

**FAKE NEWS DETECTION PROJECT**

**By**

**DEEPTHI PRAKASHAN**

**ACKNOWLEDGMENT**

I would like to sincerely thank Flip Robo Technologies for the opportunity to work in the project and the timely guidance provided.

I express my gratitude towards Data Trained for the training they have provided overtime and for the guidance and support.

I express my heartfelt support to my parents and friends who were constantly by my side providing me with motivation and inspiration to work through the completion of this project.

**INTRODUCTION**

**Business Problem**

A sensational news often causes a stir and is sold well in the society. On social media platforms, such as twitter, we often come across latest news that are designed to have clickbait titles and content of which contains no contextual details and verifiable facts or sources. However, not all such news is true and can be damaging for the reputation of the person or the institution. Fake news has become one of the biggest problems of our age. It has serious impact on online as well as offline discourse.

To combat this challenging situation, we need to build a machine learning classifier that can detect fake news. Since the problem at hand deals with text-based data, we will be using Natural Language Processing to process the data.

**Conceptual Background**

Fake news are false and misleading information or claims based on superstition, rumours, and misperceptions. Such false information can be politically motivated, instigated by religious beliefs, or based on intent to harm specific individuals among other reasons. With the easy availability of digital media platforms, there also has been an increase in fake news. It is fundamental to distinguish between the type of news. Failure to discern between the could potentially lead to a rapid spread of the news by sharing it through various platforms before appropriate action for damage control is taken.

**Literature Review**

Machine learning utilises huge data to build algorithms that can discern fake news from real news. Different social media have established approaches for the automatic detection of fake news. A very recent example of emergence of numerous false reports and news was during the COVID-19 pandemic, with media outlets releasing false remedies that can be harmful to the people. Fake case report can also induce public panic. Twitter, Instagram, Facebook uses various algorithms to reduce the spread of fake news. Write about covid report and how fake news were resolved by each platform.

**Motivation of the problem undertaken**

Nowadays, fake news proliferates at a rapid speed considering the easy access to the internet. It can cause a shift in public opinion. Since political agendas are one of the motivations for the spread of fake news, it can cause unrest in a country.

False and misleading news can lead to adverse consequences hence regulation of the spread of such fabricated news is a necessity to avoid heading towards any unprecedented situations. Thus, through this project we aim to detect fake news using machine learning and NLP.

**Analytical Problem Framing**

The objective of this project is to detect fake news from real news which is a classification problem. We are given two dataset containing fake news and real news. In true news, there is 21417 news, and in fake news, there is 23481 news. I have inserted one label column zero for true news and one for fake news and combined both datasets using pandas built-in function.

Machine learning data only works with numerical features so text data is eventually converted into numerical columns before which the text is preprocessed. In-text preprocess we are cleaning our text by steaming, lemmatization, stopwords are removed, special symbols and numbers are removed, etc. After cleaning the data, text data is converted data into numerical features using Tfidf Vectorizer.

* **Data Sources and their formats**
* The data has been received from FlipRobo Technologies
* The two dataset is in csv file format
* There as 21417 rows and 4 columns in true news dataset and 23481 rows and 4 columns in fake news dataset.
* Label which was later added, is the target column where 1 denotes fake news and 0 denotes true news.
* **Data Pre-processing**
* Labels were additionally included, where fake news were labelled as ‘1’ and real news as ‘0’
* No null values were observed.
* On performing .info(), we observed that the date column was present as object datatype. However, I faced errors while trying to convert it into datetime[ns] due to the presence of certain links and other strings.
* To filter these out I used regex, which retrieved the rows using certain patterns observed. In our case, there were few links present, and hence utilized ‘https’ to find all such rows which were then dropped.
* Following steps which we call Natural Language Processing is used to clean the text data and titles. A function was written which decontracted certain words. Numbers, punctuations, spaces, were removed. Stopwords were removed and the sentences were converted to lower case. Words were converted to their base root mode using lemmatization.
* Tfidf-Vectorizer is used to tokenize the data and gives a weightage to each word relative to the document. It gives high weight to rare terms and low weigh to common terms. This way the text data is converted to numerical data which can then be used in the building our ML models.
* **Data Inputs- Logic- Output Relationships**

With the given data, we conducted exploratory analysis.

Title, text and subject were all object datatype while date was converted to datetime[ns].For predicting fake news, we could either use title or text column. However, since I faced with MemoryError while processing the data, I went with the title data. We performed data visualization, where it was seen that all the fake news were under the subject politicsnews and world news. WordCloud did not give us relevant information.

**Hardware and Software Requirements and Tools Used**

|  |  |
| --- | --- |
| **Hardware** |  |
| Processor | Dual Core or above |
| RAM | 4GB or more |
| Cache | 1MB or more |
| Hard Disk/SSD | 180GB or more |
| **Software** |  |
| OS | Window/Mac/Linux |
| Ide | Jupiter Notebook |
| Dataset | .csv file |
| Libraries | Pandas, numpy,matplotlib,seaborn, scikit.learn, nltk, re |
| Server | Web server with HTTP process |

**Model/s Development and Evaluation**

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

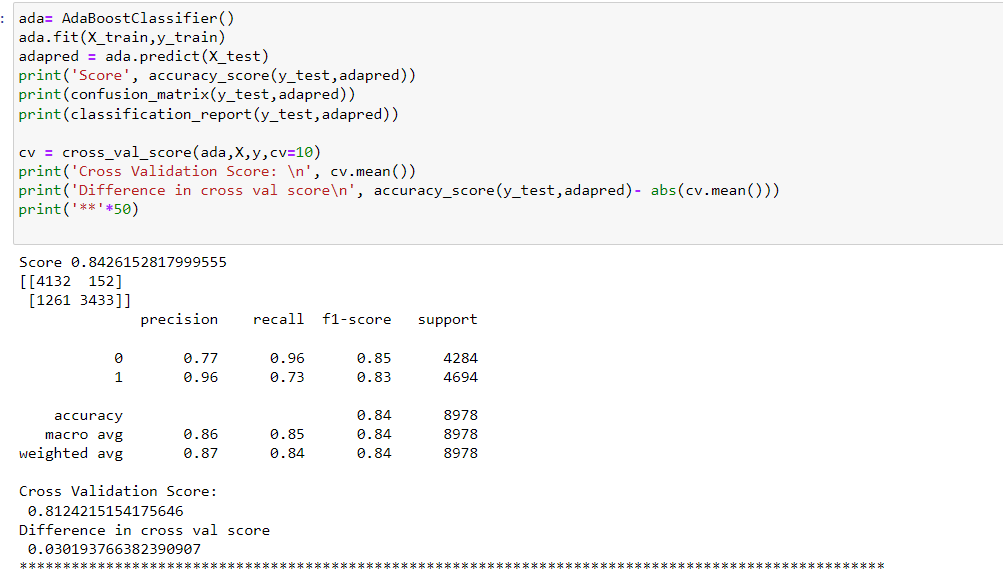
The pre-processed data was then used for training the models. The data was first split in train and test dataset. Here, 80% of the data was used for training and 20% for testing. Training the data allowed the models to learn the data and the underlying trend. Best random state was also checked, for which best accuracy score was 0.93 at random state 29. (Using RandomForestClassifier), which was later used while training the models. Model’s efficiency was then checked by predicting the fake from the test dataset. Evaluation metrics and cross-validation then helped us decide how good our prediction of fake news was and if the predicted values were not as a result of overfitting or underfitting.

The models that were used in prediction were

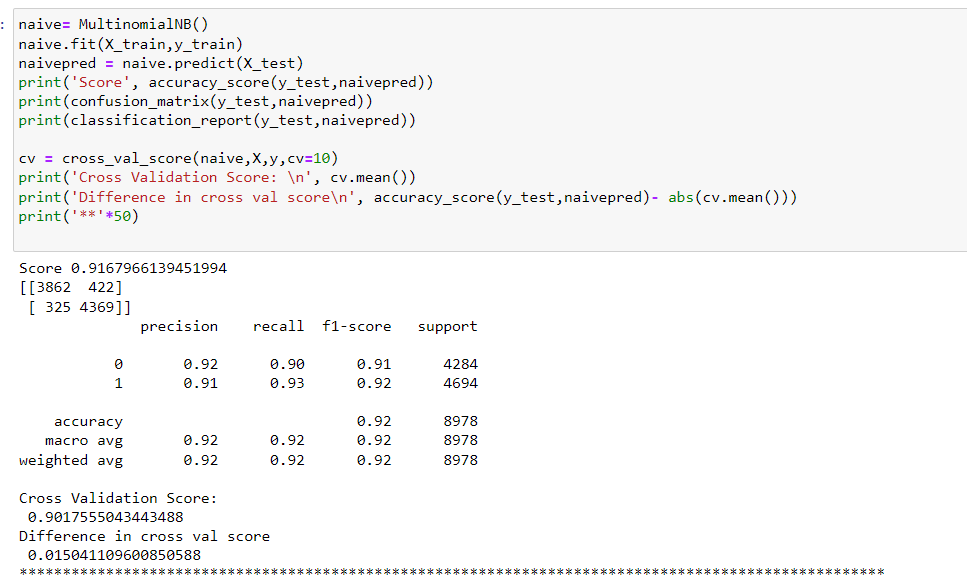
* RandomForestClassifier
* MultiNomialNB
* GuassianNB
* AdaBoostClassifier
* GradientBoostingClassifier

**Run and evaluate selected models**

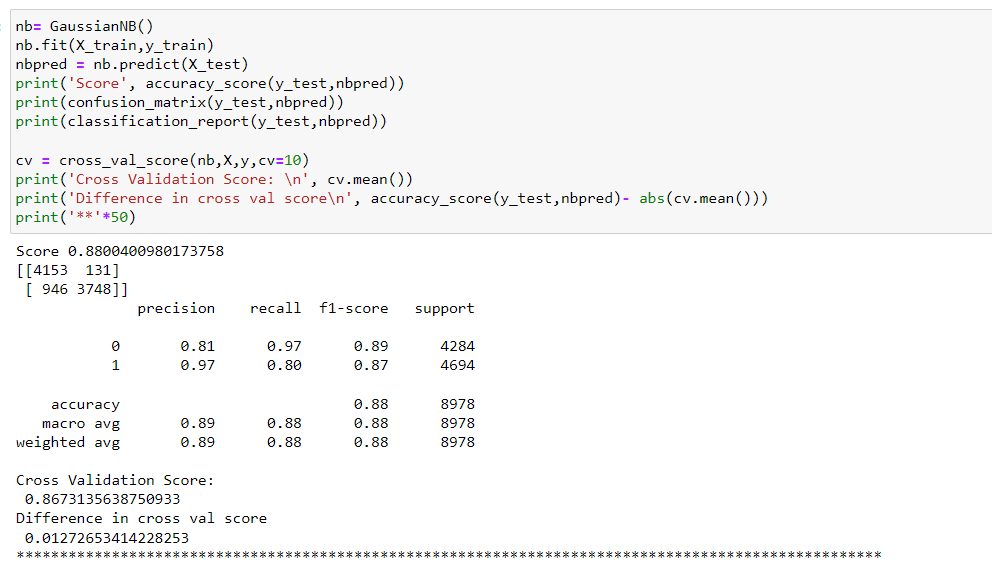
**AdaBoostClassifier**

****

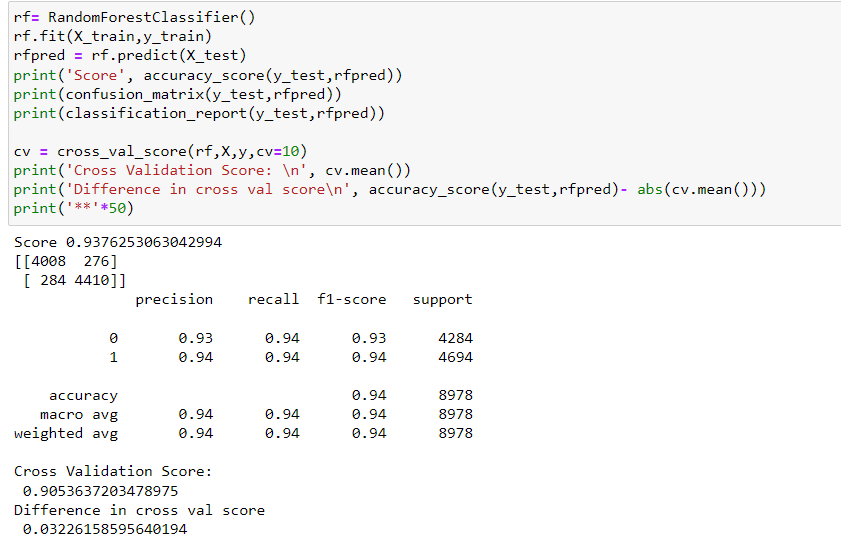
**MultinomialNB**



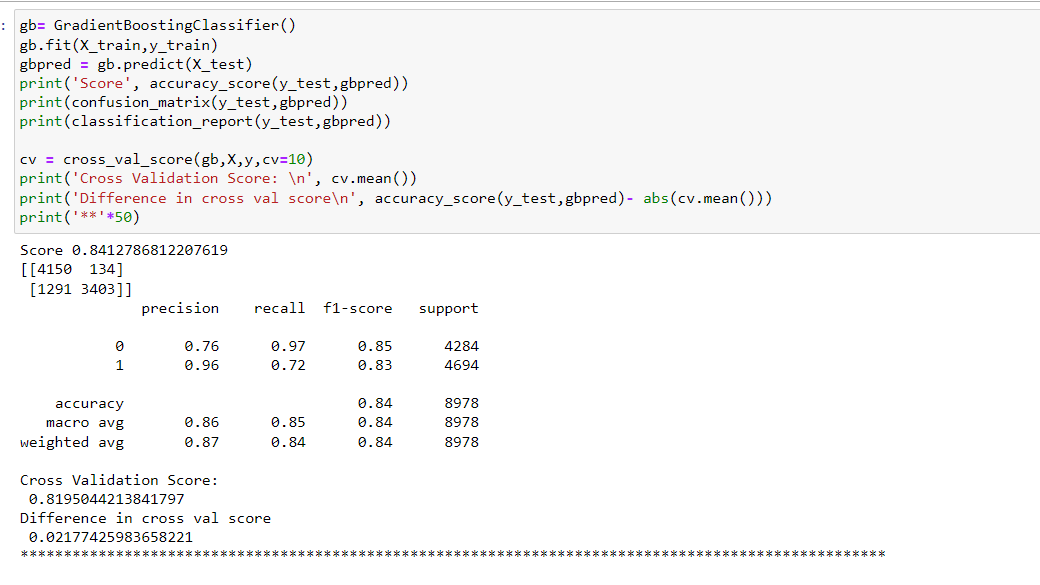
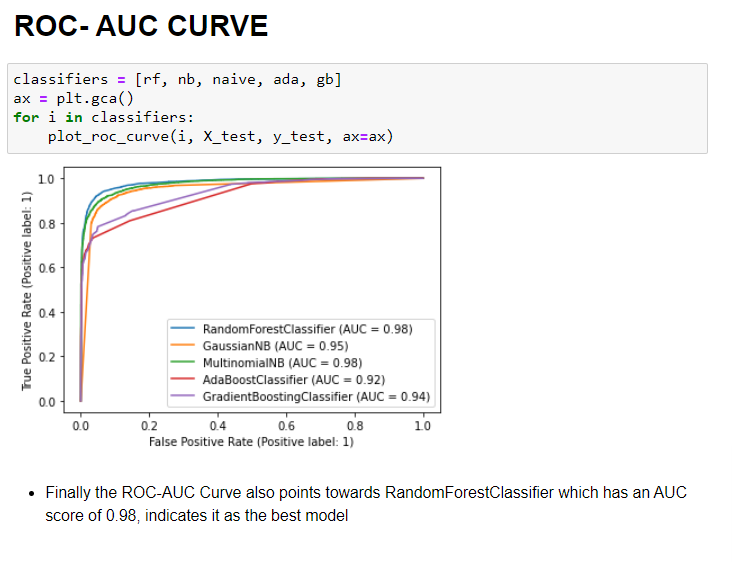
**GaussianNB**

****

**RandomForestClassifier**

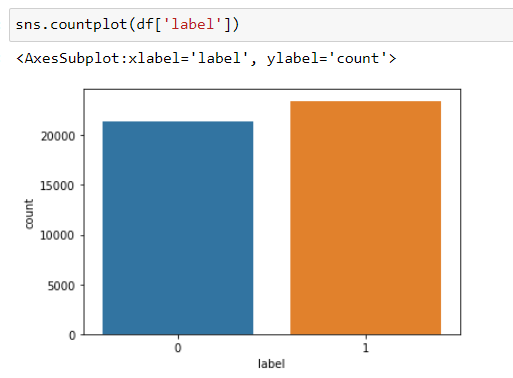
****

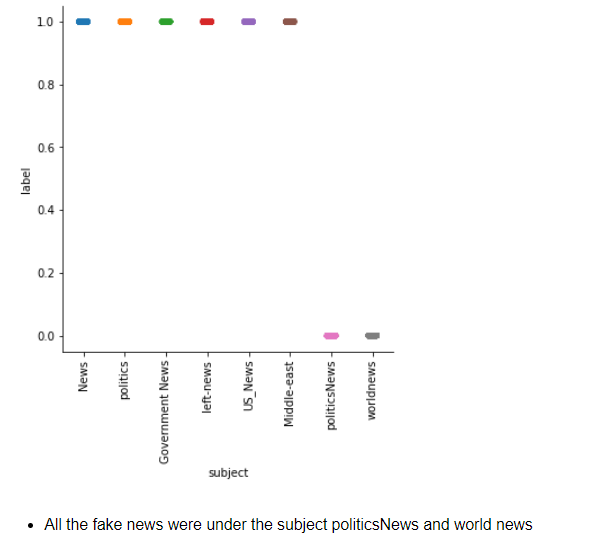
**GradientBoostingClassifier**

****

Conclusion: To select the best model, based on this problem, I considered the model that will predict more fake news, and have lower count of False positives. While the former reason is self-explanatory, the latter is because more the "fake news" are predicted as True, the worse is our model. Hence, apart from accuracy as our metric, I will also be considering precision and recall as factors for evaluating my models.

* Comparatively RandomForestClassifier is performing well in all aspects, with accuracy of 93%
* Precision for fake news being 94% and for true news to be 93%. Recall was found to be 94% for both labels.
* However, if I give more weightage to having fewer fake news presented as positive news i.e., fewer False positives then GaussianNB performs the best as it has a precision of 97% for Fake new. It has a lower accuracy of 88%. And has the drawback of predicting more True news as Fake News.

**Visualisation**

****The dataset we have here is almost balanced which is ideal for training the model.

It can be observed that all the fake news is under the subject politicsNews and world news. It can also be a bias caused during scrapping as we have received tow different datasets, each for fake and true.

**CONCLUSION**

**Key Findings and Conclusions of the Study**

* Fake news is under the subject politicsNews and world news.

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

* The best model was selected to be RandomForestClassifier that gave an accuracy, precision, recall and ROC-AUC score of 93.76%, 94%, 94%, 98% respectively.
* Few challenges that I came across while solving the problem was the word cloud visualization. There was no clarity so as distinguish between the fake news and real news using visualization. Since the data was huge, I came across MemoryError which I later tackled by reducing the max\_features.

**Limitations of this work and Scope for Future Work**

* We can go further with the text cleaning which may give better wordcloud visualization.