Big Mart Sales Prediction System using Machine Learning

A. Pradeep Kumar Reddy[1] V. Nagalakshmi[2]
[1]P.G. Student [2]Professor
[1,2]GIS, GITAM, Visakhapatnam, India

Abstract— Supply and demand are two fundamental concepts of sellers and customers. Predicting demand accurately is critical for organizations in order to be able to formulate plans. Sales Prediction is based on predicting the sales for different outlets of Big Mart companies so that they can change the business model according to performance predicted. In this paper, a new approach for demand prediction for Big Mart companies is done. The business model used by the Big Mart companies, for which the model is implemented, includes many outlets that sell the same product at the same time throughout the country where the company operates a market place model. The demand prediction for such a model should consider the price tag, outlet type, outlet location.

Keywords: Machine Learning, Big Mart Sales Prediction System

I. INTRODUCTION

Day by day competition among different shopping malls as well as big marts is getting more serious and aggressive only due to the rapid growth of the global malls and on-line 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 and transport service, etc. Present machine learning algorithm 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, addressing is done on the problem of big mart sales prediction or forecasting of an item on customer's future demand in different big mart 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 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.

II. LITERARY WORK

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 also this paper elaborates the automated process of knowledge acquisition. Machine learning is the process where a machine will learn from data in the form of statistically or computationally method and process knowledge acquisition from experiences. Various machine learning (ML) techniques with their applications in different sectors has been presented in. This work is useful in pointing out most widely used data mining technique in A Comparative Study of Big Mart Sales Prediction 3 the field of business is the Rule Induction (RI) technique as compared to other data mining techniques. Where sale prediction of a Bigmart distribution company has been described in. Also, this paper focuses on two issues: (i) stock state should not undergo out of stock, and (ii) it avoids the customer dissatisfaction by predicting the sales that manages the stock level of customer products. Handling of footwear sale fluctuation in a period of time has been addressed in. Linear and non-linear a comparative analysis model for sales forecasting is proposed for the retailing sector.

III. PROPOSED SYSTEM

Raw data collected at big mart would be preprocessed for missing data, anomalies and outliers. Then an algorithm would be trained on this data to create a model. This model would be used for forecasting the final results.

Data preprocessing is a machine learning and data mining technique used to transform sample raw data into an undesirable format. Real world collected data may be inconsistent, incomplete or contains an error and hence data preprocessing is required.

IV. ABOUT DATASET AND ATTRIBUTES

Big mart's data scientists collected sales data for the year 2013 of 1559 products across 10 stores in different cities. Also they provided definitions for certain attributes of each product and store.

1) Item_Identifier - Unique identifier for each product.
2) Item_Weight – Product weight.
3) Item_Fat_Content – Fat content of the product.
4) Item_Visibility – Percentage of total display area in a store allocated to the product.
5) Item_Type – Product category.
6) Item_MRP – List price of the product.
7) Outlet_Identifier - Unique identifier for each store.
8) Outlet_Establishment_Year – Establishment year for each store.
9) Outlet_Size - The size of the store.
10) Outlet_Location_Type - The type of city in which the store is located.
11) Outlet_Type - Whether the store is a grocery store or a supermarket.
12) Item_Outlet_Sales - Sales of the product in each store.

A. Evaluation Metrics for Regression

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

Fig. 1: Mean Absolute error

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(\hat{y}_i - y_i)^2}{n}}$$

Fig. 2: Root mean squared error

All rights reserved by www.ijsrd.com

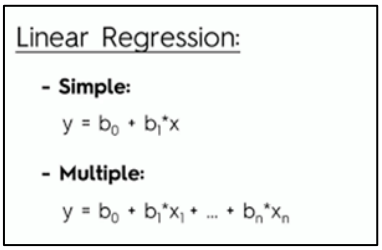

Fig. 3: Linear Regression

The basic idea of this algorithm is to fit a straight line between the selected features in training dataset and a continuous target variable i.e. sales. This algorithm finds a line that best fits the data.

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x) [3]. So, this regression technique finds out a linear relationship between x (input) and y(output).

The equation of the regression line is represented by: h(Xi)=B0+B1Xi h(Xi) is the expected value for ith observation.

1) Algorithm

Input: Dataset with proper input and output labels Output: Predict sales value and store in csv file
Begin
1) Calculate mean, variance for the list of values
Def Mean(values):
Sum(values) divided by Len(values)
Calculate Mean_x, Mean_y
Def Variance (values, Mean):
sum([(values-Mean) ^2])
Calculate Variance_x, Variance_y
2) Calculate covariance
Covar<=0
Def Covariance (x, Mean_x, y, Mean_y):
For length of x do:
Covar<= Covar + (x[i] - Mean_x) * (y[i] – Mean_y)
End
3) Estimate coefficients
B0 <= covariance (x, Mean_x, y, Mean_y) / variance (x, Mean_x)
B1 <=Mean_y – B0*Mean_x
4) Predict For every X in the test set do:
Y <=B0+B1X
End
Store predicted values in CSV file

B. Random Forest

Random forests are an ensemble machine learning algorithm for classification, regression and other tasks. It operates by constructing a many decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

Random forests (RF) construct many individual decision trees at training. Predictions from all trees are pooled to make the final prediction; the mode of the classes for classification or the mean prediction for regression. As they use a collection of results to make a final decision, they are referred to as Ensemble techniques.

V. IMPLEMENTATION AND RESULTS

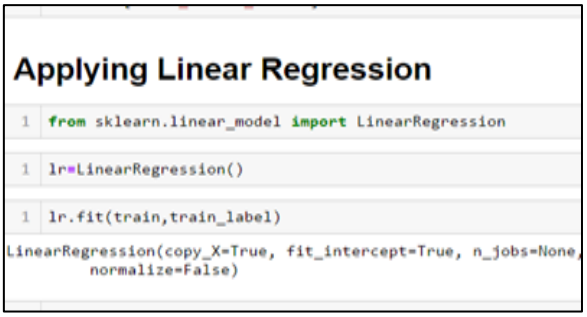

Fig. 4: Input code on Linear Regression

Fig. 5: Input code on Random Forest

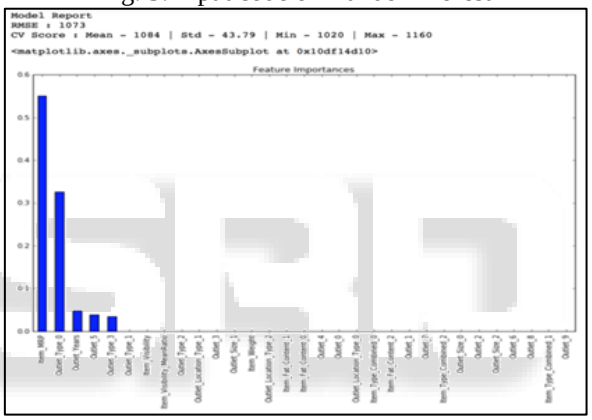

Fig. 6: Random Forest Model

VI. CONCLUSION

Hence, we propose a software tool for forecasting future sales volume based on the historical sales data. Using this tool, the accuracy of prediction for multiple linear regressions and random forests can be determined.

REFERENCES

[1] H. M. Al-Hamadi "Long-Term Electric Power Load Forecasting Using Fuzzy Linear Regression Technique", IEEE Mar.2011
[2] Yanming Yang "Prediction and Analysis of Aero-Material Consumption Based on Multivariate Linear Regression Model", 2018 the 3rd IEEE International Conference on Cloud Computing and Big Data Analysis
[3] YouLi Feng, ShanShan Wang "A Forecast for Bicycle Rental Demand Based on Random Forests and Multiple Linear Regression",

All rights reserved by *www.ijsrd.com*