**2 Part II: Building and evaluating models**

**PART B- Classification**

In this part In this part we will explore different algorithms to classify loans as delinquent or non-delinquent based on the features existing in the historical performance dataset.

**Part 1- Data Download and cleansing**

For this part, we first begin with downloading the data. For downloading data, we have created a python script which accepts the train quarter as the input and downloads data for that quarter as well as the consecutive quarter.

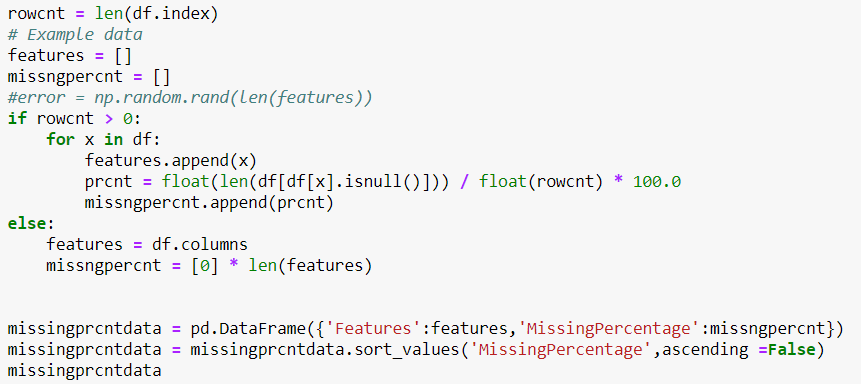
For downloading data first we run the script and save after unzipping the data.



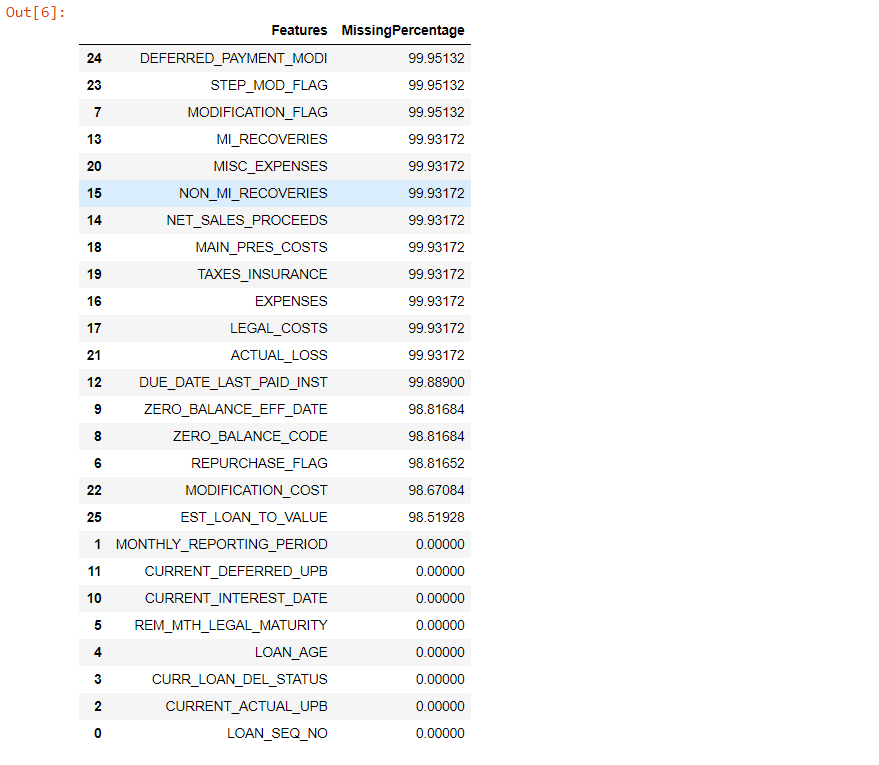
Since the data file is very big. We upload it in chunks in the pandas dataframe.

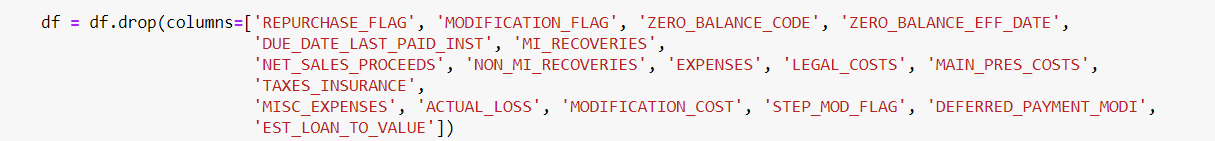


After downloading the data we check for missing values and get the values from following function



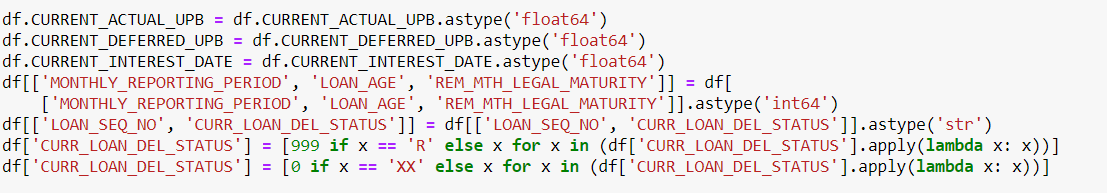
Depending on the following output, we get to know that for top 19 columns around 99 % data is missing. So, we can not use these columns for our classification algorithm.



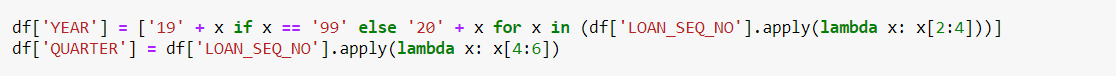


After dropping the unnecessary columns, we clean the remaining data.

First we change the data types of the columns depending upon the values.



Then we add two new columns for year and quarter for future use in classification matrix.



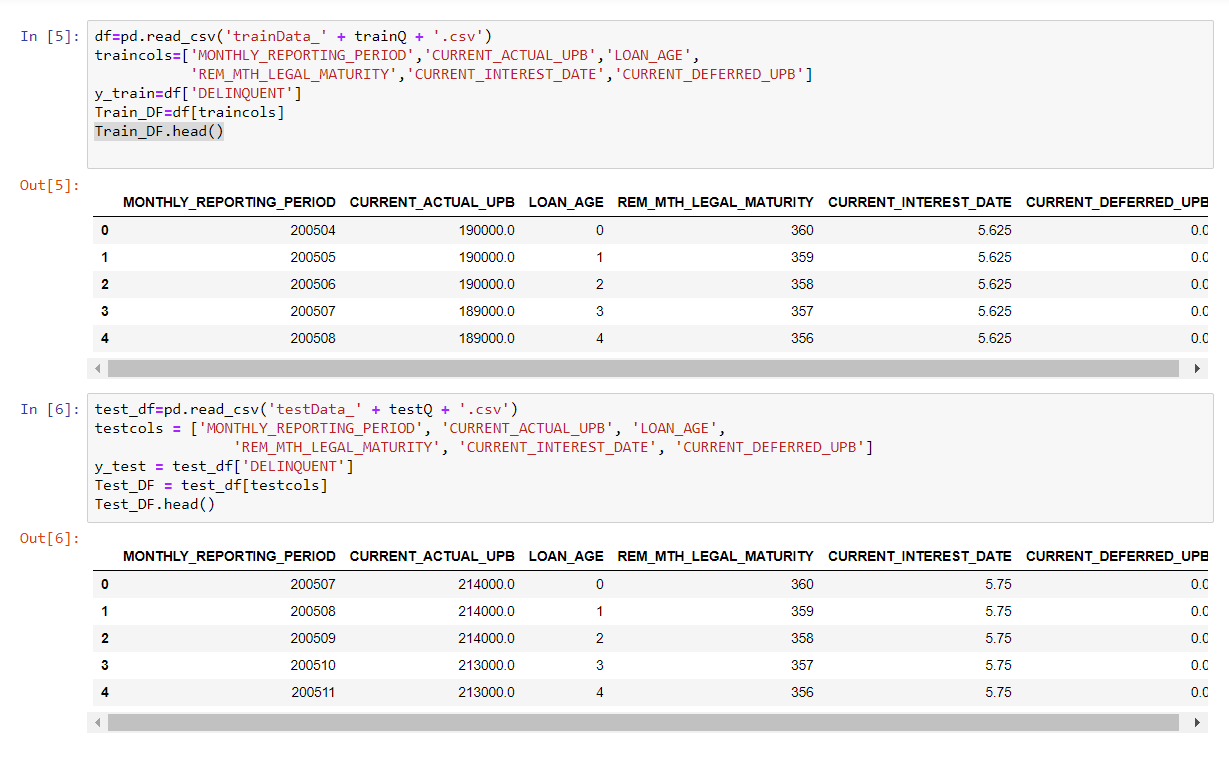
Then we add anew column as DELINQUENT depending upon the values from column - CURRENT LOAN DEL STATUS and then drop this column as this is not required in the classification data.



Now, we have clean data and we are ready for our classification algorithms.

We do the same for testing quarter data and save both in Train\_DF and Test\_DF.

We take column DELINQUENT as our Y variable.



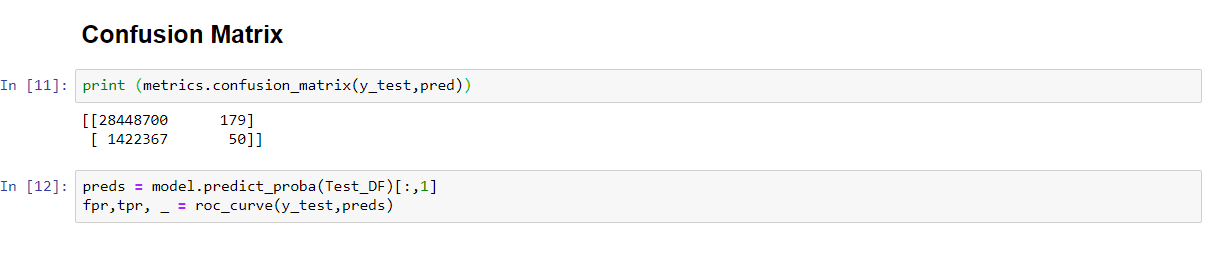
Now we run classification algorithms on these.

**Part 2- Running Algorithms**

**Logistic regression:**

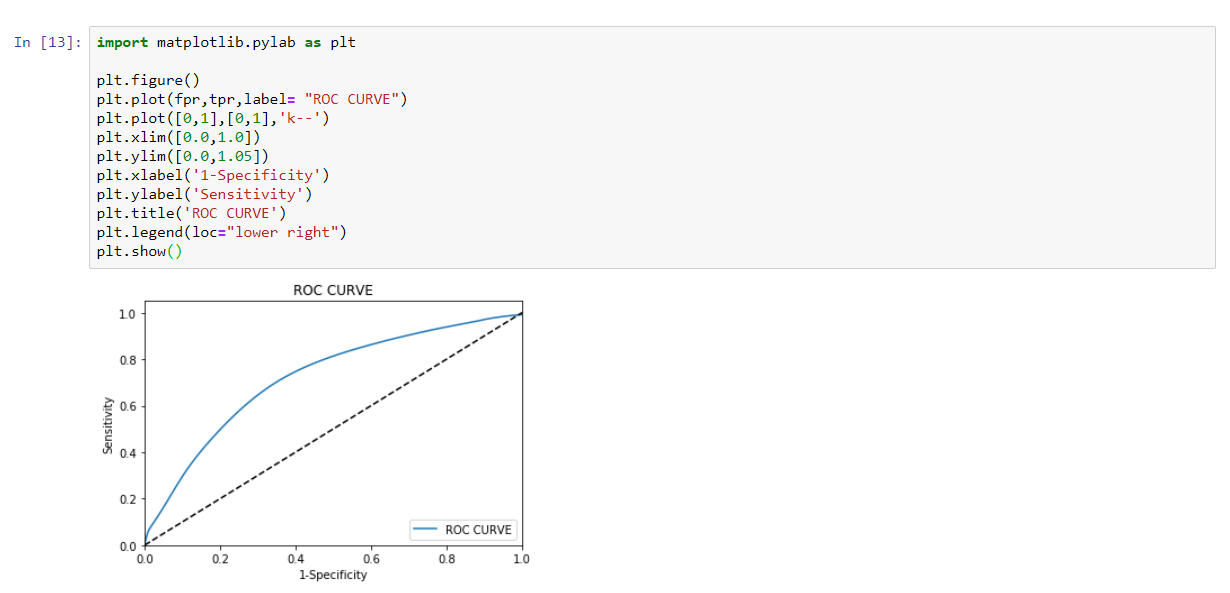
We did not select this model because although we got similar results for the accuracy, we got very different results for the confusion matrix. The number of true positives was lower by a significant amount.



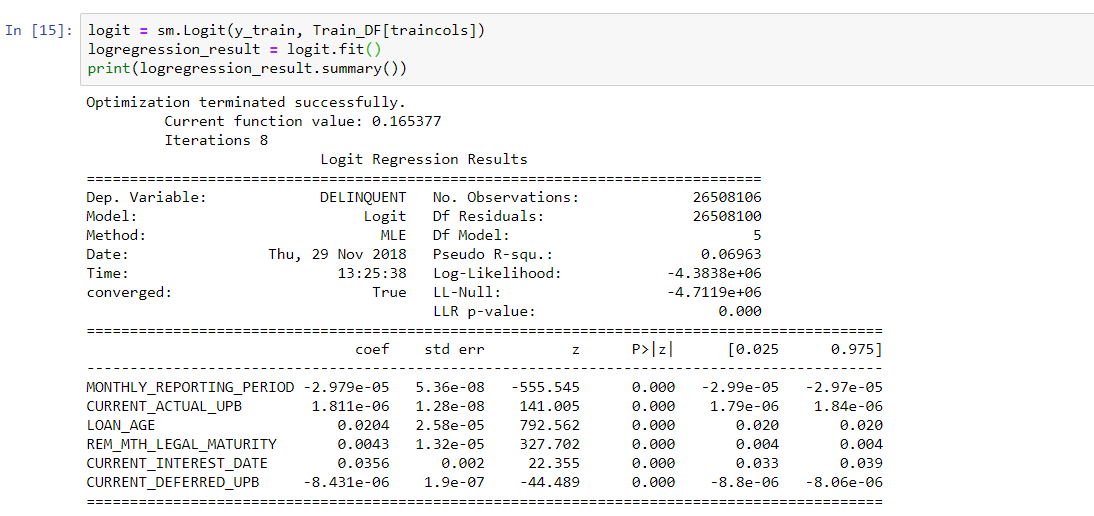


**ROC Curve**

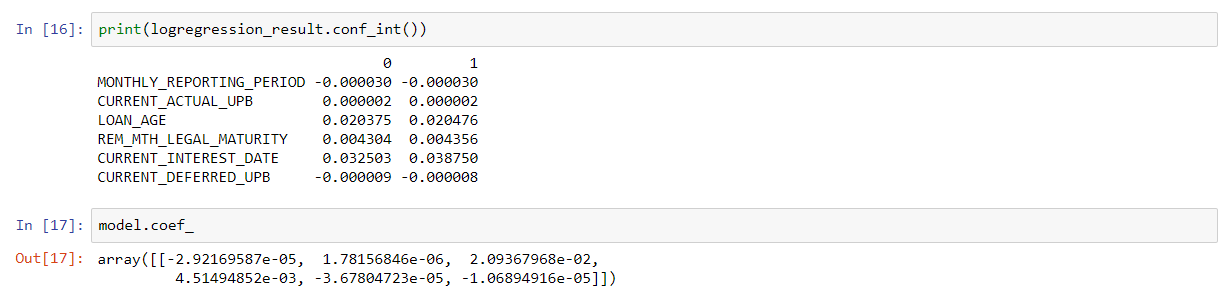
The ROC curve based on this algorithms is:



Also we used logit for detailed analysis.

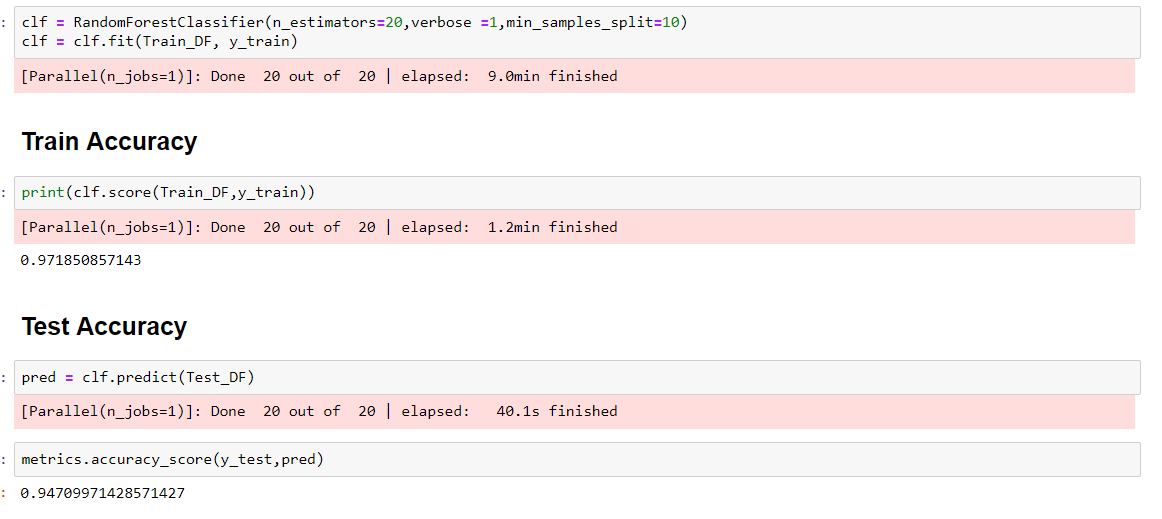


Finally, the values of other coefficients:

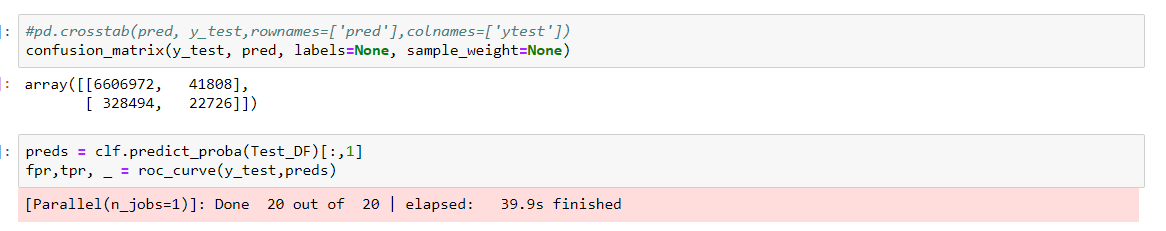


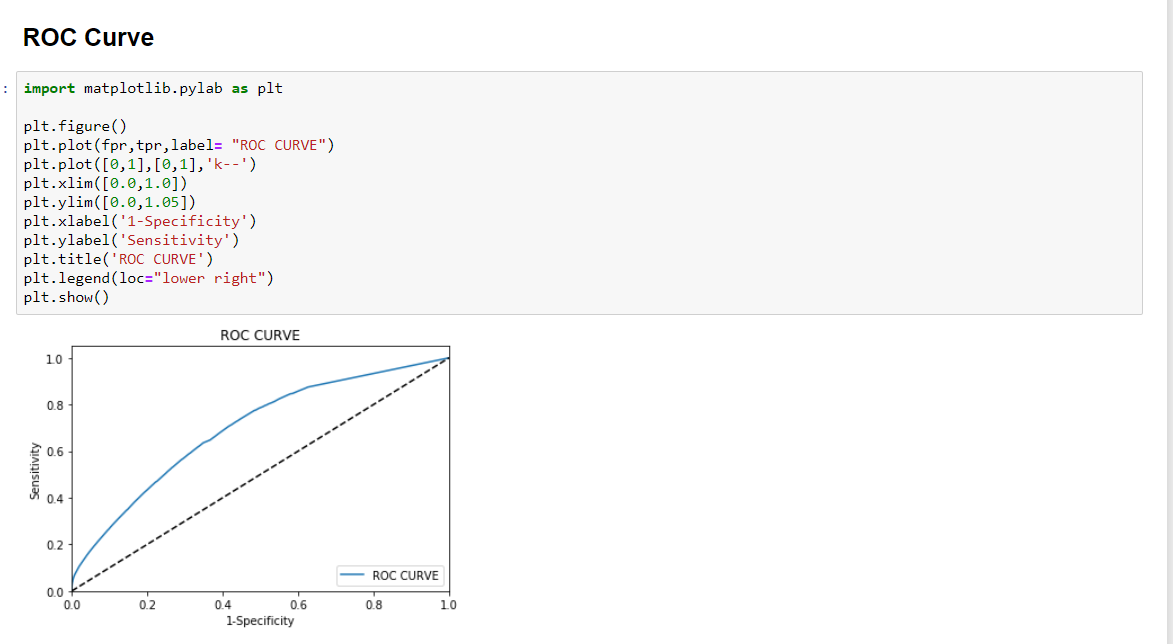
**Random Forest Algorithms:**

**We chose random forest as best algorithm due to its accuracy.**



**Confusion Matrix**





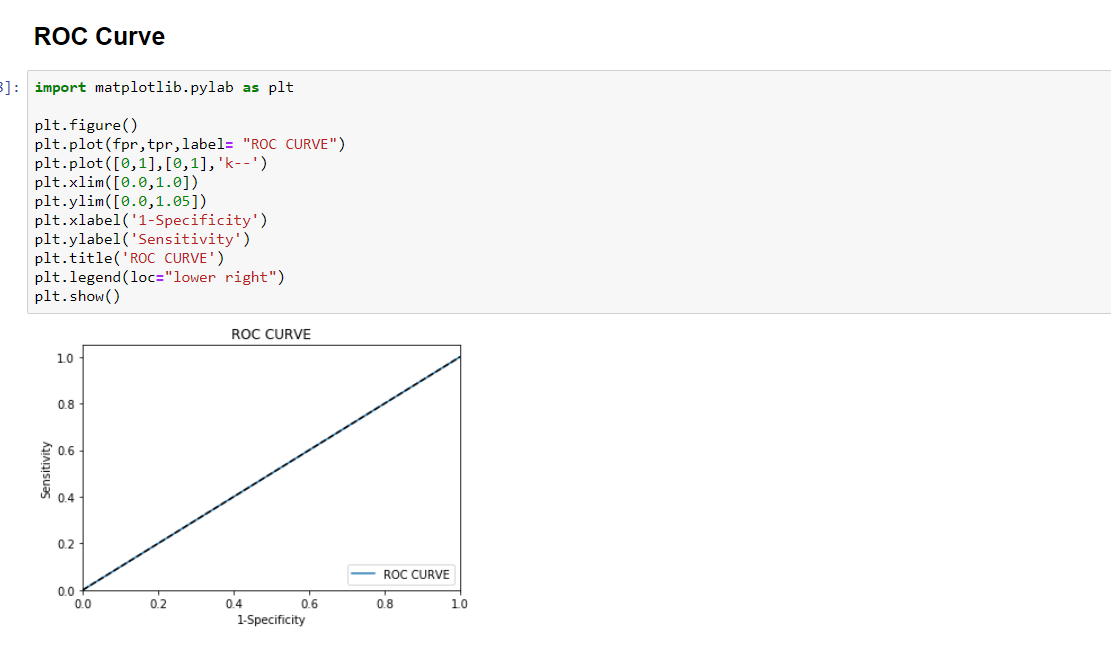
**Neural Networks:**

We tried running this algorithm with different tuning parameters which we tweaked to get results. We tried changing the number of hidden layers and the number of neurons with different learning rates.

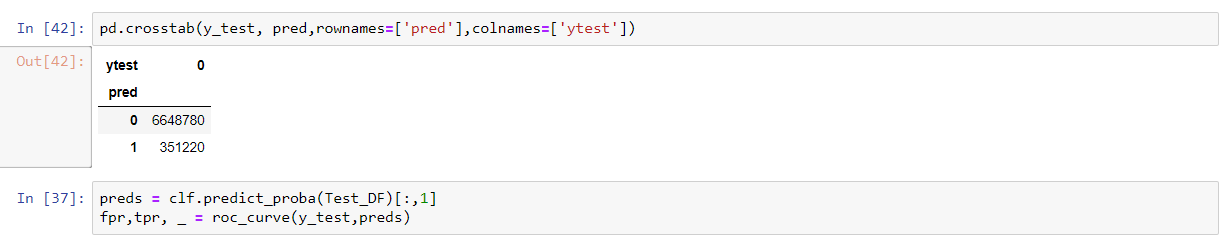
We are getting similar accuracy as the one we got from Random Forest and Logistic Regression, but the Confusion matrix shows very poor results.

Neural Network is not able to identify the delinquent loans and is predicting everything as non-delinquent. That is why we have discarded this algorithm.



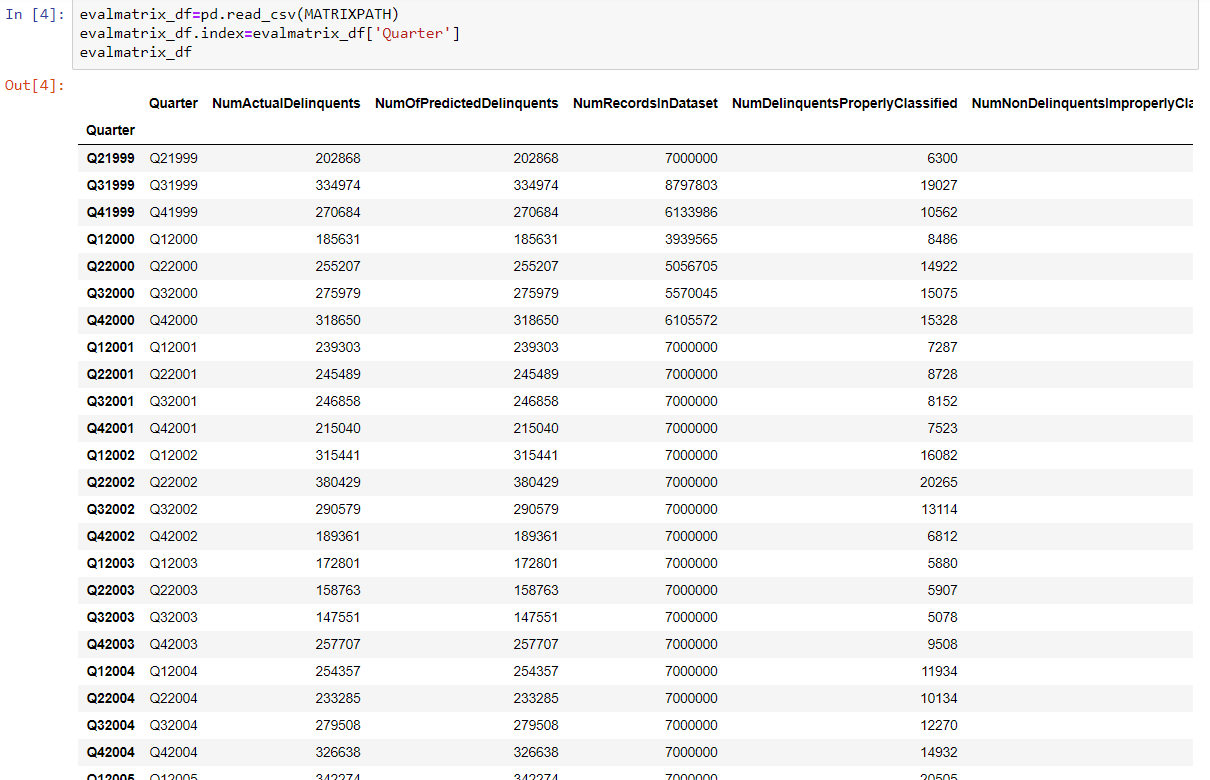


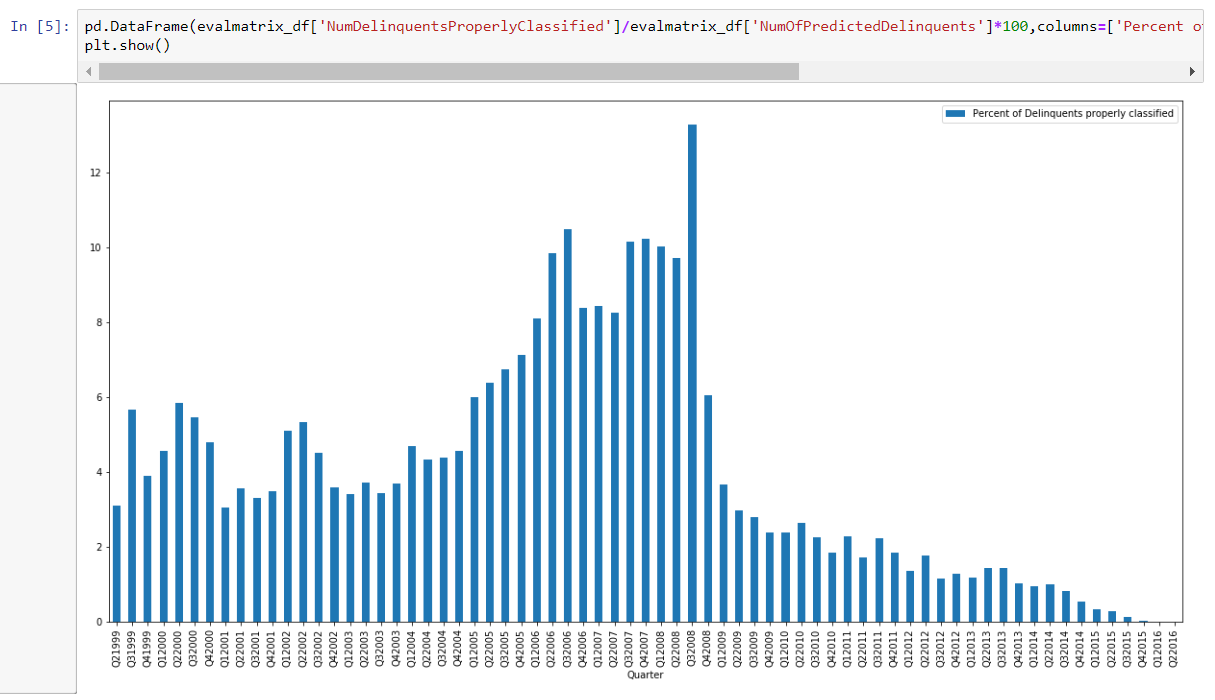
**Confusion Matrix**

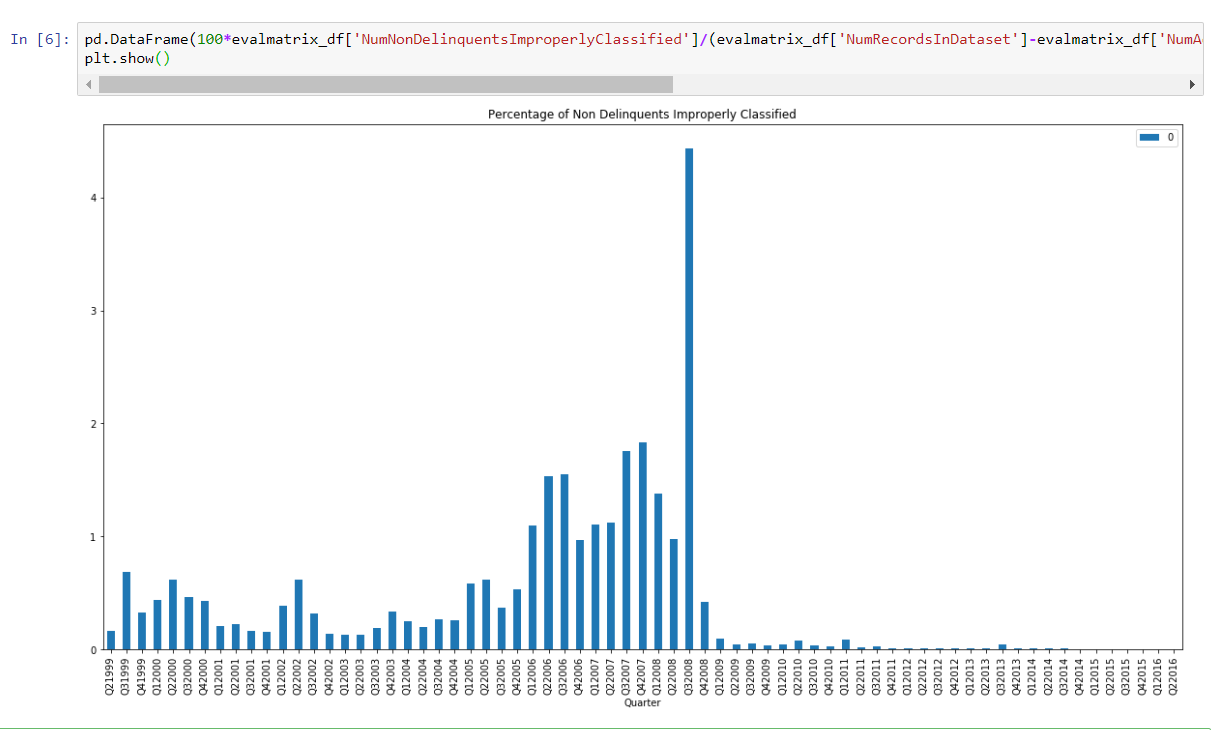


**Evaluation Matrix**

We downloaded the data from Q11999 to Q22016 and saved the evaluation matrix based on delinquents.







**Classification Summary**

**Justification on the chosen classification model**

• Similar score for all the algorithms was obtained.

• Random forest gives the best results in terms of the confusion matrix.

• Random Forest takes lesser time to train as compared to the Neural Network.

• Although the area under the ROC curve for Logistic regression is more, we have chosen the Random Forest algorithm because the number of True positives is significantly higher.

**What can be done better**

• Although we can see that the random forest provides reasonable outputs for this type of data. We can perform further analysis with more computational power to arrive at better results.

• The dataset is a biased one meaning that there are significantly more number of N values for delinquents as compared Y values. Because of which, the algorithm has a lesser True positive rate.

• We can see from the evaluation that, the algorithm performs better when trained with more delinquent data. We can re-develop this algorithm and fine tune the tree parameters for the Q12007-Q42008 period which might enable us to create an algorithm with higher accuracy as far as the confusion matrix is concerned.

• We can explore more algorithms for classification such as KNN or SVM to check the accuracy.

• There might be an optimum parameter for Neural Network that exists which we haven’t come across in our analysis.

**Summary**

We have analyzed the data given in the Freddie Mac Single Family Loans Dataset. We have seen different parameters related to the origination and performance files. These have been summarized to perform EDA on the data. Data has been automatically downloaded using HTML scraping techniques. We have performed Lasso Regression, recursive feature elimination, F regression and PCA for performing feature extraction and selection. We have analyzed different algorithms such as Regression, Random Forest and Neural Networks to perform Predictive and Classification modelling and tested the algorithm on Quarters from 1999-2016 in a rolling quarters manner. According to our analysis, Random Forest was the algorithm of choice for Prediction as well as Classification. We have done extensive analysis for the period 2007-2009 to analyze different aspects of the market crash