# Home Credit Default Risk Assessment Based on LightGBM

Hongen Tang, Zilong Pan, Xinyu Xu

——5440 Artificial Intelligence in Fintech

# CONTENTS

01

PART

# INRRODUCTION

# INTRODUCTION

### MOTIVATION
Construct a home credit risk prediction model aiming to accurately assess clients' repayment capacity.

### DATASET
11 datasets introducing clients’ personal information and credit record.

### CORE PROBLEM
Improve data utilization

——feature engineering

Enhance model performance

——model training

### FURTURE ANALYSIS
Application scenario

Performance requirements

……

| Submission and Description | Private Score | Public Score | Selected |
|---|---|---|---|
| **submission_no_stacking.csv**<br>Complete (after deadline) · now | **0.76821** | **0.76568** | ☐ |
| **submission_stacking_fe_shap.csv**<br>Complete (after deadline) · 1h ago | **0.76887** | **0.76395** | ☐ |

| Submission and Description | Private Score | Public Score | Selected |
|---|---|---|---|
| **lgbm_submission_tuned_params.csv**<br>Complete (after deadline) · now | **0.78880** | **0.79260** | ☐ |

02

PART

# DATA PREPROCESSING

# DATA PREPROCESSING

01 Missing value handling

02 Outliers handling

03 One-Hot Encoding

# DATA PREPROCESSING

## MISSING VALUE HANDLING

1. Deletion
2. Specific Value handling( e.g. 365243 )
3. Filling based on business logic
4. Filling based on numerical stability
5. Populate based on data category
6. remove illegal characters

# DATA PREPROCESSING

01 Missing value handling

02 Outliers handling

03 One-Hot Encoding

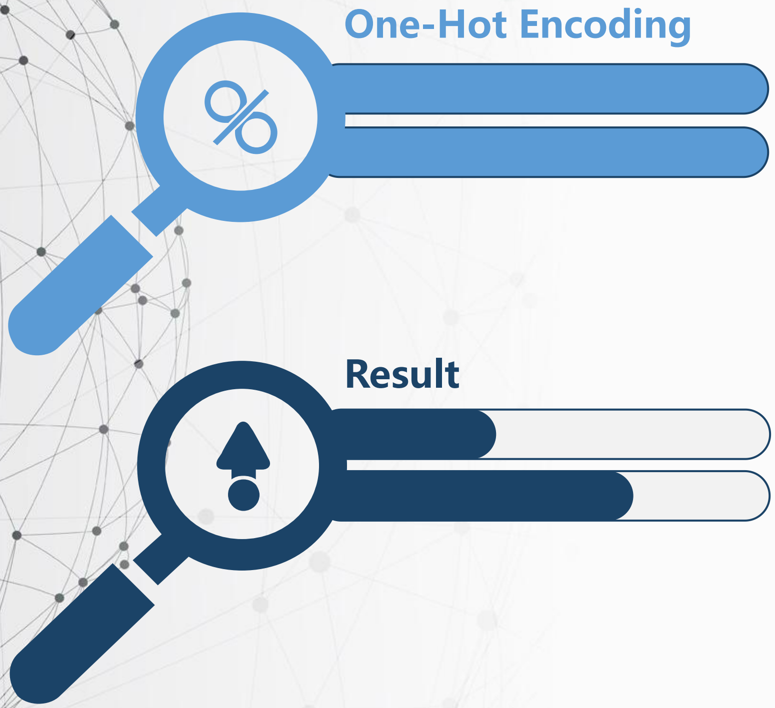
# DATA PREPROCESSING

01 Missing value handling

02 Outliers handling

03 One-Hot Encoding

# DATA PREPROCESSING



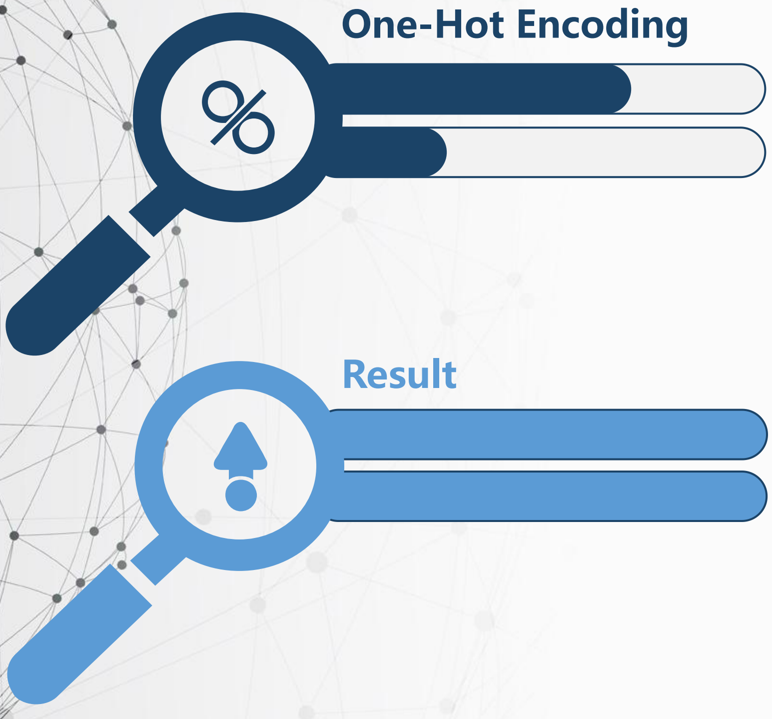

Since the raw dataset contains non-uniform data types that cannot be directly fed into the model, one-hot encoding was applied to all categorical variables using the Pandas pd.get_dummies() method to achieve numerical conversion.

# DATA PREPROCESSING

| NAME_EDUCATION_TYPE |
| --- |
| Secondary / secondary special |

| NAME_EDUCATION_TYPE_Secondary / secondary special | NAME_EDUCATION_TYPE_Higher education | NAME_EDUCATION_TYPE_Incomplete higher | NAME_EDUCATION_TYPE_Lower secondary | NAME_EDUCATION_TYPE_Academic degree |
| --- | --- | --- | --- | --- |
| 1 | 0 | 0 | 0 | 0 |

# DATA PREPROCESSING




After coding, all features are indexed by customer ID and merged with the main table through aggregation functions (such as mean, Max, min, count) to form a unified and fully numerical training set. This provides a data base with consistent structure and high quality for the subsequent feature screening and modeling.

03

PART

# FEATURE ENGINEERING

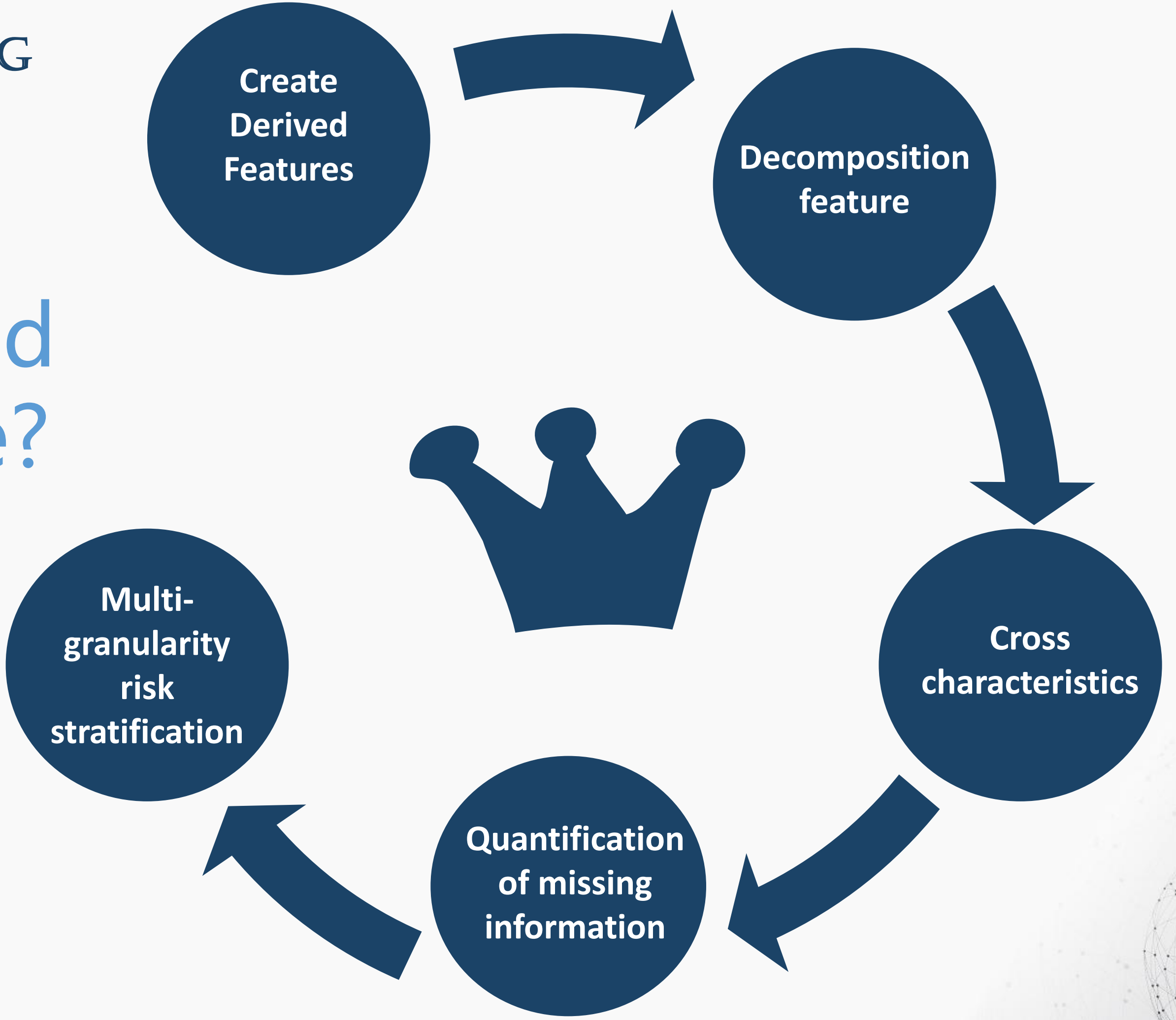
# FEATURE ENGINEERING

## WHY FEATURE ENGINEERING?

The existing indicators have the problems of low information density, vague meaning, and low matching degree with the research object. In order to further improve the value of the original data and improve the prediction ability of the indicators to the target problem, the indicators are further constructed through feature engineering on the basis of the original data.

# FEATURE BUILDING

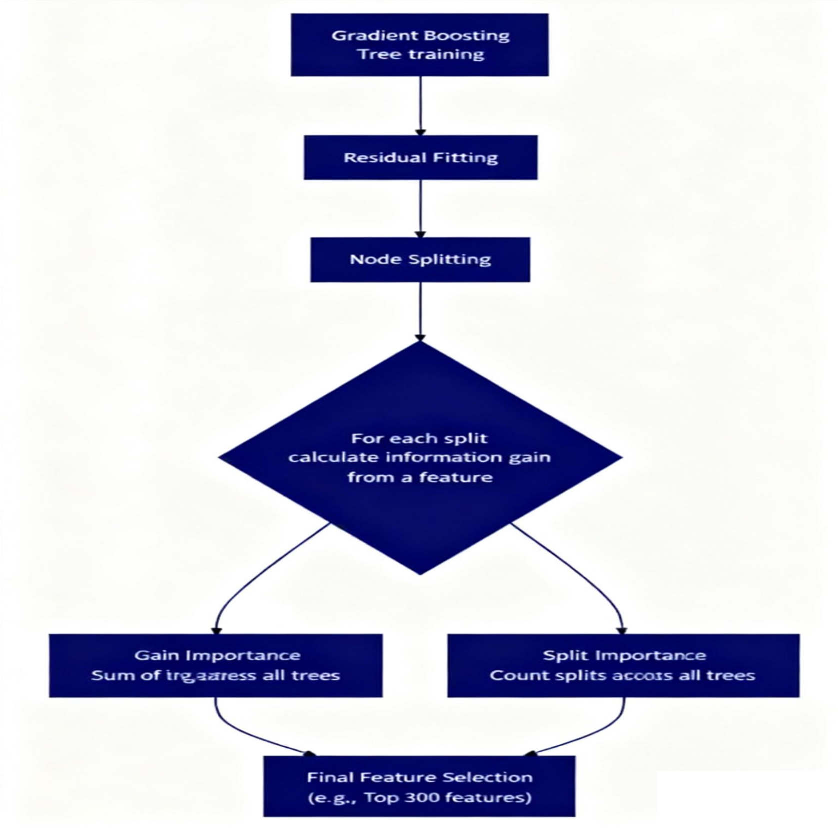
## How to build new feature?



# KEY DRIVERS OF DEFAULT RISK

| Feature | Importance |
|---|---|
| CREDIT_ANNUITY_RATIO | 355.40 |
| CREDIT_GOODS_RATIO | 210.60 |
| EXT_SOURCES_MEAN | 209.00 |
| NAME_FAMILY_STATUS_Married | 197.60 |
| AMT_ANNUITY | 185.80 |
| REGION_POPULATION_RELATIVE | 175.60 |
| EXT_SOURCES_PRODUCT | 171.60 |
| BUREAU_DAYS_CREDIT_MAX | 161.40 |
| EXT_INCOME_INTERACTION | 158.00 |
| EXT_SOURCE_3 | 146.40 |
| OBS_30_CNT_SOCIAL_CIRCLE | 143.25 |
| OWN_CAR_AGE | 143.20 |
| CODE_GENDER_F | 142.40 |
| BUREAU_DAYS_CREDIT_ENDDATE_MAX | 140.40 |
| INS_PAYMENT_RATIO_STD_MIN | 140.00 |
| EXT_SOURCES_MIN | 138.40 |
| CODE_GENDER_M | 131.40 |
| PREV_REFUSED_COUNT | 129.20 |
| BUREAU_DEBT_TO_CREDIT_RATIO | 124.60 |
| WORK_START_AGE | 120.80 |

# FEATURE SELECTION




Arrange all features in descending order of importance, and draw the cumulative contribution rate curve. The results show that the contribution of the first 300 features has exceeded 95%. Therefore, in this study, the retention threshold was set to the top 300 high-importance features, and the remaining features were excluded.

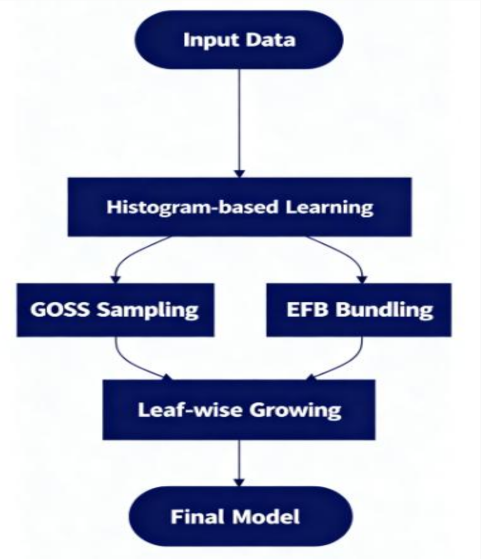
04

PART

# MODEL BUILDING

# MODEL SELECTION

# WHY LightGBM?




## Model Performance

LightGBM is good at binary classification problem based on structural data and it performs well on medium-sized data with a scale of 300000.

## Financial Data Recognition

The tree model can automatically capture the nonlinear relationship and interaction effects between variables and handle different types of feature mixing, so that it can perform well on this data set.

## Practical Applications

The wind control system in the financial industry generally uses the gradient lifting tree as the core model. LightGBM has the advantages of fast prediction, small memory footprint and high training efficiency.

# MODEL TRAINING

### 01 Sample unbalance handling

In order to improve the prediction ability of the model for the minority class, balance the proportion of positive and negative samples by SMOTE.

### 02 K-Fold Cross-Validation Training

The average performance index is calculated through five different data partitions to investigate the actual effect of the model.

### 03 Overfitting control

In the search space of Optuna, consider key parameters controlling the complexity of the model  and the L1/L2 regularization terms.

### 04 Automatic parameter adjustment

The Optuna toolkit is used to automatically find the best parameter and the machine learning model. The threshold of feature screening is also used as the hyperparameter.

# MODEL TRAINING

| Parameter | Value |
|---|---|
| feature_selection_threshold | 0 |
| learning_rate | 0.0468 |
| num_leaves | 33 |
| max_depth | 12 |
| min_child_samples | 72 |
| subsample | 0.7116 |
| colsample_bytree | 0.7563 |
| reg_alpha | 0.2442 |
| reg_lambda | 0.8161 |

Best parameters of the model

05

PART

# MODEL EVALUATION

# WHY AUC as SINGLE Evaluation Index?

### Business target matching

In practical application, the bank will set an approval threshold, and X% of the customers after the model score ranking will be rejected or manually reviewed, while AUC can well measure the ranking ability of the model .

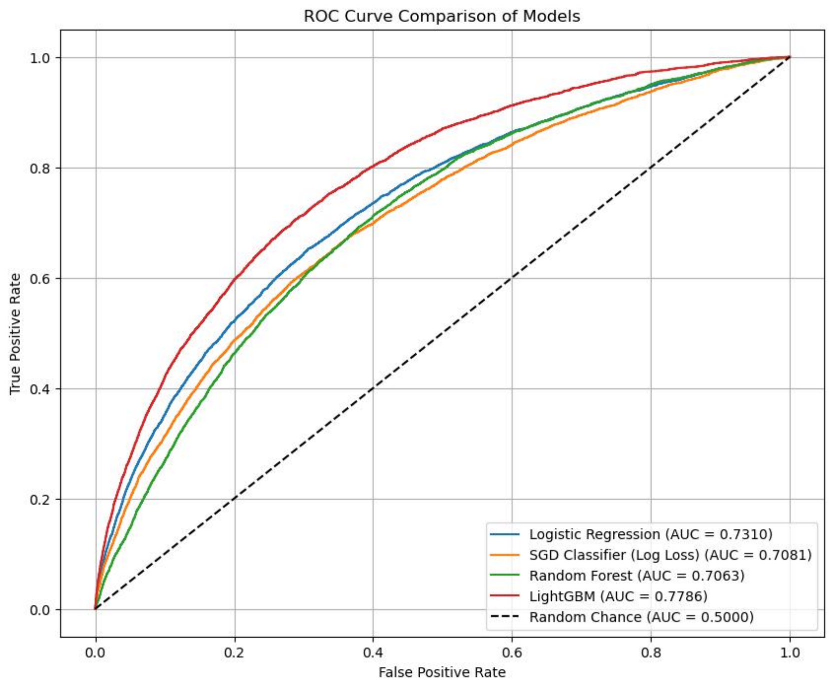
### Robust model evaluation

As credit data is a typical highly unbalanced data, the reference value of accuracy is greatly reduced, while precision and recall as a single indicator have their own drawbacks.

### Comprehensive decision support

It provides a complete perspective for business decision makers to select the classification threshold through the ROC curve, allowing companies to choose different strategies with their own risk preferences.

# CONCLUTIONS




The final running result of the model is AUC = 0.7859,

which proves that the model has a good prediction effect.

# CONCLUTIONS

| Feature | Importance |
|---|---|
| CREDIT_ANNUITY_RATIO | 355.40 |
| CREDIT_GOODS_RATIO | 210.60 |
| EXT_SOURCES_MEAN | 209.00 |
| NAME_FAMILY_STATUS_Married | 197.60 |
| AMT_ANNUITY | 185.80 |
| REGION_POPULATION_RELATIVE | 175.60 |
| EXT_SOURCES_PRODUCT | 171.60 |
| BUREAU_DAYS_CREDIT_MAX | 161.40 |
| EXT_INCOME_INTERACTION | 158.00 |
| EXT_SOURCE_3 | 146.40 |
| …… | …… |

After K-fold cross-validation,

the 20 features with the highest contribution to the model

# FURTHER ANALYSIS

## Loan repayment ability is the decisive factor

CREDIT _ ANNUITY _ RATI (1st)
CREDIT _ GOODS _ RATIO (2nd)
AMT _ ANNUITY (5th)

## External credit scoring systems are important

EXT_SOURCES_MEAN (3rd)
EXT_SOURCES_PRODUCT (7th)
EXT_SOURCE_3 (10th)
EXT_SOURCES_MIN (16th)

## Historical credit behavior has strong predictive power

BUREAU_DAYS_CREDIT_MAX
BUREAU_DAYS_CREDIT_ENDDATE_MAX
BUREAU_DEBT_TO_CREDIT_RATIO
PREV_REFUSED_COUNT

# THANKS