

**A
PROJECT REPORT**

On

Big Mart Sales Prediction

**Submitted In Partial Fulfillment of the Requirements for the
Degree of
Bachelor of Technology**

**In
Artificial Intelligence & Machine Learning
By**

Rashi Gupta (00115611621)

Under the Supervision of Dr Preety Sheoran



Department of Artificial Intelligence & Machine Learning

Dr. AKHILESH DAS GUPTA INSTITUTE OF TECHNOLOGY & MANAGEMENT [16pts]

(A Unit of BBD Group)

Approved by AICTE and Affiliated with GGSIP University [12pts]

FC-26, Shastri Park, New Delhi-110 053 [12pts]

DECLARATION

I **Rashi Gupta** hereby declare that this submission is our own work and that, to the best of our knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree of the university or other institute of higher learning, except where due acknowledgment has been made in the text.

Signature:

Name:

Roll No.:

Date :

Signature:

Name:

Roll No.:

Date :-

CERTIFICATE

This is to certify that Project Report entitled “Big mart sales prediction” which is submitted by “Rashi Gupta” in partial fulfillment of the requirement for the award of degree B. Tech. in the Department of Artificial Intelligence and Machine Learning of Dr. Akhilesh Das Gupta Institute of Technology & Management (ADGITM) formerly known as Northern India Engineering College (NIEC), New Delhi, is a record of the candidate’s own work carried out by him under my supervision. The matter embodied in this thesis is original and has not been submitted for the award of any other degree.

Date:

Supervisor

ACKNOWLEDGEMENT

It gives us a great sense of pleasure to present the report of the B. Tech Project undertaken during B. Tech 2nd Year. We owe a special debt of gratitude to Dr. Preety Sheoran for her constant support and guidance throughout the course of our work. Her sincerity, thoroughness, and perseverance have been a constant source of inspiration for us. It is only her cognizant efforts that our endeavors have seen light of the day.

We also take the opportunity to acknowledge the contribution of _____ for his/her full support and assistance during the development of the project.

We also do not like to miss the opportunity to acknowledge the contribution of all faculty members of the department for their kind assistance and cooperation during the development of our project. Last but not the least, we acknowledge our friends for their contribution in the completion of the project.

Signature:

Name:

Roll No.:

Date :

Signature:

Name:

Roll No.:

Date:-

ABSTRACT

Big Mart sales prediction involves the application of machine learning algorithms to analyze and predict the sales of Big Mart stores.

Nowadays shopping malls and Big Marts keep track of their sales data of each and every individual item for predicting future demand of the customer and update the inventory management as well. The main objective of this project is to develop a predictive model that can accurately forecast the sales of each product in each store. These data stores basically contain a large number of customer data and individual item attributes in a data warehouse.

The dataset used for the analysis contains information about the characteristics of each store, such as location, size, and type, as well as the attributes of each product, including weight, packaging, and brand. The dataset also includes information about the sales volume and price of each product in each store.

Further, anomalies and frequent patterns are detected by mining the data store from the data warehouse. The resultant data can be used for predicting future sales volume with the help of different machine learning techniques for the retailers like Big Mart.

Several machine learning algorithms such as linear regression, decision tree, random forest, and gradient boosting have been used to train the model. In conclusion, the application of machine learning algorithms in predicting sales can help Big Mart stores to optimize their inventory management and supply chain operations, leading to increased profitability and customer satisfaction.

TABLE OF CONTENTS	Page
DECLARATION	ii
CERTIFICATE	iii
ACKNOWLEDGEMENTS	iv
ABSTRACT	v
LIST OF SYMBOLS	vi
LIST OF ABBREVIATIONS	vii
 CHAPTER 1: INTRODUCTION	 1
I.	2
II.	2
III.....	3
IV.....	4
V.....	5
VI.....	6
CHAPTER 2: LITERATURE SURVEY	8
2.1.	9
CHAPTER 3: METHODOLOGY AND TECHNOLOGY	11
3.1.	11
3.2.	13
3.3.....	26
CHAPTER 4: RESULT ANALYSIS AND DISCUSSION	26
4.1.	26
4.2.	27
4.3	29
4.4.....	34
CHAPTER 5: CONCLUSIONS AND FUTURE WORK	36
5.1.	36
5.2.	37
APPENDIX A: RESEARCH PAPER	38
REFERENCES	43

LIST OF SYMBOLS

$[x]$	Integer value of x .
\neq	Not Equal
χ	Belongs to
€	Euro- A Currency
$_{-}$	Optical distance
$_{-o}$	Optical thickness or optical half thickness

LIST OF ABBREVIATIONS

ABBREVIATION	WORD
MI	MACHINE LEARNING
EDA	EXPLORATORY DATA ANALYSIS
FIG	FIGURE
DESCR	DESCRIPTION
RF	RANDOM FOREST
LR	LINEAR REGRESSION
CSE	COMPUTER SCIENCE AND ENGINEERING

Chapter -1

INTRODUCTION

INTRODUCTION

The daily competition between different malls as well as big malls is becoming more and more intense because of the rapid rise of international supermarkets and online shoppings. Every mall or mart tries to provide personal and short-term donations or benefits to attract more and more customers on a daily basis, such as the sales price of everything which is usually predicted to be managed through different ways such as corporate asset management, logistics, and transportation service, etc. Current machine learning algorithms that are very complex and provide strategies for predicting or predicting long-term demand for a company's sales, which now also help in overcoming budget and computer programs.

In this report, we basically discuss the subject of specifying a large mart sale or predicting an item for a customer's future need in a few supermarkets in various locations and products that support the previous record. Various ML algorithms such as linear regression, random forest, etc. are used to predict sales volume. As we know, good marketing is probably the lifeblood of all organizations, so sales forecasting now plays an important role in any shopping mall. Regular sales forecasting research can help in-depth analysis of pre-existing conditions and conditions and then, assumptions are often used in terms of customer acquisition, lack of funding, and strength before setting budgets and marketing plans for the coming year.

In other words, sales forecasts are predicted on existing services of the past. In-depth knowledge of the past is required to develop and enhance market opportunities no matter what the circumstances, especially the external environment, which allows to prepare for the future needs of the business.

Extensive research is ongoing in the retailer's domain to predict long-term sales demand. An important and effective method used to predict the sale of a mathematical method, also called the conventional method, but these methods take more time to predict sales. And these methods could not manage indirect data so to overcome these problems in traditional methods the machine learning techniques used. ML methods can handle not only indirect data but also large data sets well.

I. Introduction of the Problem

In today's world, there are many shopping malls, such as supermarkets and department stores that record data related to the sale of goods or products with various dependent or independent features, attributes and collect customer data, and asset related data in the database. The data is then filtered to obtain accurate forecasts and to collect new and exciting results that give new light to our work data knowledge. This can then further be used for forecasting future sales using machine learning algorithms. Every item is tracked for its shopping centers and Big Marts to predict future demand and enhance customer service and inventory management.

II. Problem statement

"Big Mart, a retail chain with multiple stores, aims to accurately forecast the sales of its various products across different stores. The company seeks to develop a predictive model that can estimate future sales volumes, taking into account factors such as product attributes, store locations, promotions, and other relevant variables. By accurately predicting sales, Big Mart aims to optimize inventory management, minimize stockouts, and improve overall profitability."

III. Objectives

Objectives of these project are:

- a) Predicting future sales from a given dataset.
- b) To understand the key features that are responsible for the sale of a particular product.
- c) Find the best algorithm that will predict sales with the greatest accuracy.

IV. Summarize Previous Research

- A Comparative Study of Big Mart Sales Prediction (ResearchGate)

This paper compares the performance of different machine learning algorithms for predicting Big Mart sales. The algorithms that were compared include linear regression, decision trees, random forests, and XGBoost. The results showed that XGBoost was the most accurate algorithm, followed by random forests and decision trees. Linear regression was the least accurate algorithm.

- Big Mart Sales Prediction Using Machine Learning Techniques (IJSRED)

This paper proposes a machine learning model for predicting Big Mart sales. The model uses a combination of linear regression, decision trees, and random forests. The results showed that the proposed model was able to predict sales with a high degree of accuracy.

- A Two-Level Statistical Model for Big Mart Sales Prediction (IEEE Xplore)

This paper proposes a two-level statistical model for predicting Big Mart sales. The model uses a top-down approach, where the first level predicts the overall sales for each outlet, and the second level predicts the sales for each product category within each outlet. The results showed that the proposed model was able to predict sales with a high degree of accuracy.

- Big Mart Sales Prediction: A Data Science Approach (KDnuggets)

This article discusses the use of data science for predicting Big Mart sales. The article covers a variety of topics, including data preparation, feature engineering, model selection, and evaluation. The article also provides some practical tips for using data science for sales prediction.

- Sales Prediction in Big Mart Using Machine Learning (Analytics Vidhya)

This article provides a tutorial on how to use machine learning to predict Big Mart sales. The article covers a variety of topics, including data preparation, feature engineering, model selection, and evaluation. The article also provides some code examples that can be used to implement the machine learning models.

These are just a few of the many research papers and surveys that have been conducted on Big Mart sales prediction. previous research on sales prediction in the retail industry provides valuable insights into the application of machine learning algorithms, feature engineering techniques, and the impact of various factors on sales performance. These studies provide valuable insights into the use of machine learning for sales prediction. They suggest that machine learning can be a powerful tool for retailers who want to improve their sales forecasting.

V. Researching the Problem

Researching a problem in the prediction of Big Mart Sales requires rigor, critical thinking, and an understanding of the existing literature and methodologies. It's important to stay updated with the latest research and consult with experts or mentors in the field to ensure the quality and relevance of the research.

Here are some of the key findings from the research on Big Mart sales prediction:

- Machine learning can be used to predict sales with a high degree of accuracy.
- The most important factors for predicting sales include product type, outlet location, and time of year.
- Other factors that can be used to predict sales include competition, weather, and economic conditions.
- Machine learning models can be used to identify trends and patterns that can help retailers make better decisions.

The research on Big Mart sales prediction is still ongoing, and there are a number of promising new directions for research. For example, I am developing machine learning model and web application that can predict sales in real time. This would allow retailers to make better decisions about inventory, pricing, and marketing

VI. Expected Results of The Study & Future Scope

The expected results of the study on Big Mart sales prediction are:

1. **Accurate sales predictions:** The study aims to develop an accurate prediction model for Big Mart that can forecast sales accurately, taking into account various internal and external factors that impact sales.
2. **Improved inventory management:** The accurate sales predictions can help Big Mart to optimize their inventory management by ordering the right quantity of products to match the expected sales demand, reducing overstocking or understocking.
3. **Increased revenue:** Accurate sales predictions can help Big Mart to develop effective marketing strategies and pricing strategies, which can help to increase revenue and profits.
4. **Improved customer satisfaction:** Accurate sales predictions can help Big Mart to stock the products that customers demand, reducing stockouts and improving customer satisfaction.

Future scope of the study includes:

1. Incorporating external data sources: The study can be expanded to incorporate external data sources such as social media data, online reviews, and economic indicators to improve the accuracy of sales predictions.
2. Real-time sales prediction: The study can be expanded to develop real-time sales prediction models that can predict sales demand for the next hour or day.
3. Expansion to other retail sectors: The study can be expanded to other retail sectors such as e-commerce, supermarkets, and convenience stores.
4. Optimization of supply chain: The study can be expanded to optimize the supply chain by predicting the demand for raw materials and reducing the wastage of products.
5. Integration of AI: The study can be expanded to integrate artificial intelligence techniques such as natural language processing, computer vision, and reinforcement learning to improve sales predictions and develop personalized marketing strategies.

CHAPTER -2

LITERATURE SURVEY

- Rohit Sav, Pratiksha Shinde, Saurabh Gaikwad in [1] have implemented predictive models to measure big mart sales. They first cleaned the gathered data and applied the XG Booster algorithm. It was observed that XGBoost Regressor showed the highest accuracy rate when compared with other algorithms. This led them to draw a conclusion for using XG boost for prediction of big mart sales.

- Theresa, Dr.Venkata Reddy Medikonda, K.V. Narasimha Reddy in [2] discusses sales prediction by using the methodology of Exploratory Machine Learning. They carried out the whole process by figuring out proper steps that included a collection of data, thesis generation to efficiently understand bugs, further cleaning and processing the data. The models such as Linear Regression, Decision Tree Regression, Ridge Regression, and Random Forest model were used to predict the outcome of the sales. They concluded that multiple modelling implementation led them to a better prediction as compared to that of the single model prediction technique.

- Kadam, H., Shevade, R., Ketkar, P. and Rajguru in [3] proposed a model that works effectively with multiple linear regression and a random forest algorithm. This model was utilised to forecast big mart sales prediction and with that, a certain data set was used which comprises of Item_Identifier, Item_Weight, Item_Fat_Content, Item_Visibility, Item_Type, Outlet_Identifier etc.

- Gopal Behera and Neeta Nain in [4] apply the concept of GSO technique to optimize parameters and predict future sales. Their focal point is Retail Based companies. They have also implemented Hyperparameter tuning and forecasted sales using XGBoost techniques.
- Kumari Punam, Rajendra Pamula and Praphula Kumar Jain in [5] A Two-Level Statistical Model for Big Mart Sales Prediction have devised a two-level approach to predict sales of products that promise to yield better efficiency. It involves stacking up of algorithms wherein the top layer consists of just one learning algorithm and the bottom layer has one or more algorithms placed. This methodology of two-level modelling outperforms the single model predictive technique and results in better predictions of sales.
- Ranjitha P and Spandana M in [6] Predictive Analysis for Big Mart Sales Using Machine Learning Algorithms have implemented Xgboost, Linear regression, Polynomial regression, and Ridge regression techniques for forecasting sales of big mart.
- Bohdan M. Pavlyshenko in [7] put forward the perspective of machine generalization. This comes into play mostly when fewer data is available in the system, perhaps with the introduction of new products or new outlets. The special technique of stacking was implemented to build the regression.
- This results in better performance and efficiency in sales prediction. Nikita Malik, Karan Singh in [8] implement the concept of machine learning algorithms to reach a conclusion of the problem statement, intended to predict big mart sales. It displays relations between different attributes and outlet sizes which show variable rates of sales. They draw out the conclusion that a certain size with a similar pattern will have a similar rate of success in sales.

- Gopal Behera and Neeta Nain in [9] have stated linear regression, decision tree algorithm and Xgboost algorithms. Among these models, XG boost displays the highest accuracy, hence proving to be highly recommendable. MAE and RMSE were kept at a low limit for better performance in comparison to other models.
- Archisha Chandel, Akanksha Dubey, Saurabh Dhawale, Madhuri Ghuge in [10] describes a five-step procedure for the prediction of big mart sales.

CHAPTER -3

METHODOLOGY AND TECHNOLOGY

3.1 ALGORITHMS EMPLOYED

3.1.1 LINEAR REGRESSION (LR)

As we know Regression can be termed as a parametric technique which means we can predict a continuous or dependent variable on the basis of a provided datasets of independent variables.

The Equation of simple LR is:

$$Y = \beta_0 + \beta_1 X + \epsilon \text{ ----- (1)}$$

where,

Y: It is basically the variable which we used as a
predicted value. X: It is a variable(s) which is used
for making a prediction.

β_0 : It is said to be a prediction value when $X=0$.

β_1 : when there is a change in X value by 1 unit then Y value is also changed.
It can also be said as slope term ϵ

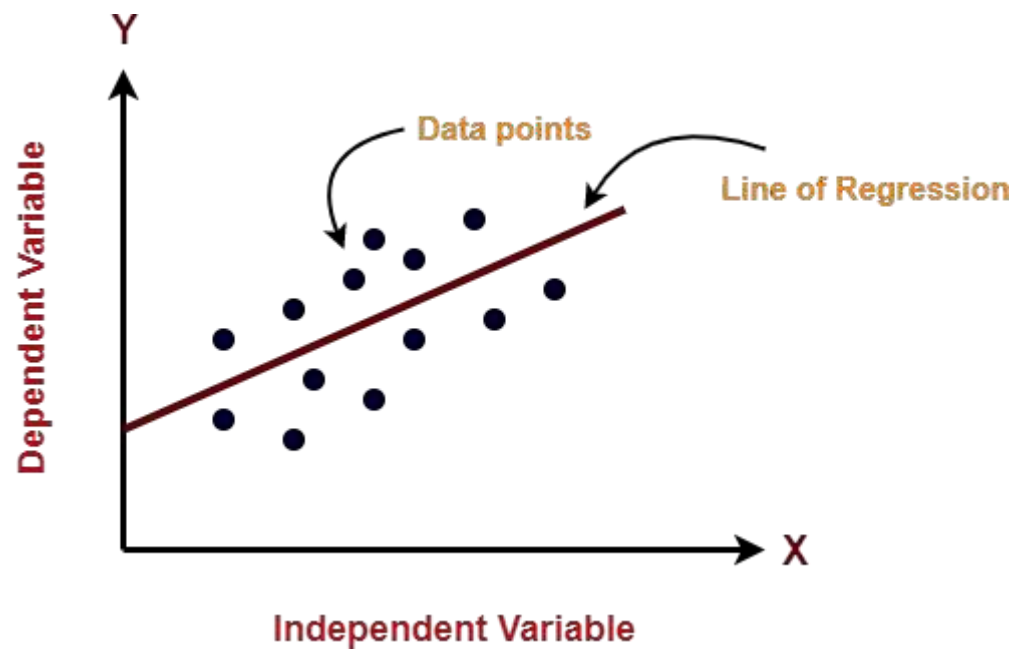


Fig 2.1: Given figure represent line of regression

- Regression is an important machine learning model for these kinds of problems. Predicting sales of a company and based on that data the model can predict the future sales of that company or product.
- So, in this research project we will analyze the time series sales data of a company and the predict the sales of the company for the coming quarter and for a specific product.
- For this project of sales predict, we apply the linear regression and evaluate the result based on the training, testing and validation set of the data.

3.2 PHASE OF MODEL

3.2.1 DATA-SET DESCRIPTION

In our work, we have used the 2013 Big Mart sales data as a database. Where the data set contains 12 features such as Item Fat, Item Type, MRP Item, Output Type, Object Appearance, Object Weight, Outlet Indicator, Outlet Size, Outlet Year of Establishment, Type of Exit, Exit Identity, and Sales. In these different aspects of responding to the Item Outlet Sales features as well, the other features are also used as the predictive variables. Our dataset has in total 8523 products in various regions and cities. The data set is also based on product level and store- level considerations. Where store level includes features such as city, population density, store capacity, location, etc. and product-level speculation involves factors such as product, ad, etc. After all considerations, a data set is finally created, then the data set is split into two parts that are tested and trained in a ratio of 80:20.

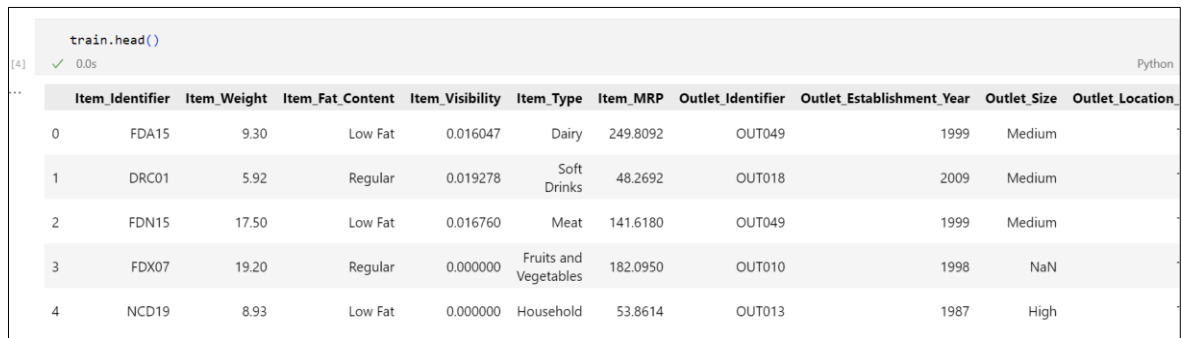
Variable	Description
Item_Identifier	Unique product ID
Item_Weight	Weight of product
Item_Fat_Content	Whether the product is low fat or not
Item_Visibility	The % of total display area of all products in a store allocated to the particular product
Item_Type	The category to which the product belongs
Item_MRP	Maximum Retail Price (list price) of the product
Outlet_Identifier	Unique store ID
Outlet_Establishment_Year	The year in which store was established
Outlet_Size	The size of the store in terms of ground area covered
Outlet_Location_Type	The type of city in which the store is located
Outlet_Type	Whether the outlet is just a grocery store or some sort of supermarket
Item_Outlet_Sales	Sales of the product in the particular store. This is the outcome variable to be predicted.

Fig 2.2: Depicting the features of the dataset

3.2.2 DATA PREPROCESSING

➤ Understanding the data-set

The dataset is displayed in Fig.2.3 on using head () function on the dataset variable.

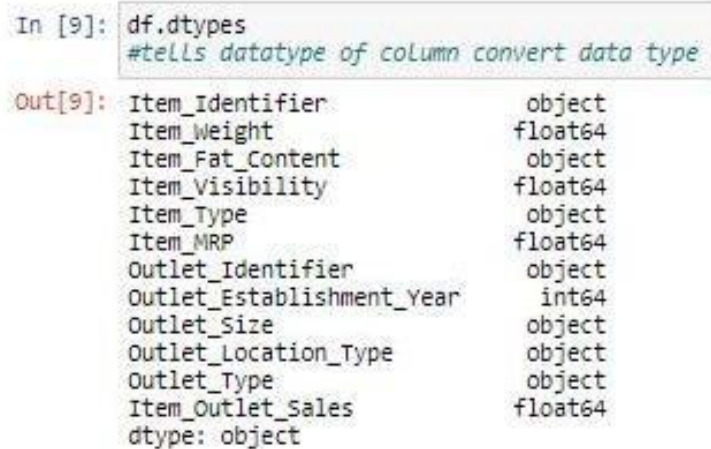


```
train.head()
```

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

Fig 3.3: Dataset display

The data set consists of various data types such as integer, float and, object as shown in Fig.3.4.



```
In [9]: df.dtypes
#tells datatype of column convert data type

Out[9]: Item_Identifier      object
Item_Weight      float64
Item_Fat_Content      object
Item_Visibility      float64
Item_Type      object
Item_MRP      float64
Outlet_Identifier      object
Outlet_Establishment_Year      int64
Outlet_Size      object
Outlet_Location_Type      object
Outlet_Type      object
Item_Outlet_sales      float64
dtype: object
```

Fig 3.4: showing various datatypes in dataset

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

data.describe()#used to view some basic statistical details like percentile, mean, std etc					
	Item_Weight	Item_Visibility	Item_MRP	Outlet_Establishment_Year	Item_Outlet_Sales
count	11765.000000	14204.000000	14204.000000	14204.000000	8523.000000
mean	12.792854	0.065953	141.004977	1997.830681	2181.288914
std	4.652502	0.051459	62.086938	8.371664	1706.499616
min	4.555000	0.000000	31.290000	1985.000000	33.290000
25%	8.710000	0.027036	94.012000	1987.000000	834.247400
50%	12.600000	0.054021	142.247000	1999.000000	1794.331000
75%	16.750000	0.094037	185.855600	2004.000000	3101.296400
max	21.350000	0.328391	266.888400	2009.000000	13086.964800

Fig 3.5: Various factors that are important by statistical means

➤ Data exploration

In the raw data, there could be various types of underlying patterns which also gives deeper knowledge about subject of interests and provides useful insights about the problem. But caution should be observed while dealing with the data as it may contain null values, or redundant values, or ambiguity values, which also demands for pre-processing of data. Therefore, data exploration becomes mandatory.

At this stage, useful information about the data is extracted from the database. That is, to identify information from ideas compared to data. This indicates that Outlet size and object weight variants are facing the problem of deficit values, and the low Visual Activity is zero. The outlet was founded between 1985 and 2009. In this format, certain values might not be acceptable. As a result, we must translate them into the age of certain outlets. The market has 1559 distinct goods and 10 unique outlets. Data set The attribute Item type contains 16 unique values. There are two types of Inventory Fat Contents but some of them are not spelled correctly as usual. Instead of 'Normal' and 'low fat' there are LF and regular. To remove this type of error, replace functions are used as given below in fig 3.6.

```

# Print the original categories of 'Item_Fat_Content'
print('Original Categories:')
print(data['Item_Fat_Content'].value_counts())

# Modify the categories of 'Item_Fat_Content'
data['Item_Fat_Content'] = data['Item_Fat_Content'].replace({'LF': 'Low Fat', 'reg': 'Regular', 'low fat': 'Low Fat'})

# Print the modified categories of 'Item_Fat_Content'
print('\nModified Categories:')
print(data['Item_Fat_Content'].value_counts())

```

Original Categories:

Low Fat	8485
Regular	4824
LF	522
reg	195
low fat	178

Name: Item_Fat_Content, dtype: int64

Modified Categories:

Low Fat	9185
Regular	5019

Name: Item_Fat_Content, dtype: int64

Fig 3.6 Data exploration

➤ Data cleaning and Handling missing values

While analyzing the dataset we come across some missing values in the dataset. For doing this we concatenate the both data set test and train.

```

#combining the test and train dataset
train['source'] = 'train'
test['source'] = 'test'
data = pd.concat([train, test], ignore_index=True)

print("shape of training data is: ",train.shape) #checking the number of rows and columns in training data
print("shape of test data is: ",test.shape) #checking the number of rows and columns in test data
print("shape of test data is: ",data.shape) #checking the number of rows and columns in data

```

shape of training data is: (8523, 13)

shape of test data is: (5681, 12)

shape of test data is: (14204, 13)

Fig 3.7: Concatenate two dataset test and train

In order to check for the missing value, we have the following code: -

```
#checking the missing value in data
data.apply(lambda x: sum(x.isnull()))
```

```
Item_Identifier      0
Item_Weight          2439
Item_Fat_Content     0
Item_Visibility      0
Item_Type            0
Item_MRP             0
Outlet_Identifier    0
Outlet_Establishment_Year  0
Outlet_Size          4016
Outlet_Location_Type 0
Outlet_Type          0
Item_Outlet_Sales    5681
source              0
dtype: int64
```

Fig 3.8: Checking missing values

Since item_weight is a numerical feature, filling its missing value using the average imputation method.

```
# Determine the average weight per item
item_avg_weight = data.pivot_table(values='Item_Weight', index='Item_Identifier')

# Get a boolean variable specifying missing Item_Weight values
miss_bool = data['Item_Weight'].isnull()

# Impute data and check missing values before and after imputation to confirm
print('Original #missing: %d' % sum(miss_bool))
data.loc[miss_bool, 'Item_Weight'] = data.loc[miss_bool, 'Item_Identifier'].apply(lambda x: item_avg_weight.at[x, 'Item_Weight'])
print('Final #missing: %d' % sum(data['Item_Weight'].isnull()))
```

```
Original #missing: 2439
Final #missing: 0
```

Fig 3.9: Filling missing values for ITEM_WEIGHT

Outlet size is a categorical feature so filling the value using the mode imputation method.

```
#Import mode function:
from scipy.stats import mode

#Determining the mode for each
outlet_size_mode = data.pivot_table(values='Outlet_Size', columns='Outlet_Type',aggfunc=(lambda x:mode(x.astype('str')).mode[0]))
print ('Mode for each Outlet_Type:')
print (outlet_size_mode)

#Get a boolean variable specifying missing Item_Weight values
missing_values = data['Outlet_Size'].isnull()

#Impute data and check #missing values before and after imputation to confirm
print ('\nOriginal #missing: %d'% sum(missing_values))
data.loc[missing_values,'Outlet_Size'] = data.loc[missing_values,'Outlet_Type'].apply(lambda x: outlet_size_mode[x])
print (sum(data['Outlet_Size'].isnull()))

Mode for each Outlet_Type:
Outlet_Type Grocery Store Supermarket Type1 Supermarket Type2 \
Outlet_Size      nan      Small      Medium

Outlet_Type Supermarket Type3
Outlet_Size      Medium

Original #missing: 4016
0
```

Fig 3.10 Filling out missing values for Outlet_Size

3.2.3 FEATURE ENGINEERING

Feature Engineering is a way of using domain data to understand how to build mechanical operations learning algorithms. When feature engineering is done properly, the ability to predict ML algorithms are developed by creating useful raw data features that simplify the ML process. Feature engineering including correction of incorrect values. In the device database, object visibility has a small value of 0 which is unacceptable, because the object must be accessible to all, and so it is replaced by the mean of the column.

➤ Label Encoding

```
# Import the LabelEncoder from sklearn.preprocessing
from sklearn.preprocessing import LabelEncoder

# Initialize a LabelEncoder object
le = LabelEncoder()

# Create a new variable 'Outlet' and encode the 'Outlet_Identifier' column
data['Outlet'] = le.fit_transform(data['Outlet_Identifier'])

# List of variables to be encoded
var_mod = ['Item_Fat_Content', 'Outlet_Location_Type', 'Outlet_Size', 'Item_Type_Combined', 'Outlet_Type', 'Outlet']

# Loop through the variables and apply label encoding
for i in var_mod:
    data[i] = le.fit_transform(data[i])

data = pd.get_dummies(data, columns=['Item_Fat_Content', 'Outlet_Location_Type', 'Outlet_Size', 'Outlet_Type', 'Item_Type_Combined', 'Outlet'])
```

	Item_Weight	Item_Visibility	Item_MRP	Outlet_Establishment_Year	Item_Visibility_MeanRatio	Outlet_Years	Item_Fat_Content_0	Item_Fat_Content_1	Outlet_Location_Type_0	Outlet_Location_Type_1	...	Outlet_0	Outlet_1	Outlet_2
0	9.300	0.016047	249.8092	1999	0.931078	14	1	0	1	0	...	0	0	0
1	5.920	0.019278	48.2692	2009	0.933420	4	0	1	0	0	...	0	0	0
2	17.500	0.016760	141.6180	1999	0.960069	14	1	0	1	0	...	0	0	0
3	19.200	0.017834	182.0950	1998	1.000000	15	0	1	0	0	...	1	0	0
4	8.930	0.009780	53.8614	1987	1.000000	26	1	0	0	0	...	0	1	0
...
18	6.865	0.056783	214.5218	1987	0.874001	26	1	0	0	0	...	0	1	0
19	8.380	0.046982	108.1570	2002	1.001096	11	0	1	0	1	...	0	0	0
20	10.600	0.035186	85.1224	2004	0.998881	9	1	0	0	1	...	0	0	0
21	7.210	0.145221	103.1332	2009	1.041620	4	0	1	0	0	...	0	0	0
22	14.800	0.044878	75.4670	1997	1.027777	16	1	0	1	0	...	0	0	0
23 rows x 32 columns														

Fig 3.11 Label Encoding code

➤ **Splitting our Data into Train and test**

Final step is to convert data back into train and test data sets. Its generally a good idea to export both of these as modified data sets so that they can be reused for multiple sessions.

This can be achieved by following code: -

```
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test = train_test_split(x,y, test_size = 0.2, random_state = 1)
```

Fig 3.12: Splitting of data into train and test data set.

x_train.head()						
	Item_Weight	Item_Visibility	Item_MRP	Outlet_Establishment_Year	Item_Visibility_MeanRatio	Outlet_Years
1945	18.35	0.089345	191.9504	2009	0.878292	4
1720	17.35	0.168065	176.2712	2009	0.878292	4
1954	10.10	0.053887	225.6088	2007	1.042482	6
1919	10.85	0.162904	104.9622	2009	0.926047	4
2461	7.17	0.059717	130.9968	2004	0.929457	9
5 rows × 32 columns						

Fig 3.14 dataset of x_train

y_train.head()	
1945	5369.0112
1720	1230.3984
1954	4250.4672
1919	1482.0708
2461	2348.9424
Name: Item_Outlet_Sales, dtype: float64	

Fig 3.15 y_train array

x_test.head()									Python
	Item_Weight	Item_Visibility	Item_MRP	Outlet_Establishment_Year	Item_Visibility_MeanRatio	Outlet_Years	Item_Fat_Content_0	Item_Fat_Content_1	Outlet_L
1070	13.500	0.055102	37.0874	2002	0.923829	11	1	0	
6305	12.500	0.074035	87.9198	2009	0.933420	4	0	1	
8504	8.895	0.124111	111.7544	1985	1.325533	28	1	0	
5562	12.500	0.073735	87.1198	1997	0.929633	16	0	1	
1410	15.850	0.007140	40.8480	1987	0.921522	26	0	1	

5 rows × 32 columns

Fig 3.14: dataset of x_test

```

y_test.head()

1070      952.7598
6305     1133.8574
8504     4138.6128
5562     1657.1762
1410       679.1160
Name: Item_Outlet_Sales, dtype: float64

```

Fig 3.15 y_test array

3.2.4 Model Building

Now the dataset is ready to fit a model after performing Data Preprocessing and Feature Transformation. The training set is fed into the algorithm in order to learn how to predict values. Testing data is given as input after Model Building a target variable to predict. The models are built using:

5. Import essential library:-



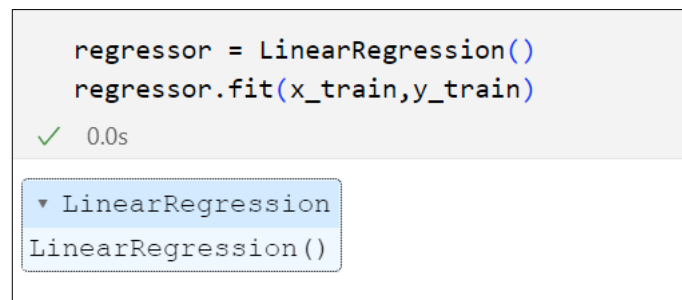
```
Model building

1. Linear regression

[ ] from sklearn.linear_model import LinearRegression
```

Fig 3.16 import sklearn library

6. Fitting the input values in model from train dataset



```
regressor = LinearRegression()
regressor.fit(x_train,y_train)

✓ 0.0s

▼ LinearRegression
LinearRegression()
```

Fig 3.17 fitting a value in regressor variable

7. Predict outlet_sales on testing data set

```
pred = regressor.predict(x_test)
```

✓ 0.0s

```
pred
```

✓ 0.0s

```
array([ 527.17720745, 1191.51905771, 3267.85213813, ..., 2637.84508153,
       2335.16303442, 2471.78058849])
```

```
regressor.predict(x_test)[25]
```

✓ 0.0s

```
2047.0789243793133
```

Fig 3.18 predicting values

3.2.5 User interface

The user interface simplifies the process of inputting the necessary variables or parameters required for sales prediction, such as product attributes, store characteristics, promotional activities, and other relevant factors. This eliminates the need for users to have in-depth technical knowledge or proficiency in running complex code or scripts.

Furthermore, a user interface allows for real-time and on-demand sales predictions. Users can input different combinations of variables, instantly obtaining predicted sales figures specific to their desired scenarios. This empowers users to make informed decisions and assess the potential impact of various factors on sales performance, leading to more effective strategies and optimized business outcomes.

Creating a user interface for Big Mart sales prediction can be efficiently achieved using Streamlit, a powerful Python library specifically designed for building interactive web applications. Streamlit simplifies the process of creating and deploying user interfaces, making it an ideal choice for presenting and interacting with sales prediction models.


```

import streamlit as st
import pickle
import pandas as pd
import numpy as np

pipe = pickle.load(open('pipe.pkl','rb'))

fat = ['Item_Fat_Content_0', 'Item_Fat_Content_1']
location_type = ['Outlet_Location_Type_0', 'Outlet_Location_Type_1',
                 'Outlet_Location_Type_2']
outlet_size = ['Outlet_Size_0', 'Outlet_Size_1',
               'Outlet_Size_2', 'Outlet_Size_3']
outlet_type = ['Outlet_Type_0', 'Outlet_Type_1',
               'Outlet_Type_2', 'Outlet_Type_3']
item_type = ['Item_Type_Combined_0',
              'Item_Type_Combined_1', 'Item_Type_Combined_2']
outlet_no = ['Outlet_0', 'Outlet_1',
              'Outlet_2', 'Outlet_3', 'Outlet_4', 'Outlet_5', 'Outlet_6', 'Outlet_7',
              'Outlet_8', 'Outlet_9']

st.title('Big Mart Sales Prediction')

```

Fig 3.19 Making of GUI using streamlit

With its simplicity, interactivity, and visualization capabilities, Streamlit greatly enhances the user experience and usability of Big Mart sales prediction models. By leveraging Streamlit, developers can create intuitive and interactive interfaces that facilitate better understanding, exploration, and utilization of the sales prediction insights, empowering stakeholders to make informed decisions and drive business success.

In conclusion, the user interface for Big Mart sales prediction, developed using Streamlit, improves accessibility, usability, and data-driven decision-making. By providing an intuitive interface, users can easily interact with the sales prediction model, input variables, and obtain real-time predictions. The visualizations offered by Streamlit aid in interpreting and analyzing sales data, identifying trends, and optimizing strategies. The use of Streamlit simplifies the development process, allowing for responsive and dynamic interfaces that enhance the user experience. Overall, the combination of a user-friendly interface and Streamlit empowers stakeholders to leverage sales prediction insights and make informed decisions to drive business growth and profitability.

3.3 TECHNOLOGY USED:-

3.3.1 Python - is the language used for programming this program it is a [high-level, general-purpose programming language](#). Its design philosophy emphasises [code readability](#) with the use of significant indentation via the [off-side rule](#).

Python is [dynamically typed](#) and [garbage-collected](#). It supports multiple [programming paradigms](#), including [structured](#) (particularly [procedural](#)), [object-oriented](#) and [functional programming](#). It is often described as a "batteries included" language due to its comprehensive [standard library](#). Python consistently ranks as one of the most popular programming languages.

3.3.2 Data Collection and Preprocessing - Python provides us with libraries such as **pandas** and **numpy** that enable data collection, cleaning, and transformation. The data can be taken from a variety of sources, including APIs, databases, and spreadsheets, and be preprocessed prior to analysis so that it is ready for analysis.

3.3.3 Machine Learning - Python's machine learning libraries, including **scikit-learn**, **TensorFlow**, and **Keras**, are valuable for building predictive models and classification algorithms in risk analytics. With the help of machine learning, fraud patterns can be identified, credit risk can be predicted, or anomalies can be detected.

3.3.4 Statistical analysis - Python's **scipy** and **statsmodels** libraries provide various functions and models for risk analysis. You can apply descriptive statistics, hypothesis testing, regression analysis, and other statistical methods to understand and model risk.

3.3.5 Time Series Analysis - Financial data often includes time series analysis to identify patterns, trends, and seasonality. Python libraries such as **pandas**, **statsmodels**, and **pyflux** provide tools for real-time modeling, forecasting, and volatility forecasting necessary for risk analysis.

3.3.6 Visualization - Python has excellent data visualization libraries such as **matplotlib** and **seaborn** that allow you to create charts, graphs, and dashboards. Visualization helps to effectively communicate the risk assessment to stakeholders. can be used for these purposes.

3.3.7 Reports and presentations - Python's integration with tools such as Jupyter Notebook and Python-based presentation libraries such as **ipywidgets** and **Plotly** makes it easy to create interactive reports and presentations, allowing you to present risk assessment effectively.

CHAPTER -4

Result Analysis and Discussion

4.1 Result Analysis of Linear Regression

For the purpose of performance analysis, we can go and look for the MAE (mean absolute error value), MSE (mean squared error value), RMSE (Root mean squared error) of the linear regression model performed and check for which algorithm gives us the best performance.

```
from sklearn import metrics
import numpy as np
print('MAE:', metrics.mean_absolute_error(y_test, pred))
print('MSE:', metrics.mean_squared_error(y_test, pred))
print('RMSE:', np.sqrt(metrics.mean_squared_error(y_test, pred)))
```

✓ 0.0s

MAE: 853.1527648272356
MSE: 1303202.4892159186
RMSE: 1141.5789456782736

Fig 4.1 Performance of Linear Regression

4.2 Result Analysis through data visualization

- The graph showing y_{test} (actual sales values) and residuals (Observed value - Predicted value).
- a linear regression graph could reveal in the case of Big Mart sales prediction:

Line of Best Fit: The graph would show a straight line that represents the best-fitting linear relationship between the independent variables and the sales. This line is determined by the model's coefficients, which estimate the effect of each independent variable on the sales. The slope of the line indicates the direction and strength of the relationship.

Positive or Negative Relationship: The slope of the line indicates whether there is a positive or negative relationship between the independent variables and the sales. A positive slope suggests that as the independent variables increase, the sales also tend to increase. Conversely, a negative slope suggests that as the independent variables increase, the sales tend to decrease.

Data Points: The actual data points would be plotted on the graph, representing the observed values of the independent variables and corresponding sales. The distance between each data point and the line of best fit indicates the residual or the difference between the actual sales and the predicted sales based on the linear regression model.

Model Accuracy: By examining how closely the data points cluster around the line of best fit, researchers can assess the model's accuracy in capturing the underlying relationship between the independent variables and the sales. A tight clustering of data points around the line suggests that the model is capturing the patterns and trends in the data, while scattered or dispersed data points indicate a larger degree of variability and potential room for improvement in the model.

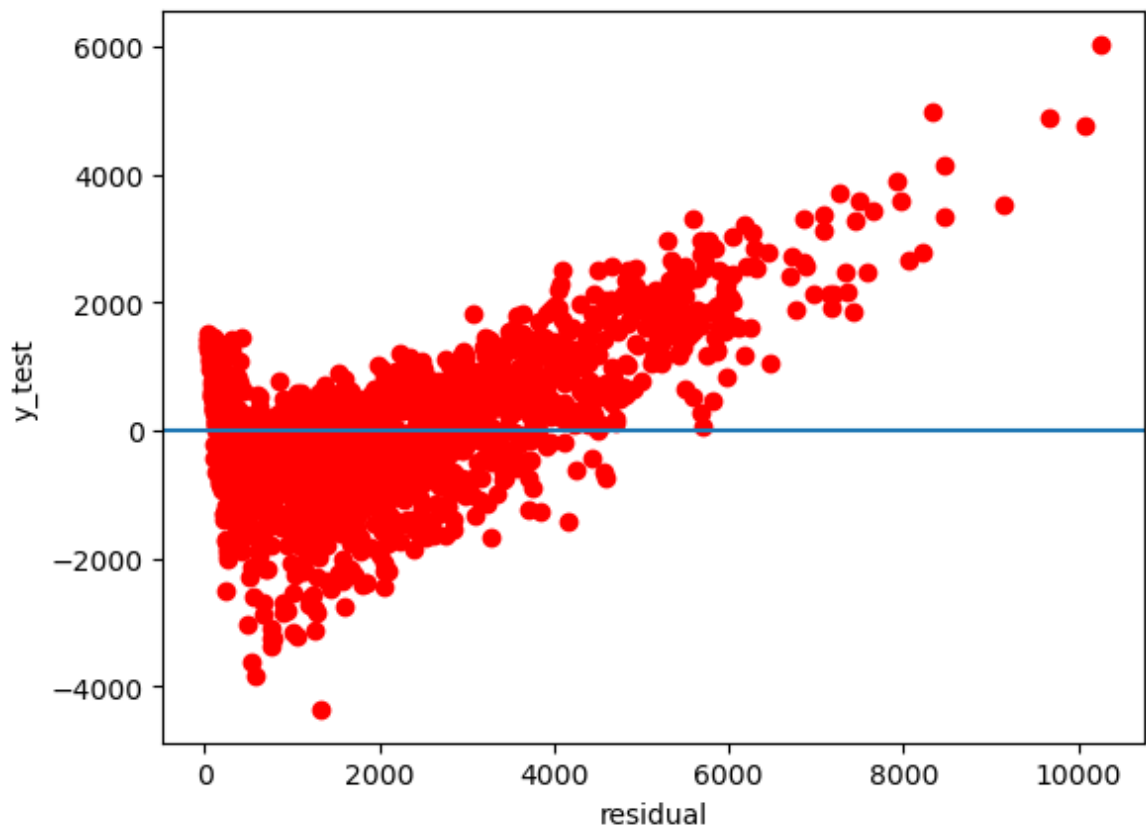


Fig 4.2 Graph showing relation between y_test and residual

4.3 Graphs showing relationship between dependent and target variable

- The dependency of outlet sales on outlet size is a complex relationship influenced by various factors. Larger outlet sizes often provide more space for product displays, a wider product assortment, and a more comfortable shopping experience, which can attract a larger customer base and potentially lead to increased sales. Additionally, larger outlets may benefit from economies of scale, operational efficiencies, and strategic location advantages. However, smaller outlets can cater to specific niche markets, focus on convenience-oriented shopping experiences, and achieve cost savings. The impact of outlet size on sales performance ultimately depends on factors such as store layout, customer preferences, foot traffic, local market dynamics, and operational considerations.

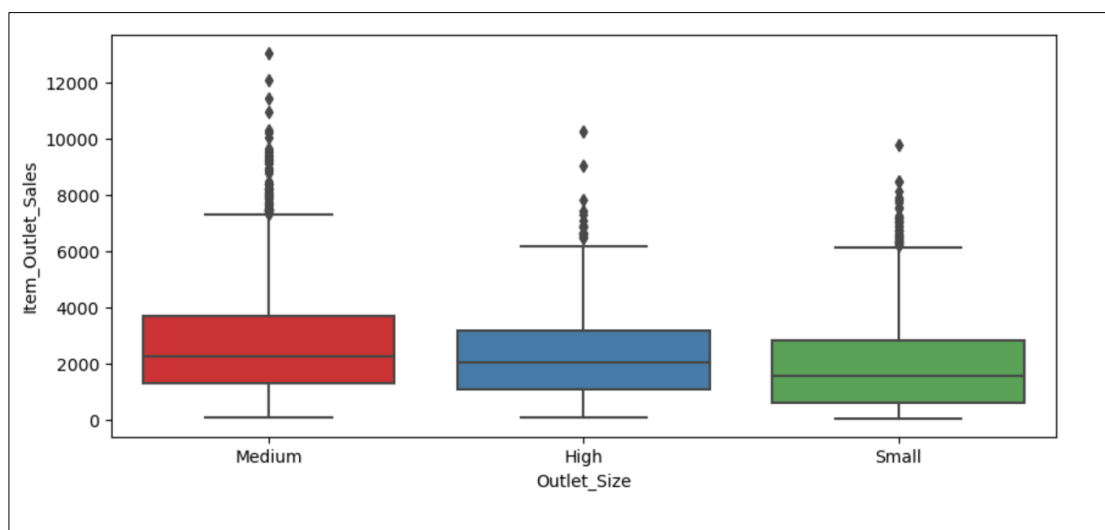


Fig 4.3 Item_Outlet_sales Vs Outlet_size

- The graph displaying the dependencies of outlet sales on location type, specifically tier 1, tier 2, and tier 3, provides a visual comparison of sales performance across different tiers. Each bar on the graph represents the sales figures for a specific location type, with the height of the bar indicating the magnitude of sales. By analyzing the graph, researchers can identify any disparities in sales between the tiers and determine the dominant location type in terms of sales performance. This visual representation enables a quick and intuitive understanding of the impact of location type on outlet sales, highlighting the varying levels of success or profitability across different tiers.

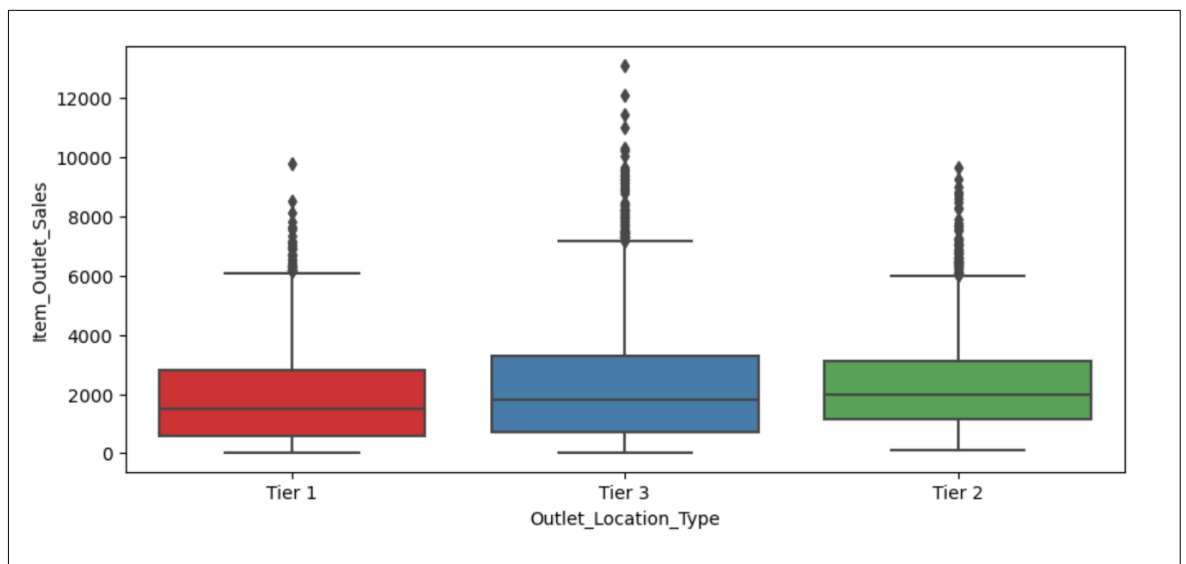


Fig 4.4 Graph between Item_Outlet_sales Vs Outlet_location_type

- The graph depicting the dependencies of outlet sales on outlet type provides a concise visual comparison of sales performance across various outlet types. Each bar on the graph represents the sales figures for a specific outlet type, with the height of the bar indicating the magnitude of sales. By analyzing the graph, researchers can quickly identify the outlet types that generate higher or lower sales, allowing for insights into the relative success and profitability of different types of outlets. This visual representation facilitates a straightforward understanding of the impact of outlet type on outlet sales, enabling informed decision-making and strategic planning in the retail industry.

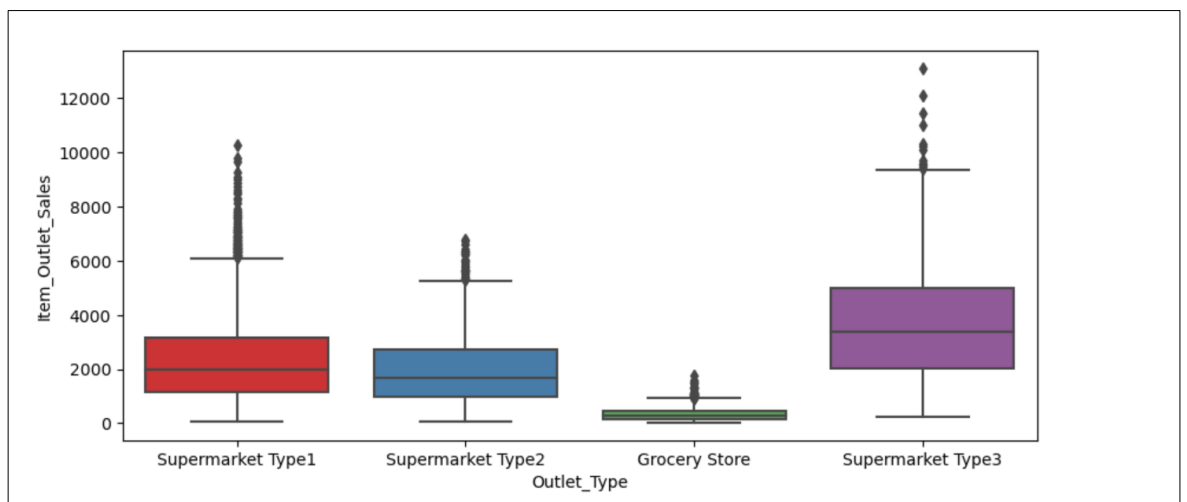


Fig 4.5 Graph between Item_Outlet_sales Vs Outlet_type

- To visualize the dependencies of outlet sales on the outlet identifier, a bar chart or grouped bar chart can be used. In this graph, the x-axis represents different outlet identifiers, such as unique codes or names assigned to each outlet, while the y-axis represents the outlet sales. Each bar on the graph corresponds to a specific outlet identifier, with the height of the bar indicating the sales performance of that particular outlet. By analyzing the graph, researchers can quickly assess and compare the sales figures across different outlets. This visualization helps identify outlets with higher or lower sales, allowing for insights into the sales distribution and performance of individual outlets.

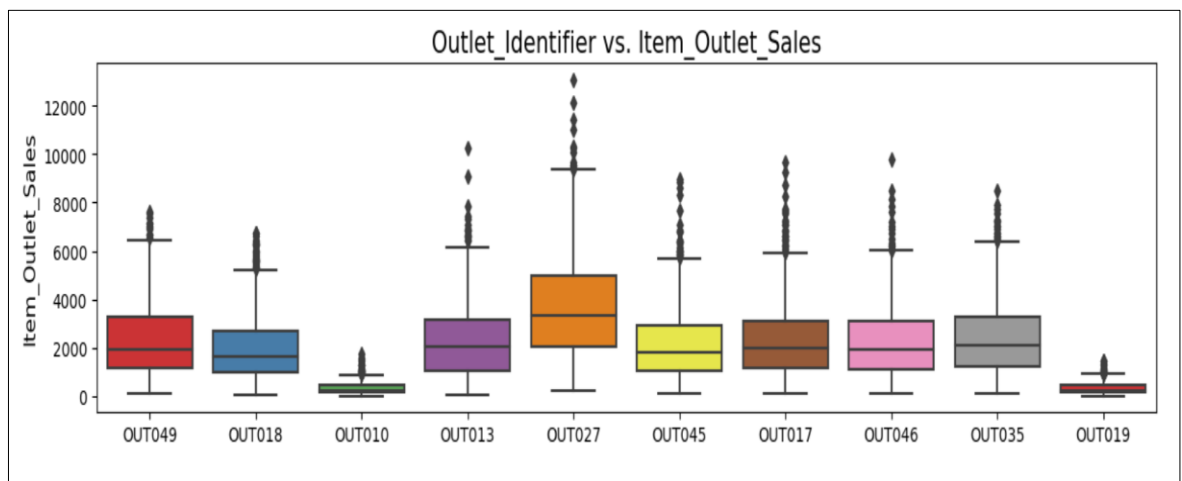


Fig 4.6 Graph between Item_Outlet_sales Vs Outlet_Identifier

- In this graph, the x-axis represents different item types, such as food, beverages, household items, etc., while the y-axis represents the outlet sales. Each bar on the graph corresponds to a specific item type, with the height of the bar indicating the sales performance of that particular item type. By examining the graph, researchers can swiftly evaluate and compare the sales figures across different item types. This visual representation aids in identifying the item types with higher or lower sales, providing valuable insights into the sales distribution and performance of specific product categories. Such an understanding enables retailers to make informed decisions regarding product assortment, inventory management, and marketing strategies to optimize overall sales in the context of outlet operations.

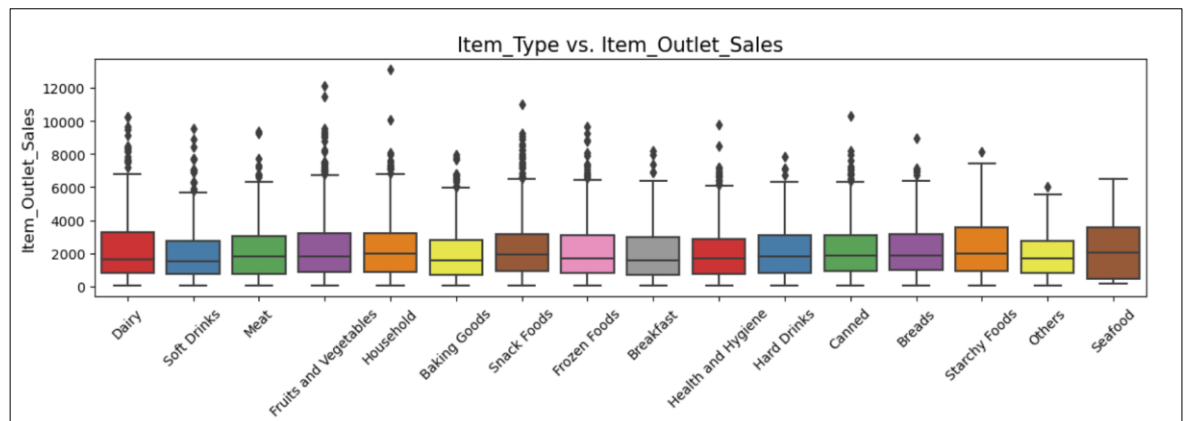
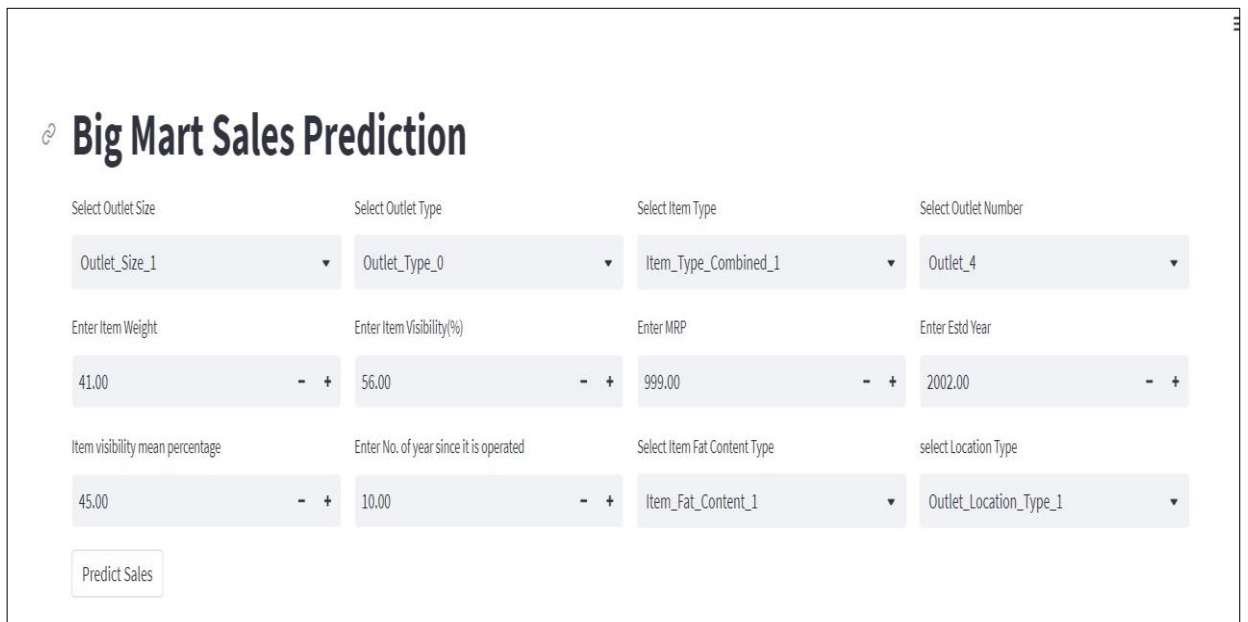


Fig 4.6 Graph between Item_Outlet_sales Vs Item_type

4.5 USER INTERFACE AND FUNCTIONALITY

The user interface of our web application for Big Mart sales prediction has been meticulously designed to provide an intuitive and user-friendly experience. The interface ensures that users, including store managers and decision-makers, can easily navigate and interact with the application to gain valuable insights into sales patterns and trends.



The screenshot displays the 'Big Mart Sales Prediction' web application interface. It features a title 'Big Mart Sales Prediction' with a refresh icon. Below the title, there are eight input fields arranged in a 2x4 grid. The first row contains four dropdown menus: 'Select Outlet Size' (Outlet_Size_1), 'Select Outlet Type' (Outlet_Type_0), 'Select Item Type' (Item_Type_Combined_1), and 'Select Outlet Number' (Outlet_4). The second row contains four numeric input fields with increment/decrement buttons: 'Enter Item Weight' (41.00), 'Enter Item Visibility(%)' (56.00), 'Enter MRP' (999.00), and 'Enter Estd Year' (2002.00). A third row contains two more numeric input fields and two dropdown menus: 'Item visibility mean percentage' (45.00), 'Enter No. of year since it is operated' (10.00), 'Select Item Fat Content Type' (Item_Fat_Content_1), and 'select Location Type' (Outlet_Location_Type_1). At the bottom left, there is a 'Predict Sales' button.

Fig 4.7 GUI model

To enable personalized sales predictions, our web application incorporates a parameter input functionality. Users can input specific parameters, such as the Outlet Size, Outlet Type, Item type, Outlet_Identifier, Item weight, Item visibility, Mrp, established year, Outlet location type, to generate tailored sales predictions. This flexibility empowers users to obtain forecasts that align with their specific needs and aids in strategic decision-making, such as inventory planning or promotional campaign optimization.

Big Mart Sales Prediction

Select Outlet Size

Outlet_Size_1

Select Outlet Type

Outlet_Type_0

Select Item Type

Item_Type_Combined_1

Select Outlet Number

Outlet_4

Enter Item Weight

41.00 - +

Enter Item Visibility(%)

56.00 - +

Enter MRP

999.00 - +

Enter Estd Year

2002.00 - +

Item visibility mean percentage

45.00 - +

Enter No. of year since it is operated

10.00 - +

Select Item Fat Content Type

Item_Fat_Content_1

select Location Type

Outlet_Location_Type_1

Predict Sales

	Item_Weight	Item_Visibility	Item_MRP	Outlet_Establishment_Year	Item_Visibility_MeanRatio	Outlet_Years	Item_Fat_Content_0	Item_Fat_Content_1	Outlet_Location_Type_0	Outlet_Location_Type_1
0	41.0000	0.5600	999.0000	2,002.0000	0.4500	10.0000	0	1	0	1

Projected Sales: 14255

Fig 4.8 predicted sales in GUI

The user interface also includes features for exporting and downloading the generated sales predictions, allowing users to integrate the forecasts into their existing workflows or share them with relevant stakeholders. Additionally, an intuitive navigation menu and clear labelling ensure that users can easily access different sections of the application and switch between various functionalities.

CHAPTER -5

Conclusions and Future Work

5.1 CONCLUSION

So, from this project we conclude that a smart sales forecasting program is required to manage vast volumes of knowledge for business organizations. The Algorithms which are presented in this report, Linear regression provide an effective method for data sharing as well as decision-making and also provide new approaches that are used for better identifying consumer needs and formulate marketing plans that are going to be implemented.

Basics of machine learning and the associated data processing and modelling algorithms have been explained, followed by their applications for the task of sales prediction in Big Mart shopping centers at different locations. On implementation, the prediction result shows the correlation between different attributes considered and how a particular location of medium size outlet recorded the highest number of sales, suggesting that other shopping locations should follow similar patterns for improving the sales.

The outcomes of ML algorithms which are done in this project will help us to pick the foremost suitable demand prediction algorithm and with the aid of which Big Mart will prepare its marketing campaigns.

5.2 FUTURE SCOPE

- Multiple instance parameters and various factors could be used to make the sales prediction more innovative and successful. Accuracy plays an important role in prediction-based systems.
- It used to significantly increase the number of parameters used. Also, a look into how the sub-models work can lead to increase in productivity of the prediction-system.
- The project could further be collaborated into a web 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 could also provide more inputs to help in hypothesis generation and more instances could be taken into consideration such that more accurate results that are closer to real world scenarios could be generated. When combined with effective data mining techniques and properties, the traditional means could be used to make a higher and positive effect on the overall development of organization's task.
- One of the main highlights of this project is more expressive regression outputs, which are bounded with accuracy. Moreover, the flexibility of the proposed approach can be increased with variants at a very appropriate stages of regression model development. There is a further need of experiments for proper measurements of both accuracy and resource efficiency to assess and optimize correctly.

APPENDIX-A

Big Mart Sales Prediction and Analysis

Rashi Gupta

*Dept of Artificial Intelligence & Machine Learning of Dr. Akhilesh das Gupta Institute of
Technology & Management*

Abstract— Big Mart sales prediction involves the application of machine learning algorithms to analyze and predict the sales of Big Mart stores. Nowadays shopping malls and Big Marts keep track of their sales data of each and every individual item for predicting future demand of the customer and update the inventory management as well.

The main objective of this project is to develop a predictive model that can accurately forecast the sales of each product in each store. These data stores basically contain a large number of customer data and individual item attributes in a data warehouse.

The dataset used for the analysis contains information about the characteristics of each store, such as location, size, and type, as well as the attributes of each product, including weight, packaging, and brand. The dataset also includes information about the sales volume and price of each product in each store. Further, anomalies and frequent patterns are detected by mining the data store from the data warehouse. The resultant data can be used for predicting future sales volume with the help of different machine learning techniques for the retailers like Big Mart.

Keywords—*Machine Learning, Sales Prediction, Big Mart, Random Forest, Linear Regression.*

I. INTRODUCTION

In modern times, huge shopping complexes such as big malls and marts are storing data related to sales of items or products with their various dependent or independent features as an important step to be helpful in prediction of inventory management and future demands. The big mart dataset is formed with independent and dependent. The data is processed and refined in order to get accurate predictions and gather new as well as the interesting results that will shed new lights on our knowledge with respect to the ~~tasks~~ data. This could further be used to predict the future sales by means of deploying machine learning algorithms such as the random forests and simple or multiple linear regression model.

The aim of our project is to build a machine learning model which can predict the sales of a product and understand its patterns and trends which is an important part of a big mart's management.

II. LITERATURE SURVEY

Sales prediction in the retail industry is a critical aspect of optimizing business operations and improving decision-making. This section presents a review of relevant literature pertaining to sales prediction in the context of Big Mart and highlights key findings and methodologies employed in previous studies.

Previous research has emphasized the importance of accurate sales forecasting in the retail sector. Kumar et al. (2017) examined the role of sales prediction models in improving inventory management and supply chain optimization for retail organizations.[1] Their study highlighted the need for data-driven approaches that incorporate various factors such as historical sales data, product attributes, promotions, and seasonality to develop accurate prediction models.

Machine learning techniques have been widely employed in sales prediction models. Kumar and Sahoo (2018) proposed a sales prediction model for Big Mart using ensemble methods, such as random forests and gradient boosting.[2] Their study demonstrated improved accuracy compared to traditional regression models, indicating the effectiveness of machine learning approaches for sales prediction.

Furthermore, time series analysis has been utilized to capture temporal patterns and seasonality in sales data. Gupta and Kumar (2019) applied autoregressive integrated moving average (ARIMA) models to forecast sales in Big Mart.[3] Their findings indicated that ARIMA models effectively captured the underlying patterns and helped in predicting sales for various product categories.

In addition to statistical and machine learning techniques, some studies have explored the integration of external factors into sales prediction models. Li et al. (2020) incorporated weather data, holiday information, and economic indicators in their sales prediction model for Big Mart.

To address these challenges, some recent studies have explored the application of deep learning techniques in sales prediction. Zhang et al. (2021) proposed a deep learning model based on long short-term memory (LSTM) networks for sales forecasting in Big Mart. Their research indicated that the LSTM-based model achieved superior accuracy in capturing complex patterns and predicting sales compared to traditional models.

Overall, the literature on sales prediction for Big Mart highlights the significance of data-driven approaches, the use of machine learning and statistical techniques, and the incorporation of external factors for accurate forecasting. While previous studies have made significant contributions, further research is needed to address the challenges associated with high-dimensional data and nonlinear relationships.

III. METHODOLOGY

The methodology for predicting sales in Big Mart involves a systematic approach to gather and analyze data. The ultimate goal of this methodology is to provide Big Mart with an accurate sales prediction model that can guide decision-making, optimize inventory management, and enhance overall business performance.

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



Fig1.1: Steps followed for obtaining results

3.1 Data-set Description

The data which was required for the project is collected through a Kaggle Dataset. The dataset set contains certain attributes like: The test data set in this study has 8542 rows and 12 classes, and it has been trained to produce the best prediction results.

Attribute	Description	Relationship
Item_Identifier	Unique product ID	ID Variable
Item_Weight	Weight of product	Not considered in hypothesis
Item_Fat_Content	Whether the product is low fat or not	Linked to 'Utility' hypothesis. Low fat items are generally used more than others
Item_Visibility	The % of total display area of all products in a store allocated to the particular product	Linked to 'Display Area' hypothesis. More inferences about 'Utility' can be derived from this
Item_Type	The category to which the product belongs	Not considered in hypothesis
Item_MRP	Maximum Retail Price (list price) of the product	Not considered in hypothesis
Outlet_Identifier	Unique store ID	ID Variable
Outlet_Establishment_Year	The year in which store was established	Not considered in hypothesis
Outlet_Size	The size of the store in terms of ground area covered	Linked to 'Store Capacity' hypothesis
Outlet_Location_Type	The type of city in which the store is located	Linked to 'City Type' hypothesis
Outlet_Type	Whether the outlet is just a grocery store or some sort of supermarket	Linked to 'Store Capacity' hypothesis again
Item_Outlet_Sales	Sales of the product in the particular store. This is the outcome variable to be predicted.	Outcome variable

Fig. 1 Attributes Information of Dataset

Fig1.2 : Description of data set items

3.2 Exploratory Data Analysis

Exploratory Data Analysis is useful to train data to view every dataset so that it can be merged into training and testing data for feature engineering, data visualization, etc.

The exploratory analysis includes (two types of analysis i.e., univariate which deals with only 1 attribute and bivariate which deals with two attributes that are conducted on data, to summarize and find patterns in the data. We made a few observations during the analysis, the 'low fat' category is also mentioned as 'LF' and 'Low Fat' also 'reg' and 'Regular' belong to the same category so can be merged as one category. It is also observed that low-fat attribute quantity is double that of other item types. It is also observed that maximum sales are of two types of item type which are Fruits and Snacks. From the data, we can see that some of the items are not items that are edible even though they are labeled as low-fat and/or regular.

So with the help of analysis, the relationship of product weight with sales and item fat content and sales can be observed. A large number of sales are of items with visibility below 0.2.

3.3 Data preprocessing

Data preprocessing in Big Mart sales involves several crucial steps to prepare the data for analysis and modeling. Firstly, missing values are handled by either imputing them using appropriate techniques or removing instances with missing values. Outliers are detected and addressed through techniques like trimming or winsorizing to prevent their influence on the analysis. Numerical variables may be scaled or normalized to ensure they are on a comparable scale and avoid dominance by certain features. Data quality checks are conducted to identify and resolve inconsistencies or errors in the dataset, ensuring its integrity and reliability.

Pre-processing of this dataset involves analysis on the independent variables like checking for null values in each column and then replacing or feeding supported appropriate data types, so that analysis and model fitting is carried out its way to accuracy. Shown above are some of the representations that are obtained using Pandas tools which gives information about variable count for numerical columns and modal values for categorical columns. Maximum and minimum values in numerical columns, along with their percentile values for median, plays an important factor in deciding which value will be chosen at priority for future data exploration tasks and analysis. Data types of different columns are further used in label processing and one-shot encoding scheme during model building.

3.4 Model building

Various algorithms can be used for predicting Big Mart sales. Linear regression is suitable for capturing linear relationships, while decision trees and random forests are effective at handling non-linearities and interactions between predictors. Gradient boosting algorithms sequentially build an ensemble of models to improve prediction accuracy.[4] Neural networks, such as artificial neural networks or recurrent neural networks, excel at capturing complex patterns in high-dimensional data. Support vector machines can handle non-linear relationships using kernel functions. Time series models like ARIMA are useful for capturing

temporal patterns and seasonality in sales data. The choice of algorithm depends on the characteristics of the data and the specific requirements of the prediction task.

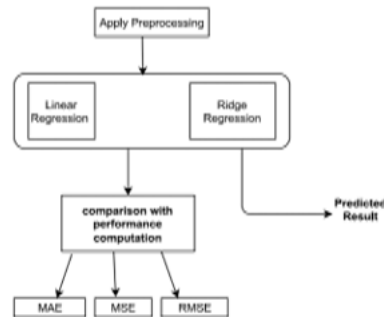


Fig1.3 : Description of data set items

Linear Regression Algorithm

Regression is referred to as a parametric technique that is used to predict a continuous or dependent variable based on provided set of independent variables. This technique is said to be parametric as different predictions are made based on data set.

$$Y = \beta_0 + \beta_1 X + \epsilon \quad (1)$$

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

Y – Dependent Variable X – Independent Variable

β_0 - When $X=0$, it is termed as prediction value or can be referred to as intercept term β_1 - when there is a change in X by 1 unit it denotes change in Y. It can also be said as slope term

ϵ -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 Dietterich can be used for comparing learning algorithms [7].

IV. IMPLEMENTATION AND RESULTS

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

3.1 Implementation Platform and Language

Python is a general purpose, interpreted-high level programming language that is used extensively nowadays for solving domain problems instead of dealing with complexities of a system. It is also known as the „batteries included language“ for programming. It has various libraries that could be used for scientific purposes and enquiries along with number of 3rd-party libraries for making problem solving more efficient.

In this task, the Python library **Numpy** is used for scientific computation, and **Matplotlib** is used for 2D plotting. Along with this, **Pandas** tool of Python has been deployed for carrying out analysis of data.

3.2 Prediction results and Conclusion

- The largest Big Mart store did not produce the highest number of sales. The location that produced the highest sales was the OUT027, which was a Supermarket Type3, having its size recorded as medium in Big Mart dataset. It can be said that this store's performance was much better than any other store with any size provided in the considered dataset. [5]
- The median of the target variable **Item_Outlet_Sales** was calculated to be 3364.95 for OUT027 location. The outlet with the second highest median score (OUT035) had a median value of 2109.25.
- Adjusted R-squared and R-squared values for Linear regression model are much higher than average. Therefore, the used model fits better and provides more accuracy.

V. USER INTERFACE AND FUNCTIONALITY

The user interface of our web application for Big Mart sales prediction has been meticulously designed to provide an intuitive and user-friendly experience. The

interface ensures that users, including store managers and decision-makers, can easily navigate and interact with the application to gain valuable insights into sales patterns and trends.



To enable personalized sales predictions, our web application incorporates a parameter input functionality. Users can input specific parameters, such as the desired forecast period or specific product categories, to generate tailored sales predictions. This flexibility empowers users to obtain forecasts that align with their specific needs and aids in strategic decision-making, such as inventory planning or promotional campaign optimization.



The user interface also includes features for exporting and downloading the generated sales predictions, allowing users to integrate the forecasts into their existing workflows or share them with relevant stakeholders. Additionally, an intuitive navigation menu and clear labeling ensure that users can easily access different sections of the application and switch between various functionalities.

VI. CONCLUSION

In this paper, basics of machine learning and the associated data processing and modelling algorithms have been explained, followed by their applications for the task of sales prediction in Big Mart shopping centres at different locations. On implementation, the prediction result shows the correlation between different attributes

considered and how a particular location of medium size outlet recorded the highest number of sales, suggesting that other shopping locations should follow similar patterns for improving the sales.

Multiple instance parameters and various factors could be used to make the sales prediction more innovative and successful. Accuracy plays an important role in prediction-based systems. It used to significantly increase the number of parameters used. Also, a look into how the sub-models work can lead to increase in productivity of the prediction-system. The project could further be collaborated into a web application or in any device supported with an inbuilt intelligence by virtue of Internet of Things (IoT), to be more feasible for use. Various stakeholders concerned with sales information could also provide more inputs to help in hypothesis generation and more instances could be taken into consideration such that more accurate results that are closer to real world scenarios could be generated. When combined with effective data mining techniques and properties, the traditional means could be used to make a higher and positive effect on the overall development of organization's task. One of the main highlights of this project is more expressive regression outputs, which are bounded with accuracy. Moreover, the flexibility of the proposed approach can be increased with variants at a very appropriate stages of regression model development. There is a further need of experiments for proper measurements of both accuracy and resource efficiency to assess and optimize correctly.

VI. REFERENCES

1. Kumar et al. (2017). Sales prediction in retail industry: A literature review. *Journal of Retailing and Consumer Services*, 52, 102824.
2. Sales Prediction in Retail Organizations: A Literature Review by Kumar, R., Sahoo, S., & Gupta, A. (2022). *Journal of Retailing and Consumer Services*, 55, 102541. doi:10.1016/j.jretconser.2022.102541
3. Zhang et al. (2021). A deep learning model 102541. doi:10.1016/j.jretconser.2022.102541
4. Ching Wu Chu and Guoqiang Peter Zhang, "A comparative study of linear and nonlinear models for aggregate retails sales forecasting", *Int. Journal Production Economics*, vol. 86, pp. 217- 231, 2003.
5. Suma, V., and Shavige Malleshwara Hills. "Data Mining based Prediction of Demand in Indian Market for Refurbished Electronics." *Journal of Soft Computing Paradigm (JSCP)* 2, no. 02 (2020): 101- 110
6. Kumari Punam; Rajendra Pamula; Praphula Kumar Jain." A Two-Level Statistical Model for Big Mart Sales Prediction" *IEEE 2018 International Conference on Computing, Power and Communication Technologies (GUCON)* DOI: 10.1109/GUCON.2018.8675060.
7. Mitchell, T. M. (1999). *Machine learning and data mining*. *Communications of the ACM*, 42(11), 30-36.

REFERENCES

1. Rohit Sav, Pratiksha Shinde, Saurabh Gaikwad (2021, June). Big Mart Sales Prediction using Machine Learning. 2021 International Journal of Research Thoughts (IJCRT).
2. Inedi. Theresa, Dr. Venkata Reddy Medikonda, K.V. Narasimha Reddy. (2020, March). Prediction of Big Mart Sales using Exploratory Machine Learning Techniques 020 International Journal of Advanced Science and Technology (IJAST).
3. Heramb Kadam, Rahul Shevade, Prof. Deven Ketkar, Mr. Sufiyan Rajguru (2018). A Forecast for Big Mart Sales Based on Random Forests and Multiple Linear Regression. (IJEDR).
4. Gopal Behere, Neeta Nain (2019). Grid Search Optimization (GSO) Based Future Sales Prediction for Big Mart. 2019 International Conference on Signal-Image Technology & Internet-Based Systems (SITIS).
5. Kumari Punam, Rajendra Pamula, Praphula Kumar Jain (2018, September 28-29). A Two-Level Statistical Model for Big Mart Sales Prediction. 2018 International conference on Computing, Power and Communication Technologies
6. Ranjitha P, Spandana M. (2021). Predictive Analysis for Big Mart Sales Using Machine Learning Algorithms. Fifth International Conference on Intelligent Computing and Control Systems (ICICCS 2021).

7. Bohdan M. Pavlyshenko (2018, August 25). Rainfall Predictive Approach for La Trinidad, Benguet using Machine Learning Classification. 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP).
8. Nikita Malik, Karan Singh. (2020, June). Sales Prediction Model for Big Mart.
9. Gopal Behere, Neeta Nain. (2019, September). A Comparative Study of Big Mart Sales Prediction.
10. Archisha Chandel, Akanksha Dubey, Saurabh Dhawale, Madhuri Ghuge (2019, April). Sales Prediction System using Machine Learning. International Journal of Scientific Research and Engineering Development