DEMOKRITOS

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
UNIVERSITY OF PIRAEUS

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, meanwhile 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 to a document in an entire corpus of documents
- Term frequency shows the frequency of a word in a document
- TF = Word occurrence / vocabulary of document
- Inverse document frequency shows the frequency of a word within the corpus of documents.
- IDF = log(Entire corpus of documents / number of documents containing specific word)
- TF-IDF score = TF * IDF
- The higher the score, the more relevant that word is in that particular document
- Sklearn TfidfVectorizer used for different case scenarios

Train test split and cross-validation

- Sklearn train_test_split was used for splitting the dataset. Different case scenarios were tested to see which one provided best results.
- But what if the training dataset includes only articles from a specific author. Then our data is biased
- This is why cross-validation is crucial
- Cross-validation (partially) was implemented with sklearn Kfold to check the accuracy on the training data and compare it to the predicted data to check if overfitting occurs
- Cross-validation computing cost issues encountered when trying to train and test the model cause of too many vocabulary features

Logistic regression metrics

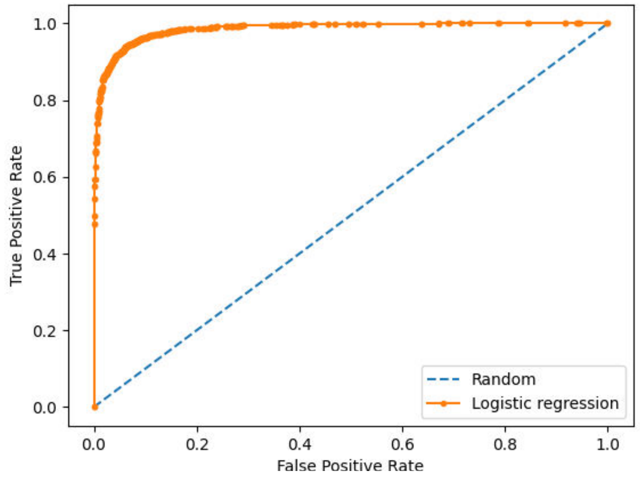
- Cross-validation mean accuracy: 93.2%
- Accuracy: 93.55%
- Confusion matrix:

	Actual class
Predicted class	[[1982 118] [150 1910]]

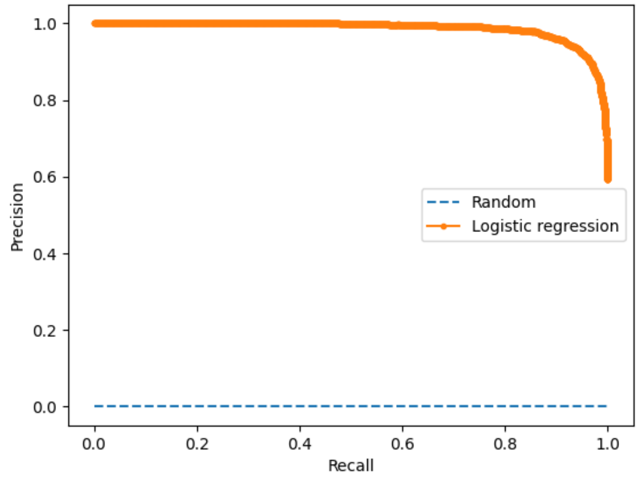
- Precision: 0.9418
- Recall: 0.9271
- Specificity: 0.9438
- False positive rate: 0.056
- F1-score: 0.9344

Logistic regression ROC/Precision recall curves

- ROC AUC: 0.9853



- ROC AUC: 0.9857



Naïve Bayes metrics

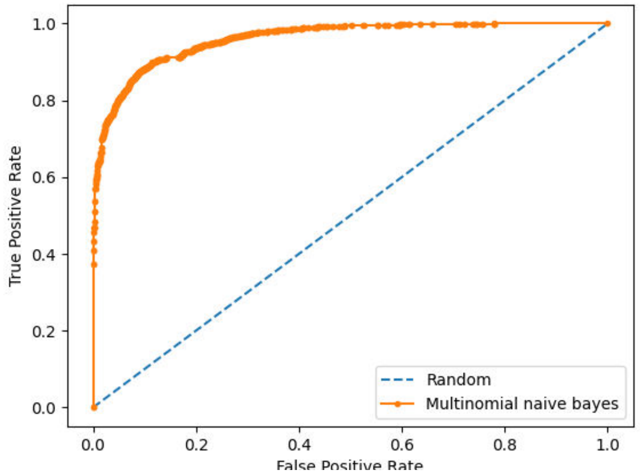
- Cross-validation mean accuracy: 87.76%
- Accuracy: 87.21%
- Confusion matrix:

$$\begin{array}{cc} & \text{Actual class} \\ \text{Predicted class} & \begin{bmatrix} 1752 & 348 \\ 184 & 1876 \end{bmatrix} \end{array}$$

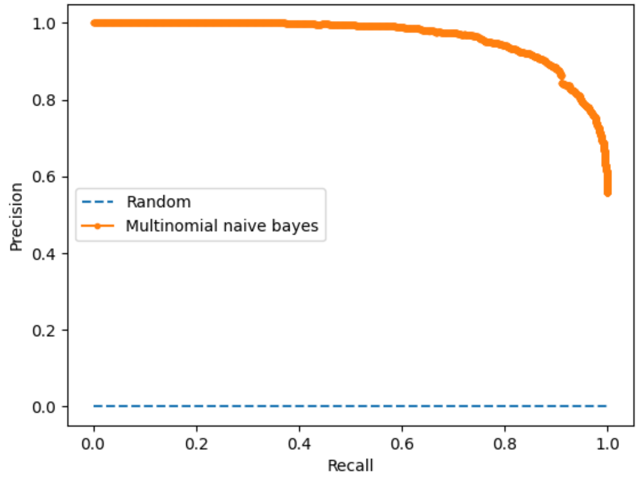
- Precision: 0.8435
- Recall: 0.9106
- Specificity: 0.8342
- False positive rate: 0.1657
- F1-score: 0.8758

Naïve Bayes
ROC/Precision recall curves

- ROC AUC: 0.9601



- ROC AUC: 0.9626



K-nearest neighbor metrics

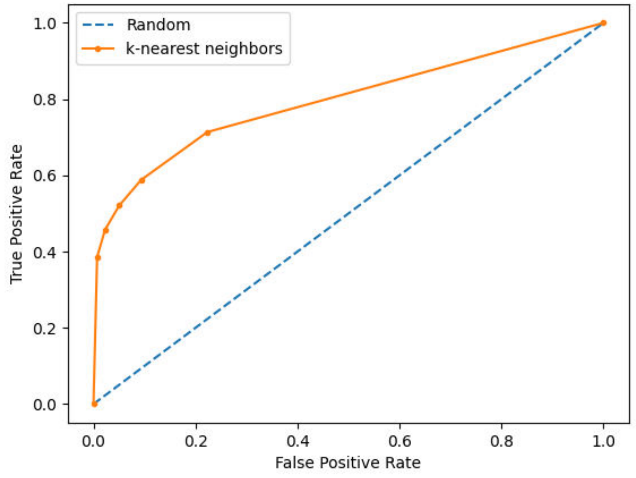
- Cross-validation mean accuracy: 72.51%
- Accuracy: 73.72%
- Confusion matrix:

	Actual class
Predicted class	[[1994 106] [987 1073]]

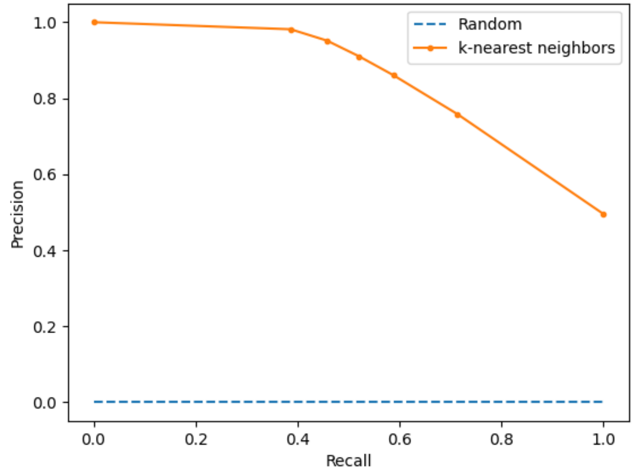
- Precision: 0.91
- Recall: 0.5208
- Specificity: 0.9495
- False positive rate: 0.05
- F1-score: 0.6625

K-nearest neighbor ROC/Precision recall curves

- ROC AUC: 0.7953



- ROC AUC: 0.8511



Decision tree metrics

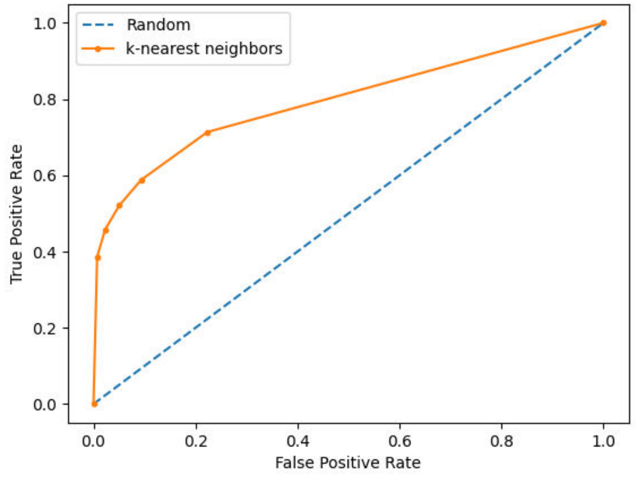
- Cross-validation mean accuracy: 87.6%
- Accuracy: 87.3%
- Confusion matrix:

	Actual class
Predicted class	[[1822 278] [250 1810]]

- Precision: 0.8668
- Recall: 0.8786
- Specificity: 0.8676
- False positive rate: 0.1323
- F1-score: 0.8727

Decision tree
ROC/Precision recall curves

- ROC AUC: 0.8714



- ROC AUC: 0.9012