Using Economic Indicators and Sentiment Analysis of Economic Policies to Build a Predictive Model of S&P 500 Stock Price

Jonathan Watkins, Jimmy Zhang, Jake Jarosik

DSCI 631 Fall Quarter 2023

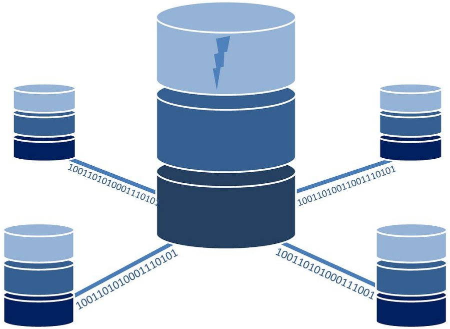
Introduction and Background

- Federal Open Market Committee (FOMC)
    - Holds 8 regular meetings a year
    - Sets federal funds interest rates and other monetary policies
    - Monitors the US economy to ensure it is working
    - They look at metrics / economic indicators like CPI and the unemployment rate
    - Previous chairman of the FOMC have shown great variance in economic philosophy
- Standard and Poor’s 500 Index Fund (S&P 500)
    - A stock index fund that is a weighted representation of the 500 largest companies traded on the US stock exchange.
    - Largest sector is currently software and technology
    - Includes: {'Communication Services', 'Consumer Discretionary', 'Consumer Staples', 'Energy', 'Financials', 'Health Care', 'Industrials', 'Information Technology', 'Materials', 'Real Estate', 'Utilities'}

Project Scope / Project Significance

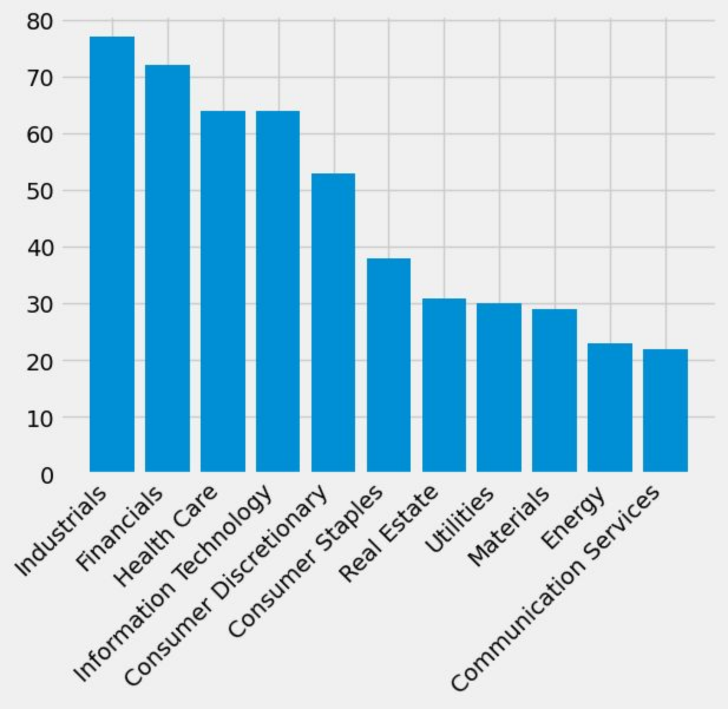
- Build a Model to Forecast the Price of the S&P 500 Index Fund
  - Exploratory Data Analysis (EDA), specifically related to S&P 500
    - Visualization of current economic climate
  - Sentiment Analysis of FOMC textual data
    - Informs predictive model of the direction of the Federal Reserve's policies
  - Time series modeling for numeric S&P 500 stock data
    - Quantify the linearity of the model as well as predict future stock prices

Data Acquisition

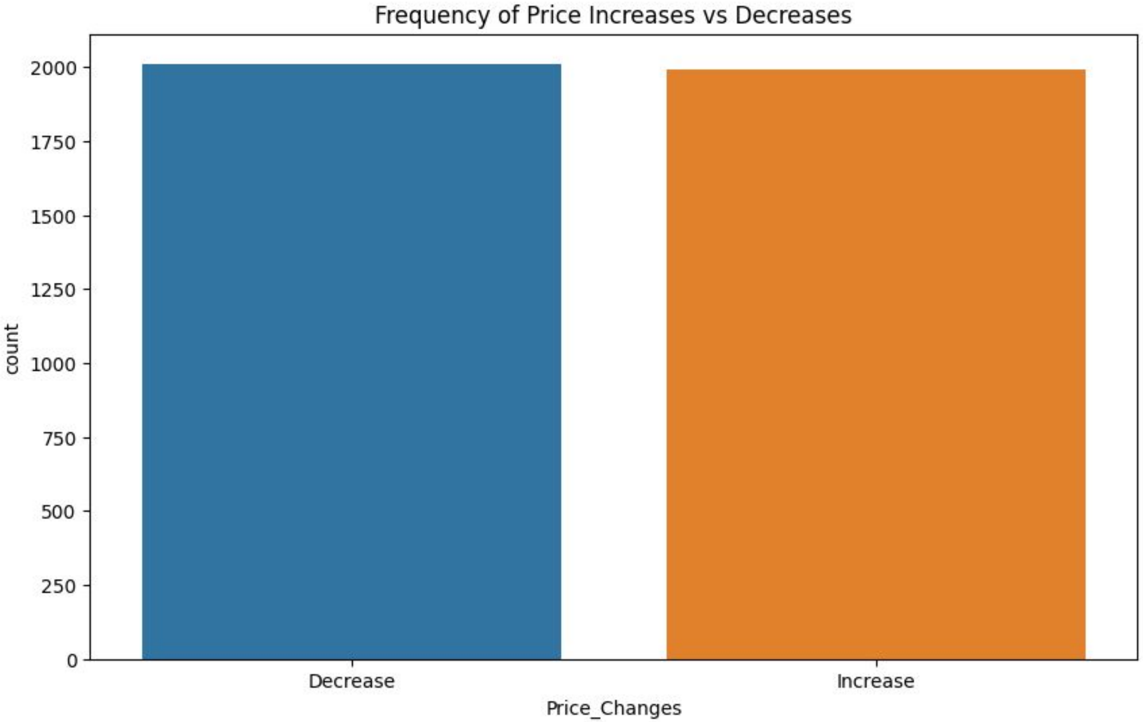


- Timeframe: From 2008 to Present
- Stock Price Data
  - Yahoo Finance API
  - Daily stock prices (Open, Close, High, Low, Volume)
  - Limited Fundamental Data (Market Cap, Moving Averages, 50 Week High and Low)
- FOMC Textual Data
  - Acquired through web scraping
  - Speeches, statements and minutes
- Federal Interest Data
- BLS Economic Data

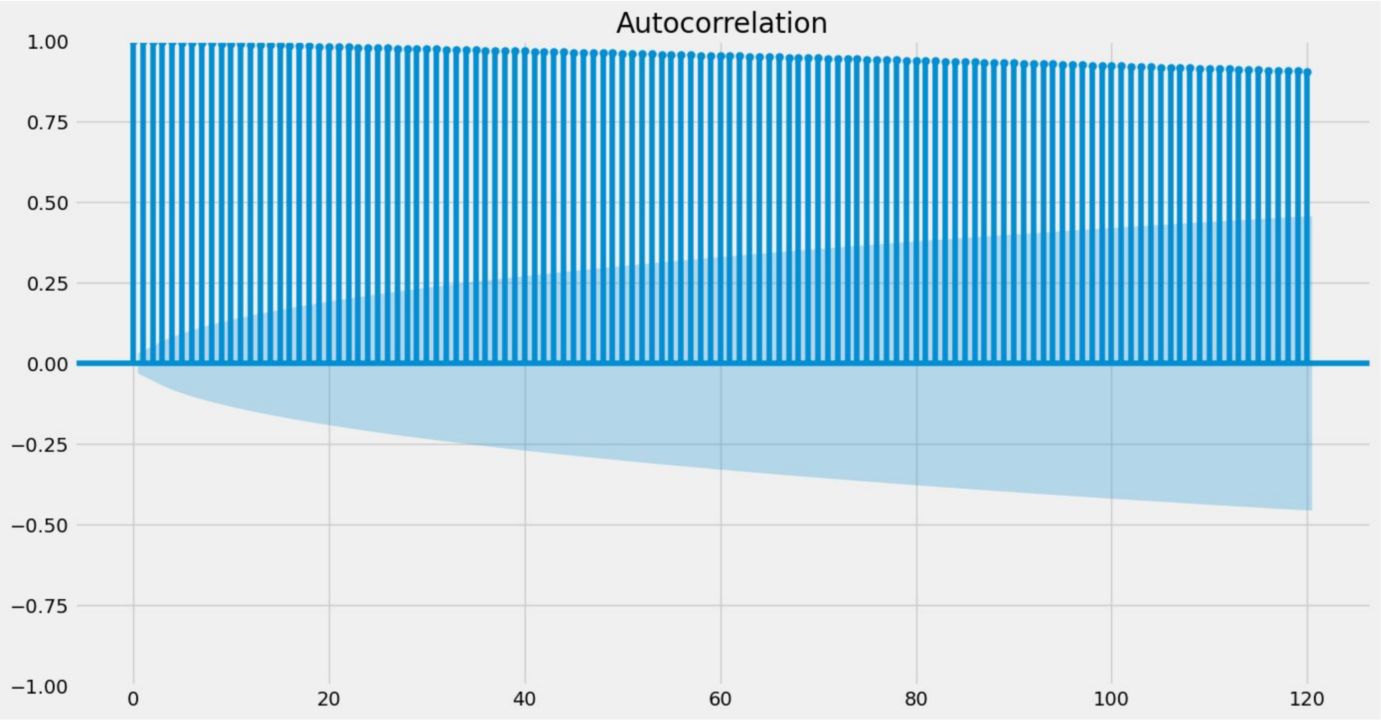
EDA: Composition of S&P 500



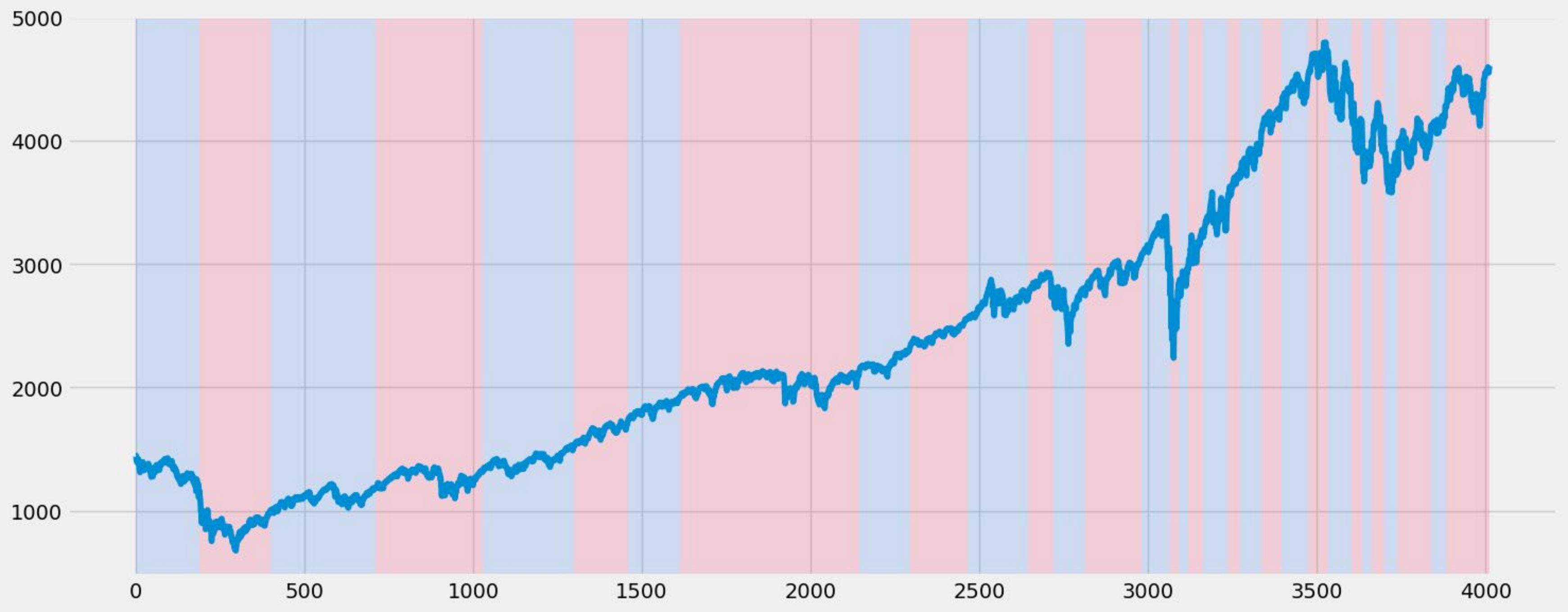
EDA: Stock Price Daily Changes



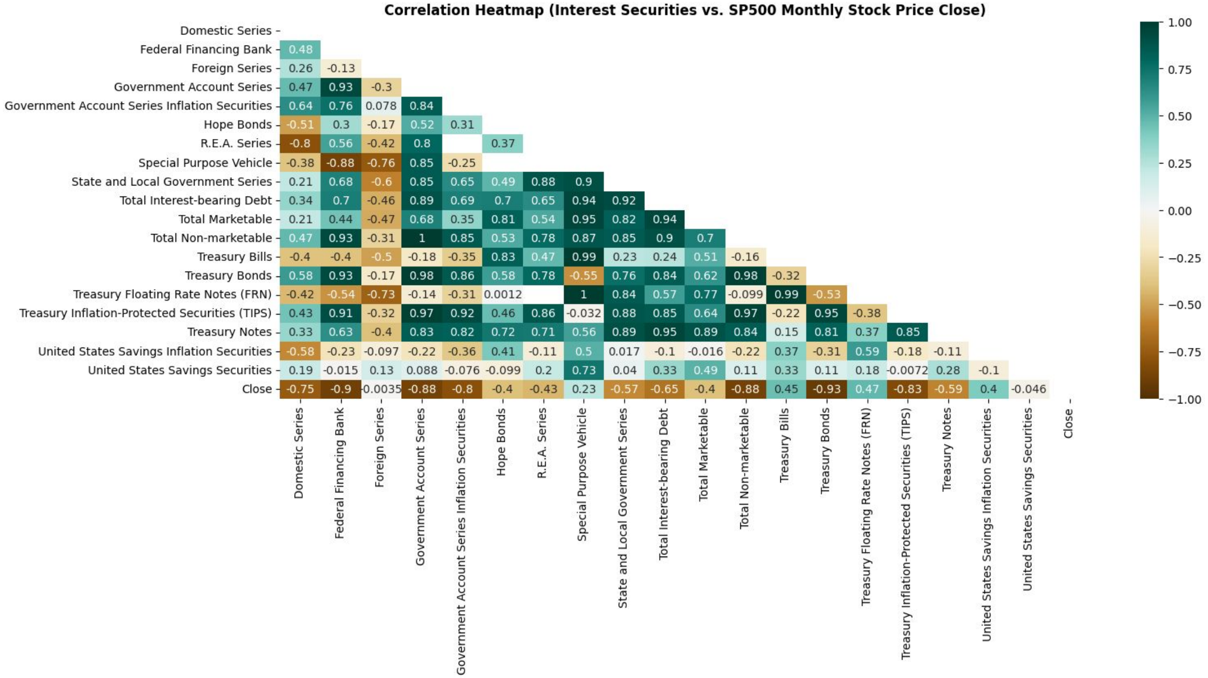
EDA: Stock Price Autocorrelation



EDA: Stock Price Change Point Detection



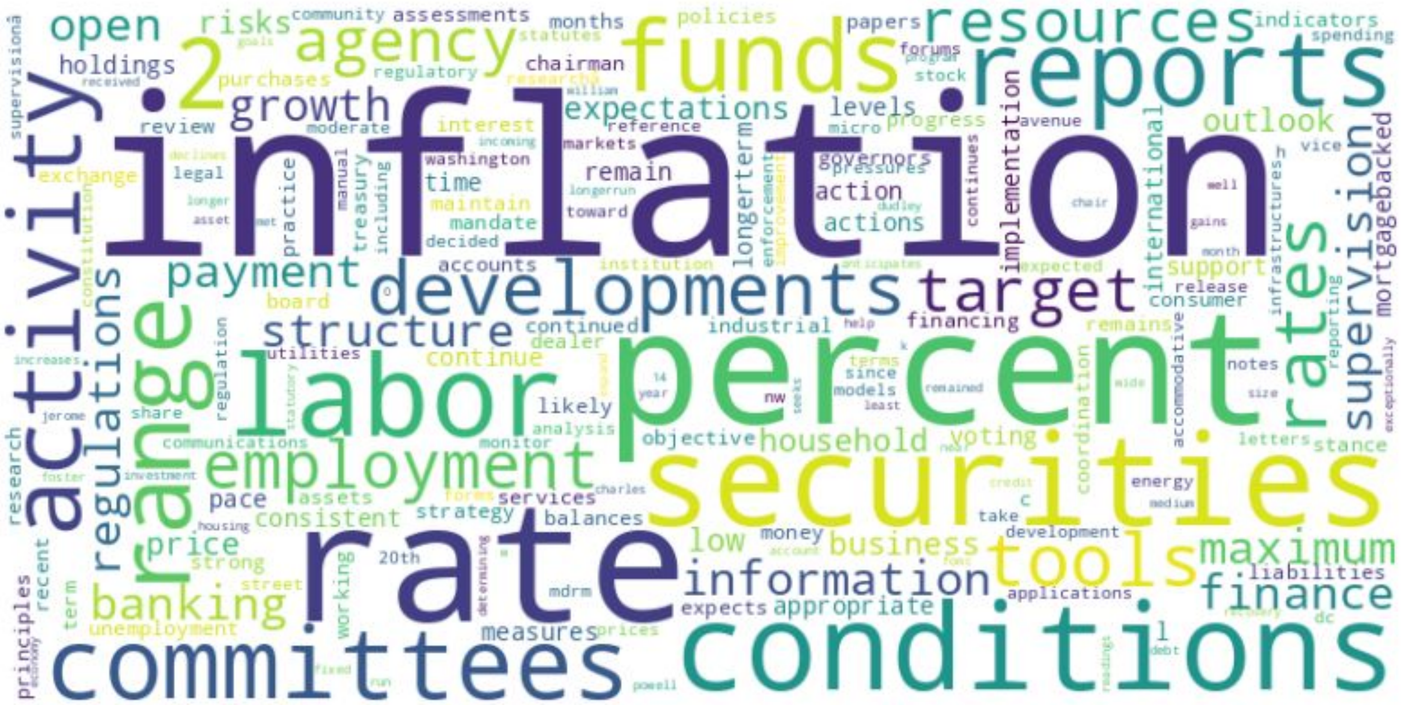
EDA: Interest Rate Correlation Matrix



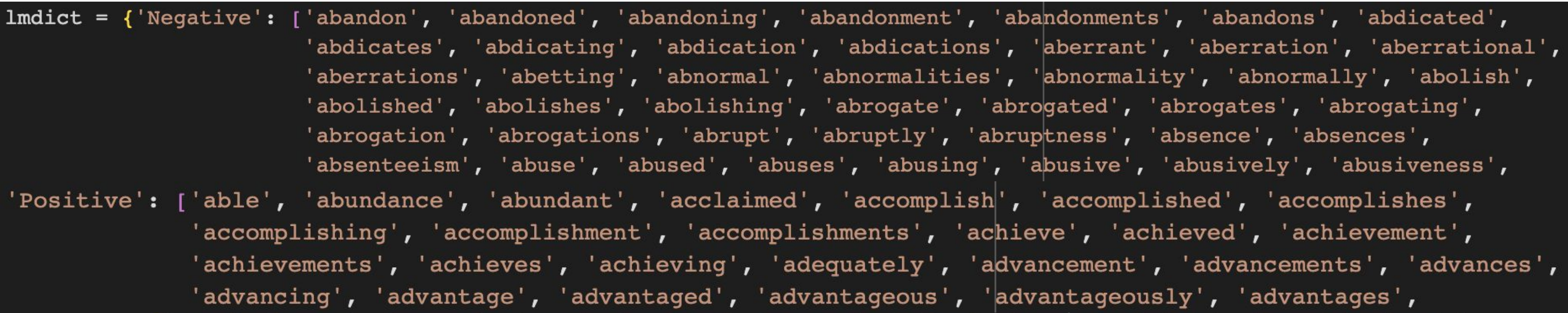
Correlation Heatmap (Interest Securities vs. SP500 Monthly Stock Price Close)

	Domestic Series	Federal Financing Bank	Foreign Series	Government Account Series	Government Account Series Inflation Securities	Hope Bonds	R.E.A. Series	Special Purpose Vehicle	State and Local Government Series	Total Interest-bearing Debt	Total Marketable	Total Non-marketable	Treasury Bills	Treasury Bonds	Treasury Floating Rate Notes (FRN)	Treasury Inflation-Protected Securities (TIPS)	Treasury Notes	United States Savings Inflation Securities	United States Savings Securities	Close
Domestic Series																				
Federal Financing Bank	0.48																			
Foreign Series	0.26	-0.13																		
Government Account Series	0.47	0.93	-0.3																	
Government Account Series Inflation Securities	0.64	0.76	0.078	0.84																
Hope Bonds	-0.51	0.3	-0.17	0.52	0.31															
R.E.A. Series	-0.8	0.56	-0.42	0.8		0.37														
Special Purpose Vehicle	-0.38	-0.88	-0.76	0.85	-0.25															
State and Local Government Series	0.21	0.68	-0.6	0.85	0.65	0.49	0.88	0.9												
Total Interest-bearing Debt	0.34	0.7	-0.46	0.89	0.69	0.7	0.65	0.94	0.92											
Total Marketable	0.21	0.44	-0.47	0.68	0.35	0.81	0.54	0.95	0.82	0.94										
Total Non-marketable	0.47	0.93	-0.31	1	0.85	0.53	0.78	0.87	0.85	0.9	0.7									
Treasury Bills	-0.4	-0.4	-0.5	-0.18	-0.35	0.83	0.47	0.99	0.23	0.24	0.51	-0.16								
Treasury Bonds	0.58	0.93	-0.17	0.98	0.86	0.58	0.78	-0.55	0.76	0.84	0.62	0.98	-0.32							
Treasury Floating Rate Notes (FRN)	-0.42	-0.54	-0.73	-0.14	-0.31	0.0012		1	0.84	0.57	0.77	-0.099	0.99	-0.53						
Treasury Inflation-Protected Securities (TIPS)	0.43	0.91	-0.32	0.97	0.92	0.46	0.86	-0.032	0.88	0.85	0.64	0.97	-0.22	0.95	-0.38					
Treasury Notes	0.33	0.63	-0.4	0.83	0.82	0.72	0.71	0.56	0.89	0.95	0.89	0.84	0.15	0.81	0.37	0.85				
United States Savings Inflation Securities	-0.58	-0.23	-0.097	-0.22	-0.36	0.41	-0.11	0.5	0.017	-0.1	-0.016	-0.22	0.37	-0.31	0.59	-0.18	-0.11			
United States Savings Securities	0.19	-0.015	0.13	0.088	-0.076	-0.099	0.2	0.73	0.04	0.33	0.49	0.11	0.33	0.11	0.18	-0.0072	0.28	-0.1		
Close	-0.75	-0.9	-0.0035	-0.88	-0.8	-0.4	-0.43	0.23	-0.57	-0.65	-0.4	-0.88	0.45	-0.93	0.47	-0.83	-0.59	0.4	-0.046	

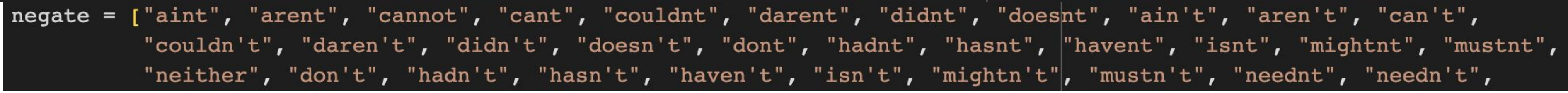
EDA: FOMC Statement Word Cloud



Sentiment Analysis: Loughran and McDonald Word List and Usage



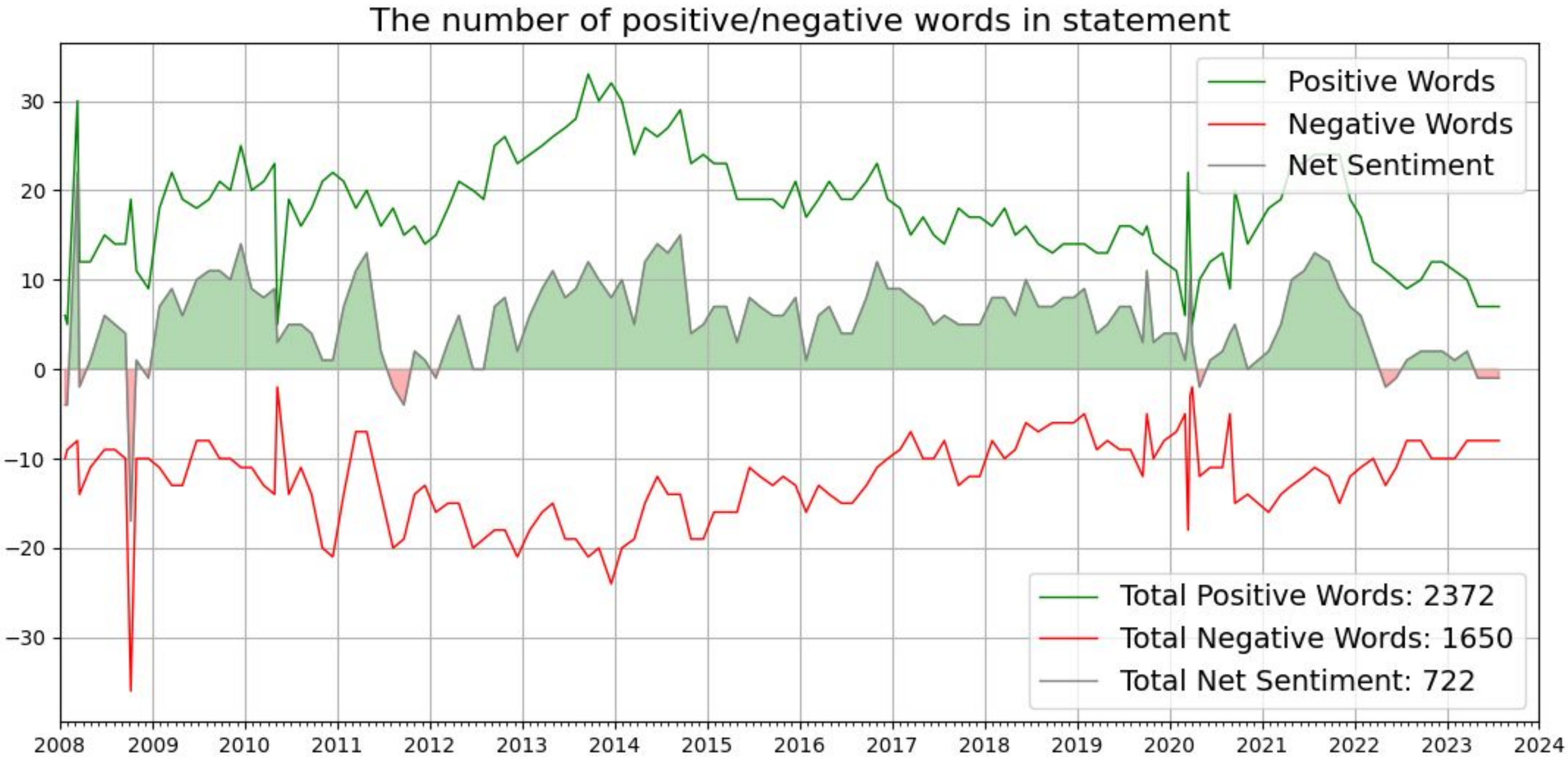
```
lmdict = {'Negative': ['abandon', 'abandoned', 'abandoning', 'abandonment', 'abandonments', 'abandons', 'abdicated',
                       'abdicates', 'abdicating', 'abdication', 'abdications', 'aberrant', 'aberration', 'aberrational',
                       'aberrations', 'abetting', 'abnormal', 'abnormalities', 'abnormality', 'abnormally', 'abolish',
                       'abolished', 'abolishes', 'abolishing', 'abrogate', 'abrogated', 'abrogates', 'abrogating',
                       'abrogation', 'abrogations', 'abrupt', 'abruptly', 'abruptness', 'absence', 'absences',
                       'absenteeism', 'abuse', 'abused', 'abuses', 'abusing', 'abusive', 'abusively', 'abusiveness',
'Positive': ['able', 'abundance', 'abundant', 'acclaimed', 'accomplish', 'accomplished', 'accomplishes',
             'accomplishing', 'accomplishment', 'accomplishments', 'achieve', 'achieved', 'achievement',
             'achievements', 'achieves', 'achieving', 'adequately', 'advancement', 'advancements', 'advances',
             'advancing', 'advantage', 'advantaged', 'advantageous', 'advantageously', 'advantages',
```



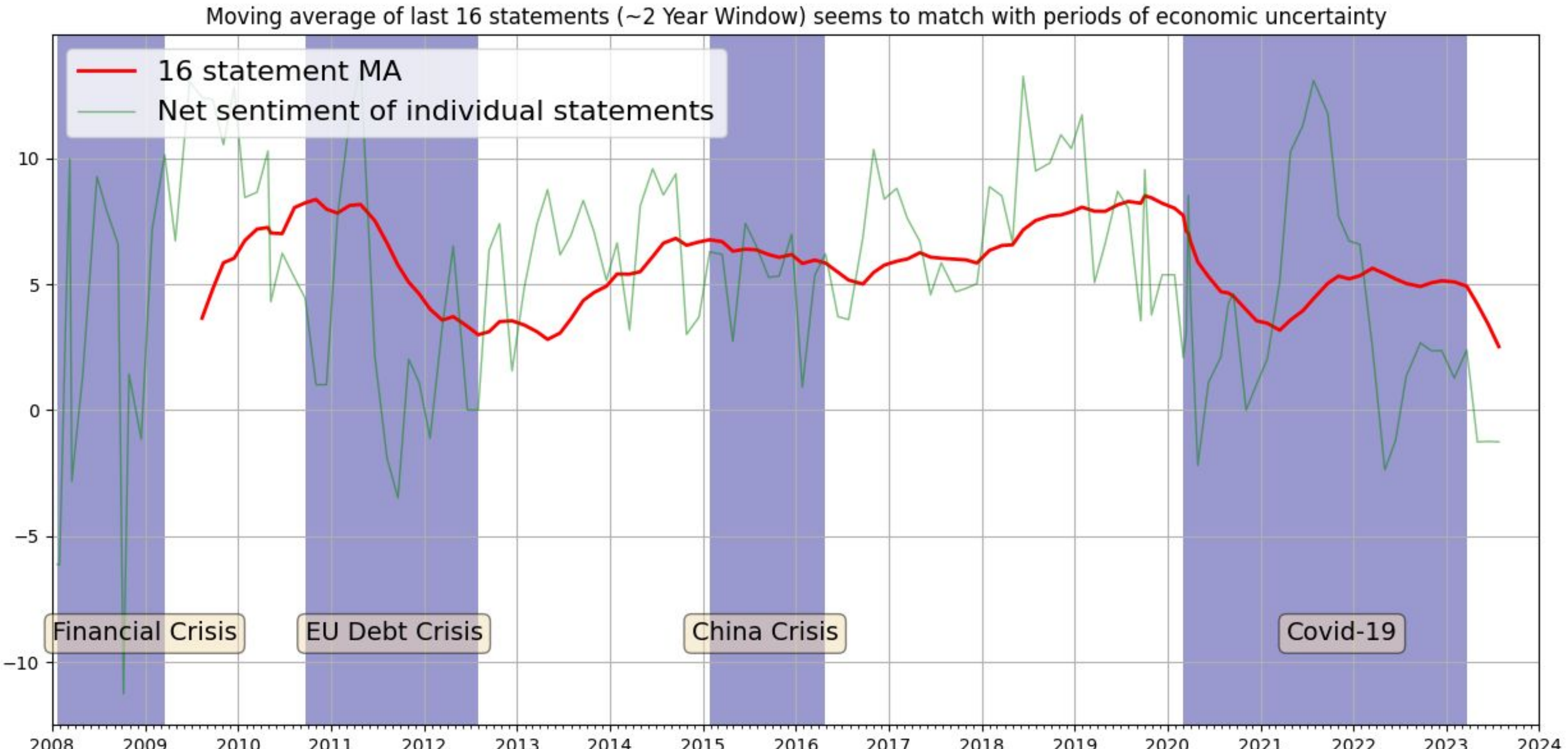
```
negate = ["aint", "arent", "cannot", "cant", "couldnt", "darent", "didnt", "doesnt", "ain't", "aren't", "can't",
          "couldn't", "daren't", "didn't", "doesn't", "dont", "hadnt", "hasnt", "havent", "isnt", "mightnt", "mustnt",
          "neither", "don't", "hadn't", "hasn't", "haven't", "isn't", "mightn't", "mustn't", "neednt", "needn't",
```

- Objective: Account for simple negation only for positive words
  - Counting Positive and Negative Words with Negation Check
  - Simple Negation:
    - Occurs when negate words appear within three words before a positive word

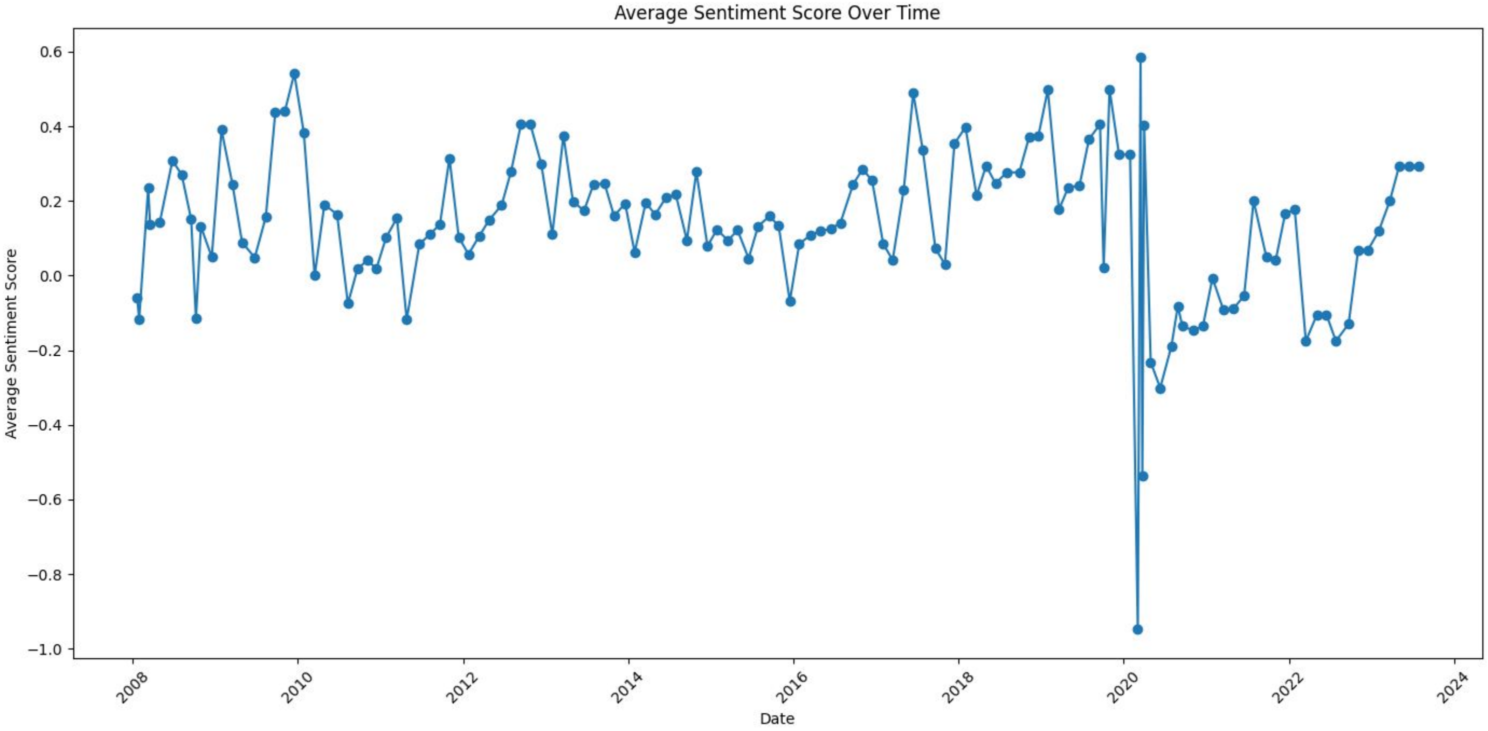
Sentiment Analysis Plots: Tracking Positive and Negative Word Counts



Moving Averages



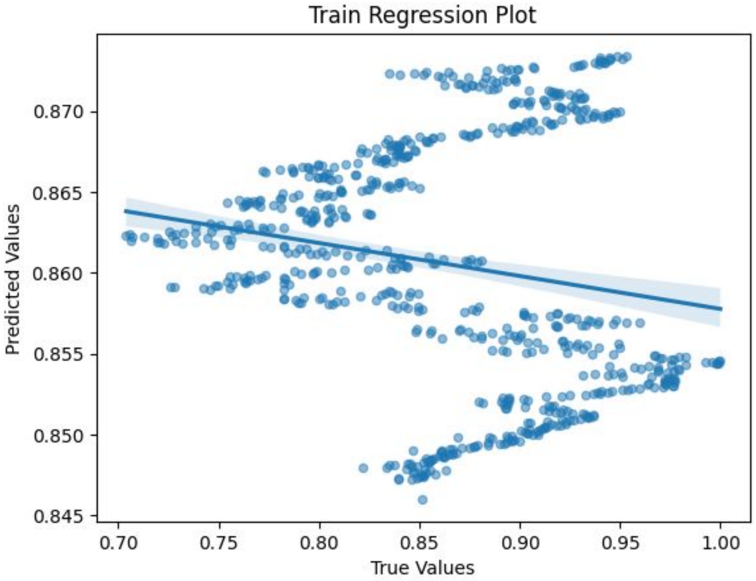
Improving upon Sentiment Analysis with FiGAS approach and with Senticnet

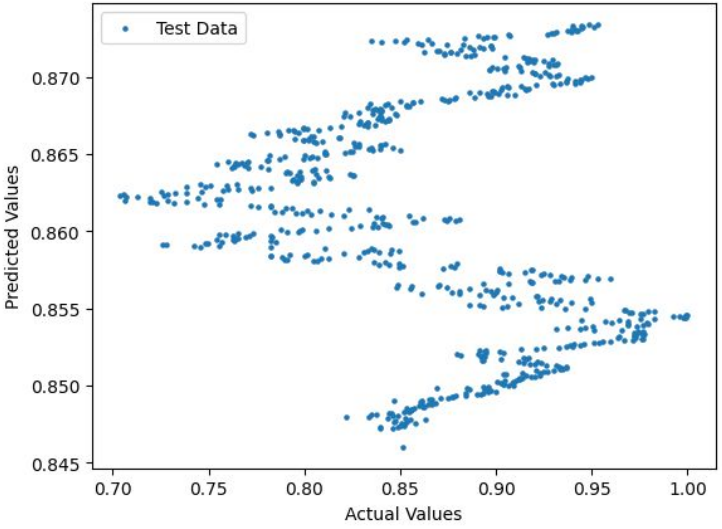


Introduction to ARIMA

- Autoregressive Integrated Moving Average
- The most simplest of modeling approach of the three.
  - Does not take additional regressors. Only fits historical values.
  - Does not account for Seasonality
  - Designed for Regular Time Spacing / Intervals
  - Assumes Stationarity

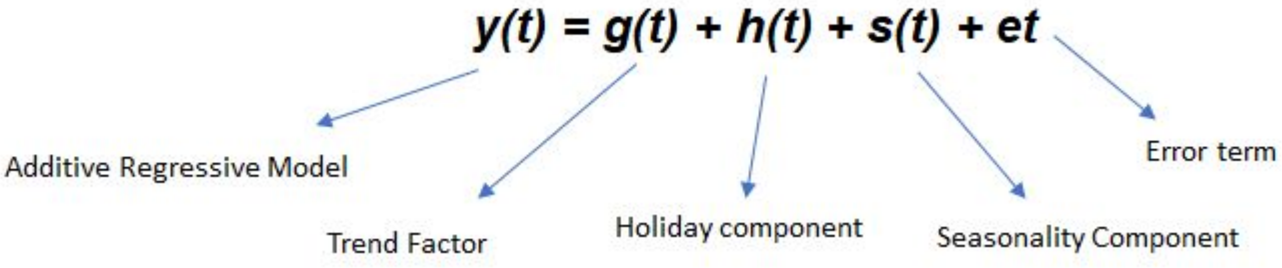
ARIMA Results





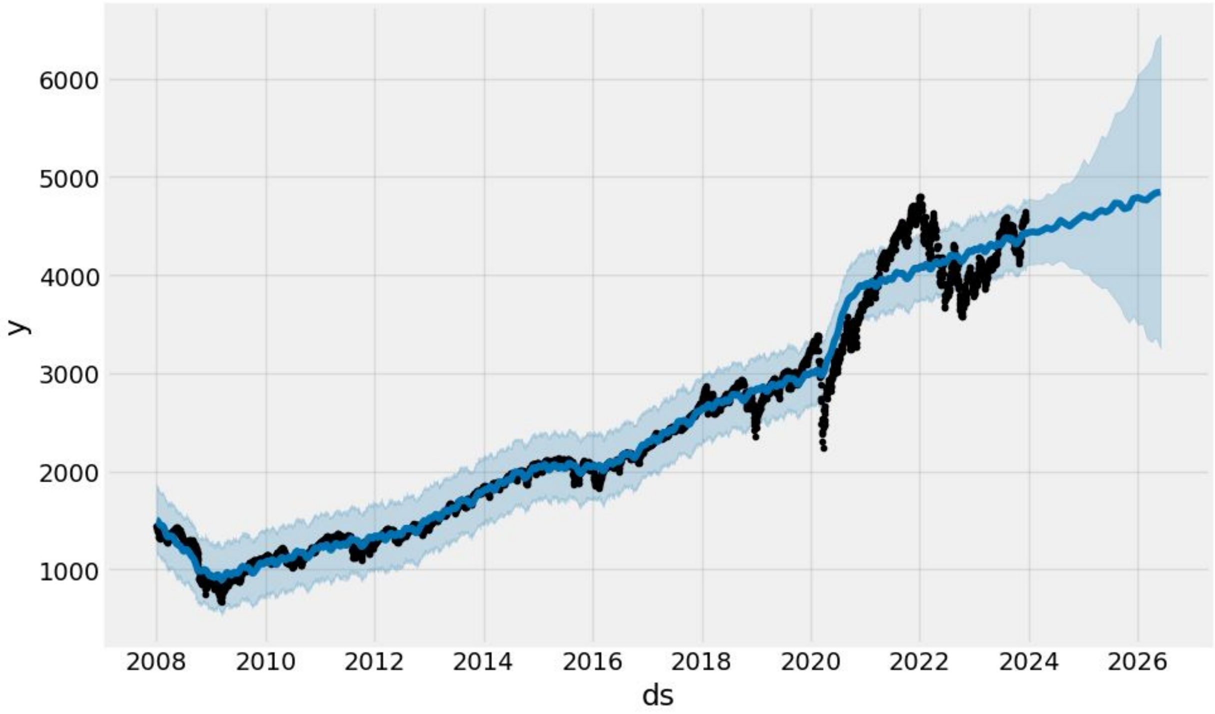
Introduction to Prophet

- Open Source Package By Meta / Facebook
- Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects.
- Simplifies Time Series Forecasting

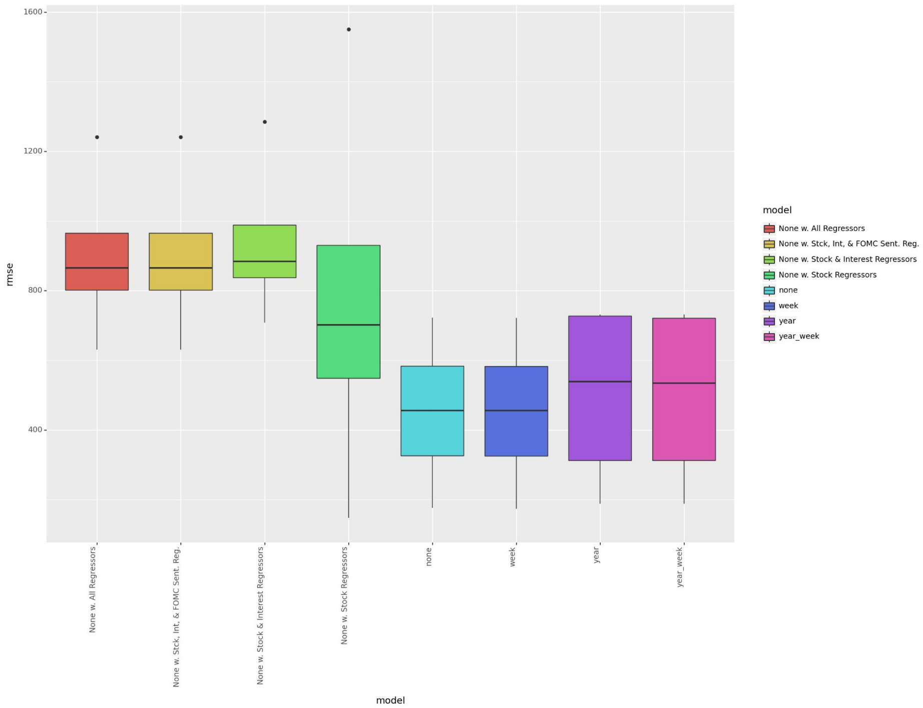


$$y(t) = g(t) + h(t) + s(t) + et$$

Prophet Results - Autoregression

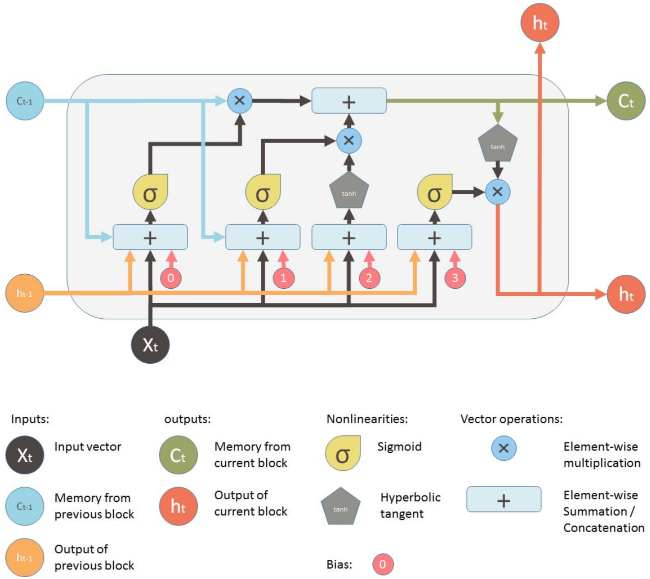


Prophet Results - Comparing Different Model Settings (n_splits=4)



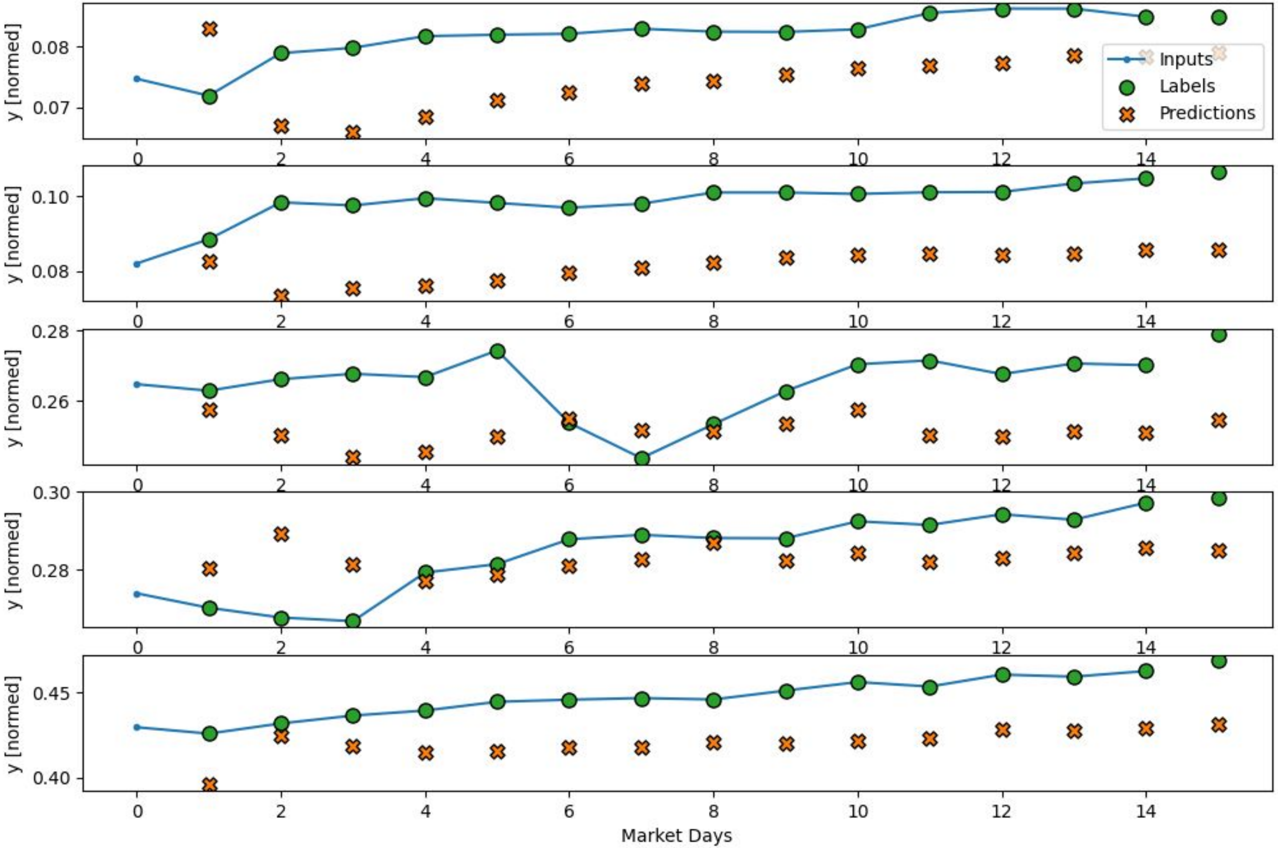
Introduction to LSTM

- Long short-term memory network is a recurrent neural network, aimed to deal with the vanishing gradient problem present in traditional RNNs.
- Excels at capturing long-term dependencies, making it ideal for sequence prediction tasks

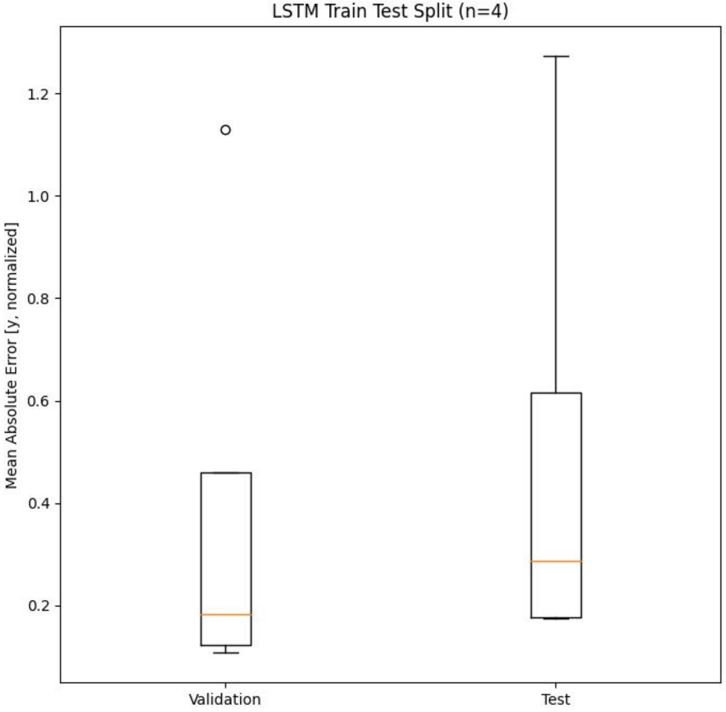


Ct-1
ht-1
Xt
ht
Ct
ht
σ
σ
tanh
σ
tanh
0
1
2
3
Inputs:
Xt Input vector
Ct-1 Memory from previous block
ht-1 Output of previous block
outputs:
Ct Memory from current block
ht Output of current block
Nonlinearities:
σ Sigmoid
tanh Hyperbolic tangent
Bias: 0
Vector operations:
× Element-wise multiplication
+ Element-wise Summation / Concatenation

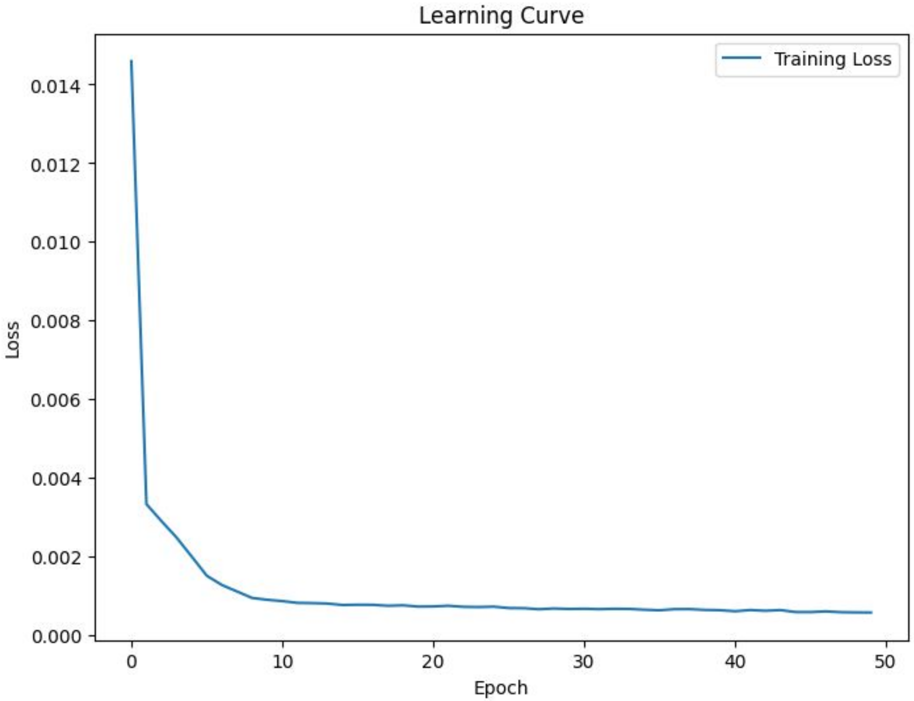
LSTM Results



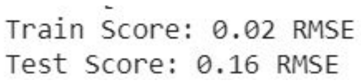
LSTM Results - Time Series Split (n_splits = 4)



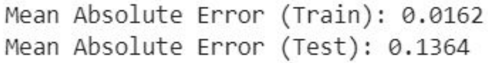
LSTM Results - Final Model



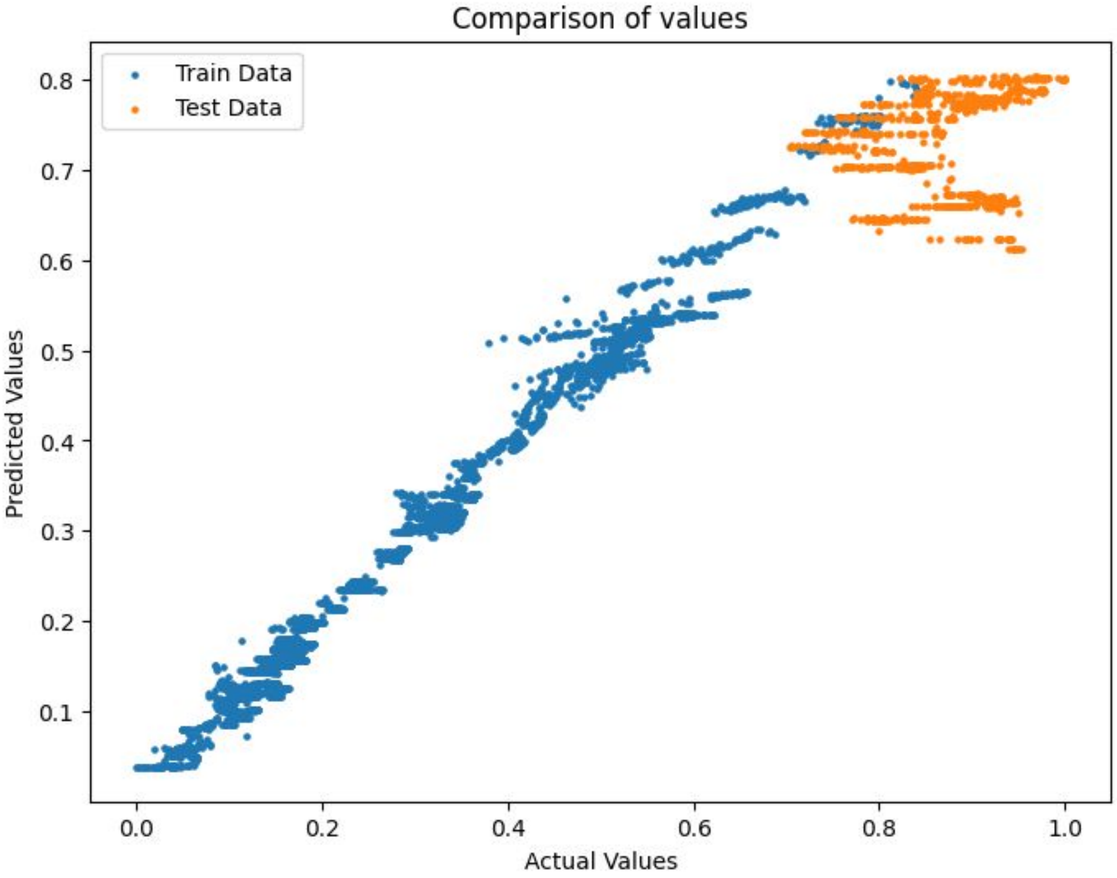
LSTM Results



```
Train Score: 0.02 RMSE
Test Score: 0.16 RMSE
```



```
Mean Absolute Error (Train): 0.0162
Mean Absolute Error (Test): 0.1364
```



Key Project Findings and Insights

- EDA on S&P 500 Index Composition
    - Provides a diverse representation of different sectors in the US stock market with a focus on software and tech
- EDA on S&P 500 Index Fund Price Data
    - High autocorrelation in daily price, low autocorrelation in daily changes in price
- Sentiment Analysis of FOMC Statements
    - Leveraged Loughran-McDonald Master Dictionary and FiGAS approach for Sentiment Analysis
- Machine Learning Models
    - Times Series Modeling requires specialized machine learning algorithms
    - While we did not see the algorithms in lecture, modeling a time series uses the same concepts

Challenges & Limitations

- Data Acquisition and Feature Selection:
  - Reliability of Free Stock Price Data and Data Retrieved via Web Scrape
- Sentiment Analysis:
  - Loughran and McDonald's Sentiment Word List still quite "basic" and misses context, intensity, offering only a limited view of sentiment through our raw scores/net sentiment
  - Flipping the sentiment for negation may not capture true causative factors behind the FED rate changes
- ARIMA
  - Model assumptions and requirements are too stringent for our Dataset
- Prophet
  - Lack of detailed documentation. While less restrictive than ARIMA, it lacks more advanced machine learning algorithms for additional regressors. And, the Holiday term is limited to model for a narrow range of events.
- LSTM
  - The most complex model, and most challenging to understand - made it difficult to use more advanced training validation methodologies (such as rolling window cross validation)
  - Shows a lot of promise, with sufficient tuning

References

Board of Governors of the Federal Reserve System. (n.d.). Federal Open Market Committee (FOMC). Federal Reserve. Retrieved [July 22, 2023], from https://www.federalreserve.gov/monetarypolicy/fomc.htm

Consoli, S.; Barbaglia, L.; and Manzan, S. (2020). Fine-grained, aspect-based semantic sentiment analysis within the economic and financial domains. In Proceedings - 2020 IEEE 2nd International Conference on Cognitive Machine Intelligence, CogMI 2020, 52 – 61.

Consoli, S.; Barbaglia, L.; and Manzan, S. (2021). Explaining sentiment from Lexicon. In CEUR Workshop Proceedings, volume 2918, 87 – 95.

Consoli, S.; Barbaglia, L.; and Manzan, S. (2022). Fine-grained, aspect-based sentiment analysis on economic and financial lexicon. Knowledge-Based Systems, 247: 108781

Conti-Brown, P. (2016). The Power and Independence of the Federal Reserve. Princeton: Princeton University Press. https://doi.org/10.1515/9781400873500

Esri. (2023). How change point detection works. *ArcGIS Pro Documentation.* Retrieved from https://pro.arcgis.com/en/pro-app/latest/tool-reference/space-time-pattern-mining/how-change-point-detection-works.html

Facebook. (2023). Quick Start. In *Prophet Documentation*. Retrieved from https://facebook.github.io/prophet/docs/quick_start.html

Hansen, K. B. (2021). Model Talk: Calculative Cultures in Quantitative Finance. Science, Technology, & Human Values, 46(3), 600–627. https://doi.org/10.1177/0162243920944225

Jabbri Maleki, F. (2023). Metrics. In *Time Series Exploration with Python: A Journey from Traditional to Advanced Forecasting Models.* Retrieved from https://faridjb.github.io/Time-Series-Exploration/chapters/chapter_09.html

References Continued

Kutzkov, K. (2023, July 31). ARIMA vs Prophet vs LSTM for Time Series Prediction. *Neptune.ai*. Retrieved from https://neptune.ai/blog/arima-vs-prophet-vs-lstm

Krieger, M. (2021, February 19). Time series analysis with Facebook Prophet: How it works and how to use it. *Towards Data Science.* Retrieved from https://towardsdatascience.com/time-series-analysis-with-facebook-prophet-how-it-works-and-how-to-use-it-f15ecf2c0e3a

Ma, E. (2018, November 3). Understand how to transfer your paragraph to vector by doc2vec. Towards Data Science. https://towardsdatascience.com/understand-how-to-transfer-your-paragraph-to-vector-by-doc2vec-1e225ccf102

Mehdian, S., Rezvanian, R., & Stoica O. (2019). "The Effect Of The 2008 Global Financial Crisis On The Efficiency Of Large U.S. Commercial Banks," Review of Economic and Business Studies, Alexandru Ioan Cuza University, Faculty of Economics and Business Administration, issue 24, pages 11-27, December.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *CoRR*, abs/1301.3781.

Mukhiya, S. K., & Ahmed, U. (2020). *Hands-on exploratory data analysis with Python: Perform EDA techniques to understand, summarize, and investigate your data*. Packt Publishing.

Takahashi, Y. (2020, November 13). FedSpeak — How to build a NLP pipeline to predict central bank policy changes: A guide to analyse policymakers' conversations by deep neural network. Towards Data Science. https://towardsdatascience.com/fedspeak-how-to-build-a-nlp-pipeline-to-predict-central-bank-policy-changes-a2f157ca0434

Tensorflow. (2023). Time series forecasting. *TensorFlow Core.* Retrieved from https://www.tensorflow.org/tutorials/structured_data/time_series

Stocker, M., Lakatos, C., Ohnsorge, F., & Kose, M. A. (2017, February 27). *Understanding the global role of the US economy*. CEPR. https://cepr.org/voxeu/columns/understanding-global-role-us-economy

Team and Contributions

- Jake Jarosik
  - ARIMA
  - LSTM

- Jonathan Watkins
  - Sentiment Analysis
  - EDA

- Jimmy Zhang
  - Data Acquisition and EDA
  - Prophet