

Final Project

Online news popularity Prediction

Chitra Ramchand Paryani | Advances in Data Science | 06/08/2018

# Contents

1. Introduction
2. Data Preparation
   1. Loading libraries
   2. Loading Dataset
   3. Exploratory Data Analysis
   4. Data Distribution
   5. Data Sampling
   6. Initial Regression Analysis
   7. Data Cleanup
3. Data Modelling
   1. Logistic Regression – All 53 Predictors
   2. Logistic Regression – Top 30 Predictors
   3. Regularization & Cross Validation
   4. SVM – All 53 Predictors
   5. SVM – Top 30 Predictors
   6. SVM – Hyper Parameters tuning & Cross Validation
   7. Random Forest – All 53 Predictors
   8. Random Forest – Top 30 Predictors
   9. Random Forest – Cross Validation
   10. Naïve Bayes – All 53 Predictors
   11. Naïve Bayes – Top 30 Predictors
   12. Naïve Bayes – Cross Validation
   13. Neural Network – Feed Forward Algorithm – All 53 Predictors
   14. Neural Network – Feed Forward Algorithm – Top 30 Predictors
   15. Bagging, Boosting and Stacking
4. Results

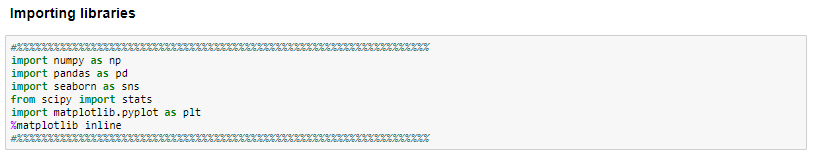
# Introduction

For performing analysis, I have used Python Programming language. All the code is embedded below along with results.

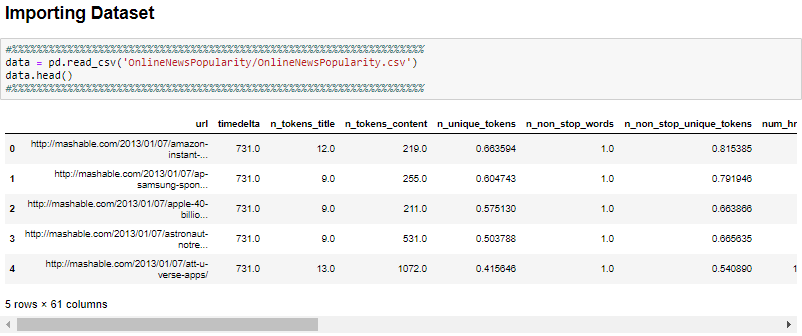
I am using Online News Popularity data set from UCI Machine Learning Repository and it consist of 61 Attributes and 39797 instances of data.

# Data Preparation

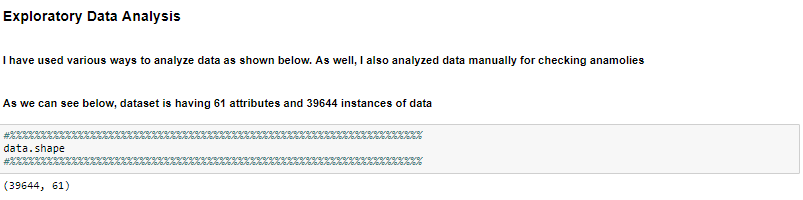
## Loading libraries

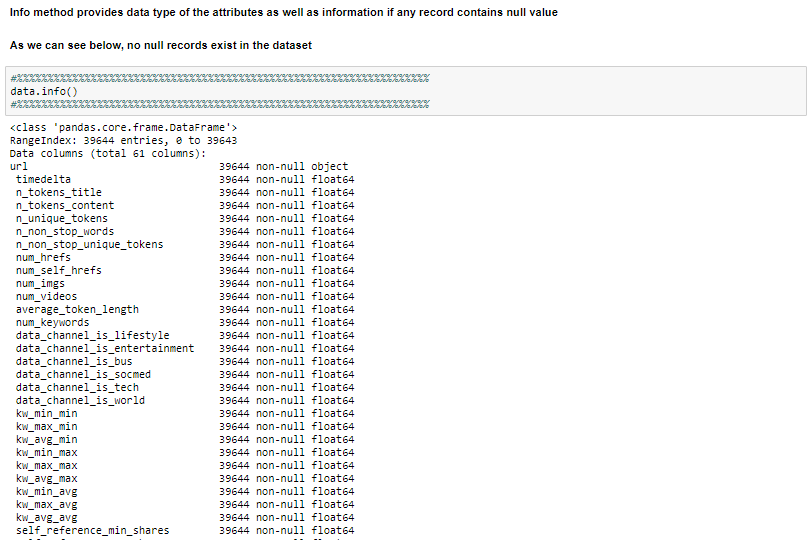


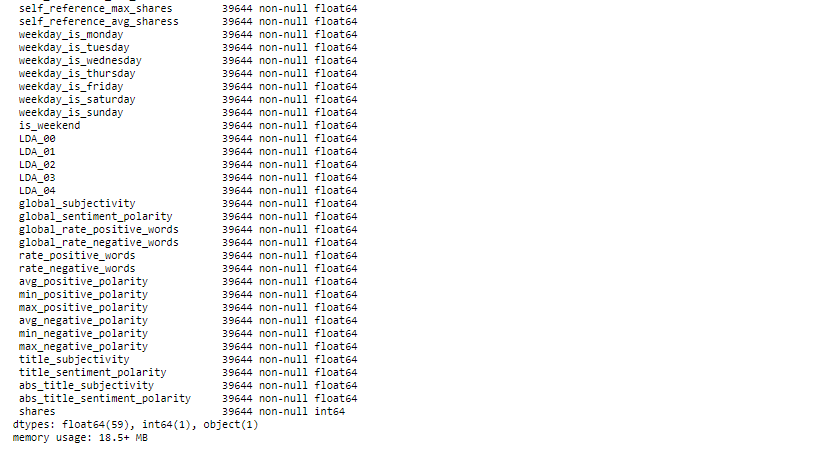
## Loading dataset

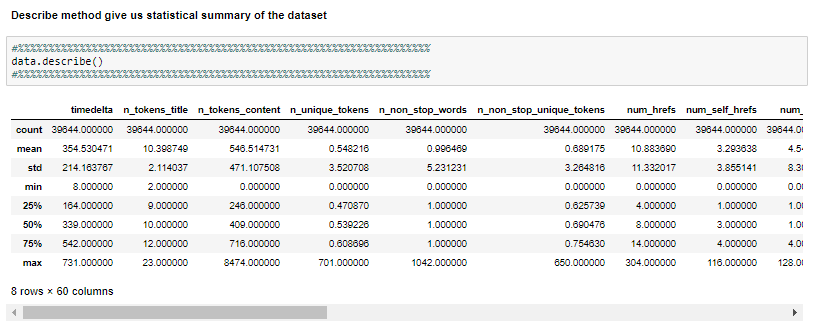


## Exploratory data analysis

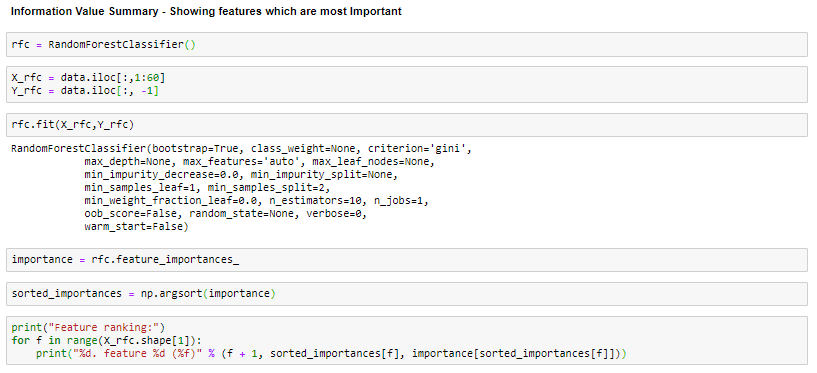


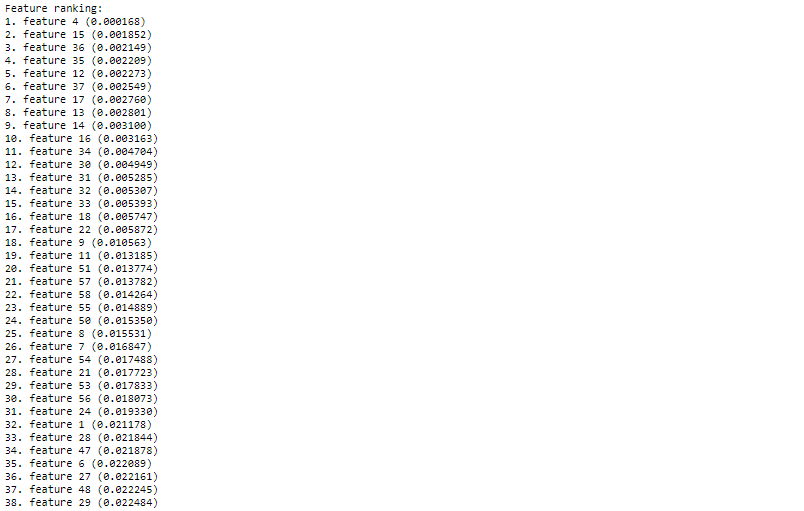




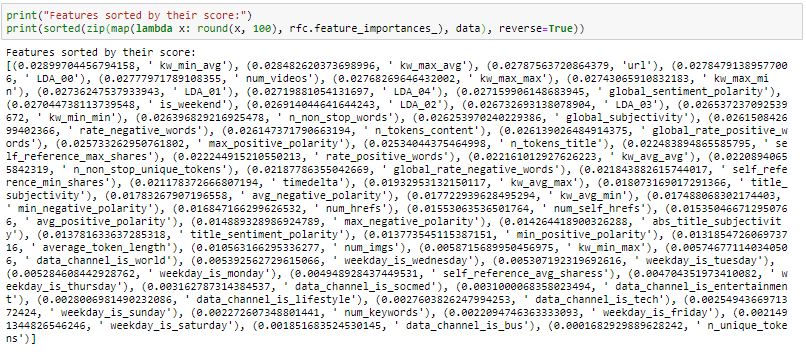


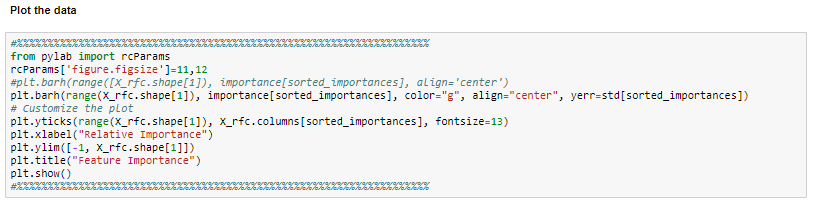
Next, predicting the features which are of highest importance

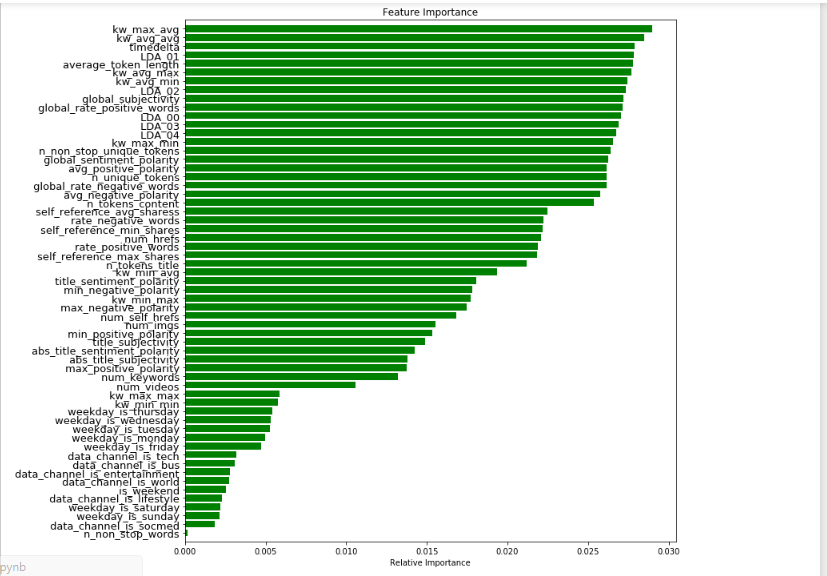




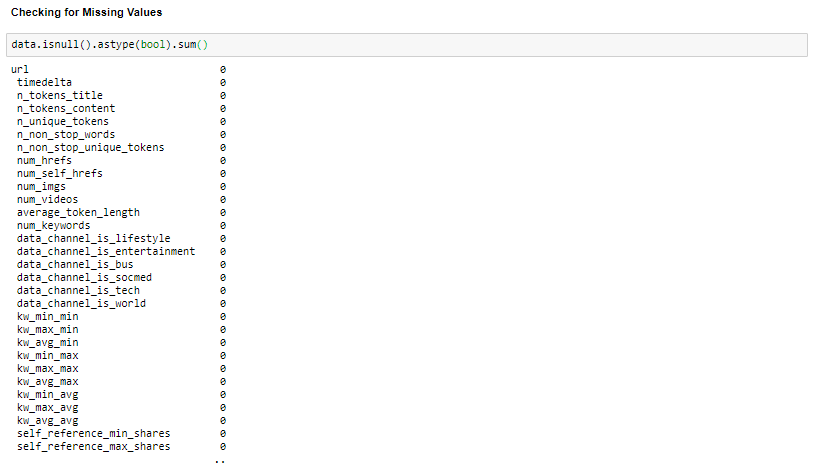


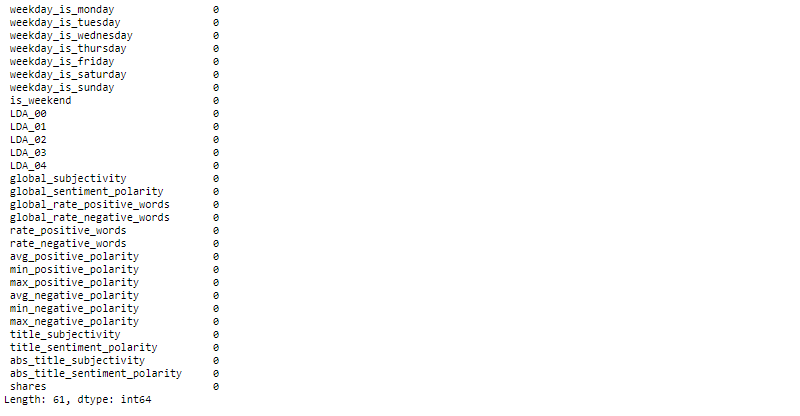




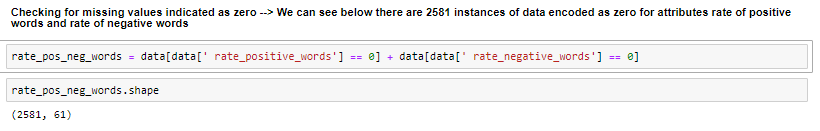


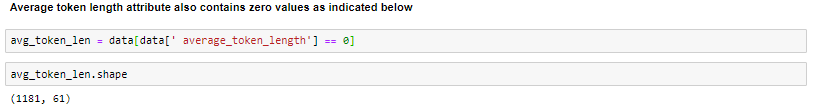
Next, checking for missing values and as we can see below data doesn’t contains any missing values



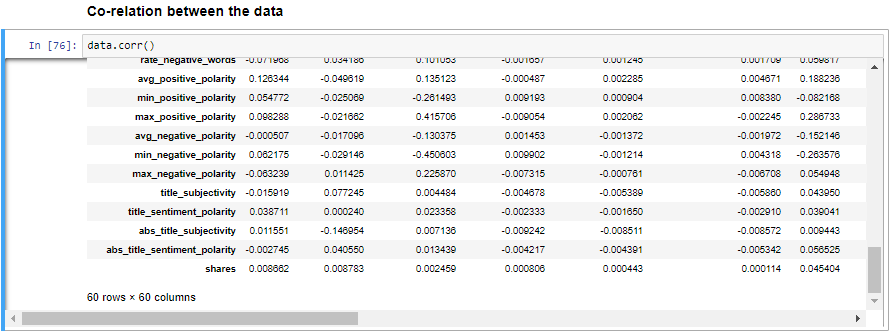


Next, checking for missing values indicated as zero

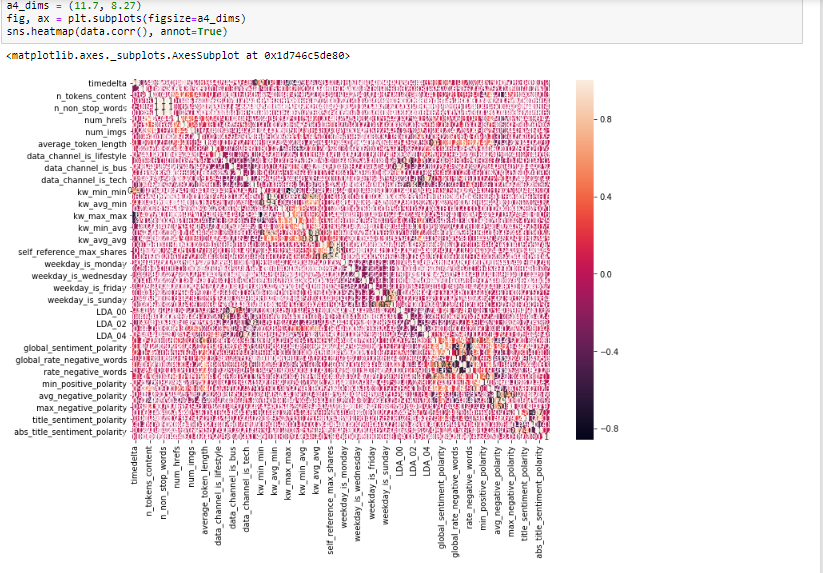




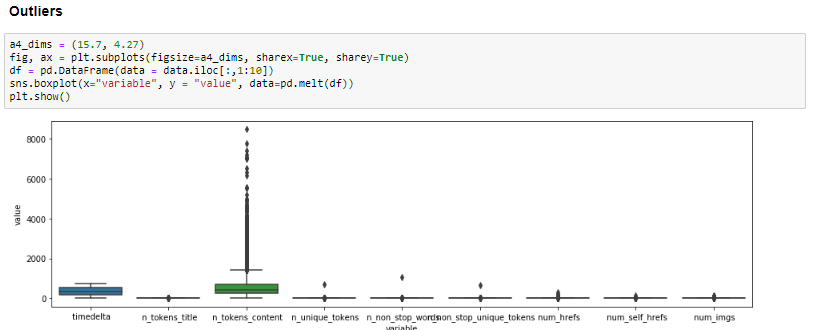
Next, checking co-relation between the data

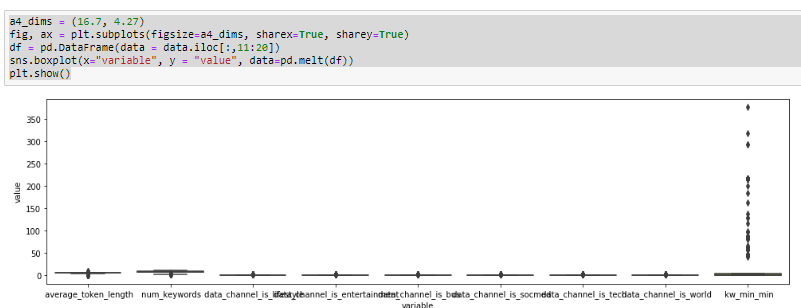


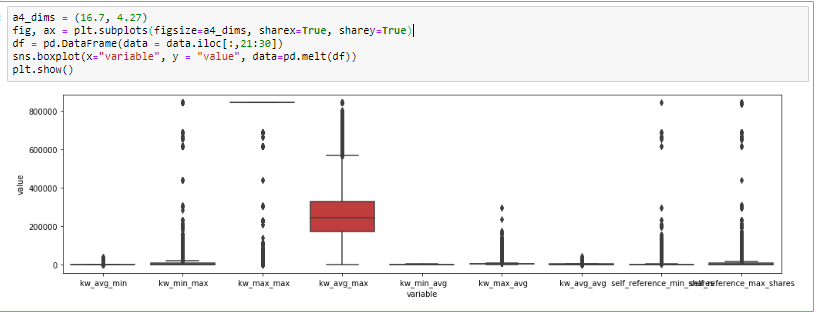
Next, generated heat map which explains data set contains few highly co-related data which we will remove while doing data cleanup

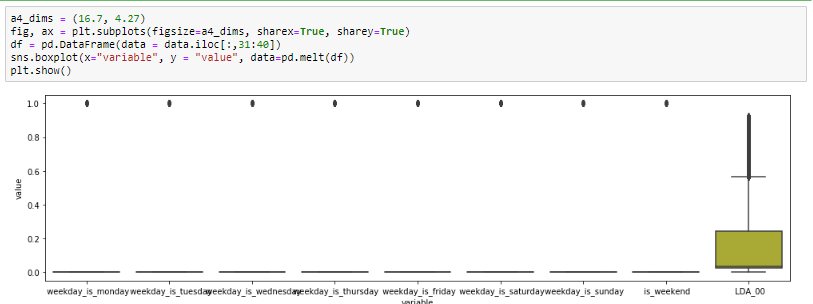


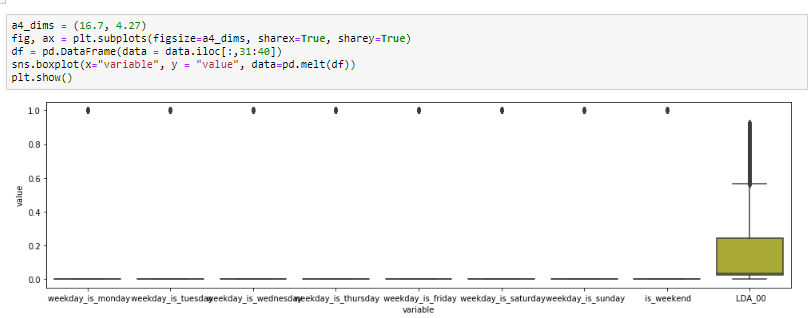
Next, checking for outliers in the data

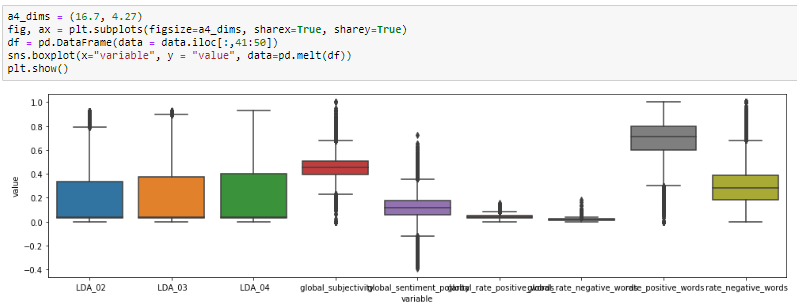


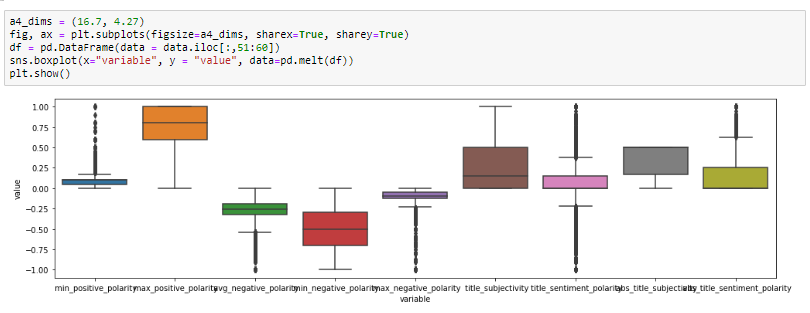




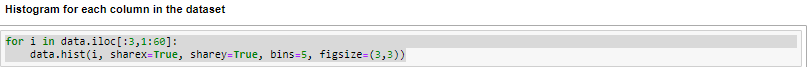


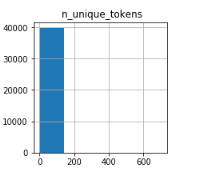
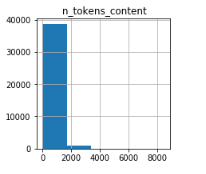
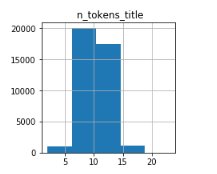
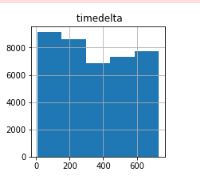


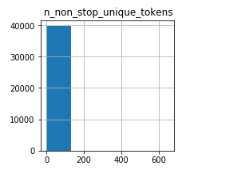
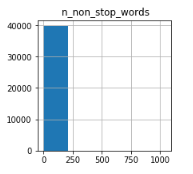


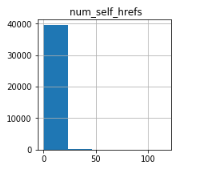
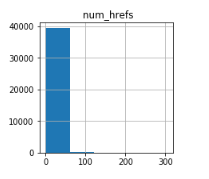


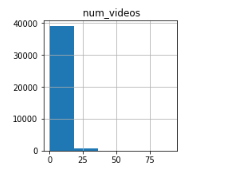
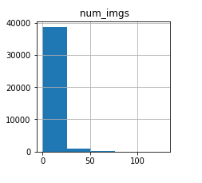
Histogram for each column in data set

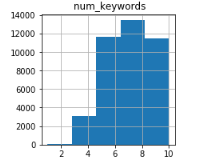
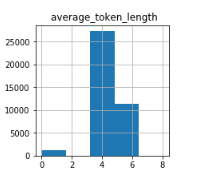


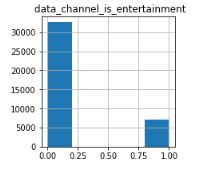
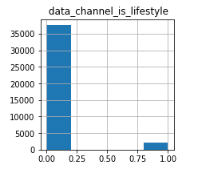


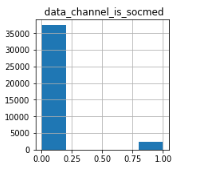
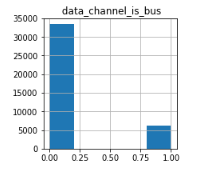


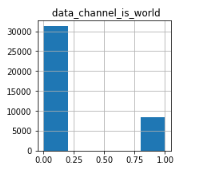
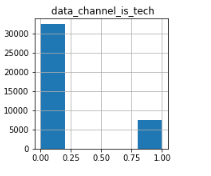


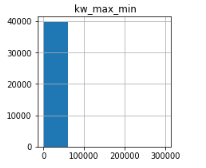
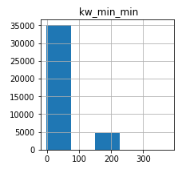


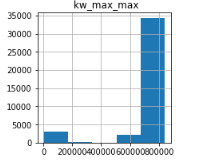
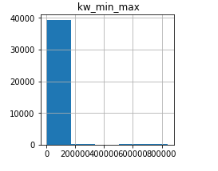


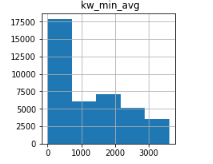
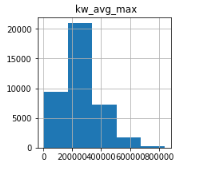


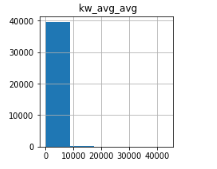
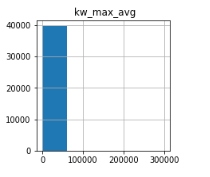


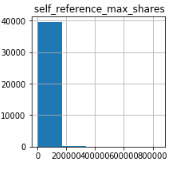
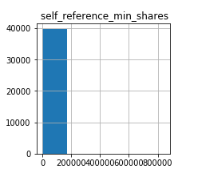


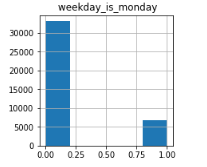
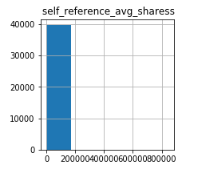


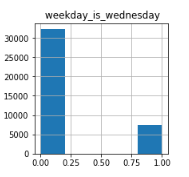
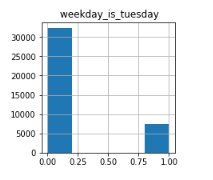


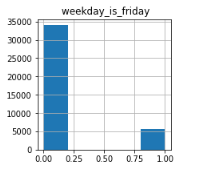
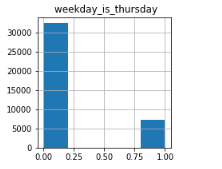


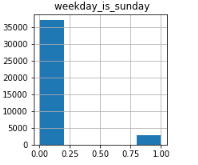
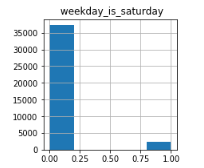


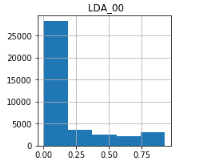
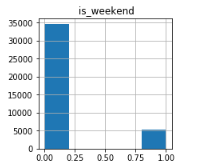


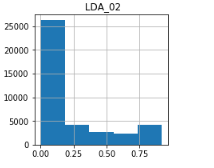
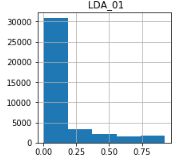


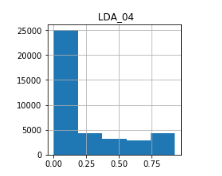
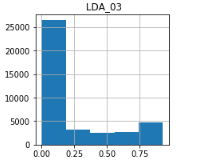


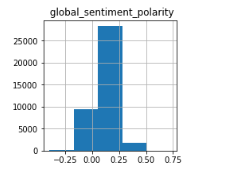
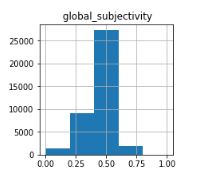


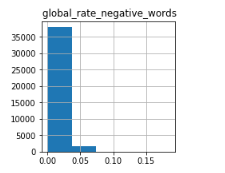
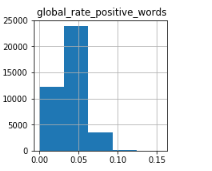


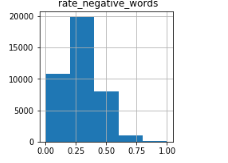
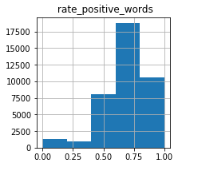


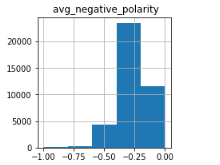
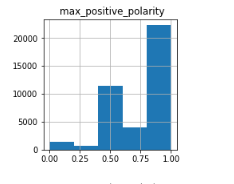
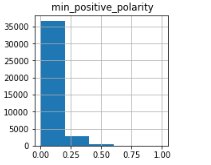
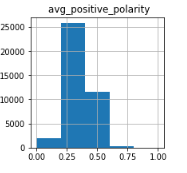


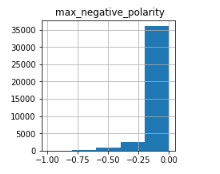
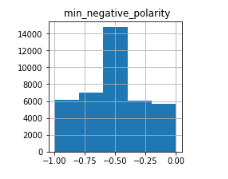


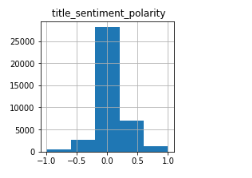
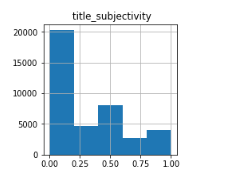


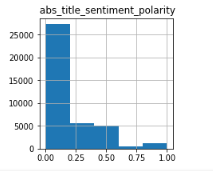
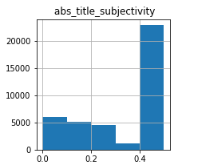






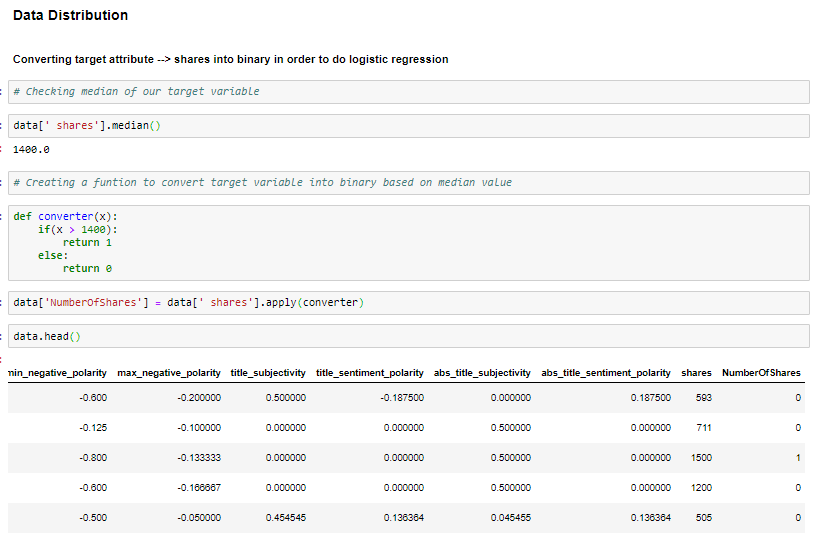


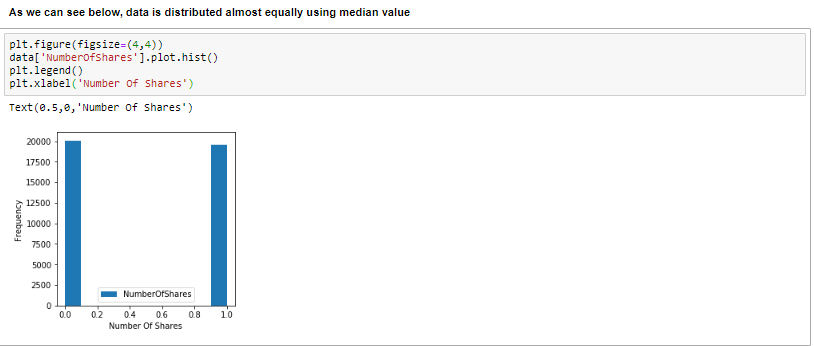




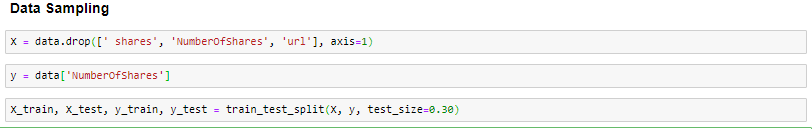
Next, I have converted target variable shares into binary by creating a function

## 2.4 Data distribution

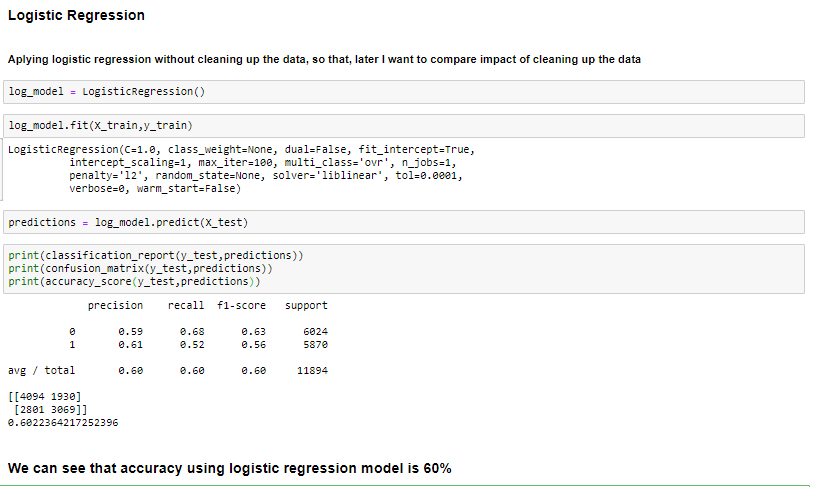




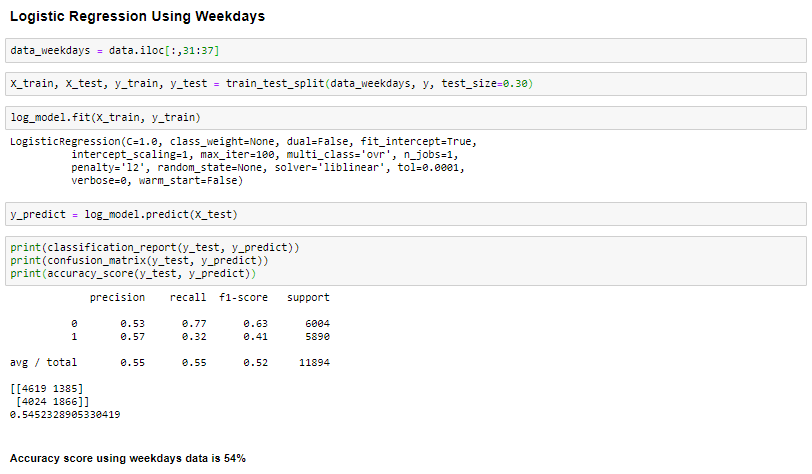
## data sampling



## logistic regression

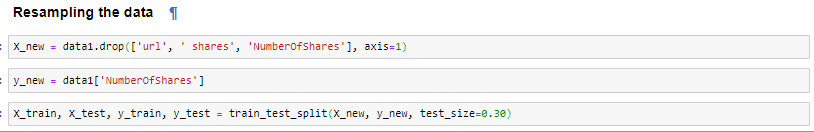






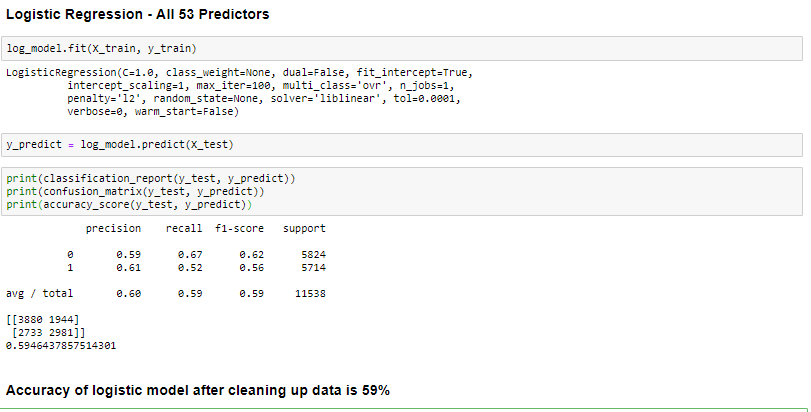
## Data clean up



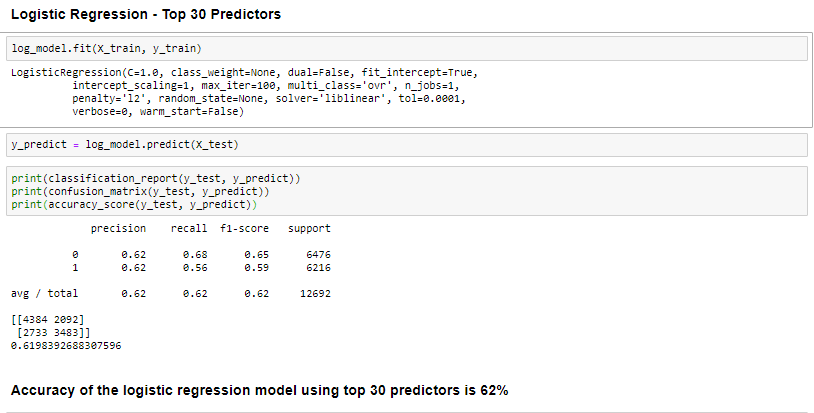


# Data Modeling

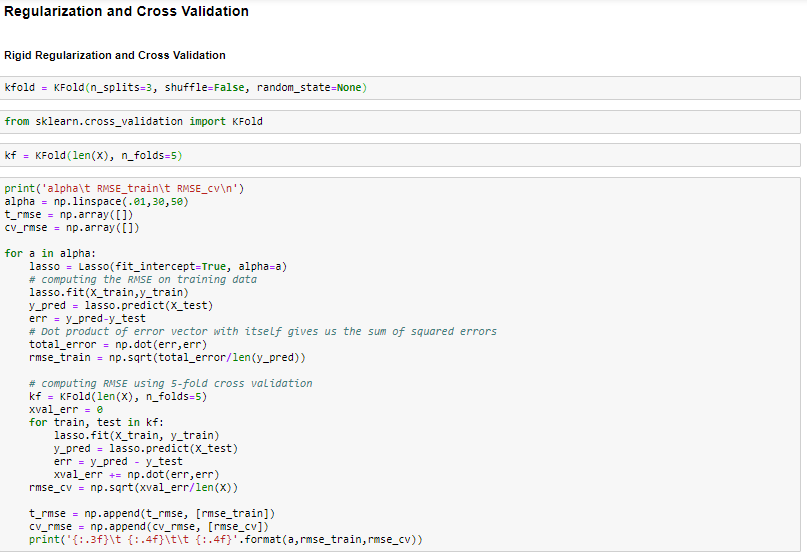
## logistic regression – all 53 predictors

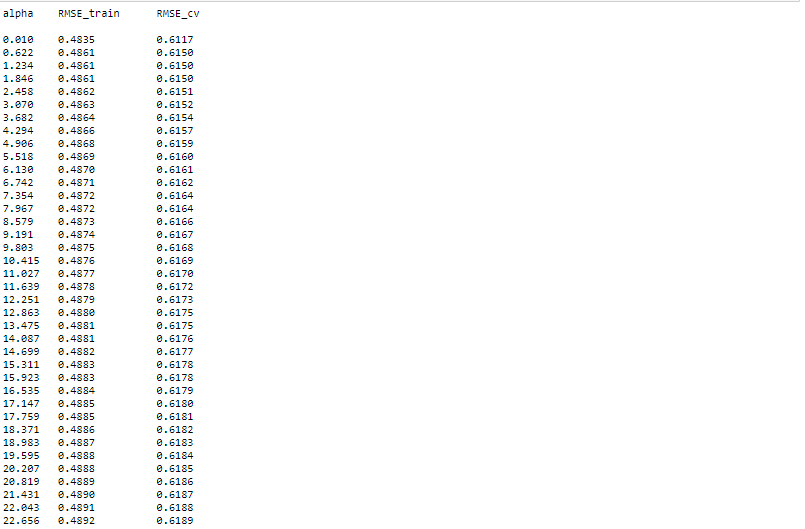


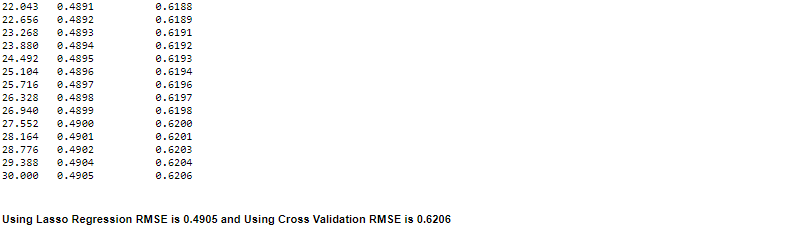
## logistic regression – top 30 predictors



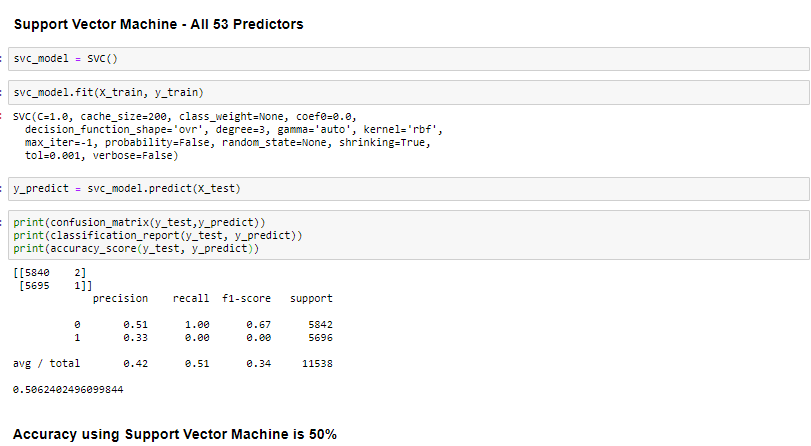
## lasso regularization and cross validation



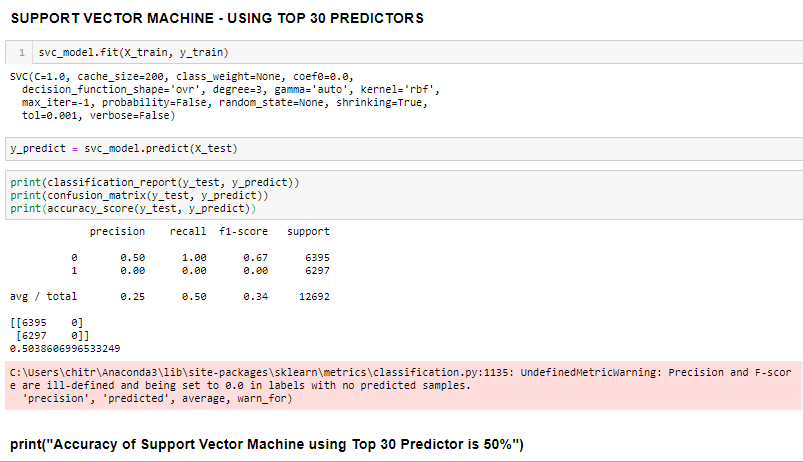




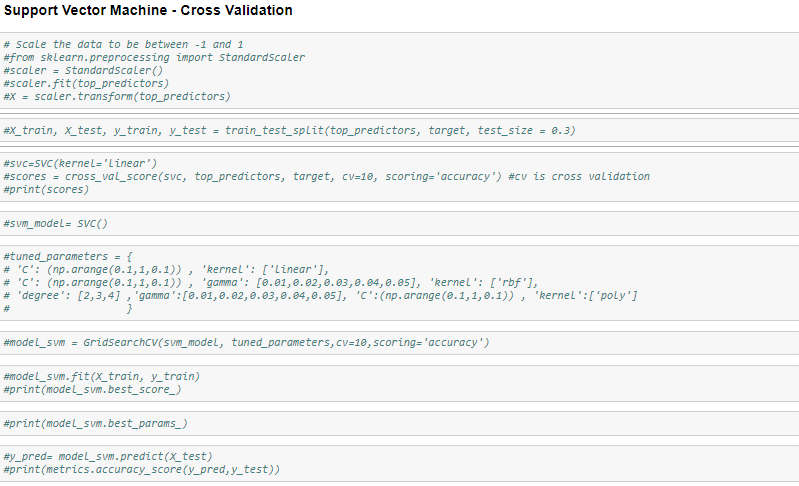
## Support vector machine – all 53 predictors



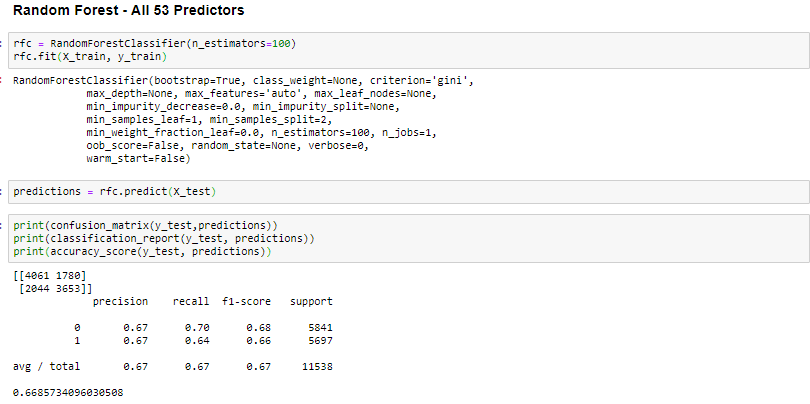
## support vector machine – Top 30 predictors



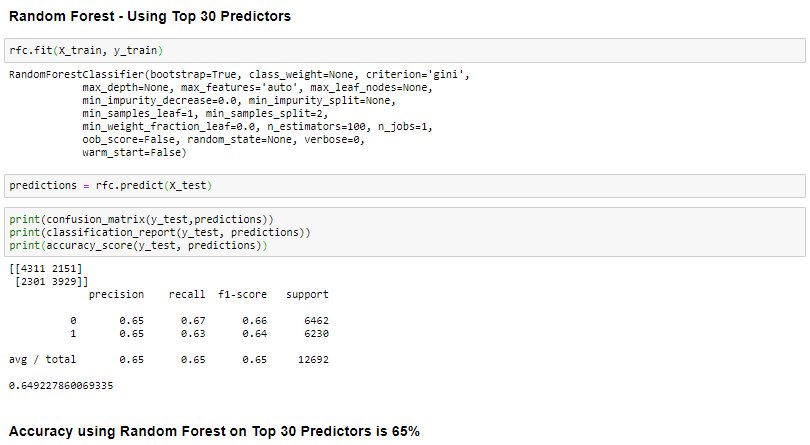
## Support vector machine – cross validation



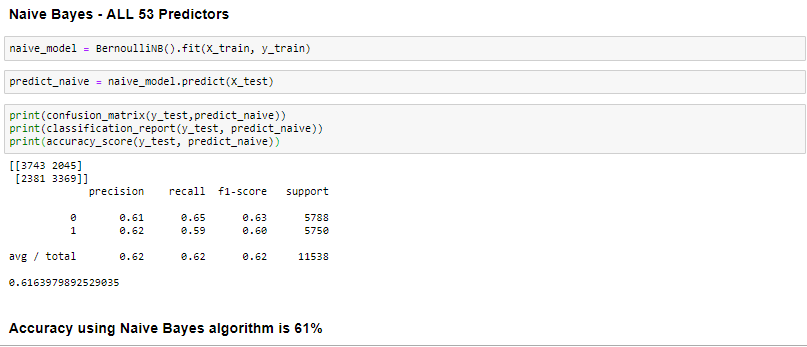
## Random forest – all 53 predictors



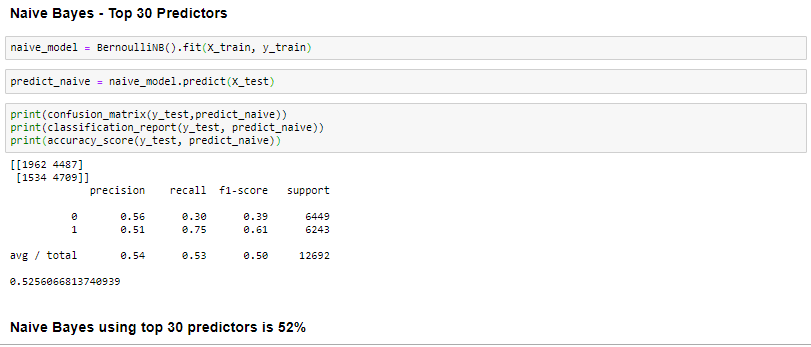
## random forest – Top 30 predictors



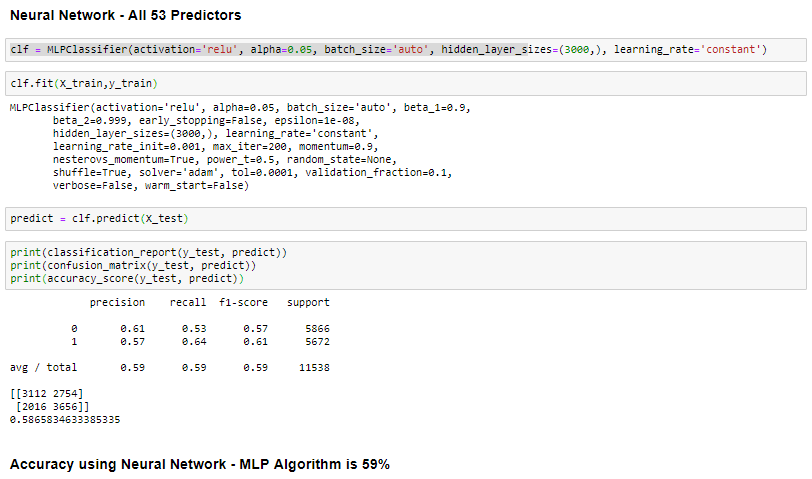
## naïve bayes – all 53 predictors



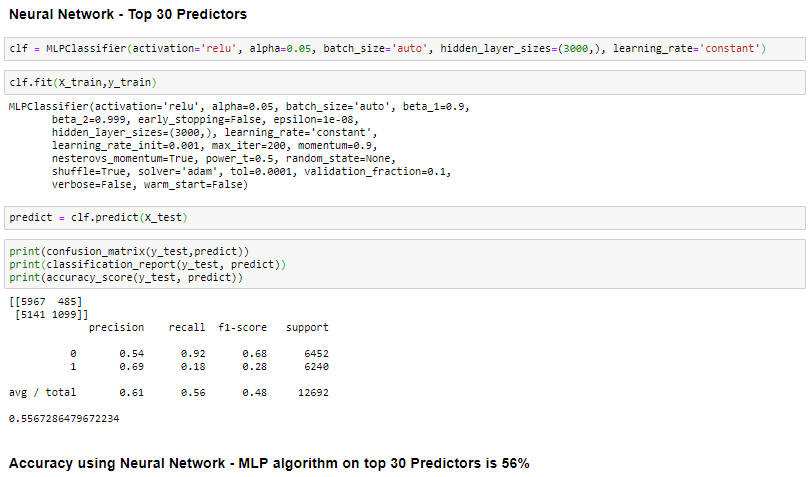
## 3.10 naïve bayes – top 30 predictors



## 3.11 neural network–feed forward algorithm -All 53 predictors



## 3.12 neural network-feed forward algorithm –top 30 predictors



# 4.Results

Accuracy is best achieved using Random Forest Classifier which is 66.85%

