SINGLE FAMILY LOAN PERFORMANCE MODEL

MOMODE

SHASHANK SHRIVASTAVA

Nov, 2016

PROBLEM STATEMENT

Financial institutions conduct “stress tests” to gauge the resilience of their balance sheets to substantial macroeconomic shocks such as rising unemployment rate or drop in gross domestic product GDP. One way to measure the performance of a financial institution is by assessing the institution’s loan portfolio loss under stressed scenarios. The loan portfolio is total of all the loans that a financial institution, or other lender, holds at a given time. Loan portfolios are the major asset of banks, thrifts, and other lending institutions. The value of a loan portfolio depends not only on the interest rates earned on the loans, but also on the quality or likelihood that interest and principal will be paid.

The first step in assessing loan loss is to estimate the Probability of default (PD)[1]. Understanding PD is necessary for the purpose of stress testing and risk management. Financial institutions may also find it beneficial as insights from default modeling can be incorporated to guide improvements on good underwriting practice and competitive mortgage pricing

The proposed model attempts to predict risk of a loan defaulting by estimating relationship between loan performance and other variables associated with loan like credit score, interest rates etc. The model will help financial institution to estimate the risk on their loan portfolio.

TARGET AUDIENCE

Though we are using a dataset provided by Freddie Mae, which is one of the biggest mortgage buyers from various mortgage banks the use of this analysis and model can also be extended to banks that hold large mortgage portfolios with the assumption that the relationships we find in this analysis are applicable to them.

The clients have incentive to identify the loans that are at risk of default and provide help to high-risk loans to minimize loss.

[1] Wikipedia: Probability of default

DATA DESCRIPTION USED FOR MODELING

This data is made available from Freddie Mae[2].

For each year from 1999 to 2015 the loan data contains a zipped file for each quarter.

For example: ***historical_data1_Q11999.zip***

Each zipped file has two datasets in a tab delimited txt file:

1. Single Family Loan-Level Dataset
2. Monthly Performance Dataset

Single Family Loan-Level Dataset has data at the time of loan was originated

- Credit score: A number, prepared by third parties, summarizing the borrower's creditworthiness.
- First time homebuyer flag: Indicates whether the Borrower, had no ownership interest in a residential property during the preceding three-year period.
- Mortgage insurance percentage (MI %) -: The percentage of loss coverage on the loan.
- Original combined loan-to-value (CLTV): The ratio is obtained by dividing the total mortgage loan amount by mortgaged property's appraised value.
- Occupancy status: Denotes whether the mortgage type is owner occupied, second home, or investment property.
- Original UPB: Total Unpaid principal balance at time of loan origination.
- Original loan-to-value (LTV): The original mortgage loan amount by the mortgaged property's appraised value.
- Original debt-to-income (DTI) ratio: The sum of the borrower's monthly debt payments, including monthly housing expenses and mortgage payment divided by the total monthly income.
- Loan sequence number: The unique identifying key

Monthly Performance Dataset has monthly performance of data from date of origination to end of 2015.

- Loan sequence number: The unique identifying key
- Monthly reporting period: The as-of month for loan in YYYYMM format.
- Current UPB: Current unpaid principal balance of loan
- Current loan delinquency status: No of days loan payment is behind its due date
- Loan age: number of months since loan originated
- Remaining months of maturity: Number of months to loan maturity date
- Repurchase flag: Indicates loans that have been repurchased or made whole

[2] http://www.freddiemac.com/news/finance/sf_loanlevel_dataset.html

- Modification flag: Indicates that the loan has been modified example refinanced.
- Zero balance code: A code indicating the reason the loan's balance was reduced to zero
- Zero balance effective date: The date on which the event triggering the Zero Balance Code took place.
- Current interest rate: Reflects the current interest rate on the mortgage note, taking into account any loan modifications.
- Current deferred UPB: The current non-interest bearing unpaid principal balance.

Sample Dataset

Single Family Loan-Level Dataset

	CREDIT SCORE	FIRST TIME HOMEBUYER FLAG	MORTGAGE INSURANCE PERCENTAGE	CLTV	DTI Ratio	ORIGINAL UPB	ORIGINAL LTV	ORIGINAL INTEREST RATE	LOAN SEQUENCE NUMBER
0	751	N	0	71	20	180000	71	6.3	F199Q1000001
1	733	N	0	51	0	116000	51	6.3	F199Q1000002
2	755	N	30	95	38	138000	95	6.6	F199Q1000003
3	669	N	0	80	33	162000	80	7.12	F199Q1000004
4	732	N	0	25	10	53000	25	6.5	F199Q1000005

Monthly Performance Dataset

	LOAN SEQUENCE NUMBER	MONTHLY REPORTING PERIOD	CURRENT ACTUAL UPB	CURRENT LOAN DELINQUENCY STATUS	REMAINING MONTHS TO LEGAL MATURITY	REPURCHASE FLAG	MODIFICATION FLAG	ZERO BALANCE CODE	ZERO BALANCE EFFECTIVE DATE	CURRENT INTEREST RATE	CURRENT DEFERRED UPB
0	F199Q1000001	5/1/02	171982.4375	0	328	F	N	0	0	6.3	0
1	F199Q1000001	6/1/02	171571.3906	0	327	F	N	0	0	6.3	0
2	F199Q1000001	7/1/02	171158.3281	0	326	F	N	0	0	6.3	0
3	F199Q1000001	8/1/02	170742.8906	0	325	F	N	0	0	6.3	0
4	F199Q1000001	9/1/02	170325.1719	0	324	F	N	0	0	6.3	0

Transforming and filling missing data

Freddie Mac publishes the previously mentioned datasets publicly. As a public dataset it has missing values and unformatted fields. We apply conversion functions defined in Python to some of the column to get the correct data types.

- The Monthly Performance Dataset is a very large file; to overcome memory issues we read the data in chunks and create dataset by concatenating the chunks.

- Adding new columns 'Year' and 'Quarter' to the dataset
  Using the 'Monthly Reporting Period' column in the Monthly Performance Dataset, we add new columns 'Year' and 'Quarter' to this dataset. This will help us do the analysis per quarter.

- Adding new column "Performing"
  With respect to this dataset, a non-performing loan is one with a value greater than zero for Current Loan Delinquency Status. We introduced new calculated column "Performing" that takes a binary value of 0 and 1. A Non Performing Loan will have value of 1.

Data problems and cleaning operations

- Missing Credit Score Number:
  A very small number of the loan is missing credit score. Credit score is key information at the loan origination level. The loans with missing credit score were removed from the dataset.

- Invalid Current Loan Delinquency Status data**:**
  A small percentage of data has invalid value for Current Loan Delinquency Status column like null values or negative number. This data is removed from the dataset.

- Foreclosed Loan:
  The intended analysis deals with loan performance and the probability of defaulting. Some of the loans that are already foreclosed also appear in the Monthly data with any change in their status. We removed those loans from the dataset.

Joining two data set into one

After converting and cleaning the two datasets the Loan-Level Dataset and Monthly Performance Dataset were joined into one dataset on loan sequence number.

Final dataset columns

After merging the two datasets, the final list of columns was:

- Loan Sequence Number
- Monthly Reporting Period
- Current Actual Unpaid principal balance (UPB)
- Current Loan Delinquency Status
- Remaining Months To Legal Maturity
- Repurchase Flag
- Modification Flag
- Zero Balance Code
- Current Interest Rate
- Current Deferred Unpaid principal balance (UPB)
- Year
- Quarter
- Credit Score
- First Time Homebuyer Flag
- Mortgage Insurance Percentage
- Combined Loan To Value (CLTV)
- Debt To Income (DTI) Ratio

APPROACH TO DEFINING A MODEL

In keeping up with the practice of banks and other financial institutions to do risk analysis quarterly, we do our data analysis and modeling by rolling up the monthly data to a quarter.
In the mortgage industry if a loan payment is past the due date it gets marked as a Non-Performing Loan.

Our aim was to find a model that will predict if a given loan will miss the payment and hence become a Non-Performing Loan in the portfolio. Since loan can fall in one of the two categories Performing and Non-Performing, the outcome is binary.

We made use of Logistic Regression to measures the relationship between the binary dependent variable and one or more independent variables by estimating probabilities using a logistic function.

STEPS TO DEFINING THE MODEL

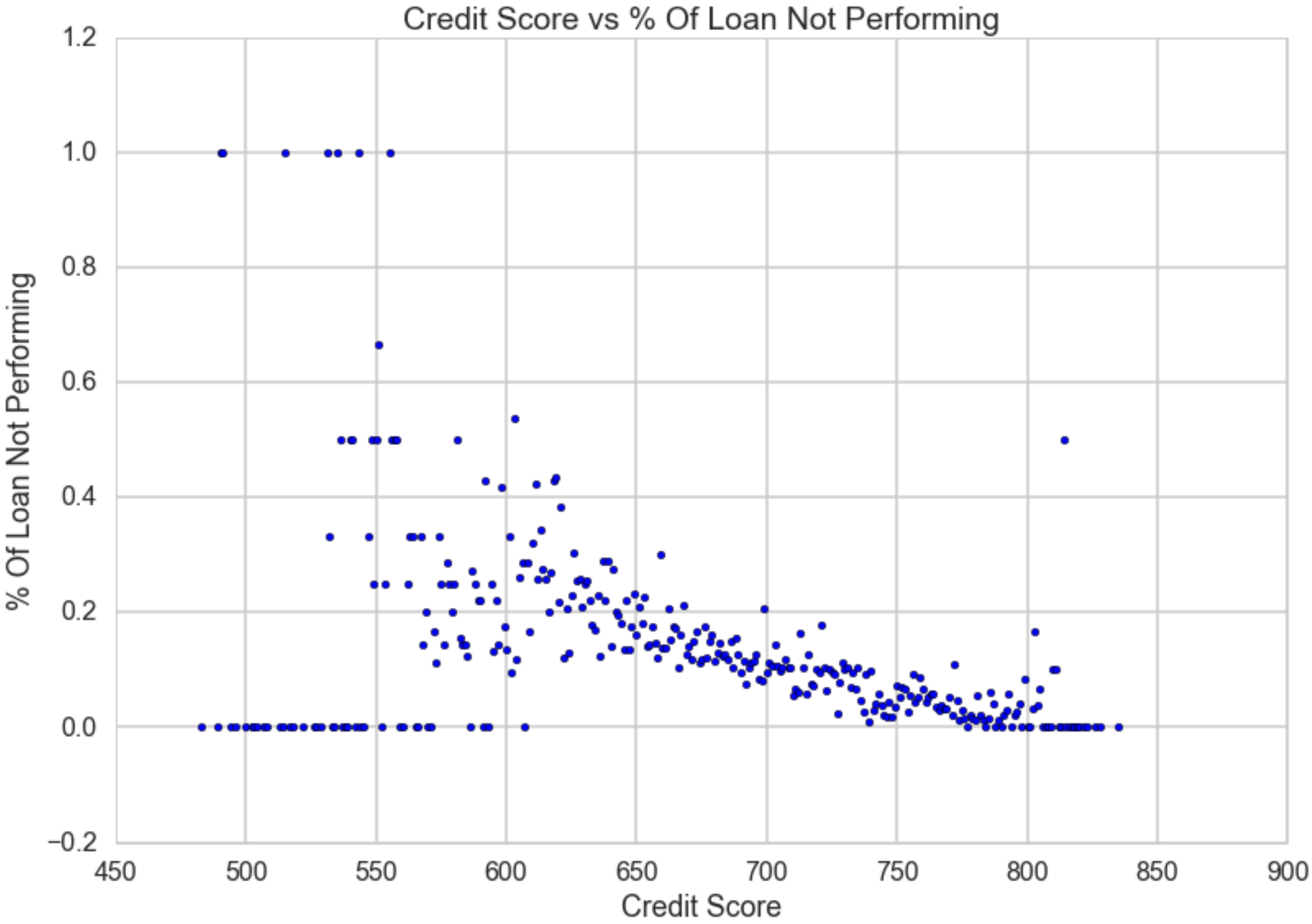
1. Feature Correlation Analysis:

   The dataset defines and captures many different aspects for a given loan. Some of the aspects are captured at loan origination and does not change for the lifetime of loan; other features are captured each month. We study the correlation of various features with the loan delinquency status (no of days loan is behind payment due date).
   The analysis is done per quarter. For this analysis we take data from 2013 Q4. The reason for choosing this data is to have a recent year data with enough monthly data.
   To analyze each feature against Loan Performance, we group the data for each feature and calculate the mean of "Performing" column, which gives us percentage of Non Performing loans in that group

- Credit Score correlation to Loan Performance

  We plot credit score grouped into group of 50, on x-axis and plot the mean of "Performing" column that represents percentage of non-performing loan in each credit score group on y-axis.

  The scatter graph gives us negative correlation between credit score and percentage of non-performing loan. Which illustrates that loans with higher credit score are less likely to be a non-performing loan.

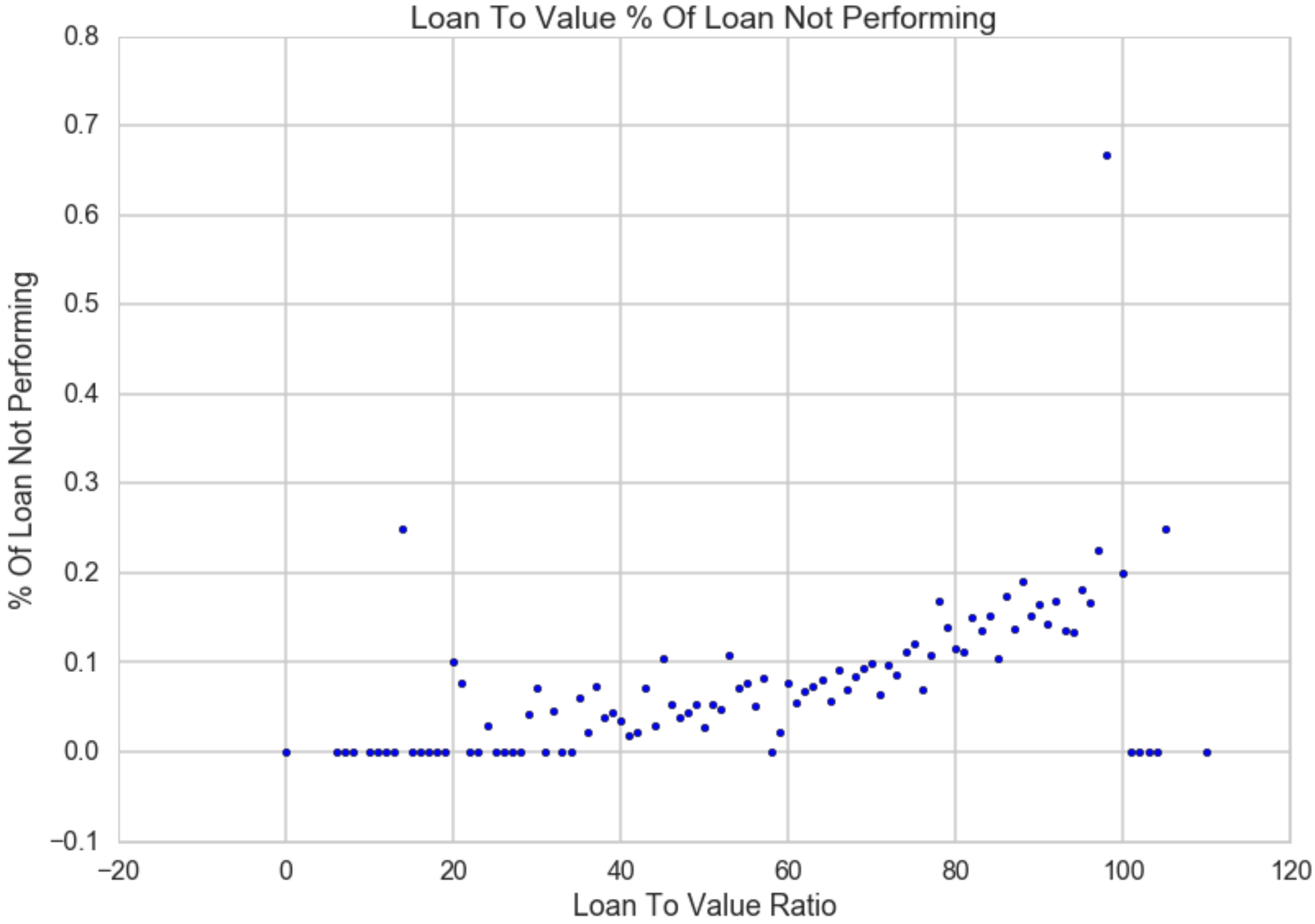


- Loan to Value (LTV) Ratio correlation to Loan Performance

  The loan to value ration varies from 0% to 100 %, with some outlier values above 100%.

  We plot loan to value ratio grouped into group of 20% on x-axis and plot the mean of "Performing" column that represents percentage of non-performing loan in each credit score group on y-axis.

  The scatter plot shows positive correlation between loan to value ratio and percentage of non-performing loan. This shows loans with higher loan to value ratio is more likely to be a non-performing loan.



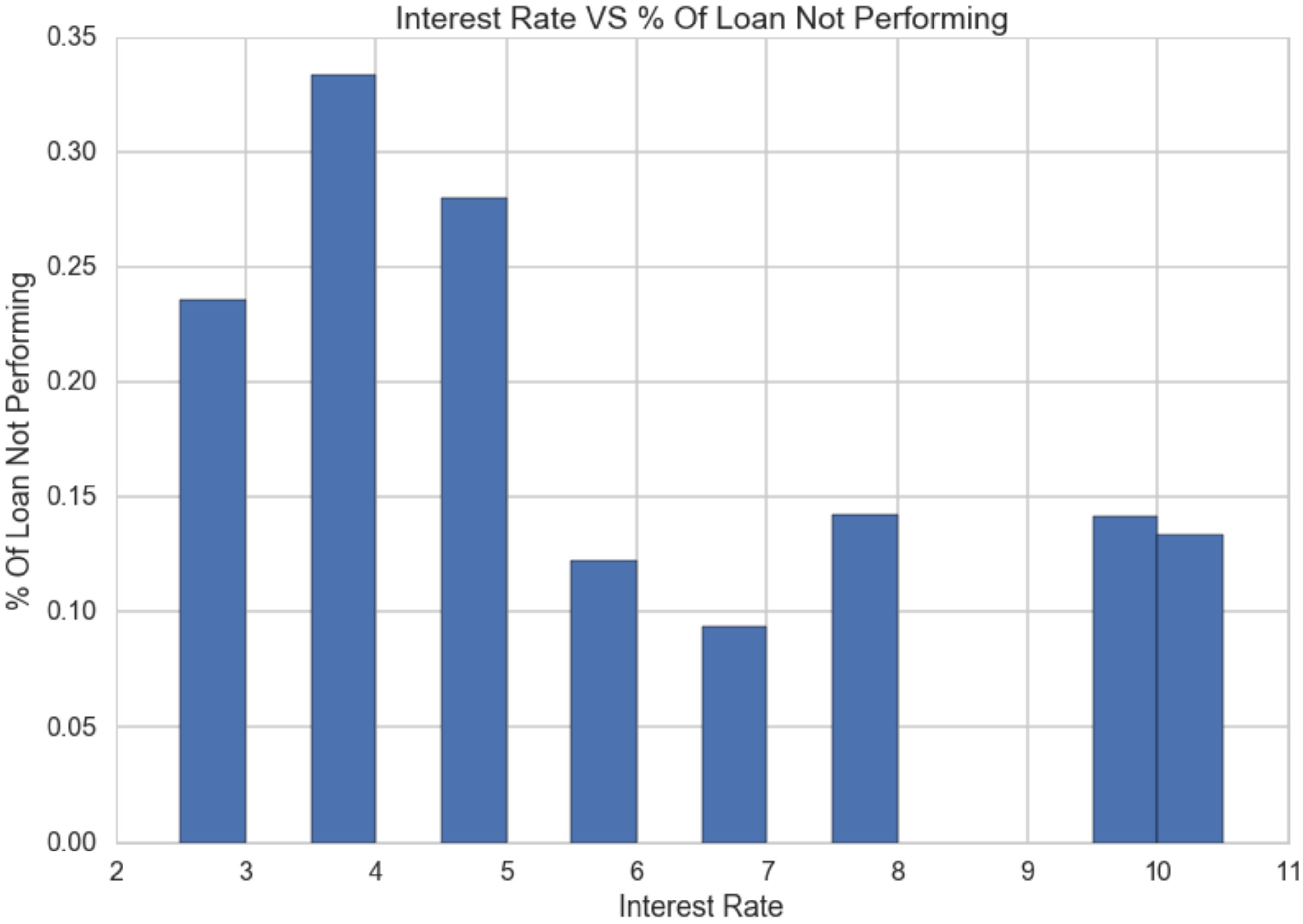
- Interest Rate correlation to Loan Performance

  The interest rates on the loan are in percentage form with continuous values varying by 0.001. To simplify the analysis we group the values in range of their integer value and give it value of integer value plus 0.5 as follows

  0% to 1 % → 0.5%
  1% to 2 % → 1.5%
  2% to 3 % → 2.5%

  We plot each interest group on x-axis of bar graph with height of each bar representing percentage of non-performing loan in that group.

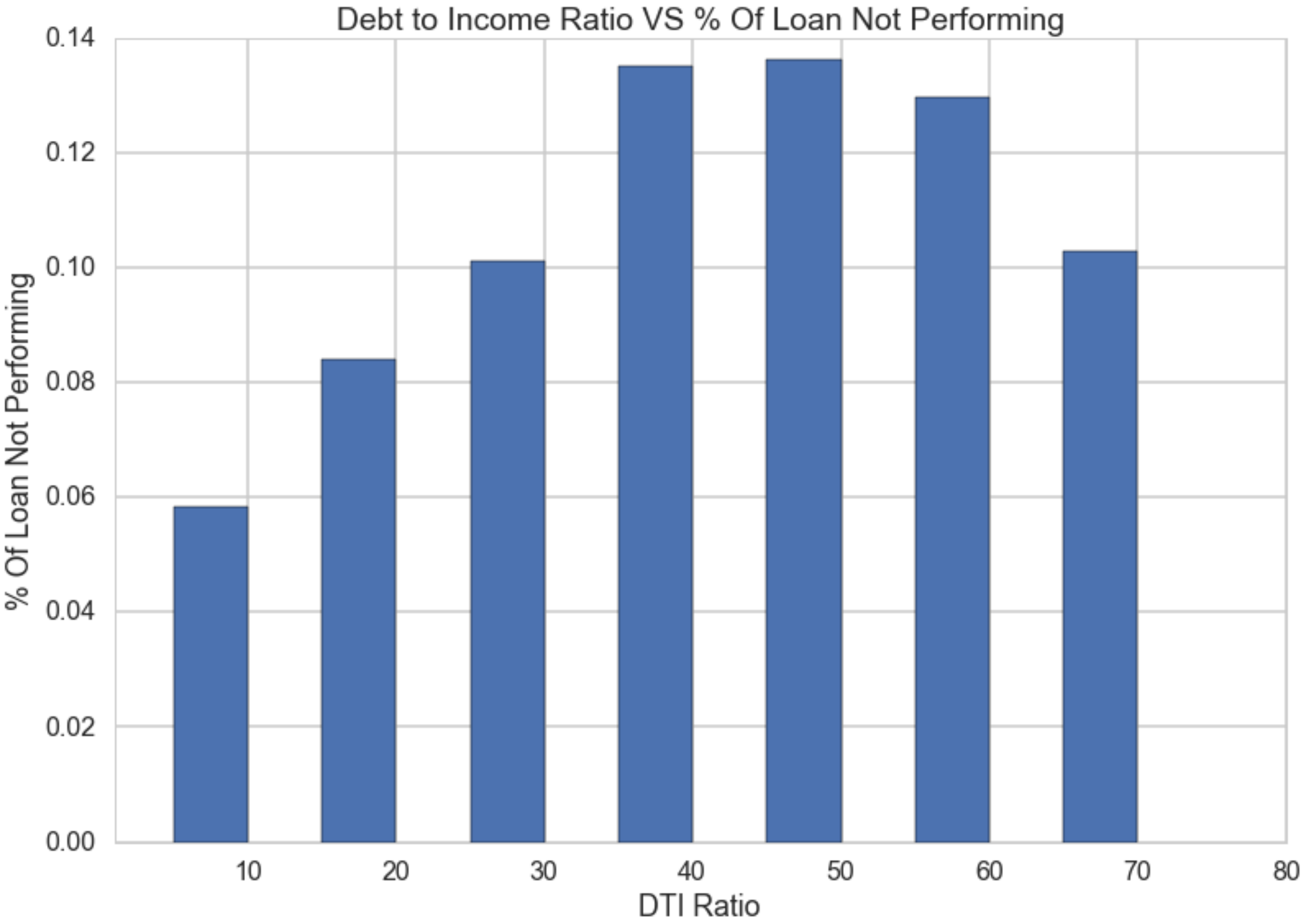
  The plot shows an interest rate range between 5% and 8% that has lowest percent of non-performing loan.



- Debt to Income (DTI) Ratio correlation to Loan Performance

Debt to Income ratio values varies from 0% to 100%. We group the values into range of 10 and plot them on x-axis of bar graph with height of each bar representing percentage of non-performing loan in that group.

The bar graph shows positive correlation between percentage of non-performing loan and debt to income ratio from 0% to 50%. Within this range higher value of debt to income ratio has high percentage of non-performing loan.



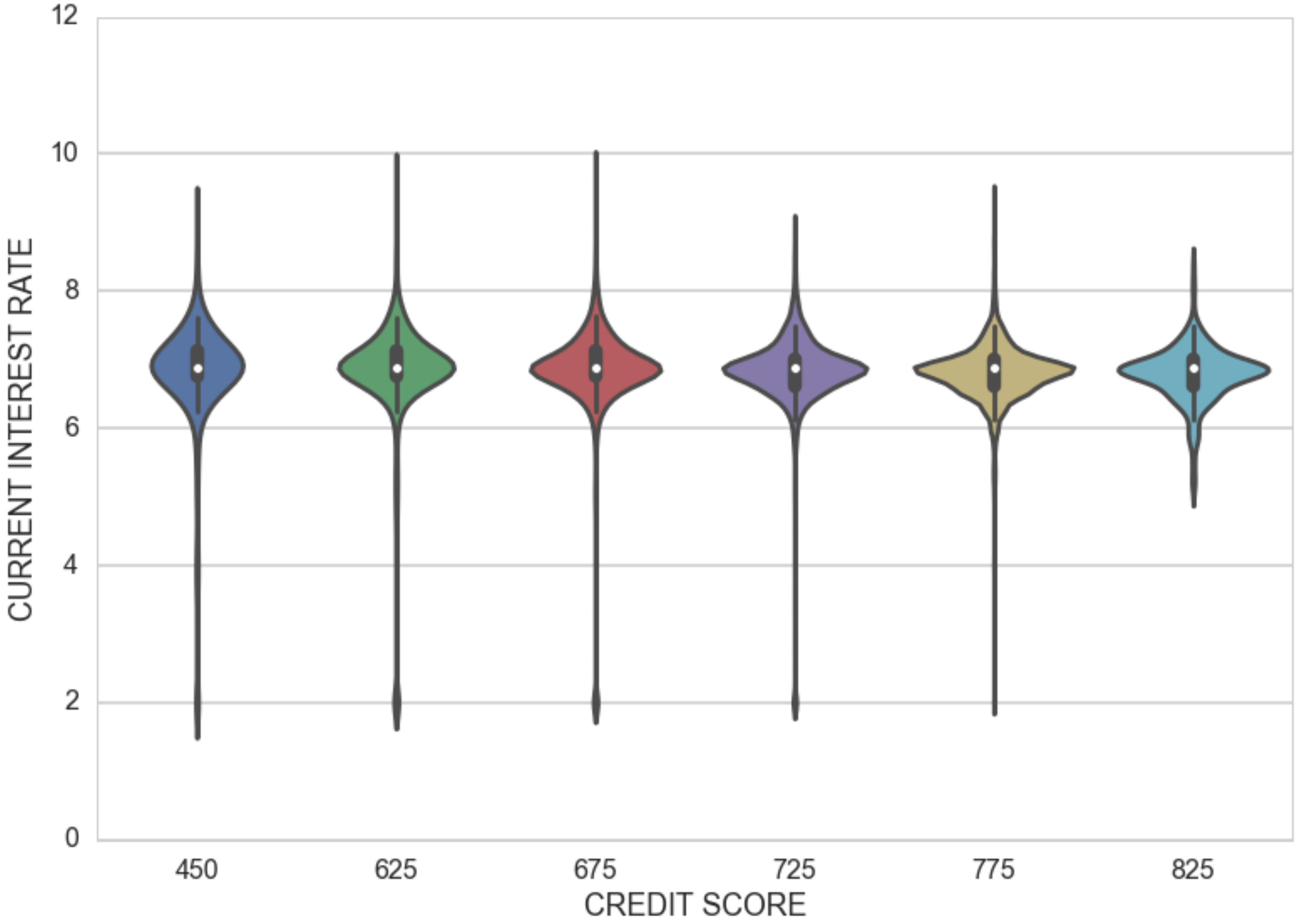
2. Checking The Correlation Between Features

We did additional analysis and constructed a correlation matrix to make sure there is no high correlation between the features.

- Credit Score and Interest Rate Correlation

  To test the correlation we plotted violin graph with credit score on x-axis and current interest rate on the y-axis.

  The violin plot suggests that interest rate is not highly correlated with the credit score but at credit score 825 the variance in interested rate is much lower. Credit score of 825 is very close to the maximum possible credit score. The plot suggests that at this level all the loans tend to have same interest rate.



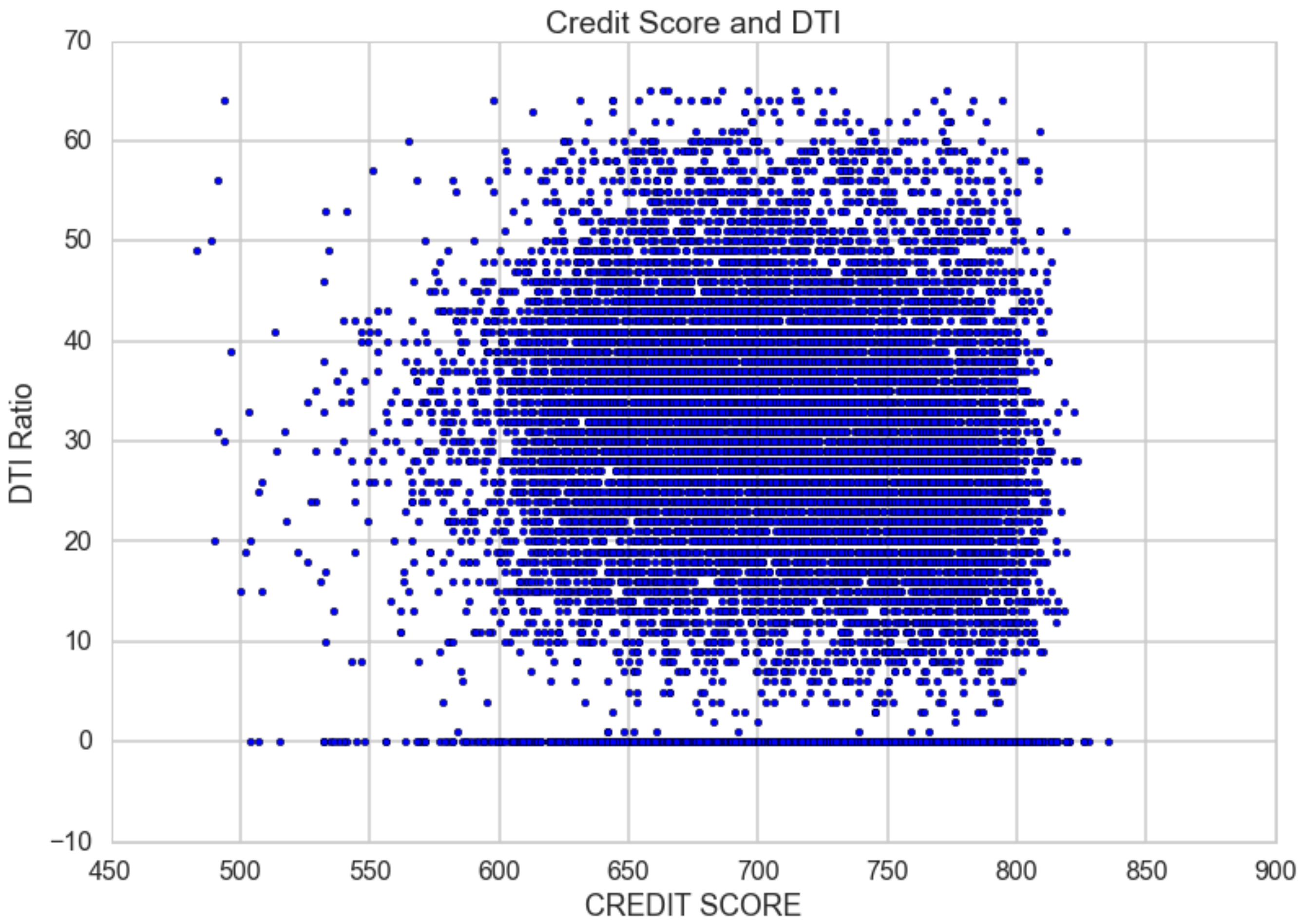
- Credit Score and Combined Loan To Value (CLTV) Ratio Correlation

  We use correlation matrix to test the correlation between credit score and combined loan to value ratio. The correlation coefficient value of -0.391254 shows weak negative correlation.

	CREDIT SCORE	**CLTV**
CREDIT SCORE	1.000000	-0.391254
CLTV	-0.391254	1.000000

- Credit Score and Debt to Income Ratio

We use scatter plot with credit score on x-axis and debt to income ratio on the y-axis. According to the scatter plot, there seems to be no correlation between cred score and DTI ratio.



The previous analysis gives us a list of features that we will use to train the logistic regression model

- CREDIT SCORE
- CURRENT INTEREST RATE
- CLTV
- DTI Ratio

TRAINING THE MODEL:

The dataset is divided into train and test data. The training data is used to define the model. The test data is then used to compare predicted output of the model with the expected output to determine best fit.

Steps To Train The Model

- The objective of the model is to predict Non Performing loan in a given quarter. Hence we take dataset for one quarter. In this case 2013 Q4 and we look at the loan delinquency at the quarter level.
- We defined a new column to capture Non Performing Loan status as binary value.
- There are 391192 loans and 72332 of them, 18.5%, are non-performing. Since the percentage is so low we rebalanced the dataset so that 50% of the loans are non-performing by sampling from the both classes with replacement.
- We use scikit-learn cross_validation to get training and test data, we then use scikit-learn GridSearchCV with 5 folds to get best regularization parameters for Logistic Regression.

The table below shows the estimated regression coefficient.

FEATURES	**COEFFICIENT**
Credit Score	-0.00963969
Current Interest Rate	0.05448207
CLTV	0.02589876
DTI Ratio	0.00706395
Intercept	4.07803818

EVALUATE THE MODEL:

After we trained our model, we run the standard evaluation and scoring function to test the quality of the model.

1 Getting key metrics using scikit-learn API

Precision = true positive / (true positive + false positive)

The precision is intuitively the ability of the classifier not to label as positive a sample that is negative. As we see the table below, the model precision for predicting non-performing loan is 0.65; it means there is 65% chance that predicted non-performing loan is correct.

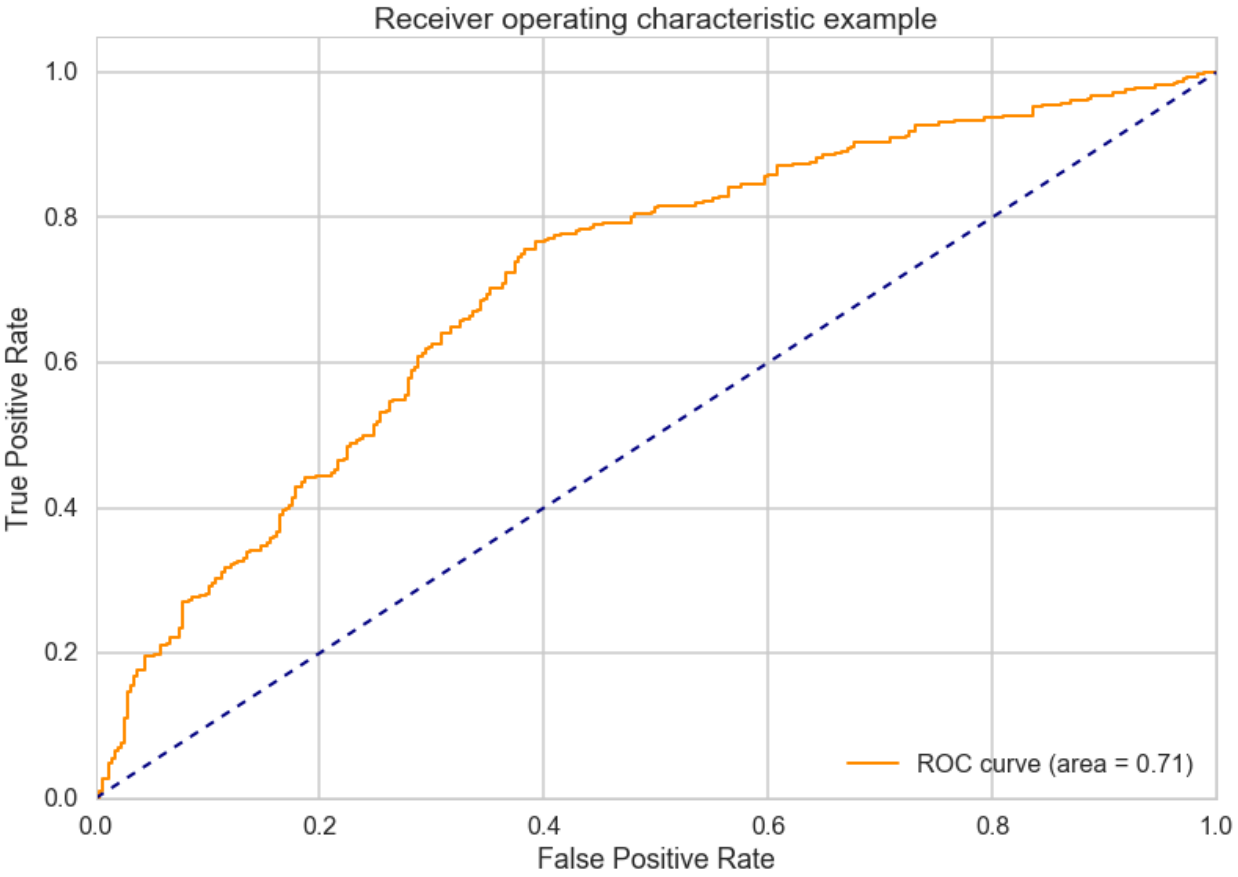
Recall = true positive / (true positive + false negative)

The recall is intuitively the ability of the classifier to find all the positive samples. As we see in the table below, model recall for predicting non-performing loan is 0.68; it means the model predicts 68% of non-performing loans correctly.

	Precision	Recall	F1-score	Support
Non-Performing	0.65	0.68	0.67	335

1. ROC

Receiver Operating Characteristic (ROC) metric is used to evaluate classifier output quality across many different thresholds. ROC curves feature true positive rate on the Y-axis, and false positive rate on the X-axis. The Area Under the Curve **(AUC)** measures the overall quality across thresholds.



The above ROC curve graph has AUC of 0.71, which means if we take random pair of loans from dataset then there is 71% chance the model will correctly classify loan as being non-performing.

4.Statistics Metrics from Statmodels API

Scikit-learn does not give us the statistics metrics of the model so we use the statmodels API. We use the training data used in logistic regression model to fit a model using statmodels API Logit Regression. The models give us summary of the statistics metrics.

Logit Regression Results

	Coef	Std err	Z	P>|Z|	[95.0% Conf. Int.]	Odds Ratio
Credit Score	-0.0062	0.000	-13.109	0.000	-0.007 -0.005	0.99381918034
Cur. Interest Rate	0.1909	0.044	4.315	0.000	0.104 0.278	1.21033841219
CLTV	0.0340	0.003	12.323	0.000	0.029 0.039	1.03458460673
DTI Ratio	0.0103	0.003	3.393	0.001	0.004 0.016	1.01035322759

Summarizing the Result

1. p-values

The p-value for above feature tests the null hypothesis that the features have no effect on loan being non-performing. A low p-value (< 0.05) indicates that you can reject the null hypothesis.

All features have low p-value (< 0.05) indicating that all the features contribute to the outcome of loan performance. Since none of the feature have p-value > 0.05, hence we cannot reject any feature.

2. Odds Ratio

Odds ratio of a feature shows how change in feature impacts odds of outcome in this case odds of loan being non performing. Credit score has odds ratio value <0.05, that means it has negative impact on the odds of loan being non-performing. The other three features, Current Interest Rate, Combined loan to value ratio and Debt to Income ratio have odds ratio >0. All these features have positive impact on the odds of loan being non-performing.

RECOMENDATION

The large mortgage banks and other financial institutes buy loans from other banks and mortgage companies. This results in them having a variety of loans in their portfolio. Using a model that can identify potential Non Performing loans in advance gives these financial institution useful metrics that they can use to perform their risk analysis.

After the mortgage crisis of 2008, consumer habits have changed. Consumers constantly make choices between paying their mortgage and walking out. Its prudent for these financial institutes to identify loans that are at risk of default and provide them with incentives and other offers to keep them current in the mortgage payment.

Loan performance is also impacted by macroeconomic factors like GDP growth, unemployment rate and share market movement. Model can be further improved by incorporating these macroeconomic factors.