FE 550 Project: Group Data Product

Credit Risk Analysis to Predict Loan Defaults

Group Members: Ho Ben Cheung, Jimit Sanghvi, Venkata Rallapalli Gouri, Xinhang Wang

Background and Problem Statement:

Credit analysis is used to determine the risk associated with repayment of the loan (the risk that an entity will default on the loan). It helps to understand the creditworthiness of a business or a person. The financial crisis hitting the US as well as spreading to other countries in 2008 emphasizing the importance of risk measurement and management. Between 2006 and 2008, hundreds of thousands of people had defaulted, causing a decrease in the value of securities and strongly impacting the global economy. Hence, it is important to do a thorough credit analysis of a person or an organization before dispatching a loan. This project/ product aims to do credit analysis of Fannie Mae Single-Family (Housing) loan with the help of data visualization.

The potential use of the product is in credit risk management by commercial banks/ federal government.

Dataset:

Major Dataset: Fannie Mae Single Family Fixed Rate Mortgage Loan

Acquisition File

Performance File

Features:

FIELD
INTEREST RATE
TERM
DEBT TO INCOME RATIO
CREDIT SCORE
ORIGINAL LOAN TO VALUE RATIO
STATE
PURPOSE
PROPERTY TYPE
OCCUPANCY STATUS

Dependent Variable: 0 = No Default, 1 = Default

Link to Dataset: http://www.fanniemae.com/portal/funding-the-market/data/loan-performance-data.html

Research questions:

1. How single-family mortgage loan performed over time?
    1. Time period: before the financial crisis (2000-2007), during the financial crisis (2008-2011), after the financial crisis (2012-2017)
    2. Features considered: debt-to-income ratio(DTI), credit scores(CSCORE), the loan age, the ratio of loan to value(OLTV), market loan to value(MLTV), loan size, zip, etc.

2. Which features are relatively more predictive/ important on forecasting loan defaults?

From a perspective of borrower credit score, ZIP code, debt-to-income ratio, the number of borrowers etc.

3. How to differentiate strategic default?

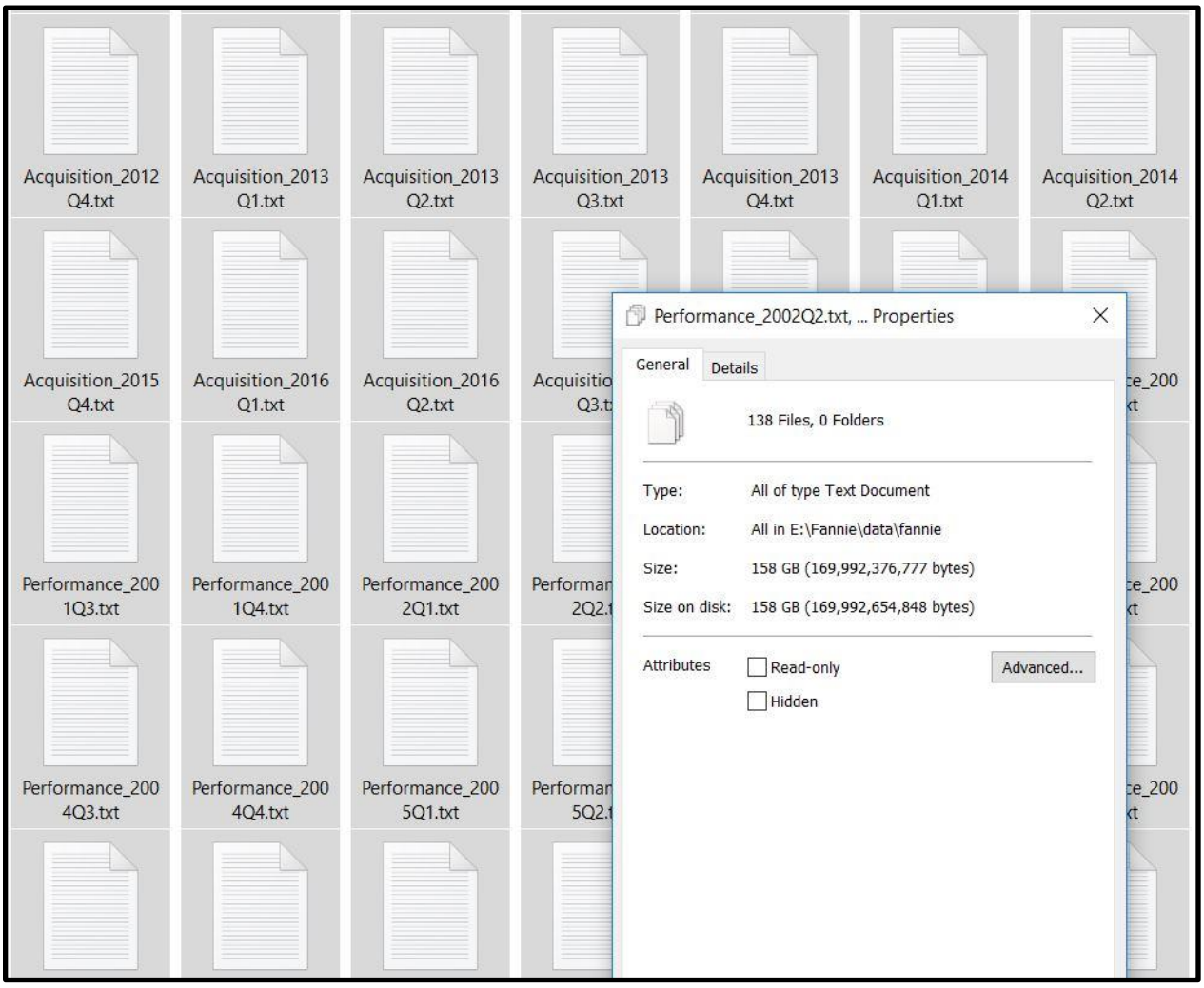
Homeowners have the ability to pay but choose to default because they have high negative equity vs lack of liquidity - homeowners no longer have the ability to pay their mortgage because they have suffered a significant negative income.

4. How is the performance of different machine learning techniques in predicting the loan defaults?

The link for the interactive dashboard is: https://rpubs.com/jimitos10/435410
For Visualization, we have used R library highcharter and plotly, matplotlib in python, and tableau. To create dashboard we have used flexdashboard in R.

Data Engineering and Data Science:

Fannie Mae provides loan performance data on a portion of its single-family mortgage loans. The Single Family Fixed Rate Mortgage dataset, which is whole performance dataset is available for free. The problem is that dataset was in text format by each quarter and is huge.



To tackle the problem, we processed the data using R and loaded the whole dataset in MySQL.

For the Data Science part to clean and process the data and get it ready for analysis and visualization we used R and Python.

Problem:

Imbalanced Data:

The loan dataset is imbalanced i.e. 95% of the loans are categorized as not default and only 5% are default. To solve this problem we used SMOTE (Synthetic Minority Oversampling Technique) in R to solve the problem. Thus, we were able to balance the dataset as 60% not default and 40% default.

Data visualization

1.How single-family mortgage loan performed over time?

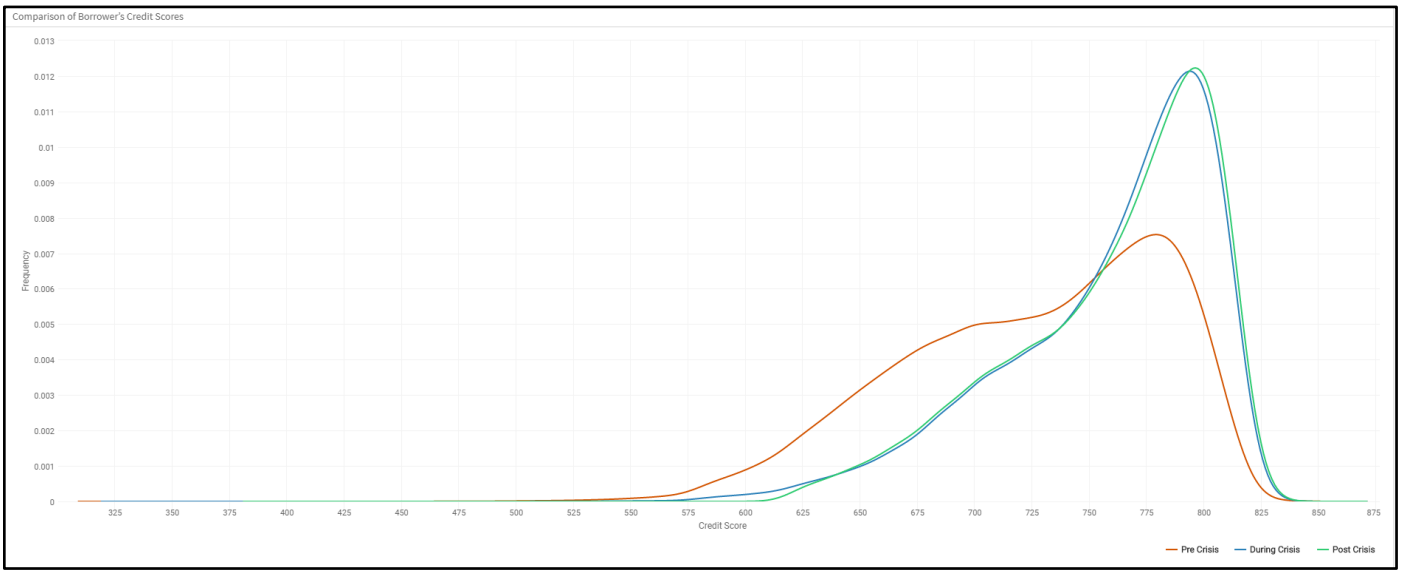


Fig 1: Comparison of Borrower's Credit Score for Pre, During and Post Financial Crisis

As we can see in the above plot, there seem a significant change in the distribution pre-financial crisis from during and post-financial crisis. This plot clearly explains the story behind the financial crisis. And we can see due to the outbreak of the financial crisis more focus was shifted to borrowers with high credit score.

The frequency of loans given to lower credit score (Subprime Mortgages) is greater before the crisis than compared to the frequency during and after the crisis. Subprime Mortgages was one of the main reason for the 2008 Financial Crisis.

For Visualization, we have used R library highcharter to create the interactive visualization (link) and we have used three different colors to differentiate the 3 time periods i.e. pre-crisis, during crisis and post-crisis.

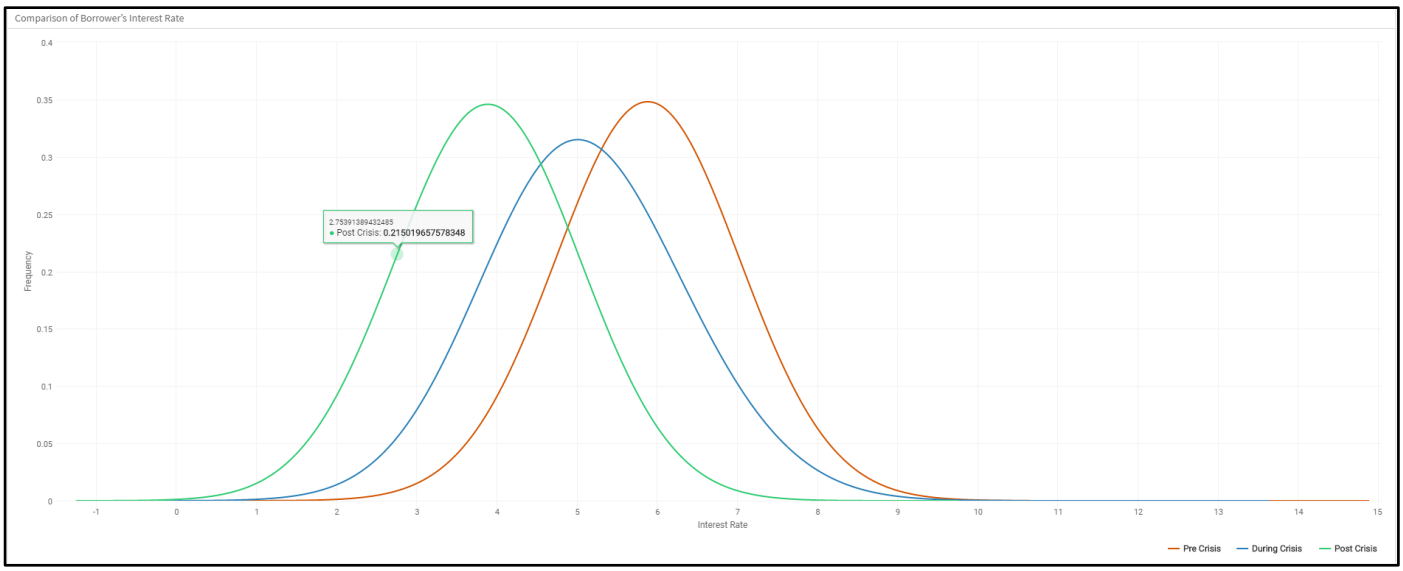


Fig 2: Comparison of Interest Rates for Pre, During and Post Financial Crisis

As we can see in the above plot, there seem a significant change in the distribution of interest rates pre-financial crisis, during and post-financial crisis. As we can see due to the outbreak of the financial crisis the interest rates reduced as fed reduced the interest rate to stimulate the economy.

For Visualization, we have used R library highcharter to create the interactive visualization (link) and we have used three different colors to differentiate the 3 time periods i.e. pre-crisis, during crisis and post-crisis.

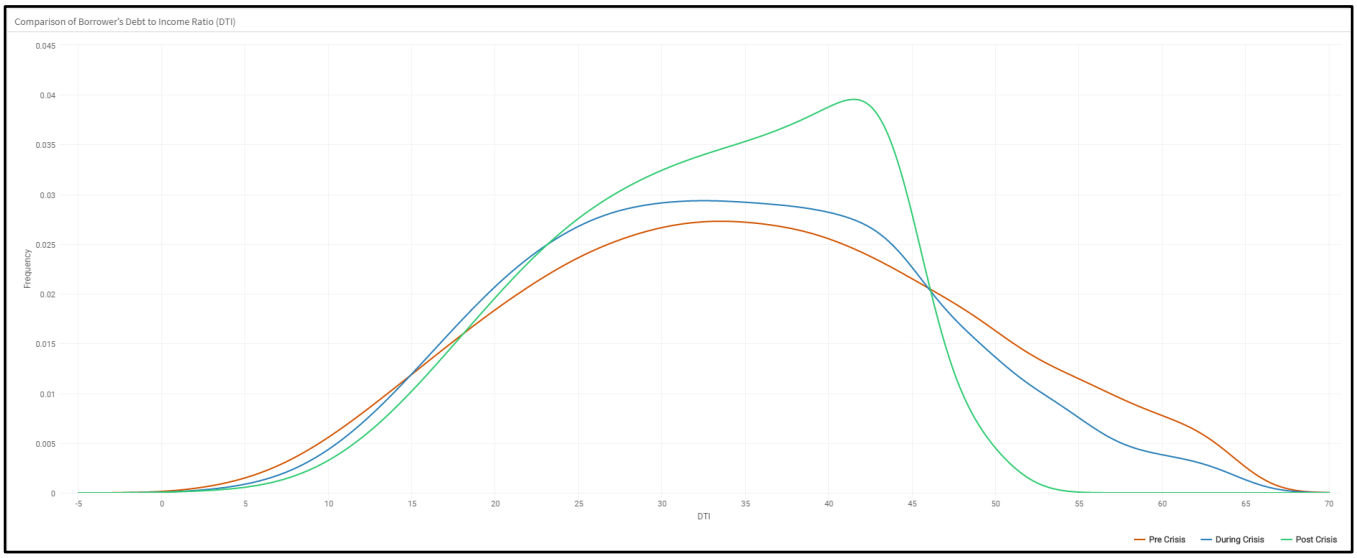


Fig 3: Comparison of Debt to Income (DTI) for Pre, During and Post Financial Crisis

As we can see in the above plot, there seem a significant change in the distribution of interest rates pre-financial crisis, during and post-financial crisis. As we can see after the financial crisis the borrowers with very high DTI were not considered for housing loans. We can see a threshold at around 50 DTI after the crisis.
For Visualization, we have used R library highcharter to create the interactive visualization (link) and we have used three different colors to differentiate the 3 time periods i.e. pre-crisis, during crisis and post-crisis.

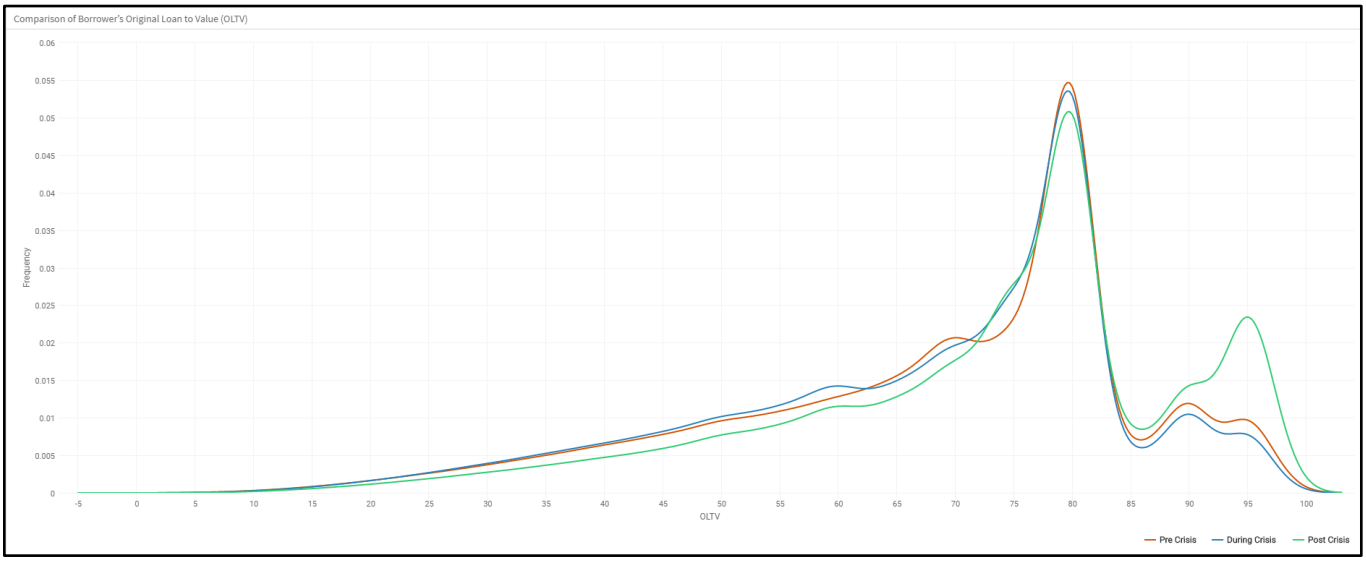


Fig 4: Comparison of Loan to Value for Pre, During and Post Financial Crisis

As we can see in the above plot, there seem a significant change in the distribution of interest rates pre-financial crisis, during and post-financial crisis.
For Visualization, we have used R library highcharter to create the interactive visualization (link) and we have used three different colors to differentiate the 3 time periods i.e. pre-crisis, during crisis and post-crisis.

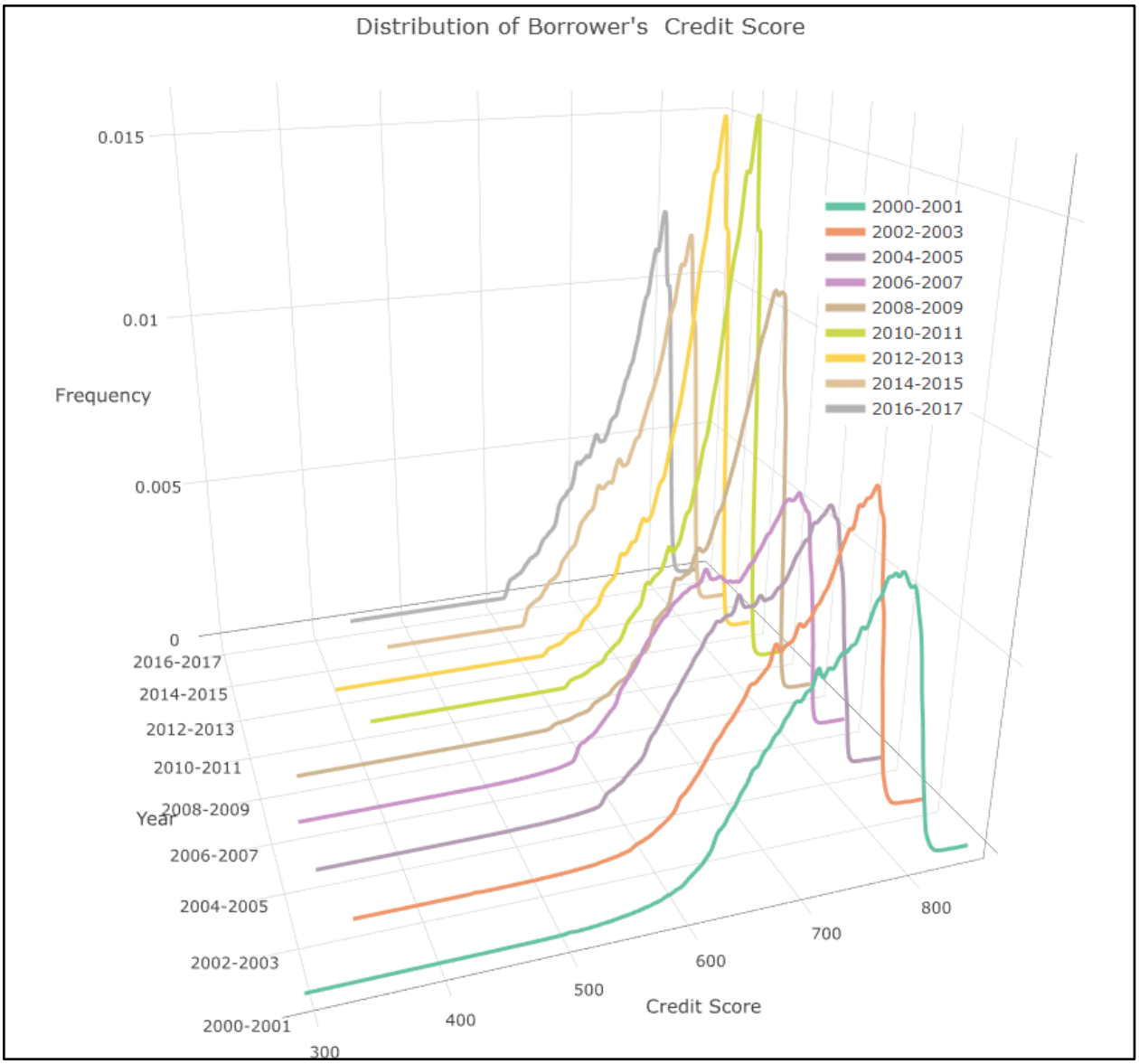


Fig 5: 3D Visualization for Credit Score to Visualize its Distribution by Years

The distribution plot shows the transition of Borrower's Credit Score from 2000 to 2017. It provides a better understanding of the 2008 Financial Crisis. The frequency of loans given to lower credit score (Subprime Mortgages) is greater before the crisis than compared to the frequency during and after the crisis.

Subprime Mortgages was one of the main reason for the 2008 Financial Crisis.

We created this plot using plotly in R. Each distribution account for 2 years. We have used different colors to differentiate the distributions by years.

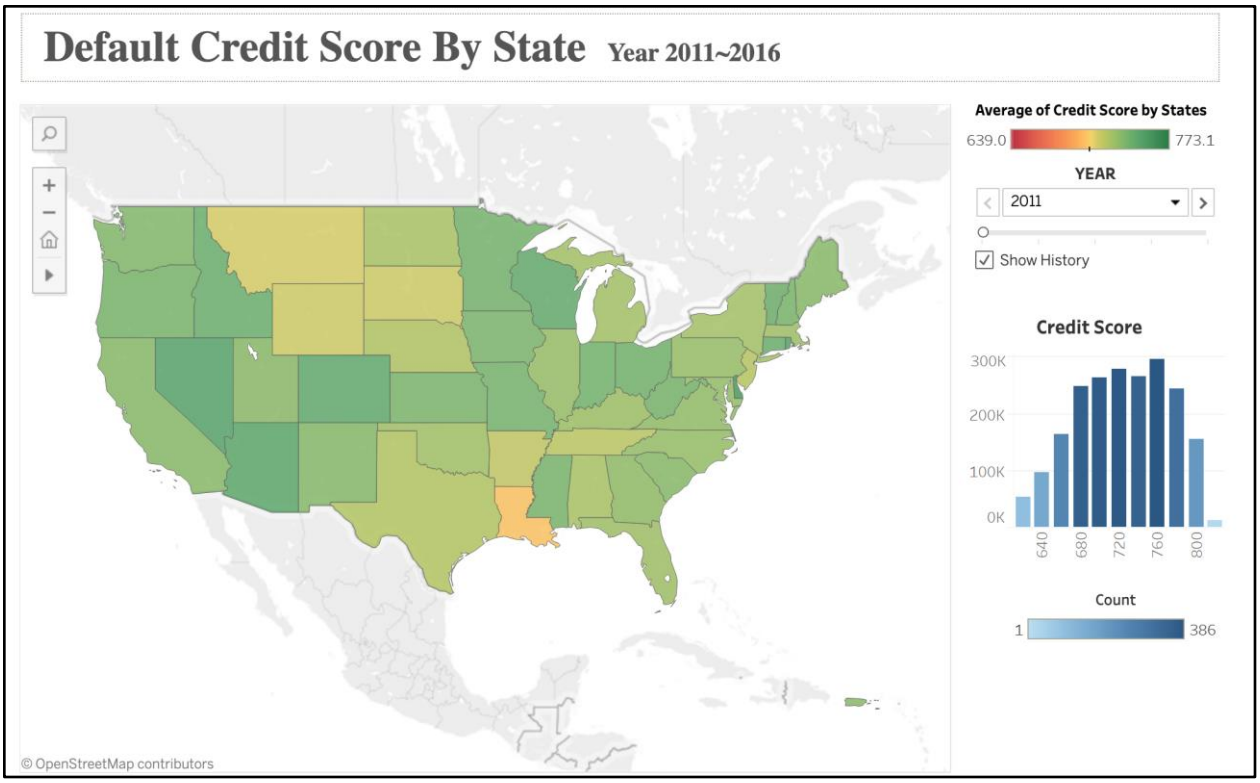
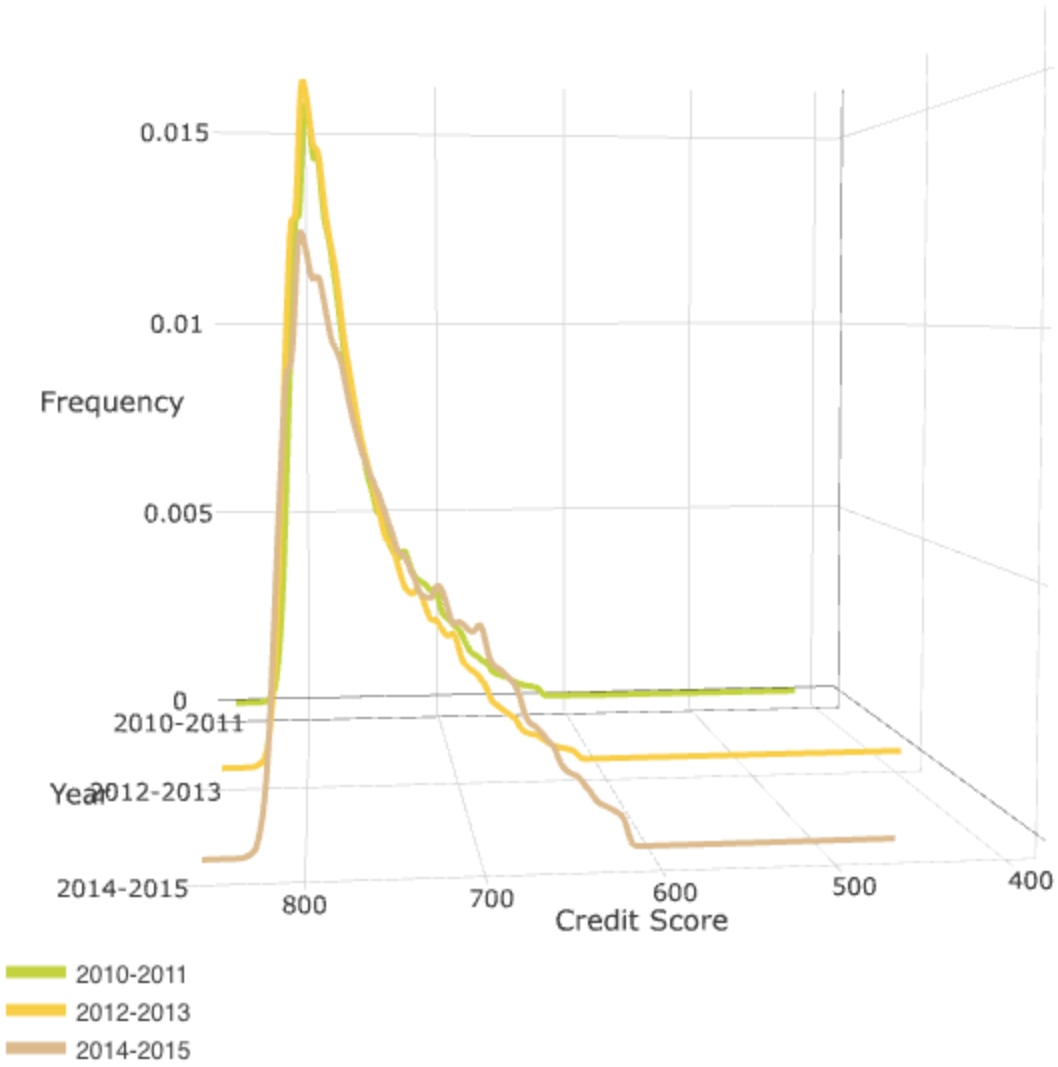


Fig 6: Interactive Map Visualization in Tableau

It is a interactive Tableau dashboard that shows the credit score of default mortgage after the financial crisis 2008 across the states (Additional information can be found at Appendix A). The color of the state is representing average credit score per state with a intuitive tone of red and green representing the two extreme of bad and good credit score respectively. On the other hand, histogram represents the density of credit score of the USA. From 2011 to 2016, the peak of the histogram graph is decreased from 770 to 660, which means the mode of credit score of default mortgage is decreasing dramatically.



However when we consider whole sample set, the peak of density plot of credit score is more stationary in the range of (790,800).

One of the interpretations is the period that just after the 2008 financial crisis, the housing market is still influence by other marco-economic feature that increase mortgage default probability of people that with both high and low credit score which we have not considered in this model. In the other words, credit score may not be a good predictor or mortgage default risk during a financial crisis.

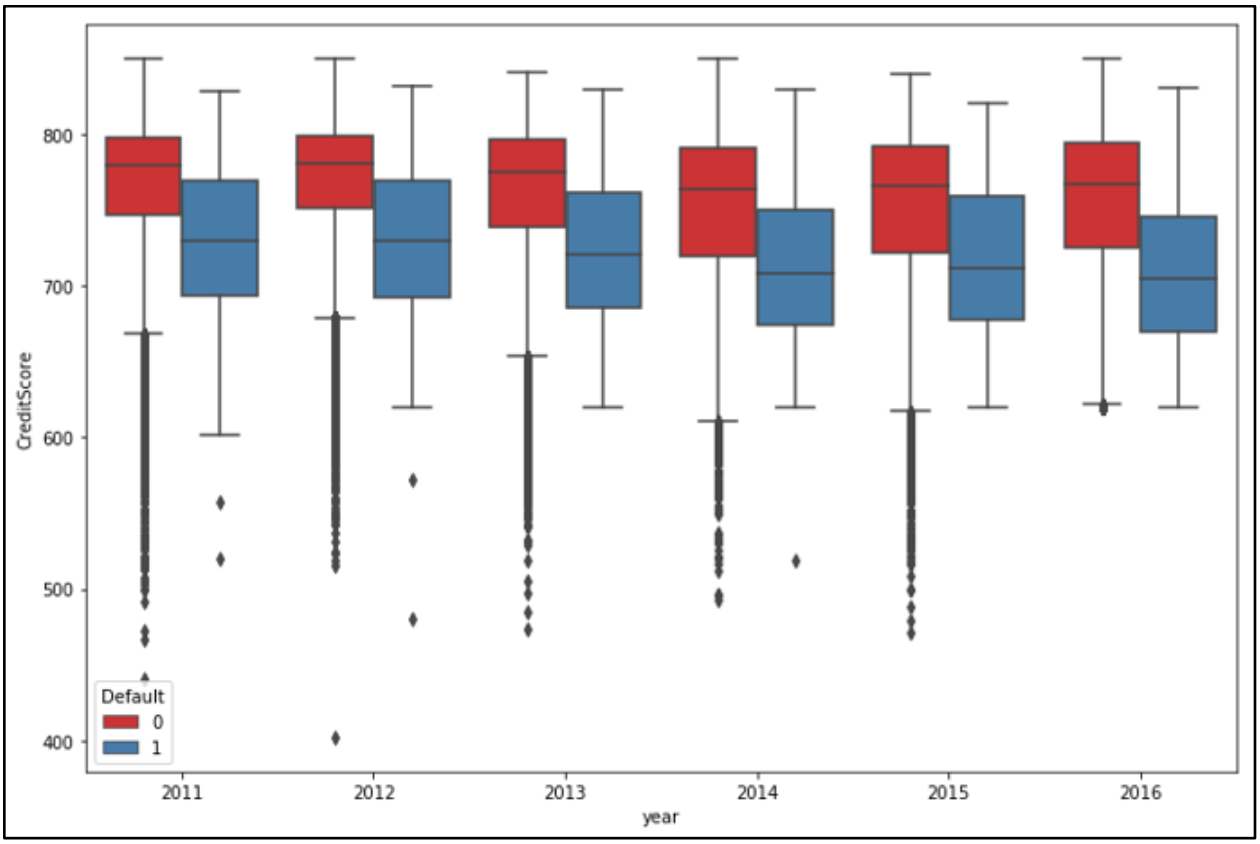
2. Which features are relatively more predictive/ important on forecasting loan defaults?

From a perspective of borrower credit score, ZIP code, debt-to-income ratio, the number of borrowers etc.

1. Analyze relationships between the status of payment and various factors (CLTV(current loan to value ratio), debt to income ratio, credit score, interest rate):

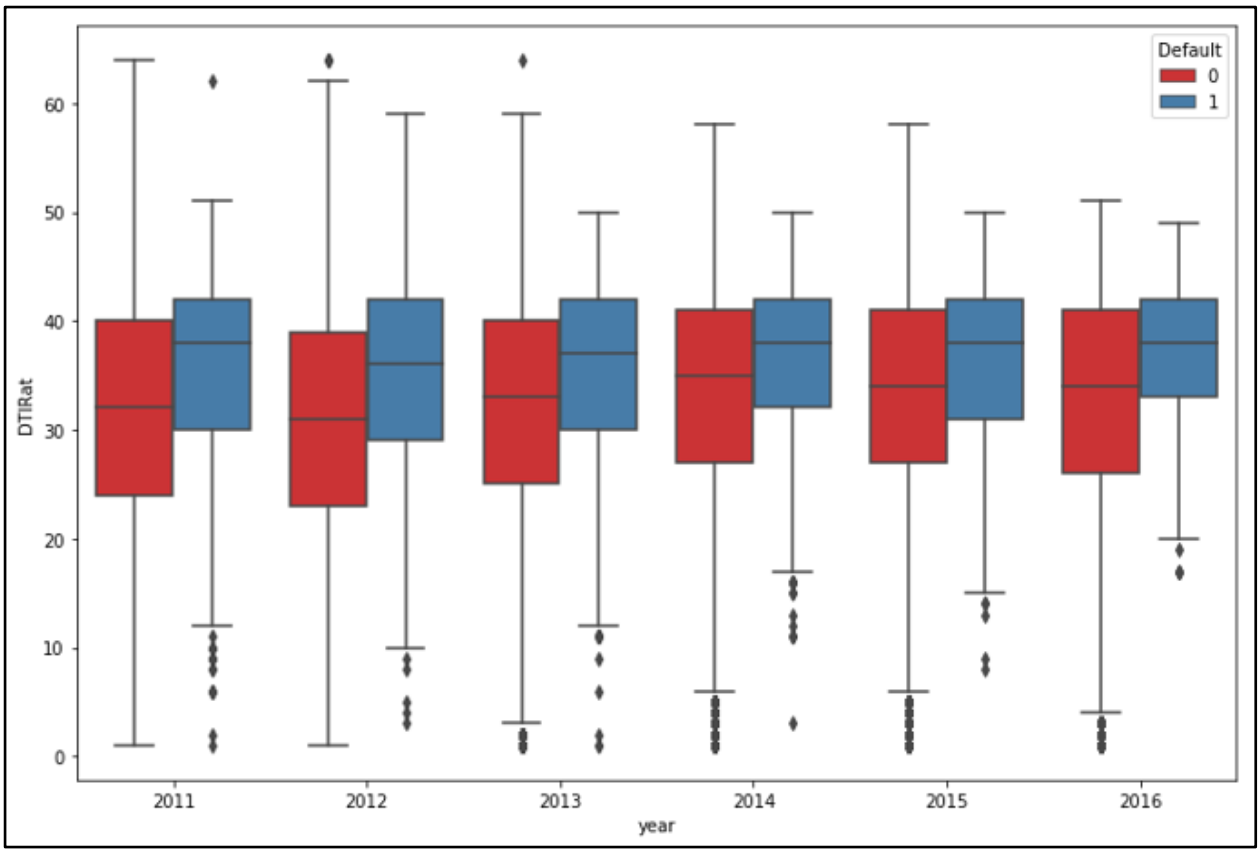
Box plot:

Credit score:



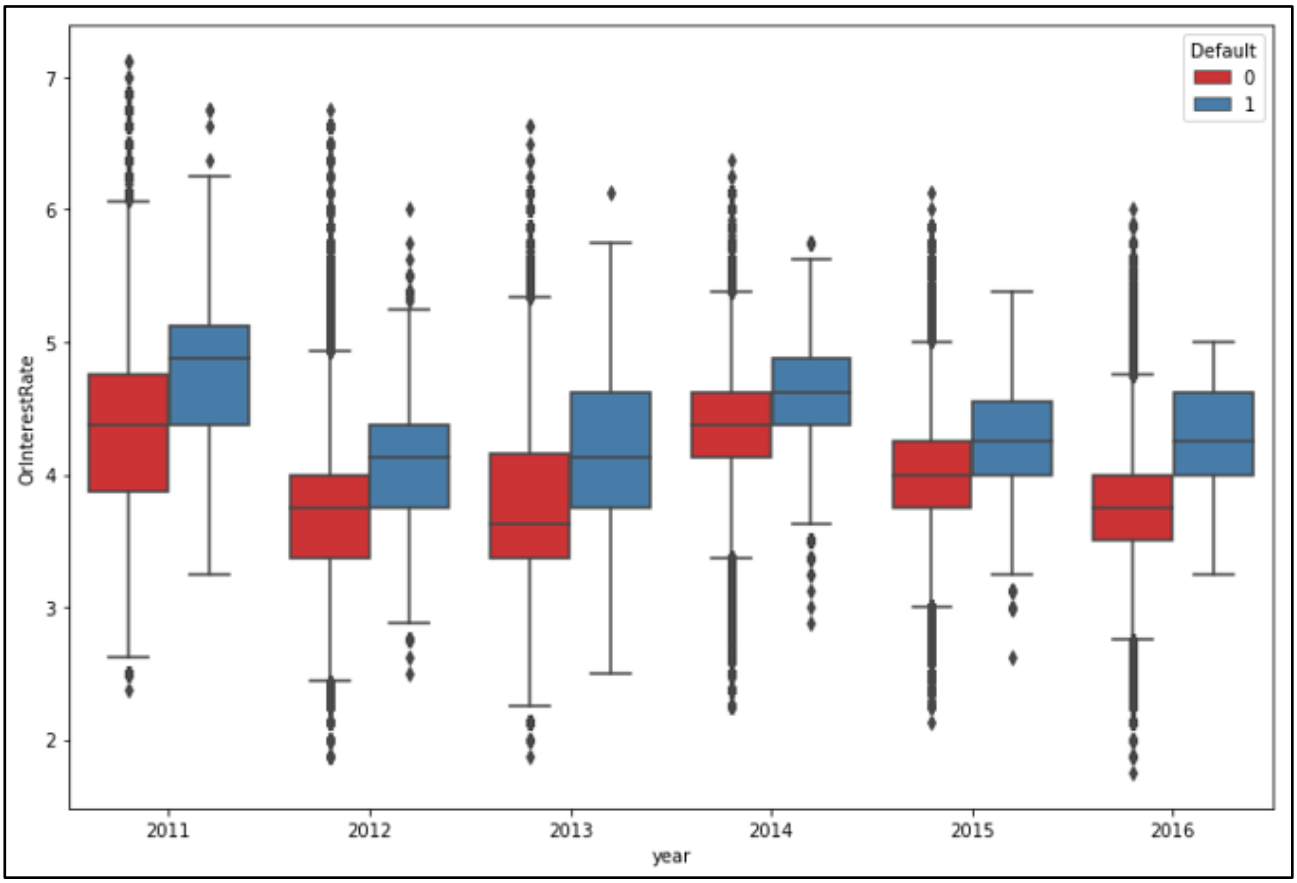
From this figure, it is clear to see that a loan default is more likely to happen with a lower credit score. This feature represent the creditworthiness of an individual or an organization and is applied to evaluate the potential risk by lending money to consumers.

Debt to income ratio:



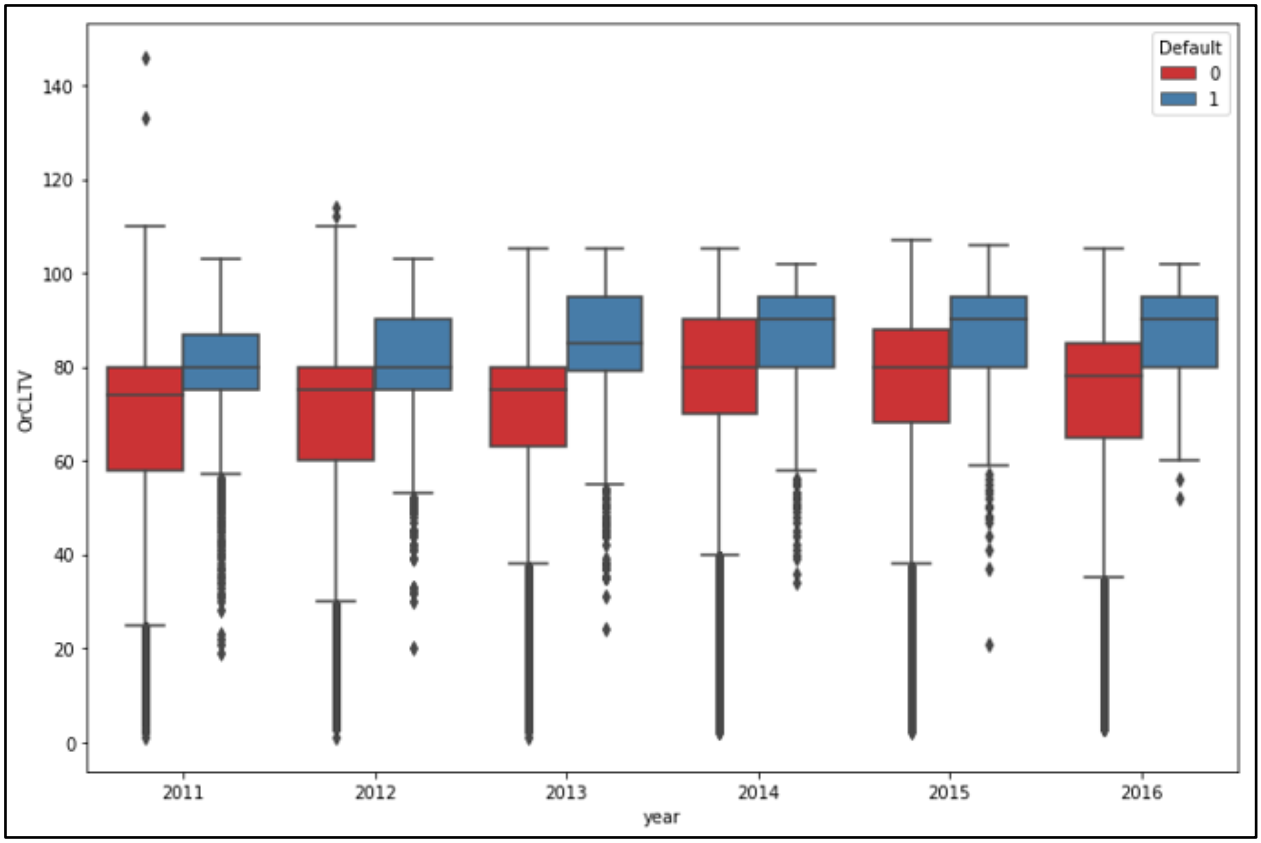
This figure displays that DTI can be an indicator of the borrower's income, and loans with higher DTI are more likely to default.

Interest rate:

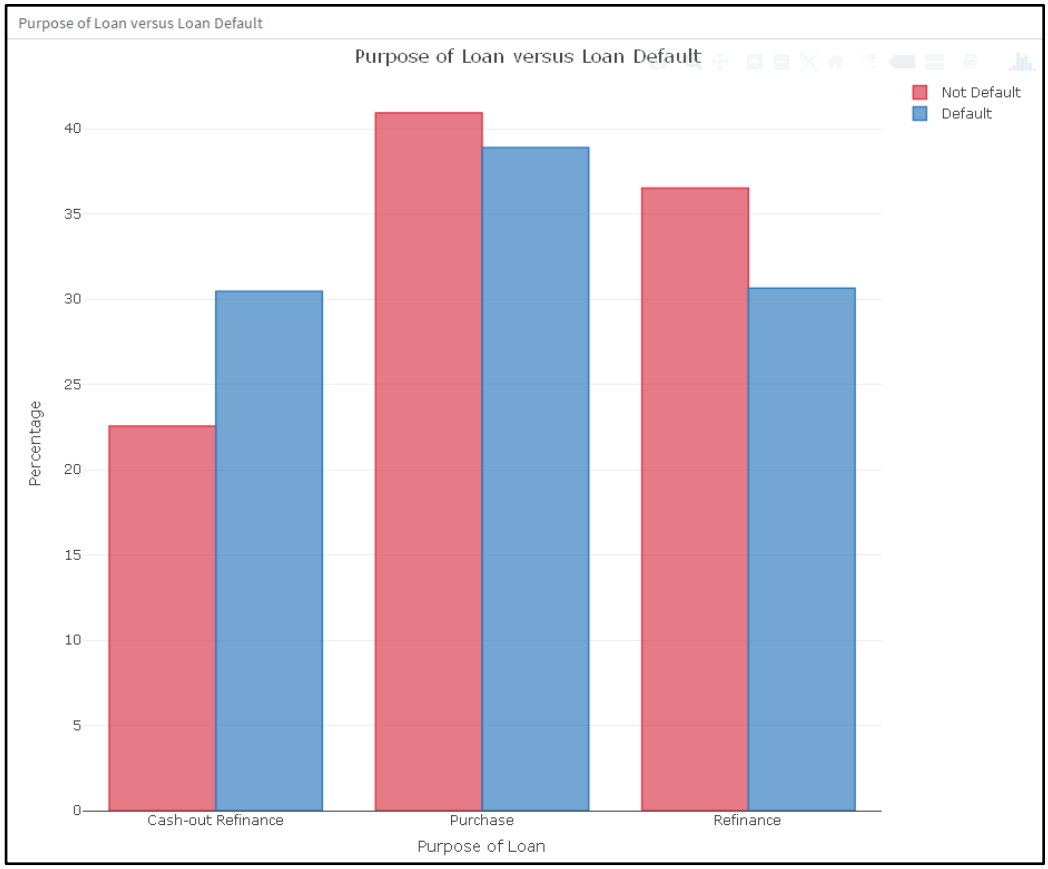


The interest rate for defaulters is higher than that for non-defaulters. Thus it can be one of the important factors to predict loan defaults.

Loan to value:

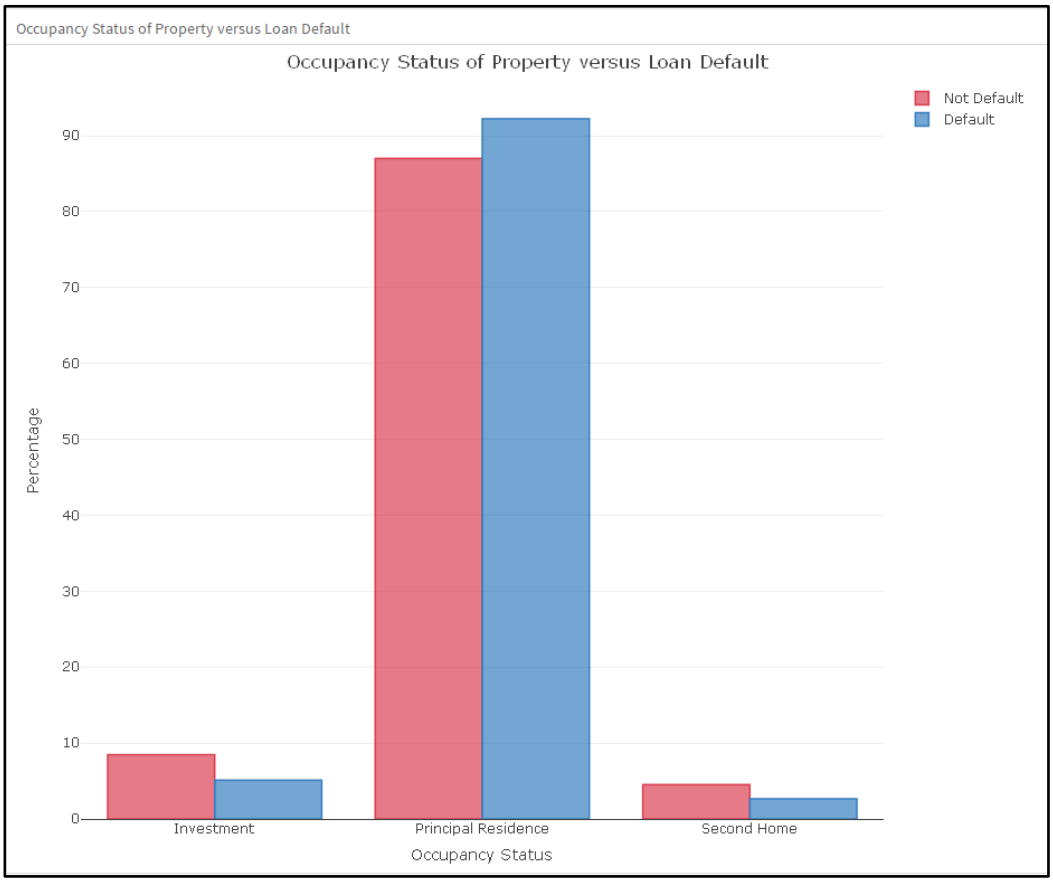


This figure displays that loans with higher LTVs are more likely to default. LTV can change over time when the local housing price goes up or down and a significant decline in housing prices lead to borrower negative equity where the value of house is less than borrower mortgage debt.



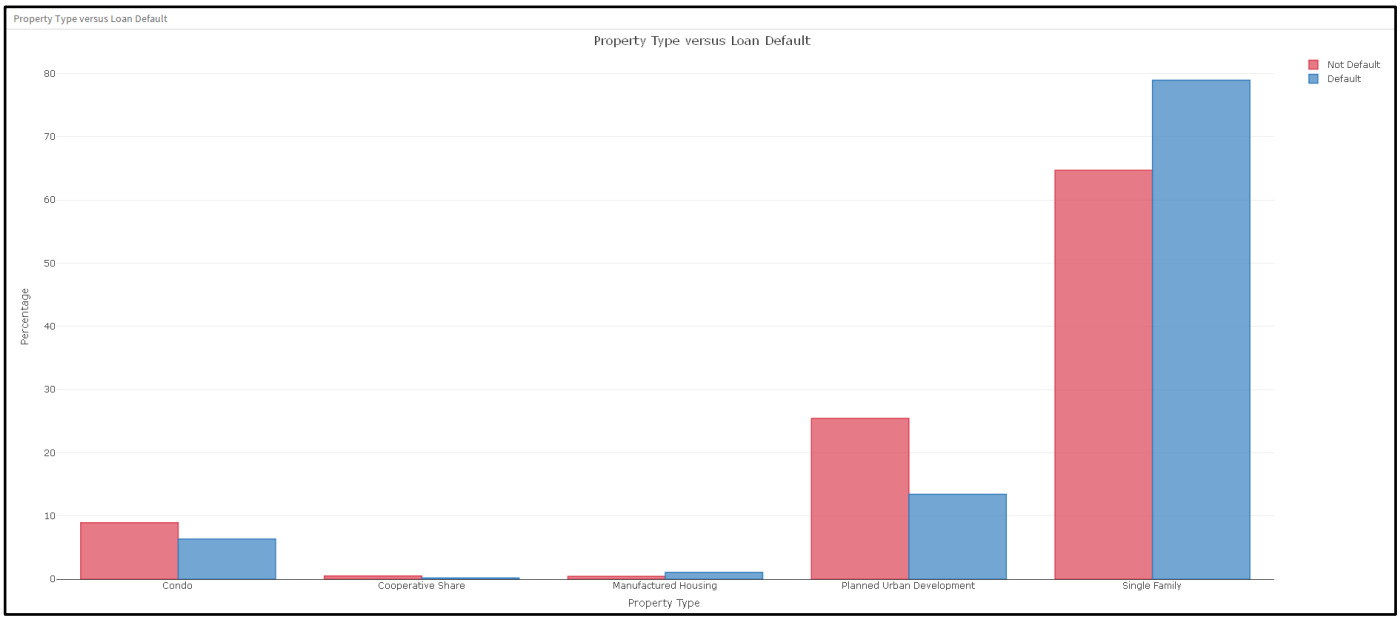
Plot Comparing Purpose of Loan and Loan Defaults

Apart from Cash-out refinance the other loan purposes have higher non-defaults compared to defaults. It can be one of the reason for Strategic Default.



Plot Comparing Occupancy Status of Property and Loan Defaults

We can clearly see Investment and second home have lesser defaults compared to not defaults. The principal residence i.e. first home buyers have higher defaults. Also, most of the loans are principal residence.



Plot Comparing Property Type and Loan Defaults

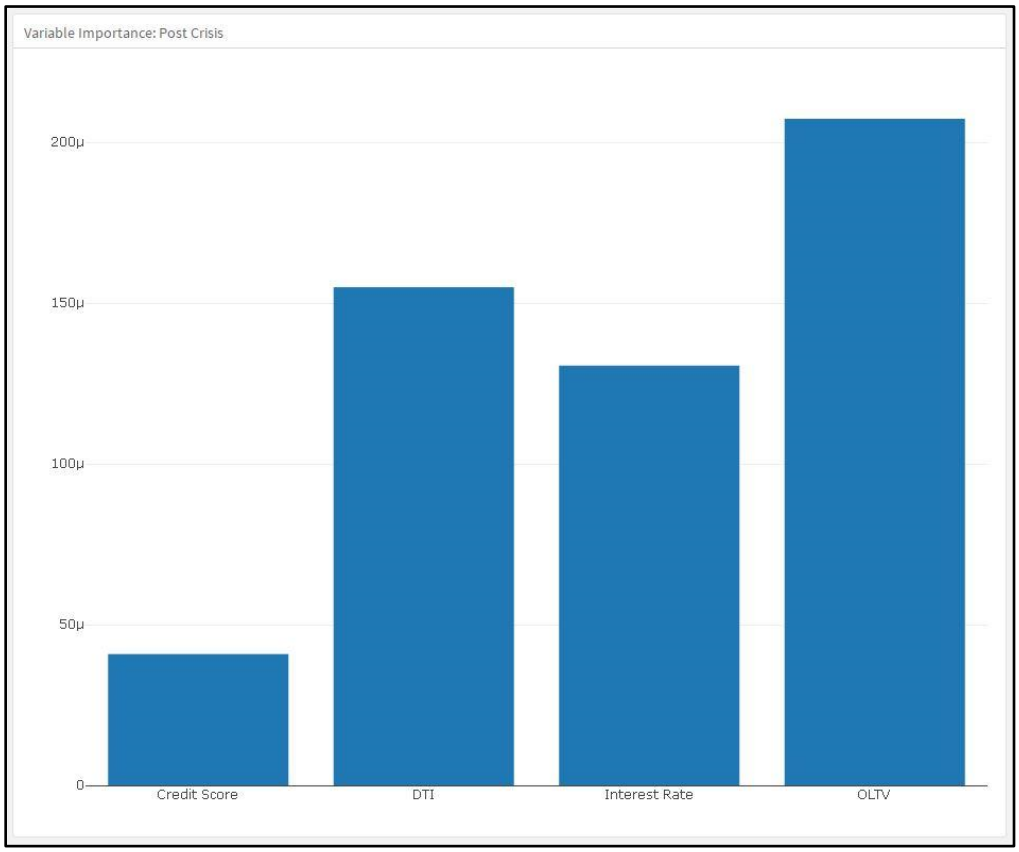
The three major property types are single home (most) followed by planned urban development and condo. Apart from single home other have less defaults compared to non defaults.

2. Analyze the importance of factors:

In this case, the tree-based classifier Random Forest is applied to solve this question. Random Forest is a machine learning algorithm which is useful for the classification on whether a variable is more or less important for building the decision tree.

In Q2, four variables are considered, which have mentioned above: LTV,

DTIRate, interest rate, credit scores.

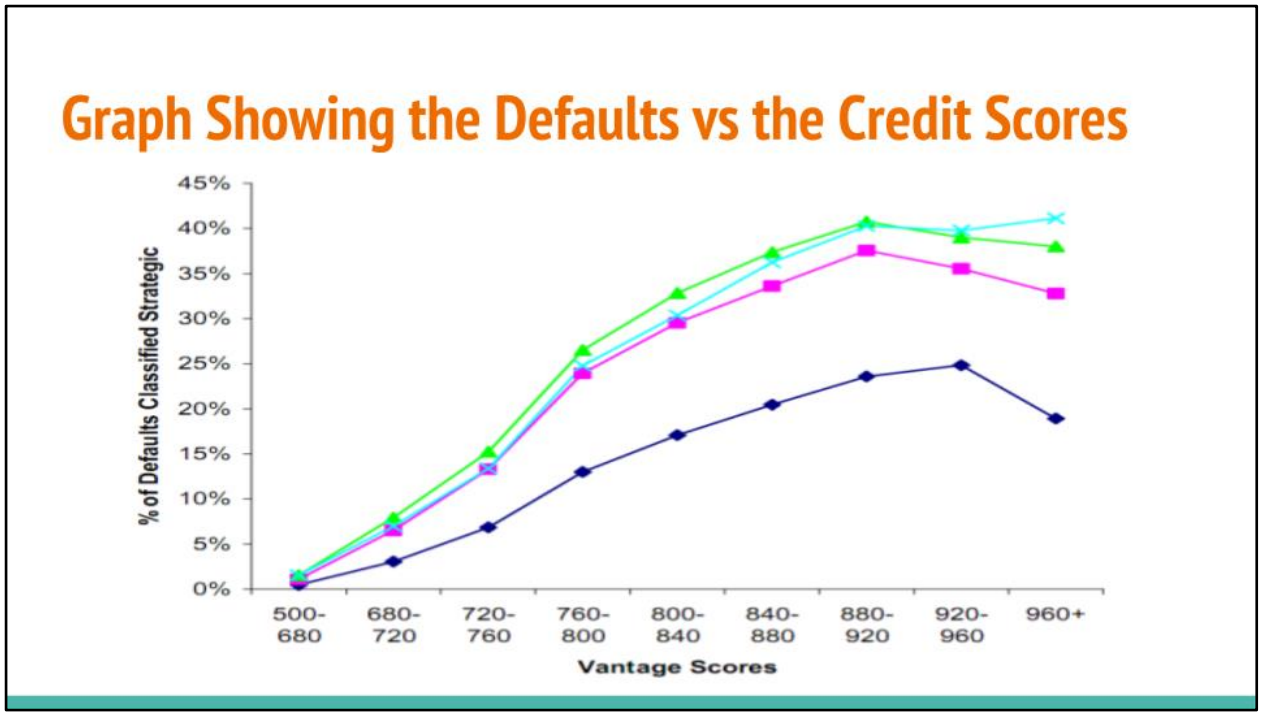


While all features show contribution to the model positively, original loan to value ratio the most significant connections to the mortgage default risk according to the model.

3. How to differentiate strategic default?

Strategic defaults are often employed by borrowers when the value of their property has dropped substantially within a fairly short time.If the value of the property dips below the mortgage balance then a strategic default provides a way to minimize the property owner's loss. Based on the data set python script would give us the loan defaulters with a threshold of credit score, so banks can negotiate with the borrowers and negotiate to lower the payment rather than allowing them to default it as this leads to a win - win situation when markets are down. As shown in below graph, we can target the borrowers with high credit score for negotiation as the strategic defaults occuring more with high credit borrowers.

As shown below we used credit scores and default % strategic defaults and it shows that majority of the borrowers with credit score higher than 750 are very likely to default in the case of down market.



4. How is the performance of different machine learning techniques in predicting the loan defaults?

Machine learning algorithm applied in this project:

1) Random forest

It is an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification).
Consider, $X = x_1, ..., x_n$ - training set, $Y = y_1, ..., y_n$ - response, the number of samples/trees: B;

Bagging repeatedly (*B* times) selects a random sample with replacement of the training set and fits trees to these samples:

For $b = 1, ..., B$:

1.Sample, with replacement, *n* training examples from *X*, *Y*; call these X_b, Y_b.

2.Train a classification tree f_b on X_b, Y_b.

After training, predictions for unseen samples from testing set can be made by taking the majority vote in the case of classification trees.

Advantages:

1. It gives estimates of what variables are important in the classification.
2. It generates an internal unbiased estimate of the generalization error as the forest building progresses.
3. It has methods for balancing error in class population unbalanced data sets.

Disadvantages:

Random forests have been observed to overfit for some datasets with noisy classification/regression tasks.

2) SVM

The SVM finds a hyperplane that separates training observations to maximize the margin. The distance between observations and the decision boundary explains how sure about prediction.
Considering a training set where m is the number of observations. Define (w, b) as the smallest margin on training observations S where w contains parameters of the hyperplane and b is the hyperplane intercept:

M = min M(i) , i = 1, ..., m

Assume the positive and negative classes can be separated by a linear hyperplane then one can write the SVM optimization problem as :

$$Max_{w,b} M$$

$$\text{St} \quad y_i(\omega^T x_i + b) \geq M, i = 1,\ldots,m$$

$$\|\omega\| = 1$$

M - the number of observations

ω - contains parameters of the hyperplane

b - the hyperplane intercept

y_i - the class of the training observation i

x_i - the feature spaces i in the training dataset

3) Logistic regression

Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome and the dependent variable is binary, i.e. it only contains data coded as 1 (TRUE, success, pregnant, etc.) or 0 (FALSE, failure, non-pregnant, etc.).

The goal is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest and a set of independent (predictor or explanatory) variables. It generates the coefficients of a formula to predict a logit transformation of the probability of presence of the characteristic of interest,

$$logit(p) = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + \ldots + b_k X_k$$

where p is the probability of presence of the characteristic of interest. The logit transformation is defined as the logged odds,

$$odds = \frac{p}{1-p} = \frac{probability\ of\ presence\ of\ characteristic}{probability\ of\ absence\ of\ characteristic}$$

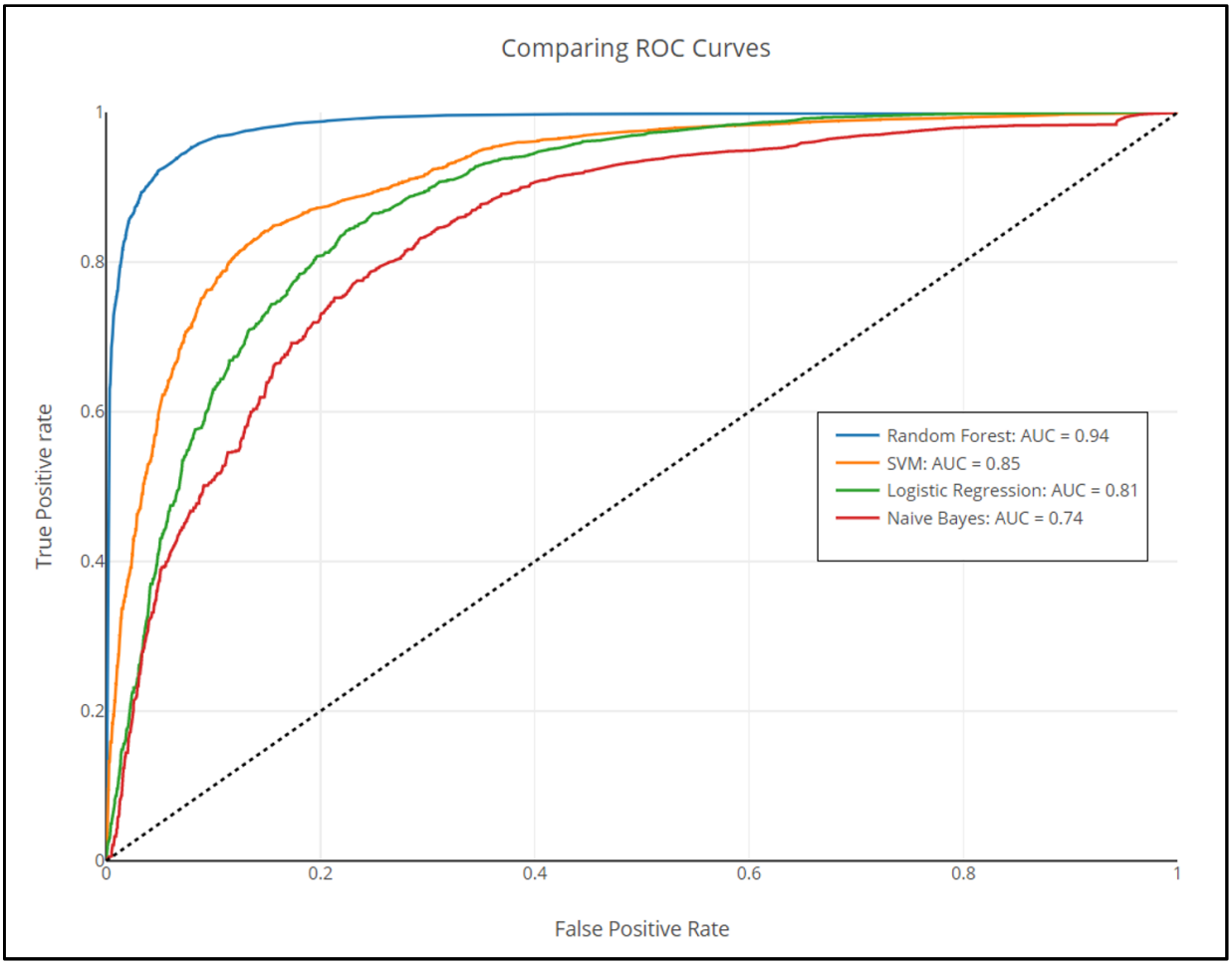
and

$$logit(p) = \ln\left(\frac{p}{1-p}\right)$$

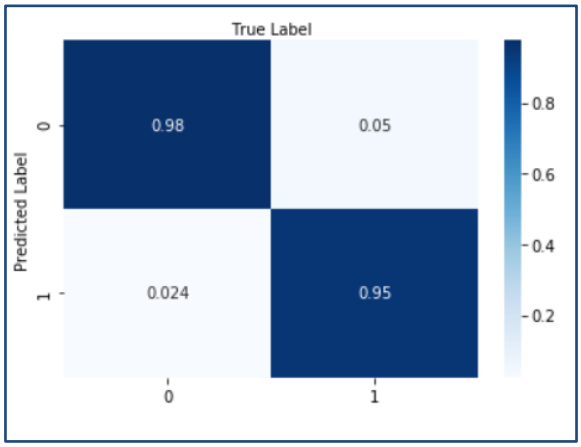
4) Naive Bayes

It is a classification technique based on Naive Bayes Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

Visualization:



ROC curve plot comparing different Machine Learning Algorithm's Performance



Confusion Matrix for Random Forest Algorithm

Result Interpretation:

The project compared result of different machine learning models and using area under the ROC as a benchmark. ROC curve, which stands for "*receiver operating characteristic curve*", is a graph that show the performance of a classification model which consists the following two parts

- True Positive Rate
- False Positive Rate

True Positive Rate (TPR) :

$$TPR = \frac{TP}{TP + FN}$$

False Positive Rate (FPR):

$$FPR = \frac{FP}{FP + TN}$$

An ROC curve plot TPR vs FPR at different classification threshold

AUC (Area under the graph) is an aggregate measure of performance of the classification model at different threshold. AUC can be interpreted as probability that the model has a higher ranks for a random positive sample over random negative sample.

Based on our calculation, Random Forest (RF) scored the best result (AUC= 0.94), and Naïve Bayes had the worst result (AUC=0.74). Logistic Regression is the common model that Office of Federal Housing Enterprise Oversight (OFHEO) and other studies to model the mortgage loan default and prepayments, which the model accuracy can be improved if supervised machine learning technique has been implemented.

Product review:

Our product is started from first identify an existing problem of the current stress test model that based on logistic regression and without constantly updated data which lead to produce inaccurate mortgage default risk for risk management purpose. The focus will be on improving the accuracy of predicting credit related default with the help with latest technology/ algorithm. From the model comparison result, random forest algorithm is outperform logistic regression which is currently used at the federal housing authorities to build their stress test model. Beside mortgage default risk, one of the potential applications is credit default swap that carries a similar problem setting as mortgage default. Moreover, our product can be make adjustment to apply other risk categories that is a classification problem such as email spam detection.

Future Scope:

Our future approach will be focusing on improving the model with taking additional features into consideration and experience with other data science technique. Since loan to value is the most important feature that contribute to the mortgage default risk, the fluencation of the value of the property is posing additional risk to the system when the volatility of the housing price starting to increase. Therefore we will consider to add macro-economic data, such as time-series of unemployment rate/ interest rate to quantify the portion of the default risk that can not be predict from the credit history.

Another problem of the stress test is the inaccurate estimation to the worse case scenario of the housing market. Traditionally, it uses method such as VARs (Value at risk) , e.tc in federal reserve board, however it has been proof during the crisis their calculated VARs had been exceeded more than six times during the crisis in 2008, which . A recent study, “Causal data science for financial stress testing” by Gelin Gao, e.tc, is aiming to solve the issue of the financial stress test scenario with the help of Suppes-Bayes Causal Networks and machine learning classification technique.

Appendix

Online Tableau dashboard login information:
https://us-east-1.online.tableau.com/en/embeddedAuth.html?path=%2Ft%2Ffanniemaemortgagedefault%2Fviews%2FDefaultCreditScoreByStateDashboard%2FDashboard2%3F%3Aembed%3Dy%26%3Adisplay_count%3Dno%26%3AshowVizHome%3Dno&siteUrlName=fanniemaemortgagedefault&siteLuid=97549c59-7e32-4887-bdd8-fcdc012f2d4f&authSettings=DEFAULT

User: wesley7d05@gmail.com
Password: mzguheK4wy4G

The link for the interactive dashboard is: https://rpubs.com/jimitos10/435410
We have attached a small portion of our dataset in CSV format as our actual dataset size is huge.

Link to Dataset: http://www.fanniemae.com/portal/funding-the-market/data/loan-performance-data.html