Mortgage-Backed Securities Prepayment Prediction: A Machine Learning Approach

Yu Dong Zhang
lz927@stanford.edu

Abstract. A mortgage loan gives the borrower the right to prepay the full or partial amount before its maturity. It is referred as mortgage prepayment. Mortgage-backed securities (MBS) are financial products structured based on cash flows of these mortgage loans. MBS is now among the largest financial sector in the United States. With this gigantic amount of MBSs on banks' balance sheet, managing prepayment is vital. In this project will be exposed aspects of machine learning regression algorithms, and how they are applied in the setting of predicting Mortgage-Backed securities prepayment, specifically, the conditional prepayment rate.

Keywords: MBS, Prepayment, Machine Learning, Neural Network

1 Motivation

Mortgage-backed securities (MBS), the "vicious" origin of 2008 financial crisis, which have severed as a crucial lever for government in managing monetary policies for nearly 40 years, are among the largest financial sector in United States. U.S. mortgage finance market had experienced structural changes after the crisis. Especially during recent couple of years, MBSs are mostly guaranteed by agencies, which lead to huge decrease in credit risk of such product, making them mostly default-free. The investing public had always been trying to solve the world that has sufficient uncertainties to generate excessive economical returns, instead of risk-free entities. This time is no exception. The elimination of credit uncertainties of MBSs has encouraged Wall Street quants to seek profits in the other key factor that drives the value of MBSs, and that is prepayment. Prepayment modelling is among the most complex and novel areas of financial modelling. Its demand has surged significantly only in recent 4 years, and the complexities involve enormous amount of data and factors, as well as the difficulties in model specification and estimation.

2 Data

The data set that is used for this project is purchased from Black Knight, a company who bought eMBS which is leading MBS data provider. The data set contains monthly conditional prepayment rate of a pool of 30-year fixed rate MBSs along with the pool's attributes such as refinancing rate, prepayment speeds, etc, which will or will not be used as independent variables. There are approximately 30 raw data attributes purely from this data set. And additionally, economical factor data is also included, such as market mortgage rate and unemployment rates.

2.1 Data Preprocessing

Since the data set is very large, significant work of data preprocessing needs to be done. First, for missing values, since these variables are time-varying, and current month of magnitude does dependent on that of previous month, linear interpolation is used to fill out missing values. Second, for outliers, since there are very few that can be spotted (after calculating their respective ranges), they are removed and they only occupy approximately 0.1% of the data.

2.2 Feature Selections

For feature selection, at this stage, only domain knowledge is used to select features. For example, refinancing rate is traditionally very relevant to prepayment rate where as issuers are not so relevant since average homeowners usually have no knowledge of what the issuers are. The most 15 relevant features are selected to proceed as to gain an insight about this data set. A further more thorough feature selection methodology is discussed as in the Next Step section.

3 Methods

3.1 Lasso Linear Regression

I first choose to use the most simplistic regression model to apply to the data set to gain some further information about the relationship between x and y. Lasso linear regression is given below:

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{n}(y^{(i)} - \theta^T x^{(i)})^2 + \gamma||\theta||_1$$

Since we have 15 features at hand, Lasso linear regression is used. By adding this lasso penalty, or L1 norm of the parameters as penalty, it aims to sparse the parameters and prevent over-fitting. In the case of multi-collinearity, as in the features may have correlations with each other, for example, refinance rate and housing turnover, Lasso could help with this issue with some θ being zeros. And also, adding the regularization term is also a good practice in terms of finding the optimal bias and variance trade-off.

3.2 Neural Network

The main method of this project is neural network. Lasso linear regression is only applied as a baseline regression model to gain insight about linearity as will be discussed in the next section. Since the neural network has not yet been completed. The specification of it will be discussed in Next Step section.

4 Preliminary Experiments

K-fold cross validation is applied to determine lasso linear regression hyperparameter γ. $k = 10$ is selected as it is a typical value to use for machine learning practitioners. The data set is randomly divided into 10 subsets, 9 of which serves as training set and the other one serves as validation set. For each of the training set, there is one trained model with a set of associated parameters, θ. For evaluation purpose, the average of the training errors over entire 9 models is computed to serve as the estimated generalization error. The model's γ that produces the smallest generalization error is chosen to perform modelling on the entire training set with hypothesis shown in last section. The data set was further split into 80% for cross-validation aforementioned and the rest 20% is for model testing or error analysis. Accuracy is used as the evaluation metric to see how lasso linear regression performs. The model had an accuracy of 40.8% which is considered poorly performed given the over 90% accuracy in the paper [2] which a neural network produces.

The reason why lasso linear regression performs such badly comparing to neural network in the paper, even when overfitting is taken care of, is that many of the risk factors that drive MBS's prepayment is highly nonlinear and also interactive. For example, as many studies suggest, prepayment due to housing turnover and prepayment due to refinancing rate have very different dependency on same risk factors. More specifically, refinancing rate increases with loan size whereas housing turnover decreases with loan size, while they are all the factors that feed into lasso linear regression model. In MBS industry, refinancing rate and housing turnover each have their own model and are estimated separate meaning they are not often related together as inputs and most of the prepayment data do not disclose the reason for prepayment.

5 Next Steps

As discussed in the previous section, lasso linear regression model is not a good choice for the data set. Thus, A more functionally rich neural network and a more thorough feature selection model are proposed in hope to achieve higher accuracy.

5.1 Feature Selection Methodology

Generally, a neural network does not require complicated feature selection methodology due to the fact that it can form proper nonlinear combination of the attributes based on the intrinsic of the data features when there are enough layers and nodes to work with. For our case, we have over 30 data features which is way too many, for the ease of computation and training time, feature selection will still be performed.

Domain Knowledge. Features will still be selected based on my knowledge on which impact prepayment the most, but the number of eliminations is limited. Only the features that with certainty are not related to the prepayment are eliminated.

Covariance Matrix. In order to tackle the problem with multi-collinearity, the factor covariance matrix is calculated explicitly to check if two factors are highly correlated. If so, only one will be included, and sign will be taken care of if there is negative correlation.

Information Theory. The information value matrix is also calculated to select features. And only factors with information value greater than 0.02 are included.

Decision Tree. A decision tree model is built to calculated mean square error and importance scores. The factors with high importance score are included.

With all four steps above, a final set of features can be selected and feed into the neural network.

5.2 Neural Network Model Specification

A neural network which is the core model to this project will take the features selected from above as input and produce conditional prepayment rate as a single numerical output. Below are the specific design regarding the neural network:

Structural Design. The neural network consists of 1 input layer, 1 output layer, and for hidden layers, the first hidden layer consists of 1024 nodes and the last has 2 nodes while in between the next hidden layer has half of the number of nodes than previous hidden layers. ReLU is used as activation function within hidden layers and sigmoid function is used on the output layer.

Functional Design. In each hidden layer, 2 sublayers are added. The first sublayer is the dropout layer since there are two many nodes in the first layers and training time could be reduced significantly with dropout sub-

layer, and also overfitting can be treated. The second sublayer is batch normalization. This sublayer normalizes the output from activation function in order to fasten parameter convergence.

Algorithm-related Design. ADAM optimizer is used to find the optimal set of weights. Grid search can be used to find the appropriate momentum rate, learning rate and learning rate decay.

With the aforementioned design, conditional prepayment rate can be predicted and also similar error analysis will also be performed and if time allows, model ensemble can also be adopted in the setting.

References

[1] Shlomo Amar. Modeling of mortgage loan prepayment risk with machine learning. 2020.

[2] J. "David" Zhang, X. "Jan" Zhao, J. Zhang, F. Teng, S. Lin, and H. "Henry" Li. Agency mbs prepayment model using neural networks. The Journal of Structured Finance, 24(4):17–33, 201