**Sales Prediction using ML Algorithms**

**Abstract:** Nowadays shopping malls and marts keep the track of their sales data for each and every individual item for predicting future demand of the customer and updating the inventory management as well. These data stores basically contain a large number of customer data and individual item attributes in a data warehouse. Further, anomalies and frequent patterns are detected by mining the data stored in the data warehouse. The resultant data can be used for predicting future sales volume with the help of different machine-learning techniques for the retailers like Walmart, and Target. In this paper, we propose predictive models like Linear Regression, Random Forest, Decision Tree, and Gradient Boosting Regressor for predicting the sales of a company like Walmart, or Hyvee and found that the model produces better performance as compared to existing models. A comparative analysis of the model with others in terms of performance metrics is also explained in detail.

**Keywords:** Mining, Data Warehouse, Machine learning Techniques, Linear Regressor, Random Forest, Decision Tree, and Gradient Boosting.

**INTRODUCTION:** Day-by-day competition among different shopping malls as well as Marts is getting more serious and aggressive only due to the rapid growth of global malls and online shopping. Every mall or mart is trying to provide personalized and short-time offers for attracting more customers depending upon the day, such that the volume of sales for each item can be predicted for inventory management of the organization, logistics, transport service, etc. Present machine learning also algorithms are very sophisticated and provide techniques to predict or forecast the future demand of sales for an organization, which also helps in overcoming the cheap availability of computing and storage systems. In this paper, we are addressing the problem of sales prediction or forecasting of an item on customers’ future demand in different malls and stores across various locations and products based on the previous record. Different machine learning algorithms like linear regression analysis, random forest, etc are used for the prediction or forecasting of sales volume. As good sales are the life of every organization so the forecasting of sales plays an important role in any shopping complex. Always a better prediction is helpful, to develop as well as to enhance the strategies of business about the marketplace which is also helpful to improve the knowledge of the marketplace. A standard sales prediction study can help in deeply analyzing the situations or the conditions that previously occurred and then, the inference can be applied about customer acquisition, funds inadequacy, and strengths before setting a budget and marketing plans for the upcoming year. In other words, sales prediction is based on the available resources from the past. In-depth knowledge of the past is required for enhancing and improving the likelihood of the marketplace irrespective of any circumstances, especially the external circumstance, which allows for preparing the upcoming needs of the business. Extensive research is going on in the retailer’s domain for forecasting future sales demand. The basic and foremost technique used in predicting sales is the statistical method which is also known as the traditional method, but these methods take much more time for predicting sales also these methods could not handle non-linear data so to overcome these problems in traditional methods machine learning techniques are deployed. Machine learning techniques can not only handle non-linear data but also huge data-set efficiently. To measure the performance of the models, Root Mean Square Error (RMSE) [15] and Mean Absolute Error (MAE) [4] are used as evaluation met metrics mentioned in two equations respectively. Here Both metrics are used as the parameter for the accuracy measure of a continuous variable.

Table

Description automatically generated with medium confidence

where n: total number of error and | xpredict – xactual | : Absolute error.

**2. RELATED WORK:** Machine Learning is defined as a computer program that learns by itself from its experience without any human interference. Research on sales prediction has been done and some of them have been discussed below: In paper[1], the general linear approach, decision tree approach, and good gradient approach were used to predict sales. The initial data set considered included many entries, but the final data set which is used for analysis as much smaller than the original as it consists of non-usable data, redundant entries, and insignificant sales data. In paper[2], the linear regression method has been organized into structured data. Then it involves modeling data for predictions using machine learning techniques where the expected accuracy was 84%. In paper[3], they used linear regression and the XG booster algorithm to forecast sales that included data collection and translation into processed data. Ultimately, they predicted which model would produce the better outcome. In paper[4], sales were predicted using three modules that are a hive, R programming, and tableau. Analyzing the store’s history helps get an understanding of the store's revenue to make some improvements to the target so it can be more successful. Within the diagram, key values are obtained to reduce all intermediate values by reducing the intermediate key feature to obtain the results. Mohit Gurnani in his research proves that composite models achieve good results in comparison to individual models. He also stated that decomposition mechanisms are far better than hybrid mechanisms [5]. J. Scott Armstrong in his research discussed predicting solutions to interesting and difficult sales forecasting problems [6]. Samaneh Beheshti-Kashi in his research reviewed different Various approaches to the predictive potential of consumer-generated content and search queries [7]. Gopal Behera has done an effective study on Big mart sales prediction and has given prediction metrics for various existing models [8]. In this paper, we use the random forest and XG booster methodology in which raw data obtained at large mart will be pre-processed for missing data, anomalies, and outliers. Then an algorithm will be used to predict the final results. ETL stands for Extract, Transform and Load and finally, we compare all the models and predict which model gives accurate results.

**3. METHODOLOGY:** Sales forecasting as well as analysis of sale forecasting has been conducted by many authors as summarized: The statistical and computational methods are studied in [2] also this paper elaborates on the automated process of knowledge acquisition. Machine learning [6] is the process where a machine will learn from data in the form of statistical or computational methods and process knowledge acquisition from experiences. Various machine learning (ML) techniques with their applications in different sectors have been presented in [2]. Pat Langley and Herbert A [7] pointed out that the most widely used data mining technique in A Comparative Study of Big Mart Sales Prediction 3 in the field of business is the Rule Induction (RI) technique as compared to other data mining techniques.

**Diagram

Description automatically generated**

Fig 1: Showing the working procedure which is followed in this paper.

**3.1 Hypothesis Generation:** “Put yourself in one’s shoes” this is to be remembered and should be implemented right at the beginning of the sales forecasting, giving a basic idea of the mart’s progress and its sales and estimations. The hypothesis is measured mainly on two criteria,

1. Store Level Hypothesis

* According to the general hypothesis, store sales depend upon whether the stores are in urban or rural areas.
* Another factor is population; it is clear that more population affects sales.
* The size of a store may also affect the sales; the more significant the store would print the store's quality on the customer's mind.
* Opponent factor how many challenges are in the market.
* Advertisement is also a pillar of sales.
* Location matters. The sale would be satisfactory if the store is near a dense population.
* Customer satisfaction.

According to the general hypothesis, store sales depend upon whether the stores are in urban or rural

− Opponent factor how many challenges are in the market.

− Advertisement also a pillar of sales.

− Location matters. The sale would be satisfied if the store is near a dense population.

− Customers satisfaction.

According to the general hypothesis, store sales depend upon whether the stores are in urban or rural

− Opponent factor how many challenges are in the market.

− Advertisement also a pillar of sales.

− Location matters. The sale would be satisfied if the store is near a dense population.

− Customers satisfaction.

According to the general hypothesis, store sales depend upon whether the stores are in urban or rural

2. Product Level Hypothesis

* People always need quality products related to some brands.
* Designing the product with good packaging also prints the excellent quality of the customer's mind.
* Stores have a related product or not., *i.e*., stores should have the product related to daily usage.
* Discount Promotions also attract customers.
* Product price.

|  |  |
| --- | --- |
| Variables | Hypothesis link |
| Item\_identifier | Unique identification |
| Item\_Weight | Absent in hypothesis |
| Item\_Fat\_Content | Relation with quality |
| Item\_Visibility | Present in hypothesis, related to the display. |
| Item\_Type | Absent in our hypothesis |
| Item\_MRP | Present in hypothesis |
| Outlet\_Identifier | Unique identification |
| Outlet\_Establishment\_Year | Absent in our hypothesis |
| Outlet\_Size | Has relation with the capacity of the store |
| Outlet\_Location\_Type | Present in hypothesis |
| Outlet\_Type | Present in hypothesis |
| Outlet\_Sales | In hypothesis |

Table1: Relation between attributes and hypothesis

Text

Description automatically generatedTable

Description automatically generated**3.2 Data Exploration:** In this phase, useful information about the data has been extracted from the dataset. That is trying to identify the information from hypotheses vs available data. Table 1: shows the relation between attributes and hypothesis. The attributes ‘Outlet size’ and ‘Item weight’ face the problem of missing values as shown in Fig2, also the minimum value of Item Visibility is zero which is not practically possible(Fig 3).If you see there are null values of ‘Outlet\_Sales’ this can be addressed with the most repeated sale values. (Fig 4) shows that there are 1559 unique products, as well as 10 unique outlets, present in the dataset. The attribute Item type contains 16 unique values. There are two types of Item Fat Content are there but some of them are misspelled as regular, and reg instead of ’Regular’ and low fat, LF instead of Low Fat. (Fig:5)

Fig2: Total null values present in data Fig3: Description of various data variables

Table

Description automatically generated

Text

Description automatically generated

Fig5: Misspelled regulations.

Fig:4 Unique ‘Item\_type’ present in the data

**3.3 Data Cleaning**: It was observed from the previous section that the attributes Outlet Size and Item Weight have missing values. In our work in the case of the Outlet Size missing value, we replace it by the mode of that attribute, and for the Item Weight missing values, we replace it by the mean of that particular attribute. The missing attributes are numerical where the replacement by mean and mode diminishes the correlation among imputed attributes. For our model, we are assuming that there is no relationship between the measured attribute and the imputed attribute.

**3.4 Feature Engineering:** Feature engineering helps us understand the data for better analysis. Here, we will create some new variables from the original data. When we explore our data for analysis, we observe that many shades have to be resolved. For example, remember talking about the concatenation of two Supermarkets because we almost think that both have the same sales. So, check whether it is true or not. The output shows the difference between supermarket sales, and there is a significant difference b/w them, that why this idea will not be valuable to combine the data of markets.

**3.4.1 Altering ‘Item\_Visbility’:** During this phase, it was noticed that the Item visibility attribute had a zero value, practically which has no sense. So the mean value of the ‘Item visibility’ variable will be used for the zero values. This makes all products likely to sell.

Graphical user interface, text

Description automatically generated**3.4.2 Grouping:** All categorical attribute discrepancies are resolved by modifying all categorical attributes into appropriate ones. In some cases, it was noticed that non-consumables and fat content properties are not specified. To avoid this we create a third category of Item fat content i.e. none. In the Item Identifier attribute, it was found that the unique ID starts with either DR or FD, or NC shown in fig:6. So, we create a new attribute Item Type Combined with three categories like Foods, Drinks, and Non-consumables.

Fig6: The table shows the data after grouping the values.

**3.4.3 Categorical Variable Transformation:** Here I use one hot encoding.Earlier in data exploration, out of 12 features there are 7 Categorical variables to build a suitable model, the variable type of the entire data set is preferred. One hot encoding can be used here to achieve this. I perform one hot encoding to convert categorical values into numerical values which will further provide ML algorithms to do a better job in prediction and result in higher accuracy.**4. MODEL BUILDING:**

After completing Data Preprocessing and Feature Transformation, the dataset is now ready to build a predictive model. Now we need to divide the files into train and test when dealing with machine learning (ML) models, however, using the built-in function of sci-kit learn lib is also a good idea. Avoiding over- and under-fitting is one of the benefits of utilizing a split function, thus do so before using any machine learning technique. The algorithm is fed into the training set in order to learn how to forecast values. After Model Building a target variable to forecast, testing data is supplied as input. The predictive models are built using

* Linear Regression
* Random Forest
* Decision Tree
* Gradient Boosting

**Linear Regression:** One of the most essential and commonly used regression techniques is linear regression. It's one of the most basic regression techniques. The simplicity with which the results may be interpreted is one of its primary merits.

𝑦 = 𝛽₀ + 𝛽₁𝑥₁ + ⋯ + 𝛽ᵣ𝑥ᵣ + 𝜀.

Where Y - Variable to be Predicted

X – Variables used for making a prediction

𝛽₀, 𝛽1…𝛽r - Regression Coefficients

𝜀 - Random Error

Regardless of how well the model is trained, tested, and validated, there will always be a variation between observed and predicted, which is an irreducible error, so we cannot rely entirely on the learning algorithm's predicted results. Data must meet several conditions for a successful linear regression model. One of them is the lack of multiple linear regression, which means that the independent variables should be correlated.

The RMSE value obtained from this algorithm is 1243.85

The MAE value obtained from this algorithm is 912.48

**Random Forest:** The random forest algorithm is a highly accurate sales prediction method. It's simple to use and comprehend for forecasting the outcomes of machine learning projects. Random forest classifiers are employed in sales prediction because they have decision tree-like hyperparameters. The tree model is similar to a decision-making tool. A random forest model is created for each individual learner using a random set of rows and a few randomly selected factors. The final forecast may be based on all of the individual learners' guesses. In the case of a regression problem, the final forecast may be the average of all previous predictions.

The RMSE value obtained from this algorithm is 1228.8

The MAE value obtained from this algorithm is 884.92

**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 and the means of execution, in a variety of areas.

The RMSE value obtained from this algorithm is 1149.45

The MAE value obtained from this algorithm is 840.94

The improvement in RSME can be obtained by tuning the parameters of max\_depth and the number of trees.

**Gradient Boosting:** Decision trees and gradient boosting are used to create the XG Boost method. The algorithm's construction was designed to maximize the efficiency of computation time and memory resources. Boosting is a sequential procedure based on the ensemble concept. This involves a group of low learners and increases the accuracy rate. At every time t, model variables are weighted depending on the impacts of the previous instant. Correctly computed findings are given a lesser weight, whereas incorrectly calculated results are given a greater weight. The Gradient Boosting model uses this method to internally perform stepwise ridge regression, which automatically selects features and eliminates multiple regression.

The RMSE value obtained from this algorithm is 1099.76

The MAE value obtained from this algorithm is 754.06

**5. EXPERIMENTS RESULTS AND DISCUSSION:** Sales prediction is conducted by using four machine learning algorithms. We use machine learning algorithms to solve our dataset. Initially, we want to predict the sales of the organization by studying the sales of different marts with specific attributes, so that is why we set the "Oulet\_Sales" attribute with the dependent variable, and we see above there are more than 10 attributes that we use as some independent variables. Our dataset contains two different train and test files; we concatenate our files to understand the data better.

A hypothesis is necessary to check the possible attributes of the data. Also, it gives an understanding between the data scientist and the prediction. Therefore, we created some hypotheses and then compared them with the existing data. We saw a little bit of difference between the hypothesis and the data, and then we adjusted the data with our hypothesis attributes. Next, we moved to explore the data. In this part, we check the basic statistics

of the dataset and the missing values in the data. Moving to the nominal variables, we checked the unique values in the data and found that there are 4-5 categorical variables. Attributes named 'Product-weight' and 'Outlet-Size' filled the missing values, 2439 and 4016, respectively. We used mean and mode for the missing values of product weight and outlet size, respectively. Feature engineering is the process of creating some new attributes for a better understanding of the data. Once we were ready with the data, we had to make a model. We used four machine learning models to predict sales and compare the results. When working with machine learning (ML) models, it is good to split the files into train and test, but using the built-in function of sci-kit learn lib is a good idea. The advantages of using a split function are avoiding over-fitting and under-fitting, so use a split function before using any machine learning algorithm.

|  |  |  |
| --- | --- | --- |
| **Algorithms** | **RSME values** | **MAE values** |
| Linear Regression | 1243.85 | 912.48 |
| Random Forest | 1228.80 | 884.92 |
| Decision Tree | 1149.45 | 840.94 |
| Gradient Boosting | 1099.76 | 754.06 |

Table 2: Comparison of RSME and MAE values with algorithms.

My results are more accurate and near the original test data. Thus, this study method performs much better in using and predicting sales. We perceive different scenarios in which different models are best among all the models. Our research only predicts sales based on specific attributes, but it is not good enough to use globally; for example, we do not include disasters in our research, so our prediction is invalid in case of disaster. For prediction, we use many machine learning models and then evaluate and compare the result of the final use. Data is always not in manageable manner, so we need to beautify our data before modeling. This paper tells us the sequence flow of research and further attains outcomes with machine learning models.

Various machine learning algorithms like Linear Regression, Random Forest, Decision Tree, and Gradient Boosting have been built to predict sales. It’s been found that the most efficient algorithm to predict sales is observed with Gradient Boosted and Decision Tree algorithms having the least RMSE and MAE values when compared with linear regression and random forest algorithms.

**6. CONCLUSIONS:** In the present era of a digitally connected world every shopping mall desires to know the customer demands beforehand to avoid the shortfall of sale items in all seasons. Day to day the companies or the malls are predicting more accurately the demand of product sales or user demands. Extensive research in this area at the enterprise level is happening for accurate sales prediction. As the profit made by a company is directly proportional to the accurate predictions of sales, marts and malls are desiring a more accurate prediction algorithm so that the company will not suffer any losses. Experiments support that Gradient Boosting and Decision Tree algorithm technique produces more accurate prediction compared to other available techniques like linear regression, and random forest algorithm. Finally, it is concluded that our model with the lowest MAE and RMSE performs better compared to existing models.