MATH6010z: Project 1

ZHAO JUNDA
MSc of Financial Mathematics
The Hong Kong University of Science and Technology
jzhaobr@connect.ust.hk

LI MINGLUO
MSc of Financial Mathematics
The Hong Kong University of Science and Technology
mlicv@connect.ust.hk

HE HAOKAI
MSc of Financial Mathematics
The Hong Kong University of Science and Technology
hheap@connect.ust.hk

HUANG WENJIN
MSc of Financial Mathematics
The Hong Kong University of Science and Technology
whuangbk@connect.ust.hk

Abstract

Home Credit try to make use of a variety of alternative data-including tel-co and transactional information-to predict their clients' repayment abilities. In this case, we use statistical and machine learning methods to make the predictions. We try to find out the best predictors according to the correlation coefficient calculation and use logistic regression to train the algorithm which could help recognize the clients capable of repayment. And with testing on Kaggle, our result can get 0.69290 private score and 0.70703 public score.

1 Introduction

Many people struggle to get loans due to insufficient or non-existent credit histories. So Home Credit strives to broaden financial inclusion for the unbanked population by providing a positive and safe borrowing experience. However, it's a huge challenge for Home Credit to recognize the clients' repayment abilities.
In this case, we try to use statistical and machine learning methods to help Home Credit make the predictions base on a variety of alternative data-including tel-co and transactional information in this project. And with trying, we focus on improve the accuracy of recognizing clients' repayment capable.

2 Credit Default Dataset (Kaggle Competition)

There are 16 categorical variables in the training dataset. 3 of which have no more than 2 unique categories and the rest have more than 2 unique categories. We use numeric value 0 and 1 to label the

1st Project MAFS 6010Z AI in Fintech (Worked by Team LI/HE/HUANG/ZHAO 2021), Hong Kong, China.

Table 1: Basic Logistic Regression Result

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
const	0.9716	0.032	30.155	0.000	0.908	1.035
EXT_SOURCE_1	-1.5501	0.052	-29.948	0.000	-1.652	-1.449
EXT_SOURCE_2	-2.4156	0.033	-73.869	0.000	-2.480	-2.352
EXT_SOURCE_3	-2.6412	0.037	-71.065	0.000	-2.714	-2.568
DAYS_BIRTH	1.692e-05	1.72e-06	9.844	0.000	1.36e-05	2.03e-05

former and one hot encoding to label the latter. After merging the training dataset with the testing dataset, there are 239 independent variables left for us to filter.
We calculate the correlation between 239 independent variables and the dependent variable namely TARGET so that we can select variables that have the highest correlation to the TARGET to build the logistic regression model. We select EXT_SOURCE_1, EXT_SOURCE_2 and EXT_SOURCE_3 who have the most negative correlation as well as DAYS_BIRTH who has the most positive correlation. After filling missing values with their medians and performing a basic logistic regression, we found their coefficients are all significant at a 0.05 level of significance and hence these variables are suitable for our model.

3 Regression on Credit Default Dataset

3.1 Methodology

For there are over hundreds of variables in the Credit Default dataset, some of them are dummy variables, and the most important thing is that the main target is a binomial variable. In this case, we choose Logistic regression to analyze the training set of the data in this project.

3.2 Analysis

The basic logistic regression model mentioned in Section 2 shows that zero TARGETs in the training set are predicted to be 0 when the default threshold is 0.5. Though only 8.7% TARGETs are indeed 1 in the training set, we think 0.5 is not an appropriate threshold. We further iterate threshold value from 0.05 to 0.95 with step size 0.05 and results are listed in Table 2.
We discard the results that all TARGETs are predicted to be 0 and select the threshold that owns the highest weighted average F1-score which is exactly 0.2 for the model mentioned above.
To further improve our model performance, we decided to introduce cubic polynomial feature to the model where we explore all polynomial combinations of 4 independent variables with degree less than or equal to 3. Similar to what we have done in Section 2, we calculate the correlation between polynomial combinations and TARGET and select combination variables whose correlation is larger than 0.15. The initial selection is given in Table 3. Then we perform basic LR for the feature set and discard features whose coefficient is insignificant at a 0.05 level of significance and finally retrieve 9 suitable features for the LR model which are presented in Table 6.
Table 4 shows that 0.2 is also the best threshold for the LR model with cubic polynomial feature when we iterate threshold value from 0.05 to 0.95 with step size 0.05. Compared with Table 2, we can observe that when threshold is 0.2, the introduction of cubic polynomial feature appreciably improve model performance in weighted average F1-score and accuracy.
Figure 1 shows that the area beneath the ROC curve of LR model with cubic polynomial feature is evidently larger than that of basic LR model.

4 Conclusion

4.1 Features Analysis

In the previous logistic regression, we totally used four kinds of data with nine different features. In this case, we want to analysis the relationship between the result and the data or the features. And with checking the meaning of the data we used from the description table, we get Table 5, and we found that most of them are from external source, and the last one is about client's age in days at the

Table 2: Basic LR Performance of Different Threshold

Threshold	Precision_0	Precision_1	Recall_0	Recall_1	F1-score Weighted Average	Accuracy
0.05	0.94	0.09	0.32	0.79	0.45	0.35
0.1	0.93	0.10	0.64	0.48	0.71	0.63
0.15	0.93	0.11	0.85	0.22	0.83	0.80
0.2	0.92	0.13	0.97	0.05	0.87	0.90
0.25	0.92	0.00	1.00	0.00	0.88	0.92
0.3	0.92	0.00	1.00	0.00	0.88	0.92
0.35	0.92	0.00	1.00	0.00	0.88	0.92
0.4	0.92	0.00	1.00	0.00	0.88	0.92
0.45	0.92	0.00	1.00	0.00	0.88	0.92
0.5	0.92	0.00	1.00	0.00	0.88	0.92
0.55	0.92	0.00	1.00	0.00	0.88	0.92
0.6	0.92	0.00	1.00	0.00	0.88	0.92
0.65	0.92	0.00	1.00	0.00	0.88	0.92
0.7	0.92	0.00	1.00	0.00	0.88	0.92
0.75	0.92	0.00	1.00	0.00	0.88	0.92
0.8	0.92	0.00	1.00	0.00	0.88	0.92
0.85	0.92	0.00	1.00	0.00	0.88	0.92
0.9	0.92	0.00	1.00	0.00	0.88	0.92
0.95	0.92	0.00	1.00	0.00	0.88	0.92

Table 3: Cubic Polynomial Feature Initial Feature Set

EXT_SOURCE_2
EXT_SOURCE_3
EXT_SOURCE_1*EXT_SOURCE_2
EXT_SOURCE_1*EXT_SOURCE_3
EXT_SOURCE_2*EXT_SOURCE_3
EXT_SOURCE_2*DAYS_BIRTH
EXT_SOURCE_1*EXT_SOURCE_2^2
EXT_SOURCE_1*EXT_SOURCE_3^2
EXT_SOURCE_1*EXT_SOURCE_2*EXT_SOURCE_3
EXT_SOURCE_1*EXT_SOURCE_2*DAYS_BIRTH
EXT_SOURCE_2*EXT_SOURCE_3^2
EXT_SOURCE_2^2*EXT_SOURCE_3
EXT_SOURCE_2*EXT_SOURCE_3*DAYS_BIRTH

time of applications.
Also, considering the nine features we chosen in Table 6, it's hard for us to explain the relationship with economics meanings since that we don't know what external source exactly is.

4.2 Kaggle Result Conclusion

Uploading our result of test data to the Kaggle, we can get a private score of 0.69290 and 0.70703 public score. To compare the expression of our algorithm, we also upload a result submission with totally random in 50-50 percent which got the result of 0.50041 private score and 0.50194 public score. In this case, it shows that our algorithm can truly help Home Credit to improve the judging accuracy of recognizing clients' repayment capable by about 20 percent.
However, compared to the top submission of the Leaderboard, there is still a large gap for us to improve. We may consider some other features in the logistic regression or use other algorithms like reinforcement learning to solve the problem. And in our opinion, the most regrettable thing is that because of the indistinct data description of some of the features we used, we cannot get a well explanation of our algorithm. There must be some underlying relation or logic between the data

Table 4: LR Performance With Cubic Polynomial Feature of Different Threshold

Threshold	Precision_0	Precision_1	Recall_0	Recall_1	F1-score Weighted Average	Accuracy
0.05	0.97	0.11	0.39	0.86	0.53	0.43
0.1	0.95	0.15	0.71	0.60	0.77	0.70
0.15	0.94	0.21	0.89	0.34	0.86	0.85
0.2	0.93	0.29	0.98	0.11	0.89	0.91
0.25	0.92	0.34	1.00	0.01	0.88	0.92
0.3	0.92	0.50	1.00	0.00	0.88	0.92
0.35	0.92	0.00	1.00	0.00	0.88	0.92
0.4	0.92	0.00	1.00	0.00	0.88	0.92
0.45	0.92	0.00	1.00	0.00	0.88	0.92
0.5	0.92	0.00	1.00	0.00	0.88	0.92
0.55	0.92	0.00	1.00	0.00	0.88	0.92
0.6	0.92	0.00	1.00	0.00	0.88	0.92
0.65	0.92	0.00	1.00	0.00	0.88	0.92
0.7	0.92	0.00	1.00	0.00	0.88	0.92
0.75	0.92	0.00	1.00	0.00	0.88	0.92
0.8	0.92	0.00	1.00	0.00	0.88	0.92
0.85	0.92	0.00	1.00	0.00	0.88	0.92
0.9	0.92	0.00	1.00	0.00	0.88	0.92
0.95	0.92	0.00	1.00	0.00	0.88	0.92

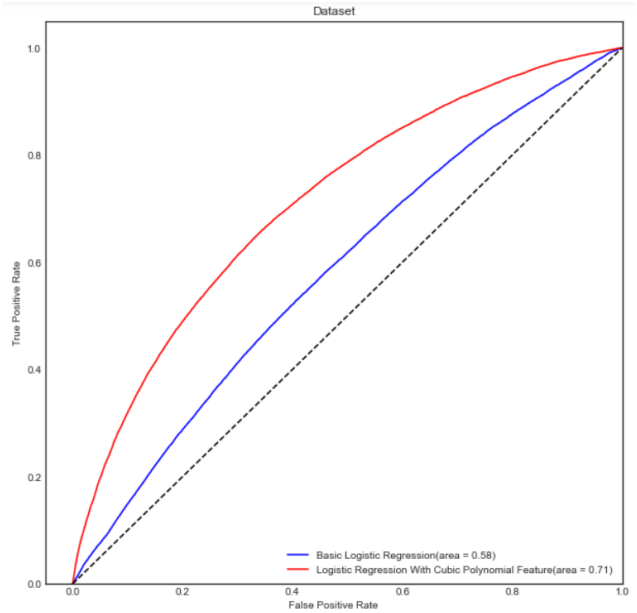

Figure 1: ROC Curve

Table 5: Feature Element Table

Data	Description
EXT_SOURCE_1	Normalized score from external data source
EXT_SOURCE_2	Normalized score from external data source
EXT_SOURCE_3	Normalized score from external data source
DAYS_BIRTH	Client's age in days at the time of application

Table 6: Features Table

EXT_SOURCE_2
EXT_SOURCE_3
EXT_SOURCE_2*DAYS_BIRTH
EXT_SOURCE_2*EXT_SOURCE_3
EXT_SOURCE_1*EXT_SOURCE_3
EXT_SOURCE_2*EXT_SOURCE_3^2
EXT_SOURCE_1*EXT_SOURCE_2*EXT_SOURCE_3
EXT_SOURCE_1*EXT_SOURCE_2*DAYS_BIRTH
EXT_SOURCE_2*EXT_SOURCE_3*DAYS_BIRTH

we use and the clients' repayment capable. We hope to find out the possible source of the external data so that we could better improve our regression method with totally understand of the algorithm logic.

5 Contribution

Python Coding: ZHAO JUNDA
Report Writing: LI MINGLUO, HE HAOKAI
Latex Coding: HUANG WENJIN