Predicting mortgages prone to foreclosure during economic recession

_____

Kayode Ayankoya, PhD

Table of Content   **Page**

1.0 Introduction

The 2008 recession was an eye-opener for the government of many countries, organizations, industries and individuals. One of the common occurrences during periods of recession is that some mortgage loans that were not at risk of default in a normal economy become susceptible to foreclosures due to significant changes in financial status of borrowers. This means that the normal risk analysis models may not necessarily identify borrowers that will default during adverse economic situations. As a result, the percentage of mortgage foreclosures increased significantly during last economic recessions that we have on record. Previous research has shown that the number of foreclosures could be more than double of the normal rate of foreclosures in a stable economy. Therefore, it is important to lenders to be able to identify the characteristics of mortgages on their books that are at risk of foreclosures in case of another economic recession.

2.0 Problem Statement

It is believed by many that another recession is imminent, although no one knows exactly when the next recession will be or what will be the triggers. Hence, several organizations are trying to prepare - by looking into their risk profiles and that of their customers. In order to provide some guidance for lenders, this work was carried out to predict mortgages that may be foreclosed during a recession. Specifically, the main problems that this study focused on is defined as:

1. What are the characteristics of mortgages at risk of foreclosure during economic recession and during the initial recovery period, typically 12 months after the last month of economic decline.
2. How far backwards or close to a recession can mortgages at risk of recession be detected?
3. To what degree of accuracy can we predict a mortgage at risk of foreclosure, in case of recession.

3.0 Data and Methodology

Fannie Mae, a government-sponsored enterprise that offers different types of mortgage loans provides open access to Single Family Mortgage dataset that was used for this study. The dataset is structured into 2 categories (Acquisition and Performance) on a quarterly basis. The acquisition dataset contains static data, while the performance dataset for each acquisition is updated monthly.

The dataset consists of nearly 22 million records with a static acquisition file and a one to many monthly performance loan-level details for each quarter of the year since inception. Using the 2008 recession as the reference point, the period of 01 January 2008 and 31 December 2010 is considered as the recession and recovery period. Thus, only mortgages foreclosed during this period were considered as affected by the recession.

4.0 Data Acquisition and Pre-processing

The acquisition and performance data is made available as separate files for each quarter on Fannie Mae's website. Furthermore, the website also provides an R script for aggregating the performance data into single metrics for each mortgage. Due to the hardware limitation it was not practicable to use the entire dataset. Therefore, this study opted to use stratified sampling to obtain a representative fraction of the dataset suitable for the study.

In order to sample from a quarter, the acquisition data was imported into postgres.hereafter, a sample of the acquisition data was extracted into a new table, followed by importing the entire performance data, and then creating a separate table for performance data of the obtained acquisition data sample. The R code is then run on the sample acquisition and relevant performance datasets to generate the final dataset. This process was carried out for quarters of loans acquired between 2005 - 2008 resulting in 266432 mortgages being included in the study. Out of these are 144 mortgages reflected loan acquisition between 1999 and 2004. These were thought to be data error, but it was decided to leave these mortgages as part of the analysis. The table below shows the yearly distribution of when the loans sampled were initiated.

Year	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
Count	14	2	13	44	272	368	45439	108395	52495	5939

5.0 Features selection and engineering
The process described above resulted in a final dataset of a total of 75 features. But after a careful examination and literature review, only 28 of these features seem relevant to this study. Furthermore, 9 of the 28 variables were dropped much later in the study because they had strong correlation with other variables. It should be noted that one of these 28 variables is a new variable created to represent the ratio of the last unpaid balance and the total loan value. The final list of features included in the study is presented in Appendix 1. One hot encoding was used to convert categorical variables into multiple binary variables,significantly increasing the number of predictor variables, especially with the state becoming 50 (minus 1) variables.

6.0 Exploratory Data Analysis
As mentioned earlier, previous statistics show that there is a significant increase in the amount of foreclosure during and immediately after a recession. The figure below shows that there is a significant increase in the number of foreclosures in 2008 and 2009. The number started to decline in 2010, but the effect can still be seen up until 2012.
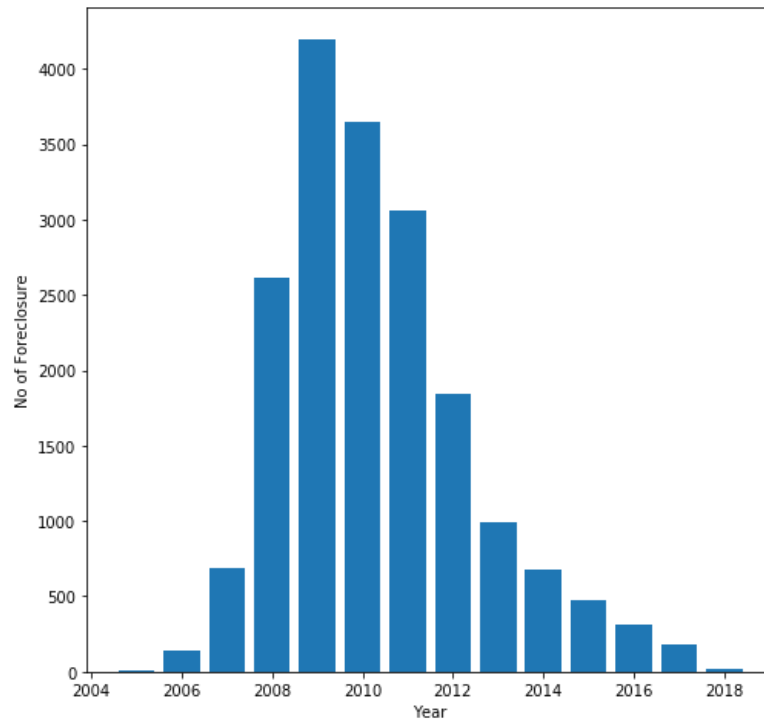
Figure 1: Distribution of foreclosures by year

During the initial stage of the analysis,the impact of variables such as interest rate, loan to value and cumulative loan to value ratio on foreclosures were also considered. Figure 2 below suggests that borrowers with high loan to value ratio (i.e low equity in the property) and with higher interest rate may be more susceptible to foreclosures in case of a recession.

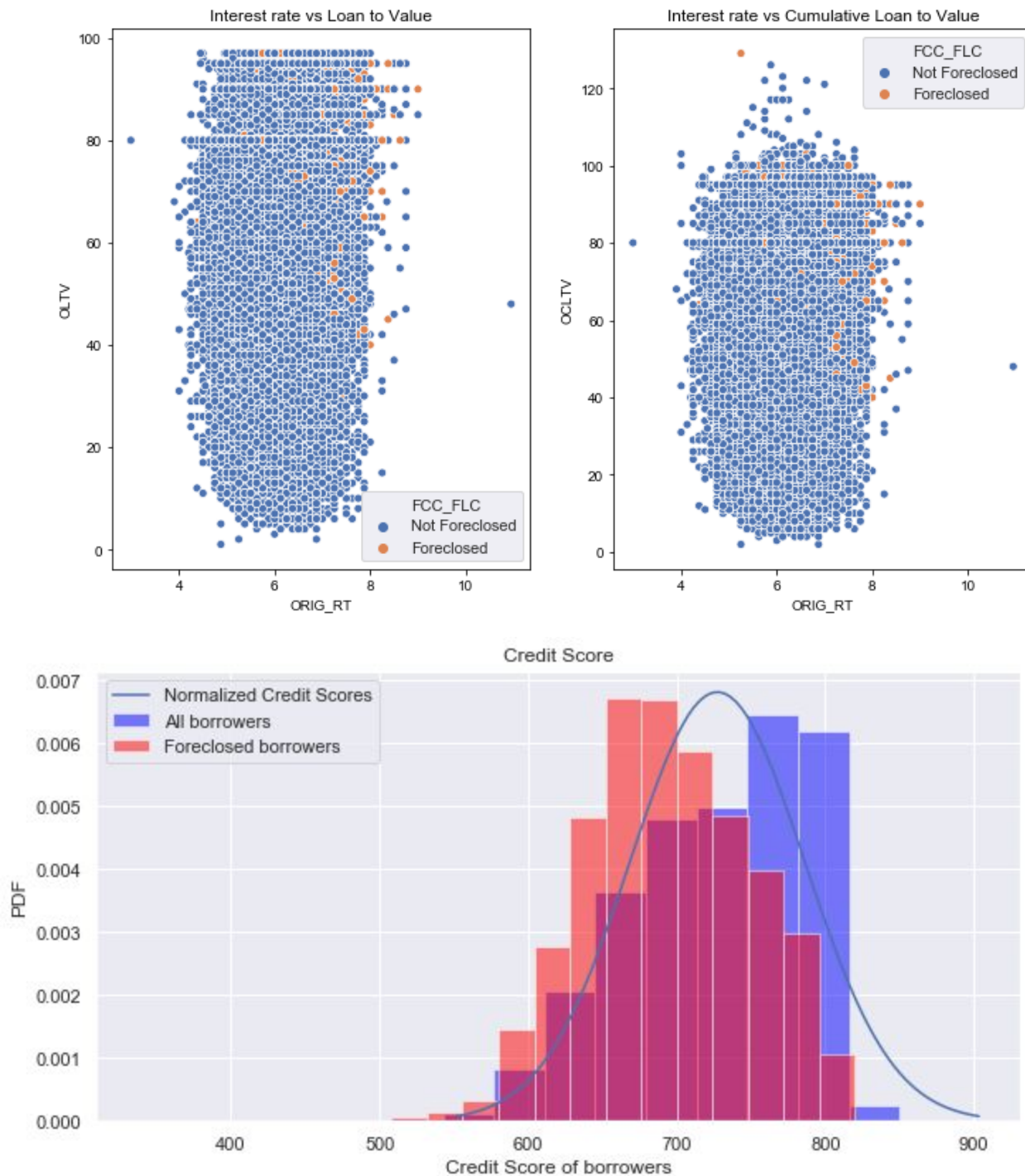Figure 2: Plot of Original Interest Rates vs Loan to Value by Foreclosures

Figure 3: Credit score of foreclosed borrowers compared to the population

Furthermore, it was of interest to find out if there is a significant difference in the original credit score of borrowers whose mortgages foreclosed compared to the others. Figure 3 shows that the borrowers with lower credit scores are more likely to become more more affected by a recession as expected. However, these

exploratory analysis are a not sufficient in understanding the characteristics of mortgages at risk of foreclosure during a recession.

7.0 Predictions

The main objective of this study is to identify mortgages that will foreclose during or immediately after a recession. With this in mind, this study used the 2008 recession as the reference point. Samples Mortgages extracted from the Fannie Mae website were initiated between 2005 - 2008 indicating that resulting model should be suitable to predict 3 - 4 years ahead. The performance of each of these mortgages was available up until 2018, but foreclosure was defined by mortgages that had their last payments between 01-Jan-2008 and 31-Dec-2010 even though their original terms extends beyond this time. All the others were considered as not-foreclosed for the purpose of this study.

7.1 Pre-processing

After all the process described in previous segments, the resulting dataset had 233883 rows and 84 columns. This was normalized for better performance of predictive algorithms using the python Standard Scaler library. Thereafter, the data was divided into training and test sets for cross validation of the model as shown
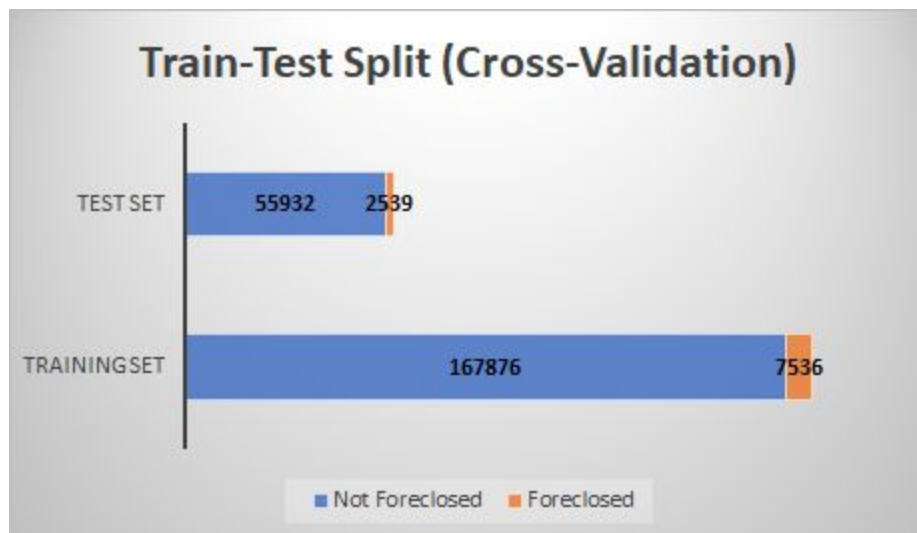

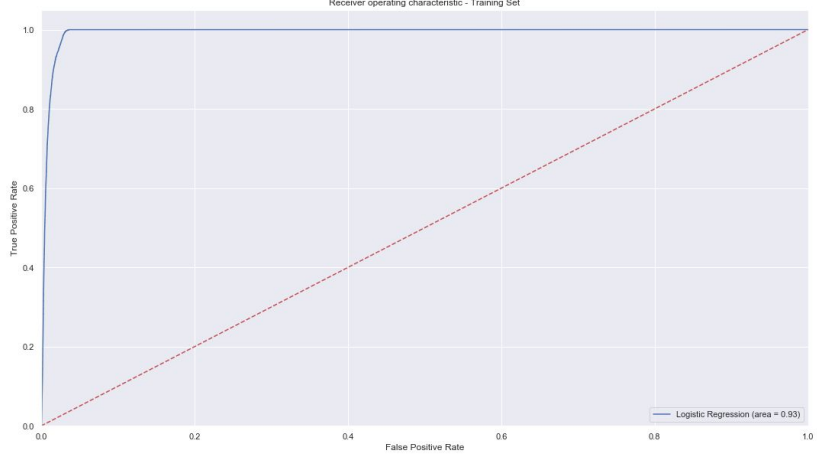
Figure 4: Cross-validation

7.2 Models

In building a baseline model to classify mortgages that will foreclose and those that will not, experiments were carried out with Support Vector Machine(SVM) and Logistic regression. However, it was decided that SVM is not the best option because it was too slow due to the volume of data, there was a need to assume linearity in order to identify feature importance and the first few experiments have only marginal increase in accuracy when compared to logistic regression.

7.3 Logistic Regression

After hyper-parameter tuning and K-fold cross-validation experiments, the resulting baseline logistic regression model using all the acquisition and 4 performance/calculated variable had 97.50% accuracy score. However, accuracy score is not a sufficient measurement of model predictive capability in this case because of the
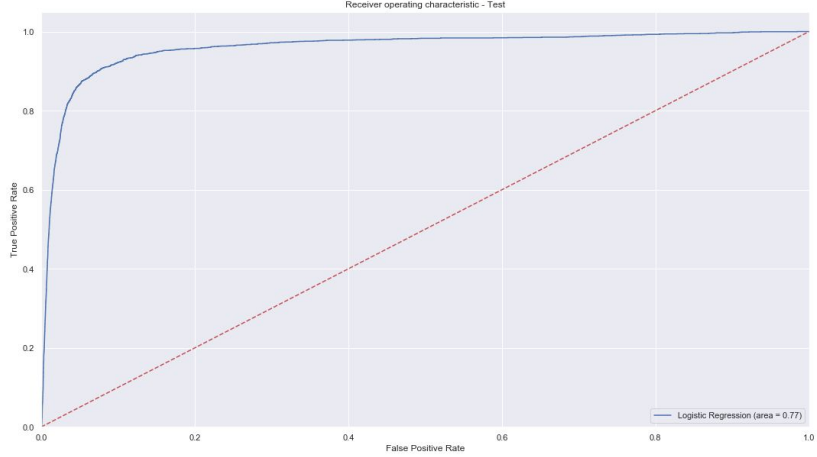
unbalanced data. Hence the confusion matrix and the percentage of precision/recall was explored. Below is the outcome of the resulting dataset.

Baseline Logistic Regression model

| Metric	Result	ROC - AUC
Accuracy	97.50%	
Precision	72%	
Recall	88%	
AUC	93%	

Based on the nature of this problem, one of the goals of this study was to characterize foreclosure/non-foreclosures based on the acquisition data available at the inception of each mortgage together with as little as possible performance data that is only available over time. Several experiments were carried out after the initial model. But based on simplicity, explainability and the problem at hand the model with all the acquisition variables, the Zero Balance code (performance variable) and the percentage of last unpaid balance to the initial value of property (calculated from acquisition and performance variables) was selected as the best. The result is shown below.

Result of Logistic Regression with all acquisition and 2 performance features

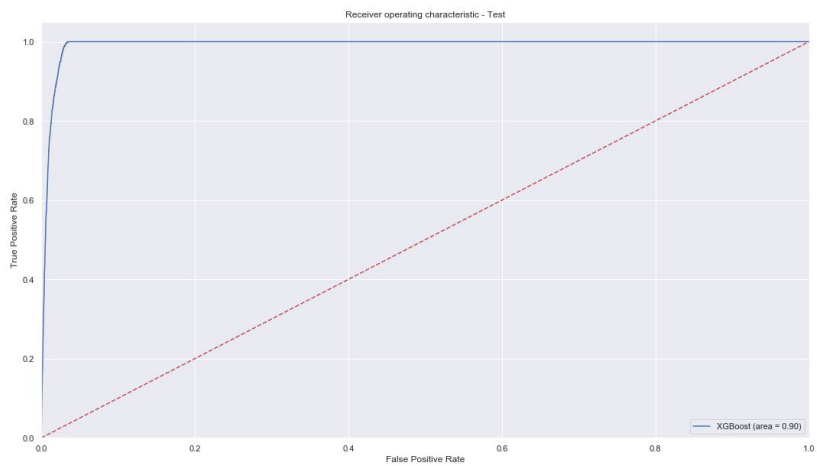| Metric	Result	ROC - AUC
Accuracy	96.78%	
Precision	67%	
Recall	56%	
AUC	77%	

One of the advantages of using logistic regression is the degree of importance of each of the input features that is provided with the coefficients of the model. Figure 5 provided below shows the top 16 features of the logistic regression. This will be discussed further later in this report.

Figure 5: Feature importance (logistic regression)
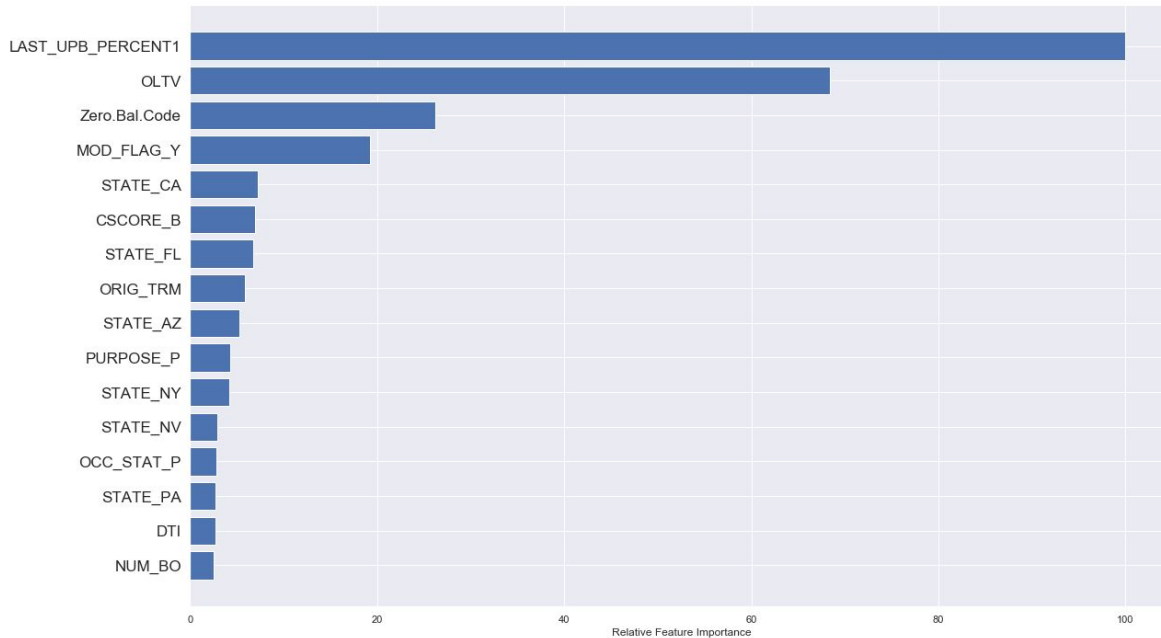
7.4 Gradient Boosting Model (XGBoost)

Gradient boosting uses an ensemble of weak decision trees for a better classification. All the experiments carried out with logistics regression were repeated with gradient boosting. The result was a much better model for predicting foreclosure in case of recession using features available during acquisition of the loan, the ratio of original property value vs last unpaid balance and the zero balance code. The performance of the selected gradient boosting model is presented below.

Metric	Result	ROC - AUC
Accuracy	97.86%	
Precision	73%	
Recall	81%	
AUC	90%	

Receiver operating characteristic - Test

This is a much better performance compared to the logistic regression model using the same set of features and perhaps the model with the best performance out more than 30 modelling experiments that were carried out.

7.5 Feature Importance

The coefficients of the logistic regression model described earlier that used the same variables as the best performing XGBoost gave an indication of how each variable impacts the predictions and effectively the characteristics of loans that will foreclose during a recession, For example, figure 6 below shows the coefficient of the top 20 variables in the logistic regression model.
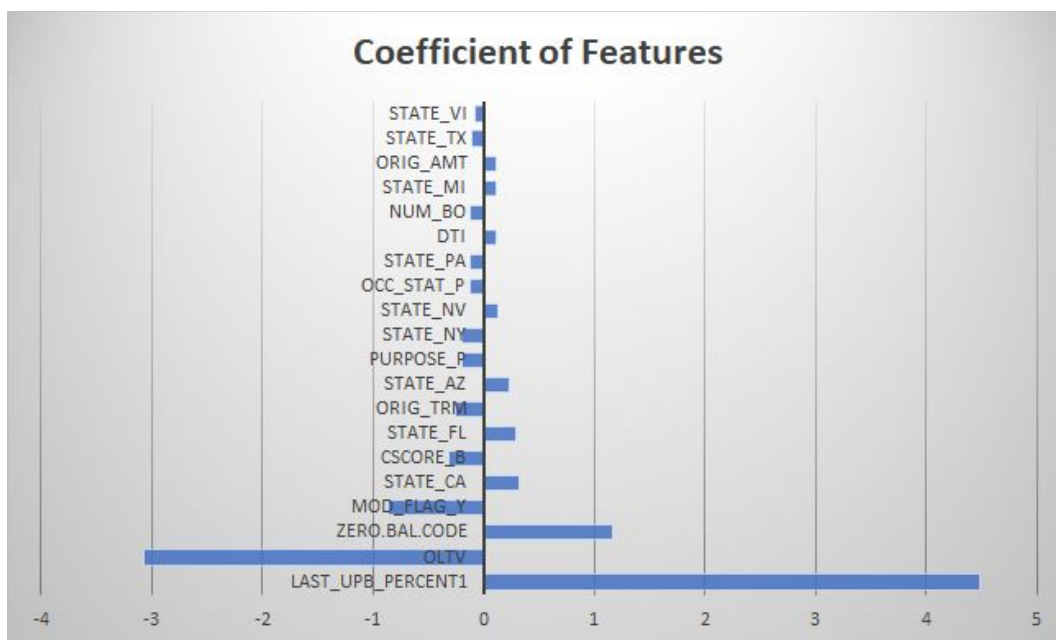

Figure 6: Coefficient of Features

The plot shows that those with higher ratio of original value to last unpaid balance (i.e low equity) are more likely to foreclose, which is also echoed in the original loan to value at the beginning of the mortgage. Interestingly, properties in California, Florida, Arizona, Nevada and Miami are more like to foreclose in that order, while there is an indication that properties New York, Pennsylvania, Texas and Virginia have a lower probability to foreclose in the case of a recession. Furthermore, it also shows that modification of properties after the initial mortgage should be? encouraged, perhaps because owners become emotionally attached, hence they do more to prevent foreclosure during recession.
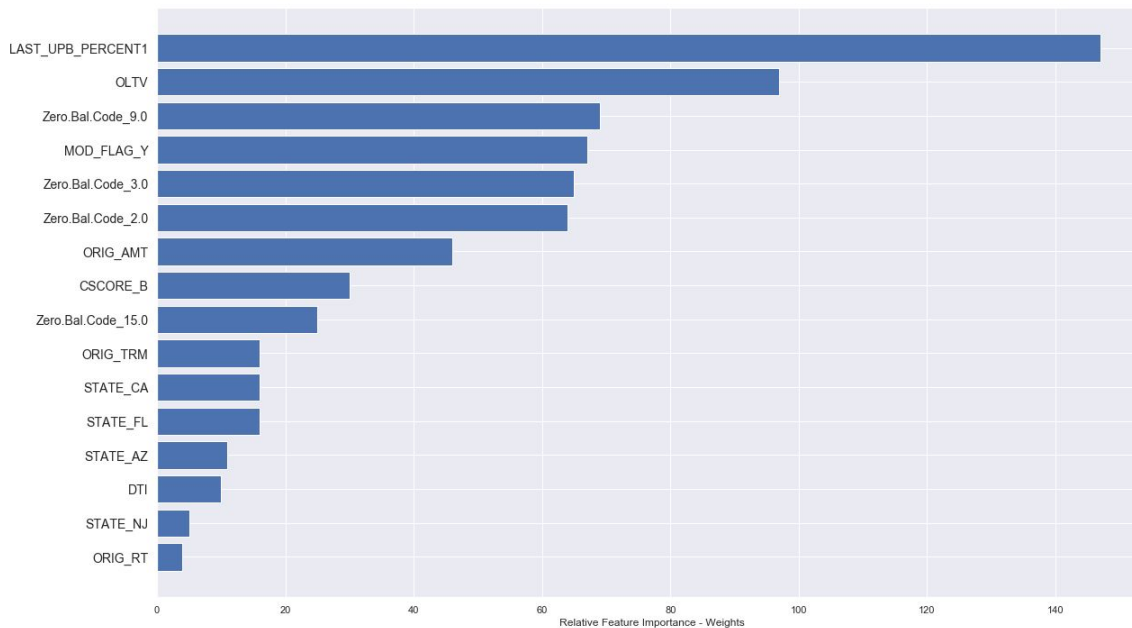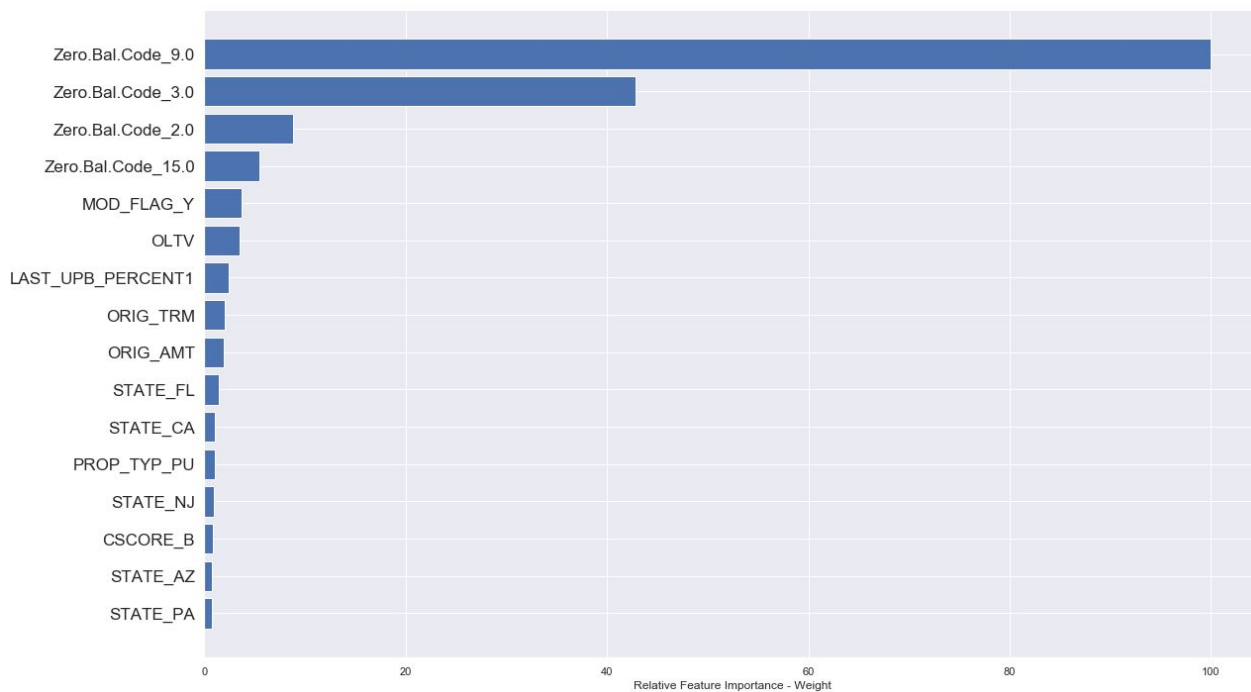

Figure 7: XGBoost Feature Importance (Weights)

Figures 8: XGBoost Feature Importance (Gain)

This fact is supported by the feature importance of the XGBoost model. Figure 7 shows the feature importance by weights, i.e. the number of times a feature appear in decision trees, while Figure 8 indicates the contribution of each feature to the accuracy of the model.

8.0 Conclusion

Different experiments were carried out using Support Vector Machines(SVM), Logistic Regression and Gradient Boosting (XGBoost). XGBoost had the best result in predicting which loan will or will not foreclose during a recession with a weighted 97.86% accuracy. Specifically, 73% accuracy was achieved in identifying loans that are prone to foreclose depending mainly .

Overall, the results show that mortgage with low equity to value ratio that the property is based in California, Florida or Arizona that has never been modified (improved/extended) are more prone to foreclosure during a recession that the others Other important factors that makes a mortgage prone to foreclosure in recession are If it is not a primary residential property, the original term is above average and the debt to income ratio of the borrower is in the upper limit of what is acceptable.

Finally, the distribution of the confusion matrix per year show that predictions are more accurate as we move closer to the year of recession.
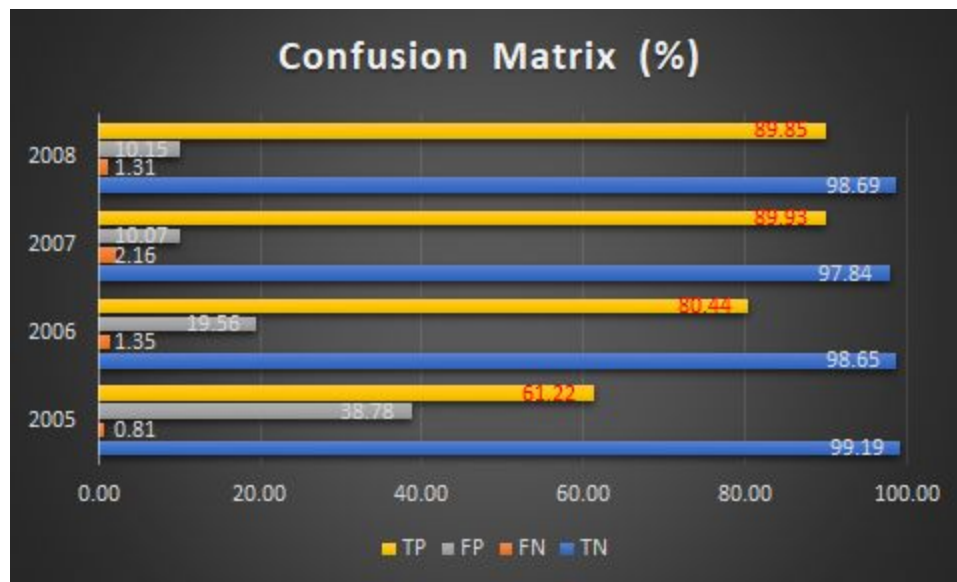

Figure 9: Distribution of Confusion Matrix per year

Appendix 1

S/N	Feature Description	Name	Source
1	CHANNEL	ORIG_CHN	Acquisition
2	ORIGINAL INTEREST RATE	ORIG_RT	Acquisition
3	ORIGINAL UNPAID PRINCIPAL BALANCE (UPB)	ORIG_AMT	Acquisition
4	ORIGINAL LOAN TERM	ORIG_TRM	Acquisition
5	ORIGINAL LOAN-TO-VALUE (LTV)	OLTV	Acquisition
6	NUMBER OF BORROWERS	NUM_BO	Acquisition
7	DEBT-TO-INCOME RATIO	DTI	Acquisition
8	BORROWER CREDIT SCORE	CSCORE_B	Acquisition
9	FIRST-TIME HOME BUYER INDICATOR	FTHB_FLG	Acquisition
10	FIRST-TIME HOME BUYER INDICATOR	PURPOSE	Acquisition
11	PROPERTY TYPE	PROP_TYP	Acquisition
12	NUMBER OF UNITS	NUM_UNIT	Acquisition
13	OCCUPANCY STATUS	OCC_STAT	Acquisition
14	PROPERTY STATE	STATE	Acquisition
15	RELOCATION	RELOCATION_FLG	
16	MINIMUM CREDIT SCORE	CSCORE_MN	Calculated
17	Ratio of Original loan amount and Last unpaid Balance	LAST_UPB_PERCENT	
18	Ratio of Original property value and Last unpaid Balance	LAST_UPB_PERCENT1	
19	MODIFICATION FLAG	MOD_FLAG	
20	ZERO BALANCE CODE(Prepaid, Short-Sale, Repurchased or Deed-in-lieu)	Zero.Bal.Code	