Tribhuvan University

INSTITUTE OF ENGINEERING

Kathmandu Engineering College

Department of Computer Engineering

Minor Project Report

On

SUPERMARKET SALES ANALYSIS AND PREDICTION

[Code No: CT654]

By

Sankalpa Pokharel - KAT074BCT064

Shameen Shrestha - KAT074BCT067

Shreya Basnet - KAT074BCT071

Subash Shrestha - KAT074BCT074

Kathmandu, Nepal

Falgun 2077

TRIBHUVAN UNIVERSITY
INSTITUTE OF ENGINEERING

Kathmandu Engineering College

Department of Computer Engineering

CERTIFICATE

The undersigned certify that they have read and recommended to the Department of Computer Engineering, a minor project work entitled "SUPERMARKET SALES ANALYSIS AND PREDICTION" submitted by Sankalpa Pokharel – 74064, Shameen Shrestha – 74067, Shreya Basnet – 74071, Subash Shrestha – 74074 in partial fulfillment of the requirements for the degree of Bachelor of Engineering.

Er. Amit Khanal
(External Examiner)
Department of Computer Engineering
Institute of Engineering (IOE)

Er. Bishon Lamichhane
(Project Coordinator)
Computer Engineering.
Kathmandu Engineering College

Er. Sudeep Shakya
(Head of Department)
Computer Engineering.
Kathmandu Engineering College

Abstract

Today's business handles huge repository of data. Most of the Supermarkets heavily depend on a knowledge base and demand prediction of sales trends. However, research has shown that companies perform better when they apply data-driven decision-making. The volume of data is expected to grow further in an exponential manner. The main idea of this project is to introduce intelligent, data-based decision models, which are comprehensive and support the interactive evaluation of decision options necessary for sales prediction in Supermarkets. In this project, we applied regression methods in machine learning and time series analysis techniques to forecast the sales amount based on several features. Time series forecasting is one of the major building blocks of Machine Learning. In this project we have used (ARIMA) for time series forecast. Data visualization is achieved in the form of pie charts, bar graphs and line graphs using a python library Plotly.

Keywords— *Machine Learning Algorithms, Regression,time series analysis, Sales forecasting*

ACKNOWLEDGEMENT

We express our sincere gratitude to Head of Department**, Er. Sudeep Shakya** and Deputy Head of Department**, Er. Kunjan Amatya** for helping us in the report. We would like to thank all the teachers of the **Department of Computer Engineering** for providing us the required guidance that we need for this project.We are grateful to **Er. Bishon Lamichhane** for the necessary templates and docs which has helped us a lot during the completion of our project.

We are grateful to people who helped us directly or indirectly to finalize this project. We must acknowledge our deep sense of gratitude to the people for their constant motivation and suggestions.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

- ARIMA - Autoregressive Integrated Moving Average
- AIC - Akaike's Information Criteria
- CSS - Cascading Style Sheets
- d - Number of nonseasonal differences needed for stationarity
- DB - Database
- DFD - Data Flow Diagram
- HTML - Hyper Text Mark-up Language
- KDD - Knowledge Discovery Database
- MA - Moving Average
- p - Number of autoregressive terms in ARIMA
- q - Number of lagged forecast errors in the prediction equation
- P,D,Q - Constant values for p, d and q used in seasonal ARIMA
- SQL -Structured Query Language
- UML - Unified Modelling Language

CHAPTER 1: INTRODUCTION

1.1 BACKGROUND THEORY

The industry is in need of data mining techniques and intelligent prediction models of sales trends with the highest possible level of accuracy and reliability. Sales Analysis and forecasting has become an essential need in industry as it enables us to make informed business decisions and predict short-term and long-term performance. Sales forecasting gives insight into how a company should manage its workforce, cash flow, and resources. One of the solutions is to use predictive analytics and rely on machine learning algorithms which will help in predicting future sales based on years of past business data.

Predictive analytics encompasses a variety of statistical techniques from data mining, predictive modelling, and machine learning, that analyse current and historical facts to make predictions about future or otherwise unknown events. Predictive analytics is used in actuarial science, marketing, financial services, insurance, travel, mobility, healthcare, child protection, pharmaceuticals, capacity planning, social networking and other fields.

Predictive Analysis Process:

1. **Define project**: Define the project outcomes, deliverable, scope of the effort, business objectives, identify the data sets that are going to be used.
2. **Data collection**: Data mining for predictive analytics prepares data from multiple sources for analysis. This provides a complete view of customer interactions.
3. **Data analysis**: Data Analysis is the process of inspecting, cleaning and modelling data with the objective of discovering useful information, arriving at conclusion
4. **Statistics**: Statistical Analysis enables to validate the assumptions, hypothesis and test them using standard statistical models.
5. **Modelling**: Predictive modelling provides the ability to automatically create accurate predictive models about the future. There are also options to choose the best solution with multi-modal evaluation.

6. **Deployment**: Predictive model deployment provides the option to deploy the analytical results into everyday decision-making processes to get results, reports and output by automating decisions based on the modelling.
7. **Model monitoring**: Models are managed and monitored to review the model performance to ensure that it is providing the results expected.

In the context of supermarkets, various machine learning processes and predictive analysis can be used for customer centric sales analysis and prediction, product centric sales analysis and prediction and location centric sales analysis and prediction.

Our project solely focuses on product centric sales analysis and prediction, we analyse the sale of various categories of products in general, compare with the previous year's sale and predict the future sales of the product category in general and also the items within the category. For prediction of sale of a product, data mining is particularly called into play since it involves mining of certain patterns which can be done by following the steps of KDD process. The steps are listed as follows:

1. Selection of Data set
2. Pre-processing
3. Transformation
4. Data Mining
5. Interpretation

1.2 PROBLEM STATEMENT:

A complexity of sales dynamics often forces decision makers to make decisions based on subjective mental models, reflecting their experiences. However, research has shown that companies perform better when they apply data driven decision making. So, we have developed a custom data analytics model to help the supermarket generate necessary insights about various aspects of sales, so that they can make the right decision and achieve growth.

1.3 OBJECTIVES:

- Overview the trend of total sales of all stores and items over time.
- Choose and reconstruct features that have impacts on Supermarket Sales.
- Take a future decision in terms of inventory management, marketing activities, schemes based on sales analysis.

1.4 SCOPE OF THE PROJECT:

This project is applicable to every supermarket as it will help the supermarket to manage its resources and help to make a future decision in terms of inventory management, marketing activities. Schemes or offers to be rolled and changes in manufacturing processes of the products if applicable. Sales analysis will also show the current market trends to the company and based on sales data.

1.5 APPLICATIONS:

1. High-Quality Lead Generation: With predictive analytics, marketers can gauge the customer's propensity to buy with greater accuracy.
2. Targeted Profiling of Customers: It gives marketers a greater understanding of how customers responded to a marketing activity, the reasons behind why they did or did not make a purchase and helps them identify how to convert a prospect into a paying customer
3. Improved Content Distribution: Predictive analytics tackles that problem head-on by analysing the types of content that most resonate with customers of certain demographic or behavioural backgrounds, and then automatically distributing similar content to leads that mirror the same demographic or behavioural habits.
4. Improved Determination of Product Fit: Equipped with historical, sales, and leads data, businesses can better understand exactly what customers' needs and wants are, which is key to developing better future products.

CHAPTER 2: LITERATURE REVIEW

2.1 How data science adds value to business

Big data science and analytics have changed the course of market strategies and paved altogether new paths for the growth and profit of the companies. We have entered the digital age in this decade and big data analysis is the latest digital technology that has accomplished even unbelievable tasks in real-time. By the end of 2020, the big data volume is going to reach 44 trillion gigabyte, breaking down all the previous trends and setting a new business world.

Data science is the most current “tool” for businesses that want to meet consumer demand. The beauty of it is that it is not based on what we may “feel” about consumers and what they want or need. It is based upon actual patterns of behaviours and trends that the facts reveal.

A high competition exists in the Fast Moving Consumer Goods (FMCG) market to increase the profits. Accurate sales forecasting is an inexpensive method to reduce lost sales, product returns and support efficient product planning. Moreover, accurate forecasts of retail sales may improve portfolio investors’ ability to predict movements in the stock prices of retailing chains. Aggregate retail sales time series are usually preferred because they contain both trend and seasonal patterns, providing a good testing ground for comparing forecasting methods, and because companies can benefit from more accurate forecasts. Retail sales time series often exhibit strong trend and seasonal variations presenting challenges in developing effective forecasting models. . Exponential smoothing and Autoregressive Integrated Moving Average (ARIMA) models are the two most widely used approaches to time series forecasting, and provide complementary approaches to the problem. While exponential smoothing methods are based on a description of trend and seasonality in the data, ARIMA models aim to describe the autocorrelations in the data. The ARIMA framework to forecasting originally developed by Box et al. involves an iterative three-stage process of model selection, parameter estimation and model checking

2.2. RELATED WORKS:

Scholars from various universities have done great work in the field. several system and application are based on sales analysis and prediction:

1.SalesPRISM:

A customer pattern recognition tool from Lattice, SalesPRISM helps brands to collect data about sales and predict potential sales leads. Every brand has a lot of data about customers and by using factors like CRM data, site traffic, and sales history, SalesPRISM also analyses external data like LinkedIn activity and LexisNexis reports.

2. Medio Platform:

With big platforms like Amazon and Flipkart, understanding the needs and habits of the customers will go a long way in helping a brand to create a comprehensive e-commerce strategy. Media Platform helps brands to analyse the problems related to customers leaving a website and to ensure corrective action.

3. TIBCO Software:

Understanding customer behaviour has always been an important aspect for the success of any brand/ organization. The predictive analytics tool in TIBCO software can effectively help brands to understand data in a much better manner, thereby enabling them to make smarter business decisions.

4. Lattice

Another great predictive analytics tool, Lattice provides immense insights into sales so as to help brands to market their products in a much better manner. By using the predictive analytics tool provided by Lattice, brands can easily find effective tools to convert their leads into sales.

2.3. RESOLUTION:

Forecasting of behavioural time series has benefitted many businesses with the aim of predicting future trends by understanding the past.

The feature of our project is that the user can analyse and compare the sales of products at the same period of time this year and of that previous year. Further, we also have year based predictive analysis in which the user can analyse and compare the sales based on the time period. Another feature in our project is a user-friendly interface. The user will not have any problems navigating the app since all the data visualizations are shown on the home page.

CHAPTER 3: METHODOLOGY

3.1. PROCESS MODEL

Incremental Model is a process of software development where requirements are broken down into multiple standalone modules of the software development cycle. Incremental development is done in steps from analysis design, implementation, testing/verification, maintenance.

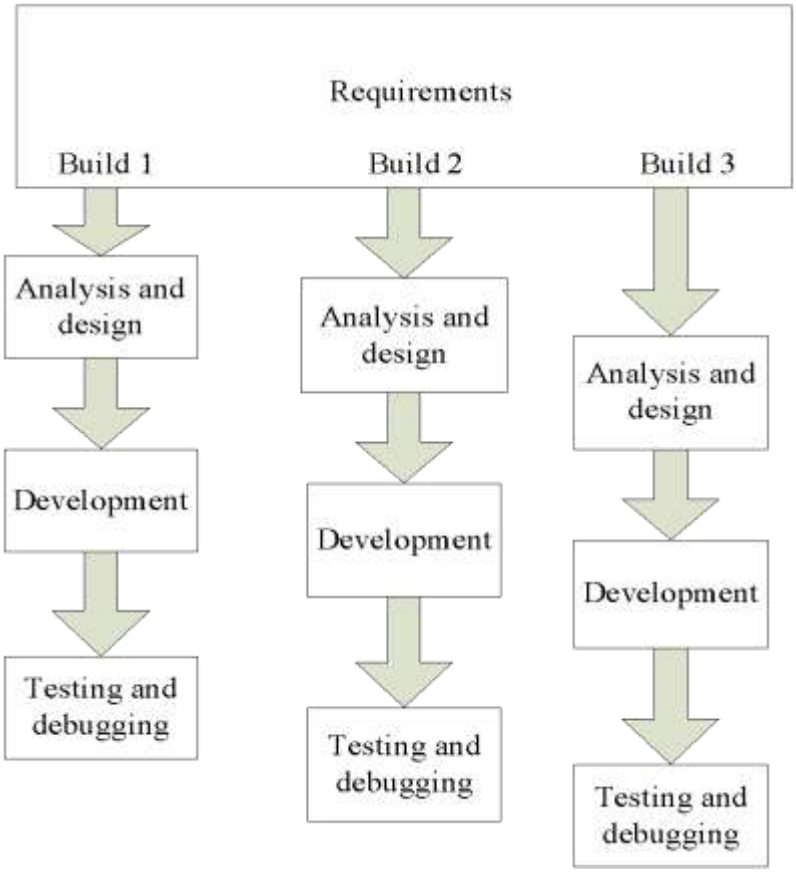


Fig 3.1.1: Representing Increment Model

Each iteration passes through the requirements, design, coding and testing phases. And each subsequent release of the system adds function to the previous release until all designed functionality has been implemented.

The system is put into production when the first increment is delivered. The first increment is often a core product where the basic requirements are addressed, and supplementary features are added in the next increments. Once the core product is analysed by the client, there is plan development for the next increment.

Advantage of Increment model:

- Requirements of the system are clearly understood.
- Project needs to be completed in a shorter time.
- This methodology is more in use for web application and product-based projects.

RELATED THEORY :

Autoregressive integrated moving average (ARIMA) model is used in time series data either to better understand the data or to predict future points in the series (forecasting). The term ARIMA can be decomposed into three parts : AR(Auto Regressive), MA (Moving Average) and I (Integrated).

Non-seasonal ARIMA models are generally denoted ARIMA(p,d,q) where parameters p, d, and q are non-negative integers, p is the order (number of time lags) of the autoregressive model, d is the degree of differencing (the number of times the data have had past values subtracted), and q is the order of the moving-average model.

Seasonal ARIMA models are usually denoted ARIMA(p,d,q)(P,D,Q)m, where m refers to the number of periods in each season, and the uppercase P,D,Q refer to the autoregressive, differencing, and moving average terms for the seasonal part of the ARIMA model.

Auto Regressive :

The AR part of ARIMA indicates that the evolving variable of interest is regressed on its own lagged (prior) values.

An autoregressive model of order can be written as

$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t,$

where ε_t is white noise. This is like a multiple regression but with lagged values of y_t as predictors. We refer to this as an autoregressive model of order p, where p is the number of lags to be used as predictors. Autoregressive models are remarkably flexible at handling a wide range of different time series patterns.

Moving Average:

The MA part indicates that the regression error is actually a linear combination of error terms whose values occurred contemporaneously and at various times in the past. A moving average model uses past forecast errors in a regression-like model.

$y_t = c + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$, ,where ε_t is white noise.

We refer to this as an MA(q) model, a moving average model of order q, where q is the number of lagged forecast errors . In this , we do not *observe* the values of ε_t, so it is not really a regression in the usual sense.

Integrated:

The I (for "integrated") indicates that the data values have been replaced with the difference between their values and the previous . d is the degree of differencing (the number of times the data have had past values subtracted),

Information Criteria

Akaike's Information Criterion (AIC), which was useful in selecting predictors for regression, is also useful for determining the order of an ARIMA model. It can be written as

$AIC = -2\log(L) + 2(p+q+k+1)$,

where L is the likelihood of the data, $k=1$ if $c \neq 0$ and $k=0$ if $c=0$.

For ARIMA models, the corrected AIC can be written as

$AIC = AIC + 2(p+q+k+1)(p+q+k+2)/(T-p-q-k-2)$

3.3. BLOCK DIAGRAM

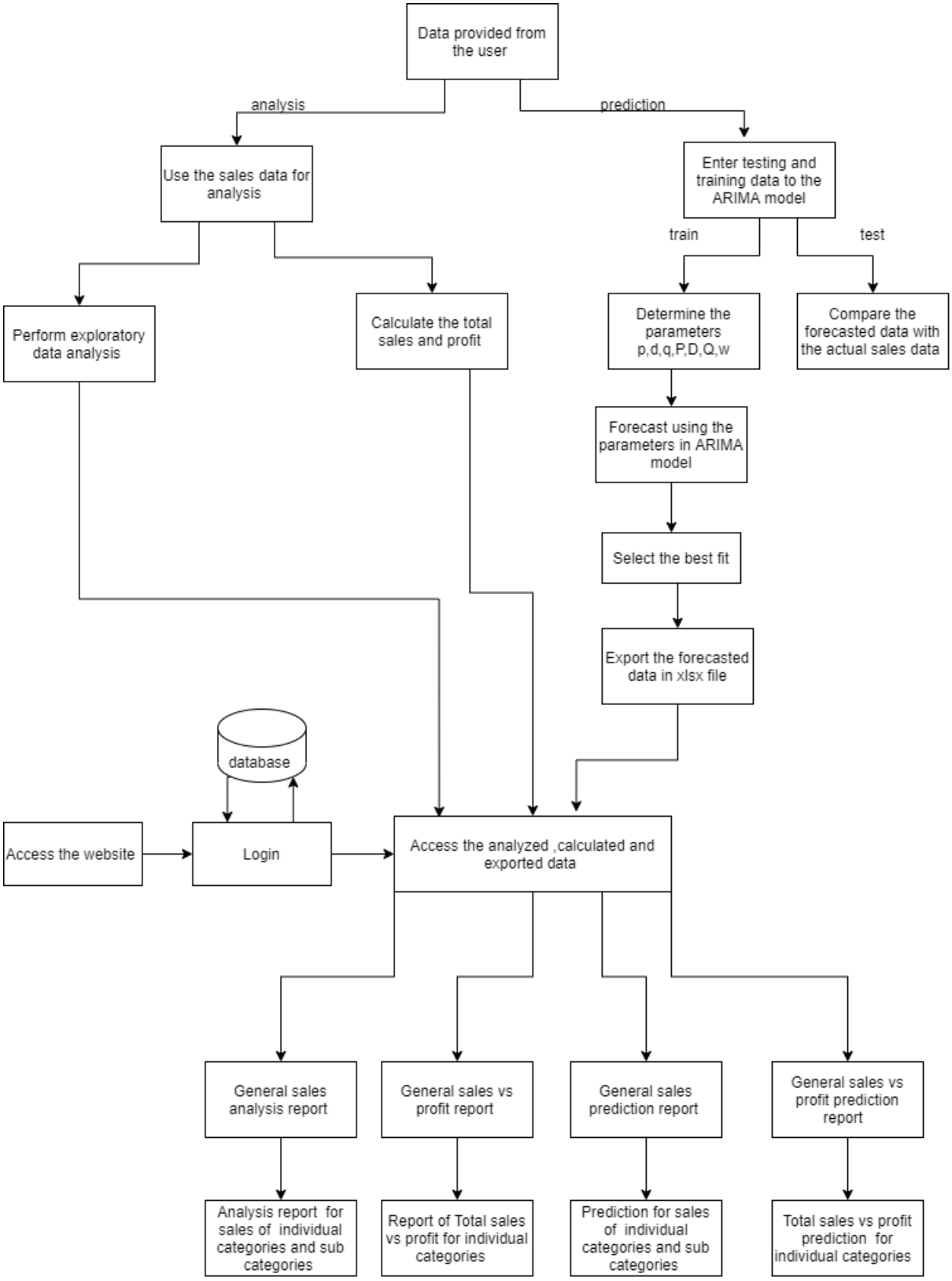


Fig 3.3.1: Block Diagram

The sales data provided from the user is used for the analysis of sales and prediction of sales for the upcoming 4 years . For the analysis we use the current sales data and perform exploratory data analysis for graphical representation of sales data. The total sales and profit of sales is calculated so that the user can have an overview of sales and profit.

For the prediction part the training and testing data is fed into the ARIMA model. From the training data the parameters p,d,q,P,D,Q,m are determined where p is the order (number of time lags) of the autoregressive model, d is the degree of differencing (the number of times the data have had past values subtracted), and q is the order of the moving-average model and P,D,Q are the autoregressive, differencing, and moving average terms for the seasonal part of the ARIMA model respectively and m refers to the number of periods in each season.

Forecasting is done by the ARIMA model and the best fit is selected by AIC(Akaike's Information Criterion). The forecasted data is exported to an xlsx file.

When the user accesses the website , the user has to insert their valid login credentials and on the dashboard the user has options on the tab to view the analysis or prediction.

If the user selects analysis, they can view the analysed number of sales of each category and subcategories in pie chart, year wise sales comparison of each category in a bar graph with slider. The user can also view the graphical representation of total sales vs profit in general and for each individual category in the form of line graphs.

If the user selects predictions, the user can view predicted sales in general and also for each category in the form of a line graph. They can view the general predicted total sales vs profit in form of line graph and date wise prediction table. To view the predicted sales for each category the user can select the category and view the sales of the selected category in the form of a line graph.

Users can also get an overview of overall estimated growth , expected growth of each category and prediction of the category that will have the highest sales in upcoming years.

3.4. ALGORITHMS

3.4.1 ALGORITHM OF THE APPLICATION

Step 1 : Start

Step 2: Provide Login Details i.e User name and password

Step 3: If Login details are correct

Go to step 4

Else

Display “Invalid Username and Password”

Step 4: If sales data is fed

Go to step 5

Else

Go to step 11

Step 5: Show options: Sales Analysis and Sales Prediction and logout

Step 6: If Sales Analysis is Selected

Goto step 7

Else If Sales Prediction is Selected

Goto Step 8

Step 7: Exploratory Data Analysis is done

Step 7.1 Show sales of categories and subcategories in pie chart

Step 7.2 : Show the sales comparison of all category according to year in bar graph

Step 7.3 : Show Interactive Total Sales vs Total Profit

Step 7.4 : Show individual sales growth of each category

Step 8: Use the forecasted data obtained from ARIMA model

Step 8.1. Show total sale vs profit prediction for upcoming 3 years

Step 8.2.Show the date wise sales prediction in a table

Step 8.3 :Show sales prediction of each category as selected by the user

Step 9: If user logs out

Goto step 10.

Step 10 : Display “User disconnected - Please login to view the success screen again”

Step 10.1 :If Go back is Selected

Goto Step 2

Else

Goto Step 11

Step 11: Stop

3.4.2 ALGORITHM OF ARIMA MODEL

Step 1: Start

Step 2: Enter the training and forecasting data to the ARIMA model

Step 3: If the model is seasonal

Go to step 4

Else

Go to step 5

Step 4: Determine the parameters p(order of auto regressive term),d(order of differentiation),q(order of moving average), P(order of auto regressive term for seasonal), D(order of differentiation for seasonal),Q(order of moving average for seasonal), w(number of periods in season)

Go to step 6

Step 5: Determine the parameters p(order of auto regressive term), d(order of differentiation), q(order of moving average)

Step 6: Determine the presence or absence of the constant term in the model

Step 7: Perform sales Forecasting using the ARIMA model

Step 8: Select the best suited structure for representation using AIC (Akaike's Information Criterion) value

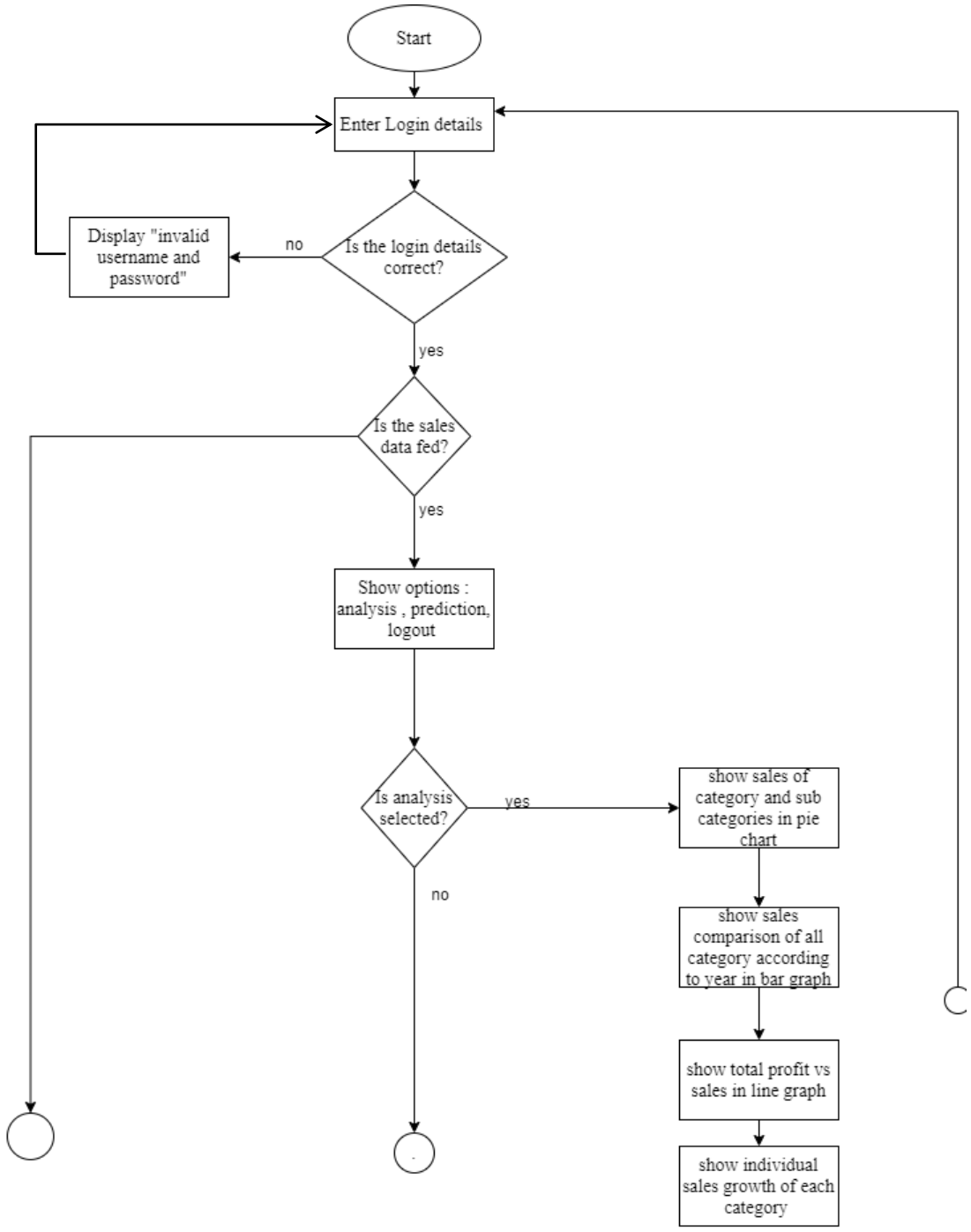
Step 9: Perform testing

Step 9.1:Compare the obtained forecasted data with the actual sales data

Step 10: End

3.5. FLOWCHARTS

3.5.1 FLOWCHART OF THE APPLICATION



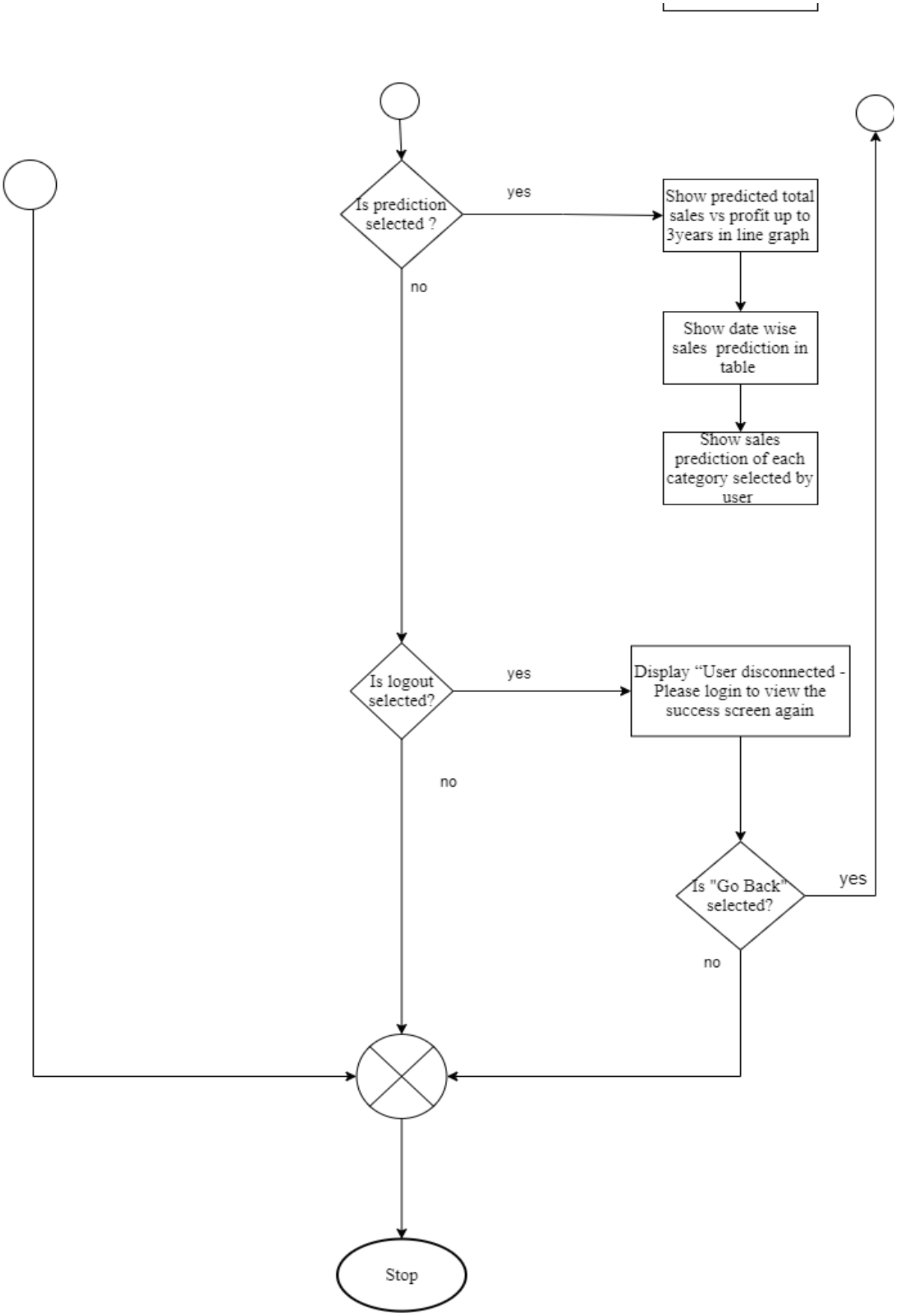


Fig 3.5.1. Flowchart

3.5.2. FLOWCHART OF ARIMA

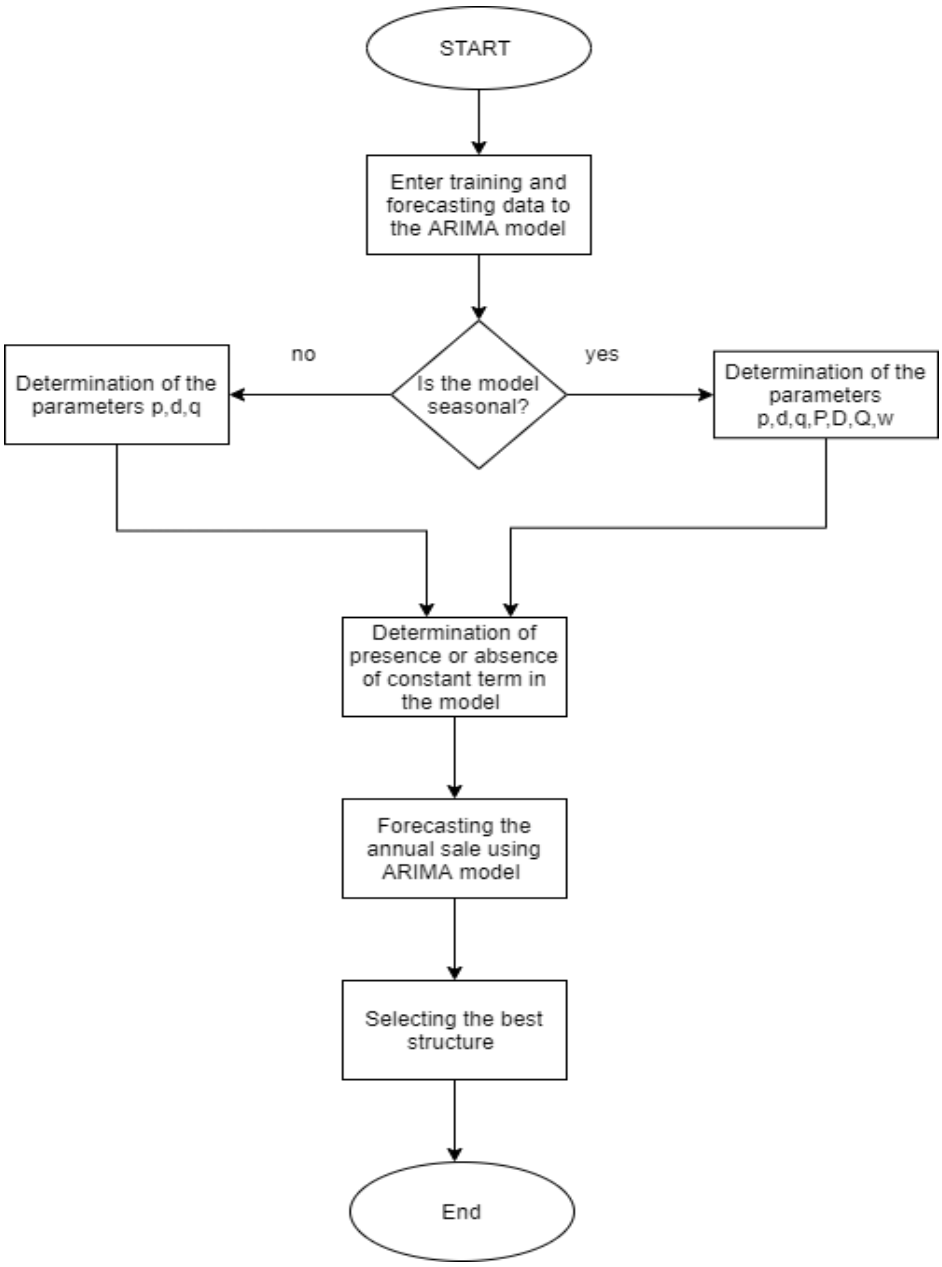


fig 3.5.2. Flowchart of ARIMA

3.6 DATA FLOW DIAGRAM

Fig 3.5.1: DFD Level 0

The DFD level 0 diagram above also represents the black box diagram of the system. The user has to provide the sales data for a significant period of time (in this case- 4 years of data) to the system. The admin reads the data from the system, analyses the data and represents it in various form into the system. The user also uses the data to predict the future sales of the supermarket with the help of ARIMA model. Then the system represents the data model prepared after machine learning algorithm is applied to it and represents it in the form of dashboard to the user.

3.7. UML USE CASE DIAGRAM

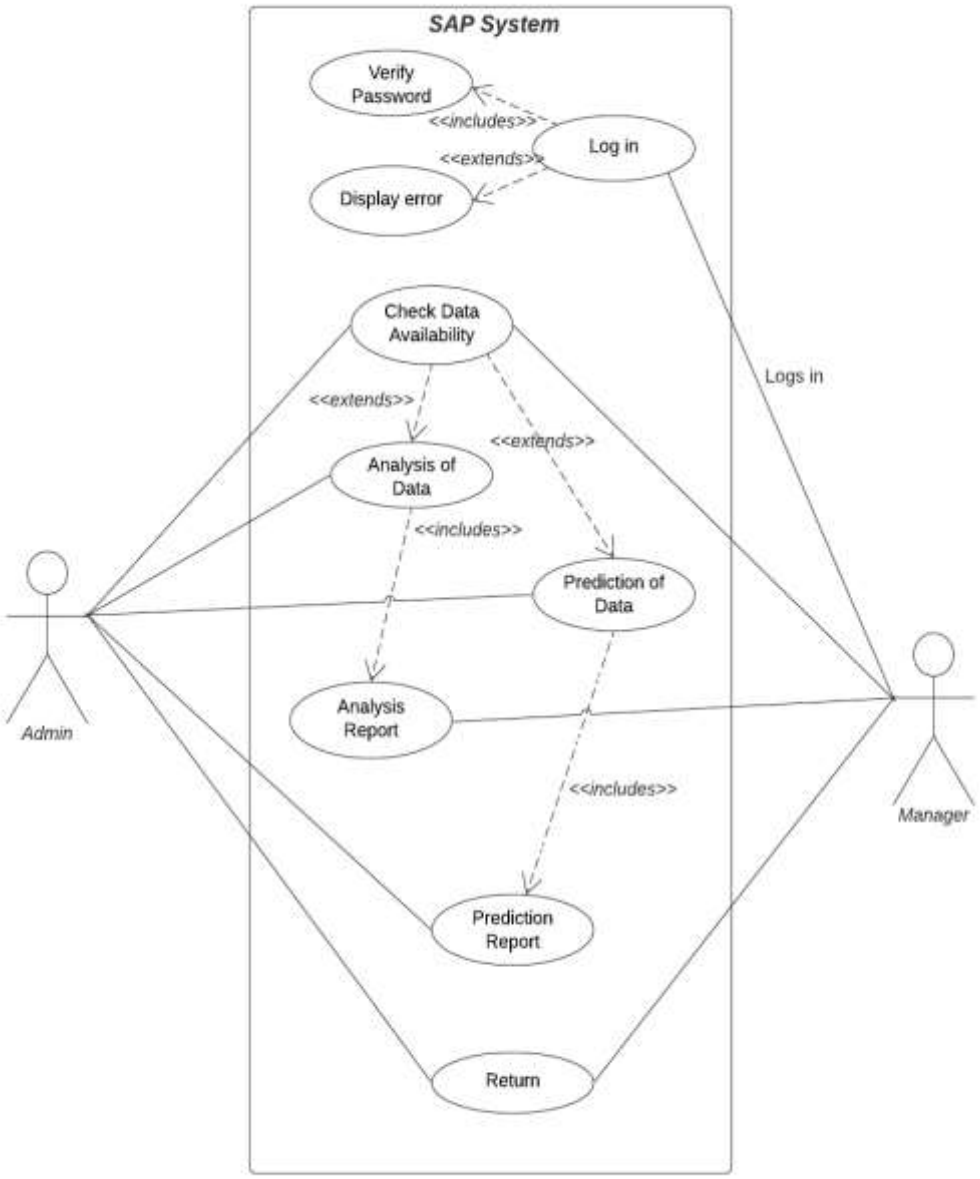


Fig 3.6: Use Case Diagram

3.8. SEQUENCE DIAGRAM FOR LOGIN

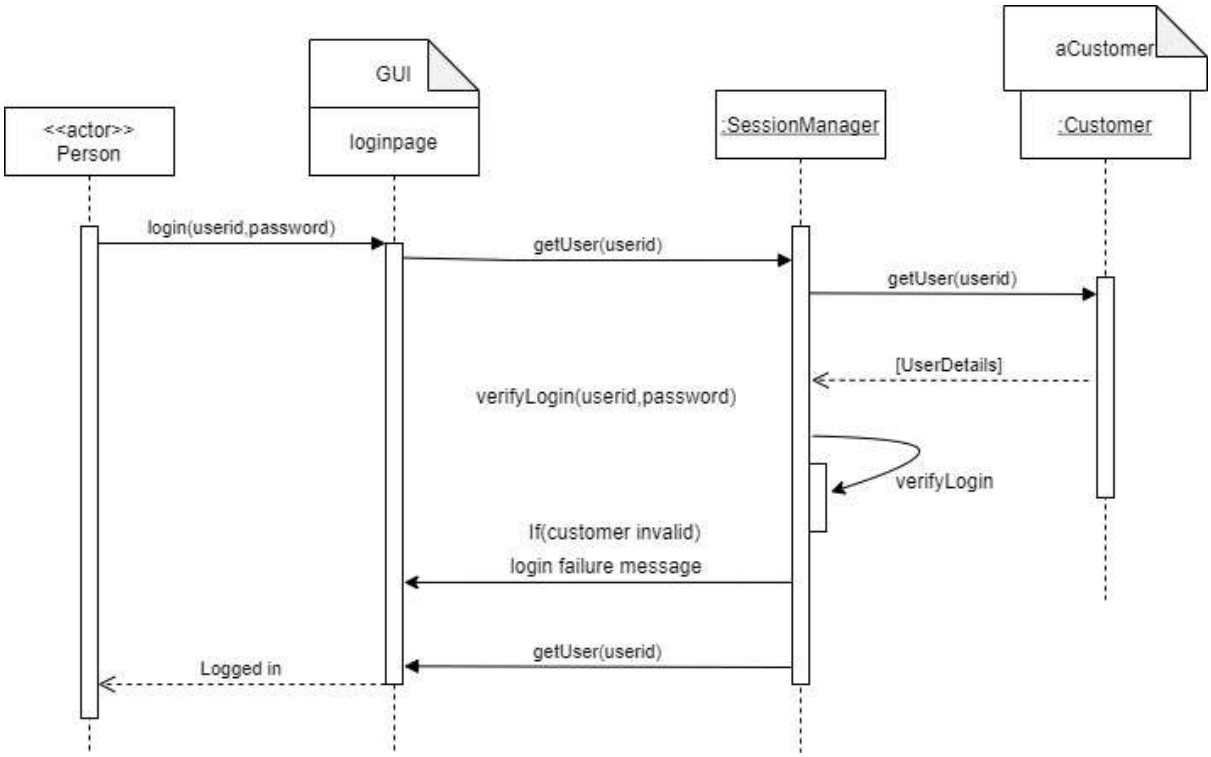


Fig 3.8. : Sequence Diagram of Login Page

The following diagram represents the sequence in which the login process is completed in the system. As the user accesses the login page, user is asked to enter their login credentials (userid and password). These credentials are passed to a function SessionManager which checks the database if the user has inserted the valid userid and password. The results from the database is also verified in this function. If the correct userid and password have been given, then the user is logged in. But if the userid and password for that user are invalid, then the user is asked to re-enter those credentials.

3.9. TOOLS USED

1. Python:

Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. It 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.

2. Dash

Dash is a user interface library for creating analytical web applications. Those who use Python for data analysis, data exploration, visualization, modelling, instrument control, and reporting will find immediate use for Dash.

3. Plotly

Plotly's Python graphing library makes interactive, publication-quality graphs. Examples of how to make line plots, scatter plots, area charts, bar charts, error bars, box plots, histograms, heatmaps, subplots, multiple-axes, polar charts, and bubble charts.

4.Pandas

Pandas is a high-level data manipulation tool developed by Wes McKinney. It is built on the Numpy package and its key data structure is called the Data Frame. Data Frames allow you to store and manipulate tabular data in rows of observations and columns of variables.

5.CSS:

CSS(Cascading Style Sheet) is a style sheet language used to format the layout of Web pages. Then can be used to define text style, table sizes, and other aspects of Web pages that previously could only be defined in a page’s HTML.

6.Bootstrap:

Bootstrap is a free and open-source CSS framework directed at responsive, mobile first front-end web development. It contains CSS and JS based design templates for forms, buttons, navigation, and other interface components.

7. ARIMA:

ARIMA is an acronym that stands for AutoRegressive Integrated Moving Average. This is one of the easiest and most effective machine learning algorithms to perform time series forecasting. It uses time series data to either better understand the data set or to predict future trends.

8.Flask

Flask is a micro web framework written in Python. It is classified as a microframework because it does not require particular tools or libraries. Flask supports extensions that can add application features as if they were implemented in Flask itself. Extensions exist for object-relational mappers, form validation, upload handling, various open authentication technologies and several common framework related tools. Extensions are updated far more frequently than the core Flask program.

9.SQLite

SQLite is a relational database management system (RDBMS) contained in a C library. In contrast to many other database management systems, SQLite is not a client–server database engine. Rather, it is embedded into the end program.

3.10. VERIFICATION AND VALIDATION

We 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 with the help of tsa.seasonal_decompose(). This process helps us determine whether to use the ARIMA model or not and to find if the trend is seasonal or not.

Fig 3.10.1 : Decomposition of Furniture Sales

Fig 3.10.2 : Decomposition of Technology Sales

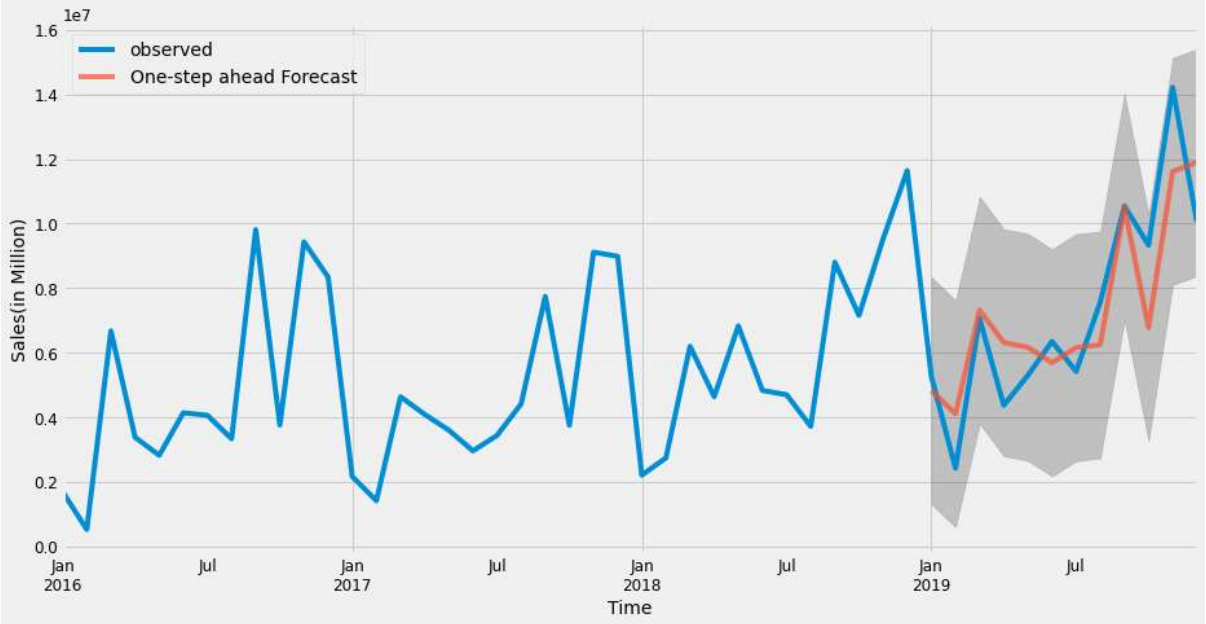
Fig 3.10.3: Decomposition of Office Supplies Sales

Fig 3.1o.4: Decomposition of Total Sales

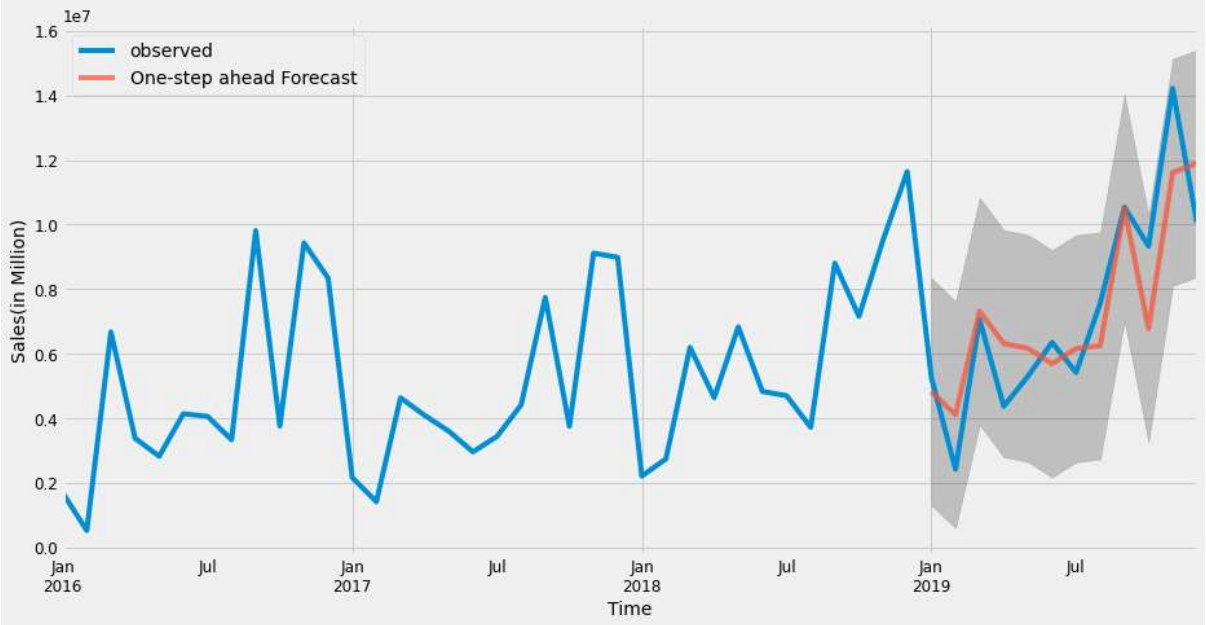
The plots above clearly show that the sales are unstable, along with its obvious seasonality. So we use Seasonal ARIMA model.

Initially we have split our data into train and test sets where train dataset is used to train our model whereas test is used to test it. Now to help us understand the accuracy of our forecasts, we compare the predicted value for the timeframe for the test dataset and compare it to real data of the test set.

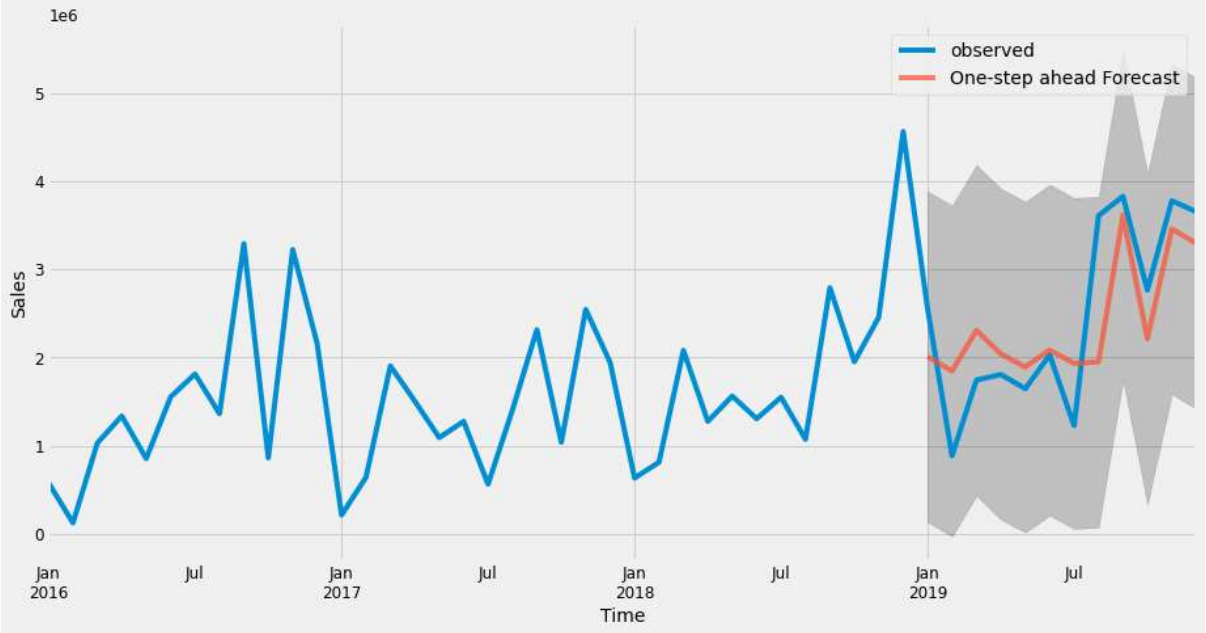
1. Furniture



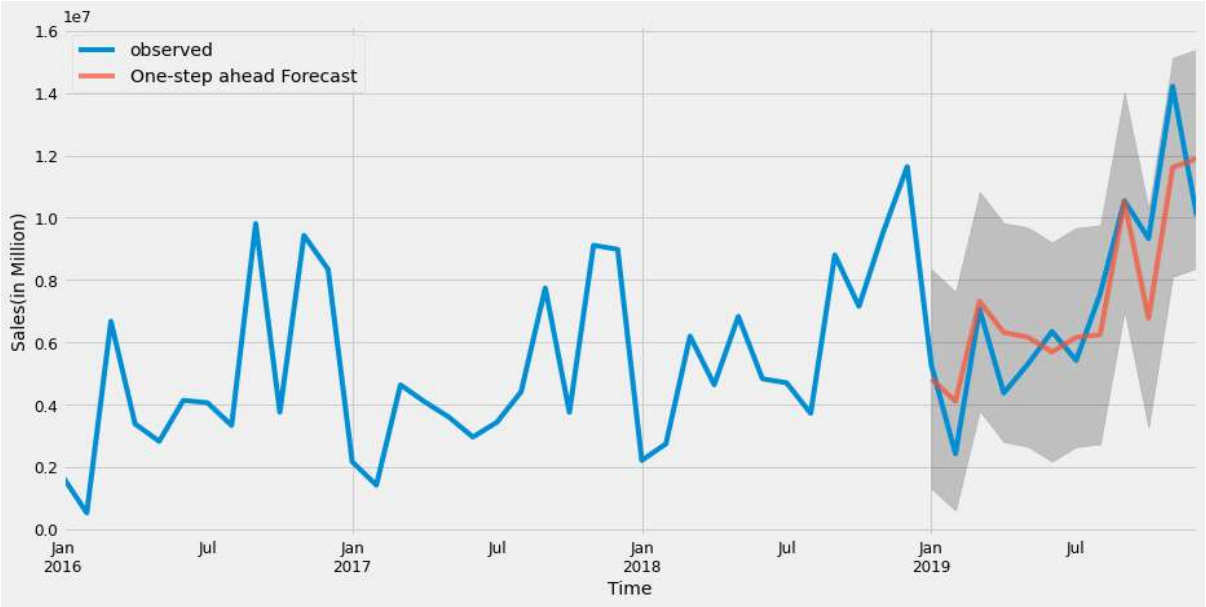
2.Technology



3. Office Supplies



4. Total sales



The line plot is showing the observed values compared to the rolling forecast predictions for all the different categories. Overall, our forecasts align with the true values very well, showing an upward trend starting from the beginning of the year and capturing the seasonality toward the end of the year.

VALIDATION

The table shows the summary of validation of the different parameters for Furniture sales, technology sales, office Supplies and Technology Sales respectively.

1.Furniture

Parameters	Coefficient	Std.Error	Z	p>|z|	0.025	0.975
ar.L1	-0.1259	0.669	-0.188	0.851	-1.437	1.185
ma.L1	-0.7952	0.272	-2.924	0.003	-1.328	-0.262
ar.S.L12	-0.7283	0.541	-1.346	0.178	1.788	0.332
ma.S.L12	0.3818	0.909	0.420	0.675	-1.401	2.164

Table 3.9.1. Validation of furniture forecast

2. Technology

Parameters	Coefficient	Std.Error	Z	p>|z|	0.025	0.975
ar.L1	-0.1342	0.497	-0.270	0.787	-1.108	0.840
ma.L1	-0.9547	0.086	-11.114	0.000	-1.123	-0.786
ar.S.L12	-0.3167	0.687	-0.461	0.645	-1.663	1.030
ma.S.L12	-0.4914	0.567	-0.867	0.386	-1.602	0.619

Table 3.9.2. Validation of technology forecast

3.Office Supplies:

Parameters	Coefficient	Std.Error	Z	p>|z|	0.025	0.975
ar.L1	0.2606	0.531	0.491	0.623	-0.779	1.301
ma.L1	-0.9896	0.070	-14.067	0.000	-1.127	-0.852
ar.S.L12	-0.3360	0.080	-0.335	0.737	-2.300	1.628
ma.S.L12	-0.4935	0.642	-0.769	0.442	-1.751	0.764

Table 3.9.3. Validation of office supplies forecast

4.For total sales:

Parameters	Coefficient	Std.Error	Z	p>|z|	0.025	0.975
ar.L1	0.1371	0.339	0.404	0.686	-0.528	0.802
ma.L1	-0.9906	0.023	-42.757	0.000	-1.036	-0.945
ar.S.L12	-0.5282	0.483	-1.094	0.274	-1.474	0.418
ma.S.L12	-0.2644	0.295	-0.897	0.370	-0.842	0.313

Table 3.9.4. Validation of total sales forecast

1. ar.L1 refers to the autoregressive term with the lag of 1
2. ma.L2 refers to the moving average term with the lag of 12
3. ar.S.L1 and ma.S.L2 refer to the seasonal ‘autoregressive’ and ‘moving average’ terms respectively with a lag of 12. All of these coefficients are part of the ARIMA equation.
4. The ‘std err’ columns is an estimate of the error of the predicted value. It tells you how strong is the effect of the residual error on your estimated parameters (the first column).As we can see the value of standard error in all this is mostly around 0.5 which is good for our model.
5. The ‘z’ is equal to the values of ‘coef’ divided by ‘std err’. It is thus the standardised coefficient.
6. The P>|z| column is the p-value of the coefficient. It is really important to check these p-values before you continue using the model. The lower the p value better the parameter.
7. [0.025 and 0.975] are both measurements of values of our coefficients within 95% of our data, or within two standard deviations. Outside of these values can generally be considered outliers.

CHAPTER 4: EPILOGUE

4.1. RESULT

After training the model with four years of past Sales record we were able to forecast the record of the following 3 years i.e. until 2022. This result can be shown in the graph below.

For Furniture:

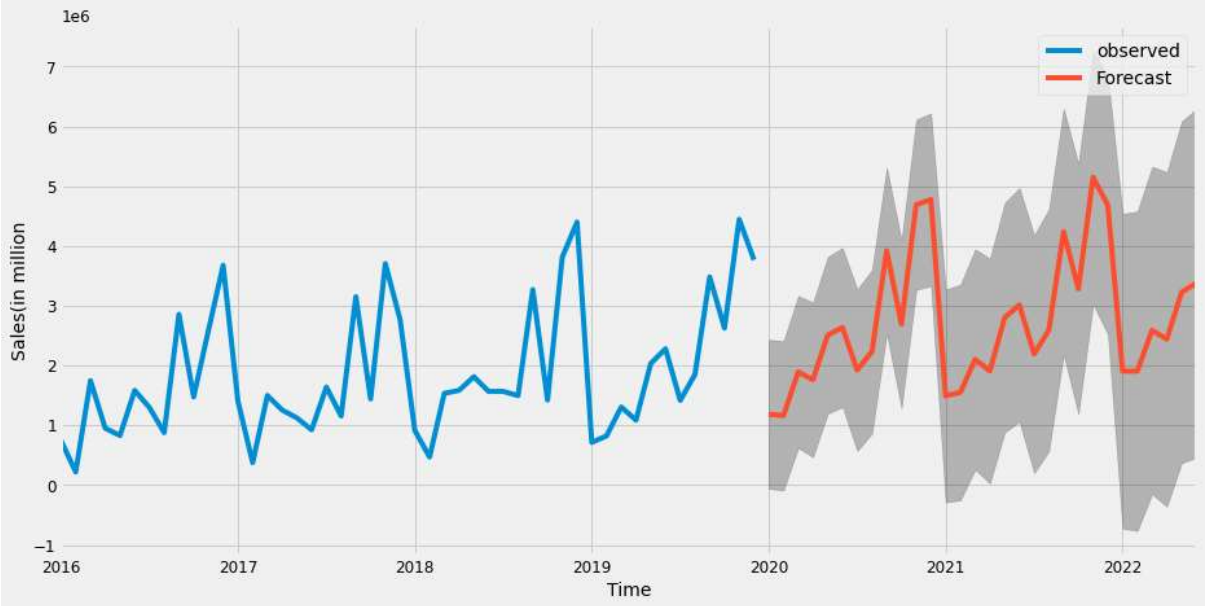


Fig 4.1.1 : Result of Furniture Sales

For Technology:

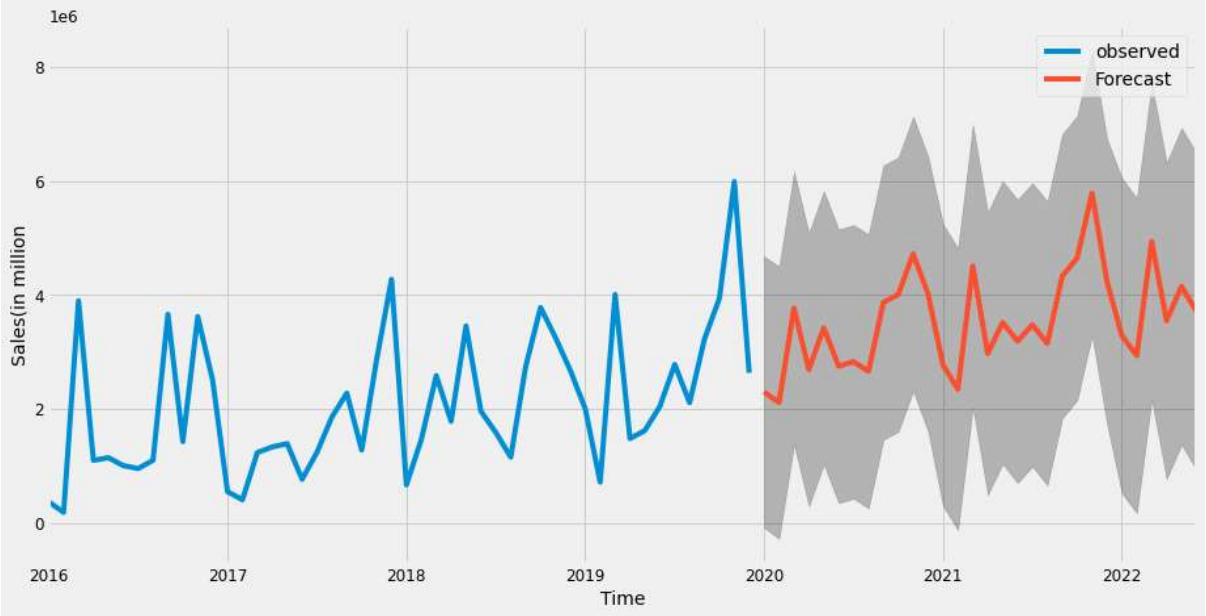


Fig 4.1.2 : Result of Technology Sales

For Office Supplies:

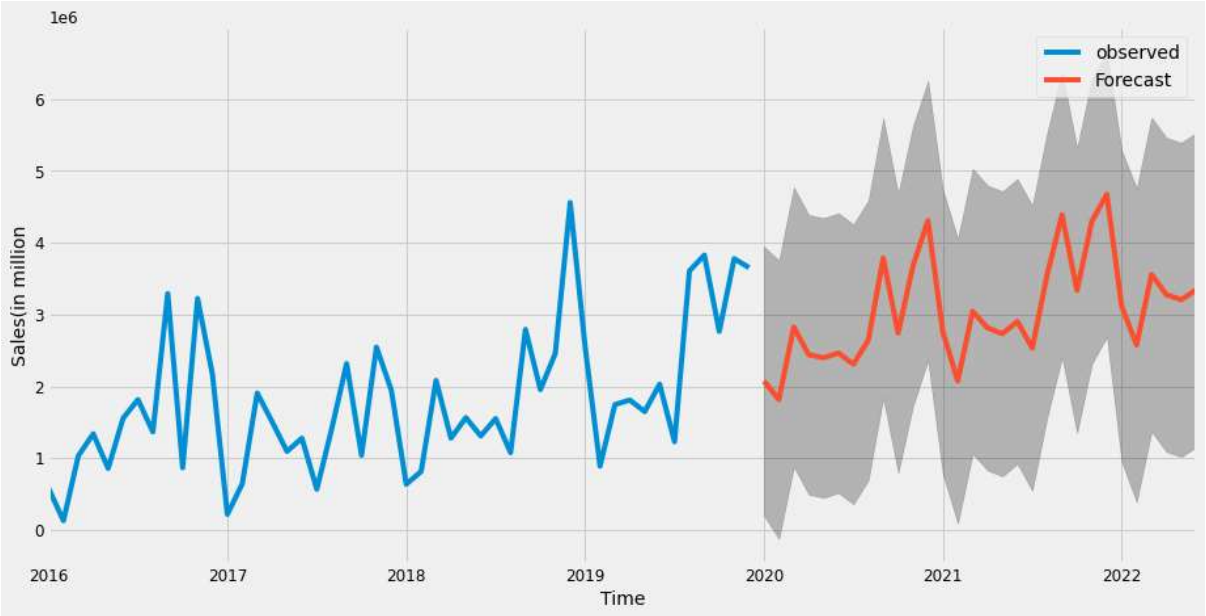


Fig 4.1.3 : Result of Office Supplies Sales

For Total Sales:

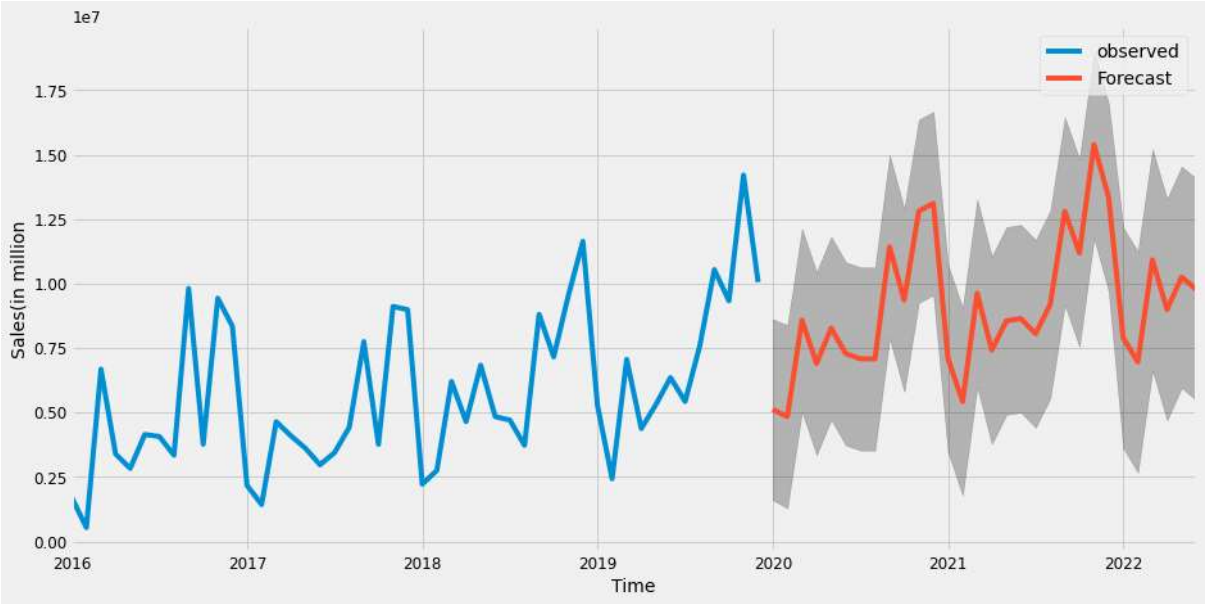


Fig 4.1.2 : Result of Total Sales

4.2. CONCLUSION

We have made an web app that takes current sales data from the user and performs exploratory data analysis and conducts prediction of sales of upcoming four years using ARIMA model

The sales data is accessed through the excel file given by the user and the analysis of sales is represented by graphical representations like pie chart and bar graph which allows the user to view the sales of each category. With the ARIMA model the sales prediction is done and the prediction is represented in a line graph. This web application allows the user to overview the sales and make future decisions based on the predictions.

4.3. FUTURE ENHANCEMENT

The following feature can be added in the webapp to make it more efficient and marketable :

Upload data to the page:

In this application for the enhancement we can add a upload data feature page in the application so that any user than upload their sales data for analysis and prediction.

REFERENCES/BIBLIOGRAPHY

[1] "A Comprehensive Guide To Predictive Analytics.." Medium20 Feb. 2019,

[2]"How Big Data is Helping Businesses Grow - Entrepreneur." 2 Jul. 2018, https://www.entrepreneur.com/article/316057. Accessed 29 Nov. 2019.

[3] "Customer Centric Sales Analysis and Prediction - International ...." https://www.ijeat.org/wp-content/uploads/papers/v8i4/D6417048419.pdf.

[4] Explaining machine learning models in sales predictions." Explaining machine learning models in sales predictions.

[5] Data Analysis :https://www.geeksforgeeks.org/python-data-analysis-using-pandas/

[6] Pandas Documentation : https://www.learnpython.org/en/Pandas_Basics

[7] Plotly Documentation : https://plot.ly/python/getting-started/

[8] ARIMA Documentation : https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python

SCREENSHOTS

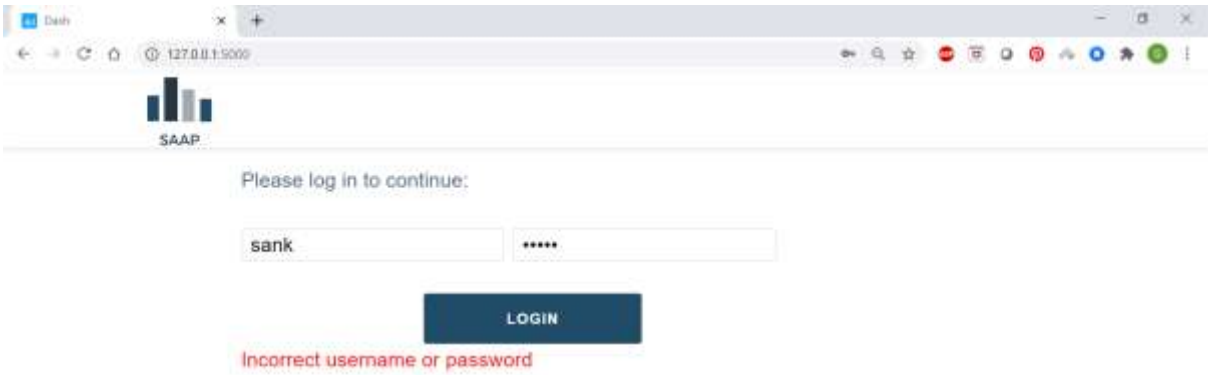


Fig 6.1 : Login page with invalid login credentials

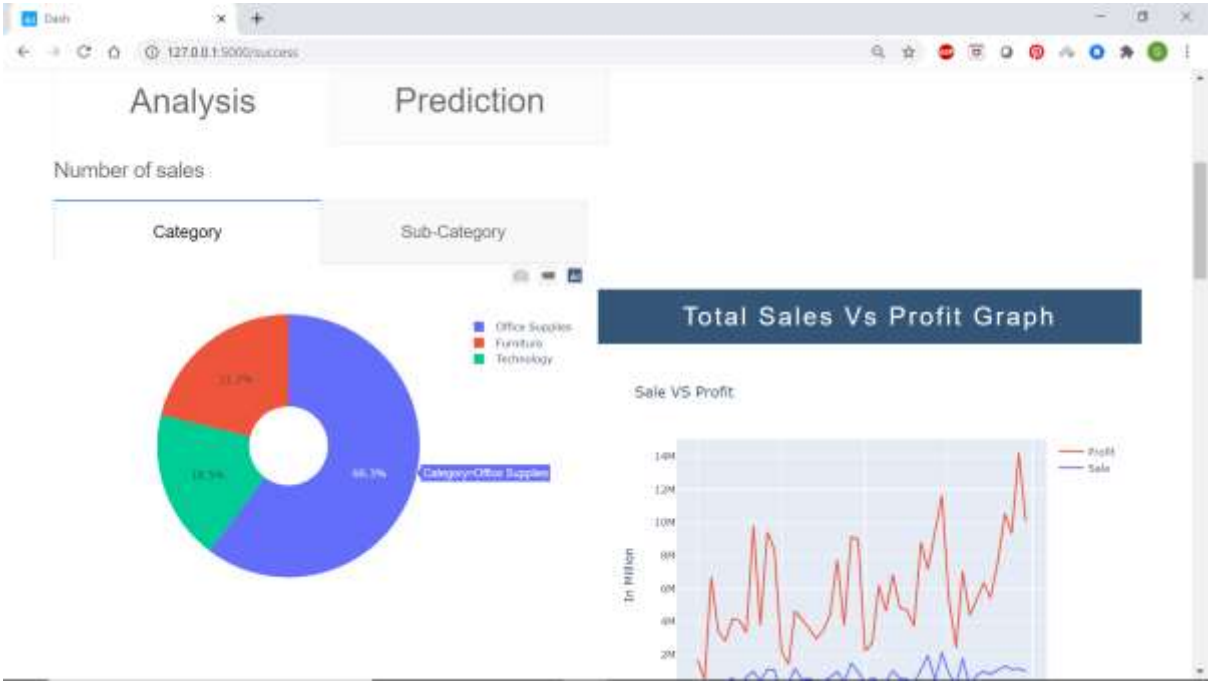


Fig 6.2 : Analysis Page

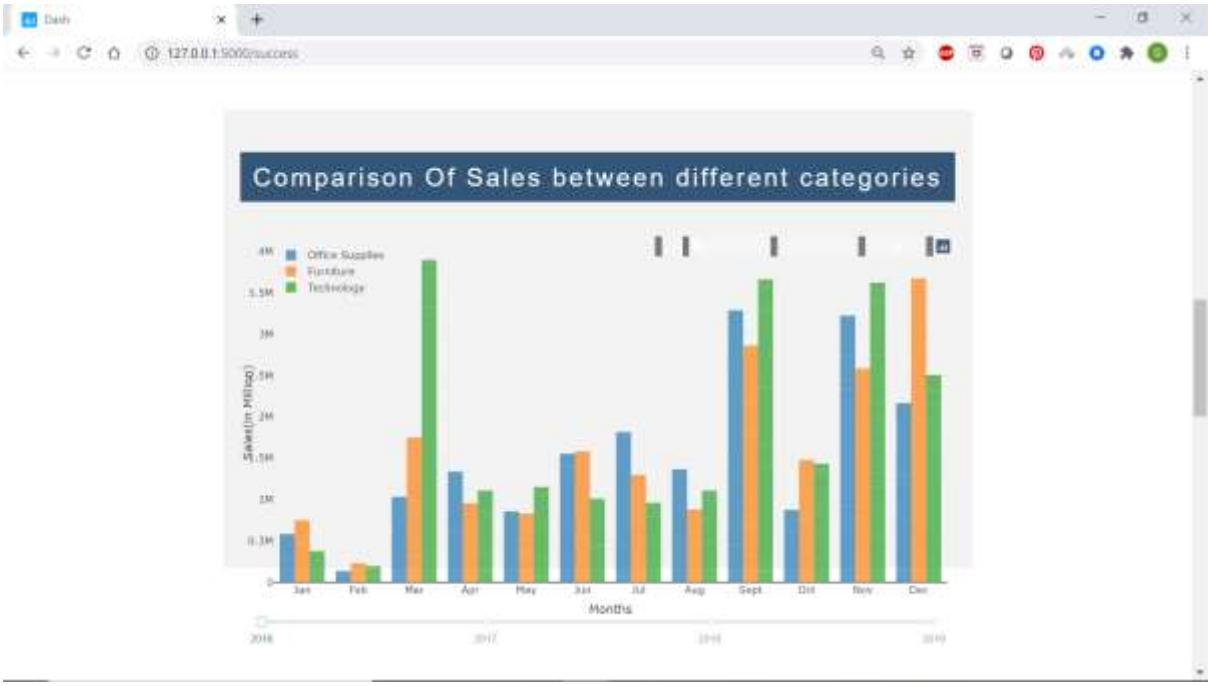


Fig 6.3 : Bar graph to view and compare Sales of categories(yearwise)

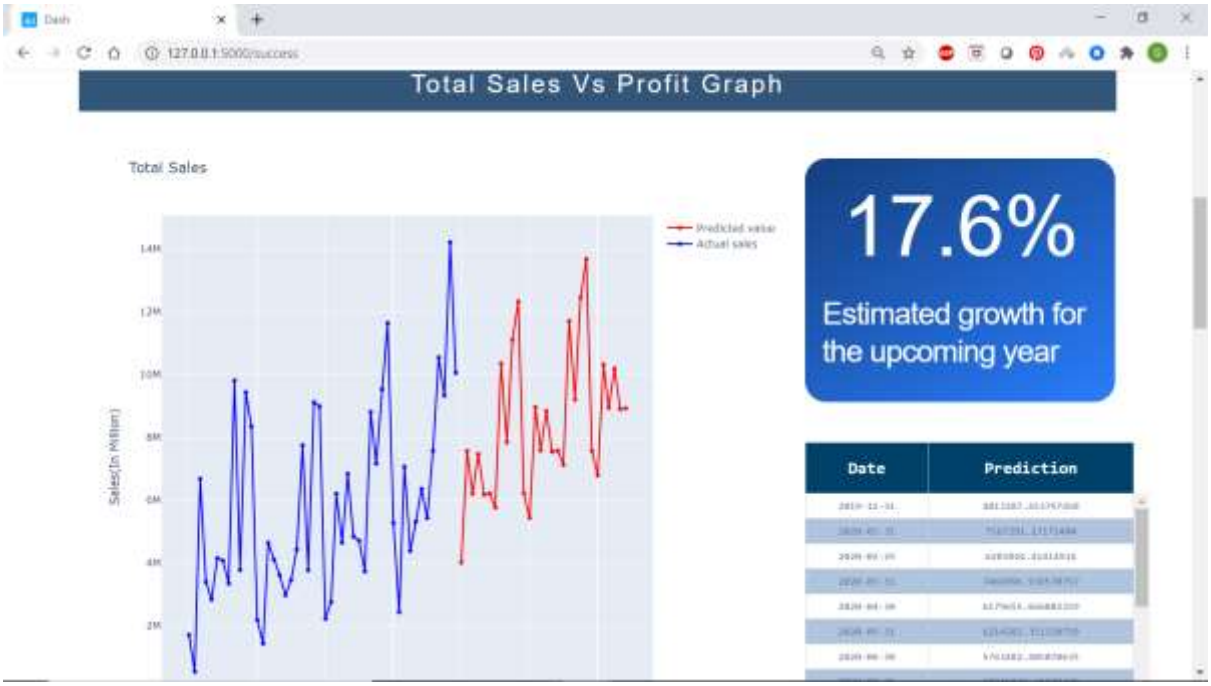


Fig 6.4 : Total sales vs Profit graph (Prediction page)

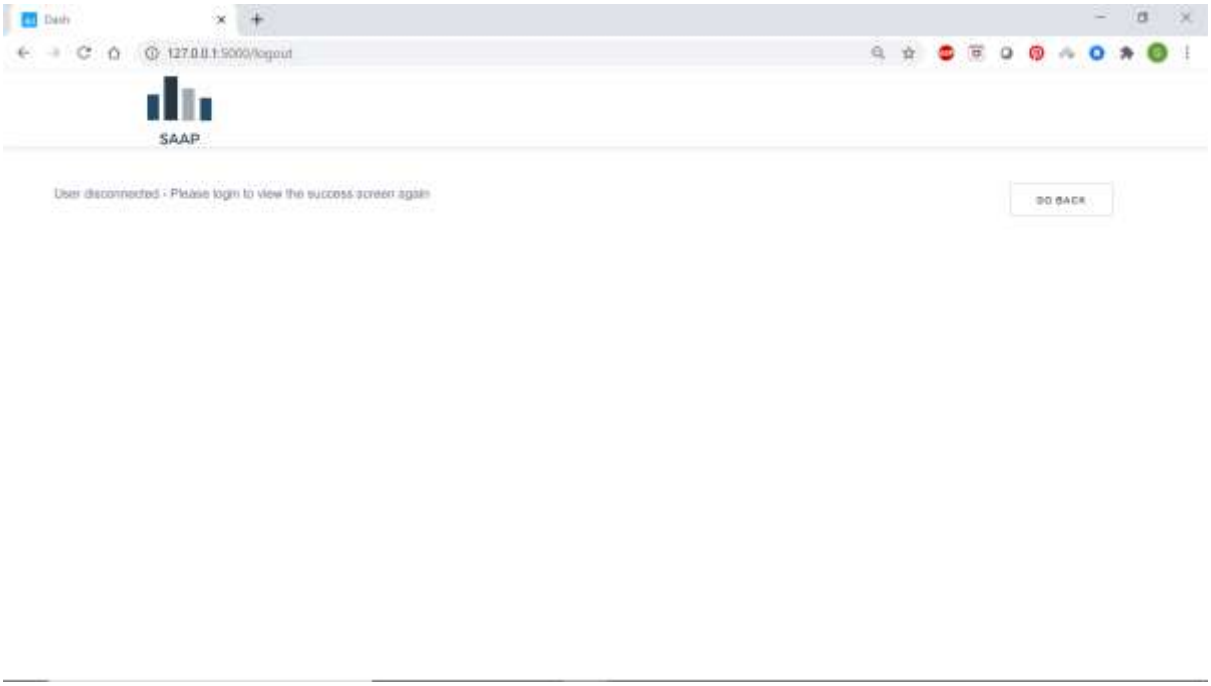


Fig 6.5 : Page displayed after logout