IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)
An International Open Access, Peer-reviewed, Refereed Journal

BIG MART SALES PREDICTION USING MACHINE LEARNING AND PYTHON

[1]E. Durga Prasad, [2]U. Harsha Vardhan, [3]M. Dhiraj Yadav, [4]Dr. Krishna Samalla, [5]Y. Srinivasulu
[1]Student, [2]Student, [3]Student, [4]Professor, [5]SAssociate Professor
[1,2,3,4,5]Department of Electronics and Communication Engineering
[1,2,3,4,5]Sreenidhi Institute of Science and Technology, Ghatkesar, Hyderabad, India

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, Linear Regression, XG Boost, Prediction.

I. 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.

II. LITERATURE SURVEY

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 used Random Forest and Linear Regression for prediction analysis which gives less accuracy. To overcome this, we canuse XG boost Algorithm which will give more accuracy and will be more efficient.

Forecasting methods and applications (2008) Makridakis, S., Wheelwrigh. S. C., Hyndman. R.J. Forecasting methods and applications contain a Lack of Data and short life cycles. So some of the datalike historical data, and consumer- oriented markets face uncertain demands and can be predicted for accurate results.

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 onBlack Friday Sales Data Used Neural Network for comparison of different algorithms. To overcome this Complex model like neural networks is used for comparison between different algorithms which is not efficient so we can use simpler algorithms for prediction. Prediction of retail sales of footwear using feed-forward and recurrent Neural Networks (2018)
by Das, P., Chaudhury. Prediction of retail sales of footwear using feed-forward and recurrent neural networks used neural networks for prediction of sales. Using a neural network for predicting weekly retail sales, is not efficient, So XG boost can work efficiently.

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.

III. PROPOSED SYSTEM

Data Processing and Methodology

a. Data Collection: We have collected the data securely in accordance with an agreed methodology. The procedure for the collected data may differ from client to client and is dependent on the type, quantity, availability, and need of data.

b. Data Cleaning and Preprocessing: The collected data is passed through a 'cleaning' process, so as to make sure that the data is segregated properly and identified gaps in the data are filled with the appropriate information, making data compatible and also fixing errors in storage systems which can cause data redundancy.

c. Data Modeling: This is primarily a process in which the given dataset and the objects in it are analyzed to get a clear view of the requirements that may help us support our business model. Based on the analysis of patterns present in the data, models are then created on the established flow of the project. This flow offers better assistance in the utilization of the previously agreed upon the semi-formal model that showcases the features of the project. It also provides guidance to follow the relation between the data objects and other objects.

d. Data Prediction: Machine Learning prediction models are trained in this process and then later on evaluated using the data. This will then be applied to the preprocessed dataset. Some of the Models to be used for the prediction are:

- Linear Regression
- Random Forest
- Decision tree
- XG Boost Regressor

e. Data Visualization: Data Analyzed is then further picturized for customers andconclusions and take effective decisions.

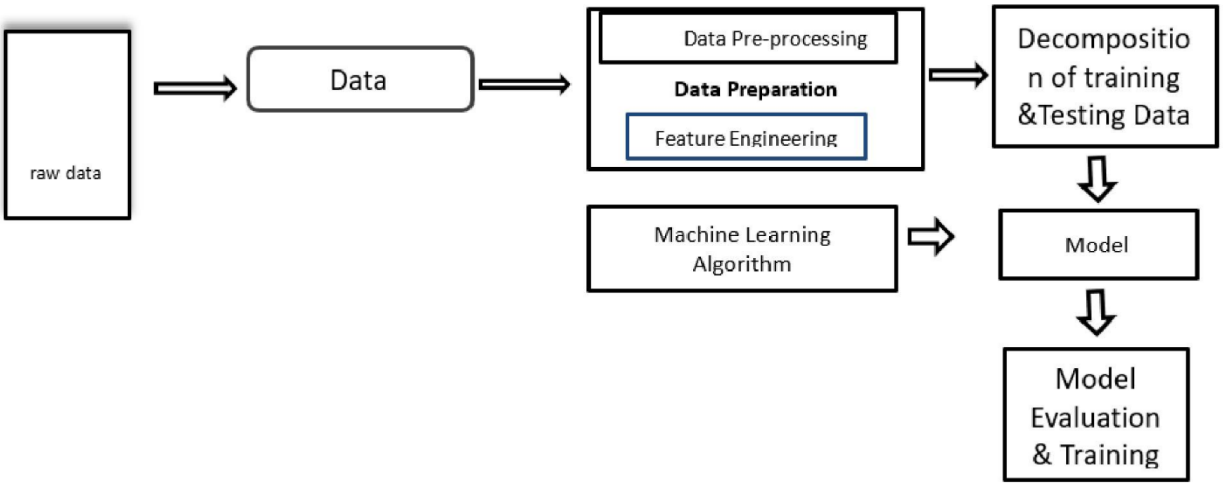

raw data
Data
Data Pre-processing
Data Preparation
Feature Engineering
Decompositio n of training &Testing Data
Machine Learning Algorithm
Model
Model Evaluation & Training

Fig:-Working procedure of proposed model

The figure represents the use case model of our system "Big Mart Sales Prediction Using Machine Learning" to find out the sales of each product at a particular store. Using this model, Big Mart will try to understand the properties of products and stores which play a key role in increasing sales. 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.

IV . METHODOLOGY

The steps followed in this work, right from the dataset preparation to obtaining results are represented in Fig.1.

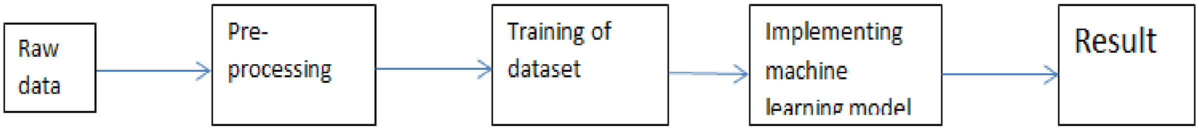
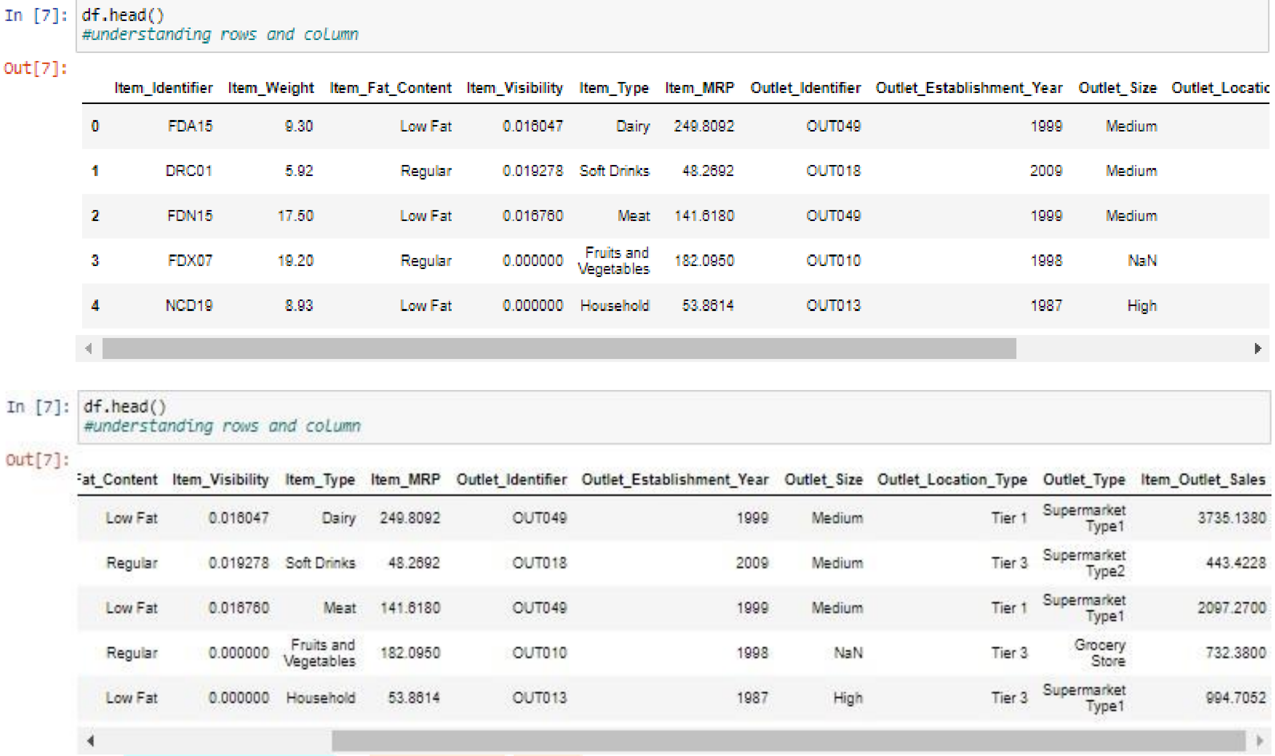


Fig1: Steps followed for obtaining results

Big Mart's data scientists collected sales data of their 10 stores situated at different locationswith 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.



```
In [7]: df.head()
        #understanding rows and column
Out[7]:
```

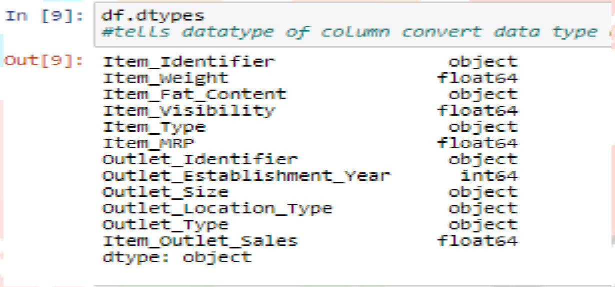
	Item_Identifier	Item_Weight	Item_Fat_Content	Item_Visibility	Item_Type	Item_MRP	Outlet_Identifier	Outlet_Establishment_Year	Outlet_Size	Outlet_Locatic
0	FDA15	9.30	Low Fat	0.016047	Dairy	249.8092	OUT049	1999	Medium	
1	DRC01	5.92	Regular	0.019278	Soft Drinks	48.2692	OUT018	2009	Medium	
2	FDN15	17.50	Low Fat	0.016760	Meat	141.6180	OUT049	1999	Medium	
3	FDX07	19.20	Regular	0.000000	Fruits and Vegetables	182.0950	OUT010	1998	NaN	
4	NCD19	8.93	Low Fat	0.000000	Household	53.8614	OUT013	1987	High	

```
In [7]: df.head()
        #understanding rows and column
Out[7]:
```

:at_Content	Item_Visibility	Item_Type	Item_MRP	Outlet_Identifier	Outlet_Establishment_Year	Outlet_Size	Outlet_Location_Type	Outlet_Type	Item_Outlet_Sales
Low Fat	0.016047	Dairy	249.8092	OUT049	1999	Medium	Tier 1	Supermarket Type1	3735.1380
Regular	0.019278	Soft Drinks	48.2692	OUT018	2009	Medium	Tier 3	Supermarket Type2	443.4228
Low Fat	0.016760	Meat	141.6180	OUT049	1999	Medium	Tier 1	Supermarket Type1	2097.2700
Regular	0.000000	Fruits and Vegetables	182.0950	OUT010	1998	NaN	Tier 3	Grocery Store	732.3800
Low Fat	0.000000	Household	53.8614	OUT013	1987	High	Tier 3	Supermarket Type1	994.7052

Fig2: Screenshot of Dataset

The data set consists of various data types from integer to float to object as shown in Fig.3.

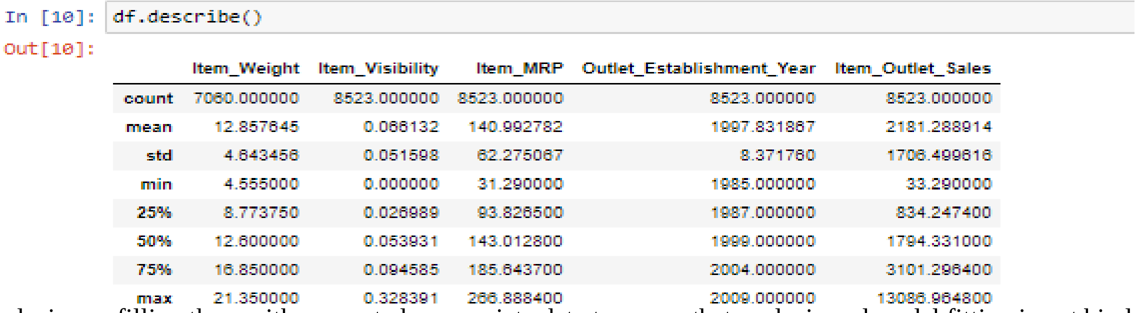


```
In [9]: df.dtypes
        #tells datatype of column convert data type
Out[9]: Item_Identifier               object
        Item_Weight                  float64
        Item_Fat_Content              object
        Item_Visibility              float64
        Item_Type                     object
        Item_MRP                     float64
        Outlet_Identifier             object
        Outlet_Establishment_Year      int64
        Outlet_Size                   object
        Outlet_Location_Type          object
        Outlet_Type                   object
        Item_Outlet_Sales            float64
        dtype: object
```

Fig3: Various datatypes used in the Dataset

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.Various factors important by statistical means like mean, standard deviation, median, count of values and maximum value etc. are shown in Fig.4 for numerical variables of our dataset.

Preprocessing of this dataset includes doing analysis on the independent variables like checking for null values in each column and then replacing or filling them with supported appropriate data types, so that analysis and model fitting is not hindered from its way to accuracy. Shown above are some of the representations obtained by using Pandas tools which tells about variable count for numerical columns and modal 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



```
In [10]: df.describe()
Out[10]:
```

	Item_Weight	Item_Visibility	Item_MRP	Outlet_Establishment_Year	Item_Outlet_Sales
count	7060.000000	8523.000000	8523.000000	8523.000000	8523.000000
mean	12.857645	0.066132	140.992782	1997.831867	2181.288914
std	4.643456	0.051598	62.275067	8.371760	1706.499616
min	4.555000	0.000000	31.290000	1985.000000	33.290000
25%	8.773750	0.026989	93.826500	1987.000000	834.247400
50%	12.600000	0.053931	143.012800	1999.000000	1794.331000
75%	16.850000	0.094585	185.643700	2004.000000	3101.296400
max	21.350000	0.328391	266.888400	2009.000000	13086.964800

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

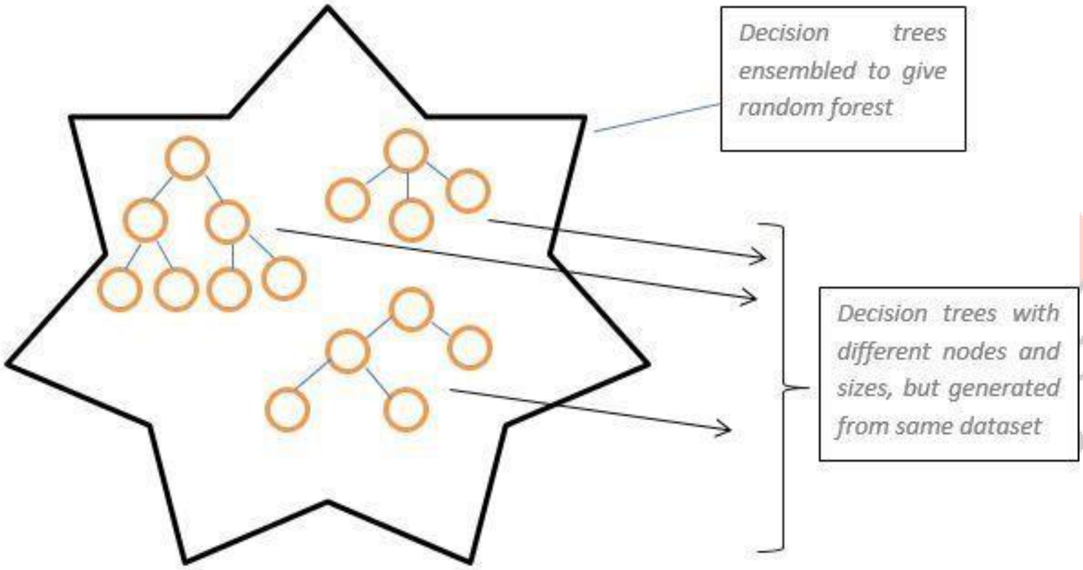
ALGORITHM EMPLOYED :

Scikit-Learn can be used to track machine-learning system on wholesome basis [12]. Algorithms employed for predicting sales for this dataset are discussed as follows:

Random Forest Algorithm:

Random forest algorithm is a very accurate algorithm to be used for predicting sales. It is easy to use and understand for the purpose of predicting results of machine learning tasks. In sales prediction, random forest classifier is used because it has decision tree like hyperparameters. Thetree model is same as decision tool. Fig.5 shows the relation between decision trees and random forest. To solve regression tasks of prediction by virtue of random forest, the *sklearn.ensemble* library's random forest regressor class is used. The key role is played by the parameter termed as*n_estimators* which also comes under random forest regressor. Random forest can be referred to as a meta-estimator used to fit upon numerous decision trees (based on classification) by taking the dataset's different sub-samples. *min_samples_split* is taken as the minimum number when splitting an internal node if integer number of minimum samples are considered. A split's qualityis measured using *mse* (mean squared error), which can also be termed as feature selection criterion. This also means reduction in variance *mae* (mean absolute error), which is another criterion for feature selection. Maximum tree depth, measured in integer terms, if equals one,then all leaves are pure or pruning for better model fitting is done for all leaves less than *min_samples_split* samples.

Fig5: Relation between Decision Trees and Random Forest


Decision trees ensembled to give random forest
Decision trees with different nodes and sizes, but generated from same dataset

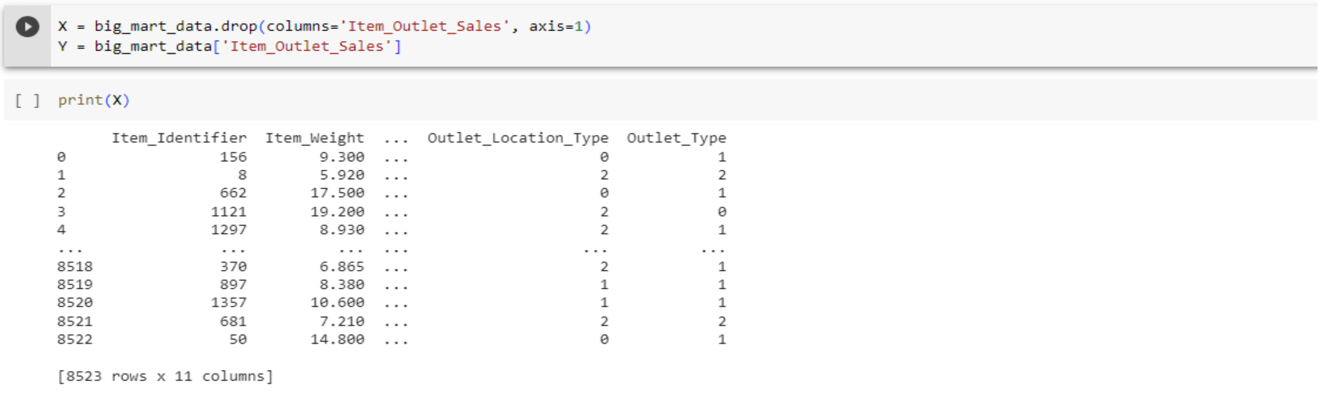
Implementation and Results:

In this section, the programming language, libraries, implementation platform along with thedata modeling and the observations and results obtained from it are discussed.

Implementation Platform and Language :

Python is a general purpose, interpreted-high level language used extensively nowadays for solving domain problems instead of dealing with complexities of a system. It is also termed asthe 'batteries included language' for programming. It has various libraries used for scientific purposes and inquiries along with number of third-party libraries for making problem solving efficient. In this work, the Python libraries of Numpy, for scientific computation, and Matplotlib, for 2D plotting have been used. Along with this, Pandas tool of Python has been employed for carrying out data analysis. Random forest regressor is used to solve tasks by ensembling random forest method.

Splitting features and Target



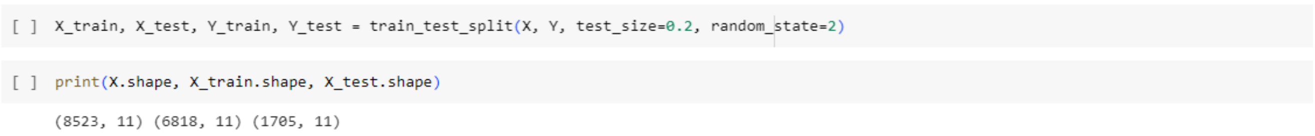
```
X = big_mart_data.drop(columns='Item_Outlet_Sales', axis=1)
Y = big_mart_data['Item_Outlet_Sales']
```

```
[ ] print(X)

          Item_Identifier  Item_Weight  ...  Outlet_Location_Type  Outlet_Type
    0                 156        9.300  ...                     0            1
    1                   8        5.920  ...                     2            2
    2                 662       17.500  ...                     0            1
    3                1121       19.200  ...                     2            0
    4                1297        8.930  ...                     2            1
    ...               ...          ...  ...                   ...          ...
    8518              370        6.865  ...                     2            1
    8519              897        8.380  ...                     1            1
    8520             1357       10.600  ...                     1            1
    8521              681        7.210  ...                     2            2
    8522               50       14.800  ...                     0            1

    [8523 rows x 11 columns]
```

Splitting the data into Training data & Testing Data



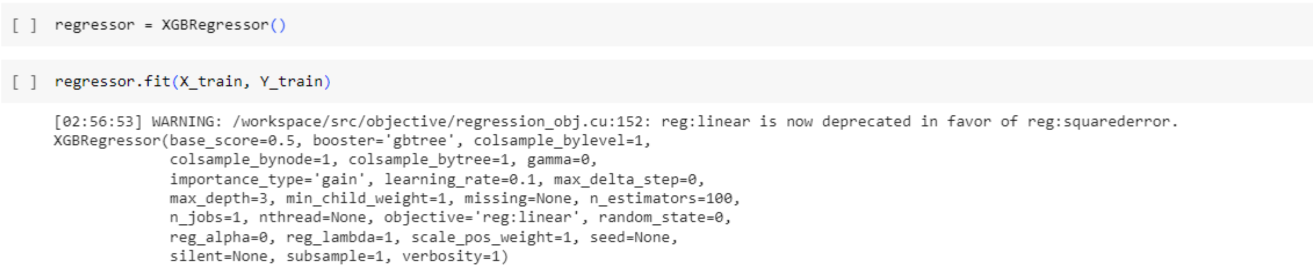
```
[ ] X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, random_state=2)
```

```
[ ] print(X.shape, X_train.shape, X_test.shape)

    (8523, 11) (6818, 11) (1705, 11)
```

Machine Learning Model Training

XGBoost Regressor



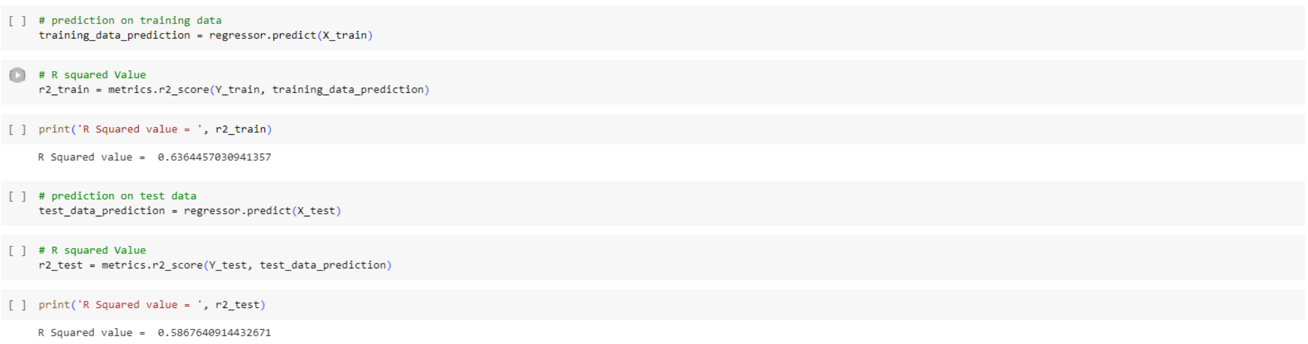
```
[ ] regressor = XGBRegressor()
```

```
[ ] regressor.fit(X_train, Y_train)

    [02:56:53] WARNING: /workspace/src/objective/regression_obj.cu:152: reg:linear is now deprecated in favor of reg:squarederror.
    XGBRegressor(base_score=0.5, booster='gbtree', colsample_bylevel=1,
                 colsample_bynode=1, colsample_bytree=1, gamma=0,
                 importance_type='gain', learning_rate=0.1, max_delta_step=0,
                 max_depth=3, min_child_weight=1, missing=None, n_estimators=100,
                 n_jobs=1, nthread=None, objective='reg:linear', random_state=0,
                 reg_alpha=0, reg_lambda=1, scale_pos_weight=1, seed=None,
                 silent=None, subsample=1, verbosity=1)
```

Result :

Evaluation



```
[ ] # prediction on training data
    training_data_prediction = regressor.predict(X_train)
```

```
    # R squared Value
    r2_train = metrics.r2_score(Y_train, training_data_prediction)
```

```
[ ] print('R Squared value = ', r2_train)

    R Squared value =  0.6364457030941357
```

```
[ ] # prediction on test data
    test_data_prediction = regressor.predict(X_test)
```

```
[ ] # R squared Value
    r2_test = metrics.r2_score(Y_test, test_data_prediction)
```

```
[ ] print('R Squared value = ', r2_test)

    R Squared value =  0.5867640914432671
```

Conclusion and Future Scope:

In this paper, basics of machine learning and the associated data processing and modeling 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 sub-models work can lead to increase in productivity of system. The project can be further collaborated in a web-based application or in any device supported with an in-built intelligence by virtue of Internet of Things (IoT), to be more feasible for use. Various stakeholders concernedwith sales information can also provide more inputs to help in hypothesis generation and more instances can be taken into consideration such that more precise results that are closer to real world situations are generated. When combined with effective data mining methods and properties, the traditional means could be seen to make a higher and positive effect on the overalldevelopment of corporation's tasks on the whole. One of the main highlights is more expressive regression outputs, which are more understandable bounded with some of accuracy. Moreover, the flexibility of the proposed approach can be increased with variants at a very appropriate stageof regression model-building. There is a further need of experiments for proper measurements of both accuracy and resource efficiency to assess and optimize correctly.

References:

[1] Smola, A., & Vishwanathan, S. V. N. (2008). Introduction to machine learning. *CambridgeUniversity, 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 preprintarXiv:0904.3664*.

[4] MacKay, D. J., & Mac Kay, D. J. (2003). *Information theory, inference and learning algorithms*.Cambridge university press.

[5] Daumé III, H. (2012). A course in machine learning. *Publisher, ciml. info*, *5*, 69.

[6] Quinlan, J. R. (2014). *C4. 5: programs for machine learning*. Elsevier.

[7] Cerrada, M., & Aguilar, J. (2008). Reinforcement learning in system identification. In *ReinforcementLearning*. IntechOpen.

[8] Welling, M. (2011). A first encounter with Machine Learning. *Irvine, CA.: University ofCalifornia*, *12*.

[9] Learning, M. (1994). Neural and Statistical Classification. *Editors D. Mitchie et. al*, 350.

[10] Mitchell, T. M. (1999). Machine learning and data mining. *Communications of the ACM*, *42*(11), 30-36.

[11] Downey, A. B. (2011). *Think stats*. " O'Reilly Media, Inc.".