Tesla Stock Price Prediction Using ARIMA-Based Models

Danning Lao, Zhanpeng Zhou, Zhengfei He

UM-SJTU JI

December 1, 2020

Table of Contents

Background

Tesla is an American electric vehicle and clean energy company which is founded in 2003 but achieve increasingly success and popularity these days. Correspondingly, as shown in Fig 1, the stock price of Tesla is irregular and unpredictable with an increasing trend.

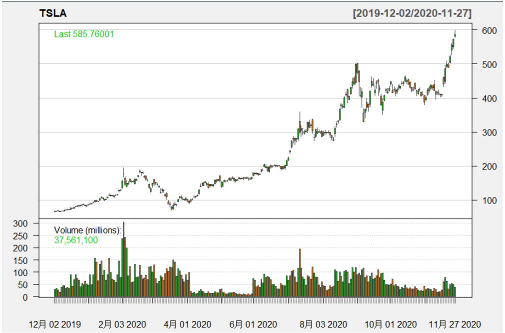


Figure: Candlestick Grpah of Tesla's Stock Price from Dec 2019 to Dec 2020

Objective

In this project, we aim to

1. Study the historical data of Tesla
2. Build appropriate explainable regression model to predict Tesla's closing price on the five trading days $7^{th} Dec$ to $11^{th} Dec$
3. Analyze and validate the model

The model building, improving, selecting and validating process will be elaborated in later sections.

Table of Contents

Flowchart of Design Overview

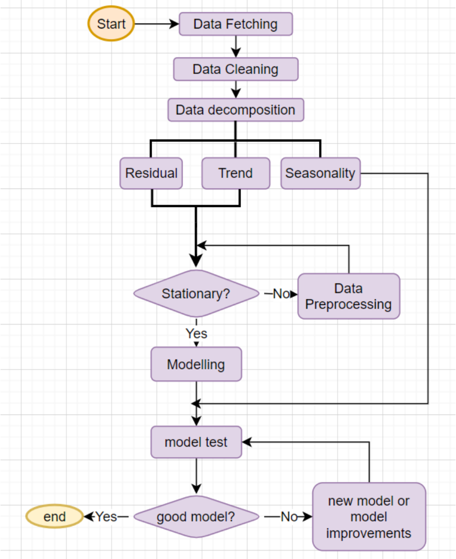


Figure: Flowchart of Design Overview

Data fetching and Data cleaning

Data fetching

1. Data Set: Yahoo Finance TSLA Closing Price
2. Data period choice: We should not choose large period of data because sudden events may become outlier of the data series and then affect stationality.
3. Special Note: We only have weekdays data. Consider the convenience of finding seasonality, data on weekend days are filled with the value of Friday. And then seasonality period is chose as 7 based on common sense of weekly data.

Data cleaning

1. Assumption: Consider the proficiency of the website that we fetch data from, there will not be simple mistakes such as typos. We only have to deal with missing values.
2. Missing values: interpolation

ADF Check

A common choice to fit time-series data is **ARIMA**, which requires a stationary input data series.

Augmented Dickey Fuller test (ADF Test)

- H_0: The time series is non-stationary
- p-value $= 0.6905480137874351$

Thus we can't reject the null hypothesis that the origin time series data is non-stationary.
To satisfy the stationary assumption, we need to do further data preprocessing.

Data Decomposition

The normal method to deal with time series data to achieve relatively high forecasting accuracy is **Data Decomposition**.

We decomposed the time series data into a merge of sub-series:

- Trend T_t, find to need further adjusted.
- Seasonality S_t, find to be stationary.
- Residual R_t, find to be further adjusted.

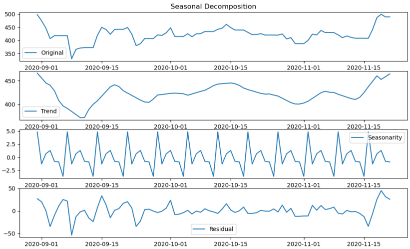
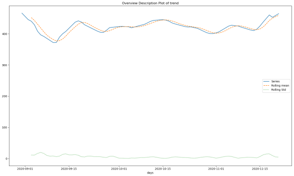


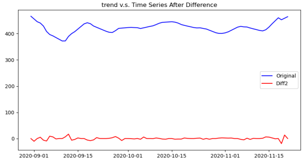
Figure: Data Decomposition in Seasonal ARIMA model

Dealing with Trend

Before using ARIMA model to fit the decomposed trend data, we need to analyze whether it's stationary, if it's not stationary, we will keep perform difference on the series data until it can pass the Augmented Dickey–Fuller (ADF) test. Based on the test result, we choose to use twice-differentiated trend data.



(a)

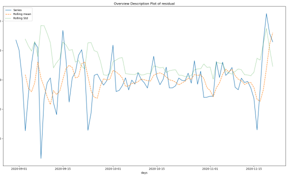


(b)

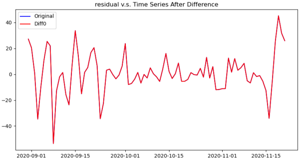
Figure: Stationary analysis of the decomposed trend data

Dealing with Residual

Similar as the procedure in Trend, we also do stationary analysis on the decomposed residual data. We choose to use the original residual data.



(a)



(b)

Figure: Stationary analysis of the decomposed residual data

Finding parameter

ARIMA(p,d,q)

1. Integrated/Mean Distributed Lag parameter d: the minimum difference order to make the sequence stationary
2. Autoregressive (AR) parameter p, indicating relationship with history value
3. Moving average (MA) parameter q, indicating relationship with AR prediction error.

(*) p, q will be selected by grid search using criteria of BIC.

Verify parameter

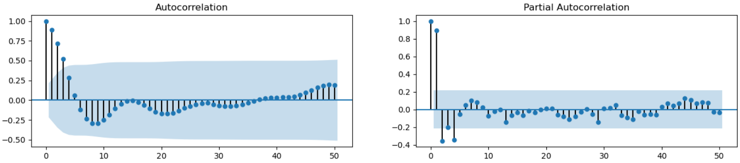


Figure: ACF/PACF plot for differentiated trend data

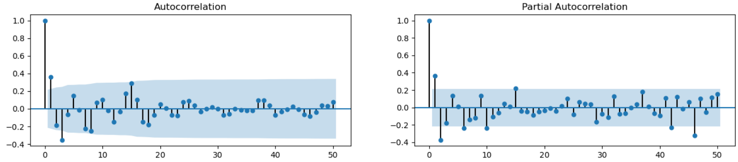


Figure: ACF/PACF plot for differentiated residual data

Seasonal ARIMA model validating

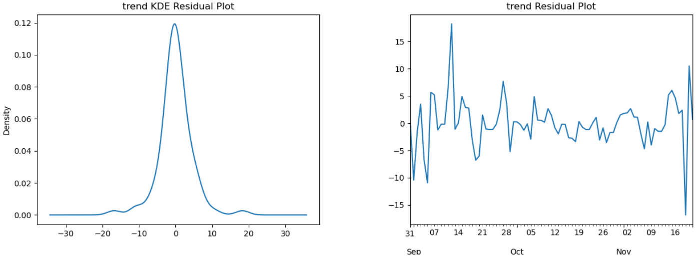


Figure: Residual Analysis of fitted trend data

Seasonal ARIMA model validating

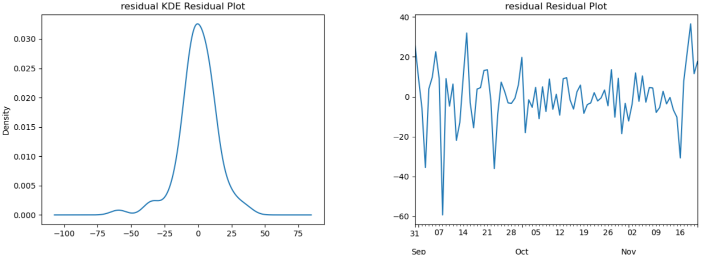


Figure: Residual Analysis of fitted residual data

Seasonal ARIMA model validating

However, when looking at the ACF plot of the model, there still exits autocorrelation, which indicates the model is not adequate.

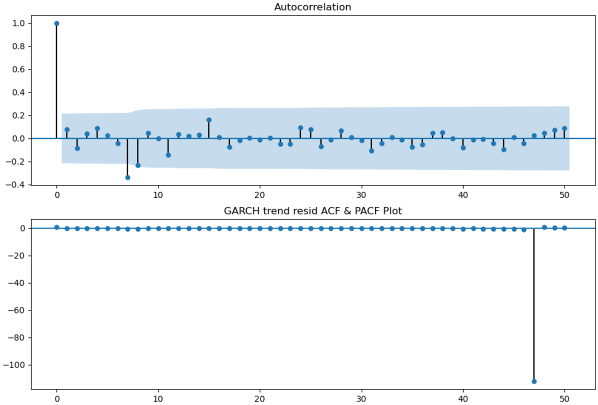


Figure: ACF/PACF residual plot for trend data

Seasonal ARIMA model validating

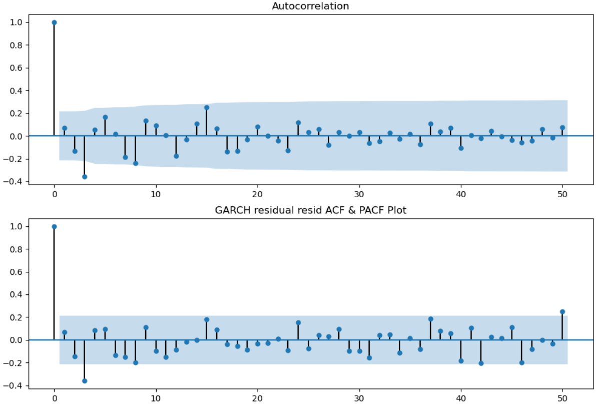


Figure: ACF/PACF residual plot for residual data

Further Thoughts

- Seems to have protential correlated errors.
- it is not safe to base your judgement on pacf graphs with no confidence bands around them. A pacf may look small in such a graph but may actually be significant. It's wiser to get numeric test statistics.

Acorr_ljungbox test:

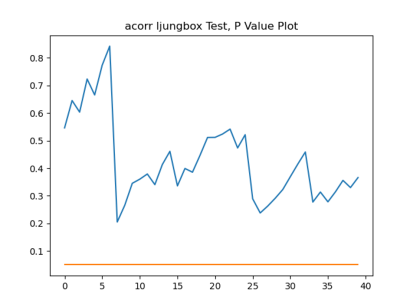


Figure: acorr_ljungbox test p-value plot

Introduction of GARCH

In this model, ARIMA explain a linear relationship between past and future data, while we have to introduce GARCH to address volatility clustering problems, reducing heteroskedasticity.

Mathematical Formulas

Where the ARIMA-GARCH model can be expressed as:
ARIMA:

$$W_t = E[T(T_t)] = \sum_{i=1}^{p} \Phi_i W_{t-i} + \epsilon_t + \sum_{j=1}^{q} \theta_j \epsilon_{t-j}$$

GARCH:

$$\epsilon_t = \sigma_t Z_t$$

$$\sigma_t^2 = \omega + \sum_{i=1}^{m} \alpha_i \epsilon_{t-i}^2 + \sum_{j=1}^{n} \beta_j \sigma_{t-j}^2$$

(*) p, q, m, n will be selected by grid search using criteria of BIC.

Assumption

Here we have two assumptions:

$$Z_t \sim N(0, 1)$$

$$\epsilon_t \ \textit{stationary}$$

And then we repeat the process with $P_t = E[T(R_t)]$, and obtain the final model:

$$Y_t = T^{-1}(W_t) + T^{-1}(P_t) + S_t$$

Modelling Result of Seasonal ARIMA-GARCH

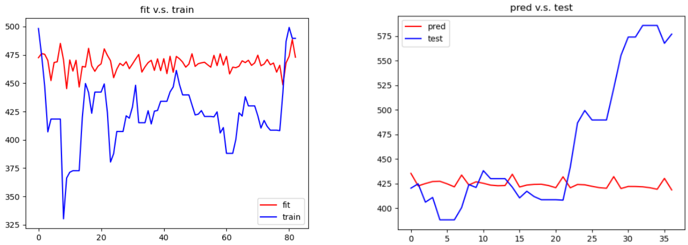


Figure: Stationary analysis of the decomposed trend data

Model Comparison and Model Selecting

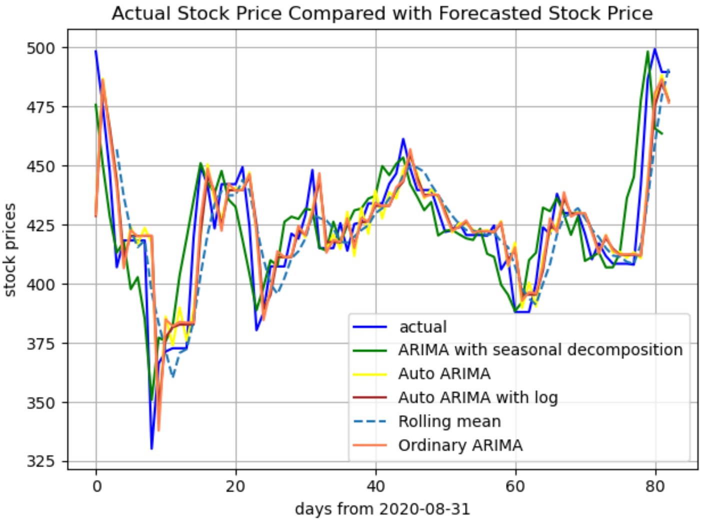


Figure: ARIMA fit result

Table of Contents

Future Work

There are two aspects we want to focus on later

1. linearity assumption of ARRIMA suppose future data is a linear combination of current, past data and errors, which may not always suit the realistic case.
2. Explainability: simply using close stock price to predict is not so meaningful as using multiple related regressors. Another issue is, We have collected multiple data, including Dow Jones Index Average, Brent Crude Oil Price, Revenue and then applies PCA to reduce dimension. However, problem comes that in order to predict Tesla's price based on these data, we need to have future value of these data, which are impossible. Later, more work will be done to fit a more explainable model with these data.

Table of Contents

Conclusion

After a comparison study among different models, based on *RMSE* result, we choose to use *Seasonal ARIMA-GARCH* model which gives the smallest RMSE $= 97.3$, indicating our model is valid and good for prediction.

Q & A

Thank you