SINGLE FAMILY LOAN PERFORMANCE MODEL

PROBLEM STATEMENT

Financial institutions conduct stress tests to gauge the resilience of their balance sheets to substantial macroeconomic shocks. One way to measure the performance of a financial institution is by assessing the institution's loan portfolio loss under stressed scenarios. The first step in assessing loan loss is to estimate the probability of default (PD).

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 default risk by estimating relationships between default risks and default determinants.

The Model will help financial institutes to run their portfolio and get understanding of risks

TARGET AUDIENCE

Though we are using dataset provided by Freddie Mae, which is one of the biggest mortgage buyers from various mortgage banks. The use of this data analysis and model can also be extended to banks that hold big mortgage portfolios.

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

DATA DESCRIPTION USED FOR MODELING

This data is made available from Freddie Mae.

http://www.freddiemac.com/news/finance/sf_loanlevel_dataset.html

For every year the loan contains zipped file for each quarter example.

Each zipped file has two dataset in tab separated 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
- FIRST TIME HOMEBUYER FLAG
- MORTGAGE INSURANCE PERCENTAGE (MI %) -
- ORIGINAL COMBINED LOAN-TO-VALUE (CLTV)
- OCCUPANCY STATUS
- ORIGINAL UPB
- ORIGINAL LOAN-TO-VALUE (LTV)
- ORIGINAL DEBT-TO-INCOME (DTI) RATIO
- LOAN SEQUENCE NUMBER

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

- LOAN SEQUENCE NUMBER
- MONTHLY REPORTING PERIOD (YYYYMM)
- CURRENT UPB
- CURRENT LOAN DELINQUENCY STATUS
- LOAN AGE
- REMAINING MONTHS OF MATURITY
- REPURCHASE FLAG
- MODIFICATION FLAG
- ZERO BALANCE CODE
- ZERO BALANCE EFFECTIVE DATE
- CURRENT INTEREST RATE
- CURRENT DEFERRED UPB

*** **LOAN SEQUENCE NUMBER** is the unique identity key

Sample Data Set

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

The above-mentioned datasets are made published publically by Freddie Mac. As a public dataset it has missing and unformatted data set.

We apply conversion functions defined in python to some of the column to get the correct data types.

- Conversion function applied while reading **Loan-Level Dataset**

```
pd.read_csv ('data/historical_data1_Q11999/historical_data1_Q11999.txt','|',
                index_col=None, encoding='utf-8', low_memory=False,usecols=fields_Origin
               , converters={'CREDIT SCORE':stringToInt,
                             'DTI Ratio':stringToFloat,
                             'CLTV':stringToFloat}
              )
```

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

- Conversion function applied while reading **Monthly Performance Dataset**

```
pd.read_csv('data/historical_data1_Q11999/historical_data1_time_Q11999.txt','|',
              index_col=None, parse_dates=['MONTHLY REPORTING PERIOD'],
              date_parser=dateparse,encoding='utf-8',low_memory=False,chunksize=10000
             ,usecols=fields_Month,converters={'CURRENT ACTUAL UPB':stringToFloat,
                                 'CURRENT LOAN DELINQUENCY STATUS':stringToInt})

df = pd.concat(chunk for chunk in reader)
```

- Adding new columns to the dataset

Using the 'MONTHLY REPORTING PERIOD' column in the Monthly Performance Dataset, we add two new columns 'Year' and 'Quarter' to this dataset. This will help us do the analysis per quarter.

Data cleaning operations

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

- Invalid Current Loan Delinquency Status data**:**
  Small percentage of data has invalid value for Current Loan Delinquency Status column. This data is cleaned from the dataset.

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

Joining two data set into one

After conversion and cleaning the two datasets Loan-Level Dataset and Monthly Performance Dataset are joined into one dataset. The join column is LOAN SEQUENCE NUMBER

```
pd.merge(dfClean,fDataClean,on='LOAN SEQUENCE NUMBER')
```

Final Dataset Columns

- 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
- Current Interest Rate

- Current Deferred UPB
- Year
- Quarter
- Credit Score
- First Time Homebuyer Flag
- Mortgage Insurance Percentage
- CLTV
- DTI Ratio

APPROACH TO DEFINE A MODEL

Most of the banks or financial institutes publish their report or analyses the risk quarterly we take do the risk analysis for each quarter.
In mortgage industry if a loan payment is past the due date it gets marked as a Non-Performing loan.

Our aim is 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 expected outcome is a binary result where loan can fall in either of the two categories.

We make use of Logistic Regression model to measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function, which is the cumulative logistic distribution

STEPS TO DEFINE THE MODEL

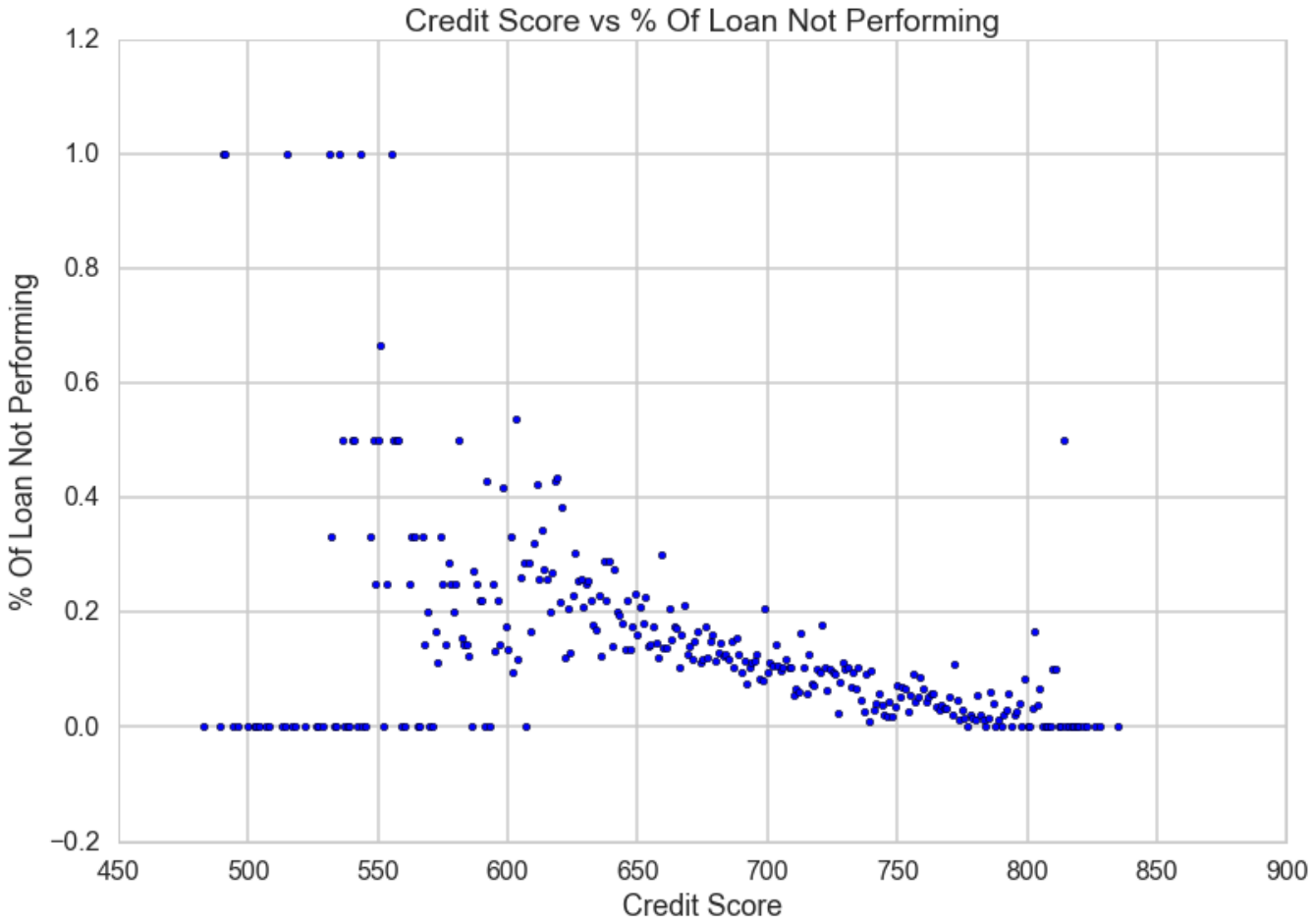
1. FEATURE CORRELATION ANALYSIS:

The dataset defines and captures many different fields for a given loan. These columns are captured at the time of loan origination and also to define monthly loan performance. We study the correlation of various features with the loan delinquency status, which tells us how many days a given loan is behind its payment date.
The Analysis is done for a quarter. For this analysis we take data from Year 2013 and Quarter 4. The reason for choosing this data is to have a recent year data with enough monthly data.

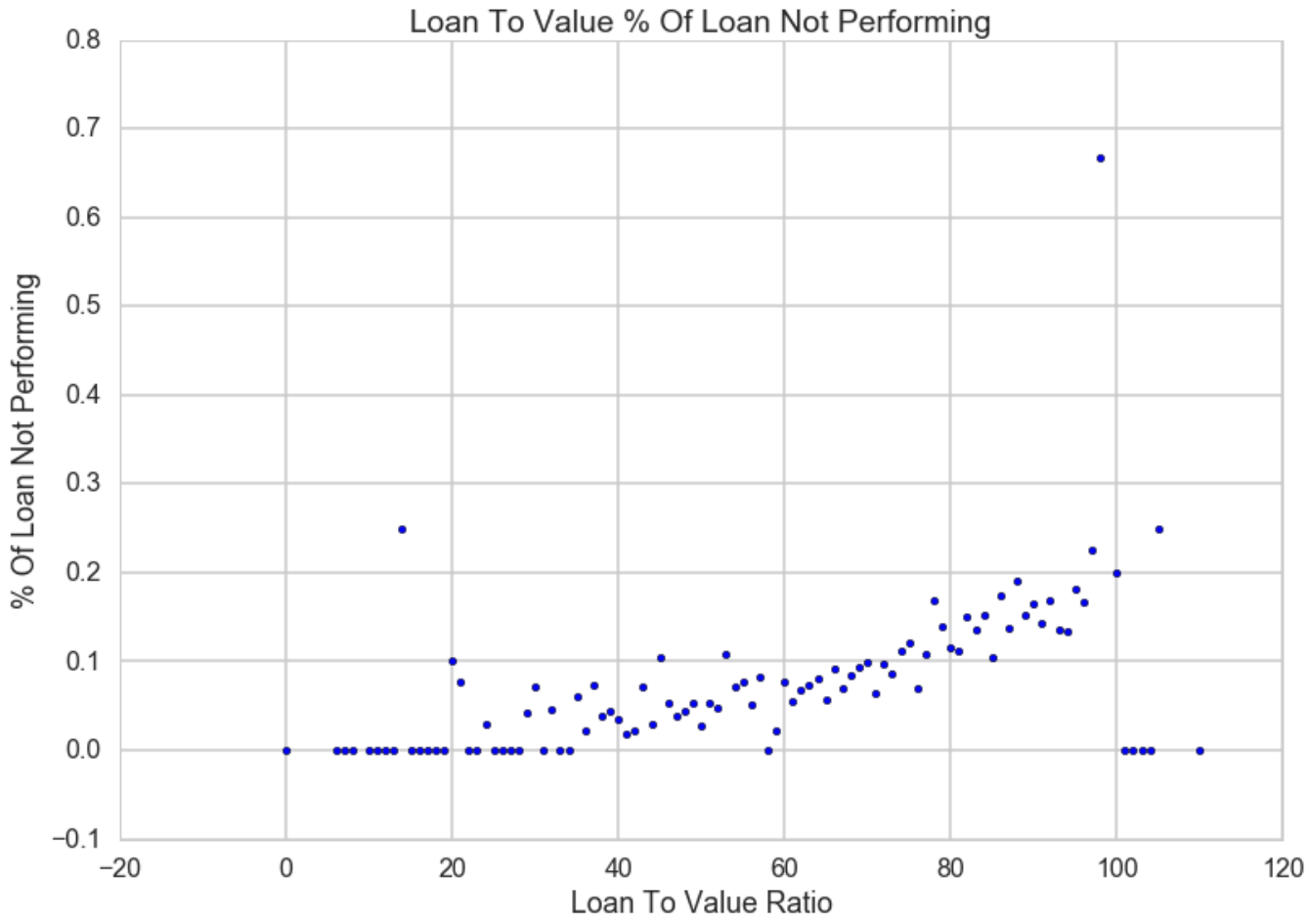
A Non Performing Loan is a loan that is behind in payments.
With Respect to data set Current Loan Delinquency Status >0 is a Non Performing LOAN.

We introduce new calculated column df['PERFORMING'] that take binary data of 0 and 1. A Non Performing LOAN will have value of 1. To analyze each features against Loan Performance, we group the data for each feature and calculate the mean of df['PERFORMING'] column , this gives us percentage of Non Performing loan in that group

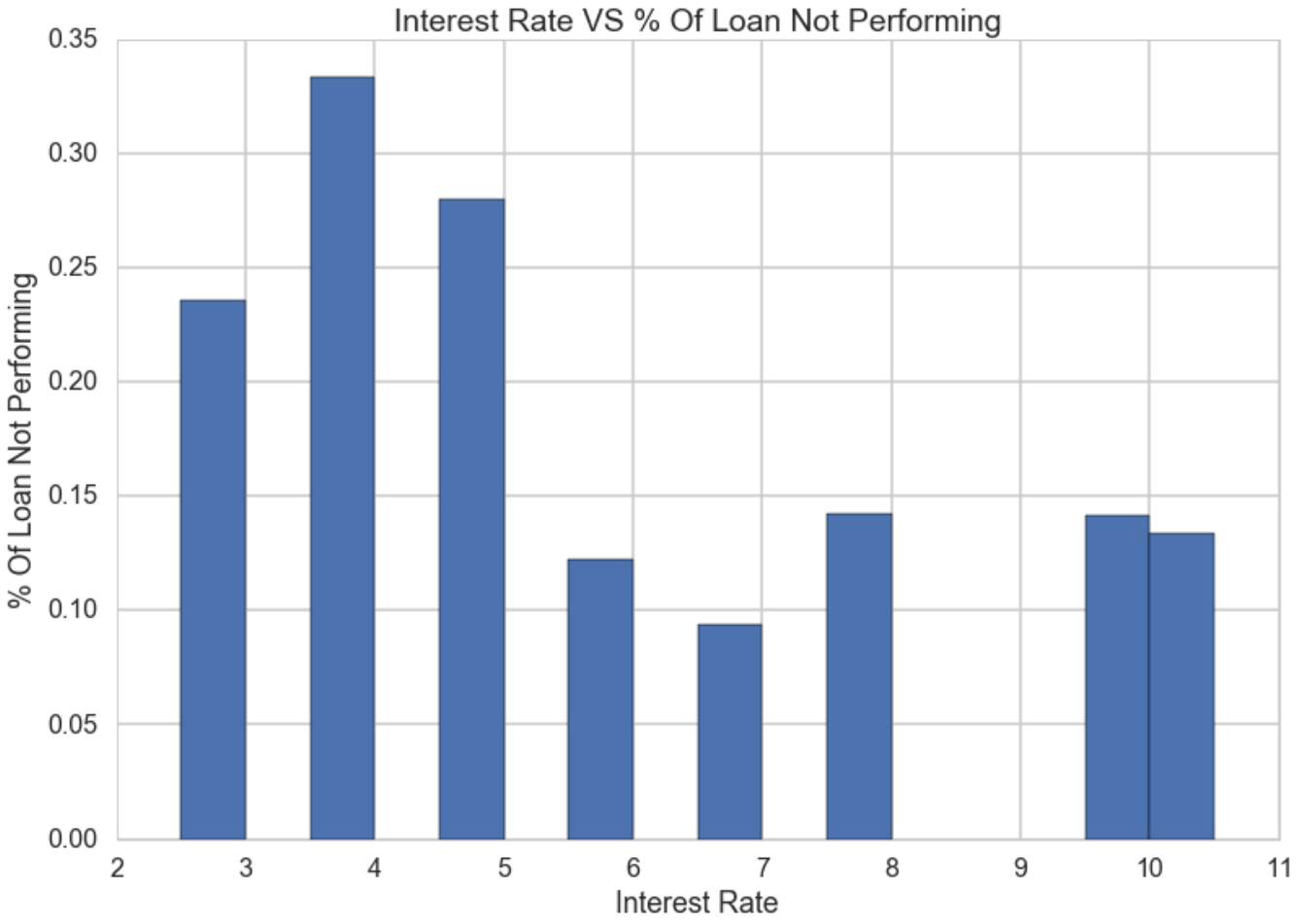
- Impact of Credit Score on the Loan Performance
  - We group the credit score into groups with range of 50 (450,500,550…)
  - We then calculate the percentage of Non Performing loans in the each range.
  - Scatter plot shows us the relationship.



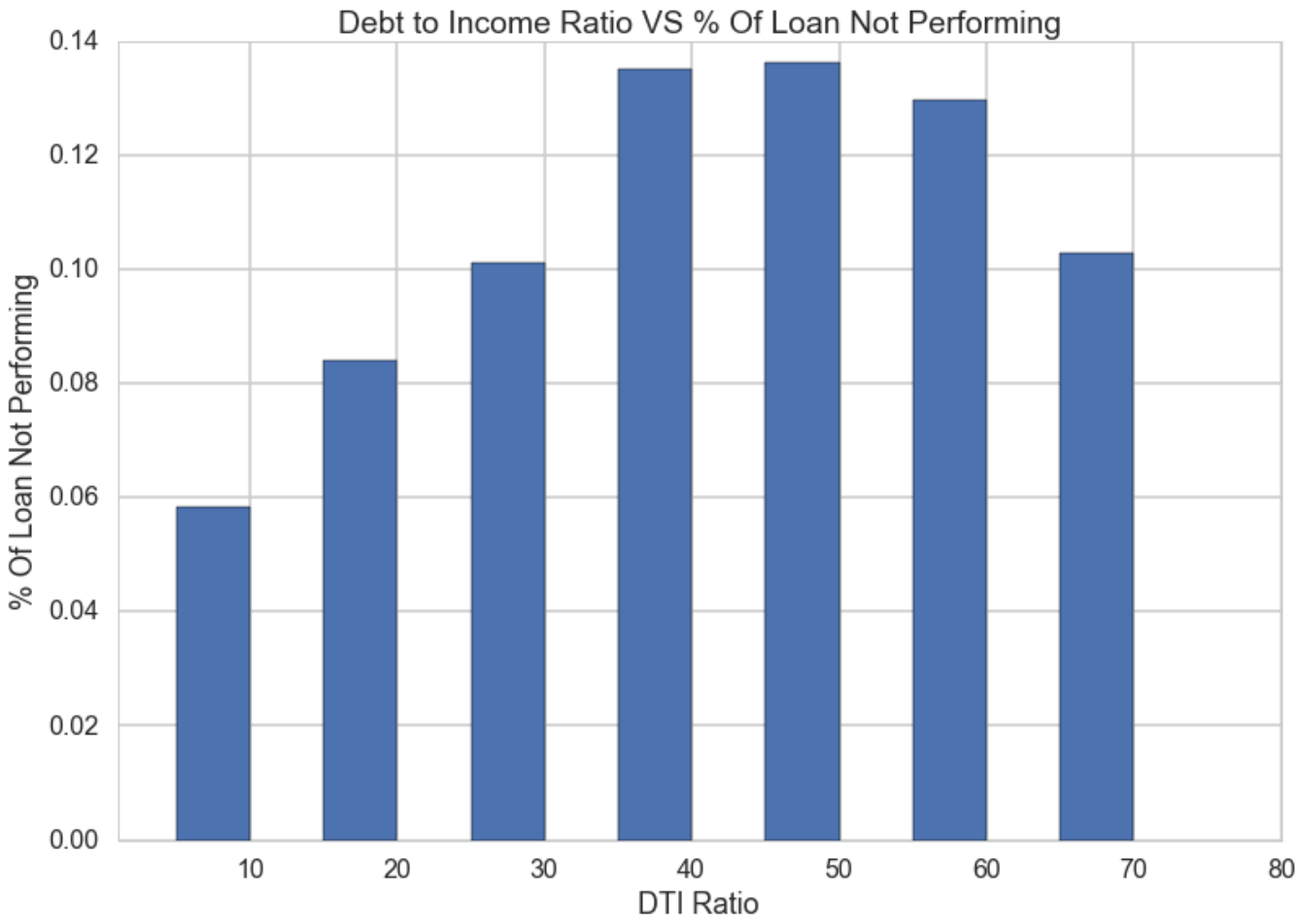
- LTV (Loan to Value Ratio) relation to Loan Performance
  - Loan to Value varies from 0% to 100% and in some case it can go above 100%
  - Data with LTV >100% are treated as outlier.



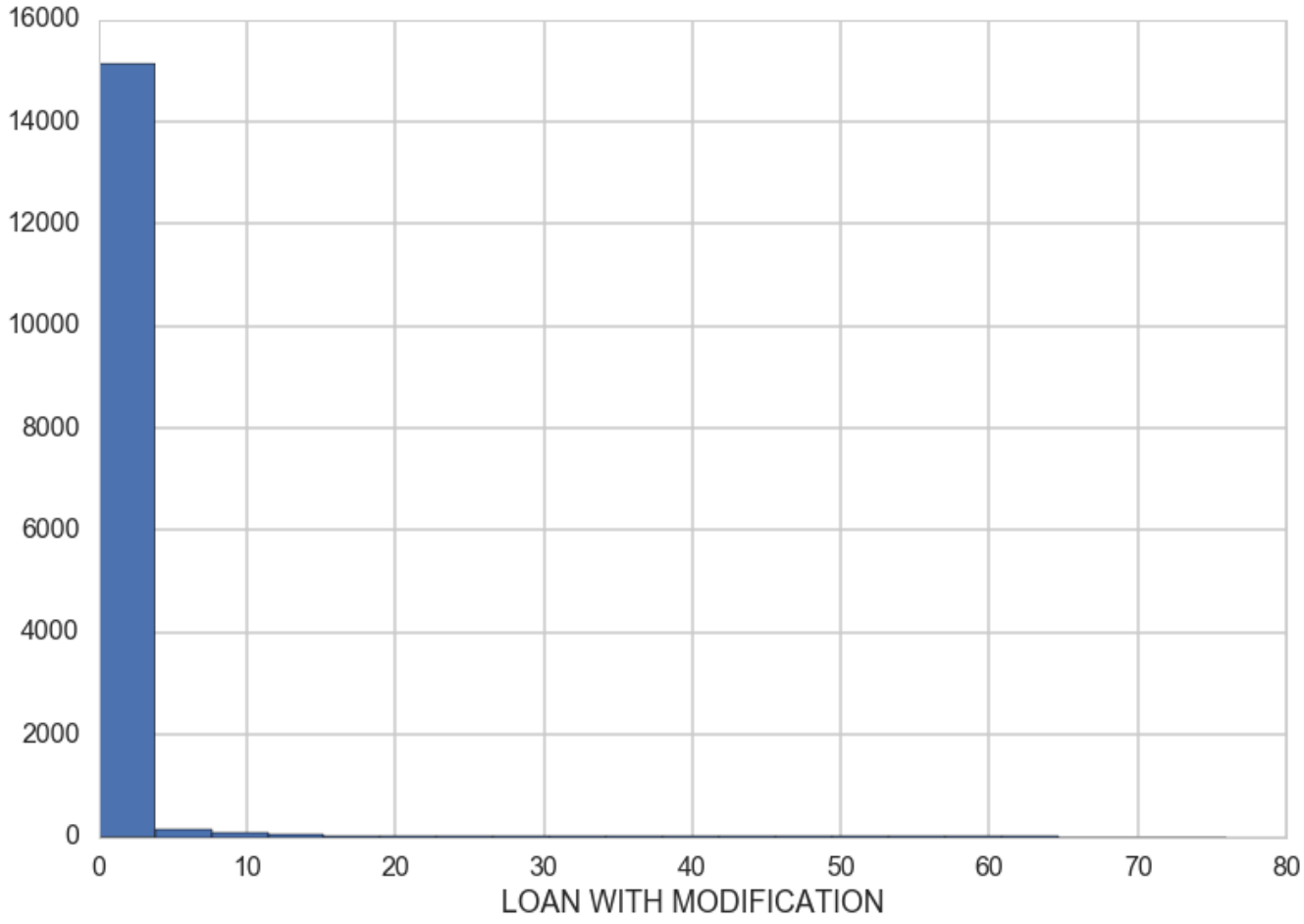
- Interest Rate relation to Loan Performance
  - The interest rates on the loan are in percentage with continuous values changing with 0.001. To simplify the analysis we group the data in range of their integer value and give it value of Integer value plus 0.5 as follows
    0% to 1 % → 0.5%
    1% to 2 % → 1.5%
    2% to 3 % → 2.5%
  - We then plot mean of df['Performing'] column for each group.
  - This plot shows that there is a interest rate range around which there is lowest percentage of loans default.



- DTI (Debt to Income Ratio) relation to Loan Performance
  - DTI measures the ability of person to pay the loan. It varies from 0 to 100%.
  - We group them in range of 10 and get the mean of df['Performing'] value.



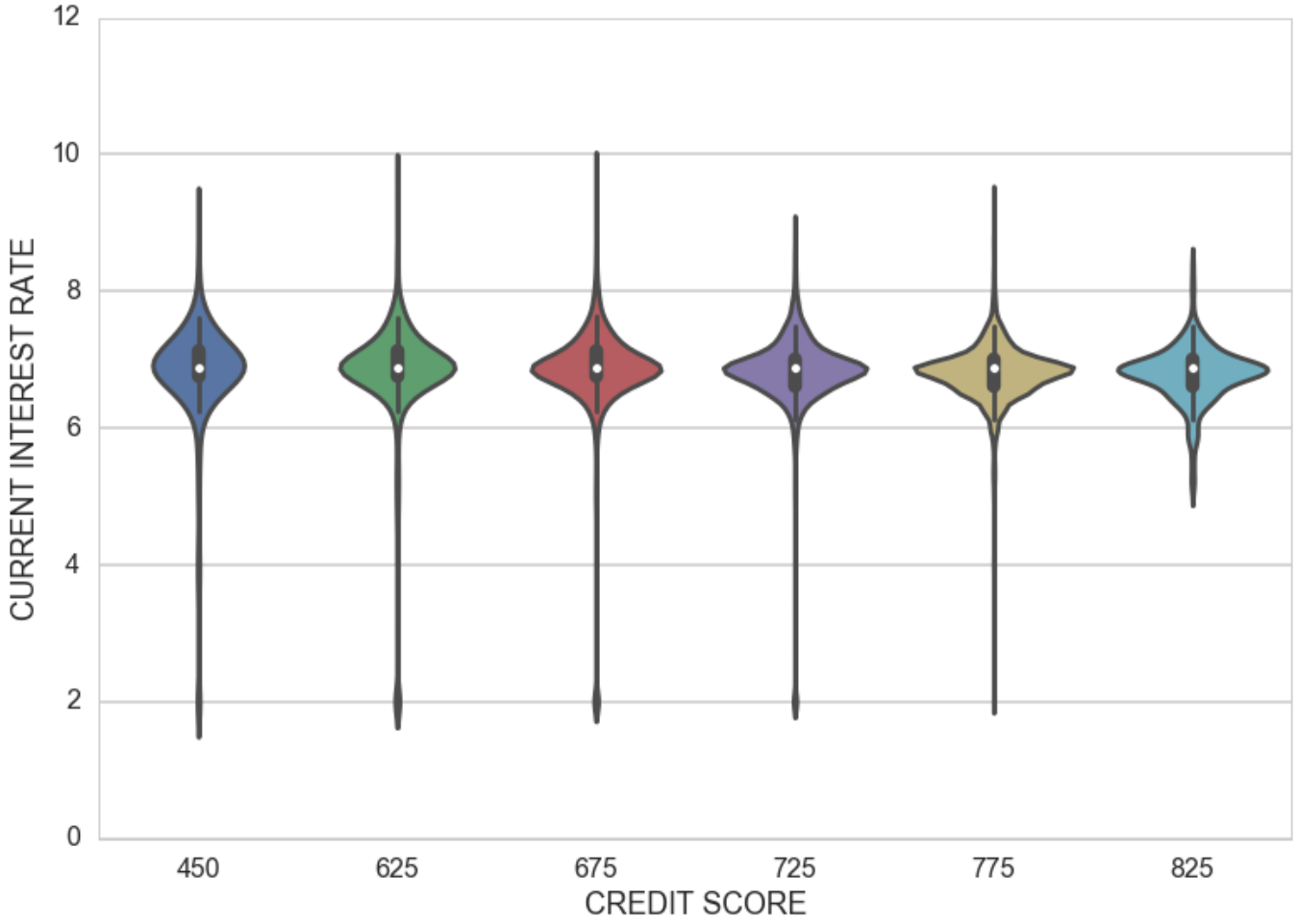
- Modification Flag in the dataset
  - Modification flag seems as good feature but we see most of the data is skewed towards non-modification.
  - We do this analysis on other Flag values and because of lack of data being updated we don't include them in the modeling.



2. Checking the Correlation between features

We use graph plotting and correlation matrix to make sure there is no high correlation between the features.

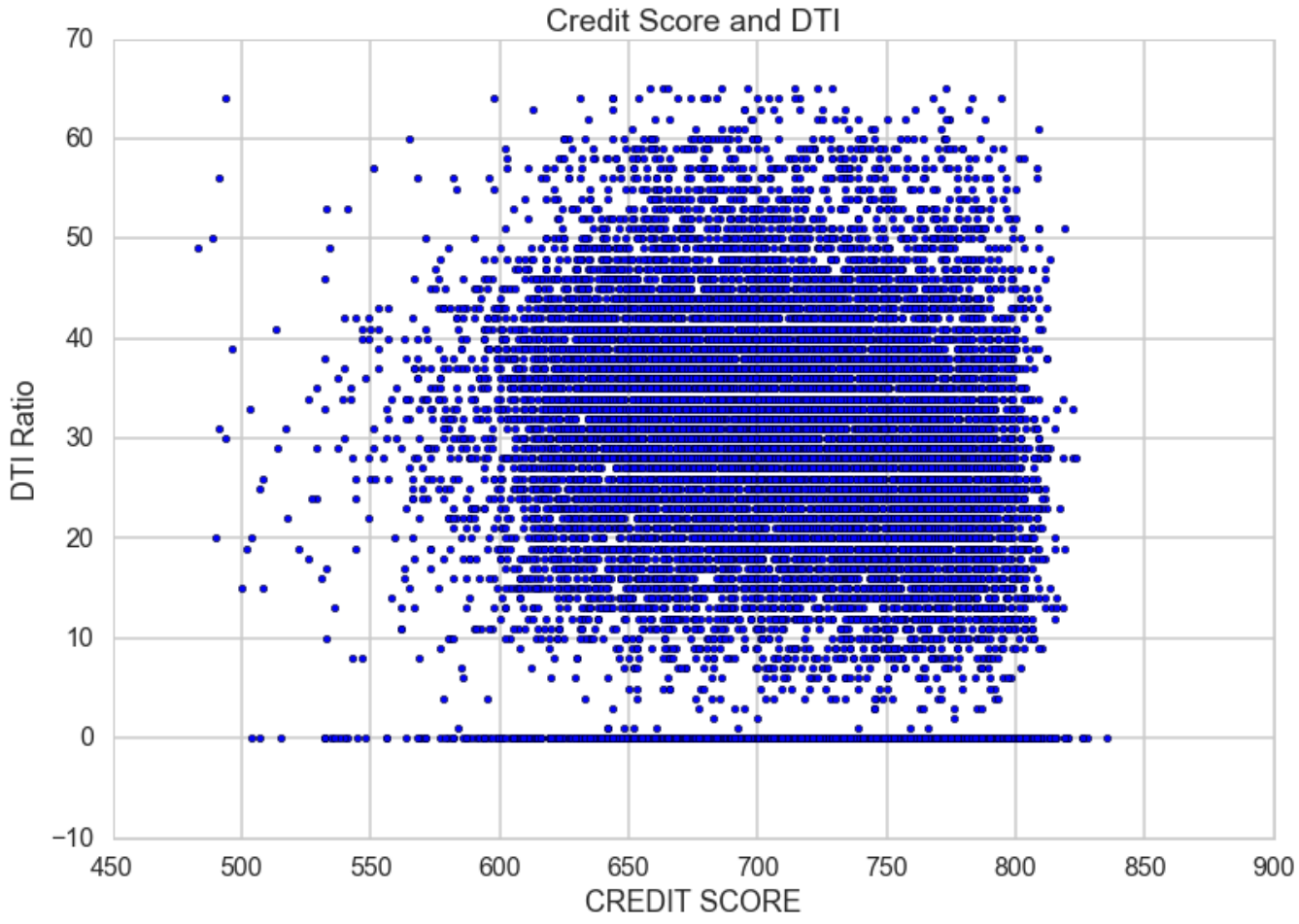
- Credit Score and Interest Rate Correlation
  - Plotting Violin Graph shows weak correlation



- Credit Score and Loan To Value Correlation
  - We use correlation matrix to test the correlation

	CREDIT SCORE	CLT
CREDIT SCORE	1.000000	-0.391254
CLTV	-0.391254	1.000000

- Credit Score and Debt to Income Ratio
  - Scatter plot shows weak correlation



Above analysis gives us list of features that we will use to train the Logistic Regression Model

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

TRAIN TEST THE MODEL:

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

Steps To Train Model

1. PREPARE DATASET:

- The objective of the model is to analysis Non Performing loan in a given quarter. Hence we take dataset for one quarter. In this case quarter 4 of year 2013

```
df1=dfloan[(dfloan['year']==2013) & (dfloan['quarter']==4)]
```

- We then look at the loan delinquency at the quarter level

```
df1=df1.groupby('LOAN SEQUENCE NUMBER').max()
```

- We need to define new column to capture Non Performing loan as binary value

```
df1['NON-PERFORMING']=['Y' if x>0 else 'N' for x in df1['CURRENT LOAN DELINQUENCY STATUS']]
```

- The number of rows with Non Performing loan out of total rows

  Total Number of Loans = 391192

  Non Performing Loan = 72332

- We get very unreliable results creating model with so small percentage (72332/391192 =0.1845) is actually non-performing.

- To create a reliable model, we create a data set from above dataset with 50% performing and 50% non-performing dataset.

- Get both two set of data with flag Y and N of equal size and merge them

```
dfNP=df1[df1["NON-PERFORMING"]=='Y']
dfP=df1[df1["NON-PERFORMING"]=='N']
dfNP=dfNP.sample(n=1700)
dfP=dfP.sample(n=1700)
df=dfNP.append(dfP)
```

- We use scikit-learn cross_validation to get training and test data. We then use scikit-learn GridSearchCV with 5 folds to get best parameters for Logistic Regression.

- Model Coefficient and intercept

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

1 Getting key Metrics using scikit-learn API

Precision = tp / (tp + fp)
The precision is intuitively the ability of the classifier not to label as positive a sample that is negative.

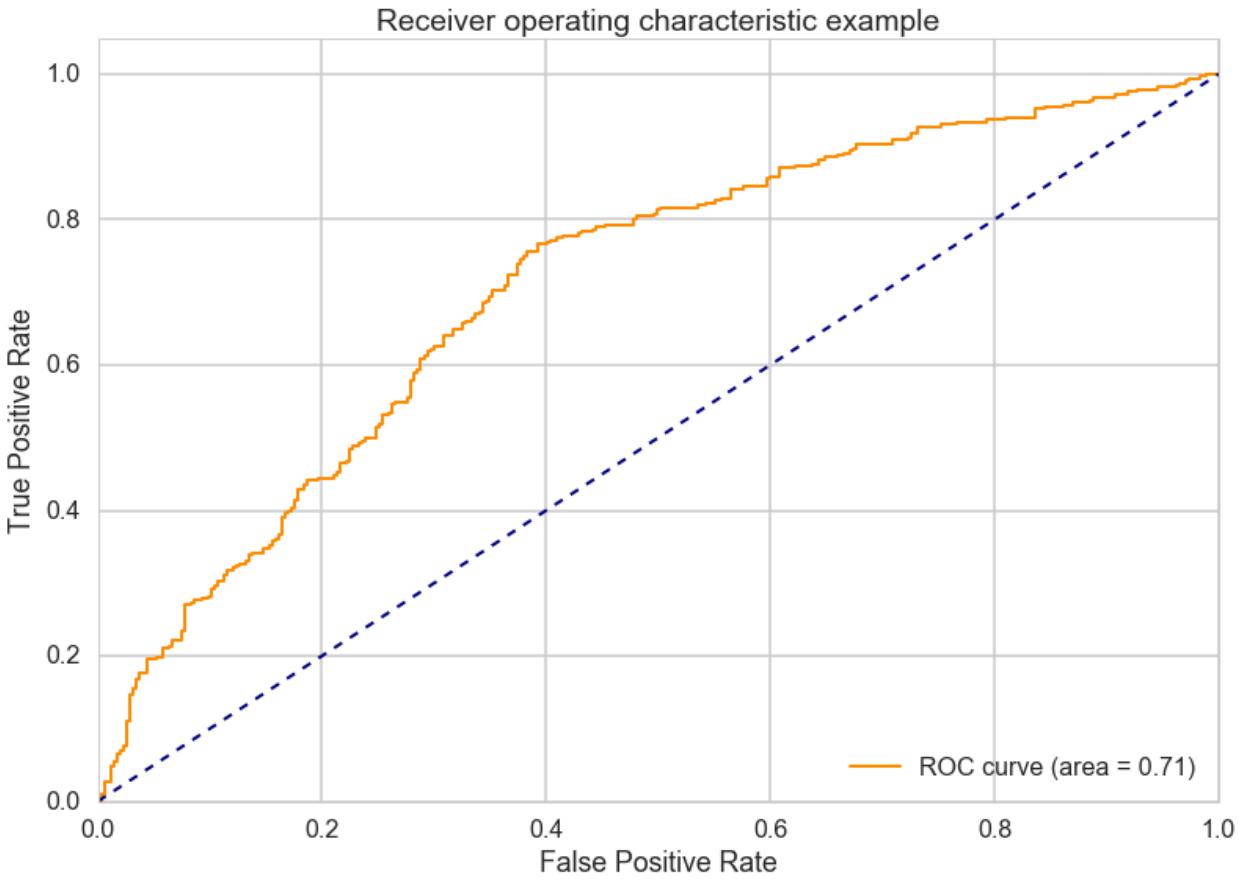
Recall = tp / (tp + fn)
The recall is intuitively the ability of the classifier to find all the positive samples.

Output 1 is the tp (True Positive) showing loan is Non Performing

	precision	recall	f1-score	support
0	0.68	0.65	0.66	345
1	0.65	0.68	0.67	335

2. ROC

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



3. Odd Ratio

Odds (success) = p/(1-p) where p is probability of success
Odd Ratio formula **np.exp (coeff)**

FEATURES	COEF	ODD Ratio
CREDIT SCORE	-0.00963969	0.990406622878
CURRENT INTEREST RATE	0.05448207	1.05599354228
CLTV	0.02589876	1.02623704698
DTI Ratio	0.00706395	1.00708895855

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

Dep. Variable:	y	No. Observations:	2720
Model:	Logit	Df Residuals:	2716
Method:	MLE	Df Model:	3
Date:	Mon, 31 Oct 2016	Pseudo R-squ:	0.07743
Time:	13:04:25	Log-Likelihood:	-1739.4
Converged:	True	LL-Null:	-1885.3
		LLR p-value:	5.506e-63

	coef	std err	z	P>|z|	[95.0%	Conf. Int.]
CREDIT SCORE	-0.0062	0.000	-13.109	0.000	-0.007	-0.005
CUR INT RATE	0.1909	0.044	4.315	0.000	0.104	0.278
CLTV	0.0340	0.003	12.323	0.000	0.029	0.039
DTI Ratio	0.0103	0.003	3.393	0.001	0.004	0.016

RECOMENDATION

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

After the mortgage crisis of 2008, the consumer habits have changed. They 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.

These financial institutes are also constantly running the risk analysis and stress test on their loan portfolio.

Using a model that can identify potential Non Performing loans gives these financial institution