Mid Term ReportSingle Family Loan Data Analysis

short line

Instructor- Srikanth Krishnamurthy

TA- Pranjal Jain

Team 9 -Tapadyuti Maiti, Khushbu Parekh

June, 2017

# Problem Statement

The United States Department of Housing and Urban Development is a Cabinet department in the Executive branch of the United States federal government. As a Data Scientist, we must understand the US housing market. We are given Single-Family Loan Data to analyze and asked to present our results. Then, using the datasets, build predictive analytics models and do the classification

Dataset:: Link for Single-Family Loan Dataset:<http://www.freddiemac.com/news/finance/sf_loanlevel_dataset.html>

# 

# Part 1: Data Wrangling

1. Data Download

Programmatically download the data: We are using Requests package of Python to save the Session Instance.The Session object allows us to persist certain parameters across requests.

The login page gets routed to the link ending with “auth.php” , So we are first routing with valid Username and password.

Few Insights of the method of the request that we are using :   
Any dictionaries that you pass to a request method will be merged with the session-level values that are set. The method-level parameters override session parameters.

with requests.Session() as s:

This will make sure the session is closed as soon as the with block is exited, even if unhandled exceptions occurred.



**Challenges Faced :**

1.Open method would not work to extract such a bigger file >1 GB

Solution:Used Bytes.IO

2.While Testing we faced repeatedly deactivation of our credentials due to excessive download activity

Solution: a.Wait for 30 min to get reactivate or get a new credentials

How to Avoid the situation : We tested by giving break condition , it helped to loop only one time to avoid this circumstance, but still you would need 2-3 ids to test it properly

# **Reference:**

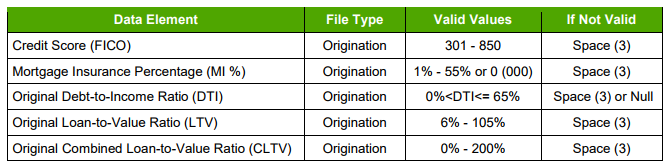
https://www.crummy.com/software/BeautifulSoup/bs3/documentation.html#The basic find method: findAll(name, attrs, recursive, text, limit, \*\*kwargs)

<https://docs.python.org/2/library/io.html>

<http://homeguides.sfgate.com/meaning-ownership-interest-real-estate-99933.html>

# **Data PreProcessing**

Important points to be considered for NULL values appearing on column :



Column 3 - FIRST TIME HOMEBUYER FLAG -

Investment Properties, Second Homes and Refinance transactions - NULL

Column 5 - METROPOLITAN STATISTICAL AREA (MSA) OR METROPOLITAN DIVISION -

Codes are available on the below sites:

<https://www.census.gov/population/estimates/metro-city/0312msa.txt>

NULL - Not in MSA or MD

Column 6 - MORTGAGE INSURANCE PERCENTAGE (MI %)-

<1 % $$ > 55 % - 000

Only primary mortgage insurance that is purchased by the Borrower, lender or Freddie Mac is disclosed. Mortgage insurance that constitutes “credit enhancement” that is not required by Freddie Mac’s Charter is not disclosed

Column 7 - NUMBER OF STOREY

Column 10 (DTI) RATIO Lesser is better

Column 12 - LTV - higher value is risk , ideally CLTV must be greater than LTV

Few Important Abbreviations:

**Real estate owned** or REO is a term used in the United States to describe a class of property owned by a lender—typically a bank, government agency, or government loan insurer—after an unsuccessful sale at a foreclosure auction

**UPB**-Unpaid Principal Balance

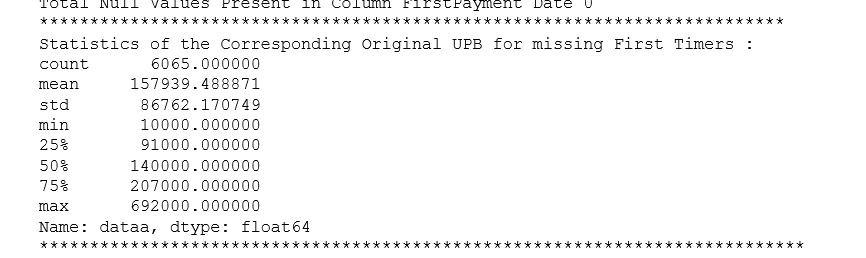
We have carefully filled the values by taking care of the above conditions :

**Original File Data Missing Handling**

We have filled the below Columns as :



We did Analysed how the data is present Original UPB for the First Timer Flag :



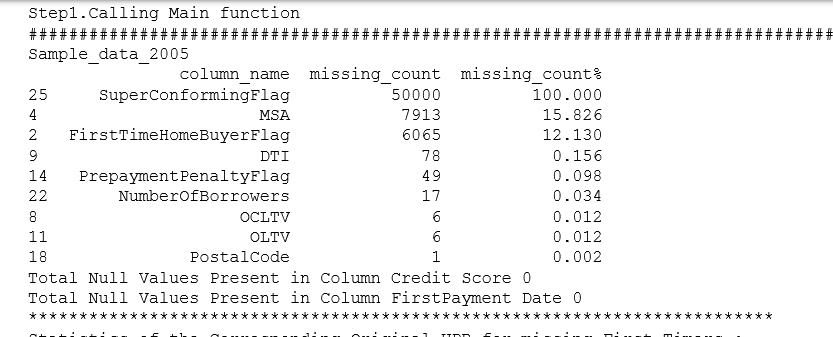
**Data Cleaning Summary Original File :**



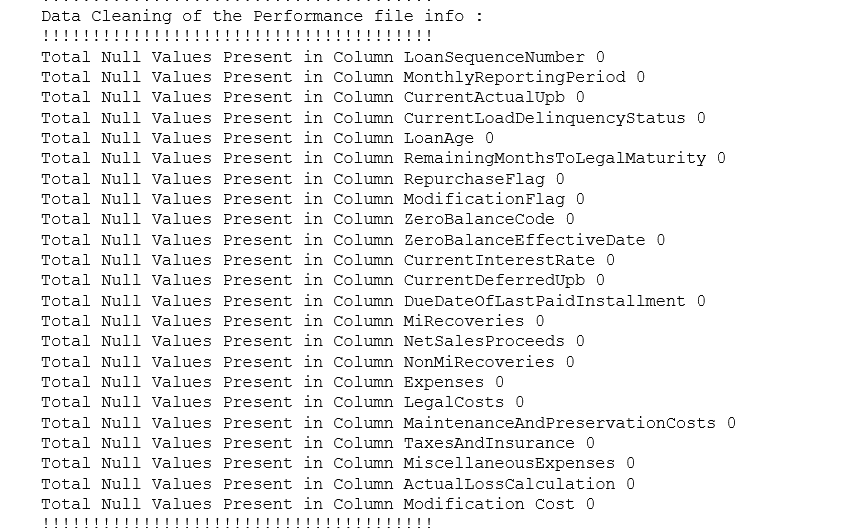
**Performance Data Missing Handling**

We have downloaded All the Sample Files from 2005 and onwards the Freddie Mac dataset :

Few Dimensions about the Null values present in the given dataset :



**Data Cleaning Summary for the Performance File :**



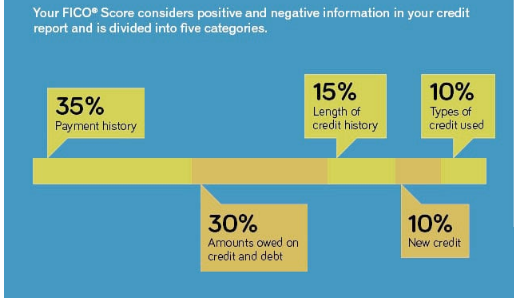
Reference:

<http://www.experian.com/blogs/ask-experian/infographic-what-are-the-different-scoring-ranges/>

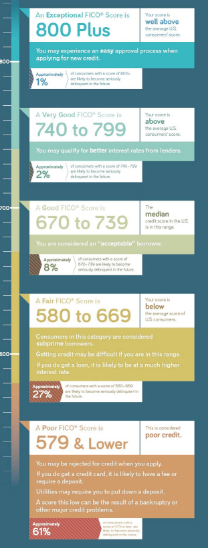
https://chrisalbon.com/python/pandas\_binning\_data.html

How the Credit Fico Scores Affects is represented in the diagram:

FICO Score Binning :



Few insights of the Fico Score , how the the Score decides the credibility :



# Exploratory Data Analysis :

We have not taken in account of the below Columns :

As we know **Actual Loss** = (Default UPB – Net Sale\_Proceeds) + **DelinquentAccrued Interest**

- Expenses – MI Recoveries – Non MIRecoveries

And *Delinquent Accrued Interest* = (Default\_Upb – Non Interestbearing UPB)\*

(Current Interest rate – 0.35) \*

( Months between Last Principal &

Interest paid to date and zero balance date ) \*30/360/100

So We have all the DAta information related to Loss that are included in the Actual Loss Calculation Column. SO we are won't be considering the below Columns:

**Expenses** = Sum(Legal Costs, Maintenance and Preservation Costs, Taxes and Insurance,Miscellaneous Expenses)

**NetSalesProceeds**

**MiRecoveries**

**NonMiRecoveries**

**Legal Costs**

**Maintenance and Preservation Costs**

**Taxes and Insurance**

**MiscellaneousExpenses**

WE are not accounting the information for the **ZeroBalanceEffectiveDate** Column as we saw

Starting date of Zero Balance Effective Date is same data as that of ending **MonthlyReportingPeriod** , as when the loan ends and no records are kept after that .

**CurrentDeferredUpb** - This is the Non-interest bearing UPB , which is used to caclulate DAI , and inturn the Actual Loss

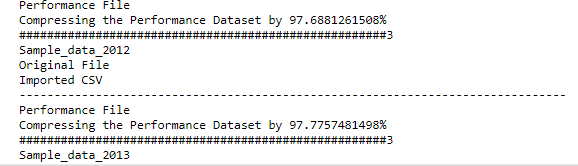
**DueDateOfLastPaidInstallment** - WE are already taking consideration by taking Column RemainingMonthsToLegalMaturity, in our case both represent the same entity

**RepurchaseFlag** - AS this will be indicated on the Zero FLag , how the Loan has ended by Code 6

**ModificationFlag** - This is related to the mortgages with loan modifications,indicates that the loan has been modified and it happens only when the loan has been repurchased , which have taken care of as mention :

.

WE have compressed the Performance File, getting only the important pramenters.

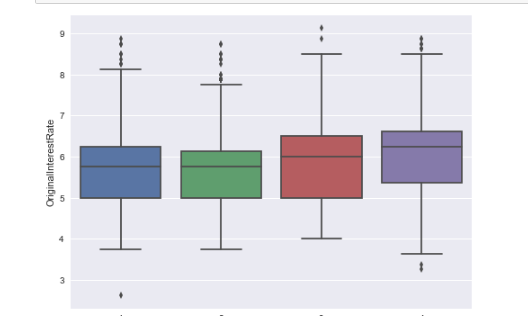


THis helped in performing the analysis .

# Summary:

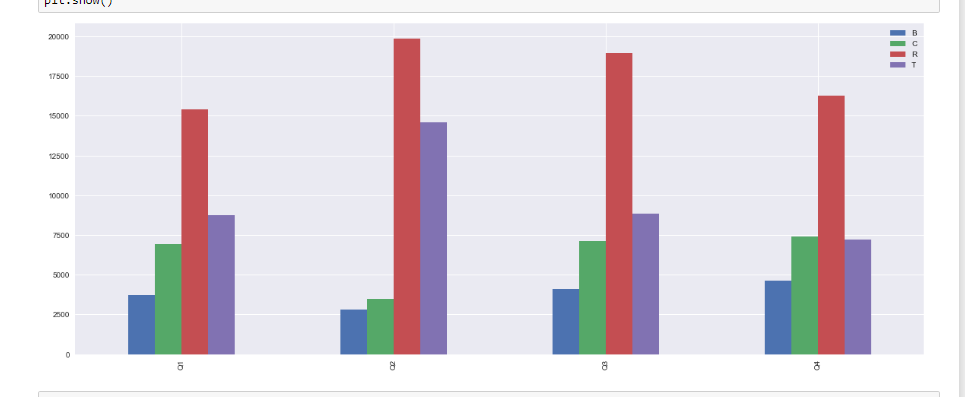
**Quartely Data 2007 - 2008 - 2009**

**Interest Rate - Quarterly Distribution - 2007-2008-2009**

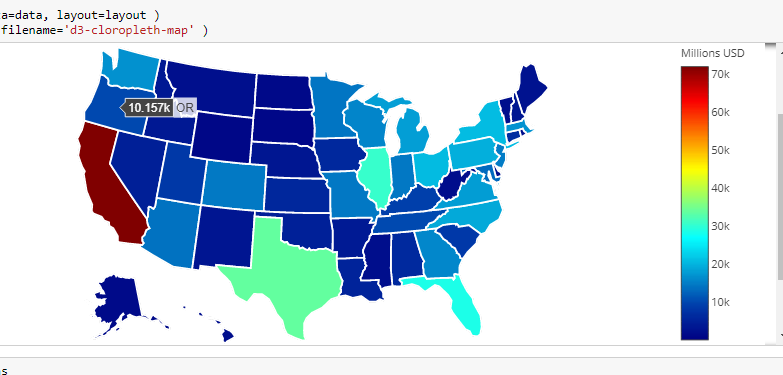


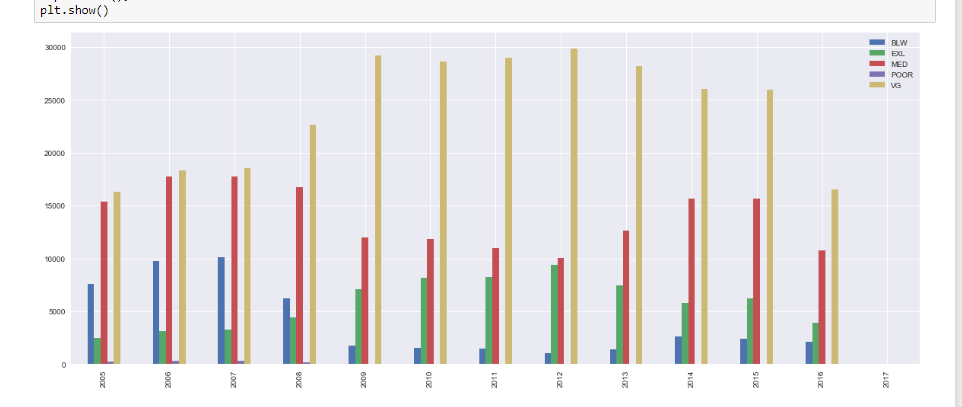
**Total Number of Loan SActioned - Quarterly 2007-2008-2009**

**Quartely Channel of Mortgage Loans Sanctioned**

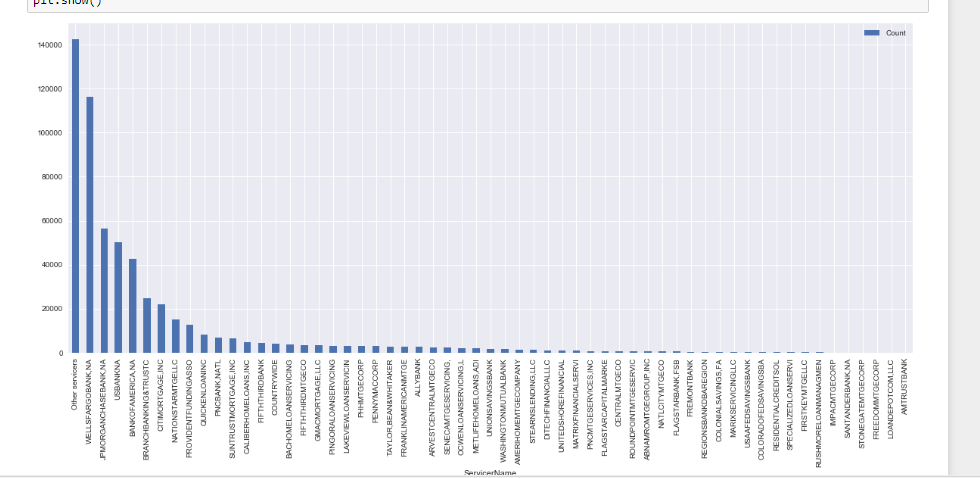


**Summary of the Different Type of AREa Wise**

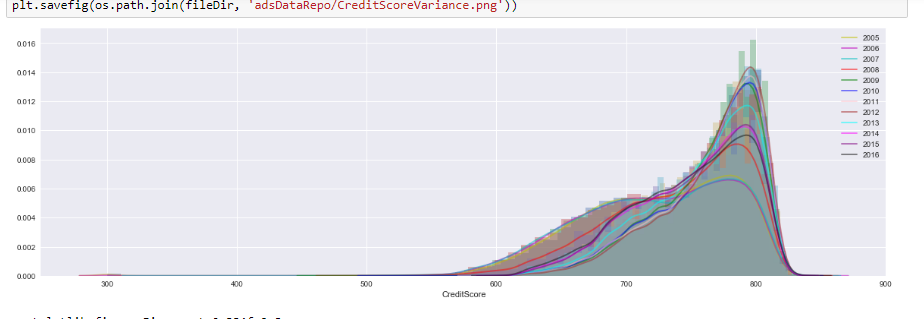




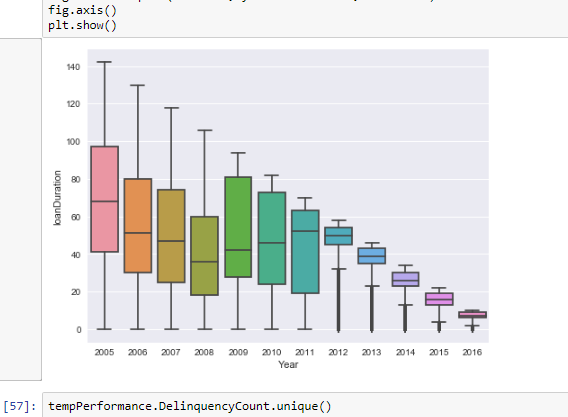
**Company wise statistics**



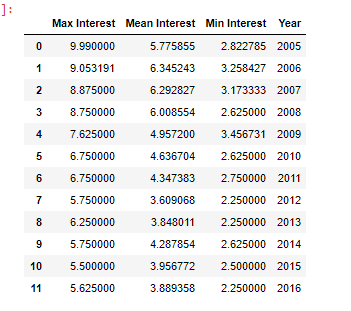
# Performance File Summary :

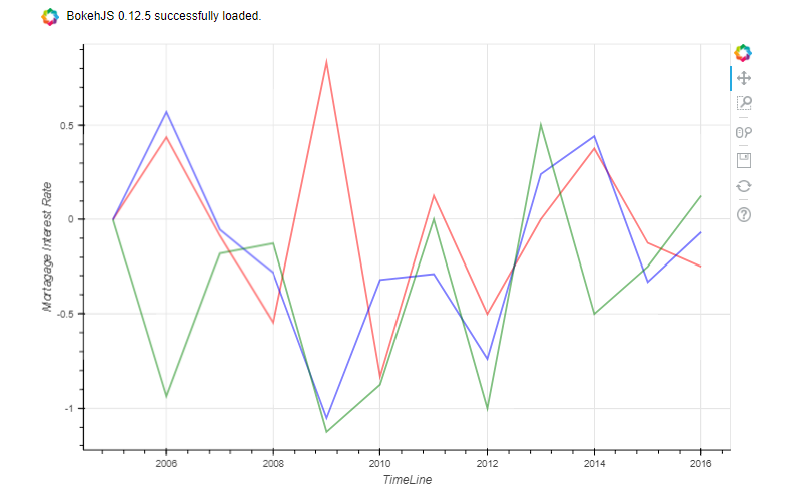


**Variation of Loan Credit Score across the years**



**Loan interest with Time Series**





**VAriance of Interest - Min red- Mean blue - Max green**

# Part 3 Classification

Following results and graphs are for the balanced Dataset where we have pruned the unwanted data so that dataset becomes balanced to make our model better . upto 99%

WE have also done the regular data Classification on all the three Algorithm.   
The best one is Random Classification.

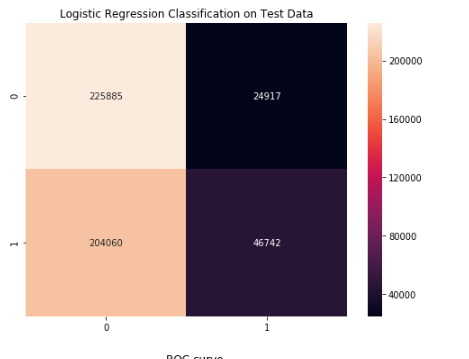
Ability to handle data without preprocessing

data does not need to be rescaled, transformed, or modified resistant to outliers

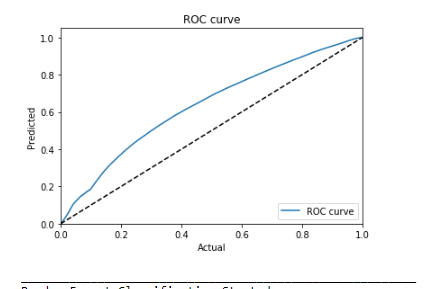
automatic handling of missing values

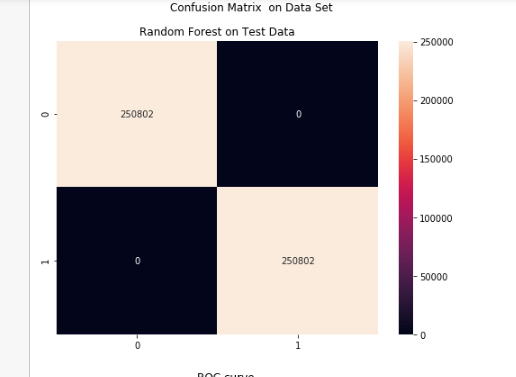
Cluster identification can be used to generate tree-based clusters through sample proximity

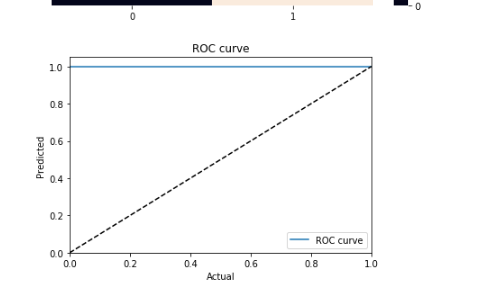
We have also added logic to keep on increasing the accuracy ,by add data on the previous .

Logistic Regression : 

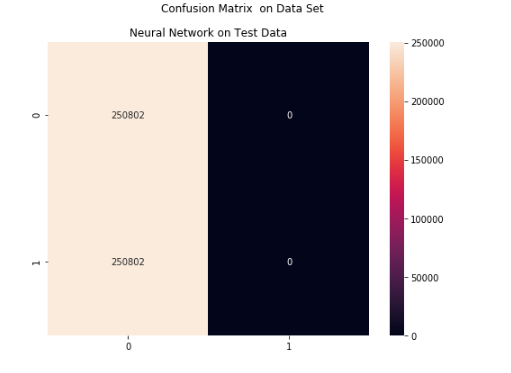
ROC Curve

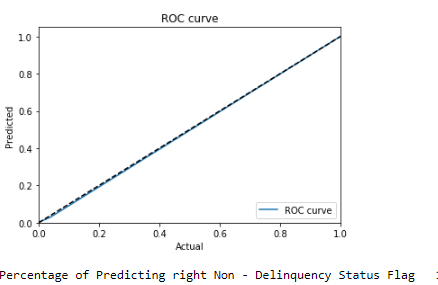




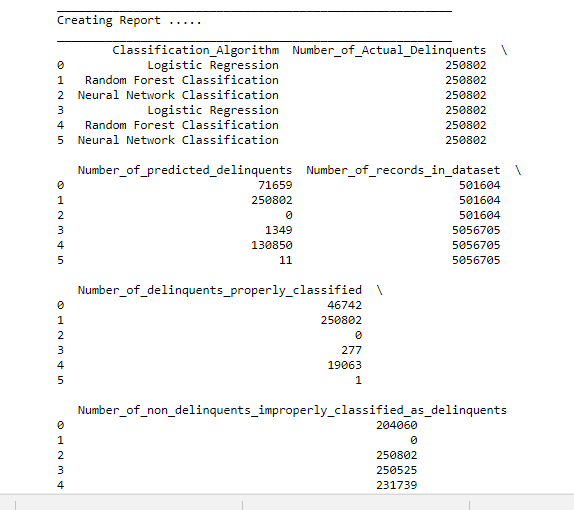


Neural Network





Score REport



We have also implied multiprocessing library :

