TIME DEPENDENT MORTGAGE LOAN DEFAULT PREDICTION WITH MACHINE LEARNING

Prosper Anyidoho, Peter Gomillion, Olivia Mwangi

1.0 Introduction

Real estate plays a fundamental role in the U.S. economy. Housings combined contribution to the U.S. GDP generally averages between 15-18%. Real estate is commonly purchased using mortgages, which are debt instruments secured by the collateral of real estate property, that the borrower is obliged to pay back with a predetermined set of payments. If the borrower stops paying the mortgage, the lender can foreclose, which is a legal process where a lender will recover the property which secured the loan by evicting the tenants and selling the property to recover the balance of the mortgage.

A foreclosure is devastating to the individual(s) borrower affecting their ability to get approval for another mortgage. Not only that but the impact of foreclosure is felt sharply by the provider of the mortgage. According to the Financial Times, three of America's biggest banks have set aside a combined $28 bn for current and future loan losses. This causes a negative impact on banks' earnings as they must set aside loan provisions as mandated by The Federal Reserve to mitigate the risk of a bank failing from too many loan losses. Also, investors of said banks are hurt as well since regulators also cap the dividends allowed given the new regulations put in place since the financial crisis of 2007-2008 to mitigate the risk that banks will not have the cash on hand to continue to provide their essential function in the economy. So, can banks mitigate the risk they endure by providing mortgages that could go into foreclosure? This paper will analyze publicly available information to understand the variables that influence loan delinquency for loans 2 years after loan origination. We will also build models to predict if someone will default in the first 2 years of the mortgage lifespan as well as the time the mortgage defaults. In our application we focused on Fannie Mae loan data from 2013Q1 to 2018Q4 to train our models. The machine learning classifiers adopted help to discriminate between default and active loans while the survival analysis methods provide time dependent analysis of probabilities.

2.0 Data Description

In this project, data from multiple sources were merged to get the final data on which the models were trained and tested. Datasets include loan mortgage data from Fannie Mae in the form of loan quarterly acquisition and performance data, data on local and national economic factors (unemployment rates, house price index, median household income, divorce rates)

2.1 Mortgage Loan Data

We narrowed our analysis to Single Family Loan Acquisition and Performance datasets which can be accessed publicly from the Fannie Mae website. The loan database consists of loans from 2000 up to date. To simplify our methods, we only focus on loans that originated from 2013Q1 to 2018Q4 with about 12million loans existing in this period. Importantly, the loan data only contains information for fully amortizing, full-documentation, single family and conventional fixed-rate mortgages(Mae, 2015).

Acquisition data files contain information about the borrower, property, and the loan itself at the origination date. The borrower characteristics has information on borrower and co-borrower's credit score at loan origination, Debt-to-income ratio, number of borrowers and loan to value ratio (OCLTV). Zip code with only first three digits and the property type describe property features. The loan features are explained by the Original loan amount, loan purpose and many more as shown in Table 1 below.
Performance data includes information such as the balance on the loan and delinquency or foreclosure status as shown in Table 1 below. The detailed explanation of the rest of the performance variables can be found at the Appendix section. Performance data continues until the loan either liquidates (matures, gets repurchased, etc.) or goes under either delinquency or foreclosure.

In order to do our analysis, we create a combined data set from the acquisition and the performance files. Static variables from the acquisition are merged with summary of monthly performance data for each loan. We derived the last status of the loan as our dependent variable from the performance file. Also, we engineered new features such as the date for 90-day delinquency of loans, the Unpaid principal balance, Foregone interest Cost, Net loss and Net severity. Details of final variables obtained from merging acquisition and performance files are shared in the Appendix. The final data prior to cleaning contains 12M loans from 2013Q1 to 2018Q4. The data cleaning process involved subsetting loans that are either current or entered 90+ days of delinquency within 24 months after loan origination. The data after this process has the current loans being 7M while the defaulted loans only amount to 11,358. We further cut down on the number of current loans to 101,556 in order to minimize the problem of imbalanced classes. This was done by using stratified sampling based on the year of loans origination to ensure we generalize our data across years to avoid bias. Also, we removed all loans with missing FICO values. Machine learning models perform better when the frequency of classes in the dependent variables are somewhat equal so we create another dataset by further under sampling the predominant class and in this case the current loans(Sealand and Group,

2018). This gives us two datasets- balanced and imbalanced dataset. A full descriptive statistic of the final data variables is shown in Tables 2 and 3

2.2 Local and economic features

We also added external data in the form of unemployment rates, house price index, and median household income(Farzad, 2018). Bureau of Labor Statistics (BLS) provides local area unemployment per year across states. We used the annual unemployment rate at the state level corresponding the year of origination for each loan. We also include Housing Price Index (HPI), obtained from Federal Housing Finance Agency, at the state level to control for the fluctuations of price of properties. Median household income data for each state is also added to the dataset and this data is obtained from Bureau of Census

Table 1: Loan mortgage and external features

Loan Variables		Local and economic variables
Acquisition variables	Performance variables	
Channel	Monthly reporting month	Unemployment rate
Seller name	Servicer name	House price index
Original interest rate	Current interest rate	Median household income
Original unpaid principal balance (upb)	Current actual unpaid principal balance	Divorce rate
Original loan term	Loan age	
Origination date	Remaining months	
First payment date	Adjusted remaining months	
Original loan-to-value (ltv)	Maturity data	
Original combined loan-to-value (cltv)	Metropolitan Statistical area	
Number of borrowers	Current loan delinquency status	
Debt-to-income ratio (dti)	Modification flag	
Borrower credit score	Zero balance code	
First-time home buyer indicator	Zero balance effective date	
Loan purpose	Last paid installment date	
Property type	Foreclosure date	
Number of units	Disposition data	
Occupancy status	Foreclosure cost	
Property state	Property preservation and repair cost	
Zip (3-digit)	Asset recovery costs	
Mortgage insurance percentage		
Product type		
Co-borrower credit score		

Table 2: Summary statistics of continuous variables

	Balanced data		Imbalanced data	
	Mean	SD	Mean	SD
Original Interest rate	4.41%	0.70%	4.13%	0.65%
OCLTV	78.61%	16.82%	74.95%	17.44%
Number of units	1.03	0.21	1.03	0.24
Original Loan Value	$295784.00	$192194.36	$313149.80	$210797
Original amount	$218312.78	$119020.83	$217426.70	$116471.60
Unemployment rate	4.87%	1.31%	5.21%	1.45%
House Price Index	$598.81	$232.72	$577.87	$226.60
Median Income	$60724.16	$8845.95	$59843.35	$8668.21
Divorce Rate	3.05%	0.56%	3.09%	0.57%
Federal funds rate	0.88%	0.70%	0.64%	0.65%

Table 3: Summary statistics of categorical variables

Variable		Balanced data		Imbalanced data	
	Levels	**No. of Obs.**	**%Obs**	**No. of obs.**	**% Obs**
Loan Status	0 Active	11358	50.00	101556	89.94
	1 Default	11358	50.00	11358	10.06
Number of borrowers	0 One	8506	37.44	53560	47.43
	1 More than one	14210	62.56	59354	52.57
Loan purpose	0 Cash-out	5560	24.48	25246	22.36
	1 Purchase	13042	57.41	60455	53.54
	2 Refinance	4114	18.11	27213	24.10
Property type	4 Single Family	14803	65.17	72236	63.97
	0 Condo	1901	8.37	10086	8.93
	1 Co-Op	78	0.34	561	0.50
	2 Manufactured	219	0.96	907	0.80
	3 Planned Urban	5715	25.16	29124	25.79
	4 Single Family	14803	65.17	72236	63.97
Occupancy status	0 Investor	1329	5.85	9594	8.50
	1 Principal	20731	91.26	98394	87.14
	2 Second	656	2.89	4926	4.36
FICO brackets	0 [0-620)	5	0.02	9	0.01
	1 [620-660)	4183	18.41	8305	7.36
	2 [660-700)	5258	23.15	15782	13.98
	3 [700-740)	4615	20.32	22741	20.14
	4 [740-780)	4464	19.65	31340	27.76
	5 [780+)	4191	18.45	34737	30.76

3.0 Analysis Methods

As already stated, we seek to predict whether a loan will experience at least 90 days of delinquency within the first two year after origination, so we deployed machine learning classifiers. The machine learning classifiers used are explained below. In the process we split our two datasets (balanced and imbalanced) into training and testing sets. We used 75% of the data for training and the other 25% to check the out of sample predictive power of our models. In each case we further conducted 10-fold cross validation on our models by splitting the training data into 10 folds. After observing none of the continuous variables are normal and are all projected on different scales, we standardized and re-train our models. For each model we used the same set of variables as well as the same training, validation and testing sets. Variables as listed in Tables 2 and 3, were selected based on initial exploratory analysis and based on literature. All categorical variables as were one hot encoded into independent variables according to the number of levels for each variable.

Also, in predicting the time of default for each loan, we resorted to survival analysis approaches. The cox proportional hazard model which is semi-parametric, models the risk of default as a function of the covariates. The cox model helps us understand the effect of each variable on the probability of loan default over the 24 months and can also be used to compare probabilities of default across risk different groups. To predict individual probabilities of defaults across 24 months, we implemented a Multi-task logistic regression.

3.1 Logistic Regression (LR)

Logistic Regression is a popular and very useful algorithm of machine learning for classification problems. The advantage of logistic regression is that it is a predictive analysis. It is used for description of data and used to explain relationship between a single binary variable and single or multiple nominal, ordinal and ration level variables which are independent in nature. The model development for the prediction is taken in account using the sigmoid function in logistic regression as the outcome is targeted binary either 0 or 1.

3.2 Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis, or LDA, is also a predictive modeling algorithm for both and multi-class classification. It can also be used as a dimensionality reduction technique, providing a projection of a training dataset that best separates the examples by their assigned class. LDA is a technique for multi-class classification that can be used to automatically perform dimensionality reduction.

3.3 K-Nearest Neighbor (KNN)

KNN is a simple, easy-to-implement supervised machine learning algorithm that is used to solve both classification and regression problem. It relies on labeled input data to learn a function that produces an appropriate output when given new unlabeled data. For a fixed value of k, KNN assigns a new observation to the class of majority of the k nearest neighbors. A new

observation was assigned to class one (y = 0) if k1 > k2 and assigned to class two (y = 1) if k1 < k2.

3.4 Gradient Boosting Machine (GBM)

Mortgage default is a binary classification problem, either a mortgage default or it does not. In gradient boosting machines, or simply, GBMs, the learning procedure consecutively fits new models to provide a more accurate estimate of the response variable. The principle idea behind this algorithm is to construct the new base-learners to be maximally correlated with the negative gradient of the loss function, associated with the whole ensemble. The loss functions applied can be arbitrary, but to give a better intuition, if the error function is the classic squared-error loss, the learning procedure would result in consecutive error-fitting. In general, the choice of the loss function is up to the researcher, with both a rich variety of loss functions derived so far and with the possibility of implementing one's own task-specific loss.

3.5 Naive Bayes classifier (NB)

A Naive Bayes classifier is a probabilistic machine learning model that's used for classification task. The crux of the classifier is based on the Bayes theorem, given a naive Bayes model, you can make predictions for new data using Bayes theorem.

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

Naive Bayes assumes label attributes such as binary, categorical or nominal.

3.6 AdaBoost (AB)

This is one of boosting algorithms. At a high level, AdaBoost is similar to Random Forest in that they both tally up the predictions made by each decision trees within the forest to decide on the final classification. There are however, some subtle differences. For instance, in AdaBoost, the decision trees have a depth of 1 (i.e. 2 leaves). In addition, the predictions made by each decision tree have varying impact on the final prediction made by the model.

3.7 Decision Trees (DT)

Classification and Regression Trees are binary decision trees, attempting to classify a pattern by selecting from a large number of variables the most important ones in determining the

outcome variable. Using these attribute values, the decision tree generates a class as the output for each input data.

3.8 Random Forests (RF)

RF is a tree-based ensemble with each tree depending on a collection of random variables, for a p-dimensional random vector representing the real-valued input or predictor variables and a random variable Y representing the real-valued response, we assume an unknown joint distribution PXY (X, Y). The goal is to find a prediction function f(X) for predicting Y. The prediction function is determined by a loss function L (Y, f(X)) and defined to minimize the expected value of the loss where the subscripts denote expectation with respect to the joint distribution of X and Y.

3.9 Extra Tree Classifier (ET)

Extremely Randomized Trees Classifier (Extra Trees Classifier) is a type of ensemble learning technique which aggregates the results of multiple de-correlated decision trees collected in a "forest" to output its classification result. In concept, it is very similar to a Random Forest Classifier and only differs from it in the manner of construction of the decision trees in the forest.

Each Decision Tree in the Extra Trees Forest is constructed from the original training sample. Then, at each test node, each tree is provided with a random sample of k features from the feature-set from which each decision tree must select the best feature to split the data based on some mathematical criteria (typically the Gini Index).

3.10 Cox Proportional Hazard

The cox proportional hazard model provides a way to model time to event data which aligns with our problem definition. It is of importance not only to be able to predict whether there will be a default or not but also to predict the time of default as this will help financial institutions to design effective measure before giving out a loan. In a typical survival analysis application, some individuals never get the chance to observe the event and such individuals are considered censored observations

The hazard function at each time t is given by:

$$h(t|X) = h_0(t)e^{\beta(X)}$$

where $h_0(t)$ is the baseline hazard function at time *t*, $e^{\beta(X)}$ is the risk associated with the covariate values, X.

The survival probability function is therefore formulated as

$$S(t|X) = S_0(t)e^{\beta(X)}$$

Where

$$S_0(t) = e^{-\int_0^t h_0(x)dx}$$

Based on the cox regression formula, we construct a partial likelihood function from the data as follows:

$$L(\beta) = \prod_{i:\delta=1} \frac{\theta_i}{\sum_{j:t_j \geq t_i} \theta_j} \quad (4)$$

Where $\theta_i = exp\beta(X)$, and we solve the optimization problem to obtain optimal values of parameters, $\hat{\beta}$. With estimate parameters, we can calculate the baseline hazard function by replacing the β with the optimal $\hat{\beta}$. The $h_0(t_i)$ is therefore written as:

$$\widehat{h_0} = \frac{1}{\sum_{j \in R(t_i)} \theta_j} \quad (5)$$

Where t_i is the event time for subject i, $R(t_i)$ is the set of all subjects who have not experienced the event and are still at risk at time t_i.

In summary, we learn the cox parameters, β_s and $\widehat{h_0}$ and for each individual in the data we estimate

$S(t|X)$ from equation 9, which is the probability loan not defaulting by time *t.* We can then estimate the probability of the loan defaulting in time, t given the loan has not defaulted in previous months as:

$$F(t|X) = 1 - S(t|X)$$

3.11 Multi-task Logistic Regression (MTLR)

Multi-task logistic regression models the survival data as a series of independent logistic regression models. The logistic regression models are built on each time step with the entire time interval to estimate the probability of event occurrence in each time period. In our case, we will model the probability of default as a sequence of binary decisions. The mathematical details of the MTLR can be found in (Yu *et al.*, 2011). The MTLR offers the advantage of estimating individual time dependent probabilities as compared to the relative risk cox model, which only provides inferences and a way to compare risk across different groups.

4.0 Results and Interpretation

4.1 Machine learning classifiers

The cross-validation results for each machine learning classifier is shown Table 4 below. A 10-fold cross validation was implemented for each model on both the balanced class data and the imbalanced class data. Also, we run each model on the original variables as well as the transformed variables (standardization). Performance of models were assessed using accuracy, precision, recall and the F1 scores. Accuracy represents the percentage of defaulted and active loans correctly predicted by the model. Precision stands for the proportion of predicted defaulted loans that are actually correct while Recall refers to the proportion of actual defaults that were correctly identified or predicted. F1 score is simply the harmonic mean of the precision and the recall. In loan mortgage prediction, recall is the most important metric as the cost of incorrectly predicting a defaulted loan is high. Models applied in this field should be able to minimize false negative rate and thus have high recall values. Also, in reality, the number of loans that enter default is generally less as compared to the active and prepaid loans so using accuracy might not be a good measure of model performance. From the cross validation results all models performed well in terms of accuracy. However, transforming the variables by putting them on the same scale improved model performance. Also, the models fitted on the balanced data performed better than models fitted on imbalanced data. The imbalanced data models have higher accuracy levels but perform poorly on recall and precision due to the problem of class imbalance. The trees models such as the random forest, gradient boosting and Ad boost seemed to perform better than other models. The full results are summarized in table 4 below. It also obvious that, the balanced transformed models performed better on recall as compared to the imbalanced transformed models.

Table 4: Cross Validation results

	Original Variables								Transformed Variables							
	Balanced				Imbalanced				Balanced				Imbalanced			
Model	Acc	Pre	Rcl	F1	Acc	Pre	Rcl	F1	Acc	Pre	Rcl	F1	Acc	Pre	Rcl	F1
LR	0.62	0.60	0.75	0.68	0.90	0.01	0.01	0.01	0.83	0.82	0.85	0.84	0.91	0.61	0.32	0.42
LDA	0.83	0.81	0.86	0.84	0.90	0.51	0.43	0.47	0.83	0.81	0.86	0.84	0.90	0.51	0.43	0.47
QDA	0.50	0.79	0.01	0.01	0.82	0.33	0.78	0.46	0.50	0.20	0.01	0.01	0.82	0.33	0.78	0.46
KNN	0.61	0.61	0.61	0.61	0.90	0.47	0.16	0.24	0.82	0.80	0.85	0.82	0.91	0.56	0.40	0.46
DT	0.78	0.78	0.77	0.77	0.88	0.41	0.45	0.43	0.78	0.78	0.77	0.77	0.88	0.42	0.45	0.43
NB	0.66	0.62	0.80	0.70	0.90	0.42	0.02	0.03	0.80	0.75	0.88	0.81	0.80	0.31	0.81	0.45
AB	0.84	0.83	0.85	0.84	0.91	0.59	0.37	0.46	0.84	0.83	0.84	0.84	0.91	0.59	0.37	0.46
GBM	0.84	0.83	0.87	0.85	0.91	0.63	0.37	0.46	0.84	0.83	0.87	0.85	0.91	0.63	0.37	0.46
RF	0.84	0.83	0.87	0.85	0.92	0.67	0.37	0.47	0.85	0.83	0.86	0.85	0.92	0.67	0.37	0.48
ET	0.84	0.83	0.86	0.84	0.92	0.63	0.38	0.47	0.84	0.83	0.86	0.84	0.92	0.63	0.38	0.47

We further tested the out of sample predicted power of our models by applying them to the test set as shown in Table 5 below. Results look similar to cross validation results in the sense that the balanced trained models averagely performed better on the imbalanced trained models. The highest recall value is 88% and 82% obtained from the Naïve Bayes algorithm for both the balanced data and the imbalanced data respectively. Due to the absence of

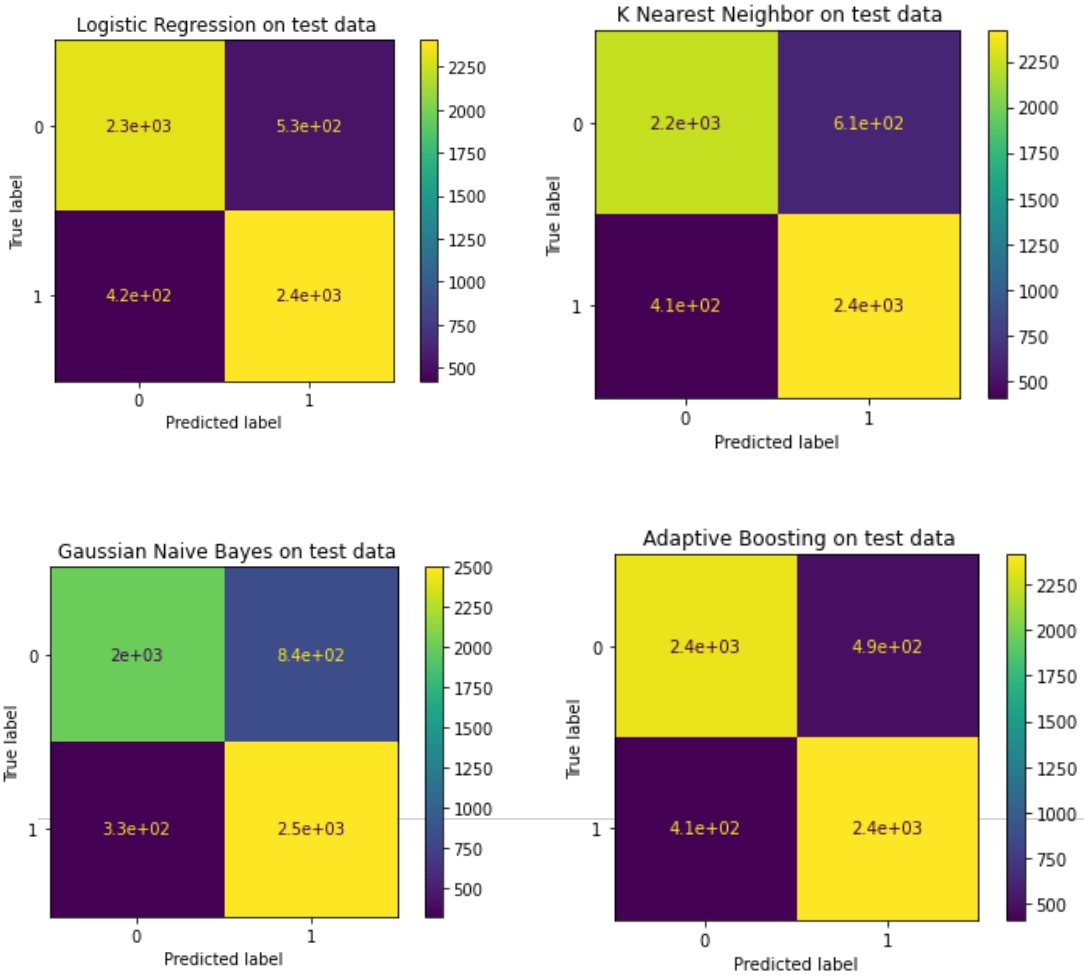
computational power, we could not optimize the hyper-parameters for each of these models. With optimal hyper-parameters we are sure to improve the predictive power of the models.

A plot of confusion matrix is also provided for selects models. The confusion matrix shows how the predicted values compare with the actual classes. In each confusion matrix, defaulted loans are labeled as 1 while active or current loans are labeled as 0.

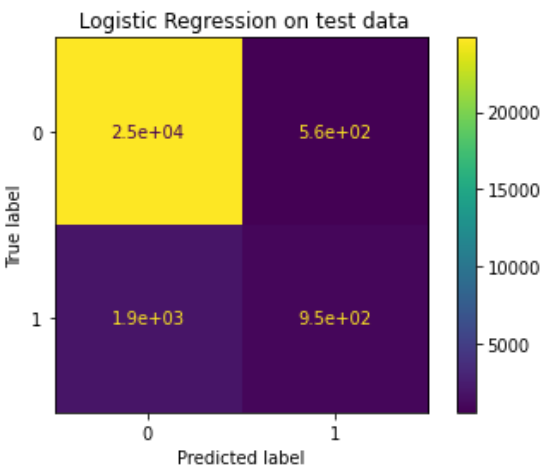
Table 5: Test results for select models for transformed models

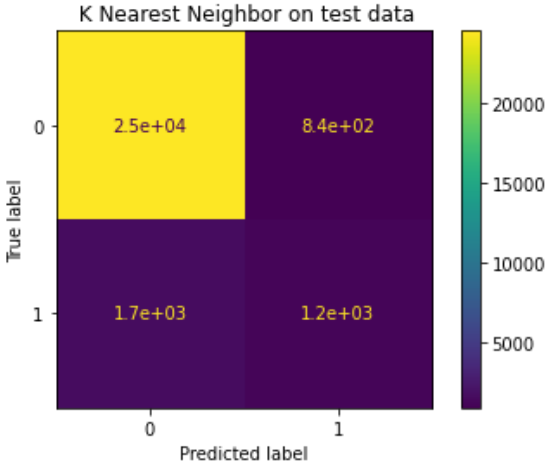
Transformed Variables								
	Balanced				Imbalanced			
Model	Acc	Pre	Rcl	F1	Acc	Pre	Rcl	F1
LR	0.83	0.82	0.85	0.84	0.91	0.63	0.33	0.43
LDA	0.83	0.81	0.86	0.83	0.90	0.53	0.44	0.48
QDA	0.5	1.00	0.1	0.1	0.81	0.33	0.78	0.46
KNN	0.82	0.8	0.86	0.83	0.91	0.58	0.41	0.48
CART	0.77	0.77	0.76	0.77	0.88	0.43	0.47	0.45
NB	0.8	0.75	0.88	0.81	0.79	0.31	0.82	0.45
AB	0.84	0.83	0.85	0.84	0.91	0.60	0.39	0.48
GBM	0.85	0.83	0.87	0.85	0.92	0.64	0.39	0.48
RF	0.84	0.83	0.86	0.84	0.92	0.67	0.38	0.48
ET	0.83	0.82	0.85	0.84	0.92	0.64	0.39	0.48

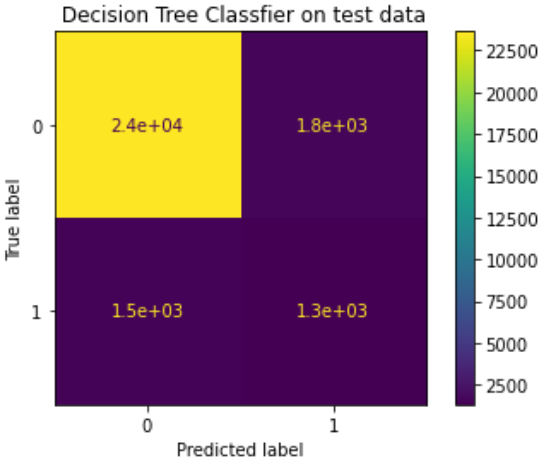
Confusion Matrix on Test Data for select models (Balanced data)

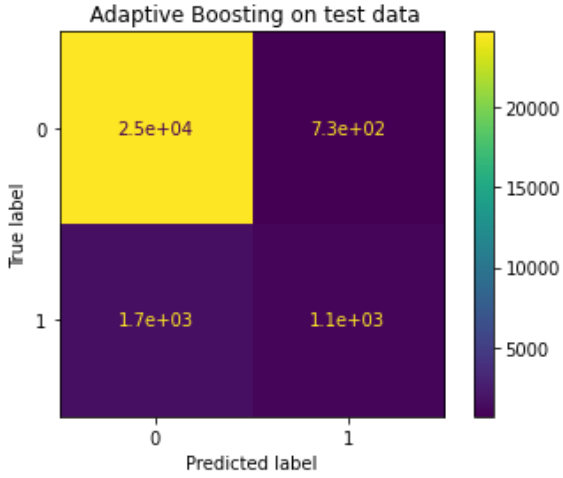


Confusion Matrix on Test Data for select models (Imbalanced data)









4.2 Cox proportional Hazard Model

For predicting time dependent mortgage default, we need to use survival analysis approaches. We run cox models on both balanced and imbalanced data to understand the impact of the variables on the probability of default over the 24 months. Table 5 below shows the estimated coefficients, p-values and the concordance index. The concordance index (C-index) provides a measure similar to the area under the ROC curve but also considering censored loans. All current loans were treated as right censored since they did not undergo default. The C-index represents an assessment of the discriminative power of the model. C-index values of 0.79 and 0.89 for balanced and imbalanced class models respectively represents a good fit for our data. Generally, variables with positive coefficients increase the probability of default while negative coefficients decrease the probability of default. For example, the Original interest rate with p-value less than 0.05 has a positive coefficient of 0.4 for the balanced data model and this shows, holding other variables constant the probability of default increases by exp (0.4). A positive coefficient was expected for this variable since loans with higher interest rate are of high risk and has higher likelihood of defaulting. It is also worth knowing that, as the FICO score of a loan increases the probability of default decreases. The probability of mortgage default increases given a loan with a higher interest rate, a high loan amount, a high loan to value ratio, a high debt to income ratio, a high House Price Index, a high Federal funds rate and also given the individual is a first-time homebuyer purchasing a single-family property and with a principal occupational status.

Divorce rate produced unexpected coefficient sign. We expected a positive sign for this variable but the model produced a negative sign. This could be attributed the usage of statewide values for each loan which might not be representative of the exact divorce rates in the applicant's locality. Also, this could be attributed to the problem of multicollinearity in our data which might interfere with model signs.

Table 5: Results of Cox proportional hazard model

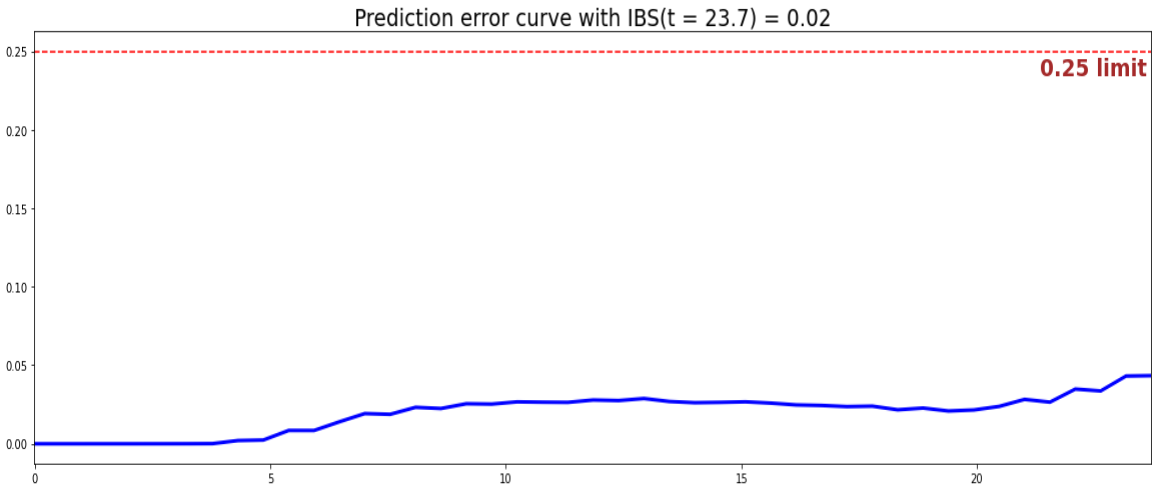
Variables	Balanced data		Imbalanced data	
	Coef.	p-value	Coef.	p-value
Original interest rate	0.4007	<0.0001	0.519	<0.0001
Original amount	0.0172	<0.0001	0.021	<0.0001
Original combined loan-to-value	0.0110	<0.0001	0.019	<0.0001
Debt-to-income ratio	0.0230	<0.0001	0.033	<0.0001
Number of borrowers	0.6672	<0.0001	0.928	<0.0001
Number of units	-0.1705	0.00770	-0.255	<0.0001
Original Loan value	-0.0082	0.00048	-0.007	0.003093
Unemployment rate	-0.1146	<0.0001	-0.149	<0.0001
House Price Index	0.0004	0.5048	0.002	0.000698
Median Income	-0.0021	0.1797	-0.007	<0.0001
Divorce Rate	-0.0358	0.0437	-0.078	<0.0001
Federal funds rate	0.1989	<0.0001	0.318	<0.0001
First-time home buyer indicator	0.1031	<0.0001	0.264	<0.0001
Purpose_Purchase	-0.2123	<0.0001	-0.315	<0.0001
Purpose_Refinance	-0.2702	<0.0001	-0.338	<0.0001
Property type_Co-Op	-0.5711	0.0342	-0.743	0.005837
Property type_Manufactured	-0.0305	0.7575	0.052	0.596544
Property type_Planned Urban	-0.0029	0.9389	0.029	0.461371
Property type_Single Family	0.0811	0.0236	0.132	0.000213
Occupancy status_Principal	0.4034	<0.0001	0.634	<0.0001
Occupancy status_Second	0.0681	0.5044	0.123	0.209331
Ficobkt_[620-660)	-0.7493	0.0953	-1.185	0.008247
Ficobkt_[660-700)	-1.1207	0.0126	-1.826	<0.0001
Ficobkt_[700-740)	-1.5874	0.0004	-2.621	<0.0001
Ficobkt_[740-780)	-2.2562	<0.0001	-3.437	<0.0001
Ficobkt_[780+)	-3.1549	<0.0001	-4.414	<0.0001
Coefficient of Concordance	0.793		0.889	

4.3 Multi-task Logistic Regression (MTLR)

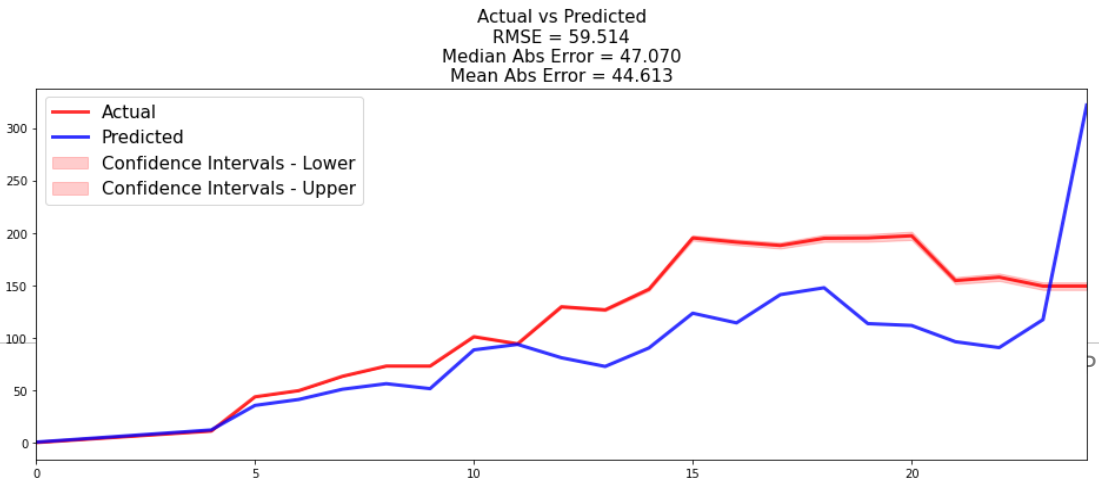
MTLR model was trained on both datasets (balanced and imbalanced) and the out of sample predictive power was estimated using the C-index and the Brier score. The Brier score measures the accuracy of the predicted survival curve at any given time t and it represents the average squared distance between the actual survival status and the predicted survival probability. A good model typically has an integrated brier score less than 0.25 and our model has a value of 0.02 and 0.0 for balanced and imbalanced data respectively which shows an excellent predictive performance. Our models also record a C-index of 0.97 and 0.95 for balanced and imbalanced data respectively, which means our models have strong discriminative power in deciphering between active and default loans.

We also estimated the number of defaulted loans in the test set over the 24 months period. This generated a curve over the 24 months and we compare this predicted curve to the actual curve as observed in the data. We compute errors of our predictions using Root mean squared (RMSE), mean absolute error and median absolute error metrics. The test set for the balanced data reported RMSE of 44 which means over the entire 24 months we over-estimated the number of defaulted loans averagely by 44. The imbalanced data however recorded an error of 87. In both cases, our models did well in predicting defaults in each month until the last month where errors were large.

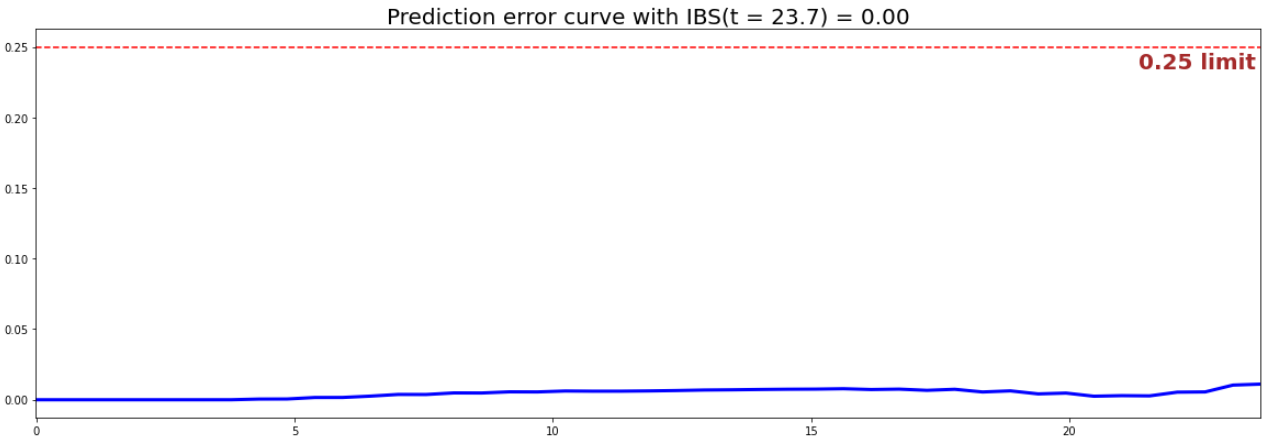
Balanced data Brier Score curve across 24 months after loan origination



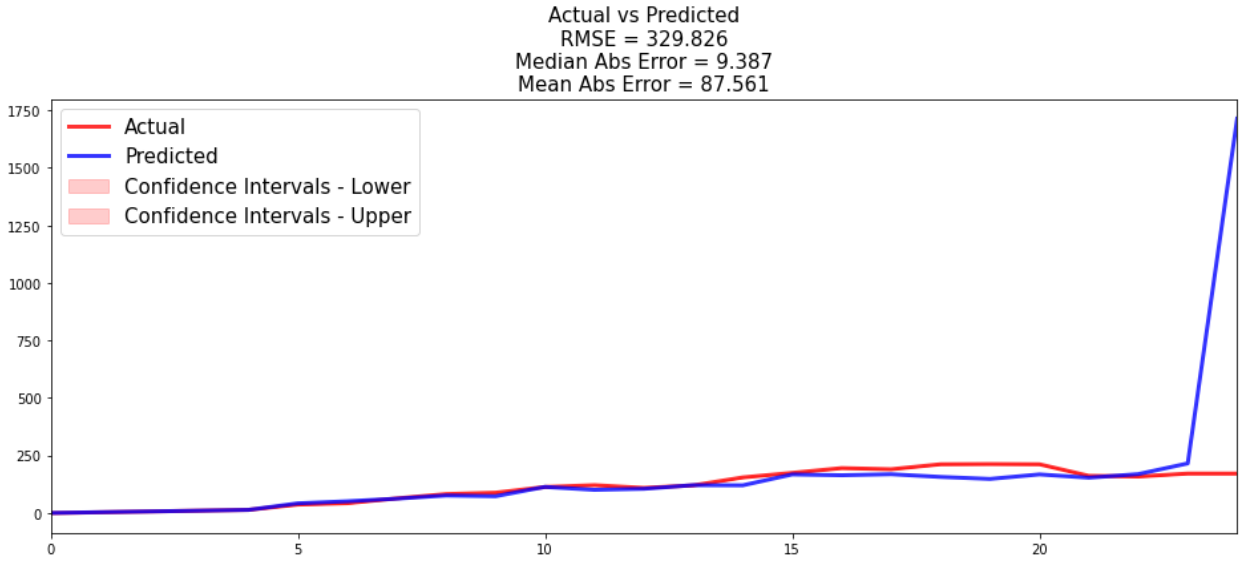
Balanced data predicted versus Actual curve across 24 months after loan origination



Imbalanced data Brier Score curve across 24 months after loan origination



Imbalanced data predicted versus Actual curve across 24 months after loan origination



5.0 Conclusion

This paper utilizes statistical and machine learning approaches to predict mortgage loan default. First, we applied Machine Learning techniques, to mortgage loan data from 2013Q1 to 2018Q4 to predict +90-day delinquency 2years after loan origination. We also applied cox proportional method, a statistical survival analysis approach, to understand the effect of each variable on the probability of default. Finally, we implemented the multi-task logistic regression model which estimates the probability of default in each of the 24 months after loan origination. Each of the machine learning models performed well in terms of predictive accuracy but for loan mortgage prediction we are mostly interested in using recall to assess our models. The Gaussian Naïve Bayes approach gave us the highest recall values of 88% and 82% for balanced and imbalanced class datasets respectively. In the cox model, we found the original loan interest rate and debt to income ratio Original interest rate, Original combined loan-to-value, Debt-to-income ratio, Number of borrowers to be positively related to the probability of default. Also, an individual's FICO score plays a crucial role on the tendency of a loan entering default. High FICO scores decrease the probability of default. In terms of individual default probability predictions, our MTLR model performed excellently with average brier score and c-index values of 0.01 and 0.96 across all models.

In order to improve model results in the future, we can tune hyper-parameters for each machine learning classifier in order to get to optimal solutions. Also, in predicting time dependent probabilities, the variables used are all static and this does not really capture what happens in reality as mortgage payments are affected by dynamic attributes. For example, external variables such as unemployment rate, divorce rate are dynamic as such future models should be able to adequately integrate the dynamic nature of these variables for reliable mortgage default prediction.

References

Farzad, T. (2018) 'Determinants of Mortgage Loan Delinquency : Application of Interpretable Machine Learning', pp. 1–33.
Mae, F. (2015) 'Loss Data Analysis Mae ' s credit risk performance data', pp. 1–59.
Sealand, J. and Group, E. I. (2018) 'Short-term Prediction of Mortgage Default using Ensembled Machine Learning Models Short-term Prediction of Mortgage Default using Ensembled Machine Learning Models Jesse C . Sealand Requirements for the degree of Master of Science (Data Analytics) in the Graduate School of Mathematics and Statistics Slippery Rock University July 2018', (July). doi: 10.13140/RG.2.2.30004.76169.
Yu, C. *et al.* (2011) 'Learning Patient-Specific Cancer Survival Distributions as a Sequence of Dependent Regressors', pp. 1–9.

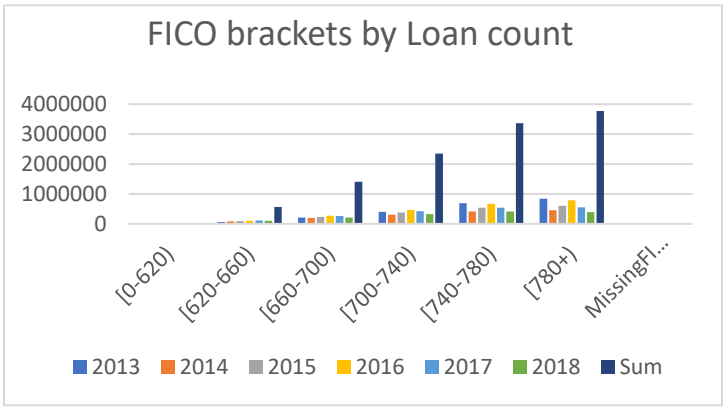
Appendix
Acquisition variables for combined data

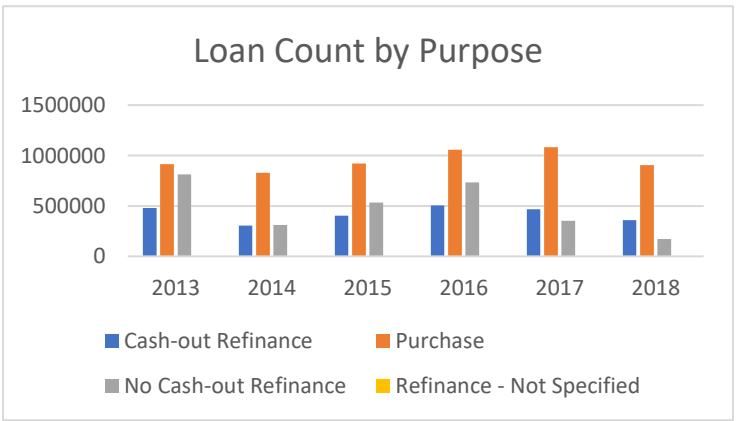
Position	Field Name	Column ID	Type	Max Length	Allowable Values
1	LOAN IDENTIFIER	LOAN_ID	Alpha-Numeric	20	
2	CHANNEL	ORIG_CHN	Alpha-Numeric	1	R - "Retail"
					C - "Correspondent"
					B - "Broker"
3	SELLER NAME	Seller.Name	Alpha-Numeric	80	Name of Seller
4	ORIGINAL INTEREST RATE	ORIG_RT	Numeric	14,10	Blank = Unknown
5	ORIGINAL UNPAID PRINCIPAL BALANCE (UPB)	ORIG_AMT	Numeric	11,2	
6	ORIGINAL LOAN TERM	ORIG_TRM	Numeric	3,0	301 - 419
7	ORIGINATION DATE	ORIG_DTE	Date	MM/YYYY	MM/YYYY
8	FIRST PAYMENT DATE	FRST_DTE	Numeric	MM/YYYY	MM/YYYY
9	ORIGINAL LOAN-TO-VALUE (LTV)	OLTV	Numeric	14,10	0 - 97% / Blank (unknown)
10	ORIGINAL COMBINED LOAN-TO-VALUE (LTV)	OCLTV	Numeric	14,10	0 - 200% / Blank (if CLTV > 200 or unknown)
11	NUMBER OF BORROWERS	NUM_BO	Numeric	3,0	1 - 10
12	DEBT-TO-INCOME RATIO	DTI	Numeric	14,10	1 - 64% / Blank (if DTI is 0 or ≥ 65 or unknown)
13	BORROWER CREDIT SCORE	CSCORE_B	Alpha-Numeric	3,0	300 - 850 / Blank (if <300 or >850 or unkown)
14	FIRST-TIME HOME BUYER INDICATOR	FTHB_FLG	Alpha-Numeric	1	Y - "First Time Home Buyer"
					N - "Not First Time Home Buyer"
					U - "Unknown"
15	LOAN PURPOSE	PURPOSE	Alpha-Numeric	1	P - "Purchase"
					R - "No Cash-out Refinance"
					C - "Cash-out Refinance"
					U - "Refinance - Not Specified"
16	PROPERTY TYPE	PROP_TYP	Alpha-Numeric	2	SF - "Single Family"
					CO - "Condo"
					CP - "Co-Op
					MH - "Manufactured Housing"
					PU - "Planned Urban Development"
17	NUMBER OF UNITS	NUM_UNIT	Alpha-Numeric	10	1 - 4
18	OCCUPANCY STATUS	OCC_STAT	Alpha-Numeric	1	P - "Principal"
					S - "Second"
					I - "Investor"
					U - "Unknown"
19	PROPERTY STATE	STATE	Alpha-Numeric	20	
20	ZIP (3-DIGIT)	ZIP_3	Alpha-Numeric	10	XXX - First three digits of property's zip code
21	MORTGAGE INSURANCE PERCENTAGE	MI_PCT	Numeric	14,10	1 - 50% / Blank (if not applicable or < 1% or > 50%)
22	PRODUCT TYPE	Product.Type	Alpha-Numeric	20	FRM - "Fixed-rate mortgage loan"
23	CO BORROWER CREDIT SCORE	CSCORE_C	Numeric	3,0	300 - 850 / Blank (if <300 or >850, unkown, or is not applicable)

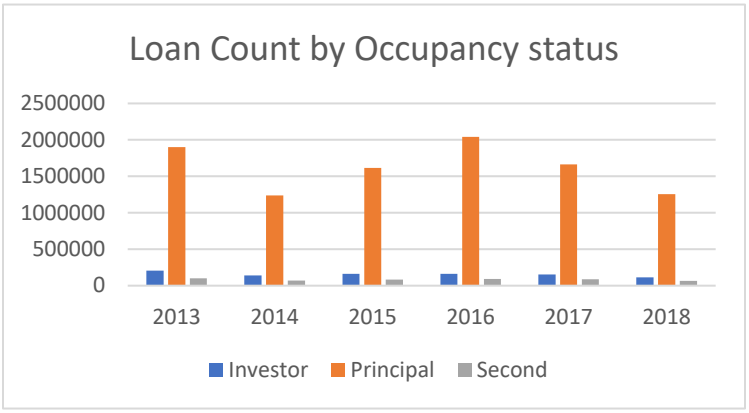
Performance variables for merged data

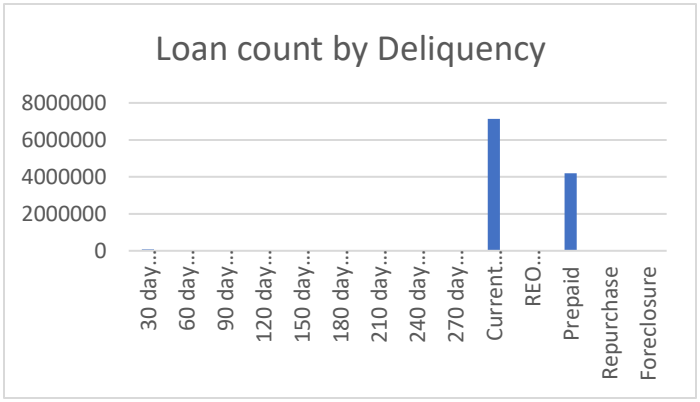
Position	Field Name	Column ID	Type	Max Length	Allowable Values
1	LOAN IDENTIFIER	LOAN_ID	Alpha-Numeric	20	
2	MONTHLY REPORTING PERIOD	Monthly.Rpt.Prd	Date	MM/DD/YYYY	MM/DD/YYYY
3	SERVICER NAME	Servicer.Name	Alpha-Numeric	80	Name of Servicer / Other / Blank (unknown)
4	CURRENT INTEREST RATE	LAST_RT	Numeric	14,10	
5	CURRENT ACTUAL UNPAID PRINCIPAL BALANCE (UP	LAST_UPB	Numeric	11,2	
6	LOAN AGE	Loan.Age	Numeric	10,0	= [Monthly Reporting Period - First Payment Date] + 1
7	REMAINING MONTHS TO LEGAL MATURITY	Months.To.Legal.Mat	Numeric	3,0	=Maturity Date - Monthly Reporting Period
8	ADJUSTED REMAINING MONTHS TO MATURITY	Adj.Month.To.Mat	Numeric	3,0	
9	MATURITY DATE	Maturity.Date	Date	MM/YYYY	MM/YYYY
10	METROPOLITAN STATISTICAL AREA	MSA	Alpha-Numeric	5	XXXXX (five-digit MSA code)
11	CURRENT LOAN DELINQUENCY STATUS	Delq.Status	Alpha-Numeric	5	0 - "Current or less than 30 days past due"
					1 - "30 - 59 days past due"
					2 - "60 - 89 days past due"
					3 - "90 - 119 days past due"
					4 - "120 - 149 days past due"
					5 - "150 - 179 days past due"
					6 - "180 Day Delinquency"
					7 - "210 Day Delinquency"
					8 - "240 Day Delinquency"
					9 - "270 Day Delinquency" / "270+ Day Delinquency"
					X - "Unknown
12	MODIFICATION FLAG	MOD_FLAG	Alpha-Numeric	1	N - "No"
					Y - "Yes"
13	ZERO BALANCE CODE	Zero.Bal.Code	Alpha-Numeric	2	01 - "Prepaid or matured"
					03 - "Short-sale, Third Party Sale, Note Sale"
					06 - "Repurchased"
					09 - "Deed-in-lieu or REO Disposition"
14	ZERO BALANCE EFFECTIVE DATE	LAST_DTE	Date	MM/YYYY	MM/YYYY
15	LAST PAID INSTALLMENT DATE	LPI_DTE	Date	MM/DD/YYYY	MM/DD/YYYY
16	FORECLOSURE DATE	FCC_DTE	Date	MM/DD/YYYY	MM/DD/YYYY
17	DISPOSITION DATE	DISP_DT	Date	MM/DD/YYYY	MM/DD/YYYY
18	FORECLOSURE COSTS	FCC_COST	Numeric	27,12	
19	PROPERTY PRESERVATION AND REPAIR COSTS	PP_COST	Numeric	27,12	
20	ASSET RECOVERY COSTS	AR_COST	Numeric	27,12	
21	MISCELLANEOUS HOLDING EXPENSES AND CREDITS	IE_COST	Numeric	27,12	
22	ASSOCIATED TAXES FOR HOLDING PROPERTY	TAX_COST	Numeric	27,12	
23	NET SALES PROCEEDS	NS_PROCS	Numeric	27,12	
24	CREDIT ENHANCEMENT PROCEEDS	CE_PROCS	Numeric	27,12	
25	REPURCHASE MAKE WHOLE PROCEEDS	RMW_PROCS	Numeric	27,12	
26	OTHER FORECLOSURE PROCEEDS	O_PROCS	Numeric	27,12	

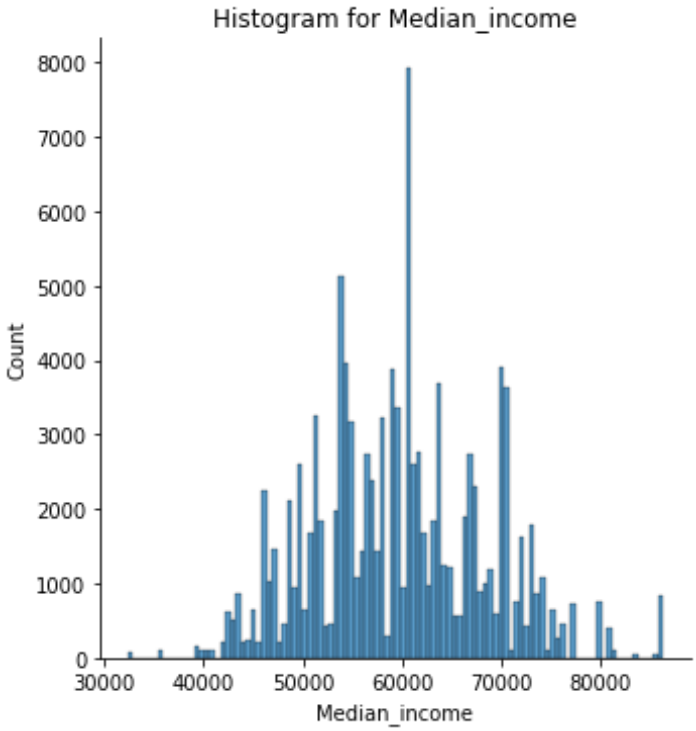
Descriptive Figures

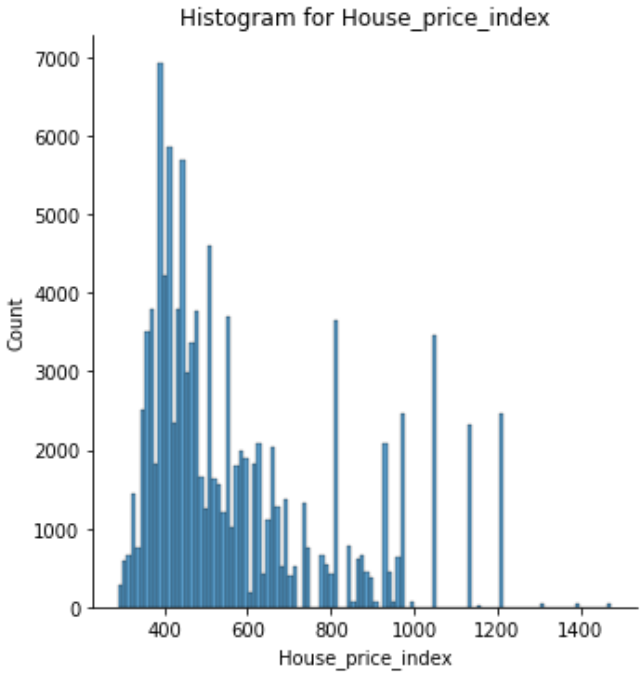


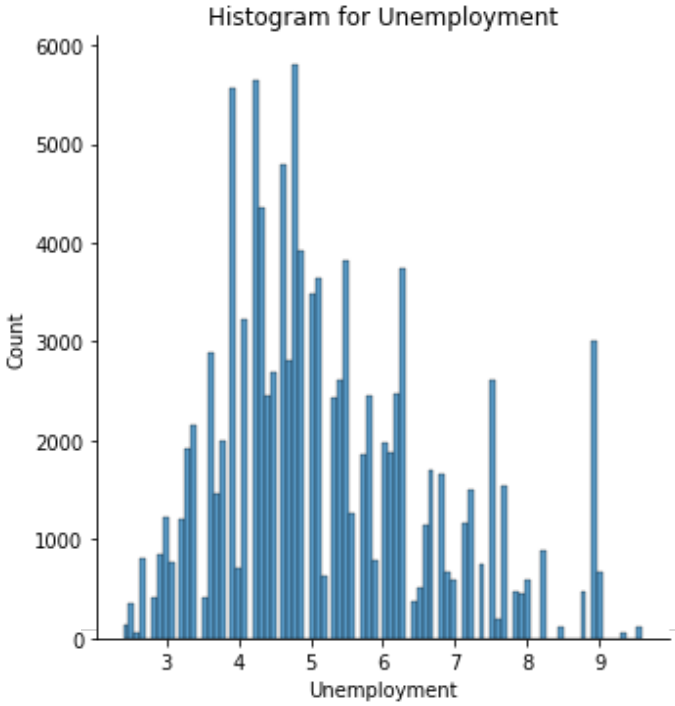


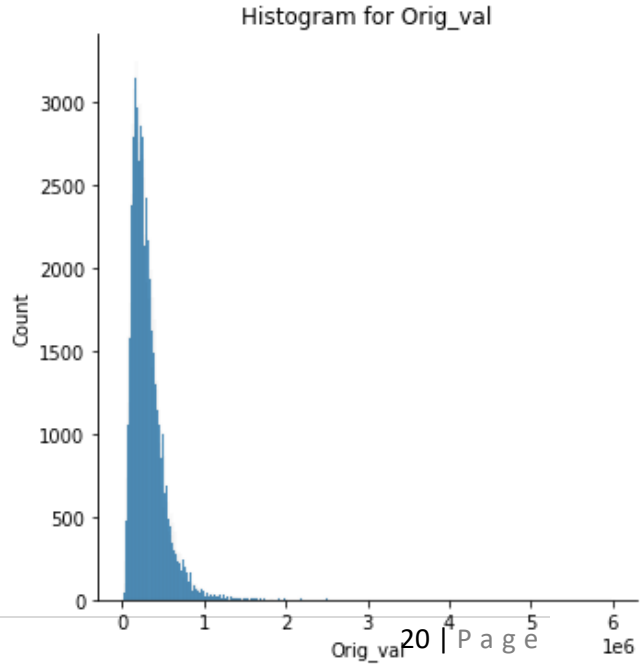

Loan Count by Occupancy status
2500000
2000000
1500000
1000000
500000
0
2013
2014
2015
2016
2017
2018
Investor
Principal
Second


Loan count by Deliquency
8000000
6000000
4000000
2000000
0
30 day...
60 day...
90 day...
120 day...
150 day...
180 day...
210 day...
240 day...
270 day...
Current...
REO...
Prepaid
Repurchase
Foreclosure


Histogram for Median_income
Count
Median_income
30000
40000
50000
60000
70000
80000


Histogram for House_price_index
Count
House_price_index
400
600
800
1000
1200
1400


Histogram for Unemployment
Count
Unemployment


Histogram for Orig_val
Count
Orig_val
1e6