Tel Aviv University

S&P 500: Time Series analysis

Roi Hezkiyahu

Roihezkiyahu@gmail.tau.ax.il

205884018

Abstract

This report presents a comprehensive time series analysis of the S&P 500's weekly closing prices from 1950 to 2020, using historical data sourced from Kaggle. The objective is to identify underlying trends, seasonal variations, and potential predictors that could influence future closing prices. The analysis includes a logarithmic transformation of the closing prices to simplify the interpretation of exponential growth trends and improve forecasting models. Seasonal decomposition reveals subtle yet consistent patterns aligning with established financial market strategies, notably the "Sell in May and go away" strategy. The findings suggest that while seasonality influences the index's fluctuations, it does not dominate the overall movement, which is primarily driven by long-term trends. This study enhances our understanding of the S&P 500's dynamics and provides a quantitative foundation for future economic and financial predictions.

1. Introduction

The S&P 500 (Standard & Poor's 500) is a market index representing the stock performance of 500 large companies on US stock exchanges. It's widely regarded as the best single gauge of large-cap U.S. equities and a key indicator of the health of the economy. This report is an analysis of the S&P 500's weekly closing prices from 1950 to 2020 taken from Kaggle [1], aiming to uncover trends, seasonal variations, and forecast future closing prices for the next year. In this work we will also include exogenous variables such as volume and elections related features to see their contribution to our predictions.

1.1 EDA

By looking at figure 1 we can see that the data has some exponential trend so we will use the Log close price which has a more linear trend (figure 2)

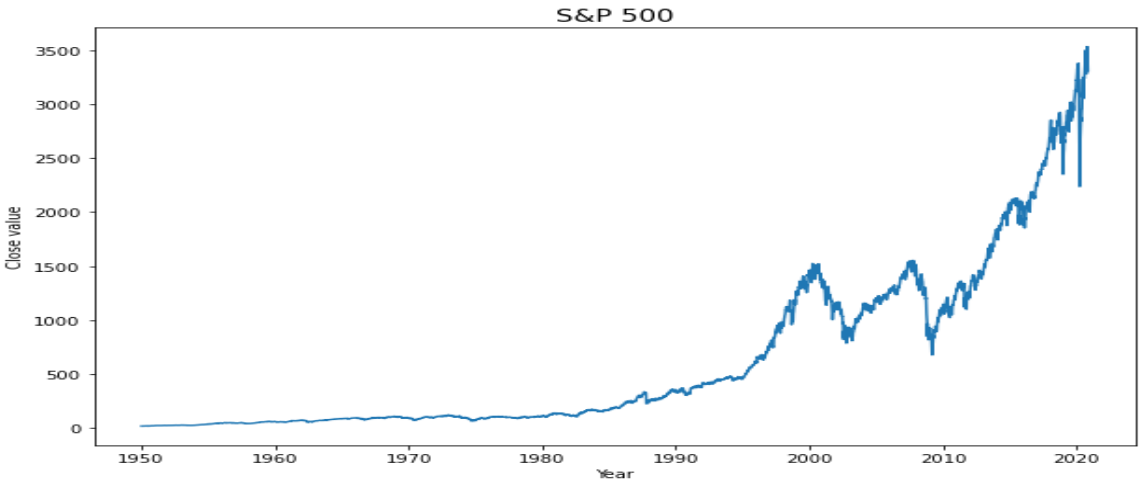

Figure 1: Close price of the S&P 500 stock index each week from 1950-2020

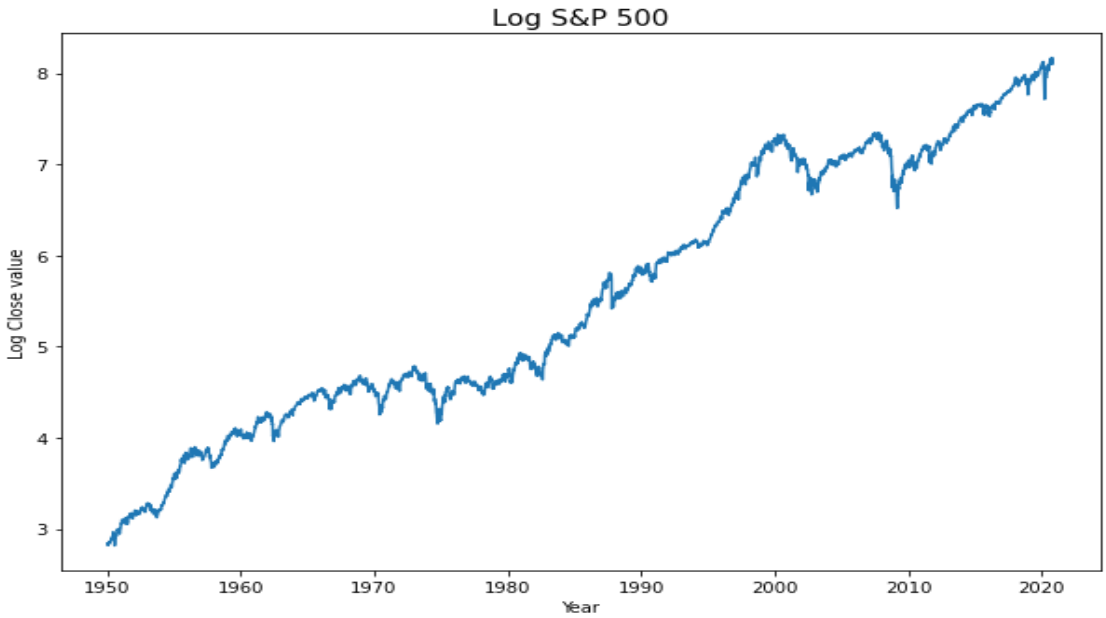

Figure 2: Log close price of the S&P 500 stock index each week from 1950-2020

By looking at the decomposition of the time series with a yearly period (figure 3) we can see that there is some seasonality over the year and that the general trend of the data is upwards with a high noise. The magnitude of the seasonality is very low (-0.02 – 0.01) which suggests that although its present it does not dominate the S&P 500 movements, the main component is the trend.

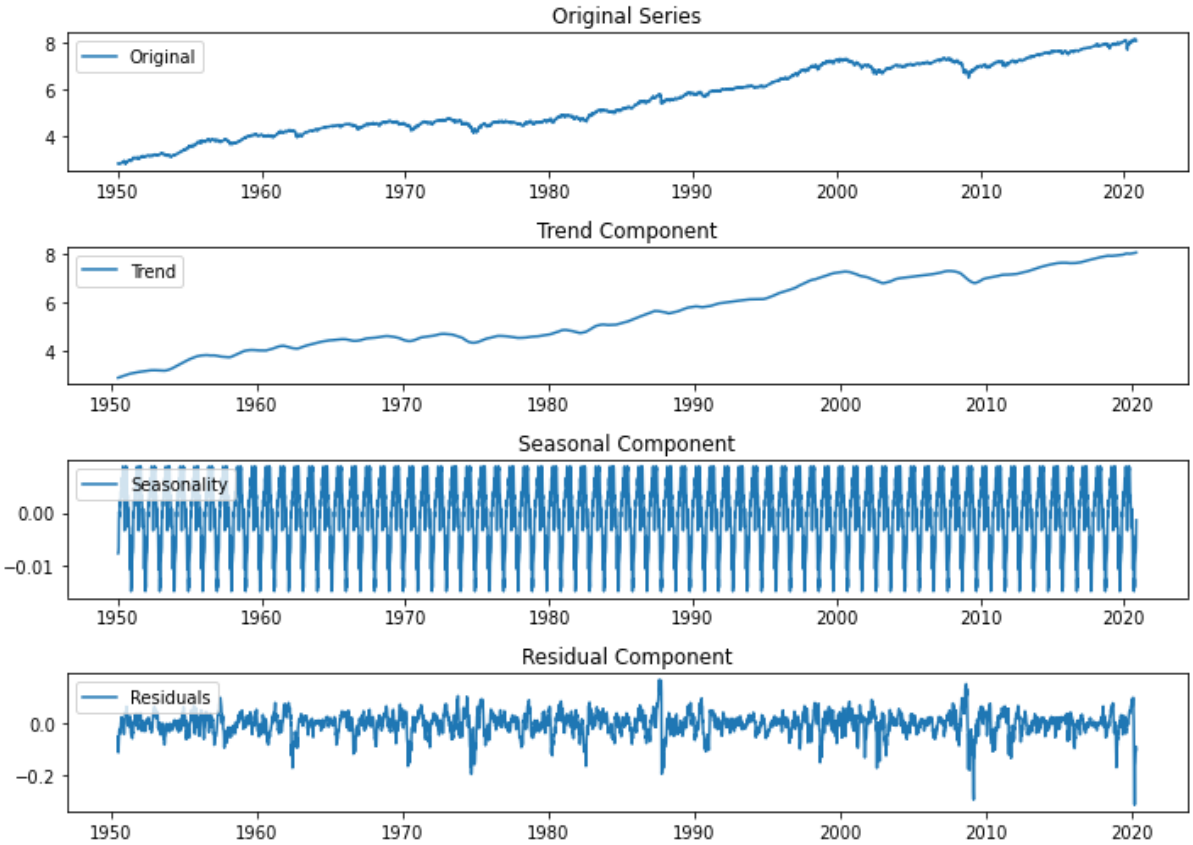

Figure 3: Break down of S&P Log Close price into Trend, Seasonality and Residual. with a period of 52 weeks.

Taking a closer look at the seasonality (Figure 4) we can see that the start of the year tends to have greater price than the end of the year, this correlated with the "Sell in May and go away" strategy [2] which claims that the stock market historically underperforms at the summer months May - October.

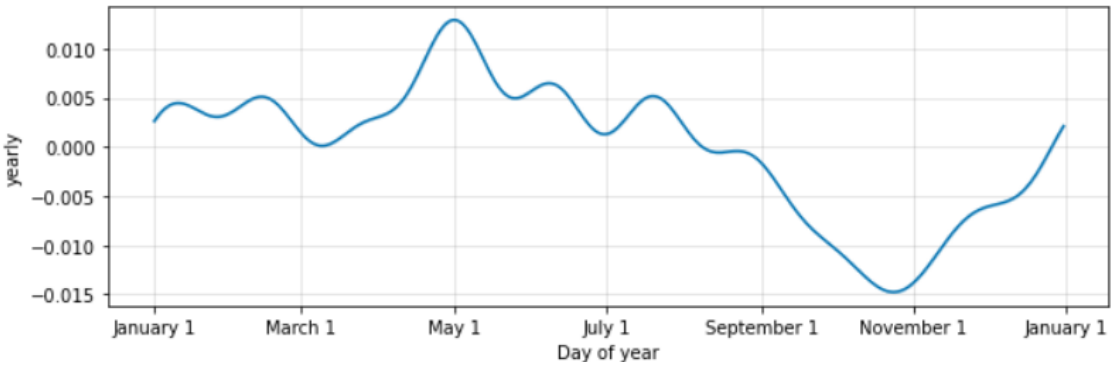

Figure 4: Monthly Seasonality

2. Methodology

To model the data, we used 2 methods: SARIMA and PROPHET, we also tried Deep LSTM which performed considerably worse so we will not discuss it in the report.

2.1 SARIMA

To choose the best SARIMA parameters we used 5-fold cross validation, checking the best MSE (predictive measure) and AIC (goodness of fit measure) across a grid of values: $p \in [1,2,3], d \in [0,1], q \in [0,1,2], P \in [0,1,2,3], D \in [1,2,3,4], Q \in [1,2,3,4], S = 4$.

Seasonality value was set to 4 because we can see some monthly seasonality according to Figure 4. Also, from Figure 5 we can see a strong PACF and ACF for past months differences.

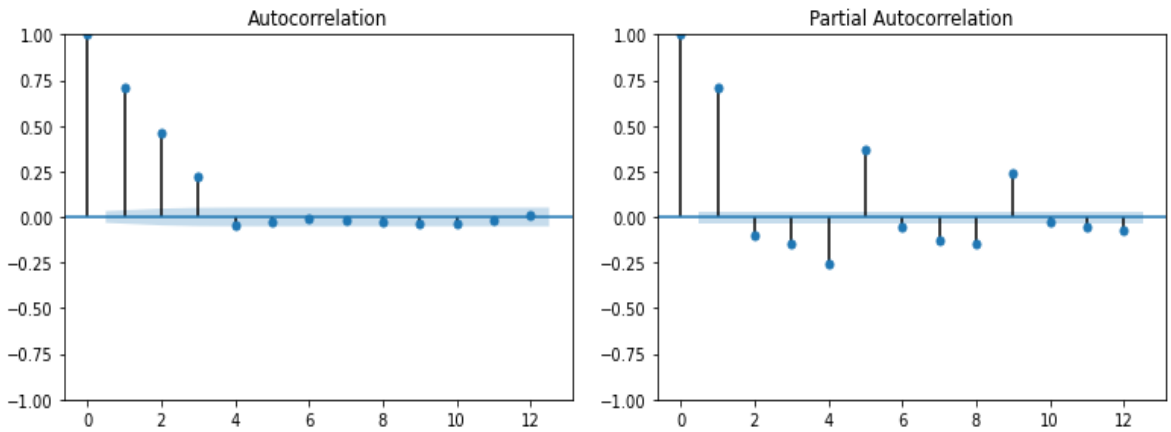

Figure 5: ACF and PACF of 4 weeks differentiated data.

2.2 PROPHET

The second model we chose to fit was prophet as it is known to perform well on time series data. Here we also used 5-fold cross validation, but because we can't measure AIC for PROPHET models, we only measured the MSE. The grid we chose was: $changepoint\ prior\ scale \in [0.01, 0.05, 0.1, 0.5, 1]$, $seasonality\ prior\ scale \in [0.005, 0.01, 0.1, 0.5, 1.0]$, $changepoint\ range: \in [0.8, 0.85, 0.9, 0.95]$.

2.3 Exogenous Variable

Out of the two methods SARIMA performed better, than we tried to improve it by inducing extra variables, we chose to include Volume which acts as kind of how certain the close price is, large volume means a higher certainty in the close price. Also, we include an indicator if the government party is Democrat or Republican, as well as a numeric variable for the number of years left until next election. There are some works suggesting the during election years the market performs better (11.28% average rise during election years compared to 10.22%), also that a Republican election is good for the market [3]

3. Results

3.1 SARIMA

order	Seasonal order	MSE	Average AIC
(3, 0, 0)	(0, 2, 4, 4)	0.00613	-9763.01
(1, 0, 1)	(0, 2, 4, 4)	0.006604	-9776.69
(2, 0, 0)	(0, 2, 4, 4)	0.0067	-9786.9
(2, 0, 0)	(1, 1, 2, 4)	0.010531	-9867.82
(1, 0, 1)	(1, 1, 2, 4)	0.011333	-9867.73
(1, 0, 1)	(0, 1, 1, 4)	0.011085	-9867.63

Table 1: Top 3 model MSE wise and top 3 models Average AIC wise

We are looking for the model with the best predictive power so we chose the following model SARIMAX(3, 0, 0)x(0, 2, 4, 4) with MSE of (0.006) and average AIC of (-9763).

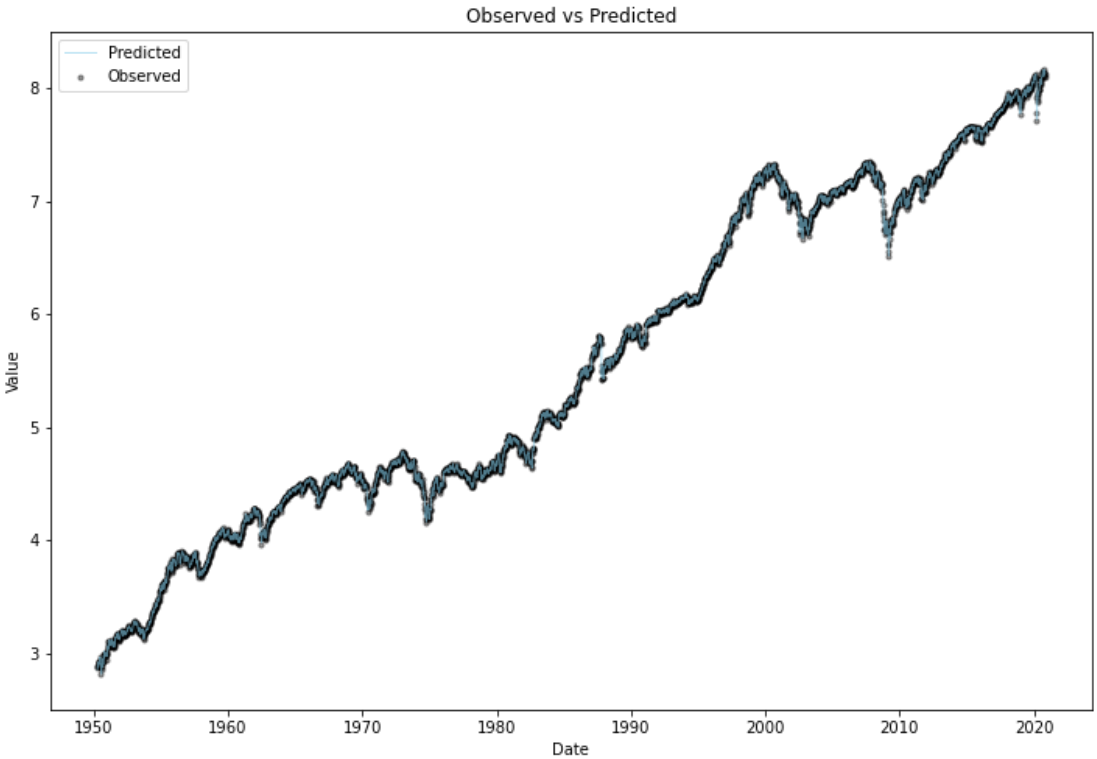

Figure6 : chosen SARIMA model predicted values compared to observed data.

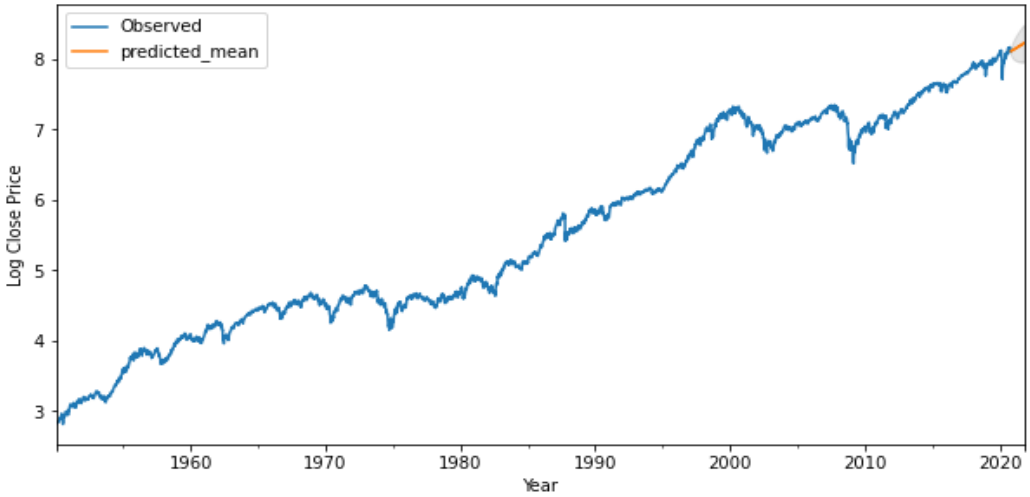

Figure 7: next year predictions based on chosen model.

3.2 PROPHET

The best prophet model with an MSE of 0.027 was: changepoint range = 0.85, seasonality prior scale = 0.01, changepoint prior scale=0.1. From Figure 8 we can see that the prophet model is kind of an exponential smoothing on the real data with some trend changes. The model averages the noise but does not learn the time series very well.

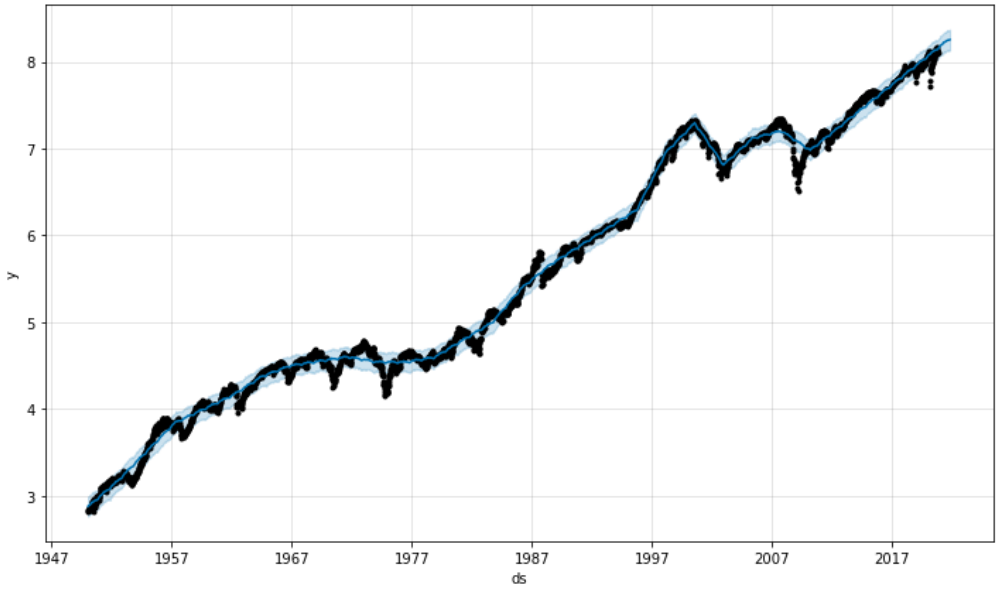

Figure 8: Prophet model predicted value compared to observed data

3.3 Exogenous Variable

We used 6 different exogenous variables: lag volume – the volume of last week close, election year – indicator if current year is an election year, years until next election: how many years are left until a reelection, democrat – indicator if the democrat party is in power. For the 2 indicators we used an interaction for the indicator and log close as well called election year log close and democrat log close.

Variable	MSE	Average AIC
Lag Volume	0.0123	-8825
Election year	0.008	-9772
Election year Log close	0.0087	-9762
Years Until Next US Election	0.0081	-9770
Democrat	0.008	-9775
Democrat Log close	0.0059	-9744

Democrat log close produces a model with a slightly lower MSE model than the original one, but impairs the AIC. The contribution is not significant.

4. **Conclusion**

This analysis of the S&P 500's weekly closing prices from 1950 to 2020 provides observation into the behavior of the index. The data reveals an upward trend, with minimal but consistent seasonal fluctuations that align with known market theory such as the "Sell in May and go away" effect. Despite the presence of seasonality, it does not significantly dictate the movements of the S&P 500. Through logarithmic transformation, we achieved a linear perspective that better facilitated our understanding of the index's long-term behavior. Predicting the S&P 500's future movements requires an understanding of both its historical tendencies and external economic factors, future works should try to include better exogenous variables, as ours did not perform very well.

Refrences

[1] https://www.kaggle.com/datasets/henryhan117/sp-500-historical-data/data

[2] Bouman, Sven, and Ben Jacobsen. 2002. "The Halloween Indicator, "Sell in May and Go Away": Another Puzzle ." American Economic Review, 92 (5): 1618-1635.DOI: 10.1257/000282802762024683

[3] https://advisor.morganstanley.com/the-ernie-garcia-group/documents/field/e/er/ernie-garcia-group/S%26P%20500%20in%20Presidential%20Election%20years.pdf