PREDICTING FUTURE SALES WITH ENSEMBLE MODELS: A HYBRID APPROACH USING RANDOM FOREST, XGBOOST, AND PROPHET

Dr.K.E.Kannammal, Dr.Y.Baby Kalpana, Mrs.P.Gomathi , Mrs.Diana Paul

Sri Shakthi Institute of Engineering and Technology, Coimbatore

ABSTRACT

Sales forecasting is the process of predicting future sales. It is an essential component of the company's financial planning. The majority of businesses rely significantly on sales projections for the future. Precise sales forecasting facilitates well-informed decision-making for organisations by predicting both short-term and long-term success. Accurate forecasting helps to prevent firms from overestimating or underestimating future sales, which can result in significant losses. The past and current sales statistics is used to estimate the future performance.

However, dealing with the accuracy of sales forecasting using conventional forecasting methods is challenging. For this purpose, various machine learning techniques have been used in recent days to analyse and maximize the online fashion store business by forecasting future sales and profit. In this work, Amazon dataset is used and a detailed analysis over the dataset has been made including the study of seasonality and trend, to identify a pattern. In this project, different machine learning algorithms like Random Forest regression, Prophet model, XGBoost regression have been implemented and evaluated. By analysing the performance, suitable predictive algorithm to the problem statement was proposed.

Keywords—Sales Forecasting, Machine Learning, Random Forecast Regression, XG Boost Regression, Predictive Algorithm, Future Sales Prediction

I. Introduction

Sales play a key role in the business. Sales forecasting is a crucial component of the business strategy and a vital source of information for decision-making inside the organisation. It is essential for organizations to produce the required quantity at the specified time. Sales forecasting may help with that by providing insight into how a company should use its resources, personnel, and budget. The business organizations can use this forecasting to estimate how many items to create, how much income to anticipate, and what kind of staff, capital, and equipment could be needed. Sales forecasting enhances corporate growth by examining future trends and requirements. The accuracy of the predictions and managing massive amounts of data are two issues with traditional forecasting systems.

To overcome this problem, Machine- Learning (ML) techniques have been implemented. These techniques helps to analyse the bigdata and plays a important role in sales forecasting. In this project supervised machine learning techniques are used for the sales forecasting. Time-series forecasting is an essential task in business analytics, where the goal is to predict future values of a time-dependent variable based on historical data. Accurate forecasting is essential for organizations to efficiently plan production, inventory management, and marketing operations in the context of seasonal item sales.

Since machine learning algorithms can learn from historical data and recognize intricate patterns, they have become widely used methods for time-series forecasting. In this project, three machine learning algorithms Random Forest regression, Prophet model and XGBoost regression have been implemented.

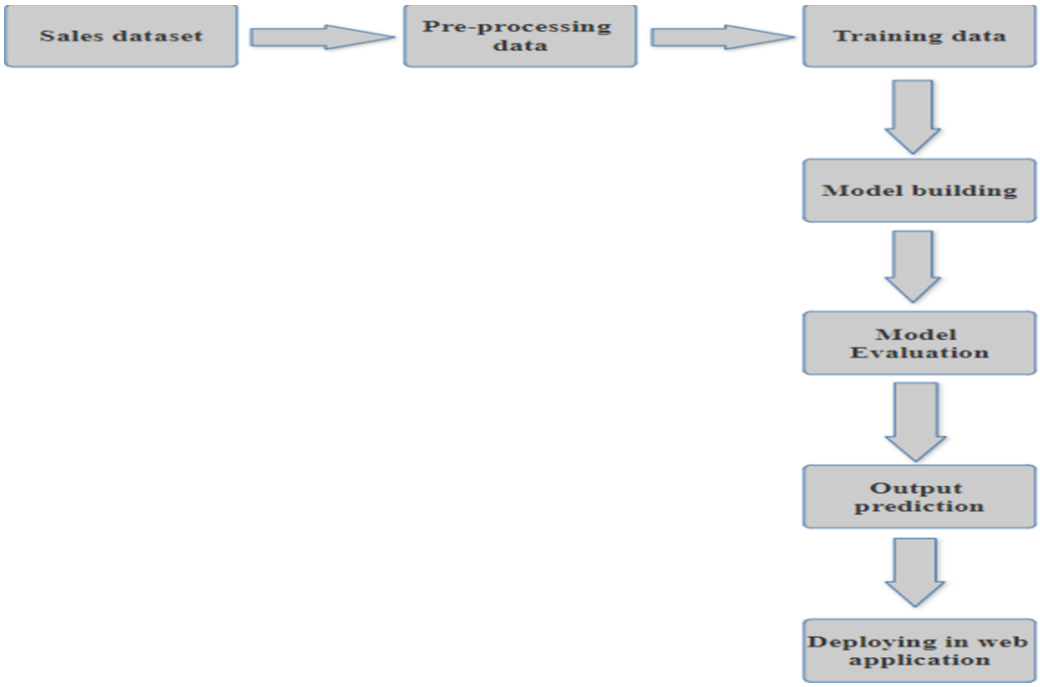
II. System Analysis

Existing System

The existing sales prediction system only utilizes traditional data and machine learning algorithms to forecast future sales trends. The system only predicts either the future sales of a particular product or the total sales of an ecommerce store. Many models that are currently in use, focus on the past sale profits and the sales volume of products using foreign datasets Walmart or bigmart store. However, there is a need to continuously improve the accuracy and scalability of the system, as well as to address challenges such as future profit and sales, model interpretability, and adaptability to changing market conditions.

Proposed System

The proposed system intend to predict profit details of the products and future sales of the different products especially in Indian Market using Amazon dataset. To forecast the future sales, Machine learning models Random Forest regression, prophet model and XGBoost Regression are used. The features from the historical sales data used are the quantity, category and date. These Machine Learning models helps to improve the accuracy and scalability of the system.



III. Module Description

1. Preprocessing and Feature extraction
2. Model Construction
3. Model Evaluation
4. Model Deployment into web application

3.1. Preprocessing and Feature extraction

Data Collection: Gather historical sales data, including dates, prices, category and corresponding profits from Kaggle dataset based on Indian sales market. Collect data from reliable sources to ensure accuracy and completeness. Ensure comprehensive coverage of the data over a sufficient time period.

Data Cleaning: Data cleaning includes handling missing values by imputation, removal, or interpolation techniques. Identifying and addressing outliers that may affect the model's performance. Data cleaning is also done by checking for inconsistencies in the data and resolve them to maintain data integrity. Normalize or scale numerical features is done if necessary to ensure consistent ranges.

Feature Engineering: Extract relevant features from the raw data. This could involve transforming the date into features like Year, Month, Day and day of the week. Performing one hot encoding for the categorical variables like category, shipping state and city and deriving additional features such as seasonality helps in improving the accuracy. Extracting the relevant features make a strong a relationship between them and improves performance of the model.

3.2.Model Construction

Model selection: In this project, XGBoost Regression, Random Forest Regression and prophet model are used for profit forecasting and sales forecasting.

Splitting the Data: The preprocessed data is split into the training set and the testing set in this stage. The training set is used to train the model, while the testing set is used to evaluate its performance. The training set typically contains a larger portion of the data, often 80%, while the testing set contains the remaining portion. This splitting ensures that the model is trained on one set of data and tested on a separate, unseen set to assess its generalization ability.

Training the Model: Once the data is split, the machine learning model is trained using the training set. Training the model involves adjusting its parameters to minimize the difference between the actual target variable values and the predicted values generated by the model. This process is often iterative, with the model gradually improving its performance as it learns from the training data. The goal is to find the optimal set of parameters that result in the best predictive performance on unseen data.

Profit Prediction: For calculating the profit prediction, the features used are date and qty and the target variable is profit, The XGBoost Regression is used to train the model based on the extracted features Year, Month, Day, Day of Week like Monday, Tuesday, Wednesday etc., from date feature, quantity and predict profit. All these features are of type numbers. The user gives year, month, day, day of week and quantity as input. The output is the profit calculated.

XGBoost Regression model: XGBoost is a powerful and effective algorithm that makes use of the boosting idea to create a robust model. When comparing these three models, XGBoost frequently offers the highest accuracy but its tuning and interpretation can be more intricate. When dataset have a lot of data and a combination of numerical and categorical variables, XGBoost performs exceptionally well.

Profit Forecasting: For calculating profit forecasting, the features used are date and profit. The Prophet and Random Forest Regression models are used for forecasting profits for a given start and end date. The model is trained using prophet for forecasting profits and Random Forest Regression is used for improving the accuracy by using as an ensemble method. The user provides the start date and end date for which they need to predict profit. The model predict profits from the given start date and end date.

Random Forest model: Random Forest is a powerful machine learning algorithm used for both classification and regression tasks. It is a part of the ensemble learning methods, where the algorithm builds multiple decision trees during the training phase and combines their predictions to improve the overall performance and robustness of the model.

In addition to sampling the data, Random Forest also introduces randomness in feature selection at each node of the decision tree. Instead of considering all features when making a split, only a random subset of features is considered. This helps in making the trees more diverse and less correlated with each other.

Sales Volume Forecasting: The features selected are date and total sales. The Prophet model is used for forecasting sales volume for a given date. After fitting the model the model is used to predict the future dates. The user gives date as input and the output is predicted sales volume. Then evaluate the accuracy of the sales volume forecasts by comparing them to actual sales volume data.

Prophet: Prophet is a forecasting tool developed by Facebook's Core Data Science team. It can handle time series data with different seasonality and strong seasonal patterns. Prophet is particularly useful for forecasting tasks in business applications where the data may have missing values, outliers, or non-linear trends. Prophet can automatically detect seasonal patterns in the data, including daily, weekly, and yearly seasonality.

Model Tuning: Fine-tune the model parameters to improve performance.

Saving the model: The model is saved using job lib module with dump function and the model is then loaded in the flask to use it for displaying the prediction

3.3 Model Evaluation

The dataset is split into train and test data. The test data is used to predict the profit forecasting to evaluate the performance. The evaluation metrics used for calculating the performance are Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), or R-squared to assess the model's accuracy and generalization ability. The prophet model and ensemble method is evaluated based on the evaluation metrics and comparing the models to check the model accuracy and select the model with higher accuracy. Visualize the predicted profit values against the actual profit values from the test dataset. Plotting these values over time can provide insights into the model's performance and any patterns or trends it captures effectively. Iterate on the model construction and evaluation process if necessary, adjusting parameters or trying different algorithms to achieve better performance.

Mean Squared Error:-

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$

Root Mean Squared Error: (RMSE)

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(\hat{y}_i - y_i)^2}{n}}$$

R-Squared:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

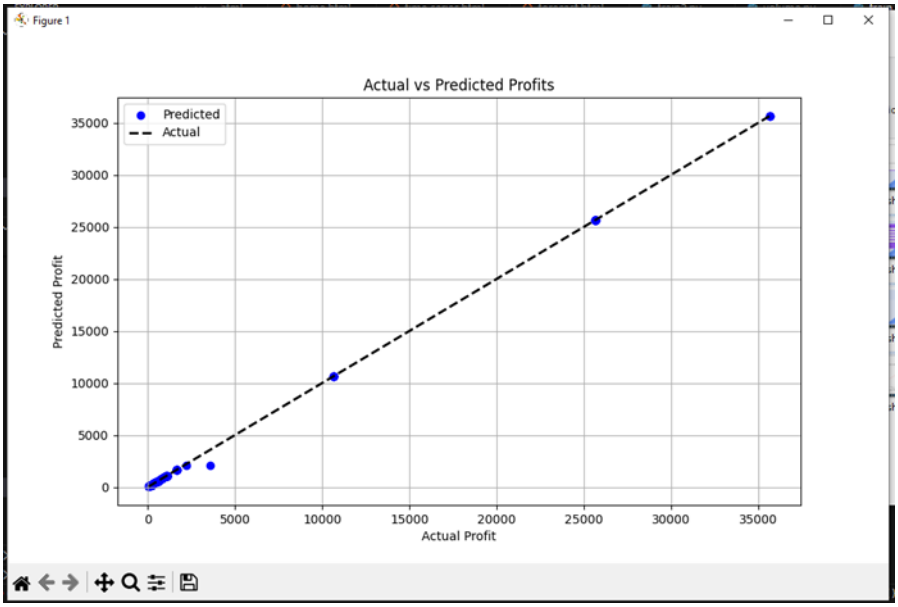


Figure 3.3.1 Actual vs Predicted

This plot is to find the difference between actual and predicted profits. When all the dots align in the line, it says that the prediction is good.

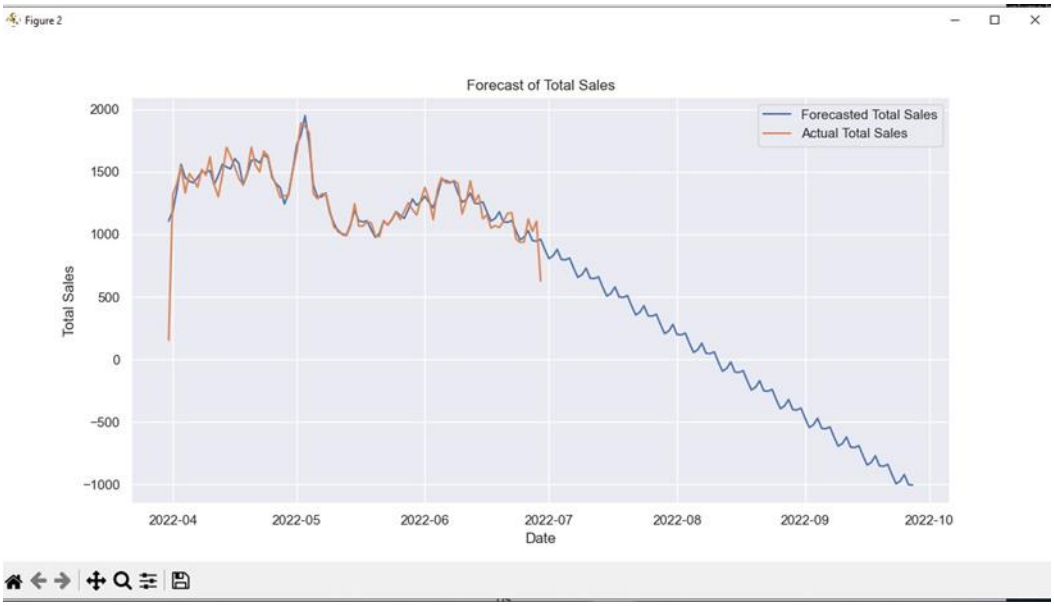


Figure 3.3.2 Forecasted sales

This plot is to find the trend in the forecasted sales in the particular dates.

Figure 3.3.3 Residuals

This represents the differences between the actual total sales values and the forecasted total sales values at each corresponding time point. Positive residuals indicate that the model underpredicted the actual values, while negative residuals indicate that the model over predicted the actual values.

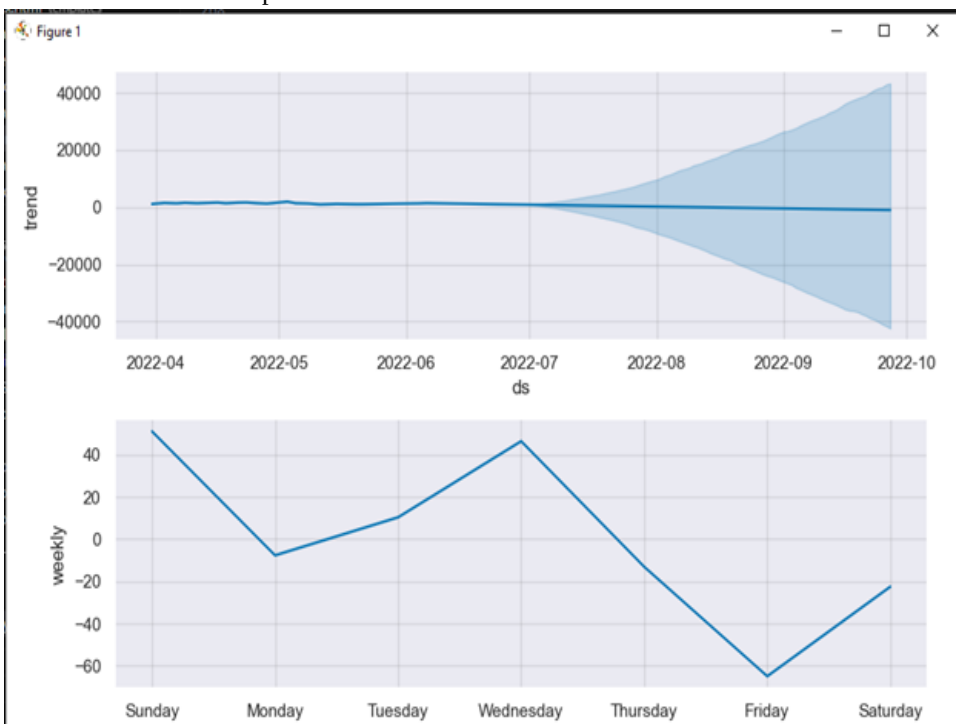


Figure 3.3.4 Trends and seasonality

This shows about the trends and seasonality for the products in the dataset.

3.4 MODEL DEPLOYMENT INTO WEB APPLICATION:

The Flask application structure is organized with appropriate directories for templates, static files. The background image is saved in the static file. Flask routes are defined to handle HTTP requests and responses inside @app.route(). The html files are rendered using routes for rendering HTML templates using render_template(). Then loading the trained models for profit prediction, profit forecasting and sales forecasting (Prophet and Random Forest Regression and XGBoost regression) within flask application helps to predict the output.

The model is loaded using load() function. The logic for forecasting the profits and sales using the loaded models is implemented. This involves processing input data from forms created using html, making predictions, and formatting the results. HTML templates are created for the web pages. These templates will display the user interface for interacting with profit forecasting application. The data given by user is sent from HTML templates to flask routes for prediction and display the predicted results to the user.

The output is displayed in different route links. The forecasted profits is displayed in route /timeseries and forecasted sales volume is displayed in route /forecast. When the user clicks any of the three buttons (profit, forecast, sales volume) it is directed to the particular page and a form is displayed. The user gives the details in the form. After filling the form, the user clicks the button, then the output is displayed.

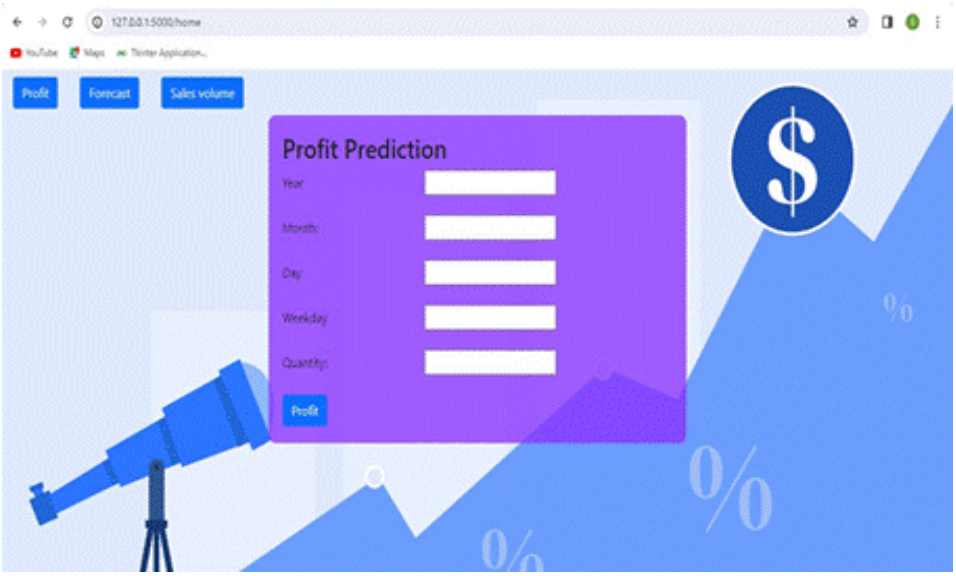


Figure 3.4.1 Profit Prediction

This is Profit prediction page where user will give details in the form and click profit button and output is displayed.

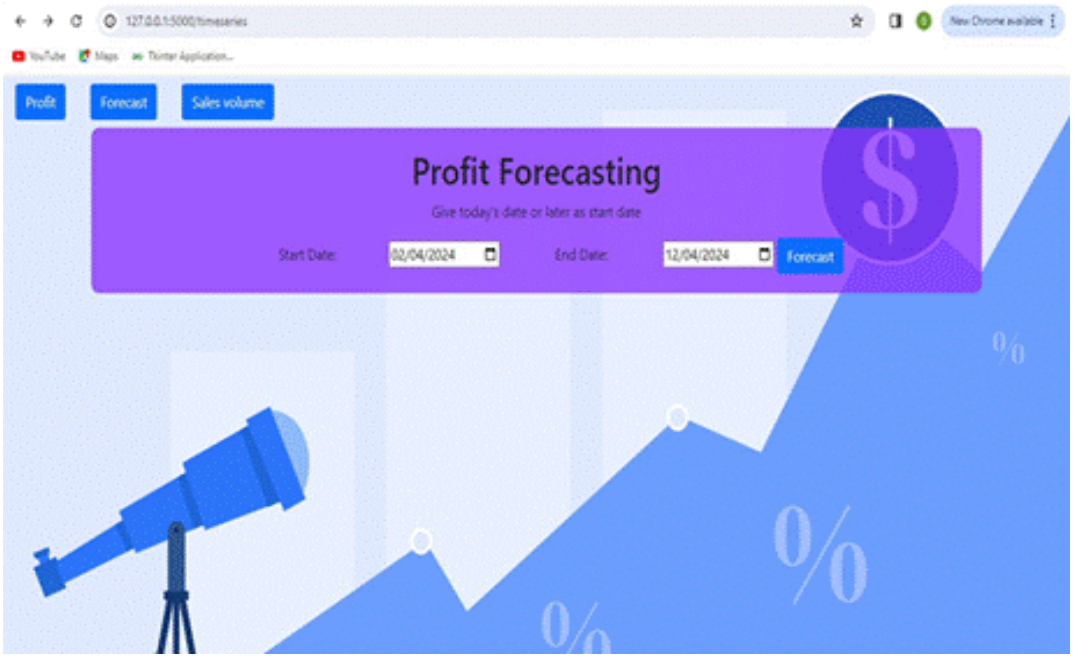


Figure 3.4.2 Profit forecasting

This is Profit forecasting page where the user gives start date and end date.

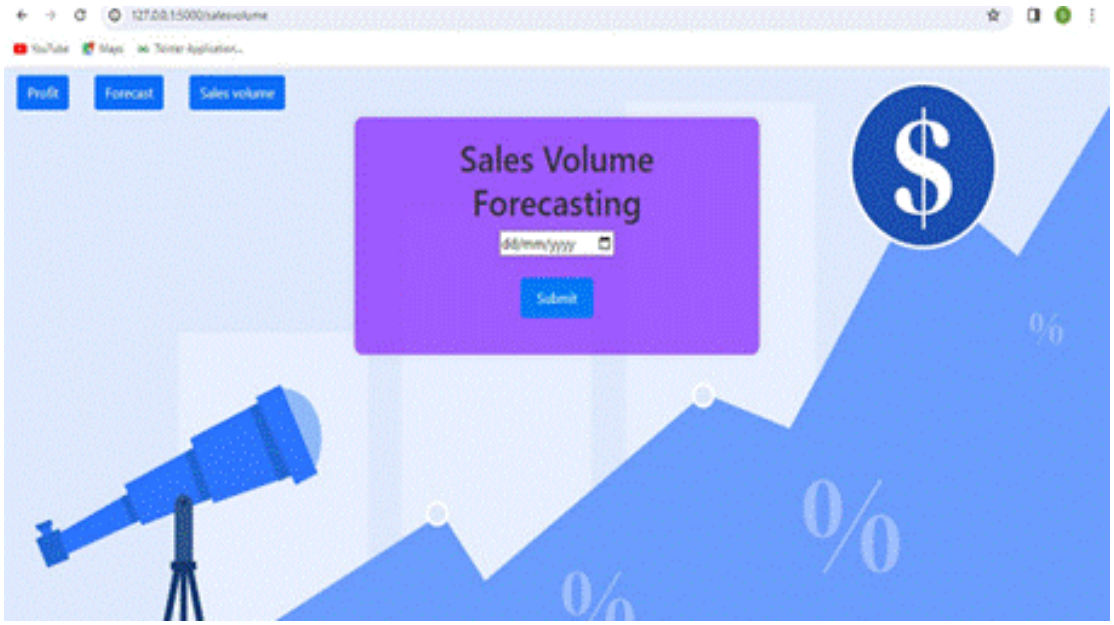


Figure 3.4.3 Sales forecasting

This is total sales forecasting page where the total sales will be forecasted for the given date.

IV. Performance Evaluation

4.1 Results and Analysis

This project predicts the profit and total sales volume for given date. The profit forecasting is done for given start date and end date and profit prediction is also done for given year, month, day, day of week using random forest regression and prophet. The total sales volume is forecasted for given date using prophet model**.**

Figure 4.1.1 Home page

Figure 4.1.1 is the Home page of the project. This page contains three buttons profit, forecast and sales volume. When the user clicks the button it is directed to the respective page.

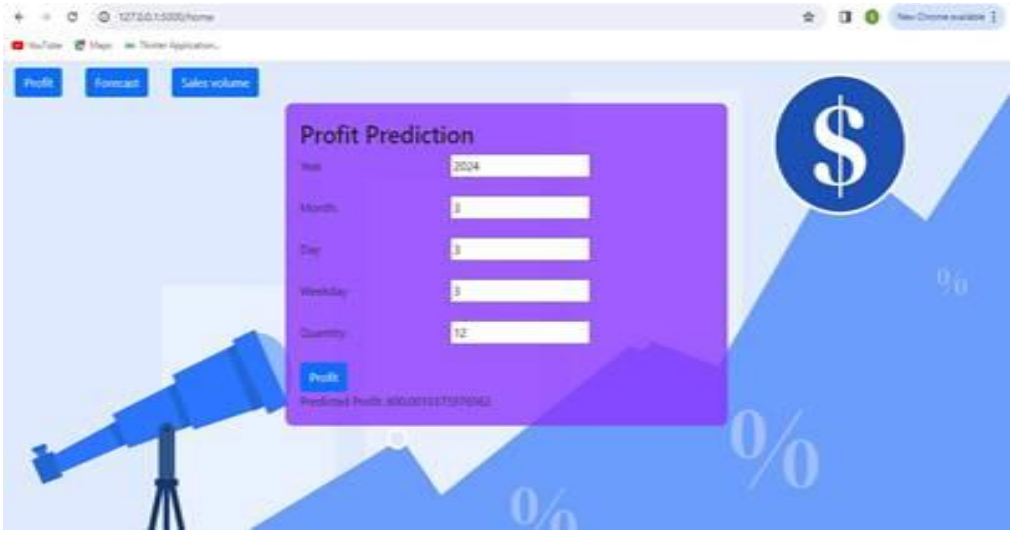


Figure 4.1.2 Profit Prediction page

This page predicts the profit for the given year, month, day and day of week. The output changes when the quantity changes. The user clicks the profit button and the profit is predicted.

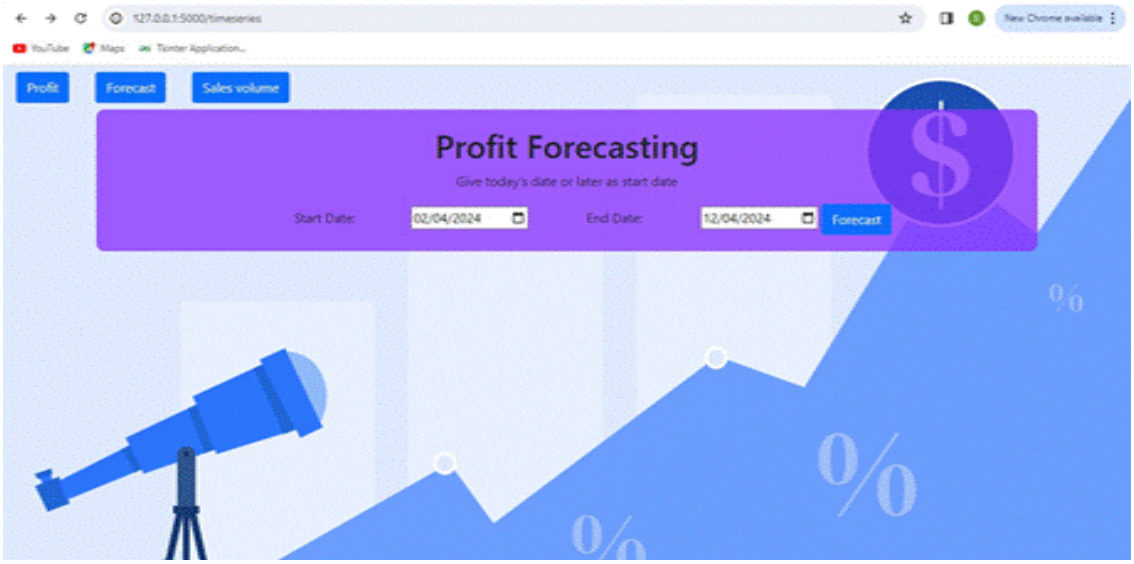


Figure 4.1.3 Profit forecasting page

This page is the input page for the dates for profit forecasting. The user gives start date and end date. The output is forecasted profits for given dates.

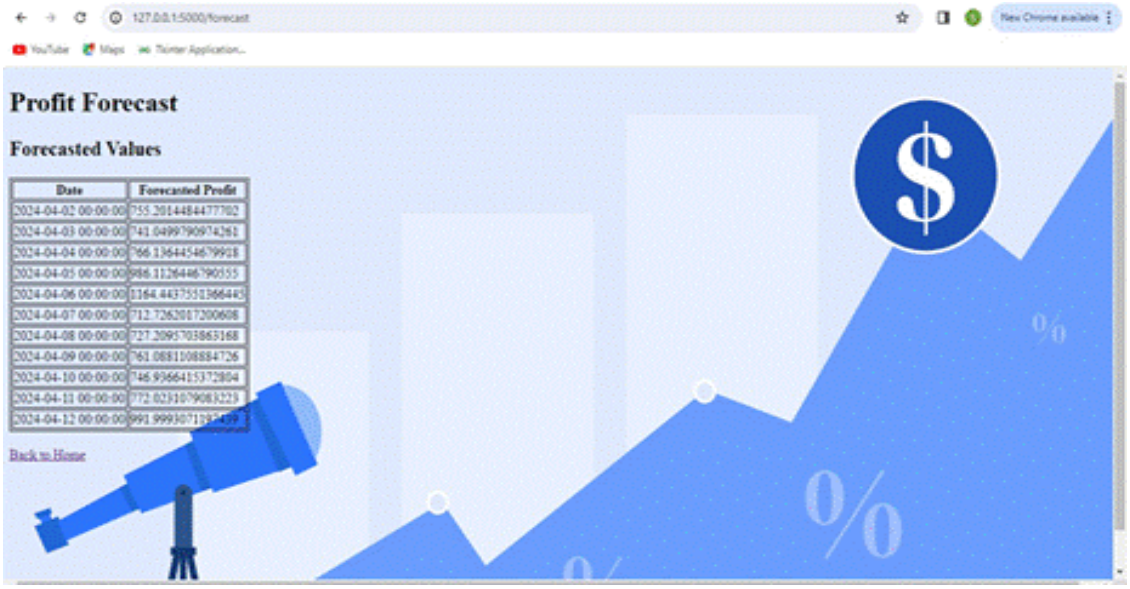


Figure 4.1.4 Forecasted profits page

This page displays forecasted profits for the given start and end date in table format.

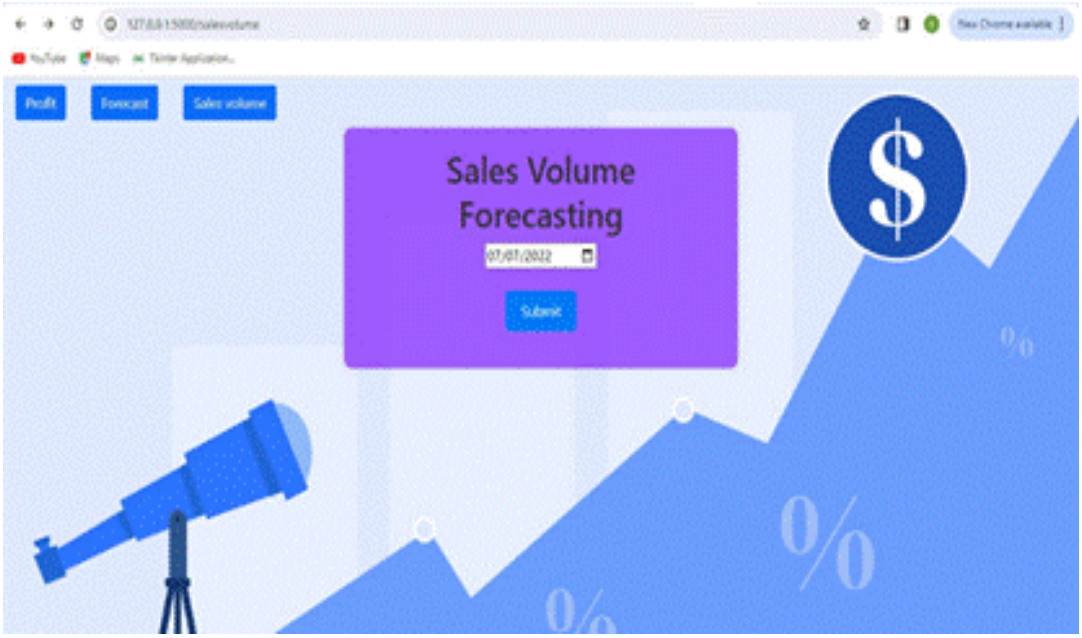


Figure 4.1.5 Sales forecasting page

This page asks for the input for total sales volume in sales forecasting. The user gives the date and clicks the submit button and the output is displayed.

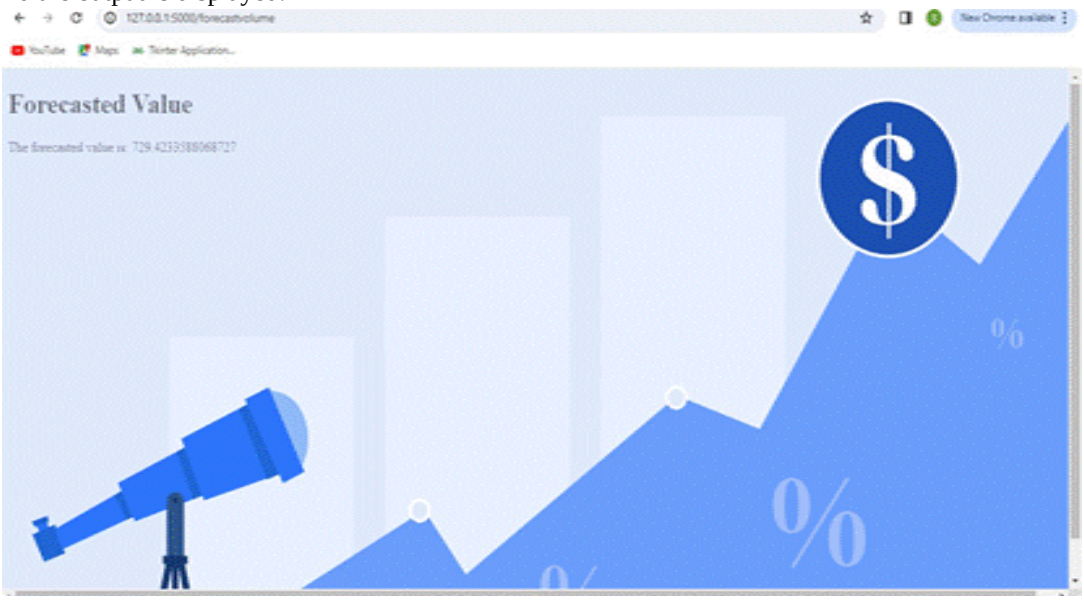


Figure 4.1.6 Forecasted sales volume page

This page displays the total sales volume for the given date.

4.2 Result Comparison:

XGBoost Regression vs. Linear Regression: XGBoost regression is identified as a superior model compared to Linear regression, particularly in terms of accuracy.
Ensemble Method (Random Forest + Prophet) vs. Individual Models: The combination of Random Forest regression and Prophet as an ensemble method is highlighted for its good accuracy in profit forecasting. Ensemble methods are known to leverage the strengths of individual models, potentially leading to improved predictive performance.

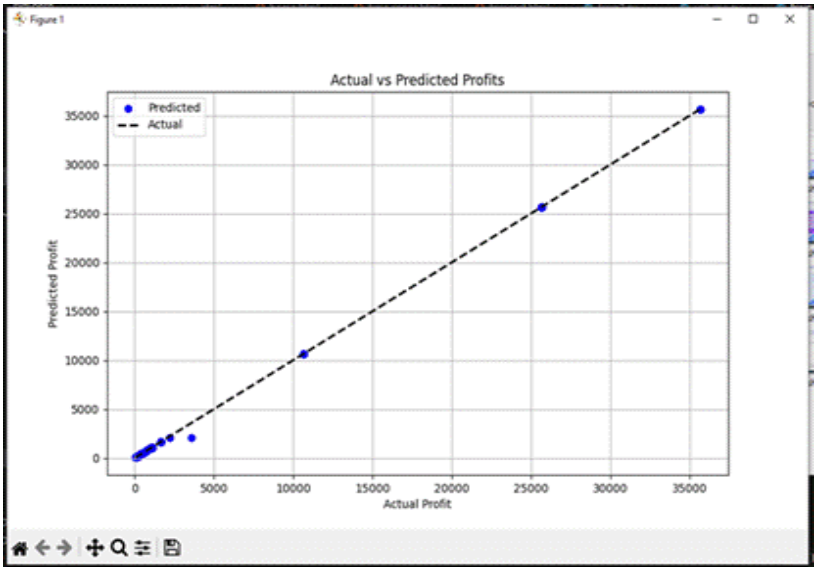


Figure 4.2.1 Actual vs Predicted values

This plot shows that the actual and predicted values are predicted well for the given dates. This model is selected as it does not overfit and the predicted values are aligned well.

Profit forecasting*:*

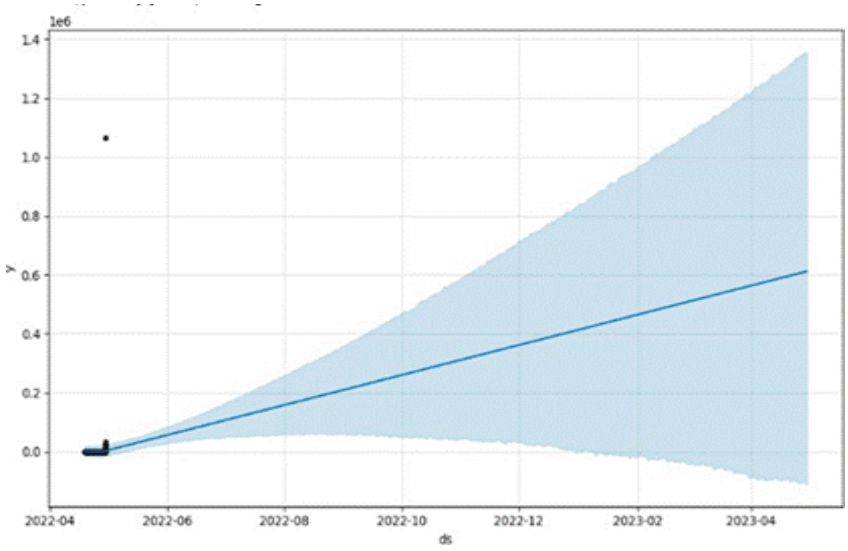


Figure 4.2.2 Profit forecasting using prophet

This plot depicts the Prophet model's forecasting, which exhibits lower accuracy and poor predictive capability.

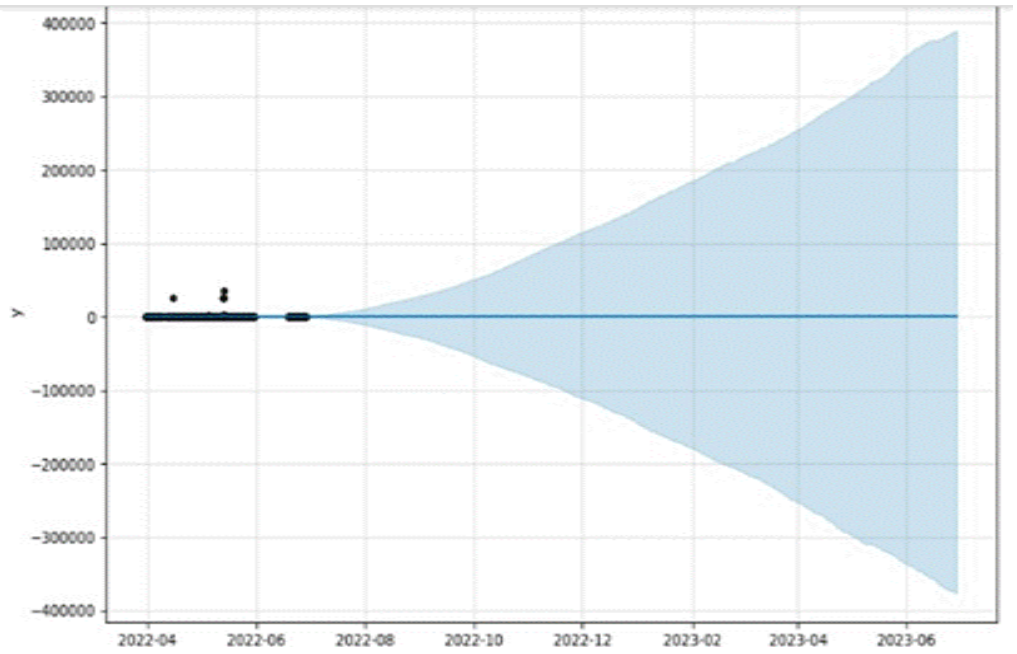


Figure 4.2.3 Profit forecasting using prophet and random forest regression

This plot illustrates profit forecasting achieved through an ensemble method,
leveraging the combined strengths of Prophet and Random Forest Regression. By integrating these models, higher accuracy is attained than when using Prophet alone. Given the improved results, this ensemble approach is chosen as preferred model for profit forecasting.

References

A. L. D. Loureiro, V. L. Migue´is, and L. F. M. da Silva, “Exploring the use of deep neural networks for sales forecasting in fashion retail,” Decision Support Systems, vol. 114, pp. 81–93, 2018.

N. S. Arunraj and D. Ahrens, “A hybrid seasonal autoregressive integrated moving average and quantile regression for daily food sales forecasting,” International Journal of Production Economics, vol. 170, pp. 321–335, 2015.

P. Ramos, N. Santos, and R. Rebelo, “Performance of state space and arima models for consumer retail sales forecasting,” Robotics and Computer- Integrated Manufacturing, vol. 34, pp. 151–163, 2015.

N. S. Arunraj, D. Ahrens, and M. Fernandes, “Application of sarimax model to forecast daily sales in food retail industry,” International Journal of Operations Research and Information Systems, vol. 7, no. 2, pp. 1–21, 2016.

G. Papacharalampous, H. Tyralis, and D. Koutsoyiannis, “Predictability of monthly temperature and precipitation using automatic time series forecasting methods,” Acta Geophysica, vol. 66, no. 4, pp. 807–831, 2018.

S. Thomassey and A. Fiordaliso, “A hybrid sales forecasting system based on clustering and decision trees,” Decision Support Systems, vol. 42, no. 1, pp. 408–421, 2006.

Application of Machine Learning Model and Hybrid Model in Retail Sales Forecast by Haichen Jiang, Jiatong Ruan, Jianmin Sun in 2021 IEEE 6th International Conference on Big Data Analytics

An Online Retail Prediction Model Based on AGA -LSTM Neural Network by Keqiao Chen in 2020 2nd International Conference on Machine Learning, Big Data and Business Intelligence(MLBDBI)

R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.