Sales Prediction Using Machine Learning

Premnath S P

Assistant Professor,

Department of Electronics and Communication Engineering,

Sri Krishna College of Engineering and Technology,

Coimbatore, India.

[premnathsp@skcet.ac.in](mailto:premnathsp@skcet.ac.in)

Uma M

UG Research Scholar,

Department of Electronics and Communication Engineering,

Sri Krishna College of Engineering and Technology,

Coimbatore, India.

[19euec167@skcet.ac.in](mailto:19euec167@skcet.ac.in)

Valarmathi E

UG Research Scholar,

Department of Electronics and Communication Engineering,

Sri Krishna College of Engineering and Technology,

Coimbatore, India.

[19euec168@skcet.ac.in](mailto:19euec168@skcet.ac.in)

Varthayini R

UG Research Scholar,

Department of Electronics and Communication Engineering,

Sri Krishna College of Engineering and Technology,

Coimbatore, India.

[19euec170@skcet.ac.in](mailto:19euec170@skcet.ac.in)

***Abstract***— **The fundamental concepts of sellers and customers are supply, demand. It is quite critical to predict demand accurately sales prediction is predicting the sales for Big mart companies through which the performance is predicted by changing the business model. In this paper, a new approach for demand prediction for Big Mart companies is proposed. The business model used by the Big Mart companies includes many outlets that sell the product which are same throughout the country at the same time where the company operates a market place model. This actually helps in business to predict realistic sales which helps in increasing the business .In this study linear regression ,gradient boosting and random forest are used to predict sales . This method is carried out on data from various companies where those are processed according the algorithms in machine learning and final data is extracted which helps in sales prediction.**

Keywords— Machine learning, Sales Prediction.

# Introduction

In current scenario, there are huge shopping centers emerging day by day such as malls ,shopping marts etc. the main motive of them is to develop business and earn money. The most important factor which helps in developing business is by predicting the sales. Various big marts has started recording their data which are related to sales of products with various factors which is either independent or dependent. A detailed knowledge of the past data can lead one to predict about the future market which will automatically lead to success regardless of any external factors. In this paper three algorithms of machine learning are used to compare them and look for the more accurate prediction. linear regression, gradient boosting and random forest are used. Initially the data mining techniques take place where the data is discovered, transformed, the feature development is done, model creation and then goes to testing. Later the algorithm is being trained to construct a model on the data which is discovered. This model is finally being used in forecasting the results. Three functions are combined in this system. In order to help such shopping centers, such models are built using machine learning so that it helps in improving their sales and also helps by saying what changes could be done.

# literature survey

Plenty of studies and researches are being made in big mart sales prediction using machine learning. Machine learning is being chosen for our project because it is such a computer programs which learns by itself and does not need any human interference.

Sales forecasts help in predicting the future sales. This is an important precondition for decision-making of enterprises. It helps in predicting the future sales thus helping in increasing the profit.

A Forecast for Big Mart Sales Based on Random Forests and Multiple Linear Regression (2018) - Kadam, H., Shevade, R., Ketkar, P. and Rajguru. A Forecast for Big Mart Sales Based on Random Forests and Multiple Linear Regression. XG boost Algorithm is used because it gives more accuracy when compared to random forest and linear regression in case of prediction analysis . Forecasting methods and applications (2008)- Makridakis, S., Wheelwrigh.S.C., Hyndman. R.J. Forecasting methods and applications contains Lack of Data and short life cycles. So some of the historical data if stored can help in accurate prediction of uncertain demands in case of consumer oriented markets. Comparison of Different Machine Learning Algorithms for Multiple Regression on Black Friday Sales Data (2018) - C. M. Wu, P. Patil and S. Gunaseelan. Comparison of Different Machine Learning Algorithms for Multiple Regression on Black Friday Neural Network was used by sales data for comparing different algorithms. To overcome such complex models simpler algorithms are used for prediction. Prediction of retail sales of footwear using feed forward and recurrent Neural Networks (2018) - Das, P., Chaudhury Prediction of retail sales of footwear using feed forward and recurrent neural networks neural network for predicting of weekly retail sales was not much efficient , So XG boost is used.

# Proposed system

The following picture Fig. 1 shows the sequence steps and the stages of the proposed prediction process. By

utilizing these steps the big mart sales prediction model is

built. In this flow diagram there are 5 major steps and each

plays a significant role in building the model.

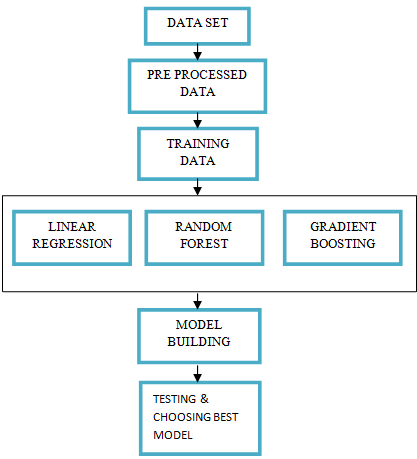


Fig. 1. Block diagram of Big mart sales Prediction.

The description of each block in this diagram is explained below in detail.

## Data Set Description

The data set of a big market is utilized for the prediction model. This dataset was drawn from the internet from Kaggle website. The Fig. 2 shows the attributes/features of the dataset and their description.

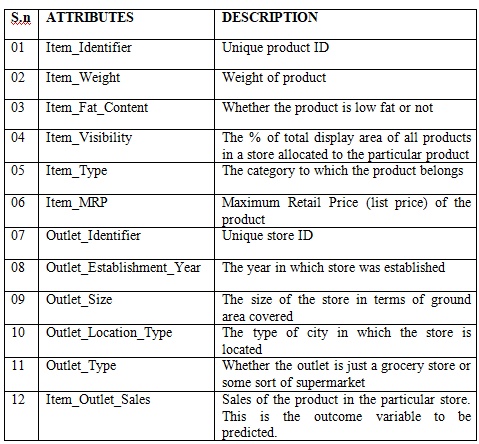


Fig. 2 Attributes and their Description of Dataset

This dataset holds the records of about 8523 items in train data and about 5681 items in test data. There are 12 features. Out of these Item\_Outlet\_Sales is the target which is going to be predicted and all others are the features which help to achieve the prediction. Test data is used to predict the sales.

## Exploratory Data Analysis and Pre Processing

This stage is used to analyze datasets and summarize the main characteristics of the dataset by visualizing things. This helps to understand the basic details which will be helpful in the upcoming stages. It will be more beneficial to merge the train and test data. In EDA, Univariate and bivariate analysis is to be done to get the data hypothesis. In this step we may notice that the categories “LF” ,”low fat” and “Low Fat” are same and “reg” and “Regular” are same and these are repeated. The relation between the bi variable features is analyzed.

The preprocessing step is the main step of predictive analysis. The data set may contain some unwanted values, missing data etc,. The aim of this step is to convert the raw data into usable form which machine learning model could use and provides more efficient and accurate results. The Statistical Values like Mean, Median, mode standard deviation, count of values and maximum value etc.., are found using data.describe() function.

Panda tools is been used to preprocess the data including analysis on the independent variables like checking for null values in each column and then replacing them with appropriate data types, so that model fitting is not hindered from its way to accuracy. The repeated features, missing values and the unwanted columns are corrected in this step.

The missing values are corrected by calculating mean and median of the corresponding features. Now the dataset can be used to train the model to predict Outlet sales.

## Model Building

After the cleaning processes of the data, now the dataset is ready to adapt for a model.

Here the model is built using three algorithms.

-- Linear regression.

-- Random forest regression.

-- Gradient boosting.

To track the machine-learning system on wholesome basis we have used Scikit-Learn here. Algorithms for Predicting the dataset are discussed below.

#### 1)Linear Regression

Linear regression tries to model the relationship between two variables by fitting a linear equation to observed data. This regression carry out the task to forecast a dependent variable value (y) based on a given independent variable (x). So, this regression technique realizes out a linear relationship between x (input) and y (output). The concept of linear regression is to find the best-fitting straight line through the given points. The formula for linear regression equation is given by:

y = a + bx + ∈ (1) Where,

y= the predicted value.

x= Independent variable.

a= Y- Intercept of line.

b= Slope of the line.

∈= the difference between actual and predicted values.

There is always a difference between actual and predicted values which is irreducible error thus we cannot rely completely on the predicted results by the learning algorithm.

*2) Random Forest Regression*

Random forest regression is a supervised learning algorithm. This algorithm is used for both classification and regression problems. The Random forest contains a ‘n’ number of decision trees on various groups of the provided dataset and extracts the mean to improve the prediction accuracy of the dataset. It works in two phase. Firstly to create the random forest by uniting N decision trees and secondly to provide prediction for several tree that are generated in first phase. This algorithm is able to handle large datasets with more dimensions.

*3) Gradient Booster*

Boosting is the process of converting weak learner to strong learner. The gradient boosted trees method is also known as ensemble learning method. In gradient boosting, learning happens by optimizing the loss function. This algorithm uses two types of base estimators one is average type model and second is decision tree with full depth.

The steps in gradient boosting are

i) The average model is created

ii) Residuals are calculated from actual and average prediction.

iii) Create another model RM1 which will take residuals as target.

iv) The new predicted residual value is obtained, and then calculates new predicted value.

v) Again the residuals are obtained (Actual – Predicted) and new model RM2 will fit again on the residues as target and will predict new residues.

Gradient boosting uses the Gradient (loss) of the model as an input to its next model and it goes on.

# EVALUATION METRICS

Evaluating the model is the major part of building a decisive model. The evaluation metric defines a model’s results. The significant feature of the evaluation metrics is the capability to differentiate between the outcomes of the model.

Here we have utilized Root Mean Squared Error (RMSE) metric for the process of evaluation. The formula of RMSE is given by

RMSE=qPN (2) Where, N= number of considerations. qP= quantitative precipitation.

# RESULTS

From the algorithms used for the model, the efficient model is being chosen and the output is fabricated from the chosen algorithm with high efficiency. The fact is that the algorithm with lower RMSE value will provide predicted output with high accuracy.

Out of the three algorithms used the highest accuracy of prediction is given by gradient booster algorithm with accuracy as 0.69 which is maximum and RMSE value as 10.343 which is minimum.

Where accuracy = (no of correct predictions/total no of predications),

RMSE = √1-(r^2)SDy

TABLE I

COMPARISON OF RESULTS OF THE ALGORITHMS

|  |  |  |
| --- | --- | --- |
| ALGORITHM | ACCURACY | RMSE |
| Linear Regression | 0.58 | 11.787 |
| Random Forest Regression | 0.649 | 10.821 |
| Gradient boosting algorithm | 0.69 | 10.343 |

Therefore the Gradient booster is been chosen and the Outlet\_sales is being predicted as shown in Fig. 3.

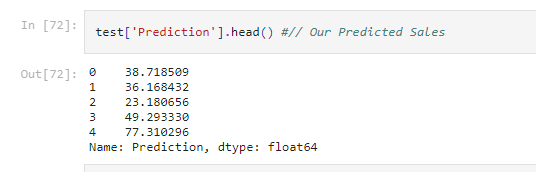


Fig. 3. Predicted Sales.

# CONCLUSION

The main motive of this project is to help in predicting sales and improve the business by giving the ideas of future sales. In this paper we have discussed how different algorithms are being used to build different models. Algorithms like linear regression , gradient boosting and random forest are used . These algorithms are applied to predict the sales of various shopping centers .It is discussed in detail how the models are built and the results are being predicted. Three algorithms of machine learning are used to build models so as to compare them and get the most accurate results among them. It is concluded that the algorithm which gave more accurate results is gradient boosting. There are still further experiments to be done so that there would be more accurate

and efficient results which would be predicted. This prediction would help big marts for predicting their futures sales so that they could improve their business thus increasing profit.

##### References

1. Gopalakrishnan T, Ritesh Choudhary, Sarada Prasad, “Prediction of Sales Value in online shopping using Linear Regression”, 20174th International Conference on Computing Communication and Automation.
2. Kadam, H., Shevade, R., Ketkar, P. and Rajguru.: “A Forecast for Big Mart Sales Based on Random Forests andLinear Regression.” (2018).
3. Kumari Punam, Rajendra Pamula, Praphula Kumar Jain, “A Two-Level Statistical Model for Big Mart Sales Prediction”, 2018 International Conference on Computing, Power and Communication Technologies (GUCON) .
4. Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques .
5. Das, P., Chaudhury: Prediction of retail sales of footwear using feed forward and recurrent neural networks (2018).
6. Natalya V. Razmochaeva, Dmitry M. Klionskiy, Vladimir V. Chernokulsky, “The Investigation of Machine Learning Methods in the Problem of Automation of the Sales Management Business-process”(2014).
7. C. M. Wu, P. Patil and S. Gunaseelan: Comparison of Different Machine Learning Algorithms for Multiple Regression on Black Friday Sales Data (2018).
8. A Forecast for Big Mart Sales Based on Random Forests and Multiple Linear Regression (2018) - Kadam, H., Shevade, R., Ketkar, P. and Rajguru.
9. Quinlan, J. R. (2014). C4. 5: programs for machine learning. Elsevier.
10. Marian Cristian Mihăescu, Carmen-Patricia Sapunaru, “Practical Machine Learning Solution for Increasing Profit”, SAI Intelligent Systems Conference 2013.
11. Forecasting methods and applications (2008) - Makridakis, S., Wheelwrigh.S.C., Hyndman. R.J.
12. Welling, M. (2011). A first encounter with Machine Learning.

Irvine, CA.: University of California, 12.

1. Comparison of Different Machine Learning Algorithms for Multiple Regression onBlack Friday Sales Data (2018) - C. M. Wu, P. Patil and S. Gunaseelan.
2. Prediction of retail sales of footwear using feed forward and recurrent Neural Networks (2018) - Das, P., Chaudhury.