Big Mart Sales Analysis

Vidya Chitre, Shruti Mahishi, Sharvari Mhatre, Shreya Bhagwat

Abstract**:** ***In the modern era of reaching new lengths of advancement, every company and enterprise are working on their customer demands as well as their inventory management. The models used by them help them predict future demands by understanding the pattern from old sales records. Lately, everyone is abandoning the traditional prediction models for sales forecasting as it takes a prolonged amount of time to get the expected results. Therefore now the retailers keep track of their sales record in the form of a data set, which comprises price tag, outlet types, outlet location, item visibility, item outlet sales etc.***

Keywords: Analysis, Big Mart, Data Science, Machine Learning, Prediction, Regression, XG Boost

I. INTRODUCTION

Customer Satisfaction and keeping up with the demand for products is very important for any store to survive in the market and to compete with other stores. And these two can only be achieved when you have a future demand figure for coming up with new plans for a flourishing business With an increased population, the number of stores and shopping malls is also increasing creating competition between different enterprises for bigger sales and popularity. Along with grocery shops and stores, even enterprises need an analysis to check about the patterns and predict future sales. Many businesses and enterprises keep track of the statistical data of their products so as to analyze their future demand in the market. The stored statistical data consists of the amount of items sold and its categories and other attribute details to provide trends and patterns to the organization regarding their supply, to grow their business, and improve their sales strategies and this might come in handy in the near future or seasonally in-order to create short-term discount offers to attract more customers towards their brand; the previous data is important as it helps in the management of logistics, inventory, and transport services, etc. And to carry out all these tasks we will use machine learning algorithms like the random forest, linear regression, decision trees, and XGBoost regressor. The aim of our project is to build a machine learning model which can predict the sales of a product and understand its patterns and trends which is an important part of a big mart's management.

Manuscript received on March 12, 2022.
Revised Manuscript received on March 21, 2022.
Manuscript published on April 30, 2022.
* Correspondence Author

Vidya Chitre, Professor, Department of Information Technology, Vidyalankar Institute of Technology, Mumbai (Maharashtra), India.

Shruti Mahishi*, Final Year Engineering Student, Department of Information Technology, Vidyalankar Institute of Technology, Mumbai (Maharashtra), India.

Sharvari Mhatre, Final Year Engineering Student, Department of Information Technology, Vidyalankar Institute of Technology, Mumbai (Maharashtra), India.

Shreya Bhagwat, Final Year Engineering Student, Department of Information Technology, Vidyalankar Institute of Technology, Mumbai (Maharashtra), India.

II. LITERATURE SURVEY

Rohit Sav, Pratiksha Shinde, Saurabh Gaikwad in [1] have implemented predictive models to measure big mart sales. They first cleaned the gathered data and applied the XG Booster algorithm. It was observed that XGBoost Regressor showed the highest accuracy rate when compared with other algorithms. This led them to draw a conclusion for using XG boost for prediction of big mart sales. Inedi. Theresa, Dr.Venkata Reddy Medikonda, K.V. Narasimha Reddy in [2] discusses sales prediction by using the methodology of Exploratory Machine Learning. They carried out the whole process by figuring out proper steps that included a collection of data, thesis generation to efficiently understand bugs, further cleaning and processing the data. The models such as Linear Regression, Decision Tree Regression, Ridge Regression, and Random Forest model were used to predict the outcome of the sales. They concluded that multiple modeling implementation 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 works effectively with multiple linear regression and a random forest algorithm. This model was utilised to forecast big mart sales prediction and with that, a certain data set was used which comprises of Item_Identifier, Item_Weight, Item_Fat_Content, Item_Visibility, Item_Type, Outlet_Identifier etc. Gopal Behera and Neeta Nain in [4] apply the concept of GSO technique to optimize parameters and predict future sales. Their focal point is Retail Based companies. They have also implemented Hyperparameter tuning and forecasted sales using XGBoost techniques. Kumari Punam, Rajendra Pamula and Praphula Kumar Jain in [5] A Two-Level Statistical Model for Big Mart Sales Prediction have devised a two-level approach to predict sales of products that promise to yield better efficiency. It involves stacking up of algorithms wherein the top layer consists of just one learning algorithm and the bottom layer has one or more algorithms placed. This methodology of two-level modeling outperforms the single model predictive technique and results in better predictions of sales. Ranjitha P and Spandana M in [6] Predictive Analysis for Big Mart Sales Using Machine Learning Algorithms have implemented Xgboost, Linear regression, Polynomial regression, and Ridge regression techniques for forecasting sales of big mart. Bohdan M. Pavlyshenko in [7] put forward the perspective of machine generalization. This comes into play mostly when fewer data is available in the system, perhaps with the introduction of new products or new outlets. The special technique of stacking was implemented to build the regression.

This results in better performance and efficiency in sales prediction. Nikita Malik, Karan Singh in [8] implement the concept of machine learning algorithms to reach a conclusion of the problem statement, intended to predict big mart sales.

It displays relations between different attributes and outlet sizes which show variable rates of sales. They draw out the conclusion that a certain size with a similar pattern will have a similar rate of success in sales.Gopal Behera and Neeta Nain in [9] have stated linear regression, decision tree algorithm and xgboost algorithms. Among these models, XG boost displays the highest accuracy, hence proving to be highly recommendable. MAE and RMSE were kept at a low limit for better performance in comparison to other models. Archisha Chandel, Akanksha Dubey, Saurabh Dhawale, Madhuri Ghuge in [10] describes a five-step procedure for the prediction of big mart sales.

Data is segregated as testing and training to obtain effective results. Data undergoes univariate and bivariate analysis. Data is later on pre-processed, modified and then transformed by using different algorithms for better results.

III. DATA AND SOURCES OF DATA

The data which was required for the project is collected through a Kaggle Dataset. The dataset set contains certain attributes like:

Table 1. Data Attributes

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 the 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 outlet 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 particular store This is the outcome variable to be predicted.

IV. EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis is useful to train data to view every dataset so that it can be merged into training and testing data for feature engineering, data visualization, etc. The exploratory analysis includes (two types of analysis i.e., univariate which deals with only 1 attribute and bivariate which deals with two attributes that are conducted on data, to summarize and find patterns in the data. We made a few observations during the analysis, the 'low fat' category is also mentioned as 'LF' and 'Low Fat' also 'reg' and 'Regular' belong to the same category so can be merged as one category. It is also observed that low-fat attribute quantity is double that of other item types. It is also observed that maximum sales are of two types of item type which are Fruits and Snacks.

From the data, we can see that some of the items are not items that are edible even though they are labeled as low-fat and/or regular. So with the help of analysis, the relationship of product weight with sales and item fat content and sales can be observed. A large number of sales are of items with visibility below 0.2.

V. DATA PROCESSING AND METHODOLOGY

A. Data Collection

We have collected the data securely in accordance with an agreed methodology. The procedure for the collected data may differ from client to client and is dependent on the type, quantity, availability and need of data.

B. Data Cleaning and Preprocessing

The collected data is passed through a 'cleaning' process, so as to make sure that the data is segregated properly and identified gaps in the data are filled with the appropriate information, making data compatible and also fixing errors in storage systems which can cause data redundancy.

C. Data Modeling

This is primarily a process in which the given dataset and the objects in it are analyzed to get a clear view of the requirements that may help us support our business model. Based on the analysis on patterns present in the data, models are then created on the established flow of the project.

This flow offers a better assistance in the utilization of the previously agreed upon semi-formal model that showcases the features of the project. It also provides guidance to follow the relation between the data objects and other objects.

D. Data Prediction

Machine Learning prediction models are trained in this process and then later on evaluated using the data. This will then be applied to the preprocessed dataset.

Some of the Models to be used for the prediction are:

- Linear Regression
- Random Forest
- Decision tree
- XG Boost Regressor

E. Data Visualization

Data Analyzed is then further picturized for customers and the admin to reach out conclusions and take effective decisions.

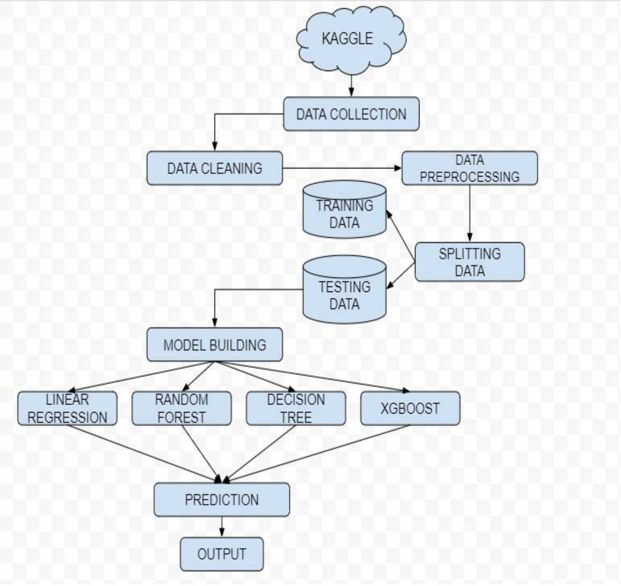

Fig.1.Flow Graph of Proposed System

VI. FEATURE ENGINEERING

This is a process that is carried out to get a finer understanding of given information and patterns in our dataset. Feature Engineering helps to build an algorithm to attain a high accuracy rate.

The prescient ability of a predictive model provides more enhancement to the machine learning procedure by creating new data features. This machine-learning procedure includes cleaning of erroneous values present in the dataset. In our dataset, there were a few flawed values present in the item visibility, item weight, and outlet size.

It was recognised that the item visibility of a few products only had a minimum value of 0 which is not suitable as the items are available for every customer. Therefore, the previous values are replaced with the mean value of the mentioned column.

The same solution is imposed on other missing or inaccurate values. As for outlet size, it is noted that the outlet type 'small' appears only in outlet size- grocery store and tier-2 type, With the help of this information we fill the missing values in the outlet size with the outlet type 'small'. Item weight, the mean value is used to correct the incorrect data.

VII. EVALUATION METRICS

Evaluation metrics are an important part while building an efficient machine learning model. It helps to get suggestive results about the model which can be compared and improved till we achieve good accuracy. Evaluation metrics are a way to concentrate on a model's results and have the ability to distinguish between model outcomes. In our project, we have used the Root Mean Squared Error (RMSE) metric for the evaluation.

VIII. MODEL BUILDING

A. Linear Regression

Linear regression is a supervised machine learning algorithm that is the most commonly used predictive analysis. This Regression model defines the relationship between the dependent as well as the independent variable. The dependent variable is also known as the response variable and can be denoted by X.

The Independent variables are an input factor and are also known as the explanatory or predictor variables and can be denoted as Y in this case. The equation for Linear Regression is as follows:

Y= a+bX RMSE Value: 1067.84

B. Random Forest

Random forest is a Machine Learning algorithm that is used for classification as well as regression problems. This ML algorithm combines multiple decision trees together to create a model that has high accuracy and more stability. Rather than depending on just one decision tree algorithm, the random forest takes the predicted outcome from each tree and also analyses the majority votes of predictions, and predicts the final output.[11]

RMSE Value: 1049.35

C. Decision Tree

A decision tree is a Machine Learning algorithm that is a tree-structured classifier, where the internal nodes represent the features of a dataset, while the branches represent the decision rules and each leaf node represents the outcome. Although it is preferred to solve classification problems, it is used for both classification and regression problems. [12]

RMSE Value: 1032.68

D. XG Booster

XG Boost stands for eXtreme Gradient Boosting. This is an efficient ensemble method of learning which makes use of a gradient boosting framework.

In this model, decision trees are created in the form of a sequence. It is a combination of weak learners to improve prediction accuracy. Suppose in case X, the outcomes are weighed based on the output of the previous instance X-1. The results predicted correctly are given a lower weight and the ones misclassified are weighted higher. [13]

RMSE Value: 1027.68

IX. RESULTS

Table 2. Results of Descriptive Statics of Study Variables

Name	Accuracy	Root Mean Square Error (RMSE) value	Cross validation score (mean)	STD
XGBoost	61.14	1027.68	55.32	0.08
Random Forest	59.49	1049.35	56.67	0.11
Linear Regression	58.05	1067.84	57.36	0.07
Decision Tree	60.76	1032.68	38.99	0.04

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

X. CONCLUSION

We have applied four algorithms XGBoost, Random Forest, Linear Regression and Decision Tree. From the results, we can conclude that among all the four algorithms XGBoost has the highest accuracy of 61.14% when distinguished together. Hence, we can say that XGBoost is the better algorithm for efficient sales analysis. This methodology is primarily used by shopping marts, groceries, Brand outlets etc. The data analysis applied to the predictive machine learning models provides a very effective way to manage sales, it also generously contributes to better decisions and plan strategies based on future demands.This approach is very much encouraged in today's world since it aids many companies, enterprises, researchers and brands for outcomes that lead to management of their profits, sales, inventory management, data research and customer demand.

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 020 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). SalesPrediction 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.
11. https://www.javatpoint.com/machine-learning-random-forest-algorithm
12. https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm
13. A. Chandel, A. Dubey, S. Dhawale and M. Ghuge,;Sales Prediction System using Machine Learning; International Journal of Scientific Research and Engineering Development, vol. 2, no. 2, pp. 1-4, 2019. [Accessed 27 January 2020].

AUTHORS PROFILE

Vidya Chitre is a Professor & DAO at Vidyalankar Institute of Technology. She holds a Phd in Information Technology and has done research in HPC from the University of Mumbai. She enjoys teaching and has been into teaching for 25+ years. Her teaching curve comprises various subjects like Database Management, Data Mining, and Data science amongst others. These subjects form the basis for AI, ML, and analytics in general and we are surrounded by them in today's world. To cope up with the pace of rapidly changing technology, she has organized and attended various FDPs, webinars, and short-term courses. Her primary focus is to make students technically sound and make them industry ready. She Encourages students to showcase their ideas, thereby contributing to the community and the world at large.

Shruti Mahishi is a Final Year Engineering Student of Information Technology at Vidyalankar Institute of Technology, Mumbai**.** Learning new technologies and tools interest her a lot. She has carried out several hands-on projects in data science and worked with BI tools like Power BI. She also has implemented Chabot using IBM Watson tool. She has her blog on wordpress and medium. She excels in public speaking and presents well in front of a large audience. Playing with data and drawing out various visuals to reach effective conclusions interests her a lot. Pursuing this passion, she wishes to continue her work and research in the domain of Analytics and Management.

Sharvari Mhatre is a Final Year Engineering Student, Information Technology, Vidyalankar Institute of Technology**,** Mumbai. Her major interest lies in the domain of Database systems and Web Development. She also enjoys Web scripting. She likes to explore the new technologies trending in the industry and implement them. She is currently learning Python and R language to implement it in her projects. She has previously worked on a python based project using turtle graphics and Tkinter to develop the classic snake game. She aspires to continue her work and research in the domain of Data science and AI. Reading books and articles in free time adds to her knowledge and interpersonal skills.

Shreya Bhagwat is a Final Year Engineering Student, Information Technology, Vidyalankar Institute of Technology, Mumbai.Her interests lie in the domain of data science, cybersecurity and database management systems. She has worked on two IoT based projects and data science projects. She enjoys applying her creative mind in her work and has great team spirit while doing group projects. She aspires to contribute to the society in manifold ways. Listening to Music in free time makes her feel good and optimistic. She has an interest in understanding and learning different languages. She is currently learning Italian language and also plans to learn German and Spanish .