MATH 4995 Mini-Project 1: Home Credit Default Risk

Liao Yi-han yliaoag@ust.hk
Department of Data Science and Technology, HKUST

1. Introduction

To help people who struggle to get loans because of insufficient credit history, Home Credit would like to evaluate their clients' repayment ability via other alternative data. With this new criteria, Home Credit gives more loans to those people but also guarantee a minimum probability of defaulting.

2. Home Credit Risk Dataset (Kaggle Competition)

This dataset contains five tables which represent different financial data for Home Credit to evaluate applicants' repayment ability.

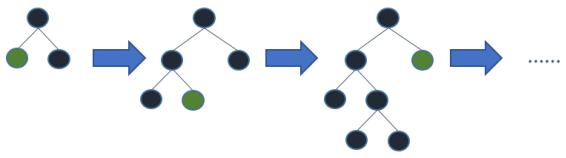
Data Preprocessing

- Convert categorical variables to dummy variables
- Complete the missing values via the imputation transformer
- Combine features from different tables

3.1 **Methodology -** Light Gradient Boosting Machine

Light GBM is a highly efficient gradient decision tree. Different from other tree learning algorithms growing trees by level(depth)-wise, light GBM grows tree leaf-wise which selects the leaf with maximum delta loss to grow. Besides, light GBM is good at dealing with a large and sparse dataset.

One useful function of light GBM is its important features scores, which helps to find out the features who have significant impact on the model.



Leaf-wise tree growth

3. 2 **Methodology – Linear Discriminant Analysis**

Based on Bayes theorem, LDA classifies a subject into the class which has the largest conditional density. In assumption, the model fits a Gaussian density to each class with same covariance matrix for all classes.

4. Prediction

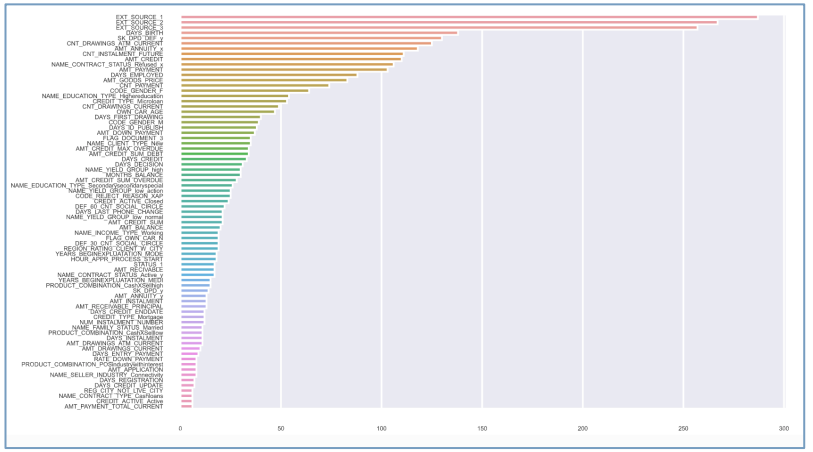
1. Train the light GBM classifier on application table and previous application table individually with 5-fold cross validation
2. Select 30 most important features from application table and 20 most important features from previous application table
3. Combine these features selected from two tables, train the light GBM classifier on this combined features with 5-fold cross validation
4. Load data from other tables (POS_CASH_balance, installments_payments, credit_card_balance, bureau), and then merge these tables with the 50-feature data above
5. Train the new dataset with 5-fold validation, and then do the prediction.
6. Get the 75 most important features in this model, use these features to train the LDA classifier for prediction

5. Analysis

With light GBM, it is efficient to know which features are important when analyzing a large and sparse dataset with high-dimension features space. It provides some concrete features as reference to evaluate client's repayment ability.

As the results show, extracting features from more tables will improve the score and no need to use all the features in one table.

In tradition, LDA is sometimes a leading model in default risk prediction. After selecting some significant features, fitting them in LDA model for further prediction gets an equivalent score as light GBM but using less features input.



Top 75 important features in all the tables

6. Results

Dataset & Features	Kaggle Score
application data	0.72696
application data (30 feats) + previous application (20 feats)	0.73885
all table data	0.75249
LDA with top 75 important features	0.75523

7. References

Microsoft Research, Peking University, Microsoft Redmond, Meng, Q., Finley, T., Wang, T., … Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree