Institute of Data Mini Project 3

Can we use Natural Language Processing and Machine Learning Models to Detect ‘Fake’ vs ‘Real’ News?

Gayle Ferguson

July 2022

The Data

- Fake.csv 23482 rows
- True.csv 21417 rows
- 4 columns: ‘Title’, ‘Text’, ‘Date’, ‘Subject’
- Target = ‘True’ (1) or ‘Fake’ (0)

- Dataset from Kaggle: https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset

- Clément Bisaillon (Owner)

Publications:
Ahmed H, Traore I, Saad S. (2018) “Detecting opinion spams and fake news using text classification”, *Journal of Security and Privacy*, Volume 1(1).

Ahmed H, Traore I, Saad S. (2017) “Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques. In: Traore I., Woungang I., Awad A. (eds) *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments.* ISDDC 2017. Lecture Notes in Computer Science, Vol 10618. Springer, Cham (pp. 127-138).

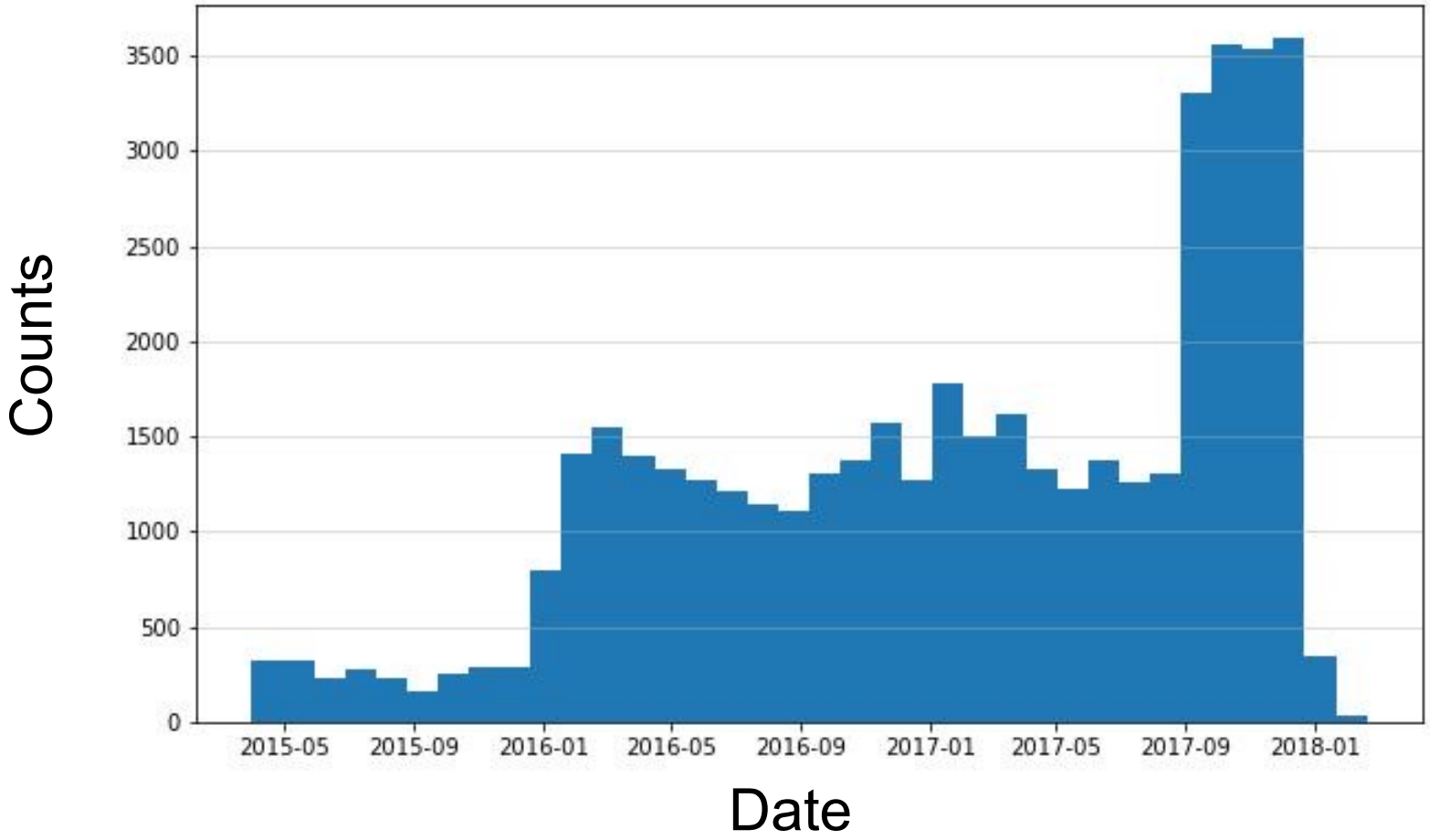
The Data

	title	text	subject	date	target
0	Donald Trump Sends Out Embarrassing New Year'...	Donald Trump just couldn t wish all Americans ...	News	December 31, 2017	0
1	Drunk Bragging Trump Staffer Started Russian ...	House Intelligence Committee Chairman Devin Nu...	News	December 31, 2017	0
2	Sheriff David Clarke Becomes An Internet Joke...	On Friday, it was revealed that former Milwauk...	News	December 30, 2017	0
3	Trump Is So Obsessed He Even Has Obama's Name...	On Christmas day, Donald Trump announced that ...	News	December 29, 2017	0
4	Pope Francis Just Called Out Donald Trump Dur...	Pope Francis used his annual Christmas Day mes...	News	December 25, 2017	0

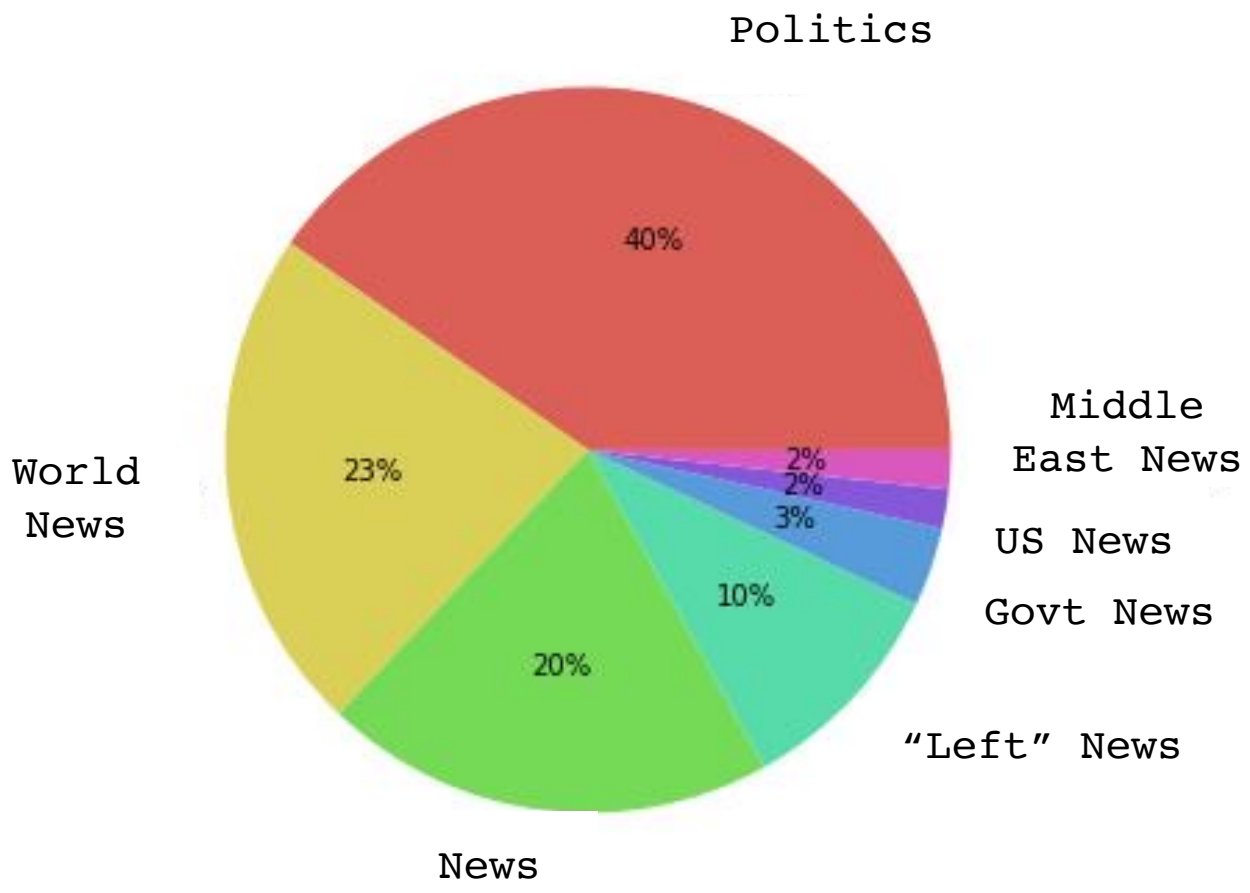
	title	text	subject	date	target
0	As U.S. budget fight looms, Republicans flip t...	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	December 31, 2017	1
1	U.S. military to accept transgender recruits o...	WASHINGTON (Reuters) - Transgender people will...	politicsNews	December 29, 2017	1
2	Senior U.S. Republican senator: 'Let Mr. Muell...	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	December 31, 2017	1
3	FBI Russia probe helped by Australian diplomat...	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	December 30, 2017	1
4	Trump wants Postal Service to charge 'much mor...	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	December 29, 2017	1

Article Counts over Time

The articles span May 2015 – Jan 2018,
with the majority from late 2017



Corpus Classification by Subject

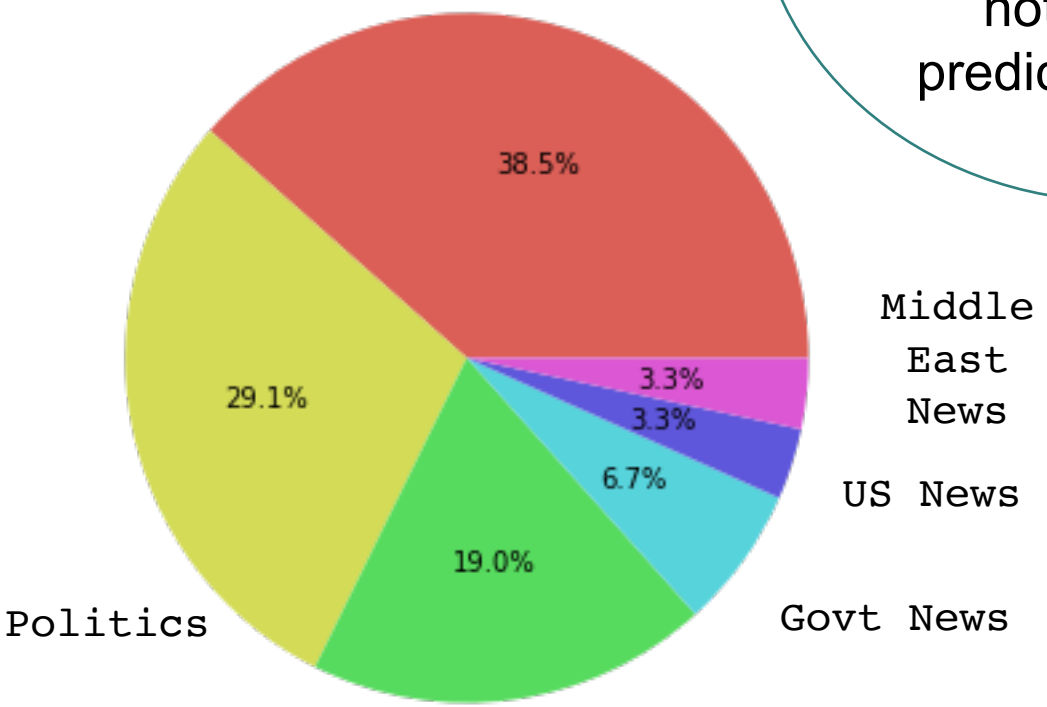


Could the ‘subject’ labels be a useful predictive feature?

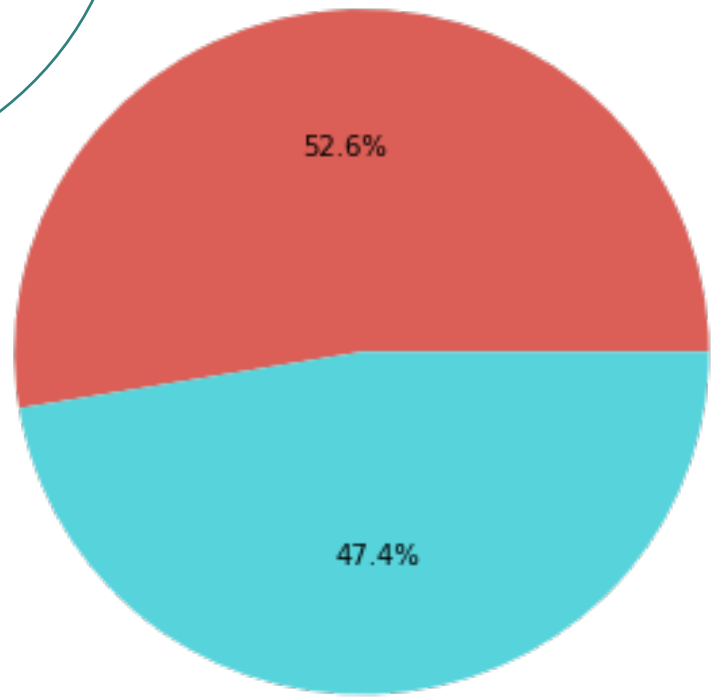
Corpus Classification by Subject

Subject labelling is inconsistent between the ‘fake’ and ‘true’ datasets. Therefore, this is not a useful predictive feature

Fake News



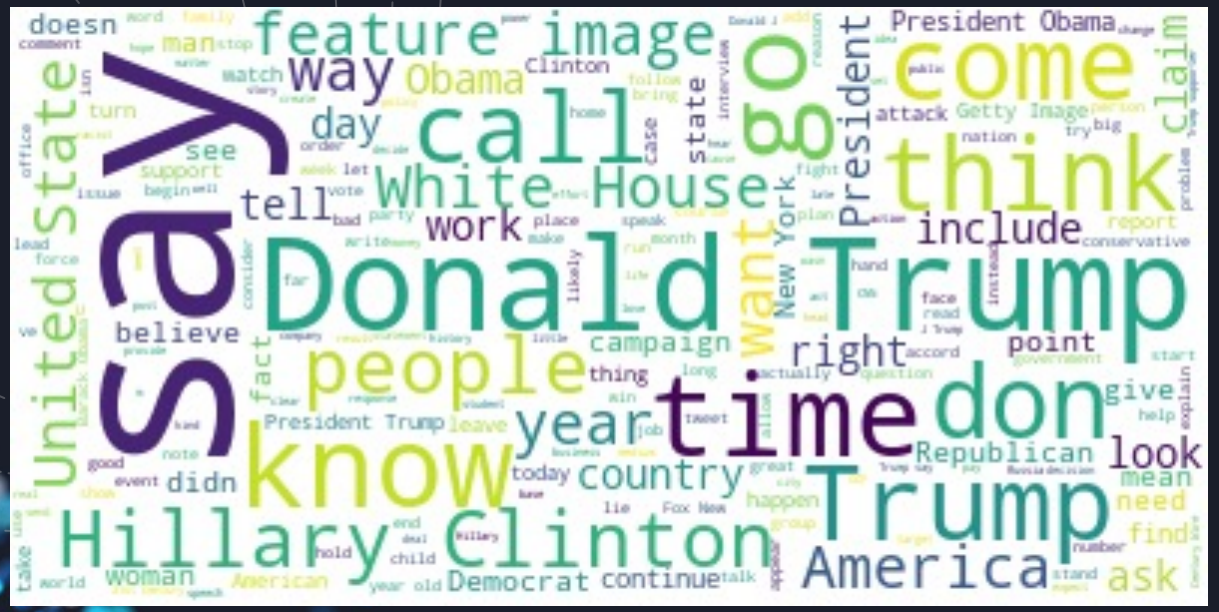
True News



Do the two classes differ in word content?

Fake News WordCloud!

38850618 words in the combination of all documents.



True News WordCloud!

34881474 words in the combination of all documents.

They differ surprisingly little!

Workflow

1. Data Cleaning

- Removal of STOPWORDS & punctuation
- Stemming
- Lemmatization

test/train/validation split (20:60:20)

2. Feature Engineering

- Word count, character count, word density, punctuation count, Title Case count, UPPER CASE count on both **'title'** and **'text'**
- Count Vectorization ('**title**' and '**text**')
- TF-IDF Vectorizer ('**title**' and '**text**')
  - bi-grams, tri-grams on both words and characters
- Topic modelling (LDA) ('**title**' only)

Dimension reduction
TruncatedSVD

Concatenation

Models

Naïve Bayes
Logistic Regression
Linear SVC
Random Forest
XGBoost

Results of preliminary modelling (non-optimized)

	Dataset 1 (X_train_combined) (26932, 460)	Dataset 2 (X_val_combined) (8978, 460)
Naïve Bayes	0.910367	0.904881
Logistic Regression	0.979578	0.979839
LinearSVC	0.969034	0.973155
SVC	0.814605	0.776331
Random Forest	0.986707	0.979060
Gradient Boost	0.988638	0.986967
XGBoost	0.994171	0.990309
Stacking	0.970370	0.974492

Accuracy Scores with Cross-Validation

Model	Set	Accuracy Score	Meta Classifier
SVC	Train	0.81 +/- 0.07	
	Validation	0.78 +/- 0.15	
Naïve Bayes (Gaussian)	Train	0.91 +/- 0.005	
	Validation	0.90 +/- 0.02	
Linear SVC	Train	0.97 +/- 0.01	
	Validation	0.97 +/- 0.008	
Stacking Classifier	Train	0.97 +/- 0.01	Logistic Regression
	Validation	0.97 +/- 0.01	
Logistic Regression	Train	0.98 +/- 0.002	
	Validation	0.98 +/- 0.004	
Random Forest	Train	0.99 +/- 0.001	
	Validation	0.98 +/- 0.002	
Gradient Boost	Train	0.99 +/- 0.002	
	Validation	0.99 +/- 0.004	
XGBoost	Train	0.99 +/- 0.0006	
	Validation	0.99 +/- 0.001	

SCORES (LOWER TO HIGHER)

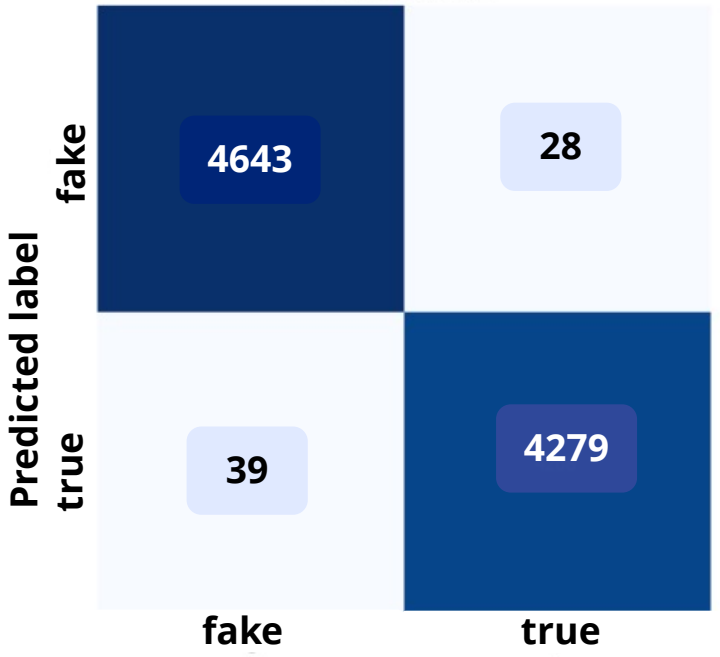
Fine-Tuning Hyperparameters produced models with >99% accuracy

Logistic Regression

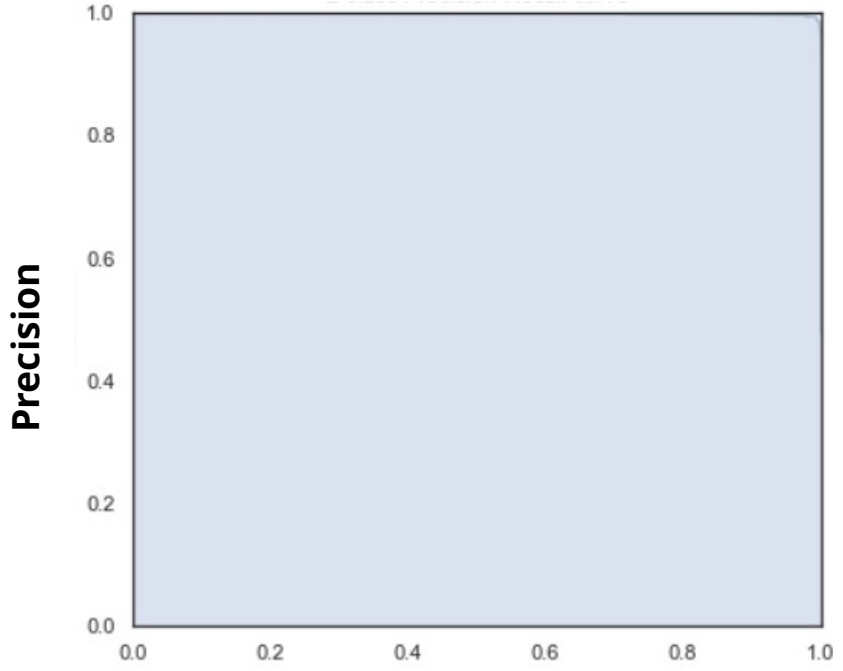
X_test

Accuracy:	0.9925
Precision:	0.9909
Recall:	0.9935

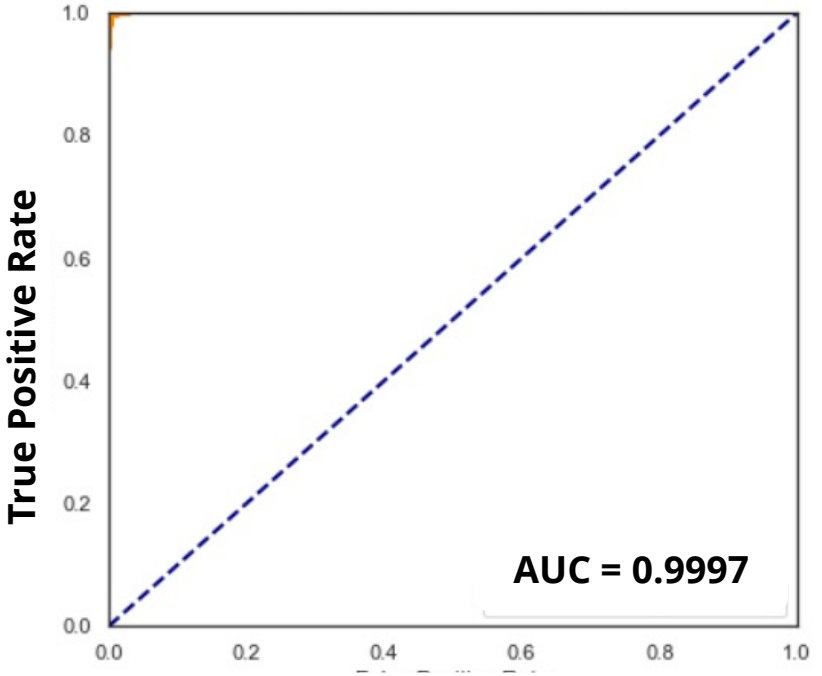
Confusion Matrix



2-Class Precision-Recall Curve



ROC Curve



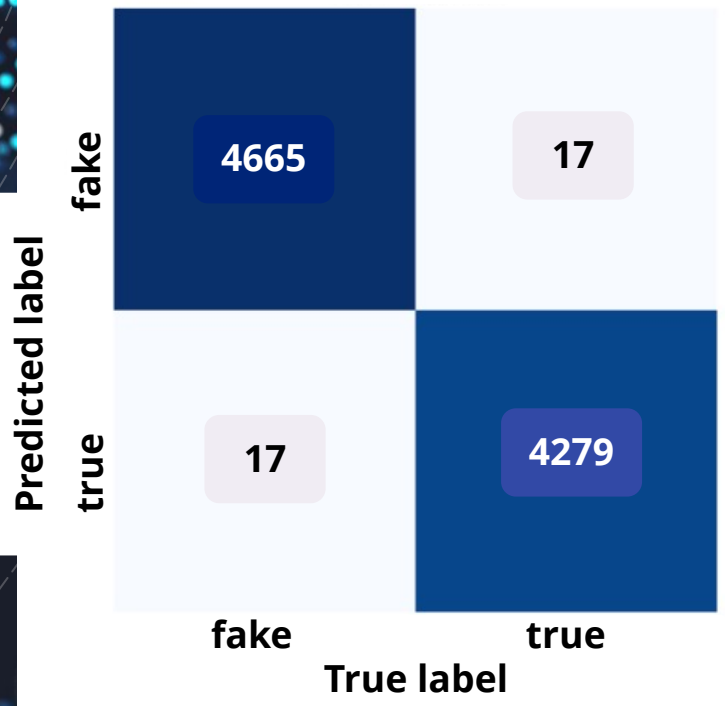
Fine-Tuning Hyperparameters produced models with >99% accuracy

XGBoost

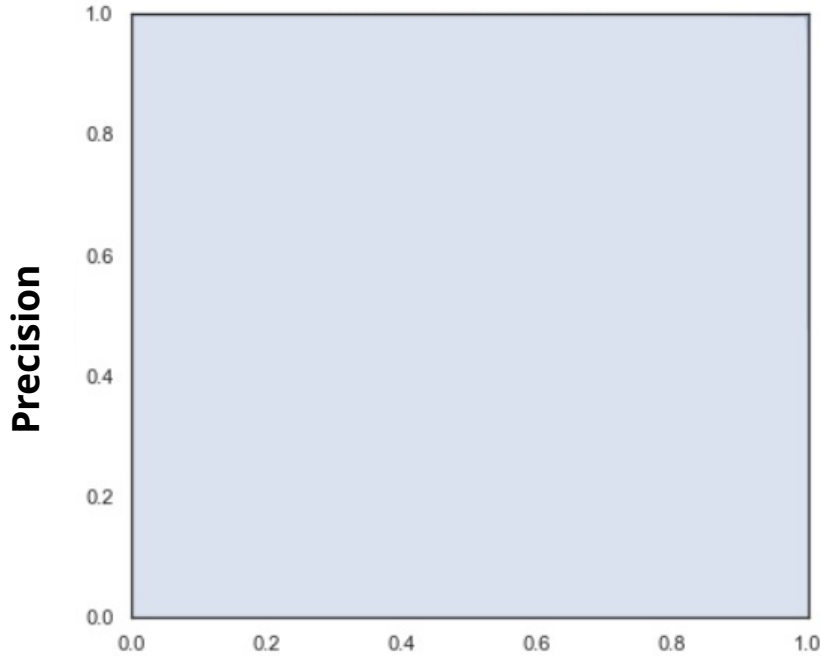
X_test

Accuracy:	0.9962
Precision:	0.9960
Recall:	0.9960

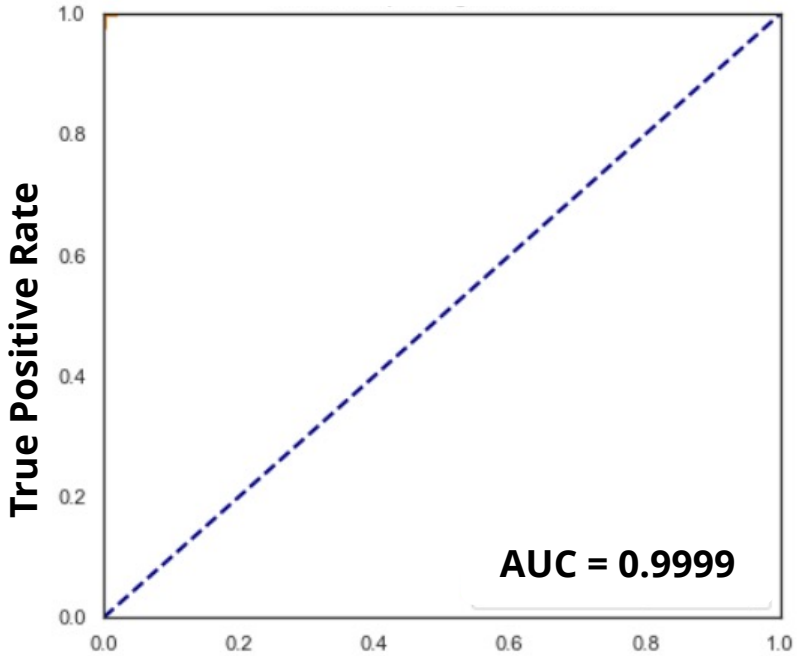
Confusion Matrix



2-Class Precision-Recall Curve



ROC Curve



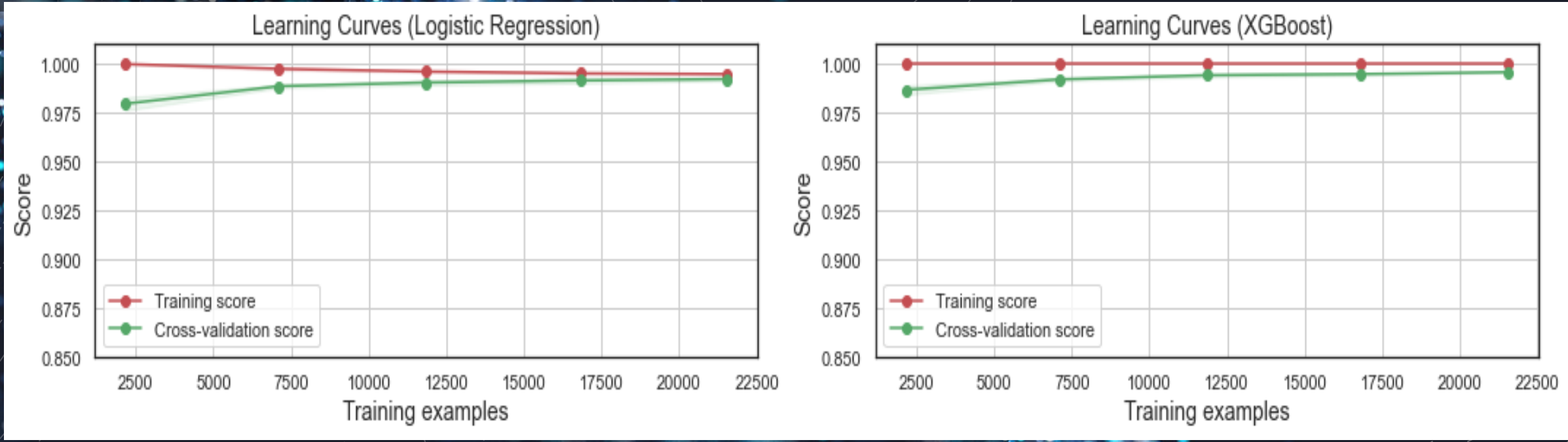
Model Evaluation
Learning Curves

Logistic Regression

Accuracy Train: 0.99 +/- 0.001
Accuracy Validation: 0.99 +/- 0.003

XGBoost

Accuracy Train: 1.00 +/- 0.0005
Accuracy Validation: 0.99 +/- 0.002

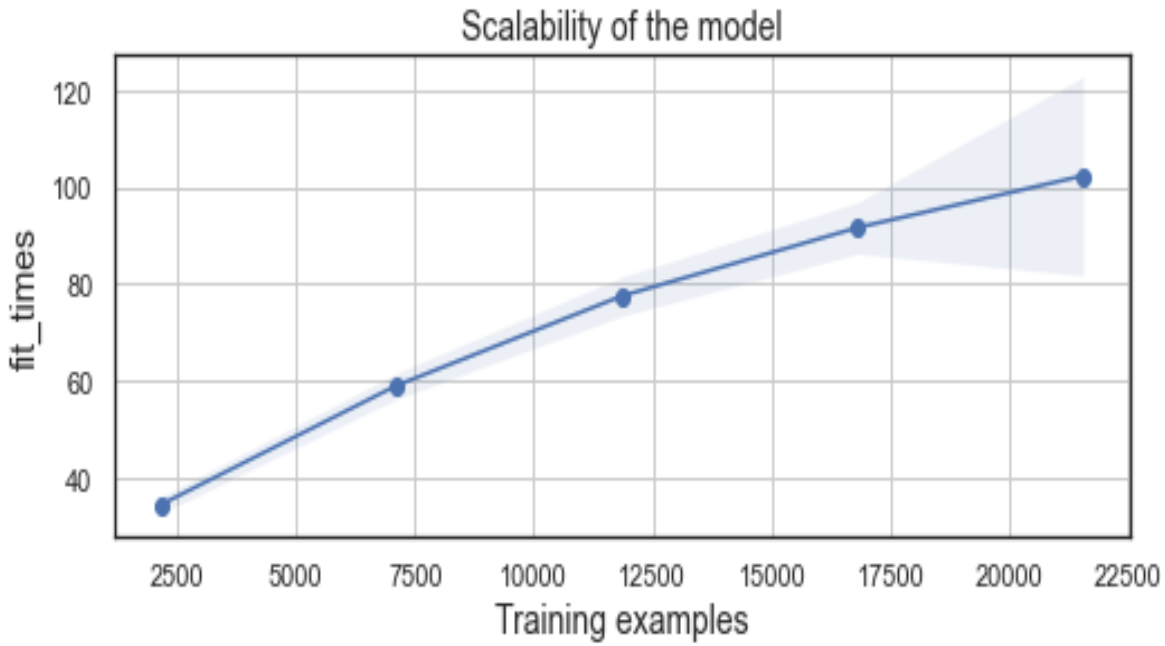


Convergence of training and cross-validation scores over training samples of increasing size indicates that the models are **not** over-fitting

Model Evaluation
Scalability

Logistic Regression

XGBoost



Fit times were on average slightly faster for the XGBoost model than for the Logistic Regression model

Conclusions

- All models (with the exception of SVC) were highly accurate, even without optimization; there was very little trade-off between precision and recall.
- XGBoost and Logistic Regression models were fine-tuned and evaluated for both accuracy (+ precision and recall) and fit speed: the models were equally fast to fit, at around 100-125 seconds for 22500 training samples, with an approximately linear relationship between sample size and fit time.
- For the highest accuracy, XGBoost is recommended: for almost 9000 samples, only 17 were incorrectly classified as ‘fake’ and 17 incorrectly classified as ‘true’. Predictions were 99.62% accurate for the test dataset.
- The most computationally-demanding step was the processing of the text data in feature engineering (in particular, stemming, lemmatization and tagging ‘parts of speech’ (POS).
- Topic modelling on the full text may improve model accuracy further, but this is unlikely to be worth the effort.