Hybird Modeling and Prediction Methods for Mortgage Loan

Yuxuan(Jonathan) Chen, Juetao Chen, Weiye Wang

This paper dives deep into a dataset containing single family loan portfolio for FNMA originating in the second quarter of 2000.
(https://www.ams.jhu.edu/ dan/ComputingForAppliedMathematics/) It includes an exploration of data with pre-processing, variable selection using multiple methods, and response prediction applying multiple models.

Applied Mathematics and Statistics
Johns Hopkins University
United States
May 5, 2021

1. Introduction

Loan is one of the most common forms in finance. The lender advances a sum of money to the borrower and receives the future repayment. To avoid a deficit, the lender needs to make some predictions based on characteristic data. There are many factors that could affect the foreclosure of a loan, and we are interested in which of these factors contribute the most. We also hope to come up with a model that fits the dataset better, thus reaching higher accuracy for predicting future loans.

For foreclosure prediction, Brown (2012) [1] compared the predictive accuracy of three supervised machine learning techniques(classification trees, SVM and genetic programming) when applied to mortgage data. Each machine learning technique achieved a classification accuracy of greater than 0.75. For other types of loans, a large number of machine learning methods have been used to predict loan default. Xiaojun Ma et al. (2018) [2] applied the modern machine learning algorithms LightGBM and XGbbost to predict Peer-to-peer(P2P) network loan default. The LightGBM algorithm achieved an accuracy of 0.801, which is better than the XGboost algorithm. For prediction of *NMONTHS* (the number of months until the mortgage is taken off the books due to foreclosure, prepayment, etc.), we did not find such papers.

We did predictions on both response variables *NMONTHS* and *FORCLOSED* in this paper. We applied linear regression to predict *NMONTHS*. For *FORCLOSED* (foreclosure), we used logistic regression and other machine learning algorithms to do so. Before modeling, we also did a brief data overview and some pre-processing.

2. Data Exploration

The data we used is a portion of the single family loan portfolio for FNMA originating in the second quarter of 2000. There are two response variables: *NMONTHS*, the number of months until the mortgage is taken off the books due to foreclosure, prepayment, etc.; *FORCLOSED*, a boolean variable that indicates whether the mortgage foreclosed (True) or not (False). The dataset contains 23 predictor variables in total, including 9 categorical variables and 14 continuous variables.

2.1 Data Pre-processing

Before data analysis, we should do some data pre-processing. There are several kinds of predictor variables in the dataset and they may relate to the others, which is fatal for linear models. Therefore, we need to explore the correlations between the variables and remove multicollinearity for the dataset.

2.1.1 Pairwise correlations

First, we plot the heat-map to explore the correlations between the 14 numeric variables. From Figure 1, we see that there are strong positive correlations between *ORIGTERM*,

REMMNTHS and *ADJRMTHS*, with all of the correlation coefficients close to 1.

To eliminate these correlations, we omit the predictor variables *REMMNTHS* and *ADJRMTHS*. Then we are left with 12 continuous predictor variables. We plot the heat-map of the correlation coefficient matrix of these 12 variables(Figure 2). By omitting those 2 variables, there doesn't exist any strong pairwise correlation between them.

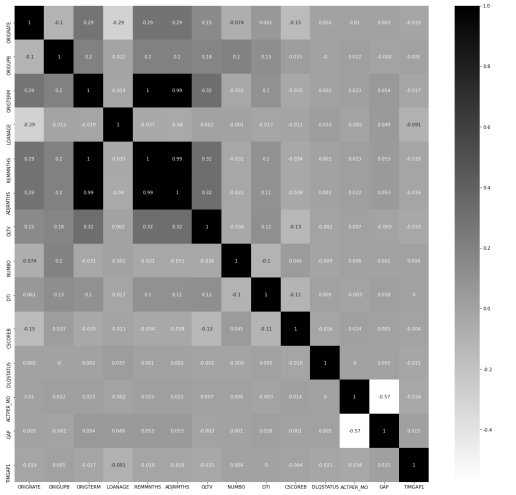


Figure 1: The heat-map of the correlation coefficient matrix(14 predictor variables)

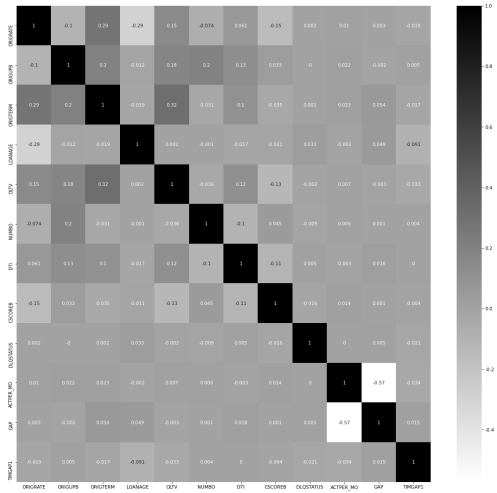


Figure 2: The heat-map of the correlation coefficient matrix(12 predictor variables)

2.1.2 Multicollinearity

Although we have eliminated strong pairwise correlations, the multicollinearity among the variables may still exist. For example, we have $X_1 = [1, 0, 2]^T$, $X_3 = [0, 2, 1]^T$, $X_3 = [1, 2, 3]^T$, where $X_3 = X_1 + X_2$. Even though $Corr(X_1, X_3) = Corr(X_2, X_3) = 0.5$, examination of pairwise correlations does not disclose multicollinearity.

A formal, widely used method of detecting the presence of multicollinearity is to use the variance inflation factors. Consider a design matrix $X = [x_1, x_2, \ldots, x_n]^T$, where x_i is a column vector. We can compute the variance inflation factor (VIF) as:

$$VIF_k = (1 - R_k^2)^{-1} \qquad k = 1, 2, \ldots . n$$

where R_k^2 is the coefficient of R-squared when x_k is regressed on the $n - 1$ other x_{-k} variables. The variance inflation factor is equal to 1 when $R_k^2 = 0$, i.e., when x_k is not linearly related to the other x variables. A VIF value in excess of 10 is frequently considered as an indication that multicollinearity may be unduly influencing the least squares estimates.

Then we compute the variance inflation factors of the 12 continuous prediction variables. From Table 1, we can find that all of the variance inflation factors are less than 10, most of which just slightly greater than 1. This means there doesn't exist serious multicollinearity problem among those 12 continuous prediction variables.

Table 1: The variance inflation factors

Name	VIF_k	R_k^2	Name	VIF_k	R_k^2
ORIGRATE	1.2821	0.2200	DTI	1.0564	0.053
ORIGUPB	1.1697	0.1451	CSCOREB	1.0533	0.0506
ORIGTERM	1.2599	0.2063	DLQSTATUS	1.0019	0.0019
LOANAGE	1.1265	0.1123	ACTPER_MO	1.5010	0.3338
OLTV	1.1647	0.1414	GAP	1.5077	0.3367
NUMBO	1.0666	0.0625	TIMGAP1	1.0132	0.0130

2.2 Description of the dataset

After eliminating strong pairwise correlations and multicollinearity among the variables, we are left with 21 predictor variables, including 9 categorical predictor variables and 12 continuous variables(Table 2). Then we perform an exploratory analysis according to the distribution histograms(Figure 3) and the boxplots(Figure 4) of *NMONTHS* of the categorical predictor variables.

From the histograms and the boxplots, we find that the response variable *NMONTHS* has different distributions under different categories of categorical variables, which means we should use them in our model to predict *NMONTHS*.

Table 2: Summary of variables

Data type	Variables
Categorical	CHNL, SELLER, FIRSTFLAG, PURPOSE, PROP, NO_UNITS, OCCSTAT, STATE, RELMORTGIND
Continuous	ORIGRATE, ORIGUPB, ORIGTERM, LOANAGE, OLTV, NUMBO, DTI, CSCOREB, DLQSTATUS, ACTPER_MO, GAP, TIMGAP1

3. Data Analysis

3.1 Variable Selection based on *NMONTHS*

NMONTHS is one of the response variables in the dataset. It gives the number of months a certain mortgage takes to be paid off. After proper data pre-precessing, we are left with 21 predictor variables, including both continuous and categorical ones. 21 is a relatively large number, but a good one for variable selection since we are able to remove some insignificant variables that do not contain much information about the response variable.

Since 21 predictor variables is a fairly large number, performing variable selection iteratively will lead to 2^{21} different possible combinations of variables. Therefore, we decided to switch to a step-wise approach. Because both continuous and categorical variables are

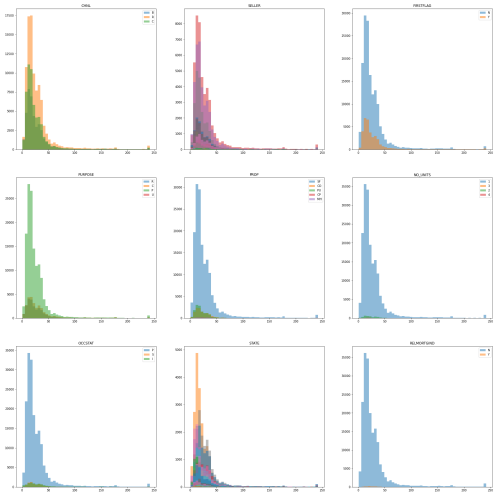


Figure 3: The histograms

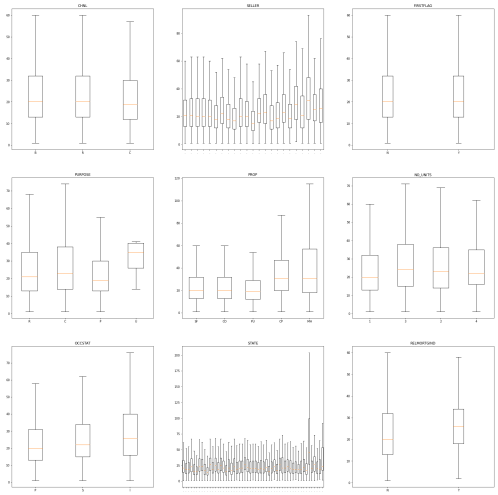


Figure 4: The boxplots

included, we could not perform step-wise variable selection based on the p-value of each variable. Each categorical variable of k different levels contains $(k-1)$ estimates and p-values. For these reasons, we tackled this problem by a step-wise variable selection based on AIC, BIC and adjusted R-squared, in both directions.

Coincidentally, the models we reached from the full model and the null model are the same, which implies an obvious difference between the significance of statistically important and unimportant variables.

3.1.1 Step-wise Approach

For AIC, forward and backward approach stops at the model with 18 variables. *RELMORTGIND*, *TIMGAP1* and *GAP* are the three variables that are omitted. Since AIC is a criteria that prefer complex models with more parameters, this result tells us these three variables are very statistically insignificant to the response variable *NMONTHS*.

Then we go to BIC, hoping to remove more variables considering the fact that BIC is penalizing complex models more harshly comparing to AIC. We were able to eliminate two more variables additionally. *DLQSTATUS* and *CHNL* are the omitted two this time. Although these two variables carry some amount of information, the information does not compensate for the extra complexity these two variables bring to the model.

Lastly, we used adjusted R-squared as our variable selection criteria. Ordinary R-squared does not work here because ordinary R-squared is strictly increasing as number of parameters increases, thus not useful to the variable elimination process. The model we end up with using adjusted R-squared has 19 variables, where *RELMORTGIND* and *TIMGAP1* are eliminated.

Although the models we obtained using the three different measurements are different, the sequence that variables are deleted follows a fixed order. We could safely conclude that the variables that are removed early from these processes are highly likely to be statistically insignificant.

3.1.2 Exhaustive Search

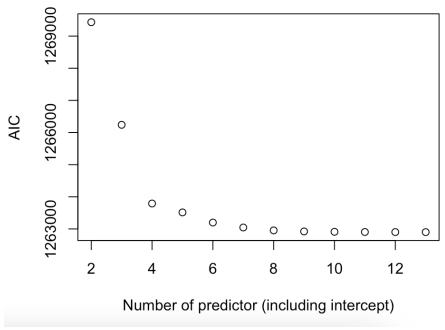
We are also interested in the significance of the continuous variables, which are 12 out of the 21 predictor variables we have. When we have 12 variables, we could apply the iterative approach to check the best sub-model of these 12 variables. We performed this method by calling regsubsets function in R, and compared the results(Figure 5) based on AIC, BIC, adjusted R-squared and Mallow's C_p.

	ORIGRATE	ORIGUPB	ORIGTERM	LOANAGE	OLTV	NUMBO	DTI	CSCOREB	DLQSTATUS	ACTPER_MO	GAP	TIMGAP1
1 (1)	" "	"*"	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "
2 (1)	"*"	"*"	" "	" "	" "	" "	" "	" "	" "	" "	" "	" "
3 (1)	"*"	"*"	" "	" "	" "	" "	" "	"*"	" "	" "	" "	" "
4 (1)	"*"	"*"	" "	"*"	" "	" "	" "	"*"	" "	" "	" "	" "
5 (1)	"*"	"*"	"*"	"*"	" "	" "	" "	"*"	" "	" "	" "	" "
6 (1)	"*"	"*"	"*"	"*"	" "	"*"	" "	"*"	" "	" "	" "	" "
7 (1)	"*"	"*"	"*"	"*"	" "	"*"	" "	"*"	" "	"*"	" "	" "
8 (1)	"*"	"*"	"*"	"*"	" "	"*"	"*"	"*"	" "	"*"	" "	" "
9 (1)	"*"	"*"	"*"	"*"	"*"	"*"	"*"	"*"	" "	"*"	" "	" "
10 (1)	"*"	"*"	"*"	"*"	"*"	"*"	"*"	"*"	" "	"*"	"*"	" "
11 (1)	"*"	"*"	"*"	"*"	"*"	"*"	"*"	"*"	" "	"*"	"*"	"*"
12 (1)	"*"	"*"	"*"	"*"	"*"	"*"	"*"	"*"	"*"	"*"	"*"	"*"

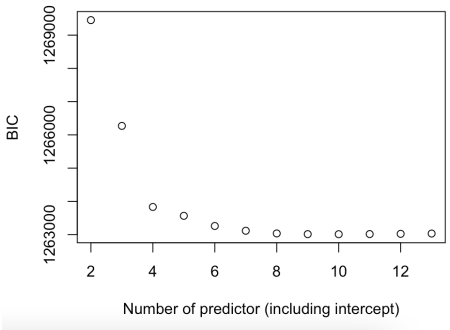
Figure 5: Continuous Variable Selection

From the four measurements(Figure 6), we reached the same conclusion that the model does not improve much once we have the first 6 significant variables among the 12, which are *ORIGRATE*, *ORIGUPB*, *ORIGTERM*, *LOANAGE*, *NUMBO*, *CSCOREB* respectively. Undoubtedly, these variables exist in the selected model in the previous step. The three continuous variables that are removed in the previous step, *DLQSTATUS*, *GAP* and *TIMGAP1* are also the first three that are removed here. The model selection procedure is relatively stable regardless of the methods and the criteria we are using.

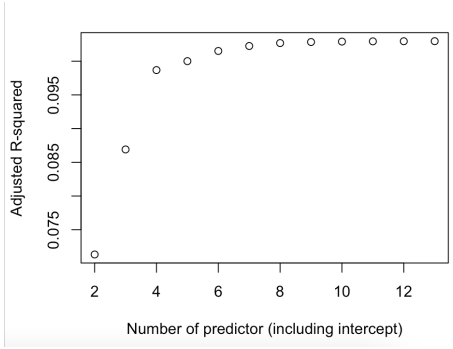
As a conclusion, we can affirm that the 6 continuous variables that are selected above, and the 7 categorical variables that are selected in the model from the step-wise approach with the least number of predictors, *SELLER*, *FIRSTFLAG*, *PURPOSE*, *PROP*, *NO_UNITS*, *OCCSTAT* and *STATE*, are significant to the response variable *NMONTHS*. Expressly, the number of month that a mortgage takes to be paid depends on a variety of factors, and



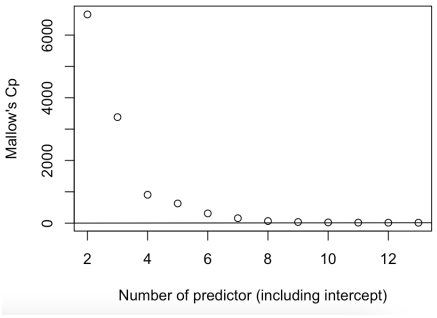
(a) AIC



(b) BIC



(c) Adjusted R-squared



(d) Mallow's C_p

Figure 6: Different measurements on the best model given number of predictors

based on the data we have, we summarized that these 13 factors could be more influential than others.

3.2 Prediction for *FORCLOSED*

In this subsection, we focus on predicting the categorical variable *FORCLOSED*'. The above heatmap of the correlation matrix shows that linear variables *ORIGTERM*, *REMMNTHS* and *ADJRMTHS* have high correlation, so we drop these three continuous variables.

First we simply use PCA to reduce the dimension. The result of function *pca.explained_v ariance_ratio* shows that the cumulative explained variance ratio is close to 1 (0.99999996) when the number of components is 3. So we choose $n\_components = 3$ and use the transformed 3-dimension data to predict the value of *FORCLOSED*.
The result of function df['FORCLOSED'].value_counts() shows that there are 186606 false responses and 2136 true responses, which implies that the classification is imbalanced. Therefore, we use the threshold-moving method to find a best threshold for the training data. For the 999 thresholds from 0.001 to 0.999, we choose the threshold that makes FPR closest to 0.5. Then it shows that the threshold = 0.992, while the false positive rate =

0.485 and the true positive rate = 0.870. We use the best threshold to train and test the training data by cross validation. The following is the ROC curve of the logistic regression.

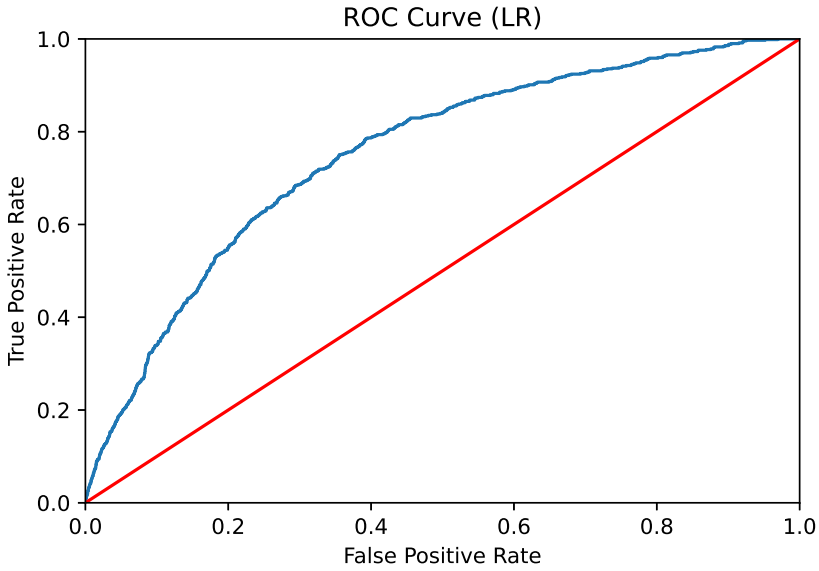


Figure 7: The ROC curve of logistic regression

We have roc_auc_score(Ytest,Ypred_prob[:,1])=0.7617. Then we use the decision tree method and the following is the corresponding ROC curve.

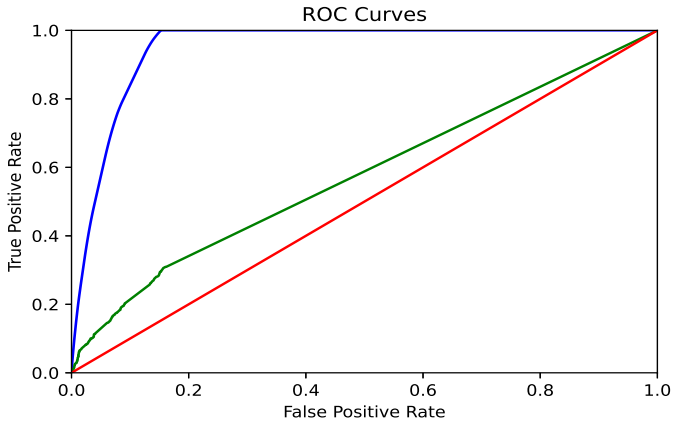


Figure 8: The ROC curve of decision tree

In figure 8, the blue line is the ROC curve of the training data and the green line is the ROC curve of the test data. The figure illustrates that the decision tree model has low AUC score, which is approximately 0.5722. Then we use the 10-fold cross-validation to get the cross_val_score. The result is [0.9883, 0.98803, 0.9892, 0.9880, 0.9879, 0.9874, 0.9906,

0.9887, 0.9892, 0.9894]. Considering that only about 1.4% data has true 'FORCLOSED' value, the result is not satisfying. Also, the AUC score is still around 0.6.
Thus, we prefer the logistic model based on the processed data.
Then we consider using the lasso regression to reduce the dimension. The following figure illustrates the distribution the coefficients.

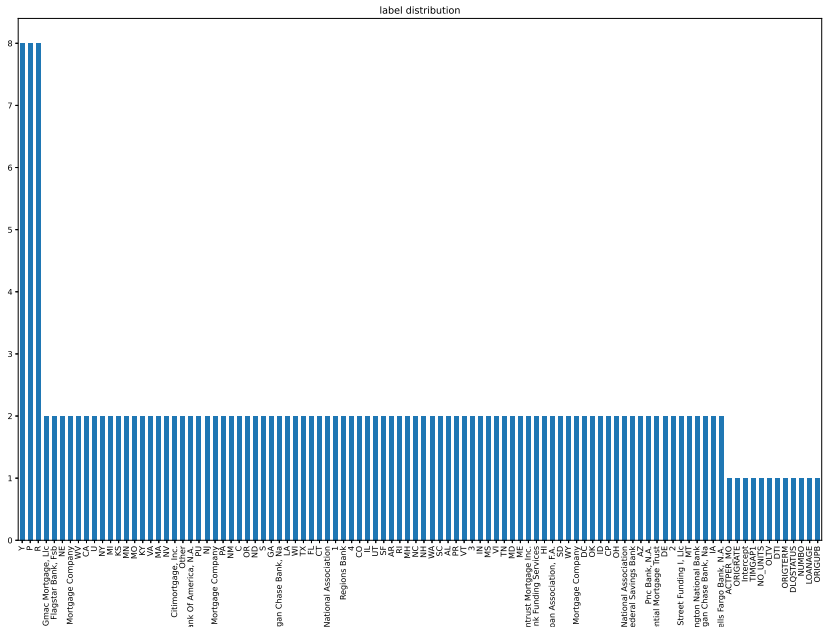


Figure 9: Coefficients distribution of lasso regression

The figure 9 shows that only the continuous variables 'ORIGRATE', 'OLTV' and categorical variable 'STATE' have coefficients larger than 1e-10. The result is consistent with the pca method.

4. Discussion

The dataset we used contains two response variables of different types, so we utilized different methods to tackle them.

We first performed multiple variable selection methods using *NMONTHS* as the response variable. We picked out 13 of the 21 valid predictor variables that we could reasonably argue that they are more statistically significant than others. Then we tried different methods to train a model that could most accurately predict the boolean response variable *FORCLOSED*. We first used the pca method to reduce the dimension of the data to 3. Then we utilized the logistic regression and decision tree method to train the model. After that, we used lasso regression to reduce the dimension and get similar results, which verifies the pca method.

Overall, our analysis on the dataset is reasonable and comprehensive. For each individual question, multiple methods are employed to cross validate the results. The methods we tried are relatively simple and basic. Nevertheless, we obtained decent outcomes in both analysis and prediction. As a future plan, we could use some more complex models(gradient descent, neural network, etc) to better model the data based on the predictor variable and thus reach higher prediction accuracy or lower error. Additionally, the dataset is only from FNMA in the second quarter of 2000. This is an earlier time and rapid changes might take place even from year to year. To make use of the results and predict the current trend, we need more dataset that comes from different times and companies.

While the two response variables are of different types, we directly used them to assess the significance of variables and accuracy of model predictions. In fact, there could be inner relationships between these two response variables that they are not independent of each other. Further studies could be done to find the mutual effect between them. Therefore, we are able to combine the separate findings on the two response variables to a single conclusion on the exact model.

References

[1] Dexter Randell Brown. "A Comparative Analysis of Machine Learning Techniques For Foreclosure Prediction". In: (2012).

[2] Xiaojun Ma et al. "Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning". In: *Electronic Commerce Research and Applications* 31 (2018), pp. 24–39.