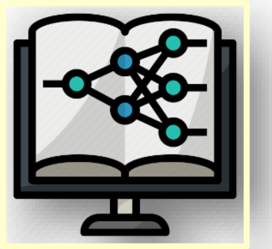
Kingdom of Saudi Arabia
Ministry of Education
College of computer
Department of Information Technology

المملكة العربية السعودية
وزارة التعليم
كلية الحاسب
قسم تقنية المعلومات



DATA SCIENCE FOR INTERNET OF THINGS

IT351

REPORT

(Sales Prediction of BigMart)

Under the supervision of:

Dr. Amal Alshargabi

Names of Students:

Ghada Aldabayan	371201914
Khuzama Alsalem	362206020
Ruba Sanad Alharbi	362216679

SEMESTER

1441-2020

INDEX

- ## Problem (Problem specification and goal)

Bigmart is a big supermarket chain, with stores all around the country and its current board set out a challenge to all Data Scientist out there to help them create a model that can predict the sales, per product, for each store. BigMart has collected sales data from the year 2013, for 1559 products across 10 stores in different cities. With this information the corporation hopes we can identify the products and stores which play a key role in their sales and use that information to take the correct measures to ensure success of their business, Also it used machine learning to predict Bigmart sales enables the data scientist to do so, as it studies the various patterns per store and per product to give accurate results .The data scientists at BigMart have collected 2013 sales data for 1559 products across 10 stores in different cities. The data also includes certain attributes of each product and store.

There is some problem for examples:

1)Traditional storage can cost lot of money to store big data.

2)Lots of big data is unstructured and Lots of effort.

3)BigMart analysis is not useful in short run. It needs to be analyzed for longer duration to leverage its benefits.

4)BigMart analysis results are misleading sometimes.

5)Speedy updates in big data can mismatch real figures.

The goal is to build a predictive model to find out the sales of each product at a particular store so that it would help the decision makers at BigMart to find out the properties of any product or store, which play a key role in increasing the overall sales.

- Literature review

A literature review surveys books, scholarly articles, and any other sources relevant to a particular issue, area of research, or theory, and by so doing, provides a description, summary, and critical evaluation of these works in relation to the research problem being investigated. Literature reviews are designed to provide an overview of sources you have explored while researching a particular topic and to demonstrate to your readers how your research fits within a larger field of study, may consist of simply a summary of key sources, often within specific conceptual categories , a summary is a recap of the important information of the source. But a synthesis is a re-organization, or a reshuffling, of that information in a way that informs how you are planning to investigate a research problem.

(Summary of three recent Literature Review):

[1] A STUDY ON SELECTION OF LOCATION BY RETAIL CHAIN: BIG MART

(Article in International Journal of Research - GRANTHAALAYAH · January 2019)

The modern-day projects, such as privatized infrastructure projects, have project life that is spread over many years. Projects are becoming larger and more complex. These projects involve the large capital investment, generate unbalanced cash flows, and involve complex contractual agreements. They encounter changing economic and financial situation, face unpredictable political environmental changes. The stability of modern projects is thus, constantly subjected to certain sensitive and volatile, external and internal environments (Mishra and Mallik, 2017). The main aim for this research it to understand how retail chains like Big Mart choose their location in order to maintain its existence in a competitive environment as well as maximize its profit. It is well known that the selection of location is vital for successful operation of stores. Especially in case of retail stores the location has huge importance. A knowhow of the approach towards selection of location can help policy makers in making land use plan for commercial purpose to attract investors. It may also aid similar commercial enterprises during location selection process.

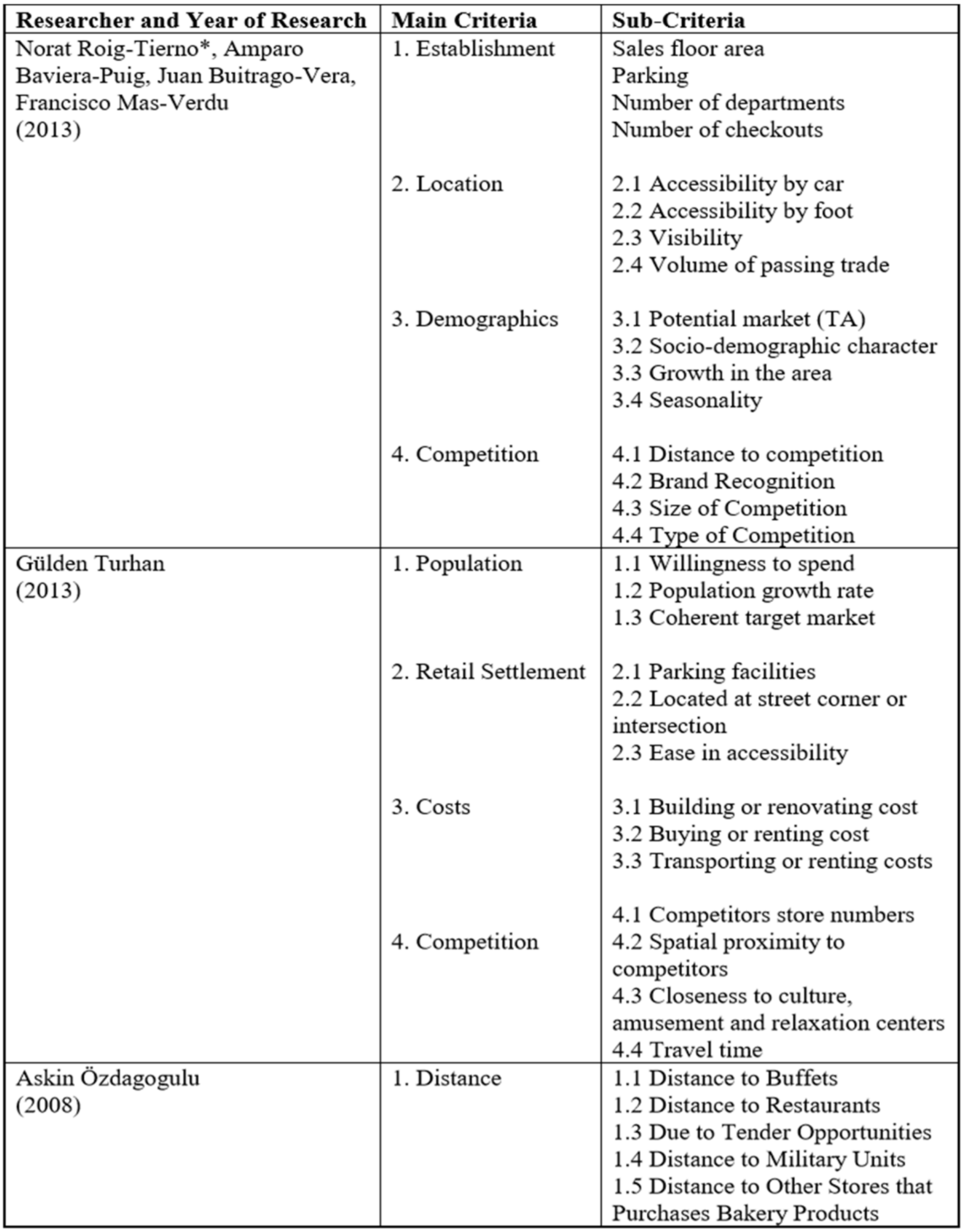
a) Research Question -How do different factors in combination affect the selection of location by retail chain (Big Mart) and what is the predominant factor?

b) Scope and Limitation-The results are overwhelming based on observation as there was no other source of reliable data. Because of the concerned authorities not entertaining any type of questions, the results are highly subjective. Since the results are based on observations of supply chain of Big Mart, it might not be sufficient enough to explain the location selection criteria for other chain outlets.

c) Literature Review -As a well-known real estate mantra is "location, location, location" and this seems to hold particularly true for retail chains as it is cited as one of the most important variable factors affecting "the profitability and sales performance of the management". Although there may be other factors that affect the success or failure of retail stores, the adverse effects of a poor location choice are formidable and non-removable. The selection of a new location for an existing store is a strategic decision requiring a long-term investment. Since new location is an additional cost there is significant financial burden on chains looking at expansion as well as the danger of their image getting damaged if the right decision in store location selection is not implemented (Turhan, 2013).

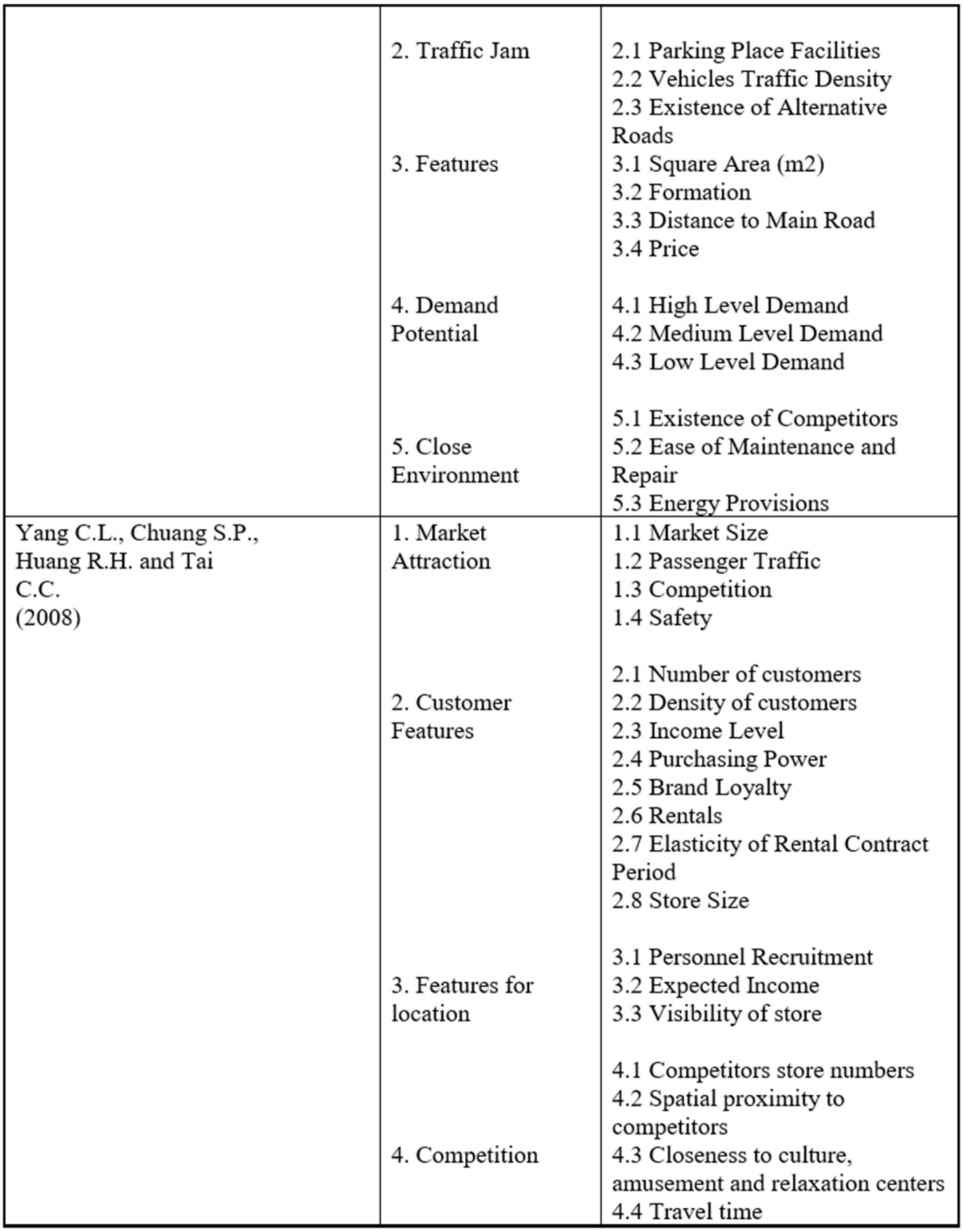
d) Parameters of Location Selection- In terms of population, the amount of money that people are willing to spend for buying retailers' goods, population growth rate and coherent target market in terms of demographics such as gender, education, age, occupation etc are studied. Retailers desire to reach people who are willing to spend money for buying goods. Since the rate of retail expenditures per capita are expected to increase by a rise in population density, the population growth is a good indicator. In addition, the coherent target market is a key to eventual success of any one location as its target market is hugely defined by where it is established .

Table 1: Researches and their criteria



Researcher and Year of Research	Main Criteria	Sub-Criteria
Norat Roig-Tierno*, Amparo Baviera-Puig, Juan Buitrago-Vera, Francisco Mas-Verdu (2013)	1. Establishment	Sales floor area Parking Number of departments Number of checkouts
	2. Location	2.1 Accessibility by car 2.2 Accessibility by foot 2.3 Visibility 2.4 Volume of passing trade
	3. Demographics	3.1 Potential market (TA) 3.2 Socio-demographic character 3.3 Growth in the area 3.4 Seasonality
	4. Competition	4.1 Distance to competition 4.2 Brand Recognition 4.3 Size of Competition 4.4 Type of Competition
Gülden Turhan (2013)	1. Population	1.1 Willingness to spend 1.2 Population growth rate 1.3 Coherent target market
	2. Retail Settlement	2.1 Parking facilities 2.2 Located at street corner or intersection 2.3 Ease in accessibility
	3. Costs	3.1 Building or renovating cost 3.2 Buying or renting cost 3.3 Transporting or renting costs
	4. Competition	4.1 Competitors store numbers 4.2 Spatial proximity to competitors 4.3 Closeness to culture, amusement and relaxation centers 4.4 Travel time
Askin Özdagogulu (2008)	1. Distance	1.1 Distance to Buffets 1.2 Distance to Restaurants 1.3 Due to Tender Opportunities 1.4 Distance to Military Units 1.5 Distance to Other Stores that Purchases Bakery Products

Table 1: Researches and their criteria



	2. Traffic Jam	2.1 Parking Place Facilities 2.2 Vehicles Traffic Density 2.3 Existence of Alternative Roads
	3. Features	3.1 Square Area (m2) 3.2 Formation 3.3 Distance to Main Road 3.4 Price
	4. Demand Potential	4.1 High Level Demand 4.2 Medium Level Demand 4.3 Low Level Demand
	5. Close Environment	5.1 Existence of Competitors 5.2 Ease of Maintenance and Repair 5.3 Energy Provisions
Yang C.L., Chuang S.P., Huang R.H. and Tai C.C. (2008)	1. Market Attraction	1.1 Market Size 1.2 Passenger Traffic 1.3 Competition 1.4 Safety
	2. Customer Features	2.1 Number of customers 2.2 Density of customers 2.3 Income Level 2.4 Purchasing Power 2.5 Brand Loyalty 2.6 Rentals 2.7 Elasticity of Rental Contract Period 2.8 Store Size
	3. Features for location	3.1 Personnel Recruitment 3.2 Expected Income 3.3 Visibility of store
	4. Competition	4.1 Competitors store numbers 4.2 Spatial proximity to competitors 4.3 Closeness to culture, amusement and relaxation centers 4.4 Travel time

e) Methodology -For this research the Analytical Hierarchy Process (AHP) has been used. This method is used for solving multi-criteria decision problems and can effectively handle both qualitative and quantitative data. The primary purpose of the research is to understand the most suitable retail location selection for Big Mart by analysis of the various criteria and attributes involved which has been made easy by the AHP method. It is necessary to compare various factors in the location selection problem. It has been possible to make comparisons and calculations of predetermined criteria. Different criteria and its attributes are arranged in a hierarchal structure. Thus, designed hierarchal structures are compared with each other in sequential order. All factors are divided into four major criteria with the help of literature study i.e. Demography, Competition, Economic factors and Features. Further these criteria are divided into three sub –criteria. These four major criteria are compared with each other to obtain a weighted value for each. Similar method is followed for sub-criteria. Relevant comparisons and results that have been obtained with the help of literature review have been prepared.

f) Analysis and Finding -As discussed in the previous section of this paper the comparison and calculations have been done using AHP by forming hierarchy for main criteria and attributes in sequential order.

1) Demography

- Income of people.
- Growth Rate of people.
- Target Population.

2) Competition

- Distance (Physical and communication).
- Brand recognition.
- Size and type of competition.

3) Economic Factors

- Building and renovating cost
- Buying or renting cost
- Transportation

4) Features

- Accessibility & visibility
- Traffic
- Services (Electricity, Communication)

[2] A FORECAST FOR BIG MART SALES BASED ON RANDOM FORESTS AND MULTIPIE LINEAR REGRESSION

(IJEDR 2018 | Volume 6, Issue 4 | ISSN: 2321-9939)

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.

A) Literature Survey -The method for long term electric power forecasting using long term annual growth factors was proposed. 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.

B) 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.

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-:

- Item_Identifier - Unique identifier for each product.
- Item_Weight – Product weight.
- Item_Fat_Content – Fat content of the product.
- Item_Visibility – Percentage of total display area in a store allocated to the product.
- Item_Type – Product category.
- Item_MRP – List price of the product.
- Outlet_Identifier - Unique identifier for each store.
- Outlet_Establishment_Year – Establishment year for each store.
- Outlet_Size - The size of the store.
- Outlet_Location_Type - The type of city in which the store is located.
- Outlet_Type - Whether the store is a grocery store or a supermarket.
- Item_Outlet_Sales - Sales of the product in each store.

c)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.

$$Y=a+b*X + e$$

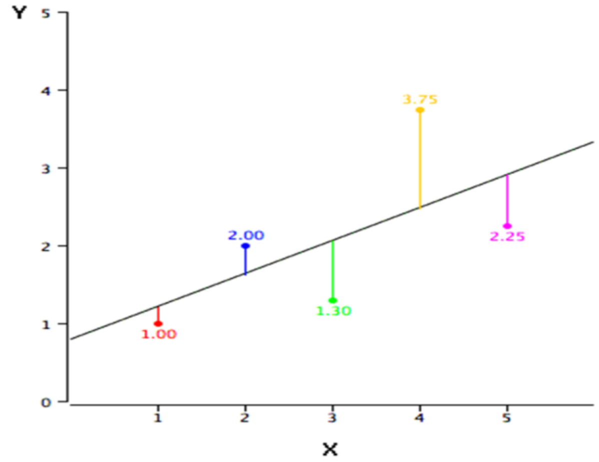


Fig.1. Linear regression

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.

[3] A COMPARATIVE STUDY OF BIG MART SALES PREDICTION

(Conference Paper · September 2019)

Day by day competition among different shopping malls as well as big marts is getting more serious and aggressive only due to the rapid growth of the global malls and on-line shopping. Every mall or mart is trying to provide personal- ized and short-time offers for attracting more customers depending upon the day, such that the volume of sales for each item can be predicted for inventory management of the organization, logistics and transport service. Sales forecasting as well as analysis of sale forecasting has been conducted by many authors as summarized: The statistical and computational methods are studied in also this paper elaborates the automated process of knowledge acquisition. Machine learning is the process where a machine will learn from data in the form of statistically or computationally method and process know- edge acquisition from experiences. Various machine learning (ML) techniques with their applications in different sectors has been presented in.

a) Proposed System-For building a model to predict accurate results the dataset of Big Mart sales undergoes several sequences of steps as mentioned in Figure 2 and in this work, we propose a model using Xgboost technique.

- Dataset Description of Big Mart - Item Fat, Item Type, Item MRP, Outlet Type, Item Visibility, Item Weight, Outlet Identifier, Outlet Size, Outlet Establishment Year, Outlet Location Type, Item Identifier and Item Outlet Sales.

- Data Exploration -In this phase useful information about the data has been extracted from the dataset. That is trying to identify the information from hypotheses vs available data. Which shows that the attributes Outlet size and Item weight face the problem of missing values, also the minimum value of Item Visibility is zero which is not actually practically possible. Establishment year of Outlet varies.

- Data Cleaning- It was observed from the previous section that the attributes Outlet Size and Item Weight has missing values. In our work in case of Outlet Size missing.

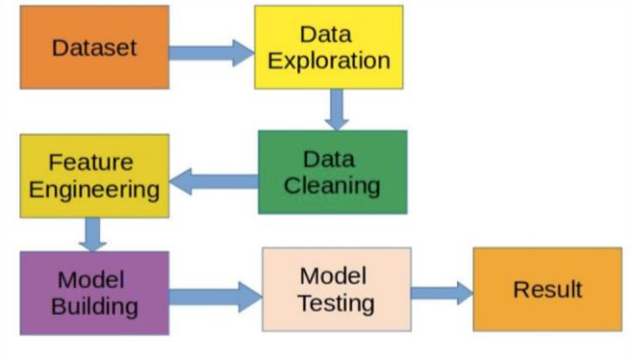
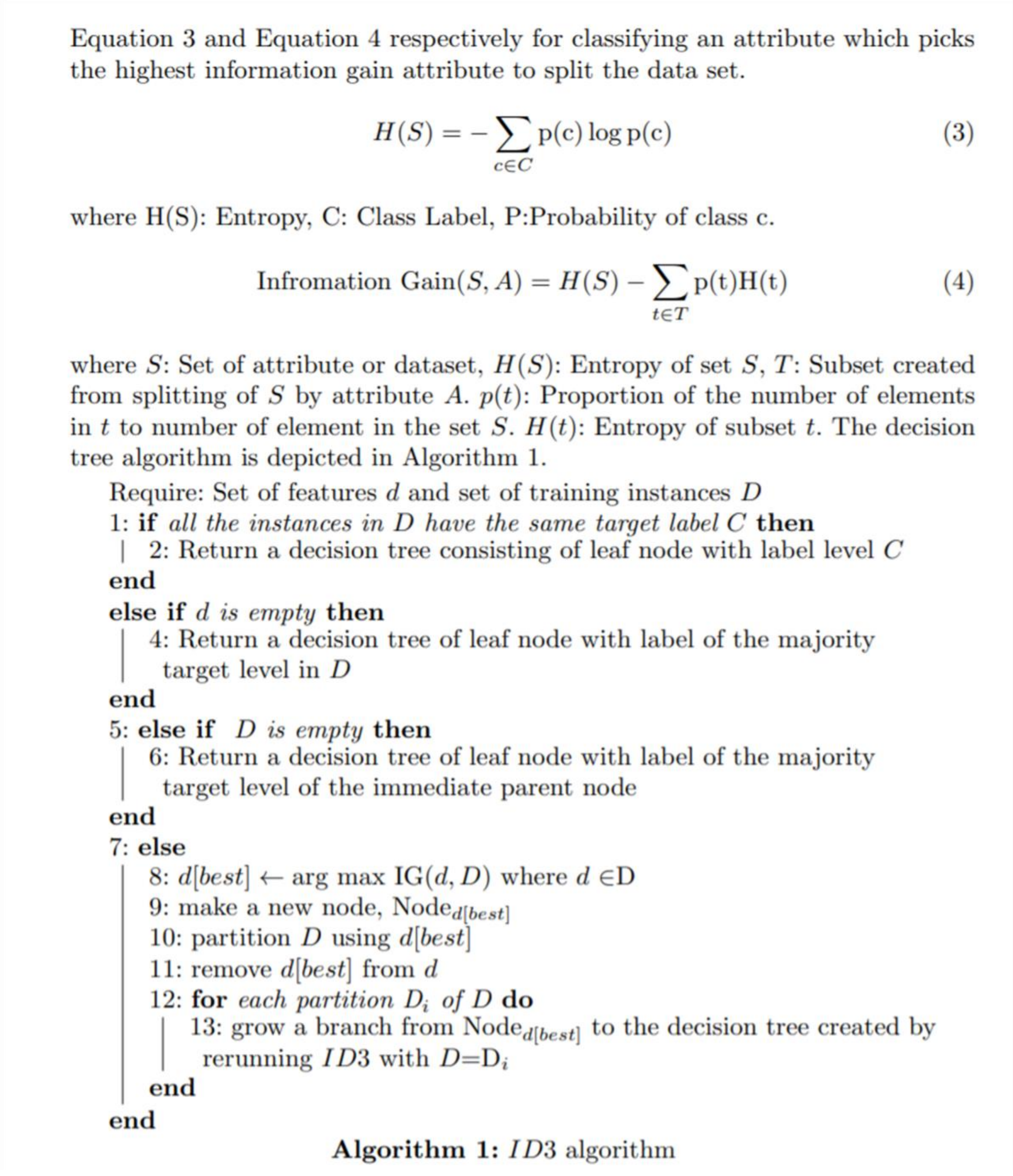


Fig.2. Working procedure of proposed model.

- Feature Engineering -Some nuances were observed in the data-set during data exploration phase. So, this phase is used in resolving all nuances found from the dataset and make them ready for building the appropriate model. During this phase it was noticed that the Item visibility attribute had a zero value, practically which has no sense. So, the mean value item visibility of that product will be used for zero values attribute. This makes all products likely to sell.

b) Model Building-After completing the previous phases, the dataset is now ready to build proposed model. Once the model is build it is used as predictive model to forecast sales of Big Mart. In our work, we propose a model using Xgboost algorithm and compare it with other machine learning techniques like Linear regression, Ridge regression, Decision tree.

Decision Tree: A decision tree classification is used in binary classification problem and it uses entropy and information gain as metric and is defined in, Equation 3 and Equation 4 respectively for classifying an attribute which picks the highest information gain attribute to split the data set.



Equation 3 and Equation 4 respectively for classifying an attribute which picks the highest information gain attribute to split the data set.

$$H(S) = -\sum_{c \in C} \mathrm{p(c)} \log \mathrm{p(c)} \quad (3)$$

where H(S): Entropy, C: Class Label, P:Probability of class c.

$$\text{Infromation Gain}(S, A) = H(S) - \sum_{t \in T} \mathrm{p(t)H(t)} \quad (4)$$

where S: Set of attribute or dataset, $H(S)$: Entropy of set S, T: Subset created from splitting of S by attribute A. $p(t)$: Proportion of the number of elements in t to number of element in the set S. $H(t)$: Entropy of subset t. The decision tree algorithm is depicted in Algorithm 1.

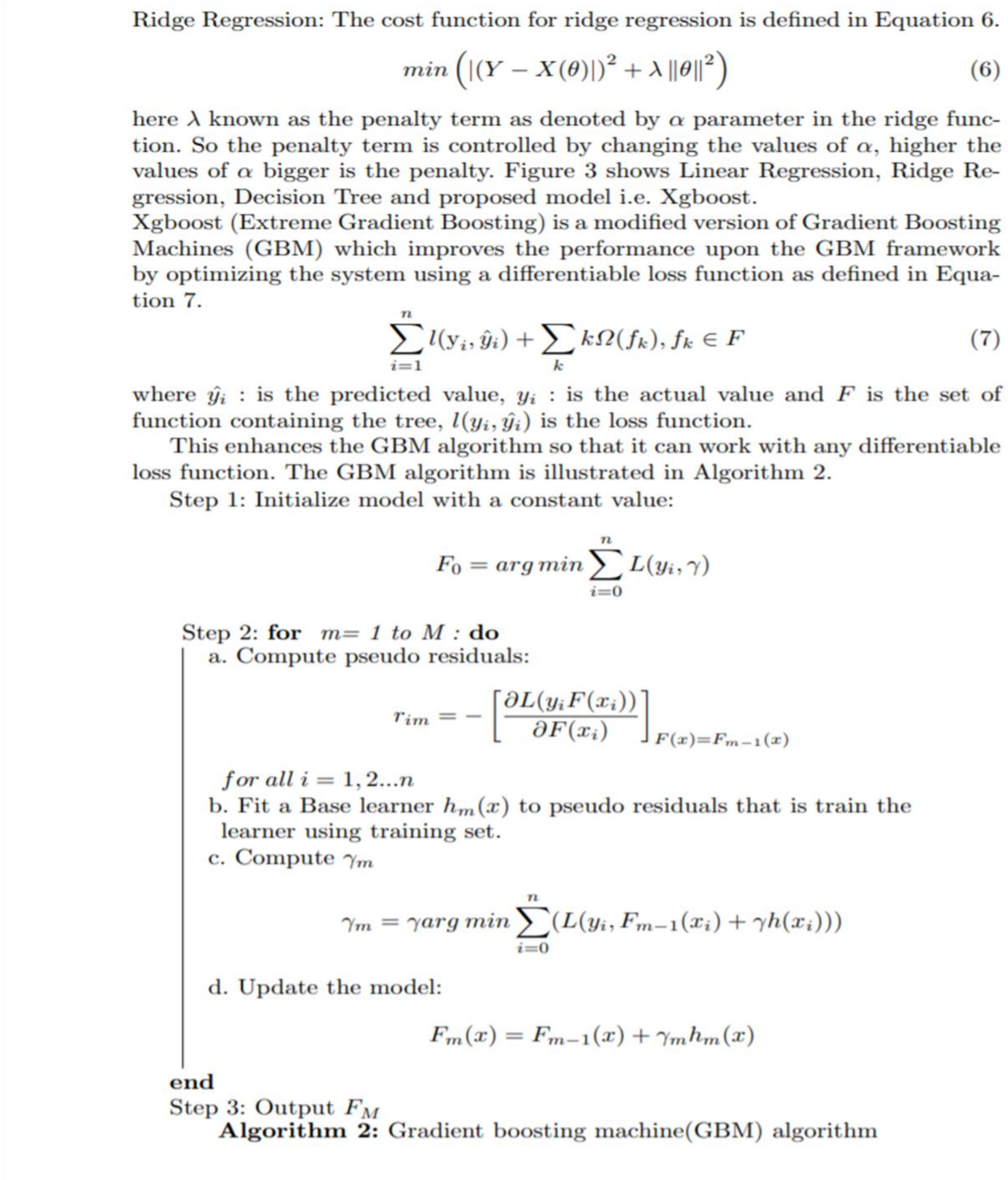
```
Require: Set of features $d$ and set of training instances $D$
if all the instances in $D$ have the same target label $C$ then
    Return a decision tree consisting of leaf node with label level $C$
end
else if $d$ is empty then
    Return a decision tree of leaf node with label of the majority
    target level in $D$
end
else if $D$ is empty then
    Return a decision tree of leaf node with label of the majority
    target level of the immediate parent node
end
else
    $d[best] \leftarrow$ arg max IG$(d, D)$ where $d \in$D
    make a new node, Node$_{d[best]}$
    partition $D$ using $d[best]$
    remove $d[best]$ from $d$
    for each partition $D_i$ of $D$ do
        grow a branch from Node$_{d[best]}$ to the decision tree created by
        rerunning $ID3$ with $D$=D$_i$
    end
end
```

Algorithm 1: $ID3$ algorithm

Linear Regression: A model which create a linear relationship between the dependent variable and one or more independent variable, mathematically linear regression is defined in Equation 5

$$y = \boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x} \tag{5}$$

where y is dependent variable and x are independent variables or attributes. In linear regression we find the value of optimal hyperplane w which corresponds to the best fitting line (trend) with minimum error. The loss function for linear regression is estimated in terms of RMSE and MAE.



Ridge Regression: The cost function for ridge regression is defined in Equation 6.

$$min\left(|(Y - X(\theta)|)^2 + \lambda \|\theta\|^2\right) \tag{6}$$

here λ known as the penalty term as denoted by α parameter in the ridge function. So the penalty term is controlled by changing the values of α, higher the values of α bigger is the penalty. Figure 3 shows Linear Regression, Ridge Regression, Decision Tree and proposed model i.e. Xgboost.
Xgboost (Extreme Gradient Boosting) is a modified version of Gradient Boosting Machines (GBM) which improves the performance upon the GBM framework by optimizing the system using a differentiable loss function as defined in Equation 7.

$$\sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k} k\Omega(f_k), f_k \in F \tag{7}$$

where $\hat{y_i}$: is the predicted value, y_i : is the actual value and F is the set of function containing the tree, $l(y_i, \hat{y}_i)$ is the loss function.

This enhances the GBM algorithm so that it can work with any differentiable loss function. The GBM algorithm is illustrated in Algorithm 2.

```
Step 1: Initialize model with a constant value:
    $F_0 = arg\,min \sum_{i=0}^{n} L(y_i, \gamma)$
Step 2: for $m$= 1 to $M$ : do
    a. Compute pseudo residuals:
        $r_{im} = -\left[\frac{\partial L(y_i F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}$
        $for\ all\ i = 1, 2...n$
    b. Fit a Base learner $h_m(x)$ to pseudo residuals that is train the
       learner using training set.
    c. Compute $\gamma_m$
        $\gamma_m = \gamma arg\,min \sum_{i=0}^{n} (L(y_i, F_{m-1}(x_i) + \gamma h(x_i)))$
    d. Update the model:
        $F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$
end
Step 3: Output $F_M$
```

Algorithm 2: Gradient boosting machine(GBM) algorithm

The Xgboost has following exclusive features:

- Sparse Aware - that is the missing data values are automatic handled.
- Supports parallelism of tree construction.
- Continued training - so that the fitted model can further boost with new data.

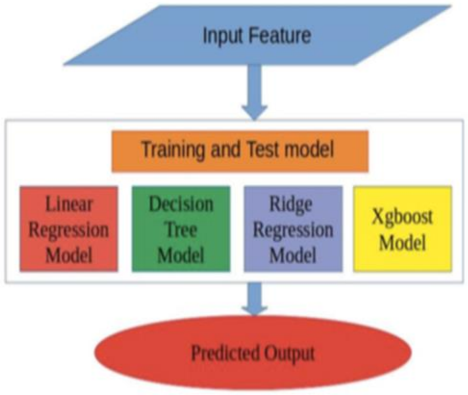


Fig. 3. Framework of proposed model. Model received the input features and split it into training and test set. The trained model is used to predict the future sales.

All models received features as input, which are then segregated into training and test set. The test dataset is used for sales prediction.

- ## Model planning

Machine learning is a large field of study that overlaps with and inherits ideas from many related fields, Types of Learning:

1.Supervised Learning.

2. Unsupervised Learning.

Supervised learning describes a class of problem that involves using a model to learn a mapping between input examples and the target variable, Models are fit on training data comprised of inputs and outputs and used to make predictions on test sets where only the inputs are provided and the outputs from the model are compared to the withheld target variables and used to estimate the skill of the model.

There are two main types of supervised learning problems: they are classification that involves predicting a class label and regression that involves predicting a numerical value.

Classification: Supervised learning problem that involves predicting a class label.

Regression: Supervised learning problem that involves predicting a numerical label.

Some algorithms may be specifically designed for classification (such as logistic regression) or regression (such as linear regression) and some may be used for both types of problems with minor modifications (such as artificial neural networks).

Why we use Regression Analysis?

The importance of regression analysis is that it is all about data: data means numbers and figures that actually define your business. The advantages of regression analysis is that it can allow you to essentially crunch the.

numbers to help you make better decisions for your business currently and into the future. The regression method of forecasting means studying the relationships between data points, which can help you to:

-Predict sales in the near and long term.

-Understand inventory levels.

-Understand supply and demand.

-Review and understand how different variables impact all of these things.

Regression model: is a predictive modelling technique. It estimates the relationship between a dependent (target) and an independent variable(predictor). It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between the dependent and independent variables, they are considering and the number of independent variables being used. Three major uses for regression analysis are-Determining the strength of predictors, Forecasting an effect, Trend for forecasting.

Comparison between Linear and Logistic regression

LINEAR REGRESSION	LOGISTIC REGRESSION
Linear Regression is continuous variables.	Logistic Regression is categorical variables.
In Linear Regression, we predict the value by an integer number.	In Logistic Regression, we predict the value by 1 or 0.
Here solve Regression problems.	Here solve Classification problems.
Here we calculate Root Mean Square Error (RMSE) to predict the next weight value.	Here we use precision to predict the ne

The linear regression model can be represented by the following equation:

$$Y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + ... + \theta_n x_n$$

where,

Y is the predicted value
θ_0 is the bias term.
θ_1,…,θn are the model parameters
x_1, x_2, … ,xn are the feature values.
The above hypothesis can also be represented by

$$Y = \theta^T x$$

Where, θ is the model's parameter vector including the bias term θ_0; x is the feature vector with x_0 =1

Y (pred) = b0 + b1*x

The values b0 and b1 must be chosen so that the error is minimum. If sum of squared error is taken as a metric to evaluate the model, then the goal is to obtain a line that best reduces the error.

$$\text{Error} = \sum_{i=1}^{n}(actual\_output - predicted\_output) ** 2$$

If we don’t square the error, then the positive and negative points will cancel each other out.

For a model with one predictor,

$$b_0 = \bar{y} - b_1\bar{x}$$

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Exploring ‘b1’: If b1 > 0, then x (predictor) and y(target) have a positive relationship. That is an increase in x will increase y, if b1 < 0, then x (predictor) and y(target) have a negative relationship. That is an increase in x will decrease y.

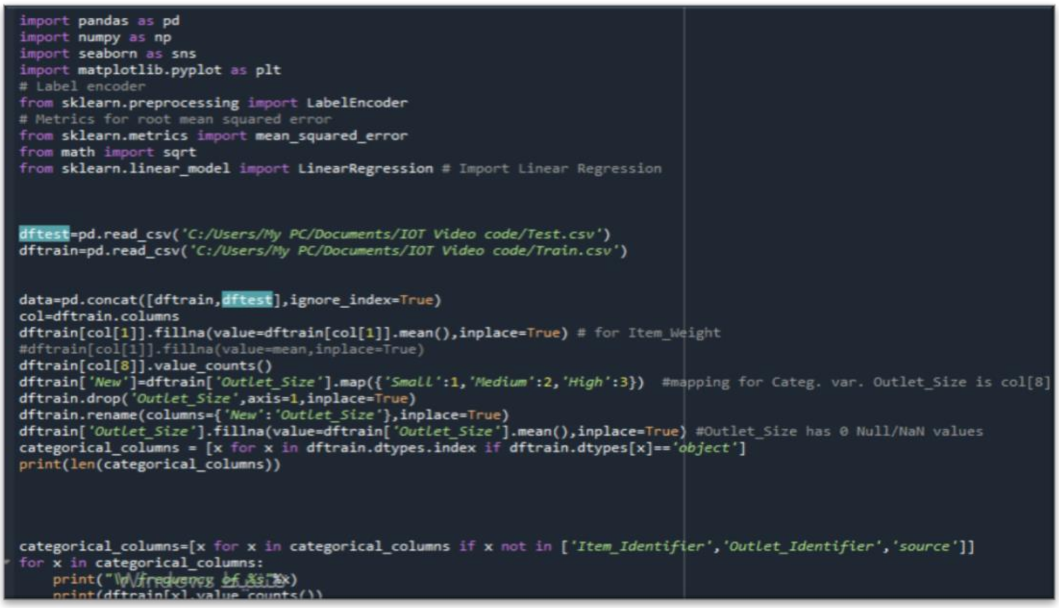
Exploring ‘b0’: If the model does not include x=0, then the prediction will become meaningless with only b0. For example, we have a dataset that relates height(x) and weight(y). Taking x=0 (that is height as 0), will make the equation have only b0 value which is completely meaningless as in real-time height and weight can never be zero. This resulted due to considering the model values beyond its scope.

If the model includes value 0, then ‘b0’ will be the average of all predicted values when x=0. But, setting zero for all the predictor variables is often impossible.

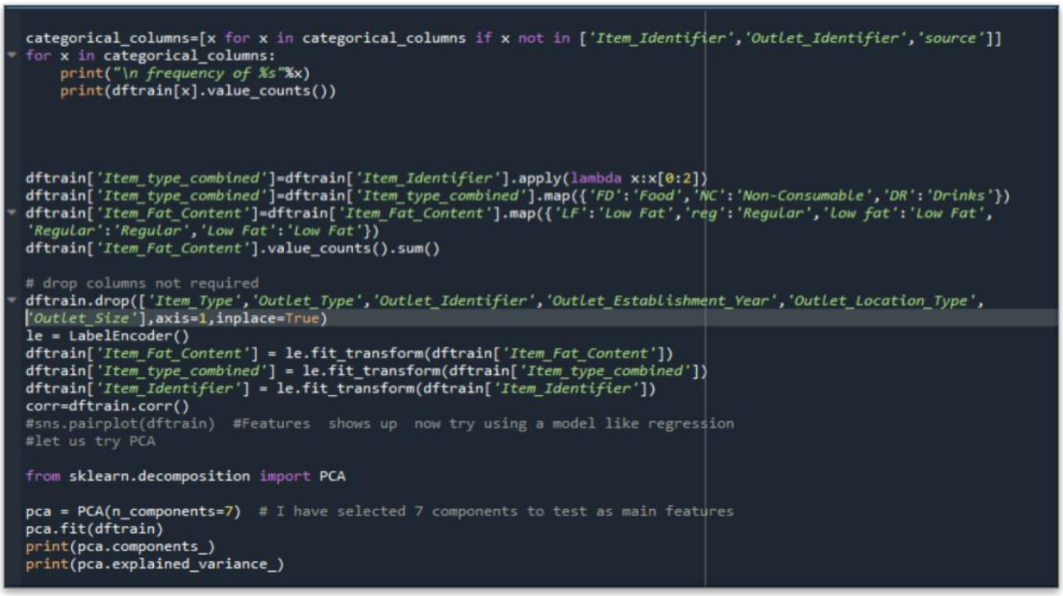
The value of b0 guarantees that the residual will have mean zero. If there is no ‘b0’ term, then the regression will be forced to pass over the origin. Both the regression coefficient and prediction will be biased.

Implementation

After completing the previous phases, the dataset is now ready to build proposed model. Once the model is built it is used as predictive model to forecast sales of BigMart.Here we will import the Linear Regression. We chose, because want to predict the sales of BigMart and it based on supervised learning that mean: helps you to predict outcomes for unforeseen data.



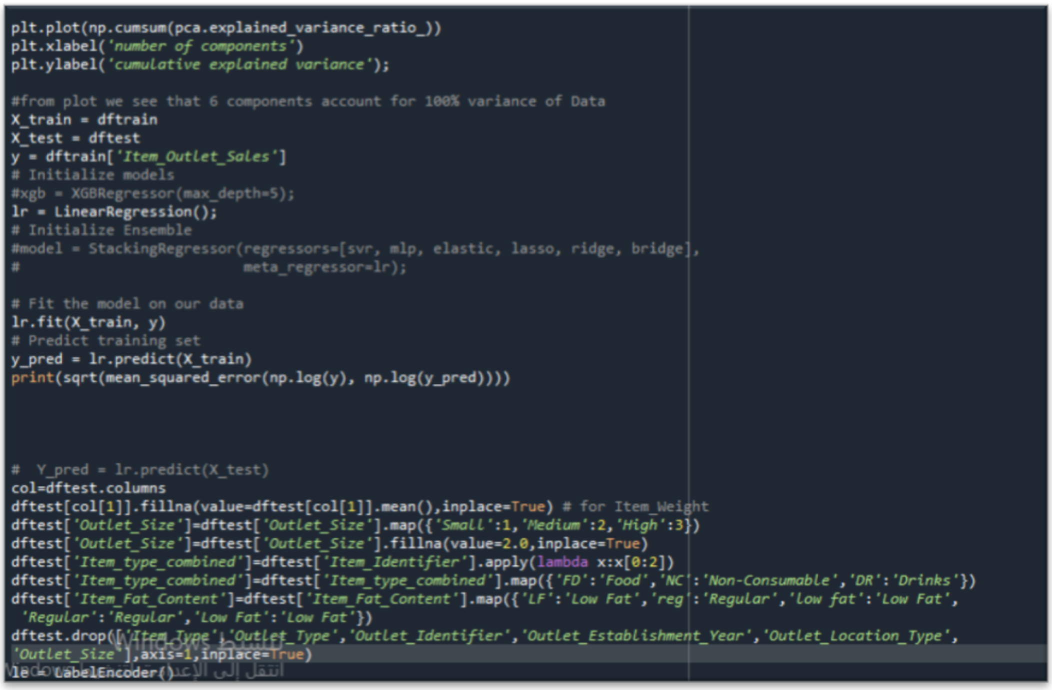
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
# Label encoder
from sklearn.preprocessing import LabelEncoder
# Metrics for root mean squared error
from sklearn.metrics import mean_squared_error
from math import sqrt
from sklearn.linear_model import LinearRegression # Import Linear Regression



dftest=pd.read_csv('C:/Users/My PC/Documents/IOT Video code/Test.csv')
dftrain=pd.read_csv('C:/Users/My PC/Documents/IOT Video code/Train.csv')


data=pd.concat([dftrain,dftest],ignore_index=True)
col=dftrain.columns
dftrain[col[1]].fillna(value=dftrain[col[1]].mean(),inplace=True) # for Item_Weight
#dftrain[col[1]].fillna(value=mean,inplace=True)
dftrain[col[8]].value_counts()
dftrain['New']=dftrain['Outlet_Size'].map({'Small':1,'Medium':2,'High':3})  #mapping for Categ. var. Outlet_Size is col[8]
dftrain.drop('Outlet_Size',axis=1,inplace=True)
dftrain.rename(columns={'New':'Outlet_Size'},inplace=True)
dftrain['Outlet_Size'].fillna(value=dftrain['Outlet_Size'].mean(),inplace=True) #Outlet_Size has 0 Null/NaN values
categorical_columns = [x for x in dftrain.dtypes.index if dftrain.dtypes[x]=='object']
print(len(categorical_columns))




categorical_columns=[x for x in categorical_columns if x not in ['Item_Identifier','Outlet_Identifier','source']]
for x in categorical_columns:
    print("\n frequency of %s"%x)
    print(dftrain[x].value_counts())
```



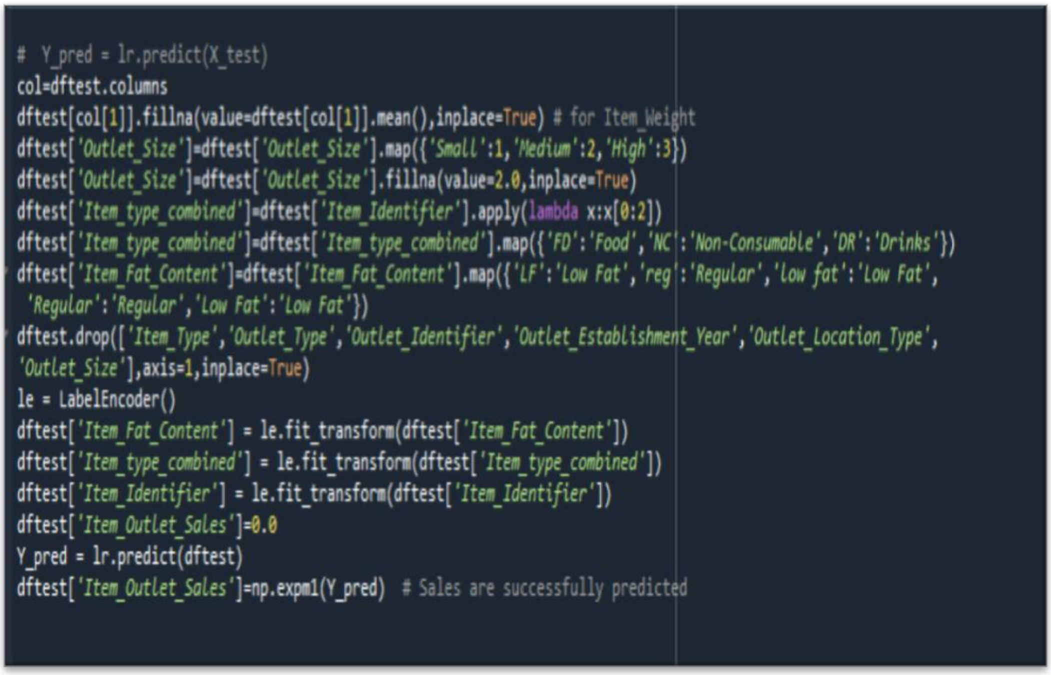
```
categorical_columns=[x for x in categorical_columns if x not in ['Item_Identifier','Outlet_Identifier','source']]
for x in categorical_columns:
    print("\n frequency of %s"%x)
    print(dftrain[x].value_counts())




dftrain['Item_type_combined']=dftrain['Item_Identifier'].apply(lambda x:x[0:2])
dftrain['Item_type_combined']=dftrain['Item_type_combined'].map({'FD':'Food','NC':'Non-Consumable','DR':'Drinks'})
dftrain['Item_Fat_Content']=dftrain['Item_Fat_Content'].map({'LF':'Low Fat','reg':'Regular','low fat':'Low Fat',
'Regular':'Regular','Low Fat':'Low Fat'})
dftrain['Item_Fat_Content'].value_counts().sum()

# drop columns not required
dftrain.drop(['Item_Type','Outlet_Type','Outlet_Identifier','Outlet_Establishment_Year','Outlet_Location_Type',
'Outlet_Size'],axis=1,inplace=True)
le = LabelEncoder()
dftrain['Item_Fat_Content'] = le.fit_transform(dftrain['Item_Fat_Content'])
dftrain['Item_type_combined'] = le.fit_transform(dftrain['Item_type_combined'])
dftrain['Item_Identifier'] = le.fit_transform(dftrain['Item_Identifier'])
corr=dftrain.corr()
#sns.pairplot(dftrain)  #Features  shows up  now try using a model like regression
#let us try PCA

from sklearn.decomposition import PCA

pca = PCA(n_components=7)  # I have selected 7 components to test as main features
pca.fit(dftrain)
print(pca.components_)
print(pca.explained_variance_)
```

- ## Implementation

"Here to out the prediction (using RMSE)"



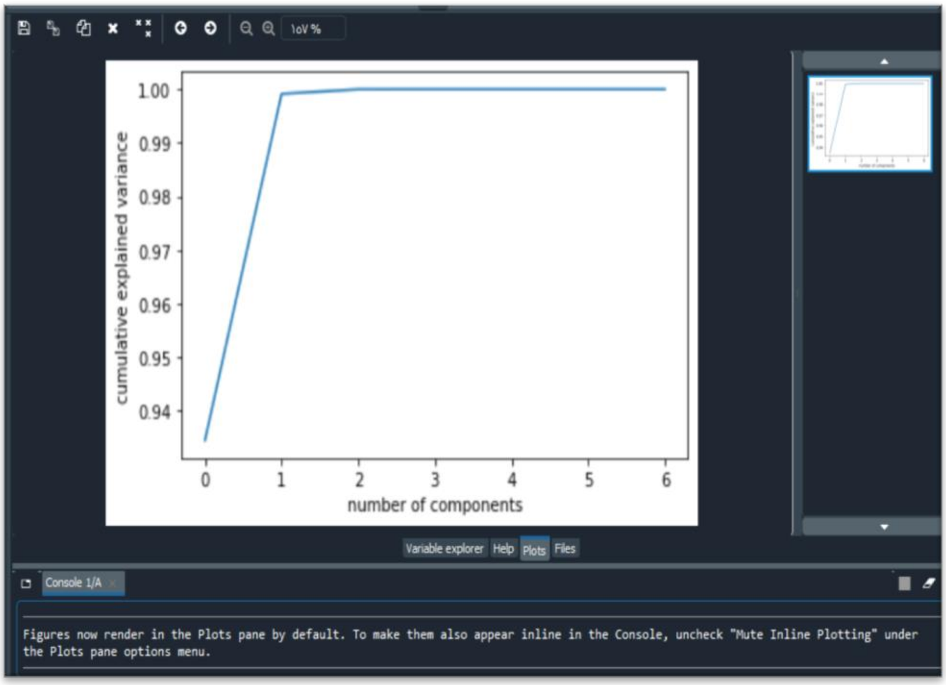
```
plt.plot(np.cumsum(pca.explained_variance_ratio_))
plt.xlabel('number of components')
plt.ylabel('cumulative explained variance');

#from plot we see that 6 components account for 100% variance of Data
X_train = dftrain
X_test = dftest
y = dftrain['Item_Outlet_Sales']
# Initialize models
#xgb = XGBRegressor(max_depth=5);
lr = LinearRegression();
# Initialize Ensemble
#model = StackingRegressor(regressors=[svr, mlp, elastic, lasso, ridge, bridge],
#                           meta_regressor=lr);

# Fit the model on our data
lr.fit(X_train, y)
# Predict training set
y_pred = lr.predict(X_train)
print(sqrt(mean_squared_error(np.log(y), np.log(y_pred))))



#  Y_pred = lr.predict(X_test)
col=dftest.columns
dftest[col[1]].fillna(value=dftest[col[1]].mean(),inplace=True) # for Item_Weight
dftest['Outlet_Size']=dftest['Outlet_Size'].map({'Small':1,'Medium':2,'High':3})
dftest['Outlet_Size']=dftest['Outlet_Size'].fillna(value=2.0,inplace=True)
dftest['Item_type_combined']=dftest['Item_Identifier'].apply(lambda x:x[0:2])
dftest['Item_type_combined']=dftest['Item_type_combined'].map({'FD':'Food','NC':'Non-Consumable','DR':'Drinks'})
dftest['Item_Fat_Content']=dftest['Item_Fat_Content'].map({'LF':'Low Fat','reg':'Regular','low fat':'Low Fat',
 'Regular':'Regular','Low Fat':'Low Fat'})
dftest.drop(['Item_Type','Outlet_Type','Outlet_Identifier','Outlet_Establishment_Year','Outlet_Location_Type',
'Outlet_Size'],axis=1,inplace=True)
le = LabelEncoder()
```



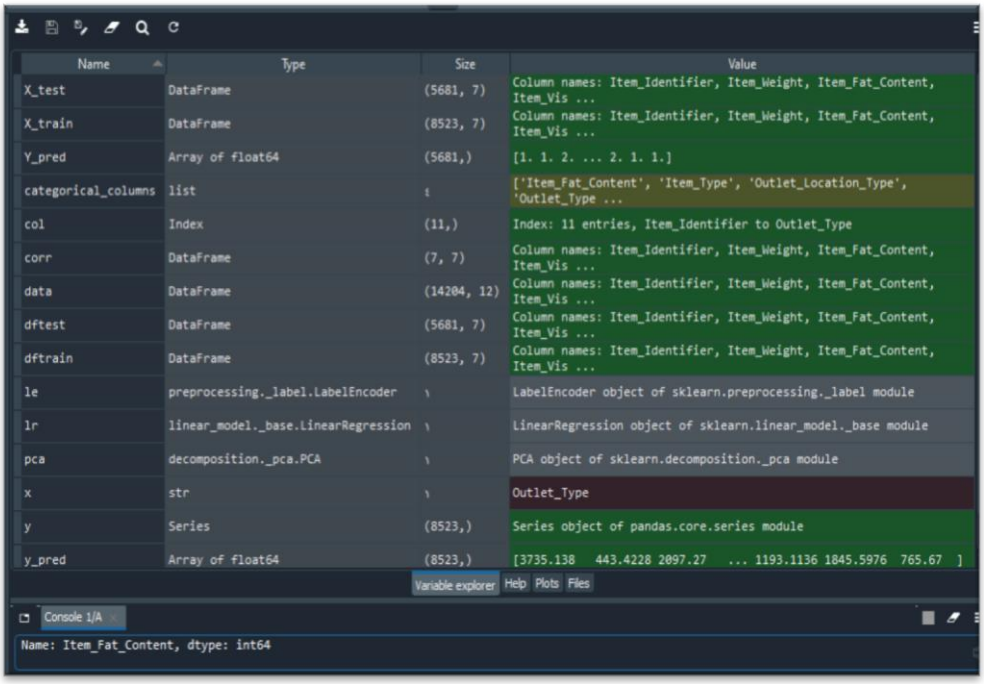
```
#  Y_pred = lr.predict(X_test)
col=dftest.columns
dftest[col[1]].fillna(value=dftest[col[1]].mean(),inplace=True) # for Item_Weight
dftest['Outlet_Size']=dftest['Outlet_Size'].map({'Small':1,'Medium':2,'High':3})
dftest['Outlet_Size']=dftest['Outlet_Size'].fillna(value=2.0,inplace=True)
dftest['Item_type_combined']=dftest['Item_Identifier'].apply(lambda x:x[0:2])
dftest['Item_type_combined']=dftest['Item_type_combined'].map({'FD':'Food','NC':'Non-Consumable','DR':'Drinks'})
dftest['Item_Fat_Content']=dftest['Item_Fat_Content'].map({'LF':'Low Fat','reg':'Regular','low fat':'Low Fat',
 'Regular':'Regular','Low Fat':'Low Fat'})
dftest.drop(['Item_Type','Outlet_Type','Outlet_Identifier','Outlet_Establishment_Year','Outlet_Location_Type',
'Outlet_Size'],axis=1,inplace=True)
le = LabelEncoder()
dftest['Item_Fat_Content'] = le.fit_transform(dftest['Item_Fat_Content'])
dftest['Item_type_combined'] = le.fit_transform(dftest['Item_type_combined'])
dftest['Item_Identifier'] = le.fit_transform(dftest['Item_Identifier'])
dftest['Item_Outlet_Sales']=0.0
Y_pred = lr.predict(dftest)
dftest['Item_Outlet_Sales']=np.expm1(Y_pred)  # Sales are successfully predicted
```

- **Implementation**

OUTPUT

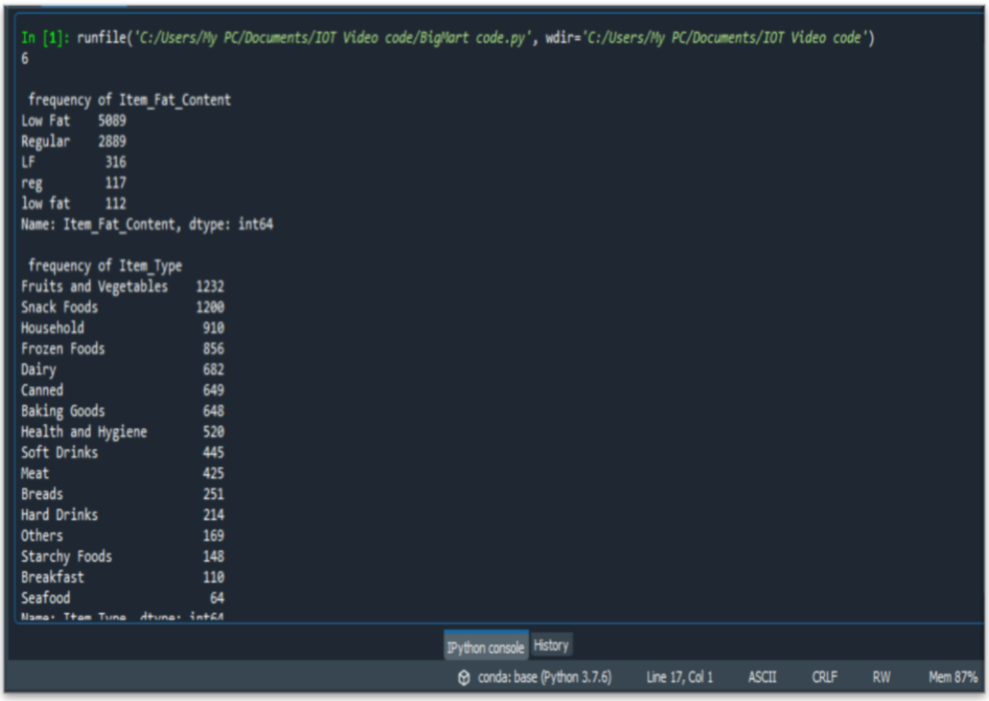


```
Figures now render in the Plots pane by default. To make them also appear inline in the Console, uncheck "Mute Inline Plotting" under
the Plots pane options menu.
```

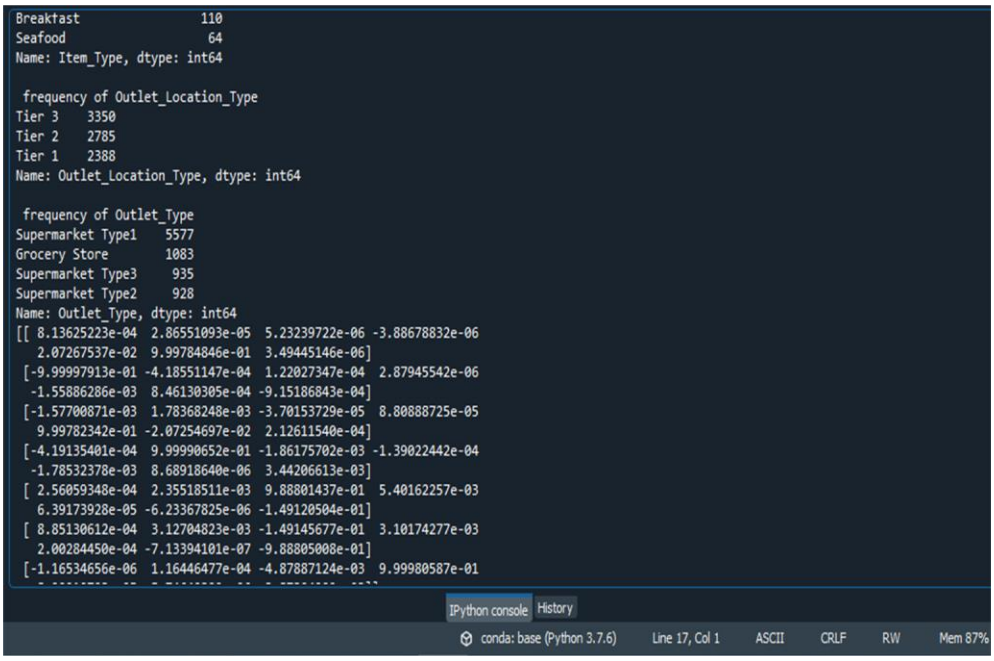


Name	Type	Size	Value
X_test	DataFrame	(5681, 7)	Column names: Item_Identifier, Item_Weight, Item_Fat_Content, Item_Vis ...
X_train	DataFrame	(8523, 7)	Column names: Item_Identifier, Item_Weight, Item_Fat_Content, Item_Vis ...
Y_pred	Array of float64	(5681,)	[1. 1. 2. ... 2. 1. 1.]
categorical_columns	list	٤	['Item_Fat_Content', 'Item_Type', 'Outlet_Location_Type', 'Outlet_Type ...
col	Index	(11,)	Index: 11 entries, Item_Identifier to Outlet_Type
corr	DataFrame	(7, 7)	Column names: Item_Identifier, Item_Weight, Item_Fat_Content, Item_Vis ...
data	DataFrame	(14204, 12)	Column names: Item_Identifier, Item_Weight, Item_Fat_Content, Item_Vis ...
dftest	DataFrame	(5681, 7)	Column names: Item_Identifier, Item_Weight, Item_Fat_Content, Item_Vis ...
dftrain	DataFrame	(8523, 7)	Column names: Item_Identifier, Item_Weight, Item_Fat_Content, Item_Vis ...
le	preprocessing._label.LabelEncoder	١	LabelEncoder object of sklearn.preprocessing._label module
lr	linear_model._base.LinearRegression	١	LinearRegression object of sklearn.linear_model._base module
pca	decomposition._pca.PCA	١	PCA object of sklearn.decomposition._pca module
x	str	١	Outlet_Type
y	Series	(8523,)	Series object of pandas.core.series module
y_pred	Array of float64	(8523,)	[3735.138 443.4228 2097.27 ... 1193.1136 1845.5976 765.67]

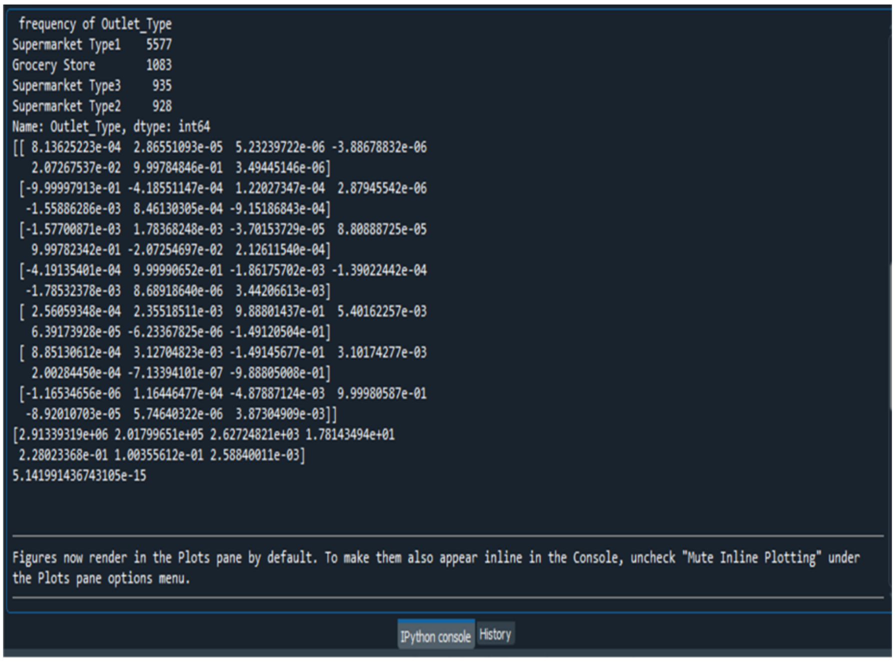
```
Name: Item_Fat_Content, dtype: int64
```



```
In [1]: runfile('C:/Users/My PC/Documents/IOT Video code/BigMart code.py', wdir='C:/Users/My PC/Documents/IOT Video code')
6

 frequency of Item_Fat_Content
Low Fat    5089
Regular    2889
LF          316
reg         117
low fat     112
Name: Item_Fat_Content, dtype: int64

 frequency of Item_Type
Fruits and Vegetables    1232
Snack Foods              1200
Household                 910
Frozen Foods              856
Dairy                     682
Canned                    649
Baking Goods              648
Health and Hygiene        520
Soft Drinks               445
Meat                      425
Breads                    251
Hard Drinks               214
Others                    169
Starchy Foods             148
Breakfast                 110
Seafood                    64
Name: Item_Type, dtype: int64
```



```
Breakfast                 110
Seafood                    64
Name: Item_Type, dtype: int64

 frequency of Outlet_Location_Type
Tier 3    3350
Tier 2    2785
Tier 1    2388
Name: Outlet_Location_Type, dtype: int64

 frequency of Outlet_Type
Supermarket Type1    5577
Grocery Store        1083
Supermarket Type3     935
Supermarket Type2     928
Name: Outlet_Type, dtype: int64
[[ 8.13625223e-04  2.86551093e-05  5.23239722e-06 -3.88678832e-06
   2.07267537e-02  9.99784846e-01  3.49445146e-06]
 [-9.99997913e-01 -4.18551147e-04  1.22027347e-04  2.87945542e-06
  -1.55886286e-03  8.46130305e-04 -9.15186843e-04]
 [-1.57700871e-03  1.78368248e-03 -3.70153729e-05  8.80888725e-05
   9.99782342e-01 -2.07254697e-02  2.12611540e-04]
 [-4.19135401e-04  9.99990652e-01 -1.86175702e-03 -1.39022442e-04
  -1.78532378e-03  8.68918640e-06  3.44206613e-03]
 [ 2.56059348e-04  2.35518511e-03  9.88801437e-01  5.40162257e-03
   6.39173928e-05 -6.23367825e-06 -1.49120504e-01]
 [ 8.85130612e-04  3.12704823e-03 -1.49145677e-01  3.10174277e-03
   2.00284450e-04 -7.13394101e-07 -9.88805008e-01]
 [-1.16534656e-06  1.16446477e-04 -4.87887124e-03  9.99980587e-01
```



```
 frequency of Outlet_Type
Supermarket Type1    5577
Grocery Store        1083
Supermarket Type3     935
Supermarket Type2     928
Name: Outlet_Type, dtype: int64
[[ 8.13625223e-04  2.86551093e-05  5.23239722e-06 -3.88678832e-06
   2.07267537e-02  9.99784846e-01  3.49445146e-06]
 [-9.99997913e-01 -4.18551147e-04  1.22027347e-04  2.87945542e-06
  -1.55886286e-03  8.46130305e-04 -9.15186843e-04]
 [-1.57700871e-03  1.78368248e-03 -3.70153729e-05  8.80888725e-05
   9.99782342e-01 -2.07254697e-02  2.12611540e-04]
 [-4.19135401e-04  9.99990652e-01 -1.86175702e-03 -1.39022442e-04
  -1.78532378e-03  8.68918640e-06  3.44206613e-03]
 [ 2.56059348e-04  2.35518511e-03  9.88801437e-01  5.40162257e-03
   6.39173928e-05 -6.23367825e-06 -1.49120504e-01]
 [ 8.85130612e-04  3.12704823e-03 -1.49145677e-01  3.10174277e-03
   2.00284450e-04 -7.13394101e-07 -9.88805008e-01]
 [-1.16534656e-06  1.16446477e-04 -4.87887124e-03  9.99980587e-01
  -8.92010703e-05  5.74640322e-06  3.87304909e-03]]
[2.91339319e+06 2.01799651e+05 2.62724821e+03 1.78143494e+01
 2.28023368e-01 1.00355612e-01 2.58840011e-03]
5.141991436743105e-15

Figures now render in the Plots pane by default. To make them also appear inline in the Console, uncheck "Mute Inline Plotting" under
the Plots pane options menu.
```

- ## Conclusion

In present era of digitally connected world every shopping mall desires to know the customer demands beforehand to avoid the shortfall of sale items. This helped from corporate profits and higher economic income. Day to day the companies or the malls are predicting more accurately the demand of product sales or user demands. Extensive research in this area at enterprise level is happening for accurate sales prediction. As the profit made by a company is directly proportional to the accurate predictions of sales, the Big marts are desiring more accurate prediction algorithm so that the company will not suffer any losses.

In this paper, we use: linear regressions experimented it on the 2013 Big Mart dataset for predicting sales of the product from a particular outlet.

Finally, experiments that our technique produce more accurate prediction compared to than other available techniques.

• REFERENCES

1. https://www.researchgate.net/publication/330994981_A_STUDY_ON_SELECTION_OF_LOCATION_BY_RETAIL_CHAIN_BIG_MART
2. https://www.ijedr.org/papers/IJEDR1804010.pdf
3. https://www.researchgate.net/publication/336530068_A_Comparative_Study_of_Big_Mart_Sales_Prediction
4. https://www.knowledgehut.com/blog/data-science/linear-regression-for-machine-learning
5. https://libguides.usc.edu/writingguide/literaturereview
6. https://www.statisticshowto.datasciencecentral.com/rmse/
7. https://www.geeksforgeeks.org/ml-linear-regression-vs-logistic-regression/?ref=rp
8. https://www.kaggle.com/brijbhushannanda1979/bigmart-sales-prediction