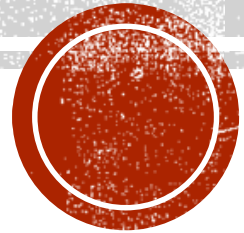
HOME CREDIT DEFAULT RISK

Predict how capable a loan applicant is of repaying the loan.



AGENDA

DATA UNDERSTANDING

DATA ANALYSIS AND VISUALIZATION

DATA PREPROCESSING

DATA MODELLING

DATA EVALUATION

FUTURE EXTENSION

DEMO

1

2

TARGET AUDIENCE

MOTIVATION

- Whenever a customer has no credit history, lenders consider it risky to lend loans as there is no track record.
- Even a trustworthy candidate might be rejected.
- This forces the customers to lend loan from small vendors with high interest rate.
- The Bank loses a good customer and the customer deals with high rates

DATA UNDERSTANDING

DATASET

122 FEATURES WITH 307511 ENTRIES

SK_ID_CURR IS USED AS A UNIQUE IDENTIFIER OF AN APPLICATION

TARGET COLUMN INDICATES IF LOAN IS REPAID OR NOT

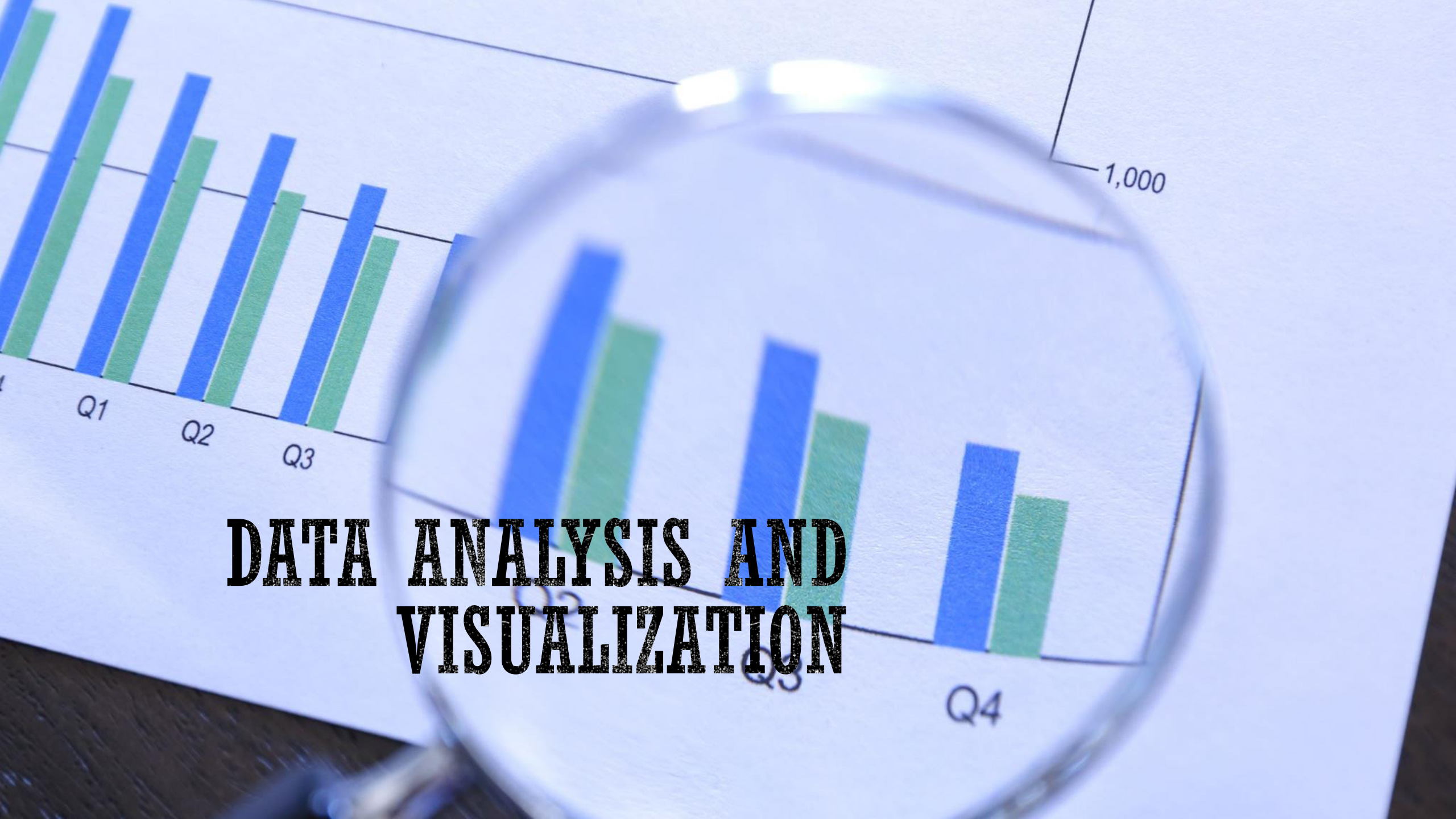
WHY DATA SCIENCE

Extracting information manually from the raw data could be tedious and error prone

High chance of mispredictions

Decision should be based on previous knowledge

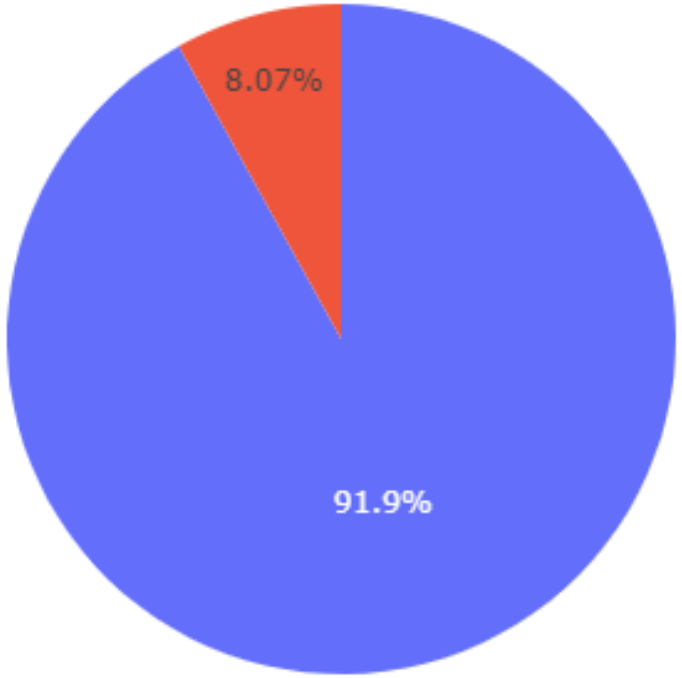
Good candidate of supervised classification problem



DATA ANALYSIS AND VISUALIZATION

DUPLICATES

No duplicates are present in the entire dataset.



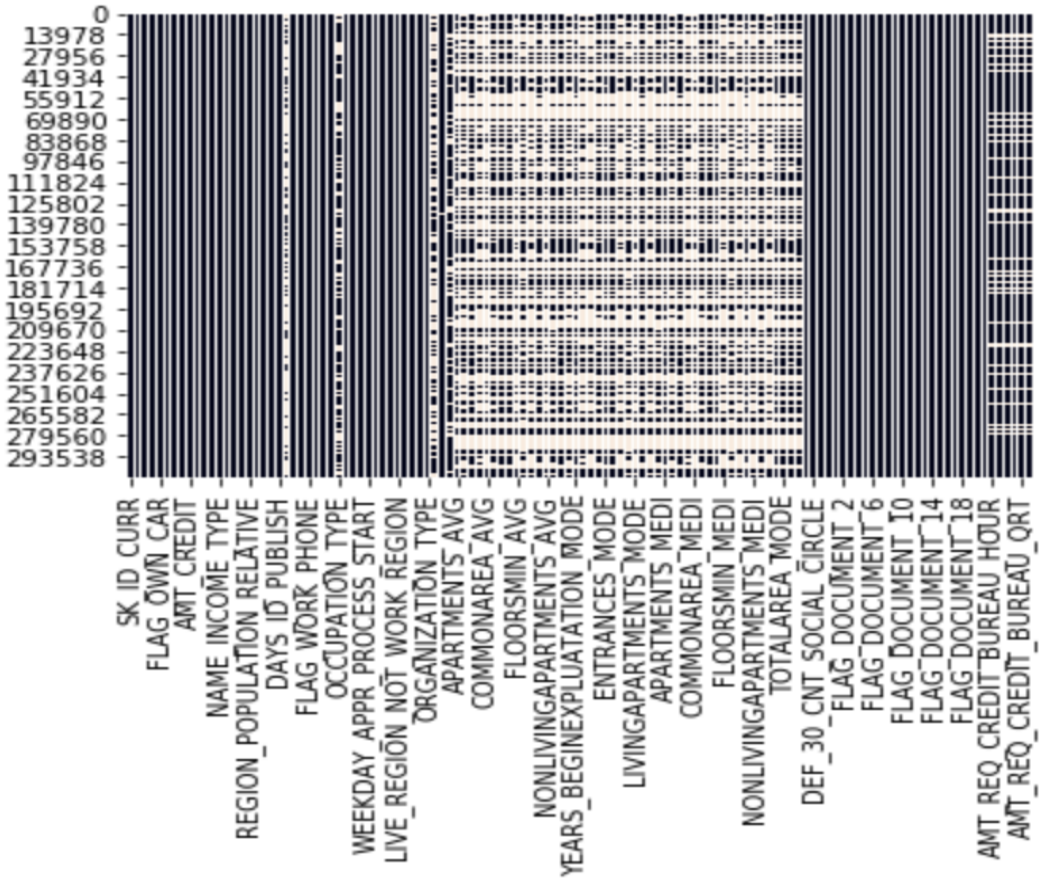
IMBALANCED DATA

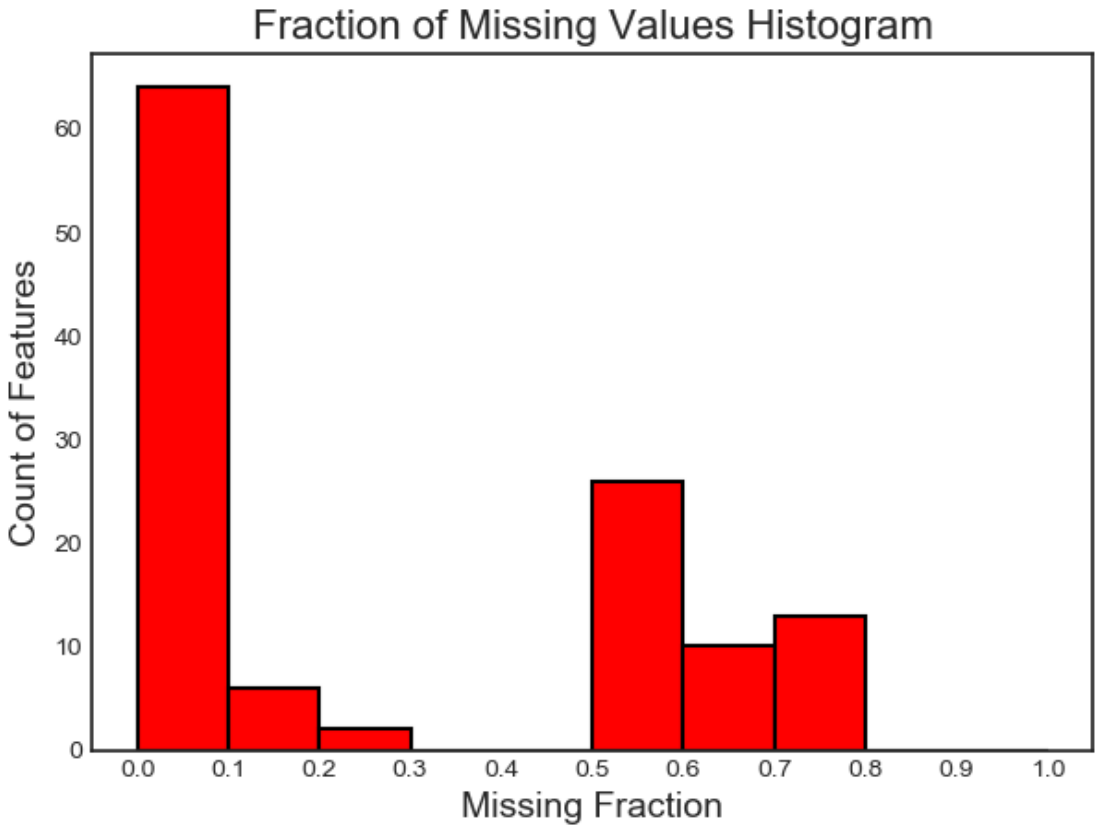
CLASS DISTRIBUTION

0 - REPAID

1 - NOT REPAID

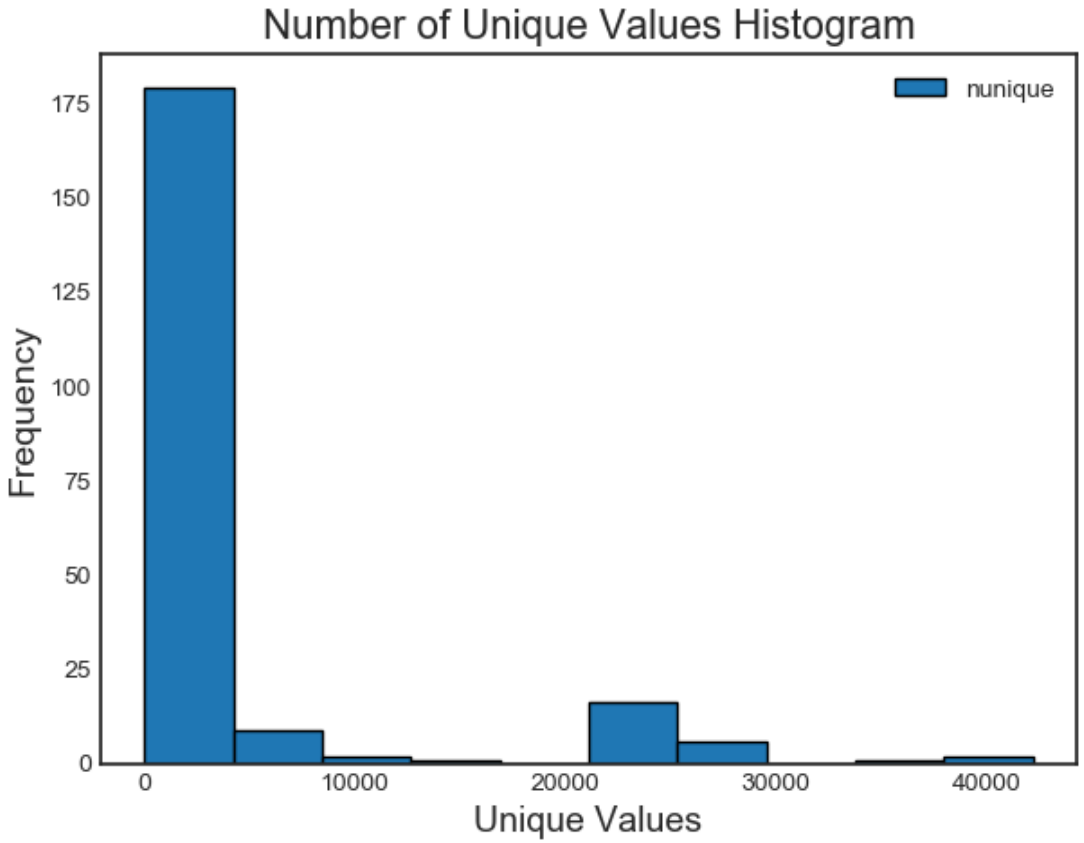
MISSING VALUES



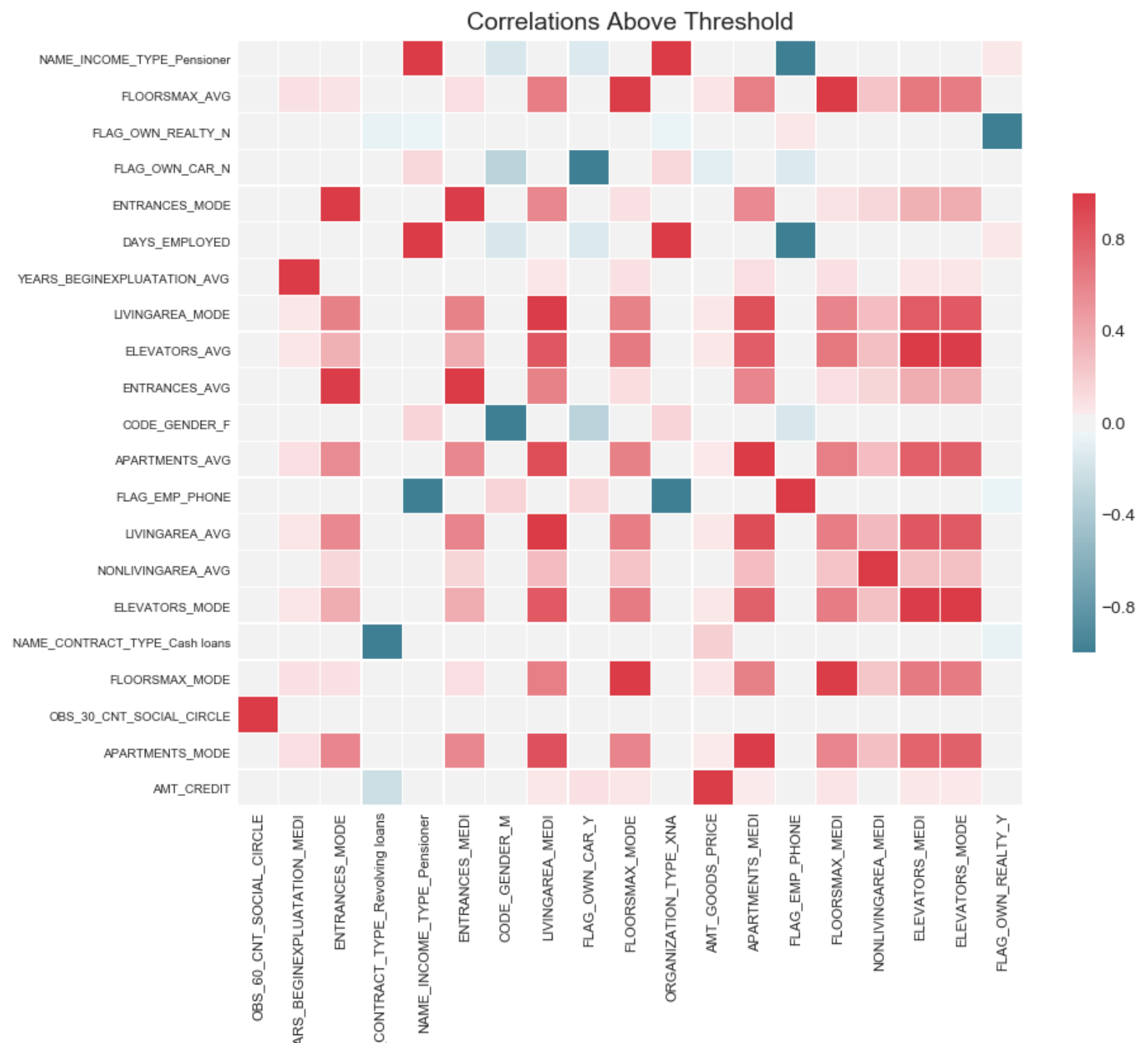


23 feature contains more than 60 percent of missing value

SINGLE UNIQUE VALUE



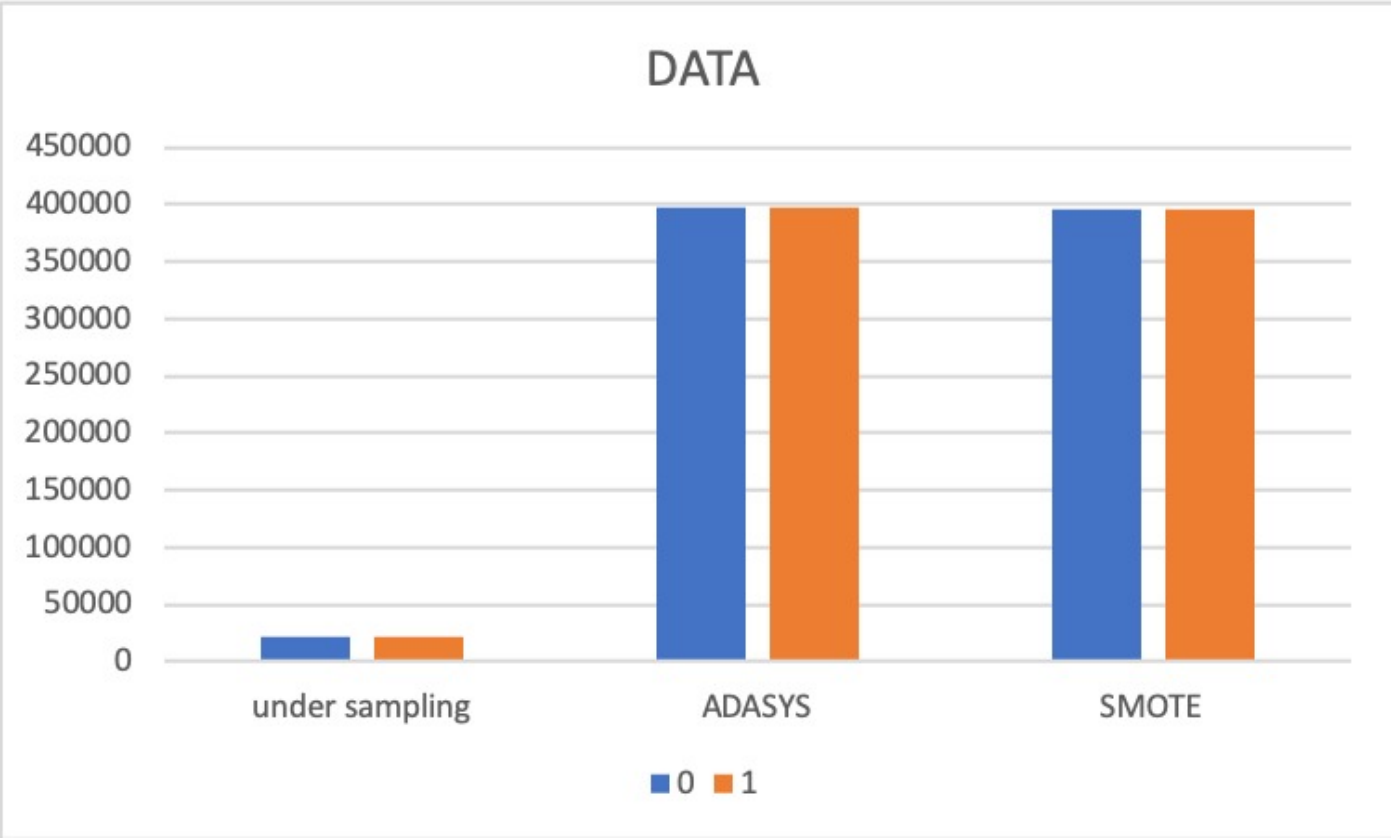
3 features with a single unique value



FEATURES CORRELATION

- Quantifies the degree to which a relationship between two variables
- Shows correlations above threshold 0.975
- There are 19 features with a correlation magnitude greater than the threshold
- Lighter green indicates the higher correlation.

DATA PREPROCESSING



DEALING WITH IMBALANCED DATA

BEST RESULT : Under Sampling

FILLING MISSING VALUES

- Mean
- Median
- Mode
- Iterative Imputation

BEST RESULT : Iterative Imputation

FEATURE SELECTION

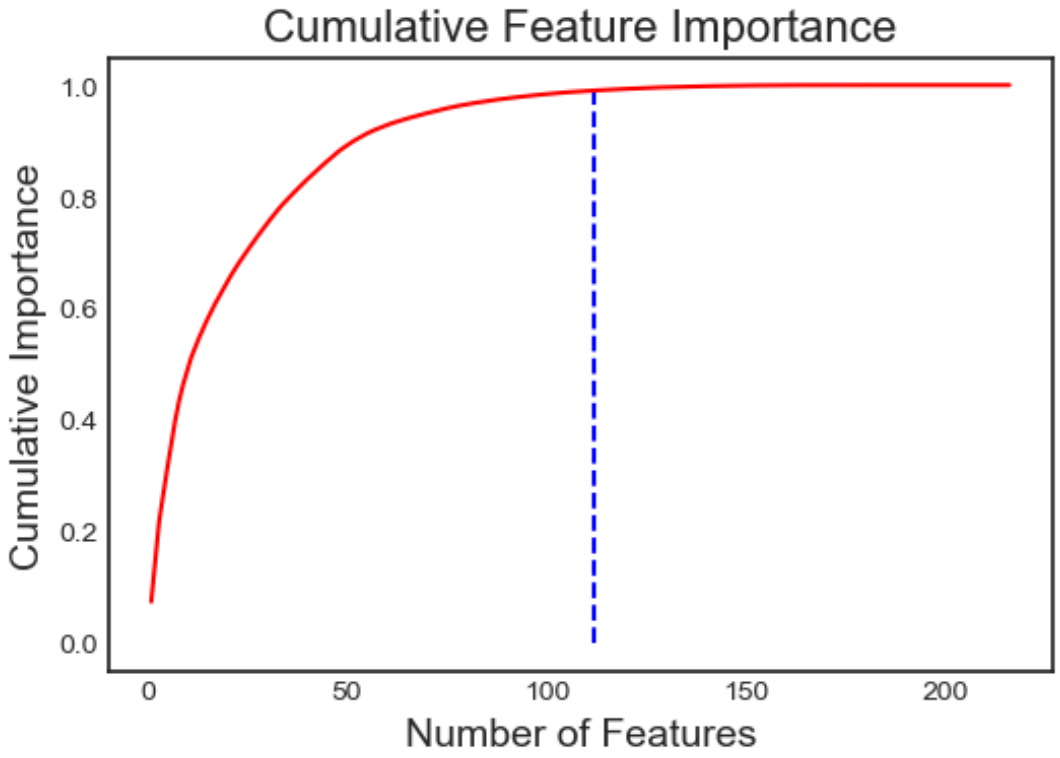
Removed *23 features which contain more than 60 percent missing value*

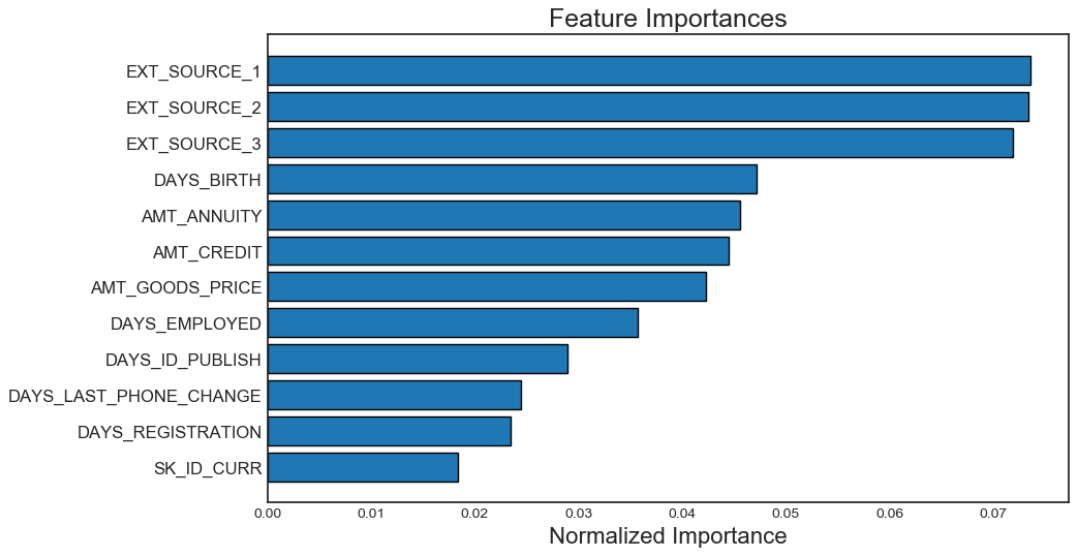
Removed *3 features with single unique value*

Dropped highly correlated features

Removed 105 features that do not contribute to cumulative importance of 0.99.

Removed 47 features of zero importance after one-hot encoding.





Most important features

ENCODING THE CATEGORICAL VARIABLE:

ONE HOT ENCODING

LABEL ENCODING

DATA MODELING

MODELS

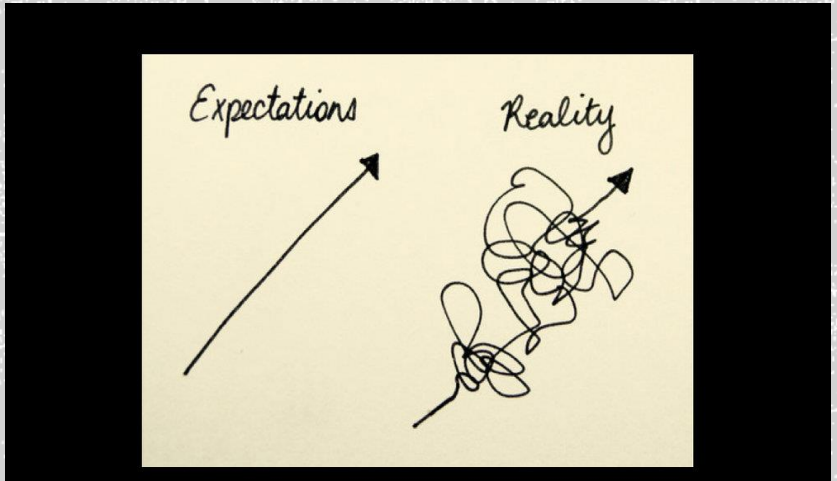
Logistic Regression	Gaussian Naive Bayes	Decision Tree Classifier	Random Forest	Bagging Classifier
Gradient Boosting	AdaBoost	Grid Search	LightGBM	Hyper-parameter Tuning using Grid Search

Stacking

- Logistic Regression
- Bagging Classifier
- AdaBoost Classifier
- RandomForestClassifier

EXCEPTATION	REALITY
Oversampling will give high accuracy	Under sampling proves out to be a better approach
Tree-based models will perform better	Best result was given by Logistic Regression
Grid Search will improve performance	Default parameters for Logistic Regression resulted in the same performance

OBSERVATIONS



DATA EVALUATION

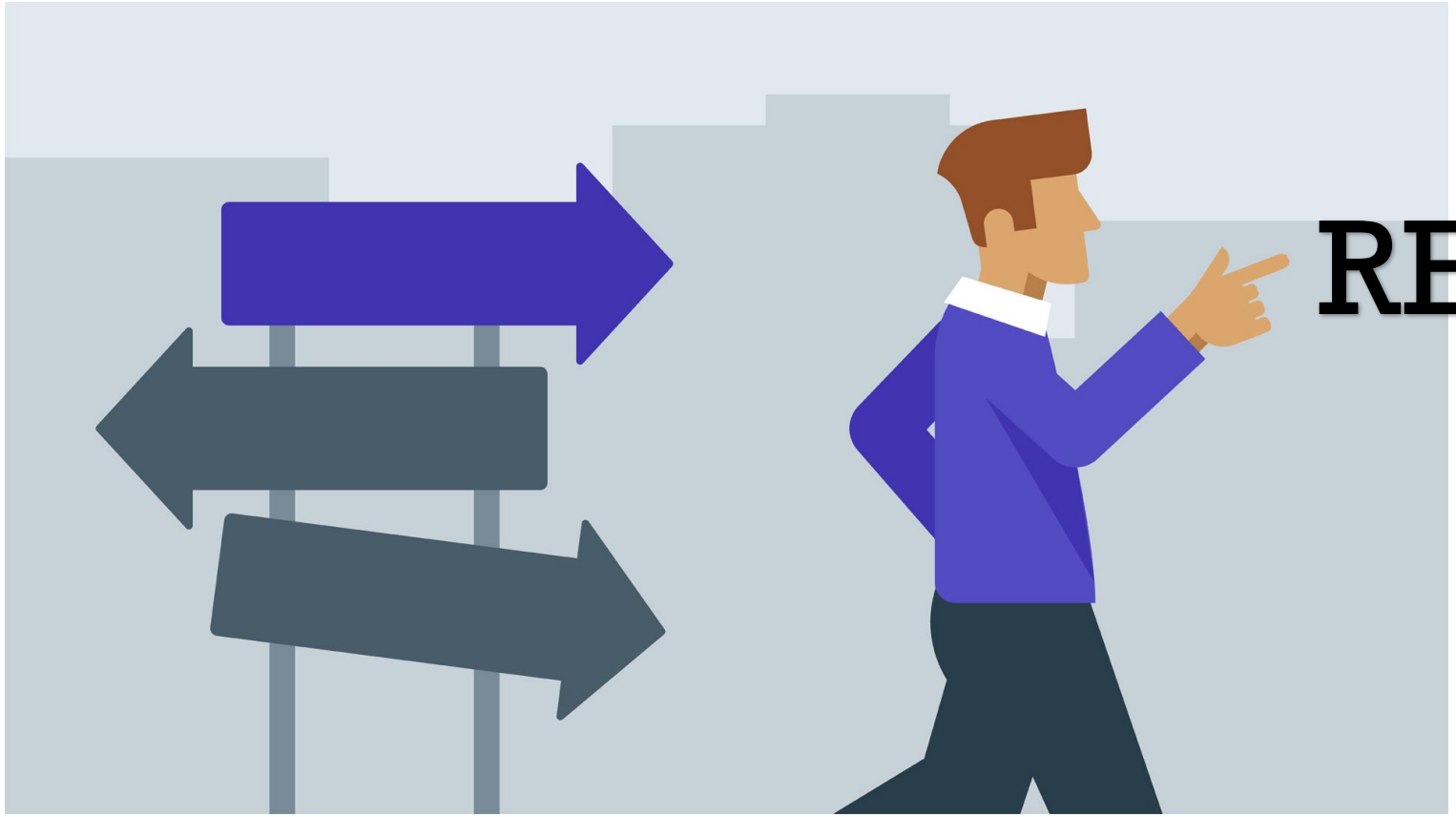
EVALUATION METRICS

- **Accuracy**
- **Precision**
- Recall
- F1 score
- ROC curve

WHICH METRIC TO USE ?

- A high precision value implies that the model returned substantially more relevant results than the irrelevant ones whereas a high recall value implies that the model returned most of the relevant results.
- Would like to have a model with higher recall value rather than the precision.
- Precision cannot be below a threshold to make sure the Home credit will not take every loaner as a person who can't have a consistent repay capability since a precision can indicate many false positives

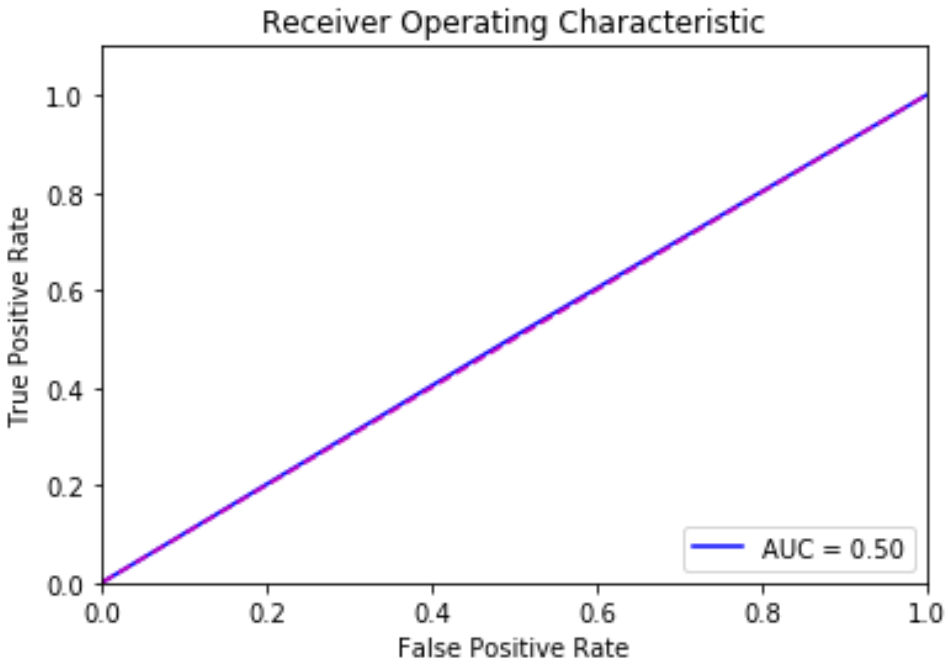
BEST EVALUATION METRIC FOR THE MODEL



RECALL

RANDOM MODEL

Accuracy	ROC	Precision	Recall
0.498	0.502	0.09	0.54



BUILDING THE CLASSIFIER

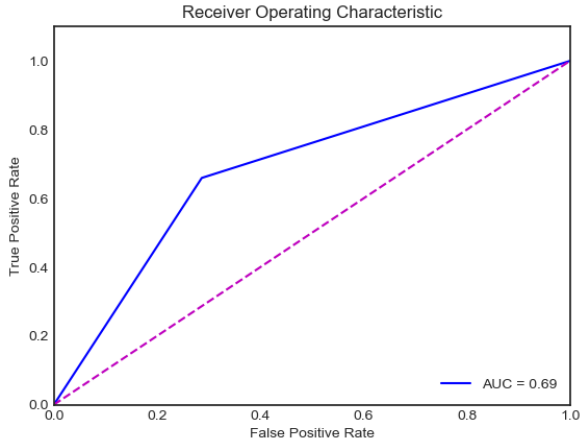
- Split the data (Stratified Shuffle Split cross-validator)
- Under sample
- One-Hot encoding of categorical features
- Impute missing values
- Scale (Standard Scaler)
- Feature Selection and filtering
- Classify

Model	Accuracy	ROC	Precision	Recall
Logistic Regression	0.688	0.682	0.16	0.68
Stacking	0.709	0.686	0.17	0.66

BEST MODELS

AUC

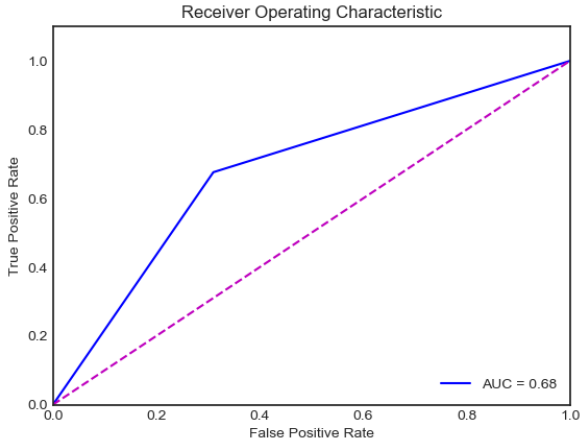
Logistic Regression



Precision score : 0.16
Recall score : 0.68

	precision	recall	f1-score	support
0	0.9604	0.6898	0.8029	42403
1	0.1607	0.6762	0.2597	3724

Stacking



Precision score : 0.17
Recall score : 0.66

	precision	recall	f1-score	support
0	0.9598	0.7135	0.8185	42403
1	0.1682	0.6595	0.2680	3724

BEST MODEL

LOGISTIC REGRESSION

FUTURE EXTENSION

- Re-sampling skewed data by the scale of client's profile or other useful information available from the data.
- Collect additional data about our features in order to train the model better
- Explore the remaining datasets available to train the model better
- Improve UI
- Move the training to distributed environment and using cloud services

TIME TO EVALUATE

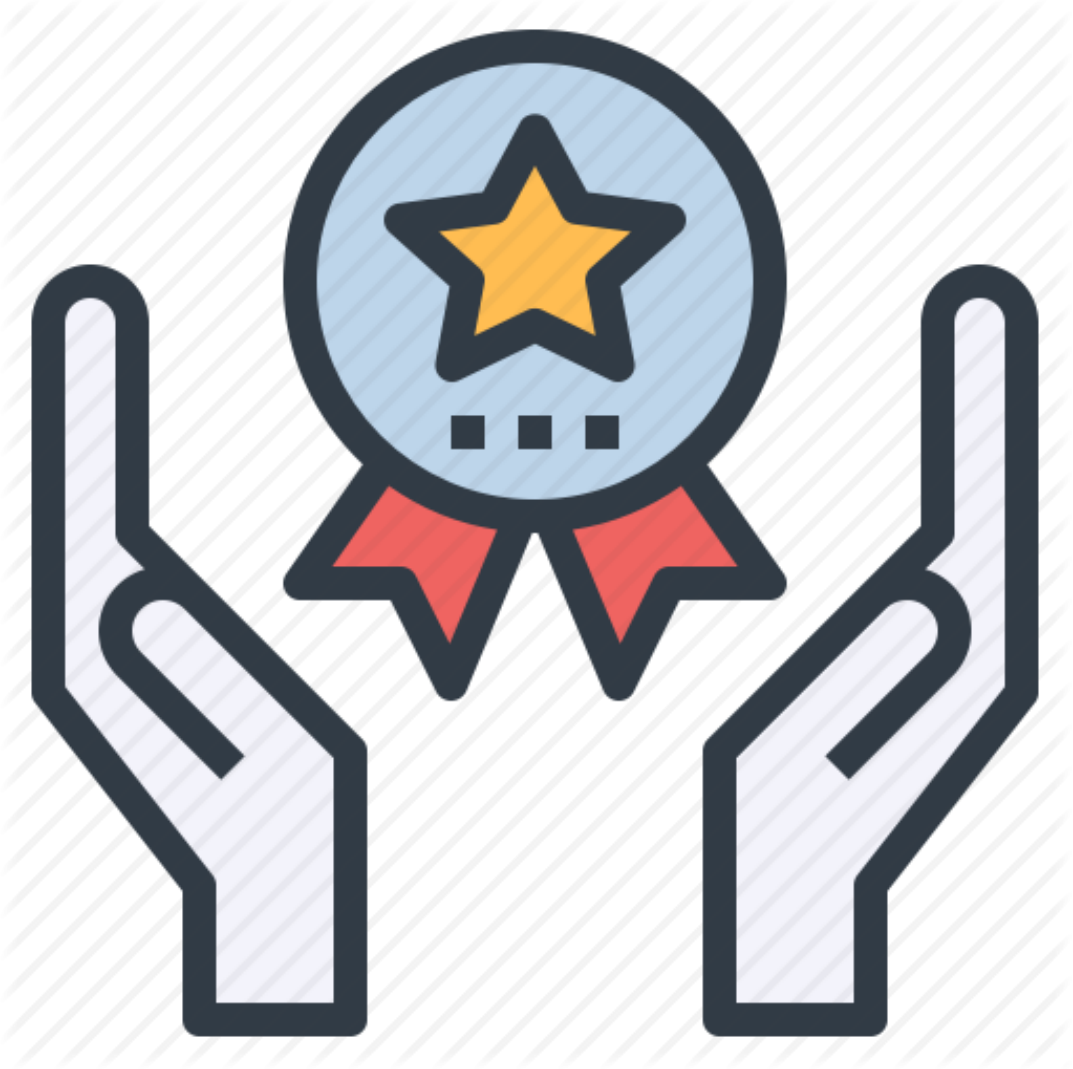
Applicant 1:

'SK_ID_CURR': 100005,
'NAME_CONTRACT_TYPE': 'Cash loans', 'CODE_GENDER': 'M',
'FLAG_OWN_CAR': 'N',
'FLAG_OWN_REALTY': 'Y',
'CNT_CHILDREN': 0,
'AMT_INCOME_TOTAL': 99000.0,
'AMT_CREDIT': 222768.0,
'AMT_ANNUITY': 17370.0,
'AMT_GOODS_PRICE': 180000.0,
'NAME_TYPE_SUITE': 'Unaccompanied',
'NAME_INCOME_TYPE': 'Working',
'NAME_EDUCATION_TYPE': 'Secondary / secondary special',
'NAME_FAMILY_STATUS': 'Married',
'NAME_HOUSING_TYPE': 'House / apartment',
'REGION_POPULATION_RELATIVE': 0.03579200000000004,
'DAYS_BIRTH': -18064,
'DAYS_EMPLOYED': -4469,
'DAYS_REGISTRATION': -9118.0,
'DAYS_ID_PUBLISH': -1623,

'OWN_CAR_AGE': nan, ‘
FLAG_MOBIL': 1,
'FLAG_EMP_PHONE': 1,
'FLAG_WORK_PHONE': 0,
'FLAG_CONT_MOBILE': 1,
'FLAG_PHONE': 0,
'FLAG_EMAIL': 0,
'OCCUPATION_TYPE': 'Low-skill Laborers’,
'CNT_FAM_MEMBERS': 2.0,
‘REGION_RATING_CLIENT': 2,
'REGION_RATING_CLIENT_W_CITY': 2,
'WEEKDAY_APPR_PROCESS_START': 'FRIDAY',
'HOUR_APPR_PROCESS_START': 9,
'REG_REGION_NOT_LIVE_REGION': 0,
'REG_REGION_NOT_WORK_REGION': 0,
'LIVE_REGION_NOT_WORK_REGION': 0,
'REG_CITY_NOT_LIVE_CITY': 0,
'REG_CITY_NOT_WORK_CITY': 0,
'LIVE_CITY_NOT_WORK_CITY': 0, ‘
ORGANIZATION_TYPE': 'Self-employed’,
'EXT_SOURCE_1': 0.5649902017969249,
'EXT_SOURCE_2': 0.2916555320093651,
'EXT_SOURCE_3': 0.432961670974407,
'APARTMENTS_AVG': nan,
'BASEMENTAREA_AVG': nan,
'YEARS_BEGINEXPLUATATION_AVG': nan,
'YEARS_BUILD_AVG': nan,
'COMMONAREA_AVG': nan,
'ELEVATORS_AVG': nan,
'ENTRANCES_AVG': nan,
'FLOORSMAX_AVG': nan,
'FLOORSMIN_AVG': nan,
'LANDAREA_AVG': nan,
'LIVINGAPARTMENTS_AVG': nan,
'LIVINGAREA_AVG': nan,
'NONLIVINGAPARTMENTS_AVG': nan,
'NONLIVINGAREA_AVG': nan,
'APARTMENTS_MODE': nan, ‘
BASEMENTAREA_MODE': nan,
'YEARS_BEGINEXPLUATATION_MODE': nan,
'YEARS_BUILD_MODE': nan,
'COMMONAREA_MODE': nan,
'ELEVATORS_MODE': nan,
'ENTRANCES_MODE': nan,
'FLOORSMAX_MODE': nan,
'FLOORSMIN_MODE': nan,
'LANDAREA_MODE': nan,
‘LIVINGAPARTMENTS_MODE': nan,

'LIVINGAREA_MODE': nan,
'NONLIVINGAPARTMENTS_MODE': nan,
'NONLIVINGAREA_MODE': nan,
'APARTMENTS_MEDI': nan,
'BASEMENTAREA_MEDI': nan,
'YEARS_BEGINEXPLUATATION_MEDI': nan,
'YEARS_BUILD_MEDI': nan,
'COMMONAREA_MEDI': nan,
'ELEVATORS_MEDI': nan,
'ENTRANCES_MEDI': nan,
'FLOORSMAX_MEDI': nan,
'FLOORSMIN_MEDI': nan,
'LANDAREA_MEDI': nan,
'LIVINGAPARTMENTS_MEDI': nan,
'LIVINGAREA_MEDI': nan,
'NONLIVINGAPARTMENTS_MEDI': nan,
'NONLIVINGAREA_MEDI': nan,
'FONDKAPREMONT_MODE': nan,
'HOUSETYPE_MODE': nan,
'TOTALAREA_MODE': nan,
'WALLSMATERIAL_MODE': nan,
'EMERGENCYSTATE_MODE': nan,

'OBS_30_CNT_SOCIAL_CIRCLE': 0.0,
'DEF_30_CNT_SOCIAL_CIRCLE': 0.0, '
OBS_60_CNT_SOCIAL_CIRCLE': 0.0,
'DEF_60_CNT_SOCIAL_CIRCLE': 0.0,
'DAYS_LAST_PHONE_CHANGE': 0.0,
'FLAG_DOCUMENT_2': 0, 'FLAG_DOCUMENT_3': 1,
'FLAG_DOCUMENT_4': 0, '
FLAG_DOCUMENT_5': 0, 'FLAG_DOCUMENT_6': 0,
'FLAG_DOCUMENT_7': 0,
'FLAG_DOCUMENT_8': 0, 'FLAG_DOCUMENT_9': 0,
'FLAG_DOCUMENT_10': 0,
'FLAG_DOCUMENT_11': 0, 'FLAG_DOCUMENT_12': 0,
'FLAG_DOCUMENT_13': 0,
'FLAG_DOCUMENT_14': 0, 'FLAG_DOCUMENT_15': 0,
'FLAG_DOCUMENT_16': 0, '
FLAG_DOCUMENT_17': 0, 'FLAG_DOCUMENT_18': 0,
'FLAG_DOCUMENT_19': 0,
'FLAG_DOCUMENT_20': 0, 'FLAG_DOCUMENT_21': 0,
'AMT_REQ_CREDIT_BUREAU_HOUR': 0.0,
'AMT_REQ_CREDIT_BUREAU_DAY': 0.0,
'AMT_REQ_CREDIT_BUREAU_WEEK': 0.0,
'AMT_REQ_CREDIT_BUREAU_MON': 0.0,
'AMT_REQ_CREDIT_BUREAU_QRT': 0.0,
'AMT_REQ_CREDIT_BUREAU_YEAR': 3.0}



BENEFITS

- Banks Risk factor reduced
- Deserving Candidates will not suffer
- One click solution
- Saves time


LIVE DEMO


QUESTIONS?
COMMENTS?