MIDTERM REPORT

## Advance in Data Sciences and Architecture

## INFO 7390

## TEAM 4:

## PRUTHVIJ THAKAR

## SUMEDH SARAF

## Data wrangling and Exploratory Data Analysis

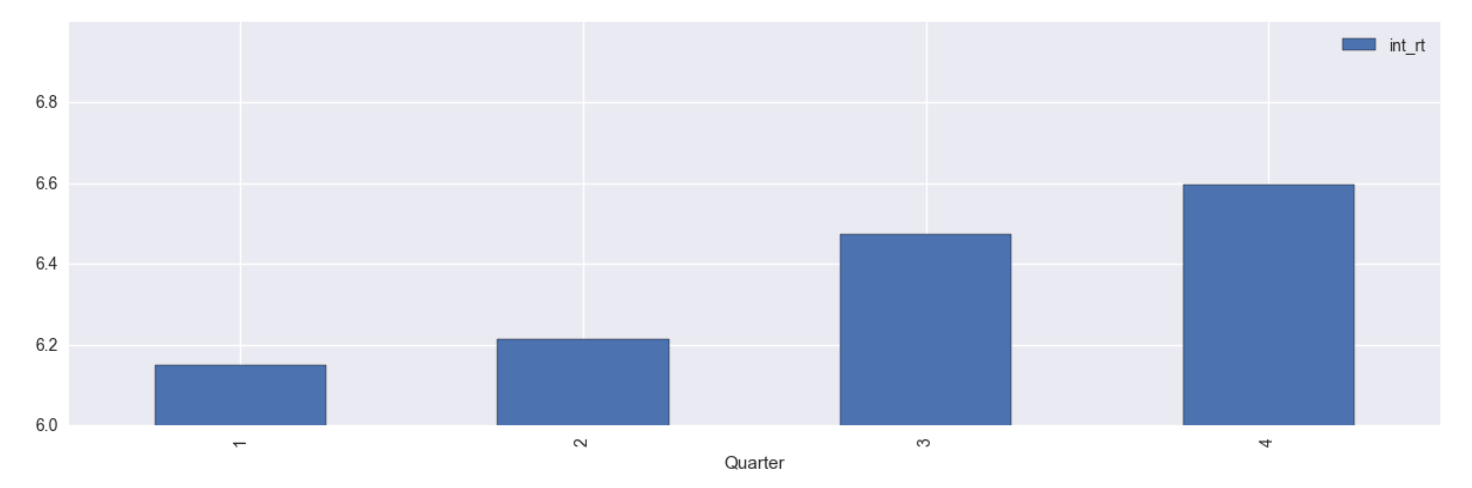
1. *Downloading and Pre-Processing:*

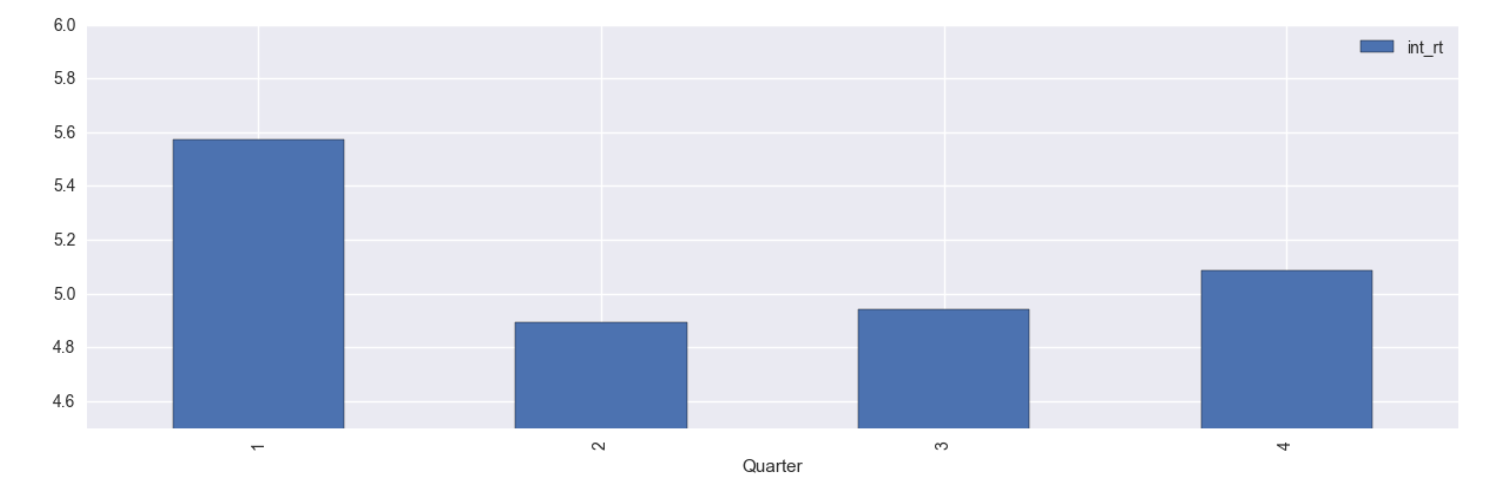
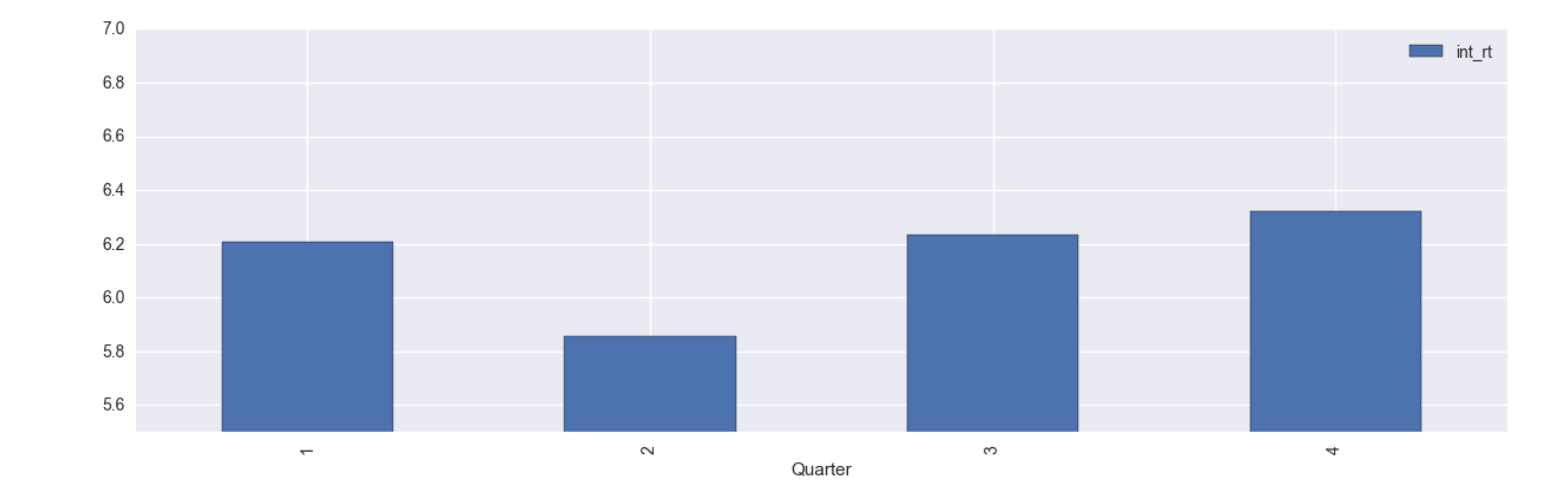
The downloading and preprocessing part will programmatically enter the Freddie Mac website and scrape the samples file from 2005 onwards and store them in the system where Docker image in run and clean them and new file with the summaries will be created which is the combined for all the origination data in one file and combined file for performance in other file.

1. *Exploratory Data Analysis:*

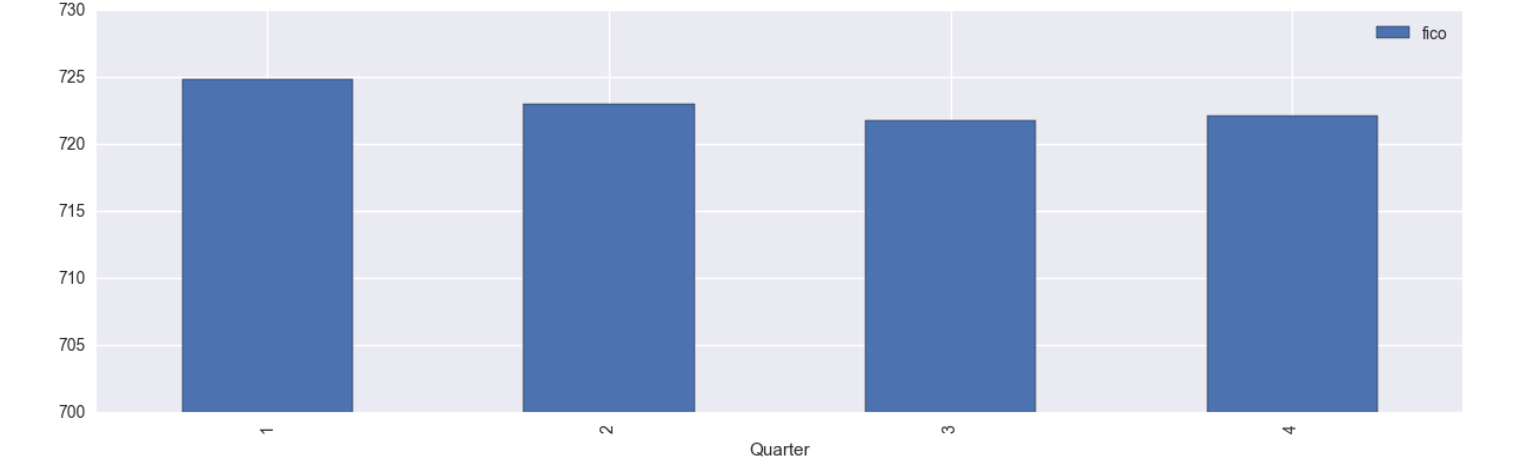
Exploring the Quarters of year 2007,2008 and 2009 for getting insights into the data and trends over these years, the following findings were discovered.

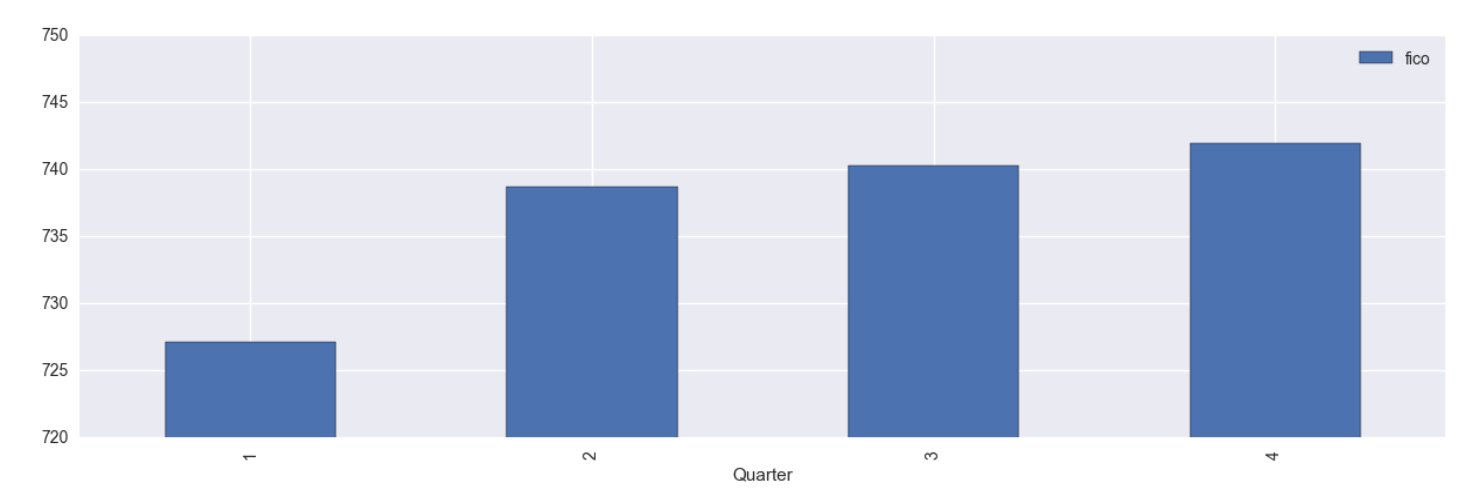
The Quarterly interest rates for year 2007,2008 and 2009 showed a trend as follows:

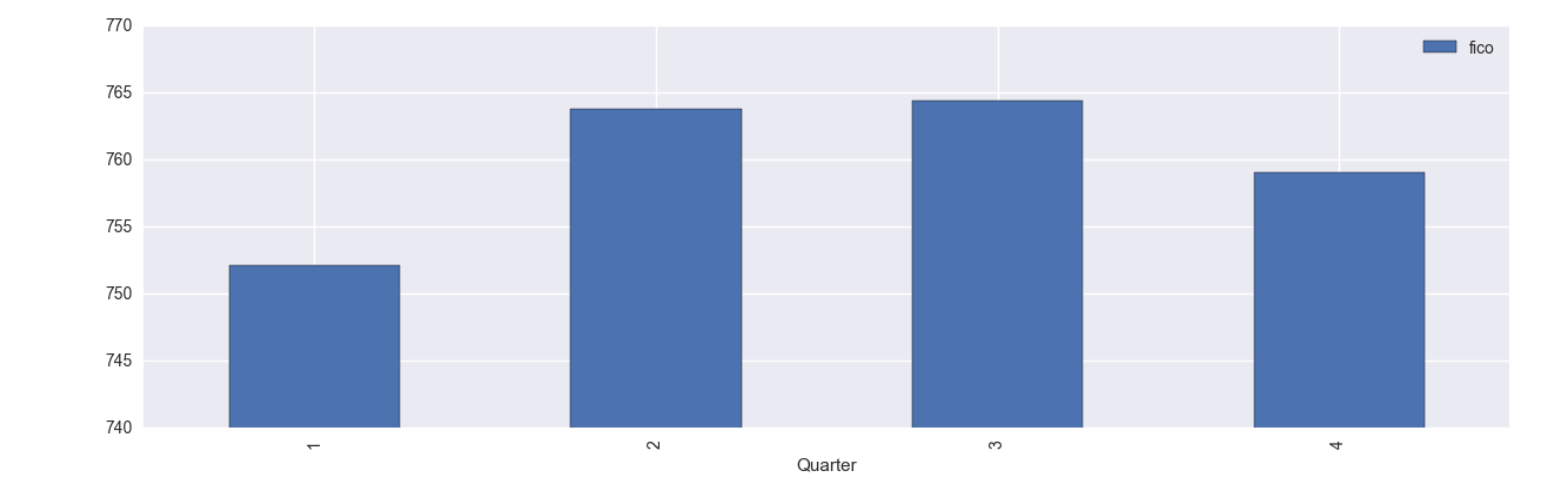




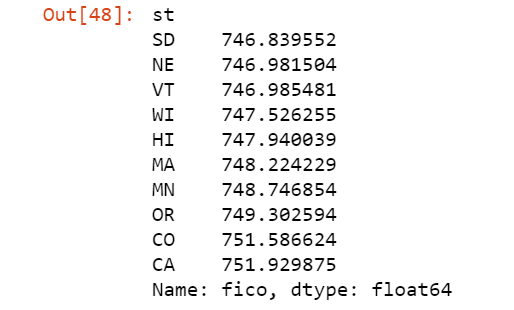
The Quarterly interest rates for year 2007,2008 and 2009 showed a trend as follows:

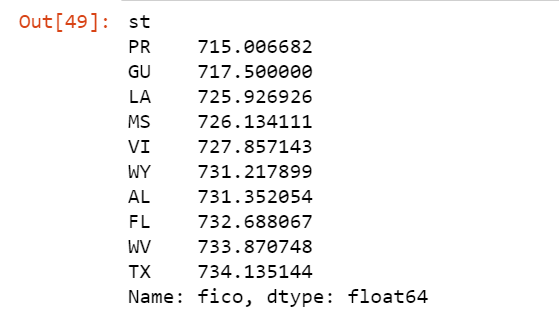




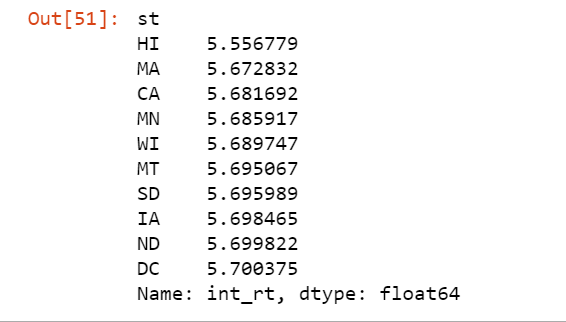
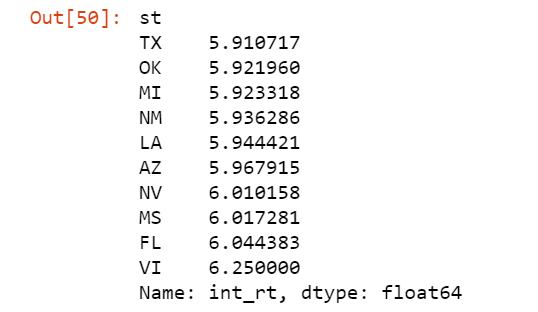


The States with maximum and minimum fico scores were as below:



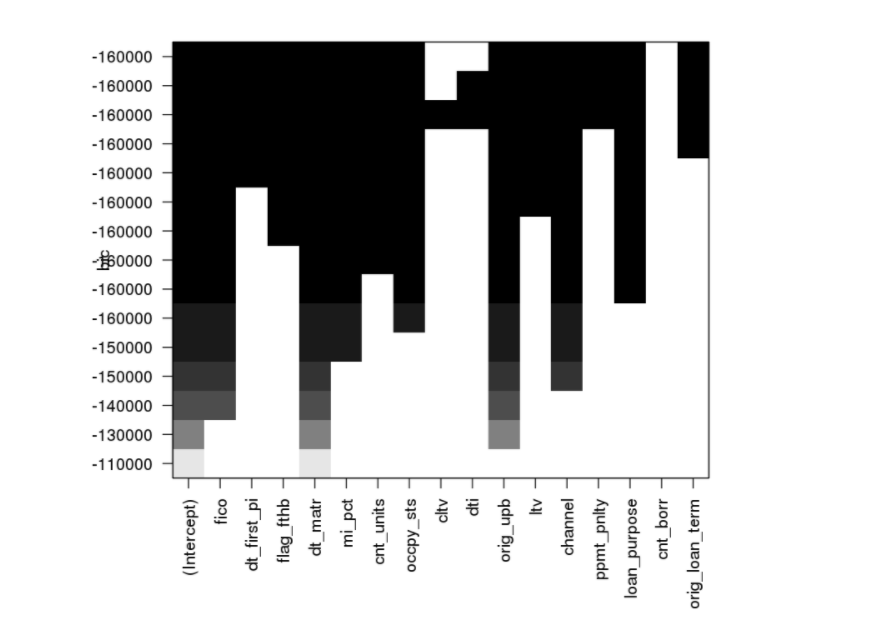
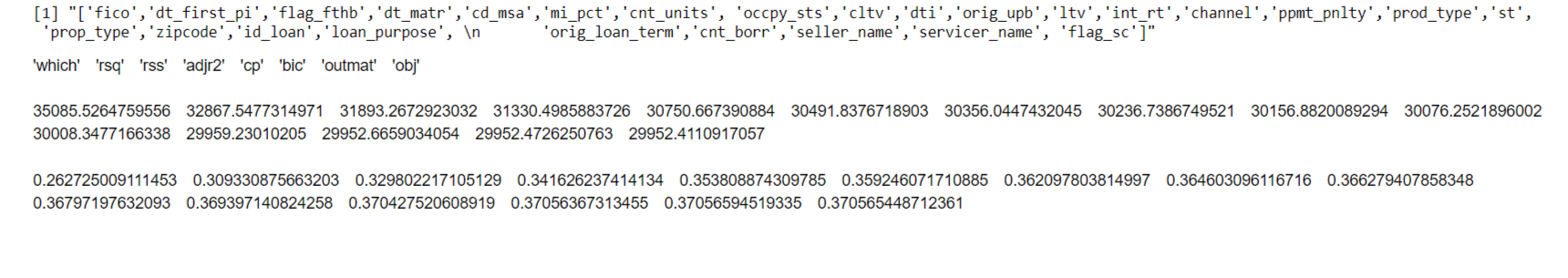


The States offering most and least interest rates were as follows:

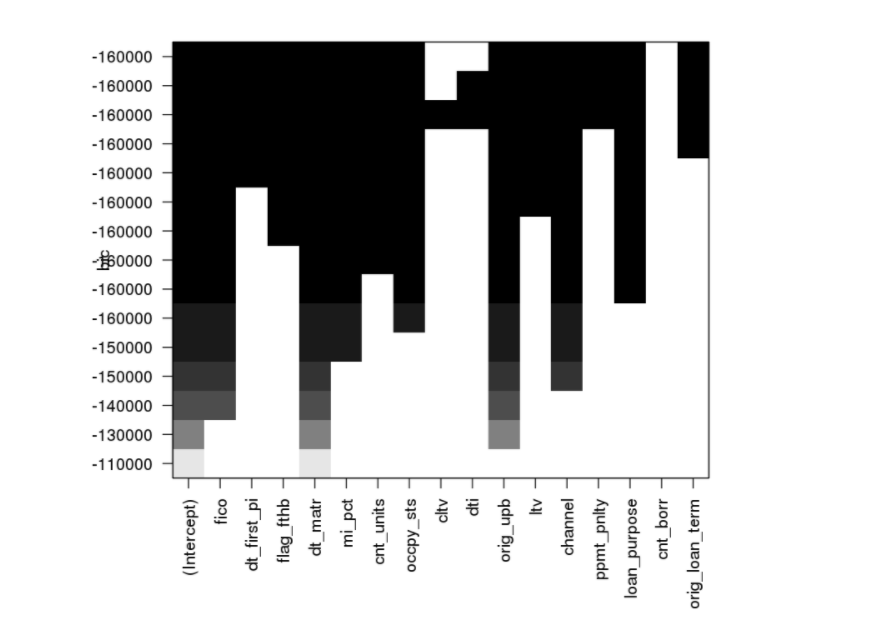
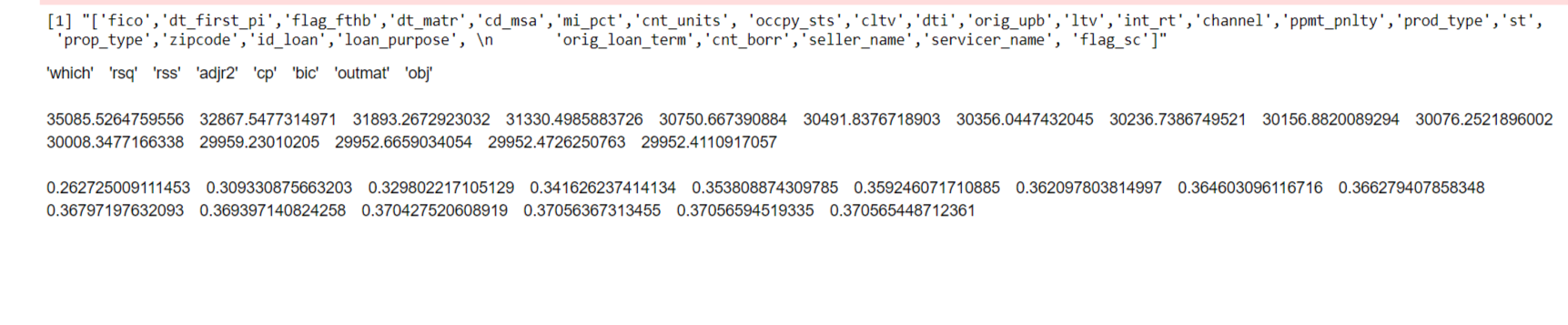


## Building and Evaluating Models:

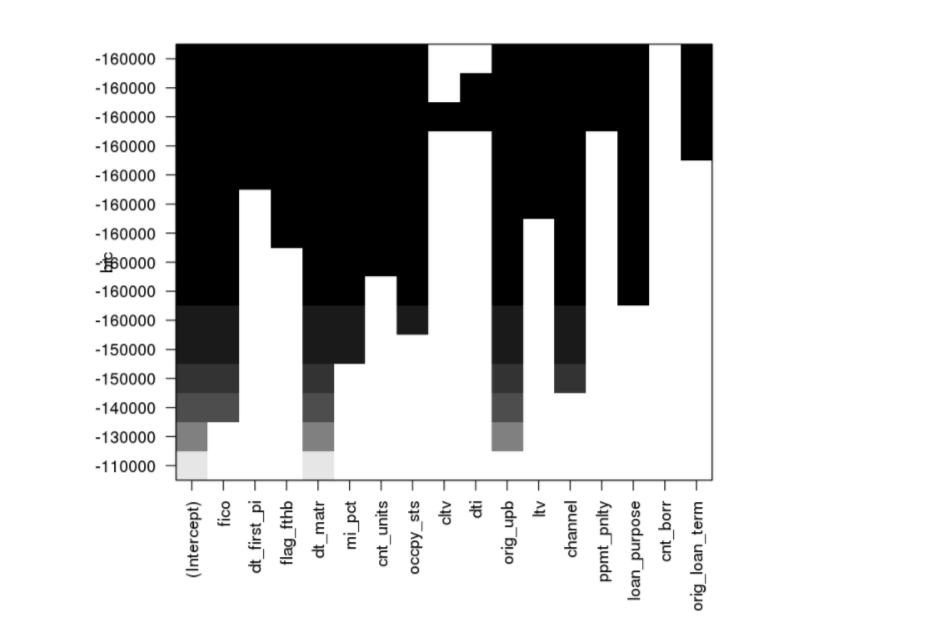
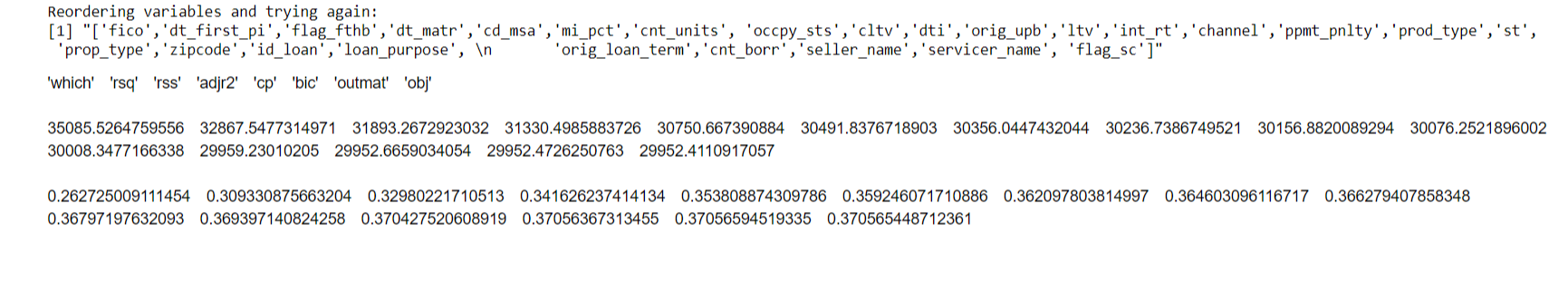
1. Prediction:



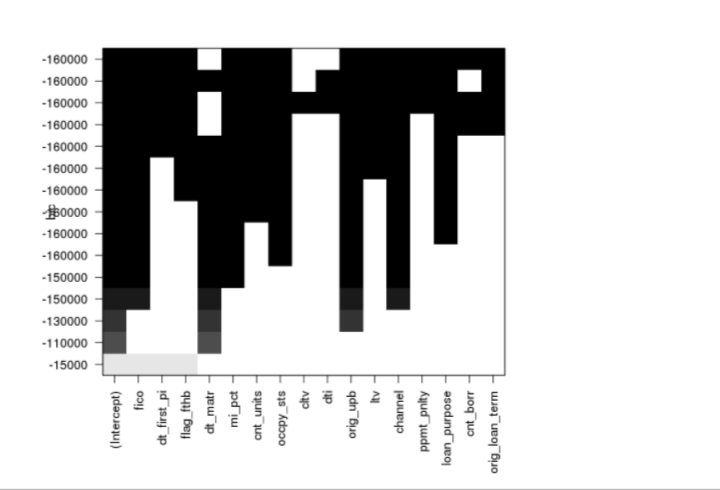
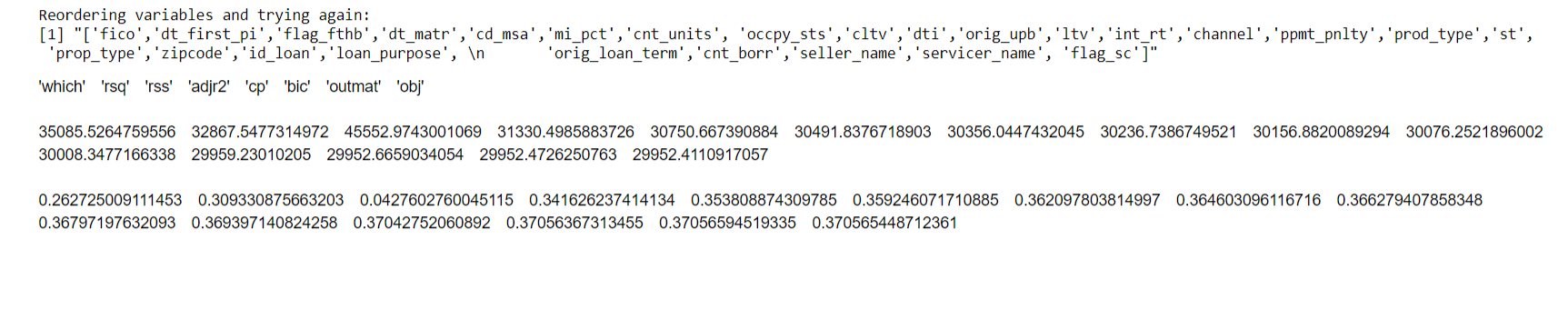
Backward Search



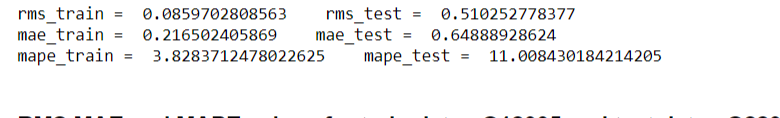
Exhaustive Search



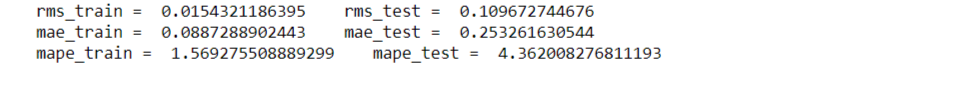
Stepwise Search:



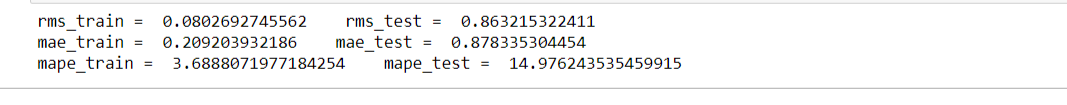
Results after deploying Regression model using the variables selected by the search methods mentioned above on Quarterly data of Q22005:



Results after deploying Random Forest model using the variables selected by the search methods mentioned above on Quarterly data of Q22005:

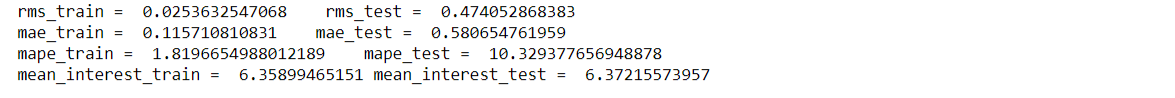
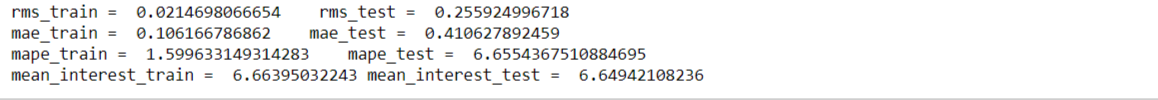
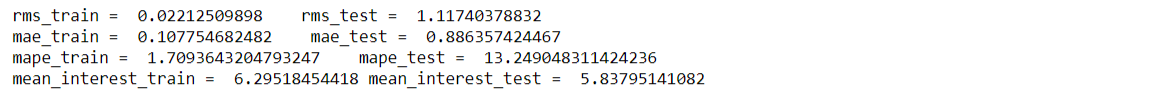
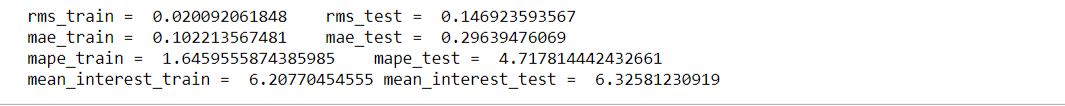


Results after deploying Regression model using the variables selected by the search methods mentioned above on Quarterly data of Q22005:



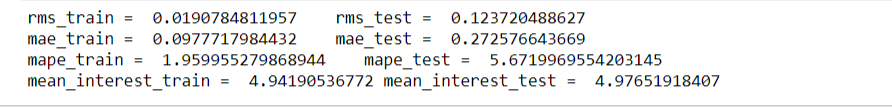
* Thus, from the above images, we can say that the most accurate algorithm after verifying their outputs is **Random Forests.**

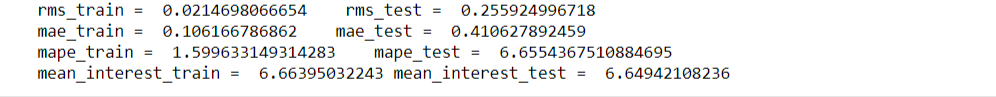
The outputs after deployment of the best algorithm on Quarters which faced the wrath of financial crisis is as follows:

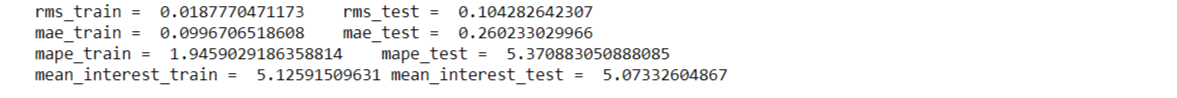


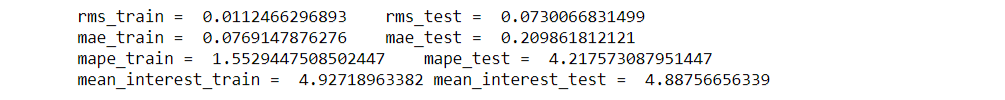
* We have used Q12007, Q22007, Q32007, Q42007 for training and predicted for the Quarters Q22007, Q32007, Q42007, Q12008 when the Financial Crisis took place. We can observe from these records that the error rate has increased on a great scale between training and testing data although the algorithm has been already tested on same type of data with other Quarter.
* If we compute our best model i.e. Random Forests on data of 2009 to check the findings if they change from 2007 or not, when in 2009, the world started recovering from the financial crisis.

So, we try to run it on Q12009, Q22009, Q32009, Q42009 as training inputs and Q22009, Q32009, Q42009, Q12010 as the testing outputs we get the following results:



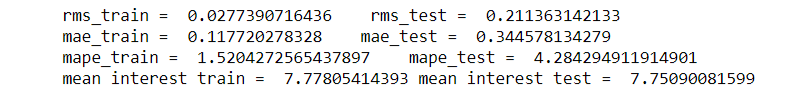
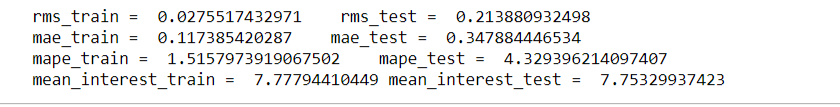
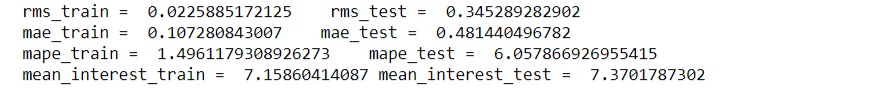
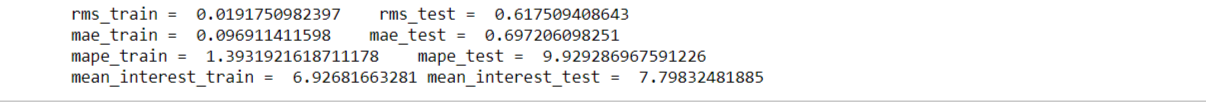






As we can observe that the error rates between the values of RMS, MAE, MAPE has decreased in a quite dramatic manner and pointing towards the stability in the market, one more thing we can observe from this is that the interest rates were fluctuating quite noticeably as it started around 4.x% for then bounced to 6.x% and again came decreasing to 4.x% for the last quarter.

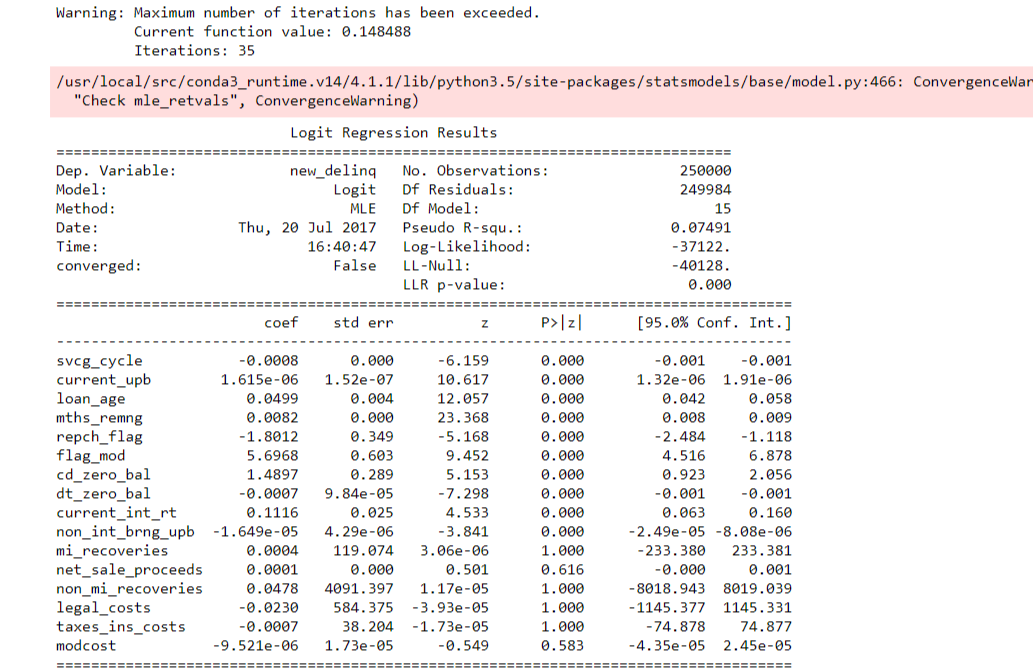
* If we compute our best model i.e. Random Forests on data of 1999 to check the findings when the world started entering Economic Boom.



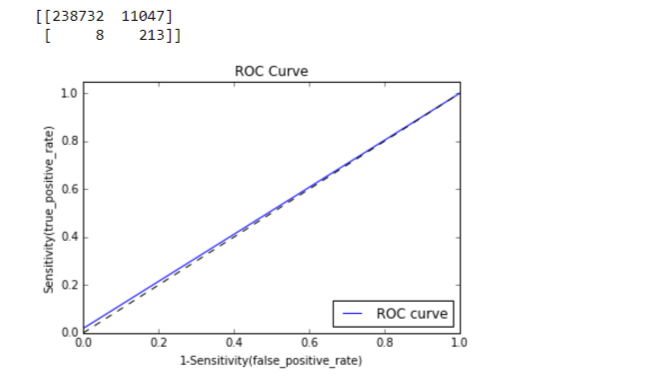
* The noticeable factor will be the interest rate which is at peak in all quarters and more than 7% where as we have observed for other quarters that it dips to 4-6% also while crisis or economic stability. The other great thing to notice in the model is it is not giving out a huge error rate in training and testing data as it did in the Recession period. We can observe the screen shots of outputs of all the 4 Quarters depicting the RMS, MAE and MAPE values of all the quarters.

1. Classification:

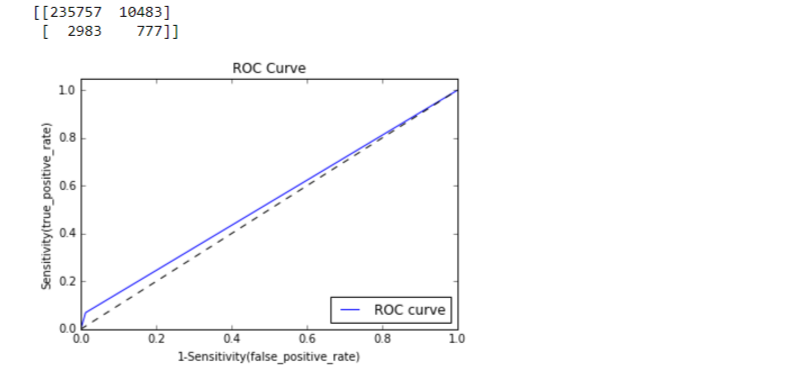
The variable selection for the models is done using the computed p values which if nearer to 0 increases the chances of dependency of variables and is good for the final model.



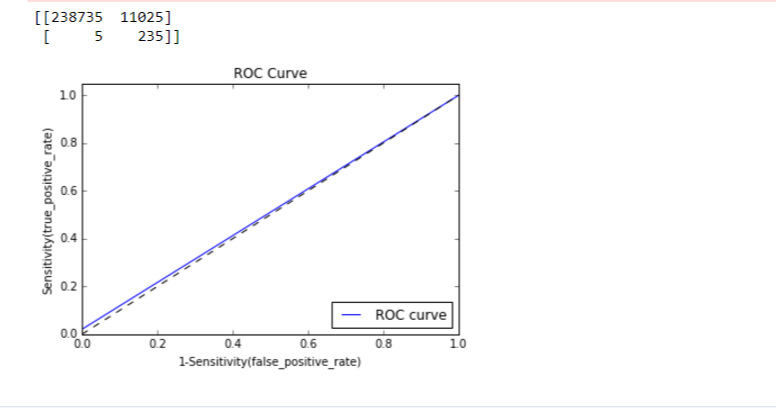
The Logistic Regression algorithm model is applied on the data of Q12005 and validated against the quterly data of Q22005 and the following findings in form of confusion matrix and ROC curve were discovered.



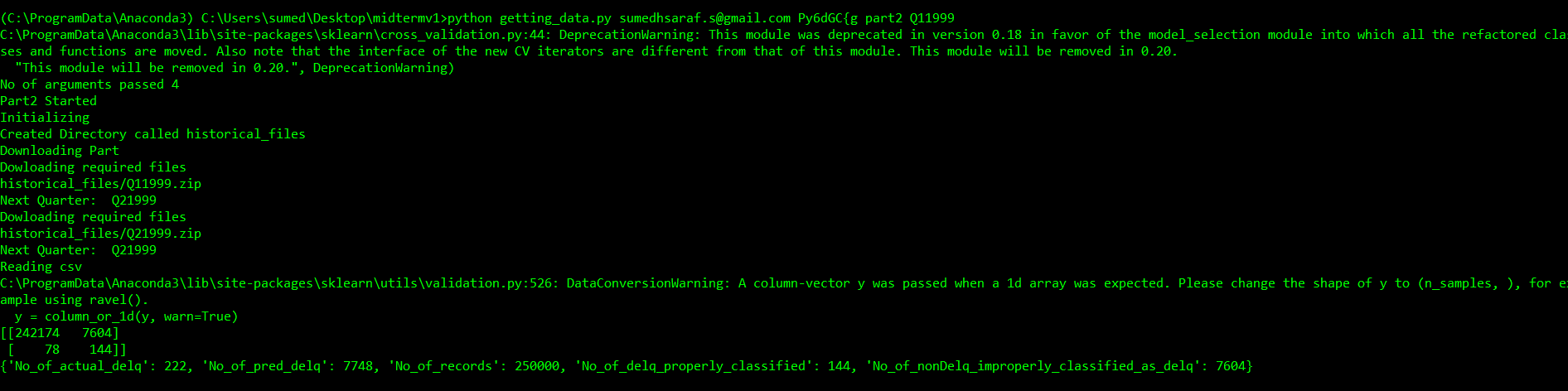
The Random Forest model is applied on the data of Q12005 and validated against the quterly data of Q22005 and the following findings in form of confusion matrix and ROC curve were discovered.



The Neural Network model is applied on the data of Q12005 and validated against the quterly data of Q22005 and the following findings in form of confusion matrix and ROC curve were discovered.



* Thus, by running the above algorithm on Q12005 dataset as training dataset and Q22005 as testing dataset, we can figure out that the Logistic Regression model is the best choice for the classification problem for this dataset.
* Parameterizing the input (e.g.: Q12005) given by user while running Docker image and giving out matrix, The output can be seen in the last line when the python file is run in command prompt, it pops out the required matrix



## Instruction to run the Docker Image:

# Instruction to Run the code:

# Download [FreddieMac\_loan\_data\_challenge](https://github.com/PruthvijThakar/FreddieMac_loan_data_challenge)folder from github in to your local machine

* To get the images in the local execute:
  + docker pull sumedh11/midtermads4:latest
  + docker run sumedh11/ midtermads4:latest usename password part1
  + docker run sumedh11/ midtermads4:latest usename password part2 Q22005

(you can specify which quarter you want to analyze while giving the input)

## Go in FreddieMac\_loan\_data\_challenge folder

-docker build -t image1 .

-Publishing an image to docker Hub

-Create a docker hub account, create a repository

-Connect to your account docker login

-Tag your image to be pushed with repository docker tag image1 sumedh11/ midtermads4:latest

-Push your image o docker hub docker push sumedh11/ midtermads4:latest

-Pull image docker pull sumedh11/ midtermads4:latest

* The commit is necessary after each time the image is run so that the data doesnt get lost.