# Running the Colab

Our Google Colab notebook can be run sequentially and consists of the following sections:

1. Loading Python Libraries \*
2. Reading raw data into Google Colab
3. Data Pre-Processing
4. Preliminary Data Analysis
5. Data Cleaning
6. Feature Engineering
7. Load df with final features for ML \*
8. Model Evaluation Functions \*
9. ML Algorithms \*
10. Final Model Performance of Best ML models \*
11. Packaged Function

The notebook can be run in two ways:

* sequentially
  + How: Run all sections in order
  + Details: Includes feature engineering and machine learning
* models only:
  + How: Run all starred (\*) sections sequentially
  + Details: oads saved dataframe with engineered features to run ML models only

The Packaged Function in section 11, requires the individual functions that were used during section 6, Feature Engineering to be run and also to have the final model trained from section 9 ML algorithms.

# References

**Data Cleaning**

<https://medium.com/mlearning-ai/nlp-a-comprehensive-guide-to-text-cleaning-and-preprocessing-63f364febfc5>

**Feature Engineering**

<https://nlp.stanford.edu/projects/glove/#:~:text=The%20GloVe%20model%20is%20trained,corpus%20to%20collect%20the%20statistics>.

<https://towardsdatascience.com/natural-language-processing-for-fuzzy-string-matching-with-python-6632b7824c49>

<https://towardsdatascience.com/the-quora-question-pair-similarity-problem-3598477af172>

<https://flavien-vidal.medium.com/similarity-distances-for-natural-language-processing-16f63cd5ba55>

**Evaluation Metrics**

<https://stackoverflow.com/questions/55892224/how-to-use-log-loss-scorer-in-gridsearchcv/55909827#55909827>

**Models**

[Ray, S. (2023, May 26). Top 10 Machine Learning Algorithms (with Python and R Codes). Analytics Vidhya.](https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/)

[R, S. E. (2023, October 26). Understand random forest algorithms with examples (Updated 2023). Analytics Vidhya.](https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/)

[Van Otten, N. (2023, October 31). Top 15 most popular Machine learning and Deep learning Algorithms for NLP. Spot Intelligence.](https://spotintelligence.com/2022/12/23/machine-learning-deep-learning-algorithms-nlp/)

[Guest\_Blog. (2023, March 30). Introduction to XGBOOST Algorithm in Machine Learning. Analytics Vidhya.](https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/)

[Ali, R., and Benjamin, R. (2017). Random Features for Large-Scale Kernel Machines](https://papers.nips.cc/paper_files/paper/2007/hash/013a006f03dbc5392effeb8f18fda755-Abstract.html)

**Hyperparameter Tuning**

[Jordan, J. (2018, December 5). Hyperparameter tuning for machine learning models. Jeremy Jordan.](https://www.jeremyjordan.me/hyperparameter-tuning/)

[Koehrsen, W. (2019b, December 10). Hyperparameter tuning the random forest in Python - towards data science. Medium.](https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74)

[Rithp. (2023, January 16). Optimizing XGBOOSt: A Guide to Hyperparameter tuning. Medium.](https://medium.com/@rithpansanga/optimizing-xgboost-a-guide-to-hyperparameter-tuning-77b6e48e289d)

**Recursive Feature Selection**

[Brownlee, J. (2020, August 27). Recursive Feature Elimination (RFE) for feature selection in Python. MachineLearningMastery.com.](https://machinelearningmastery.com/rfe-feature-selection-in-python/)

[RFECV machine learning feature selection taking far too long Python. (n.d.). Stack Overflow.](https://stackoverflow.com/questions/58797050/rfecv-machine-learning-feature-selection-taking-far-too-long-python)