Step3: [BUILDING AND EVALUATING MODELS]

Used the Forward, Backward, Stepwise and Exhaustive Search resulted in 17 variables namely "CHANNEL\_C", "PROPERTY TYPE\_MH", "PROPERTY STATE\_AR", "PROPERTY STATE\_TX", "ORIGINAL LOAN-TO-VALUE (LTV)", "OCCUPANCY STATUS\_P", "SELLER NAME\_BANKOFAMERICA,NA", "PROPERTY STATE\_NC", "OCCUPANCY STATUS\_I", "PROPERTY STATE\_MN", "ORIGINAL UPB", "ORIGINAL LOAN TERM", "LOAN PURPOSE\_P", "MORTGAGE INSURANCE PERCENTAGE", "PROPERTY STATE\_AL", "FIRST TIME HOMEBUYER FLAG\_N", "ORIGINAL COMBINED LOAN-TO-VALUE”.

Utilized these variables and run various algorithms such as Linear Regression, Random Forest, Neural Networks on Q1 2006 as training and Q2 2006 as testing. Linear Regression performed the best among other algorithms.

Further used Linear Regression to predict Q2 2007, Q3 2007, Q4 2007 and Q1 2008 on rolling basis.

From the results the RMSE, MAE, MAPE we found that as we trained more and more data the predictions improved over the Quarters.

I will not recommend using Linear Regression model because it is very user friendly for the people who are not aware of the complexities of the model. As neural networks and Random Forest gives a good accuracy will give a better accuracy but it is difficult to understand the working of the model.