MSBD5013 project - Home Credit Default Risk

Tian CHEN[1], Yi LIU[2] and Zhixuan PENG[3]{tchenaz[1], yliujt[2], zpengan[3]}@connect.ust.hk

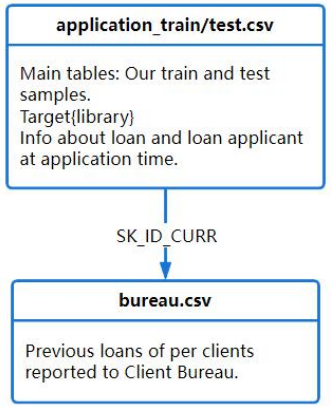
1. Introduction

In order to make sure this underserved population such as unbanked has a positive loan experience, Home Credit makes use of a variety of alternative data--including telco and transactional information--to predict their clients' repayment abilities. We design a model based on feature engineering and machine learning algorithms to predict whether or not an applicant will be able to repay a loan.

2. Data Pre-processing

- **Dataset**

The data is provided by Home Credit, a service dedicated to provided lines of credit (loans) to the unbanked population.There are 7 different sources of data, we only use the main application training and testing data and bureau.csv to generate 6 features having highest correlation with target: EXT_SOURCE_1,EXT_SOURCE_2,EXT_SOURCE_3,DAYS_BIRTH,ACTIVE_LOANS_PERCENTAGE,DAYS_EMPLOYED,TARGET.



- **Exploratory Data Analysis**

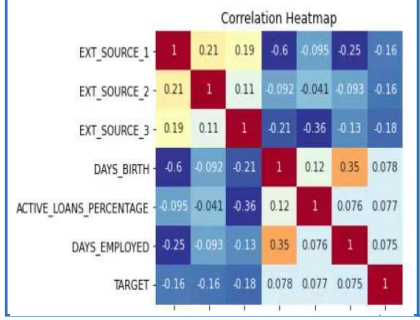
We examine the distribution of the target column, missing values, and column types, and do data cleaning using quantile to the anomalies after encoding and generate 8 features from bureau.csv to application_train.csv.

- **Encoding Categorical Variables methods**

We use Label Encoding for any categorical variables with only 2 categories and One-Hot Encoding for any categorical variables with more than 2 categories. And align the train data with test data.

3. Feature Engineering

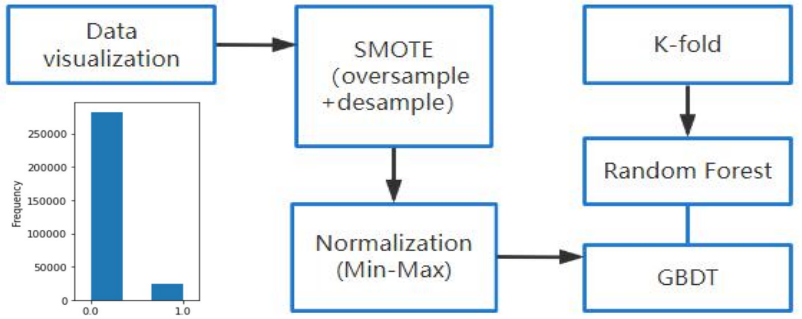
- **Data Correlation**



- **Polynomial features**

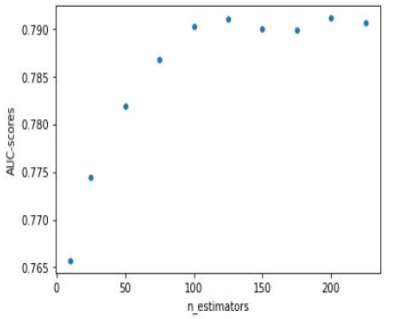
We generate the squares of the selected 6 features, one of which comes from the bureau.csv, as well as the interaction terms between them. The generated data with total 21 attributes serves as supplement to the main training data set.Then the total number of attribute is 140.

4. Model Design



parameter settings

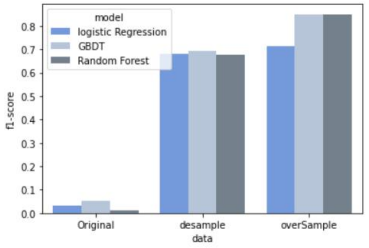
Random Forest	
estimators	50
max_depth	10
criterion	"entropy"
GBDT	
estimators	100
learning rate	0.1



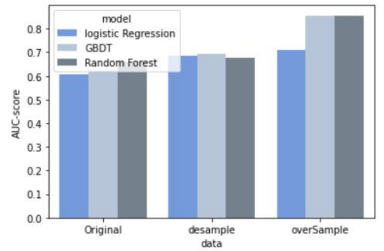
hyperparameters selection

5. Experiment Results

From visualization we find the data is highly skewed. To solve this problem we desample the majority and oversample the minority using SMOTE. Decision Tree can automatically select and group features but is a weak learner that is easy to overfit. We choose two ensemble models, Random Forest and GBD, to obtain both low variance and low bias.



f1 scores on different sampling methods



AUC scores on different sampling methods

6. Conclusion

In this project, we employed a number of feature selection methods. These methods are necessary to reduce the number of features to increase model interpretability, decrease model runtime, and increase generalization performance on the test set. For further improvement, we might actually try to add more features, instead of naively applying aggregations, think about what features are actually important from a domain point of view.

7. References

https://www.kaggle.com/willkoehrsen/start-here-a-gentle-introduction

8. Contribution

Data Processing
- Tian.CHEN

Model Design
- Zhixuan,PENG

Poster Design
- Yi,LIU