Home Credit Default Risk

Will more features help?

Leung, Ko Tsun
Cheng, Tsz Yui
Yang, Po-Yen

About

Overview

Home Credit Default Risk

Agenda

INTRODUCTION

Introduction

Home Credit is trying to provide loans to people lack of credit history by predicting his/her ability to pay back with respect to his/her other alternative or historical data.

In this project, we utilized **logistic regression** and **light Gradient Boosting Machine** (light GBM) as our prediction models.

Problem Statement

Observation

- A subset of features can obtain similar accuracy with a lower computation cost compared to using the full dataset

Amount of features

- Motivated by this, we want to study the effect of including more features in our model
- We have merged four datasets with over 600 features (6x more than the original dataset) and examined the improvement of utilizing more features during prediction

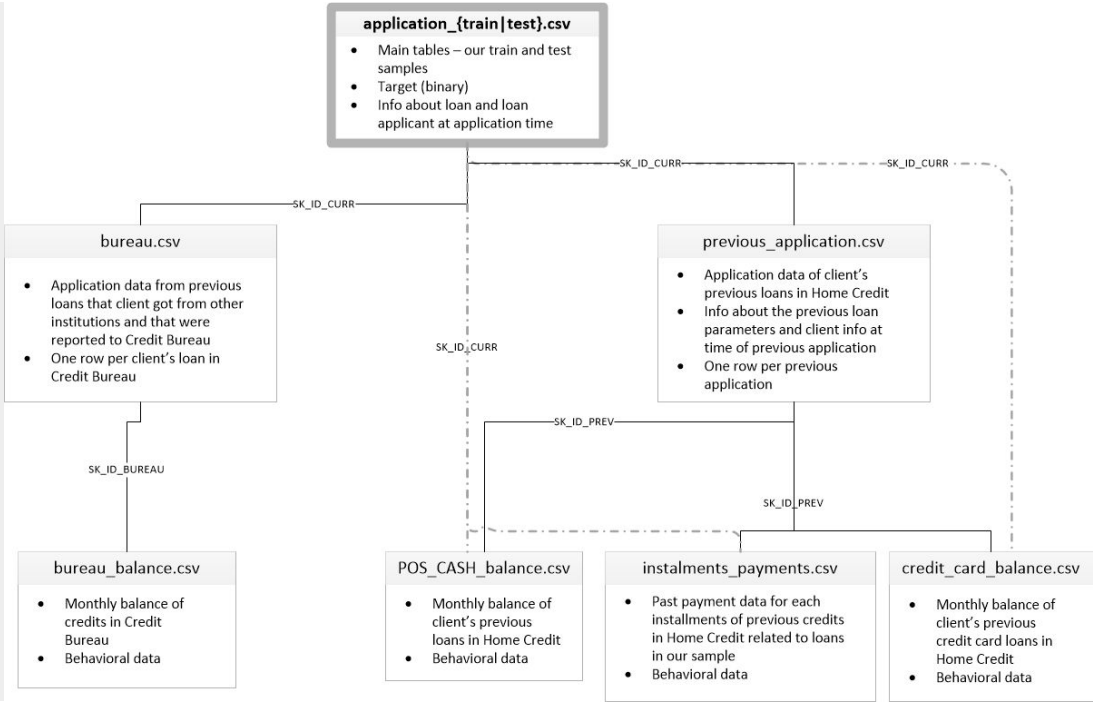
Data Overview

Full dataset consists of **seven** tables

In our model, we used the following tables:

- ➔ application_train
- ➔ bureau
- ➔ bureau_balance
- ➔ previous_application

- bureau and balance data are aggregated in the training dataset
- Test dataset merged



Data & Data Preprocessing

Various types of variables are present

- Numerical
- Categorical
- Binary

Data preprocessing method

- Label encoding is used for features having <=2 unique values
- One hot encoding is used for other features
- Label encoding has the problem of arbirary ordering significance

Data & Data Preprocessing

Missing value handling

Lots of predictor variables contain missing data

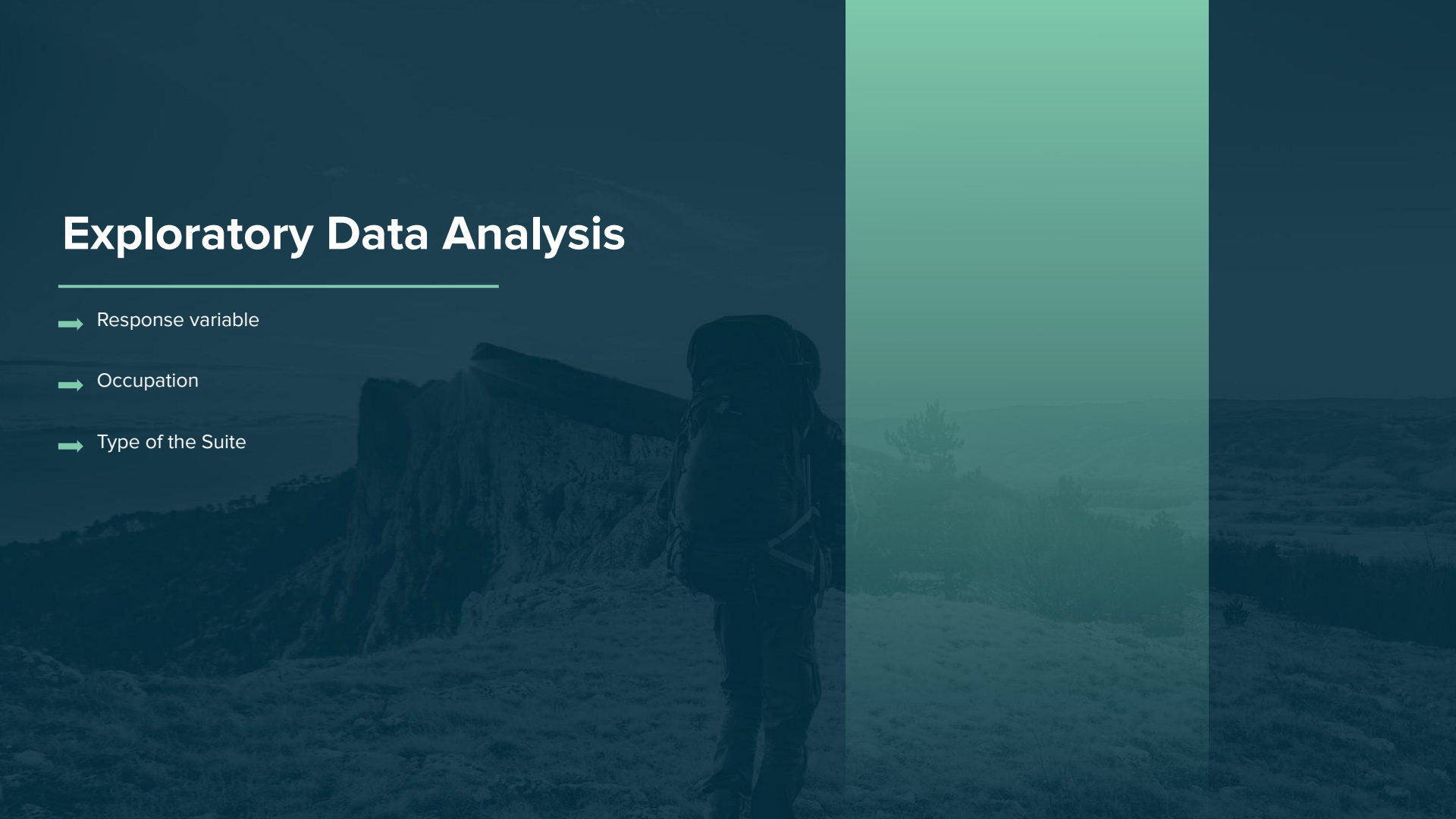
➔ Drop features with high rate of missing values (70%)

- Fill in the missing values by **median**
  - ➔ Advantage : Median is not affected by outliers
- Apply **MinMaxScalar**
  - ➔ Advantage : Compress all inliers in range [0,1]

	Missing Values	% of Total Values
COMMONAREA_MEDI	214865	69.9
COMMONAREA_AVG	214865	69.9
COMMONAREA_MODE	214865	69.9
NONLIVINGAPARTMENTS_MEDI	213514	69.4
NONLIVINGAPARTMENTS_MODE	213514	69.4
NONLIVINGAPARTMENTS_AVG	213514	69.4
FONDKAPREMONT_MODE	210295	68.4
LIVINGAPARTMENTS_MODE	210199	68.4
LIVINGAPARTMENTS_MEDI	210199	68.4
LIVINGAPARTMENTS_AVG	210199	68.4
FLOORSMIN_MODE	208642	67.8
FLOORSMIN_MEDI	208642	67.8
FLOORSMIN_AVG	208642	67.8
YEARS_BUILD_MODE	204488	66.5
YEARS_BUILD_MEDI	204488	66.5
YEARS_BUILD_AVG	204488	66.5
OWN_CAR_AGE	202929	66.0
LANDAREA_AVG	182590	59.4
LANDAREA_MEDI	182590	59.4
LANDAREA_MODE	182590	59.4

Exploratory Data Analysis

- Response variable
- Occupation
- Type of the Suite



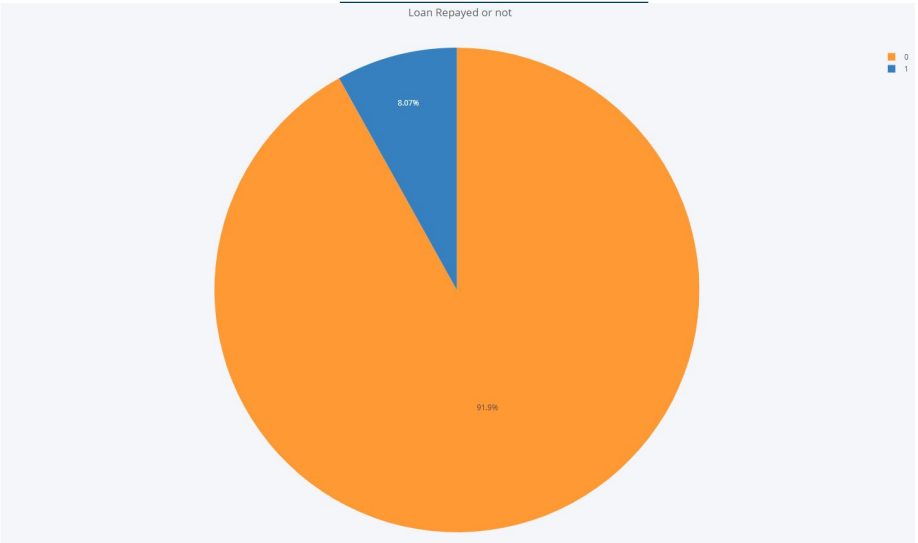
Exploratory Data Analysis

- An open-ended process where we calculate statistics and make figures to **find trends, anomalies, patterns, or relationships** within the data

Objectives:

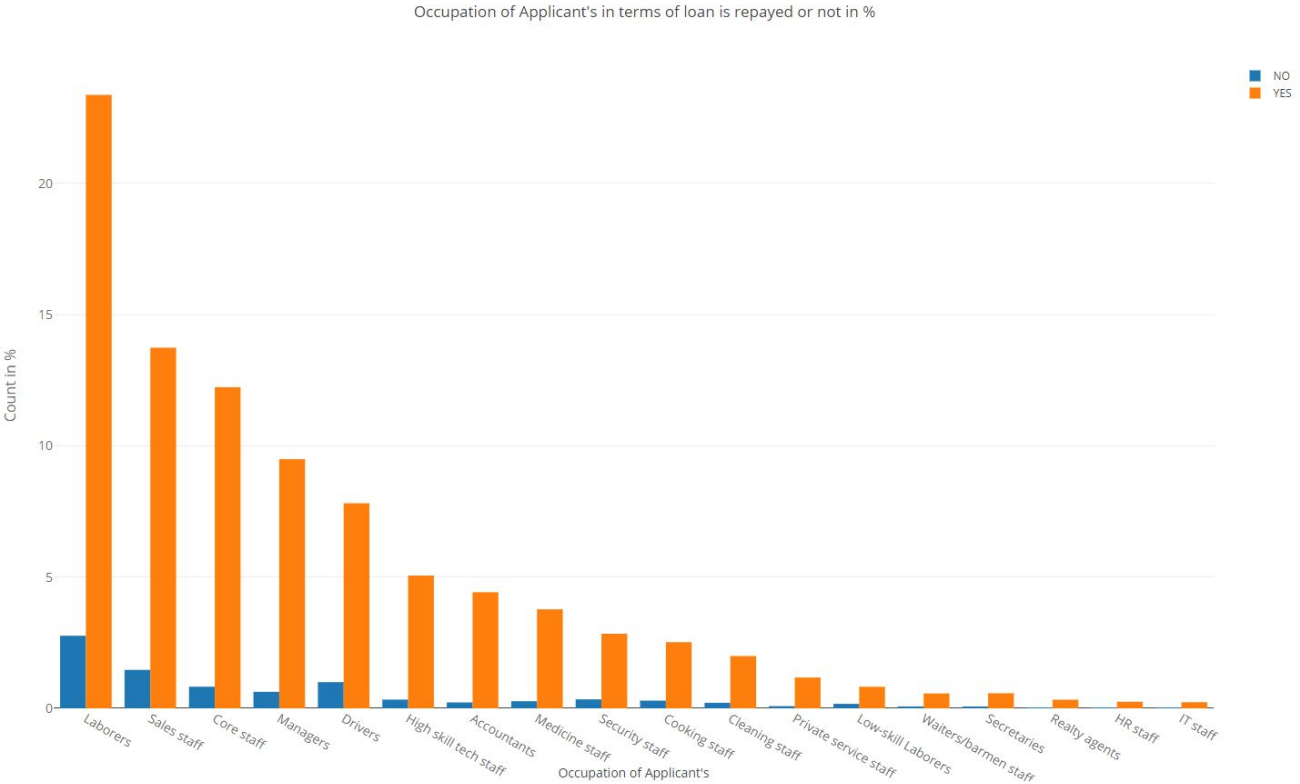
- To learn what the data can tell us
- Make decision on our model choices by helping us determine which **features** to use

EDA - Response Variable



Characterisics

- 0 for loan repaid on time
- 1 for clients having payment difficulties
- Highly imbalanced: Much more 0s than 1s
- Weighting is preferred

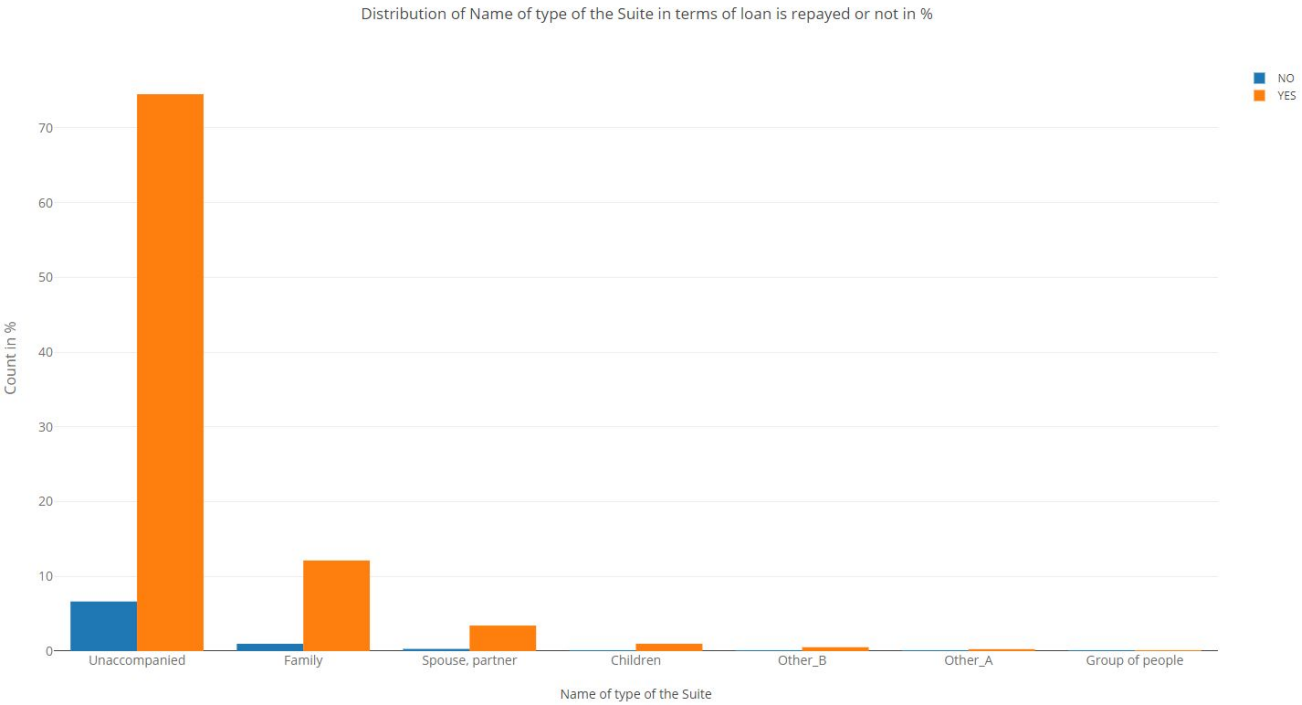


EDA

Occupation of Applicants

Applicants of most occupation are able to repay loan on time.

Among all, laborers, sales staff and drivers contain the highest proportion of applicants who are unable to repay loan on time



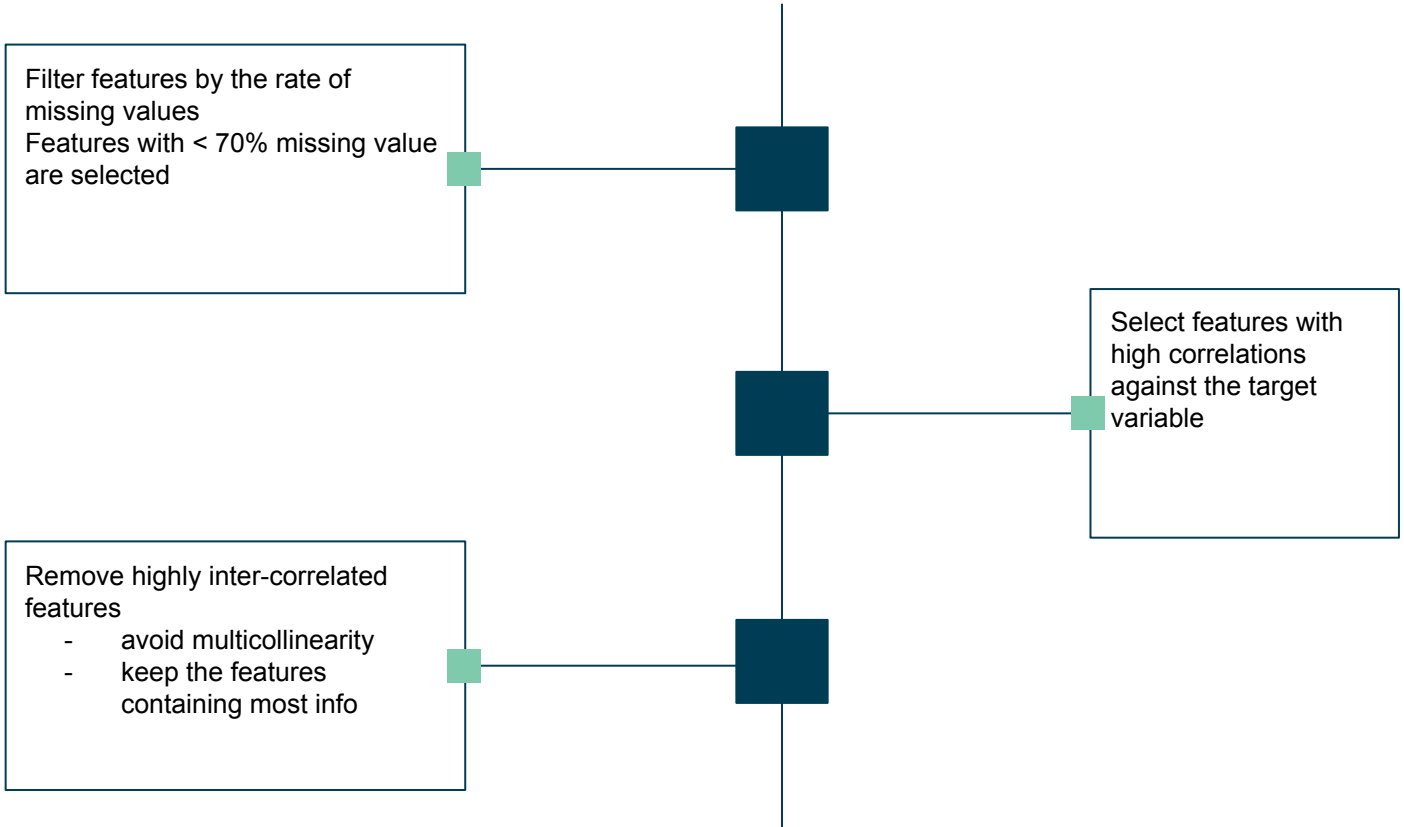
EDA

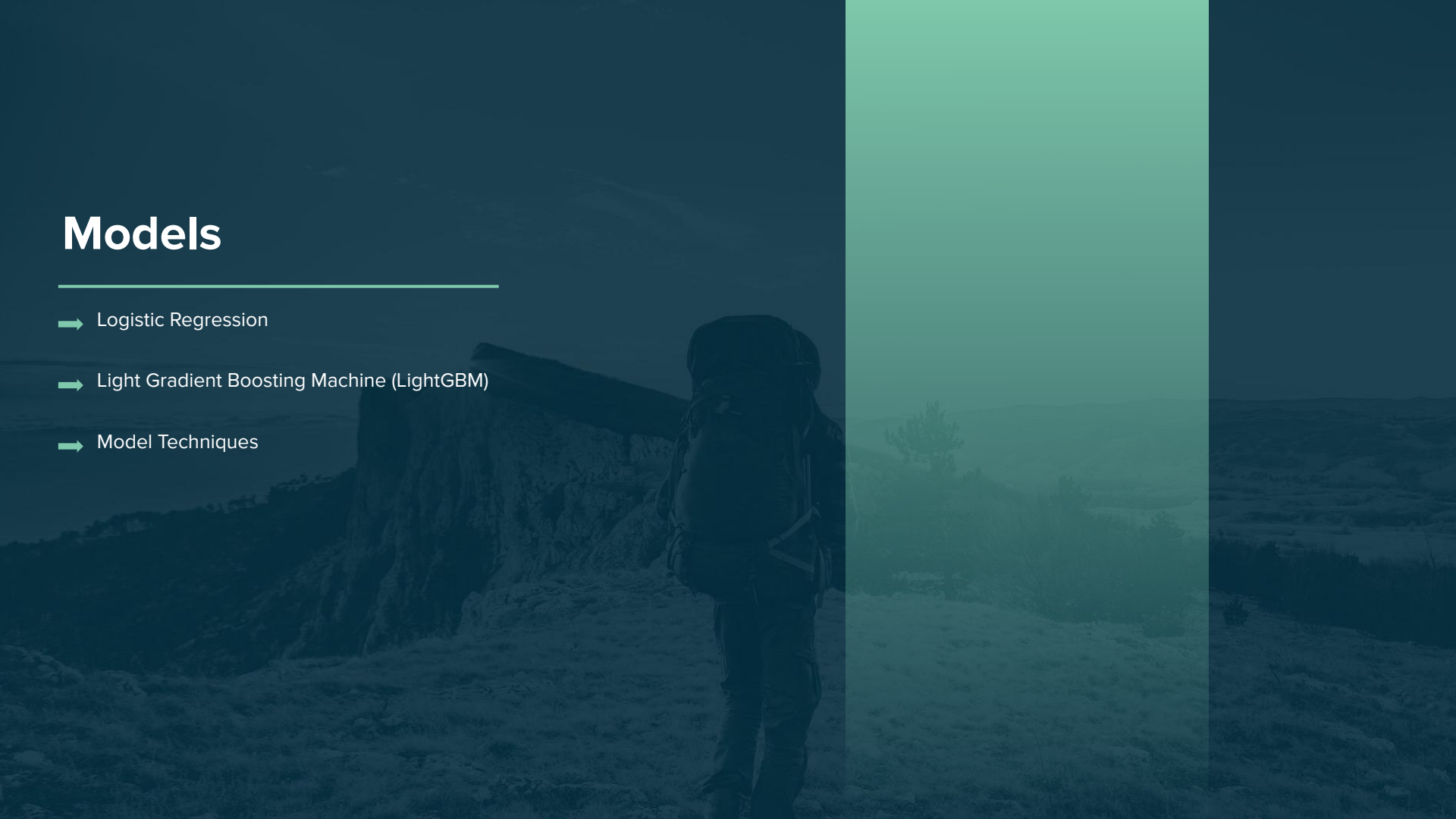
Type of the Suite

Applicants who are unaccompanied, have the highest chance of default

Those who live with family, spouse or children are more likely to repay loan. This could be explained by support from family members

Feature Selection





Models

- Logistic Regression
- Light Gradient Boosting Machine (LightGBM)
- Model Techniques

1. LOGISTIC **REGRESSION**

Useful for classifying categorical response values

- Non-linear transformation to estimate probabilities between 0 and 1
- Maximum likelihood estimation used to find parameters in model

Model assumptions

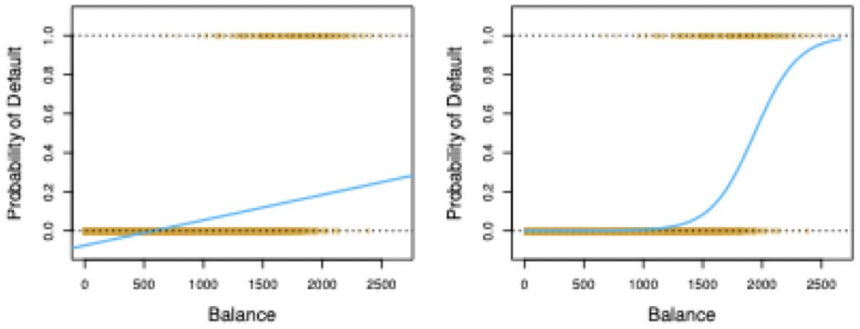
- No outliers in the data
- No multi-collinearity between the predictors
- Linearity between dependent variable(Y) and the indepdent variables(X)

Advantage of this model

- Easy to implement, interpret and efficient to train
- Less-inclined to over-fitting

Disadvantage of this model

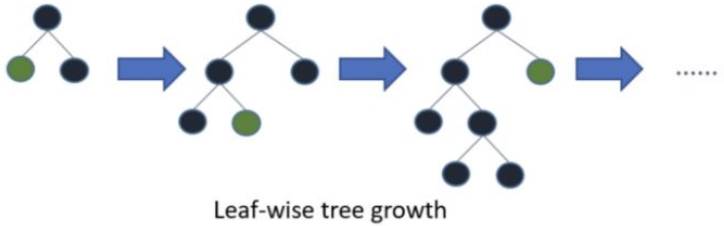
- Low model complexity to obtain complex relations
- Lower accuracy for high-dimensional dataset



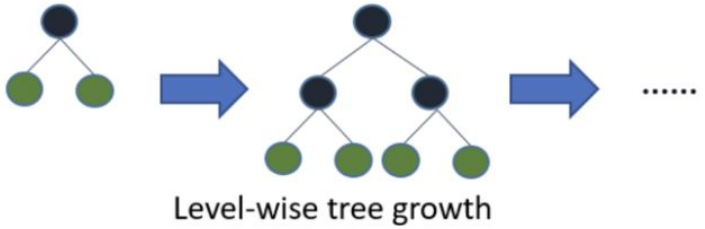
2. LIGHT GRADIENT BOOSTING MACHINE **(Light GBM)**

Revolutionary gradient boosting decision tree(GBDT) method

- Developed by Microsoft in 2017, the most popular framework in Kaggle
- Fast, distributed, high-performance gradient boosting framework based on decision tree algorithm
- GBDT: ensemble model to learn decision by fitting residual errors in each iteration
- Source code(in C++): https://github.com/microsoft/LightGBM
- An improved model compared to XGBoost in terms of hardware and algorithms
- LGBM is Leaf-Wise Tree Growth, compared to Level-Wise Tree Growth in XGBoost
- User can minimize overfitting by limiting tree depth or number of leaves



Leaf-wise tree growth



Level-wise tree growth

2. LIGHT GRADIENT BOOSTING MACHINE **(Light GBM)**

Gradient Based One Side Sampling (GOSS) & Exclusive Feature Bundling(EFB)

- We want to select the Best Split with instances of larger gradients as they are undertrained
- XGBoost uses a histogram-based algorithm with O(#bins * #feature) times, but still needs O(#data * #feature) to construct histogram
- GOSS is a novel sampling method to retain large gradient instances, while keeping the data distribution by random sampling on small gradient instances
- EFB is used to bundle mutually exclusive features to reduce the complexity to O(#data * #bundle) where #bundle << #feature
- Takeaway: LightGBM adopts an effective way to split the decision tree based on gradient

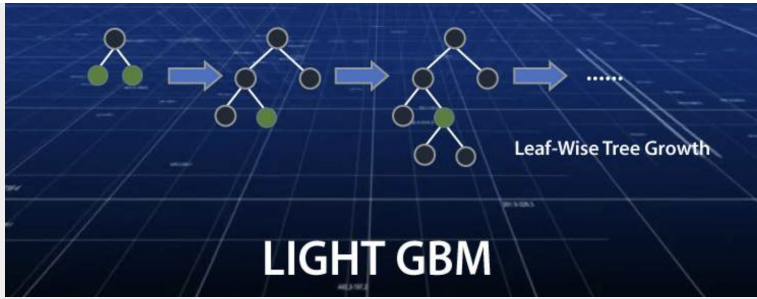
Algorithm 3: Greedy Bundling

Input: F: features, K: max conflict count
Construct graph G
searchOrder $\leftarrow$ G.sortByDegree()
bundles $\leftarrow$ {}, bundlesConflict $\leftarrow$ {}
for i ***in*** *searchOrder* **do**
    needNew $\leftarrow$ True
    for $j = 1$ ***to*** *len(bundles)* **do**
        cnt $\leftarrow$ ConflictCnt(bundles[j],F[i])
        if *cnt + bundlesConflict[i]* $\leq K$ **then**
            bundles[j].add(F[i]), needNew $\leftarrow$ False
            break
    if *needNew* **then**
        Add $F[i]$ as a new bundle to *bundles*
Output: *bundles*

2. LIGHT GRADIENT BOOSTING MACHINE **(Light GBM)**

Advantages of LightGBM

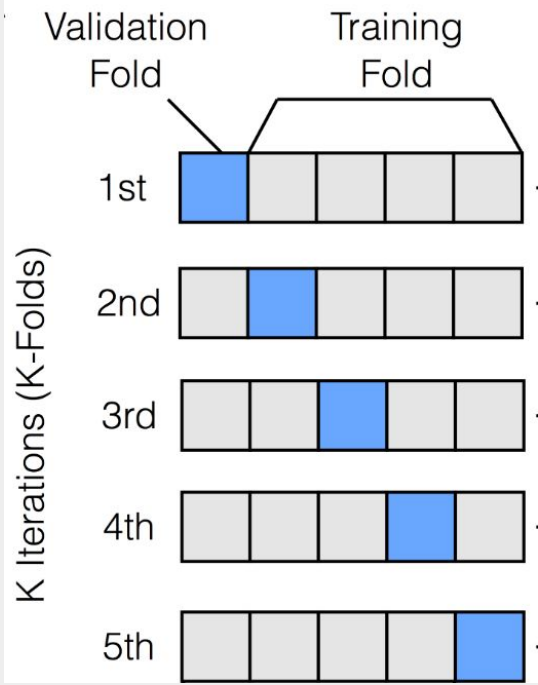
- Faster training speed and higher efficiency
- Lower memory usage
- Better accuracy than any other boosting algorithm
- Parallel learning supported
- Support categorical variables
- Compatible with large datasets



3. Model Techniques

Avoid overfitting

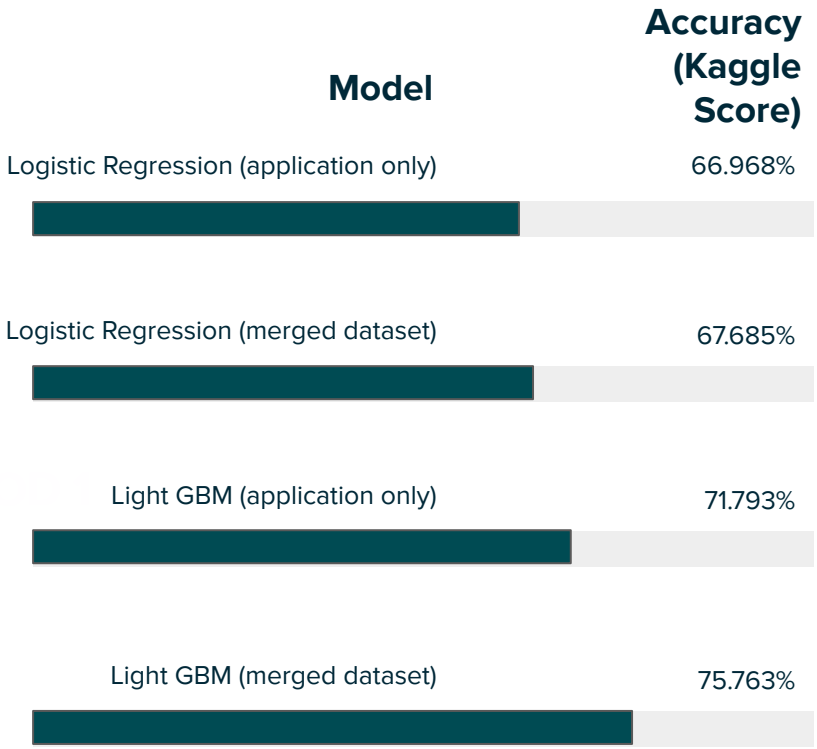
- 5-fold validation
- LASSO/Ridge Regularization
- Early Stopping
- Fine tune and grid search for hyperparameters



RESULTS

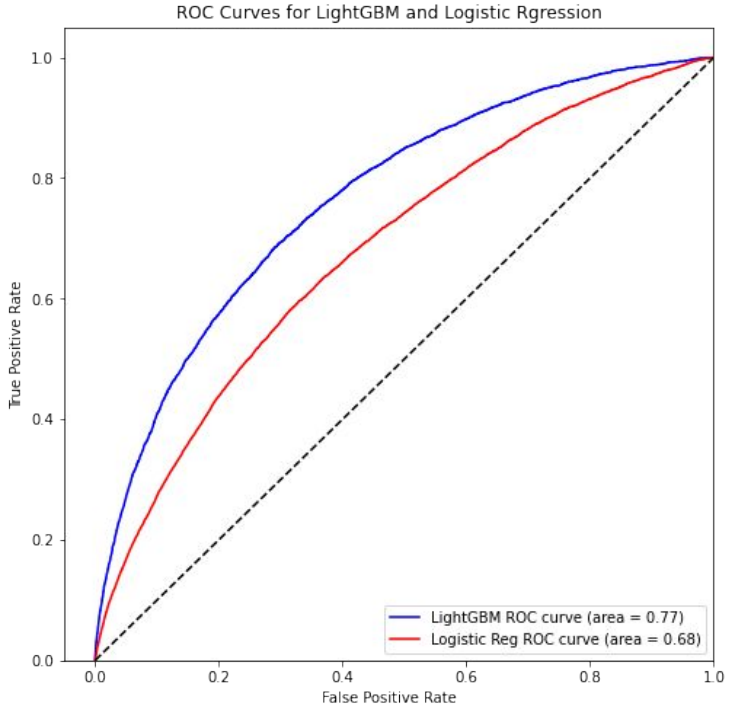
Results

The following models and architectures are used as trials. The **Light GBM** with merged dataset is chosen as our model as it achieves the highest accuracy



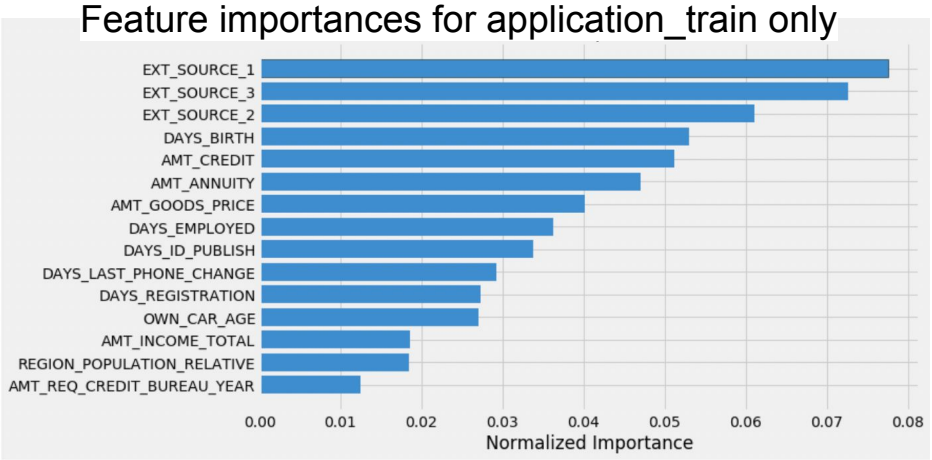
Results

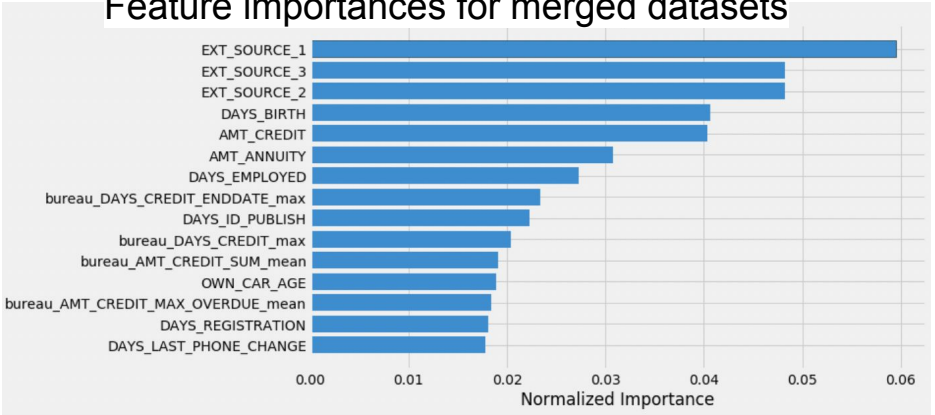
Light GBM can achieve higher accuracy than logistic regression in merged dataset scenario



Results

EXT_SOURCE_1 & 2 & 3 are the most important features.
They are credit scores provided by external organizations.
DAYS_BIRTH is the age of the lender
AMT_CREDIT is credit amount of the loan
bureau_DAYS_CREDIT_ENDDATE is the remaining duration of outstanding credit at the time of application in Home Credit recorded by Credit Bureau





CONCLUSION

Conclusion

Number of features, Model selection

1) For both models, the prediction accuracy increases along with the number of features
2) Light GBM has better performance than logistic regression in terms of
   a) Model complexity
   b) Ability to extract important features


Thank You