**PROJECT PROPOSAL FOR DS6040 – TEAM “ANT” (Andrew/Neil/Trey)**

**Fall 2022**

**The ASK: Project Presentation (25% of grade)**

* **Apply Bayesian machine learning to a real dataset in an advanced way to ensure students can apply probabilistic reasoning to a nontrivial problem of their choosing.**
* **The project will be graded based on the approach used and the demonstration of the students’ understanding of probabilistic modeling and not on the comparative performance of different techniques.**

=================================================================================

**Team ANT Project Description** :

**Use BML to determine the probability of whether a mortgage loan will prepay (or default?) in N years (N = a discrete value in the range [1:10]).**

Using publicly available loan level origination and performance data provided from FreddieMac, one of the two so-called “GSE”s (Govt Sponsored Enterprises) who essentially provide a secondary market for mortgage loans. Critically, the GSEs also provide a standard way for institutional investors to access the mortgage debt markets by issuing “MBS” – mortgage backed securities. Investors in MBS are very concerned with understanding prepayment behavior of mortgage borrowers as the early return of principal affects the cashflows and hence the yield of MBS.

The model we are building might be useful for a lender (or loan servicer) in determining how long (and how valuable) a loan servicing relationship might be when examining the potential profitability of a given loan, or perhaps adjusting the pricing on loan offers to compensate for this factor. We will use a publicly available dataset from the FHFA/Freddie Mac which represents all loans purchased by FreddieMac between [1999 – 2021]. We will try to predict the likelihood of a loan prepaying based on an analysis of loan origination details and historical monthly loan performance (payment) data for single family residential (SFR) mortgage loans. The data set can be found [here](https://www.freddiemac.com/research/datasets/sf-loanlevel-dataset) and downloaded by requesting a login for academic research.

**The Project will consist of EDA, Data wrangling, and Developing and testing Various Models**

The project will consist of various parts, including performing EDA on the mortgage loans, and analyzing some of the attributed of mortgage loans by their origination year – so-called “cohorts”. Mortgage lending collects a plethora of interesting borrower data that can be mined to develop models to analyze borrower behaviors over past historical and interest rate environments, as well as allowing for the development of predictive models that look at borrower behaviors like prepayment, but also payment default, which can be very useful for lenders in maximizing profit and minimizing credit losses in their loan portfolios.

The project will involve some wrangling of the data, both to unpack and sample/examine certain cohorts of loans but also to incorporate certain econometric data. For example, we will review the prevailing (market) rate of mortgage loans over the life of each loan, so that we can see how an individual loan was exposed to “refi incentive” – eg, the potential for the borrower to refinance their loan into a lower rate, eg., by measuring for how long and the market rate may have been at some (predefined) lower level during the life of their loan. This can be used as a predictor variable in our analysis and all else equal would be expected to increase the probability of prepayment for a given loan.

**The Datasets contain borrower and loan info at origination, and monthly performance data**

The origination dataset contains a large number of numerical and categorical predictor variables, for example the loan size, the loan’s amortization term (eg., 15/30 year), borrower credit score (eg., FICO), loan rate, etc and many other interesting regressor variables which can be examined for their usefulness as a predictor of prepayment. The origination data table for each loan (the loan data is at the individual loan level) is the following :



The monthly loan data for each loan gives the “UPB” for each month, or unpaid loan balance. If this loan balance goes to zero before the end of the loan term (or before the end of our analysis period N), we know the loan has prepaid (or defaulted, and a code is used to indicate how the loan “ended” early.

We can use the available dataset which is segmented by year of origination to build a series of models using a variety of techniques to determine the drivers of prepayments as well as some properties of the data (EDA) and perhaps we can segment the data in certain ways (eg., by state/region, or by credit score bands).

The monthly data table for each loan looks like :



**Using Test/Training sets and BMA techniques to Analyze prepay drivers and Test predictive models**

Since there is a large amount of data available, we can easily divide the dataset into testing and training datasets, and perhaps use smaller samples and Bayesian techniques along with the use of some simple priors to see if we can predict prepayment accurately, and also by using a “smallish” number of predictor variables. Since it is well established that most 30y mortgage loans (and most *are* 30y terms) do not typically “live” for 30yrs[[1]](#footnote-1) as borrowers may refinance, move, or otherwise sell their homes as they trade up or move around the country in their work or personal lives. According to a popular mortgage lender RocketMortgage “the typical mortgage length, or average lifespan of a mortgage, is under 10 years.” If this is true then there ought to be a very high proportion of paid-off mortgage loans in a given origination year after 10yrs of payments, and it would be interesting to see what proportion is paid-off earlier, say in 3, 5, or 7yrs. And how does this vary by credit score, region, or loan type?

Since mortgage prepayment is also driven by the level of interest rates (and loan age), it will be interesting to incorporate some additional data in the mix, perhaps by trying to assess the relative coupon of prevailing mortgage rates available with the original rate on a given loan as it ages. If the prevailing rate is significantly lower than the rate on the existing loan, and the balance is relatively high, there is an incentive to refinance on the part of the borrower. We can calculate how many months the loans’ life experiences such an incentive and use this as a predictor in our analysis. However, it’s key to note that in real life predictions, predicting whether a borrower will have the chance to refinance down the road is hard, so using this variable can be something we assess, etc.

**Next Steps: pulling down data, and getting into EDA and wrangling first**

In any case the work ahead seems pretty interesting and challenging and we look forward to digging into the analysis and the data. We will probably begin some data scrubbing and EDA in an attempt to skinny down the data into some manageable pieces, as we don’t envision using anything like what’s available on offer at the Freddie site – as the full dataset represents some 41 million loans! But the good news on this site is that there is random samples for each year available with 50,000 loans selected. This means we can get going right away on the analysis.

Prepayment modeling is a well known and studied field in the realm of fixed income research, and the focus there is generally ***understanding the rate of prepayment over time*** (and/or in the near future) to model/predict the cashflows (and the yield) of mortgage backed securities (MBS) investments. We are using our newly-learned Bayesian modeling techniques here to understand the probability of prepayment more generally here, but the exercise ought to be both interesting and challenging to boot.

We have already pulled some sample data and discussed our strategy for building the regression model data from the available samples. We look forward to getting the greenlight and any suggestions for making the exercise as productive and educational as we can.

1. Life and death is a theme in mortgage terminology: in the analysis of MBS and mortgage prepayment speeds there is a concept called “single monthly mortality” rate or SMM, and there is also the concept of “average life”, or the average time to receive principal payments over the life of an mortgage which is a proxy for duration. The word mortgage itself evolved from Old English for "dead pledge”. [↑](#footnote-ref-1)