**Predicting Mortgage Prepayment Using Deep Learning Network**

**Introduction**

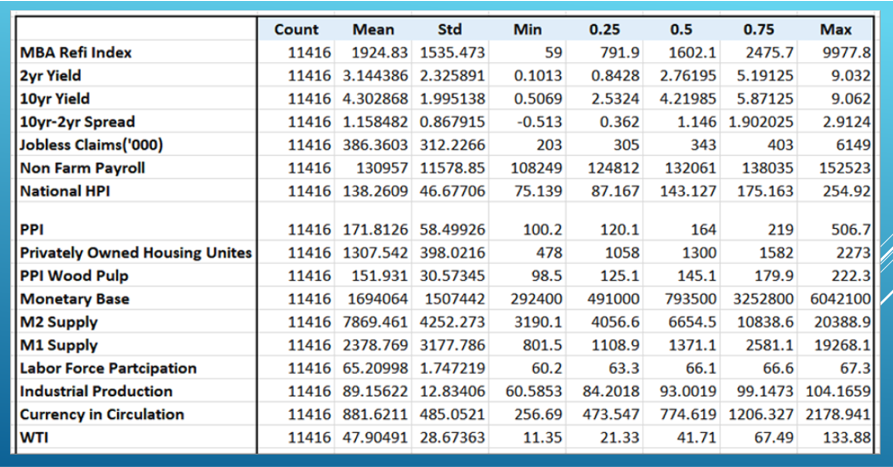
Mortgages are one of the largest financial sectors in the United States. The lending tree estimates the total value of mortgages is $10.5 trillion (about $32,000 per person in the US) (about $32,000 per person in the US) as of the third quarter of 2019. Mortgages also serve as a key lever for the federal government in managing monetary policies. When individuals buy a house, they must pay in cash or get a loan from a bank or other financial institution using the house as collateral. The mortgagor must pay specific interest rates depending on their credit to get the loan. They can get a better rate by changes in their credit, interest rate environment, or different economic or sociological factors. This is called refinancing (prepayment). Banks and investors want to know to predict when this happens to manage their funding accordingly. This prepayment modeling is vital for risk analysis.

Currently, investors (banks) pay millions of dollars to model the prepayment and are primarily based on models created by humans. Prepayments are highly nonlinear and could change because of the underlying risk factors and regime changes in credit availability and borrower behavior. Prepayment due to housing turnover and prepayment due to rate refinance have vastly different risk factor sensitivities (Yu 2018). Because of this, housing turnover and rate refinance are often estimated separately. We had planned to use a single-family loan dataset from Freddie Mac to build a neural network to model the prepayment, but we had around moved 90 million loans, and our laptop could not handle opening few data files. To mimic this, we chose to use the Mortage Banker Association Refi Index that reports the number of refinance applications submitted and is reported every Wednesday. Thus, the prepayment model we built will only be predicting the prepayment due to the rate refinance.

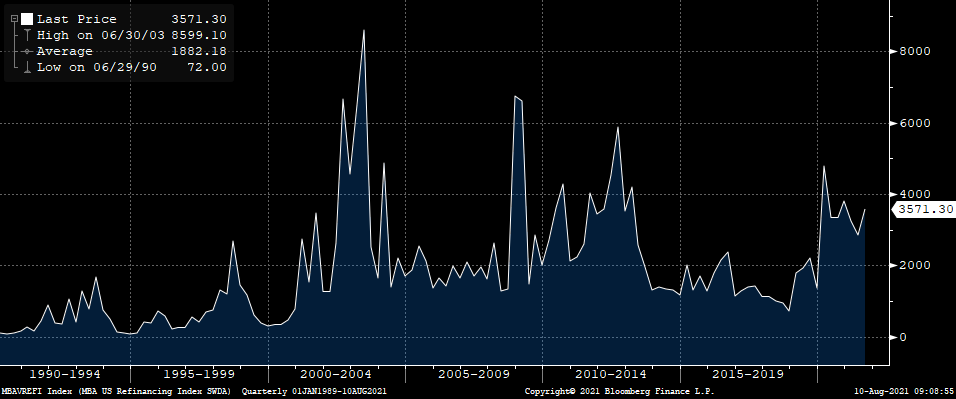
Most of the models are built using historical data and forecasting prepayment, delinquency, default, and loss probabilities. There are many models built using the interest rates dynamics modeled by a term structure model whose state variables are denoted by X1 (t), . . ., Xn (t). These models are complex with various kinds of rates curve, loan characteristics, borrower characteristics, and millions of data built to predict the refinancing of the loan. The model we built is not data-intensive, but we have done our best to create the best model.

**Data**

Most of the data was downloaded from the Federal Reserve Economic Data website (FRED) by St. Louis Fed. FRED is an online database consisting of hundreds of thousands of economic data time series from scores of national, international, public, and private sources. FRED, created and maintained by the Research Department at the Federal Reserve Bank of St. Louis. All the data except MBA (Mortgage Bankers Association) refinance index was downloaded from FRED. MBA refinance index was downloaded from Bloomberg. Some of the data were available daily, whereas some of the data were weekly or monthly. We extrapolated the monthly data to daily by using the same value daily over the month period.



**MBA refinance index:** The MBA Refinance Index is a weekly measurement put together by the Mortgage Bankers Association, a national real estate finance industry association. The index helps predict mortgage activity, and loan prepayments based on the number of mortgage refinance applications submitted.



MBA refinance index is our target variable and is used as a proxy for mortgage prepayment due to interest rates. We used a total of 26 variables, including US treasury curves, inflation, WTI (West Texas Intermediate) Crude oil, employment data, money supply data, credit data, wood pulp prices, savings rate, US National Home Price Index, and others. We checked for missing data and found none.

As MBA refinance index was started in March of 1990, we have used the data since its inception until June of 2021. There were 11,416 data points in the dataset. And from which 99% was used to train the model, whereas 1% was used to test the model.

**Exploratory Data Analysis**

The independent and dependent variables are all continuous in nature. The histogram of all ‘MBA Refi Index’, ’PPI’, ’Jobless claims’, ‘Personal Savings Rate’,’M1 supply’ are positively skewed. T

Th

The ‘CPI total’ has normal distribution. The ‘CPI Less Food and Energy’, ‘CPI’, ‘10-2yr spread’ and ‘Privately owned housing units’ have negative kurtosis. The plot for the National housing price index shows that prices are always increasing expect during the time of housing crisis in year 2008 when the prices go down. The graph for ‘MBA refinance Index’ is slowly increasing with time with sudden spikes during dotcom bubble, housing crisis and covid-19 pandemic.

**Model Selection**

This research is based on the quantitative perspective where the main goal is to predict the ‘MBA refinance Index’. Once the data is cleaned, the data is divided into training and testing sets. The training set has 99% of data with 11301 records and testing set has 1% of data with 115 records. Correlation coefficient matrix was used to find the highly correlated variables. ‘Moody BBB Yield’, ‘Moody AAA Yield’, ’10 year yield’ and ‘Labor force participation’ was having correlation greater than 0.85 which shows there is presence of multicollinearity.

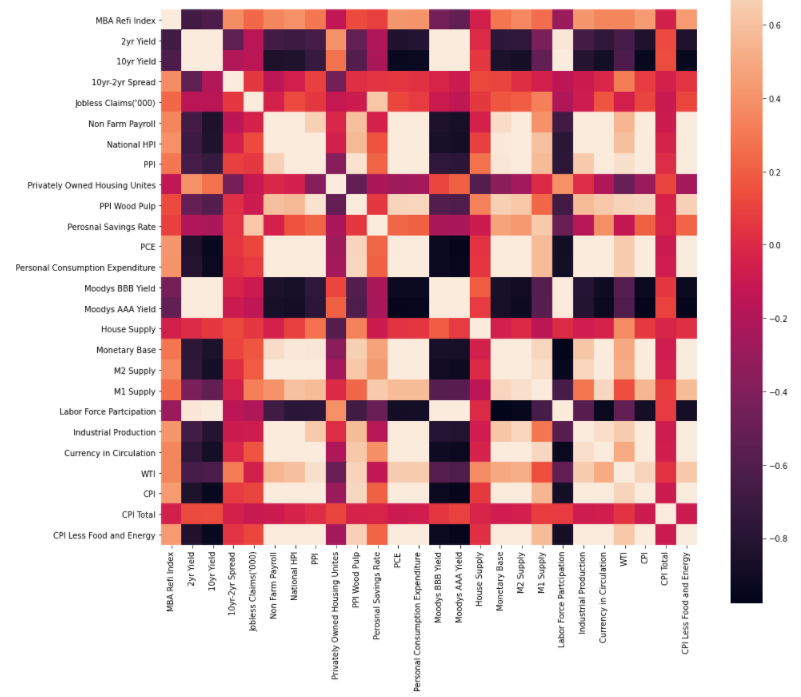


Figure.4. Correlation Matrix between all training features

The data is scaled before it is applied to the model so that each feature contributes proportionately to a fixed range. Once the data is scaled, the training data is fed into a 5-layered sequential neural network. The input layer has several neurons as a hyperparameter, it will scan between 32 and 512 in increments of 32. There are 4 hidden layers. Each of the hidden layers has 256 units. All the layers are activated with ReLU activation function and kernel initializer was normal. ReLU stands for Rectified Linear Unit which is used over other activation functions as it does not activate all the neurons at the same time. The output layer has 1 unit as there is a continuous output and was activated with linear activation function. The hyperparameter tuning learning rate is provided with three different values 1e-2, 1e-3, 1e-4.

The task of determining the best hyperparameters that minimizes a preset loss function and produces better results for a learning algorithm is known as hyperparameter optimization or tuning. Hyperparameters are significant because they regulate a machine learning model's overall behavior. For compilation, Adam optimizer is used because it combines the best aspects of the AdaGrad and RMSProp optimization methods to develop an algorithm that can deal with sparse gradients in noisy environments. First hyperparameter tuning was initiated with 10 epochs to search for an optimal number of units in the first densely connected layer and the optimal learning rate for the optimizer. Then, the model is trained with 500 epochs with a validation split of 20% and callback. Callback is used to save your model when you reach the best metrics after each successful epoch. This saves the model from overfitting as it does stop the training when the model stops improving.

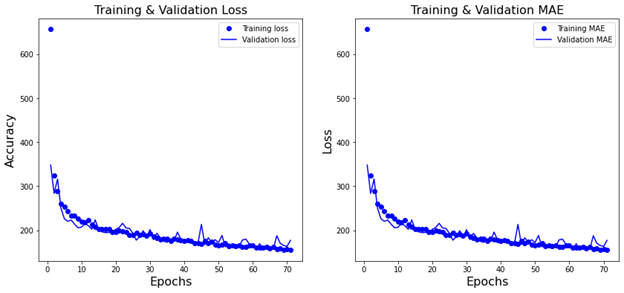


Figure.5. Training and Validation loss for Model with Hyper tuning

**Model Evaluation**

The model was evaluated by MSE (mean squared error), MAE (mean absolute error), RMSE (root mean squared error), MAPE (mean absolute percentage error) and variance score. MSE measures the average squared difference between the predicted and actual values. MAE is the mean of all errors in your prediction. The square root of the mean of the square of all errors is the root mean squared error (RMSE). The RMSE error metric is widely used, and it is regarded as the best for continuous predictions. A forecast system's accuracy is measured by MAPE. It is determined as the average absolute percent error for each time minus actual values divided by real values, and it is expressed as a percentage. The variance score is the variance explained by the model.

**Results**

The model explained above is the best model with MSE as 40556.27, MAE as 201.39, RMSE as 127.79, MAPE as 8.44 and Varscore as 0.98. These results show that a model with all the features (even considering features with high correlation) provides the best scores. Table.1. Shows the comparison of results between different models.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | MSE | RMSE | MAE | MAPE | Varscore |
| 5- layered sequential NN model(with all features) | 91096.52 | 301.82 | 174.11 | 11.07 | 0.94 |
|
| 5- layered sequential NN model (with feature selection) | 157647.36 | 434.24 | 231.35 | 13.68 | 0.92 |
|
| 5- layered sequential NN model (with hyper parameter tuning) | 40556.27 | 201.39 | 127.79 | 8.44 | 0.98 |
|

Figure.6. Comparison between different Evaluation metrics of 3 different models

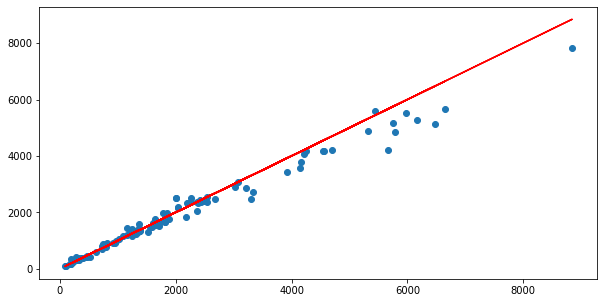


Figure.7. Predicted value and Actual values

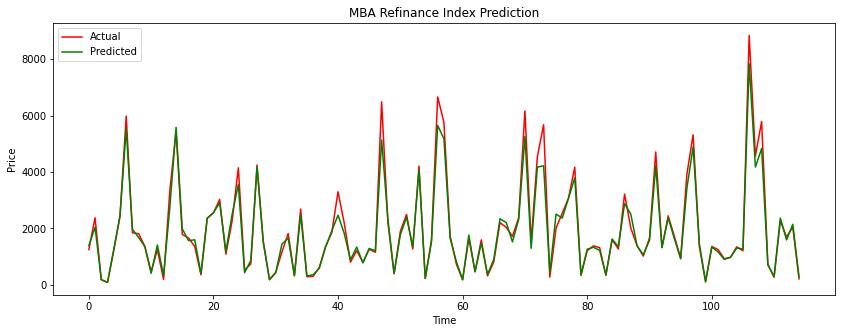


Figure.8. Prediction From March 2021 to June 2021

**Conclusion**

Out of all the models we built, the best performing in terms of MAPE was 5 – layered sequential NN model with hyperparameter tuning. From our domain knowledge of mortgage prepayment, the MAPE value of 8.44% is outstanding. Many prepayment models cost millions of dollars and have performed very poorly. The performance of these models was poor during the Covid pandemic, as we have shown in this project that the mortgage prepayment increases as the rates go down. Mortgage investors need to know the refinance activities of their mortgages as they must optimize their funding. For example, if a mortgage loan pays a 5% interest rate and is funded by 3%, an investor is earning a 2% spread on its investment. But due to lower rates, if the mortgage loan is refinanced at 2.5%, they will be losing 0.5% of their investment. Thus, if mortgage investors can predict this prepayment, they will be better placed to optimize their funding before the prepayment.

Although DNN (Deep Neural Network) is being used mostly in image recognition and NLP, it could be used to predict the quantitative data. The hyperparameter tuning was able to increase the performance of our model. As discussed earlier, we want to use this data set and use the Freddie Mac dataset to build a new model. Freddie Mac dataset has loan-level data with 62 additional features. The additional data could improve the model performance and could be used to predict the refinance and turnover aspects of the mortgage prepayment.

**References**

* Kumar, V. 2014. “Feature Selection: A Literature Review.” Smart Computing Review 4 (3) https://www.cc.gatech.edu/~hic/CS7616/Papers/Kumar-Minz-2014.pdf
* Peter G. Drembelas, Vibhor V. Khar, (2019)  
   https://cs230.stanford.edu/projects\_fall\_2019/reports/26259259.pdf
* Shlomo Amar, (2020, November)https://b-fine.be/wp-content/uploads/2020/12/Modeling\_of\_Mortgage\_Loan\_Prepayment\_Risk\_with\_Machine\_Learning.pdf
* Yu, Y. 2018. “MSCI Agency Fixed Rate Refinance Prepayment Model.” Research paper, MSCI.https://www.msci.com/www/research-paper/msci-agency-fixed-rate/01118884978
* Zhang, J. 2018a. “Is MBS Ref finance Risk Increasing?” Research paper, MSCI <https://jsf.pm-research.com/content/iijstrfin/24/4/local/complete-issue.pdf>
* Federal Reserve Bank of St. Louis. (n.d.). *Federal reserve economic data: FRED: St. Louis Fed*. FRED. <https://fred.stlouisfed.org/>.