FORECASTING OF STOCK CLOSING PRICE



Table of Contents

[Introduction and Motivation - 3 -](#_Toc531889240)

[Data Description - 3 -](#_Toc531889241)

[2001 ~ 2018 Adj Closing Price Trend - 4 -](#_Toc531889242)

[Closing Price Comparison Before and After Economic Crisis - 4 -](#_Toc531889243)

[Methodology: Description of Procedures - 5 -](#_Toc531889244)

[Regression Model - 5 -](#_Toc531889245)

[Regression Model 1 - 6 -](#_Toc531889246)

[Regression Model 2 - 6 -](#_Toc531889247)

[Other Models - 9 -](#_Toc531889248)

[Seasonal Naive Model - 9 -](#_Toc531889249)

[Evaluation of Four Models - 10 -](#_Toc531889250)

[Four models Prediction Accuracy - 10 -](#_Toc531889251)

[Residuals Comparison - 11 -](#_Toc531889252)

[Presentation of Closing Price Forecast in 2019 - 13 -](#_Toc531889253)

[Quantify the Impact of Economic Crisis - 14 -](#_Toc531889254)

[Conclusion and Actionable Recommendations - 15 -](#_Toc531889255)

# Introduction and Motivation

Predicting the behavior of stock prices has always been a popular and widely studied subject. By analyzing the pattern of stock prices, we can get a general understanding of economic development status, and investigate the impact of economic crisis, recession or expansion on stock prices; investors in the stock market can maximize their profit by buying or selling their investment. S&P 500 is a commonly used measurement to indicate stock price.

The S&P 500 Index (formerly Standard & Poor's 500 Index) is a market-capitalization-weighted index of the 500 largest U.S. publicly traded companies by market value. The index is widely regarded as the best single gauge of large-cap U.S. equities. Other common U.S. stock market benchmarks include the Dow Jones Industrial Average or Dow 30 and the Russell 2000 Index, which represents the small-cap index. The S&P 500 Index differs from other indices because of its diverse constituency and weighting methodology. It is one of the most commonly followed equity indices, and many consider it one of the best representations of the U.S. stock market, and a bellwether for the U.S. economy.

The objective of our report is to analyze historical S&P 500 index from 2001-2018, explore the trend and pattern of the index over time and make prediction for the future using time series methods. The whole process includes: build four time series models -- regression, Seasonal Naive model, Smoothing and ARIMA model; divide the dataset into training and testing data based on two partition method and run four models respectively; choose a champion model with the smallest average validation MAPE in two scenarios; use this model to forecast the S&P 500 Index in the following year.

As a result, regression is chosen to be our champion model. The model predicts that S&P 500 index would show an upward trend from December 2018 to December 2019. Apart from the variables included in the dataset, other macroeconomic factors may also be useful in predicting S&P 500 Index, because the riskiness of firms is to a large extent determined by the firm's’ exposure to macroeconomic risks. Some possible variables include Gross Domestic Product (GDP), Consumer Price Index (CPI), unemployment and interest rate.

# Data Description

The dataset we use is S&P 500 Index downloaded from Yahoo Finance (<https://finance.yahoo.com/quote/%5EGSPC/history?period1=975484800&period2=1543478400&interval=1mo&filter=history&frequency=1mo>). The dataset collects monthly S&P 500 index from 1/1/2001 to 11/1/2018. It contains 215 records and 7 columns, i.e., Date, Open, High, Low, Close, Adjusted close and Volume. The frequency of data is 12. Adj close is chosen as Y, the dependent variable that we aim to predict. The goal of our time series analysis is to explore the patterns of Adj close and the relationship between different variables.

Apart from the existing variables provided by the dataset, we also include the following possible variables which might be helpful to predict y: trend, ramp, recession(0/1 dummy variable), month, and lags.

## 2001 ~ 2018 Adj Closing Price Trend



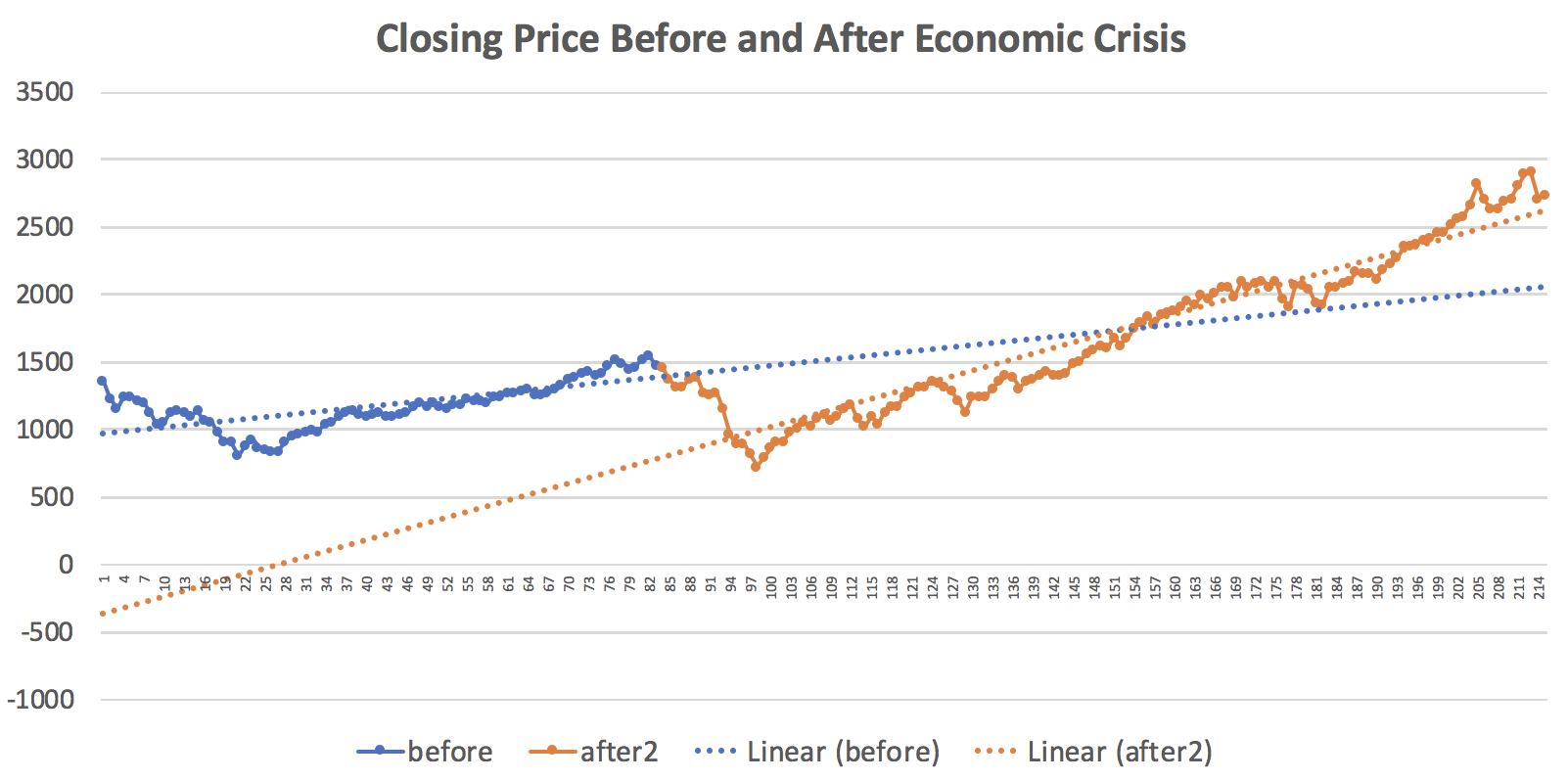
As is shown in the line chart, the Adj close price decreased from January 2001 to March 2003 and increased steadily after that. The close price started to drop sharply from December 2007 and reached its lowest point at February 2009. Then, there was an overall upward trend between February 2009 and November 2018. Based on the observations above, there is enough reason to doubt that 12/1/2007 is a month of long-term stock price change. Two-sample t test is used to quantitatively assess the significance of this abrupt change.

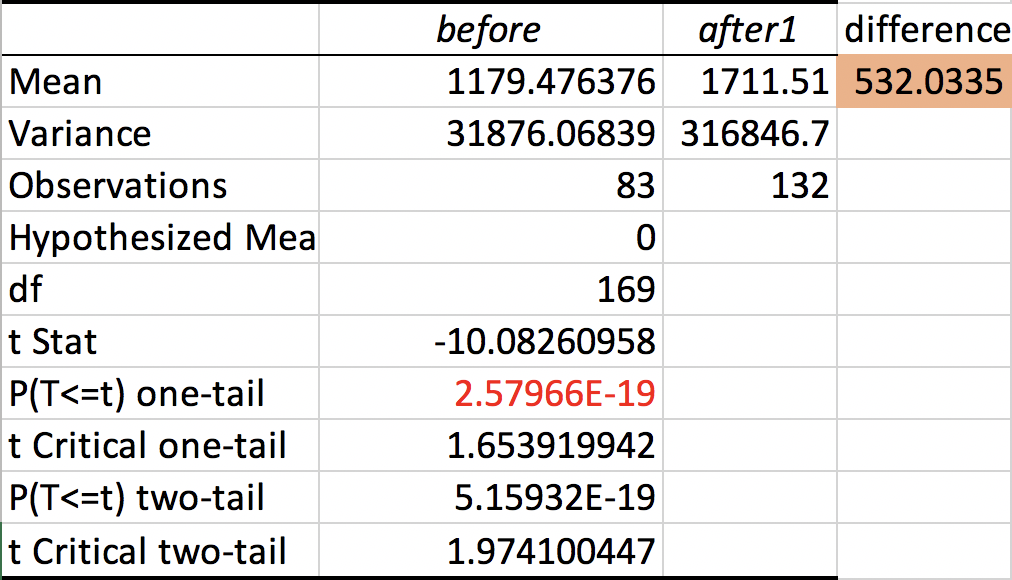
## Closing Price Comparison Before and After Economic Crisis

t test is applied to confirm that the average closing pricing before and after economic crisis had a big difference.

H0: No long-term price change after 12/1/2007

Ha: There is a long-term price change after 12/1/2007





As is shown in the output, the p-value is smaller than alpha(0.05). Therefore, we successfully reject H0, which means there is a long-term close price change after 12/1/2007. The average of Adj close price before 12/1/2007 is 1179 and the average after 12/1/2007 is 1711. The mean increases by 45% and the difference is 532. Therefore, it is statistically significant that the economic crisis had impact on the closing price of stock market.

# Methodology: Description of Procedures

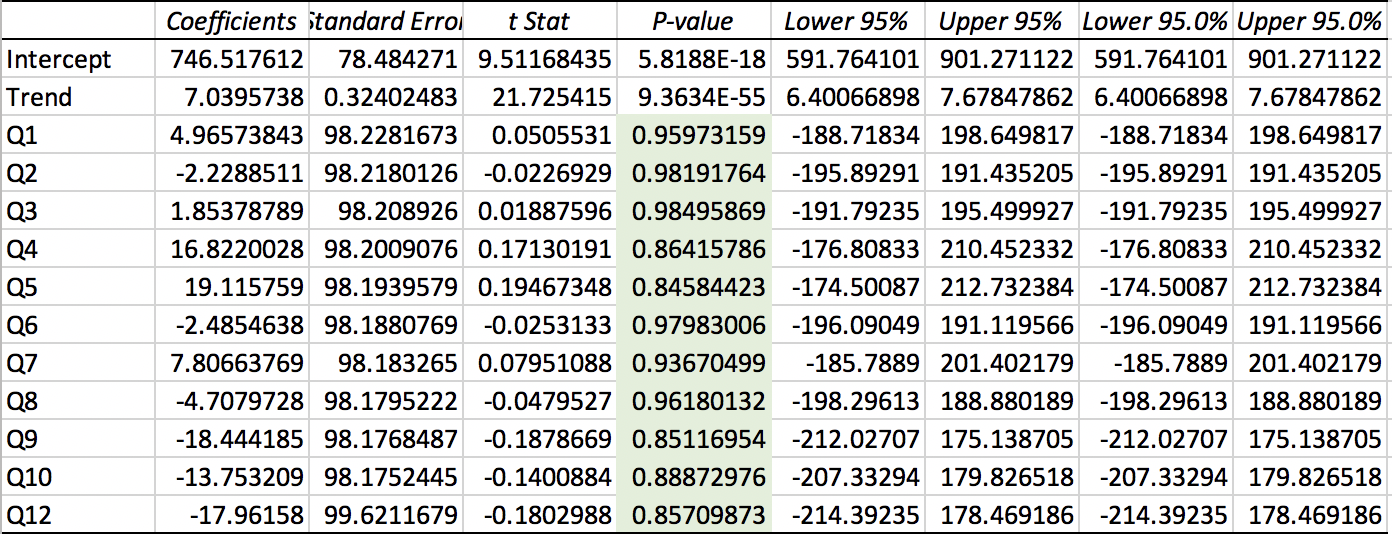
## Regression Model

We built two regression models with different independent variables based on two scenarios and selected the best regression model based on the Validation MAPE.

### Regression Model 1

We tried to build regression using time series forecasting techniques. This regression included the independent variables of “Trend”,11 “Monthly Dummies”. However, from the table of regression results below, we can find that p values for all 11 monthly dummies are larger than 0.05. That means at the confidence level of 95%, Monthly Dummies are not statistically significant in predicting the closing price. We may eliminate the Monthly Dummies from the regression and try to include other relevant indicators.

#### Regression Model 1 Results



### Regression Model 2

Our second regression model includes independent variables “Trend”, “Ramp”,”Recession” and other finance indicators, such as Open,High,Low.The description of these independent variables are shown as below.

|  |  |
| --- | --- |
| Independent Variables | Description |
| Trend | Periods counting from beginning point. |
| Ramp | Number of period after recession(0 indicate pre-recession). |
| Recession | Dummy variable. 1 indicate post-recession, 0 indicate pre-recession. |
| Open | Open price of the stock on a given day |
| High | Highest price of the stock on a given day |
| Low | Lowest price of the stock on a given day |

To better test the relevance of regression model for different time periods, we use two scenarios to build the regression with the indicator above.

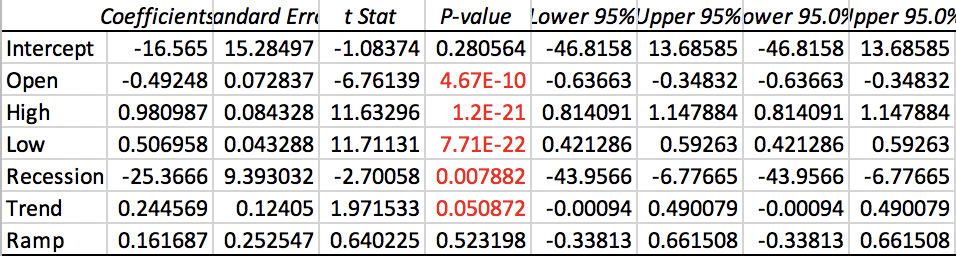
*Scenarios with different training periods and testing periods*

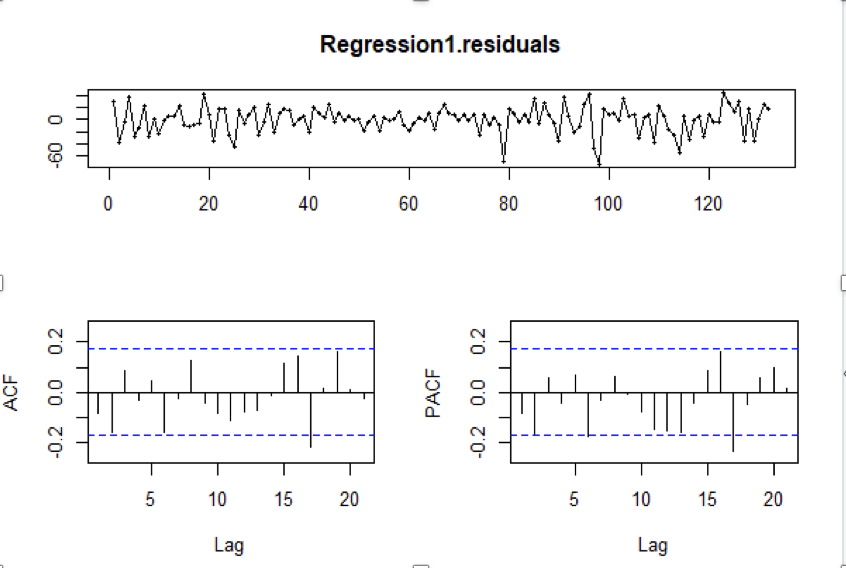
|  |  |  |
| --- | --- | --- |
|  | Training Dataset | Testing Dataset |
| Scenario 1 | 2001~2007 | 2007~2008 |
| Scenario 2 | 2001~2017 | 2017~2018 |

#### Regression Model 2, Scenario 1

We use 2001~2007 as training dataset to build the regression and 2007~2008 testing dataset to test the model. From the regression results in table below, we can find that the p-value of variables of “Open”, “High”, “Low”, “Recession”, “Trend” are less than or nearly equal to 0.05. That means at the confident level at 95%, these independent variables are statistically significant in predicting closing price. Moreover, from the result of diagnose test, we can find that the residuals look like white noise, which means there are no dependent pattern. Then we decided not to include lags in our regression model.

*Regression 2, Scenario 1: training dataset: 2001~2007; testing dataset:2007~2018*

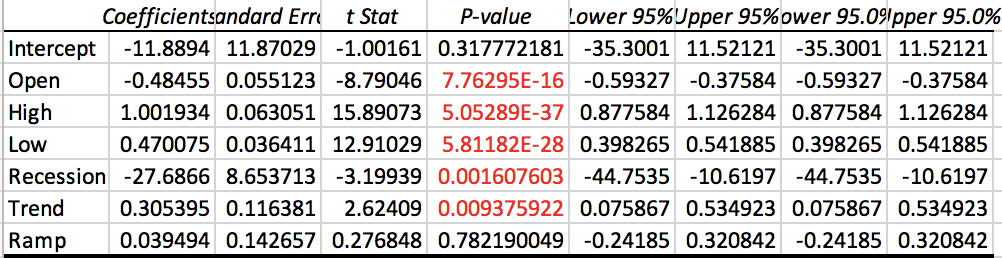


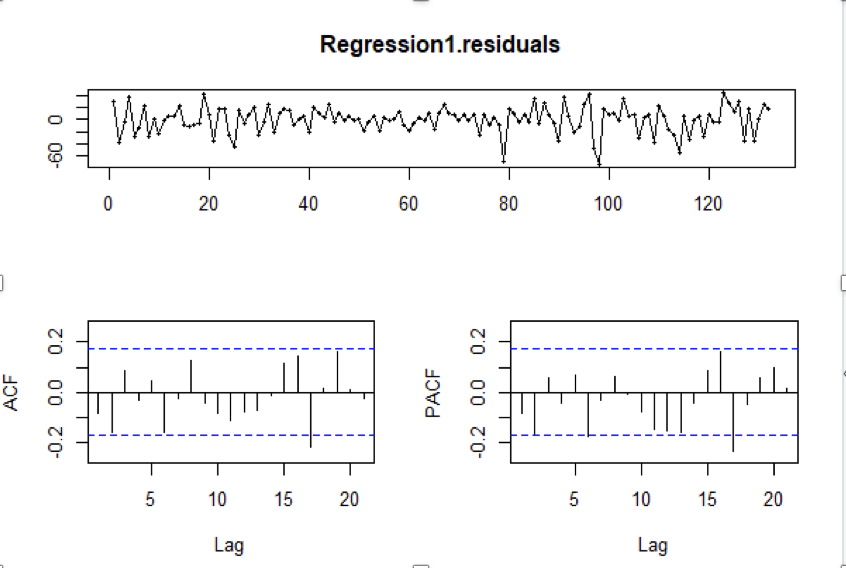


#### Regression Model 2, Scenario 2

We use 2001~2017 as training dataset to build the regression and 2017~2008 testing dataset to test the model. From the regression results in table below, we can find that the p-value of variables of “Open”, “High”, “Low”, “Recession”, “Trend” are less than 0.05. That means at the confident level at 95%, these independent variables are statistically significant in predicting closing price. Moreover, from the result of diagnose test, we can find that the residuals look like white noise, which means there are no dependent pattern. Then we decided not to include lags in our regression model.

*Scenario 2: training dataset: 2001~2017, Nov; testing dataset: 2017, Dec~2018, Dec*





Most of the independent variables in regression 2 are significant, and both of two scenarios have good performance in diagnose test. So we can get the conclusion that regression 2 with trend, ramp, recession and other finance indicators is the best regression model we can get.

## Other Models

We also tried to build Seasonal Naive model, Smoothing model and ARIMA model to predict the closing price of the stock using Scenario 1 and Scenario 2.

### Seasonal Naive Model

Seasonal Naive Model is the estimating technique in which the last period’s actuals are used as this period’s forecast, without adjusting them or attempting to establish causal factors. Seasonal Naive Model works remarkably well for many economics and financial time series.

#### Smoothing Model

Forecasting in Tableau uses a technique known as exponential smoothing. Forecast algorithms try to find a regular pattern in measures that can be continued into the future.

#### ARIMA Model

ARIMA models are, in theory, the most general class of models for forecasting a time series which can be made to be “stationary” by differencing (if necessary), perhaps in conjunction with nonlinear transformations such as logging or deflating (if necessary).

We use R to build the best models for Seasonal Naive model, Smoothing model, ARIMA model by choosing lambda="auto". Then we get Seasonal Naive model, Smoothing model with additive noise, additive damped trend and no seasonality, ARIMA(0,1,1) with no lag order, first degree of differencing, first order of moving average and ARIMA(2,2,3) with 2 lag orders, second degree of differencing, third order of moving average.

## Evaluation of Four Models

Furthermore, we use MAPE (Mean Absolute Percentage Error) of testing dataset as validation MAPE to compare the performance of these models. The table below summarizes the calculated validation MAPE for each scenario for each model. We can find that the MAPE of regression model b(trend, ramp, recession and other finance indicators) is the top model with lowest average validation MAPE, which is 2.5%. That means the regression model b has the highest prediction accuracy of a forecasting in statistics based on MAPE.

### Four models Prediction Accuracy

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | Prediction Accuracy | | |  |
| Model | Parameters/explanatory variables Technical Comments | Scenario1    Validation MAPE | Scenario2    Validation MAPE | Average    Validation MAPE |  |
| Naïve |  | 34.30% | 11.84% | 46.14% |  |
| Regression model 2 | Linear Regression with Open, High, Low, Recession, Trend, Ramp | 1.24% | 1.26% | 2.5% |  |
| Smoothing | ETS (A, Ad, N) | 35.26% | 5.94% | 41.2% |  |
| ARIMA | ARIMA (0,1,1)  ARIMA (2,2,3) | 35.19% | 2.77% | 37.96% |  |

We also perform the diagnosis test for each model to examine whether residuals are white noise. The table below shows the results of residuals comparison for each model.

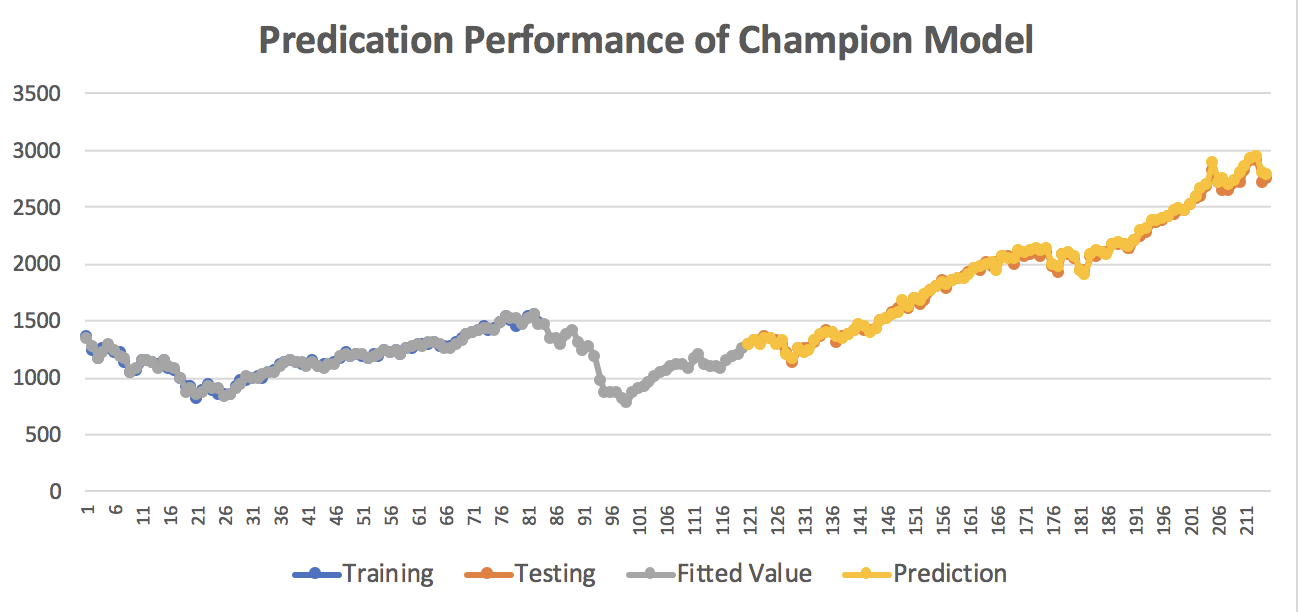
### Residuals Comparison

|  |  |  |
| --- | --- | --- |
| Model | Are residuals white noise? | Residuals Diagnosis |
| Naive | NO | No pattern looks random |
| Regression | Yes | White noise |
| Smoothing Model | NO | No pattern looks random |
| ARIMA | Yes | White noise |

From the table above, we can find that both of regression model b and ARIMA model have residuals which looks like white noise and no dependent patterns. That means both of regression model b and ARIMA model catch all lags.

Combine the results of average MAPE and Residual Diagnosis, we can get the conclusion that the regression model b with trend, ramp and economic indicators as independent variables is the champion model as it has lowest MAPE and residuals which are close to white noise.

The graph below shows the historical data in training dataset and testing dataset, fitted value and prediction using champion model. We can find that not only the fitted value if almost equal to the real value in training dataset, but also the prediction value is really close to the real closing stock price in the testing dataset. That means the champion model performs well.

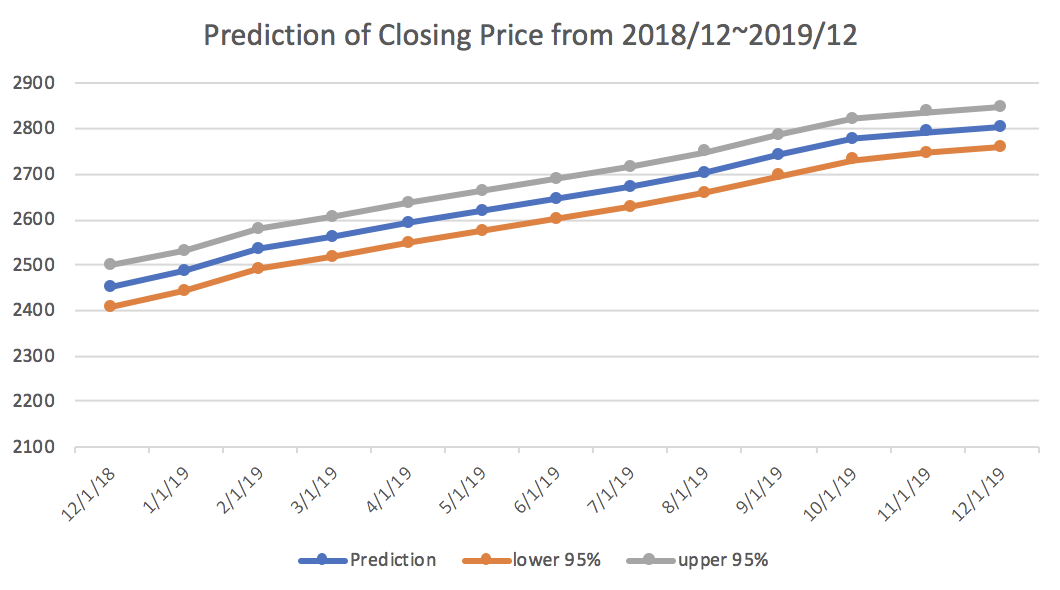


# Presentation of Closing Price Forecast in 2019

We use our champion model, which includes trend, ramp, recession and Open, high and low, to predict the future prediction for closing price in 2019. The table below shows the predicted value, 95% confidence interval with lower bond and upper bond.

|  |  |  |  |
| --- | --- | --- | --- |
| Date | Prediction | lower 95% | upper 95% |
| 12/1/18 | 2453.994 | 2409.41025 | 2498.578 |
| 1/1/19 | 2487.928 | 2443.34436 | 2532.513 |
| 2/1/19 | 2535.821 | 2491.23691 | 2580.405 |
| 3/1/19 | 2563.435 | 2518.85065 | 2608.019 |
| 4/1/19 | 2594.002 | 2549.41753 | 2638.586 |
| 5/1/19 | 2618.886 | 2574.30152 | 2663.47 |
| 6/1/19 | 2645.517 | 2600.93246 | 2690.101 |
| 7/1/19 | 2673.183 | 2628.59888 | 2717.767 |
| 8/1/19 | 2704.422 | 2659.83838 | 2749.007 |
| 9/1/19 | 2741.514 | 2696.92987 | 2786.098 |
| 10/1/19 | 2776.675 | 2732.09077 | 2821.259 |
| 11/1/19 | 2793.143 | 2748.55917 | 2837.727 |
| 12/1/19 | 2804.014 | 2759.43004 | 2848.598 |

We plot the chart graph to present the prediction of closing price with confidence interval. The prediction shows that the closing price of the stock will increases continuously from December,2018 at $2453.99 to December 2019 at $2804.01.

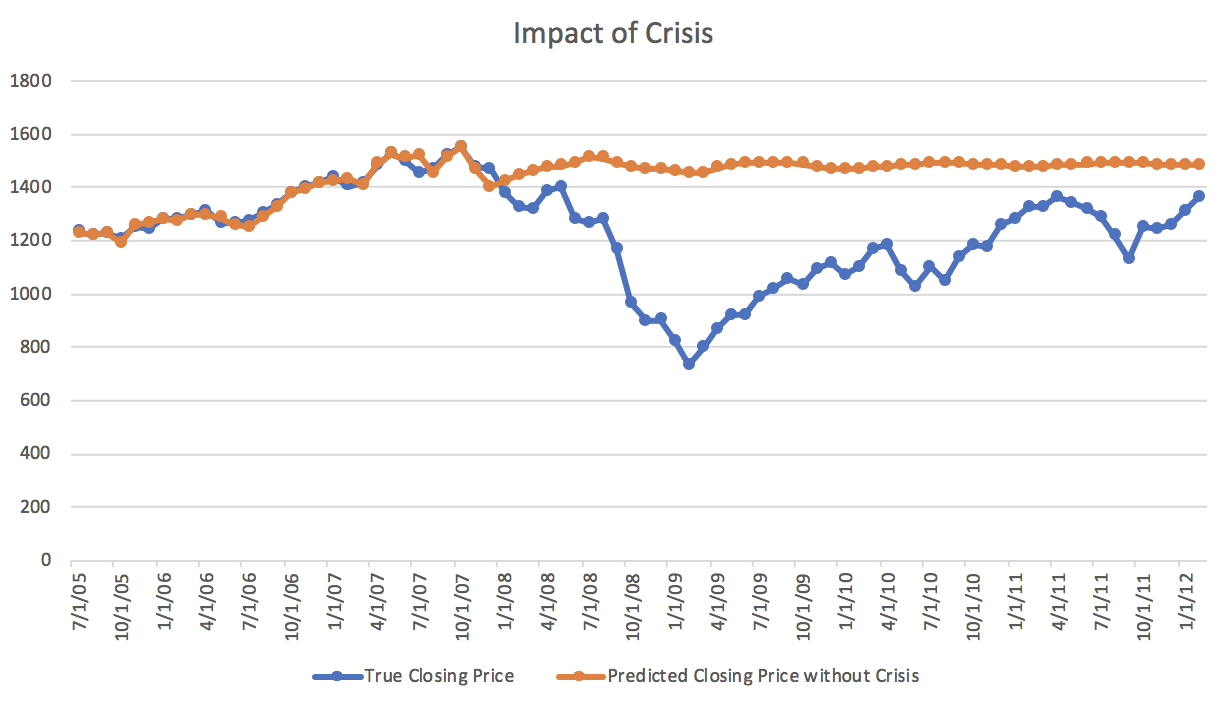


# Quantify the Impact of Economic Crisis

# 

Economics crises are very important events in the stock price history. The economics crisis of 2008 began in the US, rapidly took the form of a full blown systemic crisis in the US and almost immediately became a global phenomenon. The economics crisis and the associated decline of stock market capitalization has enormous impact on the stock market composition because recessions are associated with lower corporate earnings.

We used regression model to explore the quantitative impact of economics crisis on the closing price of the stock from 2007 to 2012. From the graph below, we can find that the stock price started drop due to economic crisis in December 2007. Moreover, the impact of economics crisis is largest in Jan 2009.



# Conclusion and Actionable Recommendations

We tried to build regression model A (includes monthly dummies), regression model B(includes trend, ramp, recession, finance indicators), Seasonal Naive model, Smoothing model, ARIMA model. We found that the regression model B (includes trend, ramp, recession, finance indicators) is the champion model as it has lowest MAPE and residuals in white noise pattern. The possible reason that regression model B performance best may because the regression model B includes recession dummy variables which could indicate the enormous impact of the economic crisis on the closing price.

Our champion model predicts that the **stock price would be increasing from December 2018 to December 2019. However, we recommend that the investors may also consider other risks when enter into the stock market.**

Predicting stock marketing is always a complicated job, with thousands of millions of earnings if predicted accurately. The novel application of this project is that people can use different models and account into various independent explanatory variables to predict future stock price and make instant decisions.