SALES PREDICTION FOR BIG MART OUTLETS

Assistant prof J.Jyothi[1], U.Chandana Madhuri[2], B.Ome Sai Srija[3], J.Sravani[4], P.Mounica[5].
Department of Computer Science and Engineering, Vignan's Nirula Institute Of Technology and Science For Women, Peddapalakaluru road,Guntur, AP, India.
jyothi.jarugula1216@gmail.com

Abstract
Machine Learning is a category of algorithms that allows software applications to become more accurate in predicting outcomes without being explicitly programmed. The basic premise of machine learning is to build models and employ algorithms that can receive input data and use statistical analysis to predict an output while updating outputs as new data becomes available. These models can be applied in different areas and trained to match the expectations of management so that accurate steps can be taken to achieve the organization's target. In this paper, the case of Big Mart, a one-stop shopping center, has been discussed to predict the sales of different types of items and for understanding the effects of different factors on the items' sales. Taking various aspects of a dataset collected for Big Mart, and the methodology followed for building a predictive model, results with high levels of accuracy are generated, and these observations can be employed to take decisions to improve sales.
Keywords: Machine Learning, Sales Prediction, Big Mart, Random Forest, Linear Regression.

Introduction

In today's modern world, huge shopping centers such as big malls and marts are recording data related to sales of items or products with their various dependent or independent factors as an important step to be helpful in prediction of future demands and inventory management. The dataset built with various dependent and independent variables is a composite form of item attributes, data gathered by means of customer, and also data related to inventory management in a data warehouse. The data is thereafter refined in order to get accurate predictions and gather new as well as interesting results that shed a new light on our knowledge with respect to the task's data. This can then further be used for forecasting future sales by means of employing machine learning algorithms such as the random forests and simple or multiple linear regression model.

1.1 Machine Learning The data available is increasing day by day and such a huge amount of unprocessed data is needed to be analysed precisely, as it can give very informative and finely pure gradient results as per current standard requirements. It is not wrong to say as with the evolution of Artificial Intelligence (AI) over the past two decades, Machine Learning (ML) is also on a fast pace for its evolution. ML is an important mainstay of IT sector and with that, a rather central, albeit usually hidden, part of our life [1]. As the technology progresses, the analysis and understanding of data to give good results will also increase as the data is very useful in current aspects.

In machine learning, one deals with both supervised and unsupervised types of tasks and generally a classification type problem accounts as a resource for knowledge discovery. It generates resources and employs regression to make precise predictions about future, the main emphasis being laid on making a system self-efficient, to be able to do computations and analysis to generate much accurate and precise results [2]. By using statistic and probabilistic tools, data can be converted into knowledge. The statistical inferencing uses sampling distributions as a conceptual key [11]. ML can appear in many guises. In this paper, firstly, various applications of ML and the types of data they deal with are discussed. Next, the problem statement addressed through this work is stated in a formalized way. This is followed by explaining the methodology ensued and the prediction results observed on implementation. Various machine learning algorithms include [3]:

- **Linear Regression**: It can be termed as a parametric technique which is used to predict a continuous or dependent variable on basis of a provided set of independent variables. This technique is said to be parametric as different assumptions are made on the basis of dataset.
- **AdaBoost regression**: Ada boost ensemble is fit on all available data, then predict() function can be called to make predictions on new data. Used to boost the performance of any machine learning algorithm.
- **Gradient Boosting Regression:** Gradient
  Boosting is one of the most powerful algorithms in the field of machine learning.
  Used not only for continuous target variable but also categorical target.

1.2 **Problem Statement:** "To find out what role certain properties of an item play and how they affect their sales by understanding Big Mart sales." In order to help Big Mart achieve this goal, a predictive model can be built to find out for every store, the key factors that can increase their sales and what changes could be made to the product or store's characteristics.

2.1 **Dataset and its Preprocessing:** Big Mart's data scientists collected sales data of their 10 stores situated at different locations with each store having 1559 different products as per 2013 data collection. Using all the observations it is inferred what role certain properties of an item play and how they affect their sales. The dataset looks like shown in Fig.2 on using head() function on the dataset variable.

	Item_Identifier	Item_Weight	Item_Fat_Content	Item_Visibility	Item_Type	Ite
0	FDA15	9.30	Low Fat	0.016047	Dairy	24
1	DRC01	5.92	Regular	0.019278	Soft Drinks	4
2	FDN15	17.50	Low Fat	0.016760	Meat	14
3	FDX07	19.20	Regular	0.000000	Fruits and Vegetables	18
4	NCD19	8.93	Low Fat	0.000000	Household	5

Figure: About the attributes in Sales Prediction.

In the raw data, there can be various types of underlying patterns which also gives an in-depth knowledge about subject of interest and provides insights about the problem. But caution should be observed with respect to data as it may contain null values, or redundant values, or various types of ambiguity, which also demands for pre-processing of data. Dataset should therefore be explored as much as possible.
Preprocessing of this dataset includes doing analysis on the independent variables like checking for null values in
The data set consists of various data types from integer to float to object as shown.

```
Item_Identifier              ob
Item_Weight                 flo
Item_Fat_Content             ob
Item_Visibility             flo
Item_Type                    ob
Item_MRP                    flo
Outlet_Identifier            ob
Outlet_Establishment_Year     i
Outlet_Size                  ob
Outlet_Location_Type         ob
Outlet_Type                  ob
Item_Outlet_Sales           flo
dtype: object
```

values for categorical columns. Maximum and minimum values in numerical columns, along with their percentile values for median, plays an important factor in deciding which value to be chosen at priority for

further exploration tasks and analysis. Data types of different columns are used further in label processing and one-hot encoding scheme during model building.

2.2 **Algorithms Employed**:

Linear Regression: It can be termed as parametric technique which is used to predict a continuous or dependent variable on basis of a provided set of independent variables. simple words, weak learners

This technique is said to be parametric as different are converted into strong ones.

assumptions are made on basis Adaboost algorithm also works of data set. on the same principle as

- **AdaBoosting Regression**: boosting, but there is a slight short for Adaptive Boosting, is a difference in working.
  Boosting technique that is used
- **Gradient Boosting**
  as an Ensemble Method

Regression:

Gradient Boosting in Machine Learning. It is called is an iterative functional Adaptive Boosting as the gradient algorithm, i.e an weights are re-assigned to each algorithm which minimizes a loss instance, with higher weights to function by iteratively choosing a incorrectly classified instances. function that points towards the Boosting is used to reduce bias negative gradient; a weak as well as the variance for hypothesis. Gradient Boosting in supervised learning. It works on Classification Over the years, the principle where learners are gradient boosting has found grown sequentially. Except for applications across various the first, each subsequent technical fields. learner is grown from previously grown learners. In

3 OUTPUTS:

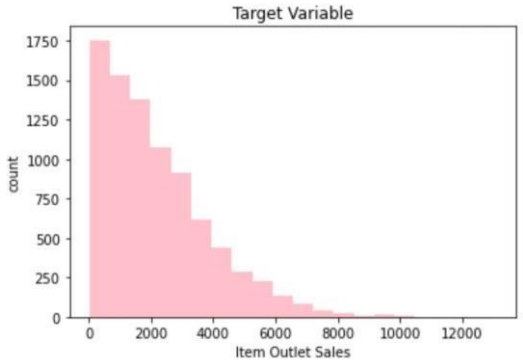

Figure: Out put for Linear Regression.

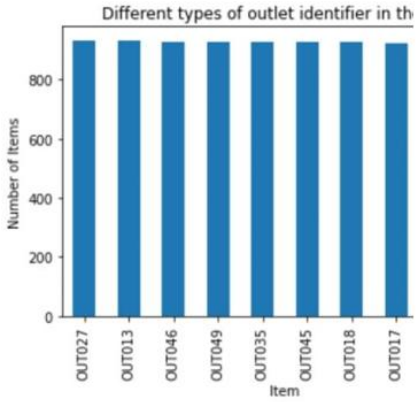

Figure: Output for the AdaBoost Regression

Here, When we use Linear Regression the value of RMSE shows some what high, in order to reduce the RMSE value we do use the other algorithms and compare them.

```
[ ] from sklearn.ensemble import AdaBoostRegressor

    model= AdaBoostRegressor(n_estimators = 100)
    model.fit(x_train, y_train)

    # predicting the test set results
    y_pred = model.predict(x_test)

    # RMSE
    mse = mean_squared_error(y_test, y_pred)
    print("RMSE :", np.sqrt(mse))
```

As, when we use the AdaBoosting Regression the RMSE will be as the high, the accurate algorithm means that which gives less RMSE value, So let's use other algorithm for more accuracy.

```
[ ] from sklearn.ensemble import GradientBoostingReg

    model = GradientBoostingRegressor()
    model.fit(x_train, y_train)

    # predicting the test set results
    y_pred = model.predict(x_test)
    print(y_pred)

    # Calculating the root mean squared error
    print("RMSE :", np.sqrt(((y_test - y_pred)**2).s

    [1674.94481982 1702.98236815 1981.59261558 ...
     2694.79004847]
    RMSE : 1236.2812557949198
```

Figure: Output for the Gradient Boosting Regreesion.

So, finally when we do use gradient boosting algorithm the RMSE value is less when compared to the linear regression, adaboosting regression, so we can say that Gradient Boosting Algorithm best suites for the Sales Prediction Of Big Mart Outlets.

CONCLUSION

And Future Scope In this paper, basics of machine learning and the associated data processing and modelling algorithms have been described, followed by their application for the task of sales prediction in Big Mart shopping centers at different locations. On implementation, the prediction results show the correlation among different attributes considered and how a particular location of medium size recorded the highest sales, suggesting that other shopping locations should follow similar patterns for improved sales. Multiple

instances parameters and various factors can be used to make this sales prediction more innovative and successful. Accuracy, which plays a key role in prediction-based systems, can be significantly increased as the number of parameters used are increased. Also, a look into how the submodels work can lead to increase in productivity of system. The project can be further collaborated in a web-based application

REFERENCE

1. Smola, A., & Vishwanathan, S. V. N. (2008). Introduction to machine learning. Cambridge University, UK, 32, 34.
2. Saltz, J. S., & Stanton, J. M. (2017). An introduction to data science. Sage Publications.
3. Shashua, A. (2009). Introduction to machine learning: Class notes 67577. arXiv preprint arXiv:0904.3664.
4. MacKay, D. J., & Mac Kay, D. J. (2003). Information theory, inference and learning algorithms. Cambridge university press.

