Money Money Money!: Forecasting the S&P 500 DAPS Final Assignment

SN:24251714

Anonymous authors
Paper under double-blind review

Abstract

This study explores the challenges and opportunities in forecasting the S&P 500, a benchmark index often used as a barometer of U.S. economic health. By leveraging seven years of historical data (2017–2024) and incorporating diverse features such as economic indicators, news sentiment, and holidays, the study aims to predict the index's movement over a two-month horizon. Using the Prophet model, a scalable time-series forecasting tool, the analysis evaluates the predictive power of various data combinations, including baseline market data, economic indices, and sentiment analysis. Results indicate that while the inclusion of multiple data sources improves general trend prediction, pinpointing precise market movements remains elusive, constrained by market noise and inefficiencies. This study highlights the potential of combining diverse datasets for directional market predictions, while emphasizing the limitations of existing models in achieving actionable accuracy. Future work could enhance predictions through refined datasets and advanced machine learning techniques.

1 Introduction

The S&P 500 is a benchmark stock market index that tracks the stock price of the 500 largest companies listed on U.S. stock exchanges. It is widely regarded as a key indicator of the overall health of the U.S. economy. The ability to predict future movements of the S&P 500 is highly valuable, as it enables investors to capitalize on index-linked returns through products such as index trackers. This study focuses on forecasting these price movements.

For this analysis, seven years of historical data (April 1, 2017, to April 1, 2024) were collected, with the objective of predicting the index's performance during the final two months of this period. While predicting daily price fluctuations can be crucial for active traders, it holds limited utility for long-term investors. High-frequency trading, dominated by sophisticated algorithms and advanced systems, creates an environment that is nearly inaccessible for individual investors.Street (2025) Instead, identifying optimal entry points for long-term returns presents a more practical and impactful goal for the average investor.

Accurately forecasting the S&P 500 requires understanding the factors that influence it—a task that is inherently complex and often imprecise.Samuelson According to efficient-market theory, asset prices already reflect all available information, including future expectations, making it theoretically impossible to consistently outperform the market. Nonetheless, this theory has faced criticism, with proponents of market inefficiency pointing to scenarios where asset prices fail to align with their true value, creating opportunities to "beat the market."Insider (2010)

2 Data Description

The primary dataset collected for this study was the S&P 500. Like individual stocks, indices have a live price that fluctuates during trading hours. Additionally, after-hours trading influences the difference between the closing and opening prices, although this information is not publicly accessible.

For this study, determining the sampling period was critical. Given the high level of noise in stock market data, both intraday and over longer periods, frequent sampling was deemed unnecessary. With the prediction horizon set at two months, the closing price was selected as the sole daily value for the S&P 500 to reduce complexity and align with the granularity of other collected data. Only the S&P 500 data from February 1 to April 1, 2024, was used, as this period matches the study's prediction window and aligns with the focus on long-term trends.

The S&P 500 reflects broader U.S. economic conditions, suggesting that related indicators might exhibit correlation or predictive power. Therefore, additional features were collected, including the 13-week Treasury Bill (T-Bill), the CBOE Volatility Index (CBOE), and the U.S. unemployment rate. The T-Bill provides insights into interest rate movements, often adjusted in anticipation of market changes. The CBOE Volatility Index gauges market sentiment and risk perceptions, while the unemployment rate serves as a statistical measure of workforce health. All these features were recorded as numeric floating-point values and were chosen as they are often used alongside the S&P 500 to measure US econmoic health.

Holidays were also included as a feature to account for potential periodic patterns in the data. Although their impact on general trends might be minimal, holidays could help the model capture recurring seasonal effects over a two-month horizon. Each holiday was recorded as a string corresponding to its date, formatted to suit the model's processing requirements, as detailed in section 7.

Another key feature was news data, which has been shown to influence asset prices over short periods (daily to weekly) but has limited evidence of effectiveness in predicting longer-term movements. Lee et al. (2014) This limitation is likely because news is quickly absorbed into market prices. However, it remains potentially useful for predictive modelling. News headlines tagged with "US Stock Market News" were collected daily during the focus period. The choice of news prompt is critical, as the composite nature of indices like the S&P 500 makes them particularly sensitive to sector-specific shocks. For instance, the technology sector accounted for 32.9% of the S&P 500 by weight in early 2025, exposing the index to volatility in this sector.Tun (2025) Exploring sector-specific prompts might be a valuable avenue for future research.

All data was stored in a tabular format using the Pandas Python package, with dates as the index. News data, initially stored separately due to the possibility of multiple headlines per day, was indexed by unique identifiers for each headline. The final dataset was combined after interpolation to address missing values, as detailed in Section 5. The resulting primary dataset comprised 2,558 rows, after interpolation, with multiple columns corresponding to the described features. While the news dataset contained 4,459 rows indexed by headline IDs.

3 Data Acquisition

Data acquisition of the S&P 500 data was more problematic than initially expected. In the initial methods attempted, data was web-scraped from public financial news sources, including Yahoo Finance and Google Finance. However, during scraping it was found that Yahoo has licensing restrictions when scraping data off their website making it technically not available data for more than personal or academic purposes. As for Google, they do not have a way to readily view all close values for the S&P 500 between the required dates. This makes it not possible to view the required data for the requested period.

Despite its issues, Yahoo Finance was chosen as the source for data acquisition as the data required was available to be viewed. Additionally, a publicly available API called "Yahoo Finance API" was the chosen method to retrieve the data from the website. This API scrapes the data from the Yahoo Finance Website given a set of parameters, such as ticker and start/end dates. Although not an official API, it is subject to the same constraints as one scraping data from the Yahoo Finance website would have. It was used over manual web scraping as it was pre-built and achieved the same level of granularity required by the study.

The T-Bill and CBOE were retrieved using the same method. Going forward, if the product were to be put into production for commercial purposes, a subscription to an official data source for stock market data would likely need to be used due to licensing issues. Several issues arose when trying to retrieve data five years or older from free sources of stock market data that was not subject to licensing issues.

US unemployment Rate data is freely available directly from the Bureau of Labour and Statistics API. Holidays for US stock market are available through the holiday's standard library in Python.

The acquisition of News data was achieved using a web scraper. Unlike Google Finance, Google News allows for an RSS feed (Really Simple Syndication Feed) to be created for all news headlines from a search query. Additionally, the headline data is not subject to any licensing restrictions so is completely free of use for all purposes. This data is easily parsable using libraries such as feedparser. As described, headlines related to "US Stock Market News" were retrieved within the required dates. Two functions were made to handle this request, one which handled pagination of the RSS feed as only so many headlines were retrievable for a given request and one which handled the URL request formatting and data parsing. This data was not subject to licensing and is not restricted from web scraping. Google News, is a good source for News data for this reason as it combines a wide variety of sources to give a more comprehensive dataset. However, it may also lack the focus and reputability that individual News outlets might provide. Unfortunately, most well-known Finance specific outlets do not have a freely available API.

4 Data Storage

Data storage was done in two parts. Once data was retrieved it was partially processed to allow for a format that could be stored within the database of choice. This processing consisted of tabularizing the index tracker, unemployment and holiday data into a single table that was indexed by date. For the News data, there was little processing at this stage. This initial step was done to ensure that the raw data retrieved did not change with further iterations and allowed the development of the data acquisition and processing to be done separately.

The next step was to re-save the data upon completion of further preprocessing. Some of the preprocessing steps such as sentiment analysis of the news headlines required significant resources and running in real-time would be a waste of resources, particularly during ablation studies. Therefore upon completion of the sentiment analysis, the data could be formatted to be a full dataset indexed by date where each column was an individual feature.

Despite the final dataset being suitable for a structured database, some of the initial raw data particularly the news headlines was not structured. Rather than swap between an unstructured and structured database, only an unstructured one was used. MongoDB was the database of choice, as it is largely free to use for small amounts of data, and allows data to be stored in key-value pair documents. Each document corresponded to a particular day where keys were a given feature. For News data, each document was an individual News article. Additionally, the code was designed not to make repeated calls to a database, as all the data was required for exploration studies and inference. Rather the database needed to act as an intermediate where all the data could be retrieved and saved at once and would then be handled accordingly. A further reason for the suitability of an unstructured database where all the data could be called at one.

5 Data Preprocessing

As highlighted in section 4 preprocessing was done in two parts. The initial part was to process all the raw data into pandas data frames and fill in the missing values. The time-series data from the market data contained many missing days. This is a result of the stock market being closed on weekends and holidays. Typically most inference models require data to be consistent with no missing values in it. Although the data could be formatted in such a way that weekends/holidays are removed, it would still interfere with the model inference as the prediction window would need to learn behaviour at weekends. Additionally, the time series would not be able to be indexed correctly on a day-by-day basis which would cause issues for the inference model chosen as described in section 7 which requires time data to be consistently spaced. Additionally,

unemployment data is only monthly frequency so this needed to be converted to a daily frequency to align with the index tracker

A linear interpolation method was used to fill the missing dates within the raw data. This method was deemed the most appropriate as zeroing the values would cause discontinuities which would interfere with the model. Additionally, more complex interpolation methods such as Lagrangian interpolation are often not necessary for small intervals with only a couple of missing values (i.e. weekends). More complex interpolation may have been suitable for the unemployment rate estimates where there are more missing values and this may be an area of improvement in further studies.

The next stage of the preprocessing was to convert the News data into a format that was useful for the model. One way to approach this problem is to view the sentiment of the news article as being either a positive, neutral or negative article on the state of the US stock market, and use this information to apply a numeric value to each of the data points. This may in turn affect how the market responds, for example, price increase for positive news and vice versa. There are many ways to perform sentiment analysis with the most primitive being hand labelling every article. Instead, a pre-trained version of BERT (Bidirectional encoder representations from transformers), downloaded from hugging face Rachid was used. BERT is a language model able to understand the context of phrases and when trained correctly perform sentiment analysis. The specific model used was tuned so to produce better results on financial news. The advantage of using "off the shelf" models is that they are quite easy to implement and produce promising results. Thus, the news sentiment headlines were passed to this model giving a label to each. Using this data, each day of the period was given a value for the number of positive, neutral and negative headlines on that day, thus resulting in three new columns one for each sentiment type. These values were stored in the same table alongside the s&P 500 data and the data was stored in the database, as per section 4

However, off-the-shelf models are not always advantageous for situations where there is a competitive advantage trying to be made. Investing in the stock market can be thought of as a game, where one is trying to beat the market and the market consists of everyone else in the world. In this sense, if everyone else in the world has the same models for sentiment analysis at their disposal then they will produce the same or similar results. This is dependent on the data fed into the model, but assuming this is fixed then simply using an off-the-shelf model will not give a competitive advantage. The difficulty lies in developing complex LLM like BERT which often requires a significant amount of resources to train and develop. For this study, the off-the-shelf model may provide some insight as to whether there is any advantage of using news data for long-term investments which has yet to be shown in the literature.

6 Data Exploration

Data exploration tasks were split between S&P 500, additional timeseries data and news data. This was done to allow a more effective comparison of the additional data to the underlying data. News data has been separated as it requires separate analysis to the rest of the data.

Firstly, the S&P 500 trend was plotted for the whole time period to visualise the general look of the data. An important trend in data of this type is to see if there is any underlying seasonality, that is caused by generic market fluctuations or caused by time-of-year events. The importance of this cannot be understated as the prediction window for this study is limited to two months by design, meaning slight periodic fluctuations may have a massive effect on the outcome. This may not be true for longer prediction horizons but is certainly true for this study. To compute the seasonality the STL function taken from the statsmodel library was used. This function decomposes the data into trend, seasonal and residual elements. A plot of an example for 2019 can be found in the appendix, Figure 4. Decompositions were made for the whole period and select years to analyse different scopes. Generally, it was found that there was a slight periodicity to the seasonal data but that it was almost always dwarfed by a residual or trend element making it difficult to provide much real predictive power. Similarly, a Fourier analysis of the timeseries was performed to try to identify if any major frequencies were making up the time series. Filtering techniques were also used to try to extract key frequencies and remove noise. As the filtering is not done in real time simply modifying the frequency domain signal by hand was done, in contrast to designing FIR or IIR filters. The results did not show that there was any underlying signal that was not already captured by the STL decomposition.

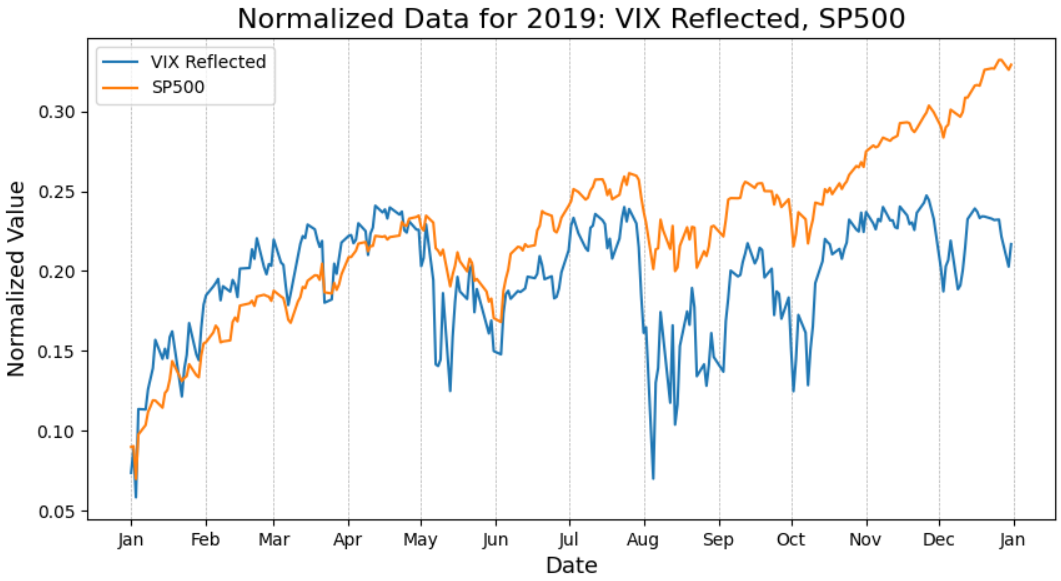

Figure 1: Plot of the CBOE (VIX) reflected and normalised to align with the S&P 500 data.

The final techniques attempted to see if there was any pattern on a daily/monthly basis by observing the mean of the values across the period and observing the changes during the month/week. Every month, the mean was calculated for each day of the year (i.e. mean of 01/01/2018, 01/01/2019, etc.), giving a singular value for each day of the year. The standard deviation was used every month to see if the price tended to move in a particular direction, this was reset after each month to ensure the spread remained within the month, a plot of this can be seen in the appendix in 5. The results of this show that the spread, even under one standard deviation, is so significant that no trends can be resolved reliably. For a daily basis, the mean price change was calculated for each day of the week by month for the whole period; this would allow us to view if there were any trends in weekly price movements, a plot of this can be found in the appendix in 6. Issues with this plot is that it does not show the spread of the price change, regardless aside from the odd anomalous day, there is little trend such as summer months experiencing more volatility during the week.

Very little predicted predictive power is gained from the S&P 500 alone, and so the next step was to analyse the effect of other stock market and market data on the S&P 500. Initially, a correlation matrix for the different data sources was computed using the normalised data and Pearson correlation for the evaluation; results can be seen in the appendix in figure 7. The results showed very little relation between the S&P 500 and the other datasets at first glance. However, after plotting the normalised data for the datasets for the entire period, Figure 8, it was noticed that the CBOE Volatility Index (VIX) is a close reflection of the S&P 500. This would make sense as it can be viewed as the markets "fear" of the S&P 500 based products, and so should have a strong correlation to the S&P 500. By manipulating the data through a reflection and replotting for the year 2019 the Figure 1 is returned. Although visually very aligned the predictive importance depends on whether the VIX movement precedes that of the S&P 500. In some scenarios, it does, such as the month of May where there is a sharp decline in the VIX before the S&P 500 adjusts. However, its composure is simply an amalgamation of the known information about the S&P 500, so it will never hold key information that will enable the prediction for long-term results.

The next dataset included was holiday data. The effect of holidays on the data could be found by overlaying the days that holidays occurred onto a graph of the S&P 500. However, it was largely found that these had little effect on the price movement, see Figure 9.

Finally, the behaviour of News sentiment data was explored. The vast majority of the labels for the dataset were of the neutral sentiment with about 66% of the total labels, leaving 13% for Negative and 21% positive.

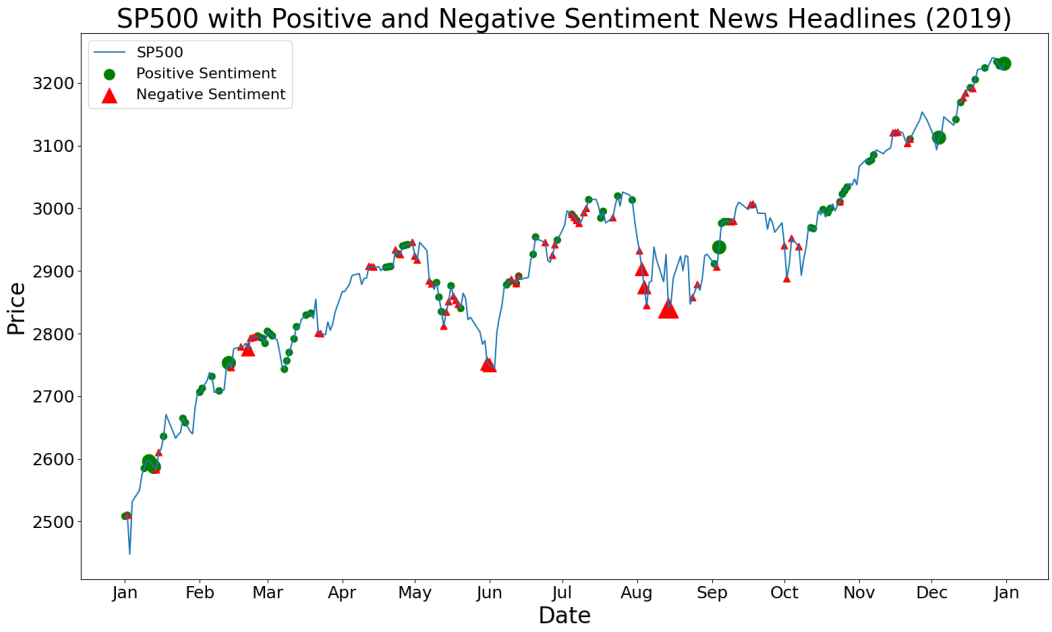

Figure 2: Plot of the S&P 500 data for the year 2019, with sentiment analysed news headlines overlaid on top.

Only the positive and negative provide useful information in this context as they would indicate times at which the price would experience increase or decrease. Additionally, having too much neutral sentiment might obscure the important data information and so was left out from the analysis. The positive and negative sentiment labels were overlaid onto the S&P 500 data, as seen in 2, in which larger markers indicate larger values for each of the sentiment labels (i.e. more positive/negative headlines in a given day, as per the preprocessing method in section 5). The results largely showed that much of the positive or negative labels fell after the event with which they appeared to be correlated. For example, there appear to be many instances where negative news headlines appear at troughs in the data. In many ways, this is as expected as the news more than often simply reports on what has already happened rather than provide critical analysis of what it thinks is going to happen and so provides little predictive power.

7 Training and Inference

Following the exploratory analysis, doubts arose about the effectiveness of using news data to predict the S&P 500's movements two months in advance. To address this, model validation was structured across three distinct approaches: a baseline model using only S&P 500 data, a model incorporating additional indices data (e.g., CBOE and T-Bill), and a model utilizing news sentiment data.

To compare the models effectively and optimize their parameters, the dataset was divided into training, validation, and test sets. Unlike some models that analyze only a limited data window, the chosen model allowed for all historical data to be used during training. Given the temporal nature of the data, methods like k-fold cross-validation were unsuitable. Instead, a single validation period, covering the two months immediately preceding the test period, was selected (i.e., validation: December 1, 2023, to January 31, 2024; test: February 1, 2024, to April 1, 2024). This ensured maximum utilization of prior data while preserving a logical chronological flow for forecasting.

Rather than relying on commonly used machine learning models, which often require significant tuning and extended training times, this study employed the Prophet model. Developed by researchers at MetaTaylor

& Letham (2017), Prophet is a robust and rapid forecasting tool designed for time series data, particularly for applications like sales predictions. It models seasonality and external factors such as holidays, making it an excellent choice for this study. While advanced AI models like recurrent neural networks (RNNs) might produce superior results with ample resources and time, Prophet's simplicity and efficiency made it more practical given the study's constraints.

The prophet model is relatively simplistic in its operation. The model requires a pandas dataframe indexed on the date of each data point with the training data, and validation or test data input as separate inputs. Additional data is then added using adding them as regressors to the model. Holiday data can also be included as part of the model and has been included for all models in this study. There are a limited numbrt of hyperparameters that can be tuned and these mainly consist of prior scales which essentially act to change the flexibility of the model to adapt to certain inputs.

For the base model, only the S&P 500 data was used. For the indices model, despite not showing significant correlation both the Treasury Bill and Unemployment data were included initially for experimental purposes. As for the CBOE volatility index, it was decided that initially the non-reflected data would be experimented with before trying reflected versions of the data which were found to have possibly better predictive power than the regular data. It should be noted that Prophet does not require data to be normalized before use. As specified, news sentiment data was included in a separate model to evaluate its performance individually. After the three initial models were implemented, it was explored what effects using the reflected VIX alone would have along with using all the available data.

Three metrics were used to measure the performance of each of the models, each with its own advantages and disadvantages, these being the MAE, RMSE and MAPE. A plot for each of the validation predictions was made to visualize the results, the baseline model can be found in the appendix in figure 10. Additionally, a scatter plot of the predictions vs the observed or real value was made for both training and validation/test set to better visualize the relation between the two, an example of the test data can also be found in the appendix in Figure 11.

Table 1: Validation and Test Results for Different Models

Model	**MAE**	**RMSE**	**MAPE** (%)
Base Model Validation	246.51	257.22	5.17
All Indices Model Validation	184.83	201.99	3.86
Sentiment Model Validation	243.50	254.69	5.10
Reflected VIX Model Validation	305.32	319.55	6.40
All Features Model Validation	178.74	195.72	3.73
All Features Model Test	280.83	295.65	5.48

The results for the different models were only compared using validation data; comparing them using test data would technically introduce a form of data leakage. It was found that the model using all the features available performed best compared to the other models tested, as shown in Table 1. A plot of the predictions for the test data is shown in Figure 3. The results show that all the models do not perform particularly well. However, they all did manage to predict the general direction and movement of the S&P for the next two months of validation, which in turn would lead to a net return if the prediction was used to direct the investment strategy. So in some way they are successful models in that regard. However, none of them can track it accurately so would not be able to inform us of when the best point to invest is by buying at the lower point within the two-month prediction window. It is hard to explain exactly why the model using all features performs best; it may be linked to how the regressors add additional information but is more likely the model struggles to use all the information available and so aligns more strongly with a general trend then being influenced by seasonality which would play a larger roll with less data. It just so happens that the actual values for the validation and test data have very little seasonality components to them and so the model that does not have as heavy a seasonal component to its prediction appears to be the better performing model. Therefore, until tested on additional test data it would be wrong to conclude which is the best model.

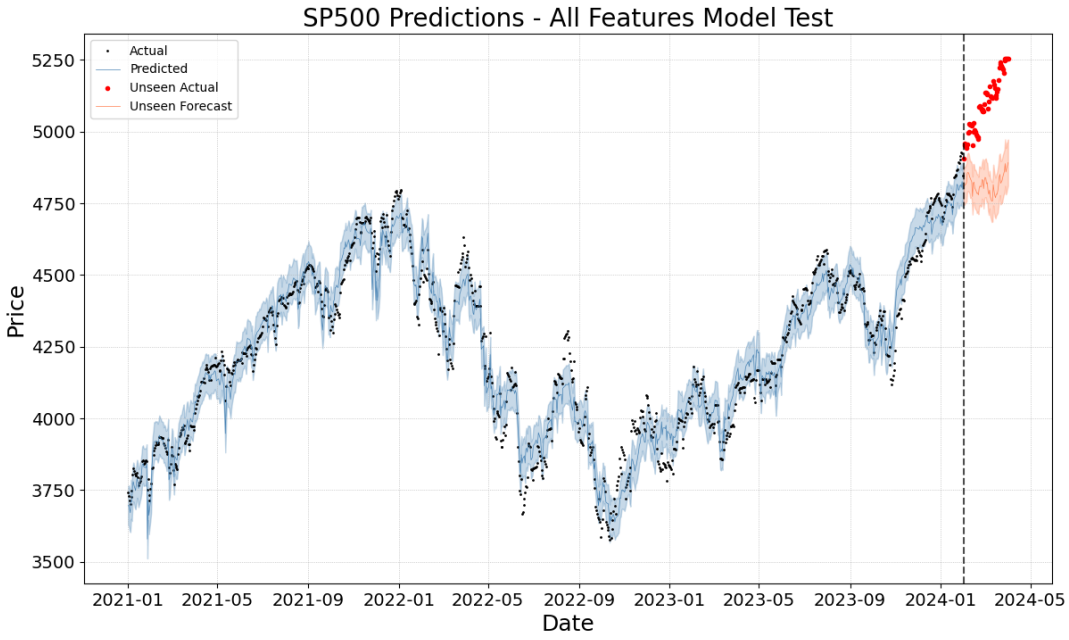

Figure 3: Plot of indices model predictions on the test dataset, predicting the two months of February and March. Sample of two years prior of the training data shown.

8 Conclusions

Predicting the S&P 500's long-term trends involves navigating the complexities of market inefficiency—where the goal is to uncover information the market has not yet fully reflected. The essence of this study lies in leveraging diverse datasets and methodologies to understand these inefficiencies and evaluate their predictive power. This analysis has underscored two critical pathways to success: either identifying overlooked market factors or challenging prevailing assumptions, both of which hinge on accessing unique insights.

While the study's approach, including the integration of economic indicators, sentiment analysis, and advanced time-series modelling, yielded some success in predicting general market trends, it also highlighted the inherent limitations. Notably, the all-features model outperformed simpler alternatives, reflecting the value of combining multiple data sources. However, its predictions, though directionally accurate, lack the granularity required to pinpoint optimal investment timing—a crucial limitation for practical application.

The results affirm that achieving precise predictions of market movements remains a formidable challenge. The interplay of seasonality, noise, and the efficient-market hypothesis contributes to this difficulty. As illustrated by the study, even advanced models such as Prophet can struggle to provide actionable foresight beyond general trends.

In conclusion, while perfect foresight of the S&P 500 remains elusive and perhaps undesirable under the principles of market efficiency, this study demonstrates that models can still offer meaningful insights for strategic decision-making. Future research should focus on refining datasets and exploring more robust machine learning architectures to better capture the nuanced dynamics of financial markets. This will not only improve predictive capabilities but also deepen our understanding of the ever-evolving interplay between market behaviour and external influences.

References

Business Insider. Warren buffett on efficient market hypothesis, December 2010. URL https://www.businessinsider.com/warren-buffett-on-efficient-market-hypothesis-2010-12. Accessed: 2025-01-18.

Heeyoung Lee, Mihai Surdeanu, Bill MacCartney, and Dan Jurafsky. On the importance of text analysis for stock price prediction. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 1170–1175, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL https://aclanthology.org/L14-1048/.

Ahmed Rachid. Financialbert-sentiment-analysis. https://huggingface.co/ahmedrachid/FinancialBERT-Sentiment-Analysis. Accessed: 2025-01-20.

Paul A. Samuelson. *Proof that Properly Anticipated Prices Fluctuate Randomly*, chapter Chapter 2, pp. 25–38. doi: 10.1142/9789814566926_0002. URL https://www.worldscientific.com/doi/abs/10.1142/9789814566926_0002.

Jane Street. Overview of what we do, 2025. URL https://www.janestreet.com/what-we-do/overview/. Accessed: 2025-01-19.

Sean J Taylor and Benjamin Letham. Forecasting at scale. *PeerJ Preprints*, 5:e3190v2, 2017. URL https://peerj.com/preprints/3190v2/.

Zaw Thiha Tun. Top 25 stocks in the s&p 500 by index weight for january 2025, 2025. URL https://www.investopedia.com/best-25-sp500-stocks-8550793. Accessed: 2025-01-20.

A Appendix

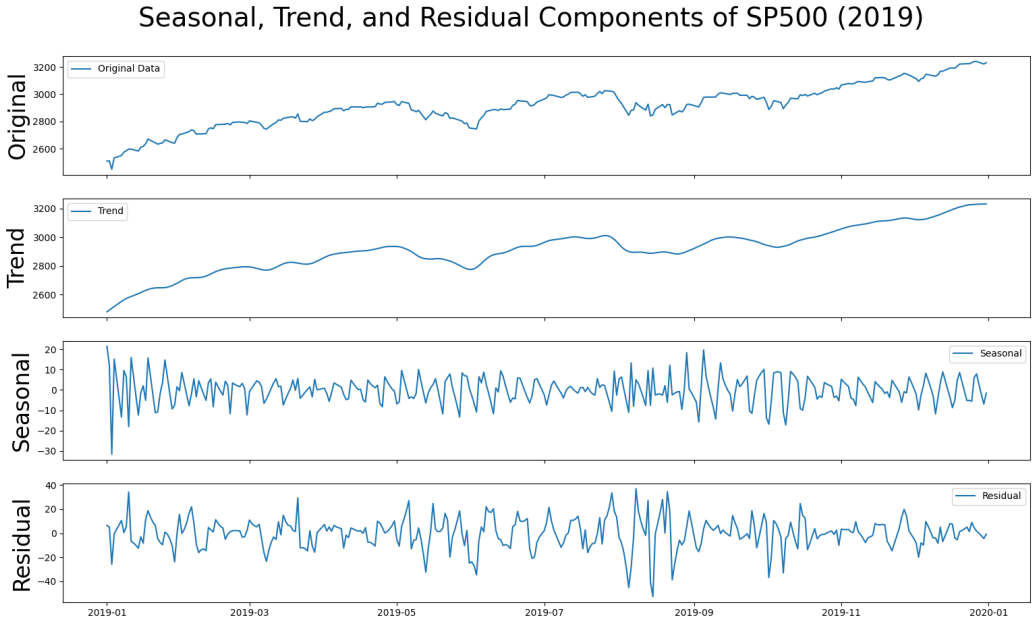

Figure 4: STL decomposition of S&P 500 for 2019.

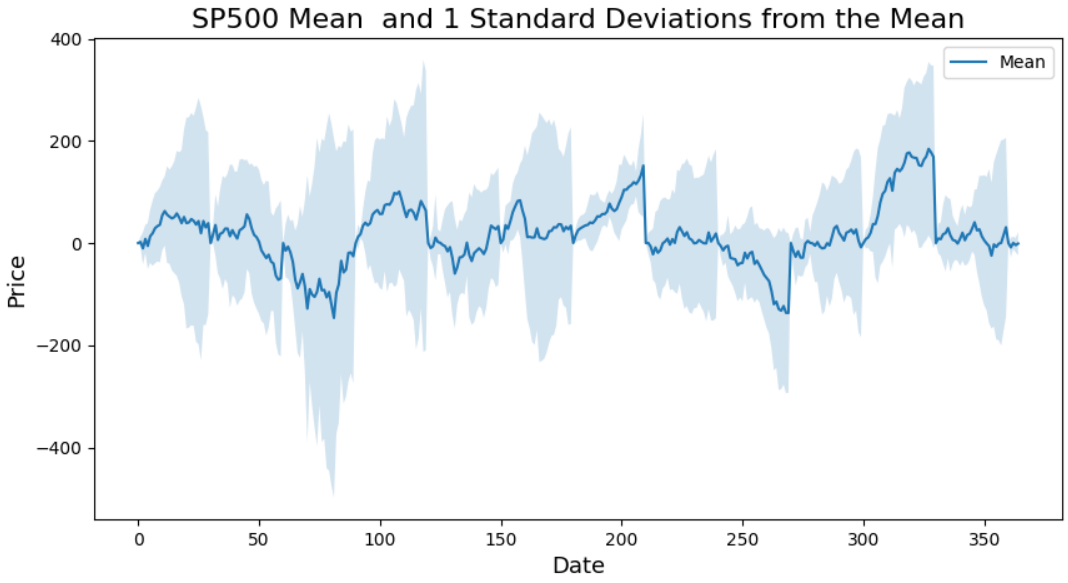

Figure 5: Mean of all values for each day of the year for the S&P 500 between 2017 and 2024. Spread is calculated on a monthly basis resetting to zero each month.

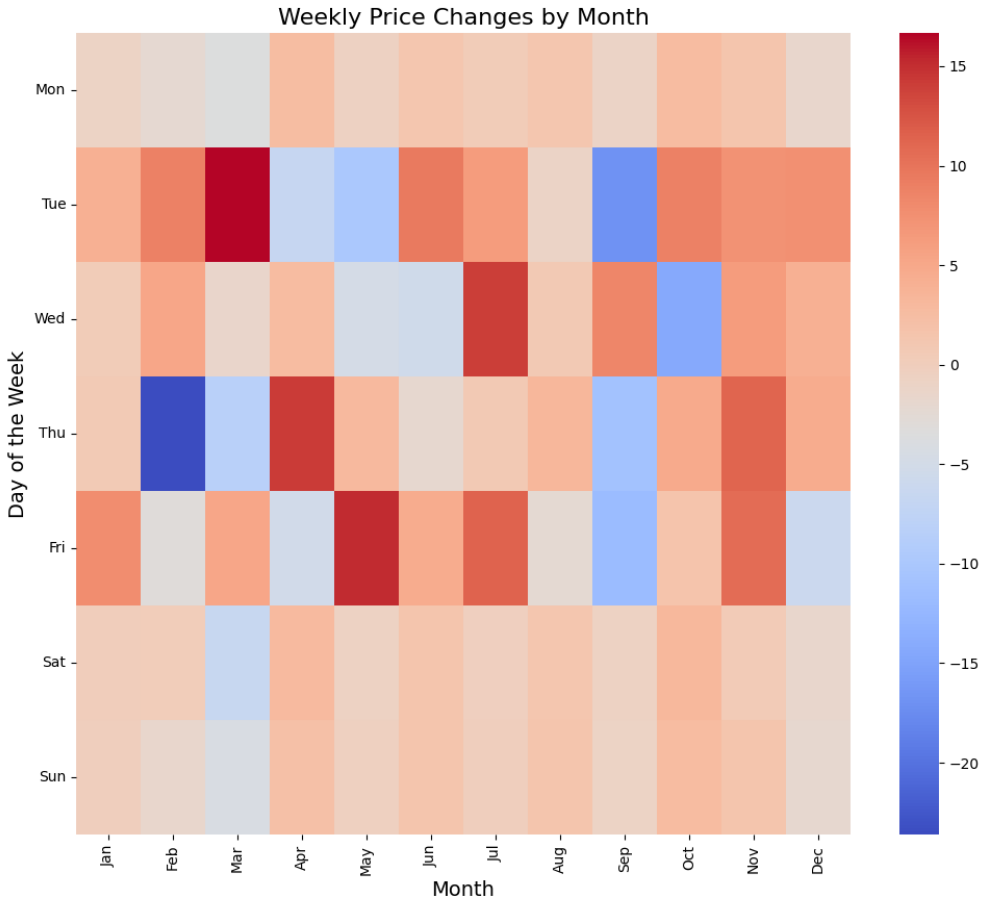

Figure 6: Change in price by day for all days of the year for the S&p 500 between 2017 and 2024.

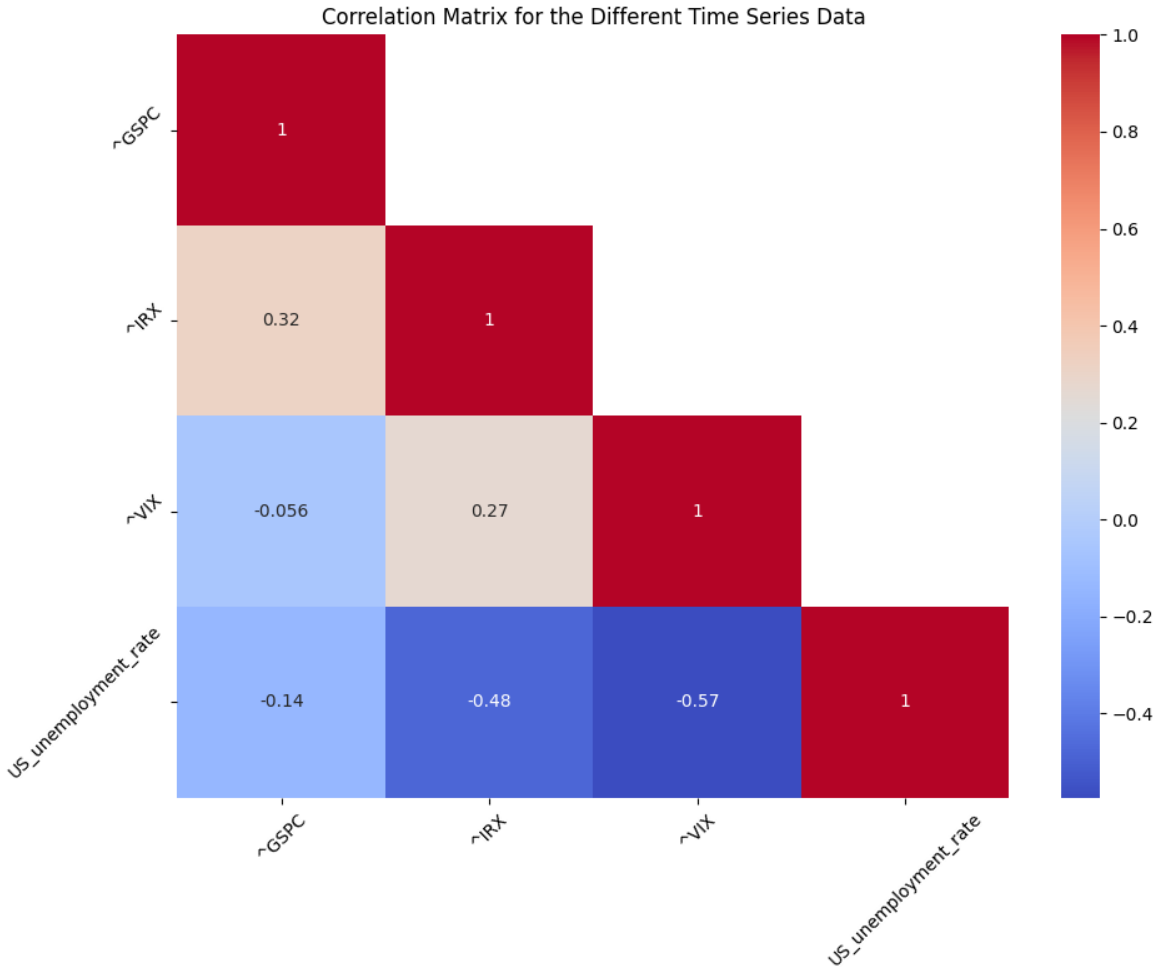

Figure 7: Correlation matrix for the different data sources calculated using Pearson correlation coefficient on normalized data. ĜSPC (S&P 500), ÎRX (Treasury Bills), V̂IX (CBOE Treasury Index).

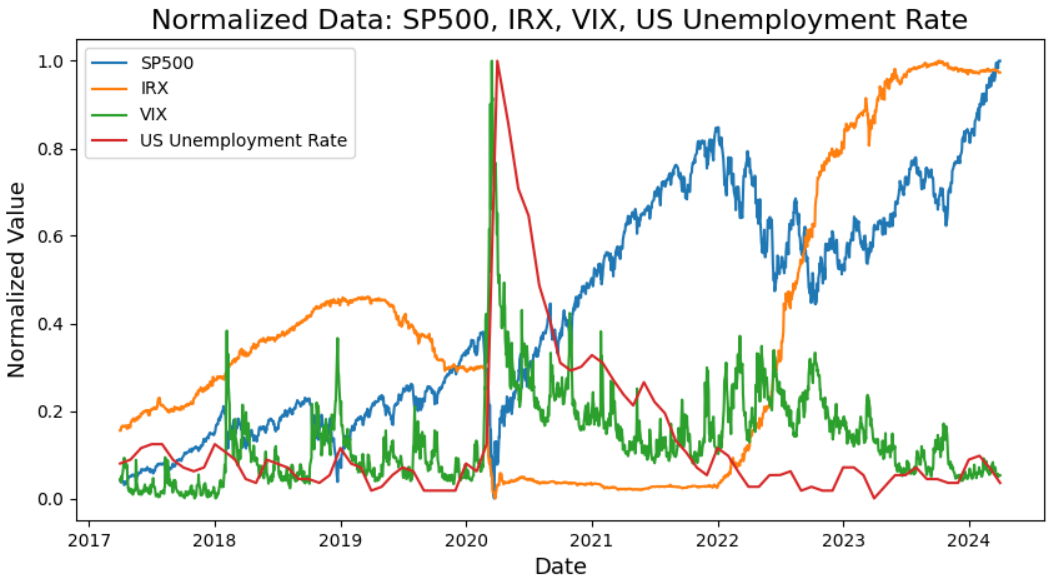

Figure 8: Plot of the Normalized time series from each of the time series data sources for the entire period.

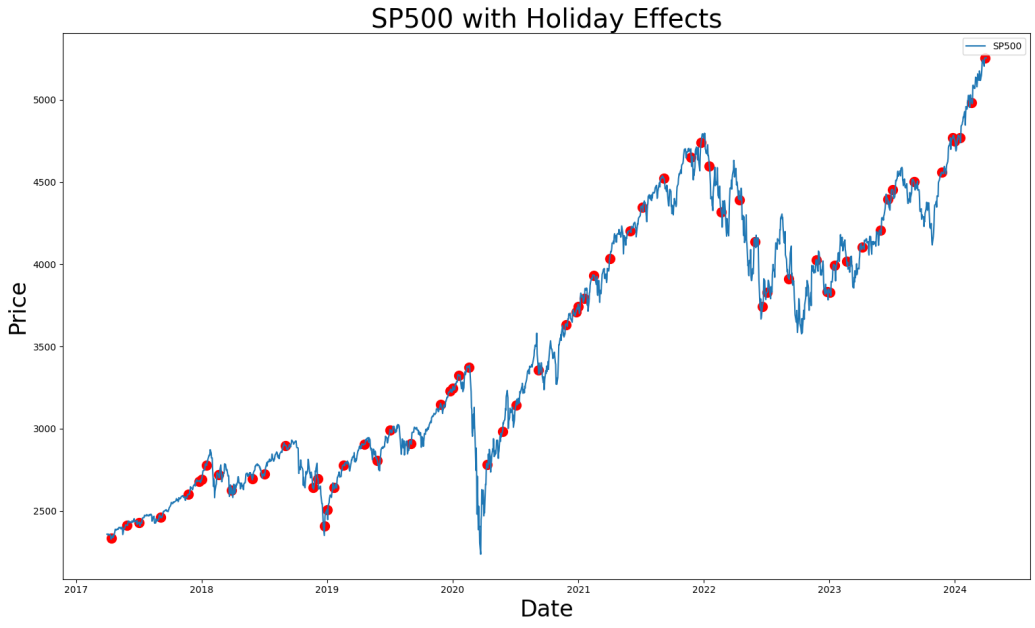

Figure 9: Plot of S&P 500 with Markers for Holidays overlaid.

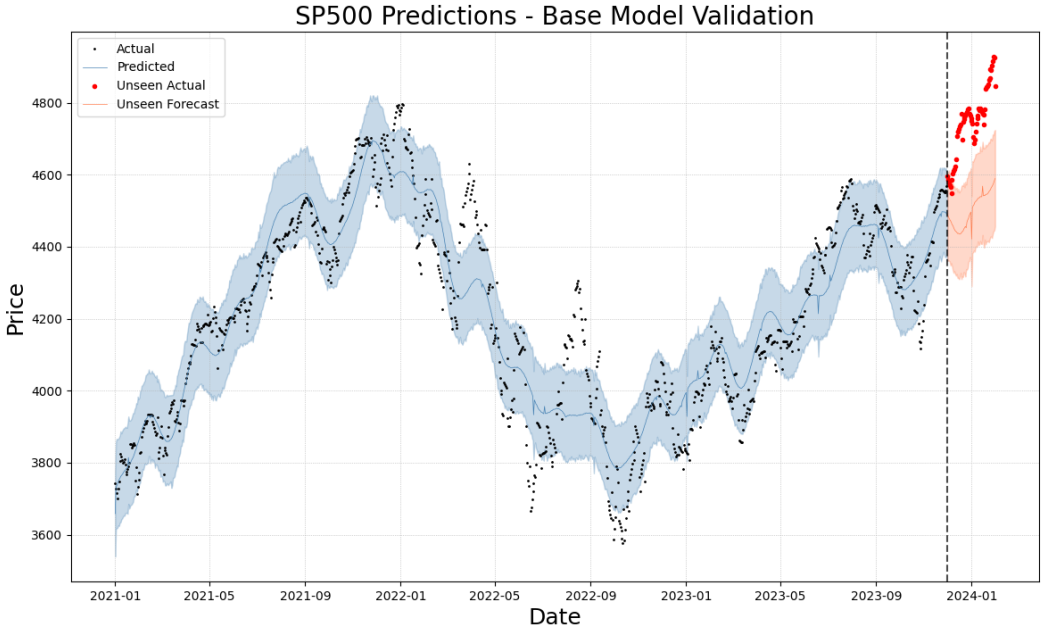

Figure 10: Plot of prophet predictions using only the S&P 500 data.

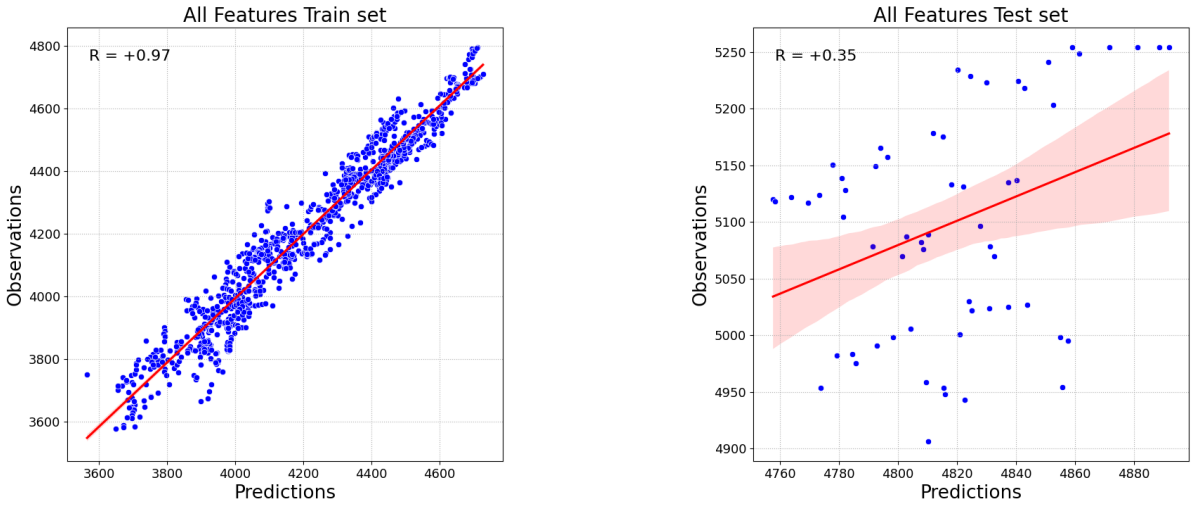

Figure 11: Scatter plots of the correlation between the predicted and observed values for the train and test data on the indices model, with correlation coefficient R.