Data science coding exercise

Objective:

Developing a Machine Learning (ML) model that predicts whether a news article is related and important to the real estate industry or not. You are provided a dataset of previously tagged news articles that are fetched and related to the industry. You should develop a web API through Flask that receives the news article's title and link as input and predicts whether the news is related using the previously developed machine learning model.

Further, with each passing day, as new training data becomes available, the ML model should be retrained using the newly available data to make it more relevant.

You could use any other data available at your disposal that seems to have a positive impact on the quality of your output. For example, it is possible to fetch the news contents through automatic tools.

Dataset:

The dataset consists of two files: *train.csv* and *test.csv*. You can train and optimize your model on the *train.csv* file, and you should submit your results on the news articles in the *test.csv* file. Each row in these two files corresponds to a news article published on the web. The news articles are fetched daily. Each news article contains the following info:

- *Title*: The fetched title for the news article.
- *URL*: The web address for the article.
- *Author:* The news website or website(s) publishing this news. Only one of the URLs for the news publishers are available.
- *Snippet:* A short part of the news article returned from the search engine.
- *Related*: Whether the news article was related and important to the industry or not.

Deliverables:

You should email us a *prediction.csv* file containing these data for each of the rows in the *test.csv* file:

- *Title*: The fetched title for the news article in *test.csv*.
- *URL*: The web address for the article in *test.csv*.
- *Predicted*: Whether the news article was predicted by the ML model as being related.

Also, you should upload your final solution developed in Python to a private repository on Gitlab and share it with @ehsanroomvu on the platform.

Finally, you should provide a written report to us of what you did to achieve the objectives of this exercise.

Evaluation:

Your output's quality is assessed through these criteria:

- The exploratory data analysis steps you performed along appropriate data visualizations to supplement your decisions and approaches.
- The suitable selection of metric(s) to evaluate the performance of your ML model against other candidate ML approaches.
- Hyper parameter optimization of your model.
- Overall architecture of training and productionizing your ML model.
- The cleanliness and readability of your codes.
- The proper use of some automatic testing tools to enforce code standards and making sure the code works as intended. These tools include but are not limited to black, flake8, pylint, pytest, mypy, and pytype.
- The use of Gitlab CI/CD tool to incorporate the above mentioned automatic testing tools, especially for the CI part since you do not want to deploy your solution.

Enquiry:

If you have any questions regarding this coding exercise do not hesitate to contact ehsan@roomvu.co for further guidance.

Best of luck,
Roomvu