International Journal of Engineering Applied Sciences and Technology, 2021
Vol. 6, Issue 7, ISSN No. 2455-2143, Pages 87-97
Published Online November 2021 in IJEAST (http://www.ijeast.com)

SALES FORECASTING USING BOX JENKINS METHOD BASED ARIMA MODEL CONSIDERING EFFECT OF COVID -19 PANDEMIC SITUATION

Bhanuj Nigam,
Research Scholar, Ujjain Engineering College, Ujjain-456010 (MP) India

Dr. A. C. Shukla,
Professor, Department of Mechanical Engineering, Ujjain Engineering College, Ujjain-456010 (M.P), India

ABSTRACT - **This paper presents Box-Jenkins method used to forecast the future demand in a two wheeler industry. An automated technique in machine learning with the help of python language has been developed and used to analyze time series data and ultimately fit the model for future demand projection. The time series data is collected for the Royal Enfield bikes' monthly sale available at the official website of Eicher motors ltd. The resulting pattern found in time series data is used to forecast the future behavior, knowledge of which will help to maintain the appropriate inventory and to reduce the risk in terms of changing customers preferences, resource availability etc. Also the effect of covid-19 pandemic has been captured to visualize its impact. The results thus obtained will be useful to understand the pattern if it occurs again in future. This method provides superior results and can be widely used in various forecasting scenario.**

*Keywords**: Time Series, Box-Jenkins, Forecast, Machine Learning, Royal Enfield***

I. INTRODUCTION

Strategic planning depends on the authentic forecasting is the main part for the successful business management within the company. This is the fact that decides the future existence of the company. The fact is true for all types of sectors, especially for automobile sectors. Authentic forecasting cannot be done by just guessing about the market development but mathematical models need to be used to make the prediction for the future probable outcomes and their accuracy and efficiency are then compared. The modern era of computer provides the tools and techniques to perform the operations, mathematical calculation which is accurate and less time consuming, and works with less human interference and thus avoiding the risk of error.

Since the market is increasing, demand and sales of the automobile vehicle are increasing and so the planning plays very important role in any automobile industry. Planning of the sales activities depend on the accurate forecast, which helps the manufacturer to develop the strategies to overcome from the problem of short inventory or over inventory. Accurate forecasting helps the manufacturer to predict the future market pattern and so permits him to increase his market performance, to develop new policies and hence to compete with the competitor in the market by gaining more profit (Gao et al.,2018, Vahabi et al.,2016).

The era of industry 4.0 enables the entire use of appliances in terms of their quality improvement, life cycle, reduction of defects, usability, and energy saving (Singaravel et al.,2016). Industry 4.0 shows the applicability of data transfer/exchange and automation in manufacturing technology and processes including cloud computing (CC), big data analytics (BDA), Internet of Things (IoT), artificial intelligence (AI), cyber–physical systems (CPS) (Tao et al.,2019). In manufacturing AI plays very important role which are based on different AI technologies like quality inspection, automation in different process, intelligent sensors, predictive analytics etc (Li et al.,2017). Machine learning [ML] is one of the most appropriate technologies of AI which provides the potential to develop the strategies and optimize the process to make the manufacturing and development of the product more efficient (Ishino et al.,2006).

II. LITERATURE REVIEW

Box Jenkins methodology

Box Jenkins method has a complex mathematics and number of steps required to reach an ultimate result, and so use of computation makes this process easy and appropriate to use that produces an autoregressive, integrated and moving average model which sets the seasonal and trend factor, finds the weighting parameter and test the model and again the cycle repeats as appropriate (Gaynor et al.,1994).

There are three processes behind the cohort of time series data.

1. Autoregressive process –past values of error term generate the data.
2. Moving average method – error/ disturbance term generate data.
3. ARMA – data is generated by the combination of autoregressive and moving average term.

ARIMA model is the most general class of model that analyze the time series in time domain but not in frequency domain (Gottman JM,1981). 'I' stands for the order of integration of the series which means how many times the series are differenced to make it stationary.

The ARIMA model has shown to have better forecasting accuracy than the other time series approach (Chu F-L.,2009, Oh et al.,2005, Dharmaratne GS,1995, Chu F-L.,1998).

The general form of ARIMA model given as-

$$X_t - \sum_{i=1}^{p} \varphi_i X_{t-i} = a_t - \sum_{j=1}^{q} \Theta_j a_{t-j}$$

p, q are the order of autoregressive and moving average and Θ & φ are the model parameter.

The ARIMA model needs to be check for stationarity because it does not work well with non stationary data. Box Jenkins method is considered as best framework that deals with stationary data(Andreoni et al.,2006).

1. Stationarity- In a stationary process the mean, variance, autocorrelation structure do not change over time. Informally, the series which is flat looking having no trend and constant variance & autocorrelation structure with time and having no periodic fluctuation is called stationarity of a series, so before performing any operation the series is checked for non stationarity, and if it needed transformed to stationarity. To check stationarity of the data values autocorrelation graph with varying time lags are commonly used (Andreoni et al.,2006, Krasić D et al.,2009, Hamed MM.,1999, Tsai et.al.,2011). (Hamed MM.,1999, Andreoni et al.,2006) shows the use of ACF and PACF plots to check for non stationarity. (Bougas,2013) used plotting time series and (Krasić D et al.,2009) uses ACF and PACF plots to check the mean and variance stability. (Tsui et al.,2011) considered the data variance non stationarity based on time series plot and human expert.
   The augmented Dickey fuller (ADF) test is also used to check for data stationarity with the existence of unit root (Krasić D et al.,2009, Lim C and McAleer M.,2002, Tsui et al.,2011, Bougas,2013).(Oh et al.,2005) uses Dicky fuller test to check for data stationarity with graphical judgment. (Krasić D et al.,2009) conclude that when data is stationary ,there will always be the low Akaike criteria and Schwartz criterion and high values in case of determination coefficient.
   The graphical analysis and the combination of graphical and statistical analysis is widely used to check for data stationarity (Anvari et al.,2016), but these analysis depends on human skills ,so causing of human risk making errors increases , also the good knowledge of reading and understanding the plots are needed. In the logarithmic transformation in which each value of series is substituted with the natural logarithm (Andreoni et al.,2006, Krasić D et al.,2009, Tsai et.al.,2011, Bougas,2013), which is used to remove variance non stationarity. Normal differencing is most popular and widely used method to remove mean non stationarity of the series (Andreoni et al.,2006, Bougas,2013) in which values of point is substituted with its difference with the values of next points.

2. Seasonality- When a certain pattern repeated at a regular time interval and there is a constant change found in a series then series is said to have seasonal behavior. (Bougas,2013) used the Osborn, chui, smith and Birchenhall (OSCB) test to find that if seasonal differencing is needed and they have used the dummy regressor to check for

monthly effect. Most of the researcher used the run sequence plot and autocorrelation plot to detect seasonality (Hamed MM.,1999, Tsai et al.,2011, Bougas,2013) . The method is quite popular and when autocorrelation at seasonal lag do not decreases rapidly indicates seasonality. Hamed (Hamed MM.,1999) used the general representation of Box Jenkins method for non stationarity series with seasonality. (Tsai et al.,2011) used the SARIMA model.

III. METHODOLOGY

This paper is about the sales forecasting for the Eicher Company which manufactures the transportation vehicles but the data analysis has been done for the two variant of royal Enfield bikes i.e. 350cc and 500cc for the year of April 2013 to January 2021. As the whole world had faced the pandemic in 2020 due to corona virus, the effect of the virus can be visualize in the sales of the bikes and so the ultimate result is divided into two category.

1. Forecasting without covid-19 pandemic taken into consideration.
2. Forecasting with covid-19 pandemic taken into consideration

Now in order to develop ARIMA model for time series forecasting, it is used to define the parameters.
P: it defines the lag order which means the numbers of lag observations involved in the model.
d: it defines the number of times the series are differenced in order to get differenced series.
q: it is called as the order of moving average.

Further the box Jenkins method involve 3 basic steps-

1. Check for data stationarity.
2. Model selection and estimation.
3. Forecasting errors estimation.

1. Check for data stationarity – The Augmented Dickey Fuller (ADF) Test is used to check for the series stationarity because we only need to difference the series if the series is non stationary ,else if d=0, no differencing is needed. We are taking the null hypothesis for the ADF test is that the data is non stationary and if the p-value for the test is below the significance level (0.05) then we reject the null hypothesis and we infer that the time series is indeed stationary and if in case the p-value is greater than the significance value (0.05), the series would be differenced until the value comes below the significance level and in this way the order of d would be decided. Sometimes the series is over differenced and it will affect the model parameter even though it shows the stationarity behavior. The minimum differencing required to get the near stationary series that roams around the define mean and auto correlation function (ACF) reaches to zero fairly quick is defined as the right order of differencing. If the autocorrelation is positive for number of lags, then the series needs further differencing, and if lag 1 autocorrelation itself is too negative, then the series is probably over differenced.

2. Model selection and estimation – most of the researchers uses ACF and PACF plots to find the appropriate model order for which high level of expertise of user is needed and as the human involvement, the risk of accuracy increases. Alternative approach is to use an iterative cycle of selection which involve estimation, judgment and then again reselection which is time consuming and inefficient and difficult to follow as a computer program. In order to avoid the drawback of these models , we simplify the selection of ARIMA model based on enumeration. In enumeration we used the fact that the order of autoregressive (AR) and moving average (MA) term is less than 3 and 4 respectively and with the help of python program we construct all the possible values with autoregressive (AR) and moving average (MA) values less than 3 and 4 respectively . Then with the help of Akaike information criterion (AIC) values for all the models, we selected the models with minimum AIC values and ultimately fit the model.

3. Forecasting error estimation – now in order to check the accuracy the data is divided into two categories i.e. - train data and the test data.
Once the model is selected, some part of the data is selected for training and fitting the model into it and the rest part of the data is being tested for already selected model and ultimately on the basis of the fitting the accuracy is determined.
The accuracy of the forecasting can be decided on the basis of the values determined by mean absolute error (MAPE). MAPE values lies between 0 to 1 and in this way one can judge, how good is the forecast irrespective of the scale of the series.

IV. RESULT

The result contains the analysis of two categories which is, with the Covid affect and without the Covid affect. As mentioned above that the analysis has been done on the two variant of the bikes and ultimately the overall sale of the bikes. The analysis has been done for the time period of April 2013 to January 2021. As the data collected is a real world data and it contains the Covid effect and in order to study the affect without the Covid situation, the data has been modified on the basis of studying the graphs up to the extent the affect can be seen. As the pandemic strikes India in January 2020 and its affect on the manufacturing industries can be seen from March 2020 to till the time data have been collected. So for the analysis considering that there is no Covid situation, the actual data is being modified from March 2020 to January 2021 by taking the mean of the January 2020 and February 2020. And also in the actual data, there was no any selling of bikes in may 2020 and sale was totally zero, so In order to have mean absolute percentage error in defined form the zero value has been taken and modified to 1, which shows no any major affect on the analysis.

Now the graphs shown in Figure 1 to figure 6 show the sale of different models of bikes from May 2013 to January 2021 including Covid situation and also if Covid situation is not considered.

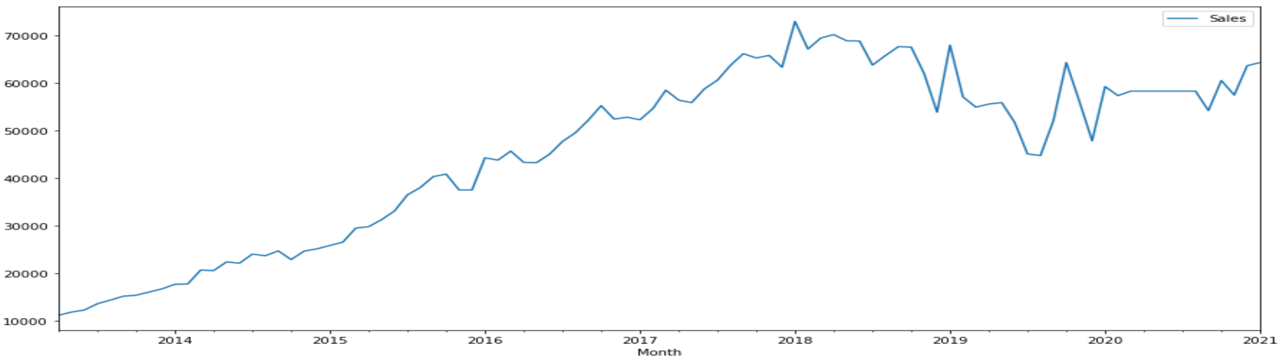

Fig.1: Data Variation of 350cc bike considering without Covid situation

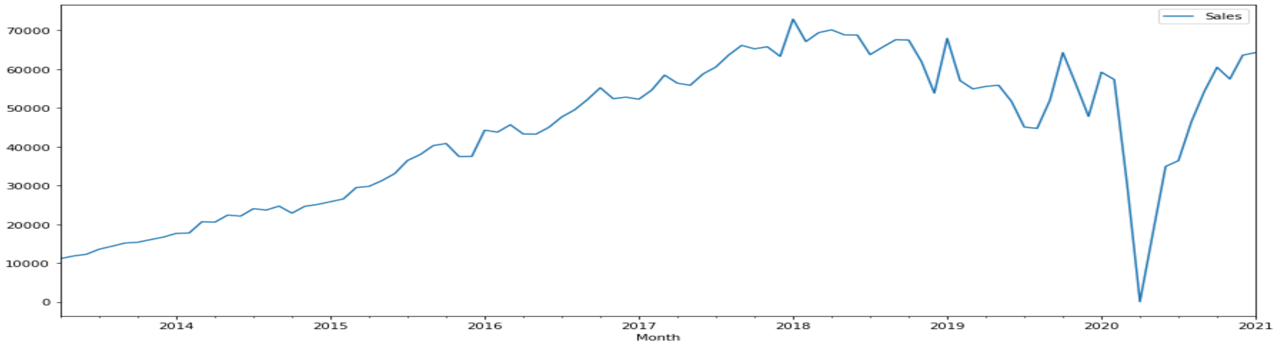

Fig 2: Data Variation of 350cc bike considering with Covid situation

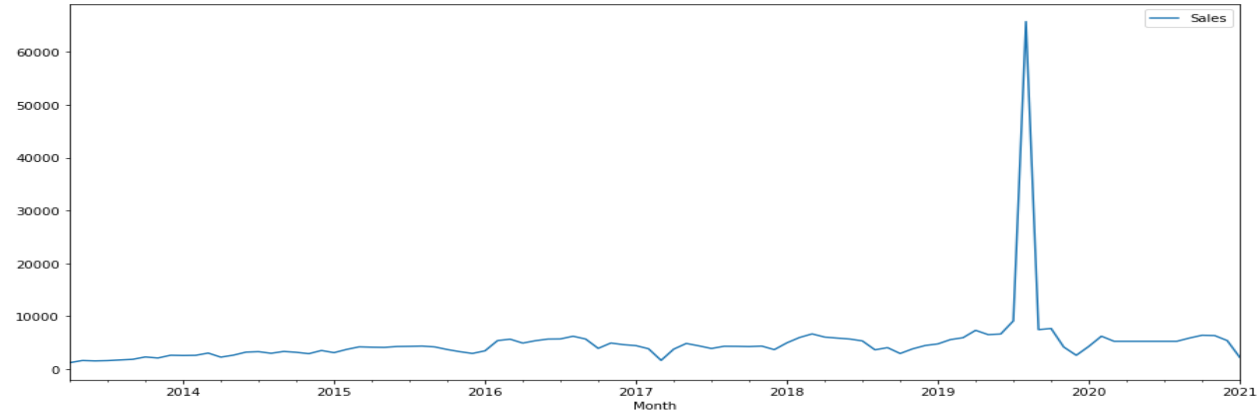

Fig 3: Data Variation of 500cc bike considering without Covid situation

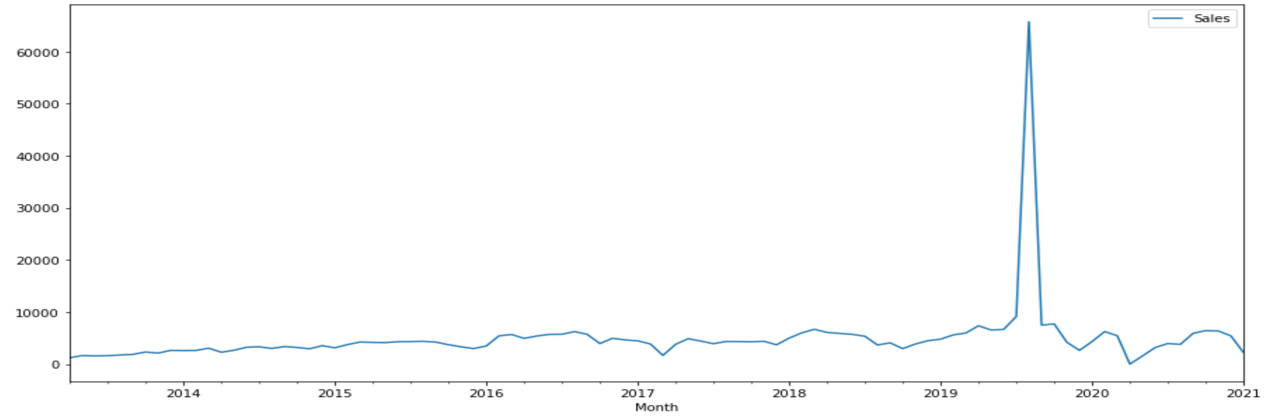

Fig 4: Data Variation of 5000cc bike considering with Covid situation

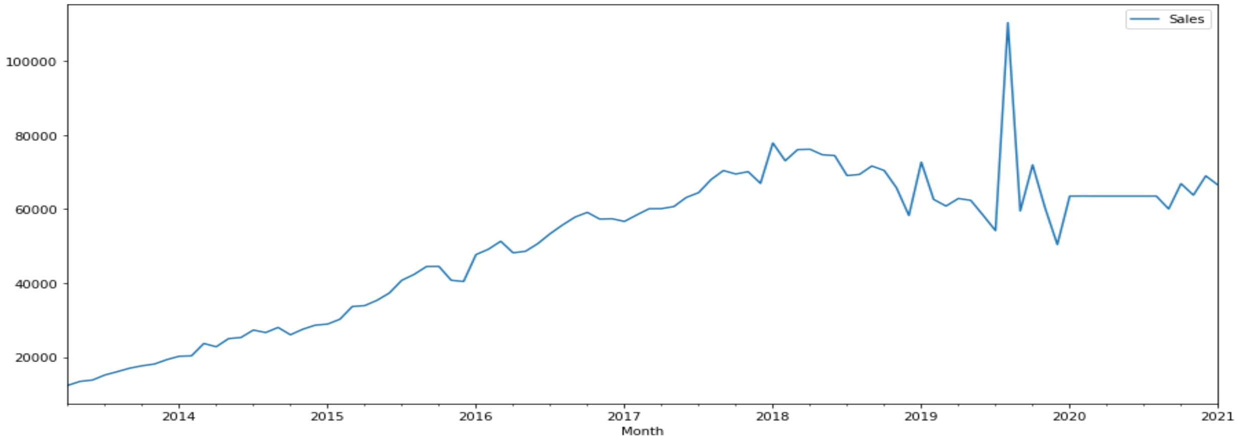

Fig 5: Data Variation of total sale of bikes considering without covid situation

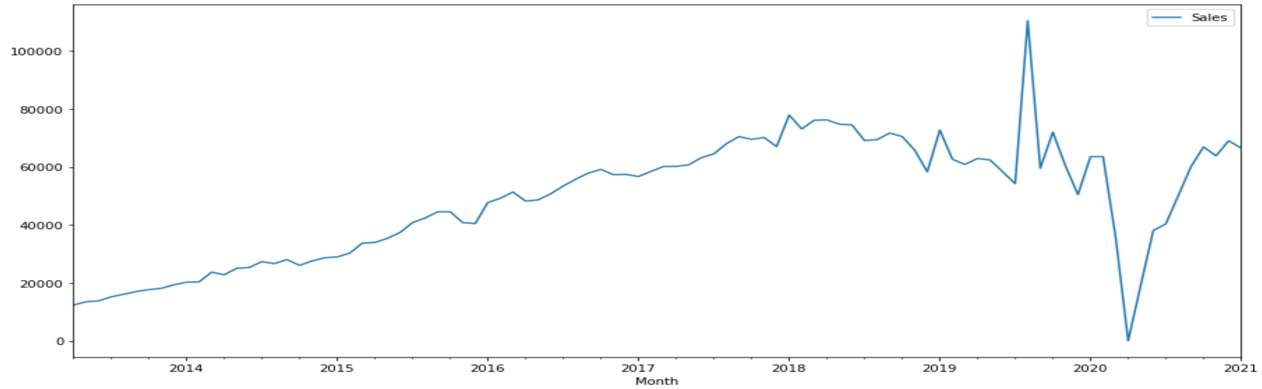

Fig 6: Data Variation of total sale of bikes considering with Covid situation

Now the time series data needs to be differenced and it required one time differencing to get stable so value of d is taken as 1.

With the help of Akaike information criterion (AIC) values for all the models, we selected the models with minimum AIC values. Below tables shows the possible models with their AIC values for both the categories with all the variants and their total sale.

The highlighted models with their AIC values are selected to fit the models.

Table -1 ARIMA Models for 350cc, 500cc and Total Sale with AIC values without Covid situation.

350cc		500cc		Total Sale	
ARIMA MODEL	AIC VALUE	ARIMA MODEL	AIC VALUE	ARIMA MODEL	AIC VALUE
(1,1,1)	1802.662	(1,0,1)	1923.962	(1,1,1)	1915.380
(0,1,0)	1806.589	(0,0,0)	1962.843	(0,1,0)	1957.666
(1,1,0)	1801.970	(1,0,0)	1941.486	(1,1,0)	1921.858
(0,1,1)	1799.008	(0,0,1)	1951.219	(0,1,1)	1913.760
(0,1,0)	1806.554	(2,0,1)	1926.240	(0,1,0)	1956.076
(0,1,2)	1798.704	(1,0,2)	1926.449	(0,1,2)	1921.399
(1,1,2)	1702.745	(0,0,2)	1947.487	(1,1,2)	1918.124
(0,1,3)	1799.378	(2,0,0)	1935.719	(0,1,1)	1917.566
(1,1,3)	1801.111	(2,0,2)	1927.995	NIL	NIL
(0,1,2)	1802.637	(1,0,1)	1922.429	NIL	NIL
NIL	NIL	(0,0,1)	1921.019	NIL	NIL
NIL	NIL	(0,0,0)	1920.795	NIL	NIL
NIL	NIL	(1,0,0)	1920.587	NIL	NIL
NIL	NIL	(2,0,0)	1921.879	NIL	NIL
NIL	NIL	(2,0,1)	1923.914	NIL	NIL

Table -2 ARIMA Models for 350cc, 500cc and Total Sale with AIC values with Covid situation.

350cc		500cc		Total Sale	
ARIMA MODEL	AIC VALUE	ARIMA MODEL	AIC VALUE	ARIMA MODEL	AIC VALUE
(1,1,1)	1896.772	(1,1,0)	1925.932	(1,1,1)	1981.174
(0,1,0)	1898.426	(0,0,0)	1961.412	(0,1,0)	1989.595
(1,1,0)	1899.172	(1,0,0)	1941.300	(1,1,0)	1981.492
(0,1,1)	1897.354	(0,0,1)	1950.300	(0,1,1)	1979.186
(0,1,0)	1897.166	(2,0,1)	1927.672	(0,1,0)	1987.887
(2,1,1)	1891.672	(1,0,2)	1926.861	(0,1,2)	1980.640
(2,1,0)	1889.625	(0,0,2)	1946.860	(1,1,2)	1981.976
(3,1,0)	1891.620	(2,0,0)	1936.058	(0,1,1)	1977.970
(3,1,1)	1893.621	(2,0,2)	1929.293	(1,1,1)	1979.770
(2,1,0)	1888.799	(1,0,1)	1923.279	(0,1,2)	1980.050
(1,1,0)	1897.721	(0,0,1)	1921.888	(1,1,0)	1980.117
(3,1,0)	1890.732	(0,0,0)	1921.758	(1,1,2)	1981.418
(2,1,1)	1890.803	(1,0,0)	1921.433	NIL	NIL
(1,1,1)	1895.192	(2,0,0)	1922.753	NIL	NIL
(3,1,1)	1892.654	(2,0,1)	1924.790	NIL	NIL

Now as the models have been selected, the next section includes the fitting of models into the time series and forecasting. There are total 94 data collected on which the first 69 values are used to train the data(from April 2013 to December 2018) and then the last 25 data values are used to test the data(from January 2019 to January 2021). After training and testing the data the forecasting that we have achieved is shown below.

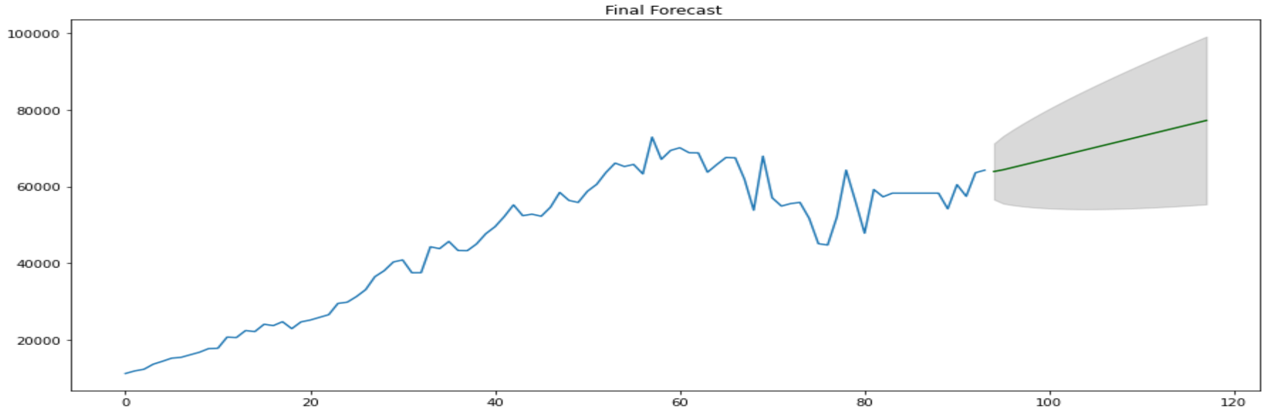

Fig7: Forecasting of 350 cc bike without the Covid Situation.

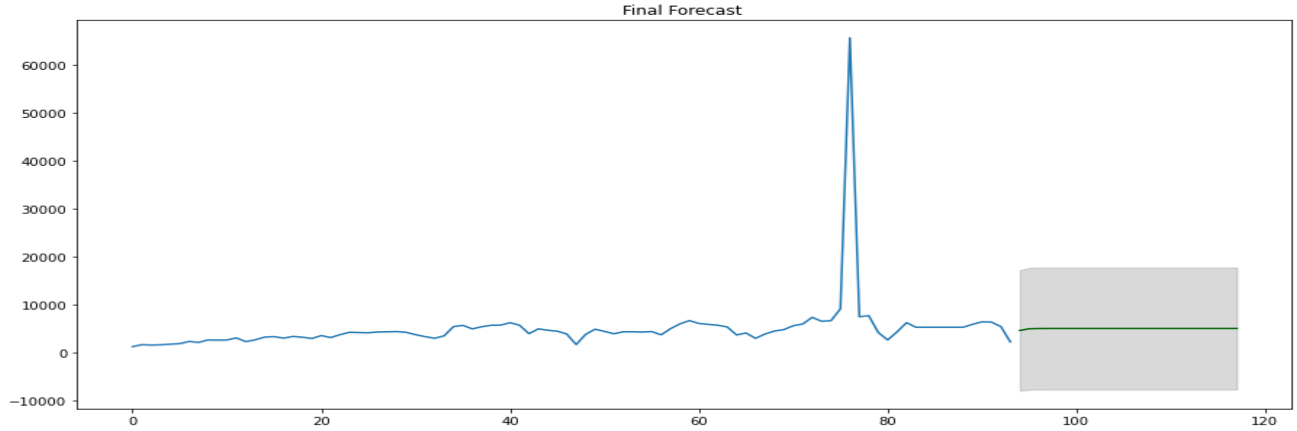

Fig8: Forecasting of 500cc bike without the Covid Situation.

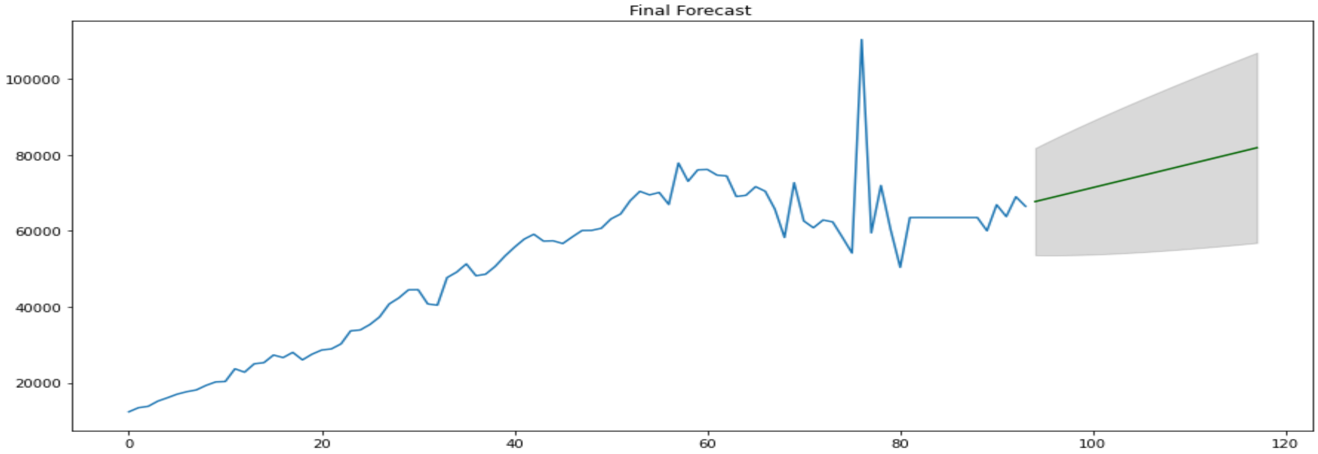

Fig 9: Forecasting of Total Sale of bikes without the Covid Situation.

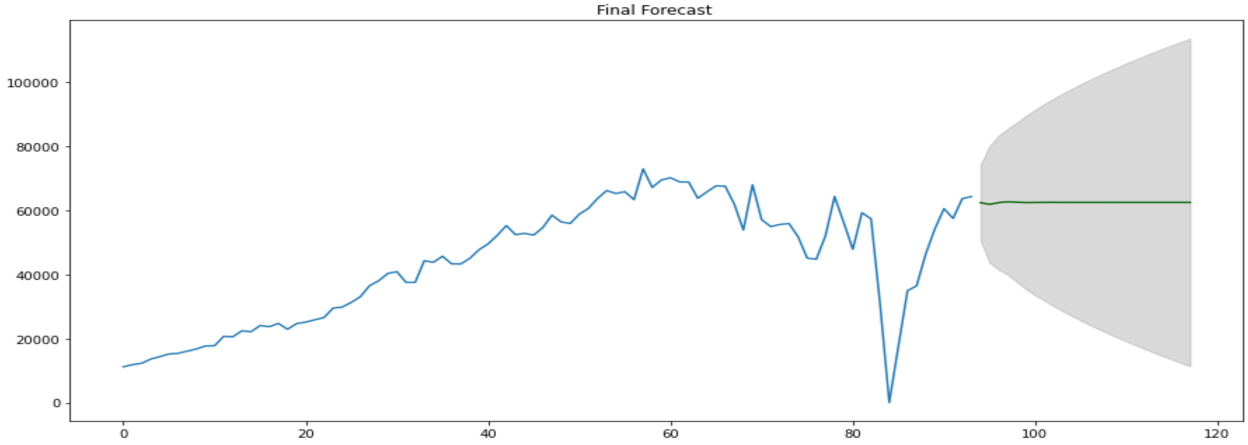

Fig 10: Forecasting of 350cc bike bikes with the Covid Situation.

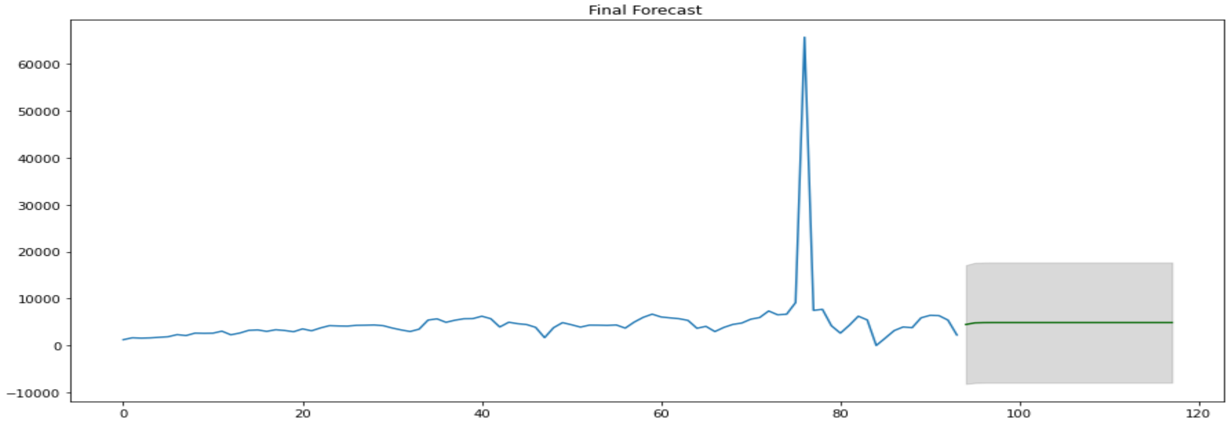

Fig 11: Forecasting of 500cc bikes with the Covid Situation.

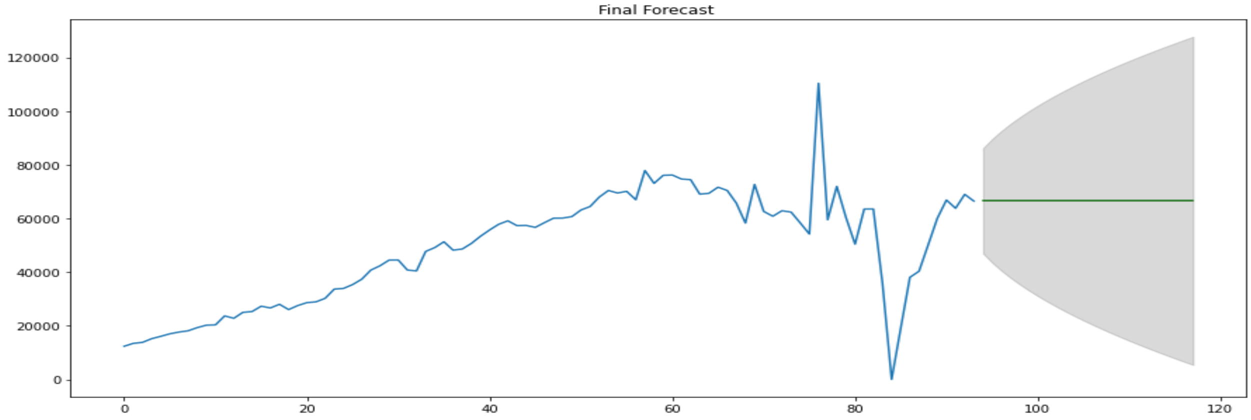

Fig12: Forecasting of Total Sale of bikes with the Covid Situation.

MAPE is used to capture the absolute error to show forecast error size. This criteria is expressed as a percentage which makes it appropriate when reporting the model accuracy. It is very easy to interpret and many researchers used this criteria in the literature. [21,14].

The achieved MAPE in case of without covid situation is 18.03% for 350cc bikes, 15.51% for 500cc bikes, and 17.95% for total sale of the bikes.

If the covid situation taken into consideration, the MAPE achieved is 26.01% for 350cc bikes, 2.03% for 500cc bikes and 18.73% for the total sale of both the variants.

V. OBSERVATIONS

Machine learning is a technology which provides the best result when the big chunk of data is available. In this research as the data collected is real world data, the result gained in the term of MAPE is acceptable and it proves that ML provides a bunch of libraries which makes the data analysis easy, and less time consuming. The result gained in this research could be improved further as the data set would be big. This research only having 94 data which is very less in order to train the data for further forecasting but ARIMA model fits well and which shows the model could give the best solution for the real world data.

Fig 2 shows that from 2013 to till 2018 the sale of the bikes increases continuously and suddenly a sudden fall in sale can be seen and again in may 2020 the sales falls to zero, the huge variation impacts the future behavior of the data while forecasting and as this is the one time situation arises once in a decade due to different reasons the affect can be ignored while analysis and data could be modified considering the constraints which occurs randomly.

VI. REFERENCE

[1] Junjie Gao, Yanan Xie, Xiaomin Cui, Han Yu, and Feng Gu, "Chinese automobile sales forecasting using economic indicators and typical domestic brand automobile sales data: A method based on econometric model", Advances in Mechanical Engineering, 2018, Vol. 10, No.2, pp.1-11.

[2] Amirmahmood Vahabi, Shahrooz Seyyedi Hosseininia, Mahmood Alborzi, "A Sales Forecasting Model in Automotive Industry using Adaptive Neuro-Fuzzy Inference System (Anfis) and Genetic Algorithm(GA)", International Journal of Advanced Computer Science and Applications, 2016, Vol.7, No.11, pp.24-30.

[3] S. Singaravel, J. Suykens, P. Geyer, Deep-learning neural-network architectures and methods: Using component-based models in building-design energy prediction, Adv. Eng. Inform.38(2018)81–90, https://doi.org/10.1016/j.aei.2018.06.004.

[4] Fei Tao, Qinglin Qi, Lihui Wang, A.Y.C. Nee, Digital Twins and Cyber–Physical Systems toward Smart Manufacturing and Industry 4.0:

[5] Correlation and Comparison, Engineering5(4)(2019)653–661 https://linkinghub.elsevier.com/retrieve/pii/S209580991830612Xhttps://doi.org/10.1016/j.eng.2019.01.014.

[6] B.-H. Li, B.-C. Hou, W.-T. Yu, X.-B. Lu, C.-W. Yang, Applications of artificial intelligence in intelligent manufacturing: a review, Front. Inf. Technol. Electron. Eng. 18 (2017) 86–96, https://doi.org/10.1631/FITEE.1601885.

[7] Y. Ishino, Y. Jin, An information value based approach to design procedure capture, Adv. Eng. Inf. 20 (2006) 89–107, https://doi.org/10.1016/j.aei.2005.04.002.

[8] P.E. Gaynor, R.C. Kirkpatrick, Introduction to Time Series Modeling and Forecasting in Business and Economics, McGraw-Hill,1994.

[9] Gottman JM. Time-Series Analysis: A Comprehensive Introduction for Social Scientists volume 400 Cambridge University Press: Cambridge, 1981.

[10] Chu F-L. Forecasting tourism demand with ARMA-based methods. Tourism Management 2009; 30(5): 740–751.

[11] Oh C-o, Morzuch BJ. Evaluating time-series models to forecast the demand for tourism in Singapore comparing within- sample and postsample results. Journal of Travel Research 2005;43(4): 404–413.

[12] Dharmaratne GS. Forecasting tourist arrivals in Barbados. Annals of Tourism Research 1995;22(4): 804–818.

[13] Chu F-L. Forecasting tourist arrivals: nonlinear sine wave or ARIMA? Journal of Travel Research 1998; 36(3): 79–84.

[14] Andreoni A, Postorino MN. A multivariate ARIMA model to forecast air transport demand. Proceedings of the Association for European Transport and Contributors, pages 1–14, 2006.

[15] Krasić D, Gatti P. Forecasting methodology of maritime passenger demand in a tourist destination. PROMETTraffic & Transportation 2009; 21(3): 183–190.

[16] Hamed MM. Stochastic modeling of airlines' scheduled services revenue. Journal of Air Transportation World Wide 1999; 4(2–1999): 32–48.

[17] Tsui ,Wai Hong Kan Balli, Hatice Ozer Gower, Hamish ,Aviation Education and Research Proceedings. Forecasting airport passenger traffic: the case of Hong Kong International Airport, 2011.

[18] Constantinos Bougas. Forecasting air passenger traffic flows in Canada: an evaluation of time series models and combination methods. Master's thesis, Laval University, 2013.

[19] Lim C, McAleer M. Time series forecasts of international travel demand for Australia. Tourism Management 2002;23(4): 389–396.

[20] Oh C-o, Morzuch BJ. Evaluating time-series models to forecast the demand for tourism in Singapore comparing within-sample and postsample results. Journal of Travel Research 2005; 43(4): 404–413.

[21] Saeedeh Anvari, Selcuk Tuna, Metin Canci and Metin Turkay. Automated Box–Jenkins forecasting tool with an Application for passenger demand in urban rail systems. Journal of advanced transportation J. Adv.Transp.2016;50:2–49

[22] Tsapakis I, Schneider WH, Nichols AP, Haworth J. Alternatives in assigning short-term counts to seasonal adjustment factor groupings. Journal of Advanced Transportation 2014; 48(5): 417–430