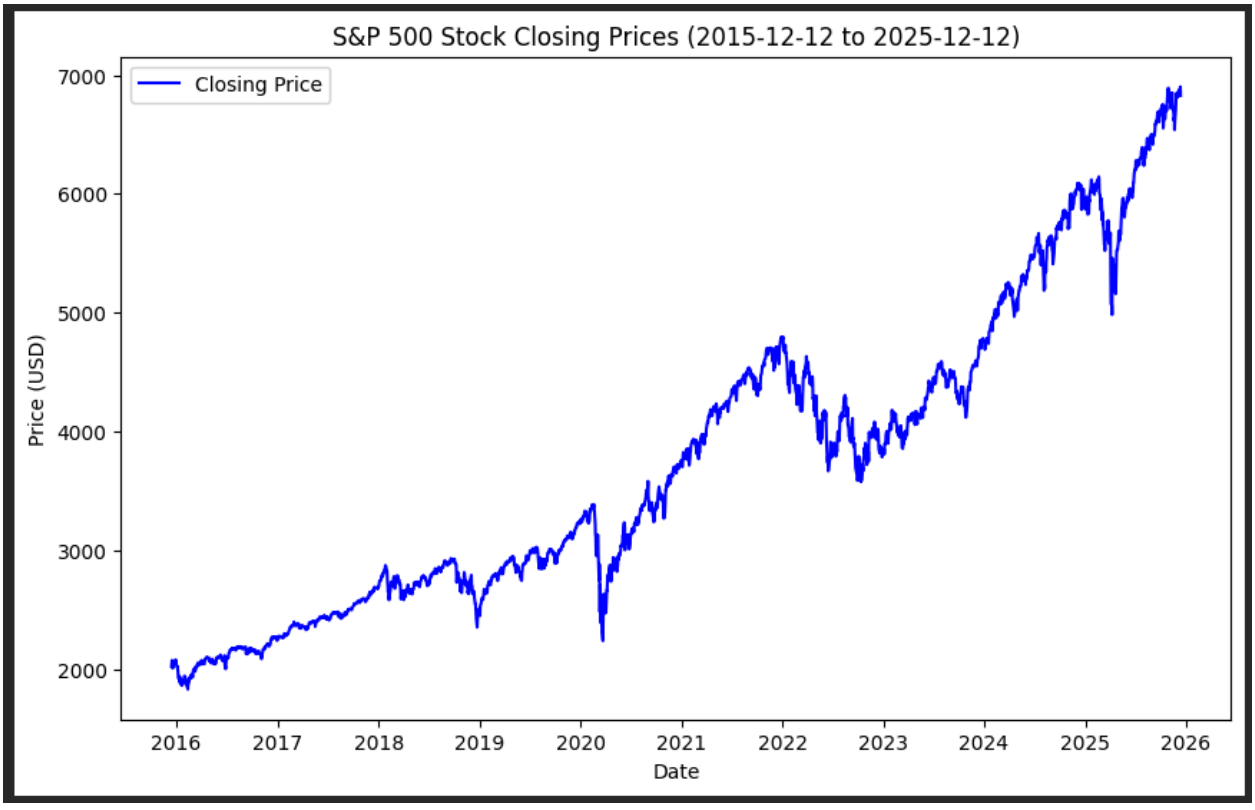
Alvina Lin

Data Bootcamp

Final Project: S&P 500

Goal

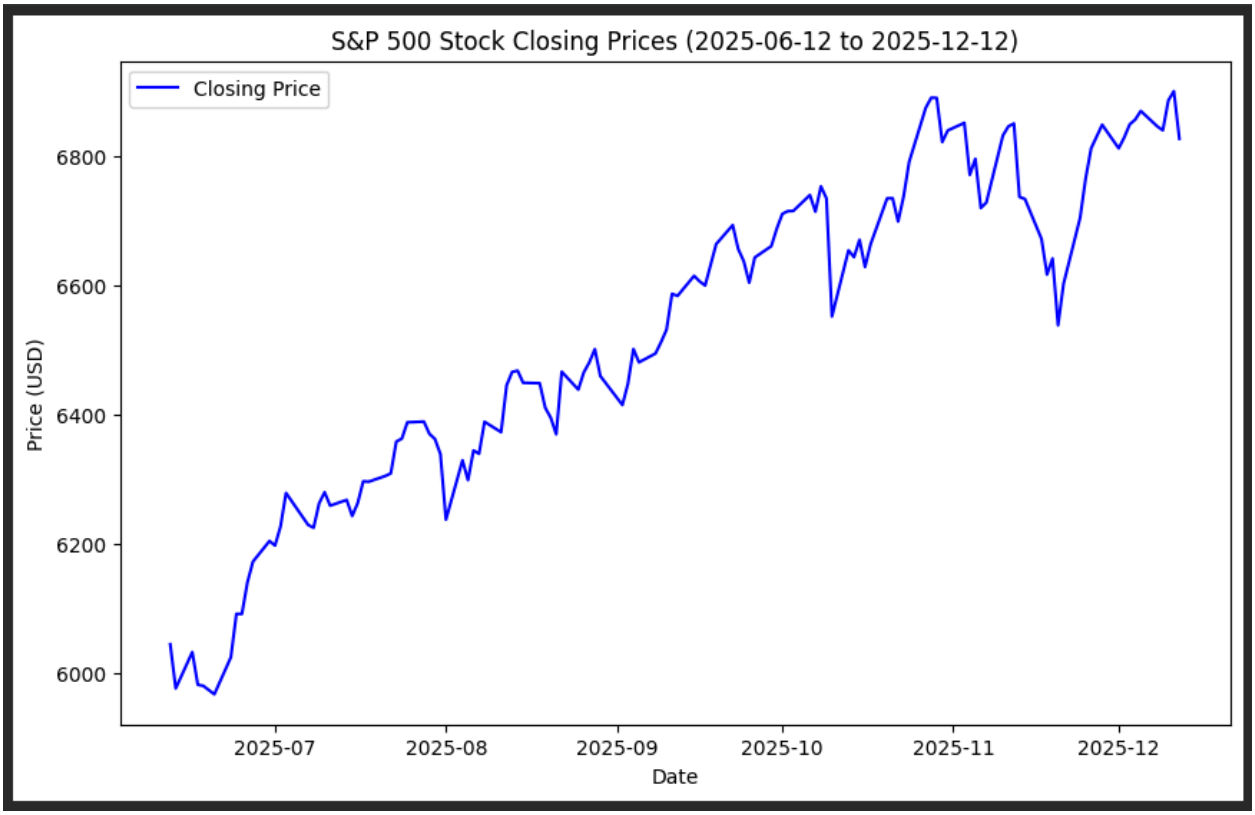
For the final project, my goal is to predict the S&P 500 closing price across two forecasting horizons: a short-term prediction 15 days into the future and a long term prediction 1 year into the future. This requires two different modeling strategies. For initial baselines, I first developed two models: a linear regression model for the 15 day forecast and a polynomial regression model for the longer 1 year forecast. Both models use only the days/time as the primary independent variable. Next, recognizing the limitations of simple time based modeling, I made a sequential model using multiple variables (Days/Time, Open, High, Low, Average Percent Change, and Moving Average Convergence Divergence (MACD) signal) to predict 15 days into the future. These models are based on data from the Yahoo Finance API and are not set to a specific date



but rather the latest market date. For this paper, the forecasting models were run on Dec. 12, 2025 (the latest date).

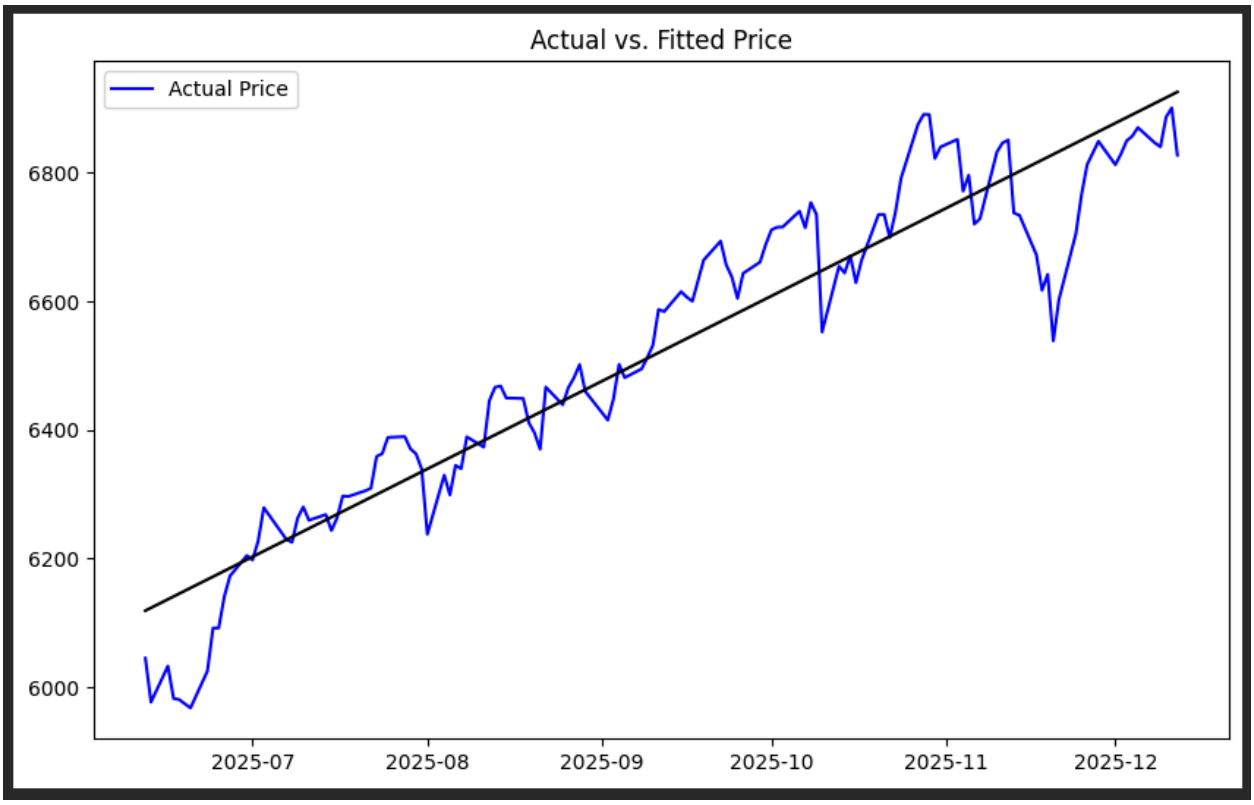
15 day forecast (Linear Model)

For the short-term 15 day forecast, I first decided to create a linear model. My linear model was created using only the most recent past 6 months of data. This is because since 15 days later is only a short time away, recent changes and patterns in closing prices are a lot more impactful on the market's current trajectory. Including data from further away in the past would have significant noise and historical breaks, introducing too much complexity that would actually decrease the accuracy of the 15 day forecast. Older information would dilute the linear relationship established by the immediate trend, decreasing precision and relevance of the short-term forecast. Therefore, the linear model serves as a strong baseline that captures the current trend.

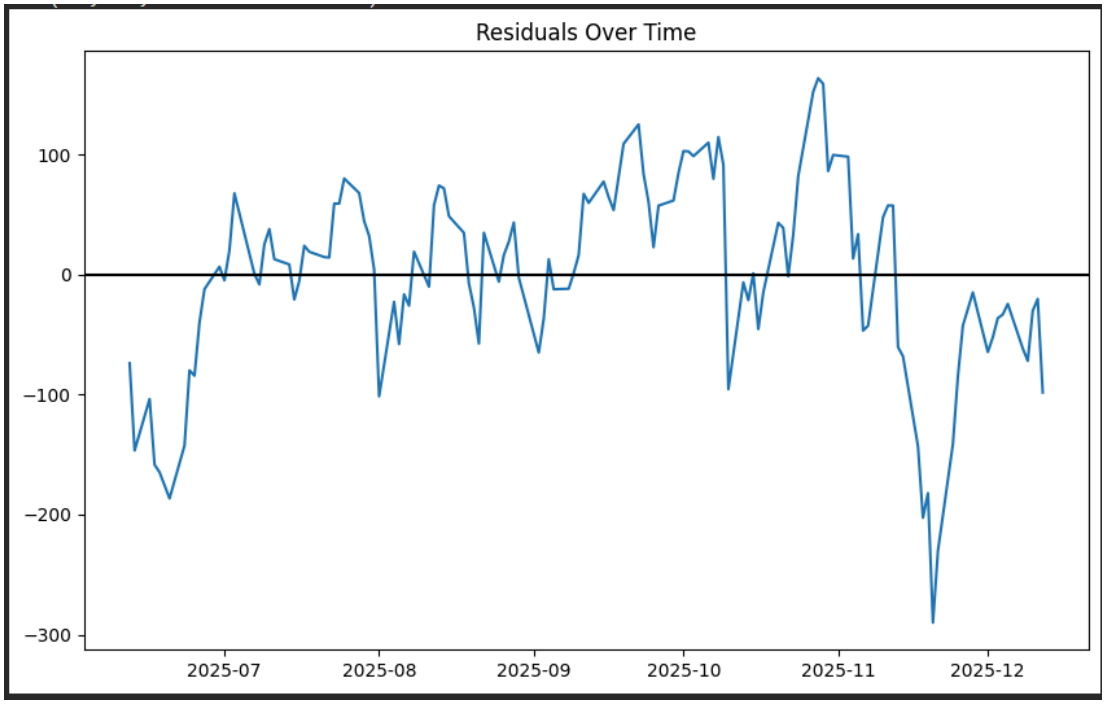


The visualization of the S&P closing price over the last 6 months shows the upward trajectory that the initial model is designed to capture. This consistent upward trend reinforces the use of a simple linear function, y = mx +b, for the short term forecast, where x equals the number of days since the beginning of the used training data. After fitting the linear regression model, it produced the following linear equation.

Linear equation: y = 4.41x + 6119.21



Linear Model Analyzation and Performance:




```
Linear Model Performance (most recent 6 months)
R-squared (R²): 0.8935
Mean Squared Error (MSE): 6,469.20
Root Mean Squared Error (RMSE): 80.43
```

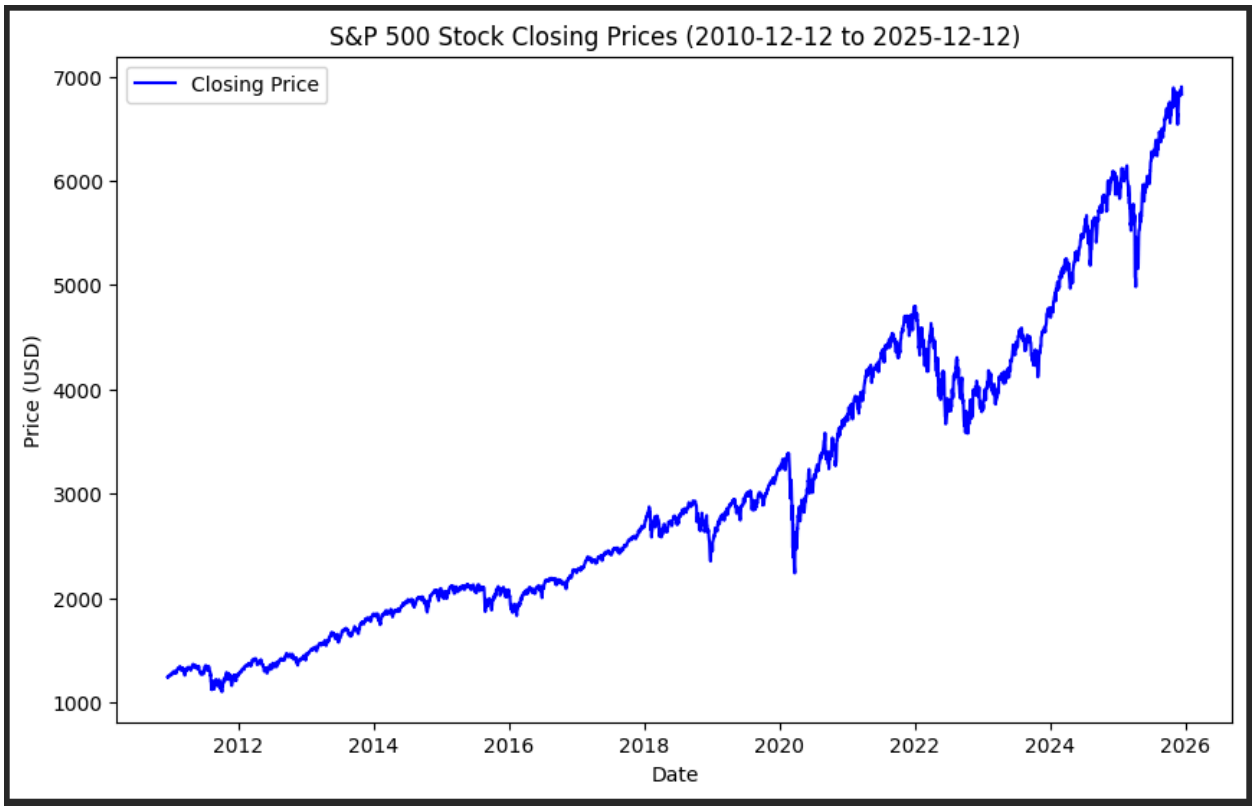
A R^2 of 0.8935 and RMSE of 80.43 shows a strong performance by the linear model. It means that on average the model was only off by 80.43 from the actual closing price value.

Using this model, it predicted a closing price of $6991.87 fifteen days later.

```
Latest Actual Closing Price on 2025-12-12: $6,827.41
Predicted Closing Price on 2025-12-27: $6,991.87
```

1 year forecast (Polynomial Model):

For the 1 year forecast, a different approach was required compared to the short-term linear model. Since the forecast horizon is significantly further away from the latest data point, short-term trends and volatility become irrelevant. Over such a long period of time, the overall market index tends to show non-linear, compounding growth, which appears more exponential. So to account for this pattern, I used a polynomial regression model, which was trained on data from the past 15 years. By fitting a polynomial curve, the model can capture the historical curvature and projected acceleration in the closing prices.

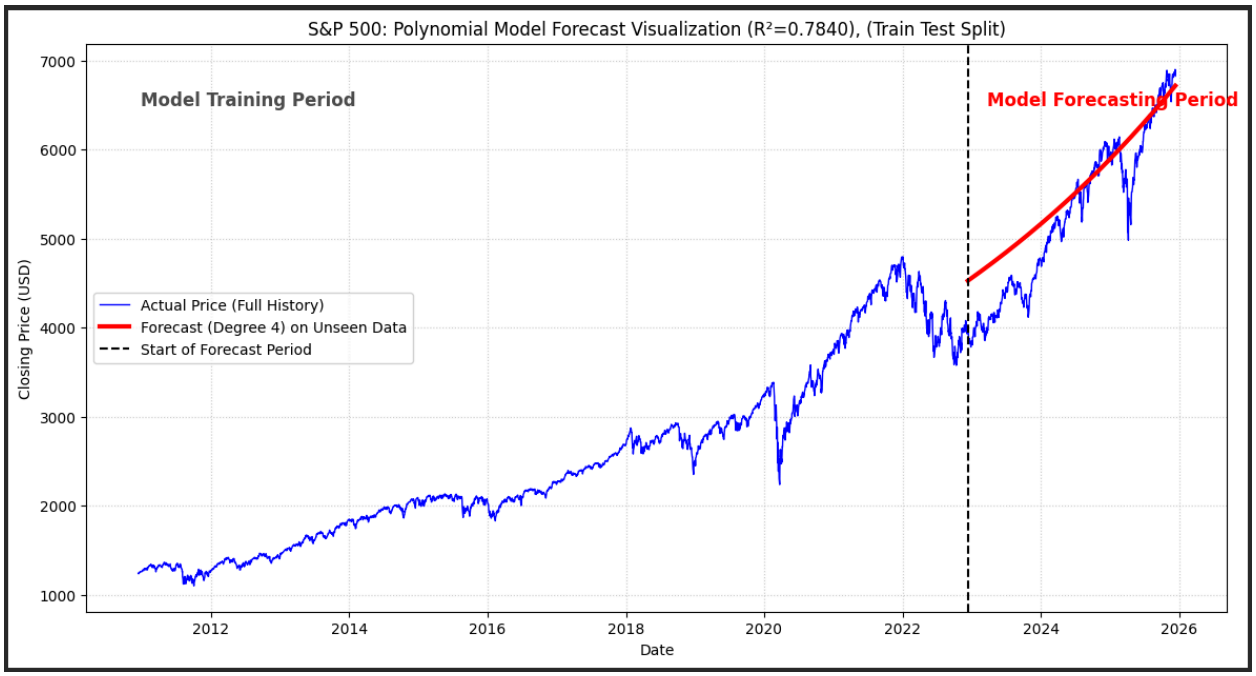


As you can see from the above graph, the closing price shows exponential growth over a long period of time.

Train Test Split:

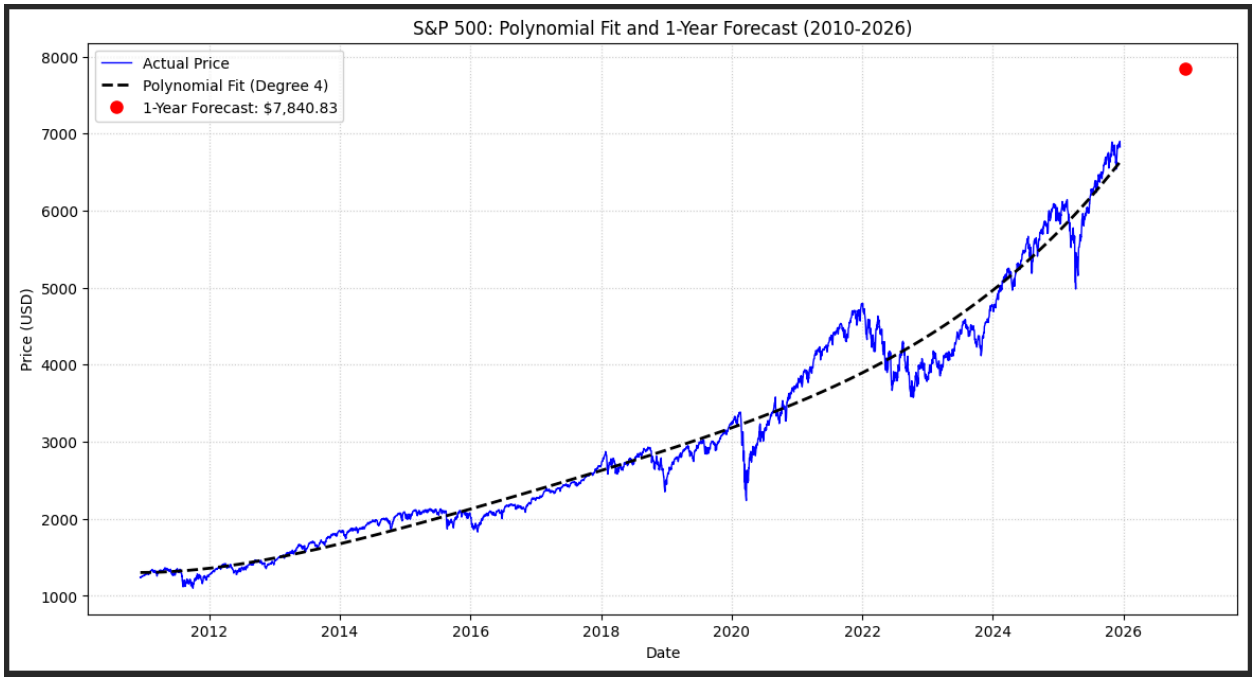
First, I used train-test-split to be able to test how well the model does on unseen data. For this, I made the first 12 years of the data the training data, and the last 3 years the testing data. After producing the model and testing it on the last 3 years, the following is the performance metrics on the unseen data.

```
Polynomial Model Performance (on Unseen Testing Data: the past 3 yrs)
---------------------------------------------------------------------
R-squared (R²): 0.7840
Mean Squared Error (MSE): 163,634.38
Root Mean Squared Error (RMSE): $404.52 USD
```



Now, to get the most accurate prediction 1 year later, I retrained the model on all the data (the most recent 15 years). From that I obtained the following results.

```
Final Polynomial Model Performance (on ALL 15 Years)
--------------------------------------------------
R-squared (R²): 0.9682
Mean Squared Error (MSE): 66,839.05
Root Mean Squared Error (RMSE): $258.53 USD
```



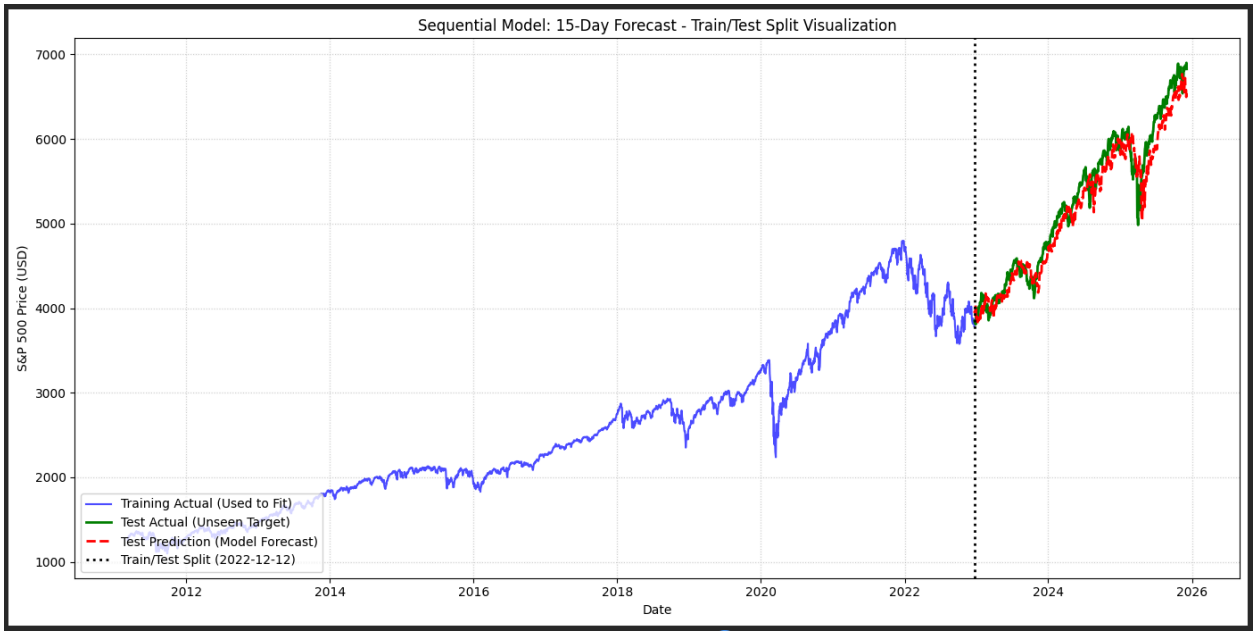
From the Polynomial model trained on the full past 15 years, it predicted the closing price to be 7840.83 one year later.

```
1-Year Polynomial Forecast
Prediction Date: 2026-12-12
Predicted Closing Price: $7,840.83
```

Sequential Model (15 day forecast)

Finally, I made a sequential model using data from the past 15 years. Instead of only using one variable (the days/time), this model uses multiple variables to hopefully create a more accurate prediction. The input variables for this sequential model include: the Days/Time variable, Opening Price, High and Low prices, Average Percent Change, and the Moving Average Convergence Divergence (MACD) signal. The MACD signal helps measure the momentum helping the model spot shifts in the stock's trend. Looking at all these variables together allows the model to look at both short and long term trends.

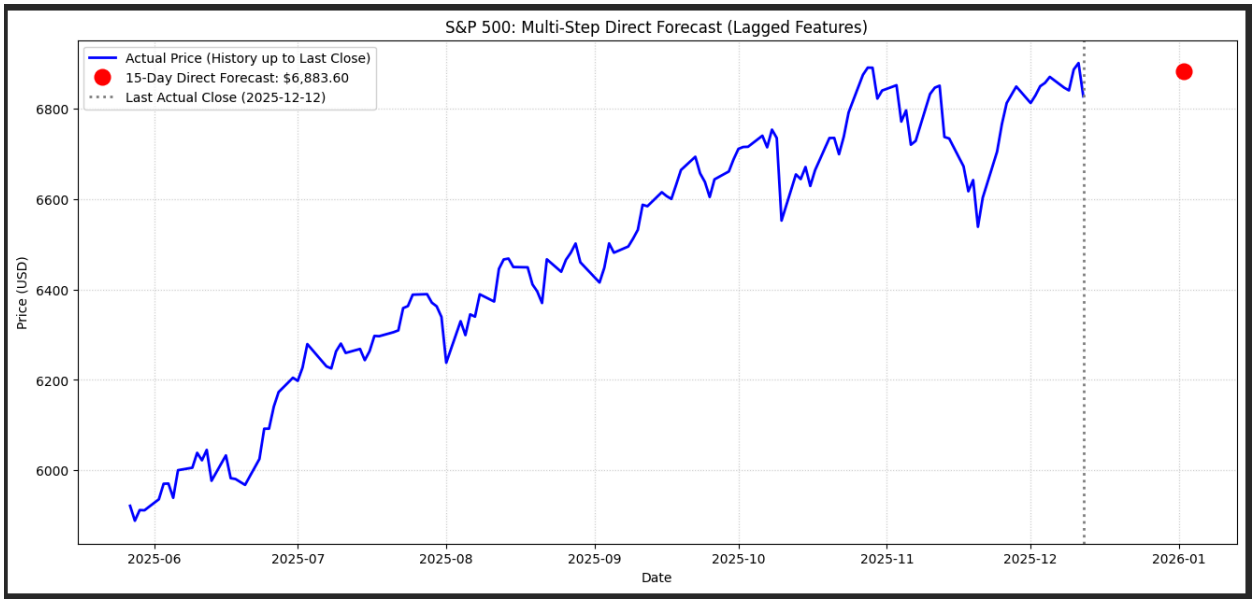
Train Test Split:




```
Sequential Model Performance Metrics (Train/Test Split)
--------------------------------------------------------
R-squared (R²): 0.9465
Mean Squared Error (MSE): 39,523.01
Root Mean Squared Error (RMSE): $198.80 USD
```

The model uses the everyday unseen data to continuously predict 15 days later everyday.

After the Train Test Split, the data was trained on all data producing the following results:



Based on this, the sequential model predicted a Closing Price of 6883.60 for 15 days later.

```
Sequential Model (multiple variables)
--------------------------------------------------------------
R-squared (R²): 0.9927
Mean Squared Eror (MSE): 15,195.17
Root Mean Squared Error (RMSE): $123.27 USD
Predictions:
Latest Actual Closing Price (2025-12-12): $6,827.41
Final 15-Day Forecast Price (2026-01-02): $6,883.60
```

Sequential Model Limitations

While the sequential model proves highly effective for the short-term 15-day forecast, relying heavily on momentum and detailed daily price movements, it would be a poor choice for a 1-year forecast. This is due to its features. Indicators like the Moving Average Convergence Divergence (MACD) and Average Percent Change are designed to capture short-term market noise, speed, and immediate changes in direction. When these features are used to predict one

year into the future, their relevance quickly diminishes. The noise from short-term daily changes would overwhelm any genuine long-term signal. Even if we were to remove these momentum variables, and only use the simple variables: Open, High, Low, and Days, the model would still lack the context to project long-term economic growth, relying too much on recent data. A reliable 1-year forecast requires capturing the structural, long-term economic growth, which is why the polynomial model, which only relies on the general compounding effect of time, is a much more appropriate baseline for a one year forecast.

Conclusion

Prediction date: 15 days later (12/27/25)	**Linear Model (trained on past 6 months)**	**Sequential Model (trained on past 15 years)**
R-Squared	0.8935	0.9927
Mean Squared Error (MSE)	6469.20	15195.17
Root Mean Squared Error (RMSE)	80.43	123.27
Latest Closing Price (12/12/2025)	6827.41	6827.41
15 Day Forecast Prediction	6991.87	6883.60

Prediction date: 1 year later (12/12/26)	**Polynomial Model**
R-Squared	0.9682
Mean Squared Error (MSE)	66,839.05
Root Mean Squared Error (RMSE)	258.53
1 year Forecast Prediction	7,840.83

Looking at the data, for the short term 15 day forecast, models focusing on immediate market trends were superior. The Linear Model, trained on only the last 6 months to only look at the immediate trend, served as a baseline with an RMSE of $80.43 and a prediction of $6991.87. However, this was significantly improved by the Sequential Model, which used features like MACD and Average Percent Change to capture momentum, leading to a stronger fit (R-squared of 0.9927) despite the higher RMSE. By capturing short term momentum and detailed market structure, the sequential model has higher accuracy, generating a final 15 day prediction of $6883.60. For the 1 year forecast, the sequential approach was not suitable due to its reliance on short-term market noise and momentum indicators. Instead, a polynomial model was used, which was trained on the past 15 years of data to capture the compounding growth of the market. This model produced a final 1-year prediction of $7840.83.

Therefore, the best strategy when predicting S&P 500's closing price is to use a momentum based sequential model for short term predictions and a polynomial model for long term predictions. To make the forecasts even better, we could have made the sequential model using a Recurrent Neural Network, which is better at learning patterns in time-series data and should help reduce the current RMSE. For the long term Polynomial Model, we could add outside data such as GDP growth, which would help make the model more accurate.