A Survey of the Big Mart Sales Prediction using Machine Learning

[1]Shruti Mahishi, [2]Sharvari Mhatre, [3]Shreya Bhagwat, [4]Dr Vidya Chitre

[1,2,3]UG Student, [4]Assistant Professor, Vidyalankar Institute of Technology, Mumbai, India,

[1]mahishishruti@gmail.com, [2]sharvarimhatre77@gmail.com, [3]shreyabhagwat04@gmail.com, [4]vidya.chitre@vit.edu.in

Abstract - Shopping malls and Big marts keep a record of sales data of their customers. This helps them to analyse the pattern and make predictions. Data warehouses typically store a large number of customer data and item attributes pertaining to each individual customer. This paper gives an overview of how various people have been able to predict big mart sales. We shall also see which machine learning algorithms and methodologies they have used for the same.

Keywords —Accuracy, Big mart, Machine learning, Regression, Sales prediction, XG Boost

I. INTRODUCTION

In the era of a digitally connected world many shopping centres, grocery stores, and sales outlets need to grasp their customer's demands beforehand to avoid the dearth of goods in any season. This includes the process of noting the sales of a specific item and its increase or decrease in demand. Sales prediction is a crucial part of contemporary business intelligence. There lies a drawback, particularly within the case of lack of knowledge, missing information, and the presence of outliers. The Sales Prediction model predicts the sales of products sold at various retailers in several cities of an enormous sales outlet of a Company. Because of the volume of merchandise, maintenance of exponential growth of their products by hand becomes cumbersome for the retailers. Predicting the proper demand for a product is important for the sellers in terms of management of stock, time, and cash. Companies tend to have restricted time or have to be compelled to sell their merchandise to avoid any unwanted consequence due to storage and cash restrictions. Therefore, the demand for a product depends on several factors like worth, popularity, time, outlet type, outlet location, etc.

Demand prediction is additionally closely associated with Sales revenue. If sellers store way more products than they need, this could result in a surplus. On the contrary, storing fewer products to avoid wastage of inventory may raise less revenue once the demand increases. Thus, Sales Prediction helps the businesses to store merchandise in accordance with the expected sales for a particular region or an outlet kind. Thus, providing corporations with the expected sales for their merchandise helps corporations to formulate a correct business model that in turn guides them to arrange and dispatch additional batches of their product expeditiously for better revenue generation. In this paper, we focus on the information and other related research works available regarding the considered subject matter and thus plan on building a more accurate model that will overrule the flaw reviewed in other works.

II. LITERATURE SURVEY

Rohit Sav, Pratiksha Shinde, Saurabh Gaikwad in [1] have used XGBoost Regressor for the implementation of their predictive model. In Order to carry out their idea, they followed a procedure that includes gathering of data, then cleaning it, which then led to feature Engineering to clear all the nuances, building the model before it is led to the last model testing phase. They observed a higher accuracy rate while using XGBoost Regressor in comparison with other algorithms. Therefore, they pointed out that the current model should be preferred over other models.

Inedi Theresa, Dr Venkata Reddy Medikonda, K.V. Narasimha Reddy in [2] talked about executing their problem statement with the help of Exploratory Machine Learning. To achieve their expected result, they followed a ruled-out procedure which includes data collection, Hypothesis Generation to help them analyse the bugs, which then follows with data exploration and cleaning, leading to the use of four predicting models- Linear Regression, Decision Tree Regression, Ridge Regression, and Random Forest model. They concluded that this method led them to a better prediction as compared to that of the single model prediction technique.

Kadam H., Shevade R., Ketkar, P. and Rajguru in [3] proposed a model that essentially works with multiple linear regression and a Random Forest algorithm. This model was expected to forecast accurate sales prediction for The Big Mart and to confined with that certain data set was used which comprises of Item_Identifier, Item_Weight, Item_Fat_Content, Item_Visibility, Item_Type, Outlet_Identifier etc.

DOI : 10.35291/2454-9150.2022.0003

Gopal Behera and Neeta Nain in [4] Grid Search Optimization (GSO) Based Future Sales Prediction for Big Mart have proposed GSO technique to optimize parameters and predict future demand of sales in retail-based companies. Hyperparameter tuning is performed and sales forecasting is done by XG Boost techniques.

Kumari Punam, Rajendra Pamula and Praphula Kumar Jain in [5] A Two-Level Statistical Model for Big Mart Sales Prediction have used a two-level approach for prediction of sales of the product which yields better performance and efficiency. The two-level approach involves the stacking of algorithms one above the other. The bottom layer consists of one or more than one learning algorithm and the top layer consists of one learning algorithm. This methodology of two-level statistical models outperformed the other single model predictive techniques and gave better predictions to the big mart dataset.

Ranjitha P and Spandana M in [6] Predictive Analysis for Big Mart Sales Using Machine Learning Algorithms have built the predictive model using Xgboost, Linear regression, Polynomial regression, and Ridge regression techniques for predicting the sales of big mart.

Bohdan M. Pavlyshenko in [7] has used machine learning techniques to implement his model. This Paper talks about the effect of machine generalization. The effect talked about here can be used when there is less data in the system because of a new product or new outlet. Here, a different outlook was set on the models by using a stacking technique to build regression. The end product of this model proves to be more useful for better performance for sales time series forecasting.

Nikita Malik, Karan Singh in [8] discuss the basis of machine learning and the procedure associated with it to get the result. After the execution of the models, it displays a correlation among different attributes and the outlet sizes showing different rates of sales, suggesting that outlets of a certain size will have the same rate of success in sales as they might follow a similar pattern.

Gopal Behera and Neeta Nain in [9] have proposed different algorithms such as linear regression, decision tree algorithm and xgboost algorithm. Xgboost was highly recommended because of its accuracy. It also mentioned that MAE and RMSE were kept at a low limit for better performance as compared to other existing models.

Archisha Chandel, Akanksha Dubey, Saurabh Dhawale, Madhuri Ghuge in [10] essentially reviewed the five-step procedure being carried out to acquire the result with higher accuracy. The procedure included the collection and division of testing and training data. The data goes through univariate and bivariate analysis as well. In the later stages of the procedure, the data is pre-processed, modified and then transformed by using different algorithms for a better result.

III DATASET OVERVIEW

VARIABLE	DESCRIPTION
Item_Identifier	Unique product ID
Item_Weight	Weight of product
Item_Fat_Content	to check If product is low fat or not
Item_Visibility	The % of total display area of all product 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	to check if the outlet is just a grocery store or some sort of supermarket
Item_Outlet_Sales	Sales of the product in the particular store This is the outcome variable to be predicted.

Table 1: Description of the attributes

IV. METHODOLOGY

1. Data cleaning and pre processing

Values in the dataset need to be replaced with relevant values, the missing values will be replaced with appropriate numerical or categorical values depending on the type of feature. The unnecessary information in the dataset will be discarded.

2. Data modeling

Based on patterns and features, models will be created about the working and process of the project and how the sources of data will fit together and flow into one another.In our work, we will be proposing a model using the Xgboost algorithm and comparing it with other machine learning techniques like Linear regression, Ridge regression, Decision tree.

3. Data prediction

Machine learning models are trained and evaluated using the data. These are then implemented on the pre-processed dataset.Some Models used for the prediction are : Linear Regression, Lasso Regression ,Ridge Regression Model ,XGBoost Regressor and Decision Trees.

4. Data visualization

Data Analysed is further picturized for customers and the admin to analyse and take appropriate decisions on the subject matter. The results are then presented to the client in a way that answers the challenges set for the project allowing the client to implement the findings.

V. CONCLUSION

After getting an analysis of all the papers and their conclusions regarding all the models they have implemented we can conclude that big mart sales prediction remains a wide topic for research. It is advised that one should select an algorithm based on the dataset provided and analyse how well different algorithms respond to the dataset. Also, the more the number of attributes and number of rows in the dataset, the more is the accuracy of the implementing model. Hence it is advised to perform the algorithm which provides you with an optimum accuracy according to the chosen dataset.

REFERENCES

[1] Rohit Sav, Pratiksha Shinde, Saurabh Gaikwad (2021, June). Big Mart Sales Prediction using Machine Learning. *2021 International Journal of Research Thoughts (IJCRT).*

[2] Inedi. Theresa, Dr. Venkata Reddy Medikonda, K.V.Narasimha Reddy. (2020, March). Prediction of Big Mart Sales using Exploratory Machine Learning Techniques. *2020 International Journal of Advanced Science and Technology (IJAST).*

[3] Heramb Kadam, Rahul Shevade, Prof. Deven Ketkar , Mr. Sufiyan Rajguru (2018). A Forecast for Big Mart Sales Based on Random Forests and Multiple Linear Regression. *(IJEDR).*

[4] Gopal Behere, Neeta Nain (2019). Grid Search Optimization (GSO) Based Future Sales Prediction for Big Mart. *2019 International Conference on Signal-Image Technology & Internet-Based Systems (SITIS).*

[5] Kumari Punam , Rajendra Pamula , Praphula Kumar Jain (2018, September 28-29). A Two-Level Statistical Model for Big Mart Sales Prediction. 2018 International conference on on Computing, Power and Communication Technologies

[6] Ranjitha P, Spandana M. (2021). Predictive Analysis for Big Mart Sales Using Machine Learning Algorithms. *Fifth International Conference on Intelligent Computing and Control Systems (ICICCS 2021).*

[7] Bohdan M. Pavlyshenko (2018, August 25). Rainfall Predictive Approach for La Trinidad, Benguet using Machine Learning Classification. *2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP).*

[8] Nikita Malik, Karan Singh. (2020, June). Sales Prediction Model for Big Mart.

[9] Gopal Behere, Neeta Nain. (2019, September). A Comparative Study of Big Mart Sales Prediction.

[10] Archisha Chandel, Akanksha Dubey, Saurabh Dhawale, Madhuri Ghuge (2019, April). Sales Prediction System using Machine Learning. *International Journal of Scientific Research and Engineering Development.*

DOI : 10.35291/2454-9150.2022.0003