Super Market Sales Prediction

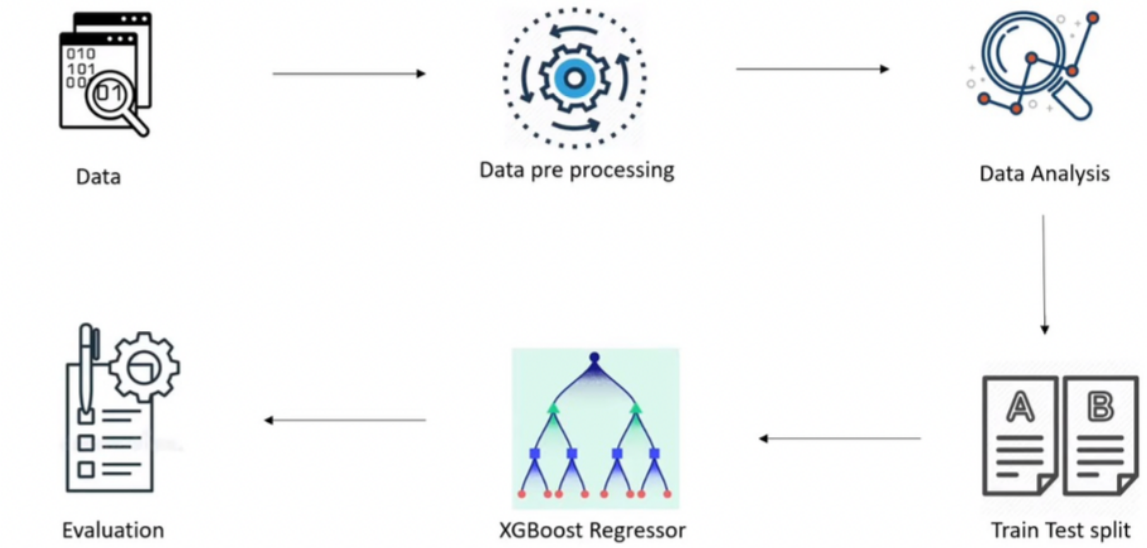
MILESTONE 3

Tushar. (E20CSE113), Advait Kumar (E20CSE117), Vaibhav Singh (E20CSE112)

Bennett University, Greater Noida – 203206

Abstract :

In this day and age of Automation and Extreme Sales Prediction, almost every supermarket and e-market keeps track of its sales data of every purchase, be it for inventory management or for predicting future trends in consumer behaviour. They usually maintain an all-encompassing database of customer data, individual item attributes like MRP, Weight, Manufacturing Date, Size, etc. Predictive Patterns and anomalies can be detected and be used for predicting future sales, with the help of different Machine Learning Techniques. In this Project, we propose a sales predictive model using Xgboost technique. A comparative analysis of the used Model with other performance Metrics will also be included.



I. Introduction :

Step by step contest among various shopping centres shopping edifices, as well as a few shopping centres and large match getting more genuine and forceful because of the fast development of the worldwide disconnected and internet shopping. Each shopping store is attempting to give customized Short time offers to drawing in more clients relying on the day, to such an extent that the volume of deals for everything can be anticipated for stock and deals the executives of the specific shop or legates say shopping complex, we are resolving the issue of large match deals expectation of gauging of a thing on clients' future interest in various huge Mart stores across different area in our nations and item founded on the past record. Different AI Oracle M like straight relapse investigations, arbitrary woods and so on are utilized for expectation of gauging of deals volume. As great deals at the existence of each association so the gauging of deals assumes a significant And extremely fundamental part of each association.

II. Block Diagram –

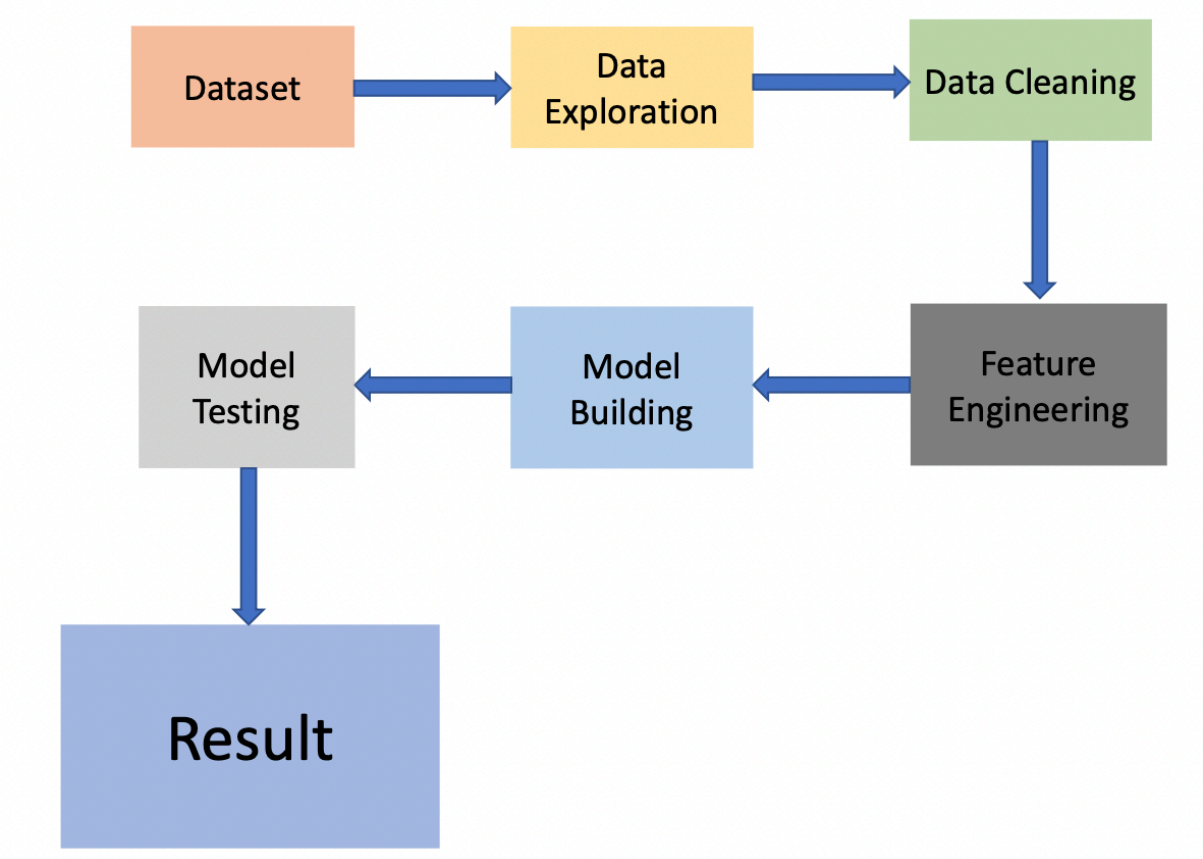


Fig. 1 Working of Procedure proposed model

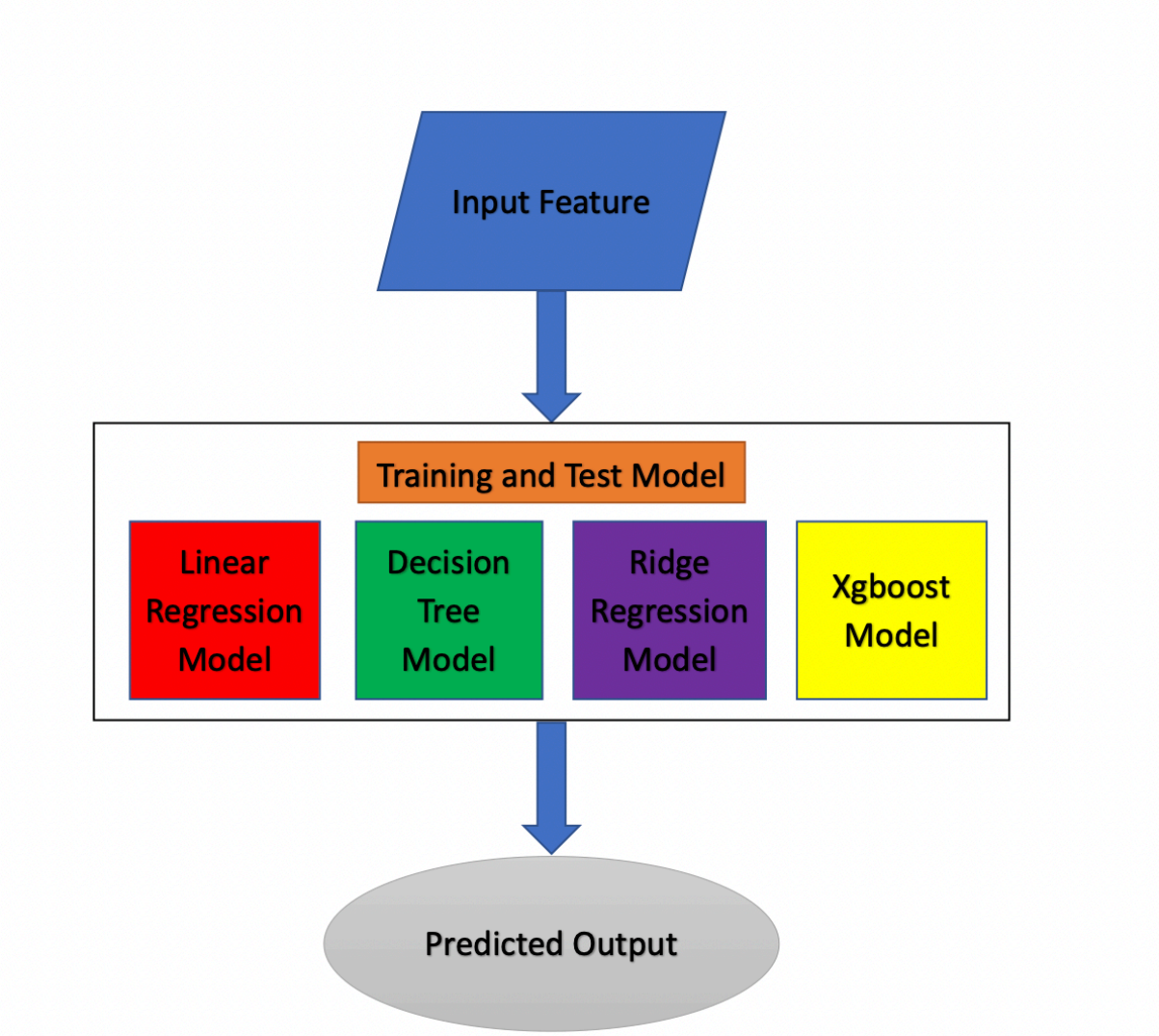


Fig. 2 Framework of proposed system

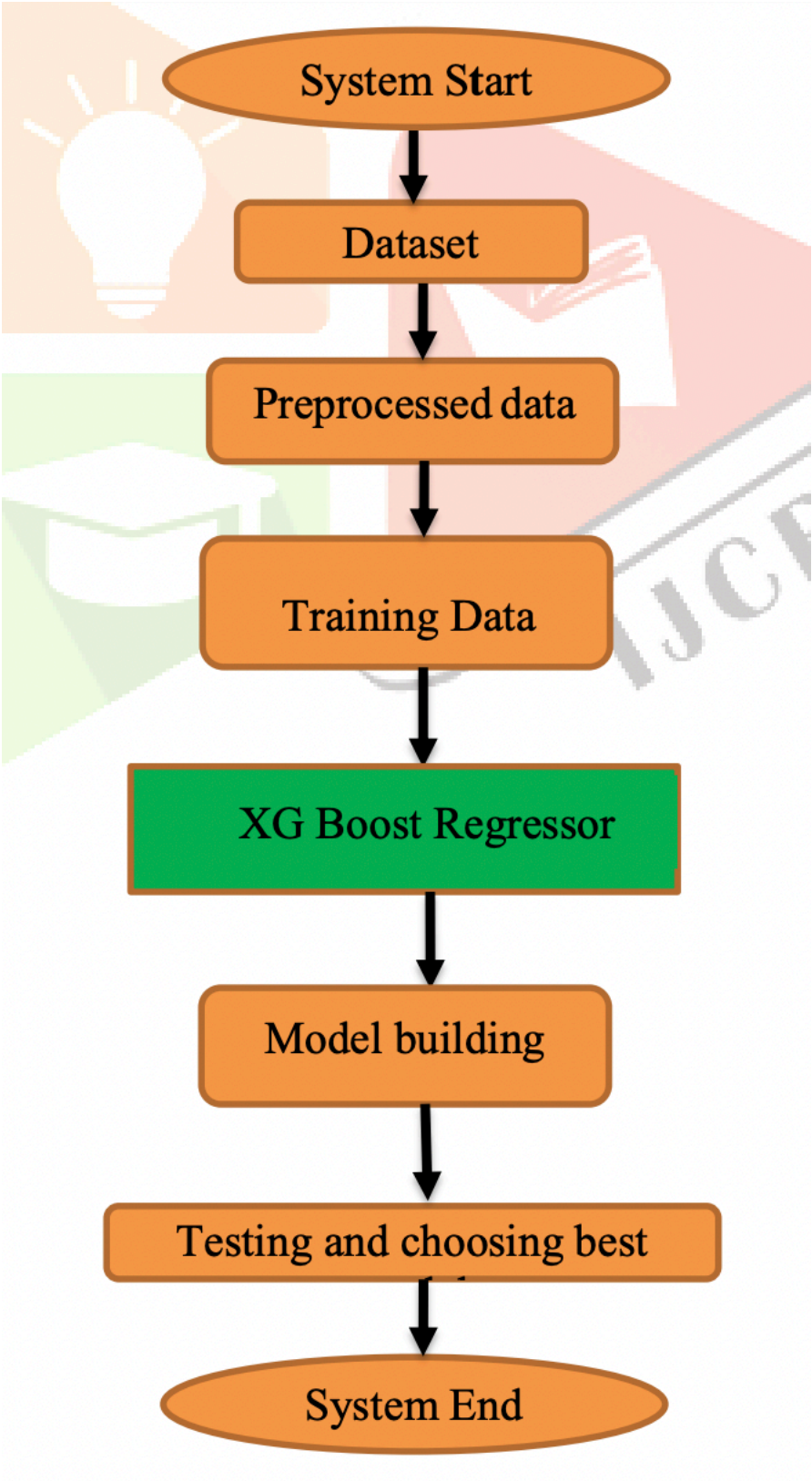


Fig. 3 Flow of XGBoost Regressor

III. Proposed Methodology -

The steps followed during this work, right from the dataset preparation to obtaining results are represented in Fig. 4

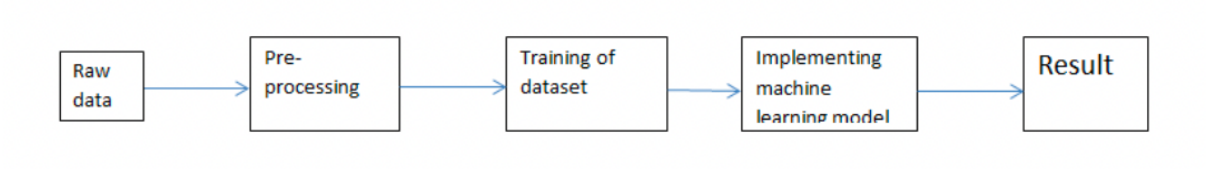


Fig. 4 Steps followed for obtaining results

3.1 Dataset and its Pre-processing

Super Market's data specialists assembled bargains data of their 10 stores organized at different regions with each store having 1559 novel things as indicated by 2013 data collection. Using all of the insights it is accumulated which work explicit properties of a thing play and what they mean for their arrangements. The dataset seems to be shown in Fig. 5 on using head() work on the dataset variable.

```
[3] # first 5 rows of the dataframe
    super_market_data.head()
```

	Item_Identifier	Item_Weight	Item_Fat_Content	Item_Visibility	Item_Type	Item_MRP	Outlet_Identifier	Outlet_Establishment_Year	Outlet_Size	Outlet_Loca
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	

```
# first 5 rows of the dataframe
super_market_data.head()
```

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

Fig. 5 Screenshot of Dataset

The data set consists of various data types from integer to float to object as shown in Fig. 6

```
[5] #tells datatype of each column
    super_market_data.dtypes

    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. 6 Various datatypes used in Dataset

In the raw data, there can be different sorts of fundamental examples which likewise gives an inside and out information about subject of interest and gives experiences about the issue. Yet, wariness ought to be seen regarding information as it might contain invalid qualities, or repetitive qualities, or different kinds of equivocalness, which likewise requests for pre-handling of information. Dataset ought to accordingly be investigated however much as could reasonably be expected.

Different elements significant by measurable means like mean, standard deviation, middle, count of values and greatest worth and so forth are displayed in Fig.4 for mathematical factors of our dataset.

```
[17] super_market_data.describe()
```

	Item_Weight	Item_Visibility	Item_MRP	Outlet_Establishment_Year	Item_Outlet_Sales
count	8523.000000	8523.000000	8523.000000	8523.000000	8523.000000
mean	12.857645	0.066132	140.992782	1997.831867	2181.288914
std	4.226124	0.051598	62.275067	8.371760	1706.499616
min	4.555000	0.000000	31.290000	1985.000000	33.290000
25%	9.310000	0.026989	93.826500	1987.000000	834.247400
50%	12.857645	0.053931	143.012800	1999.000000	1794.331000
75%	16.000000	0.094585	185.643700	2004.000000	3101.296400
max	21.350000	0.328391	266.888400	2009.000000	13086.964800

Fig. 7 Numerical variables of the Dataset

Pre-processing of this dataset incorporates doing investigation on the free factors like checking for invalid qualities in every segment and afterward supplanting or filling them with upheld proper information types, so examination and model fitting isn't upset from its way to exactness. Displayed above are a portion of the portrayals acquired by utilizing Pandas instruments which tells about factor count for mathematical sections and modular qualities for all out segments. Greatest and least qualities in mathematical segments, alongside their percentile values for middle, plays a significant element in concluding which worth to be picked at need for additional investigation undertakings and examination. Information kinds of various segments are utilized further in name handling and one-hot encoding plan during model structure.

3.2 Algorithms Applied

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 *sklearn.ensemble* library's random forest regressor class is used. The key role is played by the parameter termed as *n_estimators* which also comes under random forest regressor. Random forest can be referred to as a meta-estimator used to fit upon numerous decision trees (based on classification) by taking the dataset's different sub-samples. *min_samples_split* is taken as the minimum number when splitting an internal node if integer number of minimum samples are considered. A split's 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.

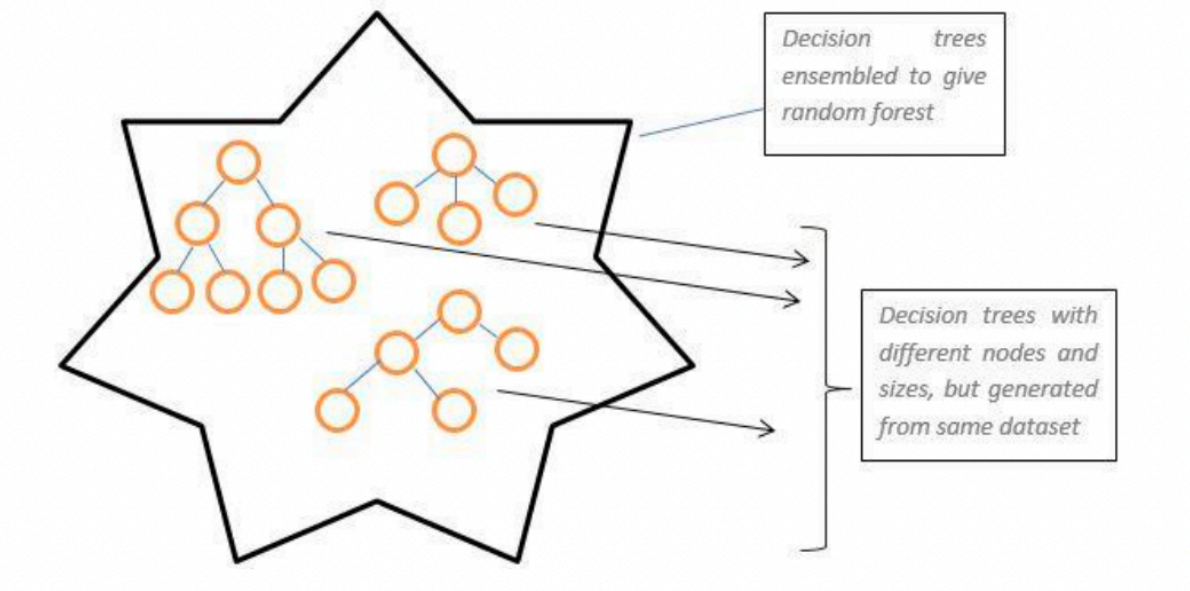


Fig. 8 Relation between Decision Tree and Random Forest

- 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.

$$Y = \beta o + \beta 1X + \epsilon \quad (1)$$

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 Dietterich can be used for comparing learning algorithms [10].

3.3 Metrics of Data Modelling

- The coefficient of assurance R^2 (R-squared) is a measurement that actions the decency of a model's fit for example how well the genuine information focuses are approximated by the forecasts of relapse. Higher upsides of R2 recommend higher model achievements as far as expectation alongside precision, and the worth 1 of R2 is demonstrative of relapse forecasts completely fitting the genuine elements. For additional improved outcomes, the utilization of changed R2 estimates does something amazing. Taking logarithmic upsides of the objective section in the dataset ends up being critical in the expectation interaction. Along these lines, one might say that on taking changes of segments utilized in expectation, improved outcomes can be reasoned. One approach to fusing change could likewise have included taking square base of the section. It additionally gives better perception of the dataset and target variable as the square base of target variable is leaned to be an ordinary dispersion.

- The error measurement is a crucial metric within the estimation period. Root mean squared error (RMSE) and Mean Absolute Error (MAE) are generally used for continuous variable's accuracy measurement. It are often said that the typical model prediction error are often expressed in units of the variable of interest by using both MAE and RMSE. MAE is that the average over the test sample of absolutely the differences between prediction and actual observation where all individual differences have equal weight. The root of the typical of squared differences between prediction and actual observation are often termed as RMSE. RMSE is an absolute measure of fit, whereas R2 may be a relative measure of fit. RMSE helps in measuring the variable's average error and it's also a quadratic scoring rule. Low RMSE values obtained for linear or multiple correlation corresponds to raised model fitting.

  With reference to the results obtained during this work, it are often said that there's no big difference between our train and test sample since the metric RMSE ratio is calculated to be adequate to the ratio between train and test sample. The results associated with how accurately responses are predicted by our model are often inferred from RMSE because it may be a good measure along side measuring precision and other required capabilities. a substantial improvement might be made by further data exploration incorporated with outlier detection and high leverage points. Another approach, which is conceptually easier, is to mix several sub-models which are low dimensional and simply verifiable by domain experts, i.e., ensemble learning are often exploited.

IV. Experimental Setup –

4.1 Dataset Description and Feature Extraction

Large Shop's information researchers gathered deals information of their 10 stores arranged at various areas with each store having 1559 unique items according to 2013 information assortment. Utilizing all the perceptions it is surmised which job specific properties of a thing play and how they influence their deals. The dataset looks like displayed in Fig.2 on utilizing head() work on the dataset variable. In the crude information, there can be different kinds of fundamental examples which likewise gives a top to bottom information about subject of interest and gives bits of knowledge about the issue. However, watchfulness ought to be seen regarding information as it might contain invalid qualities, or excess qualities, or different kinds of vagueness, which additionally requests for pre-handling of information. Dataset ought to in this manner be investigated however much as could be expected. Pre-processing of this dataset incorporates doing investigation on the autonomous factors like checking for invalid qualities in every segment and afterward supplanting or filling them with upheld suitable information types, so examination and model fitting isn't blocked from its way to precision. Displayed above are a portion of the portrayals acquired by utilizing Pandas devices which tells about factor count for mathematical segments and modular qualities for clear cut sections. Greatest and least qualities in mathematical sections, alongside their percentile values for middle, plays a significant factor in concluding which worth to be picked at need for additional investigation undertakings and examination. Information sorts of various sections are utilized further in mark handling and one-hot encoding plan during model structure.

4.2 Pre-processing

A few subtleties were seen in the informational index during information investigation stage. So this stage is utilized in settling all subtleties found from the dataset and prepare them for building the suitable model. During this stage it was seen that the Thing perceivability characteristic had a zero worth, essentially which has barely any clue. So the mean worth thing perceivability of that item will be utilized for zero qualities trait. This makes all items liable to sell. All downright characteristics errors are settled by changing all out credits into suitable ones. At long last, for deciding how old a specific outlet is, we add an extra quality Year to the dataset.
Pre-processing of this dataset incorporates doing examination on the free factors like checking for invalid qualities in every section and afterward supplanting or filling them with upheld proper information types, so investigation and model fitting isn't blocked from its way to precision. Displayed above are a portion of the portrayals acquired by utilizing Pandas devices which tells about factor count for mathematical sections and modular qualities for downright segments. Most extreme and least qualities in mathematical segments, alongside their percentile values for middle, plays a significant factor in concluding which worth to be picked at need for additional investigation assignments and examination. Information kinds of various sections are utilized further in mark handling and one-hot encoding plan during model structure.

V. Results –

- The largest location did not produce the highest sales. The location that produced the highest sales was the OUT027 location, which was in turn a Supermarket Type3, having its size recorded as medium in our dataset. It can be said that this outlet's performance was much better than any other outlet location with any size provided in the considered dataset.
- The median of the target variable Item_Outlet_Sales was calculated to be 3364.95 for OUT027 location. The location with second highest median score (OUT035) hada median value of 2109.25.
- Adjusted R-squared and R-squared values are higher for Linear regression model than average. Therefore, the used model fits better and exhibits accuracy.
- Also, model accuracy and score of regression model can reach nearly 61% if built with more hypothesis consideration and analysis.

It can be concluded that more locations should be switched or shifted to Supermarket Type3 to increase the sales of products at Super Market Any one-stop-shopping-center like Big Mart can benefit from this model by being able to predict its items' future sales at different locations.

VI. References –

1. Markakis, S., Wheelwright, S.C., Hyndman, R.J.: Forecasting methods and applications. John Wiley & sons (2008).
2. Kadam, H., Shevade, R., Ketkar, P. and Rajguru.: "A Forecast for Big Mart Sales Based on Random Forests and Multiple Linear Regression." (2018).
3. C. M. Wu, P. Patil and S. Gunaseelan: Comparison of Different Machine Learning Algorithms for Multiple Regression on Black Friday Sales Data (2018).
4. K. Punam, R. Pamula and P. K. Jain, quote; A Two-Level Statistical Model for Big Mart Sales Prediction, quote; 2018 International Conference on Computing, Power and Communication Technologies (GUCON), Greater Noida, Uttar Pradesh, India, 2018, pp. 617-620.
5. S. Yadav and S. Shukla, quote; Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification, quote; 2016 IEEE 6th International Conference on Advanced Computing (IACC)
6. V. Shrivastava and P. Arya, quote; A study of various clustering algorithms on retail sales data quote;, International Journal of Computing, Communications and Networking, vol. 1, no. 2, pp. 1-7, 2012.