Math 5470: home credit kaggle challenge

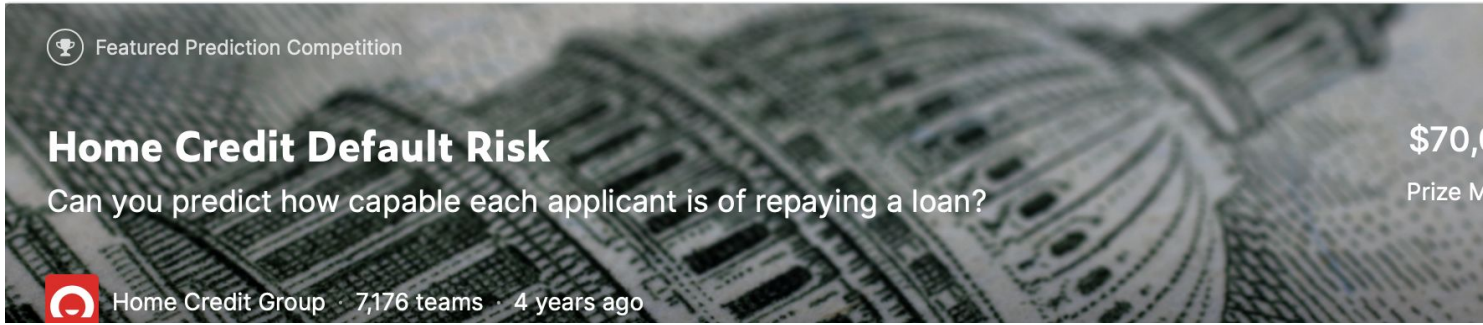
Haoran Li, Jiaxin Bai, Qi Hu, Ying Su
2022.5.2

https://youtu.be/JEQ6ghFXJEY

Outline

- Introduction
- Problem formulation
- Feature preprocessing
- Feature engineering
- Model and ensemble
- Analysis

Introduction



- Problem
  - Insufficient or non-existent credit histories make many people difficult get loans

- Goal
  - Home Credit tries strives to provide financial services for unbaned population
  - Annouce Kaggle competition to fully tap the potential of the data

Problem Fomulation

- Data
  - Application_train(test).csv: **Main training and testing data** with information about each loan application
  - Bereau.csv and bereau_balance.csv: The clients' previous credits and monthly data about the previous credits from **other financial institutions**
  - Previous_application.csv: Information about **previous loan applications** in Home Credit.

- Task formulation
  - A supervised learning system
    - Feature preprocessing
    - Feature engineering
    - Model learning

Feature Preprocessing

- Missing Values
  - Columns dropping
  - Imputation
- Individual Records
  - Data Aggregation
- Categorical Values
  - one-hot encoding

Feature Preprocessing

- Missing Values
  - Columns dropping: drop out the whole column when the number of missing values exceeds the given threshold
  - Imputation: after columns dropping, adding the median values to those missing values.

AMT_REQ_CREDIT_BUREAU_DAY	AMT_REQ_CREDIT_BUREAU_WEEK	AMT_REQ_CREDIT_BUREAU_MON	AMT_REQ_CREDIT_BUREAU_QRT	AMT_REQ_CREDIT_BUREAU_YEAR
0.0	0.0	0.0	0.0	1.0
0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0
NaN	NaN	NaN	NaN	NaN
0.0	0.0	0.0	0.0	0.0

Feature Preprocessing

- Individual Records

  Features with individual records can be summarized when modelling a single user.

  - Data Aggregation: For rows with the same “SK_ID_CURR”, we group these records together to obtain the summary statistics like max,min,mean and sum.

	SK_ID_BUREAU	MONTHS_BALANCE	STATUS
0	5715448	0	C
1	5715448	-1	C
2	5715448	-2	C
3	5715448	-3	C
4	5715448	-4	C

	SK_ID_CURR	client_bureau_balance_MONTHS_BALANCE_count_count
0	100001	7
1	100002	8
2	100003	0
3	100004	0
4	100005	3

Feature Preprocessing

- Categorical Values
  - one-hot encoding

SK_ID_CURR	Loan type
1	home
1	home
1	home
1	credit
2	credit
3	credit
3	cash
3	cash

SK_ID_CURR	credit count	cash count	home count	total count
1	1	0	3	4
2	1	0	0	1
3	1	2	0	3

Feature Engineering

Feature selection:

- Correlation-based feature selection

  We first compute the correlations between the input and output variables, and remove those variables that have a low correlations.

- Backward elimination

  We first train the model with all features, then compute the feature importance. After this, the features that do not have significant effects on the output are eliminated. Finally we use the remaining features to train the model again. This process can also be conducted multiple times until all the parameters are statistically significant.

Feature Engineering

- Feature Generation
  - Polynominal Features

	TARGET	EXT_SOURCE_1	EXT_SOURCE_2	EXT_SOURCE_3	DAYS_BIRTH
TARGET	1.000000	-0.155317	-0.160472	-0.178919	-0.078239
EXT_SOURCE_1	-0.155317	1.000000	0.213982	0.186846	0.600610
EXT_SOURCE_2	-0.160472	0.213982	1.000000	0.109167	0.091996
EXT_SOURCE_3	-0.178919	0.186846	0.109167	1.000000	0.205478
DAYS_BIRTH	-0.078239	0.600610	0.091996	0.205478	1.000000

```
['1',
 'EXT_SOURCE_1',
 'EXT_SOURCE_2',
 'EXT_SOURCE_3',
 'DAYS_BIRTH',
 'EXT_SOURCE_1^2',
 'EXT_SOURCE_1 EXT_SOURCE_2',
 'EXT_SOURCE_1 EXT_SOURCE_3',
 'EXT_SOURCE_1 DAYS_BIRTH',
 'EXT_SOURCE_2^2',
 'EXT_SOURCE_2 EXT_SOURCE_3',
 'EXT_SOURCE_2 DAYS_BIRTH',
 'EXT_SOURCE_3^2',
 'EXT_SOURCE_3 DAYS_BIRTH',
 'DAYS_BIRTH^2']
```

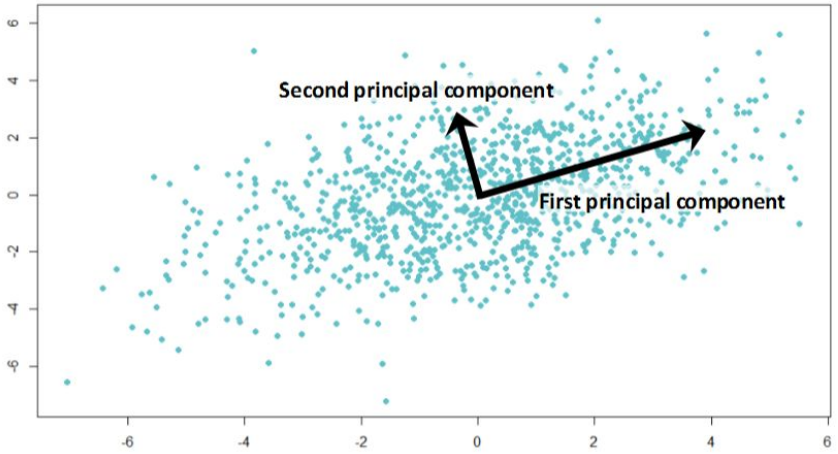
Feature Engineering

- Feature Generation
  - Domain Knowledge Features

    - CREDIT_ANN_INCOME_PERCENT: the percentage of the credit amount relative to a client's annual income
    - CREDIT_GOODS_PERCENT the percentage of the credit amount to a client’s asset
    - CREDIT_INCOME_PERCENT: the percentage of the credit amount relative to a clien’s total income
    - ANNUITY_INCOME_PERCENT: the percentage of the loan annuity relative to a client's income
    - DAYS_EMPLOYED_PERCENT: the percentage of the days employed relative to the client's age
    - PHONE_BIRTH_RATE: the rate of last phone change time to a client’s age
    - CAR_IMPLOY_RATE: the rate of own car time to a clients’ age

Feature Engineering

- Feature Generation
  - PCA based Features
    - Feature dimension reduction by unsupervised learning



Model and Ensemble

- Model
  - Trees
    - RandomForest
    - ExtraTress
  - Light GBM

- Ensemble
  - Averaging the prediction scores

Analysis

Model and Ensemble

- Boosting method is better than tree-based method
- Ensemble achieves performance gain over individual models

Model	Private Score	Public Score
RandomForest	70.88	71.10
Extra-Trees	71.95	71.61
Light GBM	77.78	77.41
Ensemble	77.76	78.00

Analysis

Data Merging

Data Constitution	Private Score	Public Score	Feature Dim
Train	74.61	74.45	241
Train+Bureau	75.86	75.75	452
Train+Bureau+Previous	77.96	78.02	1411

Table 2: Ablation study on data merging.

Analysis

Filter threshold	Num Features	CV Train	Private Score	Public Score
0.000	243	0.763	0.753	0.754
0.003	182	0.763	0.753	0.754
0.006	149	0.762	0.753	0.753
0.009	128	0.762	0.753	0.753
0.012	104	0.762	0.753	0.754
0.015	92	0.753	0.734	0.731

Table 3: Correlation-based feature selection results

Eliminated Features	Num Features	CV Train	Private Score	Public Score
0	243	0.763	0.753	0.754
30	213	0.763	0.753	0.754
60	183	0.763	0.753	0.754
90	153	0.762	0.753	0.754
120	123	0.763	0.753	0.754
150	93	0.763	0.753	0.754

Table 4: Backward elimination filtering results

Analysis

Data Generation

Data Constitution	Private Score	Public Score	Feature Dim
Original	77.61	77.59	285
Original + Polynominal	77.65	77.65	315
Original + Polynominal + Domain	78.25	78.43	333

Table 6: Ablation study on data data generation.

Analysis

Feature selection

- PCA dimension reduction

Exp #	Private Score	Public Score	Feature Dim
original	77.78	77.41	1539
PCA#1	73.78	75.16	128
PCA#2	74.68	75.36	256
PCA#3	75.22	76.19	512
PCA#4	75.64	76.66	1024

Thank you!