A Novel Approach for Stock Market Price Prediction Based on Linear Regression and ARIMA

Chetan Kumar

Abstract -

The study focuses on the application of R programming language towards stock price prediction which is a complex problem with no coherent solution. The stock market is a high risk and possibly high-profit entity and that is why predicting accurate stock prices could help address prevalent issues in the finance sector. Stock markets being highly stochastic in nature, predicting stock prices accurately poses an exceptionally difficult challenge. Hence all known models to predict the stock prices should be meticulously analyzed to provide state-of-the-art forecasting for further explorations; hence we will analyze different research papers which will include analysis of many learning methods like Linear Regression, ARIMA (Auto Regressive Integrated Moving Average) model and compare the results of the same. Performance measures are analyzed in this work with the TSX Composite Index using statistical methods in the R environment.

I. INTRODUCTION

S&P/TSX composite Index (Figure 1) is Canadian equity market's best-known index. It summarizes a lot of data, share price of 233 Canada's largest public companies. Most equity indexes are cap weighted. Where the cap is short for market capitalization. Indexes play an important role in the overall analysis of the equity markets. Indexes and their movements provide a great deal of insight into the economy, inversing the public's risk appetite and the trends in investing diversification. Understanding what it indexes is, making following the financial market more interesting and more fun.

The goal of this study is to determine the best model to predict the future stock Index price of the TSX Composite using linear regression and ARIMA using OHLC (Open, High, Low, Close and Volume) data. These features are used by various prediction models which can be seen in various research papers. Enormous research work has been concentrated on the feature prediction of stock prices based on historical prices and volume. To back our research and R code we will do some literature review.

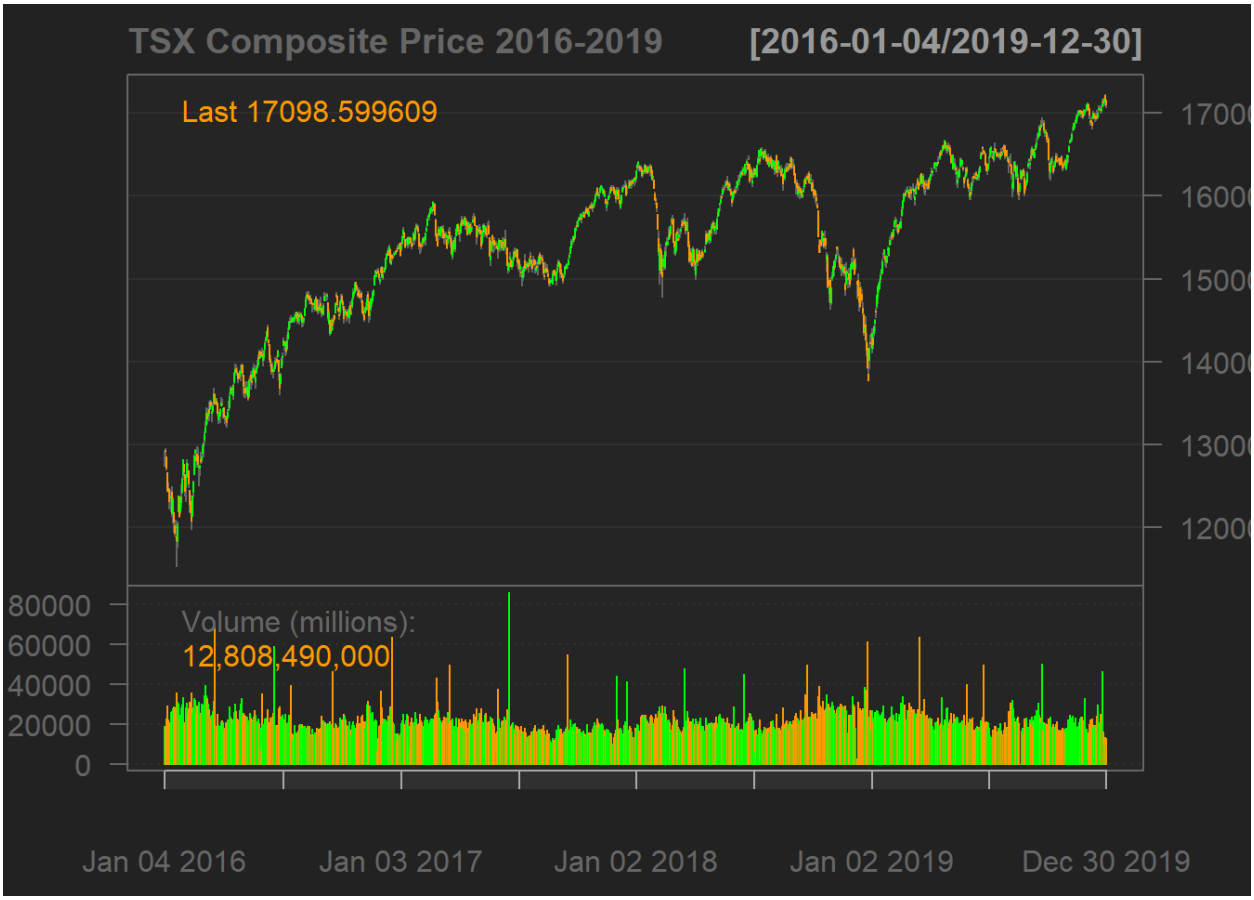

Figure 1. Bar Chart: TSX Composite Index close price

1.1 Financial data as a time series

Stock prices are treated as time series data. Stock market provides stock quotes or stock price such as Open, Close, Low, High and Volume etc., along with stock symbol and transaction date. These basic quotes give information such as high and low prices of stock in a day or its change in the value.

These financial indicators can be used directly in prediction models as a dependent or independent variable. Such as Close price is used as a dependent variable or label in prediction models (**Lin, Yang & Song, 2009; Rustam & Kintandani, 2019**). Moreover new features are also derived from the existing one such as gain in (**Garcia-Lopez, Batyrshin & Gelbukh, 2018; Mourelatos et al., 2018**)

1.2 Time series data

Overall, we can see that there are upward trends on the time series data and seasonal patterns in the plot (Figure 2). A time series consists of four components: Trend (T), Seasonal (S), Cyclic (C), and Irregular (I). All these four components are combined using additive or multiplicative models to form a time series (**Idrees, Alam & Agarwal, 2019**). To make sure of the components

of the data, we will decompose it. We will focus only on the closing price and change the periods to monthly data first.

The output shows four plots of our closing price data, which are:

- **Observed**: Original plot of the data
- **Trend**: Long term movements in the mean. In this plot, we can see the significant upward trend started around 2016.
- **Seasonal**: Repetitive seasonal fluctuation of the data. The closing price of the TSX composite Index tended to reach the highest in July and the lowest in December (Figure 3). Looking at this pattern, we can say that overall, the right time to sell this stock was at the middle of the year (especially in July) and the right time to buy was at the end of the year (especially in December).
- **Random**: Irregular or random fluctuation not captured by the trend and seasonal. The current situation of COVID-19 pandemic is an example of the factor that causes this random fluctuation. When a random component is dominant in a data, the forecast will be harder to be done accurately. Therefore, in this research I only use the data until 2019.

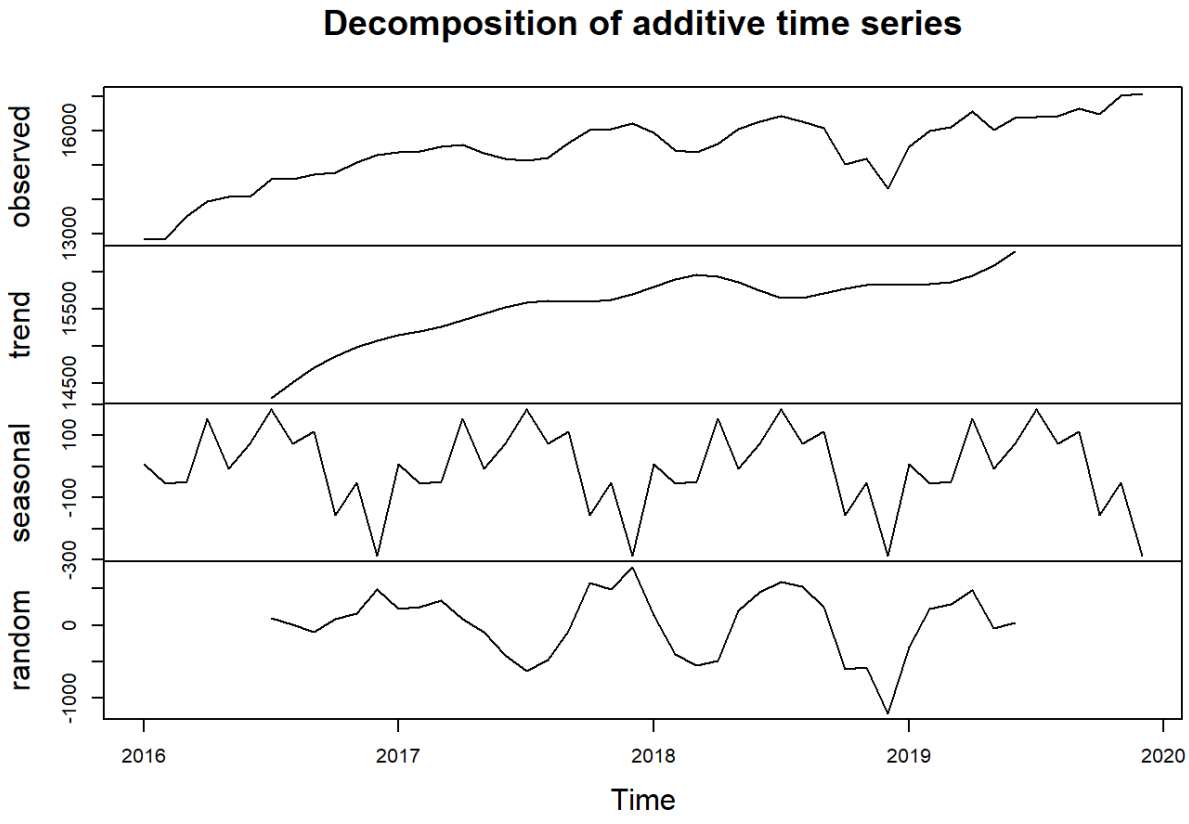

Figure 2. Decomposition of time series.

```
# Seasonal component
dc$seasonal
```

```
##              Jan         Feb         Mar         Apr         May         Jun
## 2016    8.263687  -55.147215  -48.836478  153.266319   -6.072155   72.657093
## 2017    8.263687  -55.147215  -48.836478  153.266319   -6.072155   72.657093
## 2018    8.263687  -55.147215  -48.836478  153.266319   -6.072155   72.657093
## 2019    8.263687  -55.147215  -48.836478  153.266319   -6.072155   72.657093
##              Jul         Aug         Sep         Oct         Nov         Dec
## 2016  183.922230   73.307023  113.863853 -156.771238  -51.338960 -287.114161
## 2017  183.922230   73.307023  113.863853 -156.771238  -51.338960 -287.114161
## 2018  183.922230   73.307023  113.863853 -156.771238  -51.338960 -287.114161
## 2019  183.922230   73.307023  113.863853 -156.771238  -51.338960 -287.114161
```

Figure 3. Seasonal output of the time series data.

1.3 The Efficient Market Hypothesis (EMH)

The Efficient Market Hypothesis (EMH) is a theory proposed by Professor Eugene Fama that states it is impossible to beat the market since market prices should only react to new information or changes in discount rates.

An efficient stock market is a market where stock prices reflect fundamental information about companies. In such a case, the market value of the company changes in a way very similar to that of the intrinsic value of a company. These changes are not consistent with the value and do not restrain from trading financial assets. The differences in investor awareness and uneven transaction costs prevent fundamental changes in value to be completely and immediately reflected in market prices. This means, a market is efficient when it is not possible to earn a return higher than the market return.

The efficient-market hypothesis suggests that stock prices reflect all currently available information and any price changes that are not based on newly revealed information thus are inherently unpredictable. On the same subject, the paper of **Alexandra Gabriela Titan in 2005 for "The Efficient Market Hypothesis: Review of Specialized Literature and Empirical Research"** concluded markets are not efficient and it is difficult to test and have a precise result. A hypothesis which can predict the closing price of a stock on a particular day can become very handy for successful trading.

1.4 Analysis of stock market predictor variables using Linear Regression

Selecting predictor variables for our model is the most important task. **Seethalakshmi "Analysis of stock market predictor variables using Linear Regression" in Volume 119 No. 15 2018,**

369-378 has compared two models. Model 1 with all features fitted with R2 value 0.997. This indicates open, high, low, volume and adj close are essential for predicting closing value accurately. Model 2 with open, high and low predict close value fitted with R2 value 0.992 . This indicates prediction of close value is not affected with adj close. This study reveals with open, high, low and volume itself enough for finding approximate prediction of close value. Therefore, we know that open, high, low and volume are the best predictors so we will use further statistical analysis like Welch's t-test on these to further find the best predictor variables in this research paper.

1.5 Stock Price Trend Prediction Using Multiple Linear Regression

Multiple Linear Regression is one of the versatile techniques used to predict the stock market. This technique is used by **Shruti Shakhla "Stock Price Trend Prediction Using Multiple Linear Regression "International Journal of Engineering Science Invention (IJESI), vol. 07, no. 10, 2018, pp 29-33**. The quantity to be predicted is usually referred to as the independent variable. The various factors which demonstrate a strong correlation with the independent variable are referred to as the dependent variables. The measure by which the dependent variable changes due to a unit change in the independent variable is known as the regression coefficient of that independent variable.

This model mathematically calculates a linear (straight line) relation between the dependent variable and every other independent variable. One of the popular uses of this model is considering Least Squares Regression. Least Squares Regression aims at calculating a best fit line by minimizing the residual sum of squares of the deviations of the predicted values from the corresponding data points. Squaring the deviations removes the complexity introduced in the model due to positive and negative values. The open of the market provides an insight of the variations in stock prices of the entire market. They have successfully implemented the linear regression model which produced a RMSE of 1.145040040250809.

1.6 A Prediction Approach for Stock Market using ARIMA model

Over the decade, researchers have devoted their utmost effort to come up with predictive models which are more reliable (**Ariyo et al., 2014**). This paper presents an extensive process of building ARIMA models for stock price prediction. The experimental results obtained with the best ARIMA model demonstrated the potential of ARIMA models to predict stock prices satisfactory on a short-term basis. This could guide investors in the stock market to make profitable investment decisions. With the results obtained ARIMA models can compete reasonably well with emerging forecasting techniques in short-term prediction.

To determine the best ARIMA model among several experiments performed, the following criteria are used in this study for each stock index.

- Relatively small of BIC (Bayesian or Schwarz Information Criterion)
- Relatively small standard error of regression (S.E. of regression)
- Relatively high of adjusted R2

• Q-statistics and correlogram show that there is no significant pattern left in the autocorrelation functions (ACFs) and partial autocorrelation functions (PACFs) of the residuals, it means the residuals of the selected model are white noise.

1.7 A Novel Approach for Stock Market Price Prediction Based on Polynomial Linear Regression

Using research work of (**Jayesh Amrutphale et al., 2020**), it is found that their proposed (PLR) polynomial linear regression model is far better option than simple linear regression model to develop stock market prediction system. By the analysis of historical data of the stock market, they found that the stock market is a fluctuating market. Many factors affect the stock market prices. So, it is very difficult to predict accurate value of stock prices using the SLR model. But by the implementation and experimental results, they found that the PLR model gives better prediction accuracy and results. Therefore, we can consider polynomial linear models over linear models in our future work.

1.8 Open Price Prediction of Stock Market using Regression Analysis

A study is performed to predict open price of stock market by (**Pramod Mali1 et al., 2017**). In this paper they have studied some well-known prediction algorithms concerned with regression. A comparative study has been made using tabular format. The accuracy yielded by them and the various parameters used for prediction have been stated. It is observed that prediction using multiple regression yields better results than linear regression. Further the use of neural networks for prediction sounds to be a promising field in the future and can be used for real time trading in the stock market. Similarly, we are predicting close prices in our research paper using other variables.

1.9 Stock price prediction using machine learning on least-squares linear regression basis

C. C. Emioma and S. O. Edeki 2021 has done a research to predict close price of the bank of America stock dataset using least-square linear regression technique. The steps for the prediction include Splitting the dataset into 3 (training, validation & testing datasets), Training the model with the training dataset, calculating the errors mean absolute percentage error (MAPE), root mean squared error (RMSE), Tweaking the parameters on the validation dataset to achieve the lowest errors, and Predicting the test dataset. The essence of the study was to create and test a machine learning model to aid in the analysis and prediction of a stock trend pattern. The model successfully predicted the results with a mean absolute percentage error of 1.367% and a root mean squared error of 0.512. The proposed model can be used by modifying only the training data for any other stock market in other countries.

1.10 Stock Price Prediction using KNN and Linear Regression

Poornima S P et al., 2019 predictions of the stock prices based on KNN algorithm and linear regression. They have compared these two methods on bases of confidence value and analyzed that linear regression provides the best result compared to KNN model. The outcome of these 2 techniques have been compared based on the Confidence value. By using R2 (Coefficient of determination) they found the accuracy. KNN-Algorithm shows 63% of accuracy, whereas in Linear regression 98% of accuracy has been shown for daily stock prices.

2 DATA ANALYSIS

2.1 Data description

The dataset for this index is obtained from Yahoo Finance using the Quantmod package in R. The timeline of the data is from 01-Jan-2016 to 31-Dec-2019. After downloading the data we have 7 features: Open, High, Low, Close, Volume, Adjusted Close Price (of numeric class) and Date (of Date class).

The open is the price at which the asset started the day at. For example, if you are analyzing the TSX Composite index, the open price will be where the price starts the day. This price is determined by the demand and supply of the asset at the start of the day. High is the highest level the asset has reached. For example, if the TSX Composite index opened the day at $20,200, rose to a high of $20,270, and then started moving downwards, then the high of the day will be $20,270. As the name suggests, Low is the lowest price of the day. Finally, Close is the final price where the asset trades at. Volume is counted as the total number of shares that are traded (bought and sold) during the trading day or specified set period of time. It is a measure of the total turnover of shares. Adjusted closing price represents stock value which is just the cash value of the last transacted price before the market closes. The adjusted closing price factors in anything that might affect the stock price after the market closes. "Date" represents the date when the stock was traded.

Column Name	Description	Class
Open	The Open price of the particular day	Numeric
High	The High price of the day	Numeric
Low	The Low price of the day	Numeric
Volume	Total Volume Traded on the day	Numeric
Adjusted Close Price	Adjusted Close Price at the	Numeric

	end of the day	
Date	Date when stock was traded	Date

2.2 Data Normalization

There are a couple important details to note about the way the data must be preprocessed to be fit into regression models. Firstly, dates are normally represented as strings of the format" YYYY-MM- DD" when it comes to database storage. This format must be converted to a single integer to be used as a column in the feature matrix. This is done by using the date's ordinal value.

2.3 Statistical analysis for feature selection (SELECTING VARIABLE)

Stock market close price is an important piece of information that is very useful for every trader. The close prices are very important, especially for swing traders and position traders. It also has implications for practical day trading in many day trading systems. The stock market close price level provides very important information about the general mood of investors. It tells a lot about the thinking of big investors that allocate large amount of money into the stock market for their asset management purposes.

To select the best independent set of features we will do Welch's t-test for independence to select variables for our linear regression model. Since Open price, Low Price, High Price are almost equal in mean and variance of close price so we will use only open price and Volume for t-test.

A Welch's independent t-test showed that the mean difference in groups given Close Price and Volume in the sample was statistically significant, $t(1002) = 98.938$, $p < 0.05$, with the effect of Volume over Close Price. We can conclude that there is a significant difference in the mean of close price and Volume. So, Volume can be considered while plotting a linear model.

A Welch's independent t-test showed that the mean difference in groups given Close Price and Open Price in the sample was statistically insignificant, $t(2004) = 0.030677$, $p > 0.05$, with the effect of Close Price over Open Price. We can conclude that there is an insignificant difference in the mean of close price and Open Price. So, Open Price cannot be considered while plotting a linear model.

Therefore, for the linear regression model we will include close price, volume along with date.

3. METHODOLOGY

3.1. REPRESENTATION OF MULTIPLE LINEAR REGRESSION MODEL

Linear regression comes under the category of supervised learning of machine learning algorithms. Regression is used while predicting the continued values, forecasting is used while predicting the future values based on past and values. In the statistics field, Linear regression was often used as a model for realizing the correlation between input and output numerical variables. This has been acquired by machine learning.

Multiple regression is an extension of simple linear regression in which more than one independent variable (X) is used to predict a single dependent variable (Y). The predicted value of Y is a linear transformation of the X variables. Multiple Regression Equation having "K" independent variables is given by:

$Y = b0 + b1X1 + b2X2 + b3X3 + \ldots\ldots bkXk$

After running a linear regression on the hypothesis stated it was observed that the hypothesis actually does not gain much support from the model. The predicted line is:

$y = (1.957 * X1) + (-2.499e\text{-}08 * X2) - 1.835e+04 + Error$

Here, the slope Date is 1.957, Volume Slope is -2.499e-08 and the intercept of -1.835e+04. Normal Q-Q graph's (Figure 2) points fall along the line in the middle but curve off in the extremities. This means it has more extreme values.

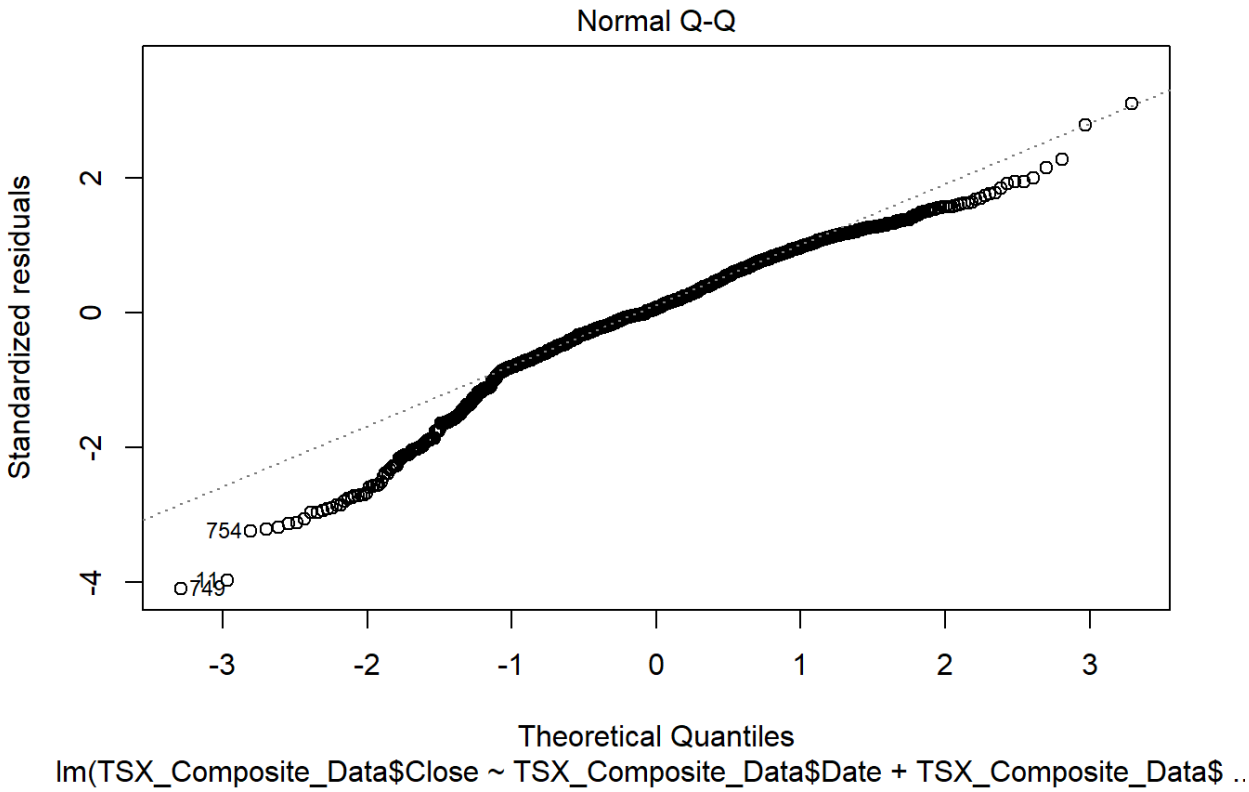

Figure 4: Normal Q-Q graph

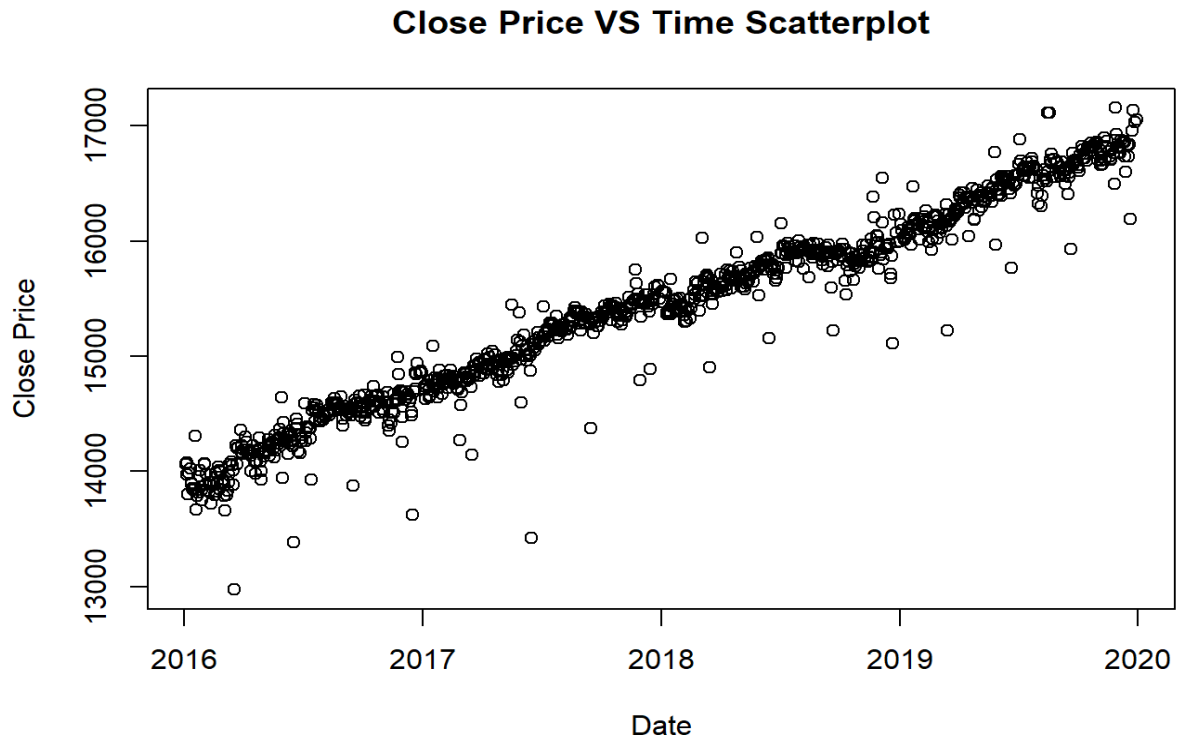

Figure 5: Close Price on Fitted Model scatter plot

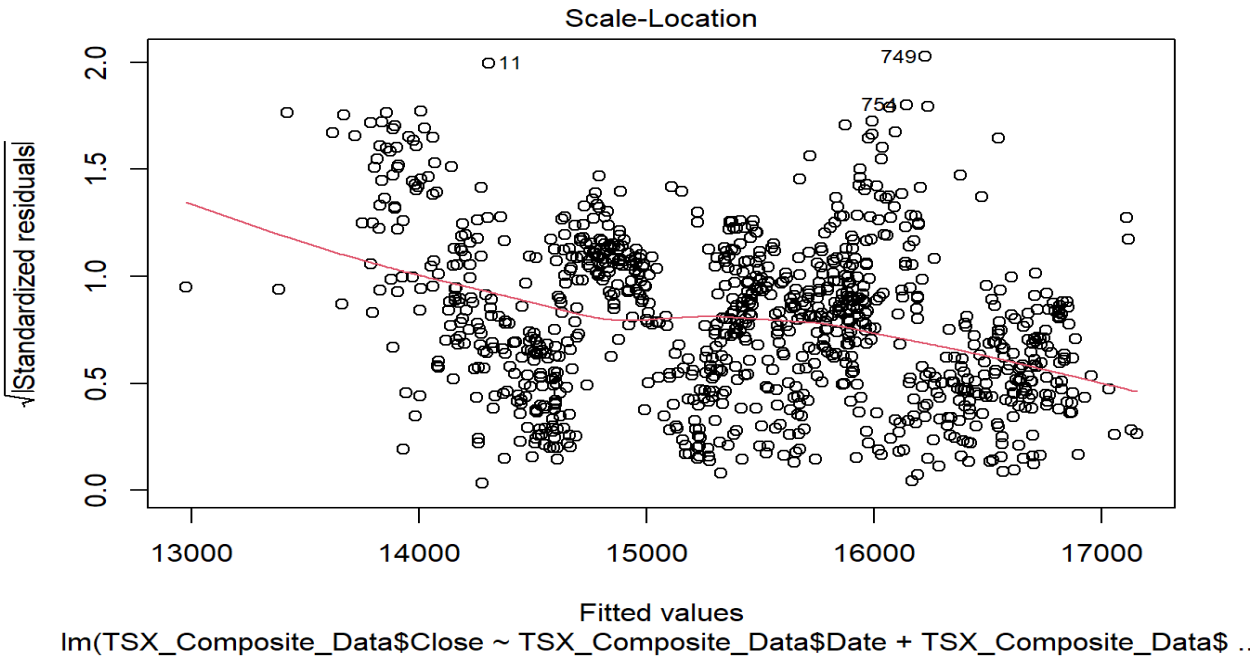

Figure 6: Scale Location

We can say that close prices have a linear trend by looking at “Close Price VS Time Scatter Plot” (Figure 5). From the beginning of 2016 and the end of 2019 there has been a gradual increase in close price.

We can study the plot and see that the “scale location” (Figure 6) plot suggests some non-linearity here, but what we can also see is that the spread of magnitudes seems to be lowest in the fitted values close to 14000, highest in the fitted values around 15500. This suggests heteroskedasticity.

We can see that the R-squared (also called the coefficient of determination) value is 0.6708 which is moderate. Date has a positive slope of 1.957 and Volume has a negative slope of -2.499e-08. Moreover, high MAPE and RMSE values of 354745.49 and 595.60 indicate that the linear model is not the best fit to predict it as there are high errors. This seems obvious for linear regression to behave on the stock price in the market for a long period of time which does not allow the graph to be linear.

3.2. REPRESENTATION OF ARIMA MODEL

Auto-Regressive Integrated Moving Average (ARIMA) model is a combination of Auto-Regressive model, Integration of differencing, and Moving Average model. The Auto-Regressive (AR) model defines the relation of an observation and some lagged observations. In order to use the ARIMA model, the time-series data has to become stationary, and this can be achieved by differencing the data. Last, the Moving Average (MA) model defines the relation of an observation and residual error of moving average model on the lagged observations.

ARIMA models are generally denoted by ARIMA(p,d,q) where p, d, and q are the parameters with positive values. The order of AR (p) is the number of lag observations in the model. For differencing, we have d as the number of times the actual data are differenced to become stationary. Usually, the maximum times of differencing to achieve stationarity is 2 times. The order of MA (q) is the size of the moving average window.

There are ways to determine appropriate values of those parameters for our model, but they are difficult. Fortunately, in R we have auto.arima() function that will do the job for us.

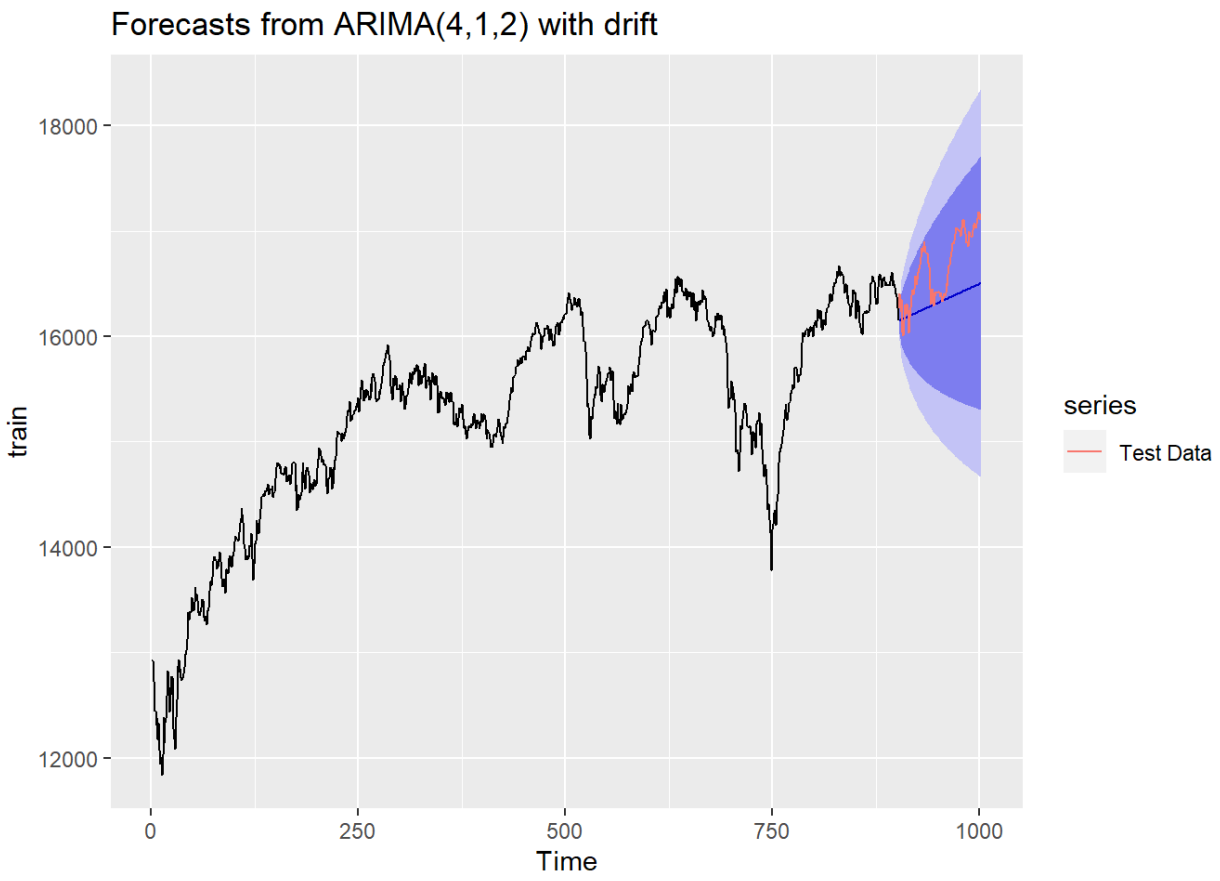

Figure 7: ARIMA model

The result using seasonal ARIMA is an upward trend data (Figure 7). However, if we compare it to the actual test data, there are still some differences between them.

After forecasting our data, the last step is to evaluate our forecast. Forecast evaluation is done by checking whether the residuals meet the residual assumptions and comparing the accuracy metrics.

Check Residuals

Residual is the difference between the forecast data and the actual data. A good model must have a randomly distributed residual without an obvious pattern. Here are some residual assumptions that have to be met:

- Normally distributed (mean = 0) that can be checked using a normal curve. The normal curve has to be a bell-shaped one.
- Have a constant variance that can be checked using residual plot. Constant variance is shown in a constant fluctuation of the data.

- Have no autocorrelation that can be checked using ACF plot and Ljung Box test. Autocorrelation can be detected in an ACF plot when there are lines beyond the upper or lower bound. In the Ljung Box test, to meet this assumption, we must have a p-value that is greater than 0.05. If the result in the ACF plot and Ljung Box test is difference, we will prefer to use the result from the Ljung Box test.

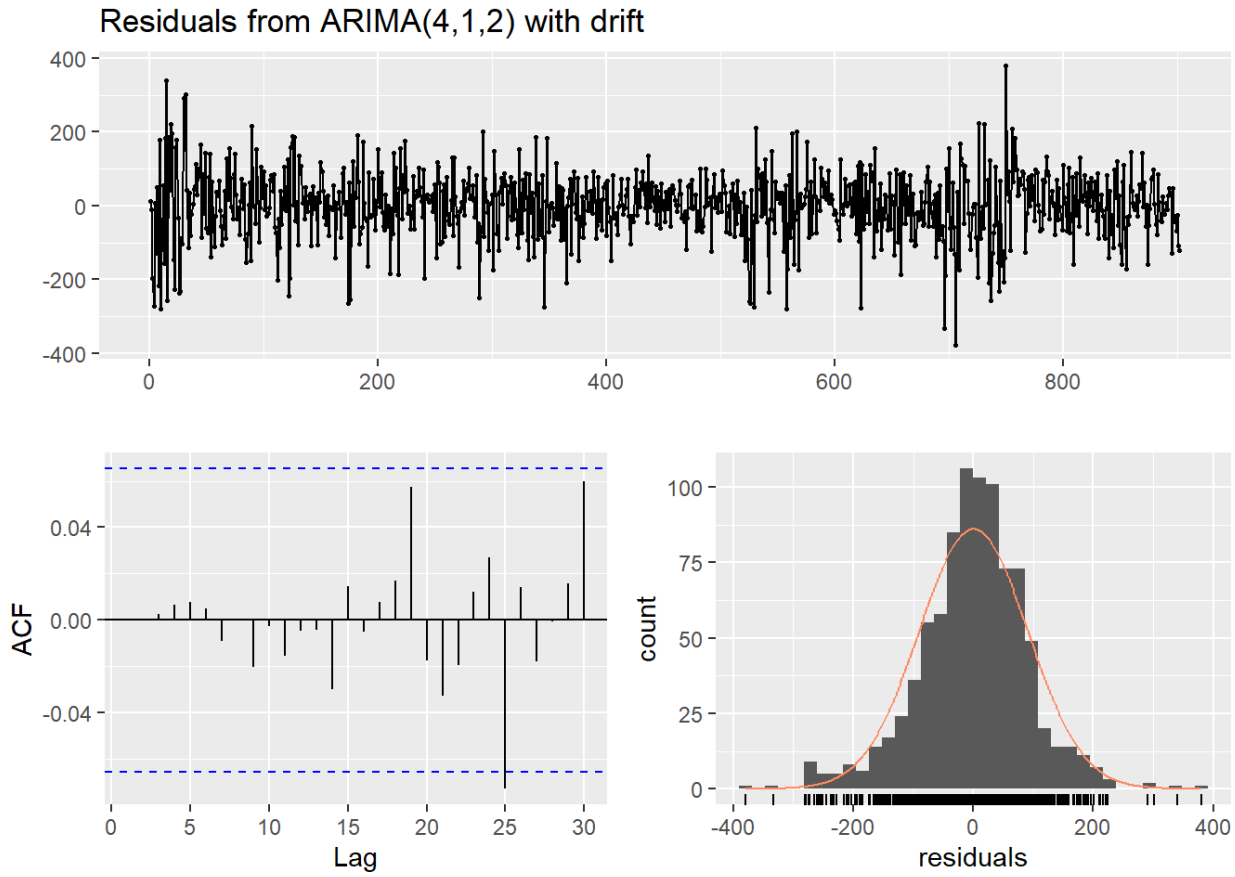

Figure 8 : Residuals from Arima(4,1,2) with drift

```
##
##  Ljung-Box test
##
## data:  Residuals from ARIMA(4,1,2) with drift
## Q* = 0.57172, df = 3, p-value = 0.9029
##
## Model df: 7.   Total lags used: 10
```

Figure 9 : Ljung-Box test

ARIMA met all of the assumptions. So based on the residual check, our ARIMA model produced better results.

4. RESULTS

In order to do accuracy metrics comparison in forecasting, we will compare Root Mean Squared Error (RMSE) of both models. Root Mean Squared Error (RMSE) is the standard deviation of the residual that measures how spread out our residuals are. A smaller value of RMSE is a sign of a better result.

Based on RMSE, we can see that the ARIMA model has a RMSE of 90.67 (Figure 10) which has better results than linear Regression with RMSE of 595.60 (Figure 11) while forecasting our TSX Composite closing price.

```
#Accuracy Metrics of ARIMA
accuracy(fc_s)
```

```
##                       ME     RMSE      MAE           MPE     MAPE      MASE
## Training set -0.001054656 90.67734 67.60009 -0.000724578 0.452702 0.9933713
##                      ACF1
## Training set 0.0001861074
```

Figure 10: Accuracy of ARIMA model

```
# Accuracy level of linear model
accuracy(model)
```

```
##                         ME     RMSE      MAE         MPE     MAPE      MASE
## Training set -1.936137e-14 595.6052 451.6092 -0.1697713 3.027041 0.5664064
```

Figure 11: Accuracy of linear model

Combining the above analysis and forecast result of stock price data with other analysis such as technical or sentiment analysis will be a good approach to analyze stock price data in order to make investment decisions.

5. CONCLUSION AND DISCUSSION

In this study we have taken original time series data such as Open, Close, Low, High and Volume etc. of the TSX Composite Index using the quantmod package in R and used the linear and ARIMA model to select the best model for prediction.

Decomposing time-series data can give us a more detailed view of the pattern of our data. We have seen repeated seasonal fluctuation of data. The closing price of the TSX composite Index tended to reach the highest in July and the lowest in December. From this decomposition, we may say that the right time to sell this stock was at the middle of the year (especially in July) and the right time to buy was at the end of the year (especially in December). Analyzing it will help us in making decisions about our data. However, decomposition in R can only be done to ts object that the frequency has been specified.

We can also conclude that the market is not efficient, with proper analysis, using advanced models and algorithms we can always beat the market and book large profits as we have seen the trend of non-linearity in various plots. With increase in momentum and volume of data traded, stock can perform better than the actual trend in the past.

Linear regression is a great tool and a simple method. However, it is suggested that we use another more robust approach to deal with other cases as its RMSE is higher. On the other hand, ARIMA models (including the seasonal one) that have less RMSE can be a good model to forecast data that are more fluctuating.

I can also further improve the modeling by comparing Polynomial Linear Regression, Support Vector Regression, Neural Networks, KNN and many other models. Here, data from 2020 is not picked as due to COVID-19 stocks prices declined which again reduces the effectiveness of using linear regression.

REFERENCES

Lin, Yang & Song, 2009; Rustam & Kintandani, 2019

Garcia-Lopez, Batyrshin &Gelbukh, 2018; Mourelatos et al., 2018

Idrees, Alam & Agarwal, 2019

Alexandra Gabriela Titan, 2005 The Efficient Market Hypothesis: Review of Specialized Literature and Empirical Research

Seethalakshmi “Analysis of stock market predictor variables using Linear Regression” in Volume 119 No. 15 2018, 369-378

Shruti Shakhla "Stock Price Trend Prediction Using Multiple Linear Regression "International Journal of Engineering Science Invention (IJESI), vol. 07, no. 10, 2018, pp 29-33

Ariyo et al., 2014 Stock Price Prediction Using the ARIMA Model

Jayesh Amrutphale et al., 2020

Pramod Mali1 et al., 2017

C. C. Emioma and S. O. Edeki 2021 Stock price prediction using machine learning on least-squares linear regression basis

Poornima S P et al., 2019