Kylee Laczkovich and Barbara Uzun

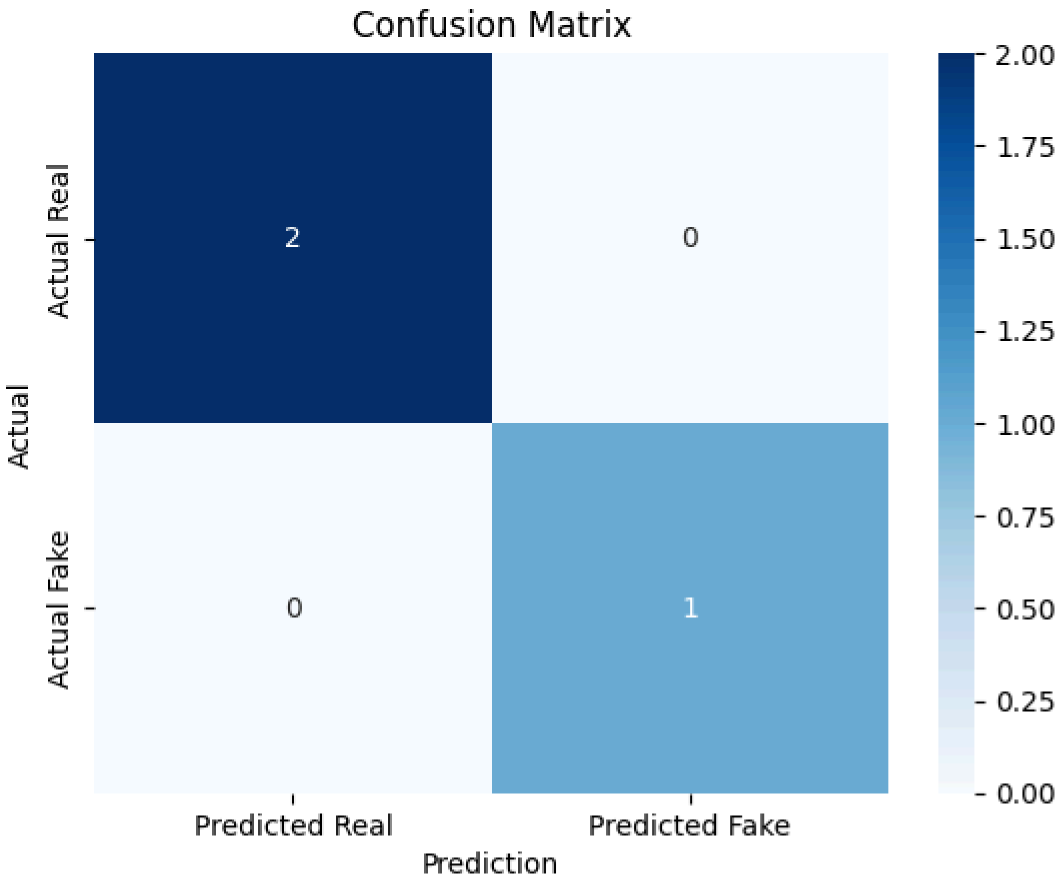
Dr. Johnson

DS 3001

30 April 2025

Machine Learning Project Results

Does the data set contain real numeric data or fabricated numeric data? This is the question we aim to answer with our model through analyzing the kurtosis, skewness, and benford’s law features of each dataset.



This is a confusion matrix for testing the accuracy of our model. Our dataset is a bit smaller than most (12 datasets) because each requires some cleaning beforehand to make sure it runs through the model. We also only focused on numeric data, as this is easier to mathematically detect fabricated data by running these well-known tests. We performed the 80-20 split of training and testing data, so this resulted in testing three datasets for one confusion matrix output. The matrix indicated that the model was highly accurate with no false positives or false negatives.

Index	**Benford Error**	**Mean Skewness**	**Mean Kurtosis**
0	0.3105	-0.0029	-1.2087
1	0.1874	24.7502	746.3266
2	0.3744	-0.0125	-1.1929
3	0.3230	0.0852	-1.0900
4	0.1406	1.6099	7.8481
5	0.0631	2.4162	7.0540
6	0.9254	-0.0362	-0.2406
7	0.4622	11.5551	223.6456
8	0.2459	2.1435	21.8211
9	0.4458	0.6098	1.8016
10	0.4800	0.4505	-0.3877
11	0.4800	0.4505	-0.3877

These are the test results for each of the datasets. This is what the algorithm used to make the classification prediction of a real or fake dataset. We did not specify an exact cutoff for a given column to indicate fake or real. The model determined this on its own through supervised learning. This is because there is no absolute rule for a cutoff and varies depending on the data being used, whereas these datasets come from different sources and topics. However, there are some general guidelines for these tests such as generally if Benford's error is greater than 0.40-0.5, this indicates that the data is fabricated. Moreover, for kurtosis, if the mean is less than zero, this suggests artificial symmetry or if there are extreme kurtosis values this suggests manipulation as well. Lastly, for mean skewness, real datasets mainly have moderate skewness between (-1,1), so extreme values outside of this are suspicious.

Accuracies:	[1.0, 1.0, 0.95, 1.0, 1.0]
Mean Accuracy:	0.99
Standard Deviation	0.020000000000000018

Another numeric summary we have is a k-fold cross validation. This method assesses the model's accuracy across multiple runs of the data. This results in reduced bias as the resampling method allows for different training and testing splits and robust performance checks. Overall, the output is highly accurate with a mean accuracy of 0.99 and 0.02 standard deviation. This is a great performance outcome and means that it could be potentially a valuable model for testing datasets in the future.

Synthetic Dataset Classifications

df_13	Fake	Confidence: 0.77
df_14	Fake	Confidence: 0.75
df_15	Fake	Confidence: 0.95

One potential criticism for the current approach would be the classification of synthetic datasets. So, we tested our model on three synthetic datasets to see if there were any patterns. For all three synthetic datasets, the model classified them as fake with relatively high confidence levels. This is an important distinction. This means the simulation of the real data does not fit the normal patterns closely enough to be considered real. Another potential concern would be human error, but this should not make the difference between a fake and real distinction since there are certain mean thresholds for each test.