Comparative Study and Prediction of Dow Jones Industrial Average & S&P 500

Summary:

Stock Indices are important indicators of the financial market performances as they aggregate the price movements of the most prominent companies with the highest market capitalizations. They are the proxy metrics of economic prosperity and informs the corporate and individual investors on decisions in asset and portfolio management. This project aims to model the movement of the monthly S&P 500 and Dow Jones Industrial Average (DJIA) by conducting univariate modeling against their past time series data between 1992 and 2020 (inclusive), as well as investigate the effect of macro-economic variables through multivariate analysis of DJIA. Our team has developed 7 univariate models and 5 multivariate models to predict the stock indices, including the conventional time series models (ARIMA, GARCH, VAR, etc.) and new-found deep learning models (LSTM, Transformers). We extracted our indices data from Bloomberg, performed data cleaning and integration with macro-economic factors and exploratory data analysis in terms of stationarity and cross-correlation. We trained most of our models mostly on original time series and some in differenced time series due to the model assumption. Based on the goodness of fit metrics from residual analysis, and prediction evaluation metrics, we compared different model's performance, found that (1) multivariate model does not necessary outperform univariate model; (2) ARMA-GARCH and eGARCH have the best predictive power with less-than-one Precision Error (PM) despite the limitation in short-run flat predictions and those from ARIMA/SARIMA follows the movement and keep up with the changes ; and (3) Prophet and LSTM might be particularly good in dealing with predicting time series with shocks.

Table of Contents

1. Introduction 1
2. Background/Purpose..1
3. Methodology.. 2
3.1 Dataset and Preparation... 2
3.2 Univariate Analysis 5
3.3 Multivariate Analysis 8
3.4 Deep Learning for Time Series 10
4. Results and Discussion 12
5. Conclusions and Future Scope 15
6. References 16
7. Appendix 16

1. Introduction

Stock market performance is a crucial indicator of economic prosperity and is critical in ensuring corporations' performance and individuals' financial achievements [1,2]. However, as investors can share in the profits of larger companies, they are also exposed to varying degrees of risk. Market fluctuations can make or break a corporation or an individual.

Stock price time series are often characterized by a chaotic and non-linear behavior, making the forecast challenging [3]. The factor that produces uncertainty in this field is complex and of different nature, from economic, political, and investment decisions to unclear reasons that, somehow, have effects and make it hard to predict how the prices will evolve. The stock market attracts investments due to its ability to produce high revenues. However, owing to its risky nature, there is a need for an intelligent tool that minimizes risks and maximizes expected profits.

Predicting stock prices using historical data of the time series to estimate future values is the most common approach in the literature. More recently, researchers have started developing Deep Learning techniques that resemble biological and evolutionary processes to solve complex and non-linear problems [4]. This work also attempts to implement this new technique to the task of forecasting the market.The application of Deep Learning algorithms can be helpful in various financial problems. It has already been applied successfully in economic forecasting, trading strategies optimization, and financial modeling [4,5,6,7].

We are interested in exploring inclusion of macro-economic factors, such as CPI (Consumer Price Index), GDP (Gross Domestic Product), Interest rates and PPI (Producer Price Index). on top of the time-series data of the Dow Jones Industrial Average (DJIA) and S&P 500 indices, in predicting the magnitude and direction of the stock market's performance [8].

2. Background/ Purpose

Market conditions follow chaotic non-linear, volatile, and complex patterns, and as such, reliable predictions are difficult to model. Specifically, the factors that most influence stock prices are complex to model. The main objective of the project is to examine and understand the market, and later forecast the movement of the market. We seek to model and predict the stock market indexes by analyzing its past data and other macro-economic indicators and proxies. Specifically, we will be examining DJIA 's and S&P 500's with their own time series and later modelling against DJIA with the addition of these factors. S&P 500 and DJIA are widely considered to be one of the most representative aggregate indicators of U.S. economic performance.

We expect that adding in the macro-economic factors to the model would improve forecasting accuracy and interpretability.

First, we start with Univariate Analysis and plan to analyze the time series of both indices by modelling against them with their own lagged values. ARIMA, GARCH and their variants will be used in this section. Second, to examine Co-integration, we attempt to understand the correlation of movement between DJIA and S&P 500 and seek to narrow our multivariate analysis to one index. It is followed by Multivariate analysis on DJIA with all the macro-economic factors. Last, LSTMs and other Deep learning techniques are applied for forecasting the time series. We will compare and discuss all models in Results and Discussions in the final part.

3. Methodology

3.1 Dataset and Preparation

3.1.1 Dataset Overview

We sourced the market data from Bloomberg, a widely used and accepted financial and economic source. For the reasons of consistency and uniformity, we depended entirely on a single data source. We began by acquiring the last 29 years of data (1992 to 2020) on the price of S&P 500 index (adjusted closing price) and Dow Jones Industrial Average index (adjusted closing price) and macro factor data which will have an impact on S&P 500 and Dow Jones Industrial Average. Monthly values dated back to 1992 were sourced across market-relevant variables. Macro factors acquired include Consumer Price Index (CPI), Producer Price Index (PPI), Real GDP, Industrial Production, Balance of Trade, M1 (Component of Money Supply), Housing Starts, Employment Report, and Treasury Yields. Table 1 gives a brief description of the time series we collected.

Series Name	Description	Last Known Value
S&P 500	S&P 500 is a stock market index tracking the performance of 500 large companies listed on stock exchanges in the United States.	4,175.20
DJIA	DJIA), Dow Jones, or simply the Dow, is a price-weighted measurement stock market index of 30 large companies listed on stock exchanges in the United States.	33,240.18
CPI	The Consumer Price Index (CPI) is a measure of the average change over time in the prices paid by urban consumers for a market basket of consumer goods and services.	1.2%
PPI	The Producer Price Index (PPI) program measures the average change over time in the selling prices received by domestic producers for their output.	1.4%
Real GDP	Real gross domestic product (GDP) is the inflation adjusted value of the goods and services produced by labor and property located in the United States.	$ 19.81 Trillion
Industrial Production	The Industrial Production Index is an economic indicator that measures real output for all facilities located in the United States manufacturing, mining, and electric, and gas utilities.	104.5853
Balance of Trade	The USA balance of trade measures the difference between the movement of merchandise trade leaving the USA (exports) and entering the USA (imports). This measure tracks the value of the merchandise trade balance.	- $ 89.18 Billion
M1	M1 is a money supply metric that consists of (1) currency outside the U.S. Treasury, Federal Reserve Banks, and the vaults of depository institutions; (2) demand deposits at commercial banks less cash items in the process of collection and Federal Reserve float; and (3) other checkable deposits.	$ 20,710.1 Billion
Housing Starts	Housing Starts are the total number of single-family houses that started construction in each month in the US.	1.79 Million
Employment Report	Employment Report also known as Total Nonfarm Payroll, is a measure of the number of U.S. workers in the economy that excludes proprietors, private household employees, unpaid volunteers, farm employees, and the unincorporated self-employed.	150.93 million
Treasury Yields	Market Yield on U.S. Treasury Securities at 10-Year Constant Maturity.	2.81 %

Table 1. Data Description

The above-mentioned macroeconomic factors impact the markets in different ways [9]. A bull market tends to raise yields as money moves from the relatively safer investments to riskier equities. However, if the inflationary pressures look up, investors tend to move back to bond markets and dump stocks. Long-term Treasury yields reflect the growth and inflation mix in the economy.

If growth is strong, there is a usual rise in bond yields. An increase in inflation increases the Treasury yields as well. However, the impact of these two factors is different for stocks. GDP is used by the Federal Open Market Committee (FOMC) as a gauge to make interest rate decisions. The gap between real GDP growth and the 10-year bond yield correlates well with stock prices. If we look at the Employment Report, it accounts for approximately 80 percent of the workers who contribute to GDP. Housing starts can be viewed as a forward-looking metric to gauge the outlook for an economy as well. This metric most notably crashed during the financial crisis of 2007-09. Housing starts reached as low as 478,000 in 2009 before beginning to recover. All in all, these macroeconomic variables are in theory closed related to the economy and stock indices.

3.1.2 Dataset Exploration and Interpretation

We then proceeded to visualize the market indices to inspect for obvious trends, patterns, or correlations. As expected, the scaled S&P 500 and Dow Jones Industrial Average closely mirrored each other's shape over time. S&P 500 and DJIA have been highly correlated in the last few decades and the correlation coefficient is found out to be 0.9933. This prompts us for further studying about co-integration (to be covered in section X). It is expected that similar factors will have similar effects on each of the S&P 500 and the Dow Jones Industrial Average. The plots of the scaled S&P 500 and DJIA are shown in Figure 1.

From Figure 1, we can see that there is an overall increasing trend but with several instances of lows. The method of splines can be used on the original time series used to obtain a non-linear trend estimation as shown in Figure 2 and Figure 3. As expected, the S&P 500 and DJIA have synchronized and almost overlapping trends.

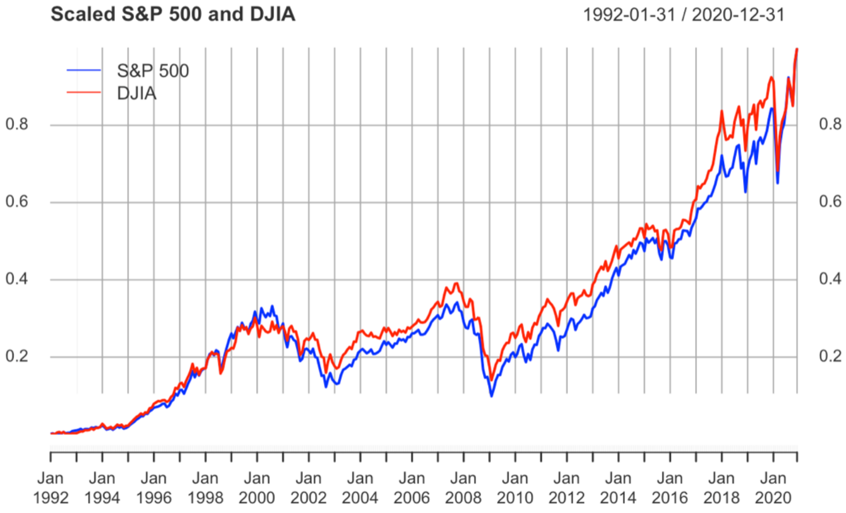


Figure 1. Time Series of scaled S&P and DJIA

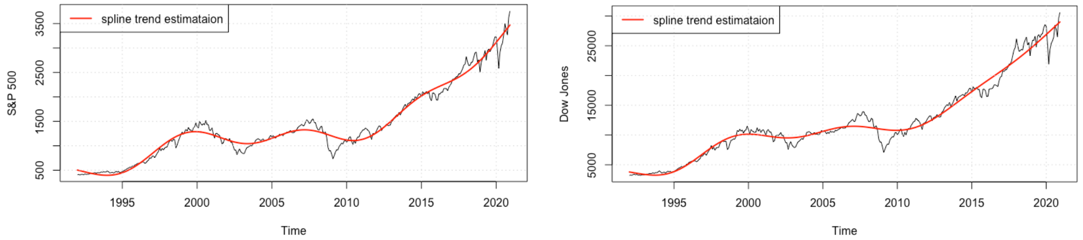


Figure 2 and Figure 3. Trend Estimation of S&P and DJIA

From figures 1, 2, and 3, we can see an upward trend in both the S&P 500 and DJIA since 1992. From 1992 to 2021, the financial markets in the USA have generated an overall return of approximately 8.46x. Investors have come a long way in the last two decades. It was defined by a booming U.S. economy, a bull market. The stock market marched steadily to new all-time highs through thick and thin. However, there were some significant interruptions along the way. These interruptions can be observed in the above figures, where there are dips in the market.

As observed in Figure 1, the first major dip was the Dot-com bubble in 2002, resulting in a crash that wiped out ~ $ 5.0 trillion from the US Markets in technology-firm market value between March and October 2002. The bubble was formed due to a surge of investments in the 1990s into anything related to the Internet and other technology stocks. During this period, several new firms came into being, most of which never generated any profit.

The next major crash, the Financial Crisis of 2007–08, came from the U.S. housing market collapse that ultimately led to the great recession. Over two years before the crisis, the Fed steadily raised the federal funds rate from 1.25% to 5.25%, which led to escalating numbers of subprime borrowers defaulting. When the resultant housing bubble finally burst, it created a domino effect that forced even large financial firms to liquidate investments in mortgage-backed securities. By mid-2009, the economy had finally begun to recover.

In 2015-16, there was a slight dip in the market due to market selloffs resulting from a series of global sell-off over an approximately one-year time frame beginning in June 2015. The market volatility initially began in China. Investors were selling shares globally amid a slew of tumultuous financial circumstances, including the end of quantitative easing in the U.S., a fall in petroleum prices, the Greek debt default, and the Brexit vote.

The next major and the most recent crash was in 2020, resulting from the outbreak of COVID-19 globally. It occurred due to panic selling following the onset of the pandemic. Airlines, cruise lines, and energy companies were particularly hard hit due to the crash because of travel restrictions countries implemented to limit the disease's spread. However, while the pandemic is still ongoing, the impact triggered by it didn't last that long. The stock market began to rebound, and by August 2020, the S&P 500 and DJIA were hitting record highs. The rapid recovery was due to the Fed, the Treasury Department, and Congress acting quickly to support the economy during the crisis by approving supplemental unemployment benefits and stimulus checks, cutting interest rates, and implementing new lending programs.

3.1.3 Checking for Stationarity

In an intuitive sense, stationarity means that the statistical properties of a process generating a time series do not change over time. To model and predict a time series, it is essential that it is stationary.

However, ARIMA and its variants, differences its data that may make the inputs stationary. In our scope, we look for our time series to be weakly stationarity. It requires the shift-invariance (in time) of the first moment and the cross moment (the auto-covariance). Thus, we expect a constant mean at all time points, and that the covariance between the values at any two time points, t and t−h, depend only on h.

The original plot of the S&P 500 and the DJIA show an obvious upward trend, which violates the stationary assumption of constant mean. Similarly, despite no observed seasonality, they have varying volatility across time, which may violate the assumption of constant variance. Their ACF plots (Appendix) have most values outside confidence bands after first few lags, which indicate severe serial correlation across time. In conclusion, neither the original S&P 500 nor DJIA are stationary.

Differencing two consecutive observations may be one way to make the data stationary. The series are thus, differenced with lag 1 and ACF plot of both differenced series are analysed. We see that the ACF plot does not have any significant lag (apart from 0) outside the confidence band. Also, from our analysis, we see that the differenced data seems to have a constant mean, but the recent few years seems to have non-constant and increasing variance. However, from visual inspection of ACF plots, we can consider the differenced data to be weakly stationary.

3.2 Univariate Analysis

3.2.1 ARIMA

ARIMA is the most common modelling approach found from our literature survey that forecasted the market. It is a generalized model of Autoregressive Moving Average (ARMA) that combines Autoregressive (AR) process and Moving Average (MA) processes and builds a composite model of the time series. It identifies the dependencies between an observation and its past observations, by measuring the differences of observations at various times. It also accounts for the dependency between observations and their residual errors. Let ‘t’ be the time point of the series X_t (DJIA or S&P 500). ARIMA (p, d, q) model can then be defined as:
$(1-B)^{d} X_t = Y_t$ and $\phi(B) Y_t = \theta(B) Z_t$, where autoregressive polynomial $\phi(z)=1-\phi_1 z- \phi_2 z^2 .. -\phi_p z^p$ and moving average polynomial $\theta(z)= 1- \theta_1 z- \theta_2 z^2 .... - \theta_q z^q$

Model: We preformed ARIMA on the actual train dataset (until 2017-December), as the differenced time series could be weakly stationary as discussed 3.1.3. AIC values for each combination of (p,d,q) model were checked, with maximum p and q equal to 10 and maximum d equal to 2. By choosing the combination of order with minimum AIC values, we select model ARIMA (0, 2, 1) for S&P 500 with minimum AIC value 3294.800, and ARIMA (2, 2, 3) for DJIA with AIC value 4633.889.

Residuals: As for the ARIMA model fitted into S&P 500, we can observe residuals with constant mean yet varying variance from the Residual Plot (Appendix). Yet, it arguably still upholds the assumption of finite variance. The ACF and PACF indicate no serial correlation. We may conclude that the residuals are plausibly stationary. With reference to the histogram, residuals are roughly normally distributed. The tails in the Q-Q plot are light and does not show sign of serious violation of normality assumption. The large p-values in Box-Ljing test and Box-Pierce test fails to reject the null hypothesis of uncorrelated residuals, so the residuals are plausibly uncorrelated. As for the

ARIMA model fitted into DJIA, we can observe similar results to the case of S&P 500. The residuals are weakly stationary and follow normal distribution. The hypothesis testing also fails to reject null hypothesis of uncorrelated residuals. In general, the residual analysis shows that the ARIMA model is a good fit to both DJIA and S&P 500 as the residuals follow stationarity and normality assumption.

Predictions: The prediction of S&P 500 from ARIMA model is good, given a less-then-one PM error in general. It shows that the residual variance is only 75% of the testing data variance. The prediction of DJIA is poorer, with a larger-than-one PM error. It is not surprising to see that the prediction accuracy of both S&P 500 and DJIA deteriorates after pandemic outbreak, with increase across all losses and PM.

3.2.2 SARIMA

Adding on to ARIMA, we would like to estimate the seasonal(S) ARIMA model, which incorporates both non-seasonal and seasonal factors. The general form of a seasonal ARIMA model is denoted as $ARIMA(p, d, q)$ X (P,D, Q) S,where p is the non-seasonal AR order, d is the non-seasonal differencing, q is the non-seasonal MA order, P is the seasonal AR order, D is the seasonal differencing, Q is the seasonal MA order, and S is the time span of repeating seasonal pattern, respectively.

Model: For both S&P 500 and DJIA, we have used the ARIMA order obtained from the ARIMA analysis performed in the previous section. Then we tested different combinations of SARIMA (P, D, Q) orders, with maximum P and Q equal to 2 and D equal to 1. By choosing the combination of orders with minimum AIC values, we select SARIMA models with seasonality factors of (1, 1, 1) for S&P 500 with minimum AIC value 3213.954, and seasonality factors of (1, 1, 2) for DJIA with AIC value 4492.507.

Residuals: From the SARIMA model fitted into S&P 500, we can observe similar results to the case of ARIMA of S&P 500. The residuals are weakly stationary and follow normal distribution skewed to the left. Here, the hypothesis testing fails to reject null hypothesis of uncorrelated residuals. As for the SARIMA model fitted into DJIA, we can observe similar results like before. The residuals are normally distributed. But, because of smaller p-values we reject the null hypothesis of uncorrelated residuals. SARIMA model is a good fit to both DJIA and S&P 500 as the residuals follow stationarity and normality assumption.

Predictions: The prediction of S&P 500 and DJIA from the SARIMA model follows a similar pattern to ARIMA. Prediction for S&P 500 is accurate, and it can be observed from its small PM error, where the residual variance is only 77% of the testing data variance. The forecast for DJIA is poor in comparison to S&P 500, with a larger-than-one PM error. The prediction accuracy of both S&P 500 and DJIA deteriorated after the pandemic outbreak.

3.2.3 ARMA-GARCH

The floating component of a time series can be modelled using ARIMA, moving average etc, but in general to account for the volatile component in financial time series ARCH/GARCH is used. GARCH here helps to predict the conditional variance of the series.

Model: We adopted the 4-step approach by alternatively selecting the orders of (p, q) and (m, n) with minimum AIC or BIC. Eventually, we select ARMA (3, 3) +GARCH (1, 1) model for S&P 500 with BIC 10.38179 and ARMA(1, 0)+GARCH(1, 1) for DJIA with BIC 14.80085.

Residuals: The ACF plots of the residuals for the ARMA-GARCH model fitted into S&P 500 and DJIA show no serial correlation. The Q-Q plots indicate that they are following closely with the normal distribution. The hypothesis testing also fails to reject null hypothesis, leading to the conclusion of plausibly uncorrelated residuals. However, the small p-values in hypothesis testing against squared residuals reject the null hypothesis in the cases of both S&P 500 and DJIA. Their conditional variance of their residuals is still likely serially correlated, thus failing the ARCH test.

Predictions: The prediction from ARMA-GARCH is in generally better than ARIMA models if we only consider the evaluation metrics. The PM of prediction from ARMA-GARCH is only 0.51 for S&P 500 and 0.82 for DJIA, which significantly smaller than one. It indicates the relatively small variance of residuals compared to testing data variance. Again, pandemic outbreak does weaken the prediction power of the ARMA-GARCH model but the deterioration in performance is not as severe as ARIMA model. However, if we look at the plot of the prediction from ARMA-GARCH, we can observe almost an upward linear prediction. It can be explained by the fact that ARMA-GARCH also model against variance to fit the data on top of mean estimation, and the above predictions only utilize the mean prediction of the models. To conclude, the ARMA-GARCH model that also model against variance has a better predictive power than ARIMA model, but the predictions are steady and stalbe, giving less information about the short-run variations.

3.2.4 EGARCH

Adding to GARCH, we implemented an EGARCH model. The formulation of EGARCH allows the sign and magnitude of the shocks to affect the volatility differently. EGARCH allows the variance to have asymmetric behavior to reflect any asymmetries in the data.

Model: For both S&P 500 and DJIA, we selected the ARMA-GARCH orders from the above ARMA-GARCH analysis and used the same for eGARCH. So, we used ARMA(3, 3)+GARCH(1, 1) model for S&P 500 and ARMA(1, 0)+GARCH(1, 1) for DJIA.

Residuals: The ACF plots of the residuals for the eGARCH model fitted into S&P 500 and DJIA show no serial correlation. The Q-Q plots indicate that they are following closely with the normal distribution. Like ARMA-GARCH, the conditional variance of the residuals is still likely serially correlated, thus failing the ARCH test.

Predictions: On prediction on the test data, we find that eGARCH was better than ARIMA and SARIMA models. The PM of forecast from eGARCH is only 0.76 for S&P 500 and 0.85 for DJIA, which is smaller than one indicating the relatively small variance of residuals compared to testing data variance. After the pandemic outbreak, the prediction power of the eGARCH model deteriorated but was not as severe as ARIMA and SARIMA models. But similar to ARMA-GARCH model, the predictions here are steady and stable with less prediction abut short-run variation.

3.2.5 Prophet Model: Forecasting at Scale

The Forecasting at Scale[10] is a novel approach for analysis of business time series. The forecasting tool, Prophet Forecasting Model, is based on an additive model which is capable of considering nonlinear trend, seasonality using standard Fourier series as well as holidays and events. It is mentioned that this model works very well with the strong seasonal effects and is robust to outliers and missing data in the model. In general, the model is represented as $y(t) = g(t) + s(t) + h(t) + \varepsilon$ where $g(t)$, $s(t)$ and $h(t)$ are representing trend function, periodic changes or seasonality and effect of holidays with irregular schedule respectively. Following definitions are showing the general formulas that were used to define each of the components.
For the trend part, a logistic model is used as follows,

$$g(t) = \frac{C}{1 + exp(-k(t-m))}$$

Where C , k and m are representing carrying capacity, growth rate and offset parameter respectively. The proper values for these parameters are computed through algorithm. The periodic component is computed using the standard Fourier series as follows,

$$s(t) = \sum_{n=1}^{N} \left(a_n \cos\left(\frac{2nt}{P}\right) + b_n \sin\left(\frac{2nt}{P}\right)\right)$$

Where P accounts for the period for time series. Given the length of the time series, N, we can construct matrix of seasonality vectors using the Fourier basis as

$$X(t) = \left[\cos\left(\frac{2nt}{P}\right), \sin\left(\frac{2nt}{P}\right)\right]_{n=1}^{N}$$

And hence the seasonal component is defined as $s(t) = X(t)\beta$ where β is the parameter for smoothing prior on the seasonality and is estimated using AIC model selection criterion.
Finally, for the last component $h(t)$, we define D_i as the set of past and future dates for the holiday i. Then the following matrix can be defined:

$$Z(t) = [1(t \in D_1), \cdots, 1(t \in D_l)]$$

and finally, $h(t) = Z(t)k$ and k is the set of parameters that will be estimated.

Residuals and Predictions: We used this model for univariate analysis for both S&P 500 and DJIA. The residual analysis also implies that there is no serial correlation. However, the residuals seem to be non-stationary from the ACF plot. Forecasting results shows that this model works well for univariate analysis, especially for the S&P 500. By adjusting the smoothness of the trend in coding, we could come up with an accurate trend estimation. Finally, using the software provided for this method, we did a change point detection analysis to understand the dynamics of the data. Based on the results, we get precision measure equal to 0.71 and 2.22 for S&P 500 and DJIA respectively.

3.3 Multivariate Analysis

Co-integration: Both DJIA and S&P 500 has been analyzed from a univariate perspective and the trend and movement of both were found to be similar to one another. To make such a statement, with statistical basis, we check for co-integration. We check if two series x_t and y_t are cointegrated, i.e., we check if there exists a parameter α such that $u_t = y_t - \alpha x_t$ is a stationary process. On performing the test, we notice that Test Statistic = -2.28 smaller than the 5% critical point (-1.95) but larger than the 1% critical point (-2.58). Hence, we reject the null of no- cointegration at 5% significance level. Therefore, we find that S&P 500 and DJIA are co-integrated.

We now, choose just one of the two market indices, i.e., DJIA to perform multivariate analysis using the macro-economic factors.

3.3.1 VAR Model

VAR model is one of the models that are the easiest to implement for multivariate time series. It is an extension of the univariate autoregressive model on multivariate time series. The basic p-lag vector autoregressive ($VAR(p)$) model for a vector of endogenous variables Y_t can be formulated as $Y_t = c + \Pi_1 Y_{t-1} + \Pi_2 Y_{t-2} + \ldots + \Pi_p Y_{t-p} + \epsilon_t$, $\epsilon_t \sim \mathrm{WN}(0, \Sigma)$ with each Π being a square matrix of the order of the number of series.

Model: As the implementation of VAR requires the input to be a stable multivariate time series, we use the differenced data as the input to the model. We applied the VAR model to the multivariate time series including all 10 economic indicators using the differenced training data. The AIC and BIC orders obtained were 3 and 1 respectively. As the selected order using AIC is larger than the selected order than selected using BIC, we apply the Wald test to evaluate whether a smaller order than the one selected with AIC would be a better choice for DJIA, meaning the smaller order model would perform similarly than the larger order model for the time series DJIA. From the wald test, we see that p-value > 0.05, i.e., we fail to reject the null hypothesis. Hence, a smaller model BIC would be sufficient for DJIA. Further, a stepwise model selection analysis was performed using stepwise regression (backward) to select for DJIA. The significant factors along with their lags are as observed in Table 2. The restricted VAR BIC model of DJIA simplifies to a bivariate model relationship of lag 1.

Model	Significant* Factors with Lag (of the total possible 10 x p combinations)
VAR (3) AIC Model	Industrial.Production.l1, Industrial.Production.l2, Housing.Starts.l1 ,M1.l3
VAR (1) BIC Model	Industrial.Production.l1
Stepwise Model	Industrial.Production.l1, Industrial.Production.l2,DJ.l2 out of the entire AIC model

Table 2. Significant Macro-economic factors * p-value<0.05

Granger Causality: We now use Granger Causality analysis using Wald test to evaluate whether any of the economic indicators lead DJIA. We see that only Balance of Trade granger causes DJIA and would thus, help in predicting or explaining DJIA for next months. However, this feature is not found to be significant in any of the models above.

Residuals: We performed goodness of fit for the above models using the multivariate ARCH test, the Jarque-Bera test and the Portmanteau test that led us to see that the VAR model does not have constant variance, nor satisfies the normality condition and has correlated errors. However, from the residual analysis for the DJIA component of VAR, we see that it is weakly stationary, satisfies normality condition and has uncorrelated errors.

Predictions: Both models of VAR and stepwise give similar predictions, for 3 months rolling forward. Like all the models, the prediction scores deteriorated after the pandemic outbreak. We see that PM ~1.1 for the overall test period and it reduces to <1 for the pre-pandemic period. However, most of the pre- pandemic predictions are flat with barely any changes captured.

3.3.2 ARIMAX Model

ARIMAX model is an extension of ARIMA which includes exogenous variable X as well. The model can be defined generally as follows,

$$Y_t = (\varphi_0 + \varphi_1 Y_{t-1} + \cdots + \varphi_p Y_{t-p}) + (Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}) + \beta X_t + \varepsilon_t$$

Where X_t stands for all exogenous factors that can impact the model. Implementing this model is similar to ARIMA model and the p , d and q orders based on a model selection approach like AIC. Values of parameters will also be estimated using MLE approach.

Models: For ARIMAX analysis, we tried three different models to observe the effect of different macro-economic factors on the analysis of target time series. We tried the following models:

- **Model 1:** For this model we include all exogenous factors that we have in our data set. We do not consider any lagged values of factors. Orders (2,1,2) are selected.
- **Model 2:** For this model we consider six macro-economic factor including CPI, Industrial Production, Balance of Trade, M1, Employment Report and Treasury Yields among all the possible factors with corresponding appropriate lags. We use outputs of VAR model as well as some economical knowledge background to select these sic factors and we are using the significant parameters from VAR output, we decided the proper lags we going to use for this model. The final selected order is (4,1,1).
- **Model 3:** For this model we only use one factor, Balance of Trade, with lag 1 for analysis. This factor is the significant factor based on the results from the Granger Causality analysis. Final order for this case is (3,1,3).

Residuals: The residual analysis for all the models above shows stationarity, normality, and uncorrelated residuals.

Predictions: All three models perform well based on the prediction measures. Among these three models, model 3 results in smaller precision measure. Also, from looking at plots, we can observe that model 3 provides more robust and reliable predictions. We can conclude that including only one economic factor, Balance of Trade, results in more precise predictions for our project.

3.4 Deep Learning for Time Series

Sequence Modelling using deep learning predict the next word in sentence by learning sentences as sequences of words. This architecture can be transferable to learn the time series data and predict the next current value. All models will train on past 24 time series values and predict next 3 future time series.

3.4.1. LSTM

Long Short-Term Memory (LSTM) model is a kind of recurrent neural network architecture, which continually process each data point in a sequence once at a time and update a hidden state to be used in processing the next-in-sequence data point. It also includes gates to regulate flow of information through long sequences.

Univariate Model: Due to long time in training and exponential number of combinations of multiple parameters in LSTM, we only fine-tune for values such as batch size and hidden size for certain values for lowest MAPE. LSTM's primary parameter – hidden size, known as the size for feature maps for each hidden RNN layer, is chosen to be 64 and batch size of training is chosen to be 16. Detailed configuration can be found in appendix.

Univariate Prediction: The prediction of S&P 500 is near perfection with MAPE of only 6.02% and Precision Measure as low as 0.57. The prediction is almost following the movement of test data. The prediction of DJIA is slightly poorer but still accurate. It results in a MAPE is only 5.87% and the prediction also follows the actual trend, despite over-amplification of the change magnitude. Notably its prediction's PM in post-covid period is smaller than pre-covid period.

3.4.2 Transformer

Transformer computes the multi-head attention which quantizes the intra-dependences within the input vector (past time series values) and output vector (predicted future time series values) and inter-dependence between input vectors and output vectors. The prediction of the next-in-sequence time series value will be calculated based on the attention scores and the input past time series.

Univariate Model: Upon limited grid search, we chose 64 as the size of expected features in the input and 4 attention heads. While the default number of heads for NLP is 8, it might be considered too complex for numerical time series data which is less complex than language token.

Univariate Prediction: Transformer's prediction of both S&P 500 and DJIA is poor, with almost constant prediction. As a result, it has the largest losses and Precision Measure among all models. We may conclude that transformer it not suitable for our time series prediction.

Multivariate Model: We have tried to increase the number of attention heads gradually to 16 due to higher complexity of multivariate analysis, but it does results in obvious in increase predicative performance. Again, due to limit in computing power and training time, we cannot exhaust all combinations of parameters to find the best model.

Multivariate Prediction: Same as univariate analysis, transformer's prediction is almost constant, despite intake of additional time series of macro-economic as a part of data input. The loss metrics and PM are still one of the largest among all models.

3.4.3 Temporal Fusion Transformers (LSTM + Transformer)

Temporal Fusion Transformer (TFT) model combine LSTM and multi-head attention queries from transformer in its basic architecture with some other innovations [11].

Multivariate Model: For the fair comparison, 16 attention heads are selected for transformer and 64 is selected to be the hidden size of LSTM.

Multivariate Prediction: TFT's prediction is better than pure transformer model as it is able to predict the big drop in covid outbreak roughly ahead of times. Yet, it cannot accurately capture the actual short-run movement and results in very large error metrics and PM as well.

4. Results and Discussion

We have analyzed and modelled the DJIA and S&P 500. We now wish to compare the performance across models and make observations from our modelling. **Note: Only Pre-Pandemic Error Measures shown. Complete table can be found in Appendix.*

Among Univariate Models: Table 4 and Figures 4-5 shows us that among univariate models, although ARMA-GARCH and eGARCH consistently yields the lowest PM, their prediction demonstrates a linear upward trend without predicting the imminent reaction to the current time and time-dependent fluctuations. Their use case will be limited to long-term indices forecasting but not short-term prediction. In contrast, ARIMA and SARIMA can predict somehow a short-term movement following the current momentum.

Between Univariate and Multivariate Models: As observed in the prediction forecasts from Figures 5-8 and error measures in Tables 4-7, contrary to our intuition, multivariate models (considering other macro-economic time series data) do not necessarily predict better univariate model. The minimum MAPE among multivariate model is 8.45% and the minimum PM is 0.9935, which is just below 1. On the other hand, several univariate models can achieve MAPE below 7% and PM below 0.90. While it can be considered from the results, that that these macro-economic factors do not have a lead relationship over stock indices, economic theory shows otherwise. From our modelling, several of the macro-economic factors seem to not help in forecasting and only add more noise the prediction of the model (during Pandemic). It could be that the inherent relationships between them are hard to capture. The stock indices may also have a lag relationship over other macro-economics, it deserves a separate study beyond our current report.

We see that the traditional models can output coefficients for different terms and time series, as well as their p-value for hypothesis testing of coefficient significance. They enable us to evaluate the how many lags of stock indices still affect its future values, and which macro-economic factors are contributing to prediction. In addition, we can evaluate the goodness of fit of the model by residuals analysis to verify the creditability of the above feature importance.

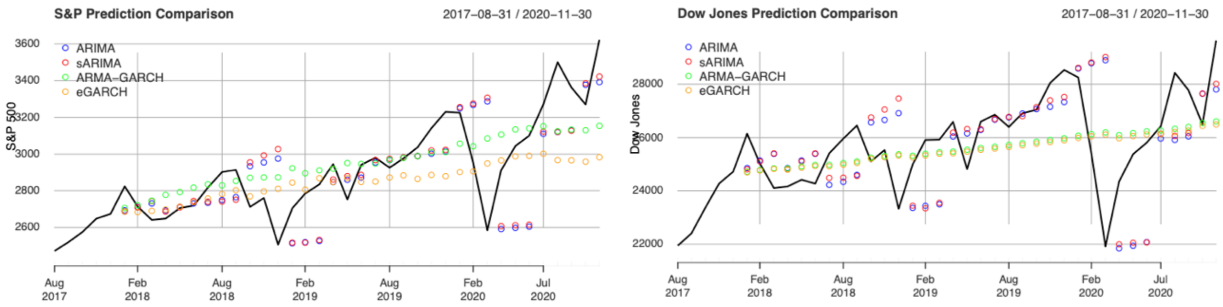


Figure 4 and 5. Univariate: S&P 500 & DJIA Prediction Comparison

Target Index	Model	Weakly Stationarity	Normality	Uncorrelated Residuals
S&P 500	ARIMA	Satisfied	Satisfied	Satisfied
	SARIMA	Satisfied	Satisfied	Satisfied
	ARMA-GARCH	Satisfied	Satisfied	Satisfied
	eGARCH	Satisfied	Satisfied	Satisfied
	Prophet model	Not Satisfied	Satisfied	NA

DJIA	ARIMA	Satisfied	Satisfied	Satisfied
	SARIMA	Satisfied	Satisfied	Not Satisfied
	ARMA-GARCH	Satisfied	Satisfied	Satisfied
	eGARCH	Satisfied	Satisfied	Satisfied
	Prophet model	Not Satisfied	Satisfied	NA

Table 3. Univariate Analysis: Goodness of Fit

Target Index	**Model**	**MSPE**	**MAE**	**MAPE**	**PM**	**PM Pandemic difference (post - pre)**
S&P 500	ARIMA	28549	127.9194	0.0458	1.0569	0.2736
	SARIMA	31733	132.3648	0.0475	1.1748	0.128
	ARMA-GARCH	17356	101.8394	0.0367	0.6426	0.304
	eGARCH	18884	105.9201	0.0368	0.6991	0.8591
	Prophet model	31439	157.25	0.057	2.5951	-1.6701
	LSTM	30887	142.9024	0.0507	1.0959	-0.245
	Transformer	137118	331.2657	0.1138	4.8649	0.7803
DJIA	ARIMA	2032968	1147.226	0.0452	1.2995	0.5082
	SARIMA	2223501	1164.096	0.0461	1.4213	0.3557
	ARMA-GARCH	1251858	935.8728	0.0359	0.8002	0.1245
	eGARCH	1324322	961.0194	0.0367	0.8465	0.1017
	Prophet model	746628	2520.3	0.0996	8.9334	-6.963
	LSTM	2355331	1172.0894	0.0465	1.4428	-0.2964
	Transformer	10960774	3019.0584	0.1149	6.7143	0

Table 4. Univariate Analysis: Prediction Results Pre-Pandemic

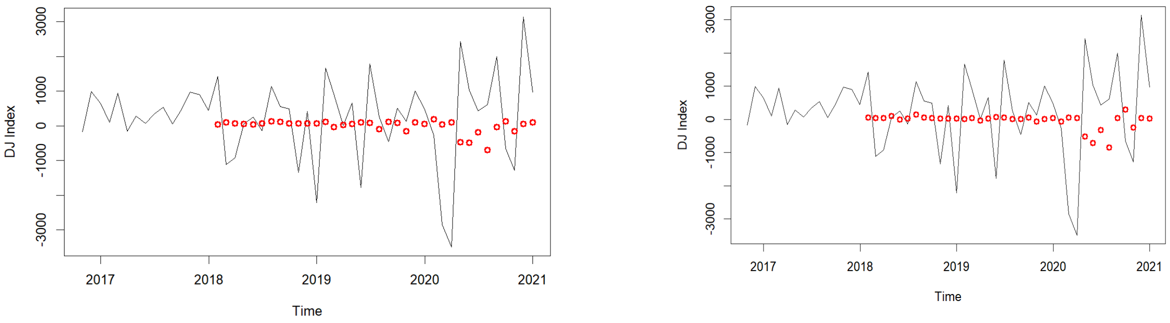


Figure 6 and 7. (Differenced) Multivariate: DJIA Prediction Unrestricted(left) and Restricted

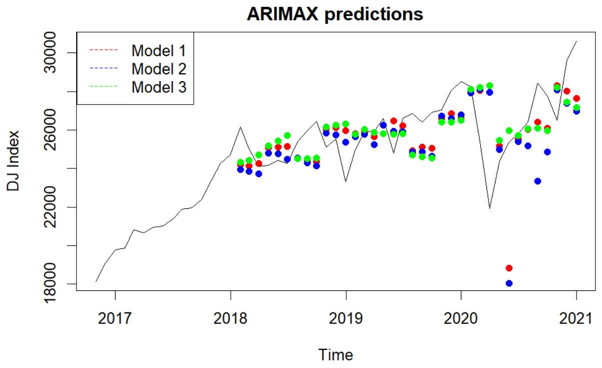


Figure 8. ARIMAX: DJIA Prediction

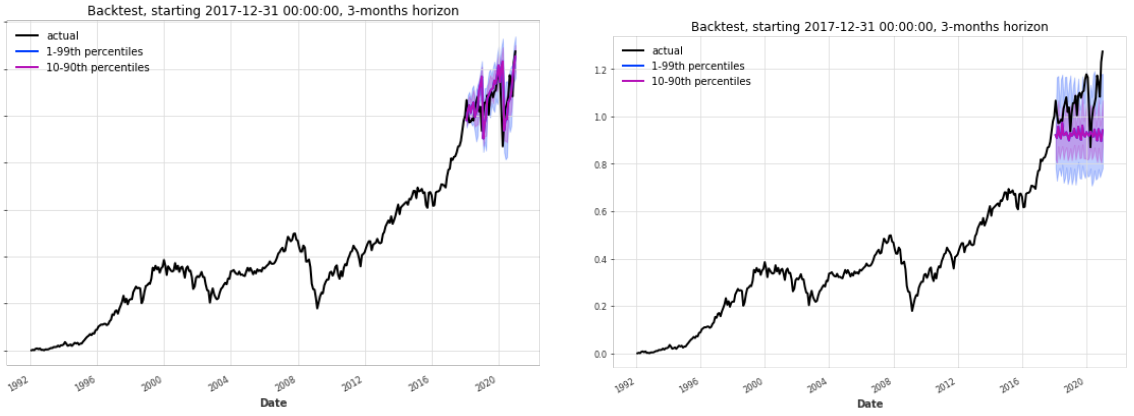


Figure 9 and 10. Univariate LSTM and Multivariate Transformer Prediction for DJIA

VAR -DJIA	Weakly Stationarity	Normality	Uncorrelated Residuals
VAR	Satisfied	Satisfied	Satisfied
Restr.VAR	Satisfied	Satisfied	Satisfied

Table 5. Multivariate Analysis: Goodness of Fit

ARIMAX Model	Weakly Stationarity	Normality	Uncorrelated Residuals
Model 1	Satisfied	Satisfied	Satisfied
Model 2	Satisfied	Satisfied	Satisfied
Model 3	Satisfied	Satisfied	Satisfied

Table 6. ARIMAX Analysis: Goodness of Fit

Model	MSPE	MAE	MAPE	PM	PM Pandemic difference (post - pre)
VAR	1100581	830.1811	0.9292	1.0311	0.1078
Rest. VAR	1060431	826.6203	0.9727	0.9935	0.1703
Transformer	8676462	2631.0065	0.0994	5.315	-2.0368
TFT	6429137	2119.706	0.0845	3.9383	0.4395
ARIMAX -1	1903184	1192.575	0.0476	2.2772	-0.6095
ARIMAX-2	1711852	1063.526	0.0422	2.0482	0.2414
ARIMAX-3	2111747	1293.074	0.0519	2.5267	-1.1892

Table 7. Multivariate Analysis: Prediction Results Pre-Pandemic

Between Traditional and DL models: We see that the traditional models' prediction is less accurate after a sudden shock, i.e. covid outbreak. However, univariate models of LSTM and Prophet, to the contrary, improve their performance after the shocks. It might be explained by the non-linearity of these models and able to capture more irregular movement in the data. The process of tuning the Deep Learning models is cumbersome and extremely time-consuming. Sequence modelling using Deep learning can be considered a black box. We also see that it is hard to obtain any understanding of how the variables are related in multivariate time models using transformers and obtain good prediction results with transformers (Plots-Appendix).

5. Conclusions and Future Scope

Investing in the market is very rewarding as well as risky. While economic crisis, pandemic and major events saw huge fluctuations in the market, there are some trends and fluctuations that can be captured to monetize. Forecasts of the market performance can help mitigate some of those risks by actively preparing for the changes and it can also generate capital through investments.

From our project, we see that traditional univariate analysis is sufficient and provides comparable if not better results than multivariate models. ARIMA model seems to generate satisfactory results in terms of both prediction accuracy and movement of forecasts. ARMA-GARCH and eGARCH have low PM however provide flat forecasts. The multivariate model VAR did not offer much information on the relationship between the macro-economic variables with only Balance of trade, granger-causing DJIA. Also, VAR is only able to look at their short-term effects (from orders less than 3). Modelling time series with these exogenous factors in ARIMAX, increases the model complexity and adds noise to the series. This was observed during the pandemic period prediction when several of the macro-economic variables became volatile. Hence, from traditional time series, ARIMA/SARIMA follows the movement and keeps up with the changes, which is sufficient to model our data.

Also, deep learning models, including LSTM and transformers, are backbox models where the hidden computation inside the models is not revealed or is too high-dimensional for a human to extract information. Despite the near-perfect prediction accuracy of LSTM, we cannot dissect the model to pinpoint the most important lag or the most important exogenous variables. By the same token, we cannot find out the reasons why Transformer predicts so poorly for our time series data. Therefore, deep learning models might be helpful to forecast long-term or predict the short-term movements of the indices, but it is not useful if we are undergoing a descriptive study of stock indices movement.

From this project, we have explored the art of modelling and forecasting the market indices of the S&P 500 and DJIA. We compared our model results to economic theory and tried to make sense of our results. We also explored new models like Prophet and Deep Learning architectures for time series and understood their advantages and disadvantages. **For further development**, we would like to research in detail the economic theory, contextualize the various relationships and lags between them and split the exogenous and endogenous factors to model it better. If possible, with sufficient computing power and time, we would also like to fine-tune the various hyperparameters of the Transformer and TFT to achieve better predictions in multivariate setting.

6. Reference

[1] Deb, S.G. and Mukherjee, J., 2008. Does stock market development cause economic growth? A time series analysis for Indian economy. International Research Journal of Finance and Economics, 21(3), pp.142-149

[2] Sen, J. and Chaudhuri, T., 2016, January. Decomposition of time series data of stock markets and its implications for prediction–an application for the Indian auto sector. In Proceedings of the 2nd National Conference on Advances in Business Research and Practices (ABRMP 2016).

[3] Mondal, P., Shit, L. and Goswami, S., 2014. Study of effectiveness of time series modeling (ARIMA) in forecasting stock prices. International Journal of Computer Science, Engineering and Applications, 4(2), p.13.

[4] Siami-Namini, S. and Namin, A.S., 2018. Forecasting economics and financial time series: ARIMA vs. LSTM. arXiv preprint arXiv:1803.06386.

[5] Zhang, G.P., 2001. An investigation of neural networks for linear time-series forecasting. Computers & Operations Research, 28(12), pp.1183-1202.

[6] Yadav, A., Jha, C.K. and Sharan, A., 2020. Optimizing LSTM for time series prediction in Indian stock market. Procedia Computer Science, 167, pp.2091-2100.

[7] Wang, Y. and Guo, Y., 2020. Forecasting method of stock market volatility in time series data based on mixed model of ARIMA and XGBoost. China Communications, 17(3), pp.205-221.

[8] Hussin, M.Y.M., Muhammad, F., Abu, M.F. and Awang, S.A., 2012. Macroeconomic variables and Malaysian Islamic stock market: a time series analysis. Journal of business studies quarterly, 3(4), p.1.

[9] Kirikkaleli, D., 2020. The effect of domestic and foreign risks on an emerging stock market: A time series analysis. The North American Journal of Economics and Finance, 51, p.100876.

[10] Taylor, S.J. and Letham, B., 2018. Forecasting at scale. The American Statistician, 72(1), pp.37-45.

[11] Lim, B., Arik, S.O., Loeff, N. and Pfister, T., 2019. Temporal fusion transformers for interpretable multi-horizon time series forecasting. arXiv preprint arXiv:1912.09363.

7. Appendix

The plots that were required to give further context to the analysis are referenced here.

I. S&P 500 and DJIA ACF plots

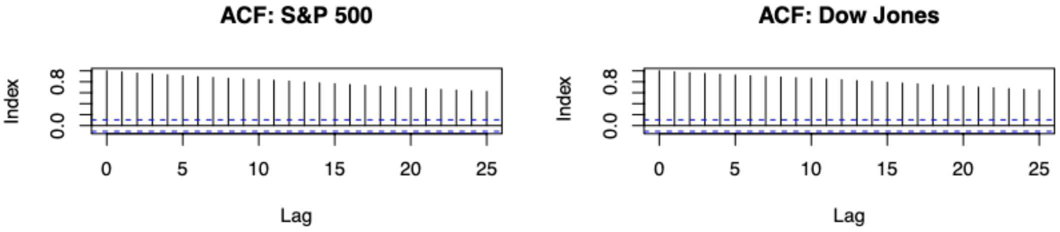


Figure. 1 ACF plots of original plot

II. Differenced S&P 500 and DJIA ACF plots

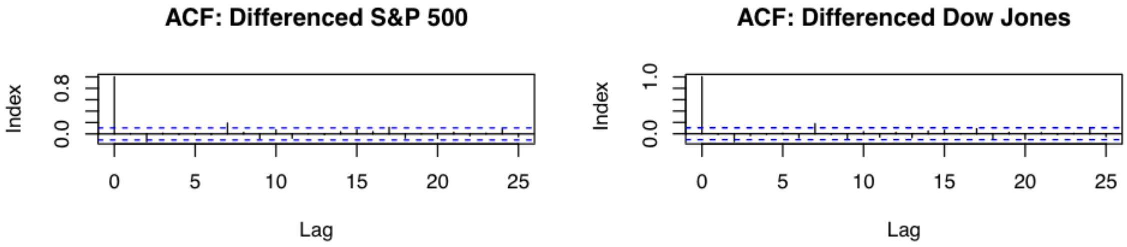
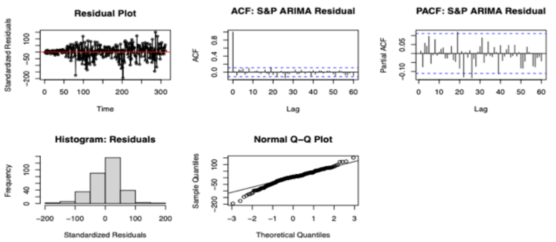


Figure. 2 ACF plots of differenced TS

III. ARMA Residual Analysis Plots for S&P 500 and DJIA



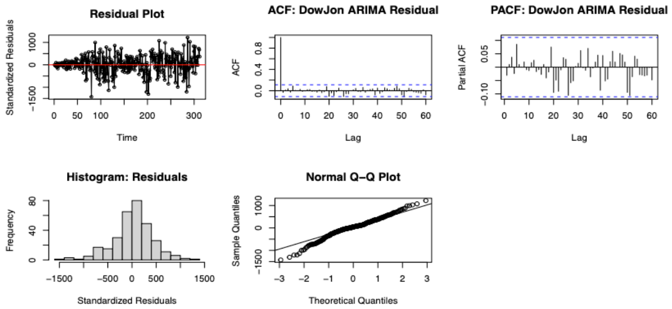


Figure. 3 ARMA Residual Analysis Plots for S&P 500 and DJIA

IV. Residual Analysis Plots for S&P 500 - SARIMA

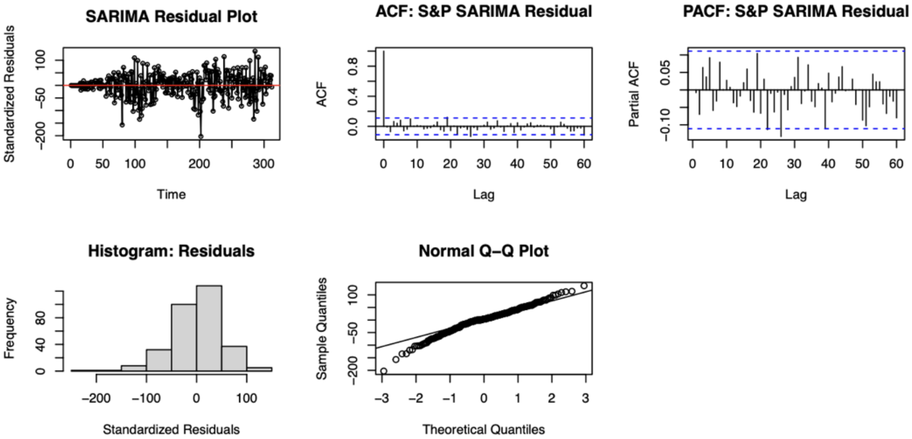
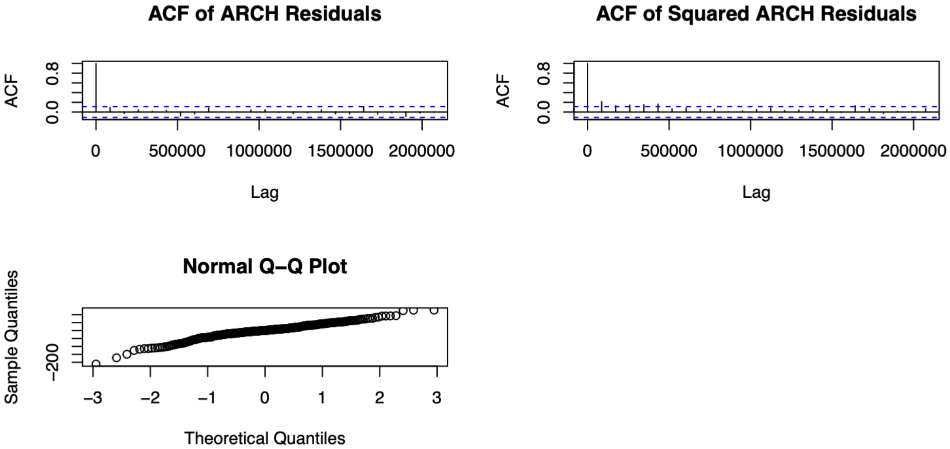
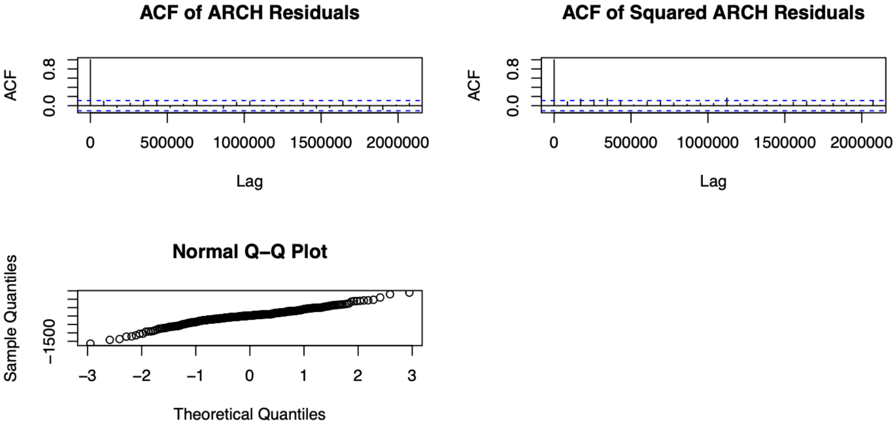


Figure 4. Residual Analysis Plots for S&P 500 - SARIMA Model

V. Residual Analysis Plots for ARMA-GARCH





VI. Multivariate: DJIA Prediction (Back to Original Scale)

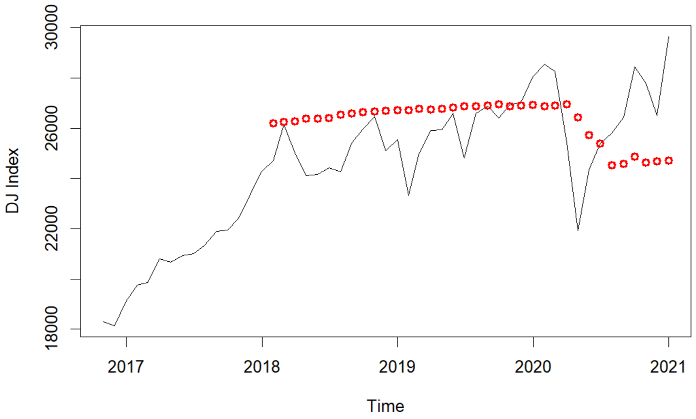


Figure 7. Multivariate: DJIA Prediction (Back to Original Scale)

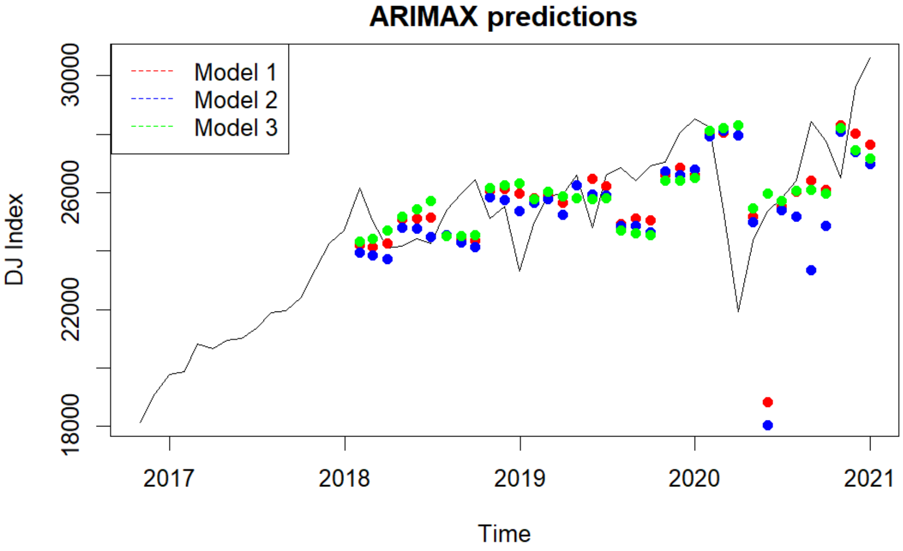


Figure 8. ARIMAX: DJIA Prediction

VII. Deep Learning Configuration

Univariate LSTM Parameter	Value
Epoch	200
Size for feature maps for each hidden RNN layer	64
Batch Size	16
Learning rate	1e-3
Dropout	0.10

Univariate Transformer Parameter	Value
Epoch	200
Number of Heads	4
Number of encoder layers	6
Number of decoder layers	6
Number of expected features	64
Dimension of Feedforward Network	256

Batch Size	16
Learning rate	1e-3
Dropout	0.10

Multivariate Transformer Parameter	**Value**
Epoch	200
Number of Heads	164
Number of encoder layers	6
Number of decoder layers	6
Number of expected features	64
Dimension of Feedforward Network	256
Batch Size	16
Learning rate	1e-3
Dropout	0.10

Temporal Fusion Transformer Parameter	**Value**
Epoch	200
Number of Heads	16
Number of LSTM layer	1
Size for feature maps for each hidden RNN layer	64
Batch Size	16
Dimension of Feedforward Network	256
Dropout	0.10

VIII. Deep Learning Univariate LSTM Prediction:

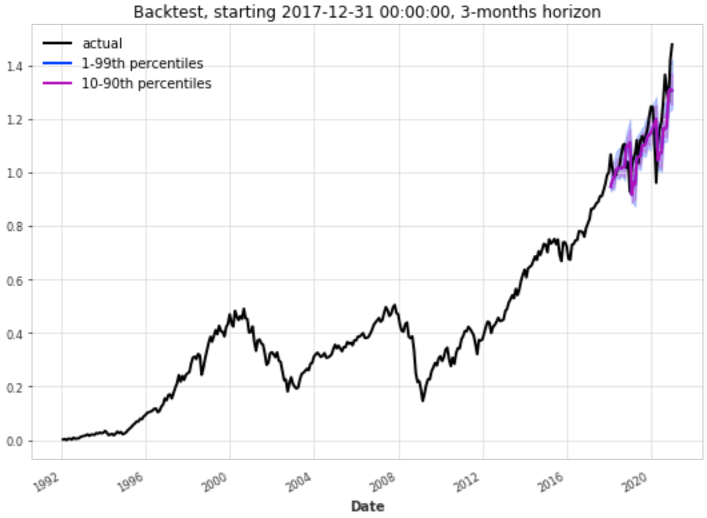


Figure X. LSTM Prediction for S&P 500

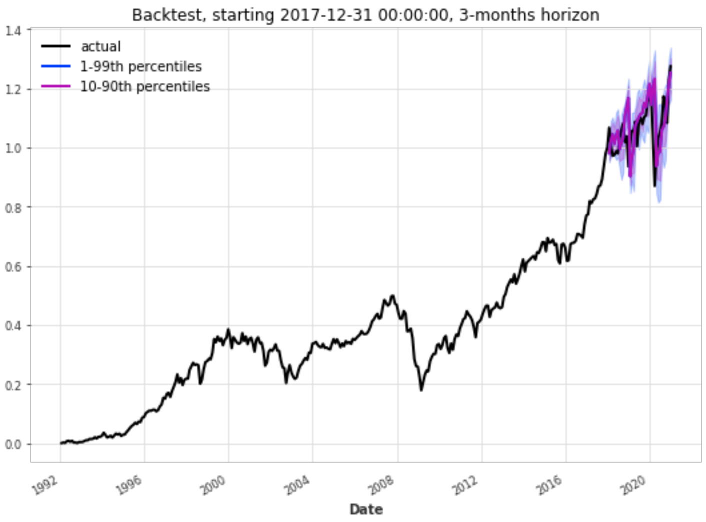


Figure X. LSTM Prediction for DJIA

IX. Deep Learning Univariate Transformer Prediction:

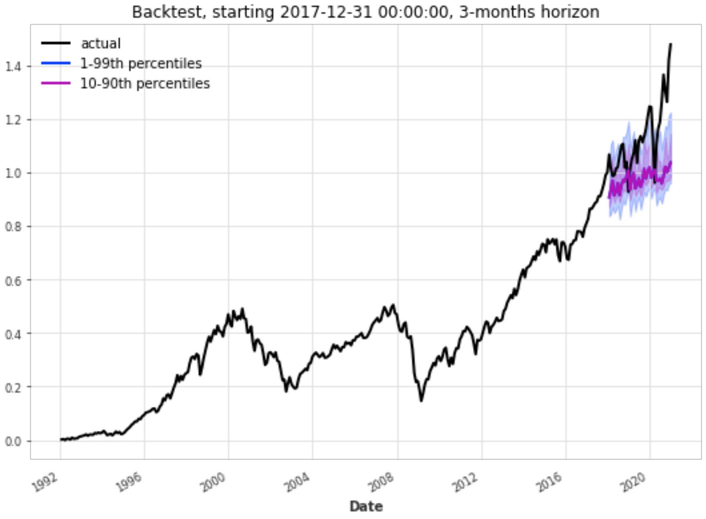


Figure X. Transformer (univariate) Prediction for S&P 500

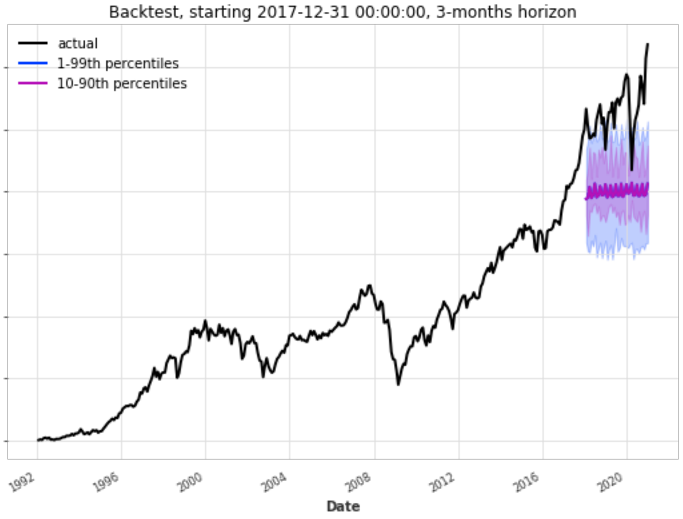


Figure X. Transformer (univariate) Prediction for DJIA

X. Deep Learning Multivariate Prediction:

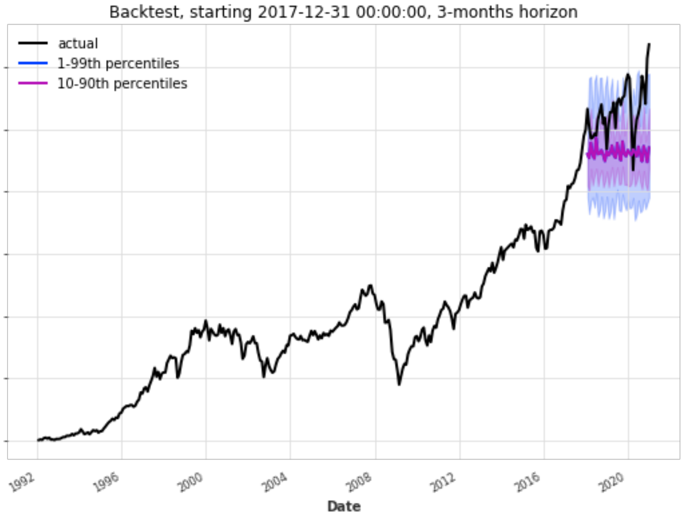


Figure X. Transformer (multivariate) Prediction for DJIA

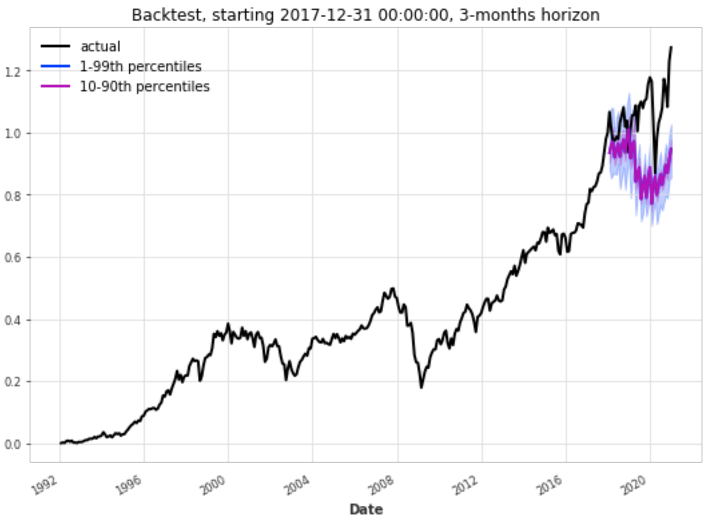


Figure X. Temporal Fusion Transformer (TFT) Prediction for DJIA

H. Complete Prediction Results: Univariate

Target Index	Model	Period	MSPE	MAE	MAPE	PM	PM difference b/t pre- and post-covid
S&P 500	ARIMA	overall	62161	189.8634	0.0644	0.7553	/
		pre-covid	28549	127.9194	0.0458	1.0569	/
		post-covid	129385	313.7514	0.1016	1.3305	+0.2736
	SARIMA	overall	63383	190.9560	0.0651	0.7701	/
		pre-covid	31733	132.3648	0.0475	1.1748	/
		post-covid	126684	308.1381	0.1001	1.3028	+0.128
	ARMA-GARCH	overall	42254	150.7378	0.0500	0.5134	/
		pre-covid	17356	101.8394	0.0367	0.6426	/
		post-covid	92051	248.5346	0.0767	0.9466	+0.304
	eGARCH	overall	63096	177.2870	0.0564	0.7666	/
		pre-covid	18884	105.9201	0.0368	0.6991	/
		post-covid	151521	320.0206	0.0958	1.5582	+0.8591
	Prophet model	overall	59244	193.6000	0.0629	0.7198	/
		pre-covid	31439	157.2500	0.0570	2.5951	/
		post-covid	73146	211.7740	0.0659	0.9250	-1.6701
	LSTM	overall	48237	179.6800	0.6020	0.5700	/
		pre-covid	30887	142.9024	0.0507	1.0959	/
		post-covid	90263	268.1480	0.0845	0.8509	-0.245
	Transformer	Overall	291030	456.9900	0.1467	3.4400	/
		pre-covid	137118	331.2657	0.1138	4.8649	/
		post-covid	598855	708.4435	0.2126	5.6452	+0.7803
DJIA	ARIMA	overall	4528758	1621.2810	0.0638	1.5136	/
		pre-covid	2032968	1147.2260	0.0452	1.2995	/
		post-covid	9520339	2569.3920	0.1009	1.8077	+0.5082
	SARIMA	overall	4601775	1609.6770	0.0635	1.5380	/
		pre-covid	2223501	1164.0960	0.0461	1.4213	/
		post-covid	9358322	2500.8370	0.0984	1.7770	+0.3557
	ARMA-GARCH	overall	2457884	1201.5310	0.0456	0.8215	/
		pre-covid	1251858	935.8728	0.0359	0.8002	/
		post-covid	4869938	1732.8470	0.0652	0.9247	+0.1245
	eGARCH	overall	2547462	1223.7420	0.0463	0.8514	/
		pre-covid	1324322	961.0194	0.0367	0.8465	/

		post-covid	4993741	1749.1880	0.0655	0.9482	+0.1017
	Prophet model	overall	666236	2311.9500	0.0878	2.2267	/
		pre-covid	746628	2520.3000	0.0996	8.9334	/
		post-covid	626041	2207.7500	0.0819	1.9704	-6.9630
	LSTM	overall	3765715	1449.8626	0.5747	1.2236	/
		pre-covid	2355331	1172.0894	0.0465	1.4428	/
		post-covid	6586485	2005.4088	0.0796	1.1464	-0.2964
	Transformer	overall	14342809	3404.8372	0.1271	4.6606	
		pre-covid	10960774	3019.0584	0.1149	6.7143	
		post-covid	21106878	4176.3949	0.1515	3.6738	

Complete Prediction Results: Univariate

VAR-DJIA	Period	MSPE	MAE	MAPE	PM	PM (diff pre & post)
VAR	overall	2100168	1134.6890	1.1002	1.0957	/
	pre-covid	1100581	830.1811	0.9292	1.0311	/
	post-covid	2599961	1286.9430	1.1857	1.1389	+0.2565
Restricted VAR	overall	2124589	1141.4550	1.1478	1.1085	/
	pre-covid	1060431	826.6203	0.9727	0.9935	/
	post-covid	2656669	1298.8720	1.2353	1.1638	+0.2626
Transformer	overall	12062349	3039.4227	0.1131	3.9196	/
	pre-covid	8676462	2631.0065	0.0994	5.3150	/
	post-covid	18834123	3856.2551	0.1394	3.2782	+0.04
TFT	overall	12669863	2991.2927	0.1113	4.1170	/
	pre-covid	6429137	2119.7060	0.0845	3.9383	/
	post-covid	25151315	4534.4661	0.1647	4.3778	+0.0802

Table . Multivariate Analysis: Prediction Results

ARIMAX Model	Period	MSPE	MAE	MAPE	PM
Model 1	overall	4166770	1487.8940	0.0579	1.3926
	pre-covid	1903184	1192.5750	0.0476	2.2772
	post-covid	5298562	1635.5540	0.0631	1.6677
Model 2	overall	5420248	1664.6500	0.0639	1.8116
	pre-covid	1711852	1063.5260	0.0422	2.0482
	post-covid	7274447	1965.2130	0.0748	2.2896
Model 3	overall	3536838	1461.7150	0.0567	1.1821
	pre-covid	2111747	1293.0740	0.0519	2.5267
	post-covid	4249383	1546.0360	0.0590	1.3375