TRIBHUVAN UNIVERSITY

INSTITUTE OF ENGINEERING

HIMALAYA COLLEGE OF ENGINEERING

A THIRD YEAR PROJECT REPORT

ON

"TIME SERIES ANALYSIS ON PASSENGER COUNT"

[CT-654]

SUBMITTED TO:

DEPARTMENT OF ELECTRONICS AND COMPUTER ENGINEERING

CHYASAL, LALITPUR

SUBMITTED BY:

Ashish Neupane (30006)

Bibek Dhakal (30008)

Bijay Aryal (300010)

Anil Shrestha (300040)

August, 2019

[Time Series Analysis on Passenger Forecasting]

A THIRD YEAR MINOR PROJECT REPORT

[CT-654]

"A Third Year Report Submitted for Partial Fulfillment of Degree of Bachelors' in Computer Engineering"

SUPERVISOR

Er. Chetraj Pandey

SUBMITTED TO:

DEPARTMENT OF ELECTRONICS AND COMPUTER ENGINEERING

CHYASAL, LALITPUR

SUBMITTED BY:

Ashish Neupane (30006)

Bibek Dhakal (30008)

Bijay Aryal (300010)

Anil Shrestha (300040)

August, 2019

ACKNOWLEGMENT

The success and final outcome of this project required a lot of guidance and assistance from many people and we are extremely privileged to have got this all along the completion of this project. This project was possible only due to such supervision and assistance and we would not forget to thank them. We respect and thank **Er. Ramesh Tamang**, for providing us such opportunity to do the project work and giving us all the support and guidance which made us to complete the project duly. We are thankful to and fortunate enough to get constant encouragement, support and guidance from all Teaching staffs of Electronics and Computer engineering department which helped us in successfully completing our project work. Also, we would like to extend our sincere esteems to all staff in laboratory for their timely support. We will like to thank our teacher **Er. Chetraj Pandey** sir for helping us to improvise different things related to this project and background.

We are thankful to our college administration for equipping us with all the resources and providing us with the pleasant environment to work on. Also, we would like to thank our seniors and colleagues for their valuable comments and suggestion throughout the making of the project.

ABSTRACT

Data science is a field of engineering which is related to the analysis and future prediction of data. Generally, Data science is a multifaceted field used to gain insights from complex data. It is a multi-disciplinary field that uses scientific methods, processes, algorithm and systems to extract knowledge and insights from structure and unstructured data. It uses the most powerful hardware and the most powerful programming systems, and the most efficient algorithms to solve problems.

Here in this project **"Time Series Analysis on Passenger Forecasting"** is used to predict the future passenger count on the basis of old data. Our project is capable of going through some clustering and analyzing process. We also conduct some cleaning process to analyze the empty places. After clustering of data, we will predict the data for future by using some probability technique. This provide us the platform to use some technique to solve time series based problems. It helps to generate sufficient theory and practice material for research-based study. This project helps us to determine the future data prediction for seven to eight months when data for two years is given. This project is analyzing the data of two years by using some clustering process, cleaning process and predict the future data by validation of hypothesis which are made on normal analysis of data.

TABLE OF CONTENTS

TABLE OF FIGURES

LIST OF ABBREVIATIONS

ACF	Autocorrelation Function
ACH	Autoregressive Conditional Heteroscedastic
AFIMA	Autoregressive Fractionally Integrated Moving Average
ARIMA	Autoregressive Integrated Moving Average
ARMA	Autoregressive Moving Average
CBM	Condition Based Maintenance
CI	Confidence Interval
CSV	Comma-Separated Values
DOM	Daphne Optimization Methodology
PACF	Partial Autocorrelation Function
RMSE	Root Mean Square Value
SARIMA	Seasonal Auto Regressive Integrated Moving Average
TSAPC	Time Series Analysis on Passenger Count
TSV	Tab Separated Values

1.INTRODUCTION

1.1 Background

Forecasting involves taking the model fit on historical data and using them to predict future observations. The skill of a time series forecasting model is determined by its performance at predicting the future. Forecasting is a method or a technique for estimating future aspects of a business or the operation. It is a method for translating past data or experience into estimates of the future. It is a tool which helps management in its attempt to cope with the uncertainty of the future.

Time series analysis is a statistical technique that deals with time series data, or trend analysis. Time series data means that data is in a series of particular time periods or intervals. Cross-sectional data: Data of one or more variables, collected at the same point in time. Most commonly, a time series is a sequence taken at successive equally spaced points in time. Thus, it is a sequence of discrete-time data. Examples of time series are heights of ocean tides, counts of sunspots, and the daily closing value of the Dow Jones Industrial Average.

More modern fields focus on the topic and refer to it as time series forecasting. Forecasting involves taking models fit on historical data and using them to predict future observations. Time series analysis comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. Time series forecasting is the use of a model to predict future values based on previously observed values. Time series analysis comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. Time series forecasting is the use of a model to predict future values based on previously observed values. Time series are widely used for non-stationary data, like economic, weather, stock price, and retail sales.

1.2 Problem Statement

In today's world the data analysis on different sector is becoming an important part. Taking these things into count data analysis project is proposed on specific field but can further be used for general purpose. Unicorn Investors wants to make an investment in a new form of transportation – Jet Rail. Jet Rail uses Jet propulsion technology to run rails and move people at a high speed. The investment would only make sense, if they can get more than 1 Million monthly users with in next 18 months. In order to help Unicorn Ventures in their decision, we need to forecast the traffic on Jet Rail for the next 7 months. The system is provided with traffic data of Jet Rail since inception in the test file.

1.3 Objectives

The general objective is to predict the future outcomes with better accuracy using SARIMA model.

The specific objectives are:

- To derive a model for Jet Railways to predict data of passengers.
- To get prediction of future passenger count in Jet Railways.

1.4 Scope of The Project and Applications

If anyone wonders about the scope of time series forecasting then they will be amazed to know there are different scientific application and there are different researches going on about the accurate predictions to be obtained. Predicting the future and taking actions according to the obtained data has been very beneficial in economical prediction and any further casualties that may occur. It is applicable in field of data science and data analysis field. In most cases this type of project is done by researcher for research purposes.

2. LITERATURE REVIEW

Various organizations / employees in Nepal and abroad have done modeling using supported time series data exploitation. The various methodologies such as statistic decomposition models, Exponential smoothing models, ARIMA models and their variations like seasonal ARIMA models, vector ARIMA models using variable time series, ARMAX models i.e. ARIMA with instructive variables has been used. Many studies have taken place within the analysis of pattern.

A long-standing interest in performance metrics can be found in forecasting and prognostics. Forecasting has a long history of employing performance metrics to measure how much forecasts deviate from observations in order to assess quality and choose forecasting methods, especially in support of supply chain or predicting workload for software development Prognostics an emerging concept in condition-based maintenance (CBM) of critical systems in aerospace, nuclear, medicine, etc. heavily relies on performance metrics [1].

A general Methodology referred to as Daphne is introduced which is used to find optimum combinations of methods to preprocess and forecast for time-series datasets. The Daphne Optimization Methodology (DOM) is based on the idea of quantifying the effect of each method on the forecasting performance, and using this information as a distance in a directed graph. Two optimization algorithms, Genetic Algorithms and Ant Colony Optimization, were used for the materialization of the DOM. Results show that the DOM finds a near optimal solution in relatively less time than using the traditional optimization algorithms [2].

In this study building on earlier work on the properties and performance of the univariate Theta method for a unit root data-generating process. A system: derives new theoretical formulations for the application of the method on multivariate time series, investigate the conditions for which the multivariate Theta method is expected to forecast better than the univariate one, evaluate through simulations the bivariate form of the method, evaluate this latter model in real macroeconomic and financial time series.The study provides sufficient empirical evidence to illustrate the suitability of the method for vector forecasting. Moreover, it provides the motivation for further investigation of the multivariate Theta method for higher dimensions [3].

All investors are very keen to know about the trend of the Gold price, whether it will rise or fall. In recent times, the price of Gold has become a hot topic for everyone, it fluctuates rapidly from last some months. In this study, we propose a time series model for forecasting the daily Gold price and use the data set of United State Dollars per ounce from Jan 02, 2014 to Jul 03, 2015 for the said purpose. By using the Box-Jenkins methodology, Autoregressive Integrated Moving Average (ARIMA) model is selected and the model selection criterion (AIC and SBC) shows that ARIMA (1,1,0) and (0,1,1) are close to each other for forecasting the daily Gold price. The forecasted values reveal that ARIMA (0,1,1) is more efficient than ARIMA (1,1,0) on the base of model selection criteria's, Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE) [4].

Looking ahead thirty years is a difficult task, but is not impossible. In this paper we illustrate how to evaluate such long-term forecasts. Long-term forecasting is likely to be dominated by trend curves, particularly the simple linear and exponential trends. However, there will certainly be breaks in their parameter values at some unknown points, so that eventually the forecasts will be unsatisfactory. We investigate whether or not simple methods of long-run forecasting can ever be successful, after one takes into account the uncertainty level associated with the forecasts [5].

The review of the past 25 years of research into time series forecasting. In this silver jubilee issue, we naturally highlight results published in journals managed by the International Institute of Forecasters (Journal of Forecasting 1982–1985 and International Journal of Forecasting 1985–2005). During this period, over one third of all papers published in these journals concerned time series forecasting. The review shows highly influential works on time series forecasting that have been published elsewhere during this period. Enormous progress has been made in many areas, but we find that there are a large number of topics in need of further development. We conclude with comments on possible future research directions in this field [6].

3. SYSTEM REQUIREMENT AND DESIGN

3.1 Software Requirement

Python is used as the core programming language. Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. Python is simple, easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance. Python supports modules and packages, which encourages program modularity and code reuse. Python is known to be an intuitive language that is used across multiple domains in computer science. It's easy to work with, and the data science community has put the work in to create the plumbing it needs to solve complex computational problems. Python is currently among the fastest-growing programming languages in the world, largely due to the ease of learning involved, the explosion of data science and artificial intelligence (AI) in the enterprise, and the large developer community around it.

The different libraries involved are Pandas, Numpy, Sklearn and Matplotlib. Pandas is used for reading and writing datafiles which is in csv format and for analysis of those data. Numpy is used to drop and add columns and perform mathematical calculations. Sklearn is used to develop model, fit the model and for predicting the outcomes. Flask is used as our application framework to connect between the user and the prediction system.

3.2 Feasibility Analysis

Economical: The project is done for research purpose. It doesn't go to high cost as we can easily approach the past dataset from the source. Thus, we can say that the project is economically feasible.

Technical: It is technically feasible as we can analyze the past data by using different analytical process available to us. Time series forecasting only involves the observable data set, researcher and analyzer as technical component.

Operational: This project is operationally feasible because we have gone through the past data and analyze them through different python libraries. The python libraries also help to predict the future data easily. Moreover, we also have taken the accurate data from the distinct source.

3.3 Sequence diagram

This sequence diagram shows the interaction between user, localhost and the prediction system. The positive scenario is assumed in which the user establishes connection and provide dataset to the localhost which is passed to the prediction system in which developer is included that does the prediction and the user can view the result. The localhost acts as the communication layer between user and prediction system.

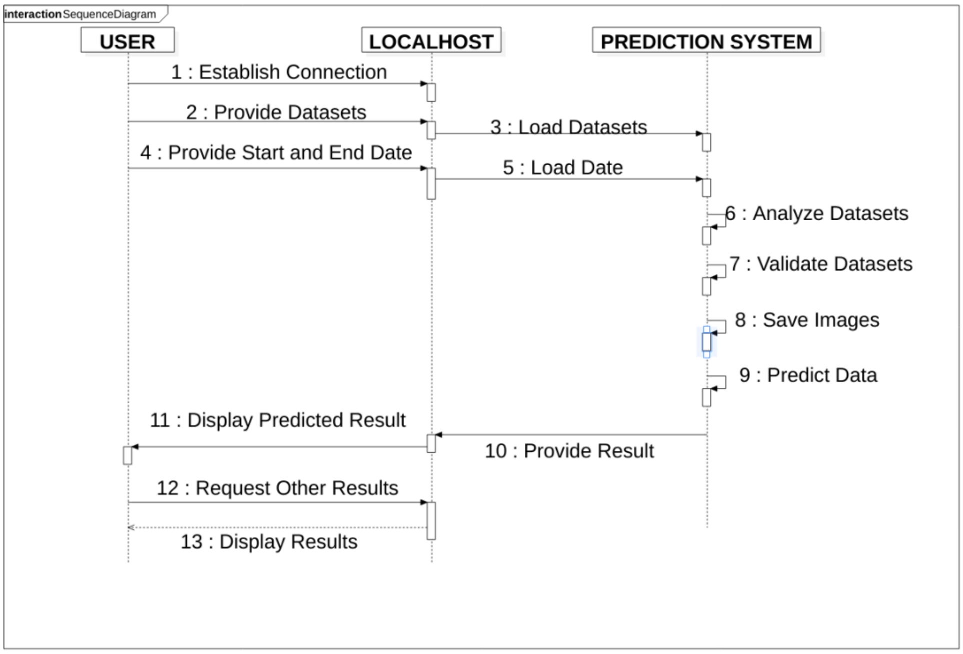

Figure 3. 1: Sequence Diagram of TSAPC.

3.4 Use Case Diagram

The user first establish connection with the system. The developer will manage the connections. Then the user will load the dataset. After that the developer will validate the dataset and generate the model. The developer will also generate results required during generation of model. The user can also ask for result by giving start and end date. After that user can view different type of results as per required.

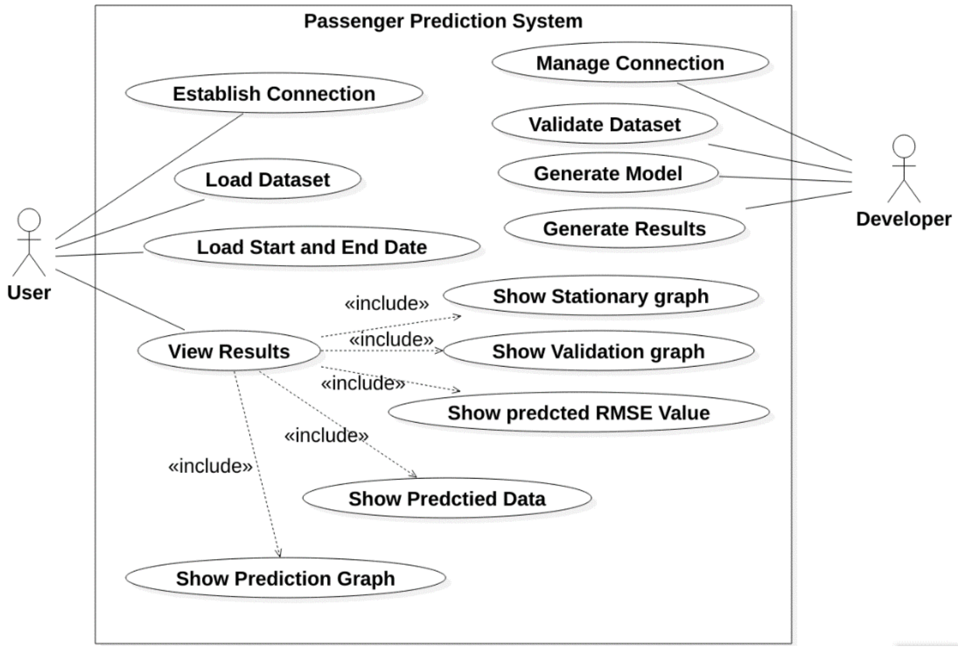

Figure 3. 2: Use Case Diagram of TSAPC.

4. METHODOLOGY

Time Series is one of the most commonly used techniques in data science. It has wide ranging applications – weather forecasting, predicting sales, analyzing year on year trends, etc. This dataset is specific to time series and the challenge here is to forecast traffic on a mode of transportation. The dataset has approximately 18000 rows and 3 columns. Time Series analysis comprises different methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. Time series forecasting is the use of a model to predict future values based on previously observed values.

Data Collection: Data is taken from provided authentic resource (Jet Railways). The data is in .csv files so that we can access this csv file in python as dictionary input.

Data Preprocessing: Data preprocessing process is conducted only after keeping data as .csv file. This step involves removing of columns which are not needed, cleaning the missing values and clustering the data accordingly.

Indexing with Time Series Data : Indexing is done with the provided data. Here, the datetime is indexed on the basis of hourly, daily, weekly and monthly basis for clear visualization.

Visualizing Time Series Data: Some distinguishable patterns will appear when the data is plotted. The time-series has seasonality pattern, such as sales are always low at the beginning of the year and high at the end of the year. There is always an upward trend within any single year with a couple of low in the mid of the year. The data can be visualized using a method called time-series decomposition that allow us to decompose our time series into three distinct components: trend, seasonality and noise.

Time Series Forecasting with SARIMA: The most widely used method for time series forecasting is applied which is known as SARIMA. SARIMA forecasting model is denoted with the notation given as SARIMA (p, d, q)(P, D, Q) s. These mentioned parameters account for seasonality, trend, and noise in data which the system had mentioned above in visualizing time series data.

Fitting of SARIMA model: The original dataset is divided into two parts namely: trained dataset and validation dataset. Trained dataset is 80% of original data from initial and validation dataset is remaining 20% of data from the last portion. Thus, model is fitted on the validation portion to validate the dataset and calculate RMSE.

Predicting and Visualizing result: After getting good RMSE value the future data is forecasted and saved in file called Sarima.csv and finally visualizing the result graphically.

4.1 Workflow diagram

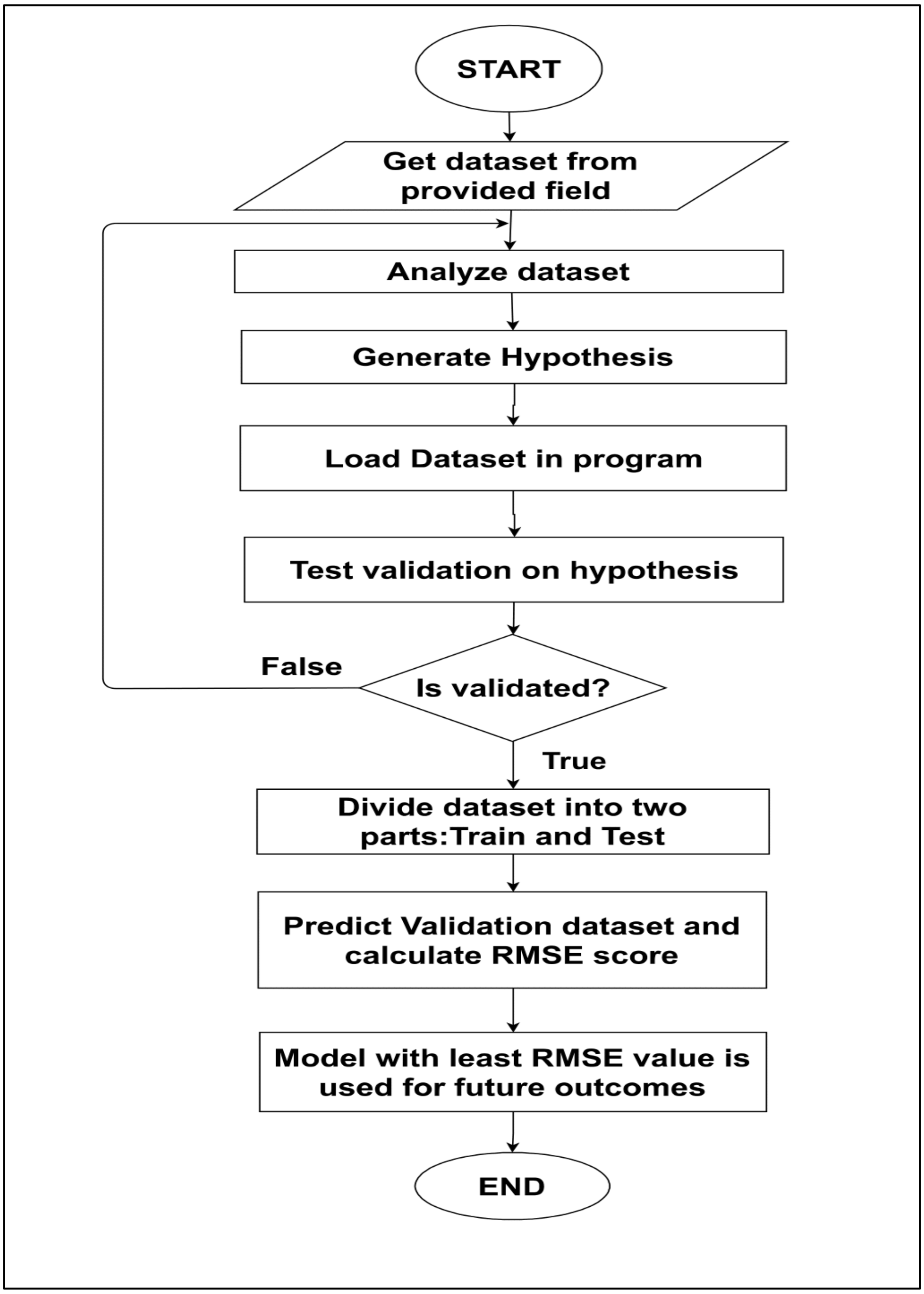

Figure 4. 1: Work flow diagram of TSAPC.

4.2 System flow diagram

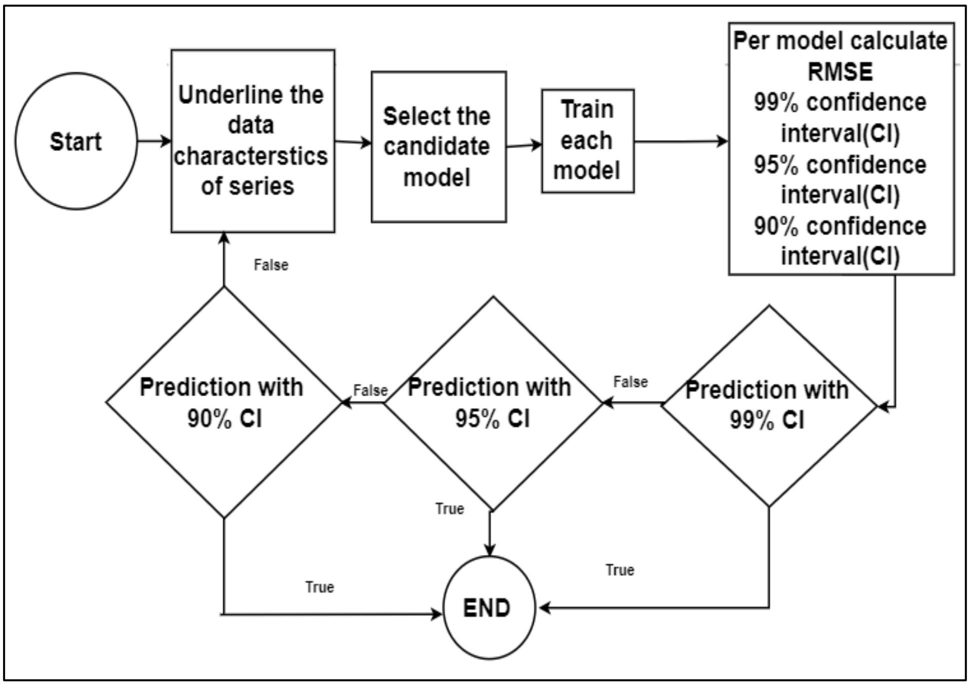

Figure 4. 2: System Flow Diagram of TSAPC.

4.3 Hypothesis

Hypothesis generation helped system to point out the factors which might affect the dependent variables. Below are some of the hypotheses which can affect the passenger count (dependent variable for this time series problem) on the Jet Railways:

- Population has a general upward trend with time, so we can expect more people to travel by Jet Railways. Also, generally companies expand their businesses over time leading to more customers travelling through Jet Railways.
- Tourist visits generally increases during the time period from May to October.
- During working days people will go to office on and hence the traffic will be more on the week days which represent time period from Monday to Friday.
- Traffic during the peak hours will be high because people will travel for college and other working purposes.

4.4 Algorithm for our SARIMA model

1. Construction of SARIMA model.
    - Stationarize the series either by differencing or logging.
    - Study the pattern of ACF and PACF plot to determine if lags of forecast errors should be included in forecasting equation.
    - Calculate ACF and PACF (tools for identifying SARIMA model).
    - Develop the forecasting equation.
2. Fit the SARIMA model.
3. Make prediction with fit model.

5.RESULT AND ANALYSIS

5.1 Result

The data will be trained by implementing that model which validate the hypothesis for this project. In this project the Time Series Analysis is performed on dataset of Jet Railways which has been launching train facilities for two years. This project allows to predict about the passenger count that will be in train at certain time and at certain date. The system took the data from the source provided by the company. After the data get trained then system will transfer the predicted data into csv file named as “Sarima.csv”. This will help to retrieve data which are meant to be forecasted.

RMSE obtained in our project is 68.8776391214778 (RMSE ranges from 67 to 70).

To calculate the RMSE, the following equation is used

RMSE = $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(F_i - O_i)^2}$

where $F_i = forecast\ value$

$O_i = observed\ value$

$n = no.of\ observation$

Residual = $(F_i - O_i)$

ID	Datetime	Count
18288	2014-09-26 00:00:00	492.219239
18289	2014-09-26 01:00:00	392.816705
18290	2014-09-26 02:00:00	332.443923
18291	2014-09-26 03:00:00	274.686536
18292	2014-09-26 04:00:00	231.204488
18293	2014-09-26 05:00:00	206.966154
18294	2014-09-26 06:00:00	222.544816
18295	2014-09-26 07:00:00	258.093206
18296	2014-09-26 08:00:00	297.210430
18297	2014-09-26 09:00:00	384.060814

Figure 5. 1: Results in sarima.csv file

5.2 Output

We have obtained the data in the form of csv format. The csv file is saved in certain directory with name Sarima.csv. The graphical representation of mentioned prediction file Sarima.csv is shown below.

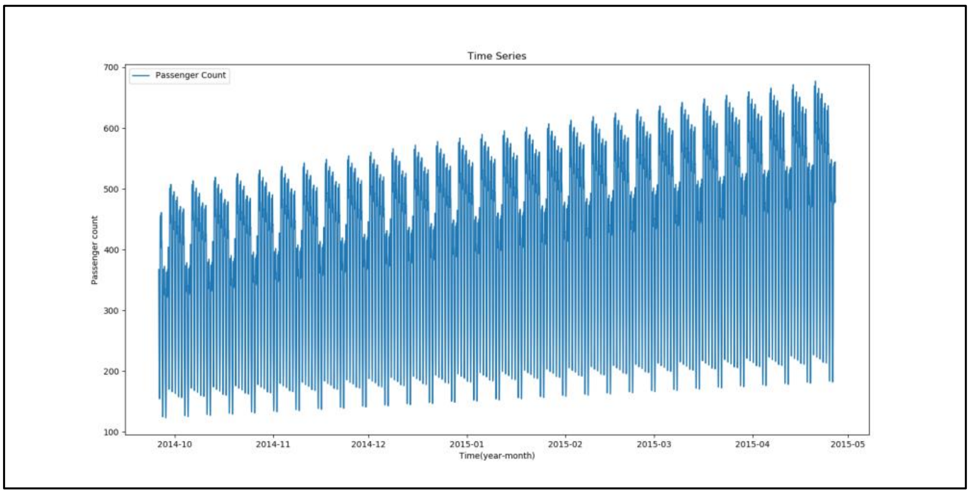

Figure 5. 2: Graphical Representation of Sarima.csv file.

5.3 Limitations

- The system does not cover live data update feature.
- The system is designed to work for only Jet Railways or similar type of data.
- Technical aspect are not considered during prediction. (example: Per Capita Income, Standard of living).

6.DISCUSSION

In this project the data was imported from the source named as "analyticsvidhya.com". After the data was taken the sample data was analyzed and according to this, the optimal hypothesis was generated. The obtained data was further divided into train and test data sets, where train data sets contained about 80% data sets and test data sets contained about the 20% data sets. After the data sets are divided then an appropriate model as per research about different models, SARIMA model was found to be more appropriate. The SARIMA model consist of seven parameters. (i.e. (p, d, q) for trend data and (P, D, Q, s) for seasonal data). Here the parameter p denotes the order of autoregressive model (number of time lags) for trend data and P for seasonal data whereas the parameter d denotes the degree of differencing(number of times the data have had past values subtracted) and q denotes order of moving average (sum of residuals) and s denotes for seasonal cycle(the time series after which it is repeated).In a general SARIMA model these above parameters are generated and equation is developed as order = (2,1,4) and seasonal order = (0,1,1,7) but, in our modified SARIMA model, system change the parameters and develop our forecasting equation as order = (3,1,4) and seasonal order = (0,1,1,7). After training each model based on our equation, then system calculate RMSE value on the confidence interval of 99%, 95% and 90%. On the basis of RMSE value calculated, system observe that the value obtained at 95% CI was found to be optimal. So, system used 5% risk for the prediction of the coming outcomes. For the prediction of a certain interval given by a user, the user gives the start date and the end date as desired and the system predicts the number of passengers on hourly basis at a given interval.

7. CONCLUSION AND FUTURE ENHANCEMENT

7.1 Conclusion

Forecasting is the best renowned method used in various firms to predict the future event. To select a suitable method among various methods is very essential for a firm. This project selects SARIMA as the best method which yields a RMSE value 68.87(RMSE value ranges from 67 to 70). As the firm does not apply any particular method for forecasting their demand, this method can be useful for them to predict their actual demand. The analysis of various predicting techniques for demand forecasting in this project can be a guide for other firms to select their appropriate forecasting method. New approach has been used for passenger prediction. Application of Time Series technique for predicting number of passengers based on the data obtained of the previous few years has been attempted. Accuracy of prediction of count of passenger, using time series is higher than that for prediction of count of passenger using residues. To increase accuracy for prediction, multivariate time series concept may be useful instead of univariate time series.

The business established by Jet railways can be continued because it is obvious that it will earn much profit. According to the analysis done with respect to the prediction system, we found that the passenger count will increase as year pass by. The major aim of our project is to be able for determining future data with respect to the hypothesis taken into the count. So that, we can say that Jet Railways should be continued for longer time.

7.2 Future Enhancement

- Live data update feature can be implemented for better visualization.
- Both Technical and Financial Analysis need to be considered for better result.
- Provided the data are available, the system can be designed to work for all companies listed for Transportation.
- The system is to be trained with healthy amount of data for better result.
- Along with statistical techniques, weather, per capita income could also be considered for better result.

REFERENCES

[1] Carbon and Armstrong, "Note. Evaluation of extrapolative forecasting methods: Results of a survey of academicians and practitioners," *Journal of Forecasting,* vol. 1, no. 2, pp. 215-217, 1982.

[2] I. Kyriakidis, K. Kukkonen, Karatzas, G. Papadourakis and A. Ware, "A Generic Preprocessing Optimization Methodology when Predicting Time-Series Data," *International Journal of Computational Intelligence Systems,* vol. 9, no. 4, pp. 638-651, 2016.

[3] D. D. Thomakos and K. Nikolopoulos, "Forecasting Multivariate Time Series with the Theta Method," *Journal of Forecasting,* vol. 34, no. 3, pp. 220-229, 2015.

[4] K.-H. Jockel and P. Pflaumer, "Die Vorhersage des Goldpreises mit dem Box-Jenkins-Verfahren/Forecasting Monthly Gold Prices with the Box-Jenkins Approach," *Jahrbücher für Nationalökonomie und Statistik,* vol. 196, no. 6, 1981.

[5] C. W. Granger and Y. Jeon, "Long-term forecasting and evaluation," *International Journal of Forecasting,* vol. 23, no. 4, pp. 539-551, 2007.

[6] R. J. Hyndman and J. G. D. Gooijer, "25 years of time series forecasting," *International Journal of Forecasting,* vol. 22, no. 3, pp. 443-473, 2006.

APENDIX

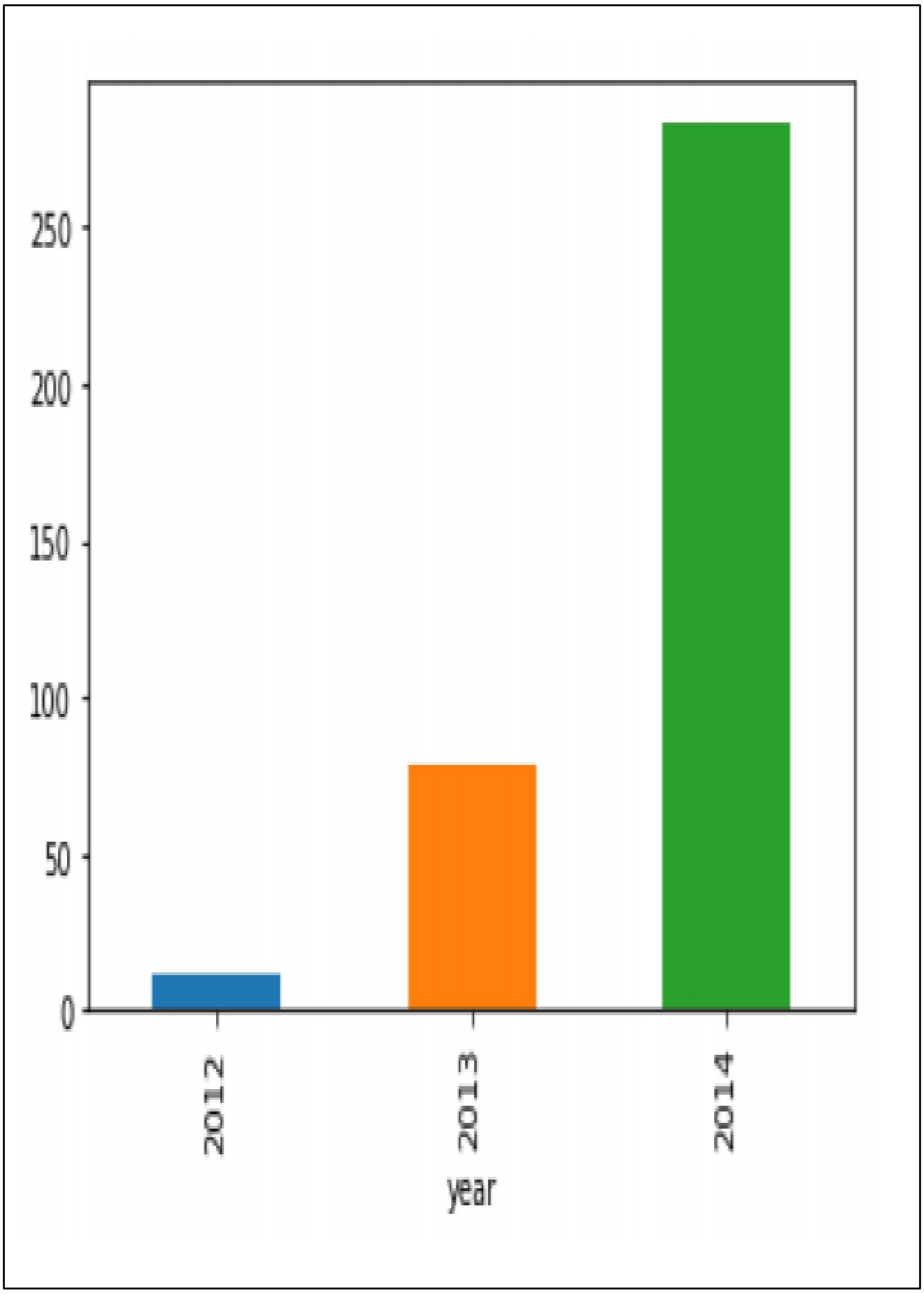

Figure 9. 1: Increasing traffic as year pass by (Hypothesis 1).

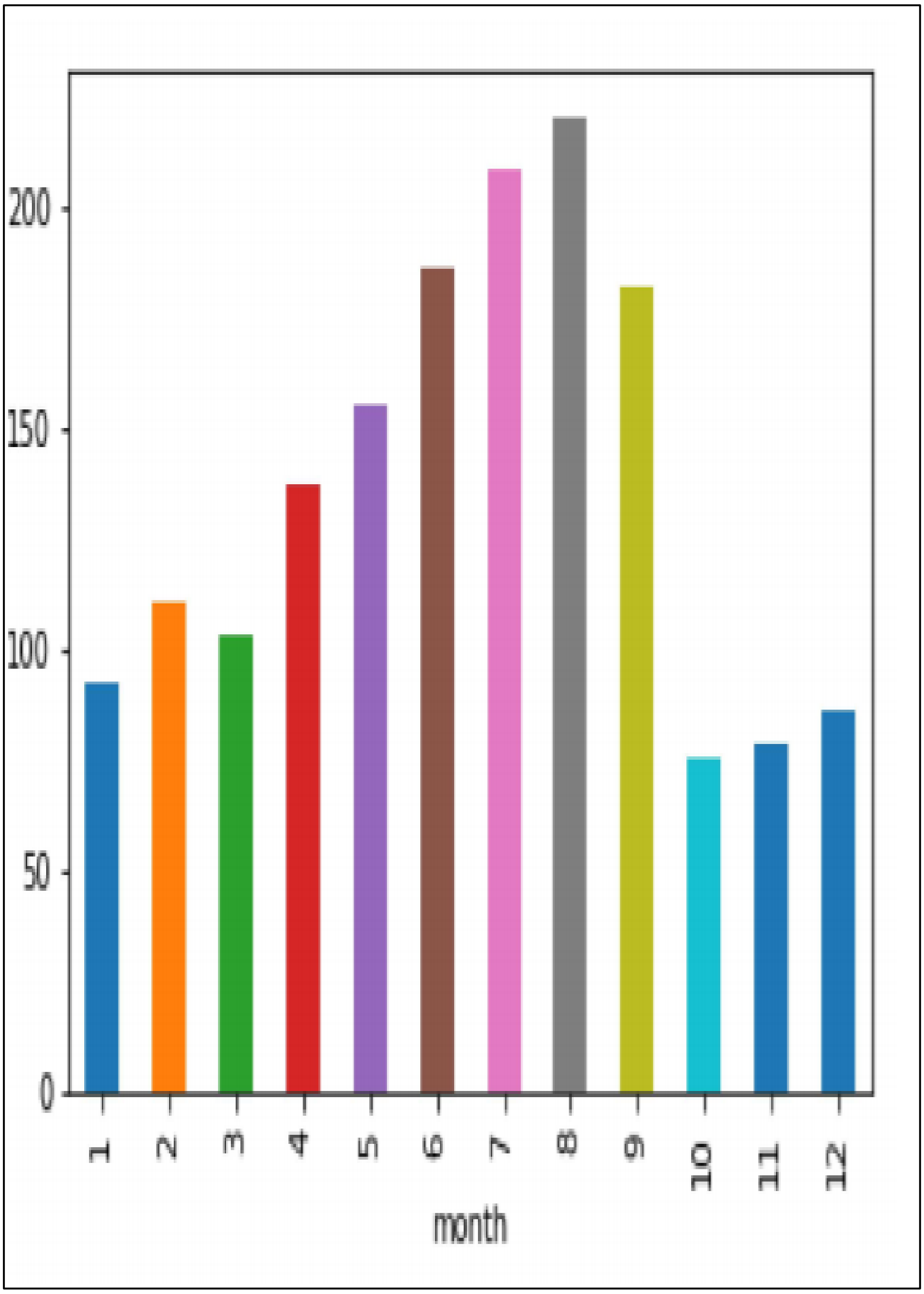

Figure 9. 2: Traffic will increase in May to October (Hypothesis 2).

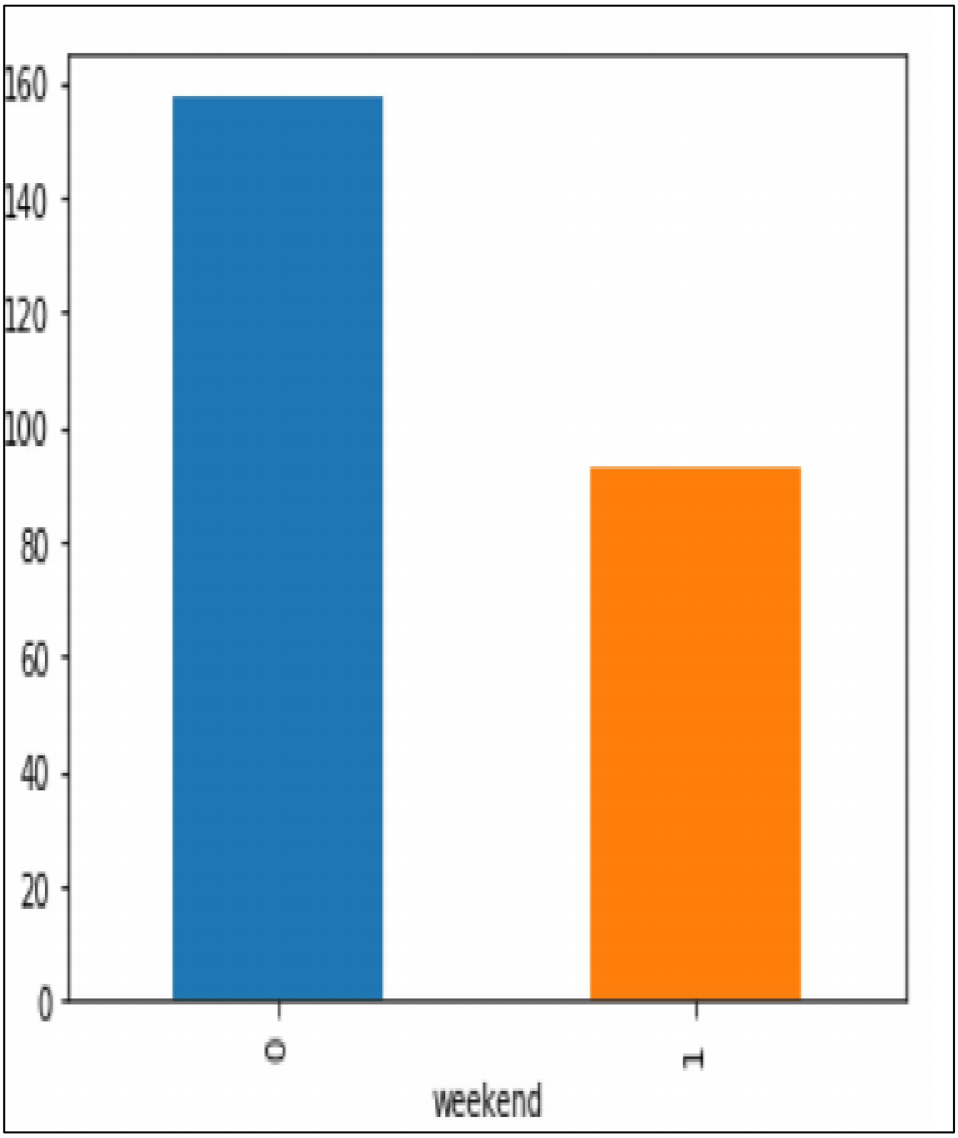

Figure 9. 3:Weekend (1) and Weekdays (0) (Hypothesis 3).

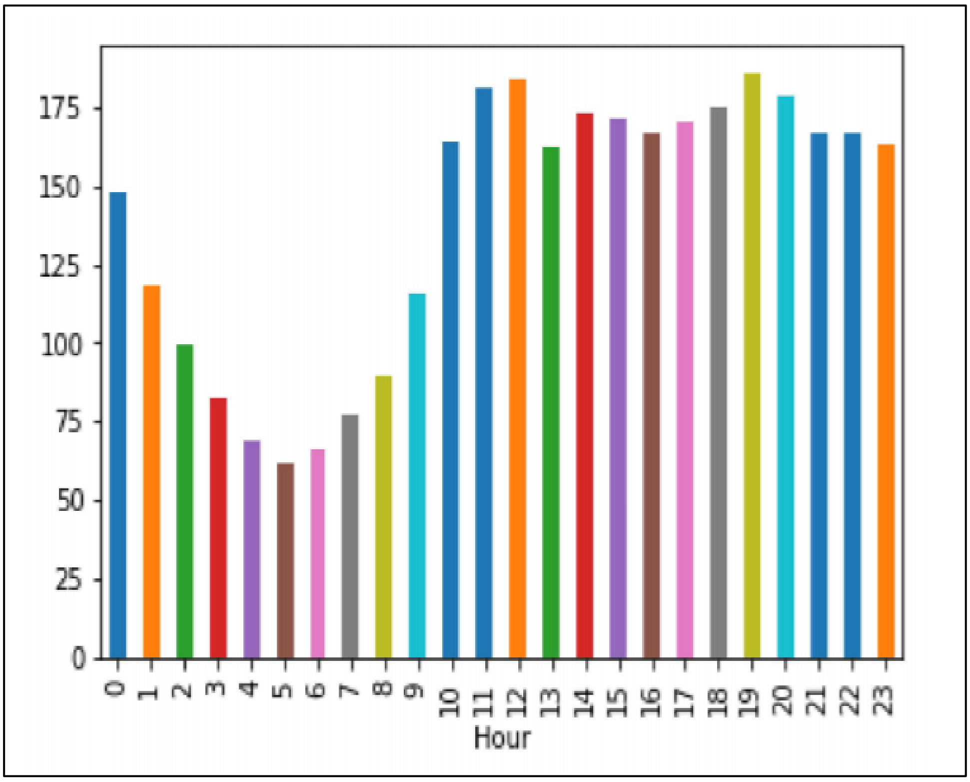

Figure 9. 4: Traffic will be more during peak hour (Hypothesis 4).

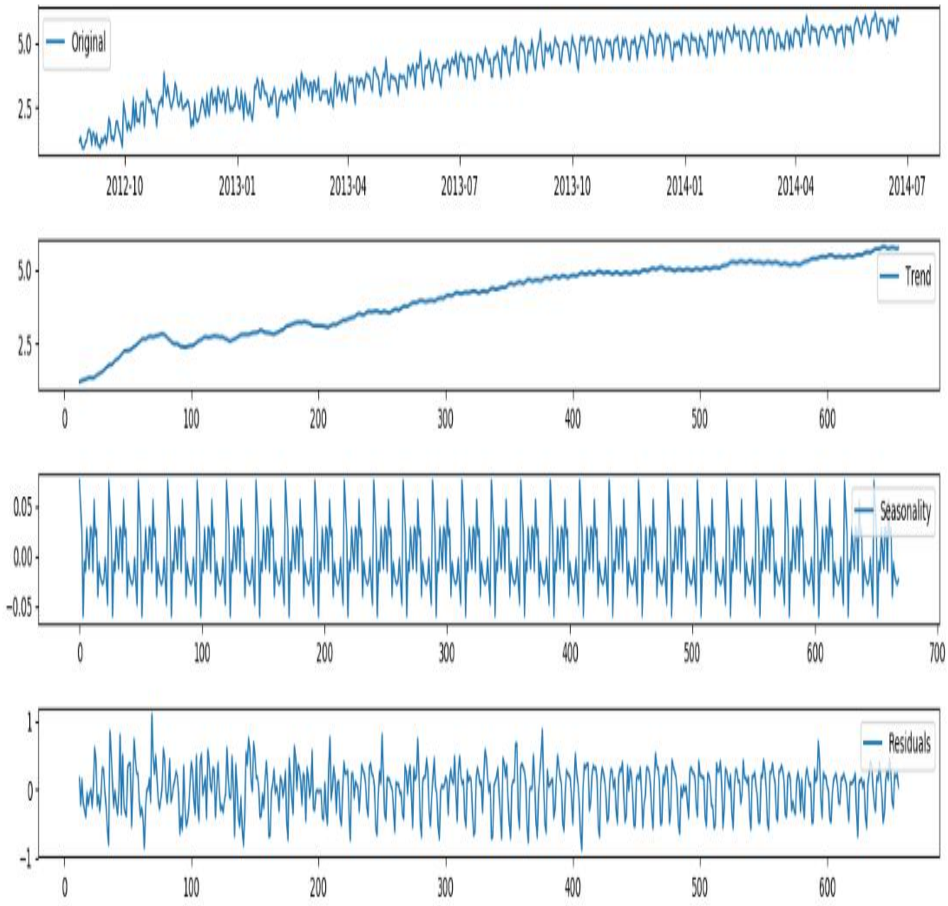

Figure 9. 5: Subplots (Original, Trend, Seasonality and Residuals).

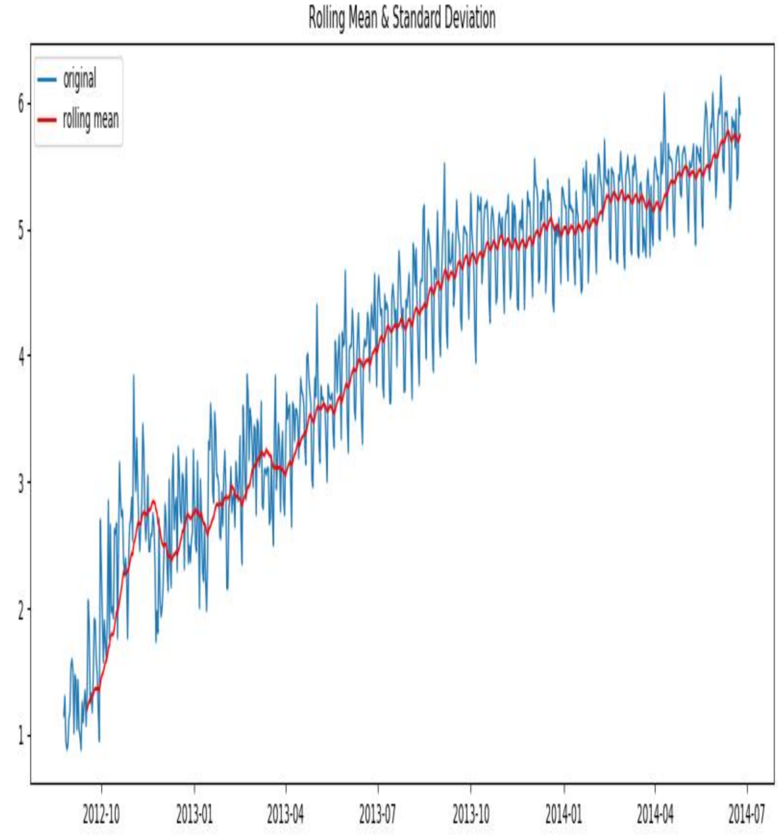

Figure 9. 6: Rolling mean of trained data set.

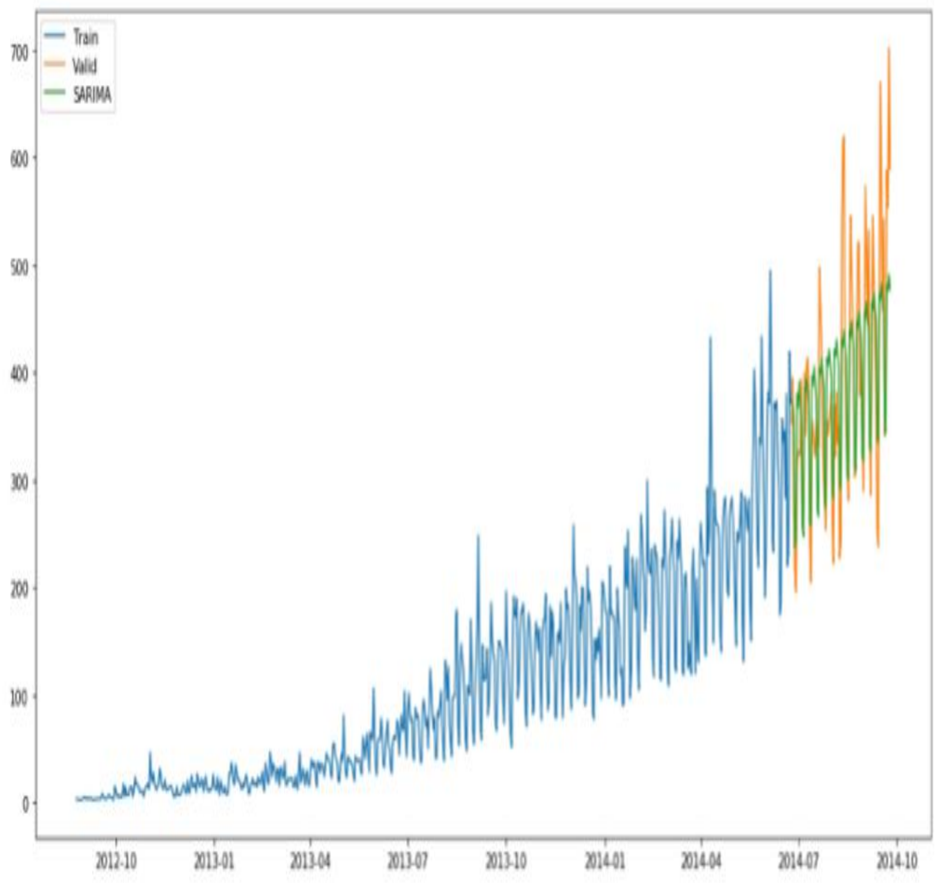

Figure 9. 7: Insertion of validation model data set in test data set.

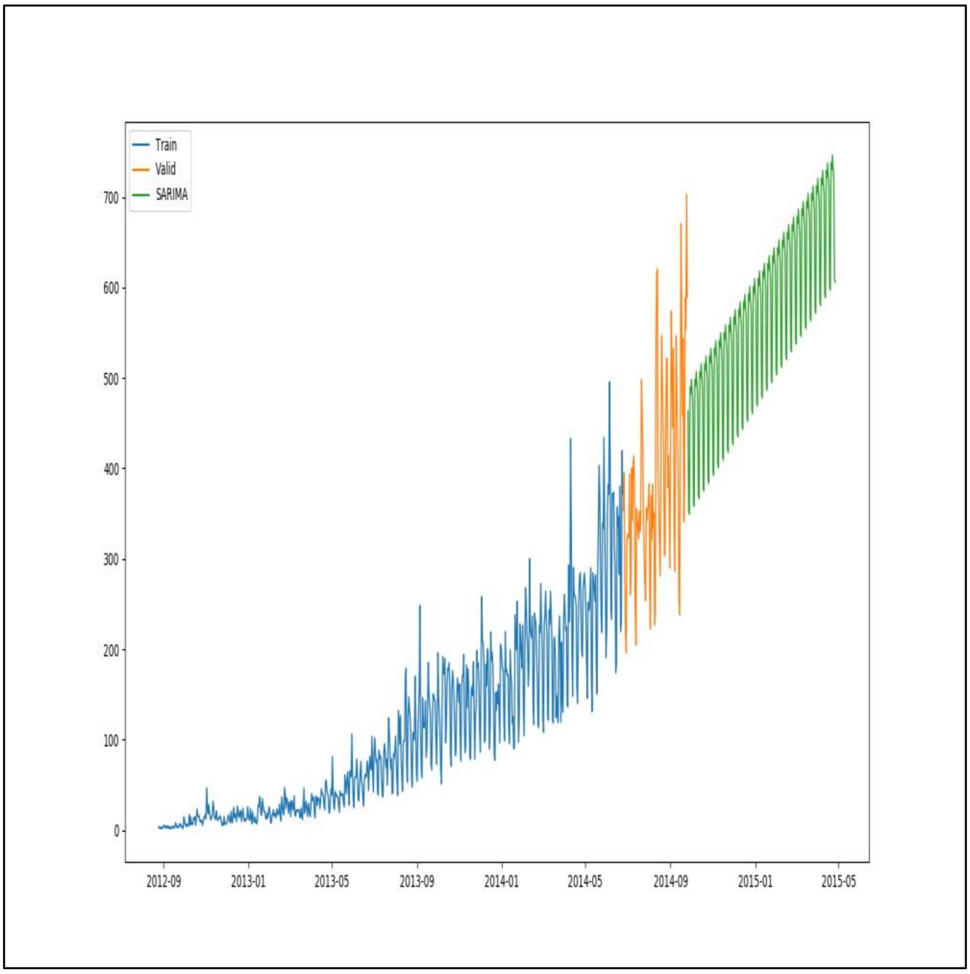

Figure 9. 8: Prediction of data according to model.