A Forecast for Big Mart Sales Based on Random Forests and Multiple Linear Regression

[1]Heramb Kadam, [2]Rahul Shevade, [3] Prof. Deven Ketkar, [4]Mr. Sufiyan Rajguru
[1] BE IT, FAMT, Ratnagiri [2] BE IT, FAMT, Ratnagiri, [3] Assistant Professor ,IT department, FAMT, [4] BE IT, FAMT, Ratnagiri

Abstract— In today's world, big malls and marts record sales data of individual items for predicting future demand and inventory management. This data stores a large number of attributes of the item as well as the individual customer data together in a data warehouse. This data is mined for detecting frequent patterns as well as anomalies. This data can be used for forecasting future sales volume with the help of random forests and multiple linear regression model.

Index Terms— Sales Forecasting, Regression, Random Forests, Machine Learning

I. INTRODUCTION

With the rapid development of global malls and stores chains and the increase in the number of electronic payment customers, the competition among the rival organizations is becoming more serious day by day. Each organization is trying to attract more customers using personalized and short-time offers which makes the prediction of future volume of sales of every item an important asset in the planning and inventory management of every organization, transport service, etc. Due to the cheap availability of computing and storage, it has become possible to use sophisticated machine learning algorithms for this purpose. In this paper, we are providing forecast for the sales data of big mart in a number of big mart stores across various location types which is based on the historical data of sales volume. According to the characteristics of the data, we can use the method of multiple linear regression analysis and random forest to forecast the sales volume.

II. LITERATURE SURVEY

The method for long term electric power forecasting using long term annual growth factors was proposed [1]. Prediction and analysis of aero-material consumption based on multivariate linear regression model was proposed by collecting the data of basic monitoring indicators of aircraft tire consumption from 2001 to 2016 [2]. The forecast for bicycle rental demand based on random forests and multiple linear regressions was proposed based on weather data [3].

III. PROPOSED SYSTEM

We propose below methodology for solving the problem. Raw data collected at big mart would be pre-processed 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.

ETL stands for Extract, Transform and load. It is a tool which is a combination of three functions. It is used to get data from one database and transform it into a suitable format. Data preprocessing is a data mining technique used to transform sample raw data into an understandable format. Real world collected data may be inconsistent, incomplete or contains an error and hence data preprocessing is required.

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. They are as follows-:

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.

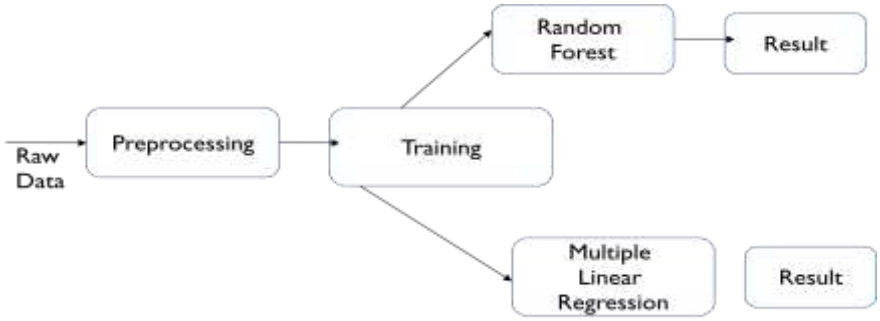

Fig.1 Block diagram of proposed system

A. MULTIPLE LINEAR REGRESSION

Multiple linear regression establishes a relationship between dependent variable (Y) and one or more independent variables (X) using a best fit straight line (also known as regression line). It is represented by an equation,

$$Y=a+b*X + e \qquad (1)$$

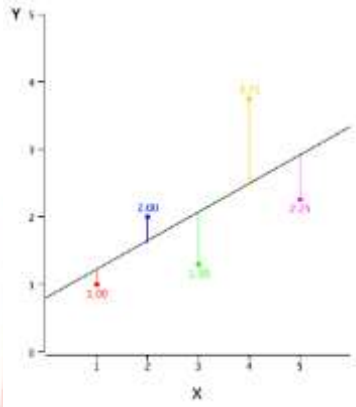

Fig.2 Linear regression

Where a is intercept, b is slope of the line and e is error term. Using this method, an accuracy can be found out. Multiple linear regression is very famous method for prediction and analysis but one drawback is it gives less accuracy.

B. RANDOM FORESTS

Random forests or random decision 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 provide solution for the decision trees habit of over fitting.

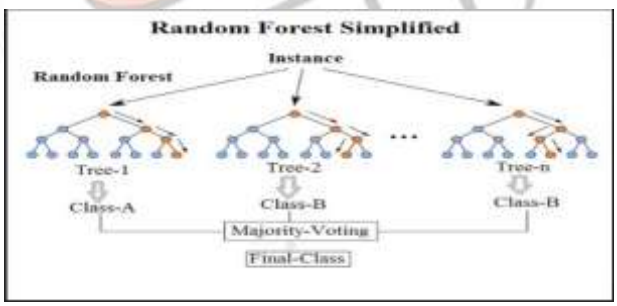

Fig.3 Random Forests

IV. 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", IEEE Dec.2013.