Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 14, Issue. 2 : 2023
ISSN : **1906-9685**

PREDICTIVE ANALYSIS FOR BIG MART SALES USING MACHINE LEARNING ALGORITHMS

[1]G.Padma, [2]S. Chandana, [3]V.Romitha, [4]Y.Vanditha Reddy, [5]M.Sukanya
[1]Assitsant Professor, [2,3,4,5]UG Students, Dept. Computer Science and Engineering-Data Science, Mallareddy Engineering college for Women, Hyderabad, India.

ABSTRACT
Everybody wants to know how to buy goods cheaper or how to advertise them at low cost. Here is the answer. That is Big Mart. Big Mart is online one stop marketplace where you can buy or sell or advertise your merchandise at low cost. The goal is to make Big Mart the shopping paradise for buyers and the marketing solutions for the sellers. The ultimate goal is to prosper with customers. The project "PREDICTIVE ANALYSIS FOR BIG MART SALES USING MACHINE LEARNING ALGORITHMs" aims to build a predictive model and find out the sales of each product at a particular store. Big Mart will use this model to understand the properties of products and stores which play a key role in increasing sales. This can also be done based on the hypothesis that should be done before looking at the data.

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.

EXISTINGSYSTEM
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.

http://doi.org/10.36893/JNAO.2023.V14I2.0142-0148

PROPOSED SYSTEM

The data scientists at Big Mart have collected 2013 sales data for 1559 products across 10 stores in different cities. Also, certain attributes of each product and store have been defined. The aim is to build a predictive model and find out the sales of each product at a particular store. Using this model, BigMart will try to understand the properties of products and stores which play a key role in increasing sales.

Advantages

This is an easily scalable model to provide detailed information and accurate predictions for sales volume for different types of products as there is a lot of data out there.

It is the percentage of display space in a store given to that particular item. Looking at the average visibility of items given in each store type and outlet.

Goals:

Building the regression models: linear and decision tree. Predicting the sales, cross validating the scores, calculating the R^2.

Classifying the training data with a decision tree and a random forest and calculating the accuracy score and the R^2.

Software Design

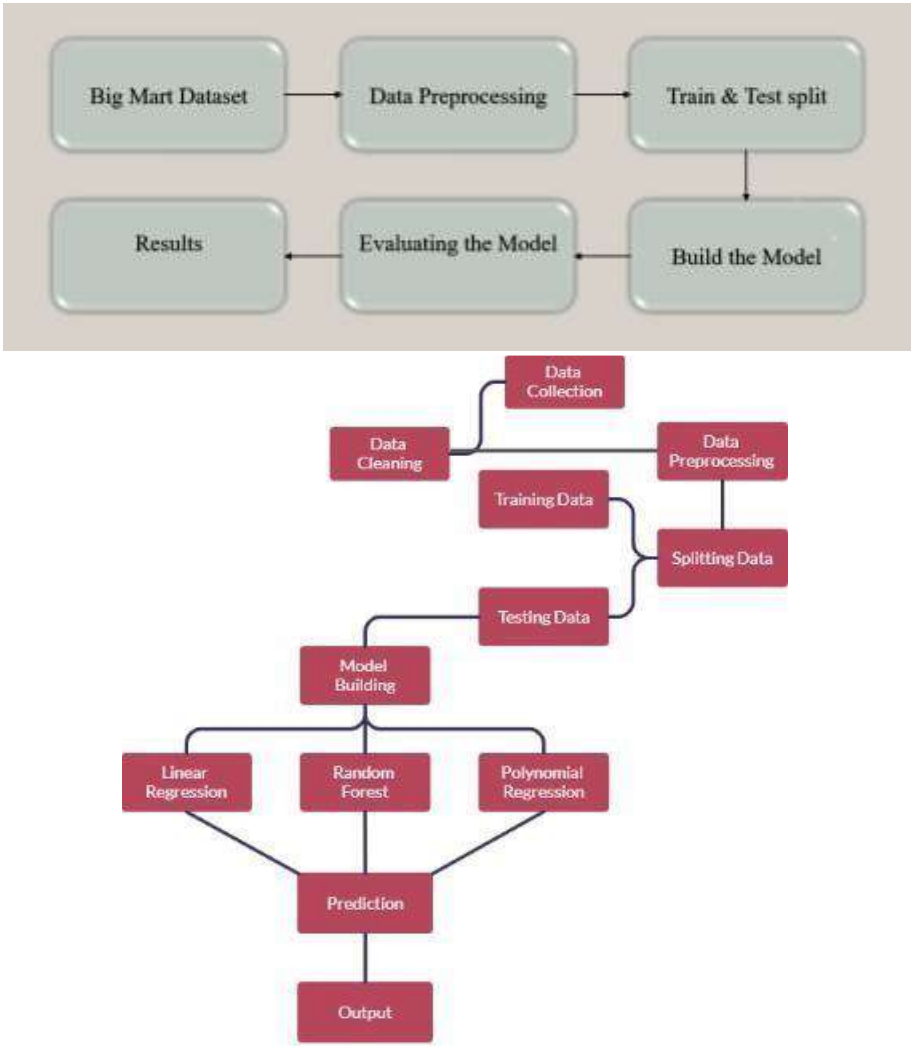


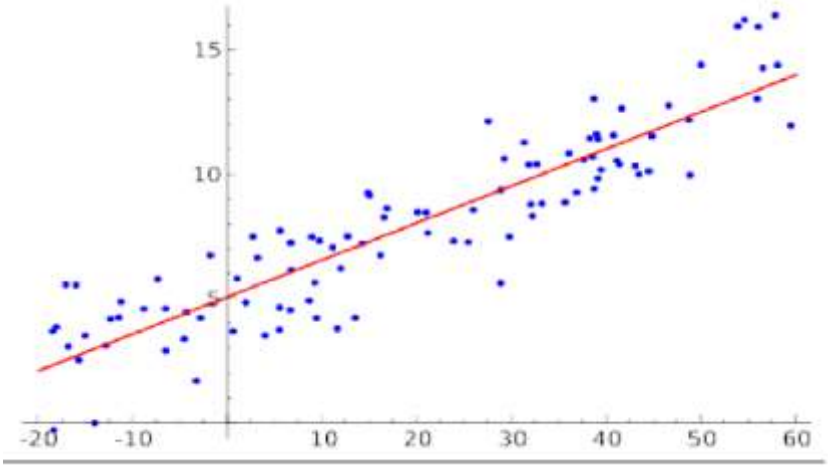
Fig.1.This block diagram describes the training phase and testing phase of Big Mart sales.

http://doi.org/10.36893/JNAO.2023.V14I2.0142-0148

Algorithms

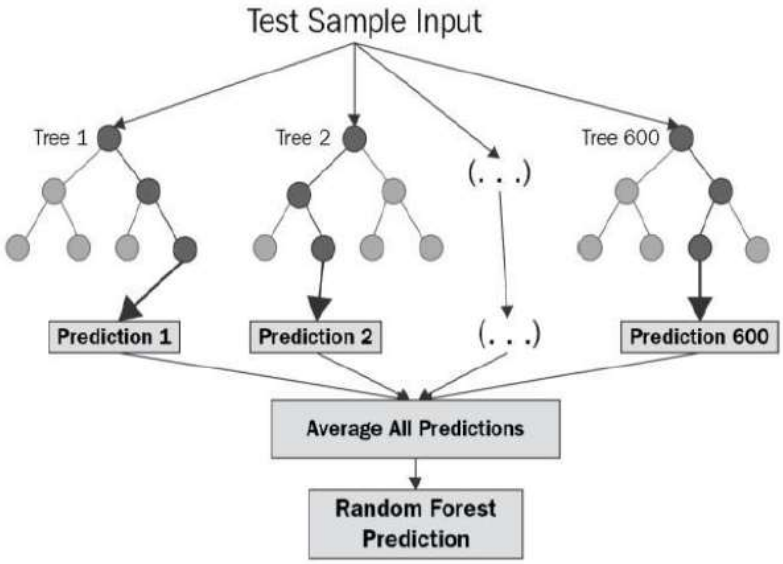
Linear Regression

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable. This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a “least squares” method to discover the best-fit line for a set of paired data. You then estimate the value of X (dependent variable) from Y (independent variable).



Random Forest Regressor

Random Forest Regression is a supervised learning algorithm that uses ensemble learning methods for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model. A Random Forest Regression model is powerful and accurate. It usually performs great on many problems, including features with non-linear relationships. Disadvantages, however, include the following: there is no interpretability, overfitting may easily occur, we must choose the number of trees to include in the model.

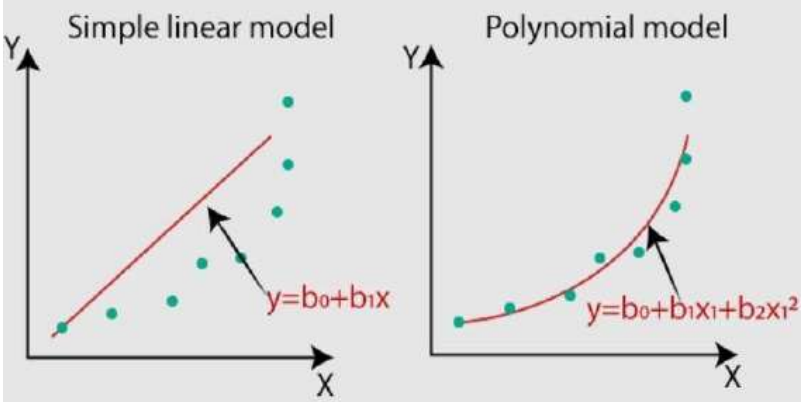


http://doi.org/10.36893/JNAO.2023.V14I2.0142-0148

Polynomial Regression

Polynomial regression is a form of linear regression in which the relationship between the independent variable(s) and the dependent variable is modeled as an nth degree polynomial. It allows for capturing non-linear relationships between variables by introducing polynomial terms.

Polynomial regression works by fitting a polynomial equation to the observed data points using the method of least squares. The degree of the polynomial determines the complexity of the curve that can be fitted to the data. During the training phase, the model estimates the coefficients that minimize the sum of squared differences between the predicted values and the actual values. This estimation is typically performed using optimization algorithms. Once the model is trained, it can be used to predict the values of the dependent variable based on the values of the independent variable(s) and the estimated coefficients.



DATA DESIGN

Data collection

Data collection is the process of gathering and measuring information on variables of interest, in an established systematic fashion that enables one to answer stated research questions, test hypotheses, and evaluate outcomes. The data collection component of research is common to all fields of study including physical and social sciences, humanities, business, etc. While methods vary by discipline, the emphasis on ensuring accurate and honest collection remains the same.

Data Preprocessing:

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model. A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data preprocessing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

It involves below steps:

➢ removing of high cardinality column

Almost all datasets now have categorical variables. Each categorical variable consists of unique values. *A categorical feature is said to possess high cardinality when there are too many of these unique values.* One-Hot Encoding becomes a big problem in such a case since we have a separate column for each unique value (indicating its presence or absence) in the categorical variable. This leads to two problems, one is obviously space consumption, but this is not as big a problem as the second problem, the curse of dimensionality**.**

Converting Data Types

Converting into data types is an important step because we cannot perform all operations on the data such as bar chart. To plot a bar chart we need both categorical and numerical variables if there are 0's and 1's in categorical variables which represents as Yes or No but as they are 0's and 1's, they are stored in INT data type. So, plotting a bar chart is not possible.

http://doi.org/10.36893/JNAO.2023.V14I2.0142-0148

Not only bar chart but there also other operations that we can't perform on the data without converting their data types. Therefore converting the data types makes data efficient.

➢ Imputing Null Values:

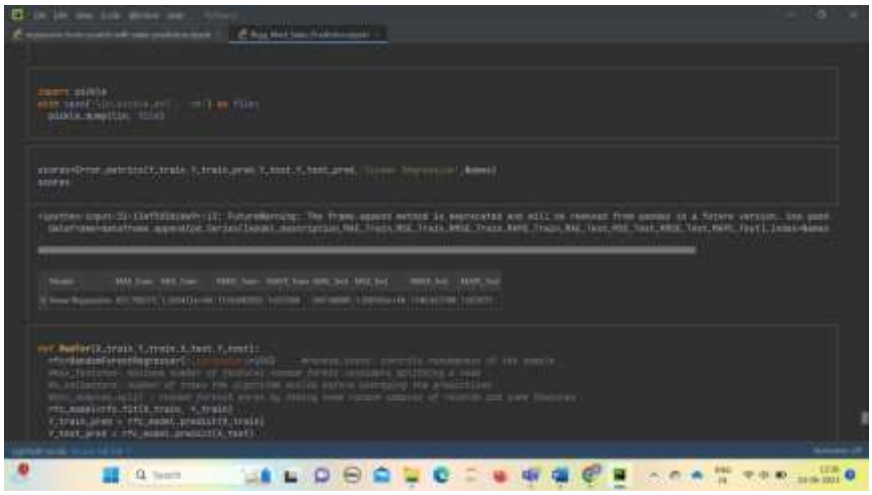
Dropping the Data Point: Sometimes Dropping the Null values is the best possible option in any ML project. One of the Efficient approach/case where you should use this method is where the number of Null values in the feature is above a certain threshold like for example, based on our domain knowledge we made a decision that if the number of null values are greater than 50% of the total number of data points then we will drop the feature. Drawback of this method is that, if you drop the column you might end up losing critical information.

Mean Imputation: This is the most common method to impute missing data. In this method we just replace the null values with the mean value of the feature. This method is used for numerical features. Although this is most common method, one should not blindly use this method because it is prone to outliers and may affect the model performance drastically.

Median Imputation: In order to overcome the drawback of the mean imputation which is that it is sensitive to outliers, one common approach which is used by ML Engineers is that rather than imputing mean value, they impute the median. Although there are no direct drawbacks of using this method but you would want to consider plotting the distribution of the feature before applying this method.

Result& Analysis

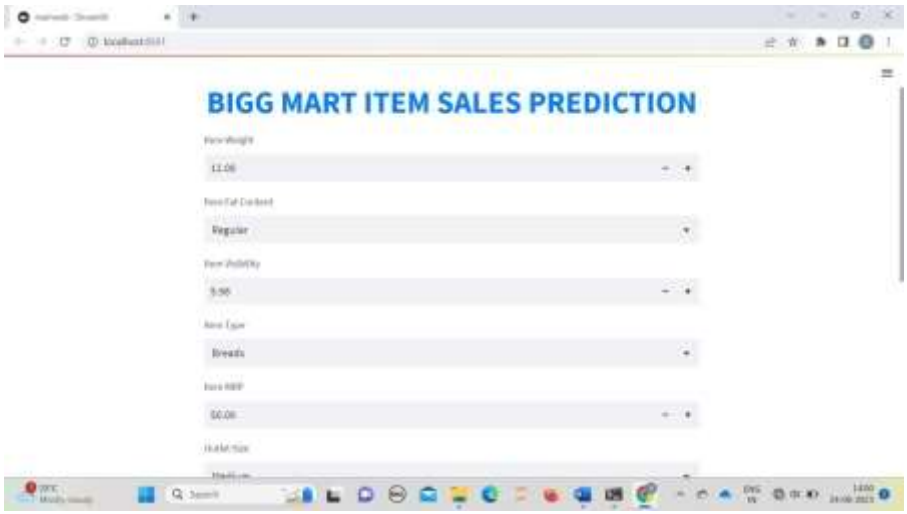
The below Pic shows the error metrics of Linear Regression.



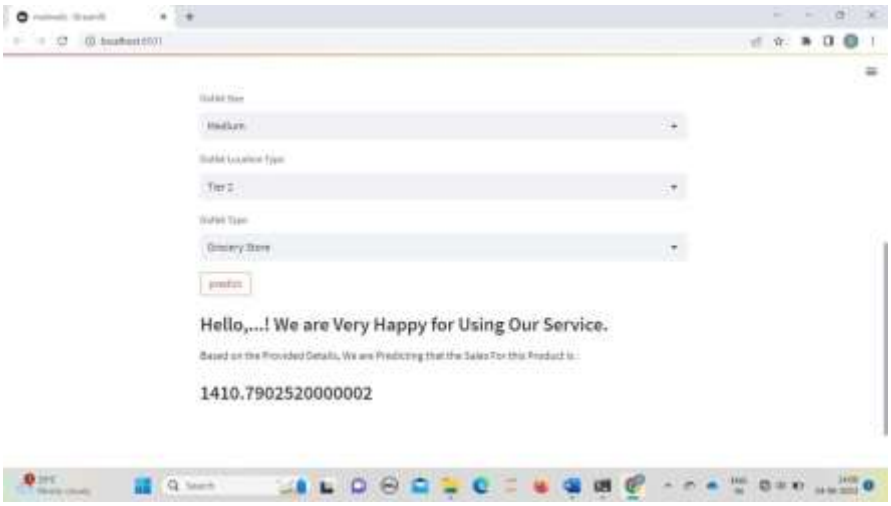
The below Pic shows the error metrics of Random Forest.

http://doi.org/10.36893/JNAO.2023.V14I2.0142-0148

The below pic is the web page where we enter input values for sales prediction.



Below pic gives the predicted sales.



Conclusion

In this project, the effectiveness of various algorithms on the data on revenue and review of, best performance-algorithm, here propose a software to using regression approach for predicting the sales centered on sales data from the past the accuracy of linear regression prediction can be enhanced with this method, polynomial regression, Random Forest algorithm can be determined.

The Machine Learning algorithm that performs the best was Random Forest which got us low RMSE (429.317334) compared to other algorithms. So, we have used Random Forest Model. So, we can conclude ridge and Random Forest algorithm gives a better prediction with respect to Accuracy, MAE, and RMSE than the Linear and polynomial regression approaches**.**

References

[1] H. M. Al-Hamadi “Long-Term Electric Power Load Forecasting UsingFuzzy 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

http://doi.org/10.36893/JNAO.2023.V14I2.0142-0148

[3] Zone-Ching Lin, Wen-Jang Wu, Multiple Linear Regression Analysis of the Overlay Accuracy Model Zone, IEEE Trans. on Semiconductor Manufacturing, vol. 12, no. 2, pp. 229 237, May 1999

[4] O. Ajao Isaac, A. Abdullahi Adedeji, I. Raji Ismail, Polynomial Regression Model of Making Cost Prediction in Mixed Cost Analysis, Int. Journal on Mathematical Theory and Modeling, vol. 2, no. 2, pp. 14 23, 2012.

[5] Xinqing Shu, Pan Wang, An Improved Adaboost Algorithm based on Uncertain Functions, Proc. of Int. Conf. on Industrial Informatics Computing Technology, Intelligent Technology, Industrial Information Integration, Dec. 2015.

http://doi.org/10.36893/JNAO.2023.V14I2.0142-0148