BIG MART SALES PREDICTION

A PROJECT REPORT

Submitted by

Vamshi Gadde
[Reg No: RA2112704010017]

Under the Guidance of

Dr. A.V. Kalpana

(Assistant Professor, Department of Data Science and Business Systems)

In partial fulfilment of the Requirements for the Degree of

M.TECH (Integrated)
COMPUTER SCIENCE WITH SPECIALIZATION IN DATA SCIENCE

DEPARTMENT OF DATA SCIENCE AND BUSINESS SYSTEMS

FACULTY OF ENGINEERING AND TECHNOLOGY

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

NOVEMBER 2022

SRM INSTITUTE OF SCIENCE AND TECHNOLOGY
KATTANKULATHUR-603203

BONAFIDE CERTIFICATE

Certified that this project report titled "**Big Mart Sales Prediction**" is the Bonafede work of **"Vamshi Gadde [Reg No: RA2112704010017"** who carried out the project work under my supervision. Certified further, that to the best of my knowledge the work reported herein does not form part of any other thesis or dissertation on the basis of which a degree or award was conferred on an earlier occasion for this or any other candidate.

Dr. A.V.Kalpana	Dr. G. Vadivu	Dr. M.Lakshmi
GUIDE	**PROGRAM**	**HEAD OF THE**
Associate Professor	**COORDINATOR**	**DEPARTMENT**
Dept. of DSBS	Dept. of DSBS	Dept. of DSBS

Signature of Internal Examiner

Signature of External Examiner

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.

ACKNOWLEDGEMENTS

We express our humble gratitude to Dr C. Muthamizhchelvan, Vice-Chancellor, SRM Institute of Science and Technology, for the facilities extended for the project work and his continued support.

We extend our sincere thanks to Dean-CET, SRM Institute of Science and Technology, **Dr T.V. Gopal**, for his invaluable support.

We wish to thank **Dr Revathi Venkataraman**, Professor & Chairperson, School of Computing, SRM Institute of Science and Technology, for her support throughout the project work.

We are incredibly grateful to our Head of the Department**, Dr M. Lakshmi** Professor, Department of Data Science and Business Systems, SRM Institute of Science and Technology, for her suggestions and encouragement at all the stages of the project work.

We register our immeasurable thanks to our Faculty Advisor**, Shantha Kumari,** Assistant Professor, Department of Data Science and Business Systems, SRM Institute of Science and Technology, for leading and helping us to complete our course.

Our inexpressible respect and thanks to my guide, **Dr. A.V. Kalpana,** Assistant Professor, Department of Data Science and Business Systems, for providing me with an opportunity to pursue my project under his mentorship. He provided us with the freedom and support to explore the research topics of our interest. His passion for solving problems and making a difference in the world has always been inspiring.

We sincerely thank the Data Science and Business Systems staff and students, SRM Institute of Science and Technology, for their help during our project. Finally, we would like to thank parents, family members, and friends for their unconditional love, constant support, and encouragement.

Vamshi Gadde

TABLE OF CONTENTS

LIST OF SYMBOLS, ABBREVIATION

ARIMA	-	Auto Regressive Integrated Moving Average.
GARCH	-	Gen-realized Auto Regressive Conditionally Heteroskedastic.
SVM	-	Support vector machine.
IOT	-	Internet of Things

CHAPTER 1
INTRODUCTION

1.1 DOMAIN INTRODUCTION

In today's modern world, huge shopping centres 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. Everyday competitiveness between various shopping centers and huge marts is becoming higher intense, violent just because of the quick development of global malls also online shopping. Each market offers personalized and limited-time deals to attract many clients relying on period of time, so that each item's volume of sales may be estimated for the organization's stock control, transportation and logistical services. The current machine learning algorithm is very advanced and provides Various methods for predicting or forecasting sales of any kind of organization, extremely beneficial to overcome low priced used for prediction. Presentmachine learning algorithm are very sophisticated and provide techniques topredict or forecast the future demand of sales for an organization, which alsohelps in overcoming the cheap availability of computing and storage systems.In this paper, we are addressing the problem of big mart sales prediction orforecasting of an item on customer's future demand in different big mart storesacross various locations and products based on the previous record. Differentmachine learning algorithms like linear regression analysis, random forest, etcare used for prediction or forecasting of sales volume. As good sales are the lifeof every organization so the forecasting of sales plays an important role in anyshopping complex. Always a better prediction is helpful, to develop as well as toenhance the strategies of business about the marketplace which is also helpful to improve the knowledge of marketplace. A standard sales prediction study canhelp in deeply analyzing the situations or the conditions previously occurred andthen, the inference can be applied about customer acquisition, funds

inadequacyand strengths before setting a budget and marketing plans for the upcoming year. In other words, sales prediction is based on the available resources from the past. In depth knowledge of past is required for enhancing and improving the likelihood of marketplace irrespective of any circumstances especially the external circumstance, which allows to prepare the upcoming needs for the busy-ness. Extensive research is going on in retailers domain for forecasting the future sales demand. The basic and foremost technique used in predicting sale is the statistical methods, which is also known as the traditional method, but these methods take much more time for predicting a sales also these methods could not handle non linear data so to over these problems in traditional methods machine learning techniques are deployed. Machine learning techniques can not only handle non-linear data but also huge data-set efficiently. To measure the performance of the models, Root Mean Square Error (RMSE) [15] and Mean Absolute Error (MAE) [4] are used as an evaluation metric as mentioned in the Equation 1 and 2 respectively. Here Both metrics are used as the parameter foraccuracy measure of a continuous variable.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|x_{\text{predict}} - x_{\text{actual}}\right| \quad (1)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\left|x_{\text{predict}} - x_{\text{actual}}\right|\right)^2} \quad (2)$$

Theremaining part of this article is arranged as following: Section 1 briefly describesintroduction of sales prediction of Big Mart and also elaborate about the evalua-tion metric used in the model. Previous related work has been pointed in Section2. The detailed description and analysis of proposed model is given in Section3. Where as implementations and results are demonstrated in Section 4 and thepaper concludes with a conclusion in the last section.

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 with respect to the tasks data.

This can then further be used for forecasting future sales by machine learning algorithms such as the random forests and simple or multiple linear regression model.

CHAPTER 2
RELATED WORKS

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 [2] also this paper elaborates the automated process of knowledge acquisition. Machine learning [6] 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 [2]. Pat Lang-ley and Herbert A [7] pointed out most widely used data mining technique in A Comparative Study of Big Mart Sales Prediction.

the field of business is the Rule Induction (RI) technique as compared to other data mining techniques. Whereas sale prediction of a pharmaceutical distribution company has been described in [12,10]. Also, this paper focuses on two issues: (i) stock state should not undergo out of stock, and (ii) it avoids the customer dissatisfaction by predicting the sales that manages the stock level of medicines. Handling of footwear sale fluctuation in a period of time has been addressed in [5]. Also, this paper focuses on using neural network for predicting of weekly retail sales, which decrease the uncertainty present in the short-term planning of sales. Linear and non-linear [3] a comparative analysis model for sales forecasting is proposed for the retailing sector. Beheshti-Kashi and Samaneh [1] performed sales prediction in fashion market. A two-level statistical method [11]is elaborated for forecasting the big mart sales prediction. Xia and Wong [17] proposed the differences between classical methods (based on mathematical and statistical models) and modern heuristic methods and also named exponential smoothing, regression, auto regressive integrated moving average (ARIMA), gen-realized auto regressive conditionally heteroskedastic (GARCH) methods. Most of these models are linear and are not able to deal with the asymmetric behavior in most real-world sales data [9]. Some of the challenging factors like lack of historical data, consumer-oriented markets face uncertain demands, and short life cycles of prediction methods results in inaccurate forecast.

CHAPTER 3

METHODOLOGIES

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

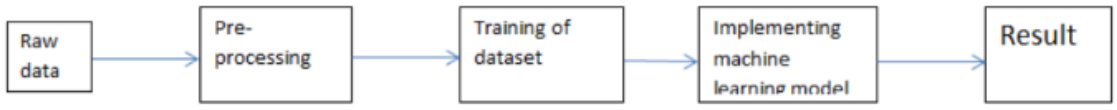


Fig1: Steps followed for obtaining results

3.1 Dataset and its Pre-processing

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.

```
In [4]: data.head()
```

Out[4]:

	Item_Fat_Content	Item_Identifier	Item_MRP	Item_Outlet_Sales	Item_Type	Item_Visibility	Item_Weight	Outlet_Establishment_Year	Outlet_Identifier	Outlet_Lc
0	Low Fat	FDA15	249.8092	3735.1380	Dairy	0.016047	9.30	1999	OUT049	
1	Regular	DRC01	48.2692	443.4228	Soft Drinks	0.019278	5.92	2009	OUT018	
2	Low Fat	FDN15	141.6180	2097.2700	Meat	0.016760	17.50	1999	OUT049	
3	Regular	FDX07	182.0950	732.3800	Fruits and Vegetables	0.000000	19.20	1998	OUT010	
4	Low Fat	NCD19	53.8614	994.7052	Household	0.000000	8.93	1987	OUT013	

```
In [16]: data.head()
```

Out[16]:

	Item_Identifier	Item_MRP	Item_Outlet_Sales	Item_Type	Item_Visibility	Item_Weight	Outlet_Establishment_Year	Outlet_Identifier	source	Item_Fat_Content_0
0	FDA15	249.8092	3735.1380	Dairy	0.016047	9.30	1999	OUT049	train	0
1	DRC01	48.2692	443.4228	Soft Drinks	0.019278	5.92	2009	OUT018	train	0
2	FDN15	141.6180	2097.2700	Meat	0.016760	17.50	1999	OUT049	train	0
3	FDX07	182.0950	732.3800	Fruits and Vegetables	0.000000	19.20	1998	OUT010	train	0
4	NCD19	53.8614	994.7052	Household	0.000000	8.93	1987	OUT013	train	0

5 rows × 37 columns

Fig2: Screenshot of Dataset

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

```
In [17]: data.dtypes
         #tells datatype of column convert data type

Out[17]: Item_Identifier              object
         Item_MRP                    float64
         Item_Outlet_Sales           float64
         Item_Type                    object
         Item_Visibility             float64
         Item_Weight                 float64
         Outlet_Establishment_Year     int64
         Outlet_Identifier            object
         source                       object
         Item_Fat_Content_0            uint8
         Item_Fat_Content_1            uint8
         Item_Fat_Content_2            uint8
```

Fig3: Various datatypes used in the Dataset

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

```
In [5]: #Numerical data summary:
        data.describe()
```

Out[5]:

	Item_MRP	Item_Outlet_Sales	Item_Visibility	Item_Weight	Outlet_Establishment_Year
count	14204.000000	8523.000000	14204.000000	11765.000000	14204.000000
mean	141.004977	2181.288914	0.065953	12.792854	1997.830681
std	62.086938	1706.499616	0.051459	4.652502	8.371664
min	31.290000	33.290000	0.000000	4.555000	1985.000000
25%	94.012000	834.247400	0.027036	8.710000	1987.000000
50%	142.247000	1794.331000	0.054021	12.600000	1999.000000
75%	185.855600	3101.296400	0.094037	16.750000	2004.000000
max	266.888400	13086.964800	0.328391	21.350000	2009.000000

Fig4: Numerical variables of the Dataset

Pre-processing 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 moral 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.

3.2 Algorithms 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. The tree 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 sclera. ensemble library's random forest regressor class is used. The key role is played by the parameter termed as 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 quality is 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.

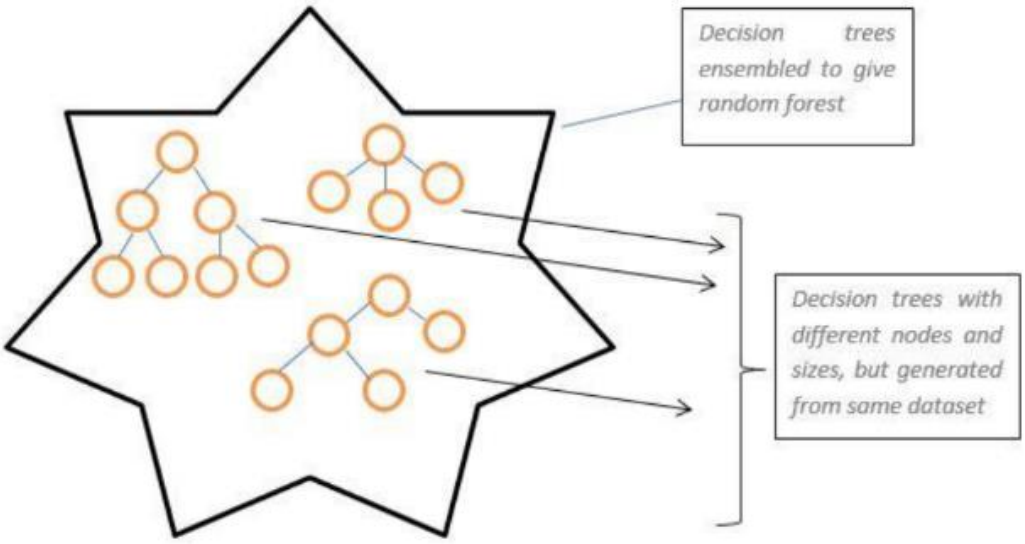


Fig 5: Relation between Decision Trees and Random Forest

- **Linear Regression Algorithm**

Regression 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 basis of data set.

$$Y = \beta o + \beta 1X + \qquad (1)$$

Equation shown in eq.1 is used for simple linear regression. These parameters can be said as:

Y - Variable to be predicted

X - Variable(s) used for making a prediction

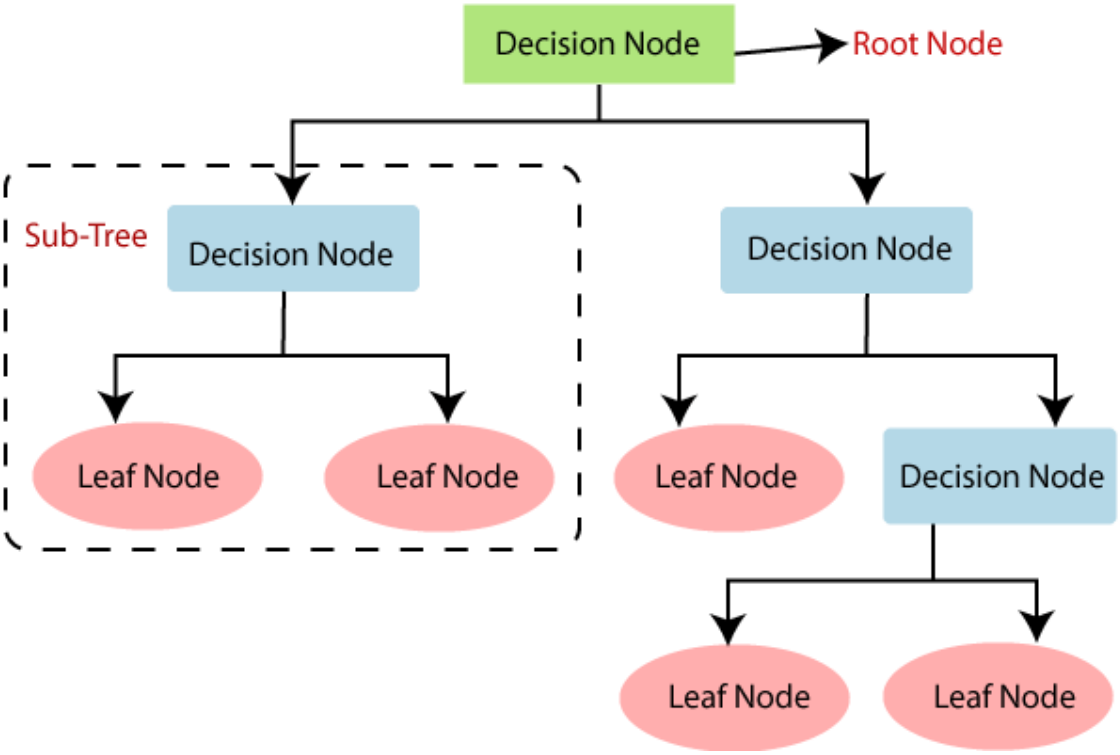
βo - When X=0, it is termed as prediction value or can be referred to as intercept term.

β1 - when there is a change in X by 1 unit it denotes change in Y. It can also be said as slope term.

∈ -The difference between the predicted and actual values is represented by this parameter and also represents the residual value. However efficiently the model is trained, tested and validated, 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. Alternative methods given by Dieterich can be used for comparing learning algorithms [10].

- **Decision Tree Algorithm**

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. The decisions or the test are performed on the basis of features of the given dataset.



It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions. It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure. A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees.

Decision tree inducers are algorithms that automatically construct a decision tree from a given dataset. Typically, the goal is to find the optimal decision tree by minimizing the generalization error. However, other target functions can be also defined, for instance, minimizing the number of nodes or minimizing the average depth.

3.3 Metrics for Data Modelling

- The coefficient of determination R2 (R-squared) is a statistic that measures the goodness of a model's fit i.e. how well the real data points are approximated by the predictions of regression. Higher values of R2 suggest higher model accomplishments in terms of prediction along with accuracy, and the value 1 of R2 is indicative of regression predictions perfectly fitting the real data points. For further better results, the use of adjusted R2 measures works wonders. Taking logarithmic values of the target column in the dataset proves to be significant in the prediction process. So, it can be said that on taking adjustments of columns used in prediction, better results can be deduced. One way of incorporating adjustment could also have included taking square root of the column. It also provides better visualization of the dataset and target variable as the square root of target variable is inclined to be a normal distribution.
- The error measurement is an important metric in the estimation period. Root mean squared error (RMSE) and Mean Absolute Error (MAE) are generally used for continuous variable's accuracy measurement. It can be said that the average model prediction error can be expressed in units of the variable of interest by using both MAE and RMSE. MAE is the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight. The square root of the average of squared differences between prediction and actual observation can be termed as RMSE. RMSE is an absolute measure of fit, whereas R2 is a relative measure of fit. RMSE helps in measuring the variable's average error and it is also a quadratic scoring rule. Low RMSE values obtained for linear or multiple regression corresponds to better model fitting.

CHAPTER 4
MACHINE LEARNING

The data available is increasing day by day and such a huge amount of unprocessed data is needed to be analyzed 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 supervise 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 basis of data set
- K-Nearest Neighbors (KNN): It is a learning algorithm which is based on instances and knowledge gained through them [4]. Unlike mining in data stream scenarios, cases where every sample can simultaneously belong to multiple classes in hierarchical multi-label classification problems, k-NN is being proposed to be applied to predict outputs in structured form [5].

- Decision tree: It is an intuitive model having low bias and it can be adopted to build a classification tree with root node being the first to be taken into account in a top-down manner. It is a classic model for machine learning [6].
- Naïve Bayes classifiers: These are based on Bayes theorem and a collection of classification algorithms where classification of every pair is independent of each other. Bayesian learning can provide predictions with readable reasons by generating an if-then form of list of rules [8].
- Random Tree: It is an efficient algorithm for achieving scalability and is used in identification problems for building approximate system. The decisions are taken considering the choices made on basis of possible consequences, the variables which are included, input factor. Other algorithms can include SVM, boost, logistic regression and so on [7].
- K-means clustering: This algorithm is used in unsupervised learning for creating clusters of related data based on their closeness to the centroid value [9].

A predictive model was developed using Xgboost, Linear regression, Polynomial regression, and Ridge regression techniques for forecasting the sales of a business such as Big -Mart, and it was discovered that the model outperforms existing models

- Linear Regression

  Build a fragmented plot.1) a linear or non-linear pattern of data and 2) a variance (outliers). Consider a transformation if the marking isn't linear. If this is the case, outsiders, it can suggest only eliminating them if there is a non-statistical justification.
- Polynomial Regression Algorithm

  Polynomial Regression is a relapse calculation that modules the relationship here among dependent(y) and the autonomous variable(x) in light of the fact that as most extreme limit polynomial. The condition for polynomial relapse is given beneath: y=b0+b1x1+b2x12+b2x13+……bnx1n

  It is regularly alluded to as the exceptional instance of various straight relapse in ML. Since we apply some polynomial terms to the numerous straight relapse condition to change it to polynomial relapse adjustment to improve accuracy.
- Ridge Regression

  Ridge regression is a model tuning tool used to evaluate any data that suffers from multicollinearity. This method performs the L2 regularization procedure. When

multicollinearity issues arise, the least squares are unbiased and the variances are high, resulting in the expected values being far removed from the actual values.

The cost function for ridge regression:

Min(‖Y−X(theta)‖^2+λ‖theta‖^2)

- Xgboost Regression

  “Extreme Gradient Boosting” is same but much more effective to the gradient boosting system. It has both a linear model solver and a tree algorithm.

  Which permits “xgboost” in any event multiple times quicker than current slope boosting executions. It underpins various target capacities, including relapse, order and rating. As “xgboost” is extremely high in prescient force however generally delayed with organization, it is appropriate for some rivalries. It likewise has extra usefulness for cross-approval and finding significant factors.

	Model	R Square
0	linear Regression	0.510595
1	KNeighbors	0.561299
2	Decision Tree	0.537335
3	RandomForest	0.548853
4	AdaBoost	0.530487
5	GradientBoosting	0.591471

Comparison of MAE, MSE, RMSE with the model

CHAPTER 5

PROJECT CODE

5.1 Algorithm

- Step 1: Importing Libraries such as NumPy, pandas, matplotlib, labelEncoder, OneHotEncoder.
- Step 2: Importing Dataset and combining into one file.
- Step 3: Analyzing the Data
- Step 4: Preprocessing the Data using Stemming, Lemmatization and removing Stop words
- Step 5: Splitting the data into training and test dataset.
- Step 6: TF-IDF Vectorizing
- Step 7: Creating Models for the evaluation of Machine Learning algorithms
- Step 8: Testing the Models

5.2Importing Libraries

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.preprocessing import LabelEncoder, OneHotEncoder
import warnings
from sklearn.linear_model import  LinearRegression
from sklearn.metrics import accuracy_score, r2_score, mean_squared_error
from sklearn.model_selection import cross_val_score
from sklearn import cross_validation, metrics
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor
```

5.3 Importing Dataset

```
#Combine test and train into one file
train['source']='train'
test['source']='test'
data = pd.concat([train, test],ignore_index=True)
print(train.shape, test.shape, data.shape)
```

```
(8523, 13) (5681, 12) (14204, 13)
```

data.head()

	Item_Fat_Content	Item Identifier	Item_MRP	Item_Outlet_Sales	Item Type	Item Visibility	Item Weight
0	Low Fat	FDA15	249.8092	3735.1380	Dairy	0.016047	9.30
1	Regular	DRC01	48.2692	443.4228	Soft Drinks	0.019278	5.92
2	Low Fat	FDN15	141.6180	2097.2700	Meat	0.016760	17.50
3	Regular	FDX07	182.0950	732.3800	Fruits and Vegetables	0.000000	19.20
4	Low Fat	NCD19	53.8614	994.7052	Household	0.000000	8.93

5.4 Analyzing the Data

In [5]:
```
#Numerical data summary:
data.describe()
```

Out[5]:

	Item_MRP	Item_Outlet_Sales	Item_Visibility	Item_Weight	Outlet_Establishment_Year
count	14204.000000	8523.000000	14204.000000	11765.000000	14204.000000
mean	141.004977	2181.288914	0.065953	12.792854	1997.830681
std	62.086938	1706.499616	0.051459	4.652502	8.371664
min	31.290000	33.290000	0.000000	4.555000	1985.000000
25%	94.012000	834.247400	0.027036	8.710000	1987.000000
50%	142.247000	1794.331000	0.054021	12.600000	1999.000000
75%	185.855600	3101.296400	0.094037	16.750000	2004.000000
max	266.888400	13086.964800	0.328391	21.350000	2009.000000

In [12]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14204 entries, 0 to 14203
Data columns (total 13 columns):
Item_Fat_Content            14204 non-null object
Item_Identifier             14204 non-null object
Item_MRP                    14204 non-null float64
Item_Outlet_Sales           14204 non-null float64
Item_Type                   14204 non-null object
Item_Visibility             14204 non-null float64
Item_Weight                 14204 non-null float64
Outlet_Establishment_Year   14204 non-null int64
Outlet_Identifier           14204 non-null object
Outlet_Location_Type        14204 non-null object
Outlet_Size                 14204 non-null object
Outlet_Type                 14204 non-null object
source                      14204 non-null object
dtypes: float64(4), int64(1), object(8)
memory usage: 1.4+ MB
```

```
data.dtypes
#tells datatype of column convert data type
```

```
Item_Identifier               object
Item_MRP                     float64
Item_Outlet_Sales            float64
Item_Type                     object
Item_Visibility              float64
Item_Weight                  float64
Outlet_Establishment_Year      int64
Outlet_Identifier             object
source                        object
Item_Fat_Content_0             uint8
Item_Fat_Content_1             uint8
Item_Fat_Content_2             uint8
Item_Fat_Content_3             uint8
Item_Fat_Content_4             uint8
Outlet_Location_Type_0         uint8
Outlet_Location_Type_1         uint8
Outlet_Location_Type_2         uint8
Outlet_Size_0                  uint8
Outlet_Size_1                  uint8
Outlet_Size_2                  uint8
Outlet_Type_0                  uint8
Outlet_Type_1                  uint8
Outlet_Type_2                  uint8
Outlet_Type_3                  uint8
Item_Type_Combined_0           uint8
Item_Type_Combined_1           uint8
Item_Type_Combined_2           uint8
Outlet_0                       uint8
Outlet_1                       uint8
Outlet_2                       uint8
Outlet_3                       uint8
Outlet_4                       uint8
Outlet_5                       uint8
Outlet_6                       uint8
Outlet_7                       uint8
Outlet_8                       uint8
Outlet_9                       uint8
dtype: object
```

CHAPTER 6

PROJECT MODELS

Linear Regression Model:

Multiple Linear Regression is an extension of Simple Linear regression as it takes more than one predictor variable to predict the response variable.

```
In [26]: # Fitting Multiple Linear Regression to the training set
         from sklearn.linear_model import  LinearRegression
         regressor = LinearRegression()
         regressor.fit(X_train, y_train)

Out[26]: LinearRegression(copy_X=True, fit_intercept=True, n_jobs=1, normalize=False)
```

Predicting the test set results

```
In [27]: # Predicting the test set results
         y_pred = regressor.predict(X_test)

In [28]: y_pred

Out[28]: array([1848.53604783, 1472.81670435, 1875.65285894, ..., 1809.18796433,
                3565.6645235 , 1267.46171871])
```

Linear Regression Algorithm model accuracy and score of regression model can reach **nearly 58%** if built with more hypothesis consideration and analysis, as shown by code snippet in the below fig.

```
In [29]: import warnings
         warnings.filterwarnings('ignore')
         # Measuring Accuracy
         from sklearn.metrics import accuracy_score, r2_score, mean_squared_error
         from sklearn.model_selection import cross_val_score
         from sklearn import cross_validation, metrics

In [30]: lr_accuracy = round(regressor.score(X_train,y_train) * 100,2)
         lr_accuracy

Out[30]: 56.36
```

Decision Tree Model:

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

```
In [37]: # Fitting Decision Tree Regression to the dataset
         from sklearn.tree import DecisionTreeRegressor
         regressor = DecisionTreeRegressor(max_depth=15,min_samples_leaf=300)
         regressor.fit(X_train, y_train)

Out[37]: DecisionTreeRegressor(criterion='mse', max_depth=15, max_features=None,
                    max_leaf_nodes=None, min_impurity_decrease=0.0,
                    min_impurity_split=None, min_samples_leaf=300,
                    min_samples_split=2, min_weight_fraction_leaf=0.0,
                    presort=False, random_state=None, splitter='best')
```

Predicting the test set results

```
In [38]: # Predicting the test set results
         y_pred = regressor.predict(X_test)
         y_pred

Out[38]: array([1673.98398729, 1349.51290433,  471.30684669, ..., 1892.06614452,
                3805.94860417, 1349.51290433])
```

Decision Tree Algorithm model accuracy and score of regression model can reach **nearly 1%** if built with more hypothesis consideration and analysis, as shown by code snippet in the below fig.

```
In [39]: tree_accuracy = round(regressor.score(X_train,y_train),2)
         tree_accuracy

Out[39]: 0.59
```

Random Forest Model:

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

```
In [45]: # Fitting Random Forest Regression to the dataset
         from sklearn.ensemble import RandomForestRegressor
         regressor = RandomForestRegressor(n_estimators=100,max_depth=6, min_samples_leaf
         regressor.fit(X_train, y_train)

Out[45]: RandomForestRegressor(bootstrap=True, criterion='mse', max_depth=6,
                    max_features='auto', max_leaf_nodes=None,
                    min_impurity_decrease=0.0, min_impurity_split=None,
                    min_samples_leaf=50, min_samples_split=2,
                    min_weight_fraction_leaf=0.0, n_estimators=100, n_jobs=4,
                    oob_score=False, random_state=None, verbose=0, warm_start=False)
```

Predicting the test set results

```
In [46]: # Predicting the test set results
         y_pred = regressor.predict(X_test)
         y_pred

Out[46]: array([1643.87106725, 1364.24193091,  603.09113992, ..., 1957.62183676,
                3698.60040819, 1290.25320329])
```

Random Forest Algorithm model accuracy and score of regression model can reach **nearly 0.65%** if built with more hypothesis consideration and analysis, as shown by code snippet in the below fig

```
In [47]: rf_accuracy = round(regressor.score(X_train,y_train),2)
         rf_accuracy

Out[47]: 0.61
```

CHAPTER 7
RESULTS

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

7.1 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 as the ‘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. As a development platform, Jupyter Notebook, which proves to work great due to its excellence in ‘literate programming’, where human friendly code is punctuated within code blocks, has been used.

7.2 Data Modelling and Observations

Correlation is used to understand the relation between a target variable and predictors. In this work, Item-Sales is the target variable and its correlation with other variables is observed.

Considering the case of Item-Weight, the feature item weight is shown to have a low correlation with the target variable Item-Outlet-Sales below.

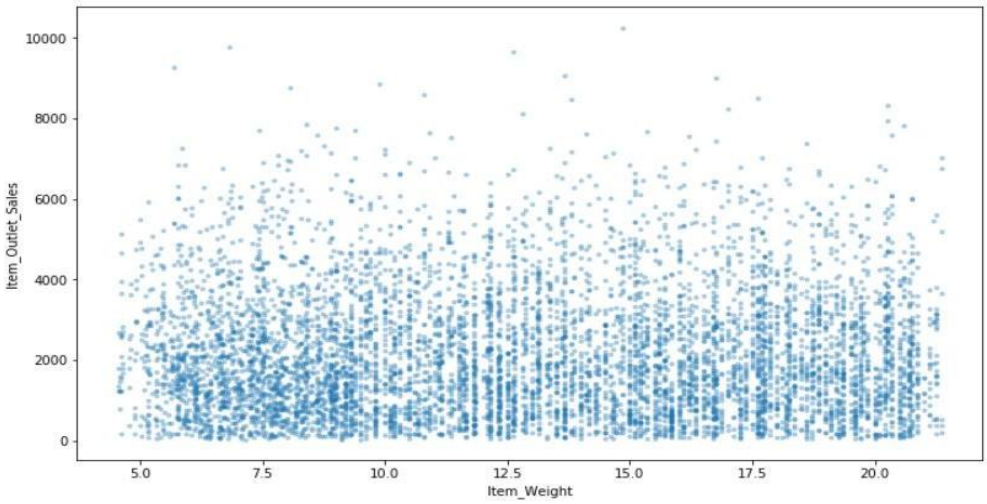


Fig6: Correlation between target variable and Item weight variable

As can be seen from Fig.7, there is no significant relation found between the year of store establishment and the sales for the items. Values can also be combined into variables that classify them into periods and give meaningful results.

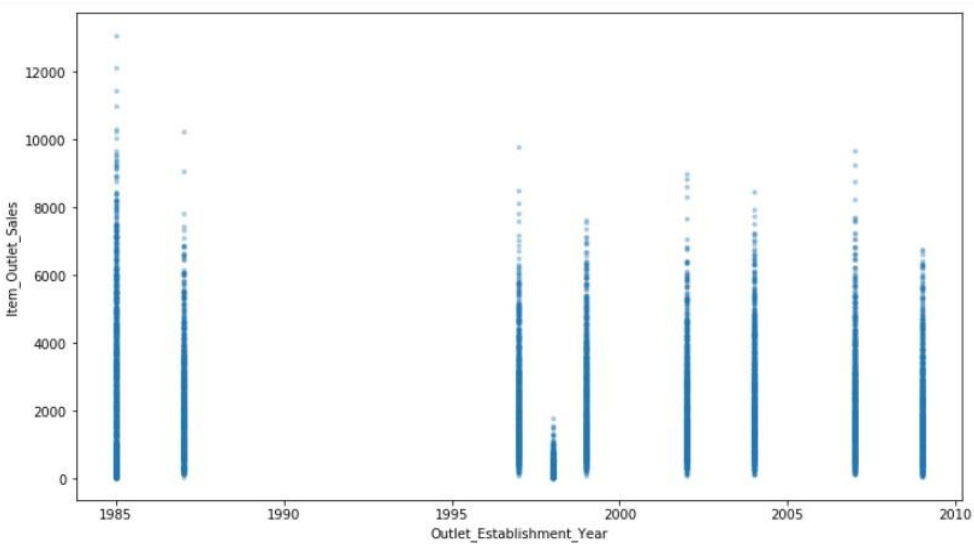


Fig7: Correlation between target variable and Outlet-establishment-year variable

The place where an item is placed in a store, referred to as Item_visibility, definitely affects the sales. However, the plot chart and correlation table generated previously show that the flow is in opposite side. One of the reasons might be that daily used products don't need high visibility. However, there is an issue that some products have zero visibility, which is quite impossible. Fig.8 shows the correlation between item visibility variable and the target variable.

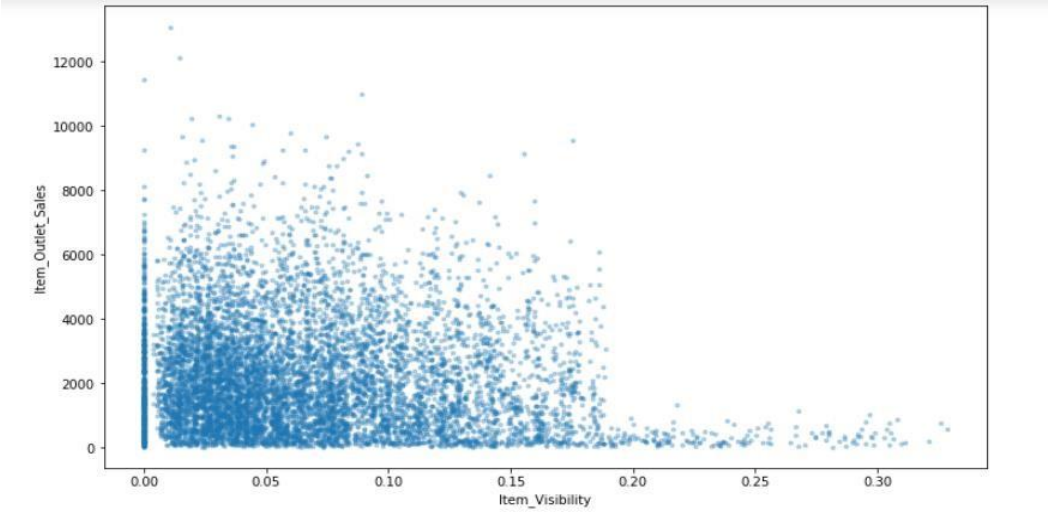


Fig8 Correlation between target variable and Item-visibility variable

Frequency for each categorical or nominal variable plays a significant role in further analysis of the dataset, thus supporting and collaborating in data exploration to be performed. As shown in Fig.9, various variables in our dataset, with their data type and categories are shown. Here, the ID column and the source column, denoting from where the test or train sample data belongs to, are excluded and not used.

```
In [18]: for col in ct:
             print(data[col].value_counts())

Low Fat    8485
Regular    4824
LF          522
reg         195
low fat     178
Name: Item_Fat_Content, dtype: int64
Fruits and Vegetables    2013
Snack Foods              1989
Household                1548
Frozen Foods             1426
Dairy                    1136
Baking Goods             1086
Canned                   1084
Health and Hygiene        858
Meat                      736
Soft Drinks               726
Breads                    416
Hard Drinks               362
Others                    280
Starchy Foods             269
Breakfast                 186
Seafood                    89
Name: Item_Type, dtype: int64
Tier 3    5583
Tier 2    4641
Tier 1    3980
Name: Outlet_Location_Type, dtype: int64
Medium    4655
Small     3980
High      1553
Name: Outlet_Size, dtype: int64
Supermarket Type1    9294
Grocery Store        1805
Supermarket Type3    1559
Supermarket Type2    1546
Name: Outlet_Type, dtype: int64
```

Fig9: Different item categories in the dataset

When a predictive model generated from any supervised learning regression method is applied to the dataset, the process is said to be data scoring. The above model score clearly infers about Data Scoring. The probability of a product's sales to rise and sink can be discussed and understood on the basis of certain parameters. The vulnerabilities associated with a product or item and further its sales are also necessary and play a very important role in our problem-solving task. Further, a user authentication mechanism should be employed to avoid access from any unauthorized users and thus ensuring all results are protected and secured.

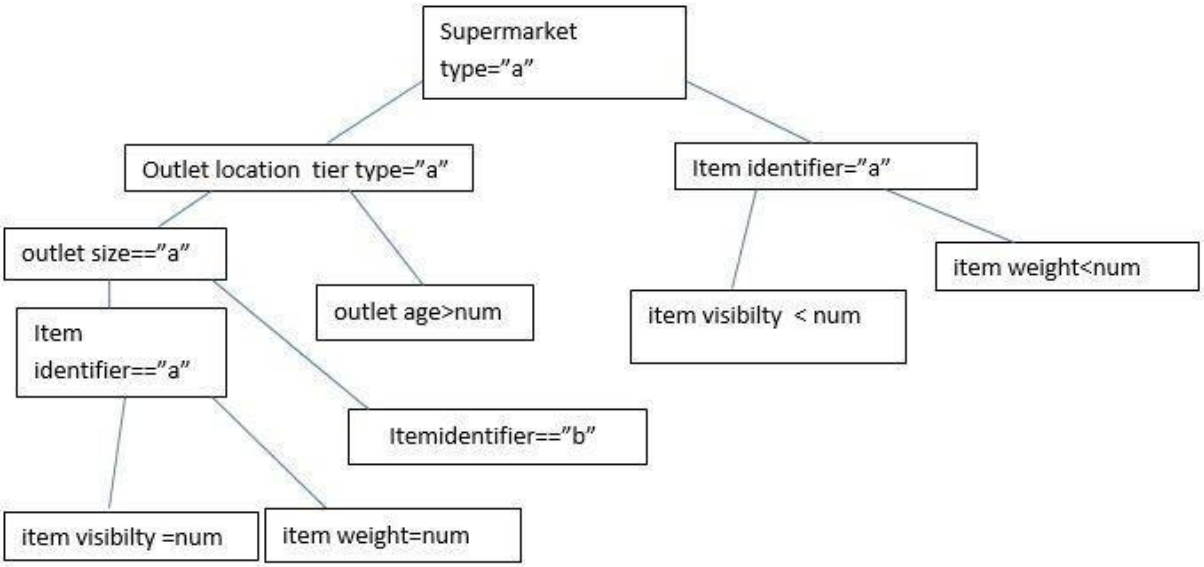


Fig 10: Flowchart for division of dataset on various factors (having proper leaves after pruning)

In Fig.10, a flowchart is represented in which the dataset has been divided on the basis of various factors. In the last stage of the flowchart, the nodes with numbers 'a' and 'b' represent some string values for distinguishing the dataset items and 'num' can be any arbitrary number. The dataset has been divided and pruning has been performed on the basis of different factors. Ensembling many such decision trees will generate a random forest model.

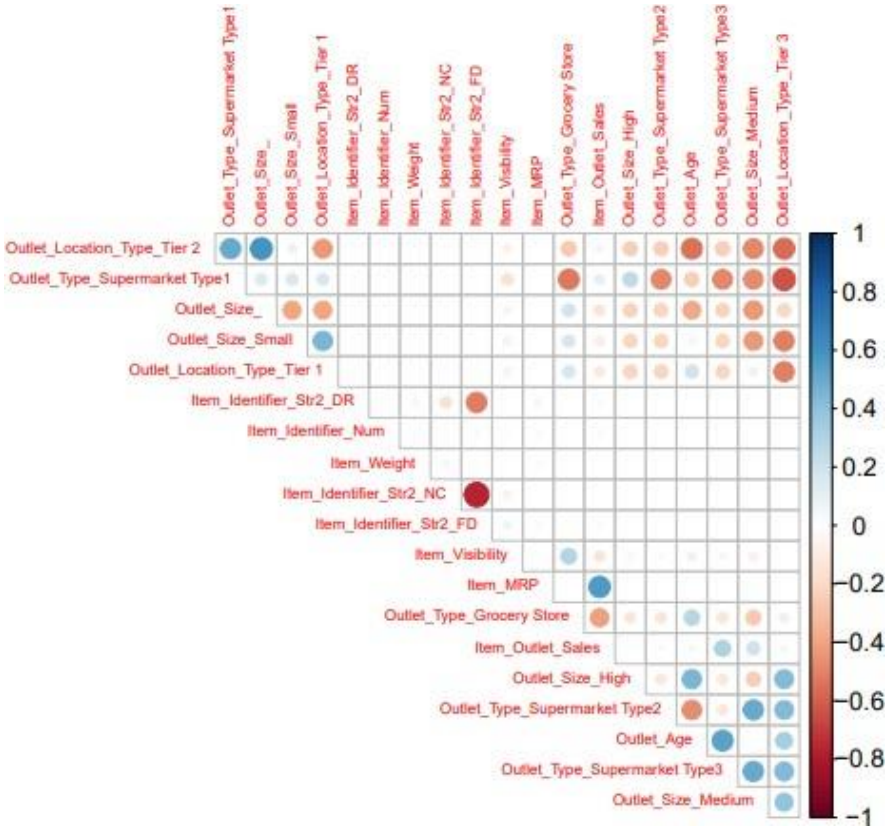


Fig11: Diagram showing correlation among different factors

From Fig.11, the correlation among various dependent and independent variables is explored to be able to decide on the further steps that are to be taken. Variables used are obtained after data pre-processing, and following are some of the important observations about some of the used variables:

•Item_visibility is having nearly zero correlation with our dependent variable item_outlet_sales and grocery store outlet_type. This means that the sales are not affected by visibility of item which is a contradiction to the general assumption of "more visibility thus, more sales".

•Item_MRP (maximum retail price) is positively correlated with sales at an outlet, which indicates that the price quoted by an outlet plays an important factor in sales.

•Outlet situated in location with type tier 2 and size medium are also having high sales, which means that a one-stop-shopping-center situated in a town or city with populated area can have high sales.

•Variation in MRP quoted by various outlets depends on their individual sales.

Fig.12 summarizes the various observations obtained from the developed linear regression model. The method used is least square method and model used is ordinary least square method (OLS).

	coef	std err	t	P>|t|	[0.025	0.975]
const	-105.2014	14.368	-7.322	0.000	-133.366	-77.037
Item_MRP	15.5564	0.197	79.109	0.000	15.171	15.942
Item_Visibility	-215.0161	257.172	-0.836	0.403	-719.136	289.103
Item_Weight	-0.5898	2.901	-0.203	0.839	-6.276	5.096
Outlet_Years	9.7876	1.569	6.237	0.000	6.712	12.864
Item_Fat_Content_0	-73.3896	14.931	-4.915	0.000	-102.658	-44.121
Item_Fat_Content_1	-31.8117	16.743	-1.900	0.057	-64.632	1.008
Outlet_Location_Type_0	-227.2672	13.187	-17.234	0.000	-253.117	-201.417
Outlet_Location_Type_1	202.3267	14.276	14.172	0.000	174.342	230.312
Outlet_Location_Type_2	-80.2609	16.687	-4.810	0.000	-112.972	-47.550
Outlet_Size_0	-89.3578	11.043	-8.092	0.000	-111.004	-67.712
Outlet_Size_1	314.4021	14.184	22.167	0.000	286.599	342.205
Outlet_Size_2	-330.2457	14.624	-22.583	0.000	-358.912	-301.579
Outlet_Type_0	-897.9003	16.858	-53.263	0.000	-930.946	-864.855
Outlet_Type_1	316.0327	15.062	20.982	0.000	286.507	345.558
Outlet_Type_2	-134.7124	16.128	-8.352	0.000	-166.328	-103.097
Outlet_Type_3	611.3787	12.762	47.905	0.000	586.361	636.396
Item_Type_Combined_0	-36.8859	29.720	-1.241	0.215	-95.144	21.372
Item_Type_Combined_1	-20.2123	20.292	-0.996	0.319	-59.990	19.565
Item_Type_Combined_2	-48.1032	24.920	-1.930	0.054	-96.953	0.747
Outlet_0	-467.5694	23.692	-19.735	0.000	-514.012	-421.127
Outlet_1	-89.3578	11.043	-8.092	0.000	-111.004	-67.712
Outlet_2	137.1682	30.199	4.542	0.000	77.971	196.365
Outlet_3	-134.7124	16.128	-8.352	0.000	-166.328	-103.097
Outlet_4	-430.3309	19.638	-21.913	0.000	-468.826	-391.836
Outlet_5	611.3787	12.762	47.905	0.000	586.361	636.396
Outlet_6	148.5790	30.608	4.854	0.000	88.579	208.579
Outlet_7	-83.4204	30.228	-2.760	0.006	-142.674	-24.167
Outlet_8	365.3278	23.434	15.590	0.000	319.391	411.264
Outlet_9	-162.2641	19.044	-8.520	0.000	-199.595	-124.933

Omnibus:	964.288	Durbin-Watson:	2.003
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2305.180
Skew:	0.669	Prob(JB):	0.00
Kurtosis:	5.169	Cond. No.	4.80e+16

OLS Regression Results

Dep. Variable:	Item_Outlet_Sales	R-squared:	0.563
Model:	OLS	Adj. R-squared:	0.563
Method:	Least Squares	F-statistic:	732.1
Date:	Fri, 19 Jul 2019	Prob (F-statistic):	0.00
Time:	10:57:12	Log-Likelihood:	-71991.
No. Observations:	8523	AIC:	1.440e+05
Df Residuals:	8507	BIC:	1.441e+05
Df Model:	15		
Covariance Type:	nonrobust		

Fig12. Summary from linear regression model

It is observed that the R-squared value is 0.563 for our dependent variable for 8523 number of observations taken under consideration. This signifies how accurately the built regression model fits.

CHAPTER 8
CONCLUSIONS

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 cantered on sales data from the past the accuracy of linear regression prediction can be enhanced with this method, polynomial regression, Ridge regression, and Xgboost regression can be determined. So, we can conclude ridge and Xgboost regression gives a better prediction with respect to Accuracy, MAE, and RMSE than the Linear and polynomial regression approaches.

In present era of digitally connected world every shopping mall desires to know the customer demands beforehand to avoid the shortfall of sale items in all sea-sons. Day to day the companies or the malls are predicting more accurately the Gopal Behera and Neeta Nain 11demand of product sales or user demands. Extensive research in this area atenterprise 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 research work, we have designed a predictive model by modifying Gradient boosting machines as Xgboost technique and ex-pedimented it on the 2013 Big Mart dataset for predicting sales of the product from a particular outlet. Experiments support that our technique produce more accurate prediction compared to than other available techniques like decision trees, ridge regression etc. Finally a comparison of different models is summa-razed in Table 2. From Table 2 it is also concluded that our model with lowest MAE and RMSE performs better compared to existing models.

CHAPTER 9
FUTURE ENHANCEMENTS

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 centres 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 concerned with 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 overall development 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 stage of regression model-building. There is a further need of experiments for proper measurements of both accuracy and resource efficiency to assess and optimize correctly.

CHAPTER 10

REFERENCES

1. https://www.researchgate.net/publication/344099746_SALES_PREDICTION_MODEL_FOR_BIG_MART

2. https://www.researchgate.net/publication/340252000_A_Comparative_Study_of_Big_Mart_Sales_Prediction