About the Task

Machine Learning Roles / Business Analyst

Important Points:
(1) Please use separate .py files to harvest ONLY the data from API or scrap from websites.
(2) Use different / alternative data for training your data, including technical features, or other country and financial features. Do not use just the history of the price as features.
(3) Use a combination of models of your choice to benchmark the accuracy, you can use both classification (up or down), or time series models. Our preferred models are MLP, RNN, LSTM, or GRU - for time series, and Logistic, Random Forest, Naive Bayes Classifier for classification. You are free to use other models
(4) Provide a report with succinct visualization of results and all your different .py scripts (class object oriented good scripting practices) and final Python notebook.
Please submit a PDF of google slides or document presenting your findings. Your evaluation criteria is partially technical and partial the ability to explain meaningful results in a presentable manner.

Time-Series Analysis

- Use any daily time series from Investing.com or similar source with a strong sample of covariates. Target commodities price like: Oil, Natural Gas, Resin, or Metal Prices.
    - Please make sure to get an extensive list of feature space, think through structural other external factors.
    - Option 1. Feature Importance. Dynamic Time Warping and/or XGBoost/Shapley Value hybrid model approach to quantify which factors influence the target positively or negative
        - **Keep the analysis focused on the feature selection and feature importance aspects**
    - Option 2. LSTM derivatives on day ahead prediction with confidence bounds
        - How would you improve and present your results with more time and resources
        - Related thinking and planning in a short report

Natural Language Understanding

- Use World Bank Projects dataset
    - Option 1. Tagging for Keyword Extraction or Named Entity Recognition (NER) using models such as YAKE, BERT-derived models, spaCy, or Google NLP API
        - Specifically identify either “sector” or “sub-sector” or entities like Government Agency, Company Name, Contractors, Investor, or unit measurements such as cost per square kilometer.
    - Option 2. Binary Classifier
        - Using the status variable build a binary classifier to predict the probability whether a project will be “closed” or “canceled/distressed”