SALES FORECASTING

A Project Report

Submitted by:

RISHI RAJ SINGH (201B213)

UDAY TAMRAKAR (201B291)

RAVINDER SINGH (201B209)

Name of Supervisor: Dr. Mahesh Kumar

in Partial fulfillment for the Award of the Degree

of

Bachelor of Technology

IN

Department of Computer Science & Engineering

at

JAYPEE UNIVERSITY OF ENGINEERING & TECHNOLOGY,

A-B ROAD, RAGHOGARH, DT. GUNA - 473226, M.P., INDIA

AUG-DEC 2022

DECLARATION

We hereby declare that the work reported in 5th semester Minor project entitled "**SALES FORECASTING**", in partial fulfillment for the award of the degree of B.Tech (CSE) submitted at **Jaypee University of Engineering and Technology, Guna** as per the best of our knowledge and belief there is no infringement of intellectual property rights and copyright. In case of any violation, we will solely be responsible.

UDAY TAMRAKAR (201B291)

RISHI RAJ SINGH (201B213)

RAVINDER SINGH (201B209)

Jaypee University of Engineering and Technology,

Raghogarh, Guna – 473226

Date: 09/12/2022

CERTIFICATE

This is to certify that the project titled "**Sales Forecasting**" is the bona fide work carried out by **Uday Tamrakar , Rishi Raj Singh , Ravinder Singh,** a student of B Tech (CSE) of **Jaypee University of Engineering and Technology, Guna (M.P)** during the academic year 2022-23, in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology (Computer Science and Engineering) and that the project has not formed the basis for the award previously of any other degree, diploma, fellowship or any other similar title.

Signature of the Guide

Jaypee University of Engineering and Technology,

Raghogarh, Guna – 473226

Date: 09/12/2022

ABSTRACT

Sales forecasting is the process of estimating future revenue by predicting the amount of product or services a sales unit (which can be an individual salesperson, a sales team, or a company) will sell in the next week, month, quarter, or year. Sales forecasts help make informed decisions about everything from staffing and inventory to new product lines and potential marketing efforts. There are three basic types qualitative techniques, time series analysis and projection, and causal models. The forecast is generated from an analysis of previous data about your sales, the sale of similar products by your competitors, and market response to your offerings. A sales forecast produces an accurate representation of your projected revenue and growth for your business. The results of the project are expressed in this report. Forecasts that simply sketch what the future will be like if a company makes no significant changes in tactics and strategy are usually not good enough for planning purposes. On the other hand, if management wants a forecast of the effect that a certain marketing strategy under debate will have on sales growth, then the technique must be sophisticated enough to take explicit account of the special actions and events the strategy entails. There are three basic types—qualitative techniques, time series analysis and projection, and causal models.

ACKNOWLEDGEMENT

We would like to express our gratitude and appreciation to all those who gave us the opportunity to complete this project. Special thanks are due to our supervisor **DR. MAHESH KUMAR** whose help, stimulating suggestions and encouragement helped us in all the time of development process and in writing this report. We also sincerely thanks for the time spent proofreading and correcting my many mistakes. We would also like to thank our parents and friends who helped us a lot in finalizing this project within the limited period. Last but not the least I am grateful to all the team members of "**SALES FORECASTING"**.

Thanking You!

RISHI RAJ SINGH (201B213)

UDAY TAMRAKAR (201B291)

RAVINDER SINGH (201B209)

LIST OF FIGURES

Table of Contents

CHAPTER-1

INTRODUCTION

1.1 PROBLEM DEFINITION

Sales forecasting is both a science and an art. Decision makers rely on these forecasts to plan for business expansion and to determine how to fuel the company's growth. For many small business-like clothing shops, grocery shops they many issues while purchasing their stocks because of in-accurate data of sales.

To overcome this problem, we will be providing them less expensive forecasting tools for them as they can't afford highly paid forecasting companies.

1.2 PROJECT OVERVIEW

The project "Sales Forecasting" mainly focuses on

- To improve the current sales trend for small business platforms. The benefits include.
- Cost reduction or Plan budgets with confidence.
- Ease in buying products for future sales.
- Reduction of human resources in purchasing and transporting goods.
- Economic opportunity.
- For effective control and evaluation.

=

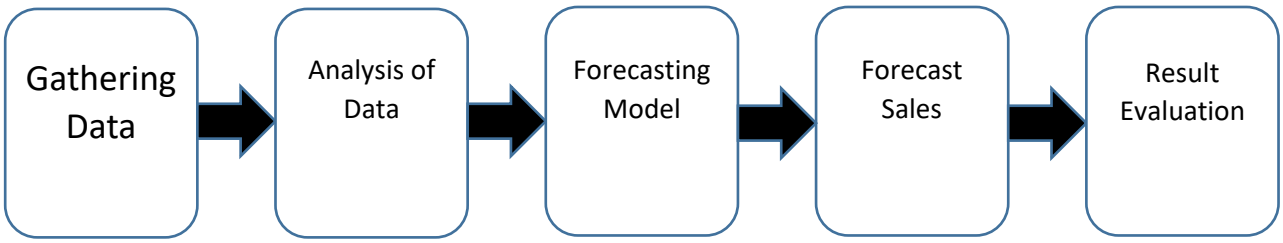

Figure 1.1 Block diagram of Forecasting [3]

The above figure 1.1 shows the flow of any forecasting technique firstly we have to gather the previous sales data for forecasting future sales. Then cleaning of data has to be done for further process, after cleaning of data we have to apply any forecasting model on the data training of data has to be done on data and testing has to be done on some last values of data. After training and testing appropriate model is used for forecasting the future values. [3]

1.1 Hardware Specification

Minimum System Requirement

- CPU (1.0 GHz or faster) or faster 64-bit Dual Core processor like Intel core-2 duo.
- Memory: 4GB (DDR4 | DDR2) RAM or more

1.2 Software Specification

- Python Interpreter
- Microsoft EXCEL
- Browser
- Operating system: windows 10 64bit

CHAPTER-2
LITERATURE SURVEY

2.1 EXISTING SYSTEM

There is some software like – Pipedrive, Anaplan, which uses best models and other stuffs to forecast sales and revenue which they provide at high prices. However, there is no such existing system to help to forecast sales at low price**.**

Pipedrive: Pipedrive's visual sales pipeline, insights, and reports prompt you to take action, remain organized, and stay in control. The revenue forecast reports predict turnover and growth to help spot problems before they arise. And the recurring revenue reports show customer payments that come in multiple installments, so you can accurately generate revenue projections.[7]

Analplan: Anaplan's sales forecasting software allows you to generate accurate, data-driven revenue predictions in real-time that improve decision-making throughout your organization. You can forecast sales by geography, product line, or account, with any level of granularity – e.g. by city or state, by specific product SKUs, or by a particular set of accounts. Plus, you can also get a clearer picture of your sales pipeline health and use the Predictive Insights to help focus sales resources on the best opportunities to drive more consistent revenue.[7]

2.2 PROPOSED SYSTEM

In the Sales Forecasting system, the process flow is initiated by gathering the previous sales data from the user which is stored in form of csv file or excel file. The adfuller test, test the data for stationarity if the data is stationary the we move forward for forecasting the data and if the data is not stationary then software apply differencing on the data with the desired shift (after how many times data is repeating). Then system find the order of the ARIMA on its own and apply the order of ARIMA if the data is seasonal software apply SARIMA. Fitting of the proposed model is done by the software and train and test model on the desired data. Then it forecasts the future values on its own and store in the excel file. The motivation behind this project is to simplify the means by

which small businesses can forecast their sales at cheaper rate as bigger companies charge high prices for the same. So, they can increase their economic growth without any difficulties. We mainly target the small companies.

2.3 FEASIBILITY STUDY

Financial Stability: The price of the software will not be too high and minimal cost will be charged from the people for the software.

Technical feasibility: The software used are freely available in the market and the technologies used are open source which means anyone can contribute in these technologies. The data collected from the user will be bot be stored in the user local system.

Economic Feasibility: Economic feasibility defines whether the expected benefit equals or exceeds the expected costs. It is also commonly referred to as cost/benefit analysis. The procedure is to determine the benefits and the savings expected from the system and compare them with the costs. A proposed system is expected to outweigh the costs.

Operational Feasibility: Operational feasibility is the measure of how well a proposed system solves the problems with the users. Operational feasibility is dependent on human resources available for the project and involves projecting whether the system will be used if it is developed and implemented. The project is operationally feasible for the users as nowadays almost all the teachers/staffs are familiar with digital technology.

CHAPTER-3

SYSTEM ANALYSIS & DESIGN

3.1 Requirement Specification

3.1.1 PYTHON

Python is an interpreted high-level general-purpose programming language. Its design philosophy emphasizes code readability with its use of significant indentation. Its language constructs as well as its object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects.

Python is dynamically typed and garbage-collected. It supports multiple programming paradigms, including structured (particularly, procedural), object- oriented and functional programming. It is often described as a "batteries included" language due to its comprehensive standard library.[4]

3.1.2 GOOGLE COLAB

Colaboratory, or “Colab” for short, is a product from Google Research. Colab allows anybody to write and execute arbitrary python code through the browser, and is especially well suited to machine learning, data analysis and education. More technically, Colab is a hosted Jupyter notebook service that requires no setup to use, while providing free access to computing resources including GPUs. [5]

3.1.3 PyCharm

PyCharm is an Integrated Development Environment (IDE) used for programming in Python. It provides code analysis, a graphical debugger, an integrated unit tester, integration with version control systems (VCSes), and supports web development with Django. PyCharm is developed by the Czech company JetBrains.

It is cross-platform working on Windows, Mac OS X and Linux. PyCharm has a Professional Edition, released under a proprietary license and a Community Edition released under the Apache License. PyCharm Community Edition is less extensive than the Professional Edition.[6]

3.1.4 Excel

Microsoft Excel is a spreadsheet developed by Microsoft for Windows, macOS, Android and iOS. It features calculation or computation capabilities, graphing tools, pivot tables, and a macro programming language called Visual Basic for Applications (VBA). Excel forms part of the Microsoft Office suite of software.[7]

3.1.5 NumPy

NumPy is library of Python programming language, adding support for large, multi- dimensional array and matrice, along with large collection of high-level mathematical function to operate over these arrays. The ancestor of NumPy, Numeric, was originally created by Jim Hugunin with contributions from several developers. In 2005 Travis Olphant created NumPy by incorporating features of computing Numarray into Numeric, with extension modifications. NumPy is open-source software and has many contributors. [8]

3.1.6 Pandas

Pandas is an open-source Python package that is most widely used for data science/data analysis and machine learning tasks. It is built on top of another package named Numpy, which provides support for multi-dimensional arrays. As one of the most popular data wrangling packages, Pandas works well with many other data science modules inside the Python ecosystem, and is typically included in every Python distribution.[8]

3.1.7 Statsmodels

Statsmodels is a Python module that provides classes and functions for the estimation of many different statistical models, as well as for conducting statistical tests, and statistical data

exploration. An extensive list of result statistics is available for each estimator. The results are tested against existing statistical packages to ensure that they are correct. The package is released under the open source Modified BSD (3-clause) license.[9]

3.1.8 Streamlit

Streamlit is an open-source app framework in Python language. It helps us create web apps for data science and machine learning in a short time. It is compatible with major Python libraries such as scikit-learn, Keras, PyTorch, SymPy(latex), NumPy, pandas, Matplotlib etc. With Streamlit, no callbacks are needed since widgets are treated as variables. Data caching simplifies and speeds up computation pipelines. Streamlit watches for changes on updates of the linked Git repository and the application will be deployed automatically in the shared link.[10]

3.2 PROJECT DESCRIPTION

Our Sales forecasting model takes the data entered by user and is being analysed by the software and processed using ARIMA model and to compare with linear regression for accuracy, Data is then given to the forecasting model which process the data to produce the desired result.

3.2.1 ARIMA MODEL

ARIMA, short for 'Auto Regressive Integrated Moving Average' is actually a class of models that 'explains' a given time series based on its own past values, that is, its own lags and the lagged forecast errors, so that equation can be used to forecast future values.

Any 'non-seasonal' time series that exhibits patterns and is not a random white noise can be modelled with ARIMA models.

An ARIMA model is characterized by 3 terms: p, d, q. where,

p is the order of the AR term

q is the order of the MA term

d is the number of differencing required to make the time series stationary

If a time series, has seasonal patterns, then you need to add seasonal terms and it becomes SARIMA, short for 'Seasonal ARIMA'.[11]

3.2.1.1 Seasonal Arima

A problem with ARIMA is that it does not support seasonal data. That is a time series with a repeating cycle therefore, SARIMA is used to forecast on seasonal data. Seasonal Autoregressive Integrated Moving Average, SARIMA or Seasonal ARIMA, is an extension of ARIMA that explicitly supports univariate time series data with a seasonal component.

To check whether the data is seasonal or not we perform Augmented Dickey Fuller test.[11]

3.2.1.2 Stationarity

A stationary time series data is one whose properties do not depend on the time, that is why time series with trends, or with seasonality, are not stationary. the trend and seasonality will affect the value of the time series at different times, On the other hand for stationarity it does not matter when you observe it, it should look much the same at any point in time. In general, a stationary time series will have no predictable patterns in the long-term.[11]

3.2.1.3 Augmented Dickey Fuller Test

The null hypothesis of the ADF test is that the time series is non-stationary. So, if the p-value of the test is less than the significance level (0.05) then you reject the null hypothesis and infer that the time series is indeed stationary. If P-value is greater than the significance level, then we have to do Differencing of data on which we are going to forecast. To import this test, we have to import adfuller from statstools.[11]

3.2.1.4 Differencing

Differencing can help stabilize the mean of the time series by removing changes in the level of a time series, and so eliminating (or reducing) trend and seasonality. A seasonal ARIMA model uses differencing at a lag equal to the number of seasons (s) to remove additive seasonal effects. As with lag 1 differencing to remove a trend, the lag s differencing introduces a moving average term. The seasonal ARIMA model includes autoregressive and moving average terms at lag s. [11]

3.2.1.5 Order of ARIMA

The auto_arima is an automated arima function of this library, which is created to find the optimal order and the optimal seasonal order, based on determined criterion such as AIC, BIC, etc., and within the designated parameter restrictions, that fits the best model to a single variable (univariable) time series. We choose that model whose AIC score is very less among all of the orders obtained from the auto_arima function.[11]

3.2.1.6 Train/Test

In Machine Learning we create models to predict the outcome of certain events. To measure if the model is good enough, we can use a method called Train/Test. Train/Test is a method to measure the accuracy of your model. It is called Train/Test because you split the the data set into two sets: a training set and a testing set. We train the model using the training set and test.[11]

3.2.1.7 Forecast Future Sales

Use the fitted model in Time Series Forecast to predict future values of time series. Compute mean squared errors of the forecast to find the accuracy of the model.[11]

3.2.1.8 Forecasting Error

Forecast error is the difference between the actual and the forecast for a given period. Forecast error is a measure forecast accuracy. There are many different ways to summarize forecast errors in order to provide meaningful information to the manager. Forecast errors can be separated into standard and relative error measures. Standard error measures typically provide error in the same units as the data. Relative error measures are based on percentages and make it easier for managers to understand the quality of the forecast. One of the most popular relative error measures is MAPE, which is the average of the sum of all the percentage errors for a given data without regard for sign. Bias, mean absolute deviation (MAD), and tracking signal are tools to measure and monitor forecast errors.[15]

3.2.1.9 Mean Square Error

Mean squared error (MSE) measures the amount of error in statistical models. It assesses the average squared difference between the observed and predicted values. When a model has no error, the MSE equals zero. As model error increases, its value increases. The mean squared error is also known as the mean squared deviation (MSD).[16]

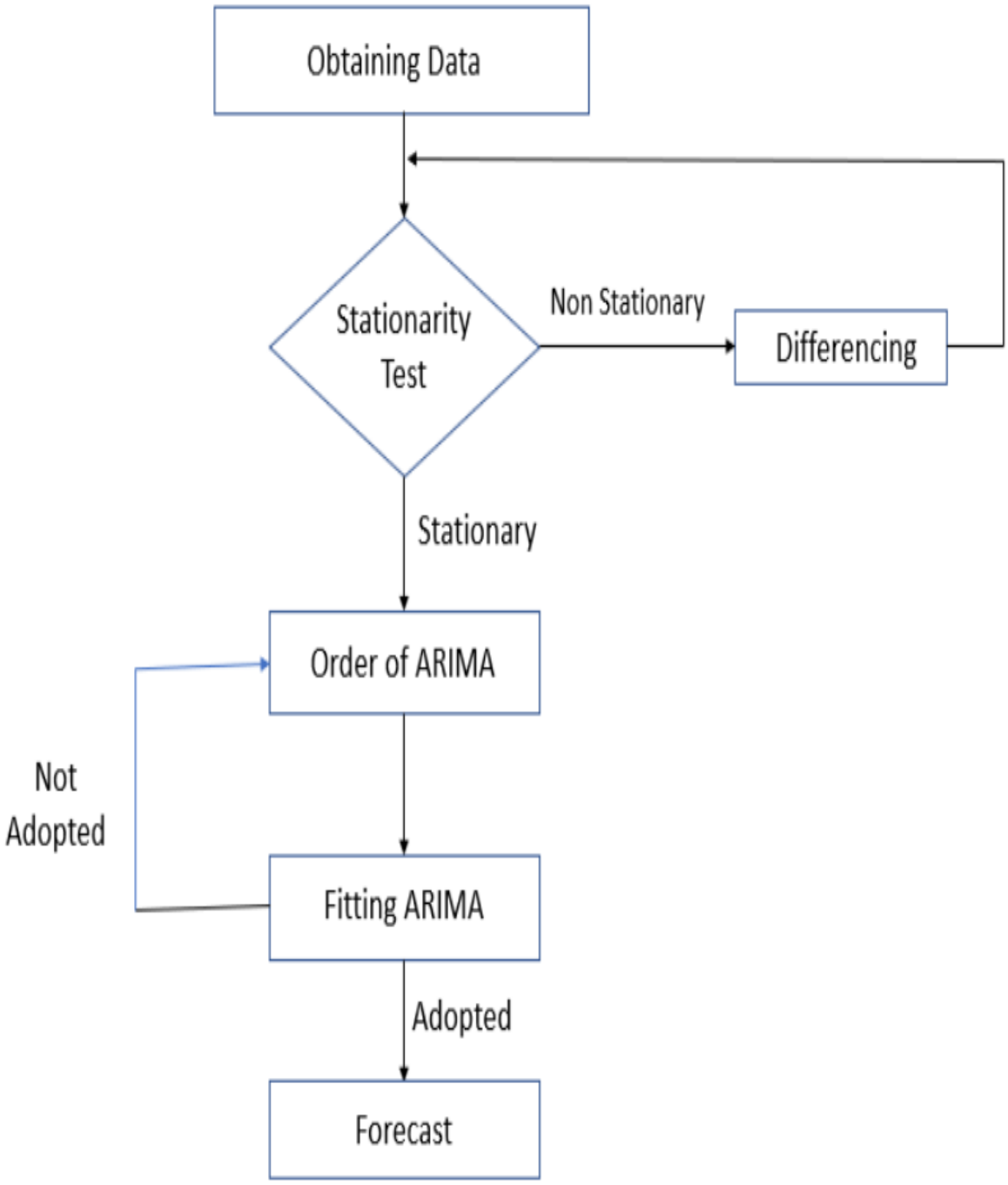

Figure 3.1 Flow Chart of ARIMA Model [2]

The above figure 3.1, shows the flow of ARIMA model first it started with fetching and cleaning the data using pandas and numpy then the stationary test Augmented Dickey Fuller test is done on the data set to test whether the data is stationary or not if the data is non-stationary then the differencing is done on the data set to make it stationary. Then order of ARIMA is calculated with the help of auto_arima function and passed in the ARIMA model, the mean squared error is calculated for checking the accuracy of the model if the accuracy of the model is best then the fitting of the model is done on the dataset. After fitting the desired model then future values are forecasted.

3.2.2 Working Of ARIMA

An **ARIMA** model is a class of statistical models for analyzing and forecasting time series data. It is really simplified in terms of using it, yet this model is really powerful. ARIMA stands for Auto-Regressive Integrated Moving Average. The parameters of the ARIMA model are defined as follows **p** (the number of lag observations included in the model, also called the lag order),**d** (the number of times that the raw observations are differenced, also called the degree of difference.),**q** (The size of the moving average window, also called the order of moving average.)[6]

An ARIMA model can be created using the statsmodels library as follows:

1. Define the model by calling ARIMA() and passing in the p, d, and q parameters.
2. The model is prepared on the training data by calling the fit() function.
3. Predictions can be made by calling the predict() function and specifying the index of the time or times to be predicted.

Finding p, q :

We can use auto_arima function of pmdarima to find the order or value of p and q. Where p is the number of observations included in the model and q is the order of moving average

Finding d :

For ARIMA first thing we do is identify if the data is stationary or non – stationary. if data is non-stationary, we will try to make them stationary then we will process further. We have to check that if the given dataset is stationary or not, For that we use adfuller test.

To identify the nature of data, we will be using the null hypothesis.

H0: The null hypothesis: It is a statement about the population that either is believed to be true or is used to put forth an argument unless it can be shown to be incorrect beyond a reasonable doubt.

H1: The alternative hypothesis: It is a claim about the population that is contradictory to H0 and what we conclude when we reject H0.

#Ho: It is non-stationary

#H1: It is stationary

We will be considering the null hypothesis that data is not stationary and the alternate hypothesis that data is stationary. If the **P-value** greater than **0.05**, which means data is accepting the null hypothesis, which means data is non-stationary that means we have to apply differencing for making the data stationary and repeat the same process till P-value becomes less than 0.05. The number of iterations used in making the data stationary is the value of d.

Define the model by calling ARIMA() and passing in the p, d, and q parameters which are estimated earlier then the model is prepared on the training data by calling the fit() function . Predictions can be made by calling the predict() function and specifying the index of the time or times to be predicted. [11]

3.2.3 Linear Regression

In statistics, regression analysis is a mathematical method used to understand the relationship between a dependent variable and an independent variable. Results of this analysis demonstrate the strength of the relationship between the two variables and if the dependent variable is significantly impacted by the independent variable.

There are multiple different types of regression analysis, but the most basic and common form is simple linear regression that uses the following equation: $Y = bX + a$ [8]

3.3 Sequence Diagram

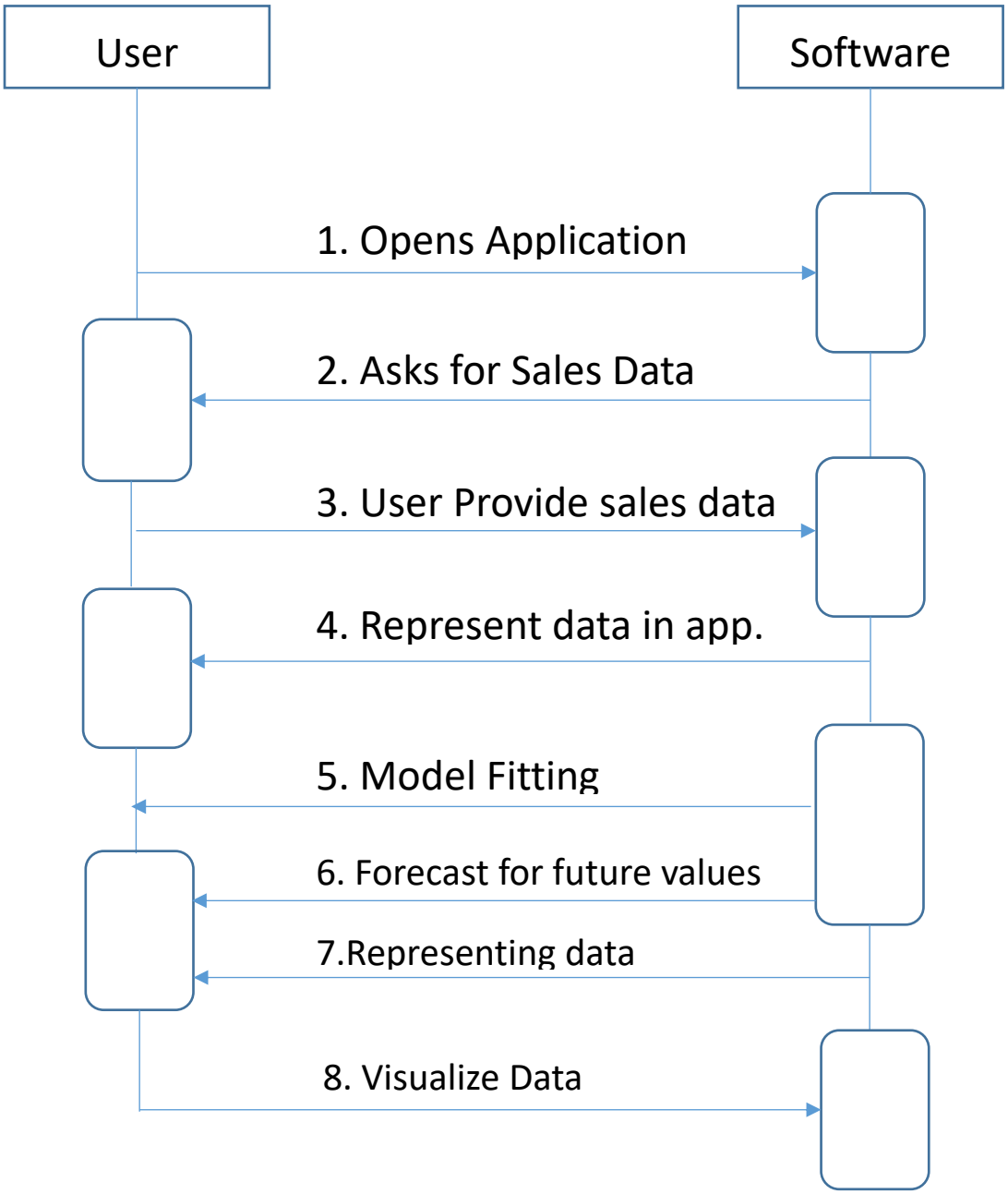

Figure 3.2 Sequence Diagram [6]

Explanation:

The use case diagram is explained as follows:

User open our application and upload its sales data in drop box which only accepts the csv or and the size of file should not exceed 200 MB. After uploading data our application analyses the data, clean the data and represents the data in form of table and chart both, for which user can visualise the data the data is represented in form of charts. The application automatically finds the order of the model used and then application fit the model using the order which is already obtained by the application on the dataset which is provided by the user and testing and training done by the application it also shows the accuracy of forecasting to the user by graphical method and also by Mean Square Error value. Then it forecast the data for desired future dates and presented in form of chats and user can toggle for how many years data should forecast and that data is also displayed using chart. The user can also save the forecasted data into the forecasted.csv file which is also provided by the user itself. The Graphic User Interface is made through streamlit which is in form of webpage.

CHAPTER-4

RESULTS/OUTPUTS

With the help of this project, person is able to forecast his data. We will be able to forecast any product sale using ARIMA on seasonal occasion. Prediction of quantity of products that a costumers will buy. Prediction of monthly sales.

SCREENSHOTS:-

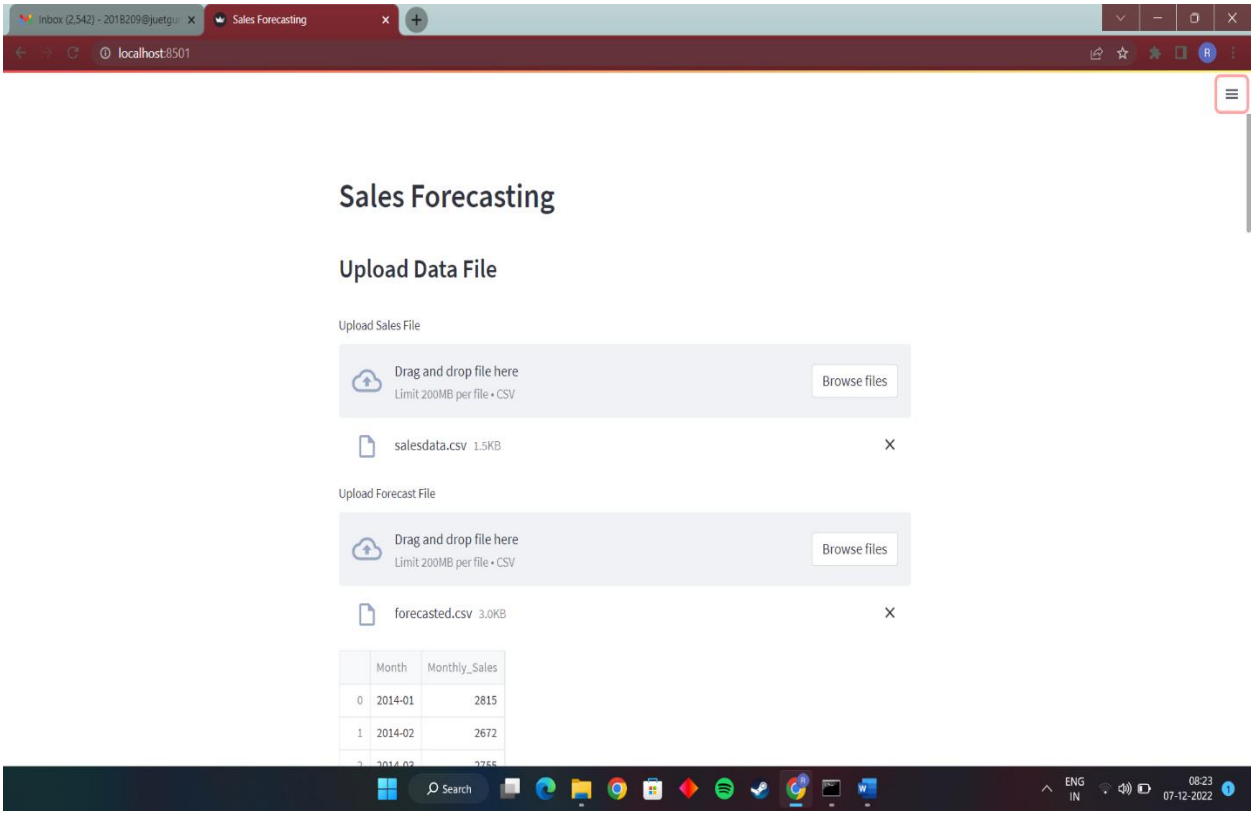

Figure 4.1 Web implementation using Streamlit

Explanation:

In above figure 4.1 we have implemented our project on web page through streamlit. Streamlit[10] is an open-source app framework in Python language. It helps us create web apps for data science and machine learning in a short time. It is compatible with major Python libraries such as scikit-learn, Keras, PyTorch, SymPy(latex), NumPy, pandas, Matplotlib etc. With Streamlit, no callbacks are needed since widgets are treated as variables. Data caching simplifies and speeds up computation pipelines. Streamlit watches for changes on updates of the linked Git repository and the application will be deployed automatically in the shared link.

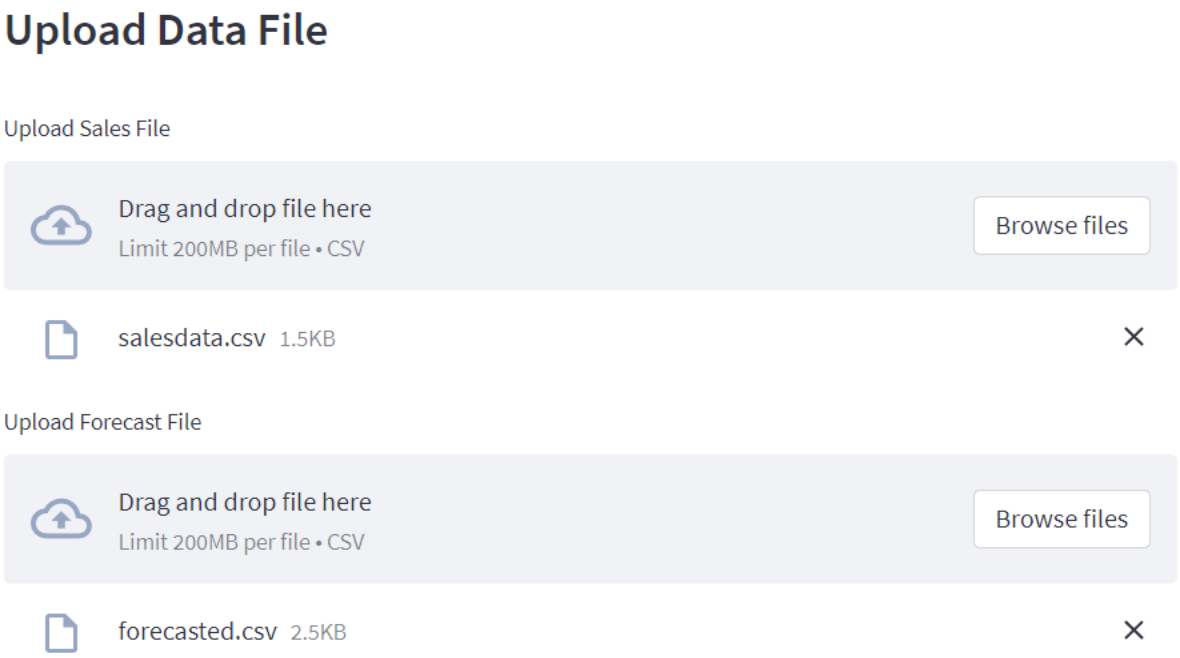

Figure 4.2 Uploading Data interface

In the above figure 4.2 user is going to upload his monthly sales data in the upload filed named as Upload Sales File and the file in which the forecasted values are stored in the field Forecast File field. The size of the data that user can upload is only 200 MB and the file which user can upload only supports csv file.

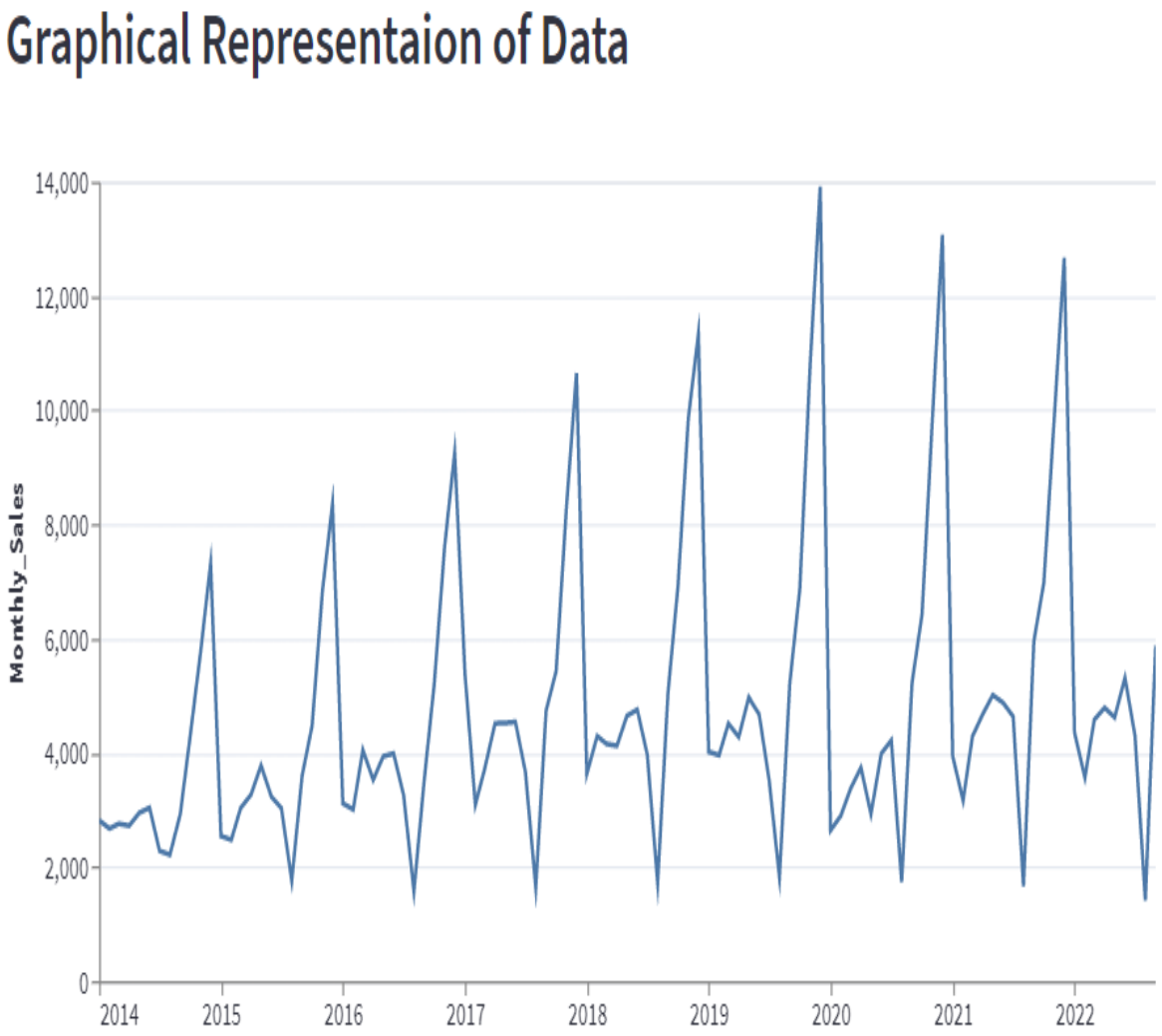

Figure 4.3 Graphical representation of data

In the above figure 4.3 monthly sales data is represented in form of line chard in our webpage where user can hover on line to see value of particular value of month. The Y axis represents the sale of the particular product and X axis represent the month and the year of the sales. The chart is made through streamlit line chart feature.

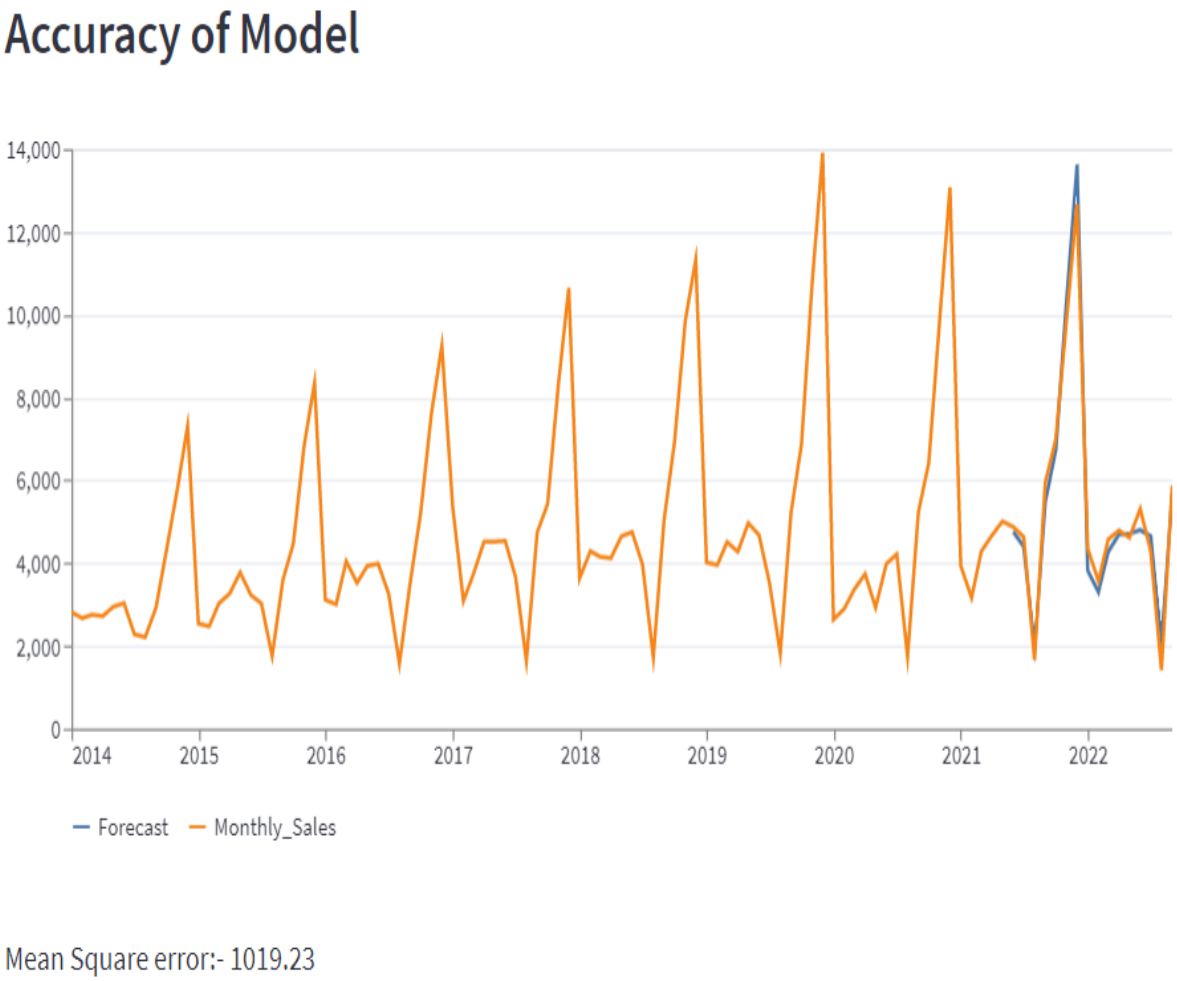

Figure 4.4 Checking accuracy of model

As we can see in the above figure 4.4 the forecasted values which is represented by blue line is almost overlapping the actual values which is in orange of color which tells us about the accuracy of the model we used and X axis represent the month and year of the sales and Y axis represent the sales of the particular product. From above figure we can conclude that model used is almost predicting accurate values. The MSE value is 1019.23 .If the value of RMSE is lower then the accuracy og the forecasted value is higher and viceversa.

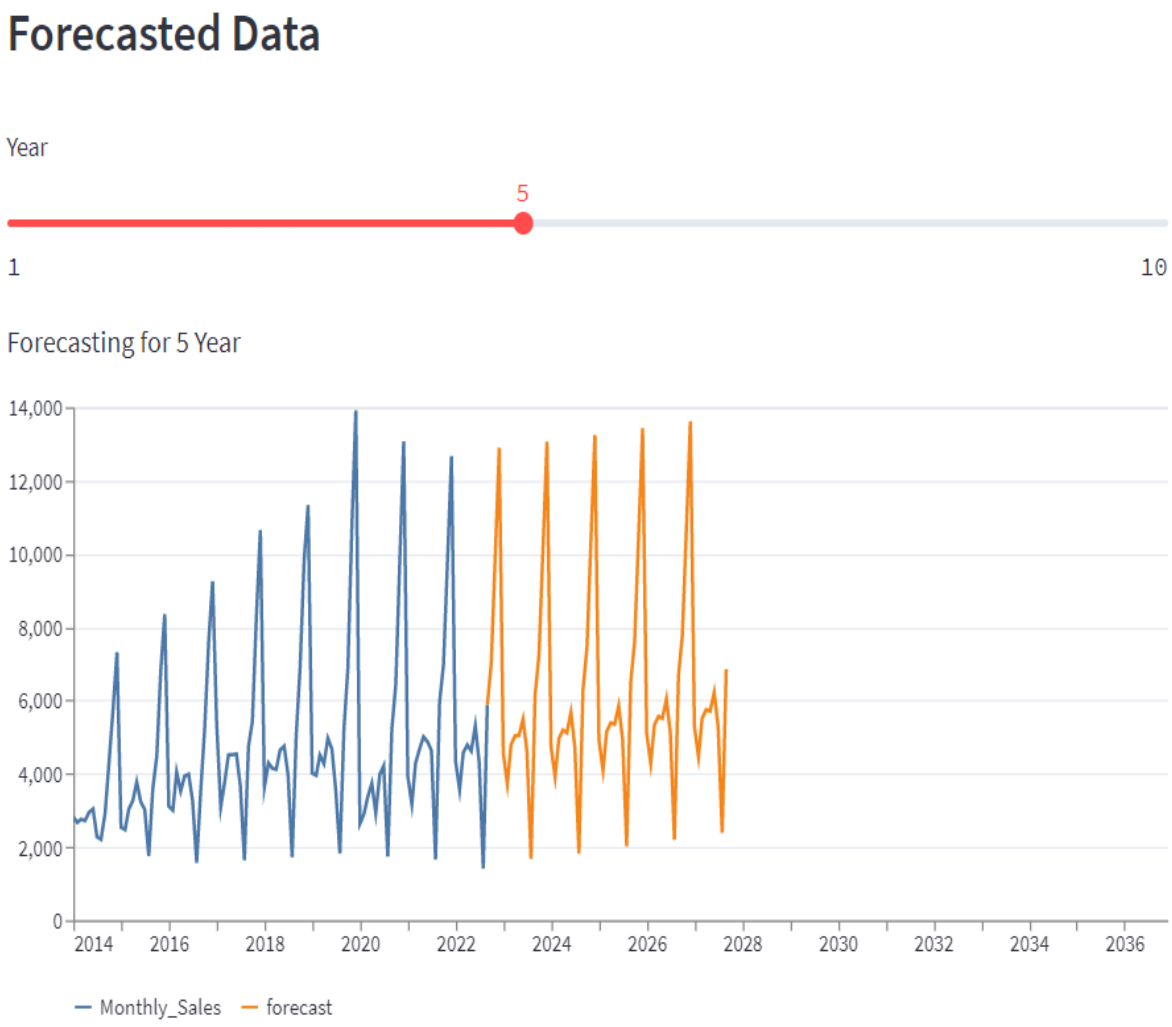

Figure 4.5 Representing Forecasted Values

In the above figure 4.5 user can forecast next 10 years sale using slider and also user can see the forecasted value by hovering over the forecasted data of the desired month. Blue line represents the data of the past monthly sales and orange line represent the forecasted data for the future months.

CHAPTER-5
CONCLUSIONS/RECOMMENDATIONS

Time Series forecasting is really useful when we have to take future decisions or we have to do analysis, we can quickly do that using ARIMA, there are lots of other Models from we can do the time series forecasting but ARIMA is really easy to understand.

• Sales forecasting provides business to increase productivity and employ more worker in future

• At the end of this project, we expect that our application is capable of giving desired output to the users and is able to help them.

• The main goal of the project is to help small business to forecast their sales.

• At the end of this project, we expect that we will be able to forecast any product sale more accurately and more efficiently.

The system can be made more flexible and scalable using these recommendations. Please note that the system implemented here is just a prototype of idea presented via this project. The recommendations are as follows:

CHAPTER-6

REFERENCES

[1] Time Series Forecasting in Python by Marco Peixeiro

[2] Data science for forecasting ,data clustering, and by Vivian Siahaan Author

[3]Forecasting: Principles and practice by Rob J Hyndman and George

[4] https://www.python.org/doc/essays/blurb/ (taken on September 2022)

[5]https://research.google.com/colaboratory/faq.html#:~:text=Colaboratory%2C%20or%20%E2%80%9CColab%E2%80%9D%20for,learning%2C%20data%20analysis%20and%20education. (taken on September 2022)

[6] https://www.jetbrains.com/help/pycharm/quick-start-guide.html(taken on September 2022)

[7] https://www.techtarget.com/searchenterprisedesktop/definition/Excel (taken on September 2022)

[8] https://www.javatpoint.com/pandas-vs-numpy (taken on October 2022)

[9] https://www.statsmodels.org/stable/index.html (taken on October 2022)

[10]https://www.latentview.com/data-engineering-lp/introduction-to-streamlit/#:~:text=Streamlit%20is%20an%20open%20source,NumPy%2C%20pandas%2C%20Matplotlib%20etc. (taken on October 2022)

[11] Time Series Forecasting by Analytic Vidhya :-

https://www.analyticsvidhya.com/blog/2021/07/time-series-forecasting-complete-tutorial-part-1/ (taken on October 2022)

[11] Arima model by Analytic Vidhya :-

https://www.analyticsvidhya.com/blog/2020/10/how-to-create-an-arima-model-for-time-series-forecasting-in-python/ (taken on October 2022)

[12] https://towardsdatascience.com/introduction-to-arima-for-time-series-forecasting-ee0bc285807a#:~:text=Auto%20Regressive%20Integrated%20Moving%20Average,to%20a%20time%20series%20data. (taken on November 2022)

[13] https://www.ibm.com/in-en/topics/linear-regression#:~:text=Resources-,What%20is%20linear%20regression%3F,is%20called%20the%20independent%20variable.

(taken on November 2022)

[14] https://spotio.com/blog/best-sales-forecasting-software/ (taken on November 2022)

[15] https://link.springer.com/referenceworkentry/10.1007/1-4020-0612-8_358#:~:text=Forecast%20error%20is%20the%20difference,meaningful%20information%20to%20the%20manager. (taken on November 2022)

[16] https://statisticsbyjim.com/regression/mean-squared-error-mse/ (taken on November 2022)

[17] https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/ (taken on November 2022)