Elinor Velasquez Capstone III Project Report

I wished to predict stock market behavior by forecasting the VWAP: Volume Weighted Average Price, the target variable to predict. VWAP is a trading benchmark used by traders that gives the average price the stock has traded at throughout the day, based on both volume and price.

My clients were Traders who traded stocks and wanted to be able to predict the behavior and outcome of stock market behavior so that their trading activity would be successful. My client may buy or sell energy stocks according to my analysis of the prediction of how the energy stocks will behave in the future.

I used a subset of the Indian stock market data, downloaded from the Kaggle website. I constrained the problem by using only data from a publicly traded Indian company which was energy-focussed: Bharat Petroleum Corporation Ltd., ENERGY, BPCL, EQ, INE029A01011. The data was time series data, and spanned approximately twenty years.

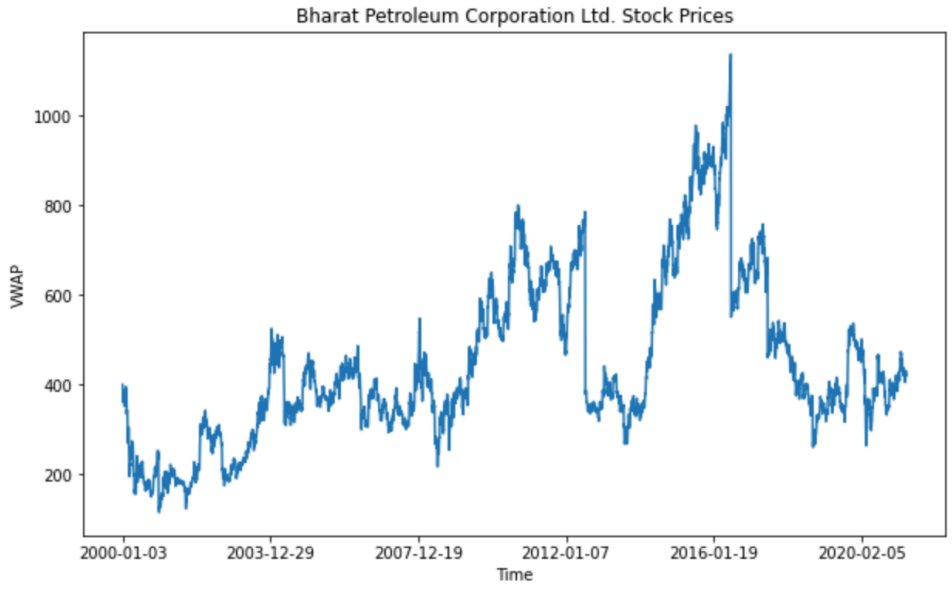

Figure 1. VWAP (Volume Weighted Average Price), 2000/01/03 - 2020/02/05.

The goal of this work was to predict or forecast the VWAP from a model describing the time series of VWAP and relevant features. Three models were investigated: a Vector Autoregressive (VAR) model, the Facebook Prophet model, and a Stacked Long Short-Term Memory (LSTM) model.

Data Wrangling/Cleaning

The data was inspected for missing values by examining the feature columns:

	count	%
Date	0	0.000000
Symbol	0	0.000000
Series	0	0.000000
Prev Close	0	0.000000
Open	0	0.000000
High	0	0.000000
Low	0	0.000000
Last	0	0.000000
Close	0	0.000000
VWAP	0	0.000000
Volume	0	0.000000
Turnover	0	0.000000
Deliverable Volume	509	9.592914
%Deliverble	509	9.592914
Trades	2850	53.712778

Figure 2. Missing Values occur in Deliverable Volume, %Deliverble, and Trades.

Deliverable Volume, %Deliverble, and Trades are three features with missing values. Using the Pandas Profiling code, histograms of each feature were created.

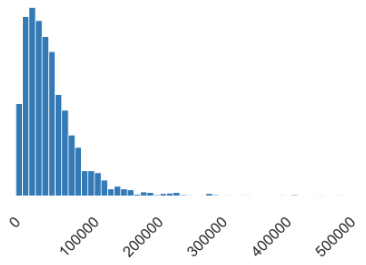

Figure 3. The feature Trades was skewed to the right.

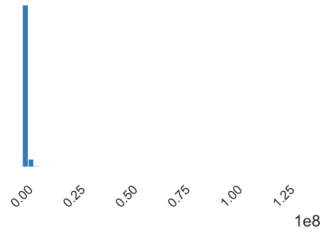

Figure 4. The feature Deliverable Volume was centered at 0.

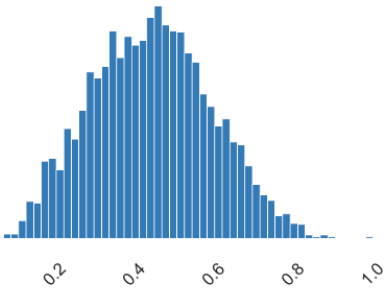

Figure 5. The feature %Deliverble was approximately normally distributed.

Trades had 53.7% missing values, so this feature was dropped from further investigation or usage. %Deliverble was approximately normally distributed, so the missing values associated with %Deliverble were imputed with the mean of %Deliverble. Deliverable Volume was not normally distributed so the missing values associated with Deliverable Volume were imputed with the median.

The categorical features, Series, Symbol, were determined to be irrelevant, so they were also dropped from the analysis.

Exploratory Data Analysis

The feature VWAP was resampled by year to create a smooth graph:

Figure 6. VWAP, resampled by year, resulting in a smooth graph.

If the years 2018 to 2021 are deleted from the graph, the plotted points demonstrate an upward trend, with an increasing mean and increasing variance, so the graph is non-stationary.

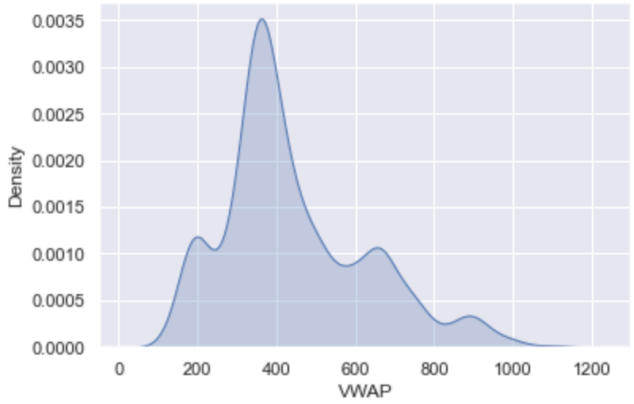

Figure 7. The kernel density plot of VWAP shows the median of VWAP is centered at approximately 350.

To determine if the VWAP was normally distributed, a Q-Q plot was calculated.

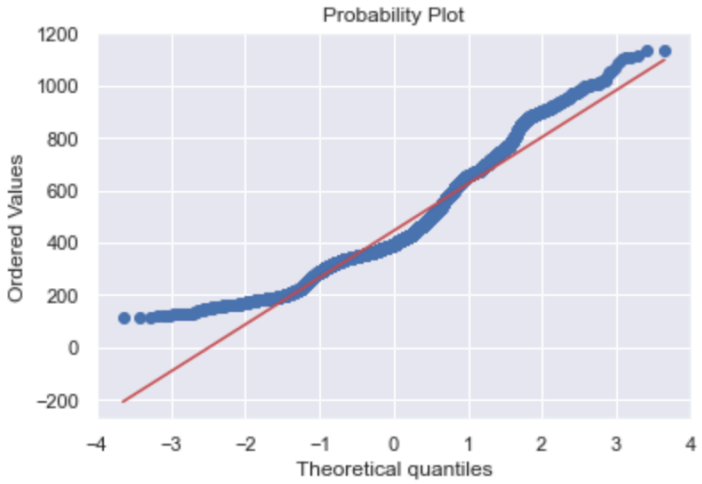

Figure 8. Q-Q plot of VWAP values

The Q-Q plot shows that VWAP is not normally distributed because the points (blue) only map to the (red) line in the middle of the graph and diverge on the ends of the red line.

The p-value for the Augmented Dickey-Fuller statistic for VWAP is computed to be 0.042. This means that the VWAP data is stationary. But Figure 1 tells a different story. Thus, performing differencing (using the diff() function from Pandas) of the VWAP causes the p-value to equal strictly 0. Perhaps it is reasonable to assert VWAP to be weakly stationary.

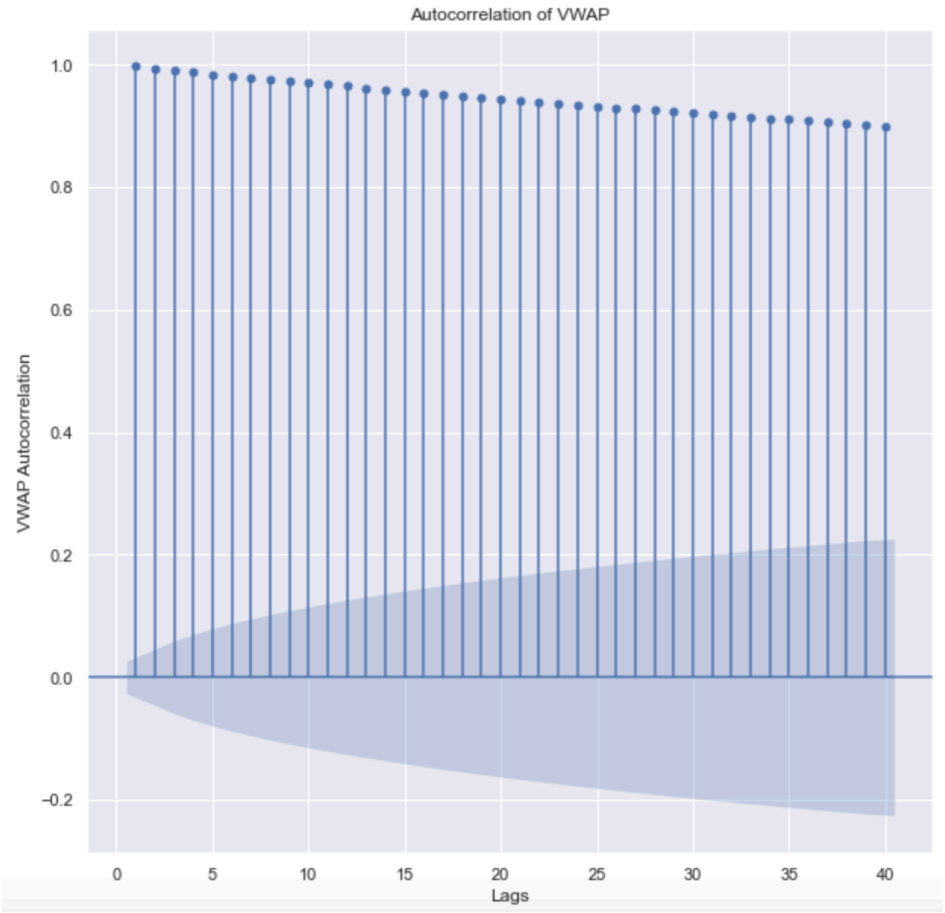

Figure 9. Autocorrelation of VWAP shows that values of VWAP are highly correlated with previous (lag) values.

The ‘spikes’ are significant for lags up to 40. To summarize, the VWAP points are highly correlated with each other. The partial autocorrelation (Figure 10) is significant at only two lags and the remaining lags are near zero, so the model should include an AR term of 2 (assume an additive model):

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t$$

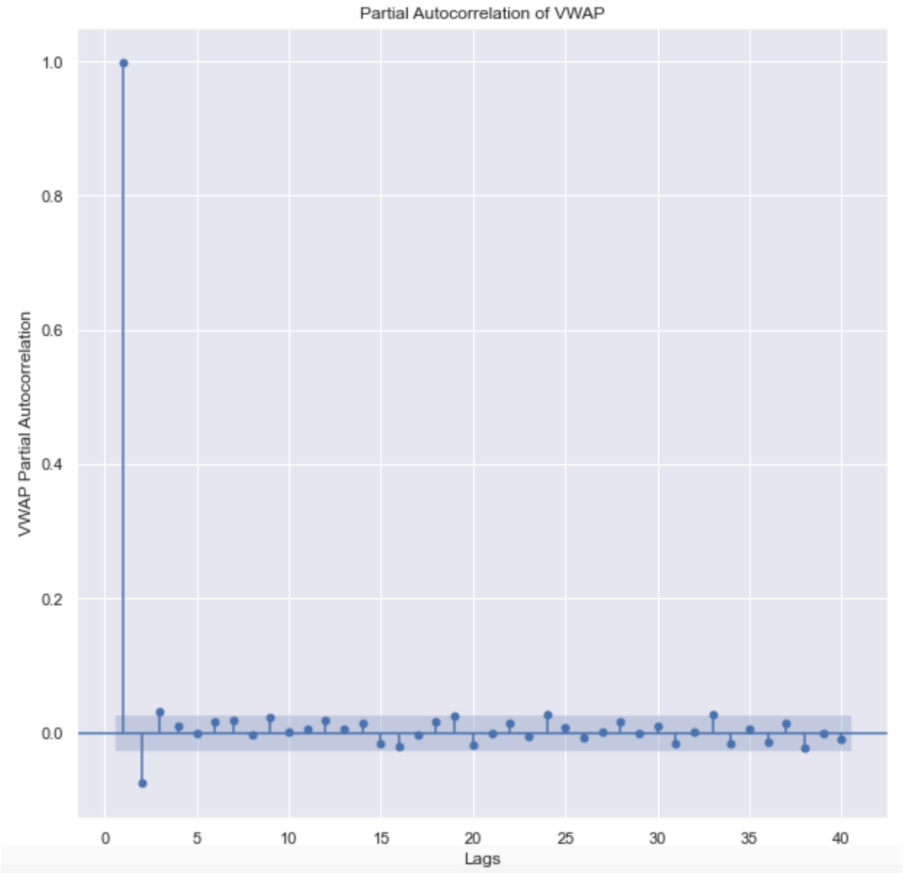

Figure 10. The graph of the partial autocorrelation of VWAP shows that only two lags are significant.

The VWAP feature may be decomposed multiplicatively and additively into trend, seasonal, and residual. The trend is slightly increasing and is equivalent for both multiplicative and additive decomposition, so VWAP has a weakly non-stationary behavior.

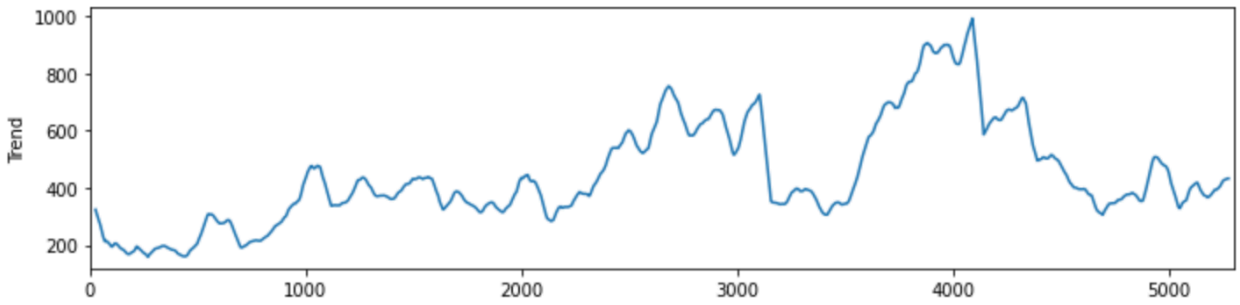

Figure 11. The trend shows the feature VWAP is weakly non-stationary, as it is slowly increasing on average.

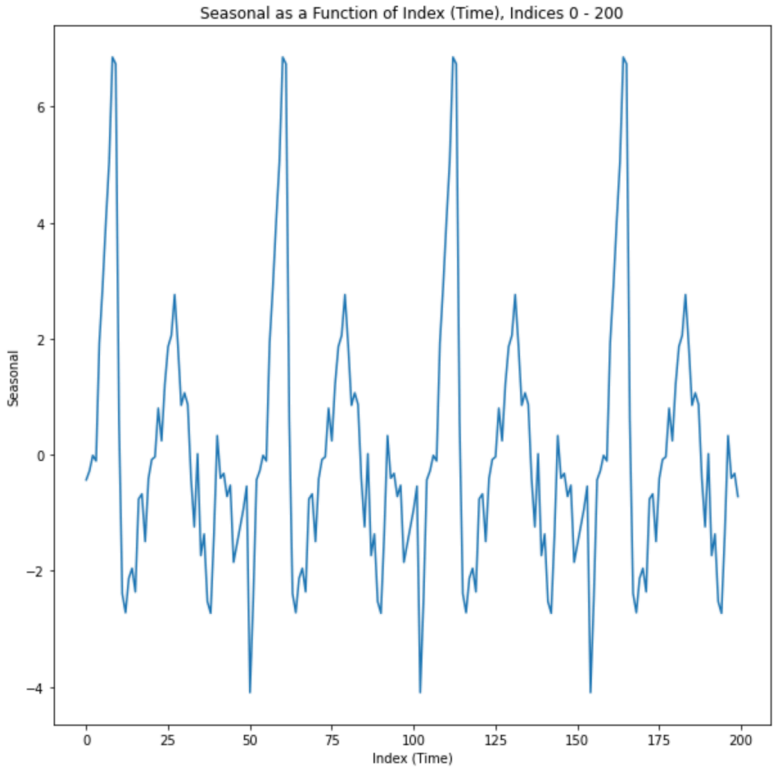

Figure 12. The seasonal component of both the additive and multiplicative decomposition is highly periodic.

For the multiplicative decomposition, the residual is centered at 1, while for additive decomposition, the residual is centered at 0.

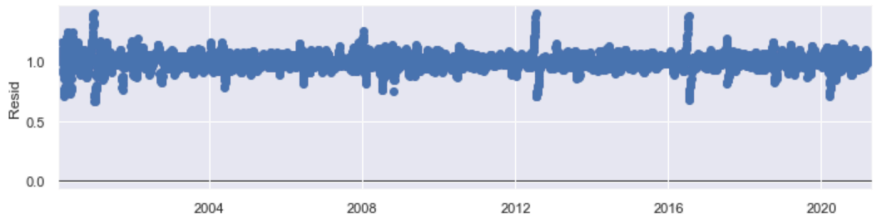

Figure 13. The multiplicative noise component or residual displays randomness.

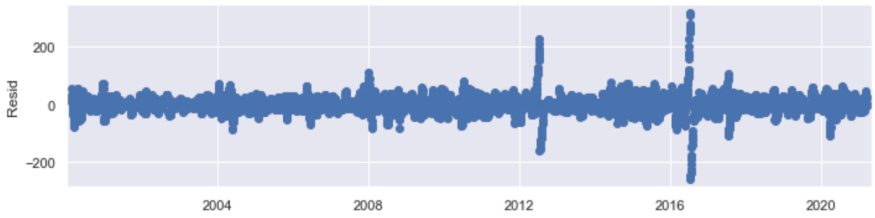

Figure 14. The additive noise component or residual displays randomness.

Modeling: Vector AutoRegression (VAR)

The Vector Autoregression (VAR) model is a classical algorithm to be used when two or more features influence each other in order to forecast. Each feature is modeled as a linear combination of past values of itself and the past values of other features in the system. Since there are multiple features that influence each other, the model is a system of equations with one equation per target (Y_1, Y_2):

$$Y_{1,t} = \alpha_1 + \beta_{11,1} Y_{1,t-1} + \beta_{12,1} Y_{2,t-1} + \epsilon_{1,t}$$
$$Y_{2,t} = \alpha_2 + \beta_{21,1} Y_{1,t-1} + \beta_{22,1} Y_{2,t-1} + \epsilon_{2,t}$$

The above is a lag 1 VAR model, with alpha and beta the coefficients to be determined (learned) and epsilon the error term. In this project, Y_1 is the VWAP price of the stock, and Y_2 is the feature “Previous Close.” The focus is on the forecast of VWAP. The idea behind Vector AutoRegression is that each of the features in the system influence each other. A feature can be forecasted using the past values of itself along with other features in the system. In the above scenario, there are two features, each of lag 1, so the order of the system or model is one.

Granger’s causality tests the null hypothesis that the coefficients of past values in the model are zero.

From Scholarpedia:
Granger causality is a statistical concept of causality that is based on prediction. According to Granger causality, if a signal X1 "Granger-causes" (or "G-causes") a signal X2, then past values of X1 should contain information that helps predict X2 above and beyond the information contained in past values of X2 alone.

The p-value generated for the Granger causality matrix between VWAP and Previous Close is 0. Thus, it is not true that VWAP does not depend on Previous Close and it is not true that Previous Close does not depend on VWAP. Thus, the Granger causality null hypothesis is rejected: The coefficients of past values are not zero.

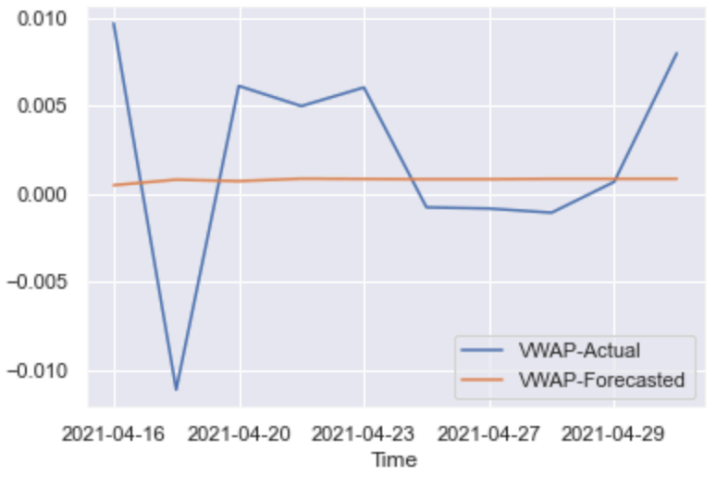

Figure 14. The actual values for the VWAP price (blue line) and the forecasted values (orange line) for the VWAP price for the VAR(5) model.

The Augmented Dickey-Fuller test (ADF) was used to test for stationarity of the data. Non-stationarity means that the means and variance vary in the dataset. The data was differenced, meaning computing the differences between consecutive observations. After one instance of differencing, the data became stationary.

```
Forecast Accuracy of: VWAP-Actual
mape :  1.1466
me :  -0.0014
mae :  0.0048
mpe :  -1.0946
rmse :  0.006
corr :  -0.3931
minmax :  2.6305
```

Figure 15. The metrics for VAR(5) of VWAP stock price.

Modeling: The Prophet Forecasting Model of Facebook

Prophet is an open-sourced library for analyzing and forecasting time series data. The Prophet model, developed by a team at Facebook, may be described as

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t$$

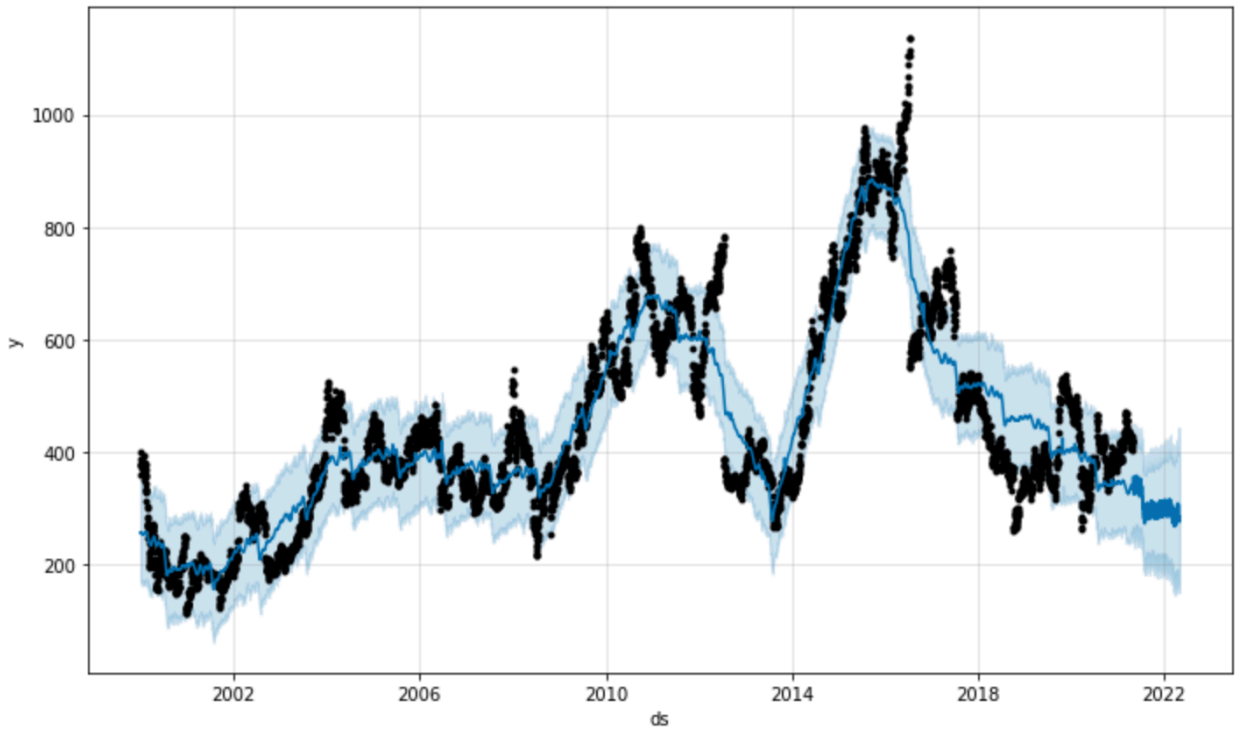

Figure 16. The Prophet model (blue) predicts the y = 'VWAP' stock price (vertical axis), with a (light blue) band describing the confidence interval, and ds, the time (horizontal axis).

with y(t) the target (VWAP stock price), g(t) the trend function, s(t) the seasonality, h(t) the effects of holidays, and ε_t the error term.

	horizon	mse	rmse	mae	mape	mdape	coverage
0	19 days	24867.956812	157.695773	114.913316	0.241599	0.161142	0.393598
1	20 days	25099.110326	158.426987	116.021946	0.244771	0.164150	0.381051
2	21 days	25827.434992	160.709163	118.011394	0.246175	0.166175	0.372258

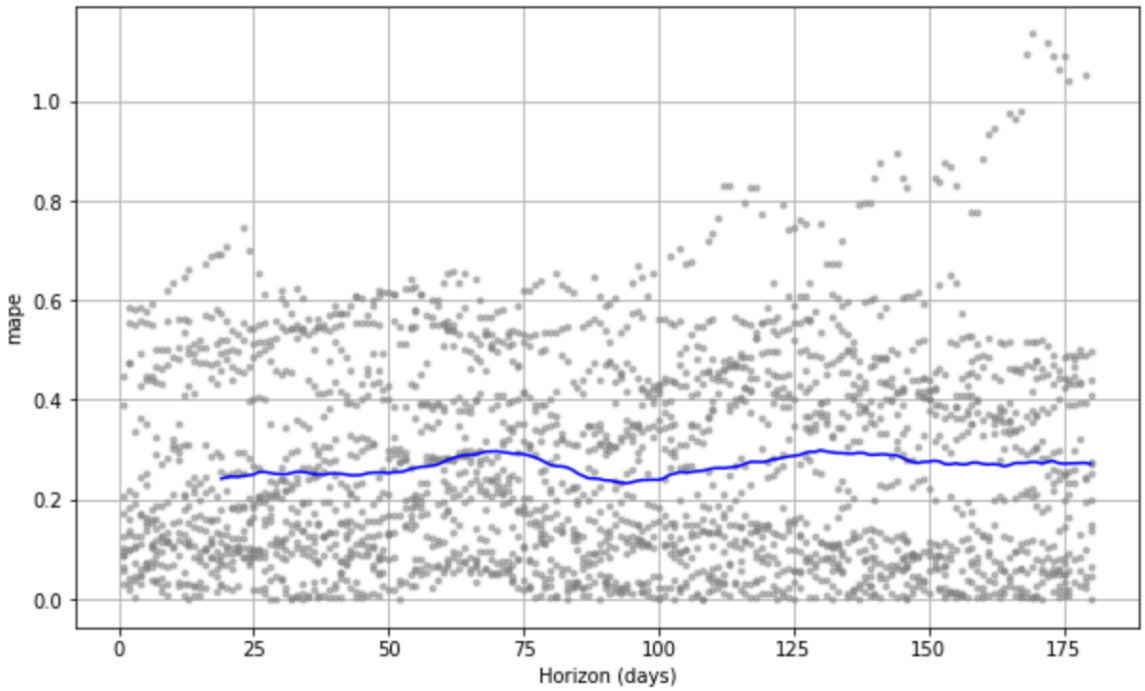

Figure 17. The mean absolute percentage error (MAPE) for the Prophet model (0.24) is significantly less than the MAPE for the VAR(5) model (MAPE = 1.15). The (blue) line is the fitted line for MAPE.

The Prophet model has less error (as measured with MAPE) than the classical VAR(5) model, which may be re-interpreted as meaning that the Prophet model is more accurate than the VAR(5).

The mode of seasonality for the model is additive.

Modeling: The Stacked Long Short Term Memory (LSTM) Model

The basic LSTM, a recurrent neural network, has an input layer, a single hidden layer followed by a standard feedforward output layer. The stacked LSTM extends the basic

model by having multiple hidden LSTM layers. Recurrent neural networks are used for sequential (ordered) data, such as times series data. A RNN hidden state maintains an internal state that captures information about the time steps it has seen at a given point.

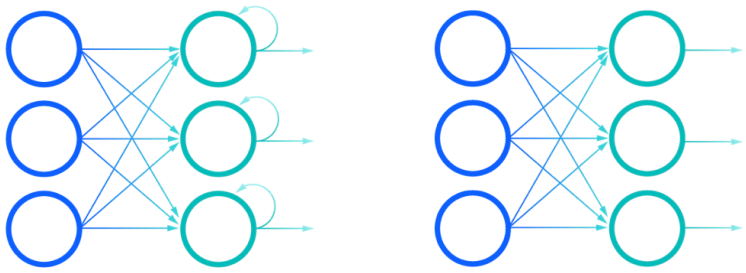

Figure 18. A recurrent neural network (left) versus a feedforward neural network (right)

The data was standardized, using MinMaxScaler. Then, the data was split into train and test subsets.

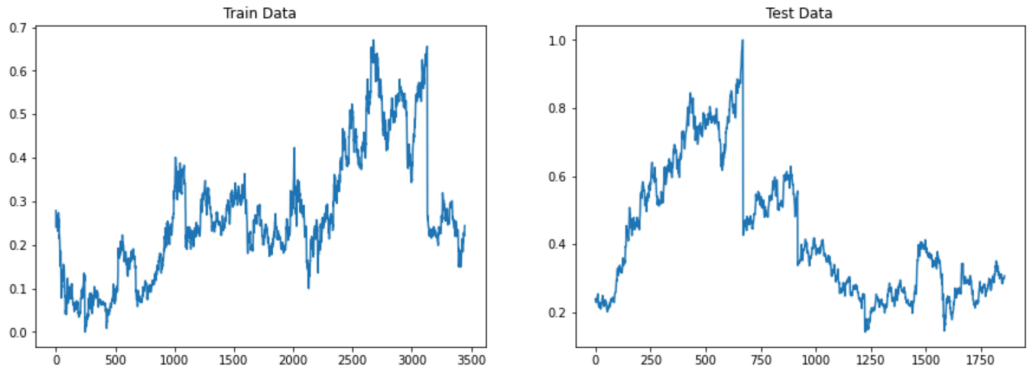

Figure 19. The vertical axis represents the standardized time series values (VWAP), and the horizontal axis represents time (days).

Because the time series is an ordered dataset, the train and test subsets needed to be created in a way so that the data order was not disturbed. The standardization of the data should have been created after the creation of the train and test subsets to avoid data linkage.

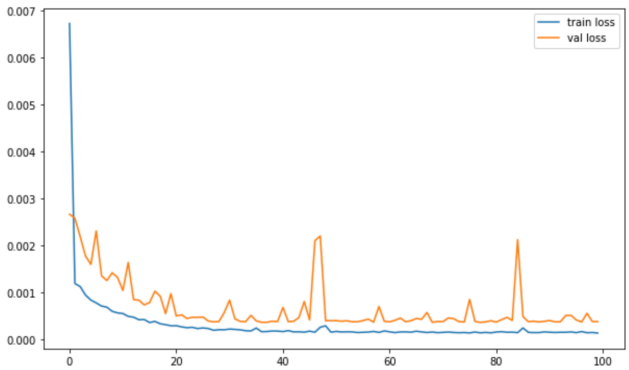

Figure 20. The loss function was chosen to be mean squared error.

The error for the trained data is shown in Figure 20 and overall, is less than the error for the test data (also shown in Figure 20).

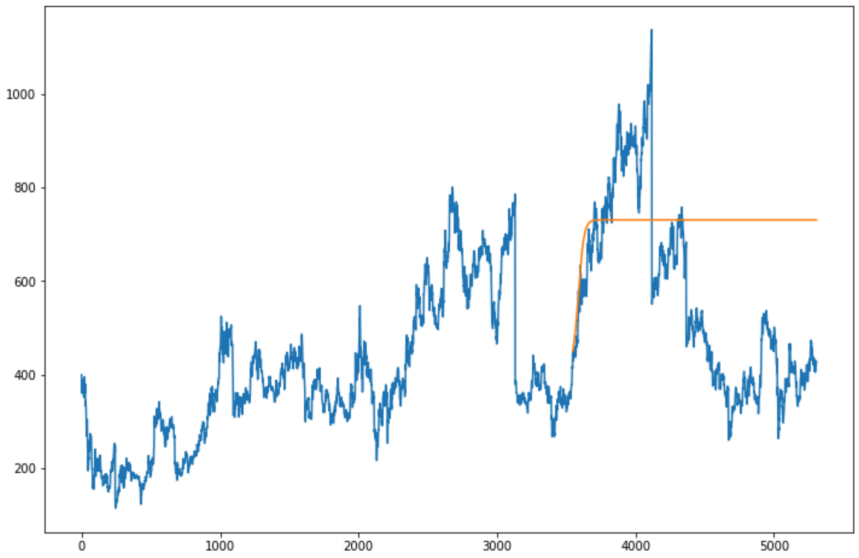

Figure 21. The VWAP price was predicted using the LSTM model. The horizontal axis represents time (days) and the vertical axis represents VWAP price. The orange line is the prediction or forecast for VWAP stock market prices and the blue line represents the (known) VWAP prices. The horizontal axis is time. The vertical axis is the forecast value of VWAP.

For the LSTM neural network, two hidden layers were used. The number of epochs defines the number of times that the chosen algorithm will pass through the trained data. One hundred epochs was used for the LSTM model.

In terms of error, the MAPE score for the LSTM model is 0.35. This means that the Facebook Prophet model (MAPE = 0.24) performs better than the VAR(5) model (MAPE = 1.15) and the LSTM RNN model (MAPE = 0.35). Thus, the best model is the Facebook Prophet model.