TRIBHUVAN UNIVERSITY

INSTITUTE OF ENGINEERING

HIMALAYA COLLEGE OF ENGINEERING

A FINAL YEAR PROJECT REPORT

ON

"Time Series Analysis on Passenger Forecasting"

[CT-455]

SUBMITTED TO:

DEPARTMENT OF ELECTRONICS AND COMPUTER ENGINEERING

CHAYASAL, LALITPUR

SUBMITTED BY:

Ashish Neupane (BCT/2073/06)

Bibek Dhakal (BCT/2073/08)

Bijay Aryal (BCT/2073/10)

Anil Shrestha (BCT/2073/50)

August,2019

[Time Series Analysis on Passenger Forecasting]

A FINAL YEAR MAJOR PROJECT REPORT

[CT-455]

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

SUPERVISOR

Chetraj Pandey

SUBMITTED TO:

DEPARTMENT OF ELECTRONICS AND COMPUTER ENGINEERING

CHAYASAL, LALITPUR

SUBMITTED BY:

Ashish Neupane (BCT/2073/06)

Bibek Dhakal (BCT/2073/08)

Bijay Aryal (BCT/2073/10)

Anil Shrestha (BCT/2073/50)

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 my project. All that we have done is 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 an 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 project and background.

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 use the most powerful hardware and the most powerful programming systems, and the most efficient algorithms to solve problems.

Here in this project in titled “Time Series Forecasting” is a model which is determined by its performance, that is predicting the future. As we know there are so many prediction problems that involve a time component. With reference to time we predict future data. Having a sample of data from some resources, 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 project is specially designed for the people who want to solve problems related to Time Series Forecasting. This provide us the platform to use some technique to solve timeseries based problems. It helps to generate sufficient theory and practice material for research based study. This project helps us to predict future in more reliable way. It develops the seasonal and trend pattern for future data prediction as well. This project helps us to determine the future data prediction for 7 to 8 months when data for 2 years is given. This project is analyzing the data of 2 years by using some clustering and 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

ARIMA	Autoregressive Integrated Moving Average
ARMA	Autoregressive Moving Average
AFIMA	Autoregressive Fractionally Integrated Moving Average
CBM	Condition Based Maintenance
ACH	Autoregressive Conditional Heteroscedastic
SARIMA	Seasonal Auto Regressive Integrated Moving Average

1.INTRODUCTION

1.1 Background

Forecasting involves taking models 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 attempts 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 we want to purpose data analysis project 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, you need to forecast the traffic on Jet Rail for the next 7 months. We are provided with traffic data of Jet Rail since inception in the test file.

1.3 Objectives

The general objective of our project is Time Series Analysis of the past data and use to predict the future outcomes with better accuracy using SARIMAX model.

The specific objective of our project:

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

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.

1.5 Feasibility Analysis

Economical: -This is a project done for research purpose, it doesn't go under high cost. As we can easily approach the past data set 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. Time series forecasting only involves 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 library also helps to predict the future data easily. Moreover, we also have taken the accurate data from the distinct source.

2. LITERATURE REVIEW

Various organizations / employees in Nepal and abroad have done modeling using supported time series data exploitation. The various methodologies viz. 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 (e.g. Carbone and Armstrong, 1982; De Gooijer and Hyndman, 2006). Prognostics an emerging concept in condition- based maintenance (CBM) of critical systems in aerospace, nuclear, medicine, etc. heavily relies on performance metrics [1].

Kyriakidis, Kukkonen, Karatzas, Papadourakis and Ware, (2015) studied 24 metrics used in air quality forecasting. De Gooijer and Hyndman (2006), in a review covering 25 years of time series forecasting, list 17 commonly used accuracy measures. Shcherbakov et al (2013) presented a survey of more than twenty forecast error measures [2].

Relative Root Mean Squared Error: RelRMSE = RMSE/RMSEb,where RMSEb is RMSE from a benchmark method, e.g. Chen, Twycross and Garibaldi (2017), Thomakos and Nikolopoulos (2015). Note that RelRMSE is also known as Theil's U or U2 (De Gooijer and Hyndman, 2006) [3].

Nonlinear time series models consider that irregularity in the observations can be attributed to nonlinear dynamics occuring on a low dimensional chaotic attractor, which can be reconstructed and used to forecast future observations under appropriate conditions. Traditional linear stochastic time series models such as ARIMA models have influenced the forecasting community significantly; however, they cannot capture the nonlinear dynamics underlying real life observations. On the other hand, many useful nonlinear time series models have been proposed.

Right after the introduction of the currently classical Autoregressive Integrated Moving Average (ARIMA) models by Box and Jenkins (1968), Carlson et al. (1970) used several stationary models of this specific family, i.e. Autoregressive Moving Average (ARMA) models, to forecast the evolution of four annual time series of streamflow processes. Today the available models for time series forecasting are numerous and can be classified according to De Gooijer and Hyndman (2006) into eight categories, i.e. (a) exponential smoothing, (b) ARIMA, (c) seasonal models, (d) state space and structural models and the Kalman filter, (e) nonlinear models, (f) long-range dependence models, e.g. the family of Autoregressive Fractionally Integrated Moving Average (ARFIMA) models, (g) Autoregressive Conditional Heteroscedastic/Generalized Autoregressive Conditional Heteroscedastic (ARCH/GARCH) models and (h) count data forecasting. The models from the categories (a)-(g) are of potential interest in hydrology, while they can be implemented for both one- and multi-step ahead forecasting [4].

For the evaluation of forecasting techniques, the indicators suggested most often (Gardner, 2006;De Gooijer & Hyndman, 2006, and others) are based

on the deviation (error) of the estimate made in the period immediately before [5].

This paper seeks to contribute to the debate around a new application by comparing output from a combination of models with that from a global model with the same specifications as the individual models, in this case the adaptive lasso applied to all regressors simultaneously (Google Trends, retail indices and SARIMA). De Gooijer & Hyndman (2006) highlight the benefits of aggregation, in particular comprehensibility where aggregated models can be easily interpreted. Here, aggregation applies to three individual models, each with their own effects: -The SARIMA model reproduces the past time series pattern: -The retail model exploits traditional retail data.

3. SYSTEM REQUIREMENT AND DESIGN

3.1 Software Requirement

Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. Python's 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's 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. It could also be that more companies are moving data projects and products into production. R is not a general purpose programming language like Python. 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.

3.1.1 Python Libraries

i. Pandas: - Pandas stands for Python Data Analysis Library. Pandas is quite a game changer when it comes to analyzing data with Python and it is one of the most preferred and widely used tools in data munging/wrangling if not THE most used one. Pandas is an open source, free to use. What’s cool about Pandas is that it takes data (like a CSV or TSV file, or a SQL database) and creates a Python object with rows and columns called data frame that looks very similar to table in a statistical software (think Excel or SPSS for example).

ii. Numpy: -Numpy is the most basic and a powerful package for working with data in python. If you are going to work on data analysis or machine learning projects, then having a solid understanding of numpy is nearly mandatory.

iii. Sklearn: - Scikit-learn provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python. The library is built upon the SciPy (Scientific Python) that must be installed before you can use scikit-learn. It is on NumPy, SciPy and matplotlib, this library contains a lot of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction. Please note that scikit-learn is used to build models.

iv. Matplotlib:-Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK+. SciPy makes use of Matplotlib. matplotlib targets many different use cases and output formats. Some people use matplotlib interactively from the python shell and have plotting windows pop up when they type commands. Some people embed matplotlib into graphical user interfaces like wxpython or pygtk to build rich applications.matplotlib targets many different use cases and output formats. Some people use matplotlib interactively from the python shell and have plotting windows pop up when they

type commands. Some people embed matplotlib into graphical user interfaces like wxpython or pygtk to build rich applications.

v. Jupyter Notebook: -Jupyter Notebook is a web-based interactive development environment. Jupyter Notebook is flexible: configure and arrange the user interface to support a wide range of workflows in data science, scientific computing, and machine learning. Jupyter Notebook is extensible and modular.

vi. Flask: -Flask is a lightweight WSGI web application framework. It is designed to make getting started quick and easy, with the ability to scale up to complex applications. Flask offers suggestions, but doesn't enforce any dependencies or project layout. It is up to the developer to choose the tools and libraries they want to use. There are many extensions provided by the community that make adding new functionality easy.

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 data has approximately 18000 rows and 3 columns. 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.

The Data: - Data are meant to be taken from some authentic resource (Jet Railways). We have kept the data in .csv files so that we can access this csv file in python as dictionary input.

Data Preprocessing: -We conduct data preprocessing process after we keep data as .csv file. In this step we include removing of columns we do not need, clean the missing values and cluster the data accordingly.

Indexing with Time Series Data: -Indexing is done with this data, therefore, we will use the average hourly values for hour, day, week and month.

Visualizing Time Series Data: -Some distinguishable patterns appear when we plot the data. 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 months in the mid of the year. We can also visualize our data using a method called time-series decomposition that allows us to decompose our time series into three distinct components: trend, seasonality, and noise.

Time Series Forecasting with ARIMA: -We are going to apply one of the most commonly used method for time-series forecasting, known as ARIMA, which stands for Autoregressive Integrated Moving Average. ARIMA models are denoted with the notation ARIMA (p, d, q). These three parameters account for seasonality, trend, and noise in data which we have mentioned above in visualizing time series data.

Fitting of ARIMA model: -We always run model diagnostics to investigate any unusual behavior. It is not perfect, however, our model diagnostics suggests that the model residuals are near normally distributed.

Validating forecasts: -We divide given trained data into two parts namely: trained data and validation data. Trained data now is 80% of previous trained data from initial and validation data is 20% of data from the last portion. So that we can predict the data from trained data and validate it with the validation data. Now we work according to the RMSE (Root Mean Square Error).

Producing and Visualizing forecast: -After getting good RMSE value we forecast the future data.

4.1 Workflow diagram

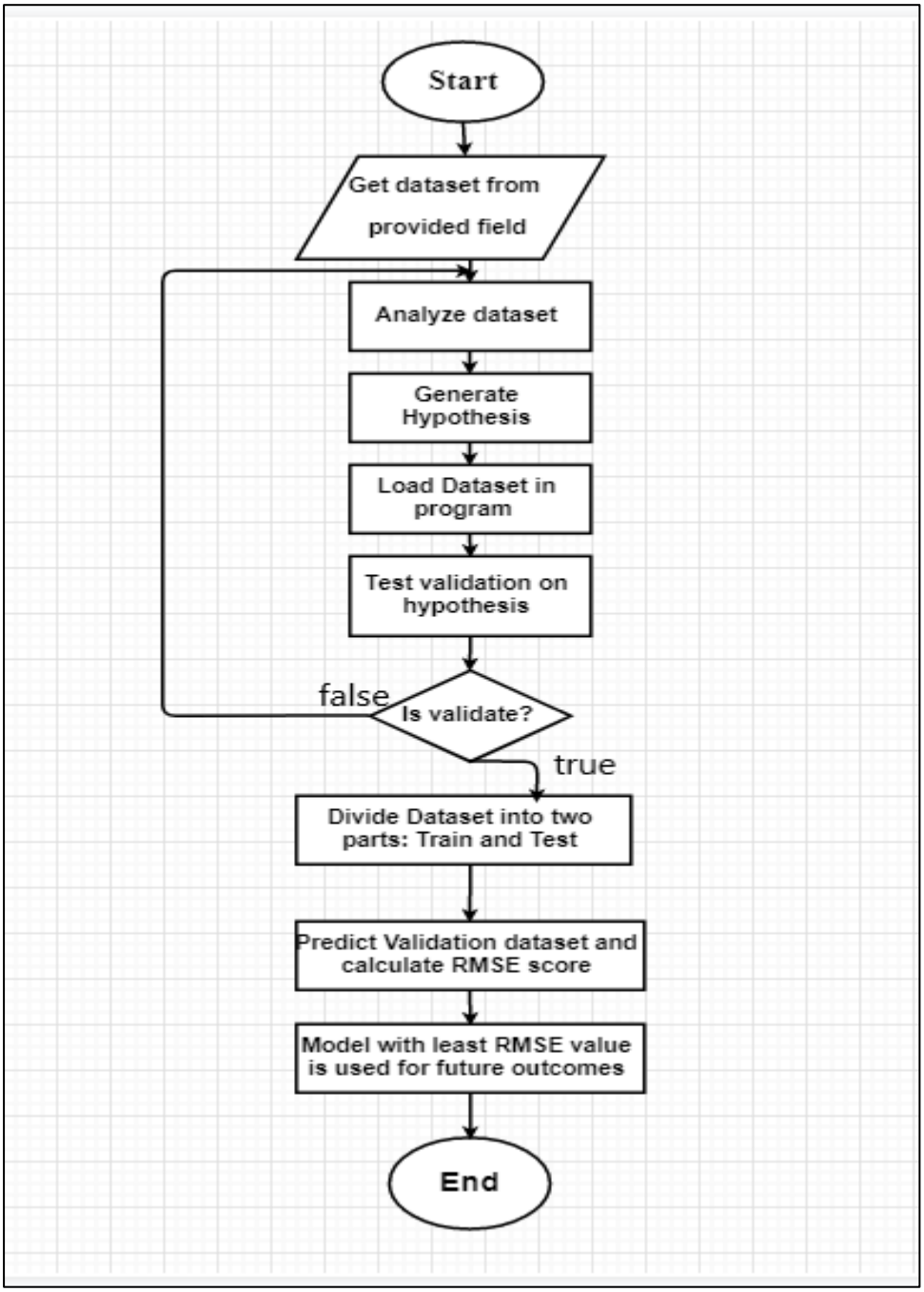

Figure 4. 1: Work flow diagram

4.2 System flow diagram

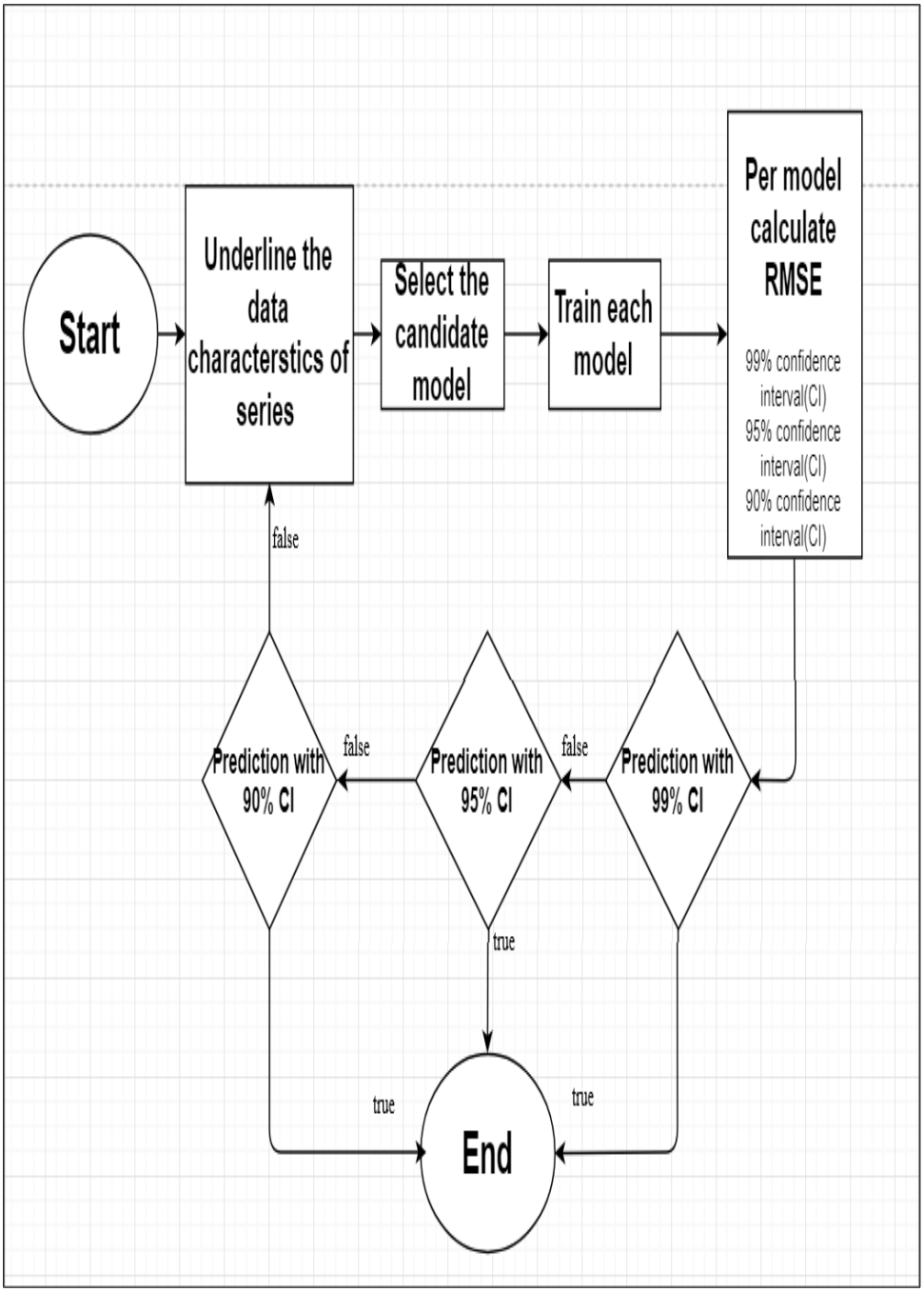

Figure 4. 2: System Flow Diagram

4.3 Hypothesis

Hypothesis testing is an essential procedure in Time Series Analysis. A hypothesis test evaluates two mutually exclusive statements about a population to determine which statement is best supported by the sample data. When we say that a finding is statistically significant, it's thanks to a hypothesis test. How do these tests really work and what does statistical significance actually mean?

Hypothesis generation helps us to point out the factors which might affect our dependent variable. Below are some of the hypotheses which I think 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 JetRail.
- 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.

5.RESULT AND ANALYSIS

5.1 Result

The data will be trained by implementing the model which has validate the hypothesis for our project. In our project we have done the Time Series Analysis on a project named Jet Railways which has been launching train facilities since 2 yrs. We have made this project which will allow us to predict about the passenger that will be in train at certain time at certain date. We took the data from the source provided by the company. After the data get trained then we will transfer the predicted data into csv file named “sarima.csv”. This will help us to retrieve data which are meant to be forecasted.

5.2 Output

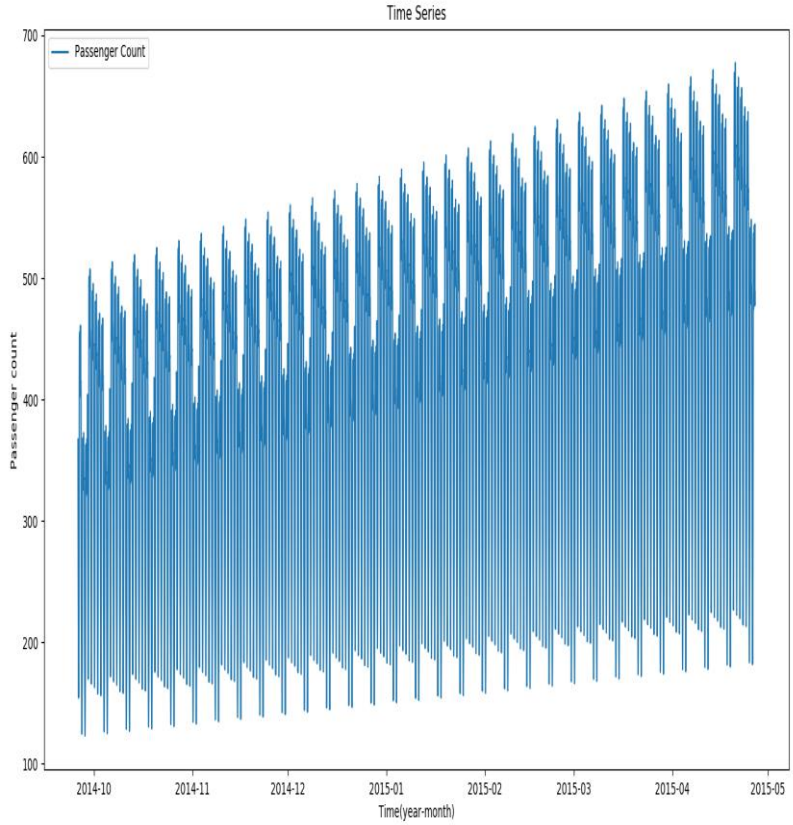

Figure 5. 1: Predicted Output

5.3 Limitations

- The system does not cover live data update feature.
- The system is designed to work for only Jet Railways.
- Only technical analysis is considered for prediction (not included factor as weather, Per Capita Income, Standard of living).
- Not enough data to train.
- For prediction, only two models: Autoregressive Integrated Moving Average (ARIMA) and Seasonal Auto Regressive Integrated Moving Average were considered.

6.DISCUSSION

After completion of this project related to time series analysis and prediction, we can conclude time series analysis is very important to each and every organization for success. It has to understand itself regarding performance, achievements, behaviorally, etc. Each and every organization should ensure use of this method when necessary of forecasting if it is to succeed. A successful business is the one which is in the position to forecast its prospects. In this case, it will be in a position to evade any possible loss that might occur. Time series is done by ensuring that proper planning and putting in place better policies. During forecasting, time series data is usually very important. Many organizations rely on data produced by accountants to forecast prospects of the firm. Company treasures this policy of forecasting since they can be in the position to remedy any future action that can be detrimental to company operations. In this way time series data is very important in forecasting of something that keeps on changing over the long period for example prices of profit, market value, etc.

In this project we found that the business established by Jet railways organization we came to know that the business can be continued because it is obvious that it will earn much profit regarding to the analysis of the time series by being related to the past data of the organization. The major aim of our project is to be able to determine and get a clue or understanding on how and for how long the observation will continue to future, So that we could run the Jet Railways for longer time.

6. CONCLUSION AND FUTURE ENHANCEMENT

6.1 Conclusion

Data science education is well into its formative stages of development; it is evolving into a self -supporting discipline and producing professionals with distinct and complementary skills relative to professionals in the computer, information and statistical science. However, regardless of its potential eventual disciplinary status, the evidence points to robust growth of data science education that will indelibly shape the undergraduate students of the future. In fact, fueled by growing students interest and industry demand, data science education will likely become a staple of the undergraduate experience. There will be an increase in the number of students majoring, minoring, earning certificates, or just taking courses in data science as the value of data skills becomes even more widely recognized. The adoption of a general education requirement in data science for all undergraduates will endow future generation of students with the basic understanding of data science that they need to become responsible citizens. Continuing education programs such as data science boot camps, career accelerators, summer schools, and incubators will provide another stream of talent. This constitutes the emerging watershed of data science education that feeds multiple streams of generalists and specialists in society; citizens are empowered by their basic skill to examine, interpret, and draw value from data. Today, the nation is in the formative phase of data science education, where educational organization are pioneering their own programs, each with different approaches to depth, breadth and curricular emphasis (e.g., business, computer science, engineering,

information science, mathematics, social science or statistics). It is too early to expect consensus to emerge on certain best practices of data science 15 education. However, it is not too early to make recommendations that can help the data science education community develop strategic vision and practices.

6.2 Future Enhancement

- Live data update feature can be implemented for better visualization.
- Provided the data are available, the system can be designed to work for all companies listed for Transportation.
- Both Technical and Financial Analysis need to be considered for better result.
- 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.

9.REFERENCES

[1] Carbon and Armstrong, "Prognostics -an emerging concept in condition based maintenance," *Gooijer and Hyndman ,* 2006.

[2] Kyriakidis, Kukkonen, Karatzas, Papadourakis and Ware, "Air Quality Forecasting," *De Gooijer and Hyndman,* 2006.

[3] Thomakos and Nikolopoulos, "Relative Root Mean Squared Error," *Thomakos and Nikolopoulos,* 2015.

[4] Box and Jenkins, "Autoregressive Moving Average," *De Gooijer and Hyndman,* 2006.

[5] Gardener, "Evaluation of forecasting techniques," *De Gooijer and Hyndman,* 2006.

10.APENDIX

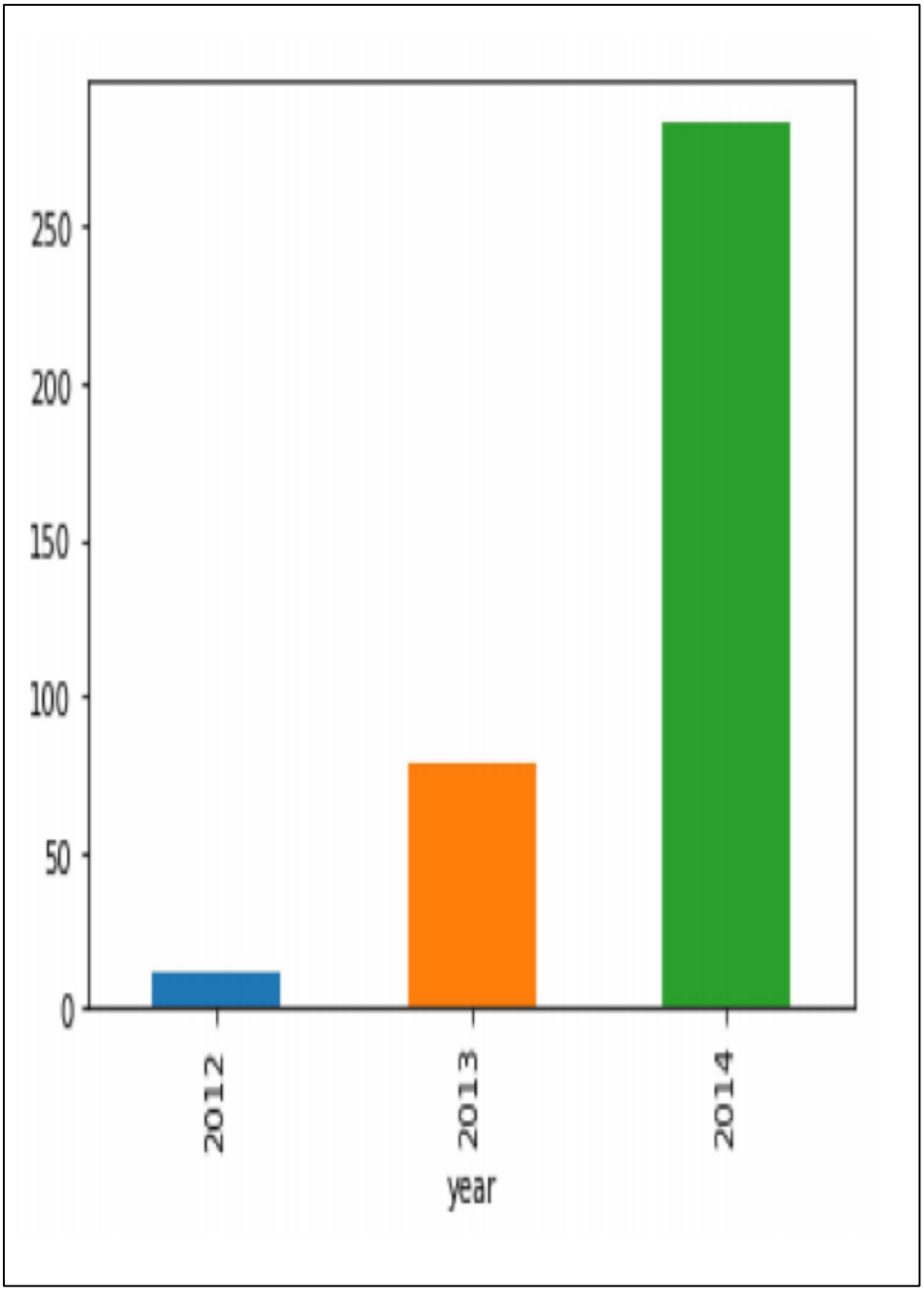

Figure 10. 1: Increasing traffic as year pass by.

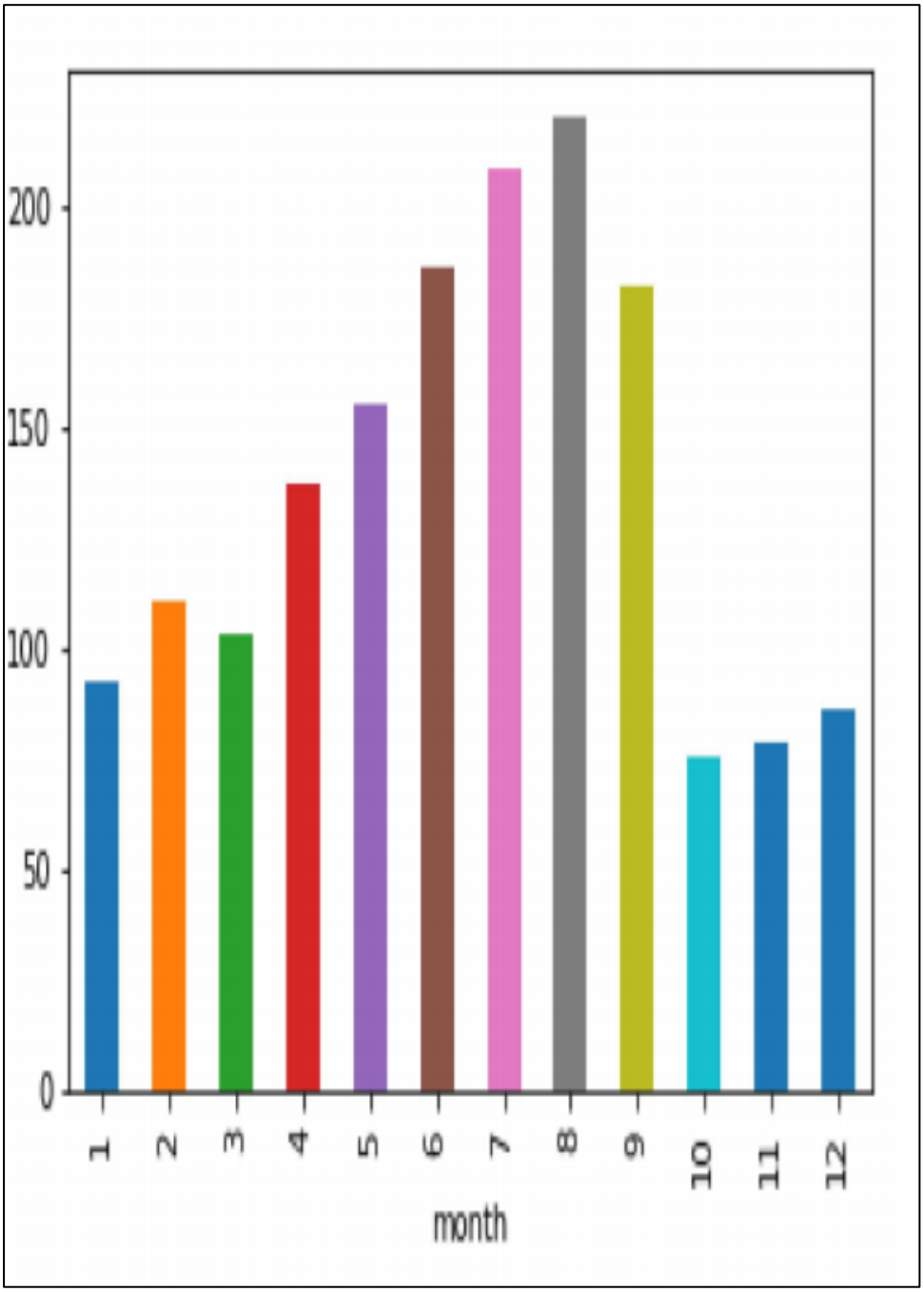

Figure 10. 2: Traffic will increase in may to October.

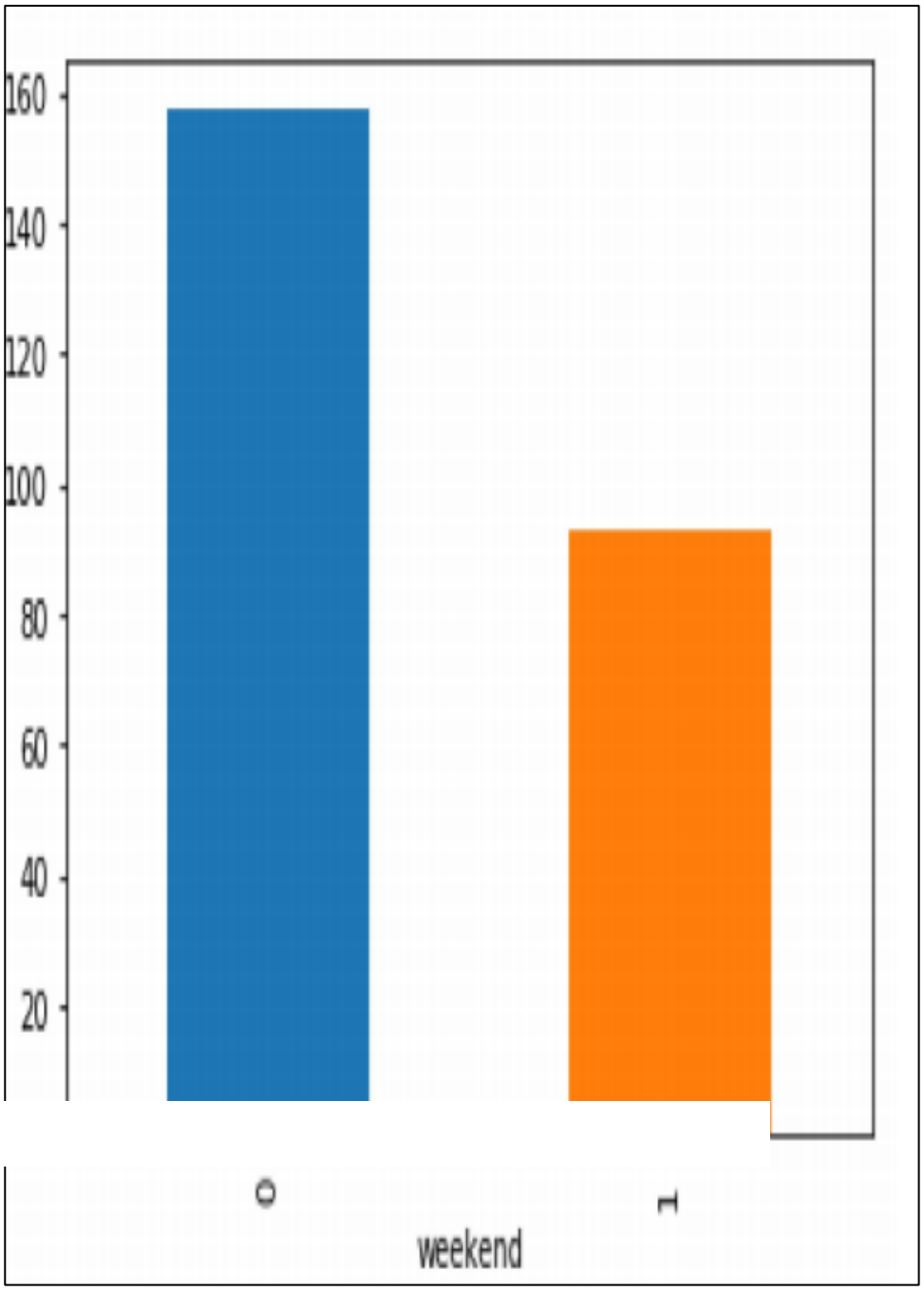

Figure 10. 3: Week end and week days

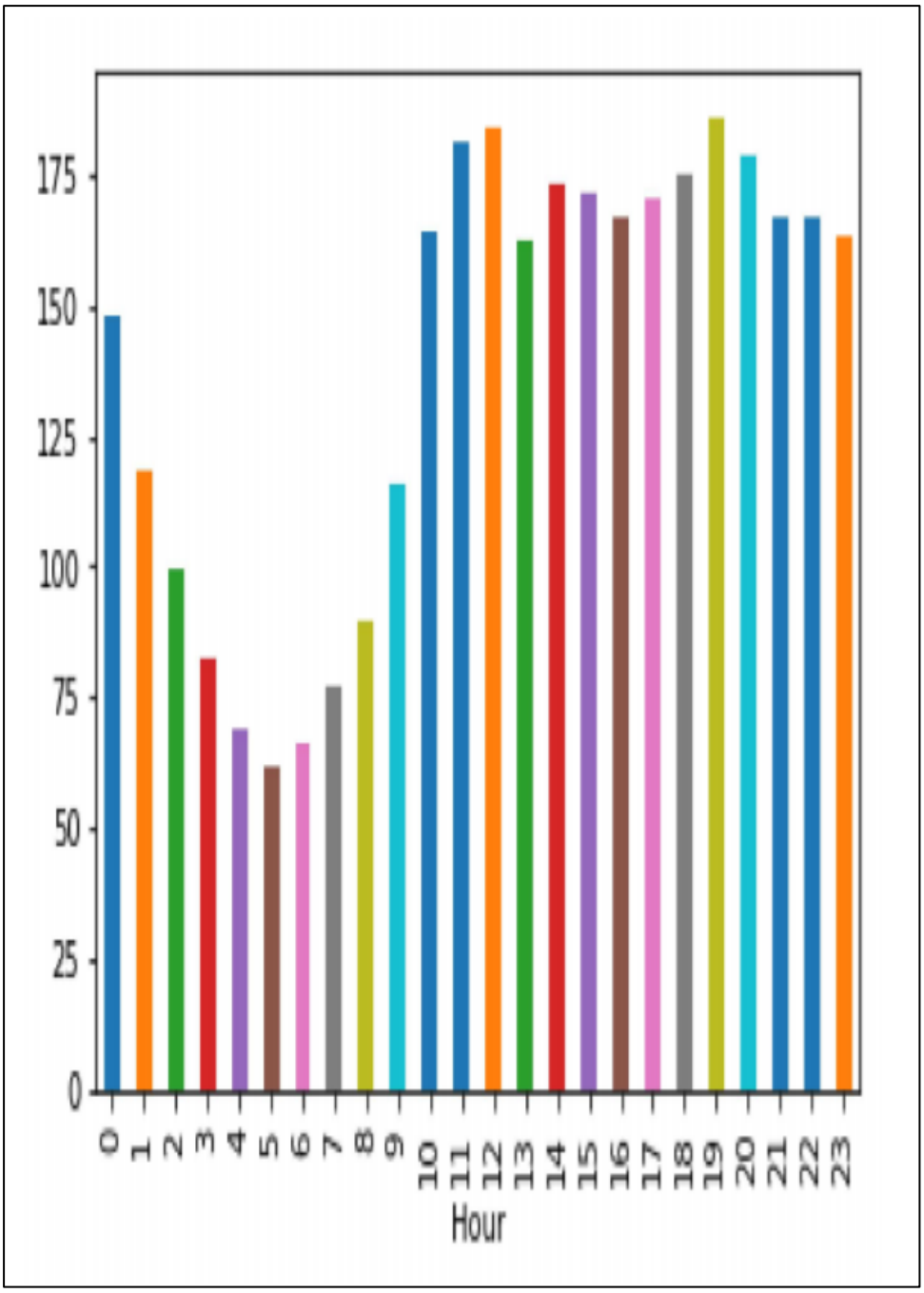

Figure 10. 4: Traffic will be more on during peak hour.

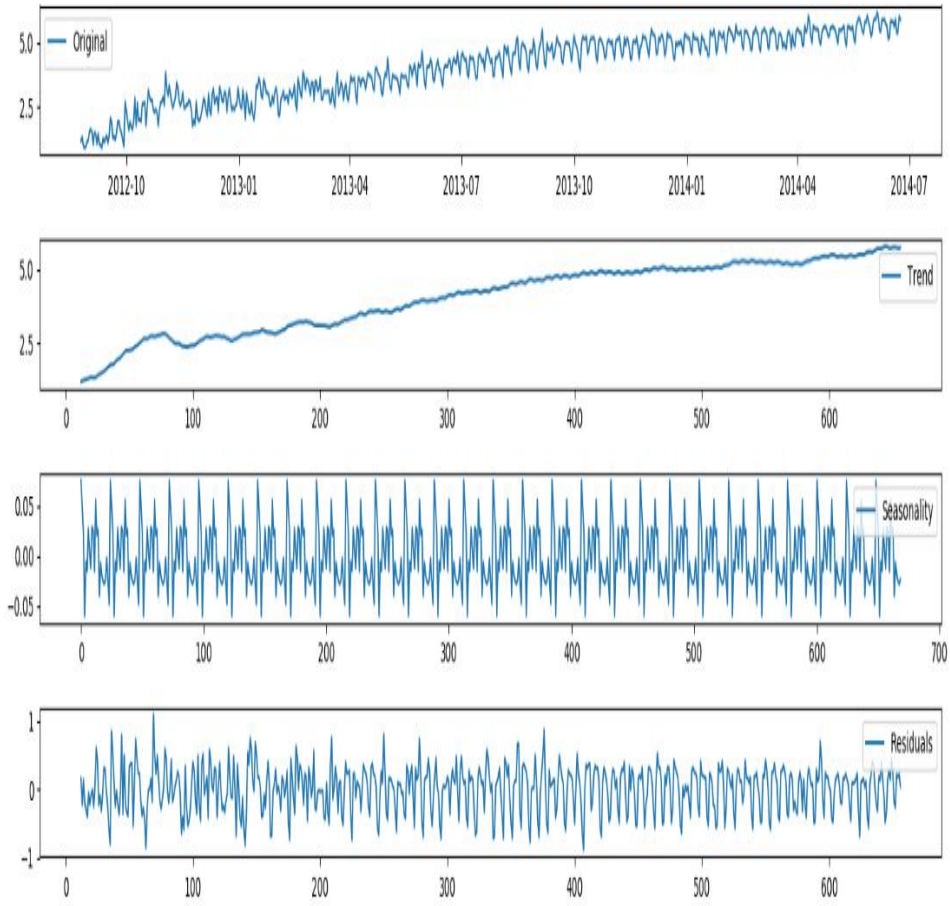

Figure 10. 5: Subplots

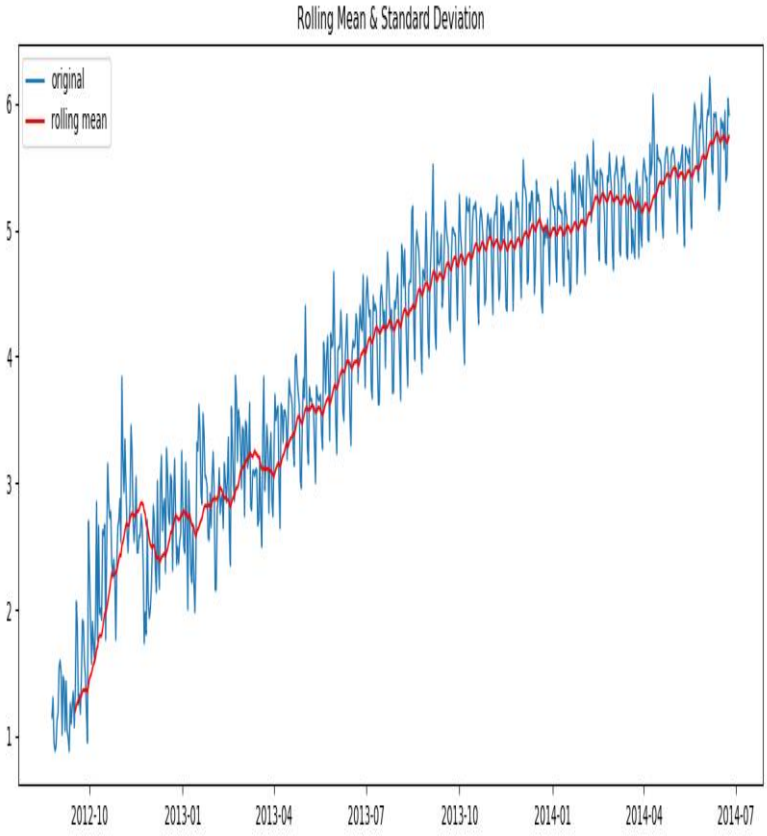

Figure 10. 6:Rolling mean

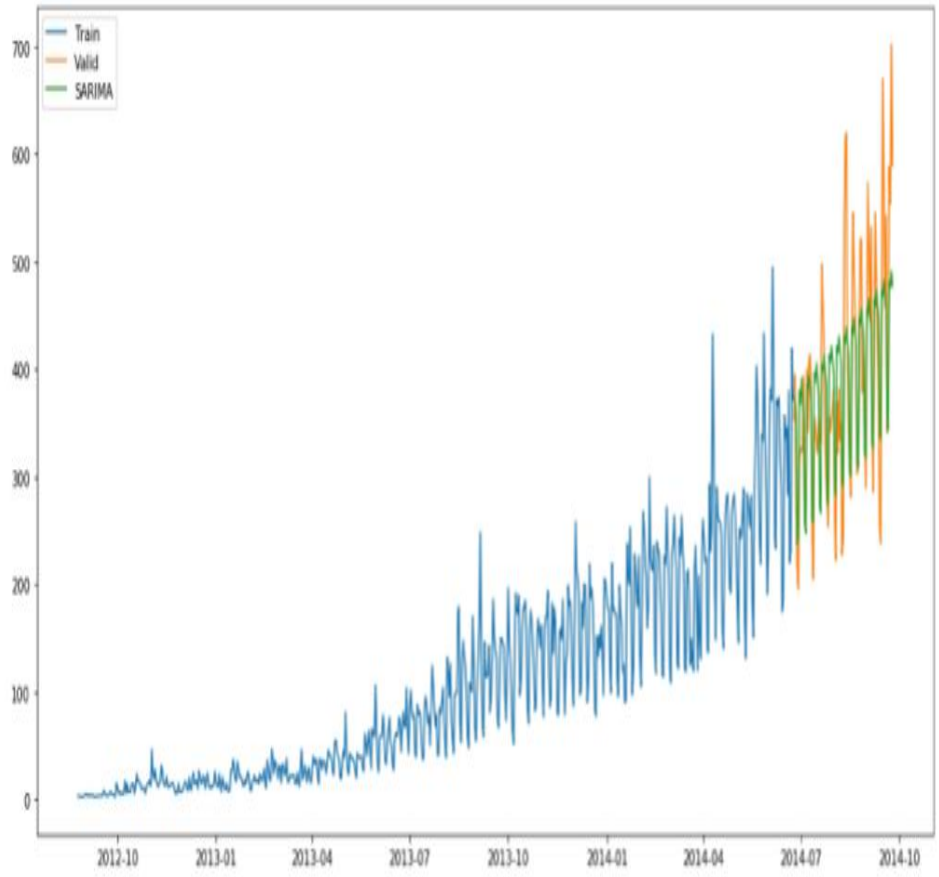

Figure 10. 7:Validation of Model

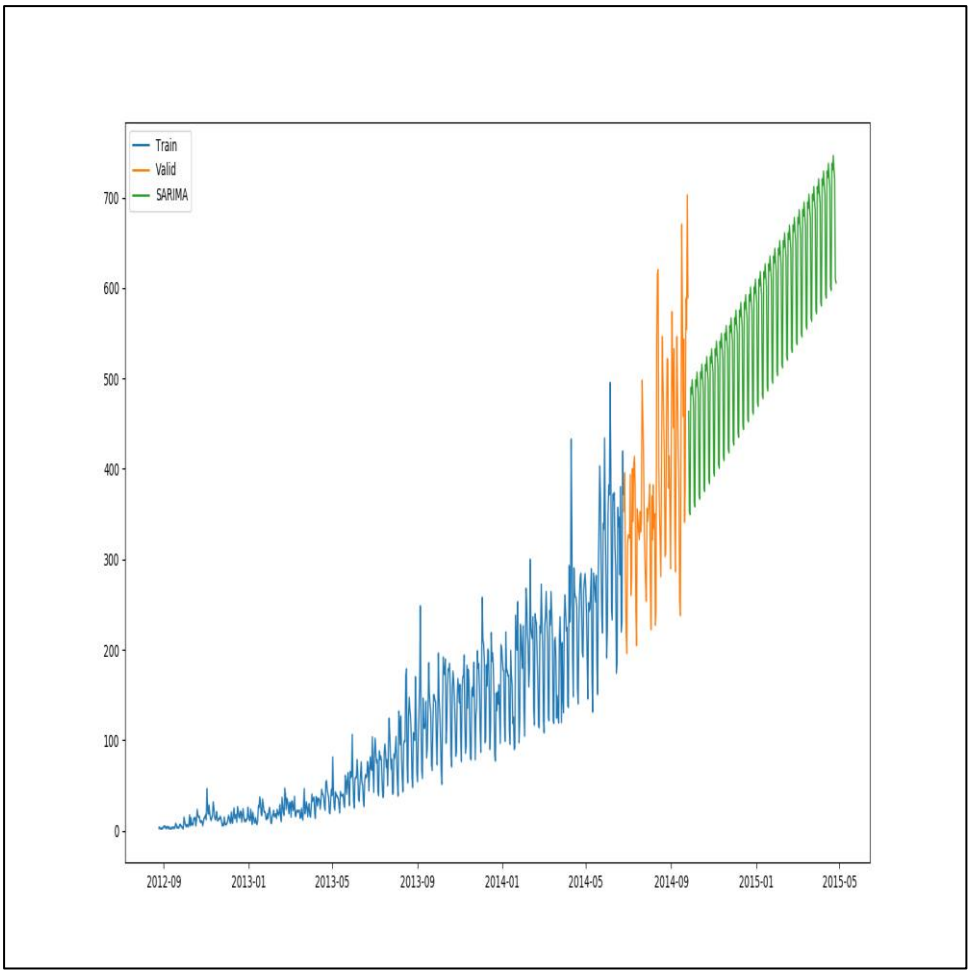

Figure 10. 8:Prediction graph