**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **Chapter No.** | **Description** | **Page No.** |
| Chapter 1 | Introduction | **2-4** |
| Chapter 2 | Literature Survey | **5-6** |
| Chapter 3 | Methodology | **7-9** |
| Chapter 4 | Result and Discussion | **10-11** |
| Chapter 5 | Conclusion and Future Work | **12** |
|  | References | **13** |

**Chapter 1**

**Introduction**

* 1. **Introduction**

Big Mart is a massive global network of stores that is facing more competition from other malls and online shopping as multinational supermarkets grow in popularity. Accurate sales forecasting is critical for advertising, retailing, warehousing, and production to remain competitive. The goal is to attract the target audience by delivering personalised and short-term promotions, enabling for accurate forecasting of sales volume for each item in order to properly manage inventory, logistics, and transportation services. Modern retail centres, such as large malls, collect sales data as well as different factors that influence them.

Machine learning algorithms have progressed, giving new tools for organisations to estimate future sales demand. This research focuses on estimating big mart sales and anticipating future demand for various items in various big mart retailers across multiple locations based on past records. This prediction is dependent on two critical factors: having the correct data and extracting relevant insights from it. These insights assist businesses in determining whether to launch new products or remove underperforming ones from the market. Analysing previous situations through sales forecasting assists in making informed judgements regarding client acquisition, funding, strengths, budgeting, and marketing plans for the following year.

Traditional sales forecasting approaches, such as mathematical methods, are time-consuming and incapable of tackling indirect data. Machine learning techniques are used to overcome these constraints. Machine learning algorithms may efficiently handle not only indirect data but also massive datasets. For continuous variables, the accuracy measures Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) are used to evaluate model performance.

Several machine learning algorithms, such as Linear Regression, Multiple Linear Regression, Random Forest, Gradient Boosted Decision Tree, and XGBoost, are used for detailed sales forecasting. A manual approach of this work could result in large errors, significantly affecting organizational management. Instead, machine learning algorithms are used to forecast sales using data mining methodologies like as discovery, data transformation, feature development, model construction, and testing. This comprises pre-processing the raw data in order to find missing data, anomalies, and outliers. The algorithm is then trained using a data warehouse's composite dataset of item attributes, customer data, and inventory management information. This enhanced data enables machine learning algorithms to make accurate predictions.

Machine learning addresses supervised and unsupervised tasks, with classification-type problems frequently contributing to knowledge discovery. It uses regression to create precise predictions about the future, with the goal of developing a self-sufficient system capable of undertaking computations and analysis to produce correct results. Statistical and probabilistic methods are used to transform data into knowledge, with statistical inference relying on sample distributions as a conceptual key.

Current machine learning algorithms are extremely complicated, providing strategies for forecasting long-term sales demand, assisting firms in overcoming budget limits, and optimizing computer programs. Accurate forecasting aids in the development of effective corporate strategies, the identification of profitable markets, and the improvement of overall market understanding. Extensive research is being conducted in the retail industry to properly estimate long-term sales demand.

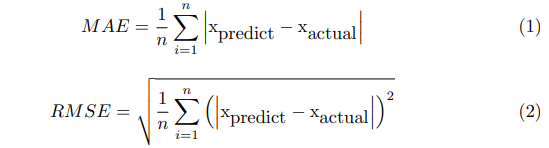


Fig 1.1 MAE and RMSE Formulas

* 1. **Problem Definition**

Due to the escalating competition in the retail industry, numerous malls and Big Mart stores are striving to maintain a competitive edge. In order to achieve this, it is critical to understand the major aspects that have a substantial impact on Big Mart's sales performance and to formulate efficient strategies for optimizing profits. Therefore, the central problem addressed in this research is to develop a robust and dependable predictive model that can effectively identify the influential elements, provide significant insights, and contribute to increased profitability of Big Mart.

* 1. **Objectives**

The central objective of this research is to leverage machine learning algorithms to develop an accurate sales prediction model for Big Mart, optimizing their inventory management and sales strategies. Our project aims at analyzing historical sales data in order to determine the key elements impacting sales patterns. This study facilitates a better understanding of the essential elements that contribute to individual product sales. It evaluates the generated model's performance using actual sales data, ensuring its credibility and identifying potential areas for improvement. The system processes and evaluates data from repositories to discover the best method for sales forecasts with maximum accuracy.

* 1. **Scope of this project**

The main emphasis of this project lies in the examination of historical sales data exclusively from Big Mart stores within a specific region. External economic issues will not be included in this analysis. To achieve this, various machine learning algorithms, comprising linear regression, multiple linear regression, random forest, gradient-boosted decision tree, and XGBoost, will be utilized to extract knowledge from the provided data. The scope of this research is to conduct a thorough investigation and comparison of these machine learning algorithms to enhance decision-making processes concerning sales prediction in BigMart stores.

**Chapter 2**

**Literature Survey**

* 1. **Related Work**

Sales prediction is a good idea for a company to analyze the sale of product. Nikita Malik [1] has discussed sale prediction using machine learning. She used Machine learning algorithms (linear regression, Random Forest, etc.). She has done analysis on a few products and established some correlation between product and store. The accuracy is between 70% to 80%. . Aditi Narkhede [1] has collected dataset of BigMart and has used ML algorithm to find RMSE value. She has done some calculation to find RMSE value and is quite easy to use. Pamula [1] has discussed about flow chart diagram to make us understand things easily. Here he used Machine learning and data mining as well. The accuracy is between 50 to 60%. He used some complex calculation to get the output which is not easy to understand. Saju Mohanan [1] has used data mining techniques and machine learning algorithm for sale prediction. He used decision tree and generalized linear model for prediction. The accuracy of the model is between 60 to 70%. He also has drawn system architecture to make thing simple but the output is in very complex form. Pavan Chatradi [1] has discussed about sale prediction using ML and the XGBoost technique. Here he has followed steps like Data cleaning, Data transformation, Data reduction. The accuracy of this method is above 80%. But the method and result shown are in complex form.

Kadam,et.al [2] have suggested when the prediction for the sales for BigMart was done using the algorithm like random forest and LR for prediction analysis it gave lesser accuracy. So to overcome this problem we can use another algorithm which is XG boost algorithm which not only gives better accuracy but also is more efficient. Makridakis, et.al [3] have suggested predicting methods and applications containing Data Lack and short life cycles. So some data like historical data, consumer-focused markets face uncertain needs, which can be an accurate predictor of outcome. A. Krishna, et.al[3] have suggested that both the normal regression and boosting algorithms were implemented and found out that boosting algorithms have better results than the regular algorithms The critical factor of every big mart or store chain is sales. The sale factor has two faces, profit and loss. Therefore, to maintain the mart's standard, the sales graph has to be good every year. In this section of the literature view, we see many previous searches and methods used to predict sales as perfectly as possible by many researchers.. Another proposed model and claims that his model is best to perform robustly and perfectly according to the customer's demand. [3].Das and Chaudhary [2] proposed a model based on a linear and non-linear statistical framework for forecasting trading. They use both regressions in the model and separate the results based on the given data sets. They are many factors that participate in forecasting. Data is directly proportional to the prediction; the prediction will be wrong if the data set is incorrect. Under the variability of items, it is difficult to predict sales. They are many factors that give a negative result in the case of uncertainty. Yuan et al. [2] deal with the effective sales of E-commerce and discuss the design and management that help predict the sales on behalf of user behaviors. One of the research is data mining on the urban traffic prediction we used in our research [3]. This paper focuses on the classifications model and studies traffic behaviors to predict traffic affairs. Even for small bossiness, sales analysis is not an easy task. This research is based on the inspection of the customer's bargain and updates the critical adjustments based on the analysis of the costumer's bargain data [4]. The authors set the plans by adding a forum in the mobile application so that customers will buy directly from sellers. Nagamma et al. [2] conduct their research to find the relationship between ticket booth revenue and the film's online viewers. They use a support vector machine (SVM) classification model to train and predict the online viewers' movie revenue.

Fawcett, Tom and Foster J. Provost[4] The method of identifying suspicious behavior using an automated prototype is described in this study. For the purpose of completing this acceptable prototype, many machine learning methods were used. (Demchenko et al.)[4]To forecast sales, a generic linear method, a decision tree approach, and a decent gradient approach were employed. The original data set evaluated included a large number of entries, but the final data set utilized for analysis was significantly less than the original since it included non-usable data, duplicate entries, and unimportant sales data. (Ragg et al.)[4] Many vendors would profit from the forecast of a single transaction rate, as shown in this study, which implies the knowledge collected may be useful for the design of a set-up that would predict a number of results. The neural network technique is used to make the prediction. They used Bayesian learning to acquire insights in this situation. (Augusto Ribeiro et al.) [4]

**Chapter 3**

**Methodology**

* 1. **Methodology used**

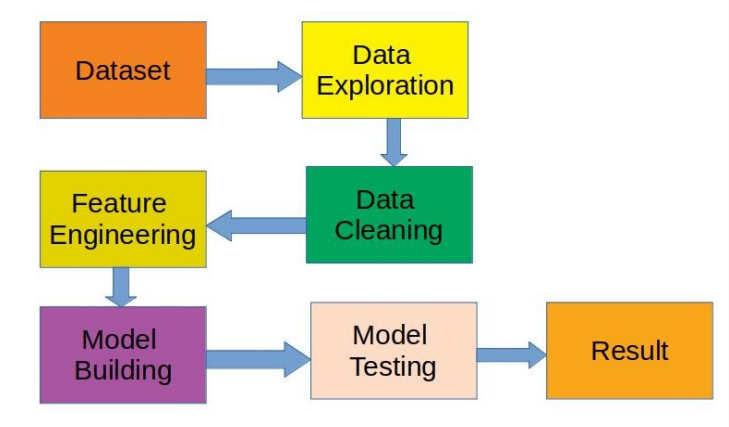
****

Fig. 1. Working procedure of proposed mode

1. **Hypothesis Generation** – understanding the problem better by brainstorming possible factors that can impact the outcome
2. **Data Pre-processing**
   1. **Data Exploration** – looking at categorical and continuous feature summaries and making inferences about the data.
3. **Data Cleaning** – imputing missing values in the data and checking for outliers
4. **Feature Engineering** – modifying existing variables and creating new ones for analysis
5. **Model Building** – making predictive models on the data
6. **EDA-** It is one of the most important parts when it comes to data analysis.To gain important insights of our data one must need to do exploratory data analysis. Here in our project we used two libraries i.e. KLib and DTale library.
   1. **Technologies used**

**3.2.1 Linear Regression**

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on the kind of relationship between dependent and independent variables, they are considering and number of independent variables being used.

Y = βo + β1X + ∈------------ (3.1)

where,

Y : It is basically the variable which we used as a predicted value.

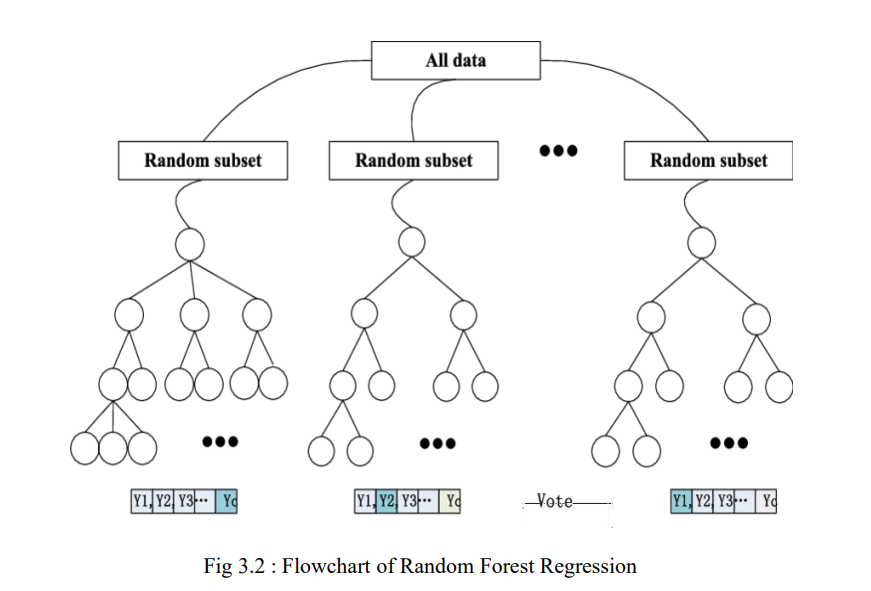
X : It is a variable(s) which is used for making a prediction.

βo : It is said to be a prediction value when X=0.

β1 : when there is a change in X value by 1 unit then Y value is also changed. It can also be said as slope term ∈.

**3.2.2 Random Forest Regression**

Random Forest is a tree-based bootstrapping algorithm based on that tree that includes a certain number of decision trees to build a powerful predictive model. The final prediction may be the function of all predictions made by each learner. In sales prediction, random forest classifier is used because it has decision tree like hyperparameters. The tree model is same as decision tool. Fig.3.2 shows the relation between decision trees and random forest. To solve regression tasks of prediction by virtue of random forest, the sklearn.ensemble library’s random forest regressor class is used**.**

****

**3.2.3 XGBoost Regression**

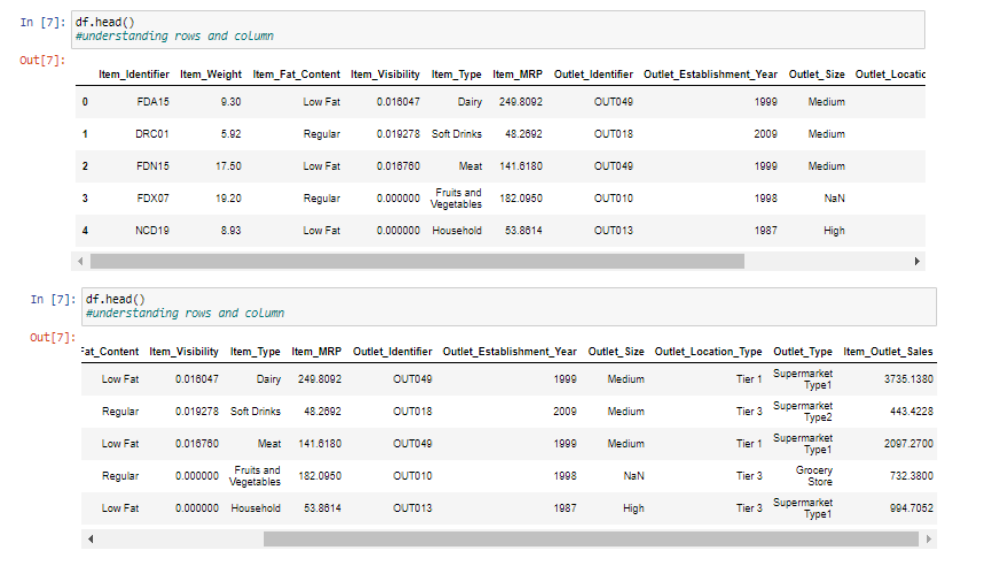
XGBoost stands for eXtreme Gradient Boosting. The implementation of an algorithm designed for the efficient operation of computer time and memory resources. Boosting is a sequential process based on the principle of the ensemble. This includes a collection of lower learners as well improves the accuracy of forecasts. No model prices n heavy for any minute t, based on the results of the previous t-speed. Well-calculated results are given less weight, and the wrong ones are weighed down. The XGBoost model uses stepwise, ridge regression internally, automatically selecting features as well as deleting multicollinearity.

**3.2.4 Decision Tree**

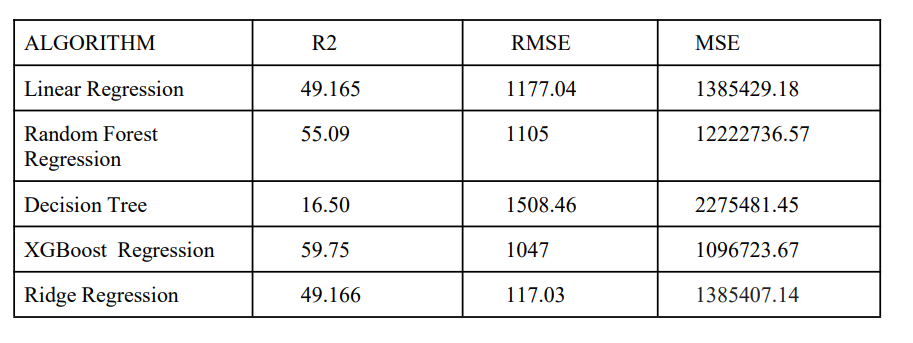
It's a simple model with little bias that may be used to create a classifier model, with the root node being the first to be considered in a top-down approach. It is a well-known machine learning model. A decision tree is referred to as a tuple recursive classifier. It is a potent approach for data mining and a powerful method of multi-variable analysis. This approach depicts the variables involved in accomplishing a particular goal, as well as the motivations for obtaining the goal .

**Chapter 4**

**Result and Discussion**

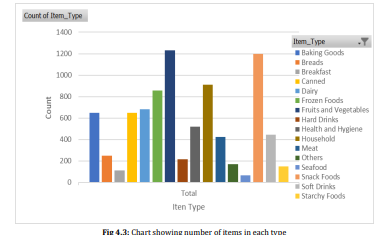
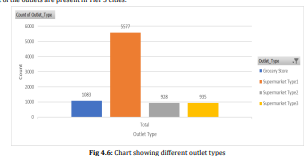
****

**Fig 4.1 Screenshots of dataset**

****

**Fig 4.2 Algorithms Performance**

To forecast BigMart’s revenue, simple to advanced ML algorithms have been implemented, such as LR, Decision Tree, RF regression and XGBoost. It’s been found that the most efficient algorithm to predict the sales revenue of Big mart is observed with Gradient Boosted Decision Tree and Random Forest algorithms having the least RMSE value among other algorithms.

The prediction results demonstrate the interrelation among different considered attributes.and A medium-sized location, for example, had the highest sales, implying that similar patterns could be implemented in other shopping locations to boost sales. Accuracy is critical in prediction-based systems, and it can improve dramatically as the number of parameters increases. The concept could be expanded to a web-based application or integrated into IoT-enabled devices with built-in intelligence to improve feasibility. Stakeholders involved in sales information can provide more inputs to facilitate with hypothesis generation, and taking into account more instances can lead to more precise outcomes that closely reflect real-world scenarios.

Notably, the regression results are more expressive and well bounded, with some degree of accuracy, making them easier to interpret. The proposed approach is flexible, and it can be improved further by incorporating variants during the regression model-building stage. Additional trials are required to quantify both accuracy and resource efficiency in order to accurately analyse and optimise the system.

**Chapter 5**

**Conclusion**

A sales forecast for Big Mart was conducted using Random Forests and Multiple Linear Regression algorithms. However, these methods resulted in lower accuracy. Upon analyzing the data, it was found that the XGBoost algorithm provided more efficient and accurate predictions with faster results. This was evident from lower RMSE, MSE, and MAE ratings and higher R2, Adjusted R-squared, and R-squared values for XGBoost compared to Linear Regression.

Based on this project, it is evident that business organizations require a smart sales forecasting program to handle vast volumes of data effectively. The algorithms presented in this report, including LR, RF regression, Decision tree, and XGBoost regression, provide efficient data sharing and decision-making approaches. They also aid in better understanding consumer needs and formulating effective marketing plans. The results of the machine learning algorithms used in this project will aid in the selection of the best demand forecast method. BigMart can plan and implement targeted marketing initiatives more successfully with this information.

**Future Work**

The potential future extension of this technology includes collaboration with other devices endowed with built-in intelligence via the Internet of Things (IoT), making it more feasible and versatile in its applications. This sales forecasting project can be improved by integrating more examples, parameters, and different factors to make it more distinctive and successful.

Experts have stressed the importance of a smart sales projection tool for business organizations to properly manage large volumes of data. The effectiveness of business assessments is determined by the speed and accuracy of the analytical techniques used to evaluate the outcomes. A substantial emphasis will be placed on using revolutionary methods for better-detecting customer demands and developing marketing strategies.

The results of machine learning algorithms will be crucial in determining the best demand prediction algorithm. Big Mart can use this data to improve its marketing activities and modify them to better meet the needs of its customers.

**References**

[1] Nikita Malik Sales Prediction Model for Big Mart, Vol 3. Issue 1 (https://msiggsip.org) , Aditi Narkhede ,Big Mart Sales Prediction Using Machine Learning Techniques (http://www.ijsred.com) ,Pavan Chatradi, Improvizing Big Market Sales Prediction, Volume 12 Issue 4, Punam, K., Pamula, R., Jain, P.K.: A two-level statistical model for big mart sales prediction. In: 2018 International Conference on Computing, Power and Communication Technologies (GUCON), IEEE (2018). pp. 617–620. **(Example: Research papers)**

[2] Asst. Prof. Keyaben patel (Author), Navneet Kumar (Author) ,Suraj Choudhari (Author), “BigMart Sale Prediction using Machine Learning ,” Volume 6, Issue 9, September – 2021 International Journal of Innovative Science and Research Technology ISSN No:-2456-2165 **(Example: Research papers)**

[3] Dr. M. Chaitanya Kishore Reddy1 , T. P. V. S. Sanjay2 , R. Geetha Meghana3 , T. Lakshmi Bhavani4’’ A Two-Level Statistical Model For Big Mart Sales Prediction,”International Journal for Advanced Research in Science and Technology ,ISSN: 2457-0362 **(Example : Journals)**

**[4]** Fawcett, Tom, Foster J. Provost. “Combining Data Mining and Machine Learning for Effective User Profiling” In KDD, pp. 8-13.1996, “Applied Linear Statistical Model”, Fifth Edition by Ragg, Thomas, Wolfram Menzel, Walter Baum and Michael Wigbers. “Bayesian learning for sales rate prediction for thousands of retailers.” Neurocomputing 43, no. 1-4 (2002):127-144**(Example: Research papers)**

**[5]** [(PDF) A Comparative Study of Big Mart Sales Prediction (researchgate.net)](https://www.researchgate.net/publication/336530068_A_Comparative_Study_of_Big_Mart_Sales_Prediction) **(Example: Website)**