Is This Real News or is it Fantasy?

Assessing Readability Scores as Potential Features in Categorizing Real from Fake News

FAKE NEWS
FAKE NEWS EVERYWHERE

Executive Summary

- Fake news is a major problem for news organizations and social media sites.
- Our machine learning model could potentially reduce churn and increase revenue by $16 million per month.

Flesch Reading Ease:

206.835 - (1.015 * **words / sentences**) - 84.6 * (**syllables / words**))

Flesch-Kincaid Grade Level:

.39 * (**words / sentences**) + 11.8 * (**syllables / words) - 15.59**

Automated Readability Index (ARI):

4.71 * (**characters** / **words**) + .5 * (**words** / **sentences**)

Coleman-Liau Index:

(.0558 * avg **letter** count / 100 **words**) - (.296 avg **sentence** count / 100 **words**) - 15.8

Gunning-Fog Index:

Index features 3+ **syllable words**

Linsear Write Formula:

Formula features 1 **syllable words**

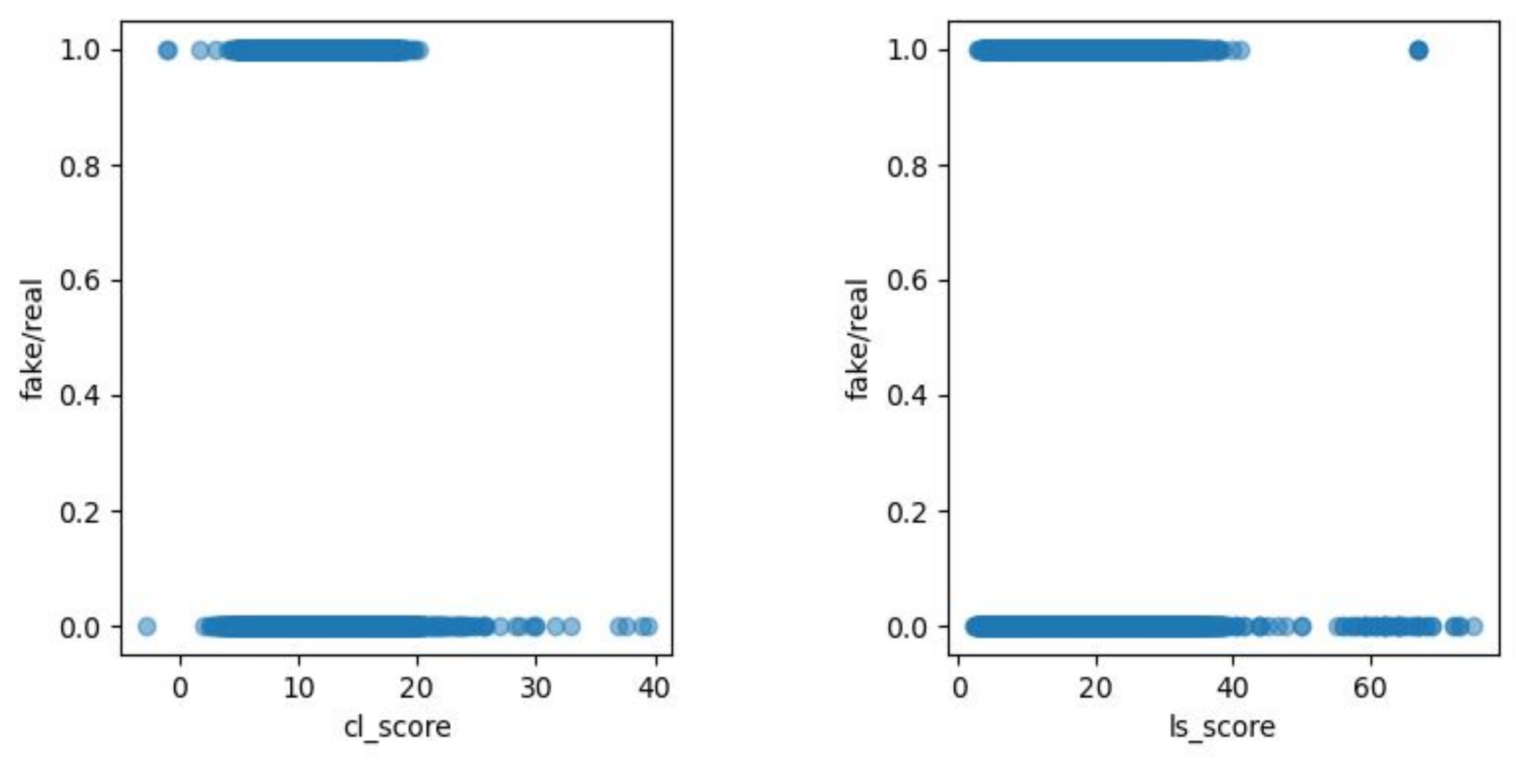
Dale-Chall Score and Spache Formula:

Both feature a word list familiar to young readers

Readability scores show promise as features in EDA:

Coleman-Liau

Linsear Write



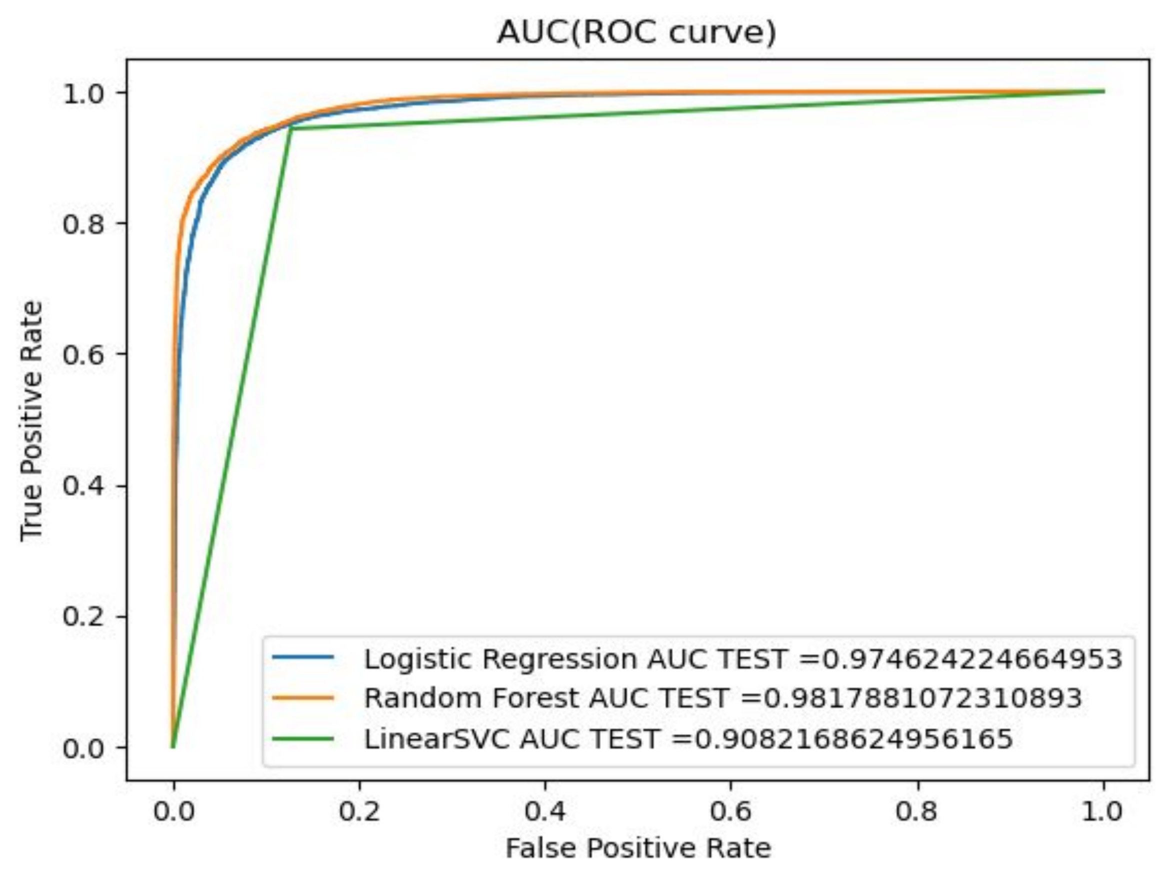
TF-IDF Vectorization:

TF = (Number of repetitions of word in a document) / (# of words in a document)

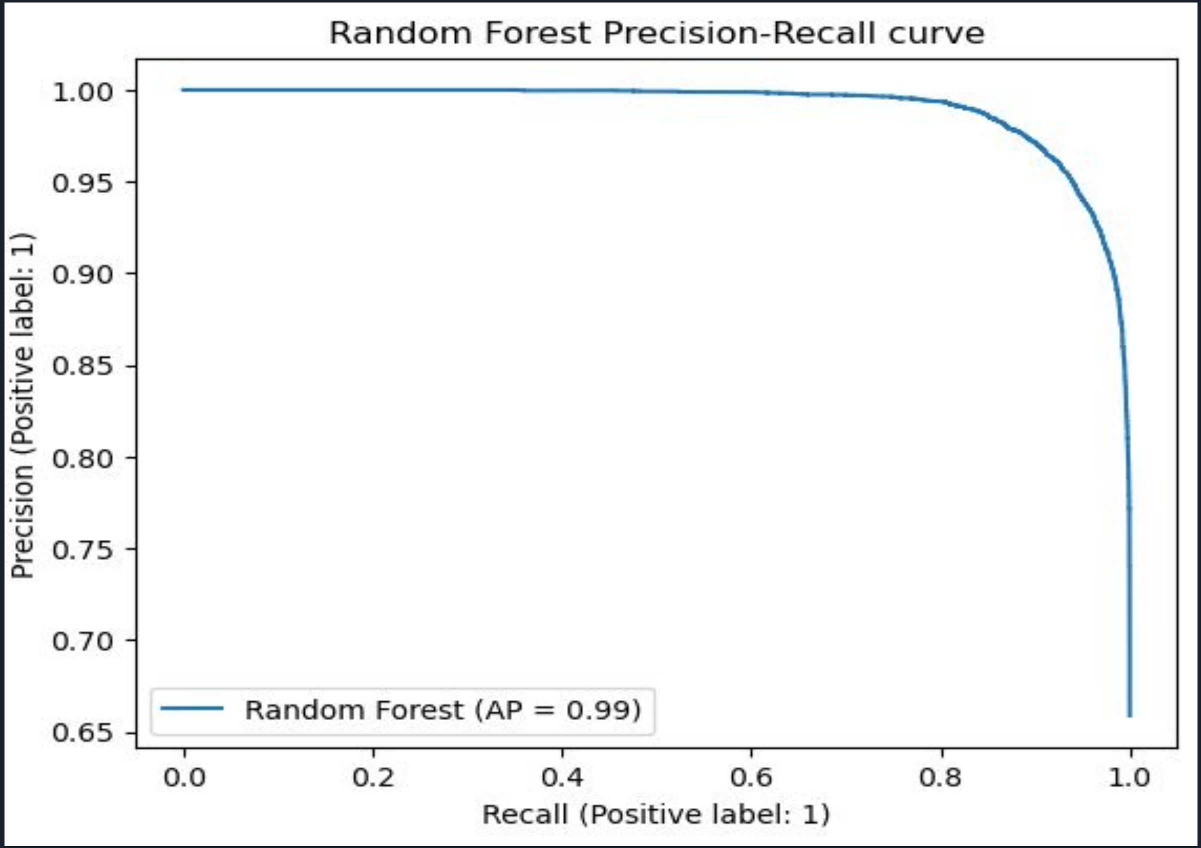
IDF =Log[(# Number of documents) / (Number of documents containing the word)]

If a term is used frequently in the document, and not frequently in the document set the TF-IDF score will be higher for that word for that document

ROC Curves


AUC(ROC curve)
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Logistic Regression AUC TEST =0.974624224664953
Random Forest AUC TEST =0.9817881072310893
LinearSVC AUC TEST =0.9082168624956165

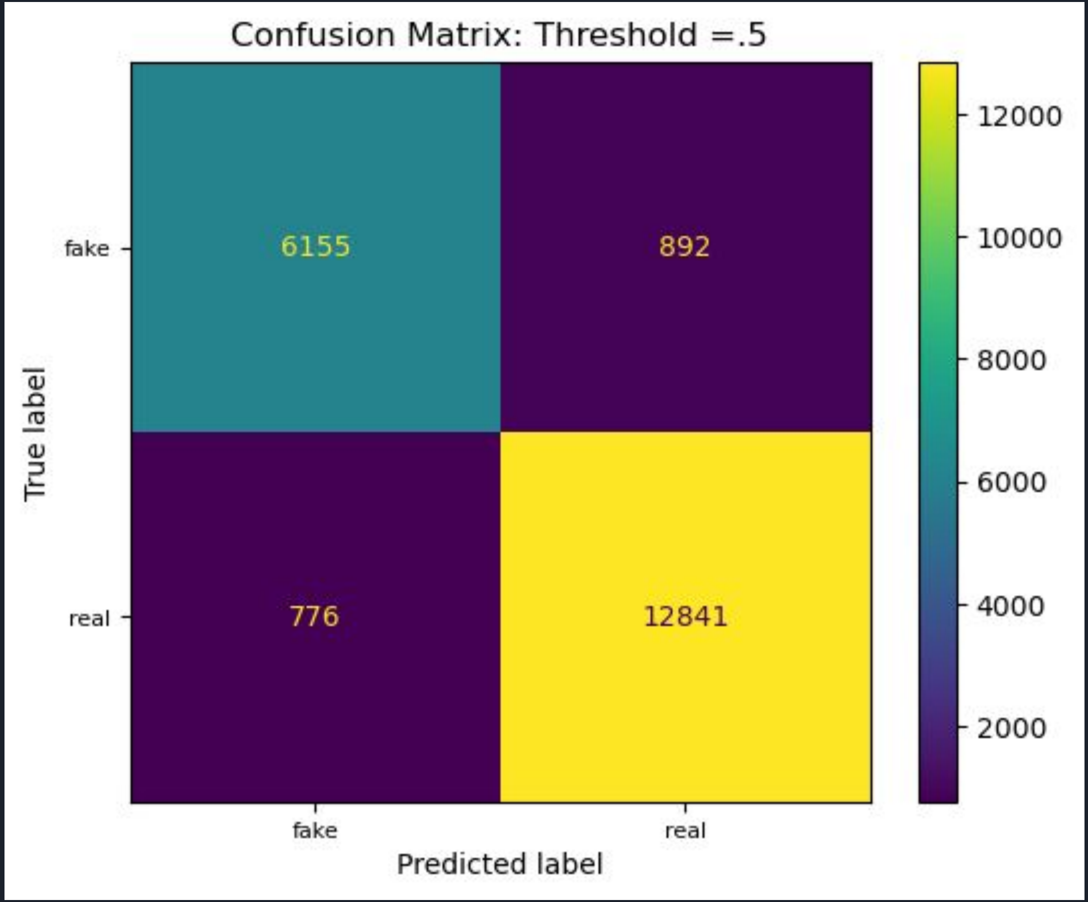
Precision Recall Curve



Top Ten Features

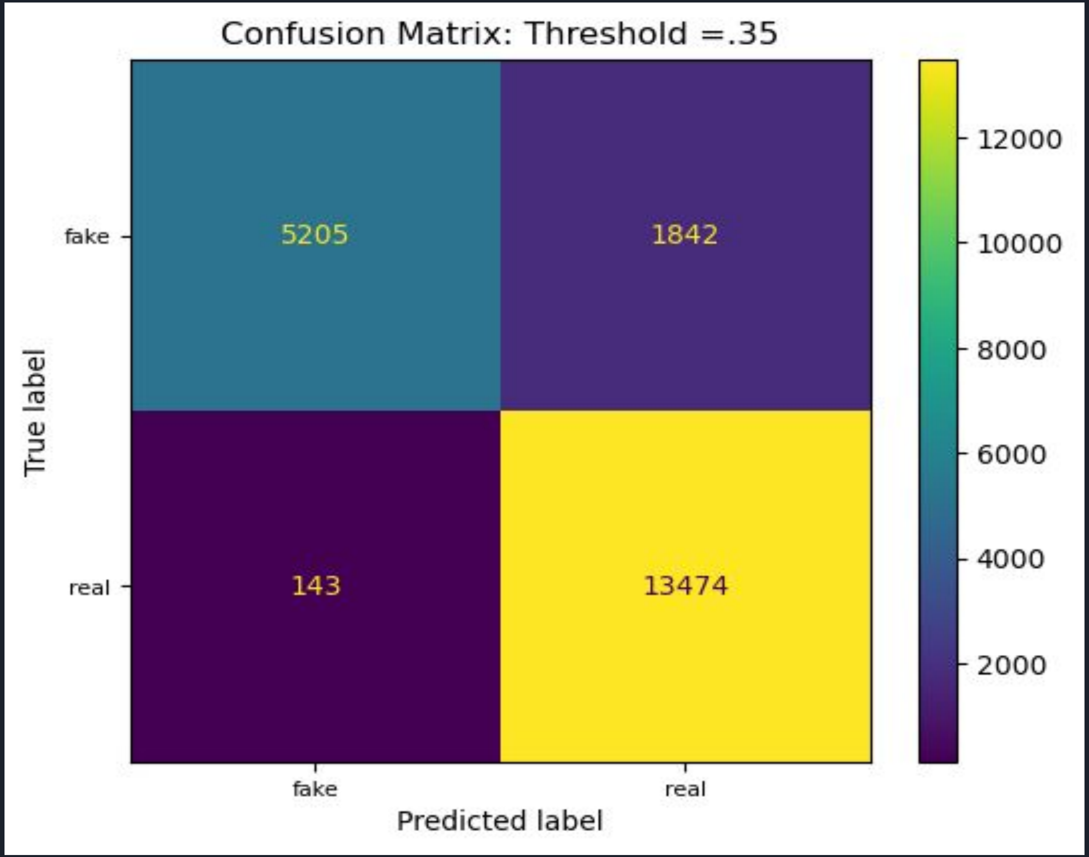
1. ‘said’
2. ‘reuters’
3. ‘featured’
4. ‘image’
5. ‘com’
6. ‘hillary
7. ‘just’
8. ‘tuesday’
9. **Dale-Chall Score**
10. ‘thursday’

Readability scores were vastly overrepresented in the top 10% of features



Confusion Matrix: Threshold =.5

True label \ Predicted label	fake	real
fake	6155	892
real	776	12841


Confusion Matrix: Threshold =.35
True label
fake
real
5205
1842
143
13474
fake
real
Predicted label
12000
10000
8000
6000
4000
2000

- Monthly churn for MyFace is estimated to be around 1-2%.
- 44% of users who stopped using a social media platform said a major reason why was fake news.
- Average Revenue Per User is estimated to be about 3.33$ per month. MyFace has about 2.96 billion monthly users.
- Incorrectly labeling real news is costly, thus the threshold choice prioritized minimizing this. This came at a cost to the recall score which is .74.
- .01 x .22 x 3.33$/month per user x 2.96 users x.74 ≈ $16 million/month