

Take Home Assignment

Overview:

You're a Machine Learning Engineer working for a new social media company called ABC Co. The data science team has uncovered a recent surge in the frequency with which news articles are shared over ABC's platform. While this is good news for customer engagement, the product team is worried about the prevalence of misinformation on the platform. To help mitigate this issue, they have asked you to design a solution for identifying which articles are authentic and which are not.

Dataset:

The product team has compiled a sample dataset to test your solution consisting of three *csv* files: *train.csv*, *test.csv*, and *labels.csv*. The training dataset has over 20,000 articles with fields for the article title, author, and text. The *label* field is the outcome variable where a 1 indicates the article is unreliable and a 0 indicates the article is reliable. The testing dataset consists of over 5000 unlabeled articles. The corresponding labels for those articles can be found in *labels.csv*.

[Dataset Link](#)

Deliverables:

We would like you to build a pipeline for predicting the reliability of the articles in the dataset using any machine learning method of your choice (you are not constrained to a deep learning approach). Please do not submit an IPython Notebook or other interactable environment. Instead, submit a folder containing your *.py* file(s).

While producing a good result on the test data is ideal, the performance is not the main focus of this assignment. We are interested in your code quality, and your design process. We would like you to justify the design decisions you made such as:

- The overall approach
- The preprocessing steps
- Evaluation procedure and metrics
- Aspects that could be improved
- etc.

These justifications can be made either with code comments, or in a README file.

To be mindful of your time, this project should not take more than 2-3 hours to complete. Use this as a benchmark to gauge the depth/detail required in your solution.

Submission:

To submit your assignment, please reply to this email with either a GitHub, GitLab, or Google Drive link to your code.

If you have any questions, please feel free to reach out.

Good luck!