Times Series Analysis as function of the Oil Price

Problem Identification

Predicting stock price is challenging due to different external factors that make this variable to move unexpectedly. Additionally, oil price is considered as a risky and volatile commodity. Many economists believe this commodity determine if the economy is going to move in a good or bad direction.

If I could create a model able to predict how different stocks may move by taking into consideration the oil price, and time; this would help investors and traders to make any investment decision on the stock exchange market

.

Considered what I mentioned above, on this capstone, I am comparing how well different times series analysis, and regressions do in order to predict the stock price as a function of the time and the oil price. The stocks I am analyzing are: Jinko Solar (JKS), American Airlines (AAL) and Tesla (TSLA).

The reasons I selected these stocks are many, but most importantly, I wanted to select stocks that I believed had a different kind of relation with the oil price. Jinko Solar would be the stock that is totally uncorrelated with the oil price since they manufacture solar panels, American Airlines would be the stock with high correlation with the oil price since oil is the main expense for the airlines, and Tesla would be the stock with negative correlation with the Oil Price due to the nature of their cars that work with electricity.

Finally, I would be making recommendations of what stocks among JKS, America Airlines and Tesla an investor or trader should buy, sell, or hold to make profits.

Data Collecting

To collect the stock prices from June the 29th of 2010 to August the 31st of 2020 of the companies mentioned before, I used yahoo finance. This is a website to check out the stock prices of public companies since the company went public. In this specific capstone.

I selected the start date of June the 29th of 2010 because even though American Airlines went public in 1980; Tesla and Jinko Solar went public in 2010, and I wanted to have the three stocks in the same time range.

Finally, I gathered the data of the oil price in the QUANDL platform and downloaded it as a csv from June the 29th of 2010 to August the 31st of 2010.

Data Wrangling

I needed to take into consideration some aspects while I was doing the data wrangling.

Firstly, I needed to check out if there were whether or not Null-Values on the data collected in yahoo finance and QUANDL, and after checking that out, I did not have any Null-Values to fill out.

Secondly, I needed to check if the dates were consistent between the stock and the oil prices. After reviewing this, all of the prices are traded in business days, so I did not have mismatching dates when I was making the merging of the four different data frames I had.

Finally, I also made sure to have indexes of the final data frame as date time format to make the times series analysis and sort the values by date in order to avoid any future overfitting and make easier the division between train and test dataset.

This is how the data frame looks like after making the merge:

Date	Oil Price	JKS	AAL	TSLA
2010-06-29	72.66	9.96	8.051013	4.778
2010-06-30	72.49	9.70	8.117004	4.766
2010-07-01	70.48	9.80	8.154712	4.392
2010-07-02	69.63	10.21	7.721052	3.840
2010-07-06	69.73	10.79	7.617351	3.222

Exhibit 1. Data Frame that contains the dates, oil, and stock prices.

Exploratory Data Analysis

For the Exploratory Data Analysis (EDA), I made autocorrelation plots to get a general idea which order I would be looking for the times series analysis. I also a graphic that shows the stocks behaviors since June the 29th of 2010. Finally, I plotted a correlation heatmap to see how correlated are JKS, AAL and TSLA stock prices with the Oil Price.

In the following Exhibit, I am showing the stocks behavior.

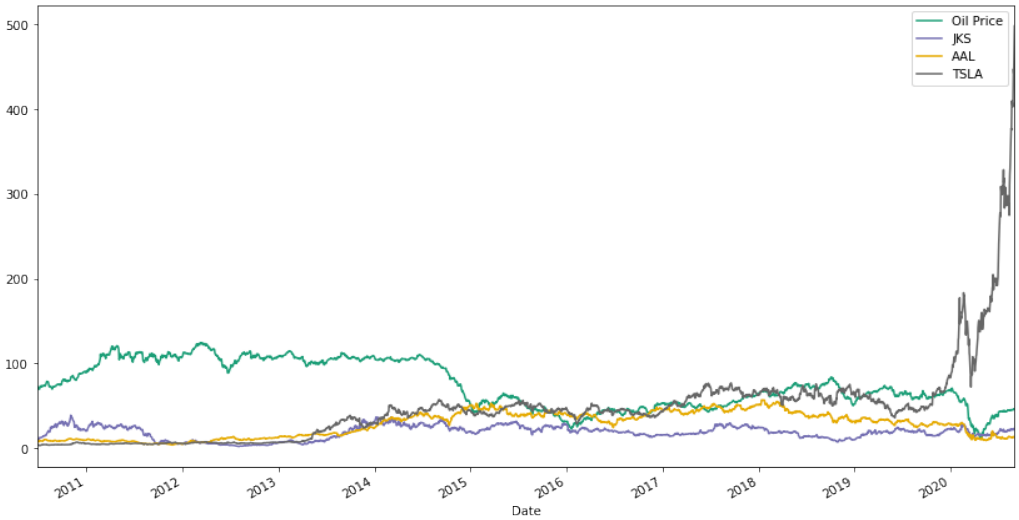


Exhibit 2. Stocks and Oil Price behavior since 2010.

From the Exhibit 2, we are seeing a tendency on TSLA stock of going up. On the other hand, Oil Price looks very volatile, and American Airlines and JKS have had a very stable behavior on their stock price.

I used three different measures of correlations. As an important mention, I used the data from 2019 to 2020 to measure the correlation because of the price gap between TSLA and Oil Price before 2019. The heatmaps are shown below:

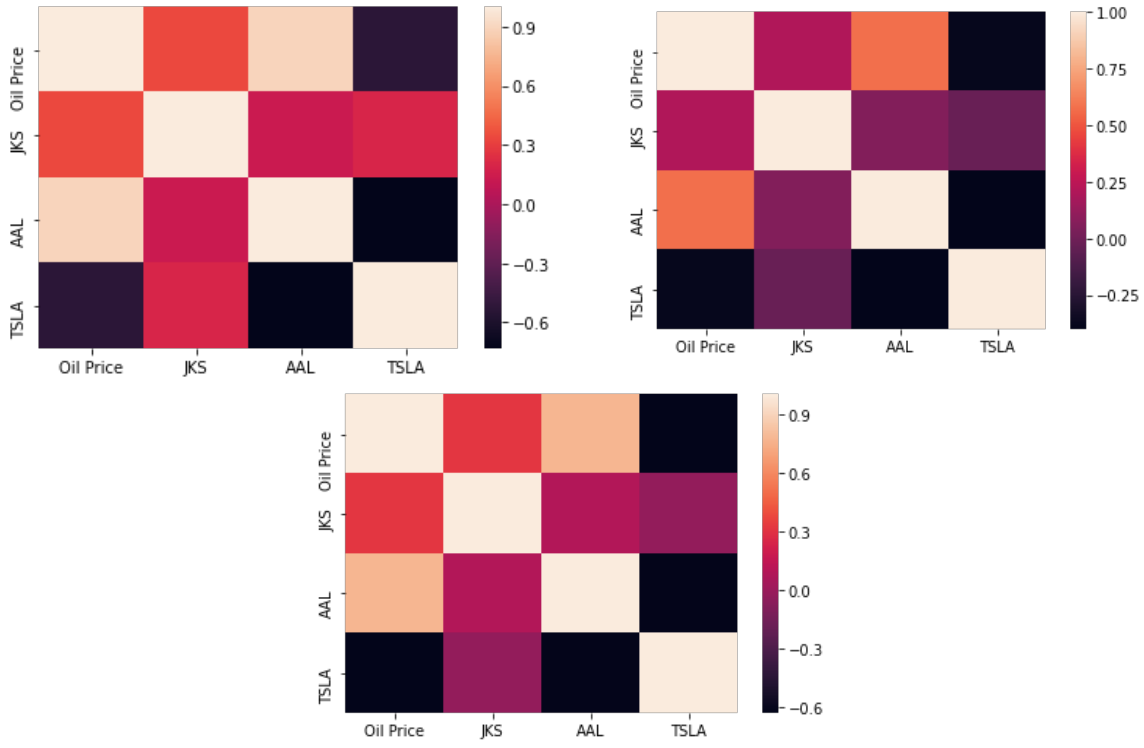


Exhibit 3. Stocks and Oil Price correlation. From right to left. A) Pearson Correlation. B) Kendall Correlation. C) Spearman Correlation

From the three different correlation heatmaps, Oil Price and TSLA are the least correlated. Oil price has been decreasing (I can also see this on the Exhibit 1) and TSLA stock price has been increasing. On the other hand, American Airlines and Oil Price have the highest correlation, and finally JKS and Oil Price are not correlated.

To finalize the Exploratory Data Analysis, I also made the autocorrelation plots for the stocks and oil price to get an idea of the order I would be using in the times series analysis. The autocorrelation plots are shown below:

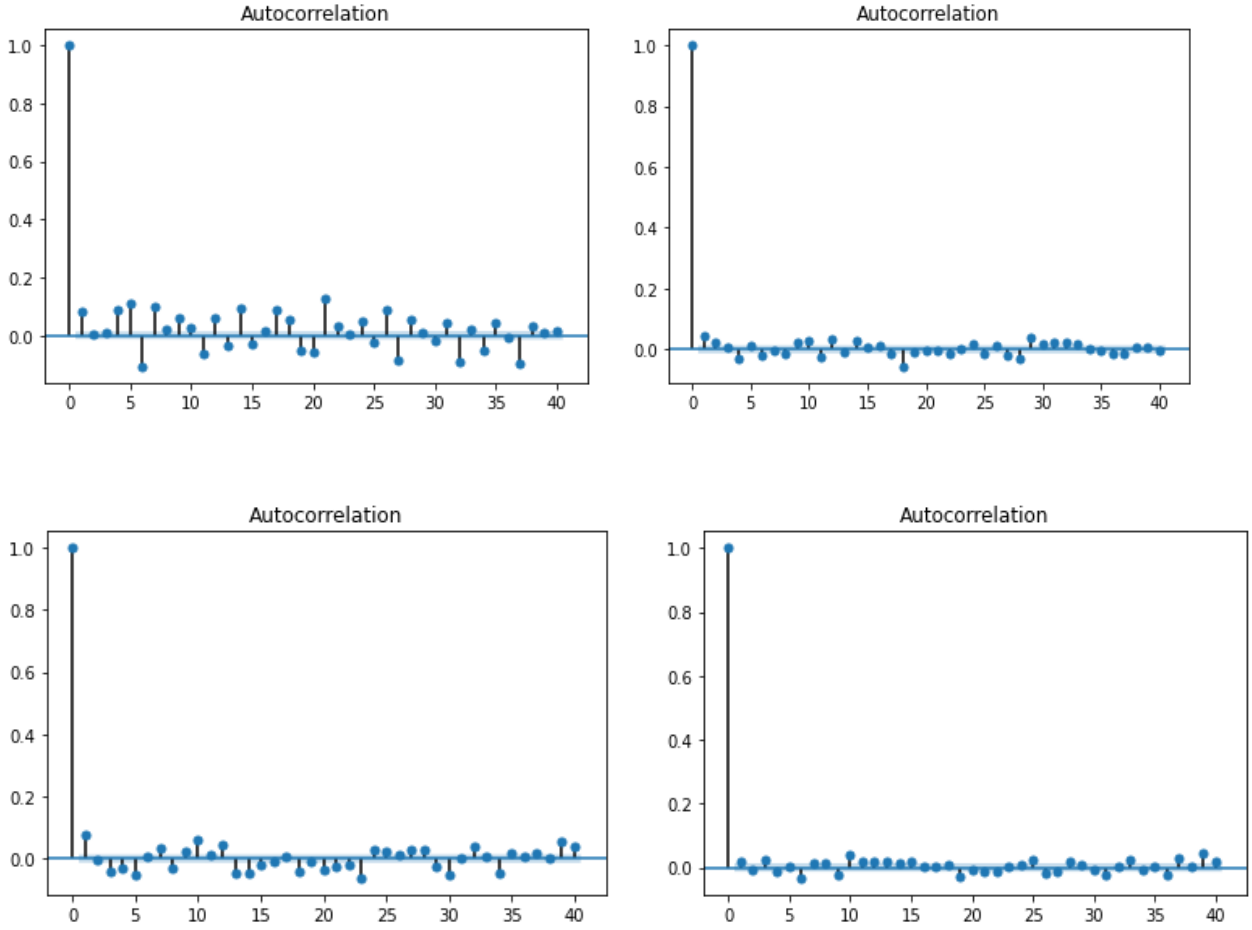


Exhibit 4. Stocks and Oil Price autocorrelation plot. From right to left A) Oil. B) JKS. C) AAL. D) TSLA

We can see a tendency on the Oil Price and the stocks of being order zero. This will give me a good initial idea of which order I should be looking on the times series analysis.

Modeling

On this section, I am explaining the different models I used to predict how stocks would be moving in the near future in order to make trading decisions.

Before I started performing models, I divided the dataset into training and testing dataset. The train dataset is used to create the model, and the test dataset to measure the performance. As an important mention, the mean squared error is the metric I used to measure how well my model was performing (Higher mean squared error means that my predictions are less accurate).
.
I made a constant model with the last stock price of the train data set, two types of regressions (Standard Linear Regression and Gradient Boosting Regression), and also performed two times Series Analysis (ARIMA and SARIMA models).

Train and Test datasets

Since this is a times series analysis, I divided the train and test dataset by taking into consideration that both should follow an order in time. The first 80% of the data would be used as train dataset (the data I will use to create the model) and the remaining 20% of the data would be used as the test dataset (the data I will use to evaluate how well is my model).

Models

1. **Constant Model**

Firstly, I made a constant model by using the last stock price of my train dataset.

At this model, I am not able to establish any relationship between the oil price and the stock prices. I am just inferring that moving forward, the stock price will not change. This will give me an idea if the stocks behavior were stable during the last years.

The comparison between the test data set and predictions for the three stocks are shown below:

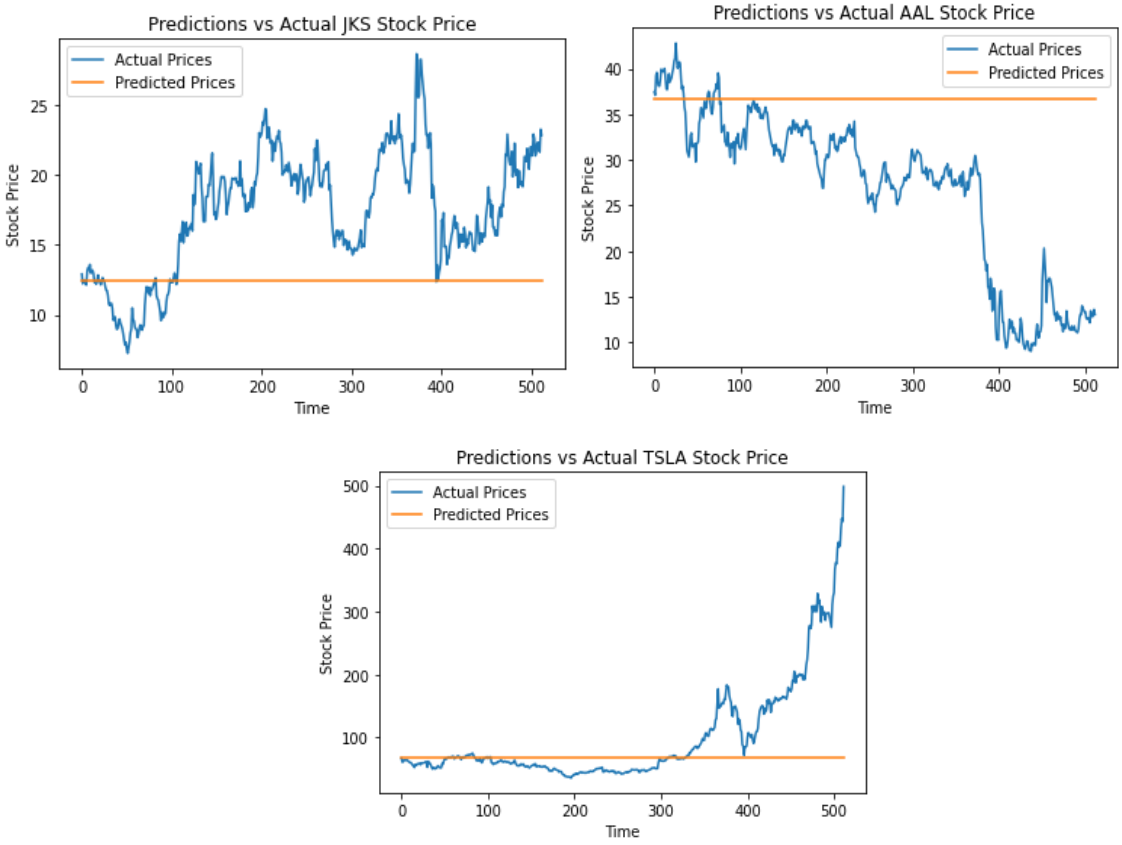


Exhibit 5. Predictions of the constant model and Actual Price comparison between JKS, AAL and TSLA Stocks.

In this particular model, it is not necessary to plot the forecast for the stocks because we are assuming the stock prices will continue having the same price than the last date of the train dataset. In the following table, I am computing the mean squared error for each stock:

Stock	**Mean Squared Error**
JKS	42.87
AAL	179.19
TSLA	1755.65

Table 1. Mean Squared Error of each stock for the Constant model

Insights

- The constant model is not the best model to anticipate if the stock price is going to increase or decrease. Though, this model is giving me a start point to know if Regression Models or Time Series Models will add or not value to my predictions by comparing their outcomes with this constant model.

- With the constant model, I can figure out which is the most and the least volatile of our stocks. Based on the results, I can conclude that JKS stocks is the one with the least volatility because we got the lowest mean squared error, and TSLA stock is the most volatile since I got the highest mean squared. This volatility would give me an idea which is the riskiest and the least risky stocks to invest as well.

2. Regression Models

For the regression models, I also divided the data into a train and test datasets. The percentages of data used to train and test the data are the same than the constant model. For the train dataset I used the first 80% of the data and to test the data I used the last 20% of the data.

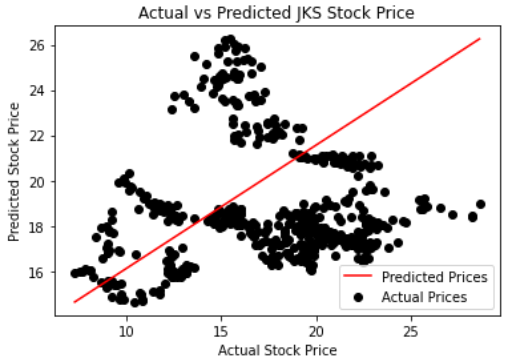
For the regression models, I had to do some preprocessing before making the models.

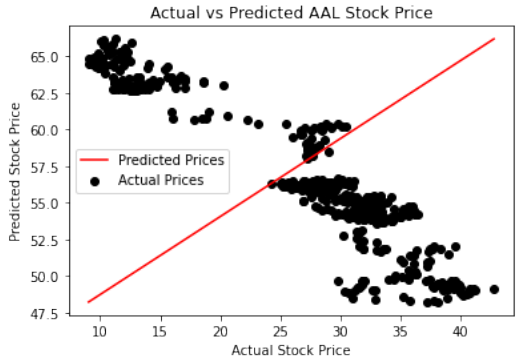
Firstly, I convert the datetimes into categorical features in order to use this date data feature in the statistical analysis.
Secondly, I needed to specify the independent variables (in this problem, those variables were year, month, date, and oil price), and the dependent variable (for each stock, the dependent variable is its price). Additionally, I standardized the data, so I was able to compare the model performance with and without standardization.

After completing the preprocessing of the data, I decided to run two different types of regressions; Standard Linear Regression and Gradient Boosting Regression for each of the stocks.

2.1 Standard Linear Regression

The results of the Standard Linear Regression without Standardization for the stocks are shown below:





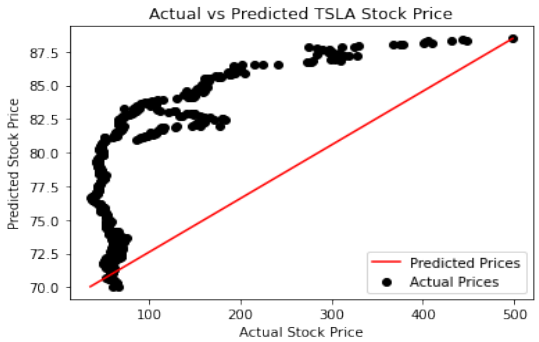


Exhibit 6. Predictions of the Standard Linear Regression and Actual Price comparison between JKS, AAL and TSLA stocks without standardization

The results of the Standard Linear Regression with Standardization for the stocks are shown below:

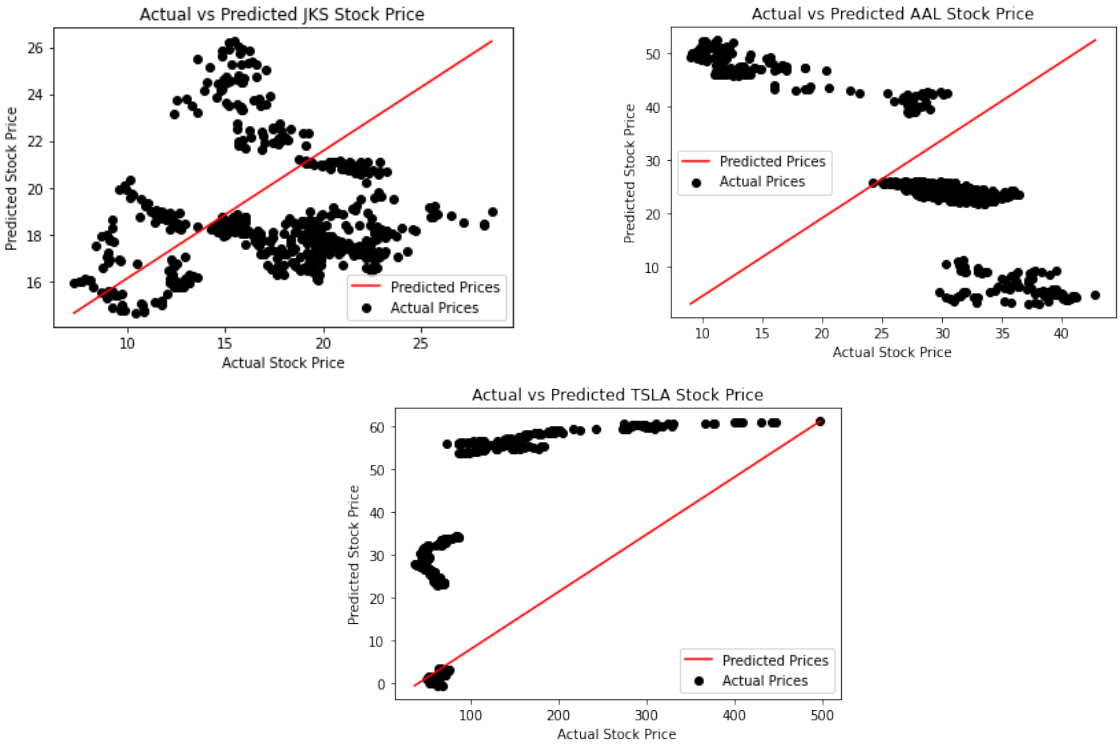


Exhibit 7. Predictions of the Standard Linear Regression and Actual Price comparison between JKS, AAL and TSLA stocks with standardization.

In the following table I am computing the mean squared error for the Standard Linear Regression models with and without standardization for the stocks.

Stocks	**Mean Squared Error without Standardization**	**Mean Squared Error with Standardization**

JKS	30.82	26.20
AAL	1080.34	520.90
TSLA	6511.29	9382.89

Table 2. Mean Squared Error of each stock for the Standard Linear Regression model with standardization

Insights

- Standardization improves the performance of the model depending on the stock behavior. For stocks with low volatility, it actually reduces the mean squared error, and for highly volatile stocks, standardization does not actually help.
- These models are implying the stock price will constantly increase as long as time passes. This is not a safe thought since stock tend to be stable, increase, or decrease its price.
- Volatile stocks such as TSLA had higher mean squared error than the stocks with lower volatility such as JKS and AAL. For regressions, it is easier to make the model if there are not huge changes between prices day after day. This also explains why TSLA had higher mean squared error; and the straight line is very separated from the actual values.
- If we compare the mean squared error of the constant model and the Standard Regression Linear model, I am not adding value to my predictions by doing a Linear Regression if we talk about American Airlines or TSLA. We could believe the regression model is adding value to JKS, but since the regression establish the stock price would always increase moving forward, this makes the model very poor at the time to predict stock prices.

2.2 Gradient Boosting Regression

For the Gradient Boosting Regression, firstly I checked out for the best parameters for each stock to run the regressions. The way I did so was by doing a Grid Search Cross Validation. In the next table I will be showing the parameters used for each stock.

Stock	**Standardization**	**Learning Rate**	**Number of estimations**	**Max Depth**	**Tolerance**
JKS	NO	0.1	100	3	0.0001
	YES	0.1	100	3	0.0001
AAL	NO	1.0	100	5	0.01
	YES	1.0	100	5	1.0
TSLA	NO	0.1	100	5	1.0
	YES	0.1	100	5	0.01

Table 3. Parameters used for the Gradient Boosting Regressor with and without standardization.

As an important mention, it normally does not matter if the data is whether standardized or not. I got the same hyperparameters in most of the parameters in both scenarios and just a slight difference on the tolerance hyperparameter.

Now talking about the outcomes, the results for the Gradient Boosting Regression of each stock without Standardization are shown below:

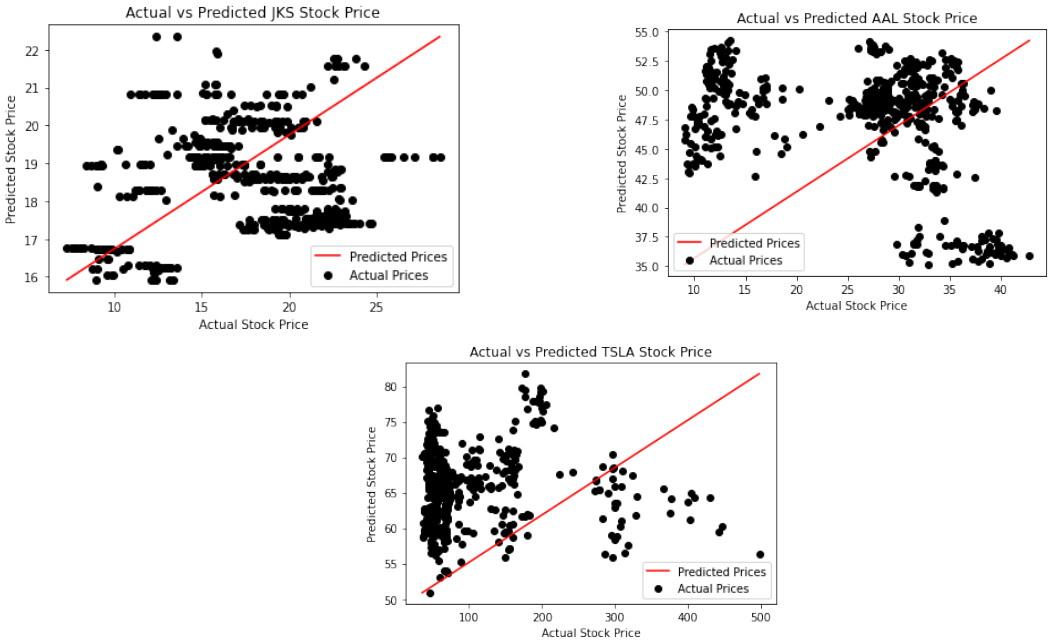


Exhibit 8. Predictions of the Gradient Boosting Regressor and Actual Price comparison between JKS, AAL and TSLA stocks without standardization.

The results for the Gradient Boosting Regression of each stock with Standardization are shown below:

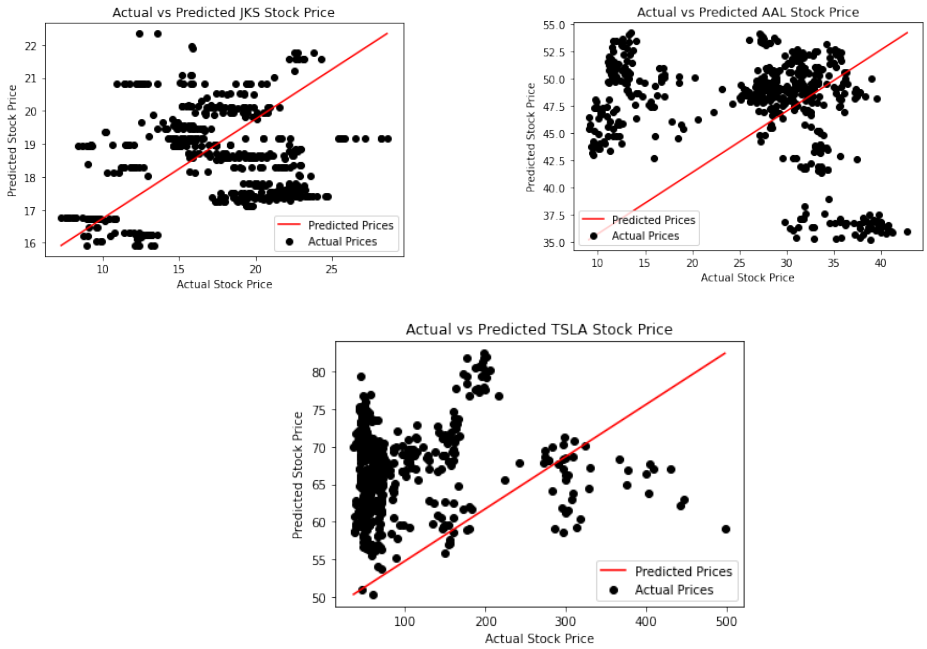


Exhibit 9. Predictions of the Gradient Boosting Regressor and Actual Price comparison between JKS, AAL and TSLA stocks with standardization.

In the following table, I am inputting the mean squared error for the Gradient Boosting Regressor with and without standardization.

Stocks	Mean Squared Error without Standardization	Mean Squared Error with standardization
JKS	20.23	20.27
AAL	548.56	548.55
TSLA	7862.27	7875.37

Table 4. Mean Squared Error for the stocks by using Gradient Boosting Regression with and without Standardization

Insights

- Gradient Boosting Regression is a very poor way to predict the stock prices. Even though we saw an improvement on the mean squared error, it was not significant and the believe the stock price will always increase is not applicable.
- Additionally, if I look at the graphics, the equation the Gradient Boosting Regressor implies does not follow the general trend of the stock prices. Therefore, I should be looking for another model to predict the stock prices as function of the time and oil price.
- Gradient Boosting Regressor still has the same problem than the Standard Linear Regression. For very volatile stocks such as TSLA, the performance is really poor because the model cannot find a proper way to get together all the dots.

Therefore, I decided to also perform Times Series Analysis.

3. Times Series Analysis

Before getting in to the outcomes of the Times Series Analysis, I am explaining some preprocessing I did before running the models.

Firstly, I needed to check if the data was whether stationary or not. Since it originally was not, I converted into stationary data by using a logarithmical function.

Secondly, I divided into a train and a test dataset. I used the data from June the 6th of 2010 to December the 31st of 2019 as the train data set and the data from January the 1st to August the 31st of 2020 as the test dataset.

After doing the preprocessing mentioned above, I went ahead and run the ARIMA and SARIMA models for the oil, and the stock prices.

3.1 ARIMA

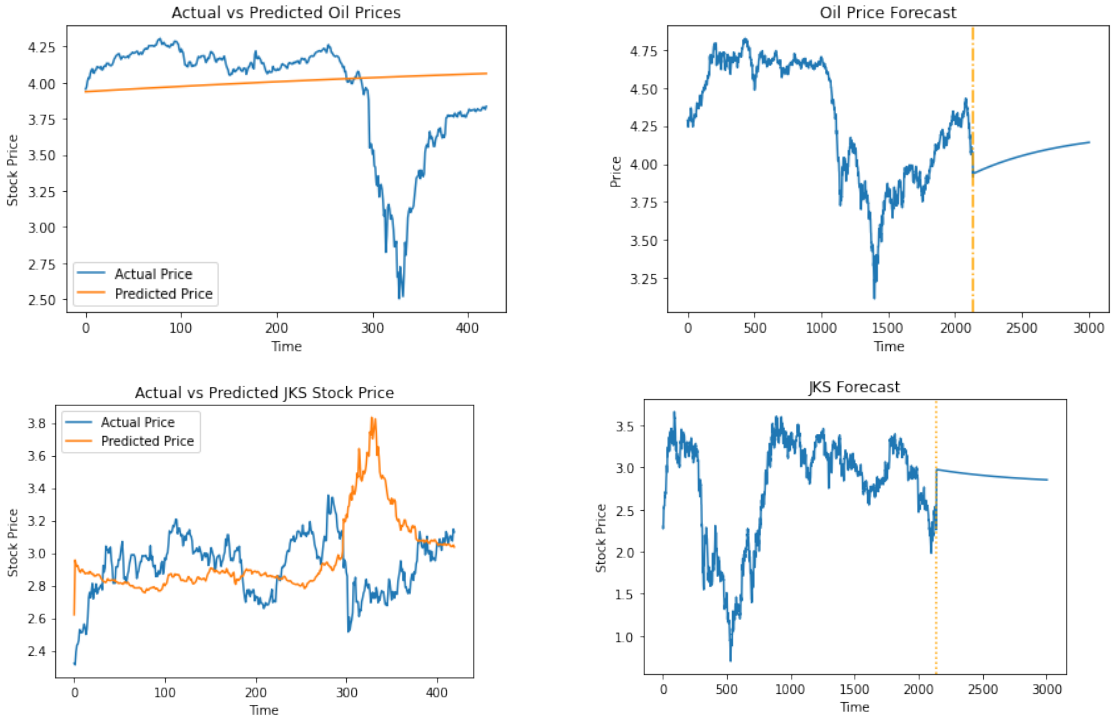
For the ARIMA model, I made a Grid Search Cross Validation in order to know the best hyperparameters. The best hyperparameters are chosen by taking just into consideration the mean squared error and how well the model follows the general trend of the test dataset.

The orders for the Oil and Stock prices are shown in the table below.

Stock	**Order**
Oil Price	(1,0,2)
JKS	(0,0,1)
AAL	(0,0,1)
TSLA	(0,0,1)

Table 5. Hyperparameters used for the Oil and the stocks prices for ARIMA model.

In the following Exhibit, the comparison between the predicted values and the actual values for the ARIMA model are shown below with the forecast.



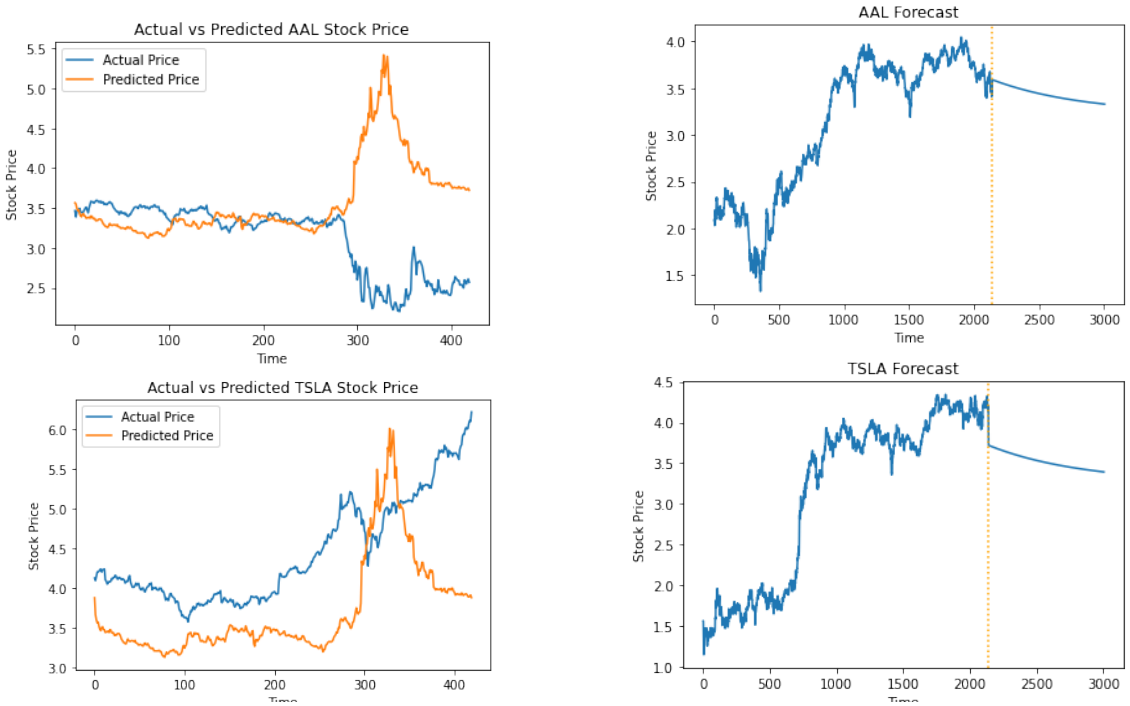


Exhibit 10. Predictions of the ARIMA and Actual Price comparison between Oil Price, JKS, AAL and TSLA stocks for the ARIMA models.

Same than the ARIMA model, I did a Grid Search Cross Validation in order to find the best hyperparameters to use in the model, and I followed the same patrons of the ARIMA model.

The orders and the seasonal orders hyperparameters are shown below:

Stock	**Order**	**Seasonal Order**
Oil Price	(0,1,1)	(0,1,1,2)
JKS	(0,1,1)	(0,1,1,2)
AAL	(0,1,1)	(0,1,1,1)
TSLA	(0,1,1)	(0,1,1,2)

Table 6. Hyperparameters used for the Oil and the stocks prices for SARIMA model.

In the following Exhibit, I can see the comparison between the predicted values of the SARIMA model and the forecast for the oil and the stocks prices.

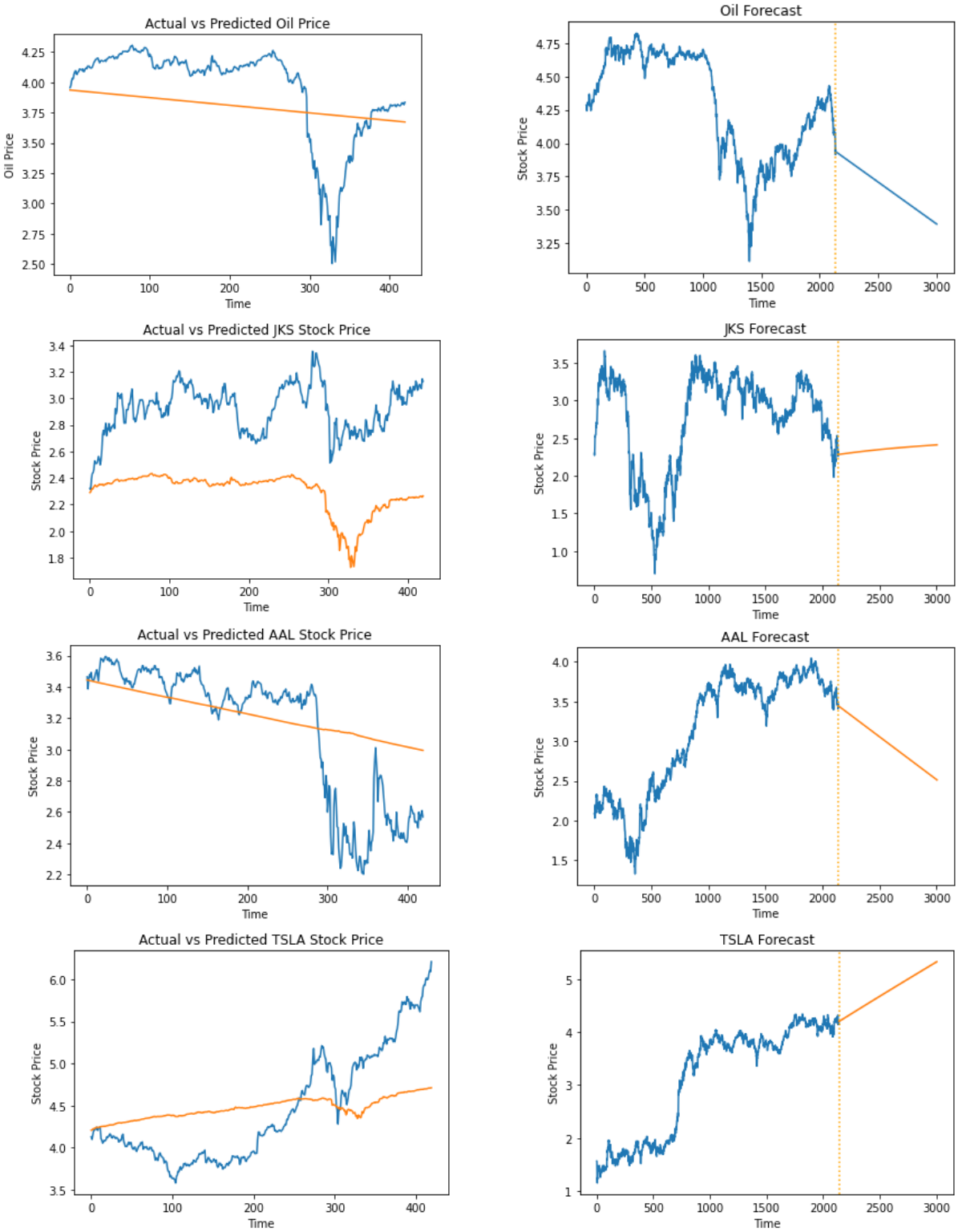


Exhibit 11. Predictions of the ARIMA and Actual Price comparison between Oil Price, JKS, AAL and TSLA stocks for the SARIMA models.

In the following table, I will show the mean squared error for the model:

Stocks	**ARIMA Mean Squared Error**	**SARIMA Mean Squared Error**
Oil Price	0.35	0.78
JKS	0.47	0.78
AAL	1.30	1.45
TSLA	2.49	8.02

Table 7. ARIMA and SARIMA Mean Squared Error for the Oil and Stocks price.

Insights from the times series models.

- Times Series model had a better performance than the regression, and constant models. We can support this because of the lower mean squared error and the models follow better the general trend of the stock prices.
- To make my conclusions if a stock price is going up or down, I am taking both (ARIMA and SARIMA) models into consideration.
- I can see some differences between ARIMA and SARIMA model predictions. While ARIMA is telling me oil price would decrease, SARIMA is telling me oil price would increase. This is when I have to make my conclusions and know which model I should rely more. I can also get some valuable information from the last trends, and not just let my thoughts go for the models with the least mean squared error.

Final Insights

My final insights will focus on which stocks if I were an investor should buy/sell or hold based on the outcomes of my models. Besides, I will share my conclusions about each stock.

JKS

- For JKS. ARIMA says the stock price may decrease and SARIMA says stock price may appreciate in the close future. According to both models, the increase/decrease is going to be very controlled which make this a very safe investment opportunity.

AAL

- This a very interesting stock since this is a company that was highly affected because of COVID-19. The American Airlines stock price has fallen down during the last months. However, it is safe to assume in the future the stock price may increase again because the world would eventually recover from COVID-19 and additionally, the normal flights will

return to normality. However, due to the last tendency, ARIMA and SARIMA are saying that the stock price may depreciate which makes this stock a very risky investment.

TSLA

- TSLA is maybe the most difficult stock to analyze because of the last company behavior regarding the stock price. The stock price has hugely increased the stock price during the last months, and SARIMA supports this last tendency. However, ARIMA says that stock price may decrease. I would say the stock price may continue appreciating since it has a momentum, but it is likely to be controlled and not dramatically as it had occurred during 2019 and 2020.

Recommendations

- JKS may be the best option to invest if you want a safe investment. American Airlines and Tesla presents more unexpected behavior.
- It is still important to mention the influence between the oil prices and the stocks behavior. JKS is not very correlated, but a decrease in oil price suppose an increase on the stock price. Now, if oil price decrease, American Airline's stock price decrease as well because they are highly correlated. Finally, Tesla and Oil Price have negative correlation between them, so a decrease in oil price suppose an increase on Tesla stock price.
- If oil price continues going down, it would be a better option to invest in Tesla or JKS rather than American Airlines. If oil price goes up, American Airlines may be the best option to invest in.
- Finally, we can get more accurate models if we introduce more variables for the study. To name few; electricity price for JKS and TSLA, value of currencies, population growth, etc.