**Future Development**

This initial analysis of the Fannie Mae data uses two modeling techniques to answer one question, whether a borrower will default. There are more opportunities for analysis and prediction using this data. There are also other methods and modifications to the existing analysis that might improve the prediction of default.

**Further Default Prediction Analysis**

There are more modeling techniques that might provide better prediction of a borrower defaulting. A Bayesian forecast technique could be implemented assuming a distribution on the current loan delinquency status from the previous month to predict whether a borrower will default.

A Random Search CV method of random forest optimization is thought to yield more robust solutions than the XGB Boost method that was used here. It would be of interest to compare the accuracy of different random forest modeling methods.

**Data Enhancements**

This study focuses on a narrow timeline to draw conclusions about defaulting early in the life of a mortgage. Defaulting is possible at any stage of a loan’s lifetime. Including data over a broader timeline would create a more robust model could predict default in more mature mortgages.

To better capture the regional differences in the economic wellbeing in the country, including an MSA level Case Shiller Index would provide another variable that likely has a high correlation with defaulting trends.

**Additional Questions the Data Can Answer**

While this study focuses on defaulting (missing 4 or more payments), an early look at delinquency could also be useful. The likelihood of missing a payment given the current loan delinquency status could be modeled using a Bayesian Logistic Regression, where the current delinquency status has an assumed distribution.

Another outcome of interest is whether a loan will be prepaid. It would be interesting to look at factors that are predictive of whether a borrower will prepay their loan.

Another way to look at the data is to consider how long it takes for a customer to default or miss a payment. Estimating the time until default could be accomplished using a hazard survival model where defaulting is the event of interest and loans that do not default in the observed time period are considered censored. Considering 3 possible ways for a mortgage to end, payment until maturity, default, and prepay, a competing risk hazard model could predict the time until any event that might result in the loan not reaching maturity.