

**Fake News Prediction Model**

**Submitted by:**

**Devanshu Suri**

**ACKNOWLEDGMENT**

I would like to express my very great appreciation to my SME Ms. Astha Mishra for his valuable and constructive suggestions during the planning and development of this research work. Her willingness to give his time so generously has been very much appreciated.

Separately, I would like to thank:

* FlipRobo Technologies team
* Data Trained Team

Research papers that helped me in this project was as follows:

* [(PDF) A Machine Learning Approach to Fake News Detection Using Knowledge Verification and Natural Language Processing (researchgate.net)](https://www.researchgate.net/publication/335191041_A_Machine_Learning_Approach_to_Fake_News_Detection_Using_Knowledge_Verification_and_Natural_Language_Processing)
* <https://theconversation.com/global/topics/fake-news-33438>

Articles that helped me in this project was as follows:

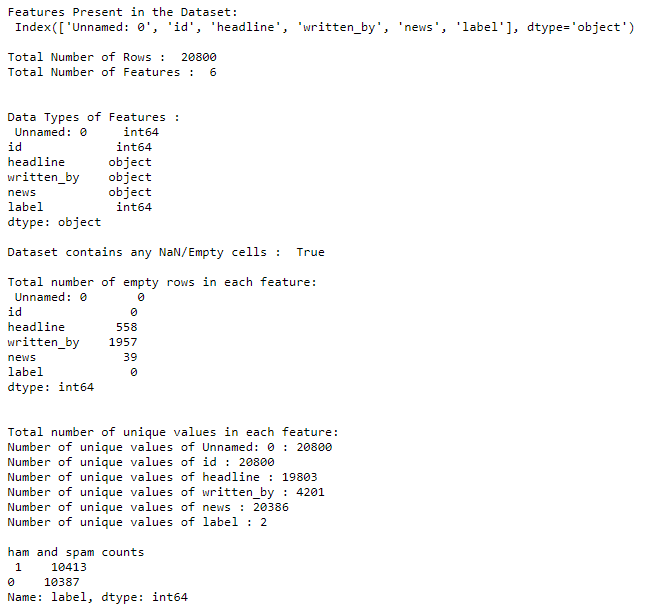
* [TF-IDF Vectorizer scikit-learn. Deep understanding TfidfVectorizer by… | by Mukesh Chaudhary | Medium](https://medium.com/@cmukesh8688/tf-idf-vectorizer-scikit-learn-dbc0244a911a)

**INTRODUCTION**

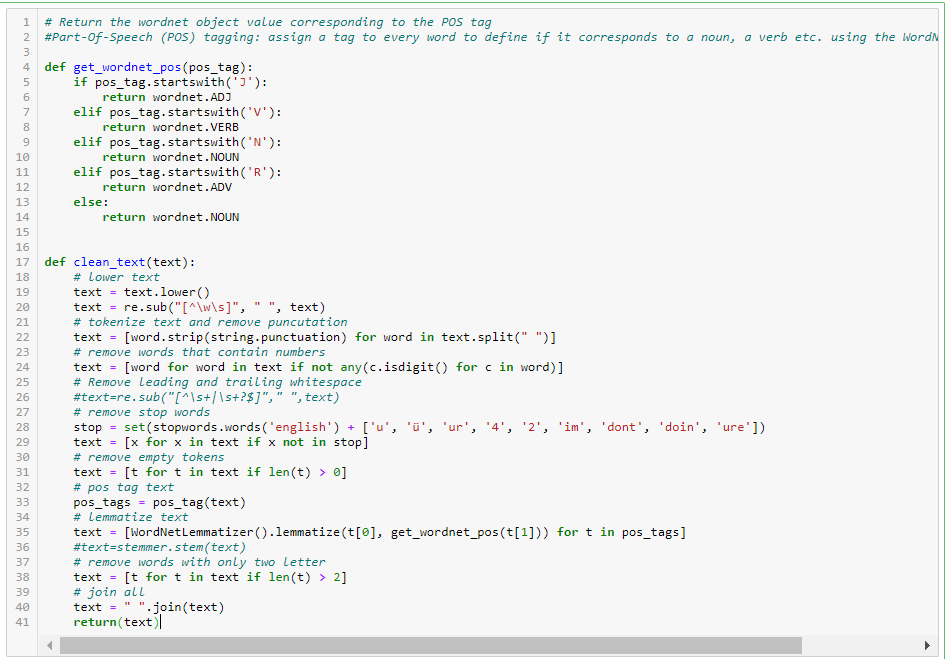
* **Business Problem Framing**
  + Fake news is a form of news consisting of deliberate disinformation or hoaxes spread via traditional news media or online social media.In this project, I have used different natural language processing (NLP) based machine learning and deep learning approaches including BERT to detect fake news from news headlines. Generally, a fake headline is a news headline which may read one way or state something as fact, but then the body of the article says something different. The Internet term for this type of misleading fake news is **“clickbait”** —headlines that catch a reader’s attention to make them click on the fake news. This type of fake news is misleading at best and untrue at worst.
* **Conceptual Background of the Domain Problem**
  + The idea of fake news is often referred to as click-bait in social trends and is defined as a “made up story with an intention to deceive, geared towards getting clicks”, Tavernise (2016). Some news articles have titles which grab a reader’s interest. Yet, the author only emphasizes a specific part of the article in the title. If the article itself does not focus on or give much truth to what the title had written, the news may be misleading. The goal of this project is to use natural language processing techniques to automate stance detection, since it is not practical for humans to fact check every piece of information produced by the media
* **Review of Literature**
  + If we look at some scholar work shows the issue that the fake news has been major concerned amongst scholar from various background. For instance, some authors have observed that fake news is no longer a preserve of the marketing and public relations departments. Instead there is a increasing risk of IT security, therefore, IT department is premised on the idea that it would help avert the various risks associated with the problem. So, if we good deeply into it we could find that the hackers use click bait with the help of fake news and make some professional of the organization downloads their malicious exploits in their system or leak sensitive information, albeit in an indirect manner. The user may, for instance, be tricked into believing that they are helping to disseminate the news further when, in the actual sense, they are providing the perpetrators with access to their emails, and we can also see that the fake news are worked extensively as they are using videos with original massage and uses their facial structure to replace the massage with false massage they want us to believe, these fake news issues is bigger day by day and we need to implement more our research and extensive knowledge to solve the problem.
* **Motivation for the Problem Undertaken**
  + This project was highly motivated project as it includes the real time problem of fake news which if we see are getting bigger, as there various concern as people do good things work hard to build a reputation, and only one false news is enough to ruin it all, it also have inverse effect on the financial market as if we observe there will a good amount of fluctuation on stock markets based on news.

**Analytical Problem Framing**

* **Data Sources and their formats** 
  + There are 6 columns in the dataset provided:
  + The description of each of the column is given below:
* “id”: Unique id of each news article
* “headline”: It is the title of the news.
* “news”: It contains the full text of the news article
* “Unnamed:0”: It is a serial number
* “written\_by”: It represents the author of the news article
* “label”: It tells whether the news is fake (1) or not fake (0).



* **Data Pre-processing Done**

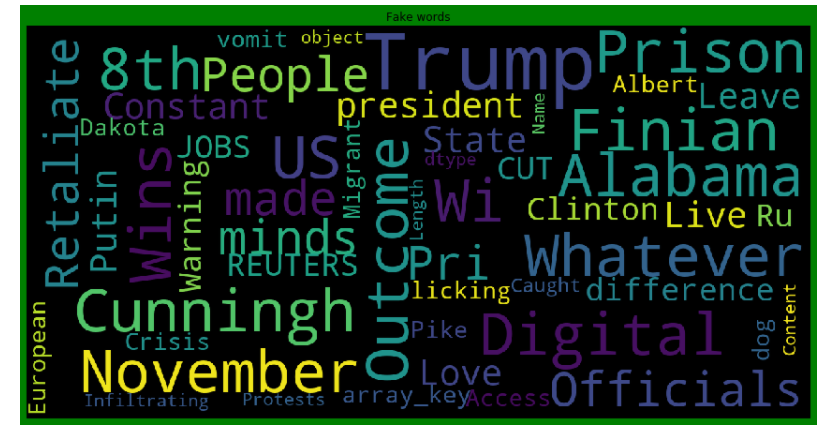
****

For Data pre-processing we did some data cleaning, where we used wordNet lemmatizer and porterStemmer to clean the words and removed special characters using Regexp Tokenizer and filter the words by removing stop words and then used lemmatizers and joined and return the filtered words.

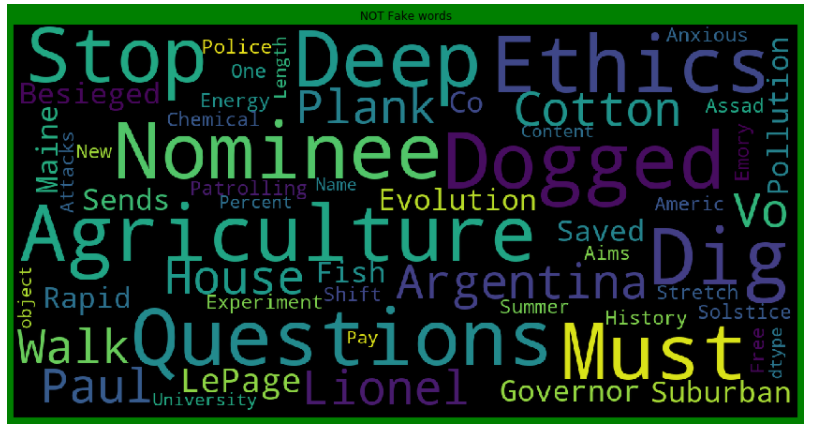
Used TFIDF vectorizer to convert those text into vectors, and split the data and into test and train and trained various Machine learning algorithms.

* **Data Inputs- Logic- Output Relationships**

**Fake Words:**

****

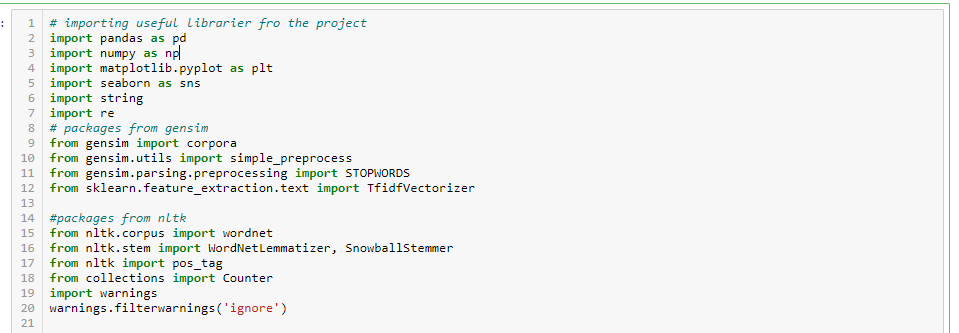
**Not Fake Words:**

****

From the above we can see that most frequent words on both labels and we can observe the words which are leading to fake new are trump, Clinton, prison,november, etc and words which are leading to real news are said, agriculture,police ,questions etc, so we can clearly see that above dataset extensively deals with news around US presidential elections between Trump and Clinton.

* **Hardware and Software Requirements and Tools Used**
  + Hardware: 8GB RAM, 64-bit, 9th gen i7 processor.
  + Software: MS-Excel, Jupyter Notebook, python 3.6.

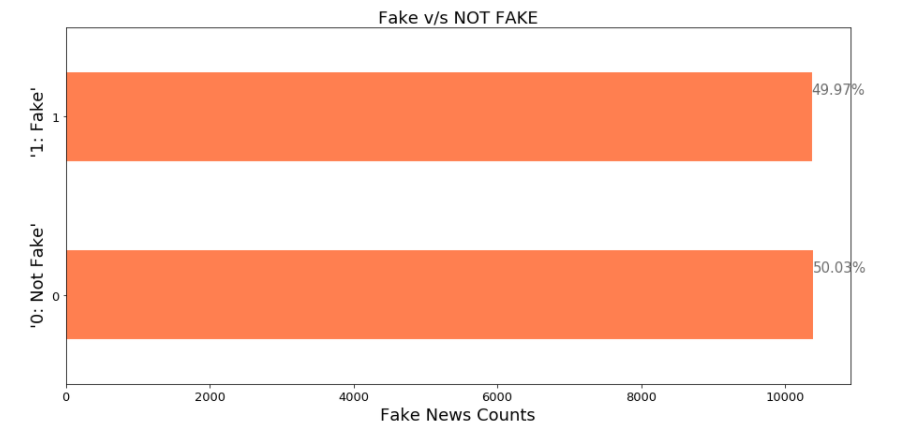
**Libraries used:-**

****

****

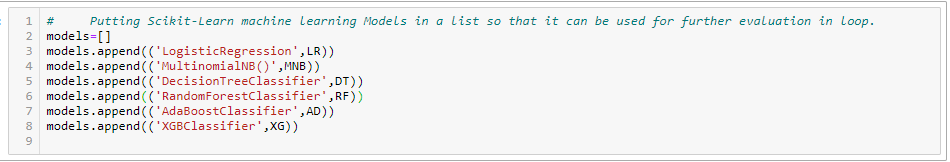
**Model/s Development and Evaluation**

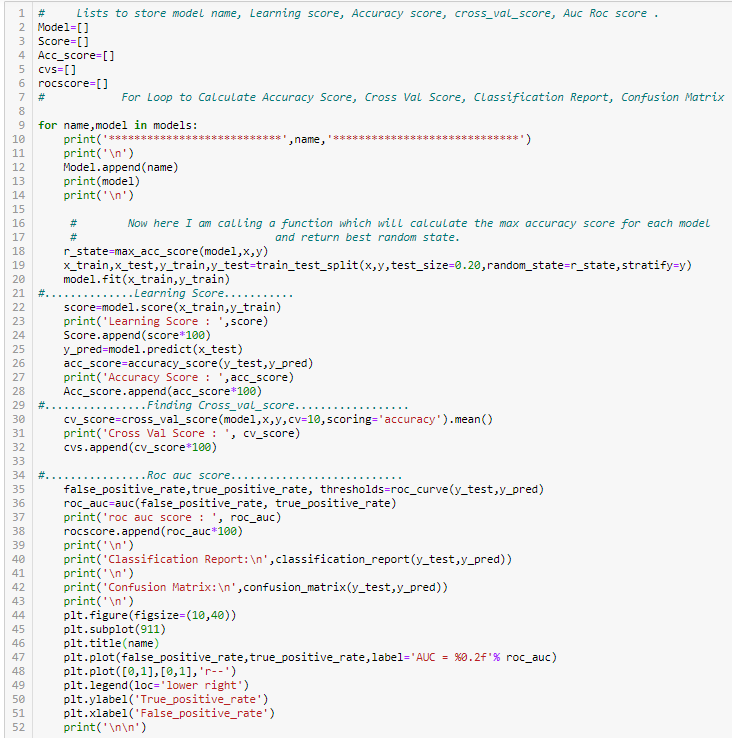
* **Identification of possible problem-solving approaches (methods).**

****

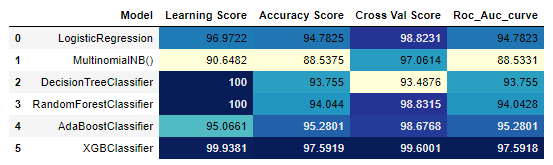
* From the above we can see that the dataset is balanced which is good as it will help our model to classify more accurately, so we should expect good accuracy score, and as the volume of data was also good.
* **Testing of Identified Approaches (Algorithms)**
  1. RF=RandomForestClassifier()
  2. LR=LogisticRegression()
  3. MNB=MultinomialNB()
  4. DT=DecisionTreeClassifier()
  5. AD=AdaBoostClassifier()
  6. XG=XGBClassifier()

**Run and Evaluated selected models**

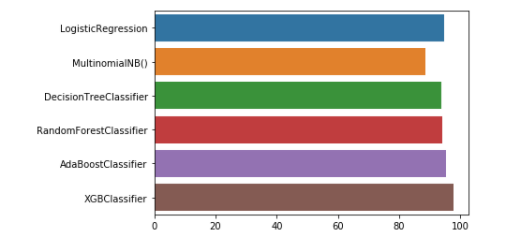




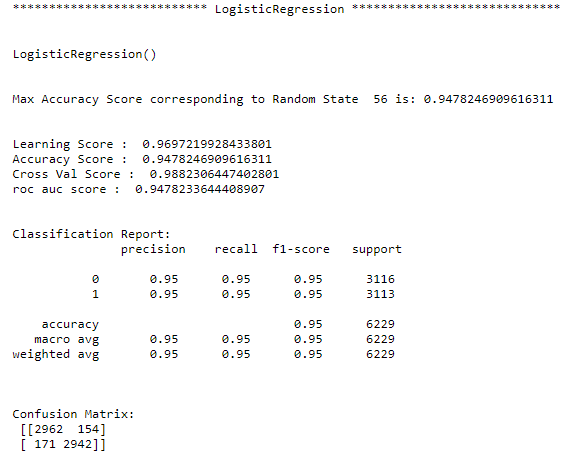
* **Key Metrics for success in solving problem under consideration**

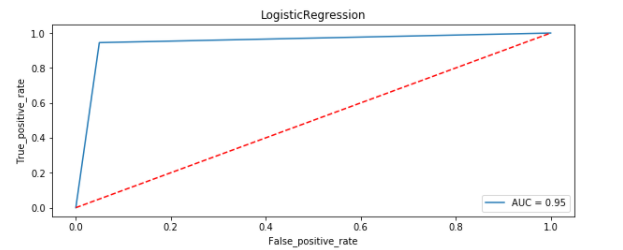


Key Metrices used were the Accuracy Score, Crossvalidation Score and AUC & ROC Curve as this was binary classification. From the above we can see that there are various models out of which we few gave good accuracy score as more than 90%,

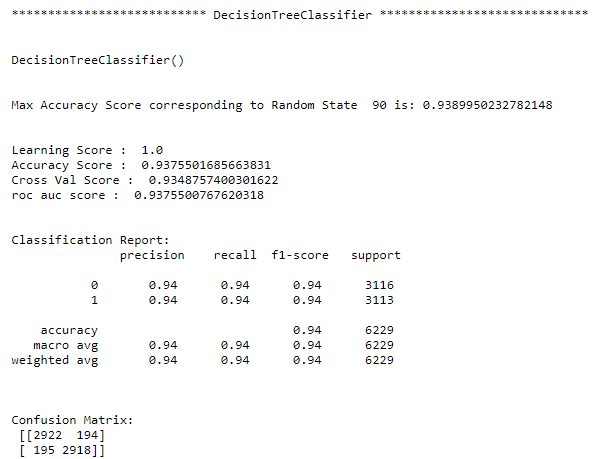


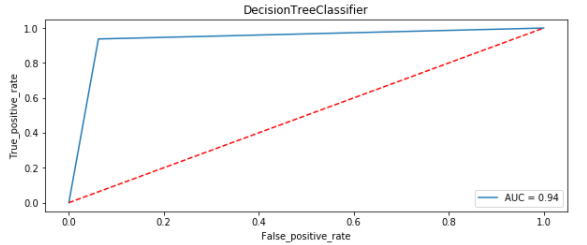
* **Visualizations:**
* Logistic regression:



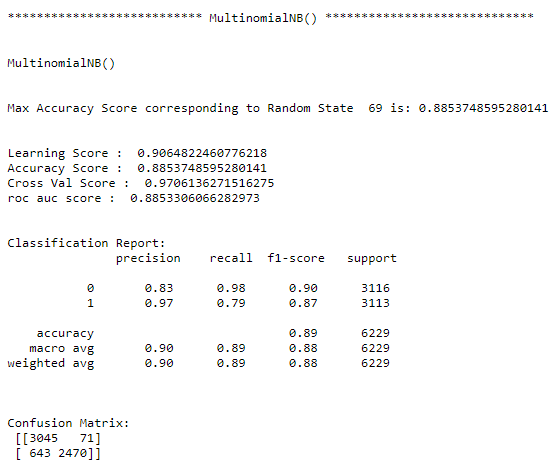


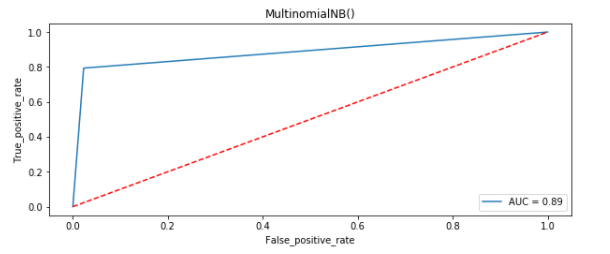
* Decision Tree Classifier:

****

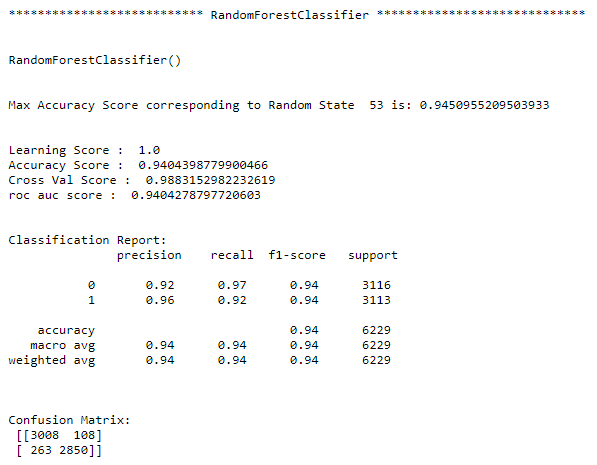
****

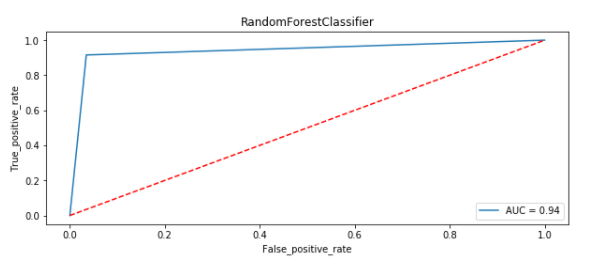
* MultiNomial NB:

****

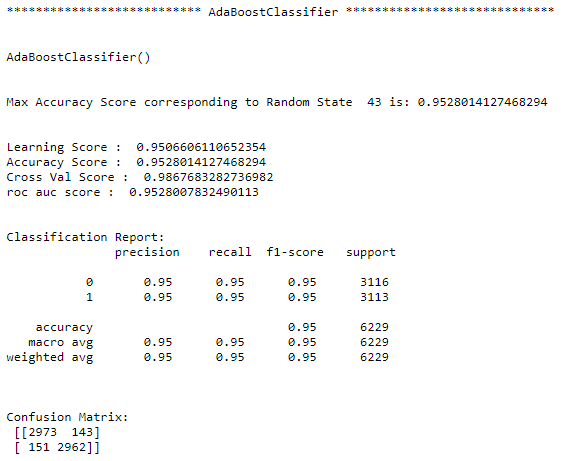
****

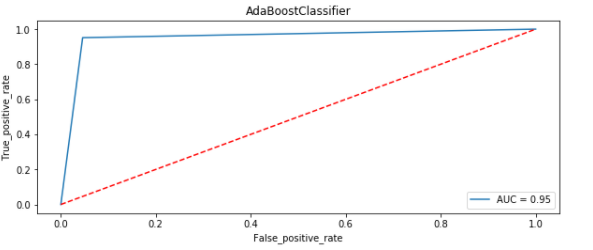
* Random Forest Classifier:

****

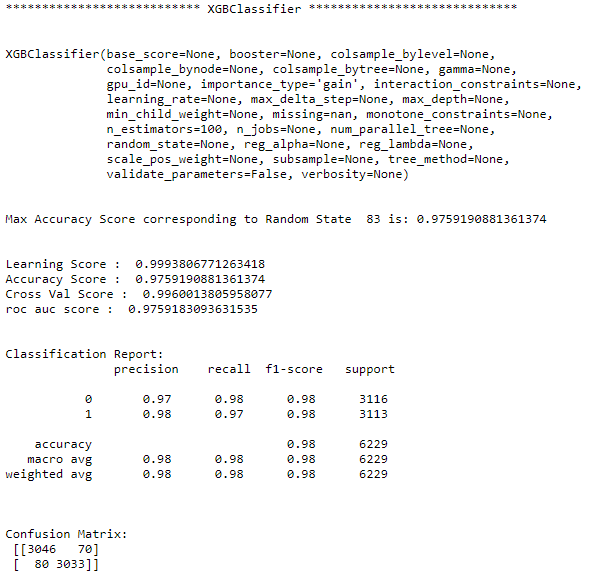
****

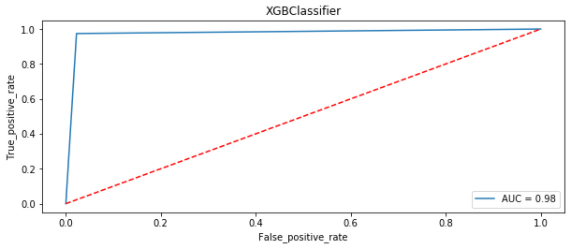
* Ada Boost Classifier:

****

****

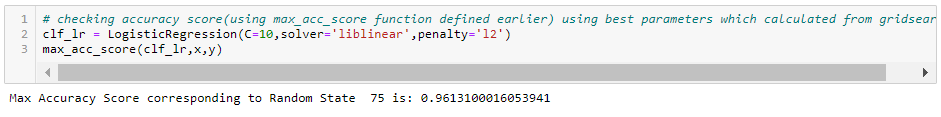
* XGB Classifier:

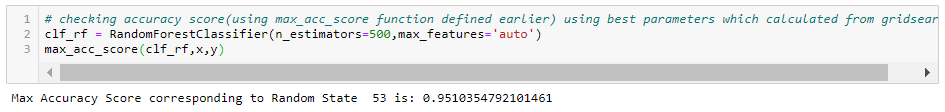
****

******

## After all this process conclusion is that XGB Classifier and Adaboost Classifier and Random Forest Classifier are performing well in terms of Accuracy score, Cross val score and Roc\_Auc score as compared to other models.

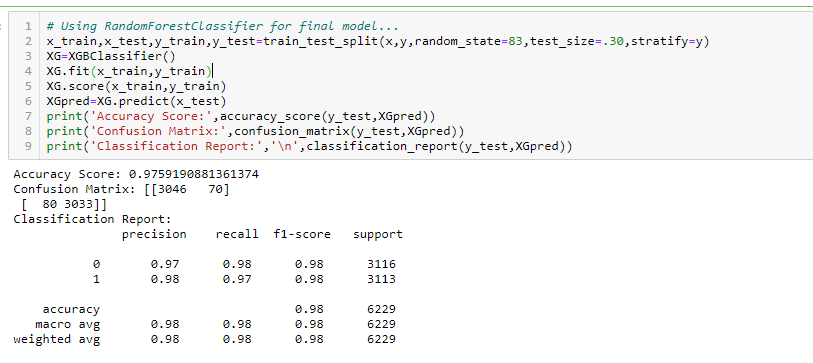
# **Hyper Parameter Tuning Results:**

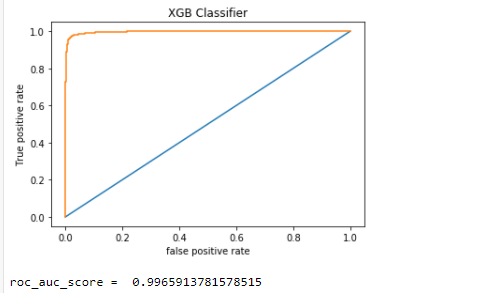




## After all this process conclusion of Hyper Parameter is that Random Forest Classifier is giving accuracy of 95.10%, but XGB Classifier is giving an accuracy of 97% without tuning. So now I am making a final model using XGB Classifier.

* **Final Model:**





From the above visualization and matrices found that the XGB Classifier performed the best 99.6% AOC\_ROC\_SCORE, with precision accuracy score of 97% and recall 98%.

* **Interpretation of the Results**
* From the above visualization and matrices found that the XGB Classifier performed the best AUC\_ROC\_SCORE **i.e. 99.6%.**

**CONCLUSION**

* **Key Findings and Conclusions of the Study**

From the whole evaluation we can see that the maximum number of words in fake news were regarding Trump, and Clinton and we can interpret that it was due to election campaign which was held during US presential election and we know these adverse effects of the voters which were influenced by the fake news and most of the real news had said, trump and president, and fake news which was cleared by trump’s campaign, but can hardly see any clarity or real news from the side of Clinton, and due to which the impact we already saw on election results and regarding the election advertisement and news Facebook’s CEO Mark Zuckerberg also got extensively question by congress.

* **Learning Outcomes of the Study in respect of Data Science**

So, from the words frequency chart we can clearly see that most of the news were related to US presedential election between Trump and Clinton, and by implementing passive aggressive algorithms we can see that the we have achieved a good score as it calculates the errors and updates its own learning rate which makes our model more reliable.

* **Limitations of this work and Scope for Future Work**
* Machine Learning Algorithms like Gradient Boosting Classifier took enormous amount of time to build the model.
* Using Hyper-parameter tuning for XGB would have resulted in some more accuracy.
* Using Deep Learning for detection fake news may get some good results.