# HOME CREDIT DEFAULT RISK

Prepared by: Ravi Shankar

## OBJECTIVE

Home Credit strives to broaden its horizon of providing credit to those people who are currently unbanked and find difficulty in getting credits. Though they might be capable of repaying their credits. To point out those peoples and their economic capabilities, Home Credit is currently using various machine learning methods to make these predictions and analysis.

## DATA SETS

1. **sample_application_train.csv Dataset**

   This contains details related to current application. Every applicant has unique SK_ID_CURR and all previous history of applicants is attached to this unique Id. Actual data set consist of 30 k observation, but it is sampled down to 10k observations. One row belongs to one loan application /SK_CURR_ID.

| | |
|---|---|
| Number of positive Target | 850 |
| Number of Negative Target | 9150 |

| | |
|---|---|
| Number of observations | 10000 |
| Number of Variables | 122 |
| Unique Key | SK_ID_CURR |
| Non-risky application (TARGET = 0) | 9150 |
| Risky application (TARGET = 1) | 850 |

2. **previous_application.csv Dataset**

   This contains details of every previous loan application applied by the person and represented by SK_PREV_ID. Every previous application has linked to unique current SK_ID_CURR belong to that person. All details of the previous application are mentioned in this data set. One row belongs to each previously applied loan application. 9480 People from our sample data has previous credit history.

| | |
|---|---|
| Number of observations | 45799 |
| Number of Variables | 37 |
| Unique application id | SK_ID_PREV |
| Unique application id belongs to | SK_ID_CURR |
| People have Previous history | 9480 |

3. **POS_CASH_balance.csv Dataset**

   This contains monthly balance snapshots of all previous point of sales and cash loans that the applicant had with Home Credit. This table has one row for each month of history of every previous credit with Home Credit.

| | |
|---|---|
| Number of observations | 283156 |
| Number of Variables | 8 |
| Row reference to | SK_ID_PRE |
| Unique SK_ID_PRE | 26351 |
| Unique SK_ID_CURR | 9422 |

4. **installments_payments.csv Dataset**

This contains instalment history for the previous credits in Home. There is one row for every payment and missed payment

| | |
|---|---|
| Number of observations | 387473 |
| Number of variables | 7 |
| Row reference to | SK_ID_PRE |
| Unique SK_ID_CURR | 9513 |

5. **credit_card_balance.csv Dataset**

This contains monthly balance of previous credit cards and all transactions that the applicant had. One row for every credit card transaction.

| | |
|---|---|
| Number of observations | 108550 |
| Number of variables | 23 |
| Row reference to | SK_ID_PRE |
| Unique SK_ID_PREV | 2879 |
| Unique SK_ID_CURR | 2861 |

6. **bureau.csv Dataset**

This contains all client's previous credits history with Credit Bureau. 8617 applicants have previous history with other credit banks out of 10000 sample applicants.

| | |
|---|---|
| Number of observations | 48368 |
| Number of Variables | 17 |
| Unique SK_ID_Bureau | 48368 |
| Unique SK_ID_CURR | 8617 |

7. **bureau_balance.csv Dataset**

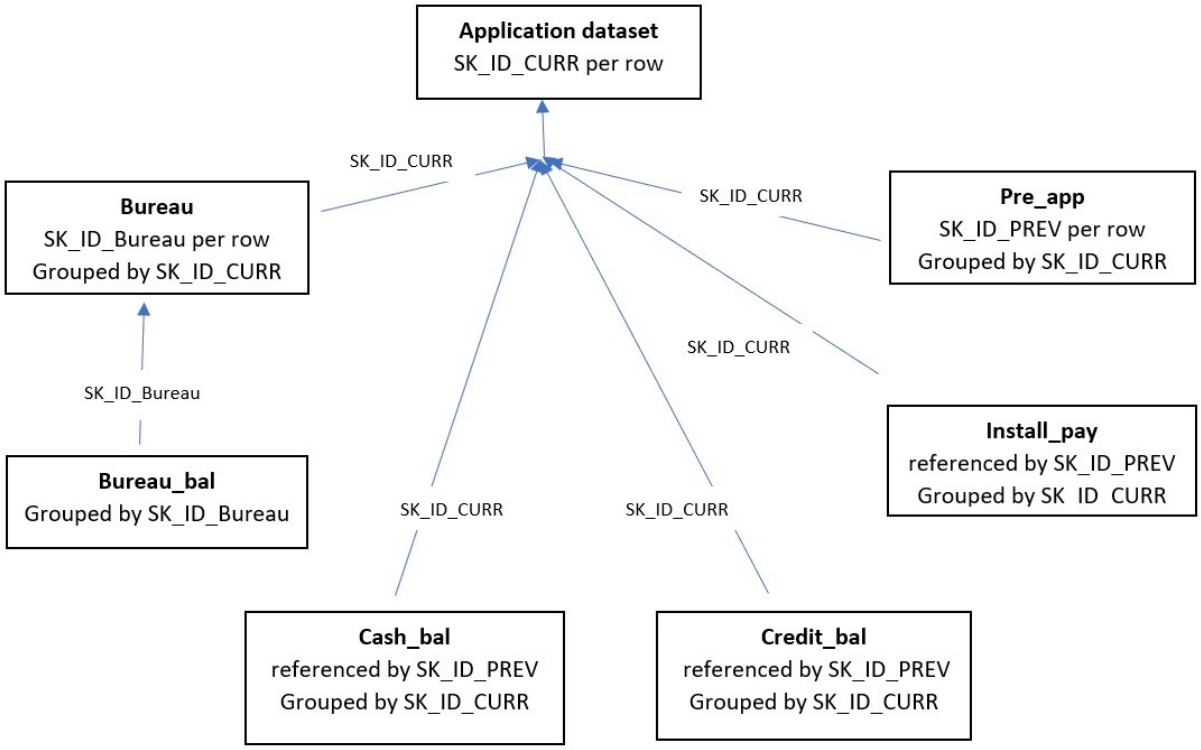
This contains monthly balances of previous credits with Credit Bureau. One row for each month of history of every previous credit with Credit Bureau. Every balance data is attached with its respective SK_ID_Bureau.

| | |
|---|---|
| Number of observations | 479014 |

| | |
|---|---|
| Number of Variables | 3 |
| Row reference to | SK_ID_Bureau |
| Unique SK_ID_Bureau | 17013 |

## DATA PREPARATION AND VARIABLE CREATION

All data preparation is done individually at table level and then merged according the requirement. Baseline for data preparation is shown in figure below.




## CHARACTER VARIABLE MANIPULATION

1. Revalue all the values of character variables
   - To easily detect source of the variables
   - To Avoid clash between variables of other datasets
   - To counter long and special characters part of the values of the character variables
2. Converting all character variables into dummy variables
   - Package used is fastdummies

## NUMERICAL VARIABLES / VARIABLE CREATION

Our main table must contain SK_CURR_ID per row . All related tables have multiple row of SK_CURR_ID attached with its respective SK_PRE_ID and SK_BUREAU_ID. So, we have

to summarise every table in the group by SK_CURR_ID . Summarization function used are maximum, minimum, mean, sum, standard deviation and distinct count.

Few other variables are also created. All relevant variables informations are mentioned in variables dictionary respectively.

## FEATURE SELECTION

Fisher score is used as feature selection method. After arranging variables in decreasing fisher score and then subsetting top N variables which are giving good results. In this case, I got N = 40. All required information about the variables are mentioned in Variables dictionary.

The broad category of selected variables are

1. Good price against which loan was granted
2. Rejection of previous applications
3. Days difference between application to the bureau credit and home credits.
4. Updates from bureau about credit end dates
5. Information provided by an external party
6. The education level of applicants
7. The way of obtaining the applicant

## BASE TABLE

Base table is exported as basetable.csv file and submitted. **All NAs are replaced by mean of the respective columns.**

| | |
|---|---|
| Number of observations | 10000 |
| Number of Variables | 40 |
| Number of character variables | None |
| Number of factor variables | None |
| Target variable | TARGET |
| Number of positive targets (1) | 850 |
| Number of negative target(0) | 9150 |

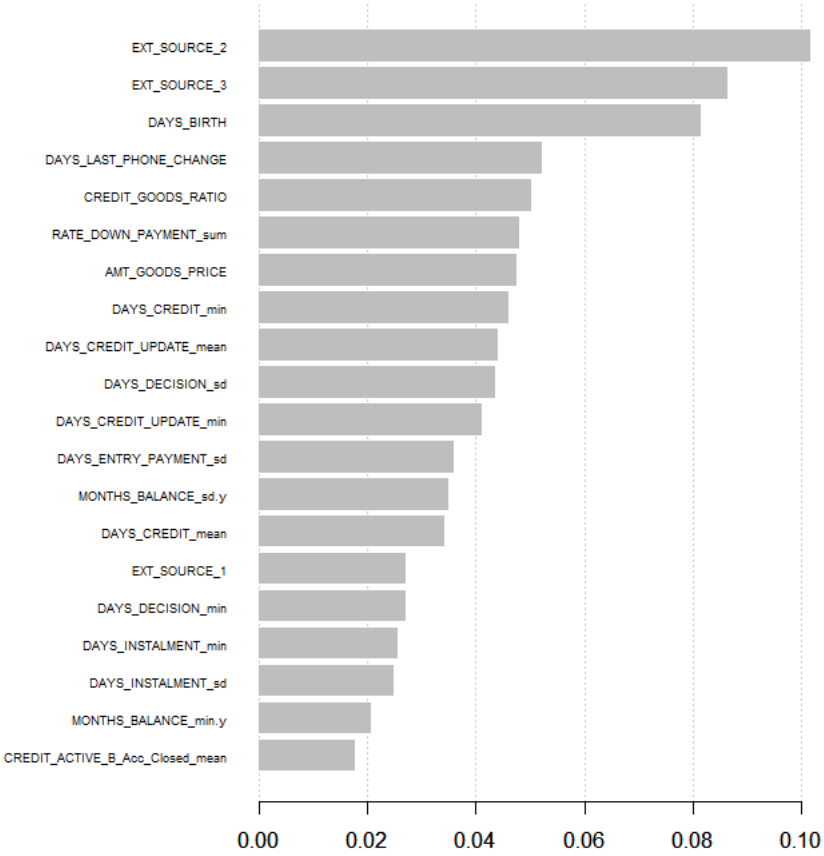
## MODELLS AND RESULTS

Resampling method: Cross-Validation 5 fold

| MODELS | AUC |
|---|---|
| Logistic Regression | 0.6763 |
| Random Forest | 0.5562 |
| Gradient Boosting | 0.7403 |
| SVM | 0.8 Unstable |

## BUSINESS MODEL

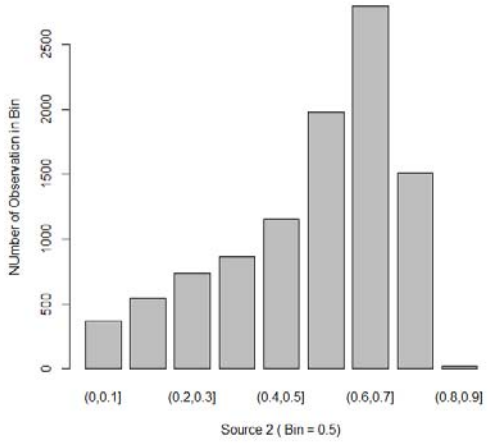
The best model is Gradient Boosting. Top 20 Features according to the importance are:

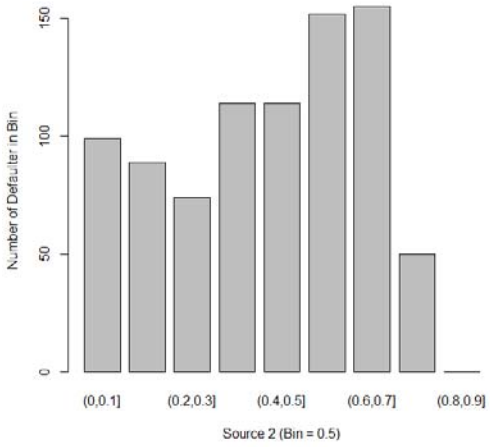



**Analysis of top features**

1. External Source 2

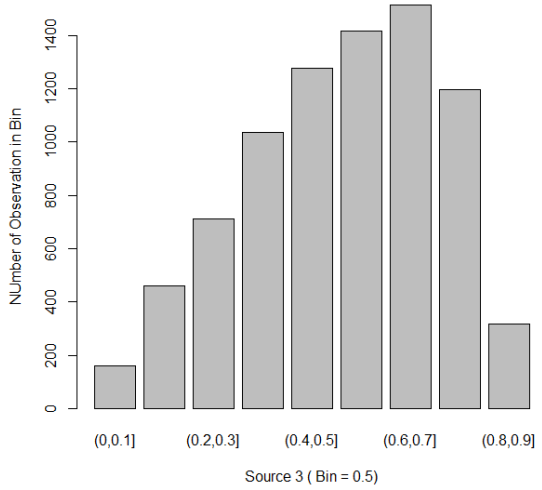
Normalize information about applicant from external source 2. After Analysis of number of applicants in the provided bin and number of defaulters in the respective bin. If external information is low, chances of default is high. With increase in value chances of defaults becoming low. There is kind of inversely proportional relationship between external information 2 and defaulters.

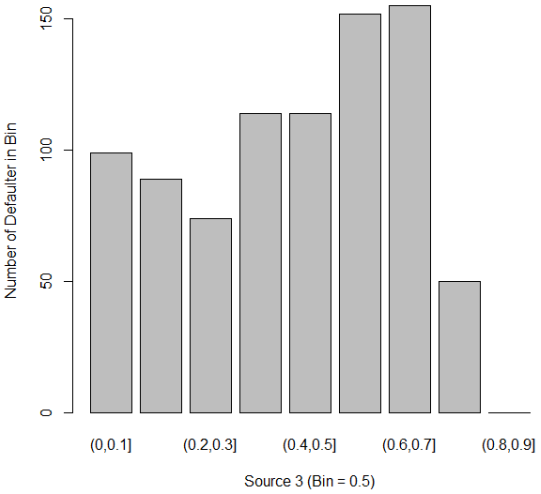


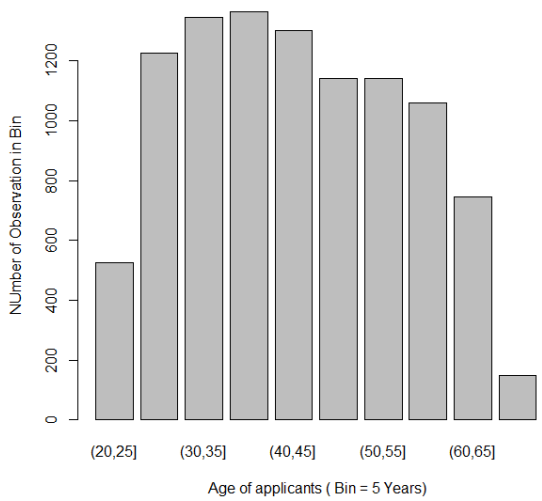

2. External Source 3

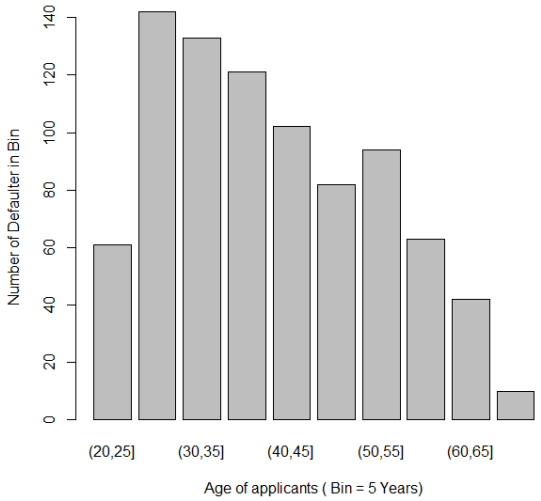
   Normalize information about applicant from external source 3. After Analysis of number of applicants in the provided bin and number of defaulters in the respective bin. If external information is low, chances of default is high. With increase in value chances of defaults becoming low. There is kind of inversely proportional relationship between external information 3 and defaulters.




3. Date of Birth
   Age of applicants is quite good predictor for defaulters. With increasing age number of defaulters get decreased.




4. DAYS INSTALLMENT MINIMUM

How recently the applicants have paid the instalment for his last credit has shown pattern with defaulters' nature. If the applicants have not recent history of paying instalments, then his chances of defaulting the credit is low.