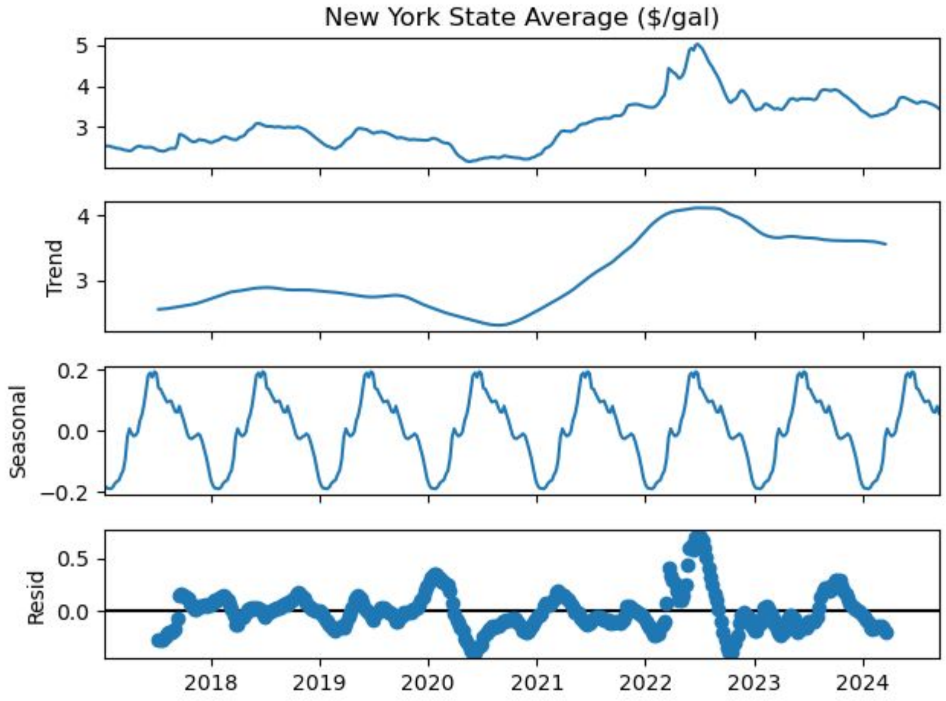
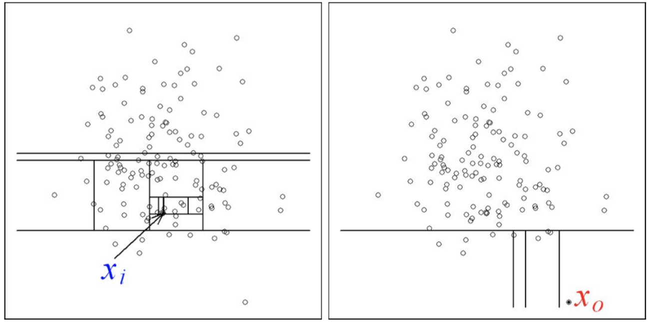
Data Preprocessing

- We remove trends and seasonality in the data
  - Trends are removed by applying a 12-week moving average (e.g., quarterly smoothing)
- Data is averaged to a weekly resolution
  - This keeps the most amount of information while ensuring equal sample spacing



Method 1: Isolation Forest

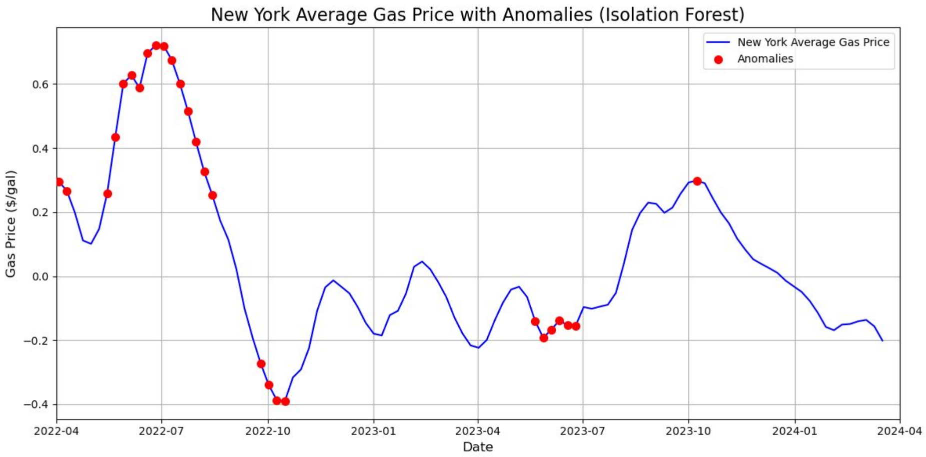


(a) Isolating x_i (b) Isolating x_o

Liu, et al. (2008)

- Isolation forests show promise in detecting anomalous patterns in time series data, Liu et al. (2008)
- This method aims to isolate anomalous points in data through recursively generating partitions of the data across all it’s features
  - Iteration stops once a maximum tree depth specified by the algorithm is reached
- On test time, anomaly scores are determined based on the normalized path lengths
  - Points requiring less path lengths to split are more anomalous, as they would be in sparser regions of the data

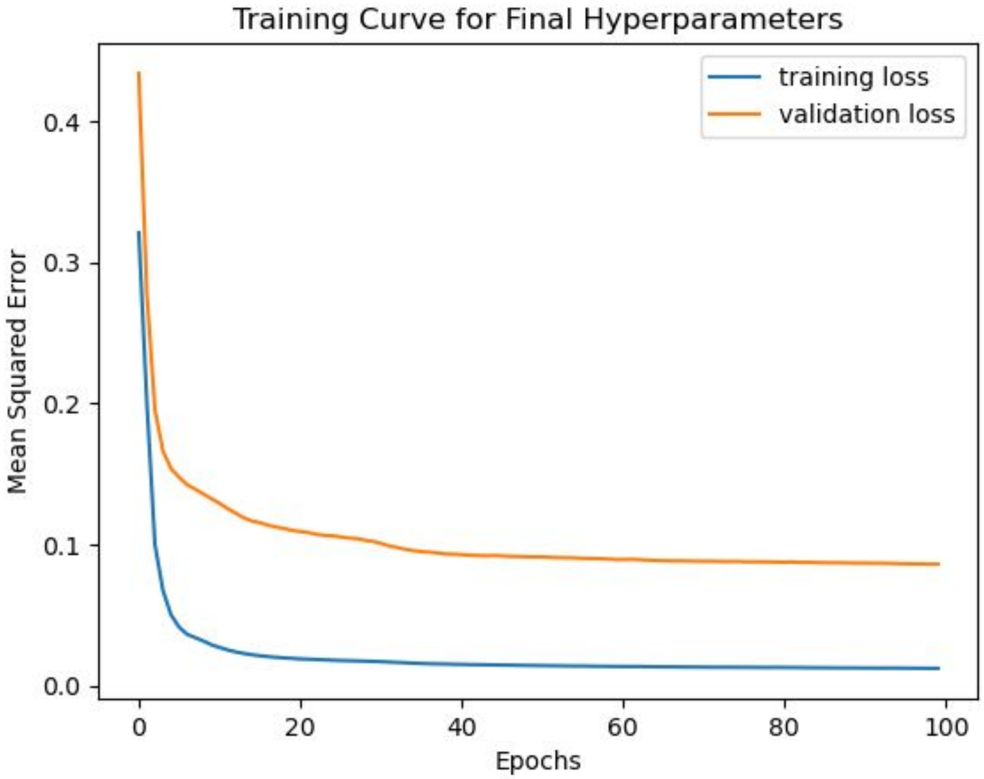
Method 1: Isolation Forest



- We filter for the 95th percentile in anomaly scores as our thresholding value
  - Model is fit to 2015-2021 data and tested on 2022-2024 data
- The fit model identifies large changes in gas price residuals as anomalous, which is behavior we would expect
- Highest concentration of anomalies in 2022, which is consistent with our 1-d trend decomposition earlier

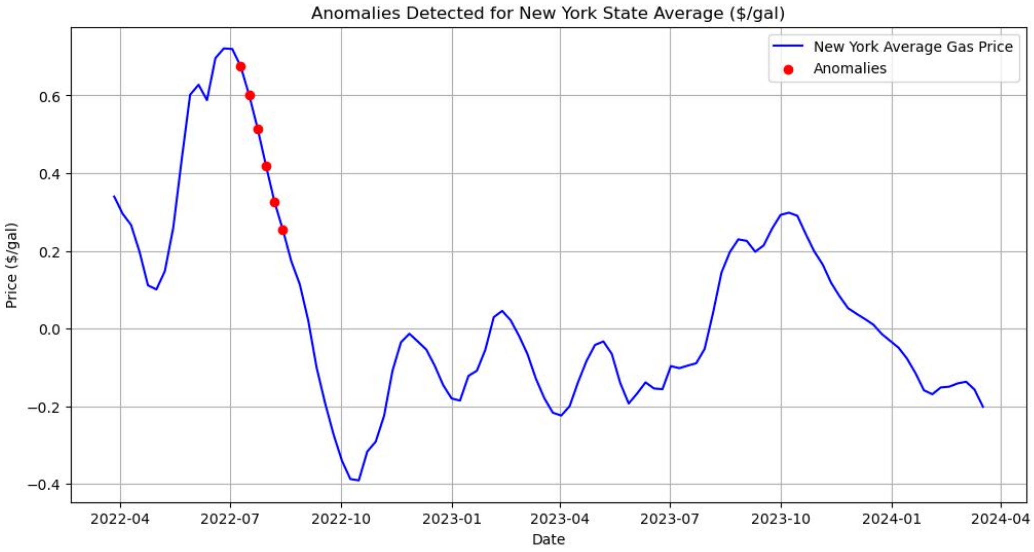
Method 2: LSTM Autoencoder

- We also use an LSTM autoencoder, trained on reconstructing the data before 2022, and tested on data from then until 2024
    - Due to size of training data, we perform 5-fold cross-validation with the data from 2015-2022
    - Hyperparameters tuned: hidden dimension and learning rate
    - Final hyperparameters used:
        - Hidden dimension: 64, learning rate: 0.001, batch size: 32, epochs: 100



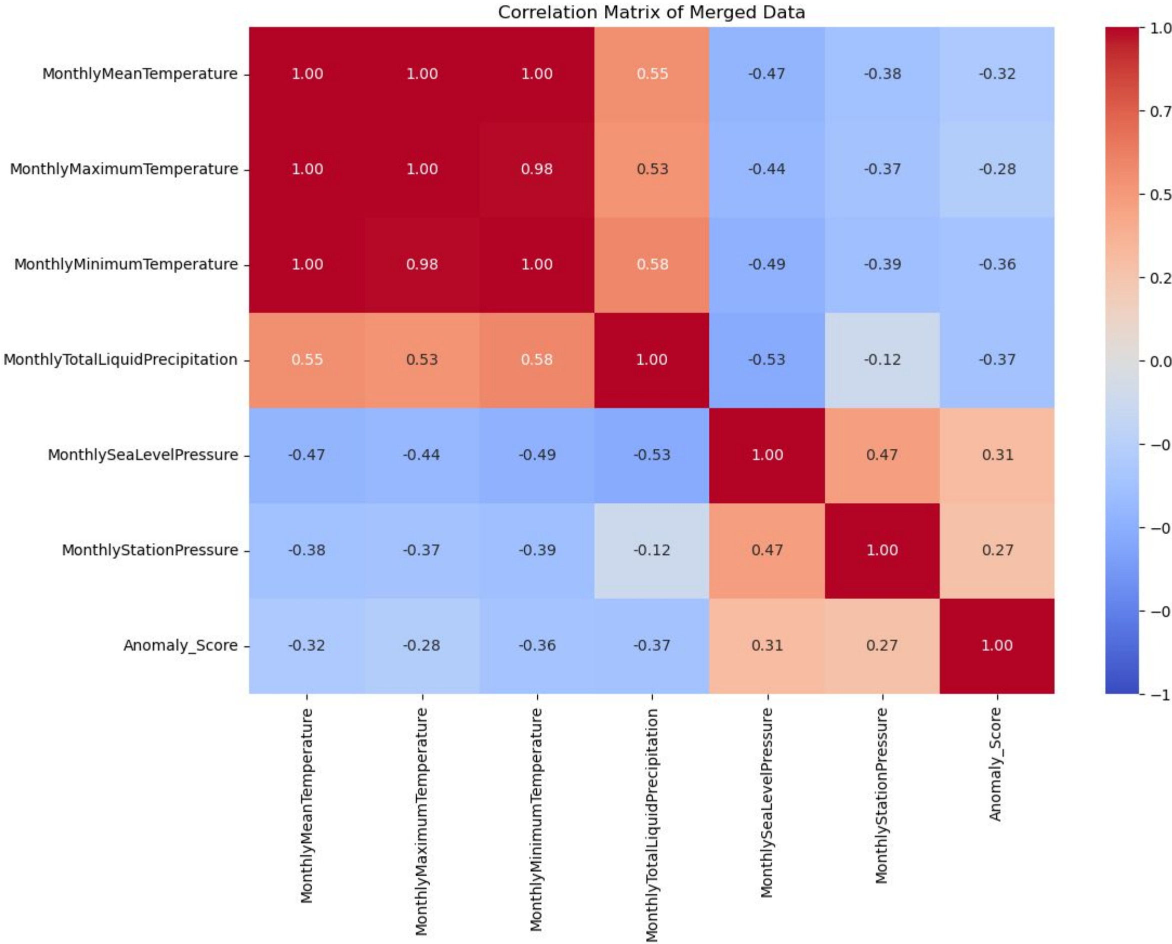
Method 2: LSTM Autoencoder

- Anomaly scores correspond to the MSE reconstruction error, tresholded according to the 95th percentile
- Compared to the isolation forest, this model is much more conservative in it’s anomaly scores
- Still flags the same downward trend in 2022 as anomalous

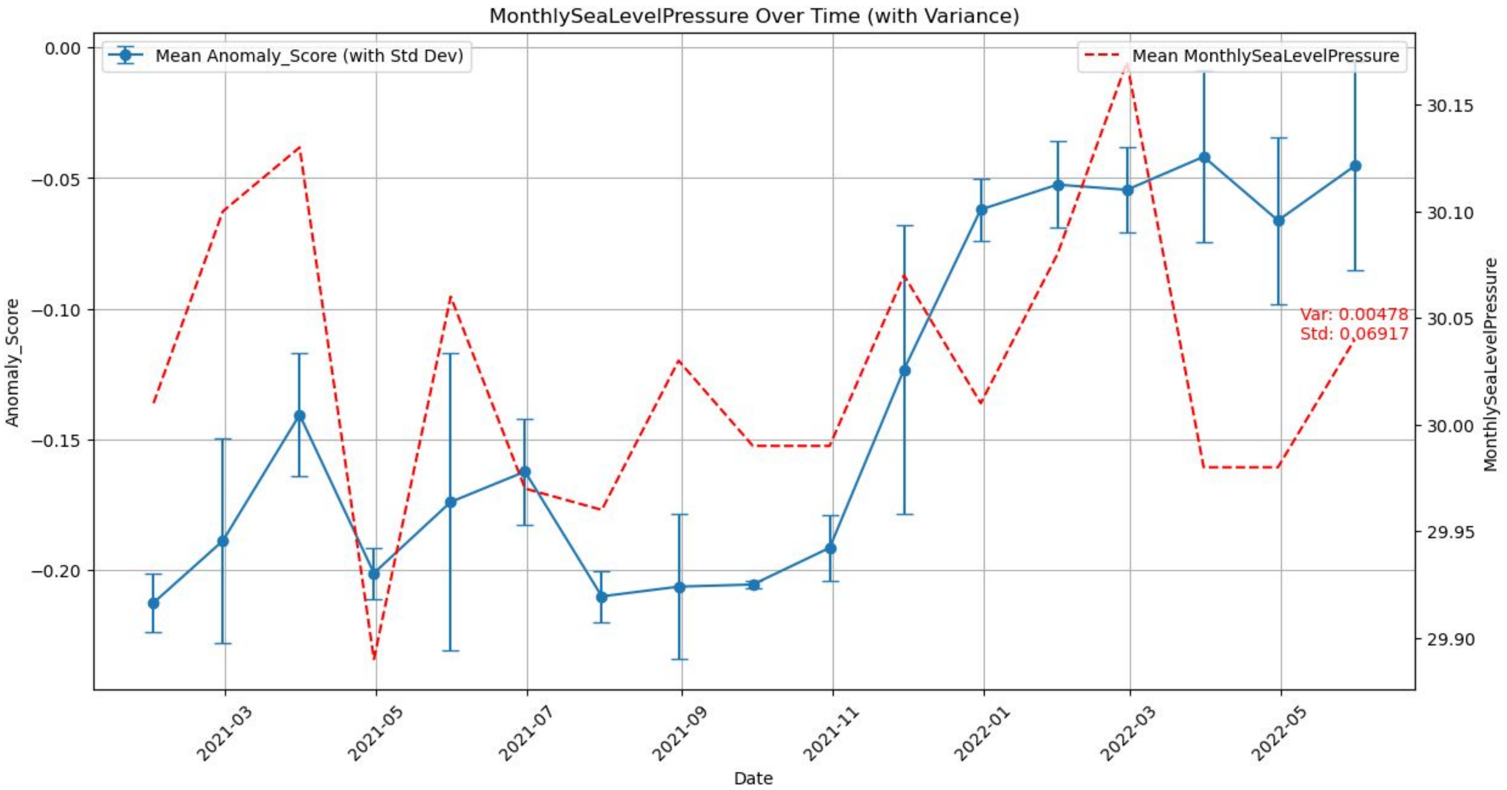


Incorporating Climate Effects

- We notice that both models indicate anomalous patterns in gas prices in 2022. One hypothesis is that this is correlated to climate patterns specific to 2022
    - To investigate further, we look into NOAA's Local Climatology Dataset (LCD), which provides station temperature and pressure at a monthly resolution
    - Specifically, we use data from NWS Albany
- The correlation matrix to the right shows possible connection between sea level pressure and anomaly score



Incorporating Climate Effects



- Sea level pressure variance is too high to make a conclusion
- Would need more data across multiple NY state stations

References

- F. T. Liu, K. M. Ting and Z. -H. Zhou, "Isolation Forest," *2008 Eighth IEEE International Conference on Data Mining*, Pisa, Italy, 2008, pp. 413-422, doi: 10.1109/ICDM.2008.17
- National Centers for Environmental Information. *Local Climatological Data (LCD) Documentation*. National Oceanic and Atmospheric Administration (NOAA), n.d., https://www.ncei.noaa.gov/pub/data/cdo/documentation/LCD_documentation.pdf.