




# Home Credit Default Risk Prediction

Supervised Machine Learning Method

Chia Yi Liaw, Mounica Subramani  
Sharyu Deshmukh, Somya Bhargava



# Table of Content

---

1. Abstract
2. Business Understanding
3. Data Understanding
4. Data Preparation
5. Modeling
6. Evaluation/Result
7. Discussion
8. Future Work



# Abstract

---

Home Credit is a finance provider that focuses on serving the unbanked population. Many people struggle to get loans due to insufficient or non-existent credit histories. And, unfortunately, this population is often taken advantage of by untrustworthy lenders.

The Home Credit Default Risk challenge is a standard supervised machine learning task where the goal is to use historical loan application data to predict their clients' repayment abilities based on datasets provided.



# Business Understanding

---

## Objective :

Ensure the clients' capable of repayment aren't rejected and the loans repayment calendar will empower the clients to be successful

## Resources:

Testing and training datasets are available from Kaggle competition

- **Supervised:** The labels are included in the training data and the goal is to train a model to learn to predict the labels from the features
- **Classification:** The label is a binary variable, 0 (will repay loan on time), 1 (will have difficulty repaying loan)

# Exploratory Data Analysis

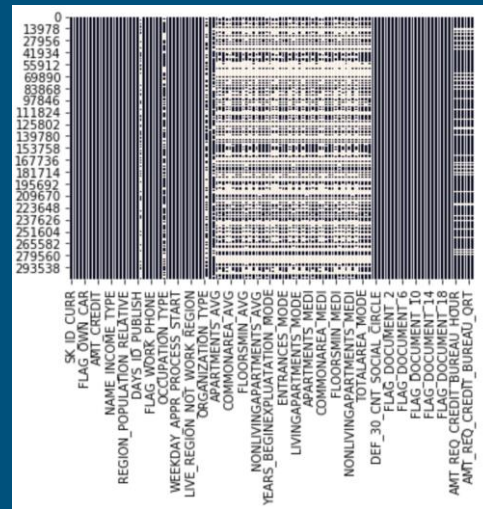
## Missing Value

The Dataset contains 122 features with 307511 entries. we found out that missing value is concentrated in several features.

In addition, there are 17 features contain more than 60 percent of missing value. Imputation and deletion will be performed in the data pre-processing stage.

COMMONAREA_MEDI	69.872297
COMMONAREA_AVG	69.872297
COMMONAREA_MODE	69.872297
NONLIVINGAPARTMENTS_MODE	69.432963
NONLIVINGAPARTMENTS_MEDI	69.432963
NONLIVINGAPARTMENTS_AVG	69.432963
FONDKAPREMONT_MODE	68.386172
LIVINGAPARTMENTS_MEDI	68.354953
LIVINGAPARTMENTS_MODE	68.354953
LIVINGAPARTMENTS_AVG	68.354953
FLOORSMIN_MEDI	67.848630
FLOORSMIN_MODE	67.848630
FLOORSMIN_AVG	67.848630
YEARS_BUILD_MEDI	66.497784
YEARS_BUILD_AVG	66.497784
YEARS_BUILD_MODE	66.497784
OWN_CAR_AGE	65.990810

Percentage of missing value

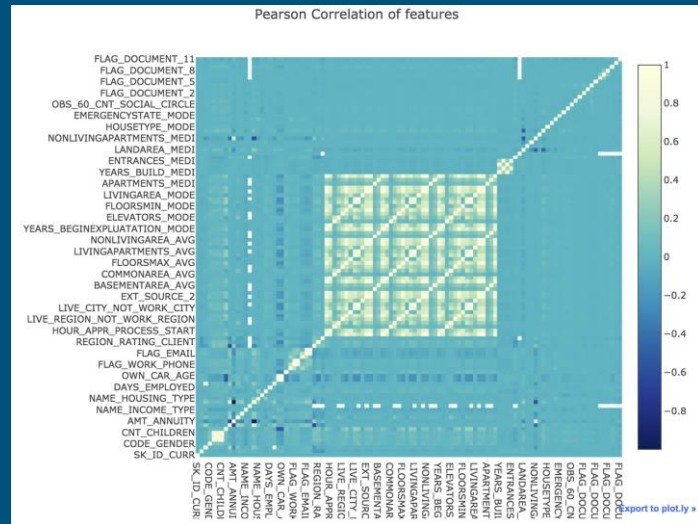


Missing Value Map

# Exploratory Data Analysis

## Feature Correlation

Pearson feature correlation quantifies the degree to which a relationship between two variables. By analyzing correlation feature, we implement the random forest method and LightGBM to improve calculation efficiency. Pearson correlation heat map shows the correlation between predictors, the lighter green indicates the higher correlation.



Feature Correlation Map

# Data Preprocessing

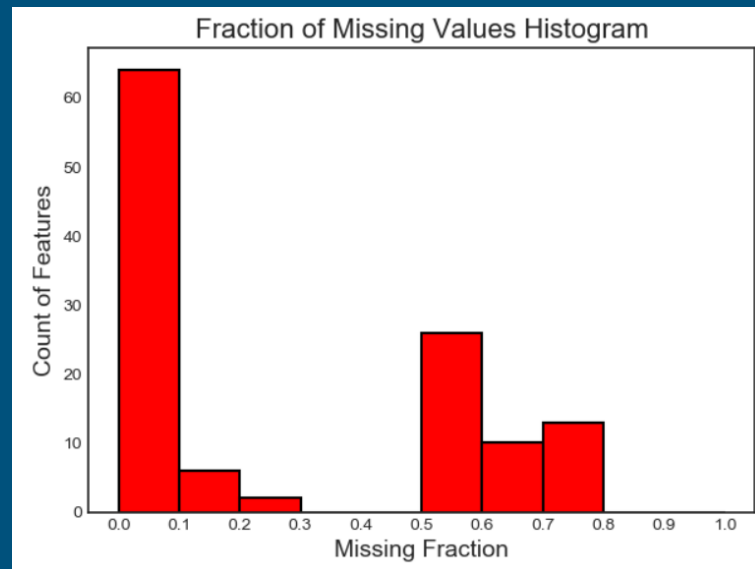
## Feature Selection

With the risk of high dimensional dataset and to reduce the variance of the model or chances of overfitting.

**Missing Value** : there are 23 features contain more than 60% missing value, hence we decide to remove those features as it's not suitable for training models.

**LightGBM** : removes the features that are collinear, low\_importance and zero\_importance.

The outputs so generated are stored in csv format so that we can train them all and compare the results.



# Data Preprocessing

---

## Encoding the Categorical Variable

**One hot encoding** for multiple categories and  
**Label encoding** for binary categorical features.

## Imputation of Missing Value

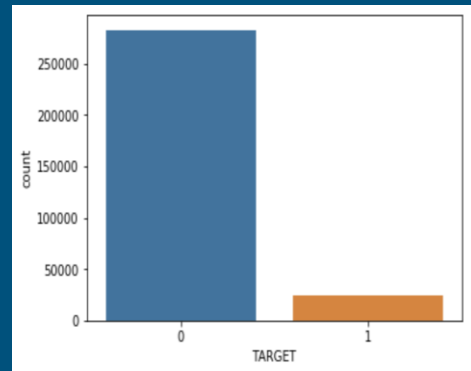
After feature selection , There are still some columns with missing data. Hence, will impute the missing value with mode which is the most frequent value in the column, moreover, it would apply well on the categorical variable.



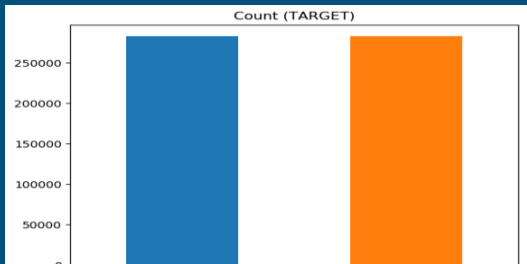
# Data Preprocessing

## Dealing with Imbalance Data

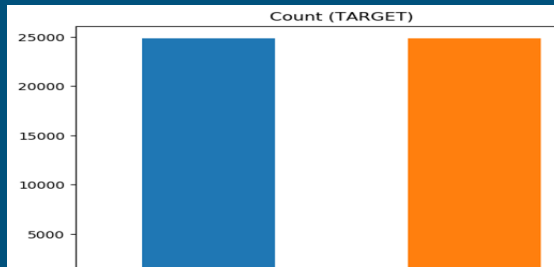
The below histogram show the imbalanced distribution of loan repayment in the dataset. The target variable defines if the loan was repaid by the borrower or not. From the graph, we find out that the data is highly imbalanced. Since, it might lead to incorrect result during modelling process, we needed resampling of the dataset. As a result, we have used both over and under-sampling to train the model. By looking at the evaluation, we will choose the better sampling method.



*Distribution of original data*



*After oversampling method*



*After undersampling method*

# Methodology

---

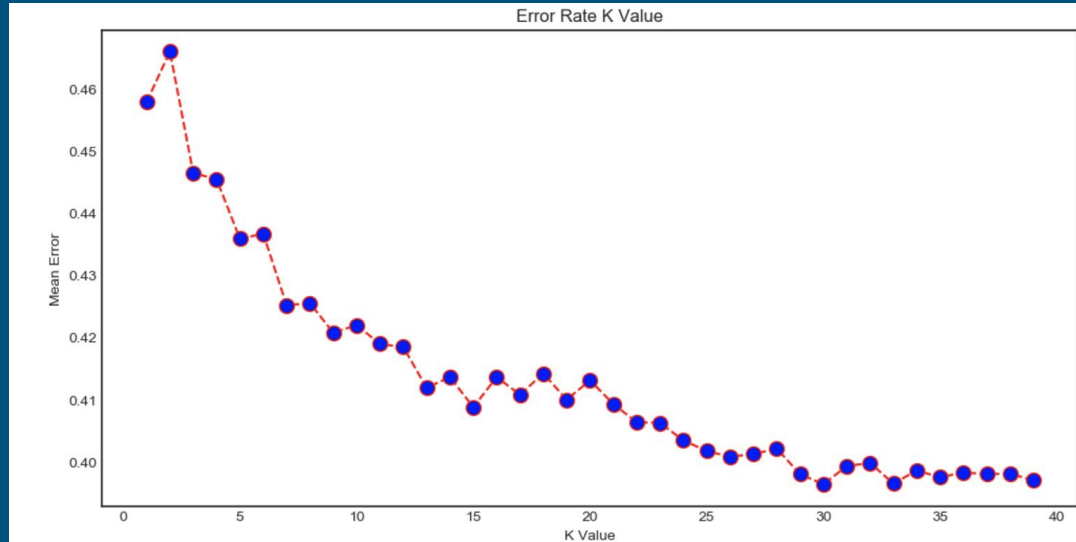
<u>Models</u>	<u>Detail</u>
Logistic Regression	Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous
K-Nearest Neighbors	KNN model is to find the K value. In general, a larger k suppresses the effects of noise, but makes the classification boundaries less distinct.
Gaussian Naive Bayes	The Naive Bayes classifiers are working based on the Bayes' theorem, which describes the probability of an event
Adaboost	Adaptive boosting creates a highly accurate prediction rule by combining many relatively weak and inaccurate rules.
Random Forest	Random Forest use the bagging idea, it resamples the data and features which will decrease the variance in the model
Support Vector Machine	SVM used to solve linear or non-linear problem which is a discriminative classifier
XGBoost	XGBoost is an implementation of gradient boosted decision trees designed for speed and performance
LightGBM	LightGBM uses Gradient-based One-Side Sampling (GOSS) to filter out the data instances to split value

# Result

<u>Models</u>	<u>Accuracy Score</u>	<u>ROC Score</u>	<u>Precision</u>	<u>Recall</u>	<u>F1 Score</u>
Logistic Regression	0.665	0.665	0.67	0.65	0.66
K-Nearest Neighbors	0.554	0.554	0.52	0.90	0.46
Gaussian Naive Bayes	0.536	0.534	0.52	0.90	0.46
Adaboost	0.676	0.676	0.68	0.67	0.67
Random Forest	0.671	0.671	0.68	0.65	0.67
Support Vector Machine	0.666	0.666	0.67	0.65	0.66
XGBoost	0.685	0.686	0.682	0.682	0.682
LightGBM	0.748	0.749	0.69	0.68	0.68

# KNN choosing K value

---



In addition to the mean error rate = 0 when K value between , adjusting k value would influence the accuracy.

# Discussion

---

1. We observed that LightGBM performs outstandingly as compared to all the other models. In general, it was observed that the tree based models are performing better on this dataset.
2. Consider a finance provider wish to evaluate the risk of loan repayment ability. Company would like to have a model with higher recall value rather than the Precision. Because company wants to make sure the clients are able to repay the loan based on given operation data.
1. Meanwhile, the 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

# Future Work

---

1. Perform manual feature engineering, make polynomial features and perform predictions based on them.
2. Collect more data about our features in order to train the model better
3. Perform Stacking/Blending for improved accuracy
4. Evaluate the performance of three different models on validation set using their precision, recall and f1-score.