Industrial resources and its impact on stock market

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

Data Science project realized at Erdős Institute
Spring 2025

Motivation

It is a fact of *Macroeconomics* that the cost of production impacts the performance of an industry. We expect **stock data** of an industry to be a valid indicator of its performance.

Thus, we use macroeconomic data and price data on industrial resources as parameters to train supervised models on the target of stock data.

In this project, we will focus solely on the technological industry. We will care about the impact of contextual variables on the **tech stock index (i.e. tech 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.

Clarification: the goal of this project 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 stock index as a representation of the state of an industry, and attempt to model the change in that state based on change of macroeconomic parameters (e.g. we may use price change of copper today as one of the parameters for the target of change of stock index of today).

Data Collection

Raw data collected:

- 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.
  Note: Companies enter and exit S and P 500 index 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 as part of the parameters indicating the cost of production.
- Daily data on raw industrial resources involved in tech industry: *gold, silver, copper, crude oil, palladium, platinum* as parameters influencing production costs.
- Daily data on federal fund rates as parameters.
- Monthly data of CPI (Consumer Price Index) as indicators of general inflation.

Data Processing

- **Classifying the companies:** Produced, in accordance to NAICS, a list of companies considered as "tech industry", and retrieved (monthly) PPI data according to the NAICS chosen.
- **Producing tech stock index (i.e. tech index):** Created a list of "stable companies" (i.e. did not exit or enter the S and P 500 index during the time frame of 2014/Jan/1 to 2024/Oct/30, the fully span of our interest), and created daily data on tech stock index:

$$\text{Tech stock index} := \text{Total index} \times \Big(\sum_{\substack{\text{company} \in \\ \text{Stable tech companies}}} \text{weight contribution of the company} \Big)$$

- **Creating proportional change daily data:** Generated daily proportional change data relative to the nth prior (trading) day (recorded as percentage data by multiplying with 100) for each of the tech stock index and the raw resources with formula:

$$\begin{matrix}\text{proportional change on day } t \\ \text{relative to the } n^{th} \text{ prior day}\end{matrix} := \frac{\text{value on day } t}{\text{value on day } t-n} - 1$$

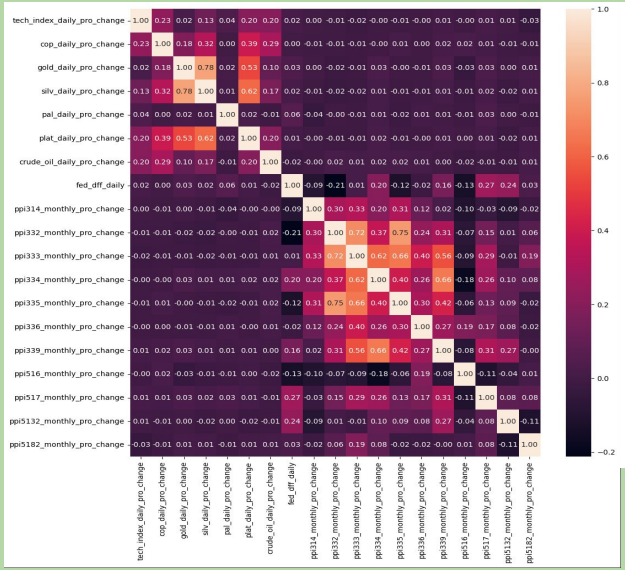
- **Create proportional change monthly data:**
  - Result: Calculated the proportional change data in the context of monthly values and "expanded" them to daily values in the following way (e.g.):

$$\text{"Proportional change" on Jan/06/2018} := \text{Proportional change on Jan/2018}$$

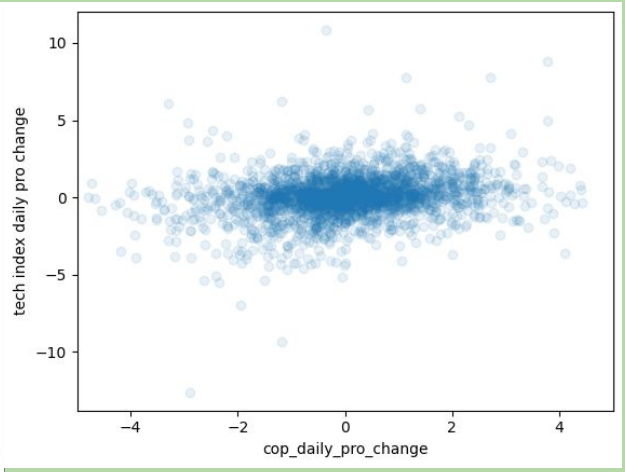
Data Format and EDA

We have created a dataset: (called **no inflation adjustment data set**)
For each row (one row for each trading day between 2014-Jan-01 and 2024-Oct-31), we have in columns:

- **Target:** proportional change of tech stock index relative to the prior 1st trading day.
- **Parameters:** proportional change of price of raw resources relative to the prior 1st trading day, proportional change of PPI's relative to the previous month, the federal fund rate.
- **Train and test split:** Test set starts on date 2024-Jan-01 (**Target has Variance: 1.6098**).



Correlation heatmap



The correlation between proportional change of tech index and copper price is small, but positive

- Low correlation between target and parameters.
- Positive (small) correlation between proportional change of tech index and parameters that indicate production pieces, this is counterintuitive (see left graph as example).
- **Conjecture:** the counterintuitive behavior can be explained by inflation.

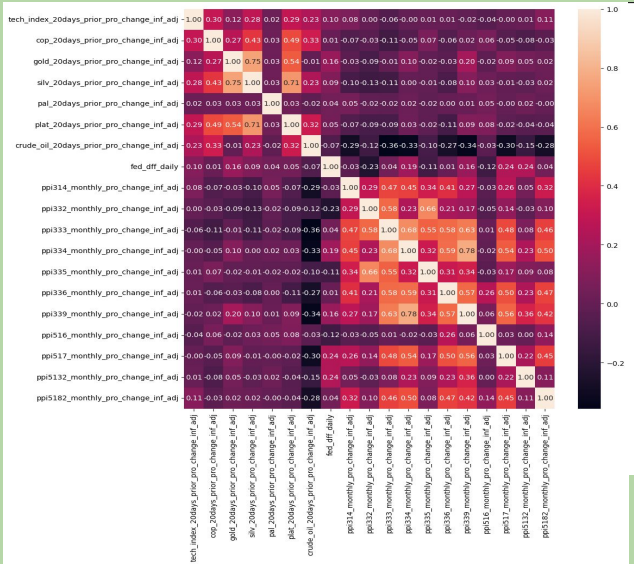
Data Format and EDA

We create **inflation adjusted dataset** with formula:

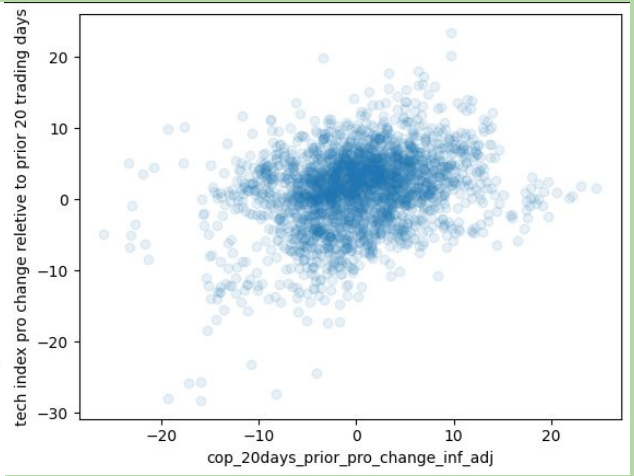
$$\text{Inflation adjusted proportional change on day } t := \frac{\text{value on day } t}{\text{value on day } t-20} \times \frac{\text{CPI previous month}}{\text{CPI current month}} - 1$$

For each row (one row for each trading day between 2014-Jan-01 and 2024-Oct-31), we have in columns:

- **Target:** inflation adjusted proportional change of tech stock index relative to the prior 20th trading day.
- **Parameters:** inflation adjusted proportional change of price of raw resources relative to the prior 20th trading day, 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 Variance: 24.1491**).



Correlation heatmap



The correlation between the proportional change of tech index and copper price increased slightly

- Some correlations increased slightly.
- The counterintuitive behavior actually intensified, disproving our conjecture.

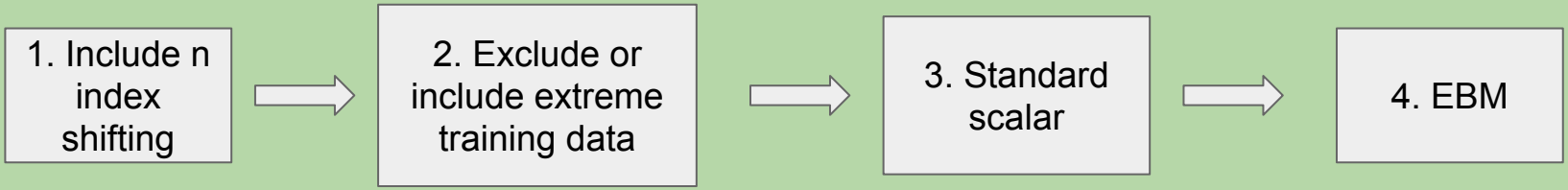
Training Models

- **EBM (Explainable Boosting Machine):**
  - EBM team: Yuan, Kayode, Wojciech.
- **NN (Neural Network)**
  - NN Team: Yuan, Kayode, Wojciech.
- **SPLINE**
  - SPLINE team: Xiangyi, Thiago, Rafael.

- ❏ We have produced models for both inflation adjusted and no inflation adjustment datasets with EBM and SPLINE, while we only produced models for inflation adjusted dataset with NN.
- ❏ We will only explicitly demonstrate the models for the inflation adjusted dataset in this presentation.
- ❏ All groups use mse (mean squared error) as metric.

Training Models - EBM - Pipeline

Pipeline:

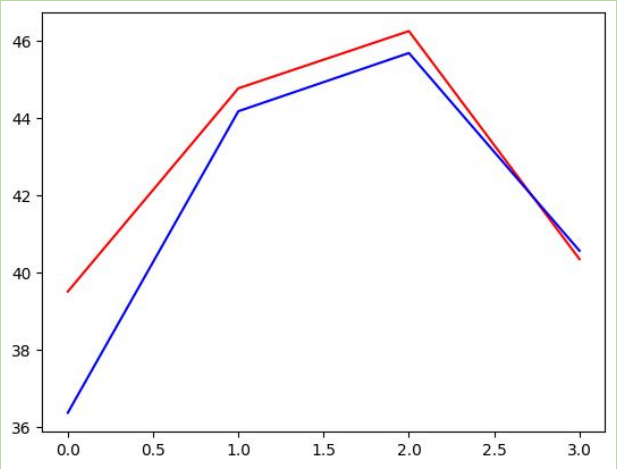


Explanation:

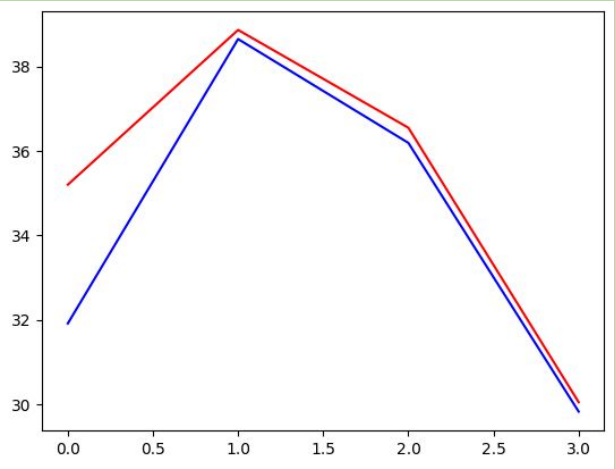
1. Include n index shifting: This step includes shifted down index of values and include them into parameters, for instance, by shifting the stock index proportional change by 1, we include the proportional change of previous trading day into the parameters.
2. Exclude or include extreme training data: A decision if we delete any training data that has too extreme values (i.e. outliers) on any of the parameters and the target.

Training model - EBM - Cross validation

(time series 5-fold) cross validation on inflation adjusted data: The following graphs show the change in mean squared error (mse) on y-axis, with x-axis being the number of shiftings of 20 index each (corresponding to 20 trading day, around 1 calendar month):



Tested on validation data with extreme values. x-axis is shifting number.



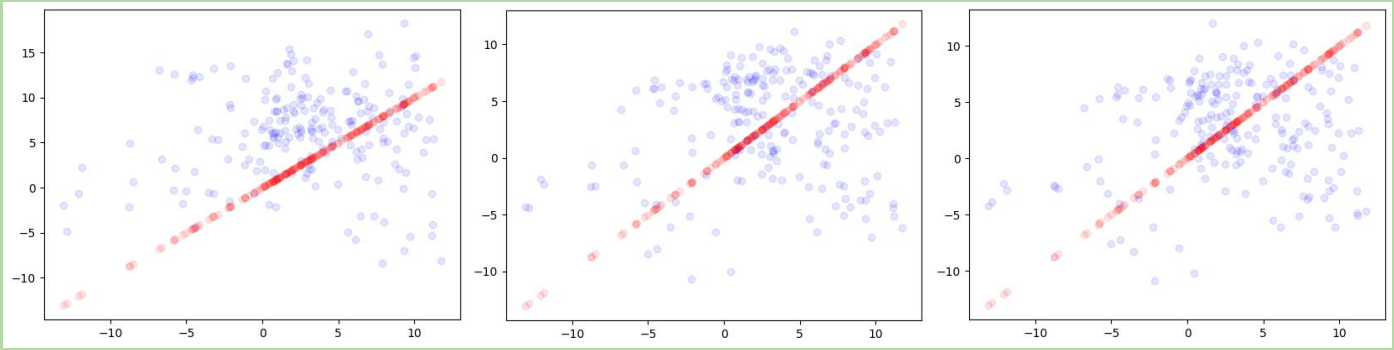
Tested on validation data without extreme values. x-axis is shifting number.

- Red line is model trained with training set including outliers
- Blue line is model trained with training set excluding outliers

It can be seen that in both cases the mse increases and then decreases, making it unclear if shiftings are preferred or not if we are testing on validation sets with or without extreme values. The difference between including and excluding extreme data when training also seems small when training with high shiftings. But one would definitely prefer training without extreme data when training with no shifting.

Training model - EBM - Result

Result on inflation adjusted data: In the following plots, red points are true values, and blue points are predicted values with x-axis representing true values:

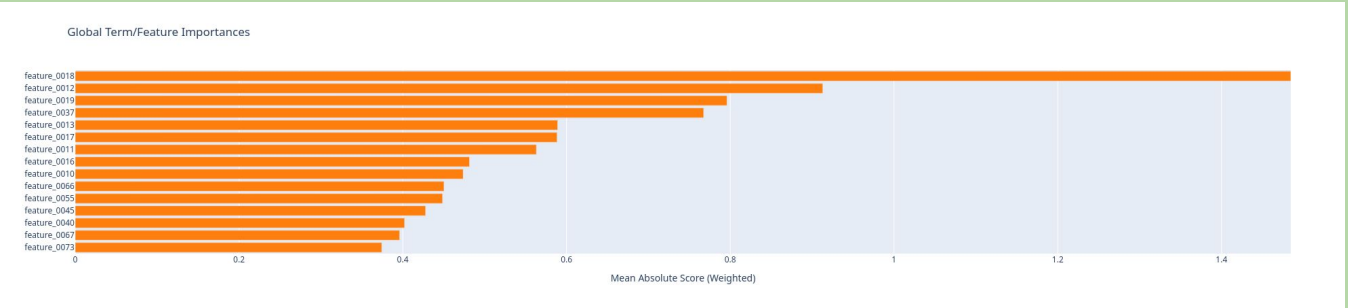


Trained without extreme data, no shifting.
mse: 54.8589
R^2: -1.2717

Trained without extreme data, 3 shiftings.
mse: 40.8505
R^2: -0.6916

Trained with extreme data, 3 shiftings.
mse: 38.4721
R^2: -0.5931

- **We prefer the model trained with 3 shiftings and extreme data included.**
  Mse: 38.4721
  R^2: -0.5931



- Tech index pro-change 20 (trading) days ago.
- Copper price pro-change
- Fed rate 20 (trading)days ago
- Tech index pro-change 40 days ago
- Gold price pro-change
- Crude oil price pro-change
- PPI-5132 (software publisher) pro-change
- PPI-517 (Telecommunications) pro-change 60 days ago
- Crude oil price pro-change 40 days ago
- PPI-339 (Miscellaneous Durable Goods Manufacturing) pro-change 40 days ago
- PPI-332 (Fabricated metal product manufacturing) pro-change 40 days ago
- PPI-5132 (Software publisher) pro-change 60 days ago
- Platinum price pro-change 60 days ago

Above is the “importance score” provided by EBM. See the features names on the right. The history of tech index plays an important role and there did not appear to be any interactions with the large weight contribution.

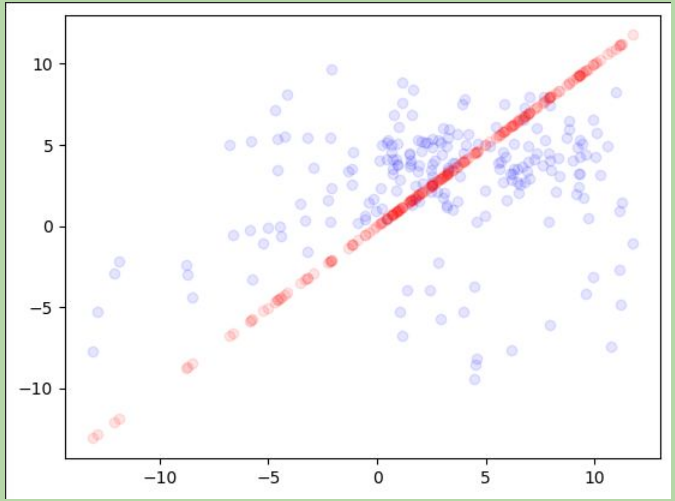
Training models - NN - Models

- **RNN (Recurrent Neural Network):**
  - **Advantage:** Take advantage of the time series nature.
  - **Cross-validation:**
    - Lag length: Determine how much history to include.
    - Epoch number: Prevent overfitting.
- **MLP (Multilayer Perceptron) with Attention:**
  - **Advantage:** Take advantage of the interactions between the parameters.
  - **Cross-validation:**
    - Lag: Determine historical data to include.
    - Epoch number: Prevent overfitting.

❏ We only produced neural network models for the inflation adjusted dataset.

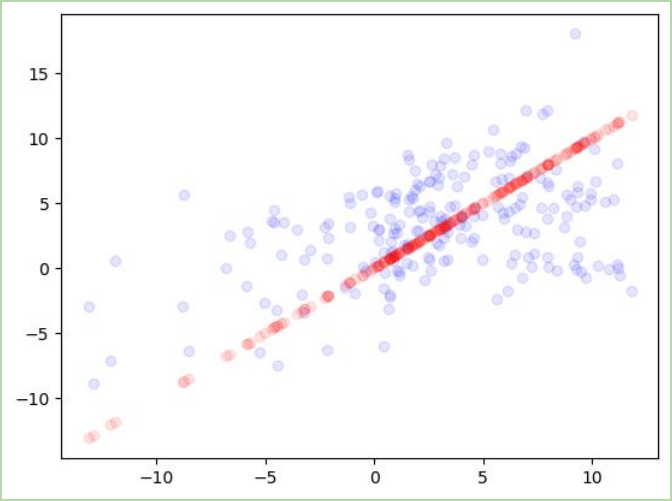
Training models - NN - Results

A recurrent neural network with 6 lag length and 190 epochs produced following result on the test set of inflation adjusted dataset:



Red points are actual values, blue points are predicted values. x-axis represents actual values.
mse: 31.4635
R^2: -0.3029

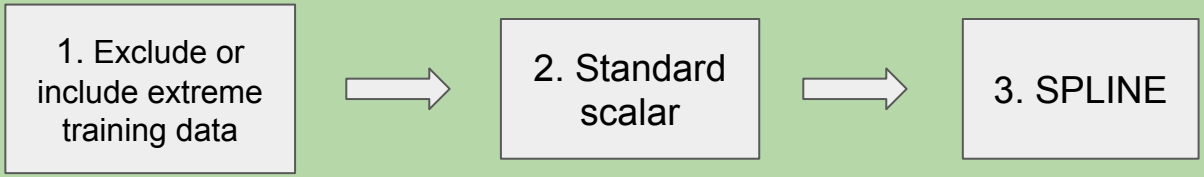
A multi-layer perceptron (MLP) with a self-attention layer with no lag and 300 epochs produced following result on the test set of inflation adjusted dataset.



Red points are actual values, blue points are predicted values. x-axis represents actual values.
mse: 23.06
R^2: 0.0451

Training Model - SPLINE - Pipeline

Pipeline:



Explanation:

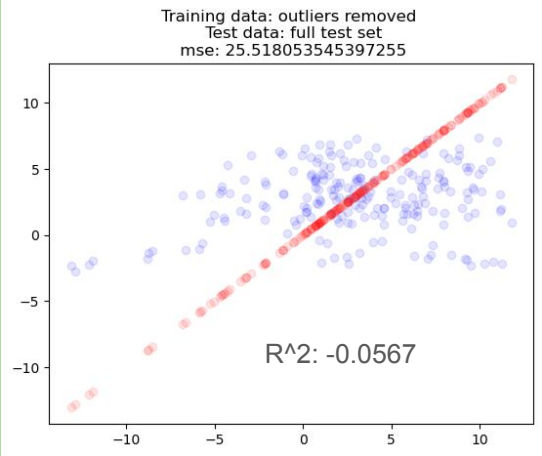
1. No shifting: The SPLINE model is trained to explain the same-day stock change from the same-day macro data.
2. Exclude or include extreme training data: A decision of if we delete any training data that has values too extreme on any of the parameters and the target.

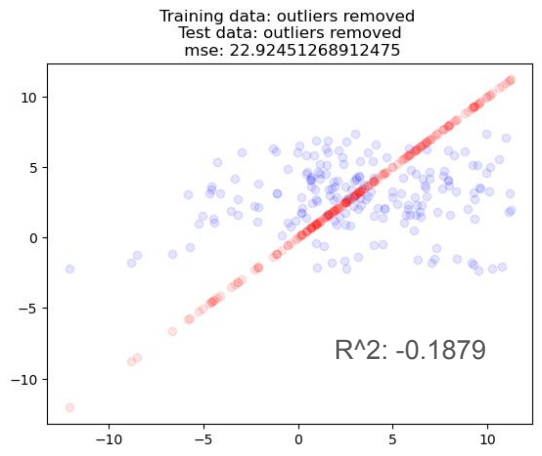
Training Models - SPLINE - Cross Validation

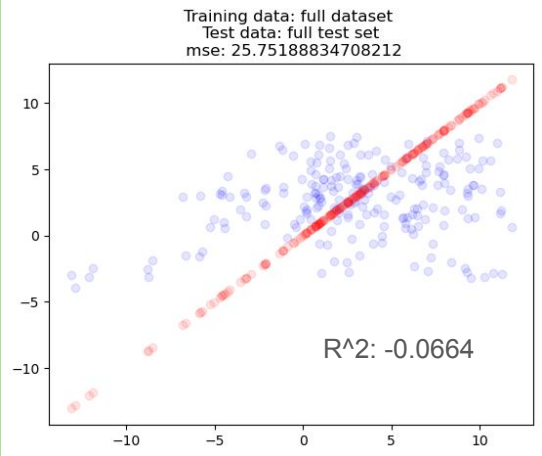
We used a five-fold time-series cross-validation to accommodate the temporal ordering of our data. We explored different values for the spline transformer's parameters (trying out several **number of knots** and **degree** settings by manually adjusting them). Through these manual checks, we found that *n_knots=2* and *degree=1* consistently yielded the lowest MSE in our cross-validation experiments. Because of that, we did not implement a systematic search (e.g., *GridSearchCV*), having already identified this configuration as reliably optimal for our purposes.

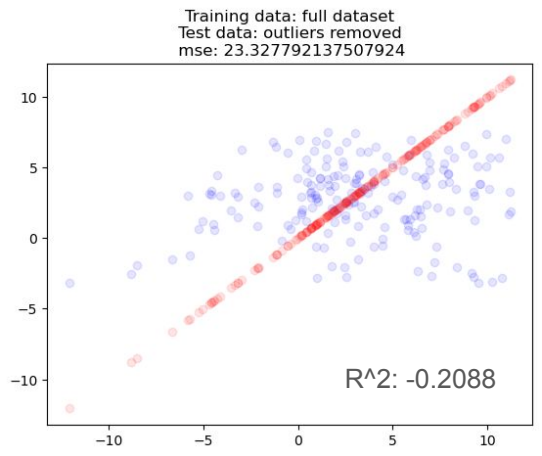
Training Models - SPLINE - Result

Result on inflation adjust data: We will provide the result of training with either including or excluding extreme training data (the outliers) and testing on test data with or without outliers. In the following, red points are true values, and blue points are predicted values with x-axis representing the true values.









Results and future work

- According to the EBM models, it seems (the collection of) historical tech stock index proportional changes seems to have more “weight” than the proportional changes of resource prices.
- Modeling the target of stock market change with the macroeconomic parameters we have is a hard task indeed.
- The goal of our project was to see the impact of price of production on an industry, and stock index is the indicator of the state of a given industry. As possible future work, we might try replacing this indicator with some other value(s).
- From another perspective, if we seek to study the behavior of stock market: with lessons learnt from recent historical events, we might consider including surveys indicating people’s view on the economy into the set of parameters.
- The “Shallow dive” into deep learning was informative, so we would like to do a “deeper dive” in the future.

Thank you!