PREDICTING SALES OF MARKETPLACE USING XGBOOST ALGORITHM BASED ON PRODUCT SALES

V Karthick, Associate Professor
Department of CSE
Rajalakshmi Engineering College
Chennai, India
vkarthick86@gmail.com

Sri sai B, UG Student
Department of CSE
Rajalakshmi Engineering College
Chennai, India
210701258@rajalakshmi.edu.in

Vimal K B, UG Student
Department of CSE
Rajalakshmi Engineering College
Chennai, India
210701309@rajalakshmi.edu.in

ABSTRACT - This paper 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 models[1]. 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.

Keywords: Marketplace, sales, Linear Regression, Random forest Algorithm, XGboost Algorithm

I.INTRODUCTION

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.

Previous trained models failed 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.

This model is going to help to increase product sales, customer satisfaction and manage the inventory, and decrease overstocking and stockouts. In the outcome of the project, it expands the body of research on sales prediction in the retail industry by effectively using these

machine learning models, and the Organization can obtain a competitive advantage, and evolve to the latest trend.

I.1 LITERATURE REVIEW

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. In another paper in my research study**[19]** N Malik, and K Singh, in this existing paper they used a RanForestrest 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 was 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 of food sales of the mart. For the business transformation, we referred to **[18]** P Majumdar and S Mitra, the main motto of this existing paper is to create a 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.**[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.

II. MATERIALS AND METHODS

The Dataset that has been used for this paper consists of 12 columns 8723 products namely Item_identifier, Item_weight, Item_visibility, Item_Fat_Content, Item_type, Item_MRP, Outlet_Identifier, Outlet_Establishment_Year, Outlet_Size, Outlet_Location_Type, Item_Outlet_Sales, Outlet_Type. The Item_Identifier column gives a unique product ID. The Item_weight column says the weight of the product. The Item_Fat_Content column tells about whether the product has low fat or not. The Item_Visibility column gives the percentage of the product, and how much quantity it is available in the particular area. The Item_Type column says which product belongs to which category. The item_MRP column gives the Market Price for the product. The Outlet_Identifier gives the Unique store ID. The Outlet_Establishment_Year tells about which

year the store was established. The Outlet_Size column tells about the size of the size based on the ground area. The Outlet_Location_Type contains the type of area in which the Store was located. The Outlet_Type column tells about whether the store is a small store or any type of market. The Item_Outlet_Sales column gives the sales of the product from this column the outcome is predicted.

HARDWARE REQUIREMENTS

A laptop or desktop computer with
8 GB RAM
QUAD-CORE PROCESSOR

SOFTWARE REQUIREMENTS

Jupyter Notebook/COLAB
Python
Web Browser(Chrome, Edge

III. 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 Proposed 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.

IV.PROPOSED SYSTEM

Our proposed solution uses LR, RF, and XGB algorithms to predict the accuracy and sales of the system. In this System, XGB was found to be more suitable with an accuracy of 81% rather than the RF algorithm. A detailed explanation of the proposed method is discussed in the following Section:

4.1 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.

Table 4.1.1 Market Report Dataset Classes

Variable	Description
Item_Identifier	Unique product ID
Item_Weight	Weight of product
Item_Fat_Content	Whether the product is low fat or not
Item_Visibility	The % of total display area of all products in a store allocated to the particular product
Item_Type	The category to which the product belongs
Item_MRP	Maximum Retail Price (list price) of the product
Outlet_Identifier	Unique store ID
Outlet_Establishment_Year	The year in which store was established
Outlet_Size	The size of the store in terms of ground area covered
Outlet_Location_Type	The type of city in which the store is located
Outlet_Type	Whether the outlet is just a grocery store or some sort of supermarket
Item_Outlet_Sales	Sales of the product in the particulat store. This is the outcome variable to be predicted.

4.2 MODEL ARCHITECTURE:

Our proposed model predicts the Market sales of the company and also uses the 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 the early prediction of the future sales of that company.

Table 4.2.1 Model architecture of the proposed system

4.2.1 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.

4.2.2 DATA EXPLORATION:

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.

4.2.3 DATA CLEANING:

Here, first, we identify and Handle the Missing values so that the performance of the Model will be enhanced. We can use several strategies for handling the missing values in the dataset. As was noted in the preceding section, there are missing values for the characteristics Item_Weight and Outlet_Size. In our work, the mode of the property for Outlet Size is used to replace missing values, and the mean of the attribute value for Item Weight is used to replace missing values.

4.2.4 DATA VISUALIZATION:

In our proposed solution, Once the above phases are completed the analysis of data would be done, In the analysis phase seaborn and Matplotlib are used for data visualization. Countplot is used for counting and categorizing the data.

4.2.5 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.

4.2.6 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, and RF Algorithm.

LR algorithm :

LR is a statistical method used to determine the relationship between a dependent variable and one or independent variables. LR aims to find the Best-fitting straight line; it describes the relationship between the predictor values and target values.

RF algorithm :

It is an ensemble learning method that combines different decision trees to improve the predicting accuracy and decrease overfitting. It uses both classification and regression.

XGB algorithm :

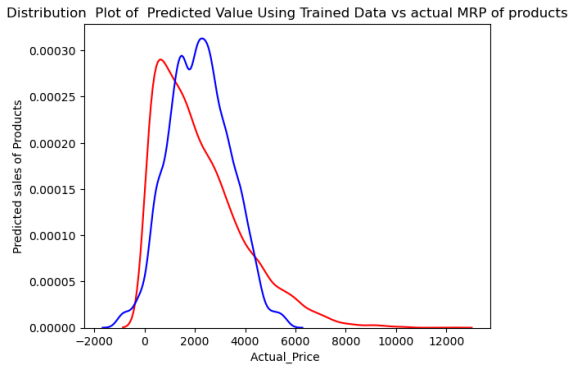
It is one of the ensemble techniques. It is a more scalable and more efficient gradient-boosting algorithm. It is supervised learning.

V RESULT AND DISCUSSIONS

5.1 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.

Figure 5.1.1 Accuracy Graph



5.2 CONFUSION MATRIX:

The proposed model is evaluated and the confusion matrix for the trained model is attached in the below figure

Figure 5.1.2 Confusion Matrix

5.3 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.

LR 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
```

Figure 5.3.1 Accuracy Score of Trained model using LR algorithm.

RF 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
```

Figure 5.3.2 Accuracy Score of Trained Model Using RF Algorithm

XGB Algorithm:

```
from xgboost import XGBRegressor

model = XGBRegressor(n_estimators = 100, learning_rate=0.05)
model.fit(X_train, y_train)
```

```
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
```

Figure 5.3.3 Accuracy Score of Trained Model Using XGB Algorithm

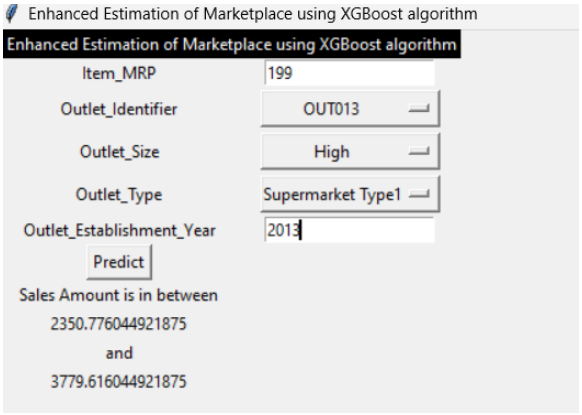
Fig 5

From the Performance analysis, we conclude that the XGB algorithm is more accurate and gives a predicted result

5.4 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.

Table 5.4.1 Prediction of Market Sales



VI 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, and store type and predict future sales. From the proposed approach it was found that for the model evaluation, the 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 an accuracy rate of 56% using the Random forest algorithm but in our proposed model we use the 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.

VII 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.