PREDICTING SALES OF MARKETPLACE USING XGBOOST ALGORITHM BASED ON PRODUCT SALES

MINI PROJECT REPORT

Submitted by

SRI SAI B	**210701258**
VIMAL K B	**210701309**

in partial fulfillment for the award of the degree of

BACHELOR OF ENGINEERING

in

COMPUTER SCIENCE AND ENGINEERING

RAJALAKSHMI ENGINEERING COLLEGE, CHENNAI

ANNA UNIVERSITY:: CHENNAI 600 025

APRIL 2024

RAJALAKSHMI ENGINEERING COLLEGE,

CHENNAI

BONAFIDE CERTIFICATE

Certified that this Report titled "**Predicting sales of Marketplace using XGboost Algorithm based on Product Sales**" is the bonafide work of **"Sri Sai B (210701258), Vimal K B (210701309)"** who carried out the work under my supervision. Certified further that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

SIGNATURE

Rahul Chiranjeevi. V
Assistant Professor,
Department of Computer Science and Engineering,
Rajalakshmi Engineering College,
Chennai – 602015

Submitted to Mini Project Viva-Voce Examination held on _____

Internal Examiner **External Examiner**

ABSTRACT

In today's business environment, it is difficult to predict the future sales of a product. It is a big task in the business world. In our proposed system it discusses the prediction model that will be able to predict sales of a particular product in the market with the help of numerous datasets. This research was for drawing better-required accuracy from the model compared to the previous models. This paper focuses on the problems that occurred while predicting product sales in Big Mart and how the different types of algorithms were used in training the model. In this project, we propose a model using the Xgboost algorithm for predicting sales of companies and found that it produces better performance and accuracy compared to existing model. This project builds a predictive model and finds out the sales of each product at a particular store. The goal of this project is to improve the satisfaction of the customers and optimize the stock levels for profitability.

ACKNOWLEDGEMENT

Initially we thank the Almighty for being with us through every walk of our life and showering his blessings through the endeavour to put forth this report. Our sincere thanks to our Chairman **Mr. S.MEGANATHAN, B.E, F.I.E.**, our Vice Chairman **Mr. ABHAY SHANKAR MEGANATHAN, B.E., M.S.,** and our respected Chairperson **Dr. (Mrs.) THANGAM MEGANATHAN**, **Ph.D.,** for providing us with the requisite infrastructure and sincere endeavoring in educating us in their premier institution.

Our sincere thanks to **Dr. S.N. MURUGESAN, M.E., Ph.D.,** our beloved Principal for his kind support and facilities provided to complete our work in time.

We express our sincere thanks to **Dr. P. KUMAR, Ph.D.**, Professor and Head of the Department of Computer Science and Engineering for his guidance and encouragement throughout the project work. We convey our sincere and deepest gratitude to our internal guide, _____Professor, Department of Computer Science and Engineering. Rajalakshmi Engineering College for his valuable guidance throughout the course of the project.

Karan Balaji R S-210701105
Kanaga Shanmugam P-210701103
Venkatesh V-210701520

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

LR	Linear Regression
RFA	Random Forest Algorithm
XGB	XGboost Algorithm
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error

CHAPTER 1
INTRODUCTION

1.1 GENERAL

In today's business environment, it is difficult to predict the future sales of a product. It is a big task in the business world. Machine learning is the trending technology in the world right now, it works with numerous data sets. Retailers can predict sales with the help of these datasets available. The dataset contains the previous sales information over some time. This research focuses on leveraging machine learning techniques to predict sales trends and optimize their strategies accordingly. Marketplace offers all kinds of products to customers in a wide range. It is a big task to efficiently maintain and manage their inventory and get the right product to their inventory which customers like to buy more.

1.2 OBJECTIVE

The main objective of our project is to predict the future sales of the product accurately. The issues behind predicting accurately overstocking or stockouts. It showed adverse effects on the revenue of the business and customer satisfaction. The main objective of the project is to develop a machine learning model capable of predicting the particular product's future sales in the Big Mart. By getting the previous sales of the various Products available in the Marketplace. Location details of the Marketplace and trending products in the market. The model is going to predict more accurately compared to the previous models.

1.3 EXISTING SYSTEM

Mostly several regression models were used to predict the sales of the market. In the paper [1] by N Malik, and K Singh in 2020 the proposed solution of this paper uses 'Big Mart' as the dataset. The Algorithms used in this system were LR, DT, and RF algorithms. Here python is used as a programming language and Jupyter Notebook is used as a tool. This Existing Solution[1] is mainly used to predict the future sales of the company with a maximum accuracy of 60.8% using the Random Forest Algorithm. In this paper[16] several algorithms were used to predict the accuracy of the system, In that RF algorithm produces maximum accuracy.

1.4 PROPOSED SYSTEM

Our proposed model predicts the Market sales of the company and also uses XGboost algorithm to enhance the accuracy of the system. The trained model is used to predict the sales of the company and also helps in early prediction of the future sales of that company. By using this prediction of Market sales we can improve the satisfaction of the customers and optimize the stock levels for profitability.

CHAPTER 2

LITERATURE SURVEY

In numerous studies regarding the sales prediction of Big Marts, the existing systems have less accuracy. In this research, my base paper is **[1]** by N Malik, K Singh. In the paper, they mainly focus on "LR, RF algorithms".The model was trained using these both algorithms, the accuracy of the model is 58 percent.

Another paper in my research study **[3]** by HV Ramachandra, this paper mainly focuses on the "RF algorithm".The model was trained using this algorithm, and the accuracy of the model is 63.3 percent.

Another paper in my research study**[19]** N Malik, K Singh, in this existing paper they were used a Random forest algorithm to predict the sales of the market and identify customer patterns.

The paper**[11]** by P Kaunchi, T Jadhav, and Y Dandawate, in this existing paper they used a Convolutional neural network to train the model and this trained model were used to predict the future sales of the company.

In this paper[**17]** D Irfan, X Tang, V Narayan, and PK Mall, in this existing paper they were used the TOR method to predict the Quality food sales of the mart.

For the business transformation, we referred**[18]** by P Majumdar, S Mitra, the main motto of this existing paper is to create dynamic platform for market forecasting for future sales prediction.

The study of these papers suggested I use "LR and RF algorithms". Another paper in my research **[5]** by, this paper discusses the "XGB algorithm ".The model's accuracy is better compared to other models using different algorithms.

This paper helped me to choose the "XGB algorithm" for my project. **[4]** by S Raizada and JR Saini, in this research, they compared algorithms such as LR, DT, and SVM. The model was trained using each of these algorithms.

In conclusion, the research says using LR and DT algorithms together helps to build a good model.**[6]** by G Behera and N Nain, this research mainly focuses on inventory management of stores. In the model, they have used the GSO algorithm to train it.

Another paper **[7]** by CK Suryadevara, the research was on selecting the best algorithm to drive greater accuracy from it. This research says that advanced algorithms can give good results. With the help of these existing models, I have chosen "LR, RF, and XGB algorithms" for my project. These algorithms give good accuracy.I have decided to develop a model using these algorithms.

CHAPTER 3
SYSTEM DESIGN

3.1 DEVELOPMENT ENVIRONMENT

3.1.1 HARDWARE SPECIFICATIONS

This project uses minimal hardware but in order to run the project efficiently without any lack of user experience, the following specifications are recommended

Table 3.1.1 Hardware Specifications

PROCESSOR	Ryzen 5
RAM	8GB
GPU	Amda Radeon Graphics
HARD DISK	6GB
PROCESSOR FREQUENCY	1.5 GHz or above

3.1.2 SOFTWARE SPECIFICATIONS

The software specifications in order to execute the project has been listed down in the below table. The requirements in terms of the software that needs to be pre-installed and the languages needed to develop the project has been listed out below.

Table 3.1.2 Software Specifications

GUI	Tkinter, joblib
LANGUAGE USED	Python
FRAMEWORKS	Pytorch, sklearn, seaborn,Tensor Flow
SOFTWARES USED	Visual Studio, Jupyter Notebook, Google Colab

3.2 SYSTEM DESIGN

3.2.1 ARCHITECTURE DIAGRAM

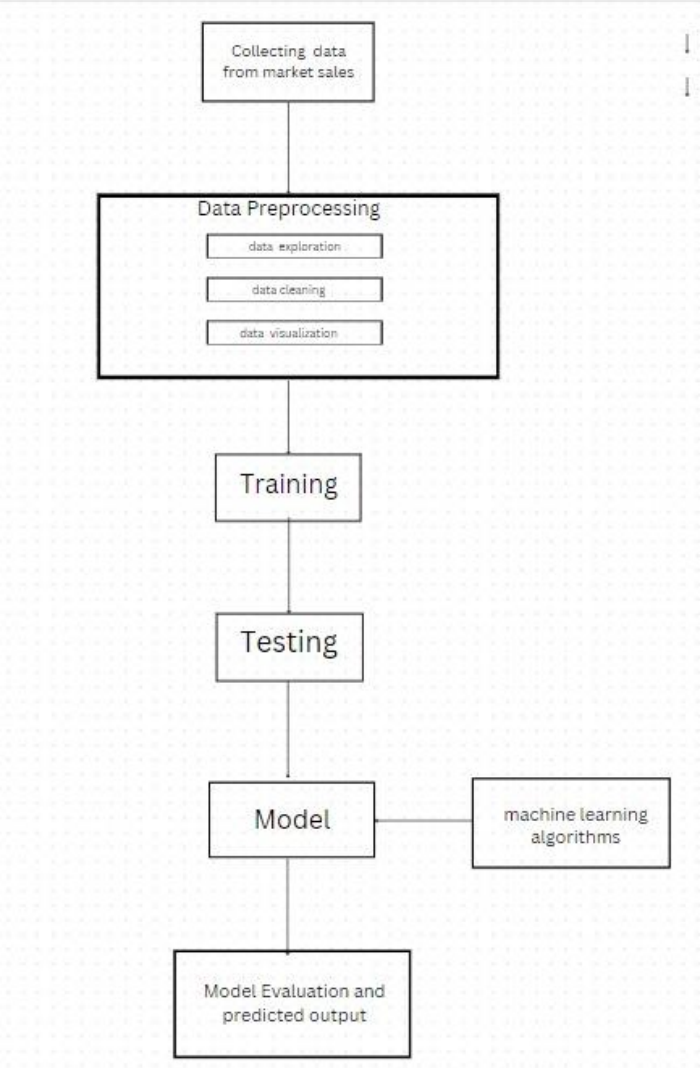

Fig 3.2.1 Architecture Diagram

CHAPTER 4
PROJECT DESCRIPTION

4.1 MODULE DESCRIPTION

4.1.1 DATASET DESCRIPTION:

Our proposed solution uses Market sales reports as a dataset. This dataset is collected from an Internet source. This is an open source platform where most of the users are allowed to choose the dataset to build and evaluate with their training and testing data. The dataset consists of 12 columns with 8723 products. It contains details about the market type, product ID, MartID, Outlet sales, Item MRP, and Location of the mart. The dataset is normalized and well-organized for the prediction.

4.1.2 DATA PREPROCESSING:

For this proposed solution data preprocessing plays an important role in preparing the dataset for the prediction. Here we showed how to preprocess the data set effectively. In our System during this stage, valuable data has been pulled from the dataset. That is, attempting to discern the knowledge derived from theories vs accessible data. In our Proposed Solution Univariate Analysis and visualization of the correlation matrix have been done.

4.1.3 SPLITTING TRAINING AND TESTING DATA:

In our Proposed solution Two different datasets are not imported for training and testing to prevent overfitting. Splitting is thus completed inside a single dataset. The information required to train the model is contained in the training dataset. Test datasets are those that have the potential to forecast test results.

4.1.4 TRAINING MODEL:

After completing all the previous phases, This dataset is ready to build models. In our proposed Solution we use the XGB algorithm and also other machine learning techniques such as LR, RF Algorithm.

4.1.5 PREDICTING SALES:

After the model was trained using the trained dataset, we conclude that the sales of the Product from the particular store have been predicted by using the Machine learning Model.

CHAPTER 5

IMPLEMENTATION AND RESULTS

5.1 IMPLEMENTATION:

To implement a our proposed model we use the XGBoost algorithm in Python, start by importing necessary libraries like pandas for data manipulation, sklearn for preprocessing and metrics, and xgboost for the model. Load the dataset and perform data cleaning by handling missing values and encoding categorical variables. Split the data into training and testing sets. Create an instance of the XGBoost regressor, configure hyperparameters, and train the model on the training data. Evaluate the model's performance on the test set to assess accuracy. Finally, use the trained model to predict sales for new data.

5.2 RESULTS:

5.2.1 TRAINING AND ACCURACY GRAPH:

Thus our model is tested against the test data and the testing and training accuracy graph has been plotted. The training and testing accuracy of the model is plotted in the format of a line graph with Actual_Price x-axis and predicted sales of products in the y-axis, where the blue line indicates the Training accuracy and the Red line indicates Testing accuracy in the below figure.

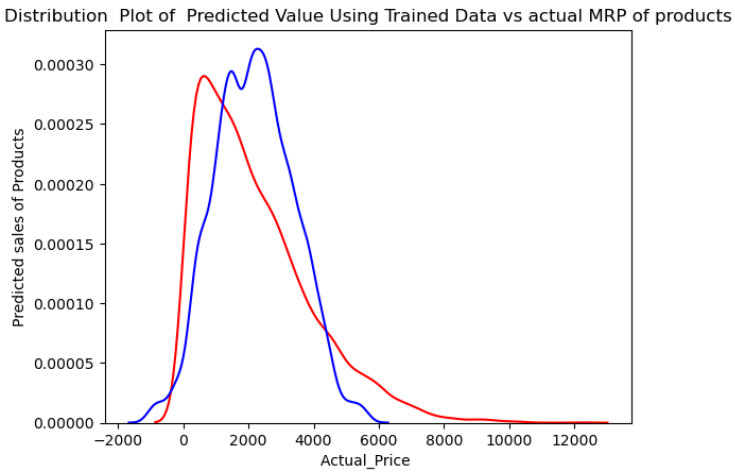

Fig. 5.2.1 Training and accuracy graph

5.2.2 PERFORMANCE ANALYSIS:

For performance analysis, we can go and look for the accuracy value of the different algorithms performed and check for which algorithm gives us the best performance with higher accuracy.

Linear regression Algorithm:

```
# Model
model = LinearRegression(normalize=True)

# Fit
model.fit(X_train, y_train)

# Predict
y_predict = model.predict(X_test)

# Score Metrics for Regression:
LR_MAE = MAE(y_test, y_predict)
LR_MSE = MSE(y_test, y_predict)
LR_R_2 = R2(y_test, y_predict)
print(f" Mean Absolute Error: {LR_MAE}\n")
print(f" Squared Mean Squared Error: {np.sqrt(LR_MSE)}\n")
print(f" R^2 Score: {LR_R_2}\n")

# Cross Validation Score check
cross_val(LinearRegression(),X,y,5)
```

```
Average LinearRegression() score: 0.4972
```

Fig 5.2.2.1 Accuracy Using LR Algorithm

Random Forest Algorithm:

```
from sklearn import metrics
from sklearn.ensemble import RandomForestRegressor

rf = RandomForestRegressor(n_estimators=400,max_depth=6,min_samples_leaf=100,n_jobs=4)

rf.fit(X_train,y_train)

rf_accuracy = round(rf.score(X_train,y_train)*100)

print(f" Accuracy Random Forest: {rf_accuracy}\n")
```

```
Accuracy Random Forest: 60
```

Fig 5.2.2.2 Accuracy Using RFA Algorithm

XGboost Algorithm:

```
y_pred = model.predict(X_train)
y_pred
```

```
array([1075.3064 , 2197.1404 , 6112.4404 , ..., 1969.081  ,  518.98914,
       3563.9968 ], dtype=float32)
```

```
model.score(X_train, y_train)*100
```

```
81.71593499623266
```

Fig 5.2.2.3 Accuracy Using XGboost Algorithm

5.2.3 MARKET SALES PREDICTION:

Thus we conclude that the sales of the Product from the particular store have been predicted by using the Machine learning Model.

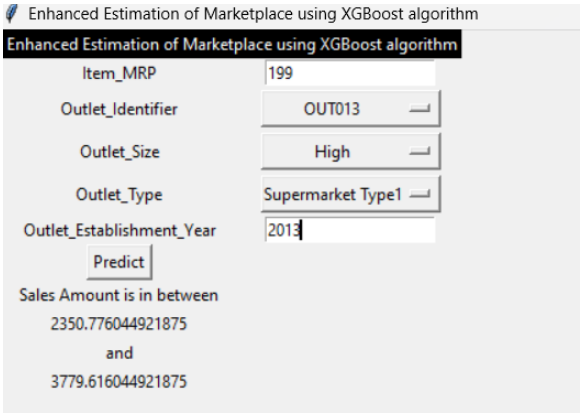

Fig 5.2.3.1 Predicting sales of Market products

CHAPTER 6

CONCLUSION AND FUTURE ENHANCEMENTS

6.1 CONCLUSION

Thus the proposed approach accurately detects and predicts the future sales of the market. By using techniques such as LR, RF, and XGB, we can analyze the location of the Market, store type and predict future sales. From the proposed approach it was found that for the model evaluation XGB algorithm is more accurate than other models and gives the predicted outcomes. This project achieves an accuracy rate of 81% when compared with the existing model[1] it produces the accuracy rate of 56% using Random forest algorithm but in our proposed model we use XGboost algorithm to increase the performance of the model by enhancing its accuracy so that there will be an enhanced estimation of the sales of a company. By using this prediction of Market sales we can improve the satisfaction of the customers and optimize the stock levels for profitability.

6.2 FUTURE ENHANCEMENTS

- **Advanced Algorithms:**

  Neural Networks: Experiment with deep learning models, such as LSTM (Long Short-Term Memory) networks, which are well-suited for time series forecasting.

  Automated Machine Learning (AutoML): Use AutoML tools to automate the model selection and hyperparameter tuning process.

- **Evaluation Metrics**: Use additional evaluation metrics such as MAE (Mean Absolute Error), MAPE (Mean Absolute Percentage Error), and R^2 score for a comprehensive assessment of the model performance in model evaluation and validation.

REFERENCES

[1] N Malik, K Singh." Sales prediction model for Big Mart", 2020.

[2] K Punam, R pamula, PK Jain. "A two-level statistical model for big mart sales prediction", 2018.

[3] HV Ramachandra, G Balaraju, A Rajashekar, H Patil," Machine learning application for black Friday sales prediction framework", 2021.

[4] S Raizada, JR Saini "Comparative analysis of supervised machine learning techniques for sales forecasting ", Journal of Advanced Computer Science, 2019.

[5] V Upadhyay, D Rathod," Location-Based Crime Prediction Using Multiclass Classification Data Mining Techniques",2022.

[6] G Behera, N Nain. "GSO based future sales prediction for big mart", Conference on Signal-Image Technology, 2019.

[7] CK Suryadevara. " Predictive Analysis for Big MartSales using Machine Learning Algorithms" 2020.

[8] P Kaunchi, T Jadhav, Y Dandawate. "Future sales prediction for Indian products using convolutional neural network-long short term memory", 2021.

[9] MA Scarpati, A Krowitz, M Tapp," Using machine learning to predict retail business volume", US Patent 11,068,916, 2021.

[10] PS Smirnov, VA Sudakov."Forecasting new product demand using machine learning" Journal of Physics: Conference, 2021.

[11] P Kaunchi, T Jadhav, Y Dandawate, "Future sales prediction for Indian products using convolutional neural network-long short term memory",2021

.

[12] D Irfan, X Tang, V Narayan, PK Mall . 'Prediction of quality food sale in mart using the AI-based TOR method", 2022.

[13] M. Singh, B. Ghutla, R. Lilo Jnr, A. F. S. Mohammed, and M. A. Rashid, "Walmart's Sales Data Analysis Big Data Analytics Perspective," 2022 4th (APWC on CSE), Mana Island, Fiji, 2022.

[14] S Malik, M Khan, MK Abid, N Aslam. "Sales Forecasting Using Machine Learning Algorithm in the Retail Sector", Journal of Computing & Biomedical,2024.

[15] I Daulat Desale, "Forecasting Using Machine Learning Algorithm", 2024

[16] A Sreelakshmi, N Padhy, "An optimized approach towards increasing the sale rate in a Grocery Mart by using Association Rule Mining Approaches", 2024.

[17] D Irfan, X Tang, V Narayan, PK Mall," Prediction of quality food sale in mart using the AI-based TOR method", Journal of Food,2022.

[18] Business Transformation Using Big Data Analytics and Machine Learning P Majumdar, S Mitra -,2022.

[19] N Malik, K Singh." Sales prediction model for Big Mart", 2020.