Predicting Loan Default

A CRISP-DM Approach

Courtney Tan

Contents

1. Business Understanding
    1.1. Business Objectives
    1.2. Analytics Goals
    1.3. Analytics Strategy
    1.4. Project Plan (include tools used)

2. Data Understanding
    2.1. Data Source
    2.2. Data Description
    2.3. Data Quality
    2.4. Preliminary Analysis

3. Data Preparation
    3.1. Data Cleaning
    3.2. Data Understanding

4. Modelling
    4.1. Modelling Techniques
    4.2. Evaluation Metrics
    4.3. Model Design (train/test split)
    4.4. Train Data Resampling
    4.5. Model Building and Assessment
        4.5.1.Logistic Regression
        4.5.2.Decision Tree
        4.5.3.Random Forest
    4.6. Final Model
5. Evaluation
    5.1. Results
    5.2. Challenges
    5.3. Improvements
6. Conclusion
7. References

8. Appendix

1. Business Understanding

1.1. Business Objectives

Loan defaults occur when the borrower fails to repay a loan per the terms of the initial agreement. This can adversely affect the financial performance of financial institutions (FI) through cash flow and liquidity constraints, and operational costs. Therefore, the objective of this project is to build a prediction model that detects loans likely to default and determines contributing factors, to ultimately reduce credit risk and enhance profitability of the FI.

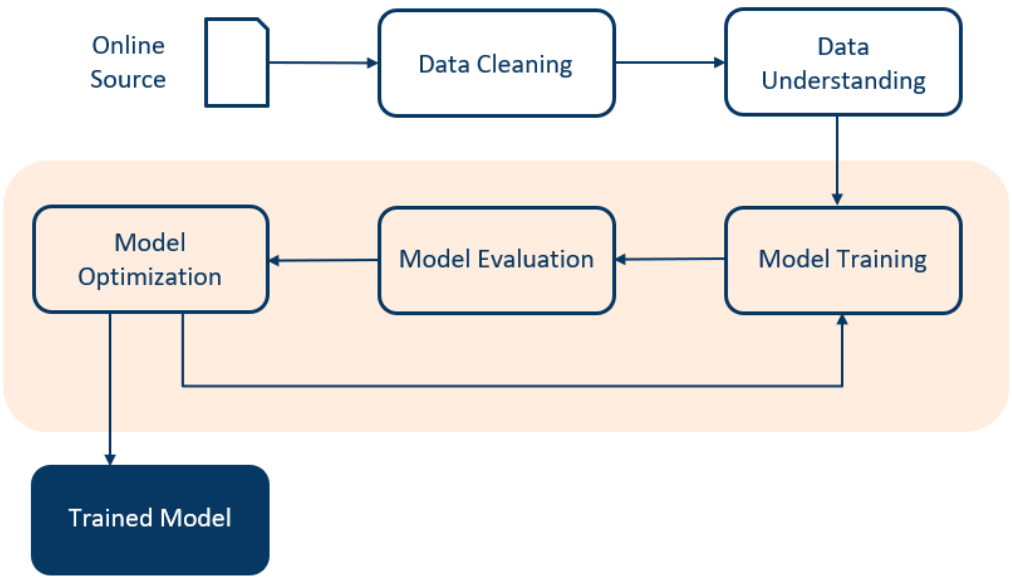
1.2. Analytics Goals

The analytics goals of this project can be divided into three categories – descriptive, diagnostic and predictive analytics. Descriptive analytics is used to determine the current loan default rate. Diagnostic analytics helps identify features that contribute towards loan default, while predictive analytics estimates the probability of loan default in existing data.

1.3. Analytics Strategy

The two primary goals of this project are to predict the probability of loan default and identify contributing features. These can be achieved by predictive models such as logistic regression, decision tree and random forest that can determine feature importance and predict a target variable. All analytics are performed using R.

1.4. Project Plan



Using a public data set, data cleaning is first performed to correct erroneous values. It is followed by data understanding to explore and visualise relationships among the data. Next comes model building, an iterative process of training, evaluating and optimizing predictive models using various techniques before achieving the final model.

2. Data Understanding

2.1. Data Source

This project uses loan application data from the Home Credit Default Risk data set sourced from Kaggle.

2.2. Data Description

```
Rows: 307511 Columns: 122
```

```
chr  (16)
dbl (106)
```

The data set has 307,511 rows and 122 columns. The response variable - 'TARGET' - takes two possible values: 0 and 1. 0 refers to no default while 1 refers to default. Column descriptions are detailed in Appendix 1.

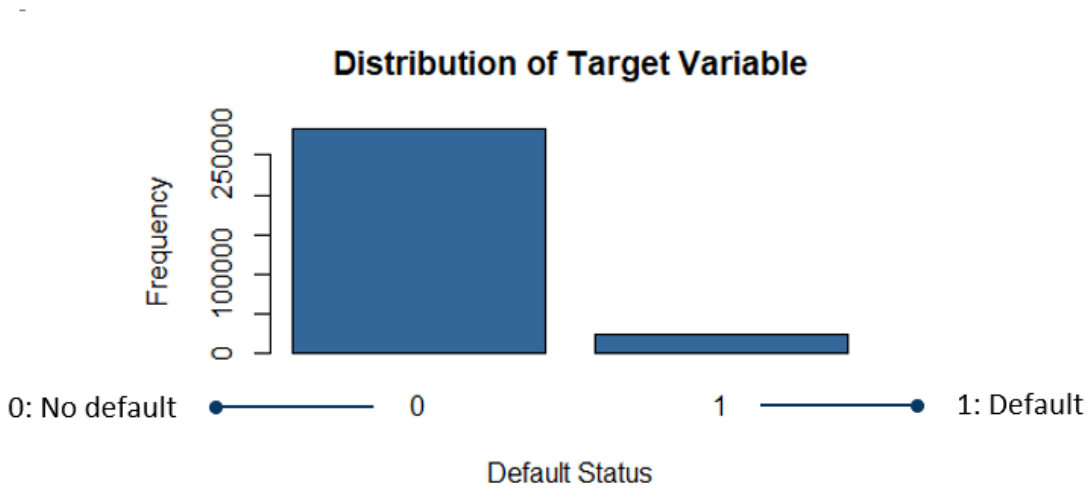
2.3. Data Quality

There were a several issues with the data set – missing values, inconsistencies and errors. Some columns comprised ≥50% missing values. The data also contained 'XNA' strings that truly meant NA, and nonsensical data such as negative years.

2.4. Preliminary Analysis

<u>Distribution of Target Variable</u>

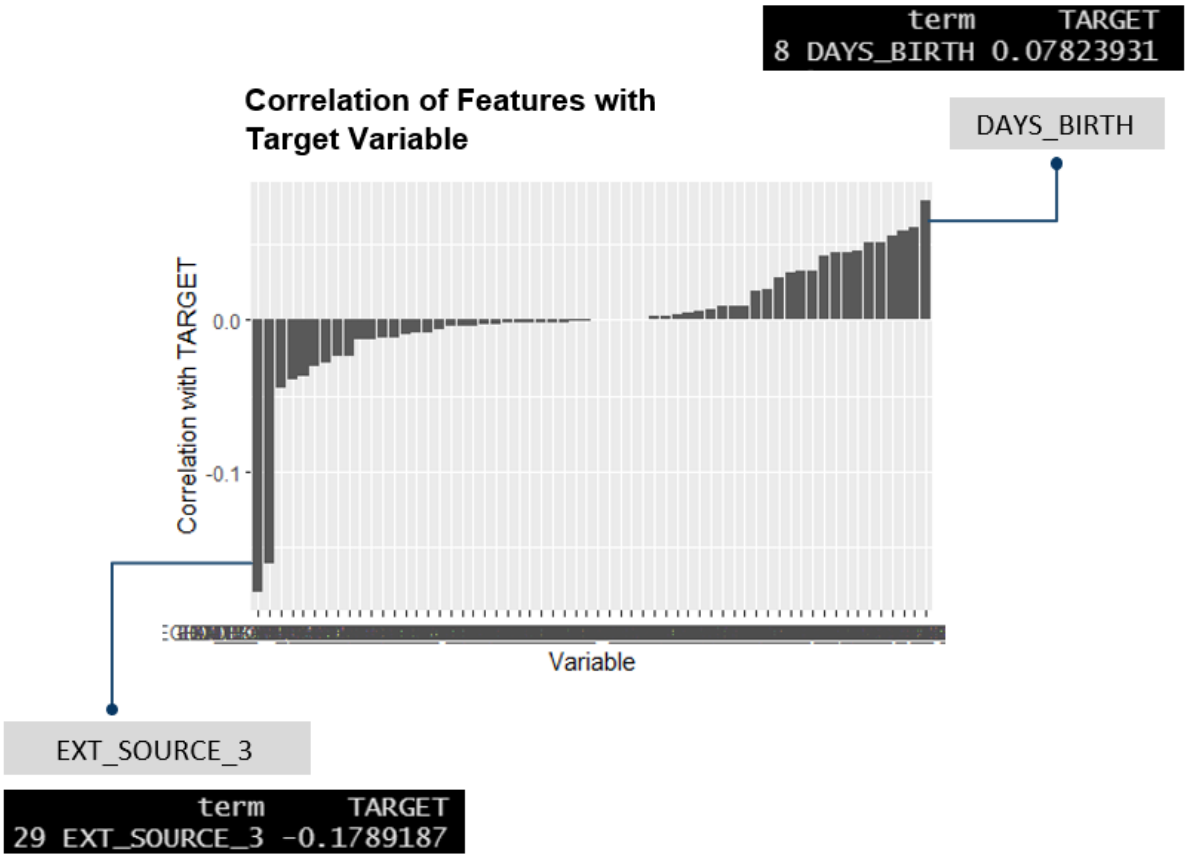
The data was highly imbalanced. The 0 class accounts for 91.9% of the observations, while the 1 class accounts for 0.08% of the observations.




```
> prop.table(table(train_set$TARGET))

        0         1
0.91910847 0.08089153
```

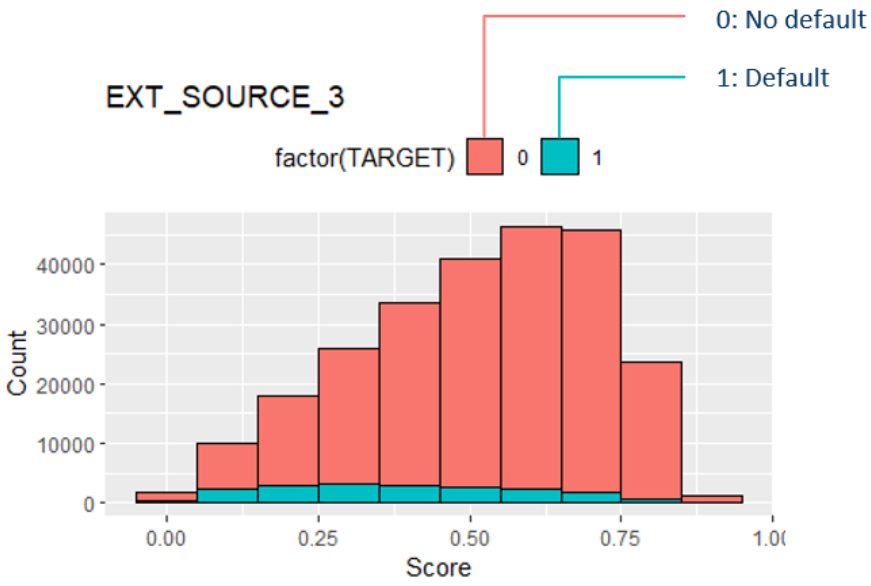
```
> table(train_set$TARGET)

     0      1
196578  17301
```

Correlation of Features against Response Variable



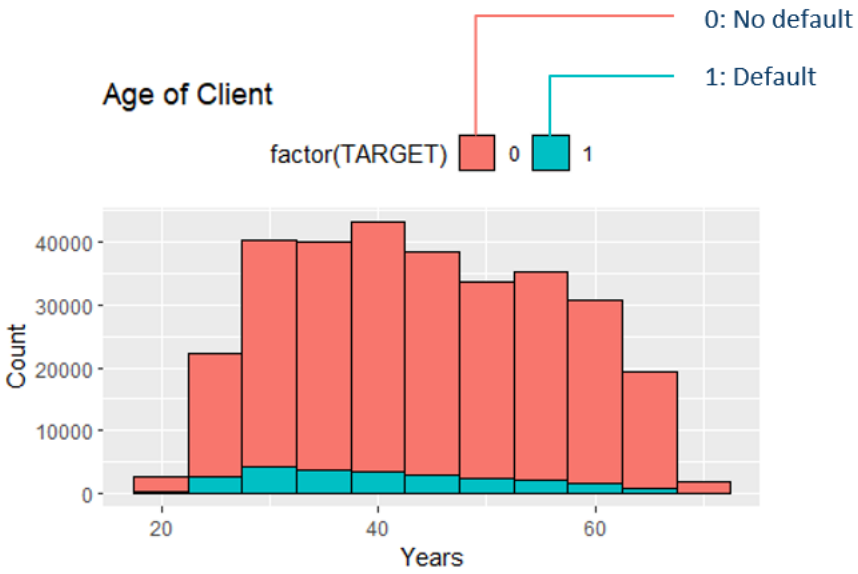
An initial correlation analysis using Pearson correlation reveals the features with highest correlations – EXT_SOURCE_3 has a correlation of -0.17 with TARGET, while DAYS_BIRTH has a correlation of 0.07 with TARGET.

EXT_SOURCE_3



EXT_SOURCE_3 is an external credit score. Its values are concentrated in the 0.5 – 0.75 range and the default observations appear the highest at approximately 0.25 in the EXT_SOURCE_3 score range.

DAYS_BIRTH



Initially referring to the number of days prior to the loan application date the client was born, the DAYS_BIRTH column was transformed to reflect the age of the client in years. Generally, the older the client, the lower the occurrence of default.

3. Data Preparation

3.1. Data Cleaning

Data Types

Continuous variables were converted to numeric type, while categorical variables were converted to factor type.

Missing Values

All columns were checked for its percentage of missing values. Those with ≥40% missing values were removed if they were not highly correlated to the response variable.

Of the remaining columns, missing values were deleted if they composed <1% of the column. Otherwise, missing values were replaced with the median if the column was numeric, while factor columns had missing values replaced with an additional factor level named 'Unknown'.

Derived Columns

Several columns – those starting with 'DAYS_*' – containing values expressed in negative days were transformed into year terms and subsequently renamed 'YEARS_*'.

Errors

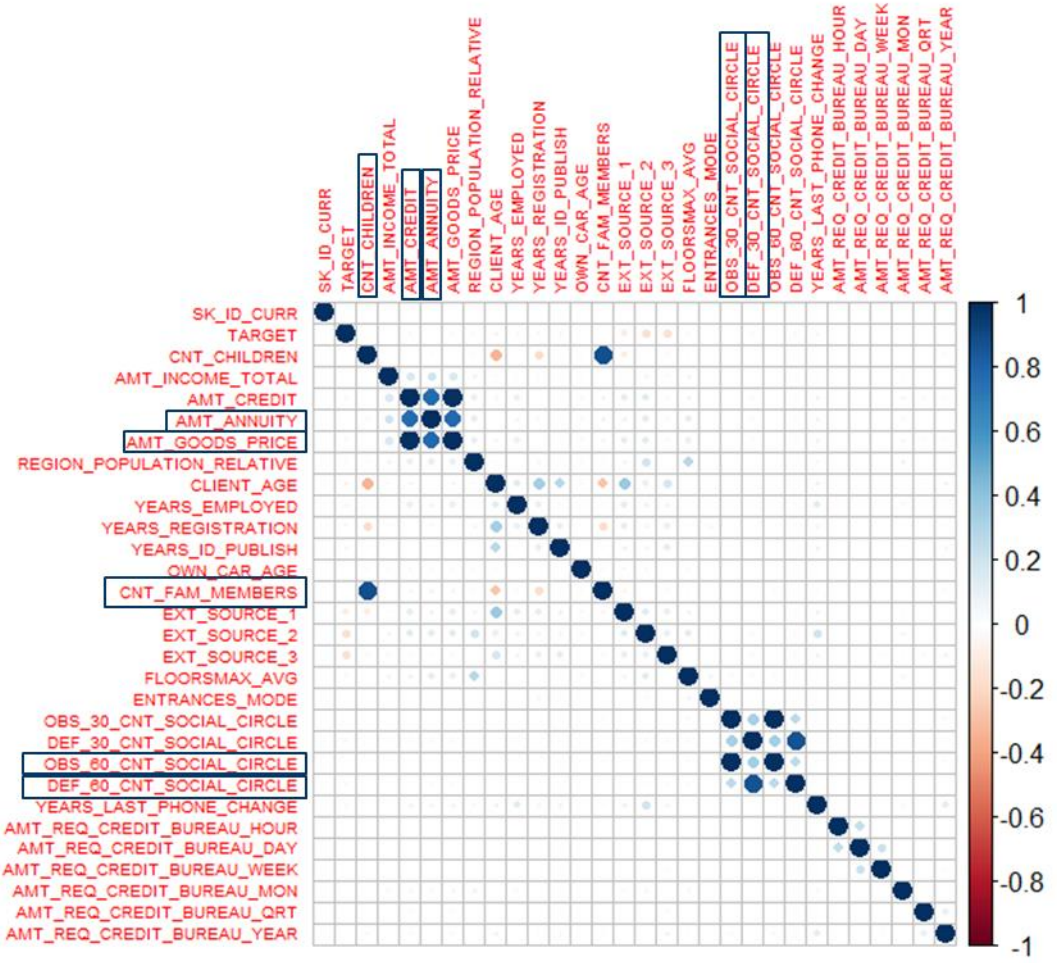
Negative values in the YEARS_EMPLOYED column were replaced with the median.

Redundant Columns

After the above cleaning steps, two factor columns - FLAG_MOBIL and NAME_FAMILY_STATUS – were made redundant with just one level remaining. They were removed along with the ID column.

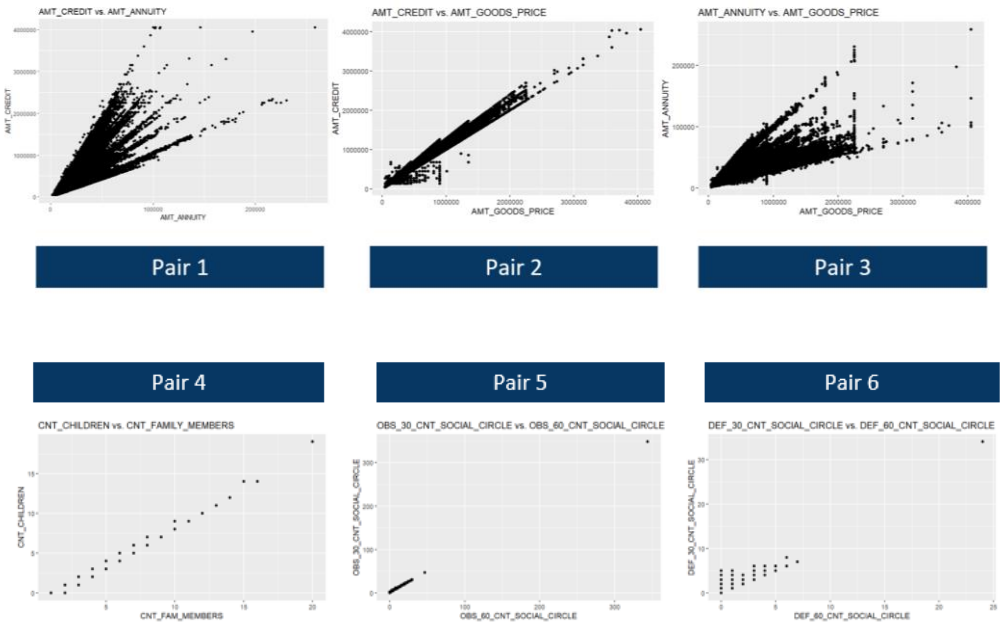
3.2. Data Understanding

3.2.1. Feature Association



To investigate pairwise relationships, a correlation analysis (Pearson correlation) was performed on the continuous variables. It revealed 6 pairs (names boxed in blue above):

Feature Association			
Pair	**Feature 1**	**Feature 2**	**Correlation**
1	AMT_ANNUITY	AMT_CREDIT	0.769
2	AMT_GOODS_PRICE	AMT_CREDIT	0.99
3	AMT_GOODS_PRICE	AMT_ANNUITY	0.774
4	CNT_FAM_MEMBERS	CNT_CHILDREN	0.88
5	OBS_60_CNT_SOCIAL_CIRCLE	OBS_30_CNT_SOCIAL_CIRCLE	0.998
6	DEF_60_CNT_SOCIAL_CIRCLE	DEF_30_CNT_SOCIAL_CIRCLE	0.86



Pair 1, 2 and 3

Pair 1, 2 and 3 are combinations of 3 features: AMT_ANNUITY, AMT_CREDIT and AMT_GOODS_PRICE. AMT_GOODS_PRICE was removed to leave the pair with the lowest correlation: AMT_ANNUITY and AMT_CREDIT.

Pair 4

Between CNT_FAMILY_MEMBERS and CNT_CHILDREN, CNT_CHILDREN was removed as the number of children would be captured in the number of family members.

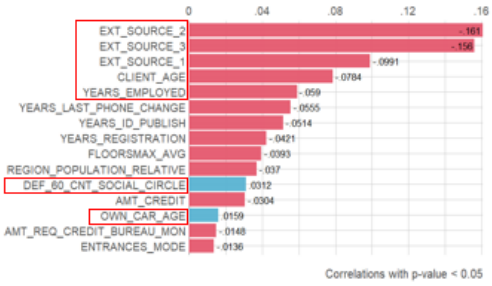
Pair 5 and 6

Pairs 5 and 6 relate to how many people in the client's social circle have defaulted. The numbers 60 and 30 refer to the number of days past due (DPD), while 'OBS' and 'DEF' refer to 'observed' and 'default' respectively. In both pairs, the 30 DPD variable was removed to consider a bigger window of missed payments.

3.2.2. Feature Importance

Continuous Variables

Pearson Correlation



EXT_SOURCE_2

EXT_SOURCE_3

EXT_SOURCE_1

CLIENT_AGE

YEARS_EMPLOYED

DEF_60_CNT_SOCIAL_CIRCLE

OWN_CAR_AGE

Categorical Variables

Chi-Squared Test

Chi-Squared test against TARGET

	Variable	p.value
1	OCCUPATION_TYPE	0.000000e+00
2	ORGANIZATION_TYPE	2.156955e-297
3	NAME_INCOME_TYPE	8.179516e-267
4	REGION_RATING_CLIENT_W_CITY	1.779467e-246
5	REGION_RATING_CLIENT	5.561268e-230
6	NAME_EDUCATION_TYPE	6.037932e-216
7	CODE_GENDER	1.865597e-203
8	REG_CITY_NOT_WORK_CITY	9.806460e-176
9	FLAG_EMP_PHONE	2.284829e-143
10	REG_CITY_NOT_LIVE_CITY	1.369267e-133

OCCUPATION_TYPE

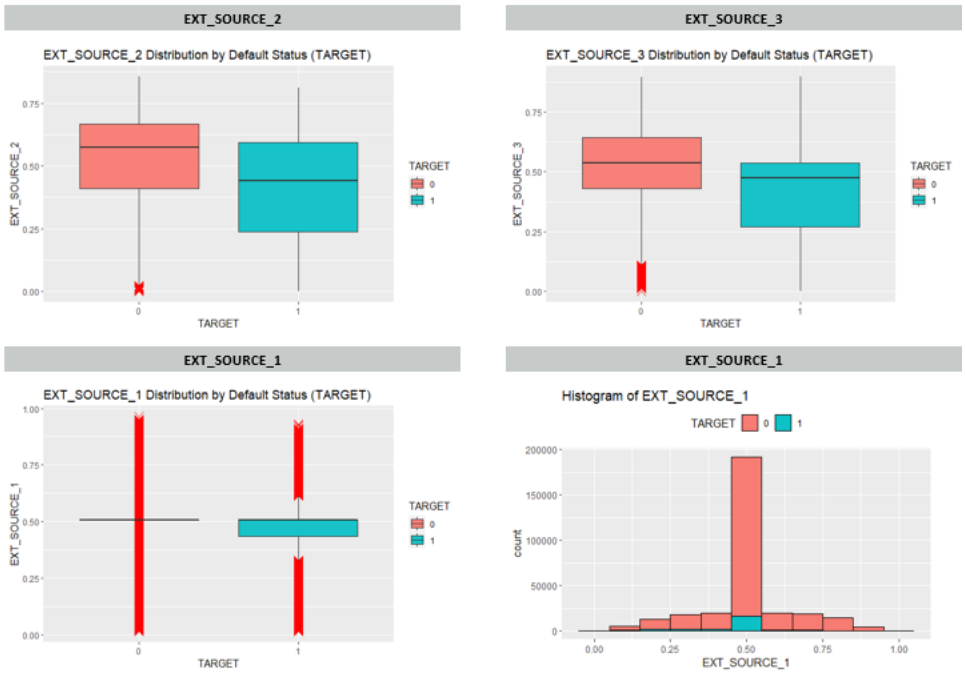
ORGANIZATION_TYPE

NAME_INCOME_TYPE

Feature importance of explanatory variables to the response variable was assessed based on Pearson correlation for continuous variables, and a Chi-Squared test for categorical variables.

3.2.3. Exploratory Plots

<u>EXT_SOURCE_2, EXT_SOURCE_3 and EXT_SOURCE_1</u>



EXT_SOURCE_2, EXT_SOURCE_3 and EXT_SOURCE_1 are normalised credit scores from external sources. Clients who default have lower scores.

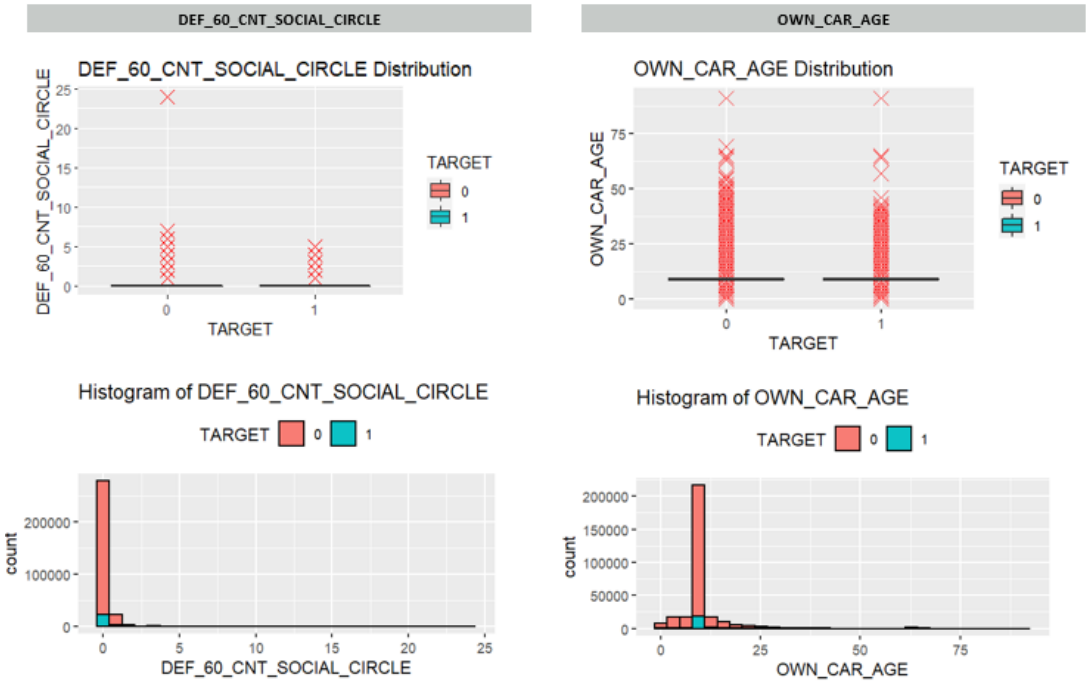
Client Age

Clients who default are younger in age.

YEARS_EMPLOYED

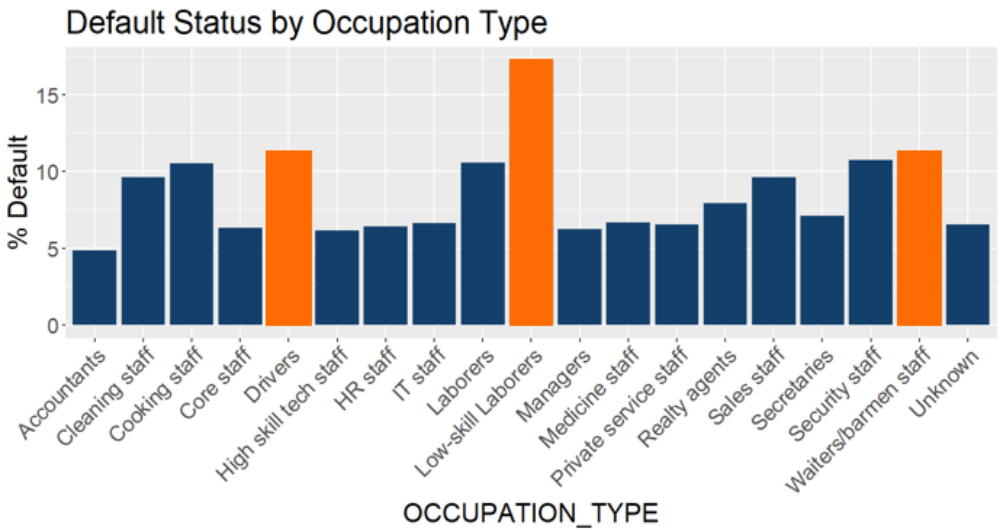
Clients who default have shorter employment periods.

DEF_60_CNT_SOCIAL_CIRCLE and OWN_CAR_AGE



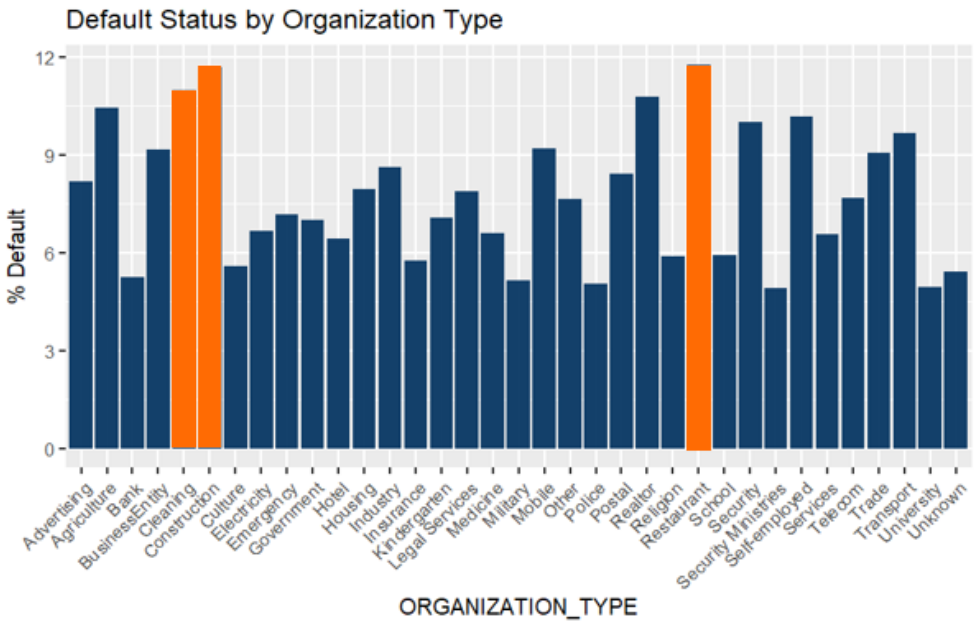
DEF_60_CNT_SOCIAL_CIRCLE and OWN_CAR_AGE are only sightly positively correlated to TARGET. There are no observable patterns between default and non-default clients.

OCCUPATION_TYPE



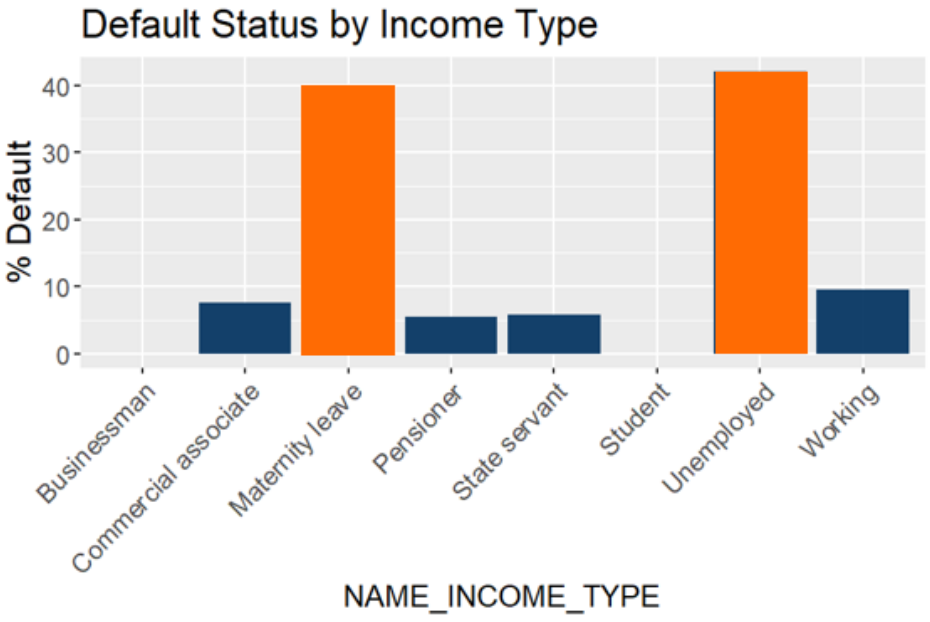
The occupation types with the highest rate of loan default are drivers, low-skill labourers and waiters/barmen staff.

ORGANIZATION_TYPE



The organization types with the highest rate of loan default are cleaning, construction and restaurant.

NAME_INCOME_TYPE



The income types with the highest rate of loan default are maternity leave and unemployed.

3.2.4. Summary

EXT_SOURCE_1	Clients who default tend to have lower external credit scores.
EXT_SOURCE_2	
EXT_SOURCE_3	
CLIENT_AGE	Clients who default tend to be younger.
YEARS_EMPLOYED	Clients who default tend to have been employed for a shorter time.
OCCUPATION_TYPE	The top 3 occupation types with the highest rate of default are drivers, low-skill labourers and waiters/barmen staff.
ORGANIZATION_TYPE	The top 3 organization types with the highest rate of default are cleaning, construction and restaurants.
NAME_INCOME_TYPE	The top 2 income types with the highest rate of default are maternity leave and unemployed.

4. Modelling

4.1. Modelling Techniques

To build a binary classification model, this project uses logistic regression, decision tree and random forest.

Logistic regression is suitable when the response variable is dichotomous. It predicts the outcome of the response variable based on probabilities and explains the relationship between the explanatory and response variables.

Decision trees and random forests are tree methods that split data into branches until a threshold value is reached. They make predictions by evaluating each variable and selecting the best attribute based on an attribute selection measure, such as information gain or the Gini Index. Advantages of these algorithms include the needlessness of heavy data cleaning and ability to handle both numerical and categorical values.

Between decision trees and random forest, random forests are a collection of decision trees. The random forest algorithm selects features randomly during the training process and combines the output of individual decision trees to generate the final output. This randomised feature selection thus generalises over the data and enables higher accuracy.

4.2. Evaluation Metrics

Accuracy

Accuracy refers to the proportion of correct predictions. It is a basic indicator with flaws when handling imbalanced data – it may not be able to predict the minority class.

Precision

Precision is the proportion of positive predictions that are correctly predicted.

Recall

Recall indicates the proportion of actual positive cases that are correctly predicted. Also known as sensitivity, this metric is prioritized as the objective of this project is to predict loans likely to default.

F1 Score

Since there is a trade-off between precision and recall, the F1 Score, being the harmonic mean of precision and recall, combines the two in a single metric.

4.3. Model Design (Train/Test Split)

The data set was split into train and validation sets in a 70/30 ratio (i.e. 70% of observations for training and 30% for validation).

Train Set		
TARGET	No. rows	%
0	196,578	91.9
1	17,301	8.09

Validation Set		
TARGET	No. rows	%
0	84,247	91.9
1	7,415	8.09

4.4. Initial Modelling

Logistic Regression

Fitting a logistic regression model started using all variables. Following that, multicollinearity was eliminated using the variance inflation factor (VIF), and insignificant variables removed based on p-value.

Model 1 (Appendix 2)

Using all variables. Running the model on the validation set, the confusion matrix is as follows:

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 84166  7330
         1    81    85

               Accuracy : 0.9191
                 95% CI : (0.9174, 0.9209)
    No Information Rate : 0.9191
    P-Value [Acc > NIR] : 0.4838

                  Kappa : 0.0189

 Mcnemar's Test P-Value : <0.0000000000000002

            Sensitivity : 0.99904
            Specificity : 0.01146
         Pos Pred Value : 0.91989
         Neg Pred Value : 0.51205
             Prevalence : 0.91910
         Detection Rate : 0.91822
   Detection Prevalence : 0.99819
      Balanced Accuracy : 0.50525

       'Positive' Class : 0
```

High recall (sensitivity), poor specificity.

Model 2 (Appendix 3)

```
## model 2 - lr2 - no NAME_INCOME_TYPE
lr2 <- glm(formula = TARGET ~ . -ORGANIZATION_TYPE, family = binomial, data = train_set)
summary(lr2)
vif(lr2)
```

Removed ORGANIZATION_TYPE.

Model 3 (Appendix 4)

```
## model 3 - lr3 - drop NAME_INCOME_TYPE
lr3 <- glm(formula = TARGET ~ . -ORGANIZATION_TYPE -NAME_INCOME_TYPE, family = binomial, data = train_set)
vif(lr3)
```

Removed NAME_INCOME_TYPE.

Model 4 (Appendix 5)

```
# model 4 - lr4 - drop REGION_RATING_CLIENT
lr4 <- glm(formula = TARGET ~ . -ORGANIZATION_TYPE -NAME_INCOME_TYPE -REGION_RATING_CLIENT,
family = binomial, data = train_set)
```

Removed REGION_RATING_CLIENT.

Model 5 (Appendix 6)

```
# model 5 - lr5 - drop REG_REGION_NOT WORK_REGION
lr5 <- glm(formula = TARGET ~ . -ORGANIZATION_TYPE -NAME_INCOME_TYPE -REGION_RATING_CLIENT
-REG_REGION_NOT_WORK_REGION, family = binomial, data = train_set)
```

Removed REG_REGION_NOT_WORK_REGION.

Model 6 (Appendix 7)

```
# model 6 - lr6 - drop FLAG_DOCUMENT_3
lr6 <- glm(formula = TARGET ~ . -ORGANIZATION_TYPE -NAME_INCOME_TYPE -REGION_RATING_CLIENT
-REG_REGION_NOT_WORK_REGION -FLAG_DOCUMENT_3, family = binomial, data = train_set)
```

Removed FLAG_DOCUMENT_3.

Model 7 (Appendix 8)

```
# model 7 - lr7 - drop REG_CITY_NOT_WORK_CITY
lr7 <- glm(formula = TARGET ~ . -ORGANIZATION_TYPE -NAME_INCOME_TYPE -REGION_RATING_CLIENT
-REG_REGION_NOT_WORK_REGION -FLAG_DOCUMENT_3 -REG_CITY_NOT_WORK_CITY, family = binomial, data = train_set)
```

Removed REG_CITY_NOT_WORK_CITY.

Model 8 (Appendix 9)

Removed insignificant variables (p-value):

AMT_INCOME_TOTAL	LIVE_REGION_NOT_WORK_REGION
CLIENT_AGE	LIVE_CITY_NOT_WORK_CITY
CNT_FAM_MEMBERS	FLAG_DOCUMENT_4
OBS_60_CNT_SOCIAL_CIRCLE	FLAG_DOCUMENT_5
AMT_REQ_CREDIT_BUREAU_HOUR	FLAG_DOCUMENT_6
AMT_REQ_CREDIT_BUREAU_DAY	FLAG_DOCUMENT_7
AMT_REQ_CREDIT_BUREAU_YEAR	FLAG_DOCUMENT_9
LAG_OWN_REALTY	FLAG_DOCUMENT_10
NAME_TYPE_SUITE	FLAG_DOCUMENT_12

NAME_EDUCATION_TYPE	FLAG_DOCUMENT_15
NAME_HOUSING_TYPE	FLAG_DOCUMENT_17
FLAG_EMAIL	FLAG_DOCUMENT_19
OCCUPATION_TYPE	FLAG_DOCUMENT_20
WEEKDAY_APPR_PROCESS_START	FLAG_DOCUMENT_21
HOUR_APPR_PROCESS_START	

Decision Tree

The decision tree model did not predict any defaults.

Confusion matrix: run model on train data

```
> confusionMatrix(dtpredt, train_set$TARGET, positive = '1')
Confusion Matrix and Statistics

          Reference
Prediction      0      1
         0 196578  17301
         1      0      0

               Accuracy : 0.9191
                 95% CI : (0.9179, 0.9203)
    No Information Rate : 0.9191
    P-Value [Acc > NIR] : 0.502

                  Kappa : 0

 Mcnemar's Test P-Value : <0.0000000000000002

            Sensitivity : 0.00000
            Specificity : 1.00000
         Pos Pred Value :     NaN
         Neg Pred Value : 0.91911
             Prevalence : 0.08089
         Detection Rate : 0.00000
   Detection Prevalence : 0.00000
      Balanced Accuracy : 0.50000

       'Positive' Class : 1
```

Confusion matrix: run model on validation data

```
> confusionMatrix(dtpred, validation_set$TARGET, positive = '1')
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 84247  7415
         1     0     0

               Accuracy : 0.9191
                 95% CI : (0.9173, 0.9209)
    No Information Rate : 0.9191
    P-Value [Acc > NIR] : 0.5031

                  Kappa : 0

 Mcnemar's Test P-Value : <0.0000000000000002

            Sensitivity : 0.0000
            Specificity : 1.0000
         Pos Pred Value :    NaN
         Neg Pred Value : 0.9191
             Prevalence : 0.0809
         Detection Rate : 0.0000
   Detection Prevalence : 0.0000
      Balanced Accuracy : 0.5000

       'Positive' Class : 1
```

Random Forest

Confusion matrix: run model on train data

```
Error matrix for the Random Forest model on train_a (counts):

      Predicted
Actual      0     1 Error
     0 196556    22   0.0
     1   5211 12090  30.1

Error matrix for the Random Forest model on train_a (proportions):

      Predicted
Actual    0   1 Error
     0 91.9 0.0   0.0
     1  2.4 5.7  30.1

Overall error: 2.4%, Averaged class error: 15.05%

Rattle timestamp: 2022-04-23 22:30:42 court
========================================================================
```

Confusion matrix: run model on validation data

```
Error matrix for the Random Forest model on validation_a (counts):

      Predicted
Actual     0  1 Error
     0 84220 27   0.0
     1  7392 23  99.7

Error matrix for the Random Forest model on validation_a (proportions):

      Predicted
Actual    0 1 Error
     0 91.9 0   0.0
     1  8.1 0  99.7

Overall error: 8.1%, Averaged class error: 49.85%

Rattle timestamp: 2022-04-23 22:29:23 court
============================================================================
```

Summary

	Train Set		Validation Set	
Logistic Regression	**Accuracy**	91.90%	**Accuracy**	91.90%
	Precision	45.07%	**Precision**	49.07%
	Recall	0.74%	**Recall**	0.71%
	F1 Score	1.46%	**F1 Score**	1.41%
Decision Tree	**Accuracy**	91.90%	**Accuracy**	91.90%
	Precision	0.00%	**Precision**	0.00%
	Recall	0.00%	**Recall**	0.00%
	F1 Score	0.00%	**F1 Score**	0.00%
Random Forest	**Accuracy**	97.55%	**Accuracy**	91.91%
	Precision	99.82%	**Precision**	46.00%
	Recall	69.88%	**Recall**	0.31%
	F1 Score	82.21%	**F1 Score**	0.62%

The logistic regression model performance is similar across the train and validation sets – it has poor precision at 45 – 49% and even poorer recall at around 0.7%. The decision tree had no predictive power – precision and recall are 0 because it predicted all records as one class – the majority class. The random forest model performed well on the train set, but not on the validation set, which suggests overfitting.

4.5. Train Data Resampling

To address the imbalanced data, a resampling technique was applied. Random oversampling examples (ROSE) is a synthetic data generation technique that uses

bootstrapping and k-nearest neighbours to balance the classes. After resampling, a 50-50 distribution of target classes is achieved.

TARGET Class Distribution - Before		
Train Set	**0**	**1**
Count	196,578	17,301
Proportion	91.9%	8.1%

TARGET Class Distribution - After		
Train Set	**0**	**1**
Count	106,824	107,055
Proportion	49.9%	50.1%

4.6. Model Building and Assessment

The following models were built using resampled data.

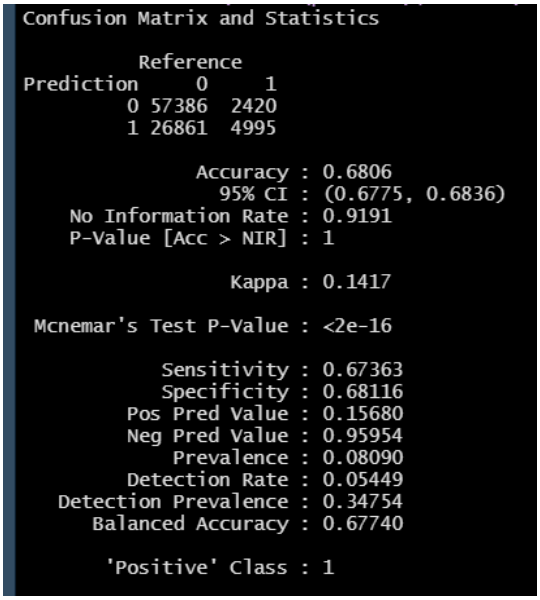
4.6.1. Logistic Regression

A logistic regression model was built using the variables determined in Model 8 above (refer to Appendix 10 for output).

Confusion matrix on train data:

```
Confusion Matrix and Statistics

          Reference
Prediction      0      1
         0 134124   5735
         1  62454  11566

               Accuracy : 0.6812
                 95% CI : (0.6792, 0.6832)
    No Information Rate : 0.9191
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1406

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

            Sensitivity : 0.66852
            Specificity : 0.68229
         Pos Pred Value : 0.15626
         Neg Pred Value : 0.95899
             Prevalence : 0.08089
         Detection Rate : 0.05408
   Detection Prevalence : 0.34608
      Balanced Accuracy : 0.67541

       'Positive' Class : 1
```

Confusion matrix on validation data:



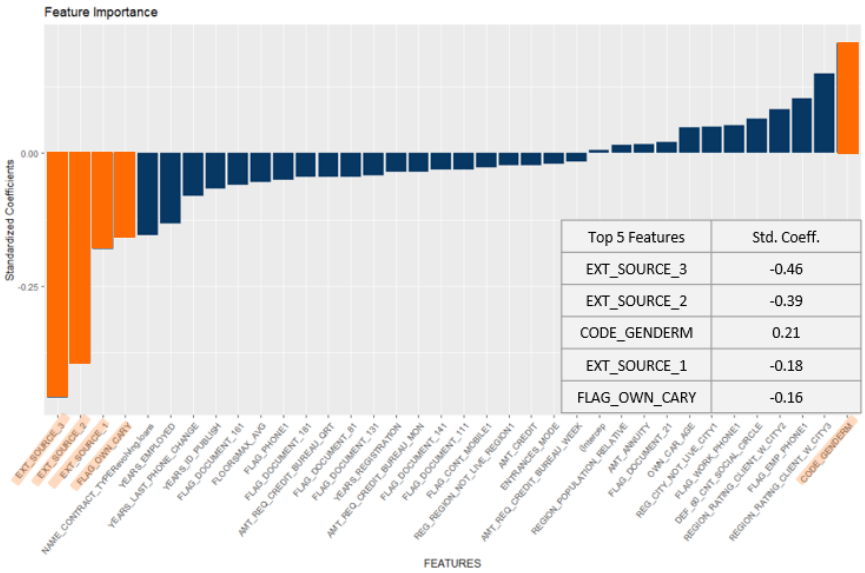
```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 57386  2420
         1 26861  4995

               Accuracy : 0.6806
                 95% CI : (0.6775, 0.6836)
    No Information Rate : 0.9191
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1417

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

            Sensitivity : 0.67363
            Specificity : 0.68116
         Pos Pred Value : 0.15680
         Neg Pred Value : 0.95954
             Prevalence : 0.08090
         Detection Rate : 0.05449
   Detection Prevalence : 0.34754
      Balanced Accuracy : 0.67740

       'Positive' Class : 1
```

Multicollinearity and insignificant variables removed

Logistic Regression	Train Set	Validation Set
Accuracy	68.1%	68.0%
Precision	15.6%	15.7%
Recall	66.9%	67.4%
F1 Score	25.3%	25.4%

Compared to the initial model, the accuracy is lower but precision and recall improved significantly

The model also indicated feature importance based on standardized coefficients (Appendix 11).



4.6.2. Decision Tree

The decision tree model was trained on resampled data using all variables.

All variables

Type: Tree Forest Boost SVM Linear Neural Net Survival All

Target: TARGET Algorithm: Traditional Conditional

Min Split: 20 Max Depth: 30 Priors:

Min Bucket: 7 Complexity: 0.0100 Loss Matrix:

Decision Tree	Train Set	Validation Set
Accuracy	83.8%	89.3%
Precision	90.9%	5.0%
Recall	75.0%	1.8%
F1 Score	82.2%	2.6%

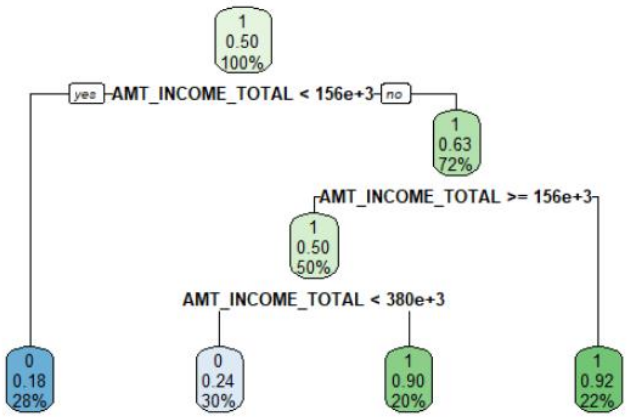
Confusion matrix on train data:

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 98799 26733
         1  8025 80322

               Accuracy : 0.8375
                 95% CI : (0.8359, 0.839)
    No Information Rate : 0.5005
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.675

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.7503
            Specificity : 0.9249
         Pos Pred Value : 0.9092
         Neg Pred Value : 0.7870
             Prevalence : 0.5005
         Detection Rate : 0.3755
   Detection Prevalence : 0.4131
      Balanced Accuracy : 0.8376

       'Positive' Class : 1
```

Confusion matrix on validation data:

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 81735  7282
         1  2512   133

               Accuracy : 0.8932
                 95% CI : (0.8911, 0.8951)
    No Information Rate : 0.9191
    P-Value [Acc > NIR] : 1

                  Kappa : -0.0168

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

            Sensitivity : 0.017937
            Specificity : 0.970183
         Pos Pred Value : 0.050284
         Neg Pred Value : 0.918195
             Prevalence : 0.080895
         Detection Rate : 0.001451
   Detection Prevalence : 0.028856
      Balanced Accuracy : 0.494060

       'Positive' Class : 1
```

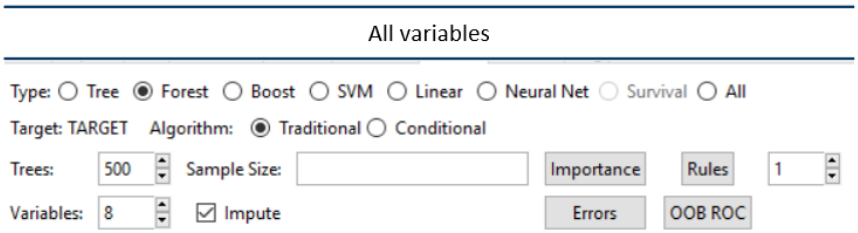
Tree:



The decision tree splits the data according to only one feature - total income amount – at different points. Looking at the first terminal node from the left, the default rate is the lowest at 18% when the total income amount is less than $156,000. On the other hand, The default rate is highest when the income is > $156,000 and <$380,000.

4.6.3. Random Forest

A random forest model was trained on resampled data using all variables and default settings.



Random Forest	Train Set	Validation Set
Accuracy	96.2%	89.5%
Precision	94.6%	24.3%
Recall	97.7%	14.0%
F1 Score	96.2%	17.8%

Running the model on the train set yielded good results, but on the validation set, the model performed rather poorly, with precision and recall at 24 and 14% respectively. This suggests overfitting.

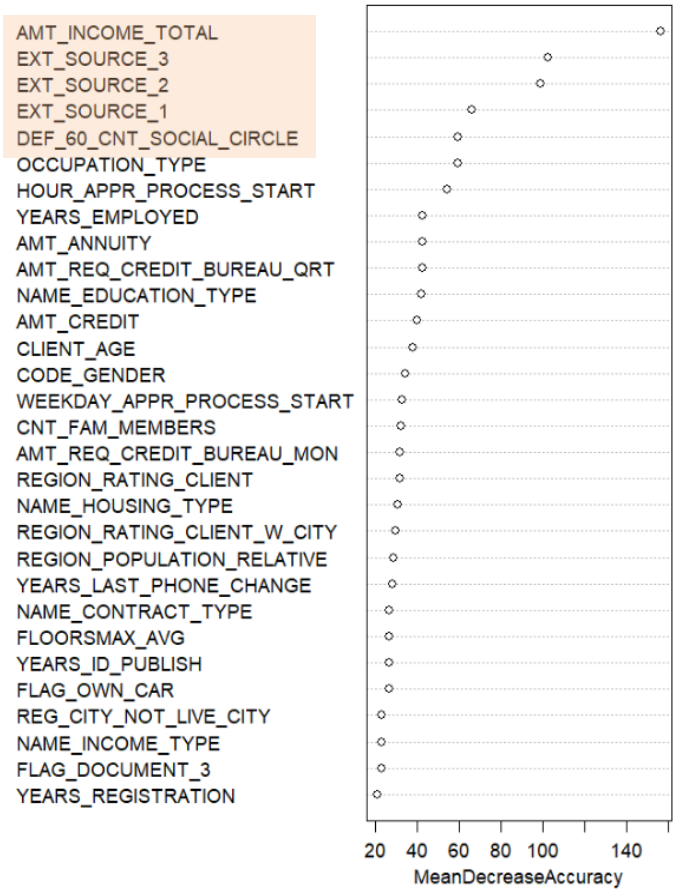
Confusion matrix on train data:

```
Error matrix for the Random Forest model on train_a_rf.rose (counts):

      Predicted
Actual      0      1 Error
     0 104462   2362   2.2
     1   5743 101312   5.4

Error matrix for the Random Forest model on train_a_rf.rose (proportions):

      Predicted
Actual    0    1 Error
     0 48.8  1.1   2.2
     1  2.7 47.4   5.4

Overall error: 3.8%, Averaged class error: 3.8%

Rattle timestamp: 2022-04-20 15:47:28 court
========================================================================
```

Confusion matrix on validation data:

```
Error matrix for the Random Forest model on validation_a (counts):

      Predicted
Actual     0    1 Error
     0 81012 3235   3.8
     1  6374 1041  86.0

Error matrix for the Random Forest model on validation_a (proportions):

      Predicted
Actual    0   1 Error
     0 88.4 3.5   3.8
     1  7.0 1.1  86.0

Overall error: 10.5%, Averaged class error: 44.9%

Rattle timestamp: 2022-04-20 14:54:15 court
```

The top five predictors of loan default as determined by the random forest model are (1) total income amount, (2) external source 3, (3) external source 2, (4) external source 1, and (5) number of 60 days past due in the client's social circle.



4.6.4. Comparison

Comparison	Logistic Regression	Decision Tree	Random Forest
Accuracy	68.0%	89.3%	89.5%
Precision	15.7%	5.0%	24.3%
Recall	67.4%	1.8%	14.0%
F1 Score	25.4%	2.6%	17.8%

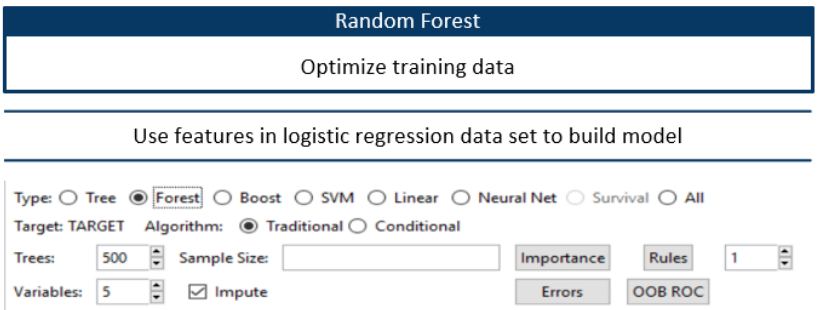
Among the three models, the decision tree can be ruled out as it has very low precision despite having high accuracy. Between the logistic regression and random forest models, logistic regression performed better in terms of recall, while random forest was superior in accuracy and precision. Judging from the F1 score, the logistic regression ultimately performed the best among the three.

However, since random forests have many advantages such as being good at handling large datasets, an attempt was made to train a random forest model using the training data determined by the logistic regression model.

4.6.5. Random Forest – Model Tuning

<u>Optimize Training Data</u>

A random forest model was built using features in the logistic regression data set – free of multicollinearity and insignificant variables.



Random Forest	Logistic Regression	Random Forest (Unoptimized)	Random Forest (Optimized)
Accuracy	68.0%	89.5%	72.5%
Precision	15.7%	24.3%	16.7%
Recall	67.4%	14.0%	60.2%
F1 Score	25.4%	17.8%	26.1%

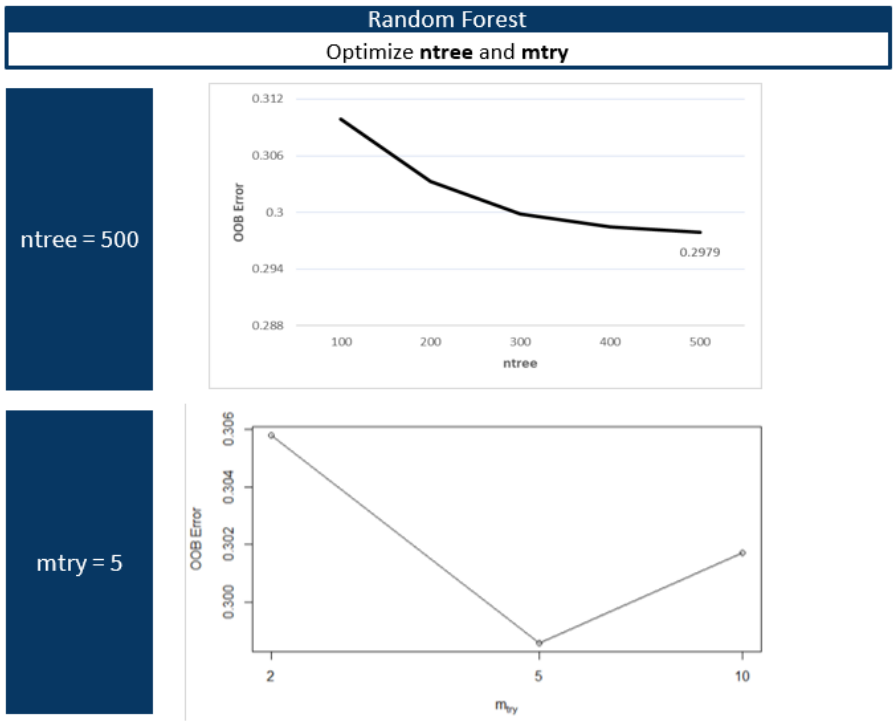
Confusion matrix on validation set:

```
Error matrix for the Random Forest model on validation_a (counts):

      Predicted
Actual     0     1 Error
     0 61950 22297  26.5
     1  2952  4463  39.8

Error matrix for the Random Forest model on validation_a (proportions):

      Predicted
Actual    0    1 Error
     0 67.6 24.3  26.5
     1  3.2  4.9  39.8

Overall error: 27.5%, Averaged class error: 33.15%
```

The F1 score improved over the previous random forest model, and is also higher than the logistic regression model.

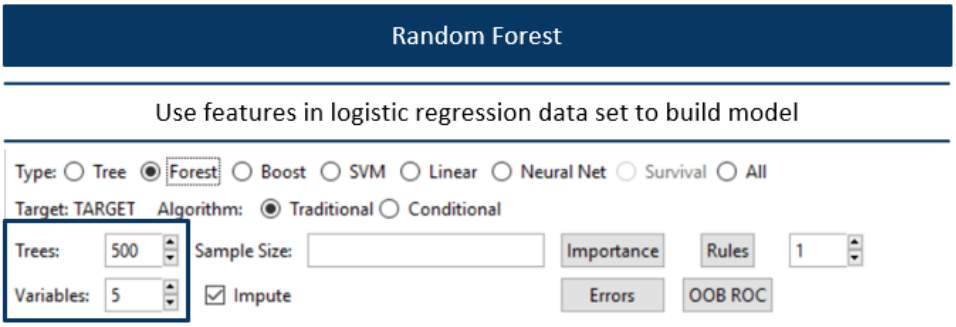
<u>Optimize ntree and mtry</u>

The number of trees grown (ntree) and number of predictors sampled for splitting at each node (mtry) were also optimized. The parameter values that minimised OOB error were ntree = 500 and mtry = 5.



4.6.6. Final Model

The final model was built features from the logistic regression data set and optimised parameters.

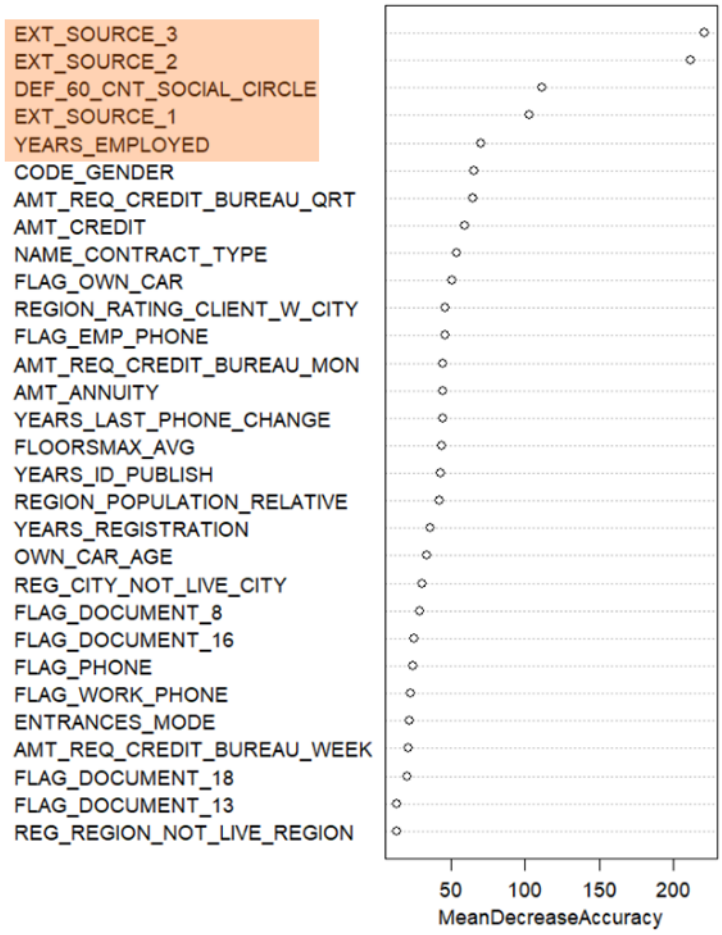


Random Forest	Train Set	Validation Set
Accuracy	100%	72.5%
Precision	100%	16.7%
Recall	100%	60.2%
F1 Score	100%	26.1%

Confusion matrix on validation set:

```
Error matrix for the Random Forest model on validation_a (counts):

      Predicted
Actual     0     1 Error
     0 61950 22297  26.5
     1  2952  4463  39.8

Error matrix for the Random Forest model on validation_a (proportions):

      Predicted
Actual    0    1 Error
     0 67.6 24.3  26.5
     1  3.2  4.9  39.8

Overall error: 27.5%, Averaged class error: 33.15%
```

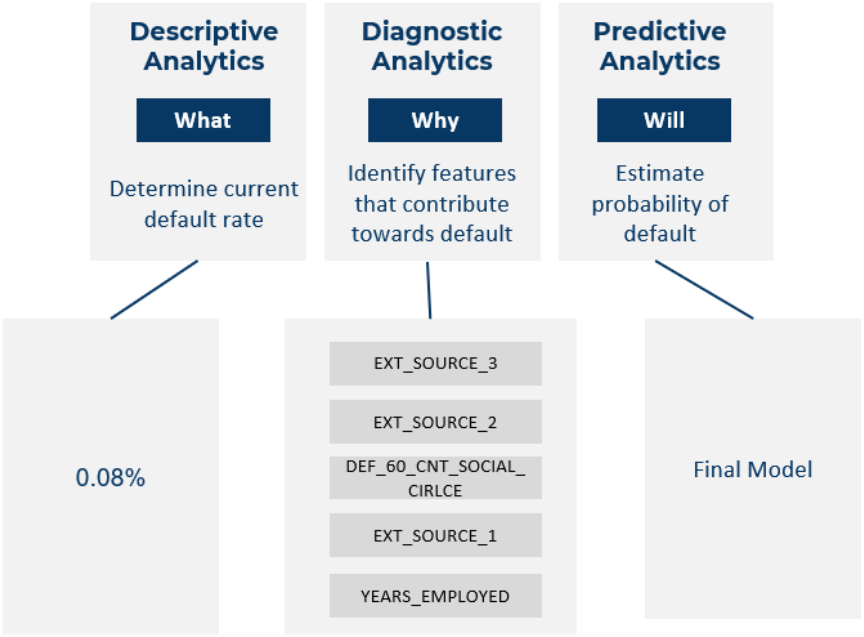
The performance of this optimized random forest model is the best among the previous models, with a recall of 60.2% and F1 score of 26.1%.



The top five features are (1) EXT_SOURCE_3, (2) EXT_SOURCE_2, (3) DEF_60_CNT_SOCIAL_CIRCLE, (4) EXT_SOURCE_1, (5) YEARS_EMPLOYED.

5. Evaluation

5.1. Results



The final model achieves all three objectives. From historical data, the observed default rate is 0.08%. The model identified important features that contribute towards default, the top 5 being EXT_SOURCE_3, EXT_SOURCE_2, DEF_60_CNT_SOCIAL_CIRCLE, EXT_SOURCE_1, and YEARS_EMPLOYED. Scoring the final model on future data will make predictions on whether a client defaults or not, allowing the FI to take mitigative action.

5.2. Challenges

Several challenges were faced during this project.

Firstly, computational limitations resulted in unnecessary delays. Model building and scoring was time consuming and computationally demanding. Secondly, vague data descriptions limited data understanding in the initial stages of the project. A deeper understanding of each column could have resulted in more meaningful manipulation of values. Lastly, high dimensionality posed some difficulty in building models – they were prone to overfitting and over complexity.

5.3. Improvements

Using more data in model building may be helpful. This project only considered loan application data, but credit bureau data previous loan data might add more insight. Another improvement is to experiment with a higher ntree value in the random forest model. This model only considered 500 due to computational constraints, but increasing it could potentially reduce error.

6. Conclusion

This project aims to build a predictive model using logistic regression, decision tree and random forest to predict loan default. The final model used a random forest algorithm, optimised in training data, ntree and mtry. It achieved an accuracy of 72.5%, precision of 16.7%, recall of 60.2% and F1 score of 26.1%. The top 5 features that contribute to loan default are EXT_SOURCE_3, EXT_SOURCE_2, DEF_60_CNT_SOCIAL_CIRCLE, EXT_SOURCE_1, and YEARS_EMPLOYED.

References

https://www.datascience-pm.com/crisp-dm-2/

https://www.analyticsvidhya.com/blog/2021/04/beginners-guide-to-decision-tree-classification-using-python/#h2_3

https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc

Appendix

1. Data Set Column Descriptions

Row	Description
SK_ID_CURR	ID of loan in our sample
TARGET	Target variable (1 - client with payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the loan in our sample, 0 - all other cases)
NAME_CONTRACT_TYPE	Identification if loan is cash or revolving
CODE_GENDER	Gender of the client
FLAG_OWN_CAR	Flag if the client owns a car
FLAG_OWN_REALTY	Flag if client owns a house or flat
CNT_CHILDREN	Number of children the client has
AMT_INCOME_TOTAL	Income of the client
AMT_CREDIT	Credit amount of the loan
AMT_ANNUITY	Loan annuity
AMT_GOODS_PRICE	For consumer loans it is the price of the goods for which the loan is given
NAME_TYPE_SUITE	Who was accompanying client when he was applying for the loan
NAME_INCOME_TYPE	Clients income type (businessman, working, maternity leave,…)
NAME_EDUCATION_TYPE	Level of highest education the client achieved
NAME_FAMILY_STATUS	Family status of the client
NAME_HOUSING_TYPE	What is the housing situation of the client (renting, living with parents, ...)
REGION_POPULATION_RELATIVE	Normalized population of region where client lives (higher number means the client lives in more populated region)
DAYS_BIRTH	Client's age in days at the time of application
DAYS_EMPLOYED	How many days before the application the person started current employment
DAYS_REGISTRATION	How many days before the application did client change his registration
DAYS_ID_PUBLISH	How many days before the application did client change the identity document with which he applied for the loan
OWN_CAR_AGE	Age of client's car
FLAG_MOBIL	Did client provide mobile phone (1=YES, 0=NO)
FLAG_EMP_PHONE	Did client provide work phone (1=YES, 0=NO)
FLAG_WORK_PHONE	Did client provide home phone (1=YES, 0=NO)

FLAG_CONT_MOBILE	Was mobile phone reachable (1=YES, 0=NO)
FLAG_PHONE	Did client provide home phone (1=YES, 0=NO)
FLAG_EMAIL	Did client provide email (1=YES, 0=NO)
OCCUPATION_TYPE	What kind of occupation does the client have
CNT_FAM_MEMBERS	How many family members does client have
REGION_RATING_CLIENT	Our rating of the region where client lives (1,2,3)
REGION_RATING_CLIENT_W_CITY	Our rating of the region where client lives with taking city into account (1,2,3)
WEEKDAY_APPR_PROCESS_START	On which day of the week did the client apply for the loan
HOUR_APPR_PROCESS_START	Approximately at what hour did the client apply for the loan
REG_REGION_NOT_LIVE_REGION	Flag if client's permanent address does not match contact address (1=different, 0=same, at region level)
REG_REGION_NOT_WORK_REGION	Flag if client's permanent address does not match work address (1=different, 0=same, at region level)
LIVE_REGION_NOT_WORK_REGION	Flag if client's contact address does not match work address (1=different, 0=same, at region level)
REG_CITY_NOT_LIVE_CITY	Flag if client's permanent address does not match contact address (1=different, 0=same, at city level)
REG_CITY_NOT_WORK_CITY	Flag if client's permanent address does not match work address (1=different, 0=same, at city level)
LIVE_CITY_NOT_WORK_CITY	Flag if client's contact address does not match work address (1=different, 0=same, at city level)
ORGANIZATION_TYPE	Type of organization where client works
EXT_SOURCE_1	Normalized score from external data source
EXT_SOURCE_2	Normalized score from external data source
EXT_SOURCE_3	Normalized score from external data source
APARTMENTS_AVG	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
BASEMENTAREA_AVG	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances,

	state of the building, number of floor
YEARS_BEGINEXPLUATATION_AVG	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
YEARS_BUILD_AVG	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
COMMONAREA_AVG	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
ELEVATORS_AVG	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
ENTRANCES_AVG	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
FLOORSMAX_AVG	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
FLOORSMIN_AVG	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
LANDAREA_AVG	Normalized information about building

	where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
LIVINGAPARTMENTS_AVG	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
LIVINGAREA_AVG	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
NONLIVINGAPARTMENTS_AVG	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
NONLIVINGAREA_AVG	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
APARTMENTS_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
BASEMENTAREA_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
YEARS_BEGINEXPLUATATION_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix),

	median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
YEARS_BUILD_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
COMMONAREA_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
ELEVATORS_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
ENTRANCES_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
FLOORSMAX_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
FLOORSMIN_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
LANDAREA_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building,

	number of elevators, number of entrances, state of the building, number of floor
LIVINGAPARTMENTS_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
LIVINGAREA_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
NONLIVINGAPARTMENTS_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
NONLIVINGAREA_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
APARTMENTS_MEDI	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
BASEMENTAREA_MEDI	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
YEARS_BEGINEXPLUATATION_MEDI	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor

YEARS_BUILD_MEDI	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
COMMONAREA_MEDI	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
ELEVATORS_MEDI	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
ENTRANCES_MEDI	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
FLOORSMAX_MEDI	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
FLOORSMIN_MEDI	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
LANDAREA_MEDI	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
LIVINGAPARTMENTS_MEDI	Normalized information about building where the client lives, What is average

	(_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
LIVINGAREA_MEDI	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
NONLIVINGAPARTMENTS_MEDI	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
NONLIVINGAREA_MEDI	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
FONDKAPREMONT_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
HOUSETYPE_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
TOTALAREA_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
WALLSMATERIAL_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size,

	common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
EMERGENCYSTATE_MODE	Normalized information about building where the client lives, What is average (_AVG suffix), modus (_MODE suffix), median (_MEDI suffix) apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floor
OBS_30_CNT_SOCIAL_CIRCLE	How many observation of client's social surroundings with observable 30 DPD (days past due) default
DEF_30_CNT_SOCIAL_CIRCLE	How many observation of client's social surroundings defaulted on 30 DPD (days past due)
OBS_60_CNT_SOCIAL_CIRCLE	How many observation of client's social surroundings with observable 60 DPD (days past due) default
DEF_60_CNT_SOCIAL_CIRCLE	How many observation of client's social surroundings defaulted on 60 (days past due) DPD
DAYS_LAST_PHONE_CHANGE	How many days before application did client change phone
FLAG_DOCUMENT_2	Did client provide document 2
FLAG_DOCUMENT_3	Did client provide document 3
FLAG_DOCUMENT_4	Did client provide document 4
FLAG_DOCUMENT_5	Did client provide document 5
FLAG_DOCUMENT_6	Did client provide document 6
FLAG_DOCUMENT_7	Did client provide document 7
FLAG_DOCUMENT_8	Did client provide document 8
FLAG_DOCUMENT_9	Did client provide document 9
FLAG_DOCUMENT_10	Did client provide document 10
FLAG_DOCUMENT_11	Did client provide document 11
FLAG_DOCUMENT_12	Did client provide document 12
FLAG_DOCUMENT_13	Did client provide document 13
FLAG_DOCUMENT_14	Did client provide document 14
FLAG_DOCUMENT_15	Did client provide document 15
FLAG_DOCUMENT_16	Did client provide document 16
FLAG_DOCUMENT_17	Did client provide document 17
FLAG_DOCUMENT_18	Did client provide document 18
FLAG_DOCUMENT_19	Did client provide document 19
FLAG_DOCUMENT_20	Did client provide document 20
FLAG_DOCUMENT_21	Did client provide document 21
AMT_REQ_CREDIT_BUREAU_HOUR	Number of enquiries to Credit Bureau about the client one hour before application
AMT_REQ_CREDIT_BUREAU_DAY	Number of enquiries to Credit Bureau about the client one day before application (excluding one hour before application)

AMT_REQ_CREDIT_BUREAU_WEEK	Number of enquiries to Credit Bureau about the client one week before application (excluding one day before application)
AMT_REQ_CREDIT_BUREAU_MON	Number of enquiries to Credit Bureau about the client one month before application (excluding one week before application)
AMT_REQ_CREDIT_BUREAU_QRT	Number of enquiries to Credit Bureau about the client 3 month before application (excluding one month before application)
AMT_REQ_CREDIT_BUREAU_YEAR	Number of enquiries to Credit Bureau about the client one day year (excluding last 3 months before application)

2. Logistic Regression Model 1: VIF

```
> vif(lr1)
                                GVIF Df GVIF^(1/(2*Df))
NAME_CONTRACT_TYPE          4.750014  1        2.179453
CODE_GENDER                 1.677649  1        1.295241
FLAG_OWN_CAR                1.314522  1        1.146526
FLAG_OWN_REALTY             1.146743  1        1.070861
AMT_INCOME_TOTAL            1.069970  1        1.034393
AMT_CREDIT                  2.472161  1        1.572311
AMT_ANNUITY                 2.587714  1        1.608637
NAME_TYPE_SUITE             1.063049  7        1.004377
NAME_INCOME_TYPE        31090404.481451  7        3.429159
NAME_EDUCATION_TYPE         1.263257  4        1.029643
NAME_HOUSING_TYPE           1.261920  5        1.023536
REGION_POPULATION_RELATIVE  1.565552  1        1.251220
CLIENT_AGE                  2.419795  1        1.555569
YEARS_EMPLOYED              1.245551  1        1.116043
YEARS_REGISTRATION          1.142272  1        1.068771
YEARS_ID_PUBLISH            1.147917  1        1.071409
OWN_CAR_AGE                 1.105162  1        1.051267
FLAG_EMP_PHONE           1968.096185  1       44.363230
FLAG_WORK_PHONE             1.247045  1        1.116712
FLAG_CONT_MOBILE            1.076890  1        1.037733
FLAG_PHONE                  1.152171  1        1.073392
FLAG_EMAIL                  1.033398  1        1.016562
OCCUPATION_TYPE            15.287082 18        1.078693
CNT_FAM_MEMBERS             1.139956  1        1.067687
REGION_RATING_CLIENT      190.016015  2        3.712766
REGION_RATING_CLIENT_W_CITY 180.511920  2        3.665443
WEEKDAY_APPR_PROCESS_START  1.027878  6        1.002294
HOUR_APPR_PROCESS_START     1.287876 23        1.005515
REG_REGION_NOT_LIVE_REGION  3.247551  1        1.802096
REG_REGION_NOT_WORK_REGION 12.137893  1        3.483948
LIVE_REGION_NOT_WORK_REGION 9.504357  1        3.082914
```

REG_CITY_NOT_LIVE_CITY	2.752893	1	1.659184
REG_CITY_NOT_WORK_CITY	7.216558	1	2.686365
LIVE_CITY_NOT_WORK_CITY	5.613121	1	2.369203
ORGANIZATION_TYPE	191054391.751262	57	1.182066
EXT_SOURCE_1	1.271194	1	1.127472
EXT_SOURCE_2	1.174974	1	1.083962
EXT_SOURCE_3	1.066905	1	1.032911
FLOORSMAX_AVG	1.119916	1	1.058261
ENTRANCES_MODE	1.017408	1	1.008666
OBS_60_CNT_SOCIAL_CIRCLE	1.114807	1	1.055844
DEF_60_CNT_SOCIAL_CIRCLE	1.108458	1	1.052833
YEARS_LAST_PHONE_CHANGE	1.119021	1	1.057838
FLAG_DOCUMENT_2	1.012362	1	1.006162
FLAG_DOCUMENT_3	13.179151	1	3.630310
FLAG_DOCUMENT_4	1.000001	1	1.000000
FLAG_DOCUMENT_5	1.838230	1	1.355813
FLAG_DOCUMENT_6	5.963673	1	2.442063
FLAG_DOCUMENT_7	1.000091	1	1.000046
FLAG_DOCUMENT_8	6.132763	1	2.476442
FLAG_DOCUMENT_9	1.212343	1	1.101064
FLAG_DOCUMENT_10	1.000000	1	1.000000
FLAG_DOCUMENT_11	1.140987	1	1.068170
FLAG_DOCUMENT_12	1.000000	1	1.000000
FLAG_DOCUMENT_13	1.013836	1	1.006894
FLAG_DOCUMENT_14	1.011866	1	1.005916
FLAG_DOCUMENT_15	1.004847	1	1.002421
FLAG_DOCUMENT_16	1.033782	1	1.016751
FLAG_DOCUMENT_17	1.002399	1	1.001199
FLAG_DOCUMENT_18	1.041414	1	1.020497
FLAG_DOCUMENT_19	1.005087	1	1.002540
FLAG_DOCUMENT_20	1.034484	1	1.017096
FLAG_DOCUMENT_21	1.012156	1	1.006060
AMT_REQ_CREDIT_BUREAU_HOUR	1.079839	1	1.039153
AMT_REQ_CREDIT_BUREAU_DAY	1.125784	1	1.061030
AMT_REQ_CREDIT_BUREAU_WEEK	1.051101	1	1.025232
AMT_REQ_CREDIT_BUREAU_MON	1.018335	1	1.009126
AMT_REQ_CREDIT_BUREAU_QRT	1.023065	1	1.011467
AMT_REQ_CREDIT_BUREAU_YEAR	1.070567	1	1.034682

3. Logistic Regression Model 2: VIF

```
> vif(lr2)
                                GVIF Df GVIF^(1/(2*Df))
NAME_CONTRACT_TYPE          4.455951  1        2.110912
CODE_GENDER                 1.622878  1        1.273922
FLAG_OWN_CAR                1.311189  1        1.145072
FLAG_OWN_REALTY             1.145595  1        1.070325
AMT_INCOME_TOTAL            1.064818  1        1.031900
AMT_CREDIT                  2.468658  1        1.571196
AMT_ANNUITY                 2.574855  1        1.604635
NAME_TYPE_SUITE             1.059253  7        1.004120
NAME_INCOME_TYPE         1777.564391  7        1.706595
NAME_EDUCATION_TYPE         1.245432  4        1.027815
NAME_HOUSING_TYPE           1.252478  5        1.022768
REGION_POPULATION_RELATIVE  1.561249  1        1.249499
CLIENT_AGE                  2.399013  1        1.548875
YEARS_EMPLOYED              1.202193  1        1.096445
YEARS_REGISTRATION          1.141466  1        1.068394
YEARS_ID_PUBLISH            1.147647  1        1.071283
OWN_CAR_AGE                 1.103993  1        1.050711
FLAG_EMP_PHONE           1492.111579  1       38.627860
FLAG_WORK_PHONE             1.243147  1        1.114965
FLAG_CONT_MOBILE            1.074951  1        1.036798
FLAG_PHONE                  1.151222  1        1.072950
FLAG_EMAIL                  1.032300  1        1.016022
OCCUPATION_TYPE             3.420291 18        1.034749
CNT_FAM_MEMBERS             1.136586  1        1.066108
REGION_RATING_CLIENT      189.298738  2        3.709257
REGION_RATING_CLIENT_W_CITY 179.843476  2        3.662045
WEEKDAY_APPR_PROCESS_START    1.023741  6        1.001957
HOUR_APPR_PROCESS_START       1.272512 23        1.005253
REG_REGION_NOT_LIVE_REGION    3.242990  1        1.800830
REG_REGION_NOT_WORK_REGION   12.105124  1        3.479242
LIVE_REGION_NOT_WORK_REGION   9.477818  1        3.078606
REG_CITY_NOT_LIVE_CITY        2.749619  1        1.658198
REG_CITY_NOT_WORK_CITY        7.202313  1        2.683712
LIVE_CITY_NOT_WORK_CITY       5.598606  1        2.366137
EXT_SOURCE_1                  1.269490  1        1.126717
EXT_SOURCE_2                  1.173293  1        1.083187
EXT_SOURCE_3                  1.064575  1        1.031783
FLOORSMAX_AVG                 1.117520  1        1.057128
ENTRANCES_MODE                1.016631  1        1.008281
OBS_60_CNT_SOCIAL_CIRCLE      1.115014  1        1.055942
DEF_60_CNT_SOCIAL_CIRCLE      1.109147  1        1.053160
YEARS_LAST_PHONE_CHANGE       1.118403  1        1.057546
FLAG_DOCUMENT_2               1.010635  1        1.005304
FLAG_DOCUMENT_3              12.189332  1        3.491322
FLAG_DOCUMENT_4               1.000001  1        1.000000
FLAG_DOCUMENT_5               1.770476  1        1.330592
```

FLAG_DOCUMENT_5	1.770476	1	1.330592
FLAG_DOCUMENT_6	5.630079	1	2.372779
FLAG_DOCUMENT_7	1.000064	1	1.000032
FLAG_DOCUMENT_8	5.727130	1	2.393142
FLAG_DOCUMENT_9	1.194721	1	1.093033
FLAG_DOCUMENT_10	1.000000	1	1.000000
FLAG_DOCUMENT_11	1.132625	1	1.064248
FLAG_DOCUMENT_12	1.000000	1	1.000000
FLAG_DOCUMENT_13	1.013663	1	1.006808
FLAG_DOCUMENT_14	1.011936	1	1.005950
FLAG_DOCUMENT_15	1.004694	1	1.002344
FLAG_DOCUMENT_16	1.033211	1	1.016470
FLAG_DOCUMENT_17	1.001436	1	1.000718
FLAG_DOCUMENT_18	1.038716	1	1.019174
FLAG_DOCUMENT_19	1.004360	1	1.002177
FLAG_DOCUMENT_20	1.033832	1	1.016776
FLAG_DOCUMENT_21	1.011131	1	1.005550
AMT_REQ_CREDIT_BUREAU_HOUR	1.078620	1	1.038566
AMT_REQ_CREDIT_BUREAU_DAY	1.123995	1	1.060186
AMT_REQ_CREDIT_BUREAU_WEEK	1.050418	1	1.024899
AMT_REQ_CREDIT_BUREAU_MON	1.016571	1	1.008252
AMT_REQ_CREDIT_BUREAU_QRT	1.022491	1	1.011183
AMT_REQ_CREDIT_BUREAU_YEAR	1.069091	1	1.033969

4. Logistic Regression Model 3: VIF

```
> vif(lr3)
                               GVIF Df GVIF^(1/(2*Df))
NAME_CONTRACT_TYPE         4.429532  1        2.104645
CODE_GENDER                1.621194  1        1.273261
FLAG_OWN_CAR               1.310739  1        1.144875
FLAG_OWN_REALTY            1.145348  1        1.070209
AMT_INCOME_TOTAL           1.051347  1        1.025352
AMT_CREDIT                 2.467430  1        1.570806
AMT_ANNUITY                2.567697  1        1.602404
NAME_TYPE_SUITE            1.058439  7        1.004065
NAME_EDUCATION_TYPE        1.240522  4        1.027308
NAME_HOUSING_TYPE          1.250347  5        1.022594
REGION_POPULATION_RELATIVE 1.561813  1        1.249725
CLIENT_AGE                 2.400115  1        1.549230
YEARS_EMPLOYED             1.194556  1        1.092958
YEARS_REGISTRATION         1.141566  1        1.068441
YEARS_ID_PUBLISH           1.147856  1        1.071381
OWN_CAR_AGE                1.103771  1        1.050605
FLAG_EMP_PHONE             3.128826  1        1.768849
FLAG_WORK_PHONE            1.236872  1        1.112147
FLAG_CONT_MOBILE           1.073766  1        1.036227
FLAG_PHONE                 1.151220  1        1.072949
FLAG_EMAIL                 1.031876  1        1.015813
OCCUPATION_TYPE            3.191177 18        1.032758
CNT_FAM_MEMBERS            1.135581  1        1.065637
```

CNT_FAM_MEMBERS	1.135581	1	1.065637
REGION_RATING_CLIENT	188.744416	2	3.706539
REGION_RATING_CLIENT_W_CITY	179.530782	2	3.660452
WEEKDAY_APPR_PROCESS_START	1.023425	6	1.001931
HOUR_APPR_PROCESS_START	1.269588	23	1.005202
REG_REGION_NOT_LIVE_REGION	3.241721	1	1.800478
REG_REGION_NOT_WORK_REGION	12.102439	1	3.478856
LIVE_REGION_NOT_WORK_REGION	9.475771	1	3.078274
REG_CITY_NOT_LIVE_CITY	2.748995	1	1.658009
REG_CITY_NOT_WORK_CITY	7.199663	1	2.683219
LIVE_CITY_NOT_WORK_CITY	5.594982	1	2.365372
EXT_SOURCE_1	1.269563	1	1.126749
EXT_SOURCE_2	1.173168	1	1.083129
EXT_SOURCE_3	1.064270	1	1.031635
FLOORSMAX_AVG	1.117139	1	1.056948
ENTRANCES_MODE	1.016506	1	1.008219
OBS_60_CNT_SOCIAL_CIRCLE	1.112776	1	1.054882
DEF_60_CNT_SOCIAL_CIRCLE	1.106798	1	1.052045
YEARS_LAST_PHONE_CHANGE	1.118361	1	1.057526
FLAG_DOCUMENT_2	1.010489	1	1.005231
FLAG_DOCUMENT_3	12.052339	1	3.471648
FLAG_DOCUMENT_4	1.000001	1	1.000000
FLAG_DOCUMENT_5	1.762331	1	1.327528
FLAG_DOCUMENT_6	5.574617	1	2.361063
FLAG_DOCUMENT_7	1.000001	1	1.000000
FLAG_DOCUMENT_8	5.673353	1	2.381880
FLAG_DOCUMENT_9	1.192714	1	1.092115
FLAG_DOCUMENT_10	1.000000	1	1.000000
FLAG_DOCUMENT_11	1.131506	1	1.063723
FLAG_DOCUMENT_12	1.000000	1	1.000000
FLAG_DOCUMENT_13	1.013379	1	1.006667
FLAG_DOCUMENT_14	1.011710	1	1.005838
FLAG_DOCUMENT_15	1.004685	1	1.002340
FLAG_DOCUMENT_16	1.032531	1	1.016135
FLAG_DOCUMENT_17	1.001499	1	1.000749
FLAG_DOCUMENT_18	1.038206	1	1.018924
FLAG_DOCUMENT_19	1.004342	1	1.002169
FLAG_DOCUMENT_20	1.033842	1	1.016780
FLAG_DOCUMENT_21	1.011112	1	1.005541
AMT_REQ_CREDIT_BUREAU_HOUR	1.078445	1	1.038482
AMT_REQ_CREDIT_BUREAU_DAY	1.124176	1	1.060272
AMT_REQ_CREDIT_BUREAU_WEEK	1.050895	1	1.025132
AMT_REQ_CREDIT_BUREAU_MON	1.016140	1	1.008038
AMT_REQ_CREDIT_BUREAU_QRT	1.022444	1	1.011160
AMT_REQ_CREDIT_BUREAU_YEAR	1.068821	1	1.033838

5. Logistic Regression Model 4: VIF

```
> vif(lr4)
                                GVIF Df GVIF^(1/(2*Df))
NAME_CONTRACT_TYPE          4.430728  1        2.104929
CODE_GENDER                 1.621027  1        1.273196
FLAG_OWN_CAR                1.310253  1        1.144663
FLAG_OWN_REALTY             1.145355  1        1.070213
AMT_INCOME_TOTAL            1.050328  1        1.024855
AMT_CREDIT                  2.466812  1        1.570609
AMT_ANNUITY                 2.565324  1        1.601663
NAME_TYPE_SUITE             1.058212  7        1.004050
NAME_EDUCATION_TYPE         1.240097  4        1.027264
NAME_HOUSING_TYPE           1.248629  5        1.022453
REGION_POPULATION_RELATIVE  1.484587  1        1.218436
CLIENT_AGE                  2.400525  1        1.549363
YEARS_EMPLOYED              1.194517  1        1.092939
YEARS_REGISTRATION          1.139792  1        1.067611
YEARS_ID_PUBLISH            1.147991  1        1.071443
OWN_CAR_AGE                 1.103458  1        1.050456
FLAG_EMP_PHONE              3.128733  1        1.768822
FLAG_WORK_PHONE             1.236853  1        1.112139
FLAG_CONT_MOBILE            1.073595  1        1.036144
FLAG_PHONE                  1.150851  1        1.072777
FLAG_EMAIL                  1.031294  1        1.015526
OCCUPATION_TYPE             3.190154 18        1.032749
CNT_FAM_MEMBERS             1.135498  1        1.065598
REGION_RATING_CLIENT_W_CITY 1.751619  2        1.150429
WEEKDAY_APPR_PROCESS_START  1.023320  6        1.001923
HOUR_APPR_PROCESS_START     1.245506 23        1.004784
HOUR_APPR_PROCESS_START     1.245506 23        1.004784
REG_REGION_NOT_LIVE_REGION  3.245814  1        1.801614
REG_REGION_NOT_WORK_REGION 12.119246  1        3.481271
LIVE_REGION_NOT_WORK_REGION 9.487719  1        3.080214
REG_CITY_NOT_LIVE_CITY      2.750164  1        1.658362
REG_CITY_NOT_WORK_CITY      7.199986  1        2.683279
LIVE_CITY_NOT_WORK_CITY     5.596833  1        2.365763
EXT_SOURCE_1                1.269361  1        1.126659
EXT_SOURCE_2                1.169821  1        1.081583
EXT_SOURCE_3                1.064154  1        1.031579
FLOORSMAX_AVG               1.114982  1        1.055927
ENTRANCES_MODE              1.016203  1        1.008069
OBS_60_CNT_SOCIAL_CIRCLE    1.113289  1        1.055125
DEF_60_CNT_SOCIAL_CIRCLE    1.107924  1        1.052580
YEARS_LAST_PHONE_CHANGE     1.118208  1        1.057454
FLAG_DOCUMENT_2             1.010399  1        1.005186
FLAG_DOCUMENT_3            12.057126  1        3.472337
FLAG_DOCUMENT_4             1.000001  1        1.000000
FLAG_DOCUMENT_5             1.762770  1        1.327693
FLAG_DOCUMENT_6             5.576032  1        2.361362
FLAG_DOCUMENT_7             1.000001  1        1.000000
FLAG_DOCUMENT_8             5.674276  1        2.382074
FLAG_DOCUMENT_9             1.192599  1        1.092062
FLAG_DOCUMENT_10            1.000000  1        1.000000
FLAG_DOCUMENT_11            1.131228  1        1.063592
FLAG_DOCUMENT_12            1.000000  1        1.000000
FLAG_DOCUMENT_13            1.013345  1        1.006650
```

FLAG_DOCUMENT_14	1.011640	1	1.005803
FLAG_DOCUMENT_15	1.004654	1	1.002324
FLAG_DOCUMENT_16	1.032464	1	1.016103
FLAG_DOCUMENT_17	1.001512	1	1.000756
FLAG_DOCUMENT_18	1.038201	1	1.018922
FLAG_DOCUMENT_19	1.004282	1	1.002139
FLAG_DOCUMENT_20	1.033858	1	1.016788
FLAG_DOCUMENT_21	1.011109	1	1.005539
AMT_REQ_CREDIT_BUREAU_HOUR	1.078464	1	1.038491
AMT_REQ_CREDIT_BUREAU_DAY	1.124114	1	1.060242
AMT_REQ_CREDIT_BUREAU_WEEK	1.050807	1	1.025089
AMT_REQ_CREDIT_BUREAU_MON	1.015994	1	1.007965
AMT_REQ_CREDIT_BUREAU_QRT	1.022432	1	1.011154
AMT_REQ_CREDIT_BUREAU_YEAR	1.068751	1	1.033804

6. Logistic Regression Model 5: VIF

```
> vif(lr5)
                                GVIF Df GVIF^(1/(2*Df))
NAME_CONTRACT_TYPE          4.430989  1        2.104991
CODE_GENDER                 1.620961  1        1.273170
FLAG_OWN_CAR                1.310247  1        1.144660
FLAG_OWN_REALTY             1.145357  1        1.070214
AMT_INCOME_TOTAL            1.050752  1        1.025062
AMT_CREDIT                  2.466806  1        1.570607
AMT_ANNUITY                 2.565250  1        1.601640
NAME_TYPE_SUITE             1.058176  7        1.004047
NAME_EDUCATION_TYPE         1.240116  4        1.027266
NAME_HOUSING_TYPE           1.248487  5        1.022441
REGION_POPULATION_RELATIVE  1.483562  1        1.218015
CLIENT_AGE                  2.400424  1        1.549330
YEARS_EMPLOYED              1.194389  1        1.092881
YEARS_REGISTRATION          1.139779  1        1.067604
YEARS_ID_PUBLISH            1.147987  1        1.071441
OWN_CAR_AGE                 1.103459  1        1.050457
FLAG_EMP_PHONE              3.128487  1        1.768753
FLAG_WORK_PHONE             1.236703  1        1.112071
FLAG_CONT_MOBILE            1.073571  1        1.036133
FLAG_PHONE                  1.150847  1        1.072775
FLAG_EMAIL                  1.031288  1        1.015523
OCCUPATION_TYPE             3.188647 18        1.032735
CNT_FAM_MEMBERS             1.135483  1        1.065590
REGION_RATING_CLIENT_W_CITY 1.750819  2        1.150298
WEEKDAY_APPR_PROCESS_START  1.023275  6        1.001919
HOUR_APPR_PROCESS_START     1.245084 23        1.004777
REG_REGION_NOT_LIVE_REGION  1.141704  1        1.068505
```

LIVE_REGION_NOT_WORK_REGION	1.119653	1	1.058137
REG_CITY_NOT_LIVE_CITY	2.678938	1	1.636746
REG_CITY_NOT_WORK_CITY	6.874906	1	2.622004
LIVE_CITY_NOT_WORK_CITY	5.375054	1	2.318416
EXT_SOURCE_1	1.269298	1	1.126631
EXT_SOURCE_2	1.169810	1	1.081577
EXT_SOURCE_3	1.064126	1	1.031565
FLOORSMAX_AVG	1.114932	1	1.055903
ENTRANCES_MODE	1.016177	1	1.008056
OBS_60_CNT_SOCIAL_CIRCLE	1.113282	1	1.055122
DEF_60_CNT_SOCIAL_CIRCLE	1.107930	1	1.052583
YEARS_LAST_PHONE_CHANGE	1.118185	1	1.057442
FLAG_DOCUMENT_2	1.010398	1	1.005186
FLAG_DOCUMENT_3	12.057955	1	3.472457
FLAG_DOCUMENT_4	1.000001	1	1.000000
FLAG_DOCUMENT_5	1.762791	1	1.327701
FLAG_DOCUMENT_6	5.576301	1	2.361419
FLAG_DOCUMENT_7	1.000001	1	1.000000
FLAG_DOCUMENT_8	5.674638	1	2.382150
FLAG_DOCUMENT_9	1.192588	1	1.092057
FLAG_DOCUMENT_10	1.000000	1	1.000000
FLAG_DOCUMENT_11	1.130486	1	1.063243
FLAG_DOCUMENT_12	1.000000	1	1.000000
FLAG_DOCUMENT_13	1.013328	1	1.006642
FLAG_DOCUMENT_14	1.011637	1	1.005802
FLAG_DOCUMENT_15	1.004654	1	1.002324
FLAG_DOCUMENT_16	1.032463	1	1.016102
FLAG_DOCUMENT_17	1.001512	1	1.000756
FLAG_DOCUMENT_18	1.038194	1	1.018918
FLAG_DOCUMENT_19	1.004282	1	1.002139
FLAG_DOCUMENT_20	1.033857	1	1.016788
FLAG_DOCUMENT_21	1.011109	1	1.005539
AMT_REQ_CREDIT_BUREAU_HOUR	1.078478	1	1.038498
AMT_REQ_CREDIT_BUREAU_DAY	1.124096	1	1.060234
AMT_REQ_CREDIT_BUREAU_WEEK	1.050758	1	1.025065
AMT_REQ_CREDIT_BUREAU_MON	1.015998	1	1.007967
AMT_REQ_CREDIT_BUREAU_QRT	1.022429	1	1.011152
AMT_REQ_CREDIT_BUREAU_YEAR	1.068744	1	1.033801

7. Logistic Regression Model 6: VIF

```
> vif(lr6)
                                GVIF Df GVIF^(1/(2*Df))
NAME_CONTRACT_TYPE          1.122675  1        1.059564
CODE_GENDER                 1.617119  1        1.271660
FLAG_OWN_CAR                1.310085  1        1.144590
FLAG_OWN_REALTY             1.145424  1        1.070245
AMT_INCOME_TOTAL            1.051529  1        1.025441
AMT_CREDIT                  2.457164  1        1.567534
AMT_ANNUITY                 2.558075  1        1.599398
NAME_TYPE_SUITE             1.058070  7        1.004040
NAME_EDUCATION_TYPE         1.239918  4        1.027245
NAME_HOUSING_TYPE           1.248701  5        1.022459
REGION_POPULATION_RELATIVE  1.483410  1        1.217953
CLIENT_AGE                  2.399121  1        1.548910
YEARS_EMPLOYED              1.194735  1        1.093039
YEARS_REGISTRATION          1.139775  1        1.067603
YEARS_ID_PUBLISH            1.148302  1        1.071589
OWN_CAR_AGE                 1.103407  1        1.050432
FLAG_EMP_PHONE              3.127042  1        1.768344
FLAG_WORK_PHONE             1.236085  1        1.111794
FLAG_CONT_MOBILE            1.022242  1        1.011060
FLAG_PHONE                  1.150088  1        1.072422
FLAG_EMAIL                  1.031225  1        1.015493
OCCUPATION_TYPE             3.175536 18        1.032617
CNT_FAM_MEMBERS             1.134745  1        1.065244
REGION_RATING_CLIENT_W_CITY 1.750884  2        1.150308
WEEKDAY_APPR_PROCESS_START  1.023318  6        1.001923
```

HOUR_APPR_PROCESS_START	1.244890	23	1.004773
REG_REGION_NOT_LIVE_REGION	1.140520	1	1.067951
LIVE_REGION_NOT_WORK_REGION	1.119676	1	1.058147
REG_CITY_NOT_LIVE_CITY	2.678066	1	1.636480
REG_CITY_NOT_WORK_CITY	6.870044	1	2.621077
LIVE_CITY_NOT_WORK_CITY	5.371533	1	2.317657
EXT_SOURCE_1	1.269312	1	1.126637
EXT_SOURCE_2	1.169903	1	1.081620
EXT_SOURCE_3	1.064060	1	1.031533
FLOORSMAX_AVG	1.115030	1	1.055950
ENTRANCES_MODE	1.016167	1	1.008051
OBS_60_CNT_SOCIAL_CIRCLE	1.113294	1	1.055127
DEF_60_CNT_SOCIAL_CIRCLE	1.107931	1	1.052583
YEARS_LAST_PHONE_CHANGE	1.117589	1	1.057161
FLAG_DOCUMENT_2	1.001811	1	1.000905
FLAG_DOCUMENT_4	1.000000	1	1.000000
FLAG_DOCUMENT_5	1.011485	1	1.005726
FLAG_DOCUMENT_6	1.597043	1	1.263742
FLAG_DOCUMENT_7	1.000001	1	1.000000
FLAG_DOCUMENT_8	1.124661	1	1.060500
FLAG_DOCUMENT_9	1.006537	1	1.003263
FLAG_DOCUMENT_10	1.000000	1	1.000000
FLAG_DOCUMENT_11	1.029484	1	1.014635
FLAG_DOCUMENT_12	1.000000	1	1.000000
FLAG_DOCUMENT_13	1.012848	1	1.006403
FLAG_DOCUMENT_14	1.010733	1	1.005352
FLAG_DOCUMENT_15	1.004585	1	1.002290
FLAG_DOCUMENT_16	1.031978	1	1.015863
FLAG_DOCUMENT_17	1.001487	1	1.000743
FLAG_DOCUMENT_18	1.037388	1	1.018523
FLAG_DOCUMENT_19	1.004190	1	1.002093
FLAG_DOCUMENT_20	1.033682	1	1.016701
FLAG_DOCUMENT_21	1.010896	1	1.005433
AMT_REQ_CREDIT_BUREAU_HOUR	1.077752	1	1.038148
AMT_REQ_CREDIT_BUREAU_DAY	1.123754	1	1.060073
AMT_REQ_CREDIT_BUREAU_WEEK	1.051249	1	1.025304
AMT_REQ_CREDIT_BUREAU_MON	1.015923	1	1.007930
AMT_REQ_CREDIT_BUREAU_QRT	1.022454	1	1.011165
AMT_REQ_CREDIT_BUREAU_YEAR	1.068584	1	1.033724

8. Logistic Regression Model 7: VIF

```
> vif(lr7)
                                GVIF Df GVIF^(1/(2*Df))
NAME_CONTRACT_TYPE          1.122670  1        1.059561
CODE_GENDER                 1.616983  1        1.271607
FLAG_OWN_CAR                1.310071  1        1.144583
FLAG_OWN_REALTY             1.145440  1        1.070252
AMT_INCOME_TOTAL            1.051655  1        1.025502
AMT_CREDIT                  2.457081  1        1.567508
AMT_ANNUITY                 2.558023  1        1.599382
NAME_TYPE_SUITE             1.058030  7        1.004037
NAME_EDUCATION_TYPE         1.239901  4        1.027243
NAME_HOUSING_TYPE           1.244145  5        1.022085
REGION_POPULATION_RELATIVE  1.483323  1        1.217917
CLIENT_AGE                  2.398591  1        1.548739
YEARS_EMPLOYED              1.192977  1        1.092235
YEARS_REGISTRATION          1.139714  1        1.067574
YEARS_ID_PUBLISH            1.148196  1        1.071539
OWN_CAR_AGE                 1.103399  1        1.050428
FLAG_EMP_PHONE              3.114354  1        1.764753
FLAG_WORK_PHONE             1.235983  1        1.111748
FLAG_CONT_MOBILE            1.022245  1        1.011061
FLAG_PHONE                  1.150023  1        1.072391
FLAG_EMAIL                  1.031221  1        1.015491
OCCUPATION_TYPE             3.172949 18        1.032594
CNT_FAM_MEMBERS             1.134012  1        1.064900
REGION_RATING_CLIENT_W_CITY 1.750207  2        1.150197
WEEKDAY_APPR_PROCESS_START  1.023233  6        1.001916
HOUR_APPR_PROCESS_START     1.244540 23        1.004767
REG_REGION_NOT_LIVE_REGION  1.140372  1        1.067882
```

LIVE_REGION_NOT_WORK_REGION	1.118598	1	1.057638
REG_CITY_NOT_LIVE_CITY	1.205062	1	1.097753
LIVE_CITY_NOT_WORK_CITY	1.150372	1	1.072554
EXT_SOURCE_1	1.269060	1	1.126526
EXT_SOURCE_2	1.169852	1	1.081597
EXT_SOURCE_3	1.064010	1	1.031509
FLOORSMAX_AVG	1.113161	1	1.055064
ENTRANCES_MODE	1.016172	1	1.008054
OBS_60_CNT_SOCIAL_CIRCLE	1.113336	1	1.055147
DEF_60_CNT_SOCIAL_CIRCLE	1.107978	1	1.052606
YEARS_LAST_PHONE_CHANGE	1.117531	1	1.057133
FLAG_DOCUMENT_2	1.001810	1	1.000905
FLAG_DOCUMENT_4	1.000000	1	1.000000
FLAG_DOCUMENT_5	1.011485	1	1.005726
FLAG_DOCUMENT_6	1.597021	1	1.263733
FLAG_DOCUMENT_7	1.000001	1	1.000000
FLAG_DOCUMENT_8	1.124623	1	1.060483
FLAG_DOCUMENT_9	1.006535	1	1.003262
FLAG_DOCUMENT_10	1.000000	1	1.000000
FLAG_DOCUMENT_11	1.029243	1	1.014516
FLAG_DOCUMENT_12	1.000000	1	1.000000
FLAG_DOCUMENT_13	1.012836	1	1.006398
FLAG_DOCUMENT_14	1.010737	1	1.005354
FLAG_DOCUMENT_15	1.004572	1	1.002283
FLAG_DOCUMENT_16	1.031953	1	1.015851
FLAG_DOCUMENT_17	1.001490	1	1.000745
FLAG_DOCUMENT_18	1.037377	1	1.018517
FLAG_DOCUMENT_19	1.004190	1	1.002093
FLAG_DOCUMENT_20	1.033681	1	1.016701
FLAG_DOCUMENT_21	1.010823	1	1.005397
AMT_REQ_CREDIT_BUREAU_HOUR	1.077737	1	1.038141
AMT_REQ_CREDIT_BUREAU_DAY	1.123747	1	1.060069
AMT_REQ_CREDIT_BUREAU_WEEK	1.051219	1	1.025290
AMT_REQ_CREDIT_BUREAU_MON	1.015922	1	1.007930
AMT_REQ_CREDIT_BUREAU_QRT	1.022444	1	1.011160
AMT_REQ_CREDIT_BUREAU_YEAR	1.068576	1	1.033719

9. Logistic Regression Model 8

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 84192  7362
         1    55    53

               Accuracy : 0.9191
                 95% CI : (0.9173, 0.9208)
    No Information Rate : 0.9191
    P-Value [Acc > NIR] : 0.5128

                  Kappa : 0.0118

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

            Sensitivity : 0.0071477
            Specificity : 0.9993472
         Pos Pred Value : 0.4907407
         Neg Pred Value : 0.9195884
             Prevalence : 0.0808950
         Detection Rate : 0.0005782
   Detection Prevalence : 0.0011782
      Balanced Accuracy : 0.5032474

       'Positive' Class : 1
```

10. Post-ROSE Logistic Regression Output

```
Call:
glm(formula = TARGET ~ ., family = binomial, data = train_a_lr.rose)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-2.979  -1.029   0.351   1.034   2.606

Coefficients:
                                    Estimate     Std. Error z value Pr(>|z|)
(Intercept)                      2.90327600145  0.11427832900  25.405  < 2e-16 ***
NAME_CONTRACT_TYPERevolving loans -0.57608187200  0.01864351534 -30.900  < 2e-16 ***
CODE_GENDERM                      0.42731328395  0.01063319017  40.187  < 2e-16 ***
FLAG_OWN_CARY                    -0.33809141013  0.01118979229 -30.214  < 2e-16 ***
AMT_CREDIT                       -0.00000005301  0.00000001356  -3.910 9.23e-05 ***
AMT_ANNUITY                       0.00000109967  0.00000038251   2.875  0.00404 **
REGION_POPULATION_RELATIVE        1.03732062276  0.37192337576   2.789  0.00529 **
YEARS_EMPLOYED                   -0.02111855584  0.00079086351 -26.703  < 2e-16 ***
YEARS_REGISTRATION               -0.00331717925  0.00045048772  -7.364 1.79e-13 ***
YEARS_ID_PUBLISH                 -0.01411685574  0.00102273595 -13.803  < 2e-16 ***
OWN_CAR_AGE                       0.00586025636  0.00059062768   9.922  < 2e-16 ***
FLAG_EMP_PHONE1                   0.28583502747  0.01456727910  19.622  < 2e-16 ***
FLAG_WORK_PHONE1                  0.12640840897  0.01251525682  10.100  < 2e-16 ***
FLAG_CONT_MOBILE1                -0.60389377373  0.10726815743  -5.630 1.80e-08 ***
FLAG_PHONE1                      -0.11402466994  0.01133266459 -10.062  < 2e-16 ***
REGION_RATING_CLIENT_W_CITY2      0.18911528326  0.01972265208   9.589  < 2e-16 ***
REGION_RATING_CLIENT_W_CITY3      0.39772044617  0.02343299200  16.973  < 2e-16 ***
REG_REGION_NOT_LIVE_REGION1      -0.18662240503  0.04040460192  -4.619 3.86e-06 ***
REG_CITY_NOT_LIVE_CITY1           0.17052275700  0.01742015617   9.789  < 2e-16 ***
EXT_SOURCE_1                     -1.11034330706  0.03073537136 -36.126  < 2e-16 ***
EXT_SOURCE_2                     -1.66751980580  0.02157388501 -77.293  < 2e-16 ***
EXT_SOURCE_3                     -2.15172847797  0.02322242148 -92.657  < 2e-16 ***
```

```
FLOORSMAX_AVG                -0.47344117164  0.04338411293 -10.913  < 2e-16 ***
ENTRANCES_MODE               -0.25831655822  0.06072161606  -4.254 2.10e-05 ***
DEF_60_CNT_SOCIAL_CIRCLE      0.14115471112  0.01045616611  13.500  < 2e-16 ***
YEARS_LAST_PHONE_CHANGE      -0.03196383411  0.00193756780 -16.497  < 2e-16 ***
FLAG_DOCUMENT_21              1.85058688363  0.46844145631   3.951 7.80e-05 ***
FLAG_DOCUMENT_81             -0.16622428108  0.01845979846  -9.005  < 2e-16 ***
FLAG_DOCUMENT_111            -0.52703329859  0.08257126443  -6.383 1.74e-10 ***
FLAG_DOCUMENT_131            -0.84550117554  0.10539895542  -8.022 1.04e-15 ***
FLAG_DOCUMENT_141            -0.64630211396  0.10487764984  -6.162 7.16e-10 ***
FLAG_DOCUMENT_161            -0.68182146483  0.05564611370 -12.253  < 2e-16 ***
FLAG_DOCUMENT_181            -0.54226174090  0.05733267044  -9.458  < 2e-16 ***
AMT_REQ_CREDIT_BUREAU_WEEK   -0.07287688989  0.02223730937  -3.277  0.00105 **
AMT_REQ_CREDIT_BUREAU_MON    -0.03940185053  0.00539370360  -7.305 2.77e-13 ***
AMT_REQ_CREDIT_BUREAU_QRT    -0.06340776324  0.00670743281  -9.453  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 296499  on 213878  degrees of freedom
Residual deviance: 262950  on 213843  degrees of freedom
AIC: 263022

Number of Fisher Scoring iterations: 4
```

11. Post-ROSE Logistic Regression Feature Importance (Standardised Coefficients)

ordered_std	FEATURES
-0.459892260	EXT_SOURCE_3
-0.393492509	EXT_SOURCE_2
-0.180098804	EXT_SOURCE_1
-0.158484164	FLAG_OWN_CARY
-0.154045065	NAME_CONTRACT_TYPERevolving.loans
-0.131992915	YEARS_EMPLOYED
-0.079887645	YEARS_LAST_PHONE_CHANGE
-0.067101462	YEARS_ID_PUBLISH
-0.059691003	FLAG_DOCUMENT_161
-0.053733145	FLOORSMAX_AVG
-0.050488681	FLAG_PHONE1
-0.045102538	FLAG_DOCUMENT_181
-0.044539959	AMT_REQ_CREDIT_BUREAU_QRT

-0.044481733	FLAG_DOCUMENT_81
-0.042234192	FLAG_DOCUMENT_131
-0.035707884	YEARS_REGISTRATION
-0.035417280	AMT_REQ_CREDIT_BUREAU_MON
-0.030678573	FLAG_DOCUMENT_141
-0.030442713	FLAG_DOCUMENT_111
-0.027206961	FLAG_CONT_MOBILE1
-0.023423034	REG_REGION_NOT_LIVE_REGION1
-0.022912579	AMT_CREDIT
-0.020083928	ENTRANCES_MODE
-0.015427111	AMT_REQ_CREDIT_BUREAU_WEEK
0.006445376	(Intercep
0.015376999	REGION_POPULATION_RELATIVE

0.017080860	AMT_ANNUITY
0.021172737	FLAG_DOCUMENT_21
0.048024375	OWN_CAR_AGE
0.050028001	REG_CITY_NOT_LIVE_CITY1
0.052056933	FLAG_WORK_PHONE1
0.064861083	DEF_60_CNT_SOCIAL_CIRCLE
0.082934463	REGION_RATING_CLIENT_W_CITY2
0.103130556	FLAG_EMP_PHONE1
0.149492807	REGION_RATING_CLIENT_W_CITY3
0.207659872	CODE_GENDERM