LOAN PERFORMANCE PREDICTION: A COMPARISON BETWEEN BAYESIAN APPROACH AND OLS

Junzhou Lin
Johns Hopkins University
Baltimore, MD 21217
jlin157@jh.edu

December 19, 2021

Abstract

In this paper, I present a Bayesian g-prior linear regression model to predict loan's performances, which is the number of months until the mortgage is taken off from the books due to foreclosure, prepayment, etc. The Bayesian inference for the model is developed using the Markov chain Monte Carlo Gibbs samplings method, then compare its performance metrics with the performance of the Ordinary least square approach. This will help us to analyze the difference between the two approaches, the benefits and shortcomings of each. Implementation of the model is calculated using actual data available to Fannie Mae's books in the second quarter of 2000.

1 Introduction

Mortgage companies and banks are always in need of a well-developed system to ensure that the loan they lend is properly analyzed and there is no loss, which means companies do not wish loans they lend go into default/foreclosure and they need to make sure the borrower has the ability to repay the loan.

Therefore, this brings statistics and probability into the play. A Bayesian model can analyze data and provide a better prediction than a company employee on whether it's a good idea to lend the loan to a certain borrower or not.

1.1 Purpose

The purpose of this project is to develop the loan lending process by making it easier and more accurate. Being able to identify correct borrowers and making the lending process quicker will potentially decrease the loss from companies or banks. This project also helps us to see the difference between the traditional linear regression OLS approach versus Bayesian approach.

1.2 Present system

The present system for a lot of mortgage companies and banks to decide whether lending a loan to a borrower is entirely depends on the intuition and perception of the employees [1], especially if the company does not have a sophisticated data scientist team. The fact that loan performance is decided solely depending on the experience of an employee made the lending process unstable and potentially hurt the company's profit and increase its loss.

2 Proposed method

The project will use two approaches to predict the performance of a loan which is the number of months until the mortgage is taken off from the books due to foreclosure, prepayment, etc. The data will be cleaned and feature engineered then fit into both models. Based on the prediction, the data will be trained and MSE will be calculated for each approach. Apart from the MSE, MAE and R^2 will also be looked at as well.

2.1 Ordinary least square

Linear regression is one of the Generalized linear models and is probably the most common machine learning model. Linear regression is used to describe the relationship between a response variable with one or more predictors. Linear regression is a robust model which usually provides a high bias and low variance result, and more often less accurate compared to other complex machine learning model such as Random forest. However, Linear regression has good explainability and is great to use as a tool to understand the data or as a baseline model.

The fundamental equation of linear regression in matrix form is: $Y = \beta X + \varepsilon$. After moving Y and ε to another side,taking square of both sides to make sure we have a monotonic increasing function and summing up all the errors $\sum_{i=1}^{n} \varepsilon^2 = \sum_{i=1}^{n} (Y - \beta X)^2 = (Y - X\beta)^T (Y - X\beta)$. Then do partial derivative with respect to each β_i. We'll find the unbiased estimate of β is:

$\hat{\beta}_{ols} = (X^T X)^{-1} X^T Y$; $\hat{\sigma} = 1/(n-p)(Y - X\hat{\beta}_{ols})^T (Y - X\hat{\beta}_{ols})$

This process of finding the unbiased estimator that minimizes the sum of errors is called Ordinary least squares estimates(OLS). Therefore, the predicted value of Y is given by: $\hat{Y}_{ols} = (X\hat{\beta}_{ols}) = MY$ where $M = (X^T X)^{-1} X^T$

2.2 Bayesian g-prior

The Bayesian analysis of linear regression is similar with the classical OLS. But instead of calculating Y with a closed-form solution, it uses Gibbs sampling method to generate $\beta_{(t)}$ and $\sigma_{(t)}$ from the normal distribution and the inverse gamma distribution.

Since we are fitting the Bayesian regression model using g-prior in this project. Following are my steps:

- Set Gibbs sampling iteration S = 2000
- Set the pior: g = n where n is number of samples; $s_0^2 = 1$; $v_0 = 2$
- $H_g = g/(g+1)X(X^TX)^{-1}X^T$; $SSR_g = Y^T(I - H_g)Y$ where I is a n*n identity matrix
- Generate S number of $s_{(t)}^2$ from inverse gamma distribution with $shape = (v_0 + n)/2$ and $scale = 1/rate = 2/(v_0 * s_0^2 + SSR_g)$
- $V_b = g/(g+1)(X^TX)^{-1}$ $\exists LL^T = V_b$ and L can be found through cholesky decomposition
- $E_b = V_bX^TY$; Generate a matrix E with n*p variables from normal distibution with $mean = 0$ and $variance = s_{(t)}^2$ where we have (p-1) features
- $\beta_{(t)} = ((EL)^T + E_b)^T$; The average of $\beta_{(t)}$ for each feature is the $\hat{\beta}_{bayes}$ found through Bayesian approach
- In the end,$\hat{Y}_{bayes} = X\hat{\beta}_{bayes}$

3 Implementation

3.1 Data

For the purpose of this project, I have acquired data from Fannie Mae mortgage company. The data set can be found on this link:

https://github.com/jlin157/JHU-632-Baysian-Final-Project-Ongoing/blob/main/Dataset.rar

The dataset has 38 columns and is filled with hard-understanding jargon. A Fannie Mae loan glossary can help with understanding the dataset:

https://github.com/jlin157/JHU-632-Baysian-Final-Project-Ongoing/blob/main/FNMA_SF_Loan_Performance_Glossary_2020.pdf

3.2 Parameters

I have dropped some of the features which I believe to be useless towards the prediction of the performance. Followings are 20 parameters that actually goes into the model:

1).$ORIG\_RATE =$ The original interest rate on a mortgage loan as identified in the original mortgage note

2).$ORIG\_UPB$ = The dollar amount of the loan as stated on the note at the time the loan was originated

3).$ORIG\_TERM$ = The number of months in which regularly scheduled borrower payments are due at the time the loan was originated

4).$LOAN\_AGE$ = The number of calendar months since the mortgage loan's origination date.

5).ADJ_REM_MONTHS = Adjusted remaning month of the loan

6).$OLTV$ = Original Loan to Value Ratio

7).NUM_BO = Number of Borrowers

8).DTI = Debt-To-Income ratio

9).$CSCORE\_B$ = Borrower Credit Score at Origination

10).$CSCORE\_C$ = Co-Borrower Credit Score at Origination

11).$FIRST\_FLAG$ = First Time Home Buyer Indicator

12).$PURPOSE$ = Loan Purpose

13).$PROP$ = Property Type

14).NO_UNITS = Number of Units

15).OCC_STAT = Occupancy Status

16).MI_PCT = Mortgage Insurance Percentage

17).IO = Interest Only Loan Indicator

18).DLQ_STATUS = Current Loan Delinquency Status (Null as status 4)

19).MI_TYPE = Mortgage Insurance Type (Null as no MI)

20).$RELOCATION\_MORTGAGE\_INDICATOR$ = An indicator that denotes whether or not the type of mortgage loan is a relocation mortgage loan

$R^2 = 1 - SSE/SSTO$

$MSE = 1/n \sum_{i=1}^{n} (Y_i - \hat{Y_i})^2$

$MAE = 1/n \sum_{i=1}^{n} |Y_i - \hat{Y_i}|$

The dataset contains 200000 rows, but due to big data overcommitting memory issues. The final training set that I use has 21000 rows and I used the rest of the data set as the testing set.

3.3 Pre processing

- Dropping any features that don't make sense to the performance of a loan or are highly correlated with other features
- Missing values are imputed by either mean or mode of its column
- Columns with above 50% missing values are dropped
- Normalizing numerical features by MinMaxScaler to enlarge any important trend
- One hot encoding is applied to all Categorical features

Table 1: Performance matrics

Matric	Bayesian	OLS
R^2	70.23%	70.24%
MAE	11.29	11.29
MSE	268.16	268.11

3.4 Training and Prediction

To ensure the consistency of the prediction, I manually implemented a 9 fold cross-validation technique for both Bayesian and OLS approaches. In this technique, I divide the data set into 9 different subsets. Each subset contains 21000 rows of data and is used as the training set, the testing set is the complement of the training set.

3.5 Performance Matrics

The goal of this project is to calculate number of months the mortgage is taken off from the books due to foreclosure, prepayment, etc.; And the performance of both model is calculated through the average of 9 folds MSE, MAE and R^2 in the testing set as shown in Table 1.

4 Conlusion

As mentioned before, it's greatly beneficial for a mortgage company or bank to be able to identify whether it's worth lending a loan to a borrower through a system rather than depend on the intuition of its employees. The above approaches are the first step that will help develop such a system and requires a lot more development and trial and error. By comparing the Bayesian and OLS approaches, I see a minimum difference between the two of them in terms of accuracy. However, the Bayesian approach is much more computational expensive due to its matrix transformation, therefore, data size is limited for casual Bayesian testing. Utilizing more data would require working in a big data environment such as Apache Spark.

References

[1] R., A. U., Kumar, C. & Kaushik, S. (2018). LOAN FORECLOSURE PREDICTION: COMPARING LOGISTIC REGRESSION AND LINEAR SUPPORT VECTOR MACHINE.*International Journal of Pure and Applied Mathematics* 1314-3395.