# Predicting Home Credit Client's Payment Abilities

BHARA YUDHIANTARA

28 SEPTEMBER 2018

# Outline

1. Understanding the Problem
2. Data Checking and Formatting
3. Exploratory Data Analysis
4. Baseline Model
5. Improved Model
6. Summary

# Understanding the Problem

- Many people struggle to get loans due to insufficient or non-existent credit histories. This population is often taken advantage of by untrustworthy lenders.
- Predicting client’s repayment abilities will help ensuring this underserved population has a positive loan experience.
- Predicting client’s repayment abilities will also ensure that clients capable of repayment are not rejected and that loans are given with a principal, maturity, and repayment calendar that will empower their clients to be successful.

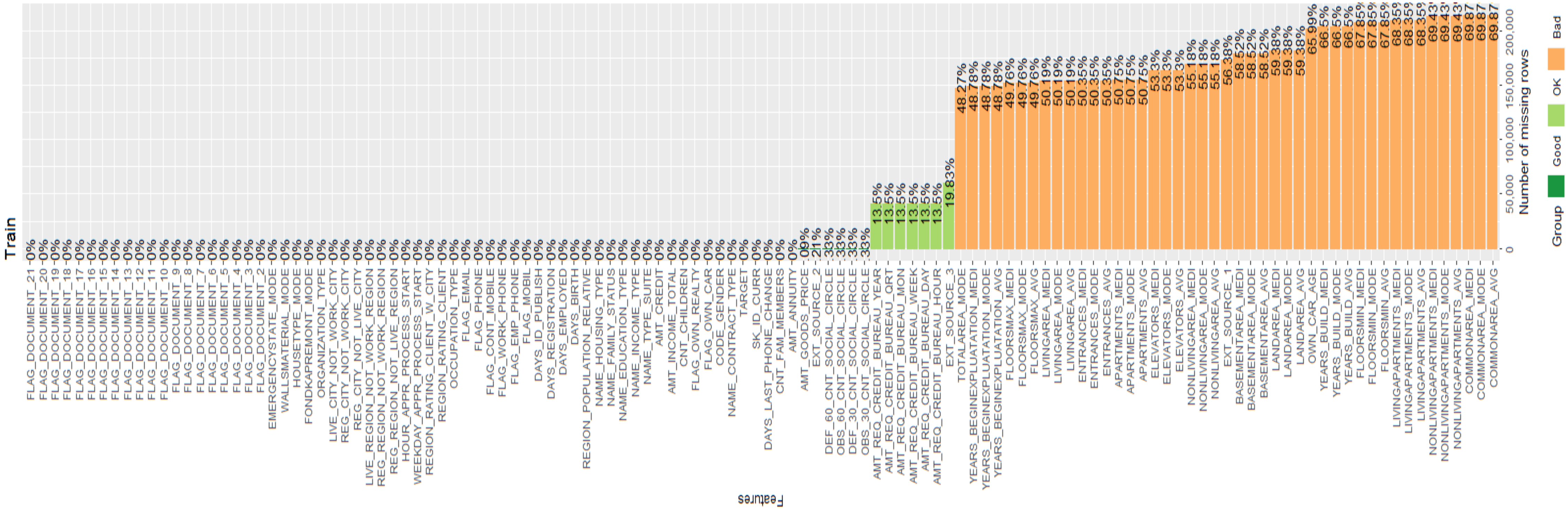
# Data Checking and Formatting

**Source of Data**

- There are 7 different sources of data, this project will only use application_train data as the baseline for analysis since it is still manageable to work with the current tool.
- Application_train: the main data with information about each loan application at Home Credit. Every loan has its own row and is identified by the feature SK_ID_CURR. The training application data comes with the TARGET indicating 0: loan repaid on time or 1: the loan was not repaid.

# Data Checking and Formatting (cont.)

**Missing Data:** it can be a great noise since we can have misleading information of the distribution of the data.



# Data Checking and Formatting (cont.)

**Anomaly Data**

DAYS_BIRTH values are negative which don't make sense. Dividing the variable with -365 will correct the data and generate new information which is approximation of client's age.

| DAYS_BIRTH <int> |
|---:|
| -9461 |
| -16765 |
| -19046 |
| -19005 |
| -19932 |
| -16941 |

6 rows | 19-23 of 122 columns

| DAYS_BIRTH <dbl> |
|---:|
| 25.92055 |
| 45.93151 |
| 52.18082 |
| 52.06849 |
| 54.60822 |
| 46.41370 |

6 rows | 19-23 of 182 column

# Data Checking and Formatting (cont.)

## Recoding Categorical Variable

There are some categorical variables that need to be recoded since they are still in string format e.g. CODE_GENDER, OCCUPATION_TYPE, etc. By recoding them, they are ready to be analyzed as categorical data (nominal and ordinal).

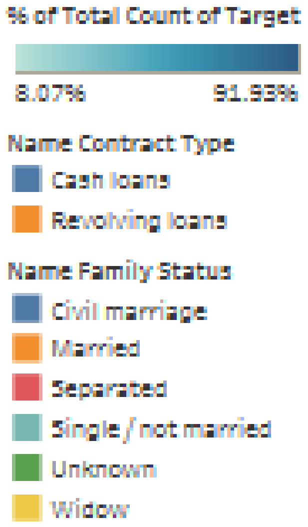
| OCCUPATION_TYPE <fctr> |
|---|
| Laborers |
| Core staff |
| Laborers |
| Laborers |
| Core staff |

| OCCUPATION_TYPE.HR.staff <dbl> | OCCUPATION_TYPE.IT.staff <dbl> | OCCUPATION_TYPE.Laborers <dbl> |
|---:|---:|---:|
| 0 | 0 | 1 |
| 0 | 0 | 0 |
| 0 | 0 | 1 |
| 0 | 0 | 1 |
| 0 | 0 | 0 |
| 0 | 0 | 1 |

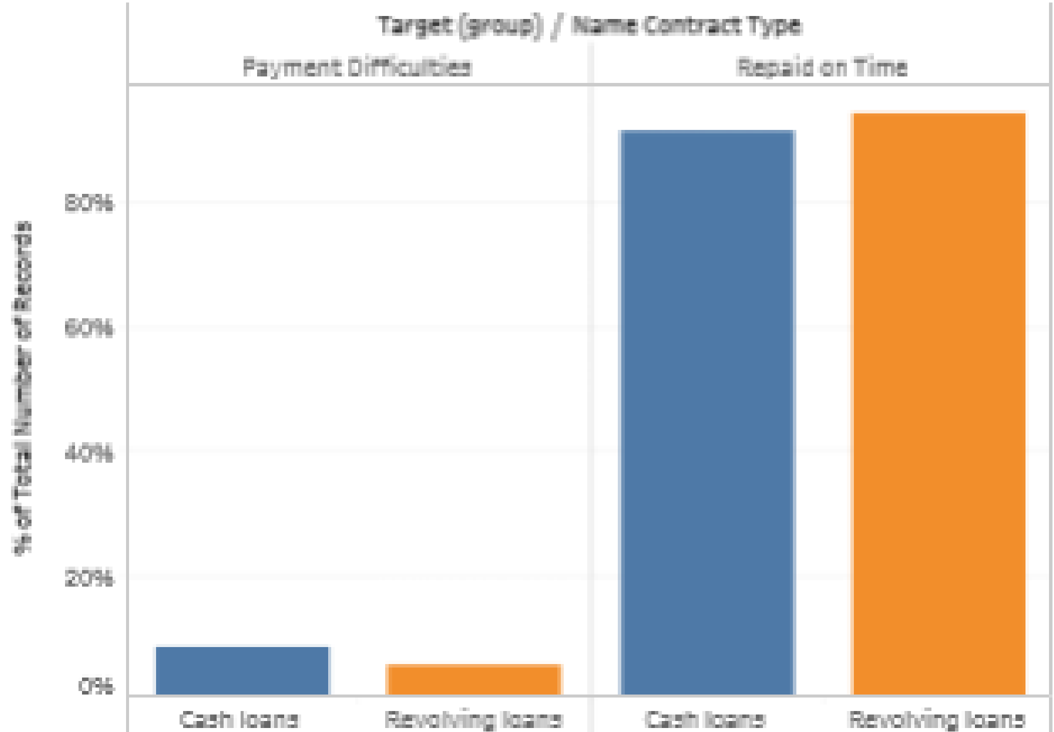
139 of 182 columns

# DATA EXPLORATORY ANALYSIS

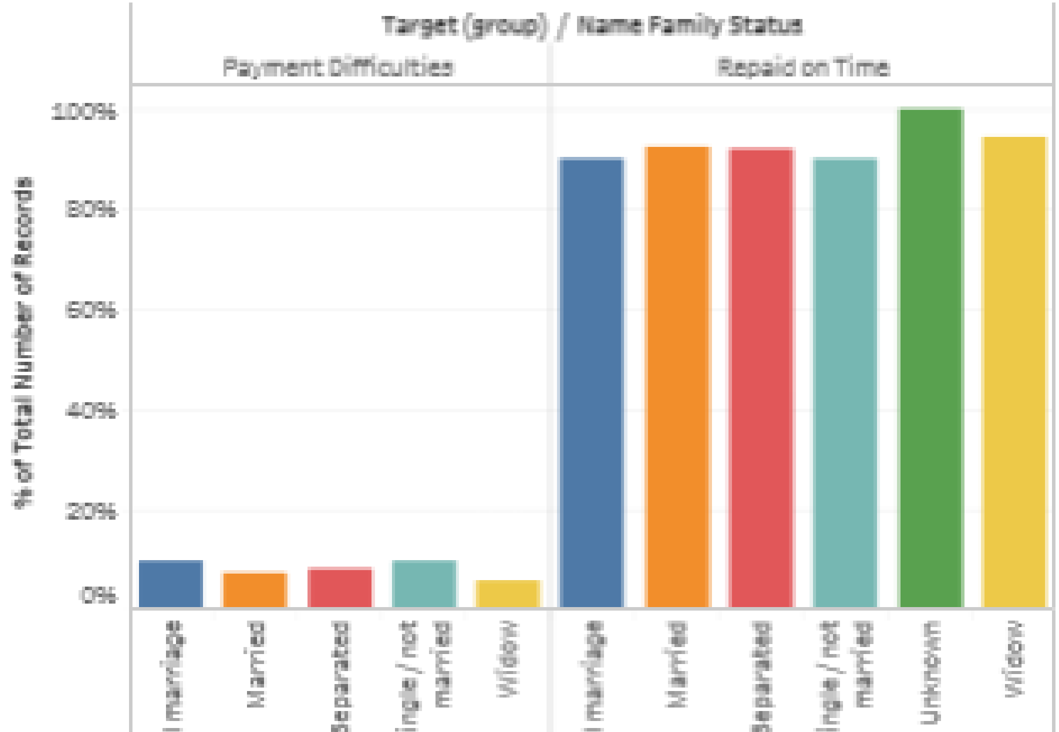
## Client Based on Target Status


## Client's Contract Type based on Target Status

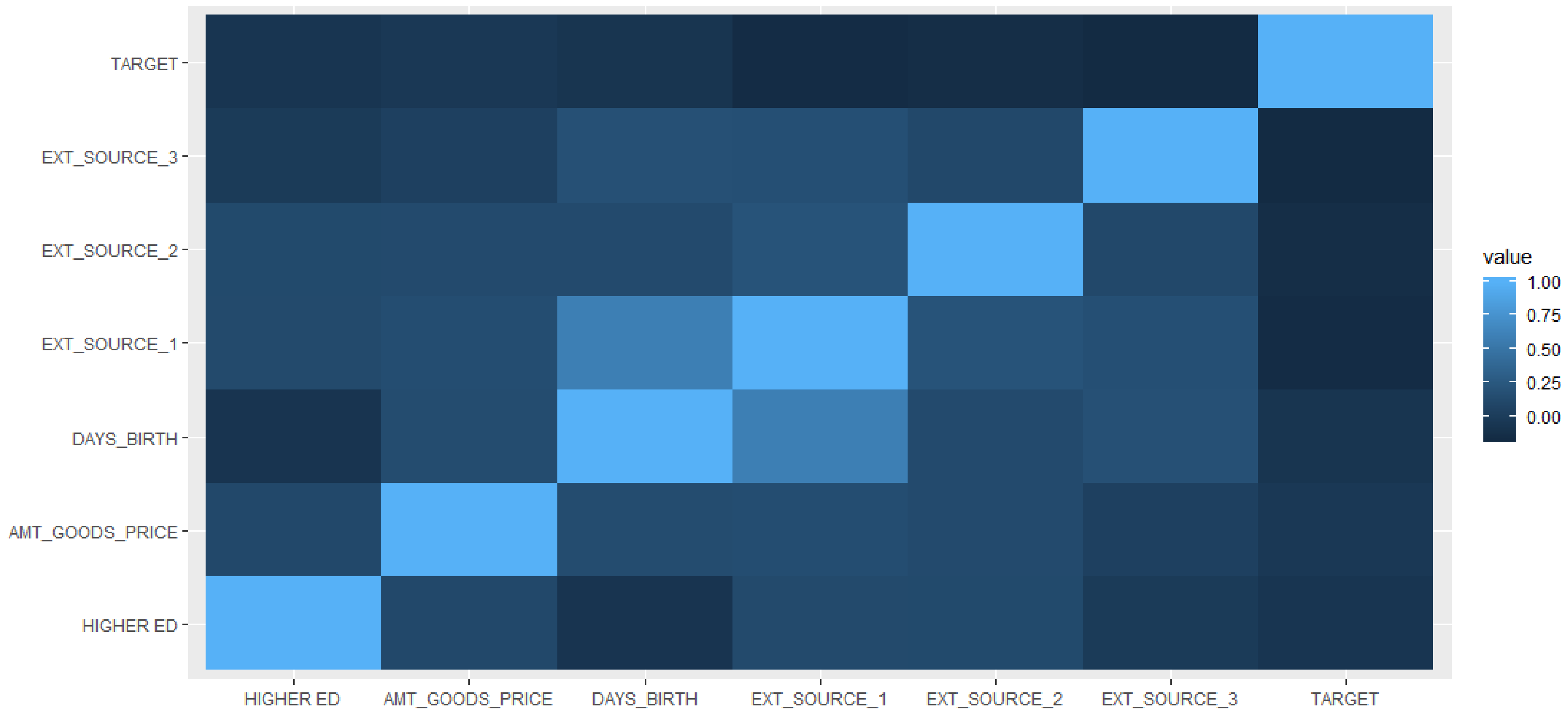



## Client's Family Status based on Target Status



# Baseline Model

## Check Correlation




From the correlation and industry-related theory, we can make the model hypothesis.

# Baseline Model (Cont.)

## Logistic Regression

```
glm(formula = TARGET ~ AMT_CREDIT + AMT_GOODS_PRICE + NAME_EDUCATION_TYPE +
    DAYS_BIRTH + EXT_SOURCE_1 + EXT_SOURCE_2 + EXT_SOURCE_3,
    family = binomial(link = "logit"), data = data, na.action = na.exclude)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2112   0.2176   0.3044   0.4197   1.5596

Coefficients:
                                                Estimate
(Intercept)                                    1.789e+00
AMT_CREDIT                                    -2.321e-06
AMT_GOODS_PRICE                                2.516e-06
NAME_EDUCATION_TYPEHigher education           -1.696e+00
NAME_EDUCATION_TYPEIncomplete higher          -1.864e+00
NAME_EDUCATION_TYPELower secondary            -2.101e+00
NAME_EDUCATION_TYPESecondary / secondary special -2.007e+00
DAYS_BIRTH                                    -1.751e-02
EXT_SOURCE_1                                   2.627e+00
EXT_SOURCE_2                                   1.980e+00
EXT_SOURCE_3                                   2.712e+00
```

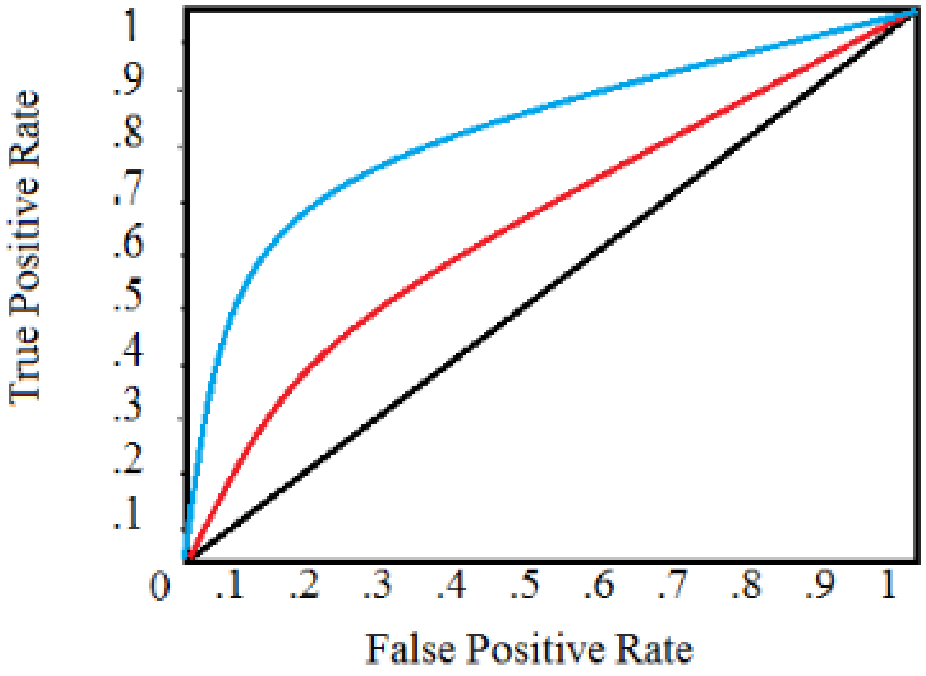
Model 1 (AIC = 63746)

```
glm(formula = TARGET ~ AMT_CREDIT + AMT_GOODS_PRICE + DAYS_BIRTH +
    EXT_SOURCE_1 + EXT_SOURCE_2 + EXT_SOURCE_3 + NAME_EDUCATION_TYPE +
    NAME_INCOME_TYPE + OCCUPATION_TYPE, family = binomial(link = "logit"),
    data = data, na.action = na.exclude)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2443   0.2159   0.3034   0.4200   1.6058

Coefficients:
                                                Estimate
(Intercept)                                    9.780e+00
AMT_CREDIT                                    -2.329e-06
AMT_GOODS_PRICE                                2.516e-06
DAYS_BIRTH                                    -1.863e-02
EXT_SOURCE_1                                   2.565e+00
EXT_SOURCE_2                                   1.970e+00
EXT_SOURCE_3                                   2.720e+00
NAME_EDUCATION_TYPEHigher education           -1.721e+00
NAME_EDUCATION_TYPEIncomplete higher          -1.869e+00
NAME_EDUCATION_TYPELower secondary            -2.073e+00
NAME_EDUCATION_TYPESecondary / secondary special -1.983e+00
```

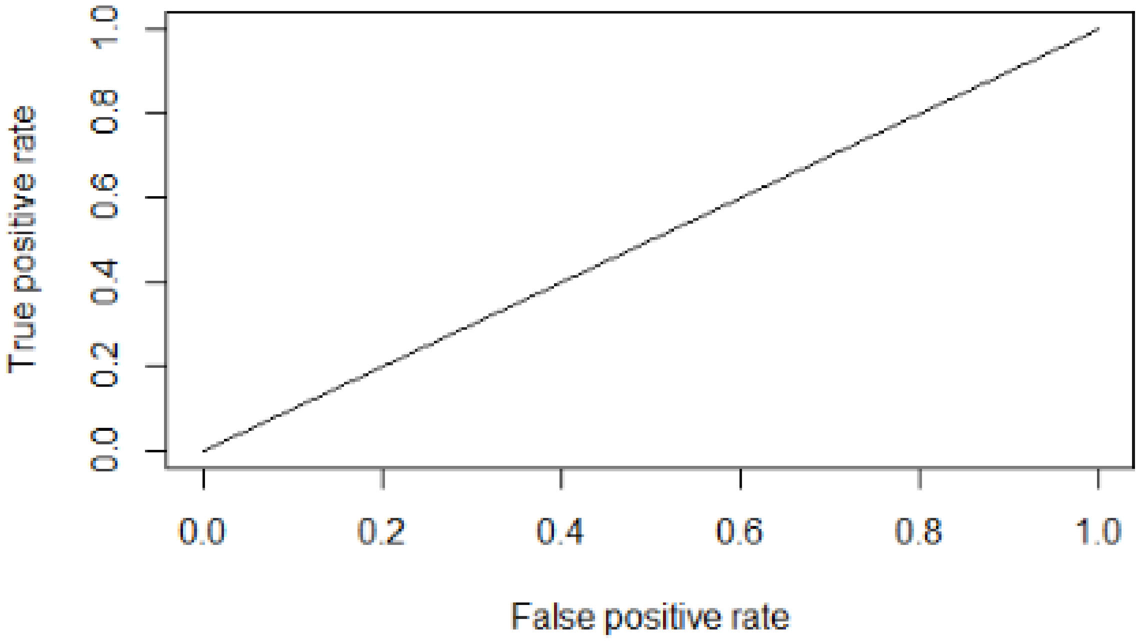
Model 2 (AIC = 63674)

# Baseline Model (Cont.)

**Idea Behind of Plot ROC**




**Logistic Regression Performance**




AUC = 0.5169

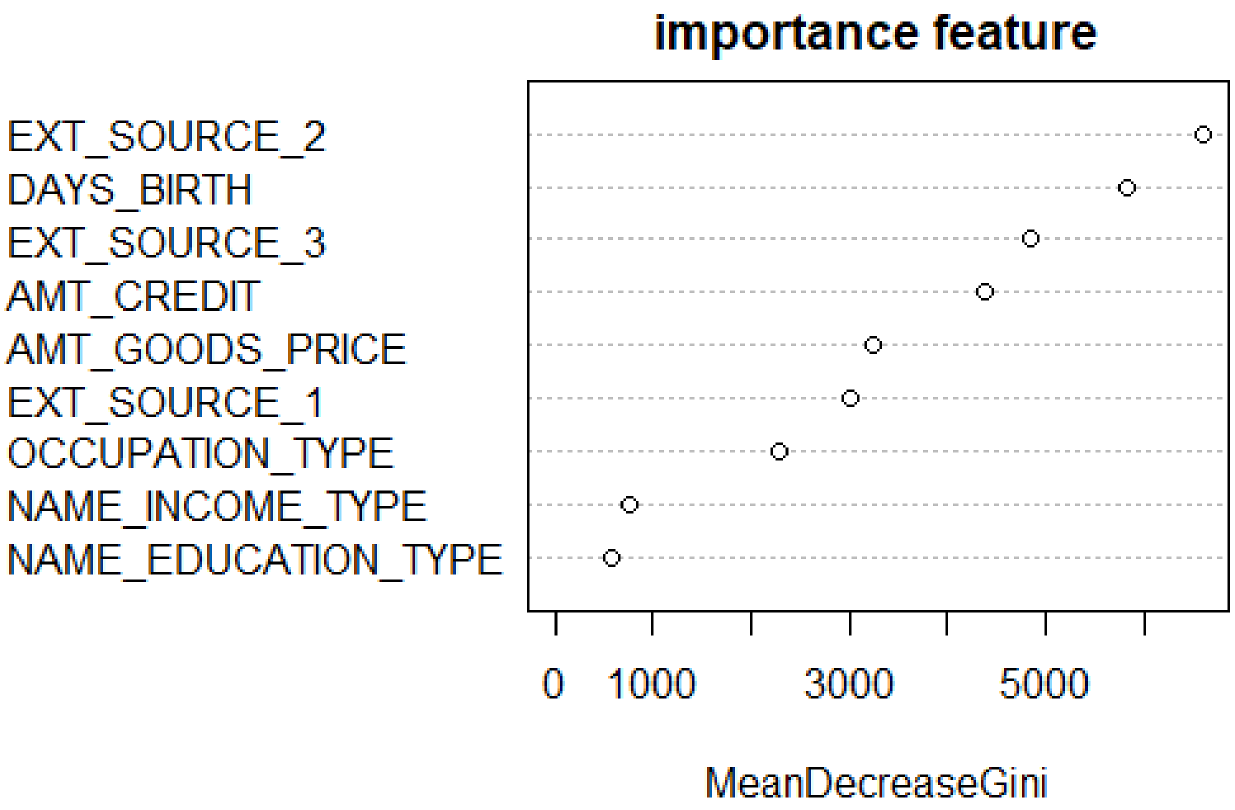
# Improved Model

**random decision forests** are an ensemble learning method for classification, regression and other tasks.

```
Call:
 randomForest(formula = TARGET ~ AMT_CREDIT + AMT_GOODS_PRICE +      DAYS_BIRTH + EXT_SOURCE_1 + EXT_SOURCE_2 +
 EXT_SOURCE_3 +      NAME_EDUCATION_TYPE + NAME_INCOME_TYPE + OCCUPATION_TYPE,      data = TrainSet, importance
 = TRUE, ntree = 50, na.action = na.roughfix)
               Type of random forest: classification
                     Number of trees: 50
No. of variables tried at each split: 3

        OOB estimate of  error rate: 8.3%
Confusion matrix:
       0    1 class.error
0 196759 1175 0.005936322
1  16698  625 0.963920799
```

AUC = 0.6899 which is better than logistic regression

# Important Feature



# Summary

- Predicting client's repayment abilities is a complex task, since the nature of the data has imbalanced class distribution.
- There is no "one clean hit" in modeling, it is a trial error process. Since, it is a computer excessive task, the right technology will improve the results.
- From baseline model, we can get the probability of client's repayment status based on the selected variables. Unfortunately, the model doesn't have good performance.
- Based on improved model, the first three of important feature is credit score from external source 2, client's age, and credit score from external source 2.
- The next question is if the client has repayment difficulties, is it genuine or a fraud attempt?

# Source


- https://www.kaggle.com/c/home-credit-default-risk
- https://www.tandfonline.com/doi/abs/10.1080/00220670209598786
- https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d
- https://medium.com/greyatom/lets-learn-about-auc-roc-curve-4a94b4d88152