Logistic Regression Modeling: Home Credit Default Risk

Xinzhen Liao(20813829): Coding
Yiyuan Shi(20745230): Report Writing
Rui MA(20736954): Coding Support
Yuan Xu(20801498): Coding Support and LaTeX Formatting
All Team Members Fully Participated
Sep. 19, 2021

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Logistic Regression Modeling: Home Credit Default Risk

***Abstract*—Home Credit, as a non-bank financial institution, has cumulatively served over 135.4 million customers. Lenders could obtain the Home Credit loan from approximate 342.7 thousand partnering retail shops and shopping centers all around the world. It holds a competition on Kaggle, hope Kagglers to help them maximize their data efficiency through the use of various statistical and machine learning methods. Home Credit offers a 2.5GB data-set, with 8 separate csv files.**

In this report, we would like to share our experience of applying the Logistic Regression Model to complete this competition. Our final score is 0.73778, as showed below. And this whole article is divided into six parts, which are Introduction, Problem Definition, Data, Feature Selection, Logistics Regression, and Model Further Improvement respectively.

I. Introduction

Due to low income, credit problems, occupational restrictions, big data abnormality, unqualified background checking, and insufficient or non-existing credit records, more and more people nowadays have difficulties in getting loans. In addition, it is extremely important for lenders to accurately predict whether or not a person is able to repay their loan on time. Followed with data analysis and machine learning, lenders recently are trying to construct a precise home credit default prediction model. In order to ensure that consumers who have sufficient ability to repay could obtain loans on time, while refuse to offer loans to those who are highly likely to default. [1]

II. Problem Definition

A series of statistics model were used to predict home credit default risk. The purpose is to figure out the best parameter portfolio of the model and test its key futures like sensitivity.

III. Data

This may be a modified version of your proposal depending on previously carried out research or any feedback received.

A. Data Description

We used data-set named application_train.csv and application_test.csv through this project, which is the main table offered by Home Default. Part of them is showed below. Although Home Credit provides lots of data-set, based on our team's observation, we considered that the application_train.csv and application_test.csv are enough. Sine both of them are complete, consisting of 122 features for each applicant. From our points of view, we agreed that it is more than enough for machine learning to find data's pattern. We also tried to use the supplementary data; however, we faced some difficulties in following data processing. Therefore, in this program, we only used application_train.csv and application_test.csv. The application_train.csv is the training data set with target used to learn process, while the application_test.csv without target is used to test models.

	SK_ID_CURR	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY
0	100001	Cash loans	F	N	Y	0	135000.0	568800.0	20560.5
1	100005	Cash loans	M	N	Y	0	99000.0	222768.0	17370.0
2	100013	Cash loans	M	Y	Y	0	202500.0	663264.0	69777.0
3	100028	Cash loans	F	N	Y	2	315000.0	1575000.0	49018.5
4	100038	Cash loans	M	Y	N	1	180000.0	625500.0	32067.0

Fig. 1. Table Offered by Home Default

B. Data Processing

1) Missing Value: i. Justify Boolean Variable From our observation, there exists lots of Boolean Variable in our data-set, for example, the Code_Gender, Flag_Own_Car, and Flag_Own_Reality. We initially transform this string into 0 and 1, where 1 represents for "Yes" while 0 represents for "No". As for missing values in this kind of variable, we filled out all missing values with 0 instead.

Take feature Flag_Own_Car as an example, 1 represents the lender owns a car while 0 indicates the lender does not have a car. If this lender's information is not provided in the data-set, we assume he or she does not own a car in default.

ii. Justify Numbers Besides Boolean Variables, there are also lots of missing numbers, like Amt_Annuity and Ant_Income_Total. As for missing values in this form, we filled out all missing values with average.

2) Dummy Variable: For variables like Weekday_Appr_Process_Start, it has more than 2 outcomes, like Monday, Wednesday, and Sunday etc. We transformed this type of variable into Dummy Variables, where it only takes the value 0 or 1 to indicate the absence or presence for the features. In logistic regression model, it could be considered as numeric representation for qualitative facts. [2]

3) Normalization: In machine learning, different features usually will have different dimension and unit, which will affect the data analysis. In order to reduce the dimensional effect among indicators, it is important for us to standardize our data to further compare and analyze. In this program, we used the method of normalization. [3]

Normalization is a widely used technique in machine learning. It creates new values which maintain the general distribution and ratios in original data, while keeps values within a fixed scale through all columns. As a result, it effectively avoids the problems of distortion differences and information lose. The first five raw of data-set after normalization is showed in figure 2.

SK_ID_CURR	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	REGION_POPULATION_RELATIVE	DAYS_BIRTH
100001	-0.559988	-0.427809	0.142475	-0.553684	-0.037477	-0.164654	-0.733477
100005	-0.559988	-0.782413	-0.804537	-0.752938	-0.839362	1.009586	-0.461392
100013	-0.559988	0.237075	0.401002	2.519991	0.497113	-0.147258	-0.917718
100028	2.260729	1.345214	2.896221	1.223579	3.303709	0.358078	0.483623
100038	0.850370	0.015447	0.297651	0.164921	0.483748	-0.775825	0.699997

Fig. 2. Data-set After Normalization

IV. Feature Selection

It is well-known that a huge number of features in data analysis is quite challenging, as a result, feature selection is a vital step in machine learning since it offers an effective way to solve this difficulty. Feature Selection aims to remove irrelevant and useless data, which highly deduct the computation time and improve the machine learning accuracy. As introduced in the Data part, there are 122 features in total, which is too much. Thus, we used Principal Component Analysis (PCA) and Forward Step wise Selection to select features. [6]

A. Principal Component Analysis (PCA)

Principal Component Analysis is a widely applied method in machine learning. It is defined as a dimensional reduction method used to reduce the dimension of large data set. Also, it transforms suits of large variables into a smaller one while preserving as much as information as possible. [4]

In this program, we initially want to use PCA to help us reduce features, we set components equal to 100,50,20 respectively and expected that PCA will automatically choose the best features that contribute the most. However, the results are all not satisfactory, the detailed scores and performance for each round are showed in the figure 3, 4, 5. [5]

Therefore, we abandon the Principal Component Analysis method since it brought negative impact on our model.

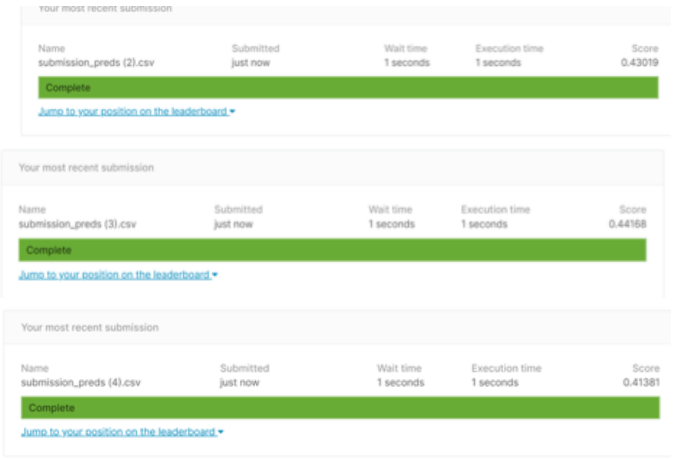


Fig. 3. Result for 100 components

Fig. 4. Result for 50 components

Fig. 5. Result for 20 components

B. Forward Step wise Selection

Forward Step wise selection is a method to select important variables. It begins with a null model with no variable and one feature will be added once a time. Based on the output indicators like adjusted R square and RSS results, we are able to choose the best model.

Since Principal Component Analysis (PCA) method does not work previously, we determined to try Forward Step wise Selection this time. Unfortunately, we ran the code for the whole night but the Kaggle website did not output anything. Followed with two more times of trying and failure, we abandon this method as well and will try to figure out the problems further.

V. Logistics Regression

A. K-Fold

In machine learning, it is common to divide the data-set into a training set and testing set respectively. The testing set is completely independent of the training set, it is only used for the evaluation of the final prediction model. Over-fitting is also a common problem in data processing, it happens when the prediction model could perfectly fit the training data, but it could not fit any data else. If we use the testing data to adjust our parameters at this time, it means that we will know part of testing data's information in advance, which will highly affect the accuracy in our final evaluation result. The normal solution to solve this problem is to separate a part of the training data as the validation data, in order to evaluate the training effect of the prediction model. The validation data is taken from the training set, but not being trained. In this way, the prediction model's degree of matching could be objectively measured. Cross validation makes effective use of limited data while the evaluation result could be as close as possible to the model performance on testing data, which could also be used as an index for model optimization.

In this project, in order to avoid the problem of over-fitting caused by excessive number of variables. We use cross-validation method to divide the original data-set into 5 groups, and make a validation set for each subset. The remaining 4 sets of subset data will be used as the training set. Finally, we will have 5 models as a result. These 5 models will be evaluated by validation set respectively. The AUC value of the fitted model for all the training data-sets is 0.744. (figure 6.)

As for the subset model, their fitting effect could be observed by the AUC value, which concentrated around [0.745,0.750]. As for the performance effect besides the model prediction, it is presented by the AUC value of the corresponding validation data set of the models in five groups. It could be seen that the AUC value for this part varies from [0.735,0.740], which is similar to training data-set's performance. The ROC curve of each subset model is showed in Appendix A:.9

Therefore, we could conclude that there does not exist the problem of low AUC value caused by over-fitting problem, since partial of the performance effect of the validation data-set and the prediction effect score of the training data set are almost at the same level.

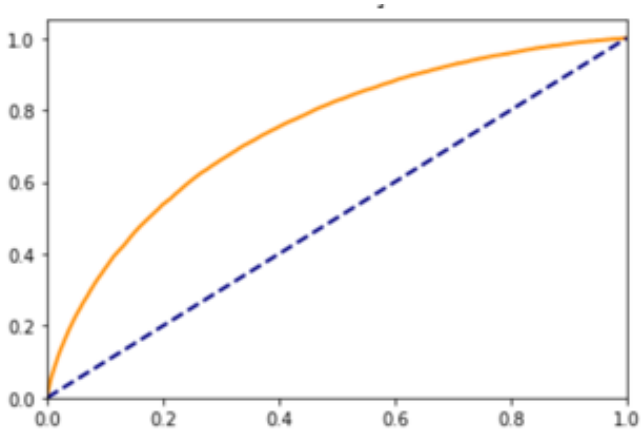


Fig. 6. ROC curve for the full training data-set

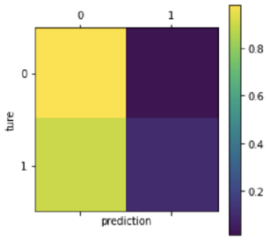
Strain	Train	Validation
$Fold_1$	0.747026	0.743900
$Fold_2$	0.746098	0.751525
$Fold_3$	0.747099	0.743660
$Fold_4$	0.748234	0.738886
$Fold_5$	0.747178	0.742988
Full	0.744185	

TABLE I. ROC CURVE

Based on the above results, it indicates that although the number of features is relatively large, there does not exist the over-fitting problem in our prediction model, which also proves that our prediction model has certain degree of stability.

B. The Confusion Matrix

A confusion matrix compares the LDA predictions to the true default statuses for the 48,744 training observations in the Application_train.csv data set. Elements on the diagonal of the matrix represent individuals who was judged to be able to get loans were correctly predicted, while off-diagonal elements represent individuals that were mis-classified. LDA made incorrect predictions for 1.65%. Individuals who were not able to get loans and for 89.72% individuals who were able to get loans. The confusion matrix is shown in figure 7.



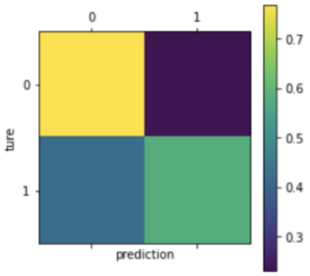
		Predicted Class	
		0	1
True Status	0	98.35%	1.65%
	1	89.72%	10.28%

Fig. 7. Confusion Matrix

The class-specific performance:

- Error rate with people who were able to get loans: 89.72%
- Sensitivity (TPR): 1 - 89.72% = 10.28%
- Error rate within people who were not able to get loans (FPR): 1.65%
- Specificity: 1 - 0.65% = 98.35%

When we change the standard threshold of whether to lend or not into 0.14, the corresponding prediction effect has been significantly improved(fig. 8):



		Predicted Class	
		0	1
True Status	0	76.89%	23.10%
	1	42.02%	57.97%

Fig. 8. Confusion Matrix(modified)

The class-specific performance:

- Error rate with people who were able to get loans: 42.02%
- Sensitivity (TPR): 1 - 42.02% = 57.97%
- Error rate within people who were not able to get loans (FPR): 23.10%
- Specificity: 1 - 23.10% = 76.89%

The process of changing the threshold involves the trade-off of sensitivity and specificity. In the process of lowering the threshold selection from 0.5 to 0.1, although specificity has been reduced to a certain extent, it is more important for banks to accurately determine whether or not to lend.

C. Meaning of Logistic Regression Coefficient

The results of our logistic regression model are showed in Appendix B.10

We take feature "days_employed" as an example, its coefficient is 3.24×10^{-6} with p-value approximates to 0, which is much lower than 0.05. Which means, we should reject the null hypothesis that H_0=0 and says that the feature "days_employed" contribute a lot to our prediction.

On the other hand, we take feature "Days_Birth" as an example, its coefficient approximates to 0,with p-value equals to 0.604, which is much higher than 0.05. Which mean, we should not reject the null hypothesis that H_0=0 and says that the feature "Days_Birth" does not contribute a lot to our prediction.

VI. MODEL FURTHER IMPROVEMENT

Since the result of dimensionality reduced by Principal Component Analysis (PCA) is not satisfactory. We want to perform other types of regression analysis for the dataset after the process of PCA, in order to improve the prediction

accuracy. We choose the XG-Boost method because of its high second-order derivative accuracy and fast-running speed.

The general steps of XG-Boost parameter are:

- Learning Rate: fluctuate between 0.05 to 0.3, usually is initially set at 0.1.
- Turning specific parameters for decision tree like max_depth, min_child_weight, gamma, subsample, colsample_bytree. In the process of determining a tree, we could choose different parameters.
- Adjustment of parameters' regulation (e.g., lambda, alpha). These parameters could reduce the complexity of the model, and thereby improve the performance of the whole prediction model.
- Reduce the learning rate and determine the ideal parameters.

We tried to use XG-Boost to fit the feature vector after PCA's dimensionality reduction, however, it the process of adjusting parameters, we found that there seems to be no global optimization solution with this data set. As a result, our program failed to find the optimal solution of corresponding parameters and prediction models during the operation.

We also tried to use Forward Stepwise Selection and Backward Stepwise Selection to select feature, however, the scores for these two methods are much lower than without using them. Thus, we still decided to keep the original method.

Our prediction model did not use supplementary datasets this time. We believe that we will have a better prediction effect after using more supplementary datasets and adding more supplementary variables with the help of machine learning to further fit.

References

[1] "Normalize Data," Jun. 5, 2019. [Online]. Available: https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/normalize-data. [Accessed: Jun. 5, 2019].

[2] "LightGBM on Home Credit Default Risk Prediction," Nov. 4, 2020. [Online]. Available: https://becominghuman.ai/lightgbm-on-home-credit-default-risk-prediction-5b17e68a6e9. [Accessed: Nov. 4, 2020].

[3] "wikipedia: Dummy variable (statistics)," Available: [Online]. https://en.wikipedia.org/wiki/Dummy_variable_(statistics).

[4] "PCA: Application in Machine Learning," Feb. 28, 2019. [Online]. Available: https://medium.com/apprentice-journal/pca-application-in-machine-learning-4827c07a61db. [Accessed: Feb. 28, 2019].

[5] "A Step-by-Step Explanation of Principal Component Analysis (PCA)," Apr. 1, 2021. [Online]. Available: https://builtin.com/data-science/step-step-explanation-principal-component-analysis. [Accessed: Apr. 1, 2021].

[6] "Feature Selection Methods in Machine Learning.," Aug. 1, 2018. [Online]. Available: https://medium.com/@sagar.rawale3/feature-selection-methods-in-machine-learning-eaeef12019cc. [Accessed: Aug. 1, 2018].

Appendix A
ROC curves of each subset model

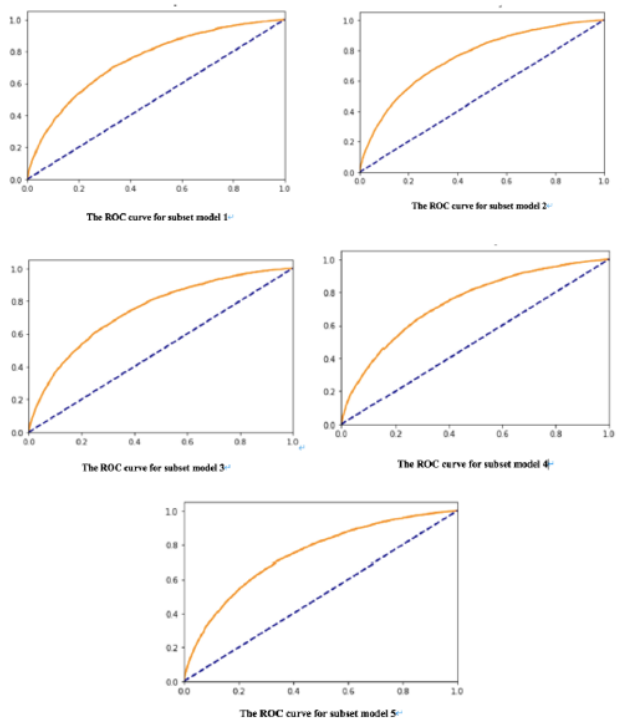


Fig. 9. ROC curves of each subset model

Appendix B
Logistic Regression Model

	coef	std err	t	P>|t|	[0.025	0.975]
CNT_CHILDREN	-0.0299	0.046	-0.645	0.519	-0.121	0.061
AMT_INCOME_TOTAL	4.992e-09	2.07e-09	2.409	0.016	9.31e-10	9.05e-09
AMT_CREDIT	1.548e-07	7.68e-09	20.161	0.000	1.4e-07	1.7e-07
AMT_ANNUITY	6.727e-07	5.5e-08	12.231	0.000	5.65e-07	7.8e-07
AMT_GOODS_PRICE	-1.829e-07	8.47e-09	-21.597	0.000	-2e-07	-1.66e-07
REGION_POPULATION_RELATIVE	0.1691	0.043	3.921	0.000	0.085	0.254
DAYS_BIRTH	9.369e-08	1.81e-07	0.518	0.604	-2.61e-07	4.48e-07
DAYS_EMPLOYED	3.242e-06	2.55e-07	12.693	0.000	2.74e-06	3.74e-06
DAYS_REGISTRATION	6.123e-07	1.48e-07	4.150	0.000	3.23e-07	9.01e-07
DAYS_ID_PUBLISH	3.069e-06	3.38e-07	9.072	0.000	2.41e-06	3.73e-06
OWN_CAR_AGE	0.0003	6.93e-05	4.050	0.000	0.000	0.000
FLAG_MOBIL	0.0456	0.264	0.173	0.863	-0.471	0.563
FLAG_EMP_PHONE	-0.0383	0.076	-0.501	0.616	-0.188	0.112
FLAG_WORK_PHONE	0.0133	0.001	9.933	0.000	0.011	0.016
FLAG_CONT_MOBILE	-0.0158	0.011	-1.387	0.166	-0.038	0.007
FLAG_PHONE	-0.0034	0.001	-2.948	0.003	-0.006	-0.001
FLAG_EMAIL	-0.0052	0.002	-2.478	0.013	-0.009	-0.001
CNT_FAM_MEMBERS	0.0311	0.046	0.671	0.502	-0.060	0.122
REGION_RATING_CLIENT	-0.0092	0.003	-2.941	0.003	-0.015	-0.003
REGION_RATING_CLIENT_W_CITY	0.0189	0.003	6.037	0.000	0.013	0.025
HOUR_APPR_PROCESS_START	-9.293e-05	0.000	-0.594	0.553	-0.000	0.000
REG_REGION_NOT_LIVE_REGION	-0.0171	0.006	-2.814	0.005	-0.029	-0.005

Fig. 10. Logistic Regression Model