# **Executive summary**: Industrial resources and its impact on stock market

Git: https://github.com/YCoeusZ/industrial-resources-and-its-impact

Contributors: Kayode Oyedele, Rafael Almeida Fernandes, Thiago Brasileiro Feitosa, Wojciech Tralle, Xiangyi Tao, and Yuan Zhang

**Motivation:**

It is a macroeconomics fact that the cost of production has an impact on the performance of an industry. We expect that the stock data of relevant industries will be a valid indicator of the performance of the industry. We, hence, look to use macroeconomic data and price data on industrial resources as parameters to train a supervised model on the target of stock data. In particular, we will focus solely on the technological industry in this project. We will care about the impact of contextual variables on the tech stock index: so instead of looking at the stock index itself, we will focus on the proportional change of the stock index relative to certain numbers of previous (trading) days; and, similarly, instead of the value of the parameters, we focus on the proportional change of those variables. However, certain variables are contextual on their own, for instance, the federal funds rate.

As a clarification: the goal is **NOT** to predict the future of the stock market (e.g. predicting the index of tomorrow with data we have today). Instead, we are using the stock index as a representation of the state of the industry, and attempt to model the change in that state based on change of macroeconomic parameters (e.g. we may use proportional price change of copper today as one of the parameters for the target of proportional change of stock index of today).

**Raw data collection:**

- Daily data on the total S and P 500 stock index between 2009 and 2025.
- Daily data of relative weight contribution of each S and P 500 companies. Notice that: Companies enter and exit S and P 500 throughout time depending on their performance.
- SIC (Standard Industrial Classification codes), and stock tickers (stock identifiers) of companies.
- The SIC to NAICS (North American Industry Classification System) crosswalk data.
- Monthly data of PPI (Producer Price Index) classified by NAICS. These data are used as part of the parameters indicating the cost of production.
- Daily data on raw industrial resources involved in the tech industry: gold, silver, copper, crude oil, palladium, platinum. These data are used as parameters influencing production costs.
- Daily data on federal fund rates. These data are used as parameters.
- Monthly data of CPI (Consumer Price Index). These data are used as indicators of general inflation.

**Data Processing:**

- **Classifying the companies:** Produced, in accordance with NAICS, a list of companies considered as “tech industry”, and retrieved (monthly) PPI data according to the NAICS chosen.
- **Producing Weighted Stock index of tech companies:** Created a list of “stable companies” that does not exit or enter the index between the start and end of interested timeframe (2014 to end of 2024), and combined the raw data on their weight contribution and total S and P index to receive the daily data on tech stock index.
- **Creating proportional change daily data:** Generated daily proportional change data (recorded as percentage data by multiplying with 100) for each of the tech stock index and the raw resources.
- **Create proportional change monthly data:** Calculated the proportional change data in the context of monthly values and “expanded” them to daily values.
- **Create inflation adjusted data:** Created inflation adjusted data relative to 20th (trading) day prior.

**We have created two datasets for modeling:**

***No inflation adjusted dataset:*** For each row (one for each trading day), we have in columns:

- **Target**: proportional change of tech stock index relative to previous trading day.
- **Parameters:** proportional change of price of raw resources relative to previous trading day, proportional change of PPI’s relative to previous month, federal fund rate.
- **Train and test split:** Test set start on date 2024-Jan-01 (**Target has (population) Variance: 1.6098**).

***Inflation adjusted dataset:*** For each row (one for each trading day), we have in columns:

- **Target**: inflation adjusted proportional change of tech stock index relative to 20th trading day prior.
- **Parameters:** inflation adjusted proportional change of price of raw resources relative to 20th trading day prior, inflation adjusted proportional change of PPI’s relative to previous month, federal fund rate.
- **Train and test split:** Test set start on date 2024-Jan-01 (**Target has (population) Variance: 24.1491**).

**Models, cross-validation, and Results:**

The metric we will be using is the mse (mean squared error).

- **EBM (Explainable Boosting Machine):**

- EBM team: Yuan, Kayode, Wojciech.
- Cross-validation: We did cross-validation on changing shifting numbers (which decide how much historical data to include, the higher the shifting number, the more history is included), and if we were to include extreme training data. The cross-validation is done with time series 5-fold.
- Results:
  - **On the not inflation adjusted data:** Our preferred final models produce: mse 1.5846 (R^2 value 0.0157) trained with outliers and 10 shiftings, and mse 1.5694 (R^2 value 0.0251) trained with outliers without any shiftings.
  - **On the inflation adjusted data:** Our preferred final model produces: mse 38.4721 (R^2 value -0.5931) trained with extreme data and 3 shiftings.

- **NN (Neural Network):**
  - NN team: Yuan, Kayode, Wojciech.
  - Cross-validation: We created two different models, one with a RNN (recurrent neural network) to include the impact of historical data, and another with an attention layer to include in the consideration of interactions. For each of these, we proceed with cross-validation by changing epoch numbers of each model. The attention layer model also includes cross-validation through including different lags to include lagged data from history. The RNN also includes lag length (determining how much past data to include).
  - Results: (**Clarification: We only worked on the dataset with inflation adjustment with NN models**)
    - **RNN:** Our preferred model produced mse 31.4635 (R^2 value -0.3029) trained with 190 epochs and 6 lag length.
    - **Attention layer model:** Our preferred model produced mse 23.06 (R^2 value 0.0451) trained with no lag and 300 epochs.
- **SPLINE:**
  - SPLINE team: Xiangyi, Thiago, Rafael.
  - Cross-validation: We used a five-fold time-series cross-validation. We explored different values for the spline transformer's parameters (trying out several n_knots and degree settings by manually adjusting them).
  - Results:
    - **On the not inflation adjusted data:** Our best model produced mse 1.4930 (R^2 value 0.0726) trained without outliers, 2 knots, and degree 1.
    - **On the inflation adjusted data:** Our best model produced mse 25.5181 (R^2 value -0.0567) trained without outliers, 2 knots, and degree 1.

**Future:**

Modeling the targets we had with the parameters we had proved to be a difficult task.

According to the EBM, the change in tech index is more related to its own historical data than other macroeconomic data. However, considering our best model (MLP with attention) was trained with no lag (i.e. it considers no historical data), this conclusion by EBM might be strictly specific to our preferred EBM models.

The goal of our project was to see the impact of the price of resources on an industry, and stock index is the indicator of the state of the industry we chose. To keep studying from this perspective, we might seek to replace the indicator with something else.

From another perspective, if we seek to study the behavior of the stock market: with lessons learnt from recent historical events, we might consider including surveys indicating people's view on the economy into the parameters.

The "Shallow dive" into deep learning was informative, we would like to do a "deeper dive" if possible.