math 6010-Zhou-Sun-Huang-Tian Project1 Report

Zhou Xiaomin 20749212
Sun Ke 20747903
Tian Xinyu 20750015
Huang Yuning 20738524

September 19, 2021

Abstract

We treat data from 'application-train.csv' as our raw dataset. After data cleaning and processing, we split data into training set and test set. Next, we use LDA, Lasso and Ridge models to fit training data and validate the models by test set, predicting the default status of each client. The performance of each classifier is evaluated on area under the ROC curve between the predicted probability and the observed target.

1 Data Processing

The raw data has 122 variables and 307511 observations.

1.1 Imputation of missing data

The first thing we need to do is examining our dataset for missing values. As Fig.1 shows, there are 65 variables having missing values, and the missing rate of more than half of them reach 1/2 or above. Therefore, we remove all the variables with missing values from the data.

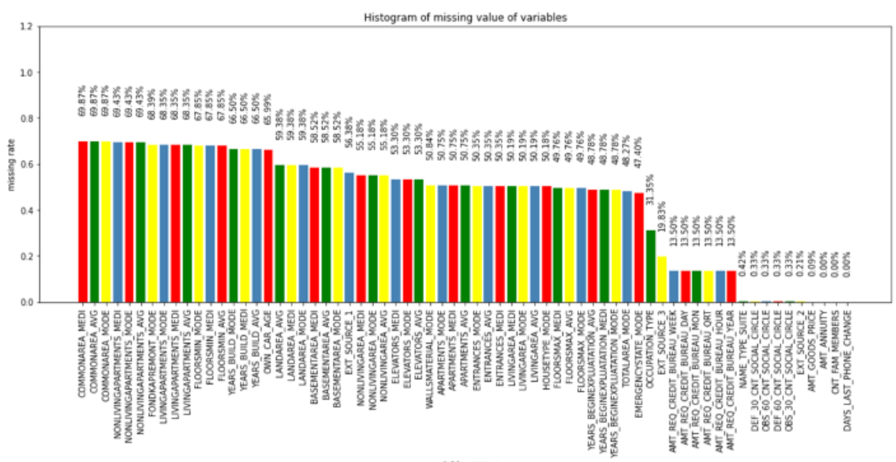

Figure 1: histogram of missing value of variables

1.2 Remove irrelative variables

After checking the description of each data, we find that variables called "FLAG-DOCUMENT-XX" are not correlated of our target since the definitions are not clear. So, we remove these 20 variables from the data.

1.3 Encode categorical variables

Since we require all input variables to be numeric, we need to encode all categorical variables. Looking at the data, we can see that ten of all the variables are categorical.
As Fig.2 shows, there are 58 categories of 'ORGANIZATION-TYPE', which may lead to too much new vector when we use the encoder to transform it. Therefore, we remove it from the data.

NAME_CONTRACT_TYPE	2
CODE_GENDER	3
FLAG_OWN_CAR	2
FLAG_OWN_REALTY	2
NAME_INCOME_TYPE	8
NAME_EDUCATION_TYPE	5
NAME_FAMILY_STATUS	6
NAME_HOUSING_TYPE	6
WEEKDAY_APPR_PROCESS_START	7
ORGANIZATION_TYPE	58

Figure 2: Categorical Variables

Then We factorize variables with 2 categories and use one hot encoding to convert the remaining categorical data to numeric form.

1.4 Deal with outliers

The definition of variable 'DAYS-EMPLOYED' is 'How many days before the application the person started current employment', which means the values should be negative. By visualizing the data, we find that there are some outliers all equals 365243. We transform them to NA.

count	307511.000000
mean	63815.045904
std	141275.766519
min	-17912.000000
25%	-2760.000000
50%	-1213.000000
75%	-289.000000
max	365243.000000

Figure 3: DAYS-EMPLOYED

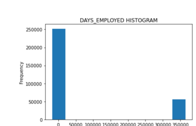

Figure 4: DAYS-EMPLOYED

1.5 Data transformation

We convert "DAYS-BIRTH","DAYS-EMPLOYED","DAYS-REGISTRATION","DAYS-ID-PUBLISH" into positive values in years. Besides, we also convert all NAs to 0.

Now we only have 62 variables in our data.

```
 18  REGION_RATING_CLIENT                                307511 non-null  int64
 19  REGION_RATING_CLIENT_W_CITY                         307511 non-null  int64
 20  HOUR_APPR_PROCESS_START                             307511 non-null  int64
 21  REG_REGION_NOT_LIVE_REGION                          307511 non-null  int64
 22  REG_REGION_NOT_WORK_REGION                          307511 non-null  int64
 23  LIVE_REGION_NOT_WORK_REGION                         307511 non-null  int64
 24  REG_CITY_NOT_LIVE_CITY                              307511 non-null  int64
 25  REG_CITY_NOT_WORK_CITY                              307511 non-null  int64
 26  LIVE_CITY_NOT_WORK_CITY                             307511 non-null  int64
 27  CODE_GENDER_F                                       307511 non-null  uint8
 28  CODE_GENDER_M                                       307511 non-null  uint8
 29  CODE_GENDER_XNA                                     307511 non-null  uint8
 30  NAME_INCOME_TYPE_Businessman                        307511 non-null  uint8
 31  NAME_INCOME_TYPE_Commercial associate               307511 non-null  uint8
 32  NAME_INCOME_TYPE_Maternity leave                    307511 non-null  uint8
 33  NAME_INCOME_TYPE_Pensioner                          307511 non-null  uint8
```

Figure 5: Information of part of variables

2 Training set and test set

We use function "train-test-split" to randomly choose 30% of the observations as the test set and 70% as the training set.

3 LDA

Firstly, we use LDA to fit the training dataset and do prediction. The result is as follows:

True default status	No	Yes
Predicted default status		
No	84909	7337
Yes	5	3

	precision	recall	f1-score	support
No	0.920	1.000	0.959	84914
Yes	0.375	0.000	0.001	7340
accuracy			0.920	92254
macro avg	0.648	0.500	0.480	92254
weighted avg	0.877	0.920	0.882	92254

Figure 6: LDA prediction– decision boundary: prob=50%

Moreover, instead of using the probability of 50% as decision boundary, we say that a probability of default of 25% is to be classified as 'Yes'. And the result is shown below. We can find that the accuracy of the model prediction has declined.

True default status	No	Yes
Predicted default status		
No	84421	7190
Yes	493	150

	precision	recall	f1-score	support
No	0.922	0.994	0.956	84914
Yes	0.233	0.020	0.038	7340
accuracy			0.917	92254
macro avg	0.577	0.507	0.497	92254
weighted avg	0.867	0.917	0.883	92254

Figure 7: LDA prediction– decision boundary: prob=25%

4 Logistic Regression with L1-Regularization—Lasso

Lasso method is a technique that constrains or regularizes the coefficient estimates, or equivalently, that shrinks the coefficient estimates towards zero.
As Fig.8 shows, some of the coefficients are now reduced to exactly zero.

	Lasso_Coefficient
NAME_CONTRACT_TYPE	-0.163517
FLAG_OWN_CAR	-0.145645
FLAG_OWN_REALTY	-0.025738
CNT_CHILDREN	0.012345
AMT_INCOME_TOTAL	0.023244
AMT_CREDIT	-0.053046
REGION_POPULATION_RELATIVE	-0.005890
DAYS_BIRTH	-0.158669
DAYS_EMPLOYED	-0.238207
DAYS_REGISTRATION	-0.058747
DAYS_ID_PUBLISH	-0.117894
FLAG_MOBIL	0.004086
FLAG_EMP_PHONE	0.000000
FLAG_WORK_PHONE	0.052513
FLAG_CONT_MOBILE	-0.019312
FLAG_PHONE	-0.048632

Figure 8: Lasso-coefficients of some variables

True default status	No	Yes
Predicted default status		
No	84914	7340

Figure 9: Lasso Prediction—decision boundary: prob=50%

As Fig.9 shows, when using probability of 50% as decision boundary, no one will default. So we implement prediction using probability of 25% as decision boundary instead.

Pred	No	Yes
True		
0	84554	360
1	7220	120

	precision	recall	f1-score	support
No	0.921	0.996	0.957	84914
Yes	0.250	0.016	0.031	7340
accuracy			0.918	92254
macro avg	0.586	0.506	0.494	92254
weighted avg	0.868	0.918	0.883	92254

Figure 10: Lasso Prediction-decision boundary: prob=25%

5 Logistic Regression with L2-Regularization—Ridge

Comparing to Lasso, Ridge method does have one obvious disadvantage, it will include all predictors in the final model. The penalty will shrink all of the coefficients towards zero, but it will not set any of them exactly to zero.

	Ridge_Coefficient
NAME_CONTRACT_TYPE	-0.164660
FLAG_OWN_CAR	-0.146597
FLAG_OWN_REALTY	-0.026652
CNT_CHILDREN	0.012958
AMT_INCOME_TOTAL	0.028046
AMT_CREDIT	-0.054089
REGION_POPULATION_RELATIVE	-0.006369
DAYS_BIRTH	-0.158262
DAYS_EMPLOYED	-0.239486
DAYS_REGISTRATION	-0.059235
DAYS_ID_PUBLISH	-0.118405
FLAG_MOBIL	0.013598
FLAG_EMP_PHONE	-0.052111
FLAG_WORK_PHONE	0.053290
FLAG_CONT_MOBILE	-0.020062
FLAG_PHONE	-0.049271

Figure 11: Ridge-coefficients of some variables

Similarly, we use probability of 25% as decision boundary here to predict. The result of logistic regression with L2-regularization is very similarly to Lasso method.

Pred	No	Yes
True		
0	84543	371
1	7219	121

	precision	recall	f1-score	support
No	0.921	0.996	0.957	84914
Yes	0.246	0.016	0.031	7340
accuracy			0.918	92254
macro avg	0.584	0.506	0.494	92254
weighted avg	0.868	0.918	0.883	92254

Figure 12: Ridge Prediction-decision boundary: prob=25%

6 ROC curve

Fig.13 displays the ROC curve for the LDA classifier, Lasso method and Ridge method on the test data. The overall performance of a classifier, summarized over all possible thresholds, is given by the area under the (ROC) curve (AUC). An ideal ROC curve will hug the top left corner, so the larger the AUC the better the classifier.

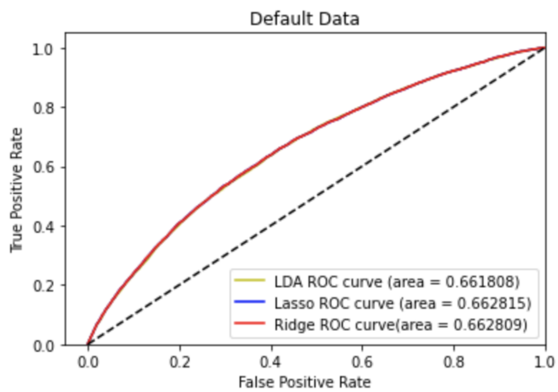

Figure 13: ROC curve of 3 classifiers

We can find that logistic regression with L1-regularization performs the best. In fact, all these three classifiers performs poorly, we need to come up with more advanced models to fit the data.

7 Contribution

Huang Yuning: Latex, report
Sun Ke: code
Zhou Xiaomin: code, report
Tian Xinyu: code, report