IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 4, April 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304339 |

E-Commerce Sales Forecasting with Big Data Analytics

Dr. Gururaj T[1], Amruth Ravindranath Hebbar[2], B R Veeresha[3], Soundarya H M[4], U K Spoorthi[5]

Associate Professor, Department of Computer Science and Engineering, Bapuji Institute of Engineering and Technology, Davangere, Karnataka, India[1]

U.G. Student, Department of Computer Science and Engineering, Bapuji Institute of Engineering and Technology, Davanagere, Karnataka, India[2 3 4 5]

ABSTRACT: This paper represents the forecasting system for sales. Nowadays, E-commerce is a valuable resource and a simple way for anyone to buy, sell, and earn money through an online platform. E-commerce will benefit both online sellers and customers who are looking to buy things. There are plethora of e-commerce websites available on the internet, which forces vendors to maintain a high level of competition for all clients looking to purchase products. e-commerce websites confront numerous challenges, such as new items, quality, and so on, all of which have an impact on sales. As a result, we require a large network for sales forecasting. And sales forecasting is a major concern for every company on the market. To address these issues, we proposed the online products sales forecasting using AIML techniques. We will use past company data sales for products specifics to construct a machine learning system to anticipate sales for this prediction of sales data.

KEYWORDS: E-Commerce, Forecast, Machine Learning.

I. INTRODUCTION

Big Data: Big data refers to vast volumes of structured and unstructured data that inundate businesses on a daily basis. This data is characterized by its volume, velocity, and variety, often exceeding the processing capabilities of traditional database systems. The term encompasses not just the data itself but also the technologies and methodologies used to analyze, process, and derive insights from it. With the proliferation of digital devices, sensors, and online activities, big data has become a cornerstone of modern business operations across various industries. One of the key challenges of big data is managing its sheer volume. Traditional data management systems struggle to handle the scale of data generated by sources like social media, IoT devices, and business transactions. As a result, organizations have turned to technologies such as Hadoop, Spark, and NoSQL databases to store and process large datasets in a distributed and scalable manner.

Machine Learning: Machine learning is a branch of artificial intelligence (AI) that empowers computers to learn from data and improve their performance without being explicitly programmed. It involves algorithms that analyze data, identify patterns, and make decisions or predictions based on that analysis. One of the key aspects of machine learning is its ability to automatically adapt and learn from experience, making it particularly valuable for tasks where explicit programming may be impractical or impossible. There are several types of machine learning approaches, including supervised learning, unsupervised learning, and reinforcement learning. Supervised learning involves training a model on labeled data, where each input is paired with the correct output. Unsupervised learning, on the other hand, deals with unlabeled data, and the algorithm must find patterns or structures within the data on its own. Reinforcement learning is a type of learning where an agent learns to make decisions by interacting with an environment and receiving feedback in the form of rewards or penalties.

Linear Regression Algorithm: Linear regression is a fundamental statistical method used for modelling the relationship between a dependent variable and one or more independent variables. It assumes that there is a linear relationship between the variables, meaning that a change in one variable is associated with a proportional change in the other(s). The primary goal of linear regression is to find the best-fitting straight line that describes the relationship between the variables. This line is determined by estimating the slope (coefficients) and intercept that minimize the difference between the observed values and the values predicted by the model. In practice, linear regression is widely

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304339 |

applied in various fields such as economics, finance, biology, and social sciences for prediction and inference tasks. It is extensively used for forecasting future trends, understanding the strength and direction of relationships between variables, and making decisions based on data-driven insights.

XG Boost: XG Boost, short for Extreme Gradient Boosting, is a powerful and efficient machine learning algorithm that has gained widespread popularity for its effectiveness in various data science tasks, particularly in structured data analysis and predictive modeling. Developed by Tianqi Chen and now maintained by the community, XG Boost is renowned for its speed, scalability, and performance in competitions like Kaggle, where it has been a staple choice for many data scientists. At its core, XG Boost belongs to the ensemble learning category, specifically boosting algorithms, which combine multiple weak learners (typically decision trees) to create a robust predictive model. Unlike traditional decision tree algorithms, XG Boost utilizes a gradient boosting framework, where subsequent trees are built to correct the errors of the previous ones, effectively minimizing the overall prediction error.

II. LITERATURE SURVEY

In recent years, there has been a variation in research focusing on enhancing sales forecasting in e-commerce through the utilization of advanced technologies such as machine learning and big data analytics. Papers [1] and [2 underscore the transformative impact of machine learning in various domains, including commerce, highlighting the need for more sophisticated approaches to sales forecasting. These advancements have led to the development of techniques like modified random forest and decision tree algorithms, as discussed in Paper [1], and hybrid machine learning methods, as explored in Paper [2].

Papers [3] and [4] delve into the integration of big data analytics into e-commerce sales forecasting, emphasizing its effectiveness in revolutionizing traditional approaches. While Paper [3] focuses on the enhanced predictive accuracy and dynamic resource allocation facilitated by big data analytics, Paper [4] highlights the challenges posed by factors such as data overload, data quality, and model explain ability. Further research directions and trends are explored in Papers [5], [6], and [7]. Paper [5] discusses the optimization of e-commerce sales force casting through data analytics, pointing out evolving technologies and strategic planning as key areas of focus. Meanwhile, Paper [6] presents a case study evaluating the impact of big data analytics on e-commerce sales force performance, emphasizing the need for addressing challenges related to data quality and model explain ability.

Additionally, Paper [7] introduces a store-sales forecasting model using machine learning, highlighting its potential limitations in capturing external factors and dependency on accurate data. In the realm of comparative analysis, Papers [8] and [9] compare traditional approaches with big data analytics techniques in e- commerce sales forecasting. While Paper [8] evaluates various data-driven approaches and emphasizes the importance of data quality and computational resources, Paper 9 provides insights into the superiority of big data analytics in enhancing sales force casting effectiveness. Paper [10] conducts a meta-analysis to investigate the overall impact of big data analytics on e-commerce sales forecasting accuracy. It identifies challenges such as heterogeneity of studies and publication bias, while also highlighting the importance of data preprocessing and model selection in achieving accurate forecasts.

III. METHODOLOGY

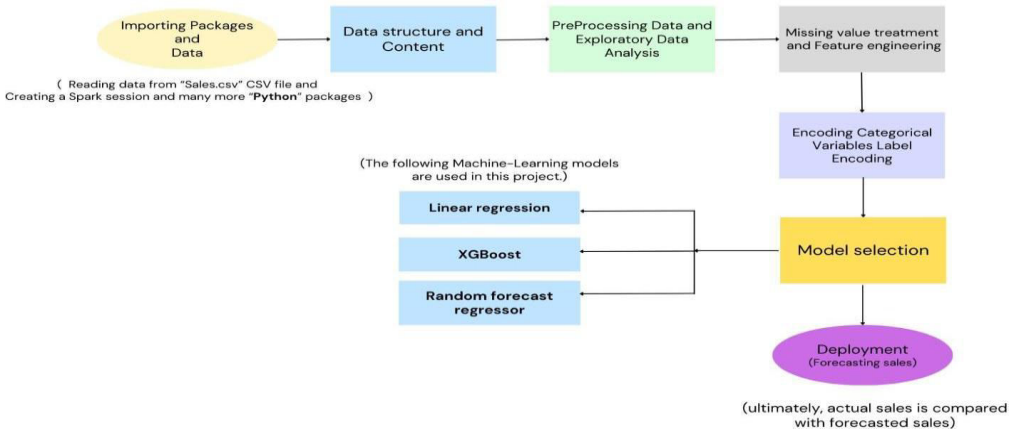

Fig 1: Methodology of Proposed System

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304339 |

Description of Methodology

1. Loading Packages and Data:

To build a project for e-commerce sales forecasting, you'll need to load relevant packages and collect datasets as shown in the fig 4.1. Below is a general guide on the steps involved in these processes without providing specific code:
Package Loading:
Pandas: Use for data manipulation and analysis.
NumPy: Essential for numerical operations.
Matplotlib/Seaborn: For data visualization

2. Dataset Collection:

Kaggle:
Explore Kaggle datasets related to e-commerce and sales forecasting. Download datasets in formats like CSV, Excel, or other commonly used data formats. Handle missing values, outliers, and irrelevant columns. Convert date columns to datetime objects. Explore and understand the structure and characteristics of the data.

3. Data structure and Content:

In the context of e-commerce sales forecasting as in fig 4.1, data can be categorized into structured, semi-structured, and unstructured formats, each offering unique insights into customer behaviour, market trends, and business performance. Structured data in e-commerce sales forecasting refers to information organized in a tabular format with well-defined columns and rows. Semi-structured data in e-commerce often involves information that doesn't fit neatly into traditional relational databases but still has some level of organization. Examples include data in JSON or XML formats, which may be encountered in web logs or API responses. Semi-structured data can contain valuable details such as user interactions, clickstream data, and customer reviews. Unstructured data in e-commerce includes information that lacks a predefined data model, making it less organized and more challenging to analyse using traditional methods. Examples of unstructured data include social media comments, images, and video content. Sentiment analysis on customer reviews, image recognition for product identification, and natural language processing for text data are techniques applied to derive meaningful information from unstructured data. Understanding customer sentiments, preferences, and brand perception from unstructured sources contributes to a holistic view of the e-commerce landscape. The content within e-commerce data encompasses the actual information contained in each data record.

4. Preprocessing data and Exploratory data analysis:

Exploratory Data Analysis (EDA) is a critical step in any data science project, including e-commerce sales forecasting. It involves analysing and visualizing the dataset to understand its characteristics, identify patterns, and gain insights. Here's a guide on performing EDA for e-commerce sales forecasting. Begin by loading your e-commerce sales dataset into a data analysis environment such as Jupiter Notebook or your preferred IDE. Check the structure of the dataset using functions like head(), info(), and describe().Understand the meaning of each column and their data types. Identify and handle missing values appropriately. You may choose to impute missing values or remove rows/columns depending on the impact on your analysis. Explore the distribution of the target variable (sales) and other relevant features. Use histograms, kernel density plots, or box plots to visualize the distribution. If the data involves a time component, plot the sales trends over time. Check for seasonality, trends, and any recurring patterns. Use correlation matrices or heatmaps to understand the relationships between numerical variables. Identify potential predictors for sales, considering factors like product attributes, marketing spend, or external events. Explore the distribution of categorical variables using bar charts. Analyse how different categories might affect sales. Conduct statistical tests to validate assumptions or identify significant differences between groups.

5. Missing value treatment and Feature engineering:

Handling missing values is a crucial step in the data preprocessing phase of any data analytics or machine learning project, including e-commerce sales forecasting. Here's a guide on how to find and treat missing values:
Identify Missing Values:

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304339 |

Use descriptive statistics or functions like is null () to identify missing values in your dataset. Visualize missing values using heatmaps or bar charts to understand the distribution of missing data across different features. Investigate why data might be missing. It could be due to data collection errors, system issues, or other reasons. Differentiate between missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) scenarios.
Feature engineering is a crucial step in the process of building effective machine learning models for e-commerce sales forecasting. It involves creating new features or transforming existing ones to enhance the predictive power of the model. Here are several ways feature engineering can be applied in the context of e-commerce sales forecasting.

6. Encoding categorical variables label encoding:

Label encoding is a technique used to represent categorical variables with numerical labels shown in fig 4.2. In e-commerce sales forecasting, categorical variables may include product categories, customer segments, or any other feature with distinct categories. It is a method of converting categorical values into numerical representations. Each category is assigned a unique integer label. This transformation is particularly useful when working with machine learning algorithms that require numerical inputs.

7. Deployment (Forecasting sales):

Sales forecasting results can be visualized in the form of pie charts or bar graphs to provide a clear and intuitive representation of the predicted values.

Pie Chart:
A pie chart is a circular statistical graphic divided into slices to illustrate numerical proportions. Each slice represents a portion of the whole, and the size of each slice is proportional to the quantity it represents. Use a pie chart to show the distribution of sales across different product categories. Each slice represents a product category, and the size of the slice corresponds to the percentage of total sales that category contributes. Imagine you have a sales forecast for an e-commerce store, and the pie chart might show the percentage of total sales coming from categories like electronics, clothing, and home goods.

Bar Graph:
A bar graph is a chart that presents categorical data with rectangular bars. The lengths of the bars are proportional to the values they represent. Use a bar graph to compare sales performance across different time periods, products, or other categorical variables. Each bar represents a category, and the height of the bar corresponds to the sales volume or revenue for that category. Consider a bar graph that displays monthly sales figures for a range of products. Each bar represents a month, and the height of the bar represents the total sales for that month.

IV. EXPERIMENTAL SETUP

Step 1: Setting up Python Environment
Install Python: Download and install Python
Step 2: Install anaconda prompt
Step 3:Install all the libaries
pip install pillow
pip install pandas
pip install django==3.0
Step 4: Data Collection: We have collected data from Kaggle.
We have performed quality checks on the collected data to ensure accuracy.
Step 5: Data Training
Import all the libraries run the command
Python training.py
Step 6: Results \ output by running the server by command python manage.py runserver

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304339 |

V. EXPERIMENTAL RESULTS

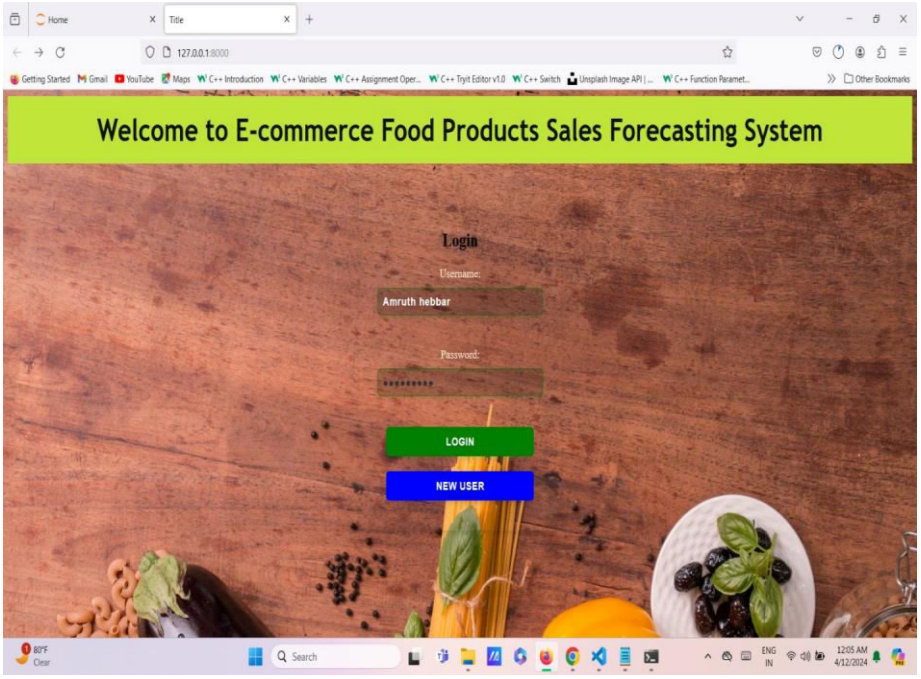

Fig 2: User Login

User have to login into the above page by entering username and password. The new user has to sign up by clicking new user button. And then again login to the above page.

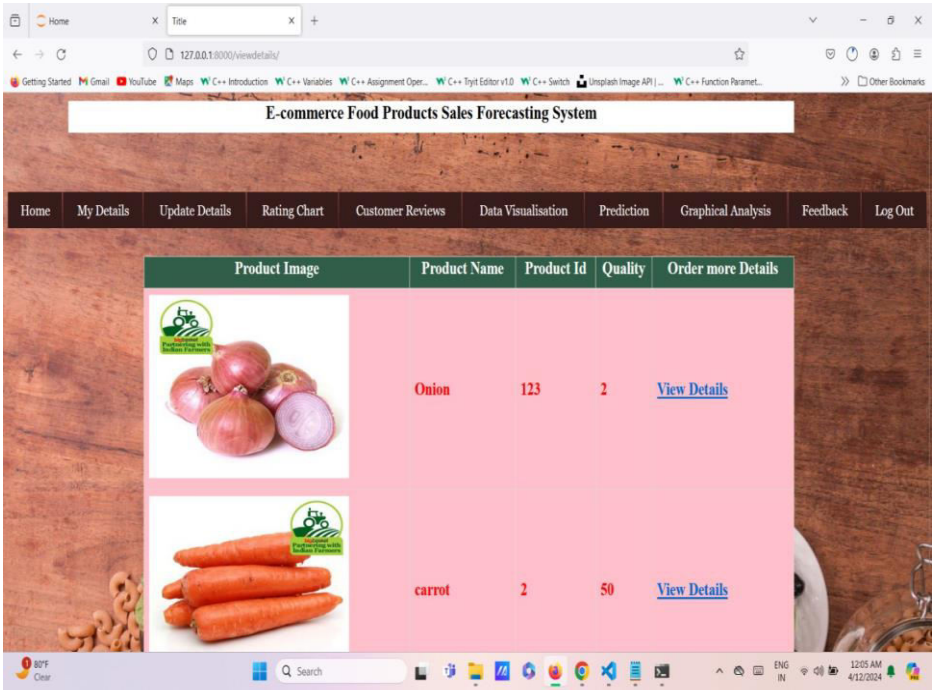

Fig 3 : Home Page

This is the home page user can see the details of all the products.

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304339 |

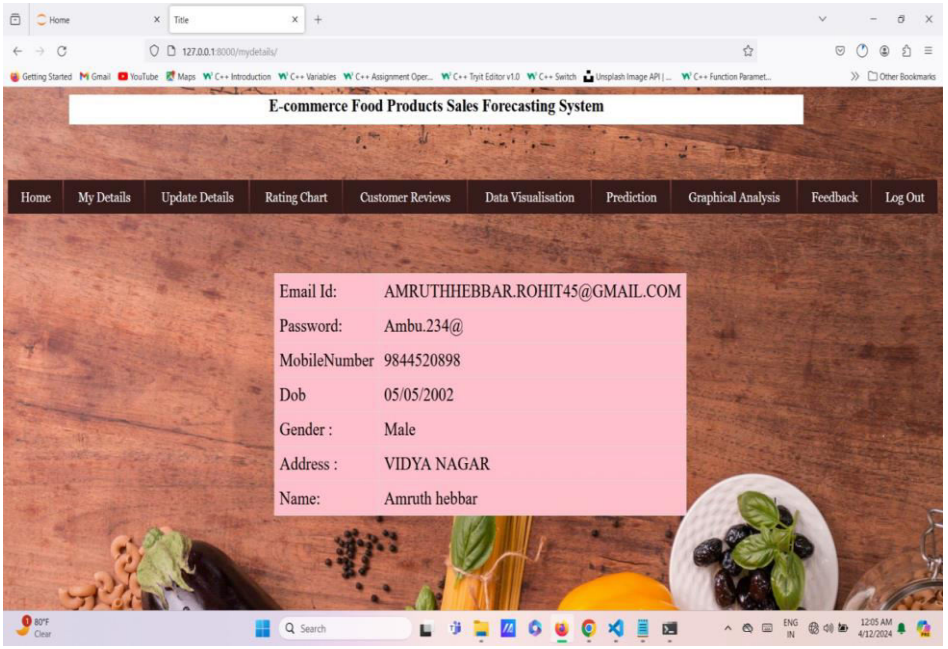

Fig 4: User Details

User can see the details including email id, password, mobile number, dob, gender, address, name

.

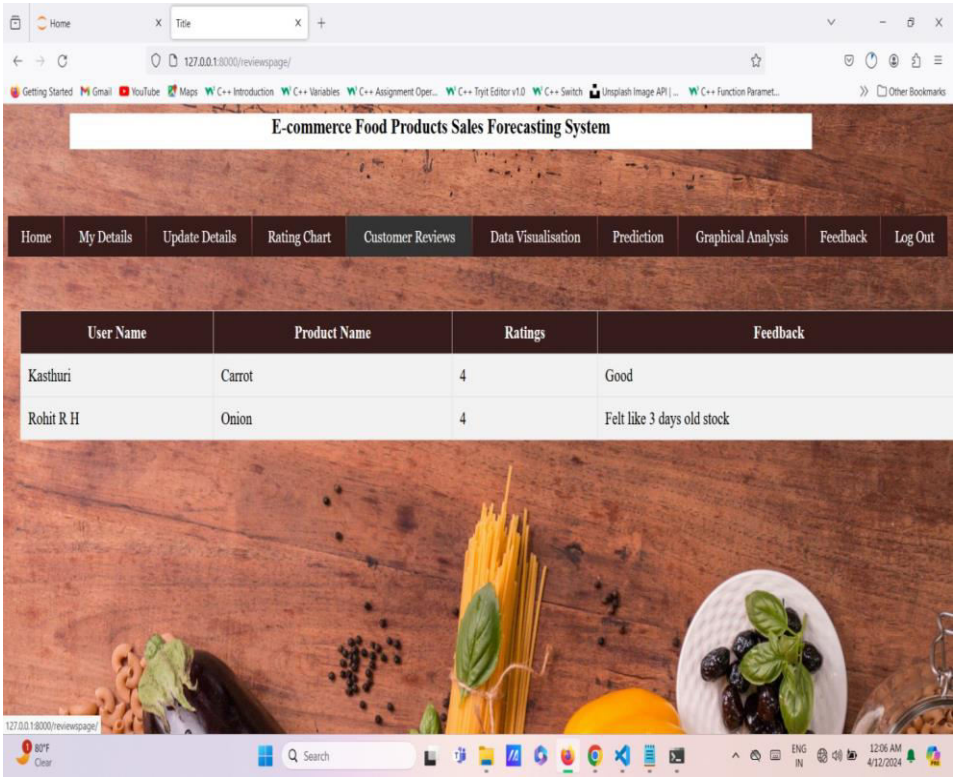

Fig 5. Customer Reviews

In this page we can see the reviews given by the customer with respect to the particular product.

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304339 |

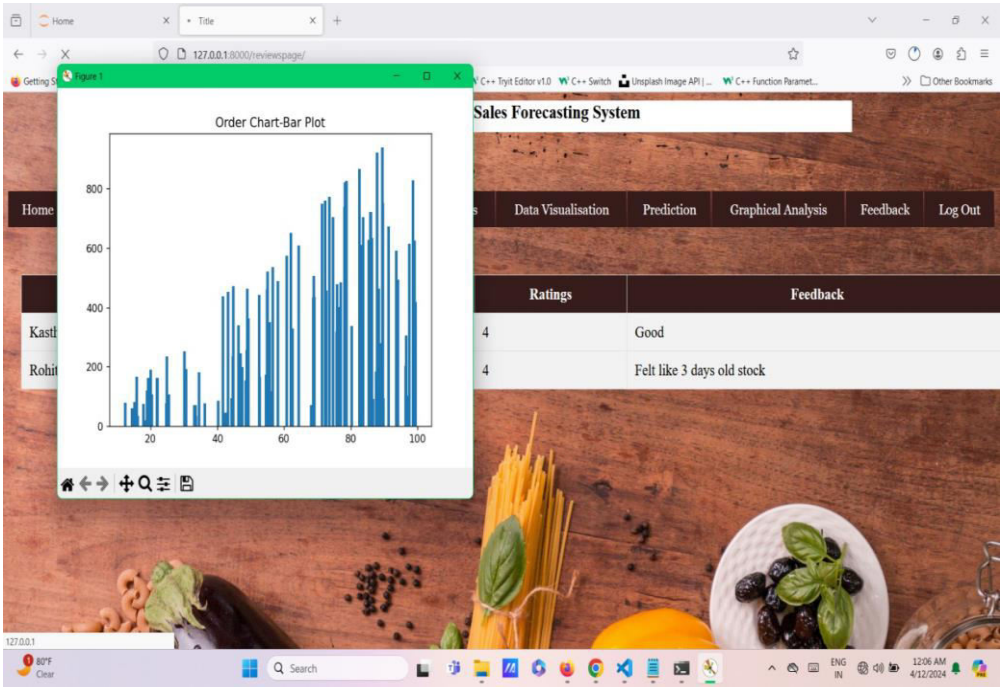

Fig 6. Before Prediction

Bar graph before prediction will be dispalyed. Including year and number of product sale.

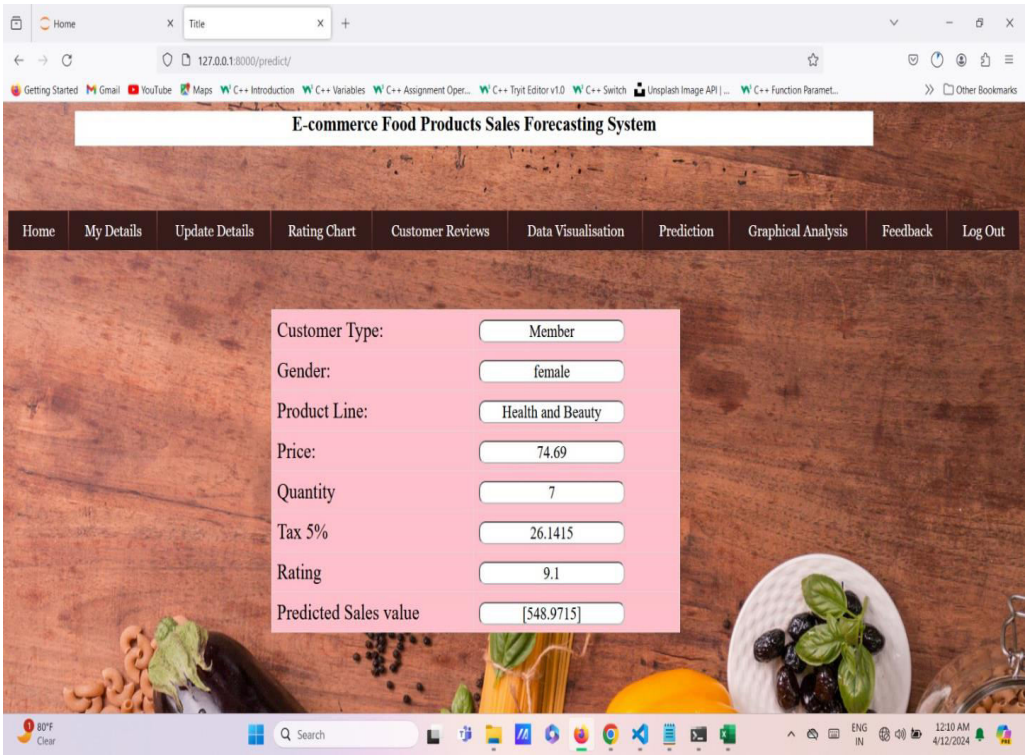

Fig 7. Forecasting

Forecasting of the particular product will be displayed with rating, price, quality and tax.

| e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304339 |

VI. CONCLUSION

Our project offers a web application to analyze the sales by using machine learning algorithm. E commerce websites will be very useful for all the customers who willing to purchase and sell the products within shorten time period. This project uses machine learning algorithm to predict the sales system. Using this sales system, we can analyze the data for all the products this helps the admin people to improve the sales system based on customer purchasing products. Here for predicting the sales we are using linear regression algorithm and plot the graph we are using matplotlib. For the visualization of the sales prediction, we interpreted results using pie chart, bar plot etc. Finally, this sales prediction system gives the better performance model.

REFERENCES

[1]. Analyzing and Predicting the Sales Forecasting using Modified Random Forest and Decision Tree Algorithm. Author: Ryali Ajay; R. Sabin Joel; P.G. Om Prakash. Year: 2023
[2]. Hybrid Machine Learning Method for Product Sales Forecasting in E-Commerce Author: Ravi Kumar. Year: 2023
[3]. Revolutionizing E-commerce Sales Force Casting: A Literature Survey on Big Data Analytics Integration. Author: Samantha Lee. Year: 2023
[4]. Big data analytics for e commerce sales forecasting: A review and research directions Author: Zhang, Y. Zheng, Y., & Yu, L. Year: 2022
[5]. Optimizing E-commerce Sales Force Casting through Data Analytics: Trends and Future Directions. Author: Jason Kim. Year: 2022
[6]. Evaluating the Impact of Big Data Analytics on E-commerce Sales Force Performance: A Case Study. Author: David White. Year: 2022
[7]. Store-sales Forecasting Model to Determine Inventory Stock Levels using Machine Learning. Author : Akanksha Akanksha, Year : 2022
[8]. A comparative study of big data analytics techniques for ecommerce sales forecasting Author: Chen, K., & Chen, T. Year:2020
[9]. A Comparative Analysis of E-commerce Sales Force Casting Techniques: Traditional vs. Big Data Analytics Approaches. Author: Michael Brown. Year: 2021
[10]. The impact of big data analytics on e-commerce sales forecasting accuracy: A meta analysis. Author: Wu, J., & Chen, X. Year:2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details