

**ADVANCED DATA SCIENCE**

Assignment 3

Midterm Case Studies

Prof. Sri Krishnamurthy



Team Members: Sayali Borse Rishi Rajani Komal Ambekar

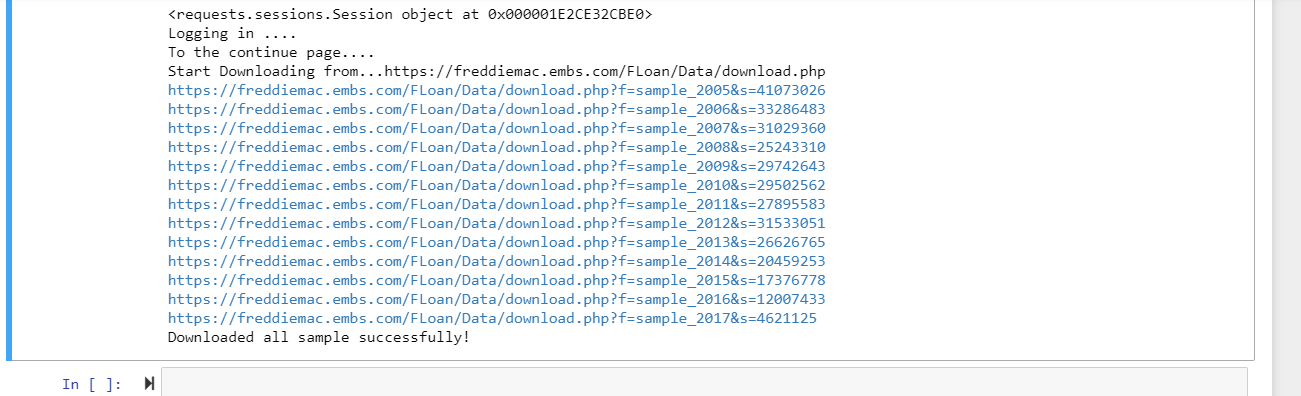
* Content:

|  |  |
| --- | --- |
| **Sr.No.** | **Content** |
| 1 | Data downloading and Pre-processing |
| 2 | Exploratory Data Analysis |
| 3 | Building and evaluating models |
| 4 | Prediction |
| 5 | Classification |

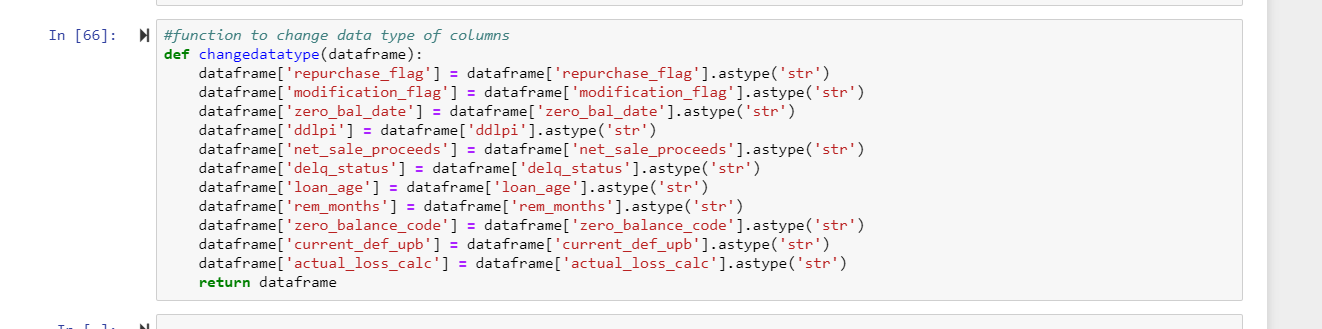
* Part 1: Data Downloading and Pre-processing
* Download Origination and performance files from [https://freddiemac.embs.com/FLoan/Data/download.php](https://freddiemac.embs.com/FLoan/Data/download.php%20%20)  downloaded using mechanicalsoup by passing and saving cookies.
* Summarizing and cleaning the data based on the user guide provided. For example: Checking the valid Credit Score, checking and replacing blank values.
* Processing big combined performance files by summarizing it with maximum no of months, maximum and minimum actual upb , getting maximum of other columns, getting minimum of non mi recoveries, expenses , legal costs and taxes and insurance.
* Following are the screen shots of the code snippets:

Programatically downloads the sample\_original and sample\_svcg files starting from 2005 to 2017 from freddiemac website.





The files are then combined to a .csv file and the datatype for the columns are changed.



Data cleaning is done on the columns.





* Exploratory Data Analysis:

It is an approach to analyze data sets to summarize their main characteristics, often with visual methods. EDA is for seeing what the data can tell us beyond the formal modeling. It is typically the first step of analysis.

Step 1: Read the combined CSV file



Step 2: Plotted graphs to ease data visualizations

* Average Credit score by year



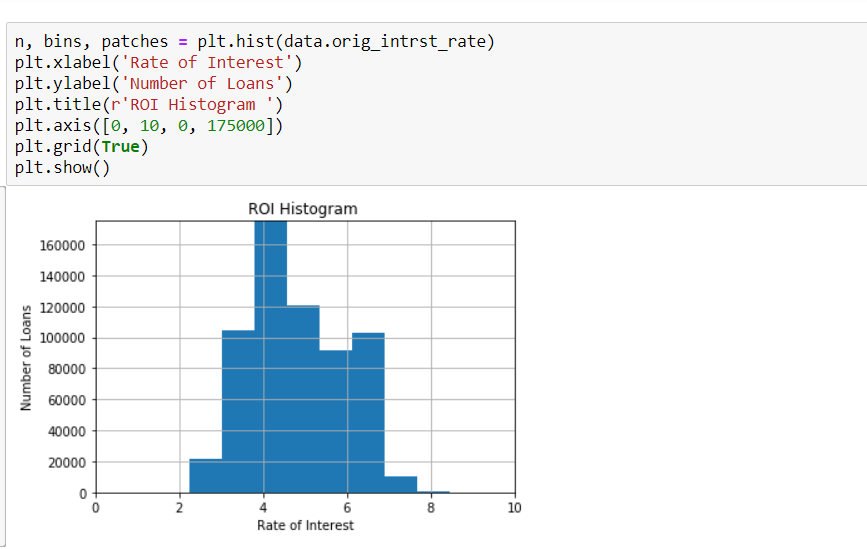
We deduce that credit scores did not have a drastic change in any of the years.

* Adding quarter number for better understanding of EDA



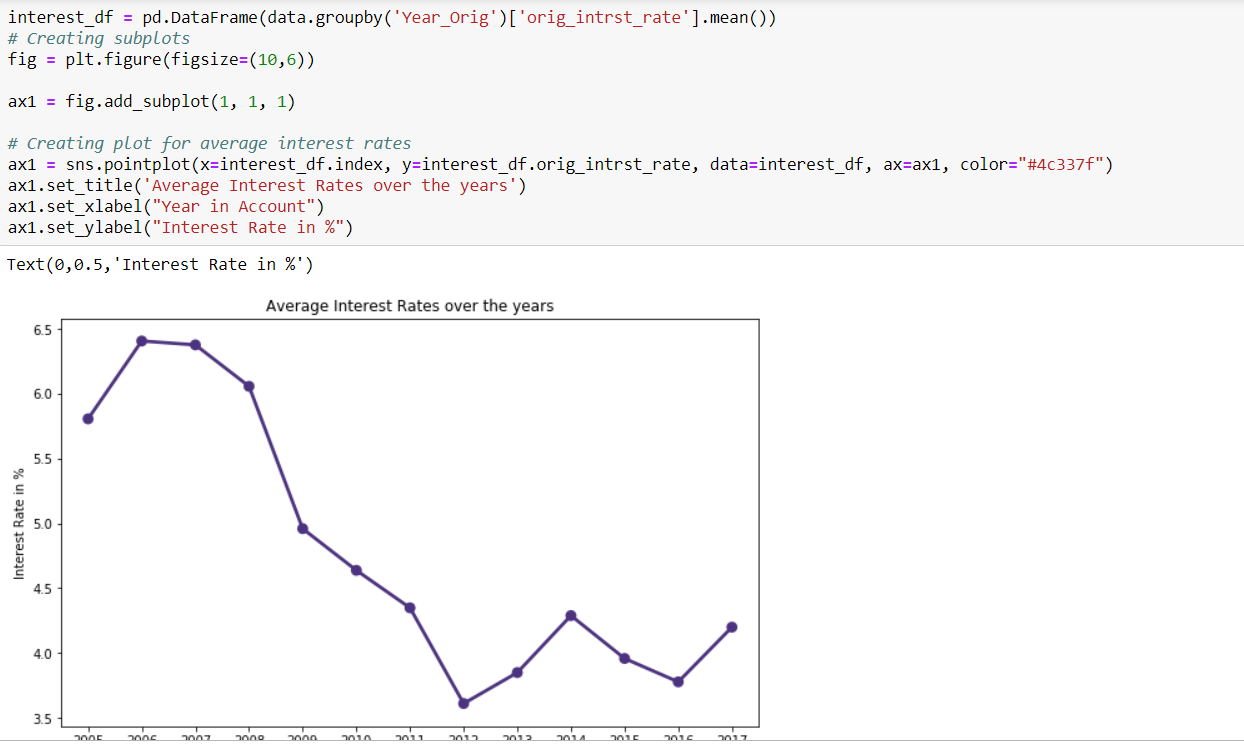
Similarly, credit scores across the quarters also remain same.

* Calculating Rate of Interest against Number of loans using Histogram



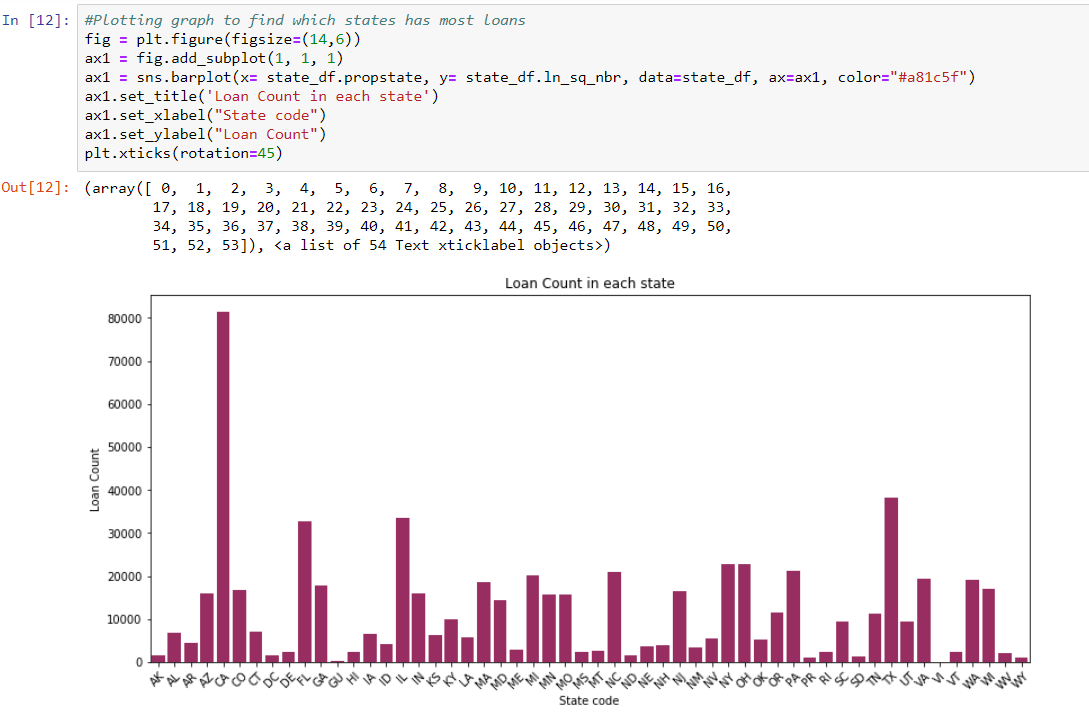
We deduce that most of the interest rates were given out at anywhere around 4-5%

* Plot for Average Interest Rates over the years



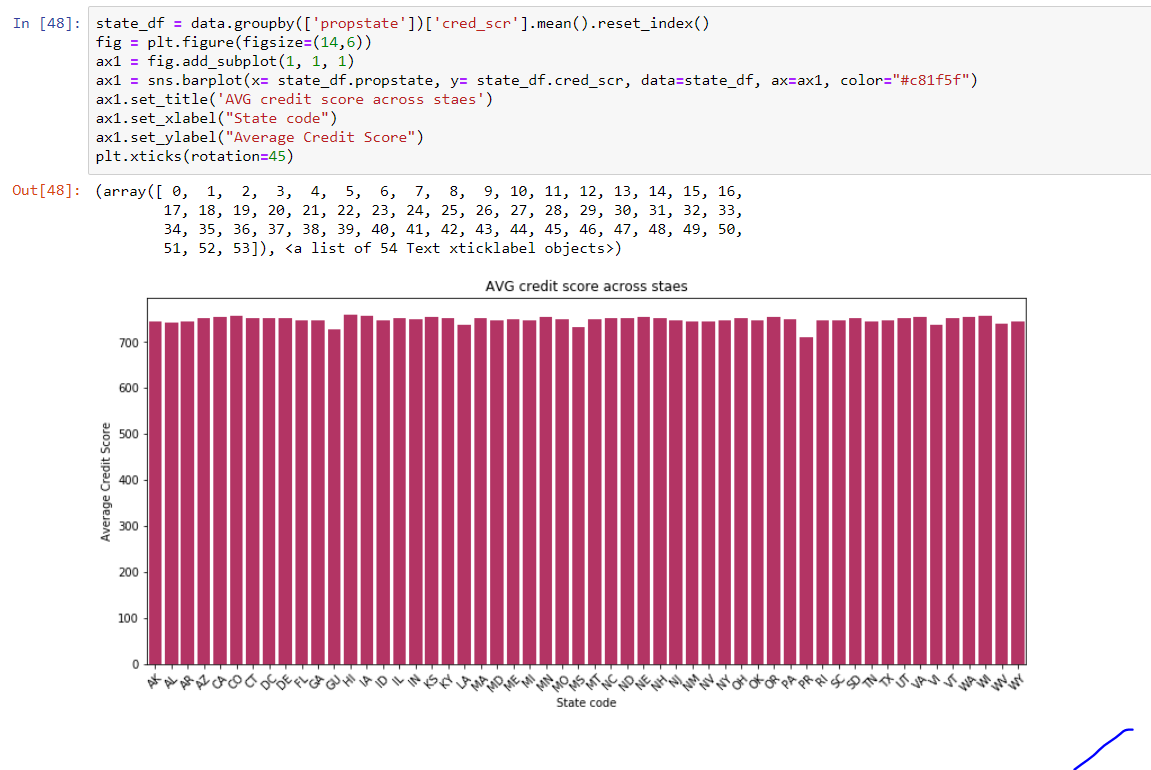
We deduce that rates peaked initially during the first few years but then fell to very low during 2013 and then rose again but not as bad as 2006.

* Plotting graph which states has most loans



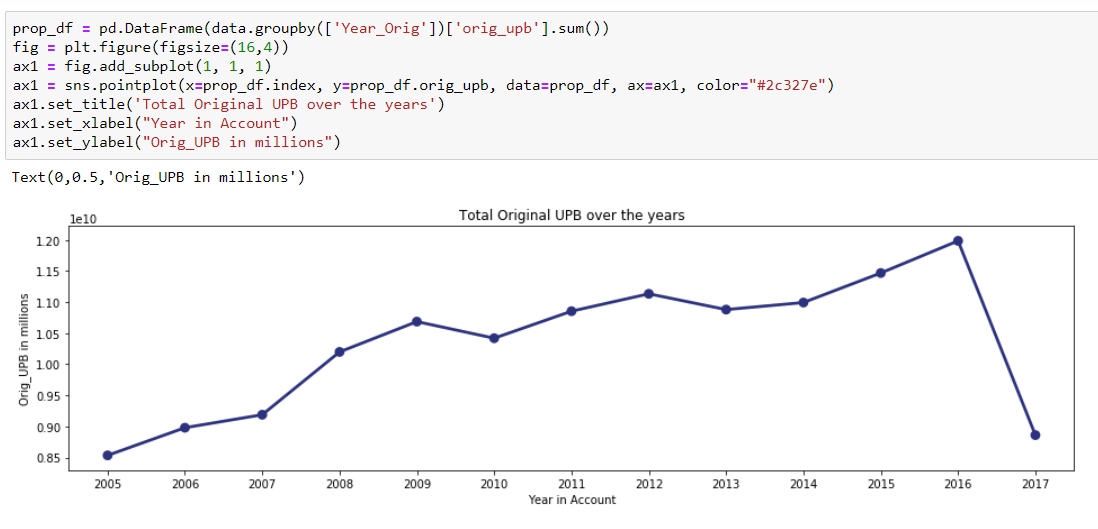
We noticed that the number loans in California were the highest.

* Average credit scores across states



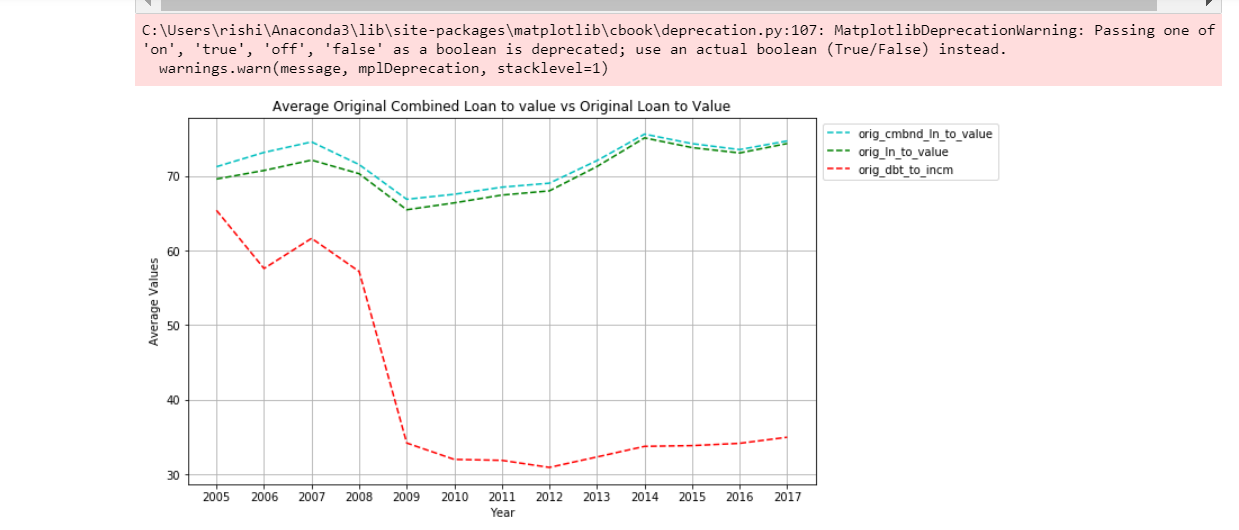
We noticed that credit scores in Georgia and Philadelphia are states where the credit scores were slightly less compared to other states.

* Total original UPB over the years



* Average original combined loan vs original loan to value





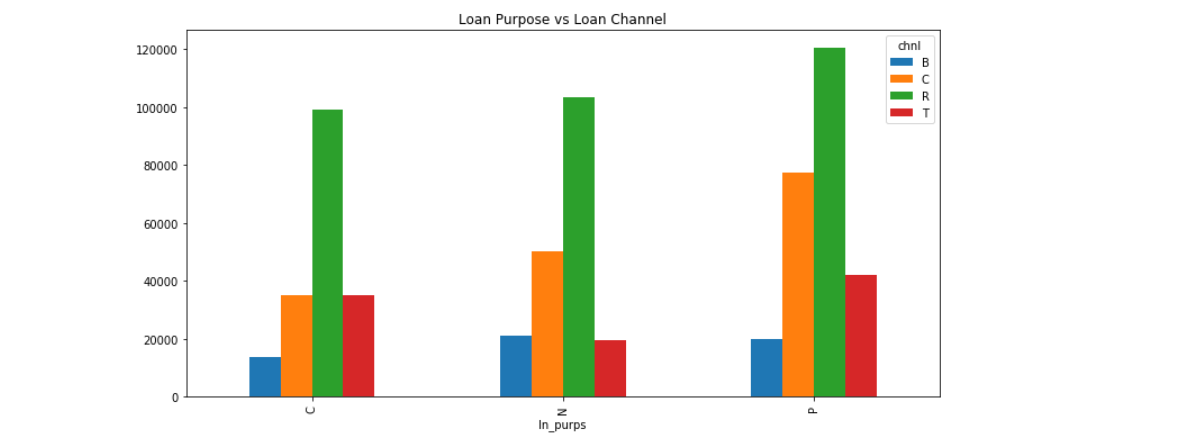
We notice that the debt to income ratio fell drastically in 2009 .

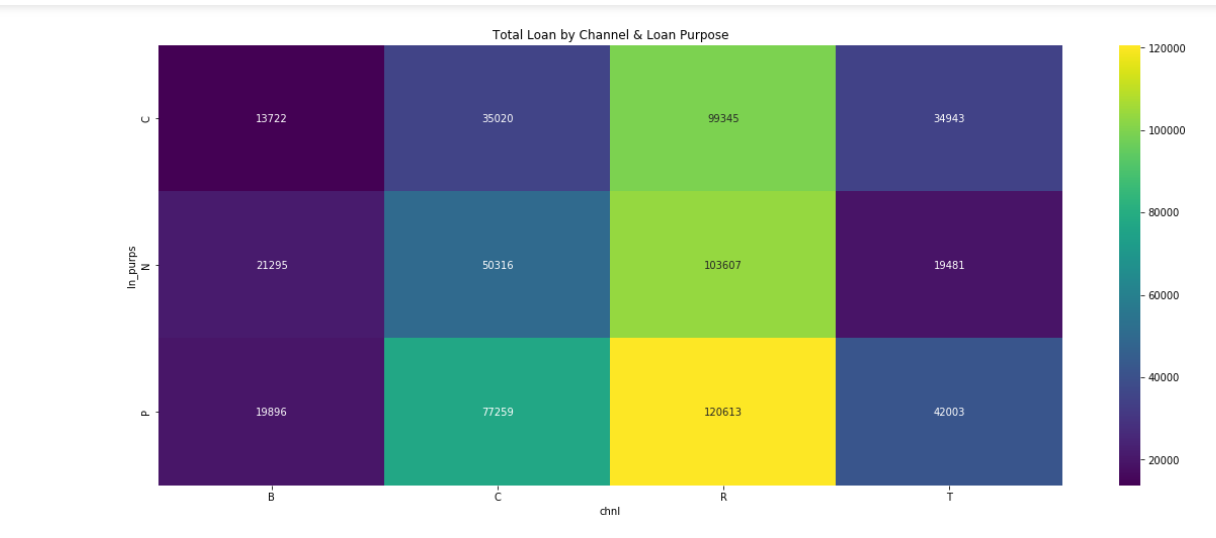
* First time loan counts



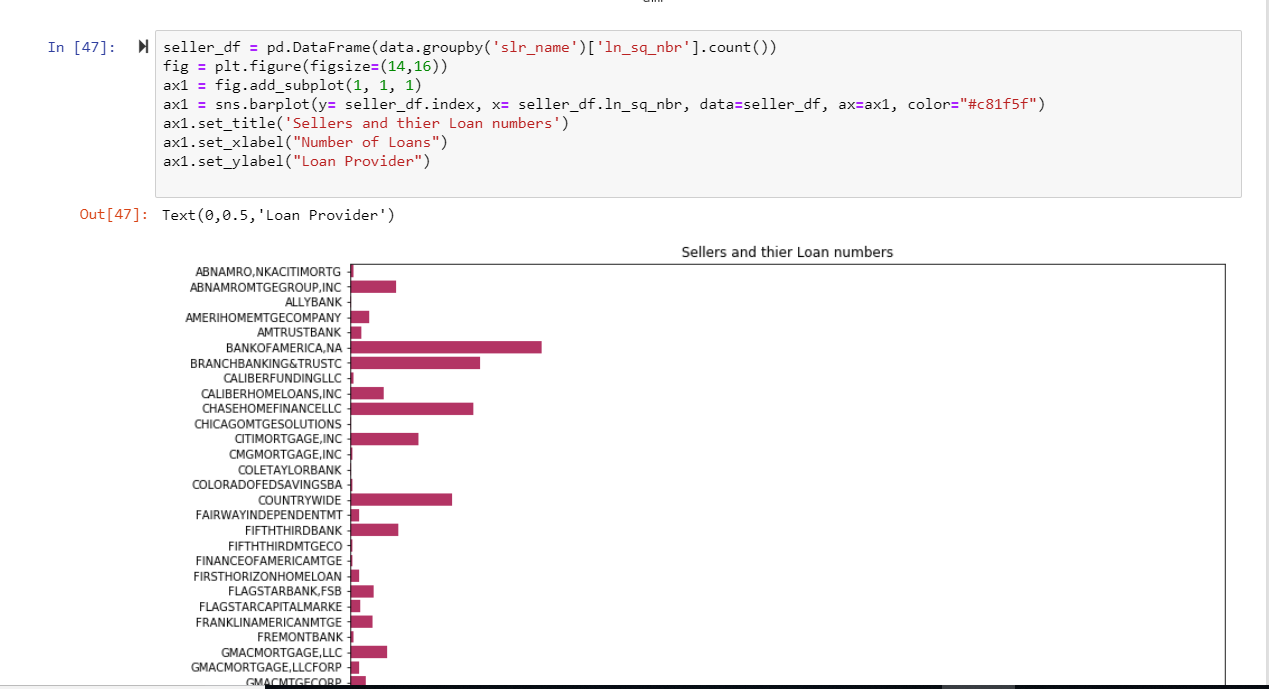
* Loan purpose vs loan channel



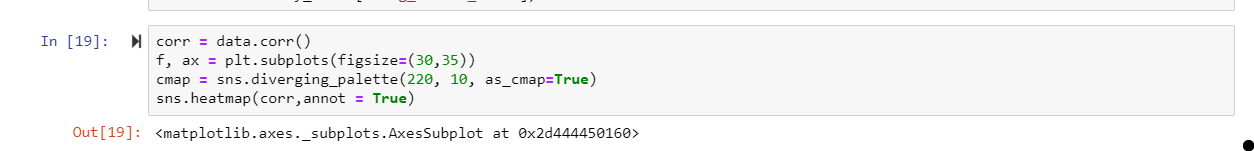


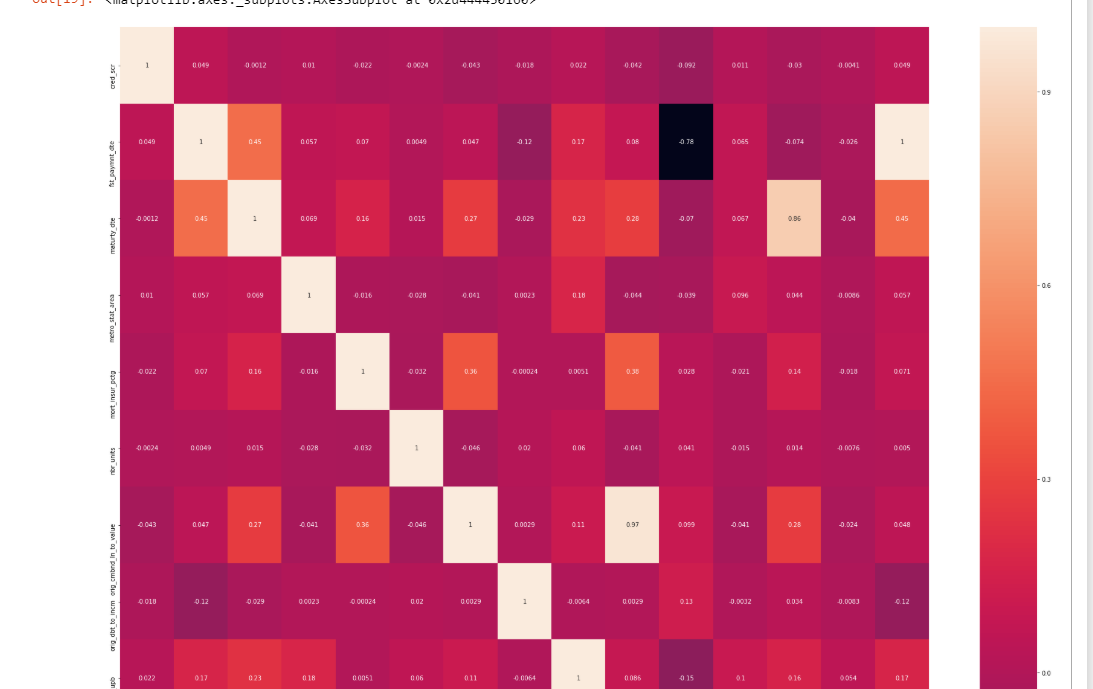


* Sellers and their loan numbers



* Correlation matrix



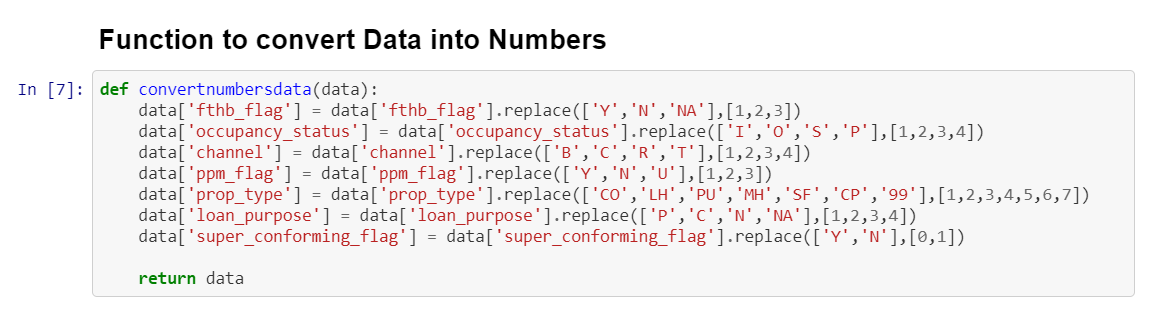


* Prediction:
* Here we will use the historical origination data to predict the interest rate for quarters.
* Here we will use Q1 2005 as a training data and we will predict the values for Q22005 quarter.
* We will calculate and evaluate different algorithms based on below parameters:
* **MAE (Mean Absolute 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.

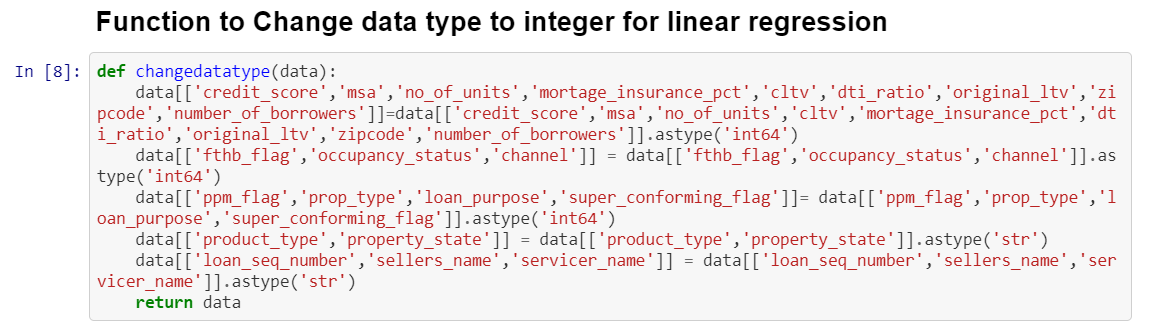
1. Building Model
2. Clean Data



1. Convert Data Into Numbers



1. Change Data Type to Integer



1. Perform Linear Regression

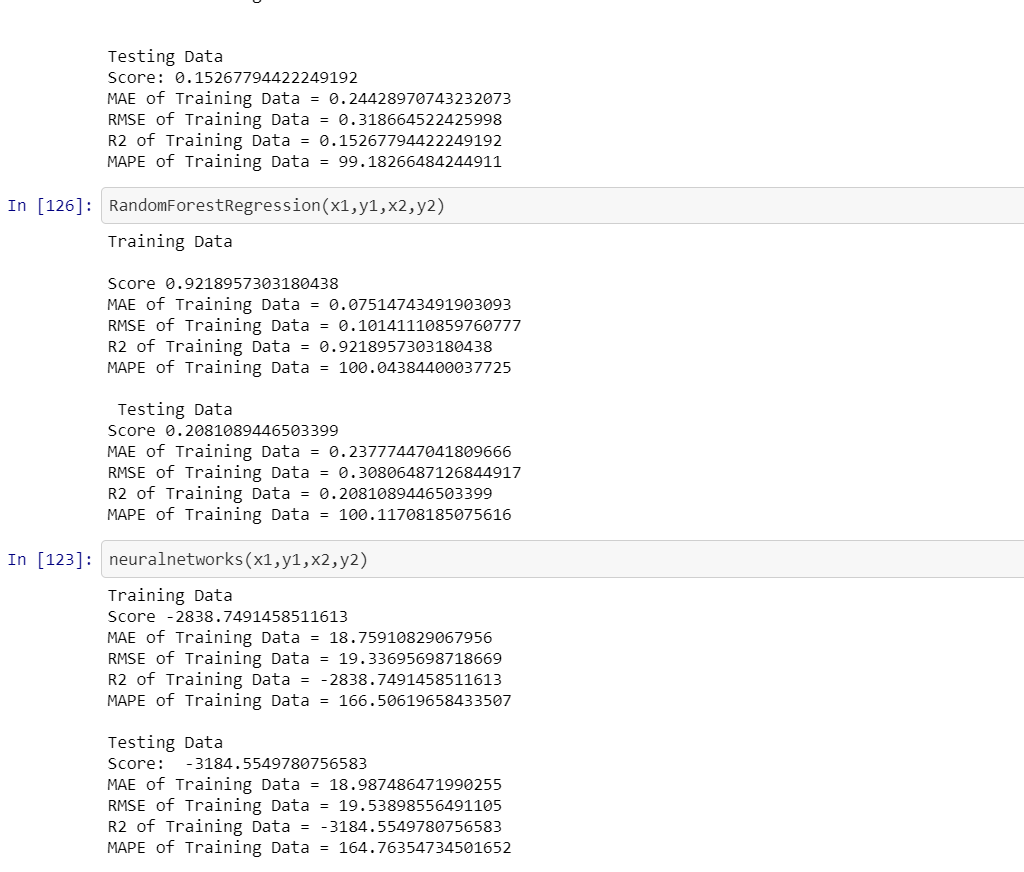


1. Perform Random Forest Regressor



1. Perform Neural Network



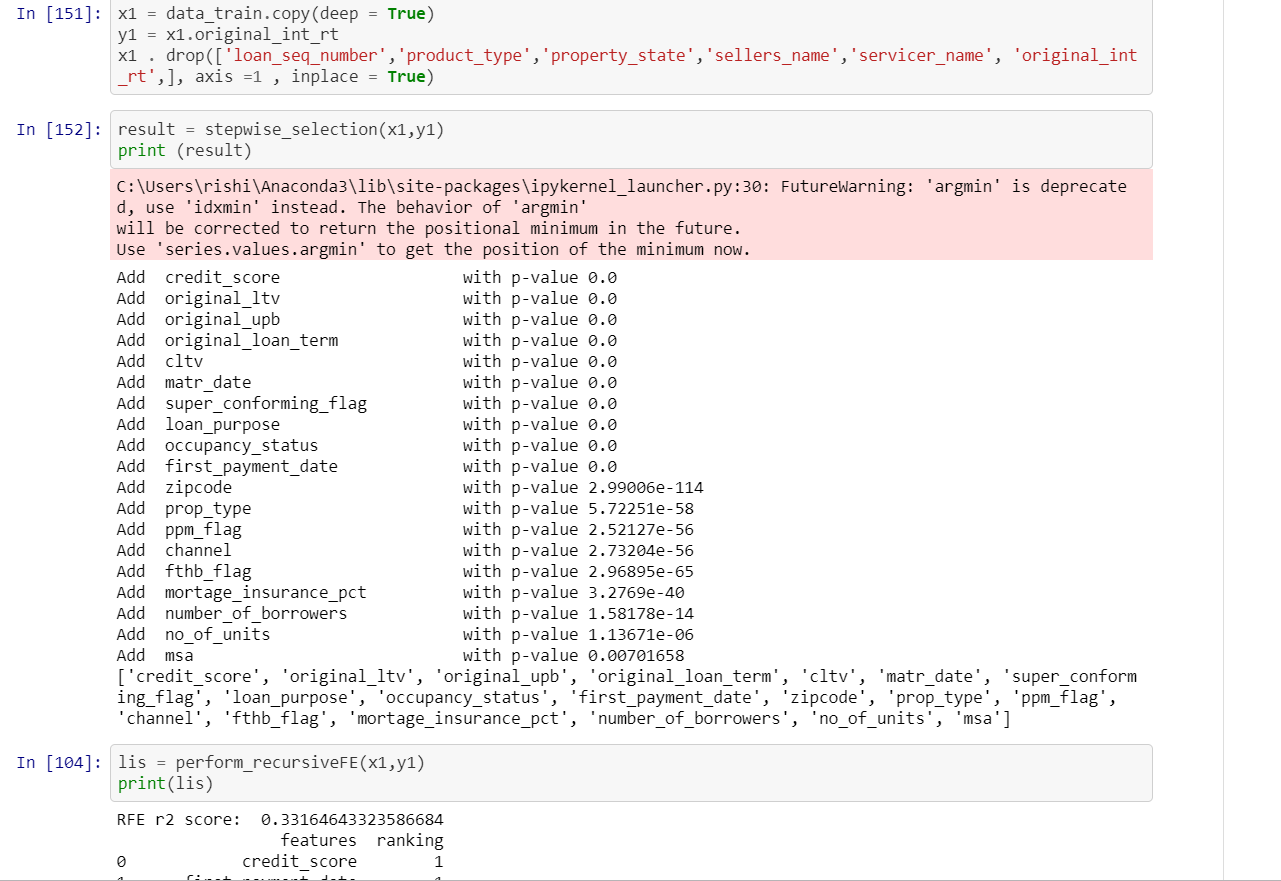


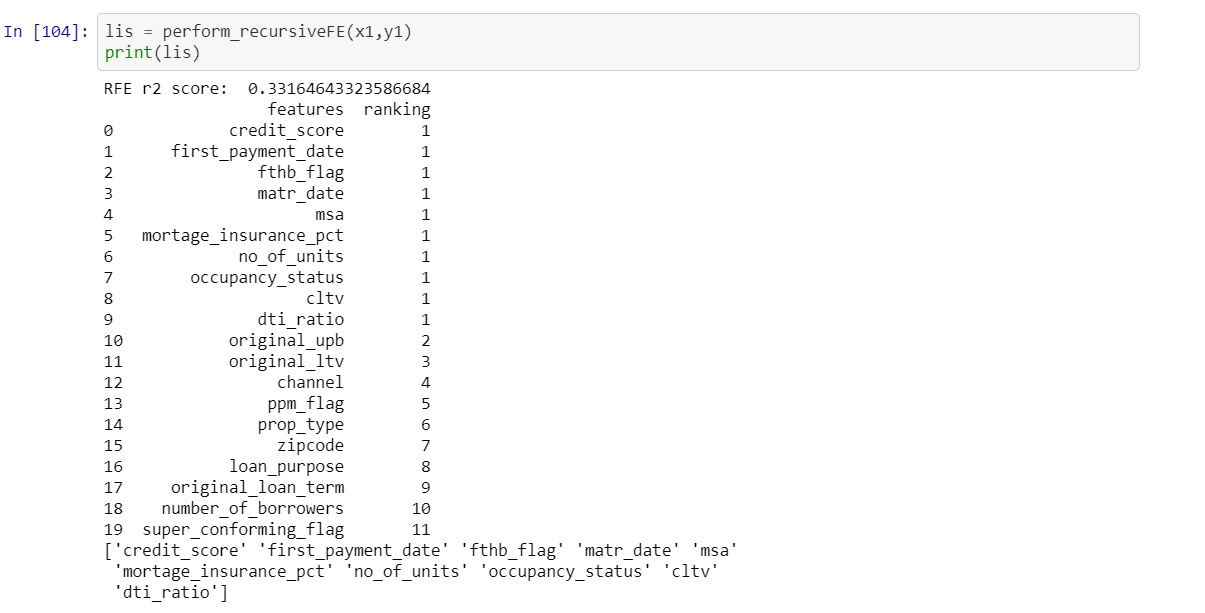
1. Feature Selection using Stepwise

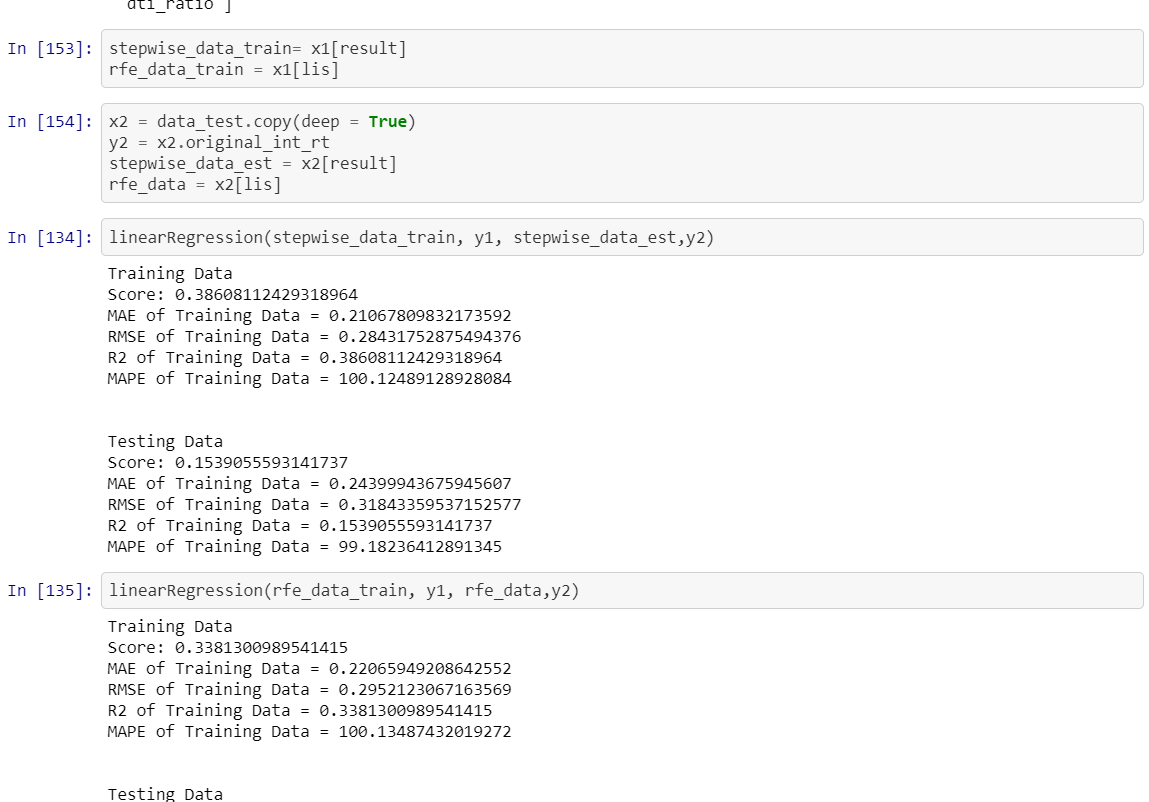


1. Feature Selection using RFE







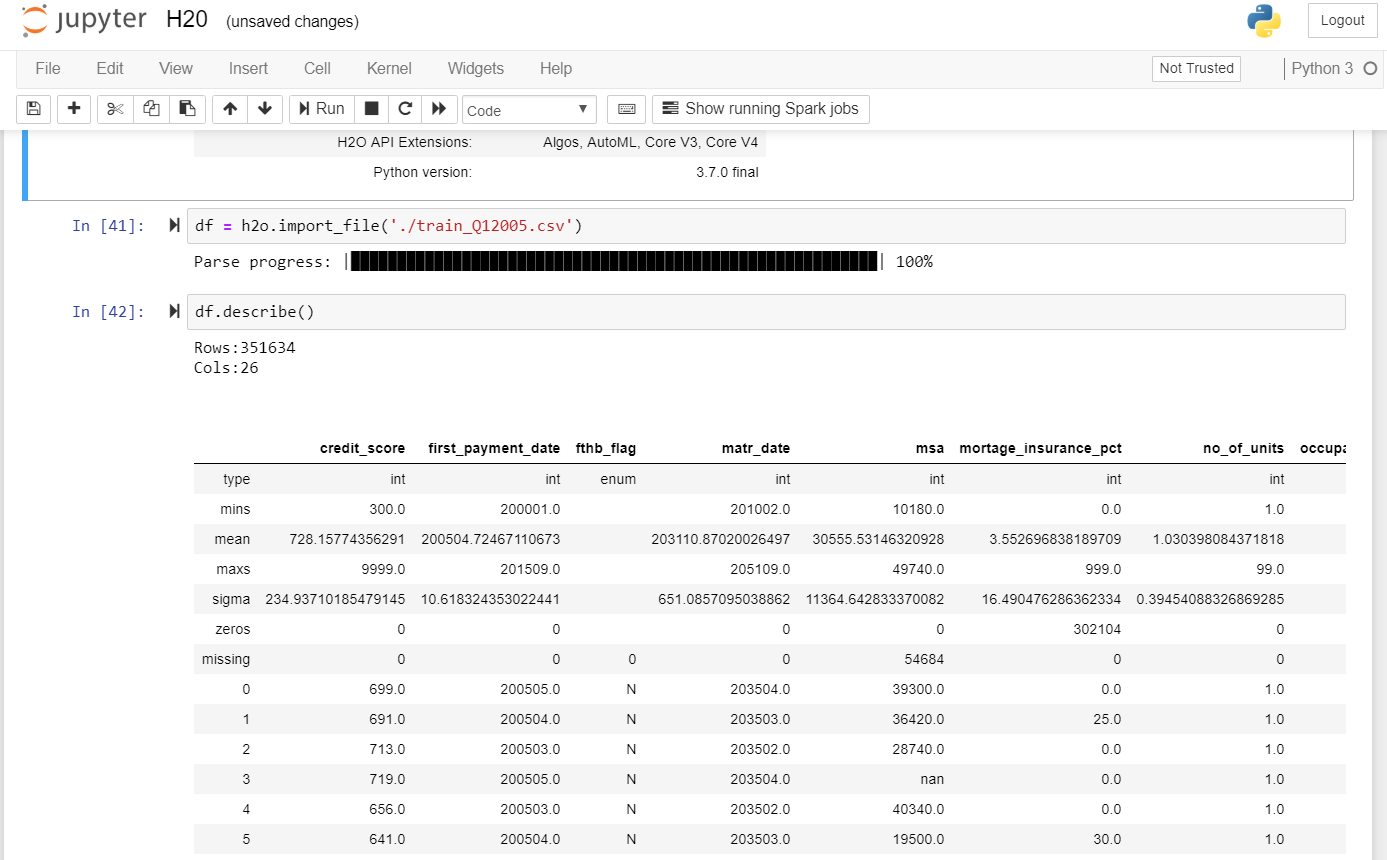


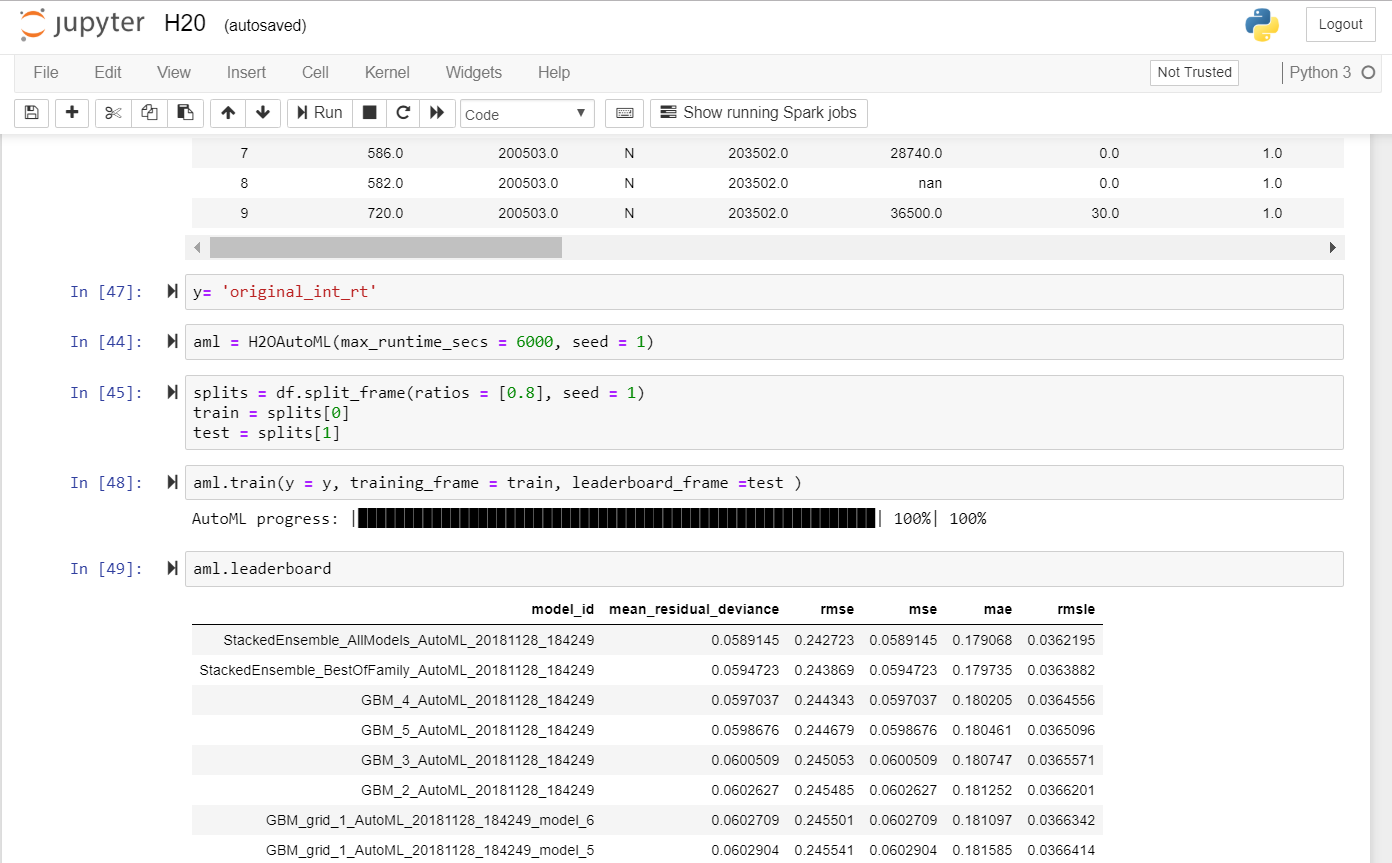
From above three algorithms we chose Random Forest because:

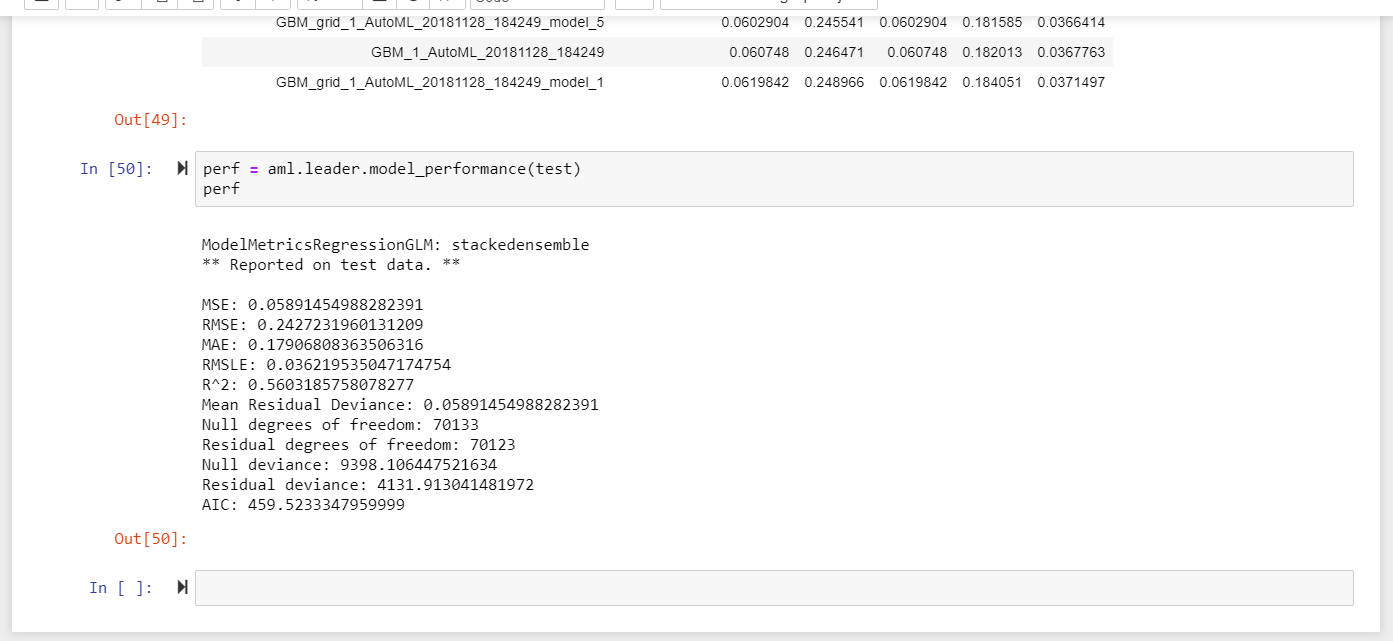
* Better results than Linear Regression
* Lot less processing time than Neural networks (Fast and scalable)

Furthermore,

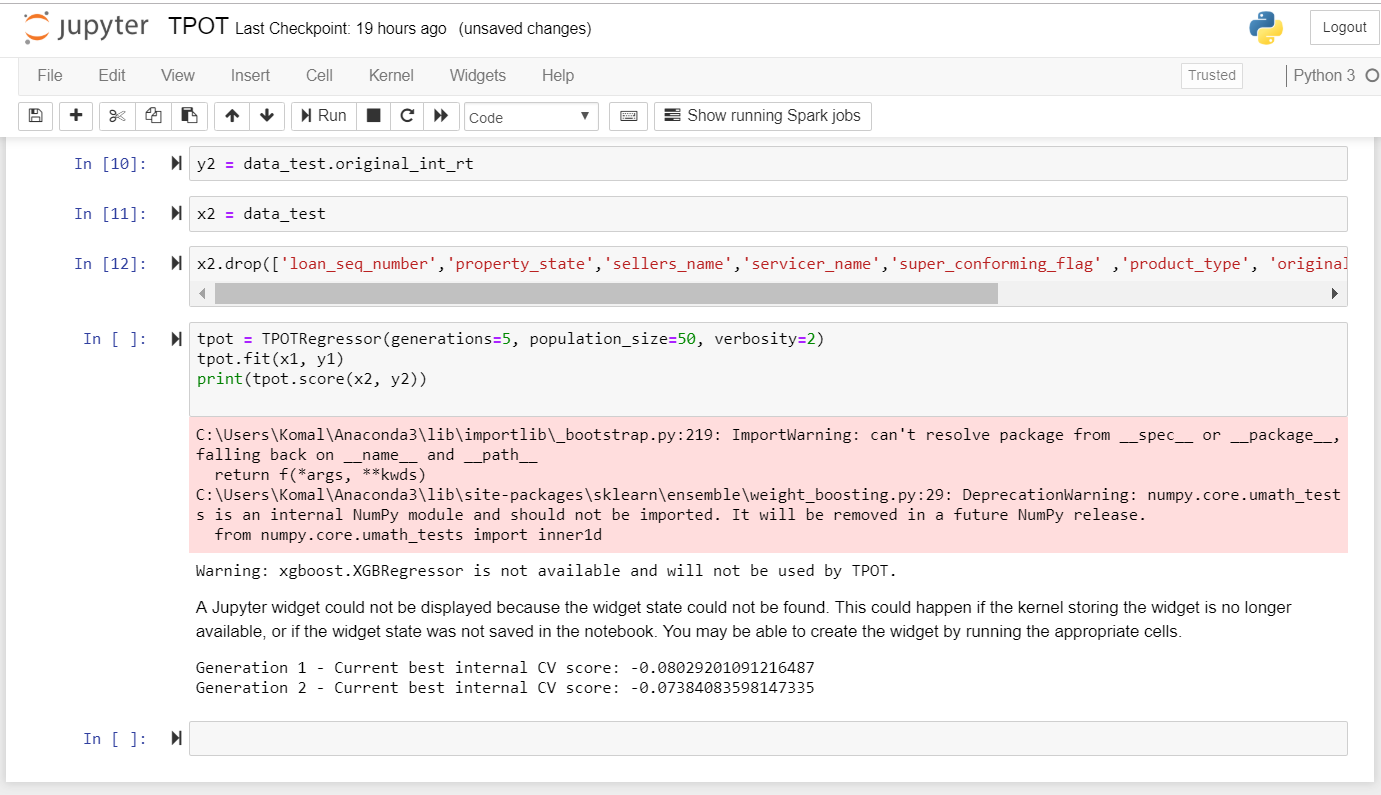
* Processing time does not increase substantially with increase in number of observations.
* Easy to interpret ,adjust (tune) parameters to achieve desired results.
* It is Non-parametric ,we don’t have to worry about outliers.
* H2O.AI







* TPOT:



* 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.

* Programmatically downloading the historical data based on user input:



* Cleaning the dataframe:



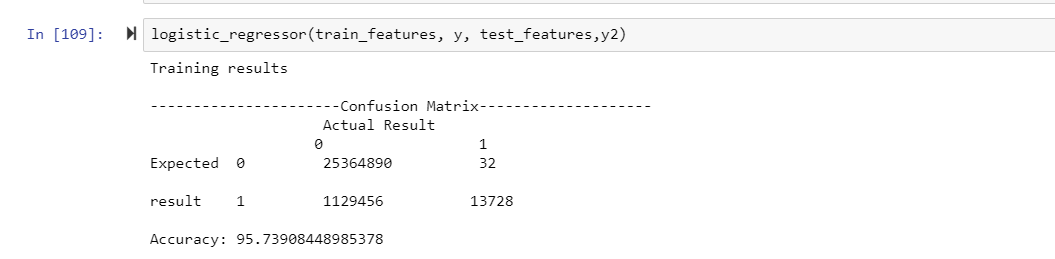
Adding delinquent column:

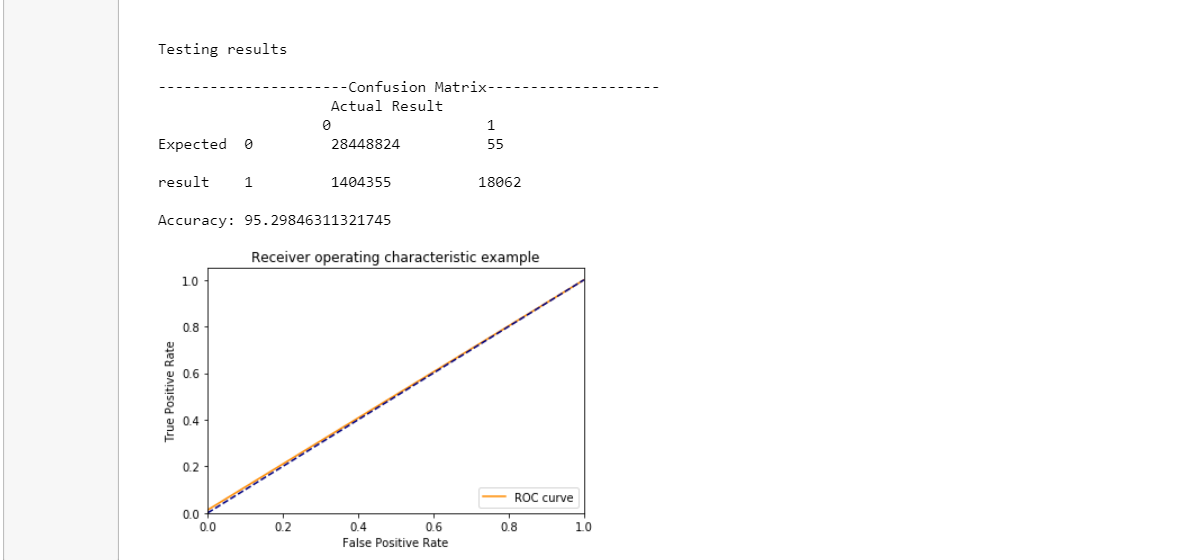
‘Deliquent’ column is added based on delq\_status. If delq\_status is greater than 0 then Deliquent = 1 else the value of Deliquent is 0.



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

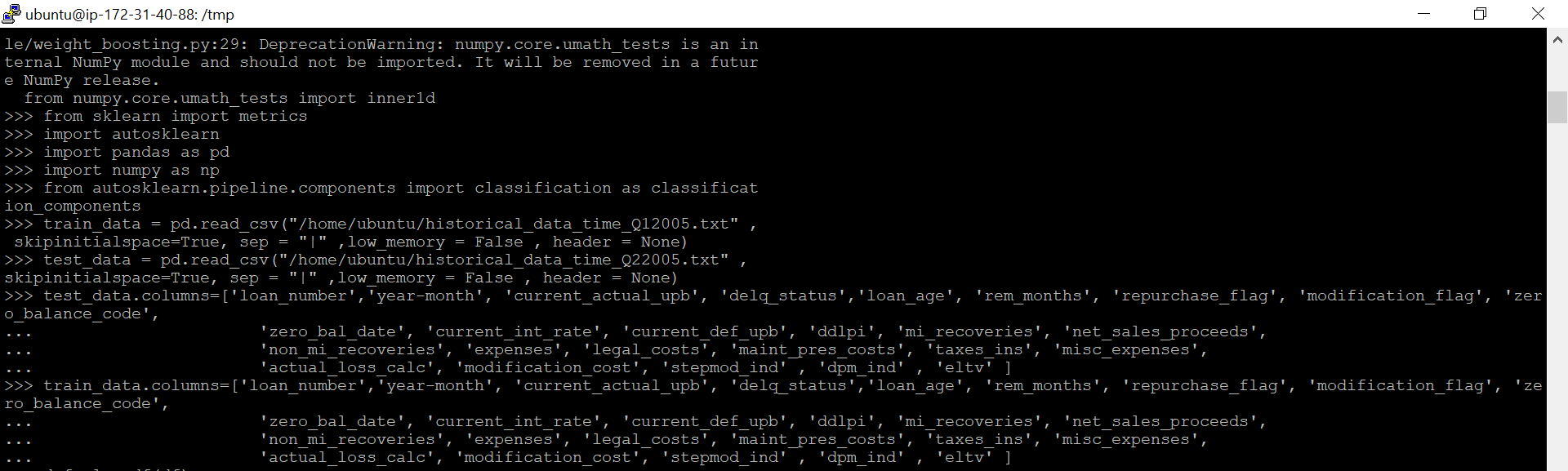
Artificial neural networks are relatively crude electronic networks of neurons based on the neural structure of the brain. They process records one at a time, and learn by comparing their classification of the record (i.e., largely arbitrary) with the known actual classification of the record.



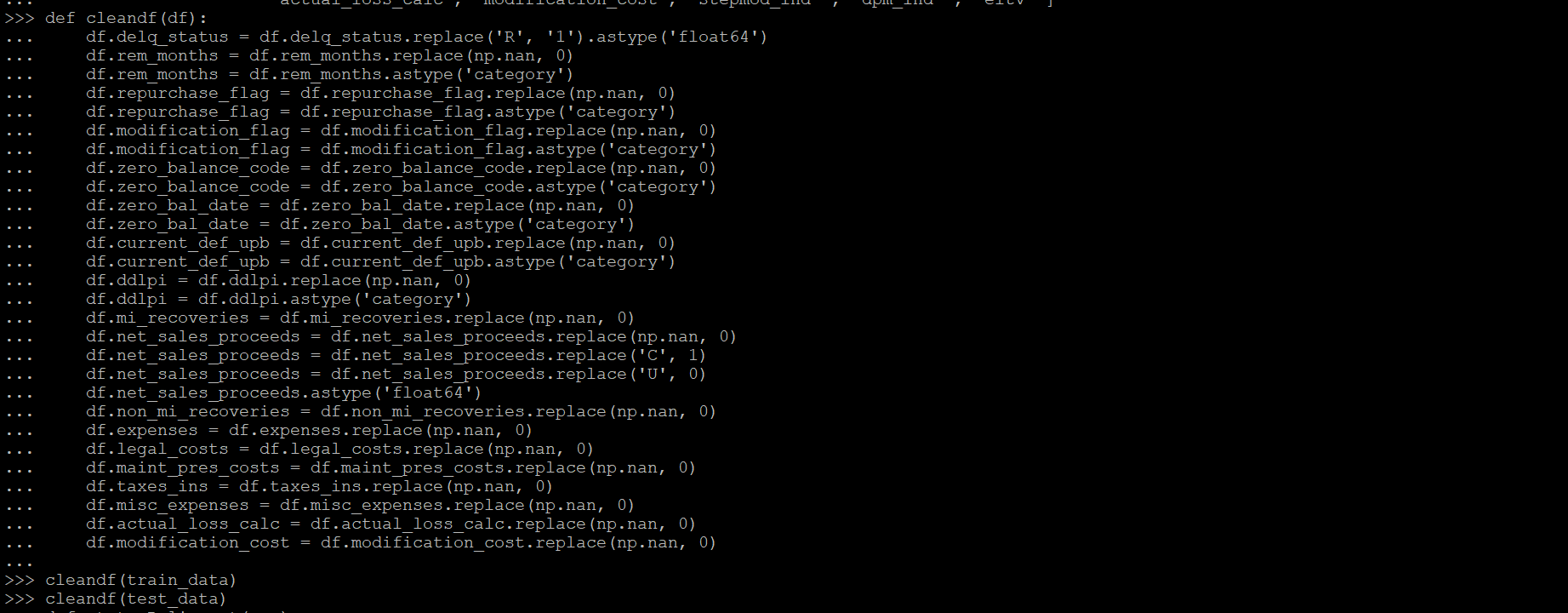
AutoSklearn:

AutoSklearn for classification can be executed on Ubuntu. This execution is done on Amazon EC2 instance on Ubuntu 2xlarge.

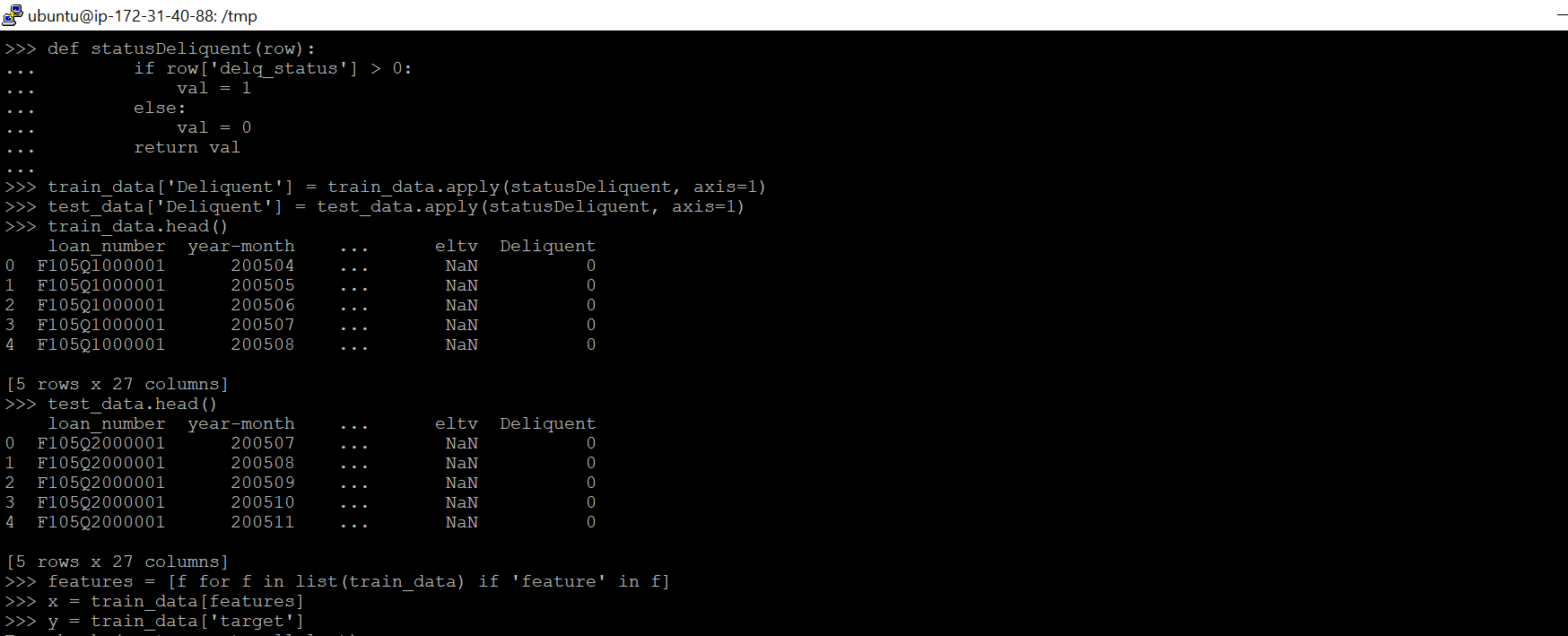
All the scikit learn libraries for classification are installed and imported. The train and test data is then read in a dataframe.

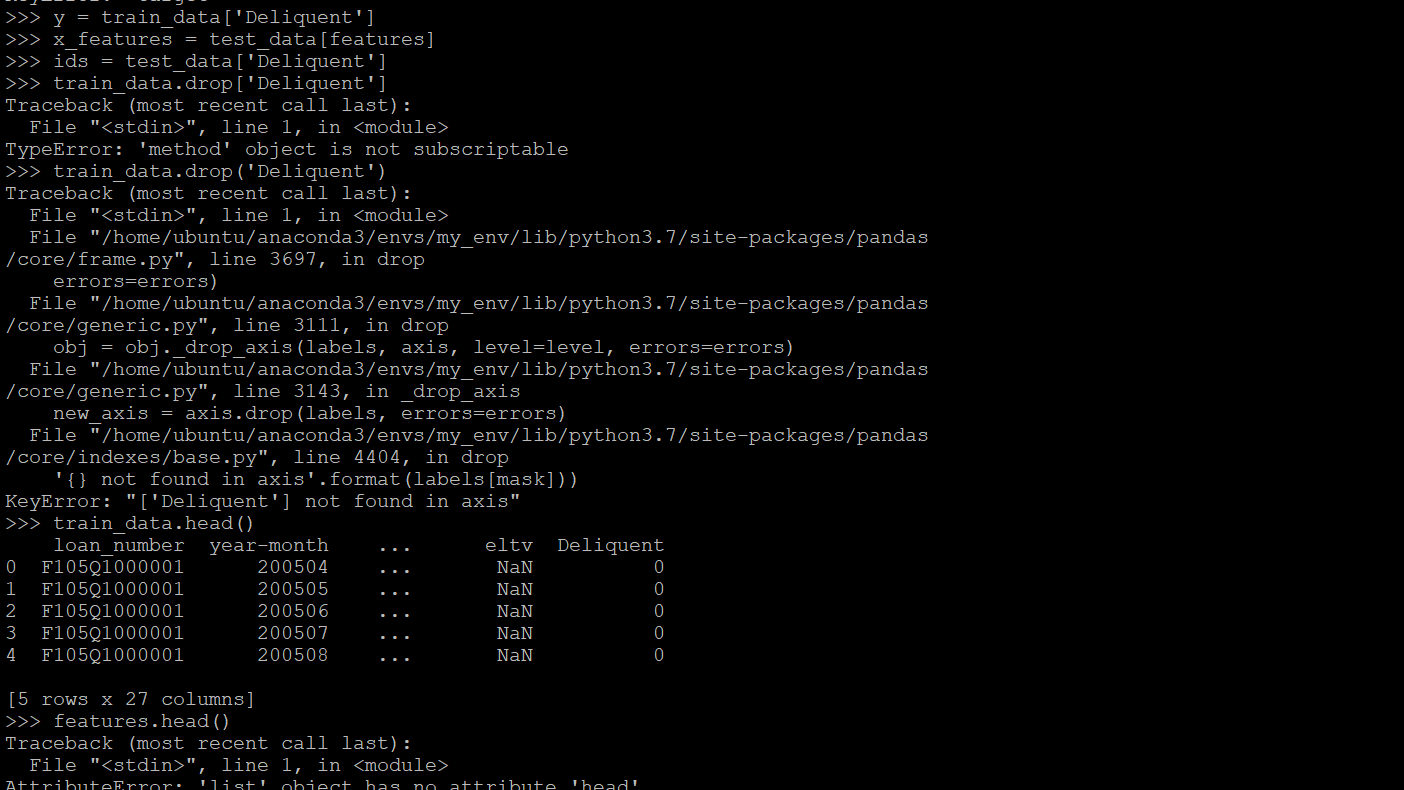


The train\_data and test\_data dataframe is cleaned with necessary datatypes.



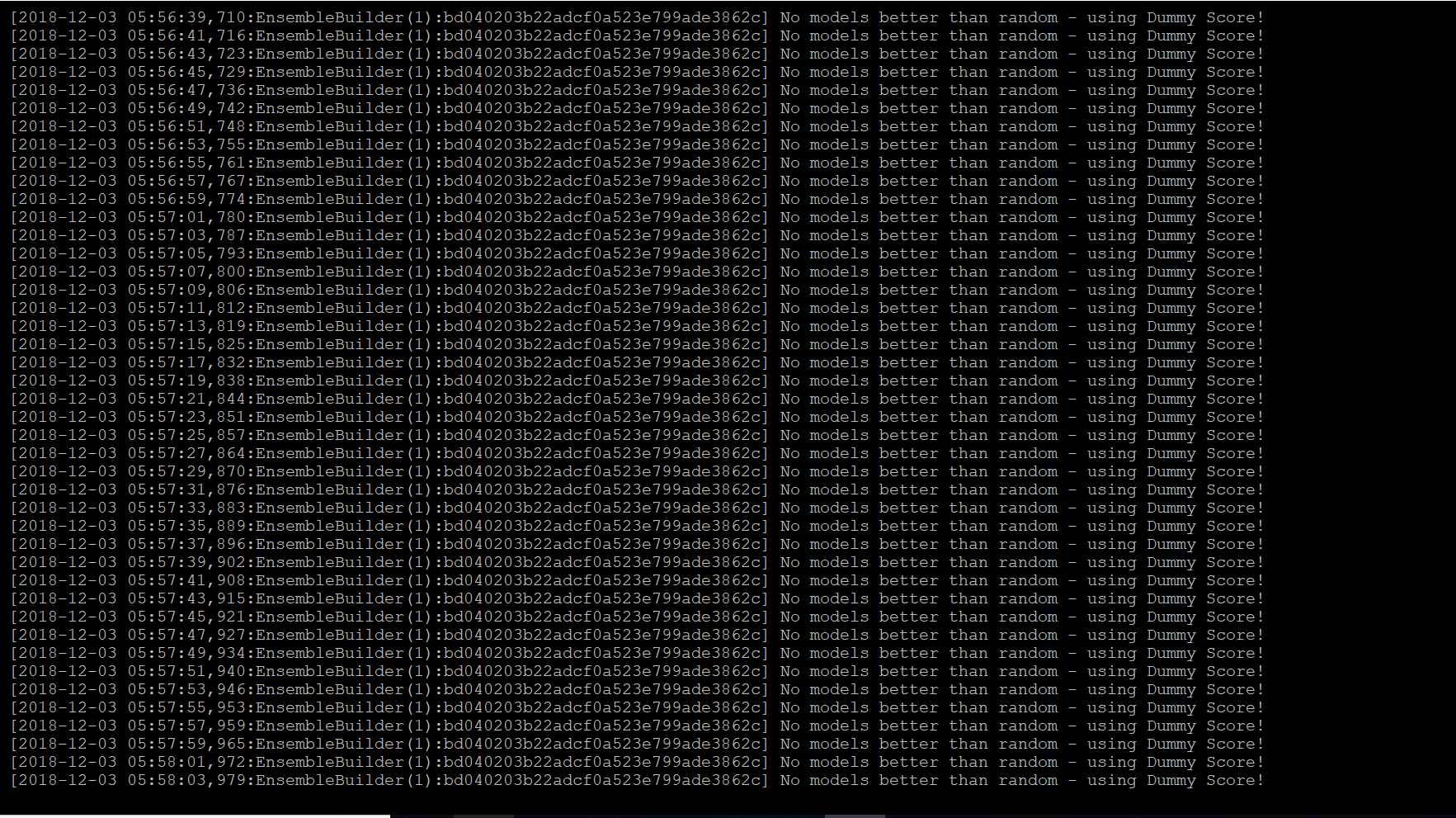
Adding the ‘Deliquent’ column to the dataframe based on delq\_status/





Fitting the variables X and Y to the model autosklearn.





Conclusion: We get the best model as Random Forest.

TPOT:

A screenshot of a cell phone

Description generated with high confidence