**Midterm- Project**

**(Advances in Data Science)**

**TEAM 6  
- Priyal Chaudhari  
- Prashant Vikram Singh**

**FreddieMac Single Family Loan-Level Dataset**

**Process Flow:**

**Part 1**

Perform EDA

Scrap Sample Data

Create Summary Files

User Login

**Part 2**

Apply Classification

Algorithms

Combine Historical Files

User Login

Scrap Historical Data

Apply Regression Algorithms

**Part 1(a): Data Wrangling**

**Step 1:**

User should have filled the login form available on Freedie Mac [website](https://freddiemac.embs.com/FLoan/Bin/loginrequest.php) and have received the login credentials on his mail address.

**Step 2:**

Programmatically downloaded the sample dataset present on Freddie Mac [download url](https://freddiemac.embs.com/FLoan/Data/download.php). For this process to be successful, user needs to provide:

* Username
* Password
* Start Year
* End Year

Using the above information, downloaded the sample data for the year range provided and stored data inside the directory /*Midterm/Data/SampleData*

Code snippets are as follows:

* Log in process



* File link collection



* Download file and unzip

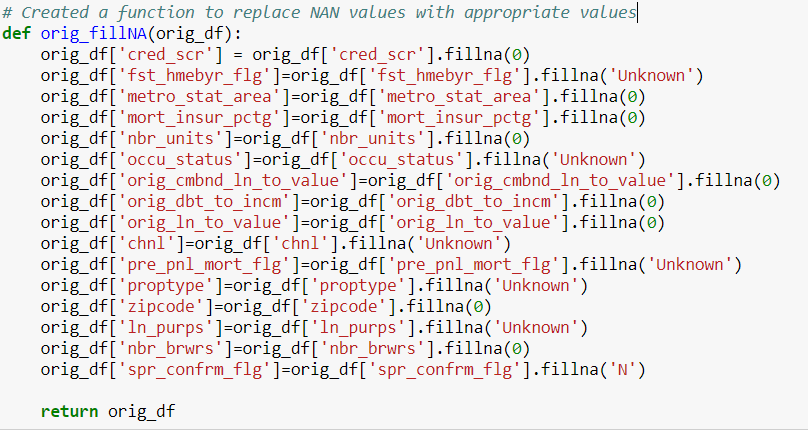


**Step 3:**

**a) Created Sample Summary File**

We followed below steps:

* Read all the files present under the directory /*Midterm/Data/SampleFiles* using os.path.walk() function.
* For each Origination files performed below steps:
  + Read the origination file using pd.read\_csv() into a DataFrame
  + Replaced NAN values with appropriate values using below technique:
    - Filled credit score with zero as they were providing values only between 301-850 and we cannot take average as the filling value as it will distort the distribution
    - Filled homebuyer flag with “Unknown” as they were giving it value of Yes/No so we filled NANs with “Unknown”
    - We followed the same approach for other fields as well, wherein we referred the “SF LLD - June 2017 User Guide .pdf” and filled the NANs with appropriate default values.
    - We didn’t used average values anywhere to fill the NANs as it will distort the distribution, which can impact our prediction and classification algorithms very severely.



* + Changed the datatype of few fields to integer to facilitate in aggregation process during EDA
  + Created a year column in the DataFrame to facilitate in the EDA process
  + Checked if the summary file is getting written for the first time or not. If writing for the first time, then added column names else skipped it.
  + Created the file using Dataframe.to\_csv() and used writing mode as ‘append’
* Code snippet is as follows:



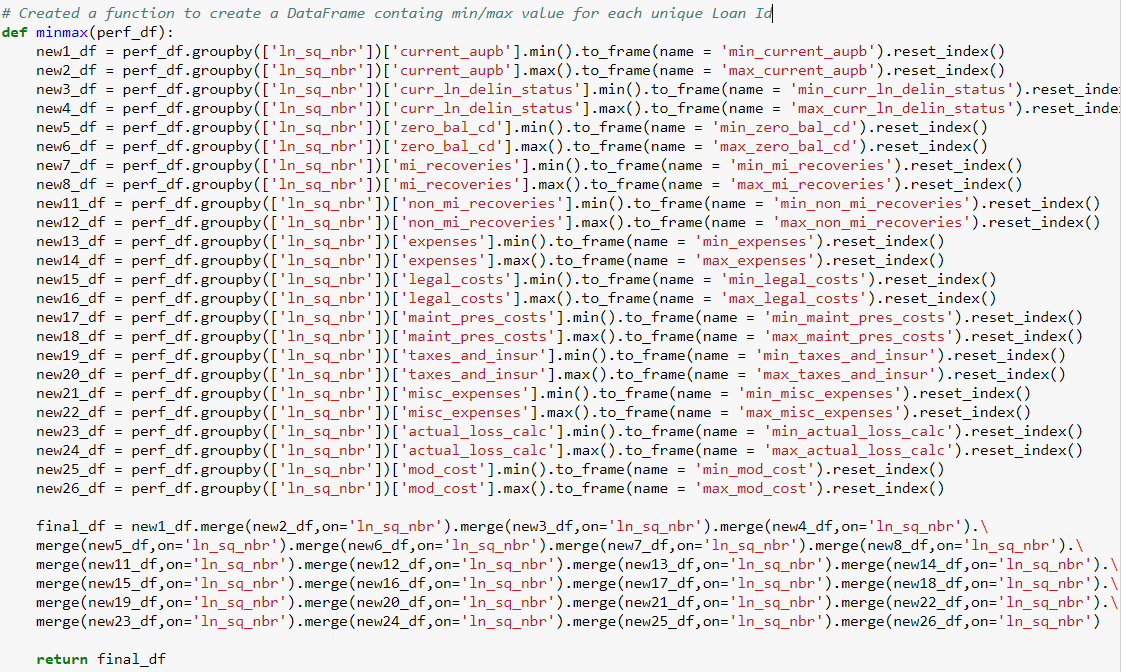
**b) Created Sample Performance File**

We followed below steps:

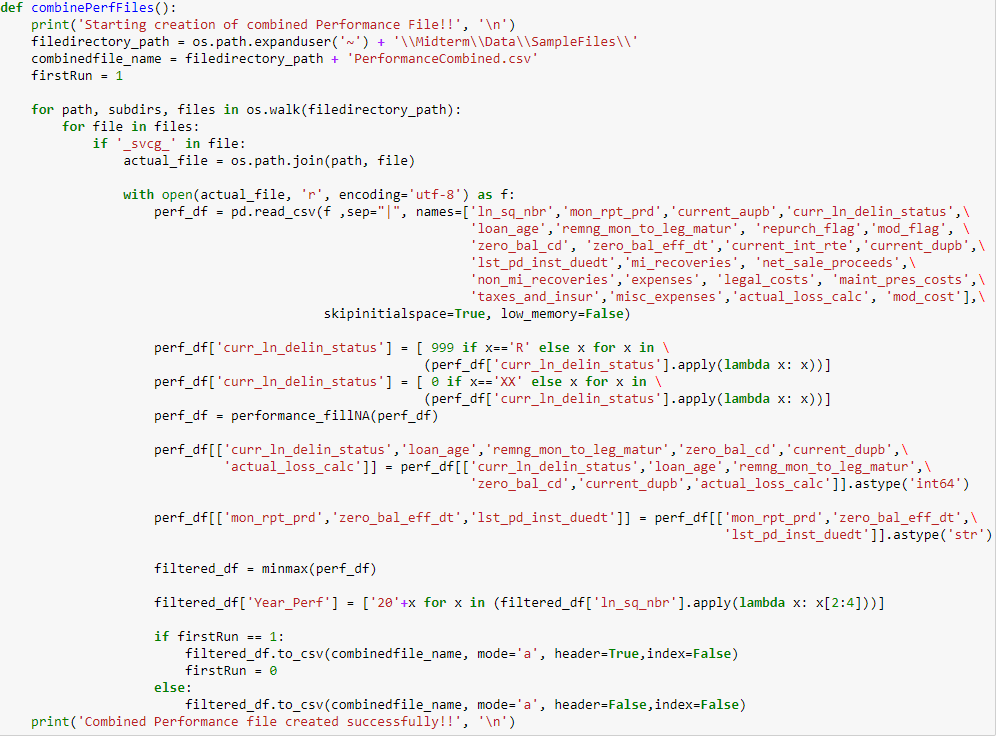
* Read all the files present under the directory /*Midterm/Data/SampleFiles* using os.path.walk() function.
* For each Performance files performed below steps:
  + Read the performance file using pd.read\_csv() into a DataFrame
  + Replaced NAN values with appropriate values using below technique:
    - As described above in the orig\_fillNA(), we followed the same approach here and filled the NANs with meaning values
    - We avoided filling NAN values with average values and referred “SF LLD - June 2017 User Guide .pdf”, and filled the NAN values appropriately



* + Changed the datatype of few fields to integer to facilitate in aggregation process during EDA
  + Created a function which returns min/max value for each unique Loan Id. The reason for doing this is:
    - We needed to join this summary performance file with the summary origination file, so we took unique loan sequence ids and then for each column calculated maximum/minimum values and created a file that contains unique loan sequence ids followed by maximum/minimum values of each column
    - We could have replaced maximum/minimum values with average values but we wanted to see the range for each column so we calculated maximum/minimum value
    - After joining this created file with the summary origination file, we can have a complete view of the data where for each unique loan sequence ids we can see contents from origination file as well as range through with the values in the performance file varies

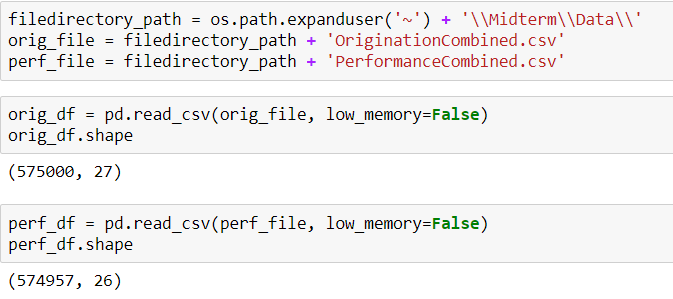


* + Created a year column in the DataFrame to facilitate in the EDA process
  + Checked if the performance file is getting written for the first time or not. If writing for the first time, then added column names else skip it.
  + Created the file using Dataframe.to\_csv() and used writing mode as ‘append’
* Code snippet is as follows:



**Part 1(b): Exploratory Data Analysis**

**Step 1:**

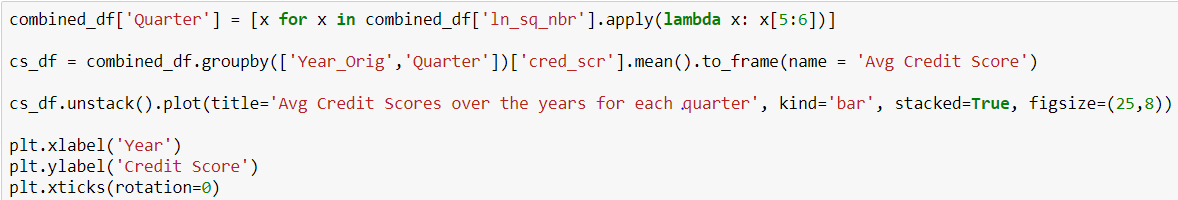
* Read the two summary files as:
* Joined the summary files to get the unified view as:

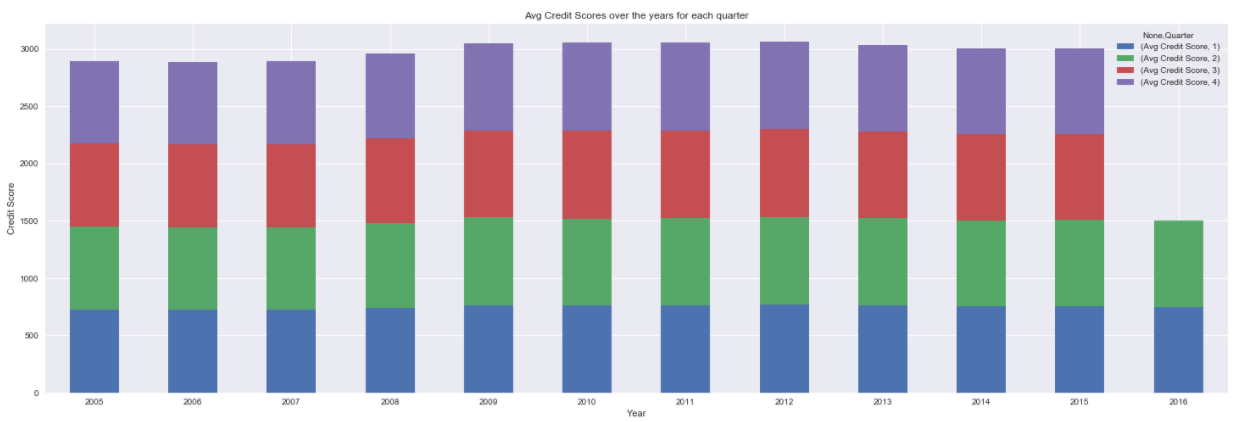


**Step 2:**

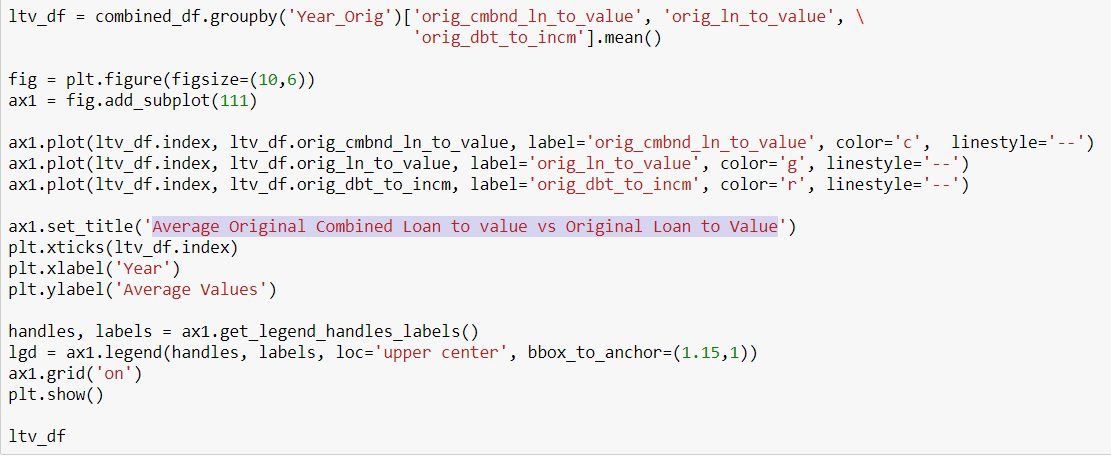
Plotted graphs to ease data visualization:

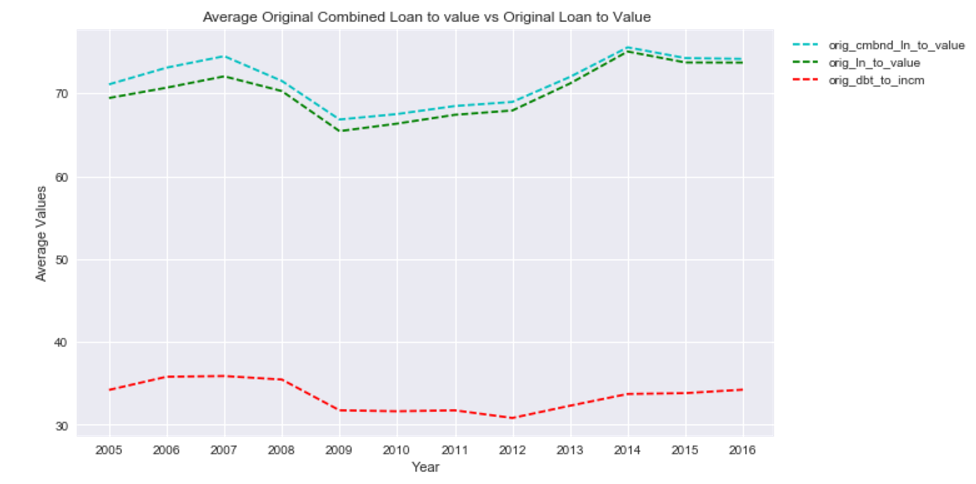
* Average Credit Scores over the years for each quarter
  + Cr



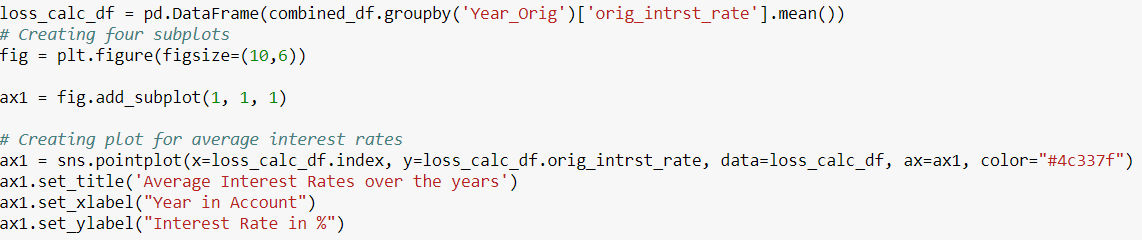
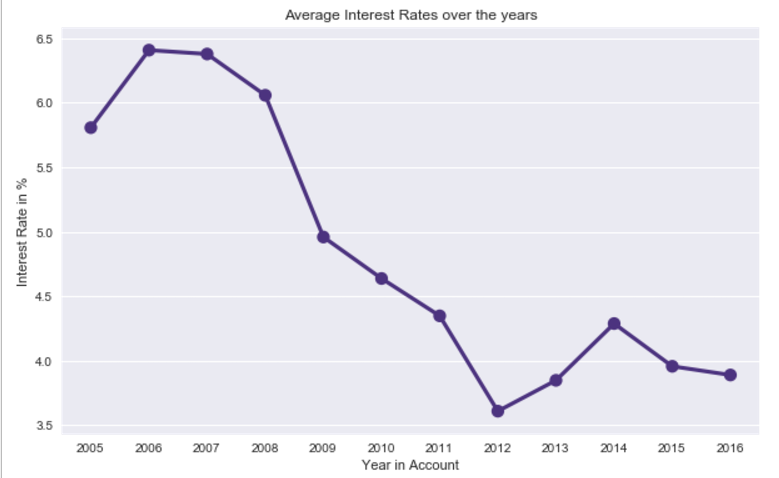


* Average Original Combined Loan to value vs Original Loan to Value

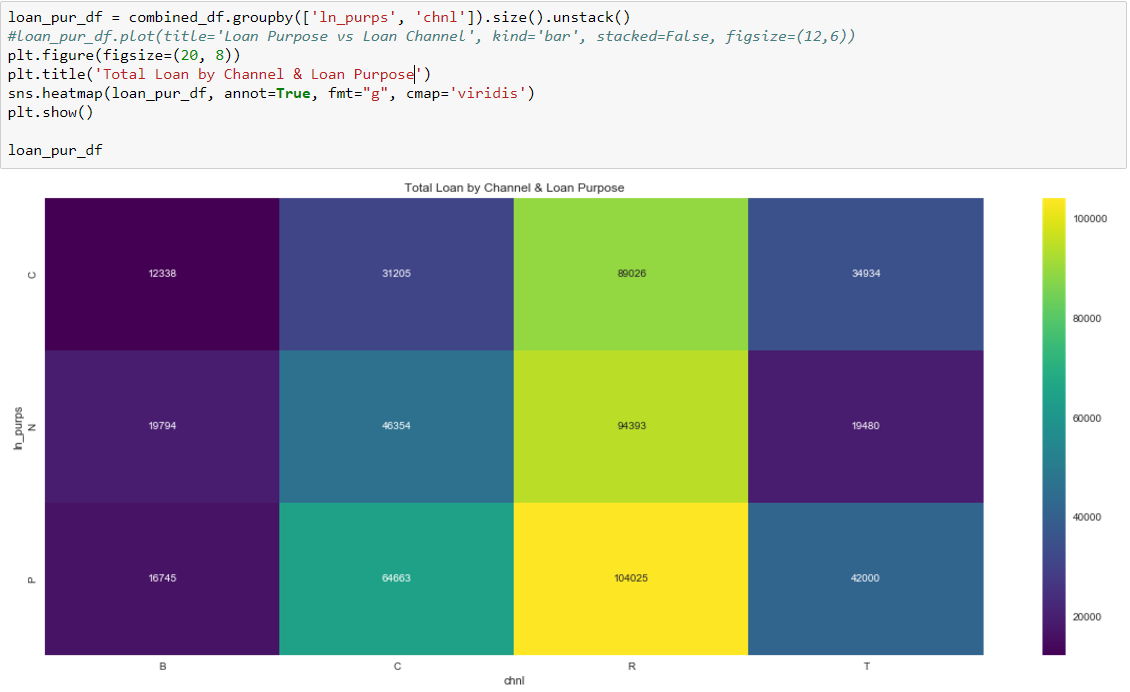




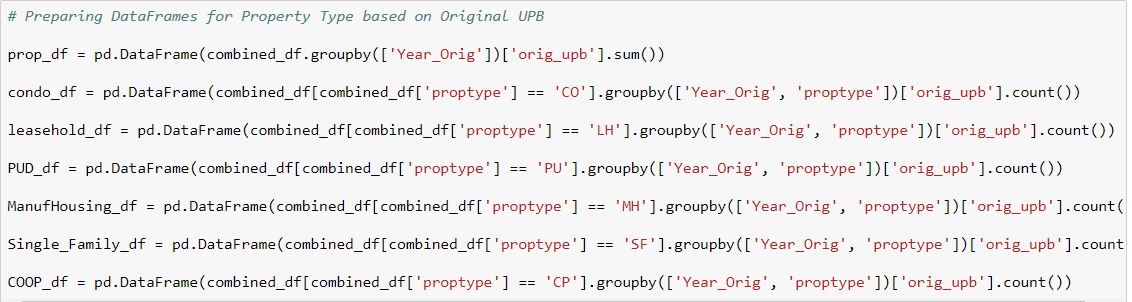
* Average Interest Rates over the years

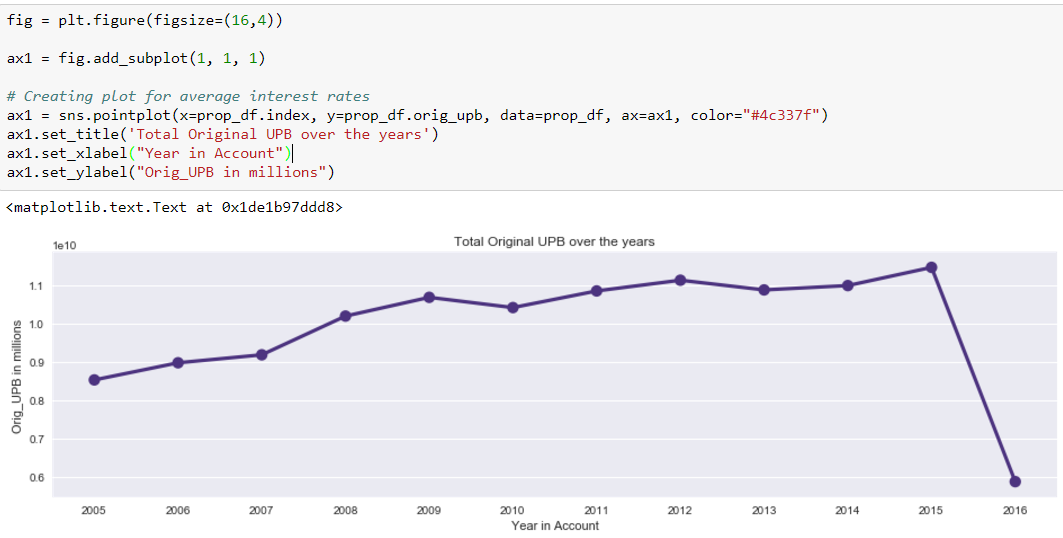
 

* Total Loan by Channel & Loan Purpose
  + Asas

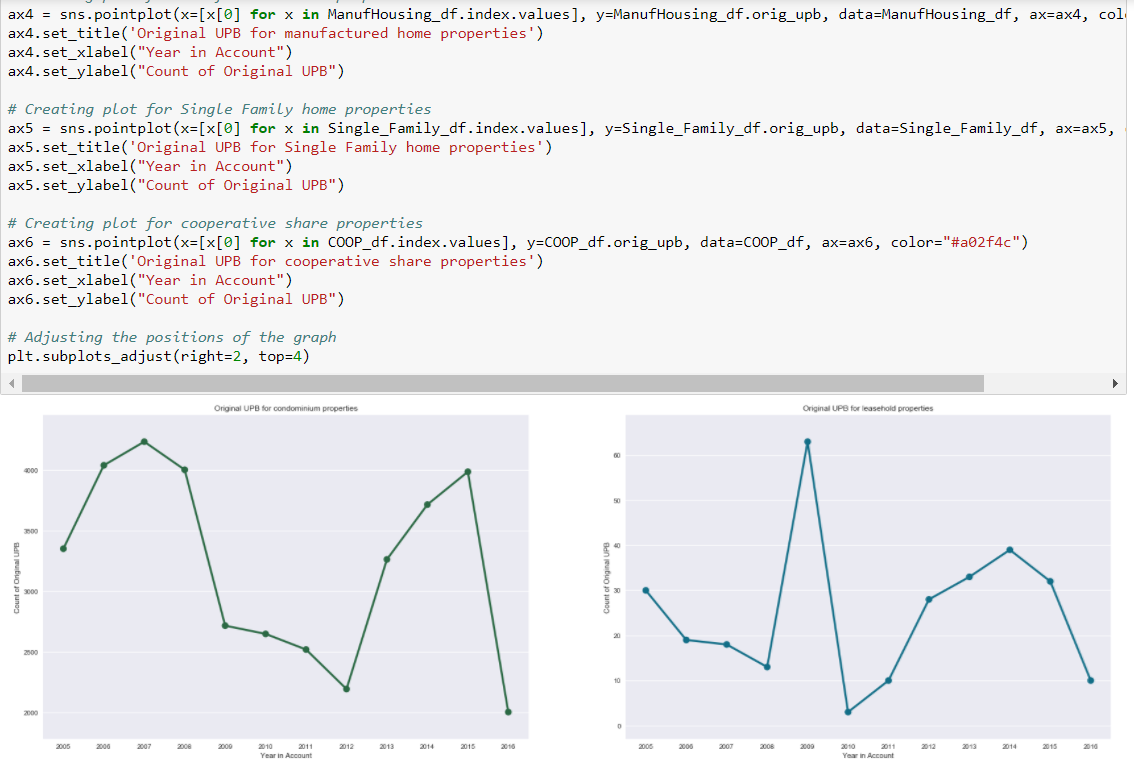


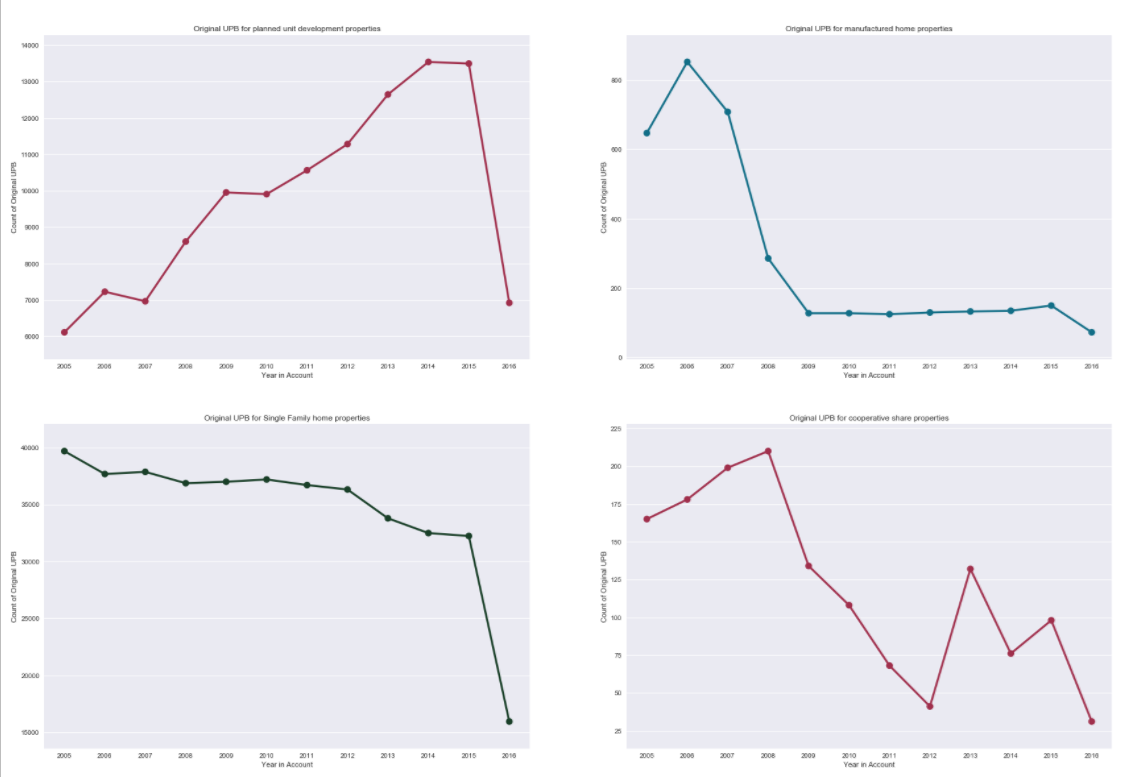
* Total Original UPB over the years
  + Sadadad



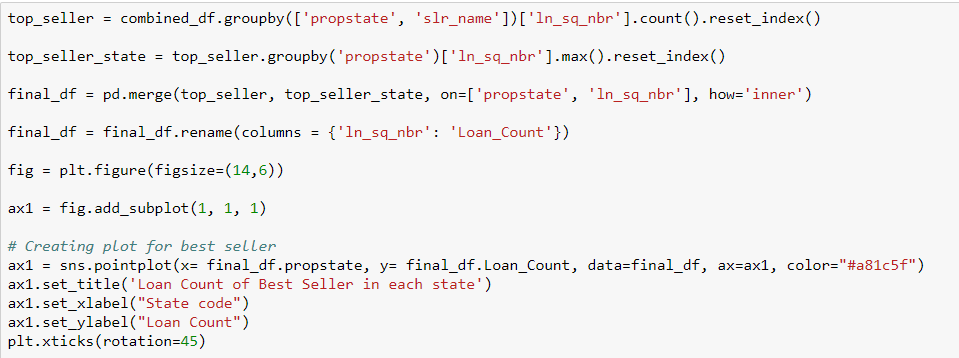


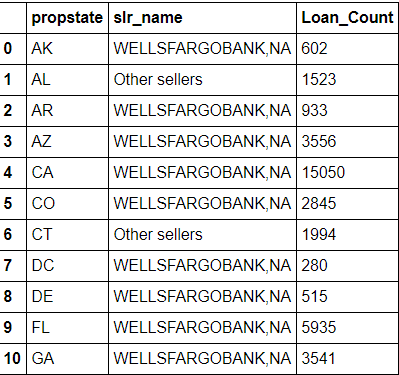


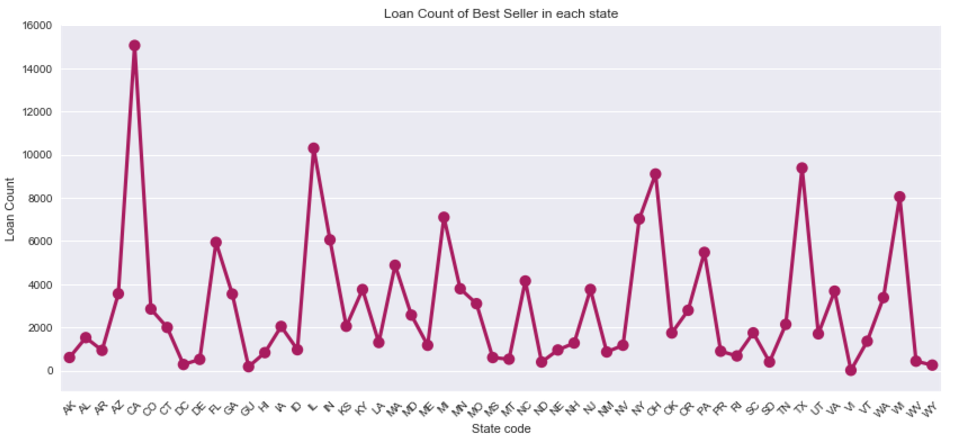




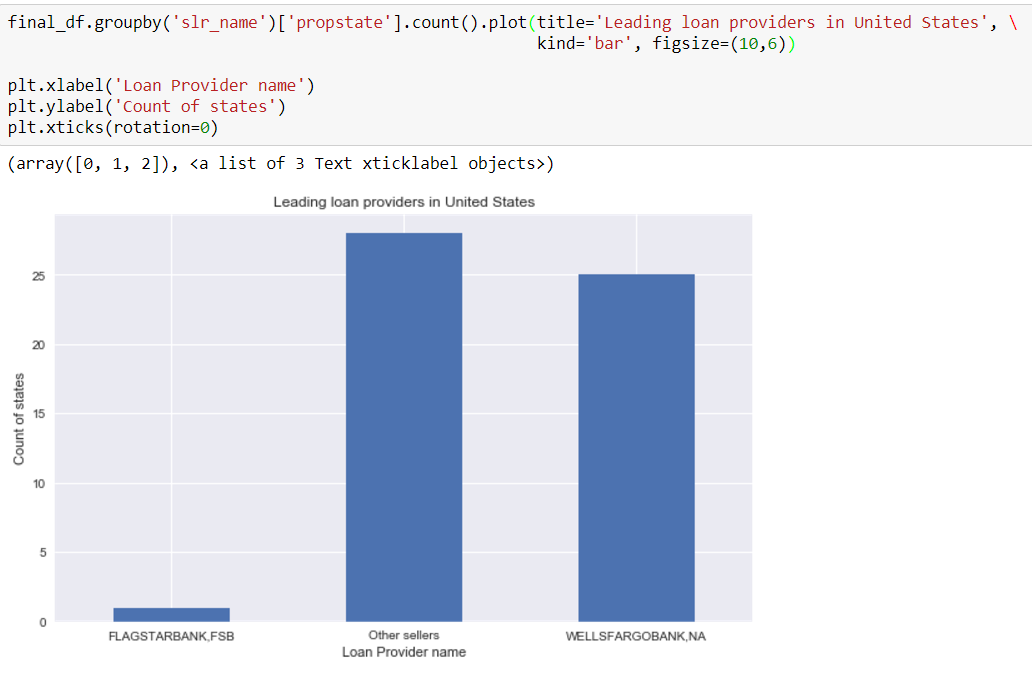
* Loan Count of Best Seller in each state







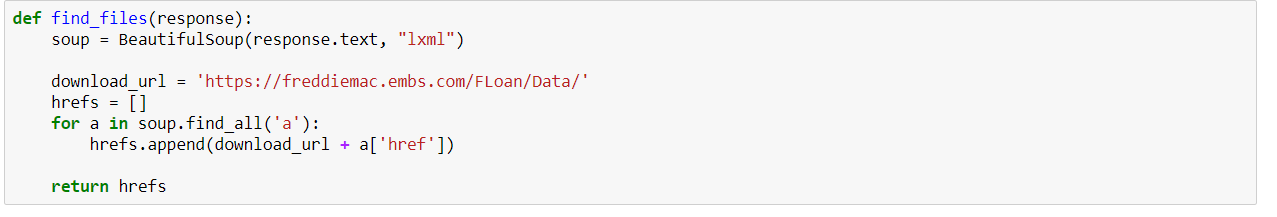
* Leading loan providers in United States



**Part 2: Building and Evaluating Models:**

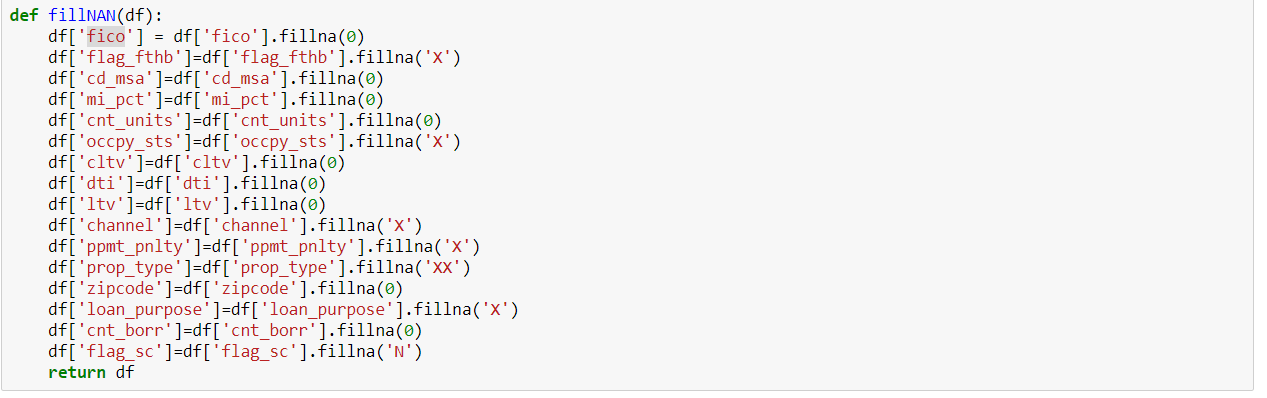
**Prediction (Interest Rate)**

* **We created a windows instance of EC2 to ease the running process of complex classification and prediction algorithm**
* Here we will use the historical origination data to predict the interest rate for quarters.
* Here we will use Q1 2005 as a training data and we will predict the values for Q22005 quarter.
* We will calculate and evaluate different algorithms based on below parameters:
* **MAE (Mean Absolute Error) -** In statistics, the mean absolute error (MAE) is a quantity used to measure how close forecasts or predictions are to the eventual outcomes.
* **RMSE (Root Mean Square Error) -** The root-mean-square deviation (RMSD) or root-mean-square error (RMSE) is a frequently used measure of the differences between values (sample and population values) predicted by a model or an estimator and the values observed. The RMSD represents the [sample standard deviation](https://en.wikipedia.org/wiki/Sample_standard_deviation) of the differences between predicted values and observed values. These individual differences are called [residuals](https://en.wikipedia.org/wiki/Errors_and_residuals_in_statistics) when the calculations are performed over the data sample that was used for estimation, and are called prediction errors when computed out-of-sample. The RMSD serves to aggregate the magnitudes of the errors in predictions for various times into a single measure of predictive power. RMSD is a good measure of [accuracy](https://en.wikipedia.org/wiki/Accuracy_and_precision), but only to compare forecasting errors of different models for a particular variable and not between variables, as it is scale-dependent
* The **mean absolute percentage error** (**MAPE**), also known as mean absolute percentage deviation (MAPD), is a measure of prediction accuracy of a forecasting method in statistics, for example in trend estimation.
* **Approach:**
* Creating the function and downloading the Input:
* We have created functions that would take hit the Freddie Mac website and download all the Historical file for our purpose. Based on the input Quarter, the file will be imported into python Data Frame. It went through following preprocessing before building model.
* First we downloaded the data form the Fredie macs website programmatically.





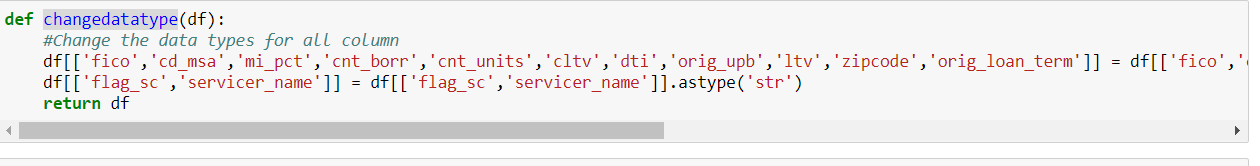
* After data is downloaded the file has some missing values . So missing values are handled.



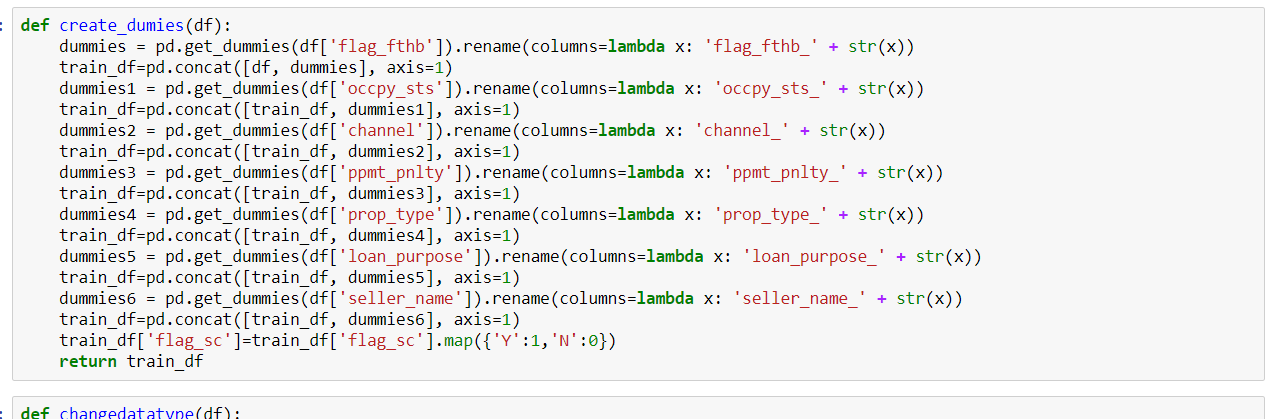
* Based on guideline given in readme files the values are handled.

1. Like for column “Credit Score” the values are in between 301 to 850. So anything below 300 or above 850 is considered ‘UNKNOWN’. So, we replaced the blanks by 0 so that calumn is converted to numeric and can be passed to algorithms.
2. In ‘First time homebuyerflag’ there are 2 valid values ‘Y’ ’N’ and unknown values are blank so we filled that blank with ‘X’ as an unknown.
3. Similarly we have handled the missing values of other columns.

* After filling missing values We have changed data type of columns to Integer to maintain the consistency in datatypes.



* After this step we have created dummies so that numerical data can be vanished and we convert each column to numerical data to provide to machine learning algo’s.



**Linear Regression:**

Regression analysis is a form of predictive modelling technique which investigates the relationship between a **dependent**(target) and **independent variable (s)** (predictor). This technique is used for forecasting, time series modelling and finding the casual effect relationship between the variables.

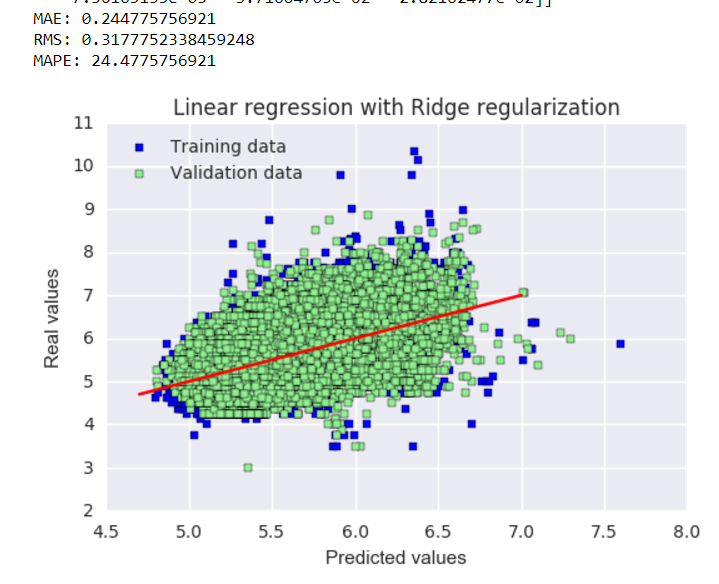
Linear Regression establishes a relationship between **dependent variable (Y)** and one or more **independent variables (X)** using a **best fit straight line** (also known as regression line).

It is represented by an equation **Y=a+b\*X + e**, where a is intercept, b is slope of the line and e is error term. This equation can be used to predict the value of target variable based on given predictor variable(s).

* **Implementation:**



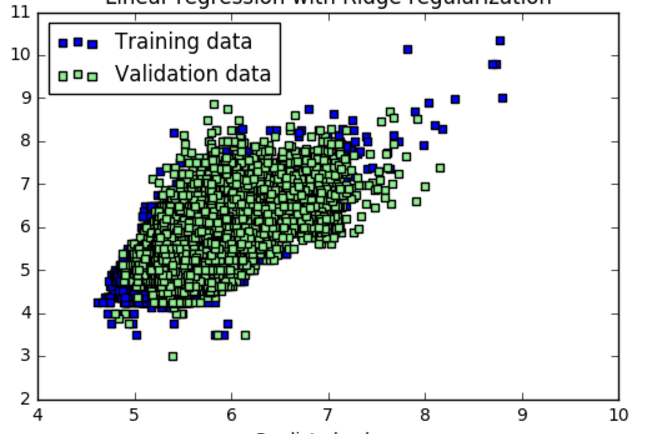
**Evaluation Parameters:**

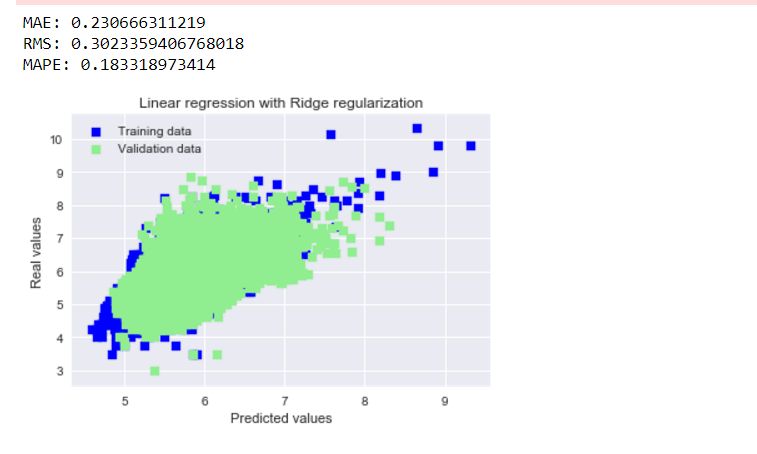


**Random Forest:**

**Random forests** or random decision forestsare an ensemble learning method for regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting mean prediction (regression) of the individual trees.

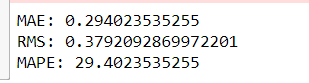
* We have used the randomForest Package and then used the two quarters to make the model and predict the value





**Neural Networks:**

An [artificial neural network](https://en.wikipedia.org/wiki/Artificial_neural_network) (ANN) learning algorithm, usually called "neural network" (NN), is a learning algorithm that is inspired by the structure and functional aspects of [biological neural networks](https://en.wikipedia.org/wiki/Biological_neural_networks). Computations are structured in terms of an interconnected group of [artificial neurons](https://en.wikipedia.org/wiki/Artificial_neuron), processing information using a [connectionist](https://en.wikipedia.org/wiki/Connectionism) approach to [computation](https://en.wikipedia.org/wiki/Computation). Modern neural networks are [non-linear](https://en.wikipedia.org/wiki/Non-linear) [statistical](https://en.wikipedia.org/wiki/Statistical) [data modeling](https://en.wikipedia.org/wiki/Data_modeling) tools. They are usually used to model complex relationships between inputs and outputs, to [find patterns](https://en.wikipedia.org/wiki/Pattern_recognition) in data, or to capture the statistical structure in an unknown [joint probability distribution](https://en.wikipedia.org/wiki/Joint_probability_distribution) between observed variables



**Conclusion:**

Based on the evaluation parameters, Random Forest Algorithm is more efficient and it produces less error % as compared to other Prediction algorithms.

**What IF Analysis**

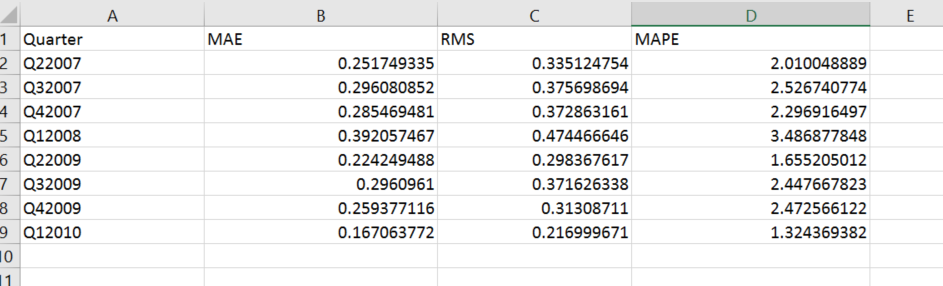
**Financial Crisis Analysis**:

* In the fall of 2007, Freddie Mac shocked the market by announcing large credit-related loses, fueling the fire for the argument that the two companies pose a tremendous risk to the entire financial system.

(<http://www.investopedia.com/articles/economics/08/fannie-mae-freddie-mac-credit-crisis.asp>)

* The Federal Home Loan Mortgage Corporation (Freddie Mac) announced that it will no longer buy the most risky subprime mortgages and mortgage-related securities.
* In July 24, 2007 Countrywide Financial Corporation warned of “difficult conditions.” This is evident from the Q32007 Testing measures as the difference between Training and Testing RMSE increased substantially by around 16%.

**Table(Prediction Parameters:)**



**Findings**:

1. There was linear increase in MAPE for year 2007, with highest shift between Q4 and Q1 of 2008.
2. The model showed no high deviation for the quarters.
3. There was a similar increase in RMSE for year 2009, in quarter Q1 and Q2.
4. The model deviated from its prediction/accuracy values for the quarter Q1 and Q2 for 2009.
5. **Ran Neural net Regression Model for every quarter for economic boom (1999 and 2013)**.

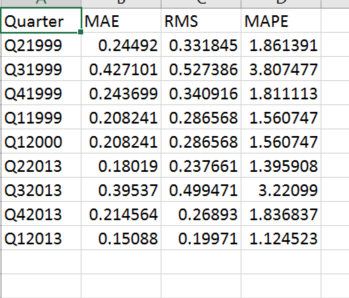
**1999:**

The easing of credit also coincided with spectacular stock market run-ups from 1999 to 2000

Freddie Mac financed homes for more than 2 million families and achieved record earnings per share of $2.96, an increase of 28 percent over 1998.

In 2013,

Mortgage rates peaked at 4.6% in August and have held steady since September and several accounting events had significant impacts on the Enterprises’ reported financial results. Fannie Mae and Freddie Mac reported levels of 2013 net income are greater than at any prior time in their respective histories. Their historically high net income was driven by reversals of previously accrued losses associated with deferred tax assets (DTA) and their allowance for loan and lease losses (ALLL)—plus revenue from legal settlements of representation and warranties claims and lawsuits regarding private-label securities that the Enterprises purchased as investments. FHFA does not expect benefits of this nature to be repeated in future years and does not expect the 2013 levels of net income to be approached anytime in the foreseeable future.



**CLASSIFICATION:**

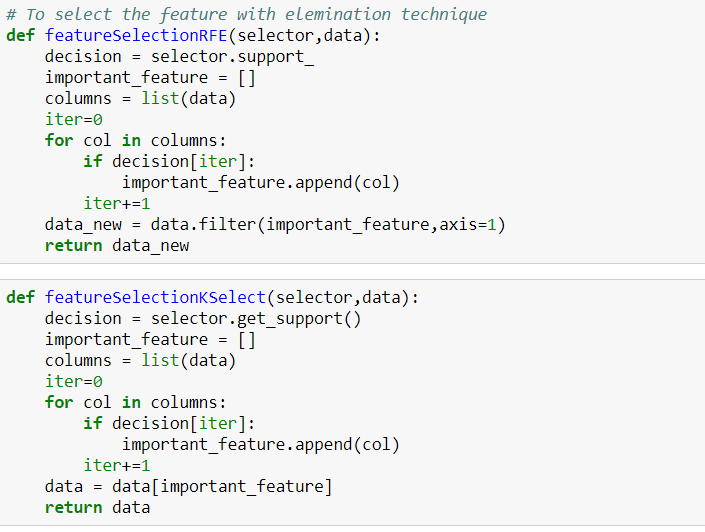
The main goal of classification is to predict the target class (Yes/ No). If the trained model is for predicting any of two target classes. It is known as binary classification. Here we are predicted the derived column Delinquent which is the target class.

We Programmatically downloaded files from the freddiemac website. The input is parameterized. The user provides two inputs one for test data and the other for train data. We have built four models namely: Random Forest, Neural Network, SVN and Logistic Regression.

FEATURE SELECTION

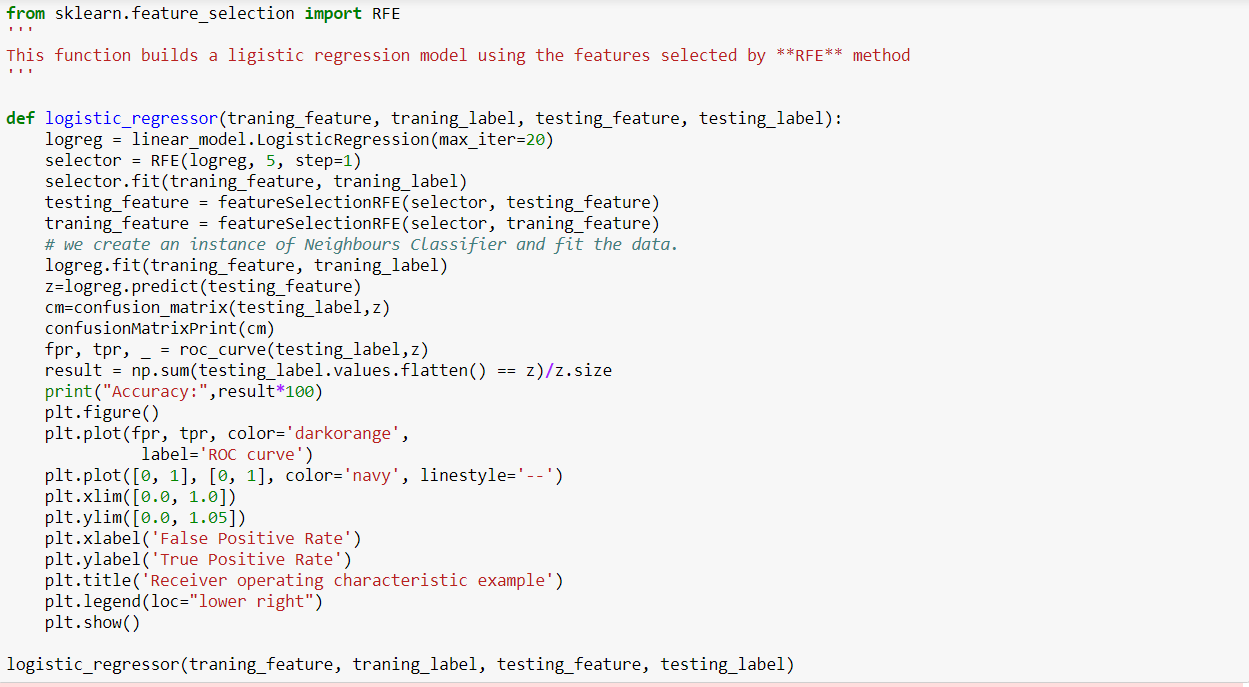
Before proceeding with our models, we have done best feature selection using three algorithms.

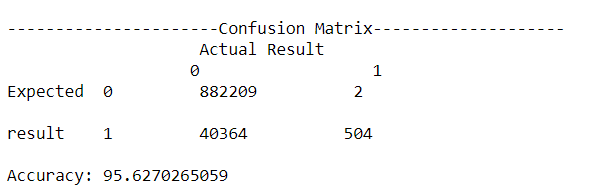
The best features that add to the predictive power of the model and irrelevant features removed from the model. We implemented following feature selection techniques in Python:

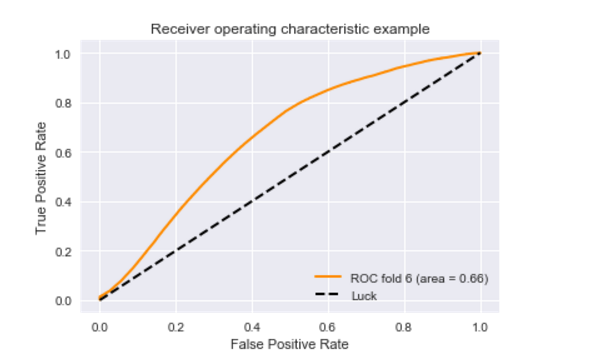


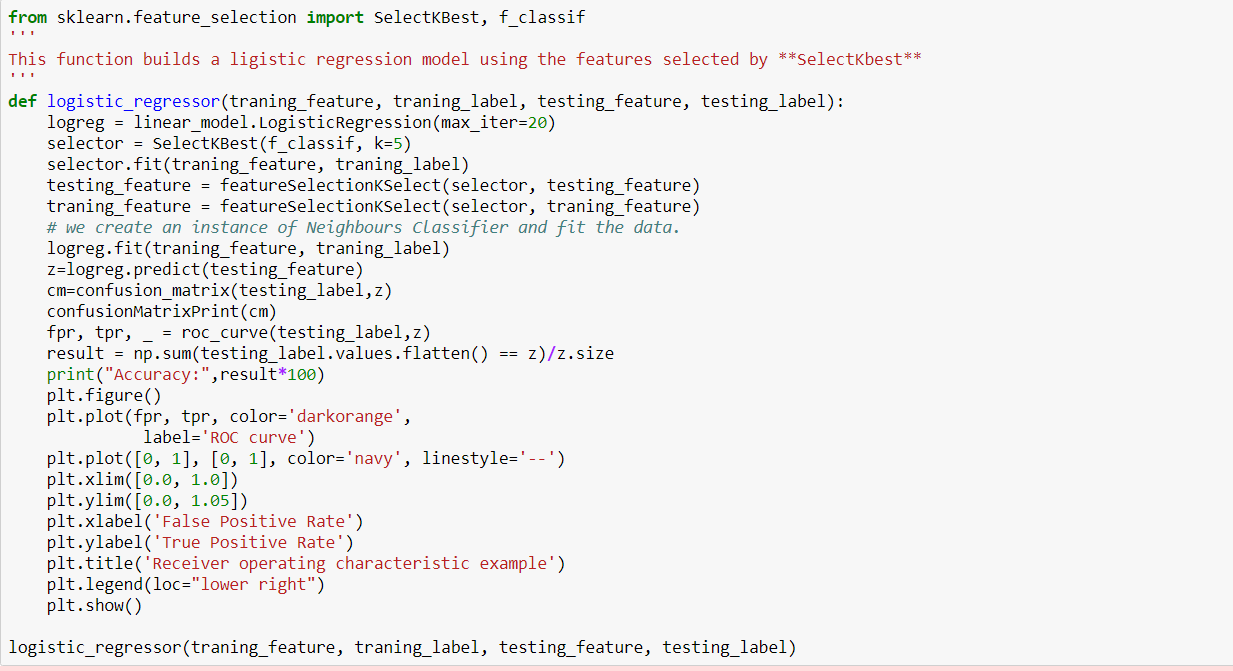
LOGISTIC REGRESSION

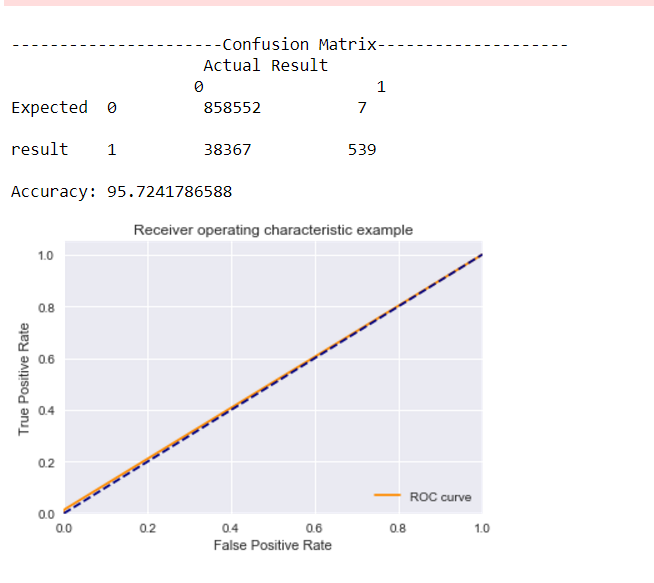
Binary Logistic Regression is a special type of regression where binary response variable is related to a set of explanatory variables, which can be discrete and/or continuous. We are using the logistic regression model for training the model for the quarter supplied and predicting the delinquency status based on the trained model.





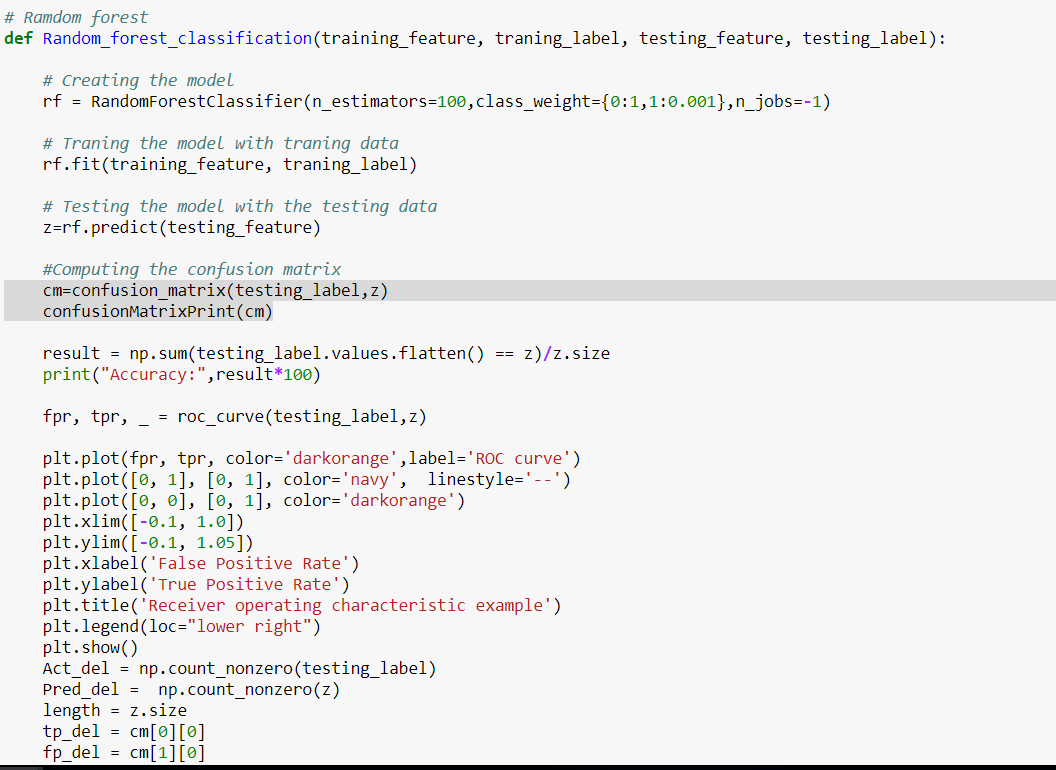


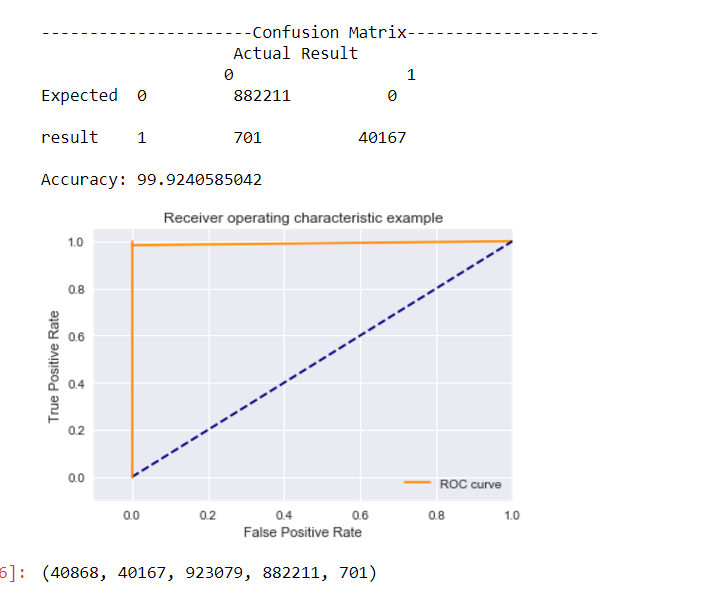




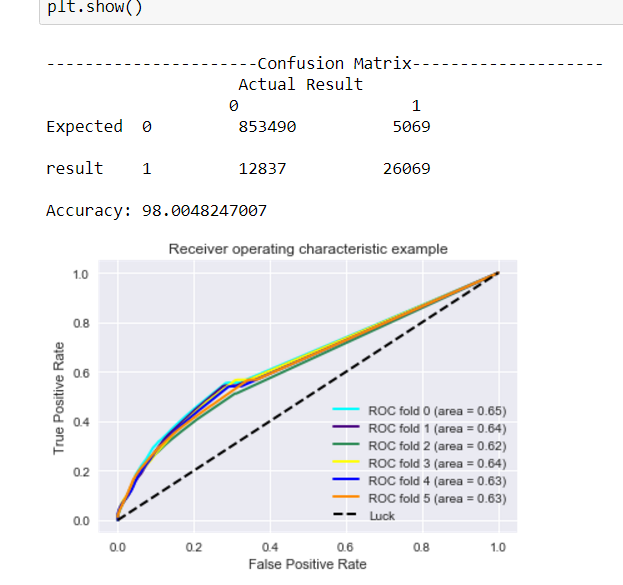
RANDOM FOREST

Random Forests grows many classification trees. To classify a new object from an input vector, put the input vector down each of the trees in the forest. Each tree gives a classification, and we say the tree "votes" for that class. The forest chooses the classification having the most votes (over all the trees in the forest).





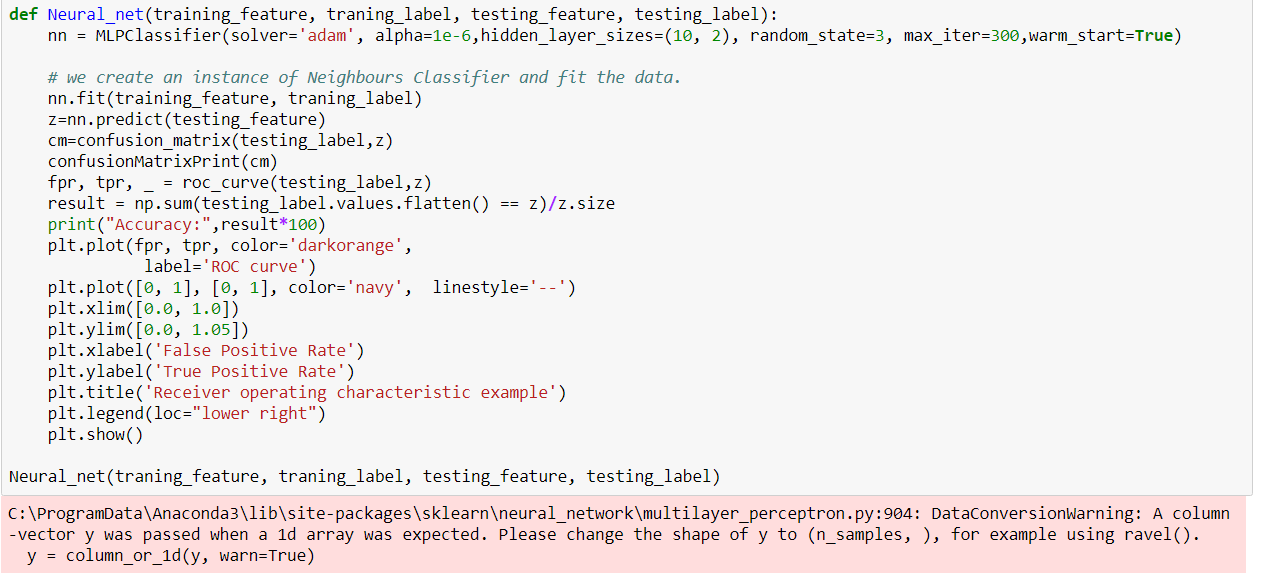


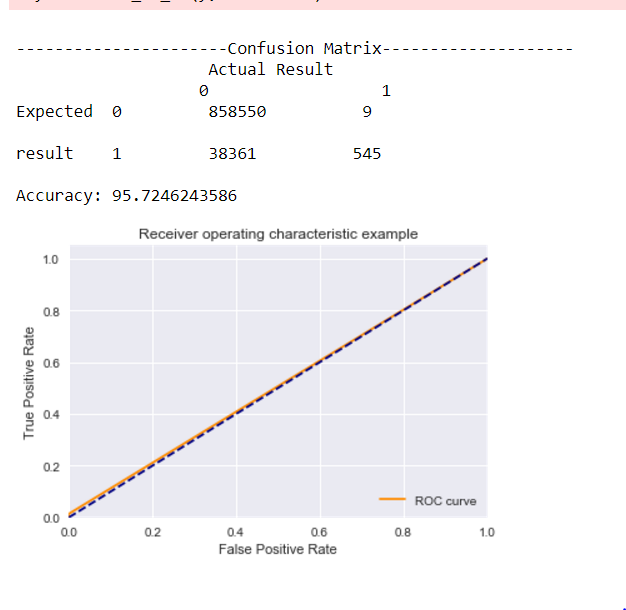


NEURAL NETWORK

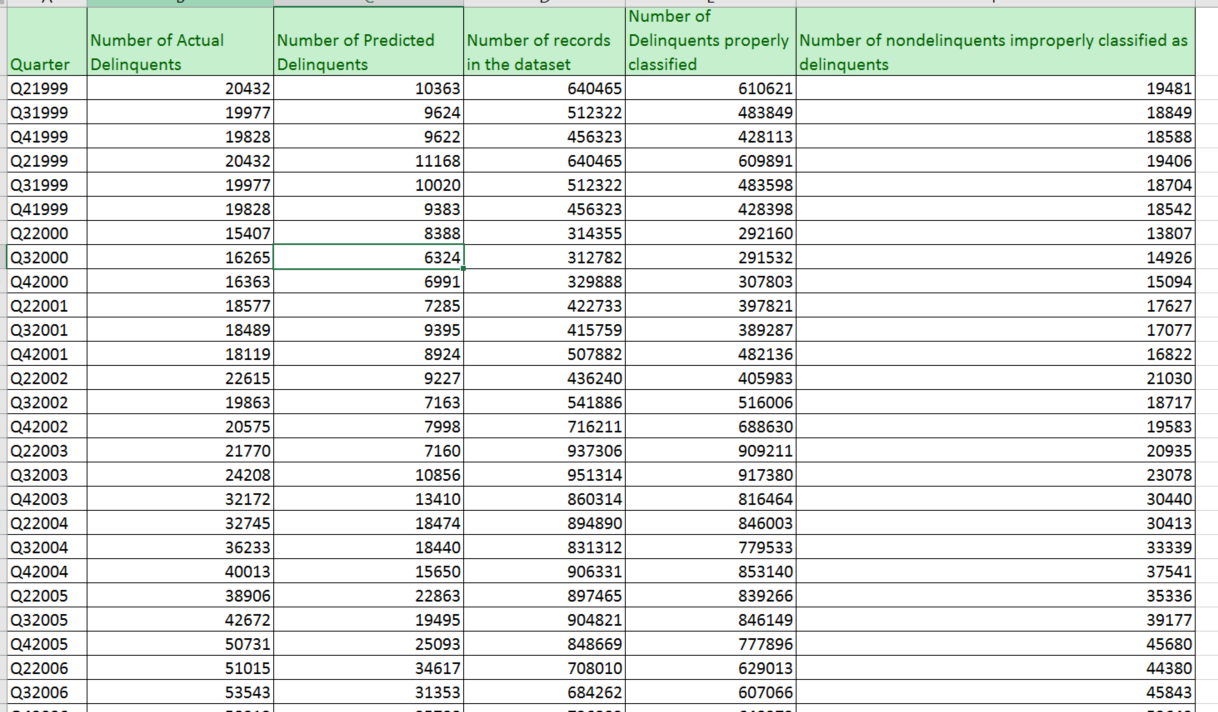


Artificial neural networks are relatively crude electronic networks of neurons based on the neural structure of the brain. They process records one at a time, and learn by comparing their classification of the record (i.e., largely arbitrary) with the known actual classification of the record.





**RUNNING FROM 1999 to 2013 for calculating following matrix.:**



**Conclusion:**

Based on accuracy, Random Forest has the maximum accuracy as compared to other classification algorithms.

* Random Forest is fast to build, even faster to predict and much more customizable
* Data doesn’t need to be rescaled, transformed or modified to outliers
* It is resistant to over training

According to the results, we feel that Random forest model has a good quality:

* The area in which it can be improved is that, we can test it on some random data as the dataset that we currently have biased values.
* So, based on classification accuracy, we will recommend using Random Forest algorithm for next quarter.