

Department of Digital Systems
MSC in Artificial Intelligence

Fake News Prediction

Dimitrios Delikonstantis

14/02/2021



Reason for subject selection

- Fake or misleading news can be dangerous
- It is also used for making money via advertising (clickbait)
- Prevalence has increased with the rise of social media
- Social media algorithms have been implicated in the spread of fake news
- Since coronavirus appeared fake news have been on the rise
- Anti-vaccination movement is growing vastly via fake news. WHO declared it one of the top ten global threats
- Therefore, vastly reducing fake news is crucial for the society's own good



The data

- Dataset retrieved from kaggle
- Used data from 20800 article news labeled 'Real' or 'Fake'



Bag-of-words model

- Text is changed into vectors of numbers to be processed by machine learning algorithms (feature encoding)
- Describes the occurrence of words within a document
- Transform each document from a corpus of documents into a vector and use it as an input to a machine learning algorithm



Feature extraction with TF-IDF

- Statistical measure that shows how important a word is in a document
- Term frequency shows the frequency of a word in a document.
- Inverse document frequency shows the frequency of a word within the corpus of documents.
- $\text{TF-IDF score} = \text{TF} * \text{IDF}$
- The higher the score, the more relevant that word is in that particular document
- Sklearn TfidfVectorizer used for different case scenarios.



Cross validation

- Initially, sklearn `train_test_split` was used for splitting the dataset.
- Can we be sure that the splitting done by `train_test_split` is unbiased?
- Cross validate with sklearn `RepeatedKFold` where different splits each repetition are produced



Logistic regression metrics

- Accuracy: 93.79%

- Confusion matrix:

	Actual
Predicted	$\begin{bmatrix} 1965 & 111 \\ 147 & 1937 \end{bmatrix}$

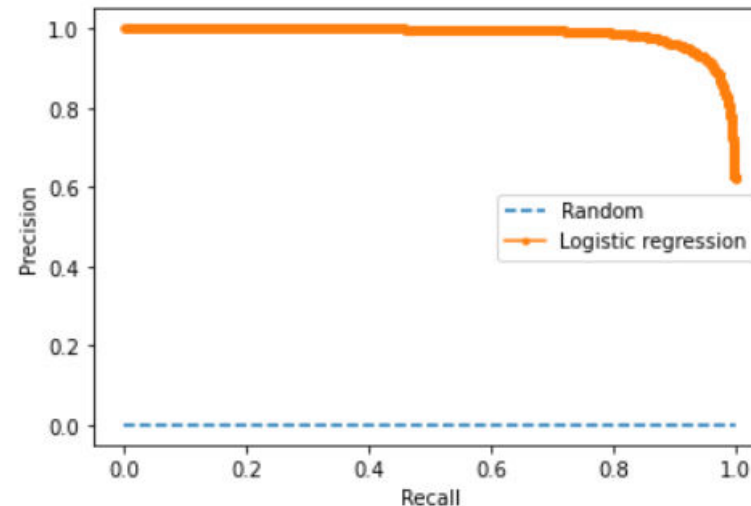
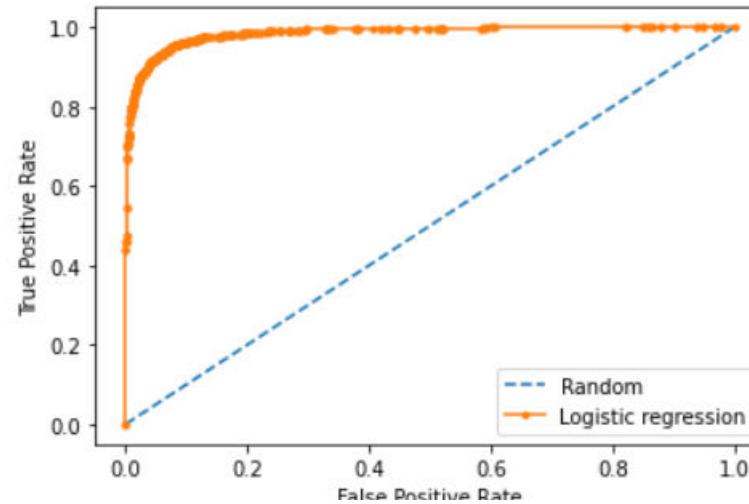
- Precision: 0.9458

- Recall: 0.9294

- F1-score: 0.9375

Logistic regression ROC/Precision recall curves

- ROC AUC: 0.9851
- ROC AUC: 0.9859





Naïve Bayes metrics

- Accuracy: 88.46%

- Confusion matrix:

Predicted

Actual

[[1802 274]

[206 1878]]

- Precision: 0.8726

- Recall: 0.9011

- F1-score: 0.8866

Naïve Bayes

ROC/Precision recall curves

- ROC AUC: 0.9535
- ROC AUC: 0.9574

