SP 500 Market Analysis

Thursday 11th December, 2025

```
- - -
bibliography: references.bib
numbering:
  heading_1: false
 - - -
```

1 S&P 500 Market Analysis

1.1 Basic Information

Authors: Arhaan Aggarwal, Brian Hwang, David Robertson, Jose Aguilar **Class & Semester:** Fall 2025, Stat 159 **Supervisors:** Fernando Pérez, Jimmy Butler, Sequoia Andrade

1.2 Introduction

Many of us wonder how to profit from the financial world, specifically the stock market. Logically, this task is very challenging and is one of the deeply researched fields in academia. According to the Efficient Market Hypothesis (EMH), asset prices reflect all available information, indicating that it is impossible to regularly “beat the market” using solely historical dataMalkiel [2003].

In our project, we will look to analyze how and what to predict the stock market through the S&P 500 index ((ĜSPC)). By utilizing historical trading data from 1990 to 2025, we aim to answer the specific questions proposed in our study:

Actually Predicting Market Movement

Can we accurately predict short-term movements in the S&P 500 using historical price and macroeconomic data?

Model Comparison

How do simpler predictive models compare to more advanced models in forecasting performance?

1.3 Data and Methods used

To begin, we needed to find a public dataset that contained the historical financial data of the S&P 500. In order to do this, we used the yfinance API to gain daily trading data from 1990 to the present Aroussi [2023]. From this data, we added extra technical indicators that gave us more information about how the prices were moving.

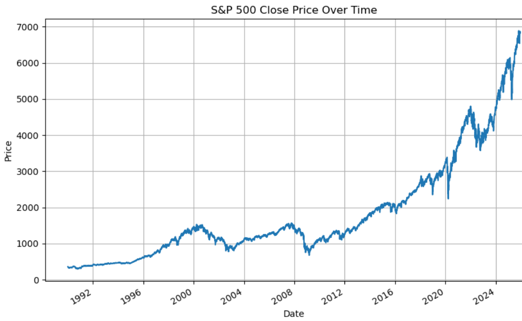
** FOR MY GROUP PLEASE EXPLAIN FURTHER about the TECHNICAL INDICATORS IF U THINK IT IS NEEDED **

We note that for our research, we do have some assumptions. First, we assume that the Yahoo Finance data itself is trustworthy. For example, for the "Adjusted Close" column Yahoo is using historical stock splits and dividend distributions to calculate this. Therefore, we are assuming that Yahoo's data is accurate, valid, and represents the true stock market. Second, we assume that financial data is non-stationary, thereby using log returns rather than normal return values. This is to ensure that the models we use are appropriate to be used as they model stationary data. Lastly, we are assuming that the EMH holds true as we are going through our research.

1.4 Exploratory Data Analysis (EDA)

Our EDA is looking to find how the statistical properties of the S&P 500 work.

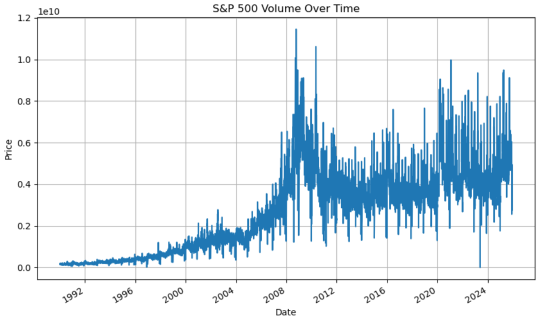
1.4.1 Trend



Our financial data does indeed have a long-term upward trend.

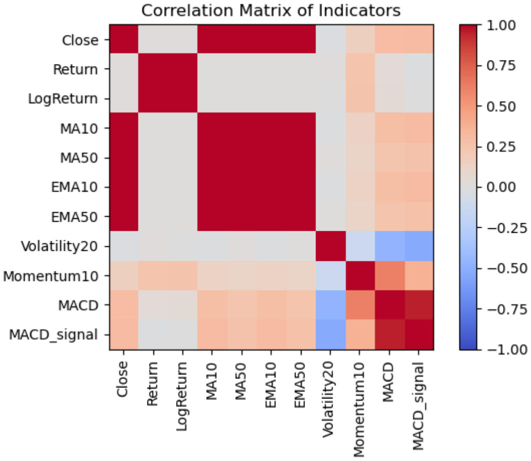
1.4.2 Volatility

Our financial data confirms that in certain years such as 2008 and 2020 we have high volatility as we had a lot economic downturn during those times.



1.4.3 Feature Correlations

We analyzed the correlation between our technical indicators and the variable we seek to predict (LogReturn). While trend indicators such as MA50 correlate with each other, their correlation with LogReturn is close to zero, helping us understand that perhaps Linear Regression model is too "simple" to capture any valid predictions in the stock market.



1.5 Our Models and the Results

1.5.1 Why we chose our 5 models

The Naive and Linear Regression models are to give us a general benchmark as if more complex models cannot outperform these simple ones, it aligns our findings with our assumption and idea that the EMH is true.

The Random Forest model was chosen to capture any non-linear relationships between technical indicators to see which ones help the most with predictions of returns.

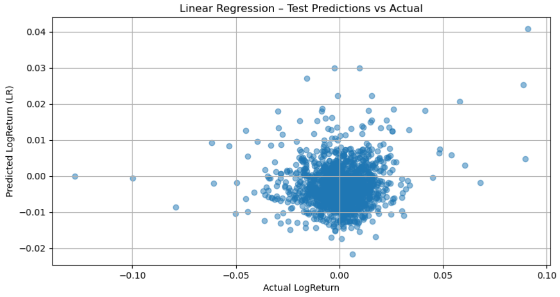
The ARIMA model was chosen to model the autocorrelation of the data itself in that instead of focusing on external technical indicators we added, it looked purely at past returns to predict the future one.

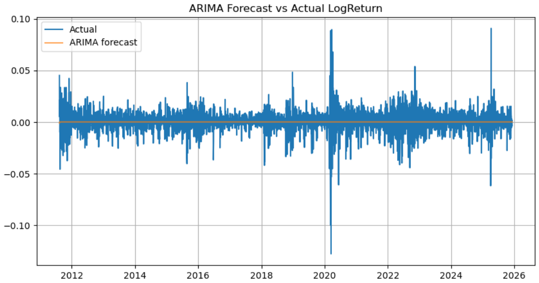
Lastly, we used LSTM model to see how we can find long-term relationships in features and the market through sequential data. For example, given the past 30 days predicting the current return would be what LSTM is used for.

1.5.2 Modeling Results

We evaluated five models with a split (Train: 60%, Val: 20%, Test: 20%) to ensure a valid modeling technique.

Below you can see the visualizations of the Linear Regression model and ARIMA model.





Evaluating these and the other models, we summarized the performance of each model on the unseen Test set.

```
visio = pd.read_csv("data/final_metrics.csv")
visio = visio.drop(columns= "Unnamed: 0")
visio.index = visio.index + 1
visio
```

	Model	RMSE	MAE	MAPE	R2	DirAcc
1	Naive	0.016941	0.011610	5.502865e+07	-0.766366	0.509162
2	LinearRegression	0.013780	0.010006	3.790434e+07	-0.168709	0.464187
3	RandomForest	0.019296	0.015812	5.571611e+07	-1.291640	0.463631
4	ARIMA(1,0,1)	0.010885	0.007145	3.433071e+07	-0.000894	0.545656
5	LSTM	0.015943	0.010461	4.436545e+07	-0.573669	0.492377

1.6 Conclusion

In this study, we saw if we could truly predict the daily S&P 500 returns using historical price data and technical indicators. The research we conducted allows to answer our initial research questions.

Can we accurately predict short-term movements in the S&P 500?

We deem that predicting daily returns is extremely difficult, if not impossible, using only historical technical data. Our Linear Regression and Random Forest models had negative R^2 scores on the test set, indicating they failed to outperform a simple mean-baseline (our naive baseline). This result sheds evidence to the Efficient Market Hypothesis (EMH), as if past price information is fully reflected in current prices, then simple regression models would fail to accurately predict anything Fama [1970]. However, our ARIMA model had a Directional Accuracy of ~54.6%. While the model could not predict the amount of the return, it correctly identified whether the market would close Up or Down more often than the random guess we initially suspected. In the context of high-frequency trading, a 54% directional accuracy could potentially be profitable.

Simpler or Advanced Models: Which one is better?

While this is a bit of a mixed area, the fact is that the assumption that "more complex is better," is false. This is as we showed that the ARIMA – a more "simpler" model outperformed complex machine learning algorithms for this job such as the Random Forest model. Moreover, the LSTM (our deep-learning model) failed to find significant sequential patterns that simpler models missed.

In conclusion, this project demonstrates that predicting the stock prices is extremely difficult and is in line with it being a random walk. Moreover, our research also suggest that financial time series analysis is perhaps more effective in managing risk and identifying small advantages like a 54% directional accuracy. We recommend that future work should look at incorporating more variables such as interest rates or news sentiment rather than relying solely on historical price data in order to predict future prices and returns.

1.7 Author Contribution

- **Arhaan Aggarwal** Enter
- **Brian Hwang** I split the main notebook containing all the code into its separate notebooks. I also structured the entire repo, moved all the functions into the scripts folder, and wrote tests for some of them. I also worked on the reproducibility aspect of the project by working on the environment.yml and make files while also beginning myST section. Lastly, drafted the main narrative notebook.
- **David Robertson** Enter
- **Jose Aguilar** Enter

References

R. Aroussi. yfinance: Yahoo! Finance market data downloader, 2023. URL https://github.com/ranaroussi/yfinance.

E. F. Fama. Efficient capital markets: A review of theory and empirical work. *The journal of Finance*, 25(2):383–417, 1970.

B. G. Malkiel. The efficient market hypothesis and its critics. *Journal of economic perspectives*, 17(1):59–82, 2003.