MAFS 6010Z Warm-up Project: Home Credit Default Risk

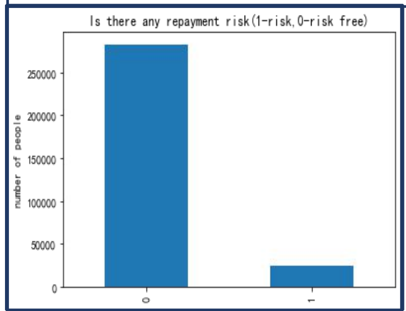
MA Rongyue 20826086; Peng Junkai 20756772
YE Mengxiang 20799762; NI Xiaohan 20825846

1. Introduction

We try to predict the repayment abilities of clients by using machine learning algorithms through information including historical loan records and the characteristics of loan users. To achieve the goal, we use two main datasets, after converting and cleaning data by label encoding and other measures, we use EDA, feature engineering methods. Then, we choose Light GBM as the primary model, and use the model to do assessment and get the feature importance.

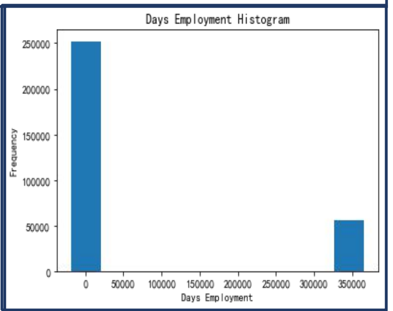
2. Exploratory Data Analysis

In this part, we will devide into two sections, first one is the distribution of target variables, and the second one is about missing values and outliers examination.



Distribution of target variable

According to the graph, we could see that there are far more loans that were repaid on time than loans that were not repaid. There are approximately 24,000 applicants(about 8% of the total loans) in the training data involved clients who don't have abilities to repay the loan.



Missing values and outliers examination

1.Drop all features which have more than 60% missing data.
2.Analyze text columns by label encoding and other methods.
3.Analyze numeric columns and drop abnormal data.

3. Feature Engineering

Domain knowledge feature:

We try to incorporate our financial knowledge by creating ratios to derive more valuable information from simple numerical data.

1. Credit Term (TERM = AMT_CREDIT / AMT_ANNUITY)
2. % of the total loan amount relative to the applicant's income
3. % of the monthly payment relative to the applicant's income
4. % of employment duration relative to age

Missing data pattern

In our perspective, we can derive some insights from the missing data pattern as people are more likely to skip the optional field when their answers are negative.
We regard the missing data pattern as a new indicator. We create a new binary indicator called "Incomplete" which will be flagged when there are more and equal to 40 blanks in their applications.

6. References

[1] oskird. (2018, June 21). EDA + baseline model using application. Kaggle. Retrieved September 19, 2021, from https://www.kaggle.com/sz8416/eda-baseline-model-using-application?scriptVersionId=4207867&amp;cellId=1.

[2] Koehrsen, W. (2018, July 31). Introduction to manual feature engineering. Kaggle. Retrieved September 19, 2021, from https://www.kaggle.com/willkoehrsen/introduction-to-manual-feature-engineering?scriptVersionId=4852087&amp;cellId=1.

4. Model Training

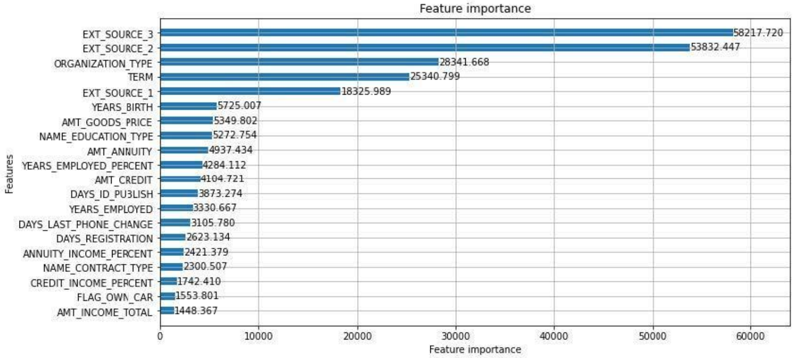
Cross-Validation :5-fold Cross-Validation to make LB (Leader board) score more similar to valid score

Light GBM: The objective of feature selection is to remove redundant variables to optimize our data set. We also set early stopping parameters to avoid overfitting. Early stopping means the training will stop when the performance of the validation is not improving after the last early stopping round. For our model, we set num_iterations as 4000 and early_stopping_round as 100.

5. Assessment

Best iteration:

Training's AUC: 0.812852
Valid_1's AUC:0.765925
AUC of Fold 5: 0.766
Valid score: 0.7644
Kaggle LB score: 0.65006



7. Contribution

Please find this part in our report