This helped to prevent bias in the model and improve its accuracy in predicting loan defaults.

```r
#Stochastic Gradient Boosting
```{r}
set.seed(100)

# Split the data into training and testing sets
train_index <- createDataPartition(data$TARGET, p = 0.7, list = FALSE)
train <- data[train_index, ]
test <- data[-train_index, ]

# Check the number of rows and columns in training and testing sets
dim(train)
dim(test)

# Check the summary of the target variable in training set
summary(train$TARGET)

# Adjust the threshold for classification to handle the imbalanced data
train$TARGET_adjusted <- ifelse(train$TARGET == 1, 0.15, 0.85)

# Check the summary of the adjusted target variable in training set
summary(train$TARGET_adjusted)

blm_model <- gbm(formula = TARGET_adjusted ~ ., data = train[, c(variables, "TARGET_adjusted")], distribution = "bernoulli", n.trees
= 200, interaction.depth = 3, shrinkage = 0.1, bag.fraction = 0.5, train.fraction = 1, n.minobsinnode = 10)

# Predict the probability of default for the test set
pred <- predict(blm_model, newdata = test[, variables], n.trees = 200, type = "response")

# Convert the predicted probabilities to binary class labels based on the adjusted threshold
pred_class <- ifelse(pred > 0.15, 0, 1)

# Compute the accuracy and AUC of the model on the test set

conf_mat <- table(pred_class, test$TARGET)
accuracy <- sum(diag(conf_mat))/sum(conf_mat)
auc <- pROC::auc(test$TARGET, pred)

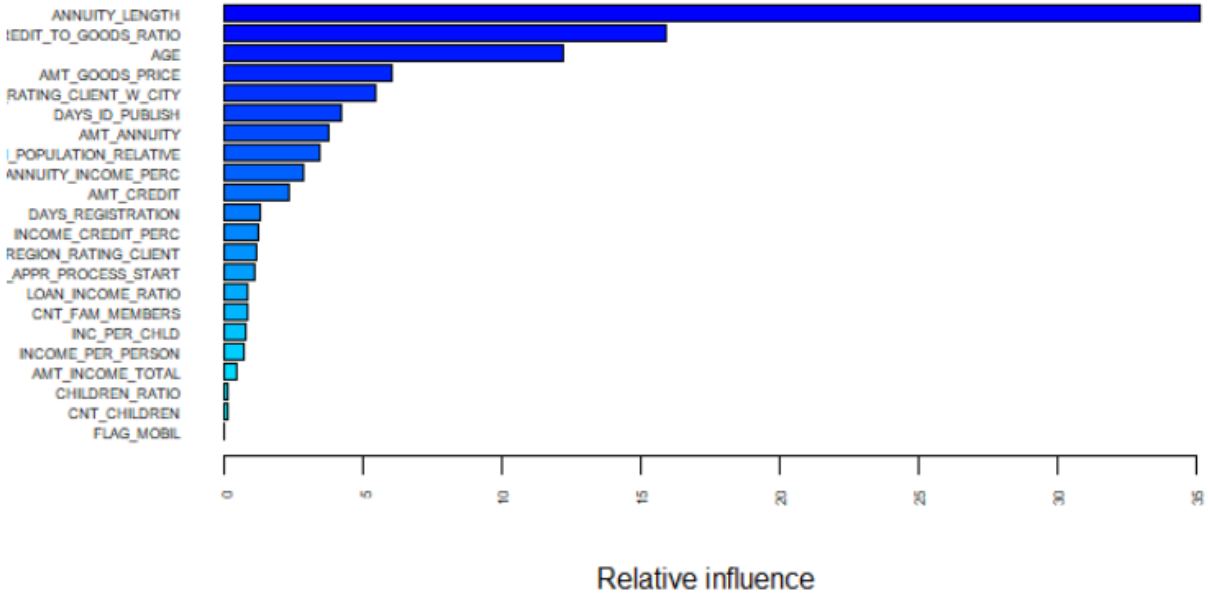
# Print the results
print(paste("Accuracy:", round(accuracy, 3)))
print(paste("AUC:", round(auc, 3)))

roc_obj <- roc(test$TARGET, pred)
plot(roc_obj, main = "ROC Curve", print.thres = c(0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9))
```
```

### 3.4.4 Result

The performance of the SGB model was evaluated based on its precision, AUC, and significance of features. The model's accuracy was 92%, which indicates that it accurately classified 92% of all cases. The AUC of 0.7 was deemed satisfactory and superior to logistic regression. In addition, the feature significance chart revealed that the five most influential related features are -

- Annuity length: the length of the client's annuity payment
- Credit to goods ratio: Ratio of the amount of credit customers apply for to the price of the goods they purchase with the credit.
- Age: Client's age number (Year)
- Amount good Price: For consumer loans it is the price of the goods for which the loan is given.
- Rating client with city: Home credit's system rating of the region where client lives with taking city into account.




In conclusion, the SGB model may be an excellent tool for financial institutions to anticipate the chance of loan default. The program can handle unbalanced data and find the most significant elements leading

to loan default. It is essential to highlight, however, that the accuracy of the model depends on the selection of model parameters, the quality of the data, and the existence of outliers. Consequently, it is necessary to examine the model's performance carefully and make improvements appropriately to guarantee accurate forecasts.

## 4. Findings

### 4.1 Business problems

The dataset titled "Home Credit Default Risk" has the potential to be used in such a way as to solve a number of important difficulties facing businesses today. One of the possible applications is increasing the proportion of loan applications that are accepted by prioritizing those from applicants with the highest creditworthiness. This may be accomplished by developing prediction models that strike a compromise between the competing imperatives of approving the maximum number of loans feasible and the possibility of default. Moreover, the information may be used to enhance customer segmentation by establishing distinct client groups based on the characteristics and behaviors of the customers, which can be done by analyzing the clients' characteristics and behaviors.

Improving risk management by determining the elements that are most closely connected with default risk is another possible use of this technology. This information may be helpful to Home Credit in developing targeted risk management measures to reduce such risks. One example of this would be altering lending rules or loan conditions depending on an applicant's risk profile. Last but not least, the dataset may be put to use in order to increase loan recovery by means of the development of more efficient recovery tactics that are individualized to the risk profile of each individual client. For instance,

proactive contact may be made with high-risk consumers, and individualized repayment plans can be devised, all with the goal of increasing the probability of successful loan payback.

### 4.2 Ethical implication

The Home Credit Default Risk dataset includes confidential information on customers, such as details about their demographics, finances, and credit histories. Even though the dataset is available to the public and Home Credit does not reveal the identities of its customers, there is still a possibility that the information could be misused for purposes that were not intended. For example, vulnerable individuals could be targeted for predatory lending practices, or decisions could be made regarding an individual's creditworthiness without their knowledge or consent. It is therefore the responsibility of data analytics professionals and executives at Home Credit to ensure that the data is used in an ethical manner and for the purpose it was intended for, and that individuals whose data are included in the dataset have provided informed consent regarding how their data will be used.

It is also essential, in order to guarantee justice and equality, that the algorithms used in the data analysis process be made completely transparent. This involves being honest about the sources of the data and the techniques used to evaluate the data, as well as being responsible for any choices made based on the analysis. Moreover, this includes being liable for any decisions made based on the analysis. In addition, it is essential to identify and eliminate any possible sources of bias that may exist within the data collection, whether such biases originate from the selection of characteristics or from the sample of persons. To guarantee that the model is both fair and equitable, it is particularly crucial to do this step if the model is going to be used to make choices about the individual's financial future.

As experts in the field of data analytics, it is essential for us to take into consideration the possible ethical ramifications of our work and to conduct ourselves in a manner that is consistent with

the Christian principles of dignity, respect, privacy, and justice. The verses in Proverbs 31:8-9 “Speak up for those who cannot speak for themselves, for the rights of all who are destitute. Speak up and judge fairly; defend the rights of the poor and needy.”. By adhering to these principles over the course of our work with the Home Credit Default Risk dataset, we may contribute to ensuring that our analysis is carried out in an ethical manner and that it is used for the benefit of society as a whole.

### 4.3 Limitations

The dataset contains a huge number of characteristics, and it is feasible that some of them may not be helpful for predicting loan defaults. It is also possible that some of the features will be helpful. In addition, the dataset is quite large and related to a wide variety of outside sources; hence, it is necessary to combine several datasets by using left join procedures. Unfortunately, the introduction of other data sources resulted in a significant increase in computing demand, which caused numerous R sessions to terminate unexpectedly. As a result, in order to finish the project, we took out any variables that came from outside sources. While we were able to finish the project more quickly as a result of this, the performance rates of our models may be considered satisfactory; yet they were not excellent. As a result, in order to get a better grasp of the benchmark scores, we studied the performance scores of other competitors on the competition website. Given the limited amount of time and resources that we had at our disposal, we were surprised to see that the performance ratings of other participants were not considerably different from our own. Although we admit that more advanced feature engineering approaches may have enhanced the performance of our models, we feel that the results we have achieved are adequate in light of the restrictions that the project imposed. The feature engineering process might benefit from more study and development in the future, which could lead to improved model performance. In the future, it would be helpful to re-run the study utilizing the whole dataset in order to obtain findings that are more accurate.