Executive Summary
Fake-News Classification

Objective:
We aimed to develop a machine-learning model capable of analyzing news articles and classifying them as "real" or "fake." Additionally, we strived to understand how synthetic features could enhance the accuracy of these models in the case where the source of the news is either unknown or purposefully misrepresented. This would be a valuable tool for new media organizations and individuals.

Data Source:
We utilized a dataset from a Kaggle competition focused on fake news detection. The dataset contained a comprehensive collection of news articles, their titles, and author(s).

Feature Exploration:
In addition to the typical text vectorization techniques, we explored various attributes that could contribute to the detection of fake news. These attributes were drawn from the existing literature and included:

- Number of grammatical and spelling errors in the article.
- Frequency of stop words used in the article.
- Emotional classification using DistillBERT.
- Number of words capitalized in the article.
- Presence of numbers in the article.
- Frequency of long words (more than 10 or 15 characters) in the article.

Modeling and Accuracy:
We employed three different machine learning algorithms: Logistic Regression, Gaussian models, Random Forest, and Naive Bayes. Our final accuracy when we didn't consider the source of the news was 0.67, which even though it was below the competitive threshold in Kaggle is better than random. Comparing the high accuracy in the kaggle competition shows the source of news is highly correlated with truthfulness.

Feature Analysis:
Through thorough analysis, we identified the most relevant attributes that significantly improved accuracy. The features that yielded the highest improvements were Emotions, Word Count and Stop-Words Counts.

Conclusion:
We successfully developed a machine-learning model capable of analyzing news articles and classifying them as "real" or "fake". Moreover, our exploration of synthetic features highlighted the importance of attributes such as grammatical errors, stop words, emotional classification, capitalized words, numbers, and long words in determining the authenticity of news articles.

Moving forward, further research and experimentation could potentially enhance our model's accuracy, bridging the gap to compete effectively in the Kaggle competition. Our findings and the developed model offer news organizations a valuable tool to detect and combat the spread of fake news, ultimately fostering a more informed society.

In summary, our project exemplifies the importance of good data science practices, which involve engaging with prior research on the topic, and thoughtfully selecting relevant features. By taking this approach, we strive to develop models that not only deliver competitive accuracy but also provide meaningful and interpretable results, ultimately contributing to the advancement of the field and its practical applications.