Udacity Machine Learning Nanodegree 2020

Capstone Proposal

Neural-Network based Mortgage Prepayment Model

Yuri Turygin

October 2020

1 Domain background

Mortgage Backed Securities (MBS) market is one of the largest fixed income bond markets in the US with a total size of about 15 trillion dollars. Its largest portion is called Agency MBS market and it consists of mortgage loans whose creditworthiness is guaranteed by the Government Sponsored Enterprises (GSEs) such as Fannie Mae, Freddie Mac, and Ginnie Mae, which are all in turn backed by the US Government. The main reason the market exists is to provide mortgage financing to US home buyers. What makes this financing possible is that an originator can to sell a newly originated loan onto the market while freeing up capital that can be used to originate more mortgage loans. The latter (the sale) is possible due to the government credit risk guarantee, i.e. a promise for a loan principal to be repaid even in case of a borrower's default, and the relative liquidity in the Agency MBS sector. The government credit guarantee is an important assuring factor that keeps investors interested in this sector, while good liquidity (i.e. low bid/offer spread) ensures that investors can deploy and withhold large amounts of capital quickly and safely in the sector. Both factors help keep mortgage bonds yields relatively low which directly translates into the lower mortgage rates. Needless to say, Agency MBS market is one of the most important markets in the US.

So, what does the Agency MBS market consists off? It consists of different types of MBS with the most basic building block for them all being a mortgage pool. What exactly is a mortgage pool? A mortgage pool is a fixed income security (a bond) whose underlying collateral, i.e. assets, is a collection of mortgage residential loans. The majority of mortgage payments paid by the borrowers whose loans are in a pool, are directed towards mortgage pool investors, and a small portion of those payments goes towards paying bank mortgage servicing fees and insurance premiums towards the government credit risk guarantee. For example, in a pool of mortgage loans whose borrowers pay 4% mortgage rates on average, all of the repaid principal and about 3% of interest is directed towards mortgage investors in a pool, while about 1% of interest payments is distributed between loan servicers and one of the GSEs (an insurance premium payment to compensate a GSE for the credit risk guarantee). What happens to the cash flows of a mortgage pool if some of the loans get refinanced/repaid? The answer is that a pool sees a spike in cash flow in a form of a loan principal being repaid as a result of which the unpaid principal balance of a pool, i.e. bond notional, is being reduced. The same happens in case of a loan default. In this case a GSE advances the principal of a loan to a pool investor as a result of the credit risk guarantee. It turns out that a fairly significant portion of mortgage borrowers are repaying their mortgages early. This can either be due to a mortgage refinance into a lower rate, a cash out refinance (taking out home equity), house sale, or a mortgage default. All these factors depend on a variety of factors such as prevailing mortgage interest rates (i.e. ability to refinance into a lower rate), home price appreciation (i.e. ability to take out home equity), and the state of the economy (a driver of mortgage defaults) to name a few. What all these factors have in common is that they all change the cash flows of a mortgage backed security, thus, making the task of predicting mortgage prepayments an important task for pricing and risk management of these securities.

2 Problem Statement

In this project we build a neural network-based prepayment model designed to predict the rates at which mortgage borrowers repay their mortgage loans depending on a variety of factors. These repayment rates are called mortgage prepayment speeds. Let us explain how exactly we calculate them.

When a borrower makes her monthly mortgage payment, a part of the payment goes towards paying off interest and another part goes towards paying off mortgage principal. Each month we know exactly how much principal

should be left after a borrower makes her payment. That is actually decided at the time of the mortgage origination and this principal repayment schedule is called mortgage amortization schedule. Suppose, the scheduled remaining principal balance after the N^{th} payment on a mortgage loan is C_N, but a borrower has made a payment higher than the necessary by an amount of p_N, then we define a Single Monthly Mortality (SMM) rate SMM_N to be

$$SMM_N = \frac{p_N}{c_N}.$$

In other words, we measure prepayment speeds as percentage points of the outstanding principal balance. Note that higher than scheduled mortgage payment does not decrease the scheduled mortgage principal for the next month, because the extra payments are applied towards the tail of a mortgage. For example, if you had 100 mortgage payments left, but one month you made two mortgage payments instead of one, then all is different is that you don't need to make your 100^{th} mortgage payments anymore. Thus, the above definition makes sense irrespective of prepayments observed in the previous months.

Although the model we are going to build is designed to predict SMM, people rarely discuss prepayment speeds in SMM units. It is much more common to discuss mortgage prepayment speeds using an annualized version of SMM called CPR (Constant Prepayment Rate). SMM and CPR are connected via the following simple formula.

$$CPR = 1 - (1 - SMM)^{12}$$

In other words, CPR shows the percentage of the principal loan balance to be gone in one year if prepayment speeds remain the same every month for a year.

3 Dataset and Inputs

The dataset we use to build our model is the mortgage pools prepayment dataset provided by Fannie Mae. They provide prepayment speeds and pool characteristics for all the pools whose credit risk they guarantee on a monthly basis. This dataset is rather big, and its total size runs well over 10GB. For the purpose of this study, we will only consider mortgage pools originated in 2010 and later, and we will further restrict our dataset to pools with at least 250 loans in them. This will also help us eliminate some of the noise in prepayment speeds. It should be intuitively evident that the more loans a pool contains, the more stable the prepayment speeds on a pool are.

Here are some of the fields provided in our dataset. Note that they typically represent the average value of a corresponding metric across all loans in a pool.

- Average mortgage rate of loans.
- Average loan size.
- Average age of loans in a pool.
- Average FICO score or borrowers.
- Loan to Value ratio (LTV), i.e. the ratio of the outstanding loan amount to the value of the house.
- What percentage of loans in a pool are investor loans?
- What percentage of loans in a pool are purchase loans?
- What is the geographical composition of a pool by state?
- Which financial institutions or banks are servicing the loans?
- The outstanding total balance of loans in a pool.

And we also supplement the Fannie Mae dataset with a few more fields with the most important of them all being Rate Incentive, which is a difference between average mortgage rates of loans in a pool and the prevailing mortgage rates in the market at the time. It should be intuitively obvious that the higher the rate incentive the higher the prepayment speeds are on a pool.

We have provided our full dataset in csv files on GitHub together with this project. The total size of this reduced dataset is just under 400 MB, which is a lot more manageable.

4 A Solution Statement

In order to solve the problem above we will construct an FF-neural network designed to predict mortgage prepayment speeds given pool characteristics and the current mortgage rate incentive. The model will first be fitted to a larger historical dataset via minimizing the RMSE of model errors. In the process we will also find the suitable hyper parameters, which result in a reasonably good model fit and do not result in overfitting of the test dataset. Then we will examine the model against our evaluation metrics (see below).

5 Benchmark Model

We are not in touch with any industry standard (not neural network-based) prepayment model for the sake of this project, so we will benchmark the goodness of our model predictions against a linear regression model. It seems like a good model candidate to use for checking if our model output makes economic sense or not. After all, any monotone changes in model inputs, i.e. monotone changes of pool attributes, should typically lead to monotone changes in the output of the model, and a linear regression model guarantees that. For example, higher average FICO score should typically lead to higher prepayments, because of borrower's easier access to mortgage credit.

6 Evaluation Metrics

The main evaluation metrics will be the root mean square error of model predictions on a test set – that is going to be the evaluation criteria during the process of hyperparameters turning. But there's much more to it than that. The model will also be evaluated based on its out of sample performance on larger pools (i.e. containing many loans) and also various sub cohorts of a substantial combined current loan balance. In other words, the model will be evaluated on various statistically significant collections of loans. In addition to that we will vary the model inputs (characteristics of a pool) and evaluate the changes in model output. We need to make sure that the changes in input parameters result in reasonable changes in predicted prepayment speeds, i.e. the changes make economic sense. For example, loans of higher loan size tend to prepay faster than loans of lower loan sizes, or pools with higher percentage of investor loans tend to prepay slower than pools with lower percentage of investors.

7 Project Design

We first downloaded the data thru a vendor data service API and stored the data in a series of cvs files. We then loaded the data into memory and examined its quality, filled in NAs, dropped the unusable data rows, and kept only the columns which could be used in a model. We then filtered out the data corresponding to pools containing at least 250 loans and stored this data in a series of csv files, which we then made available on GitHub along with this project. After that we performed a grid hyper parameter search for a series of neural network-based prepayment models and kept the most promising one. We then evaluated the model against our other evaluation metrics (see above).