IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁑ Impact Factor 7.39 ⁑ Vol. 11, Issue 5, May 2022

DOI: 10.17148/IJARCCE.2022.115106

Big Mart Sales Prediction Using Machine Learning

Nimit Jain

Student, Department of Information Technology, Maharaja Agrasen Institute of Technology(GGSIPU), NewDelhi

Abstract: Machine Learning is a category of algorithms that allows software applications to become more accurate in predicting outcomes without being explicitly programmed. The basic premise of machine learning is to build models and employ algorithms that can receive input data and use statistical analysis to predict an output while updating outputs as new data becomes available. These models can be applied in different areas and trained to match the expectations of management so that accurate steps can be taken to achieve the organization's target. In this paper, the case of Big Mart, a one-stop-shopping center, has been discussed to predict the sales of different types of items and for understanding the effects of different factors on the items' sales. Taking various aspects of a dataset collected for Big Mart, and the methodology followed for building a predictive model, results with high levels of accuracy are generated, and these observations can be employed to make decisions to improve sales.

I. INTRODUCTION

Big Mart is a big supermarket chain, with stores all around the country and its current board set out a challenge to all Data Scientists out there to help them create a model that can predict the sales, per product, for each store to give accurate results.

Big Mart has collected sales data from the year 2013, for 1559 products across 10 stores in different cities.

With this information, the corporation hopes we can identify the products and stores which play a key role in their sales anduse that information to take the correct measures to ensure the success of their business.

II. LITERATURE SURVEY

1. **Title: - A Forecast for Big Mart Sales Based on Random Forests and Multiple Linear Regression(2018)**
Author: - Kadam, H., Shevade, R., Ketkar, P. and Rajguru
Description: - A Forecast for Big Mart Sales Based on Random Forests and Multiple Linear Regression used Random Forest and Linear Regression for prediction analysis which gives less accuracy. To overcome this we can use XG boost Algorithm which will give more accuracy and will be more efficient.

2. **Title: - Forecasting methods and applications (2008) Author: - Makridakis, S., Wheelwrigh.S.C., Hyndman. R.J**
Description: - Forecasting methods and applications contains a Lack of Data and short life cycles. So some of the data like historical data, consumer-oriented markets face uncertain demands, can be prediction for accurate results.

3. **Title: -Comparison of Different Machine Learning Algorithms for Multiple Regression onBlack Friday Sales Data (2018)**
Author: - C. M. Wu, P. Patil and S. Gunaseelan
Description: - Comparison of Different Machine Learning Algorithms for Multiple Regression on Black Friday Sales Data Used Neural Network for comparison of different algorithms. To overcome this Complex models like neural networks is used for comparison between different algorithms which is not efficient so we can use simpler algorithmfor prediction.

III. PROBLEM STATEMENT

"To find out what role certain properties of an item play and how they affect their sales by understanding Big Mart sales. In order to help BigMart achieve this goal, a predictive model can be built to find out for every store, the key factors thatcanincrease their sales and what changes could be made to the product or store‟s characteristic

DOI: 10.17148/IJARCCE.2022.115106

IV. METHODOLOGY

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

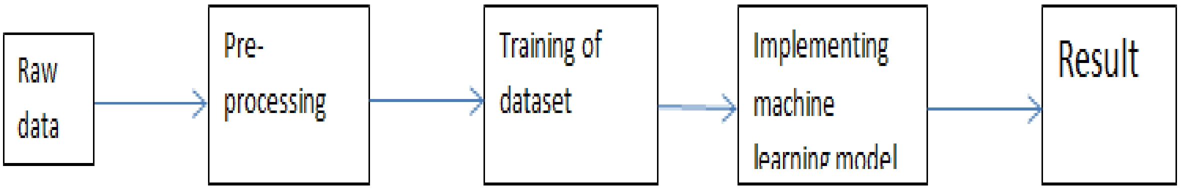
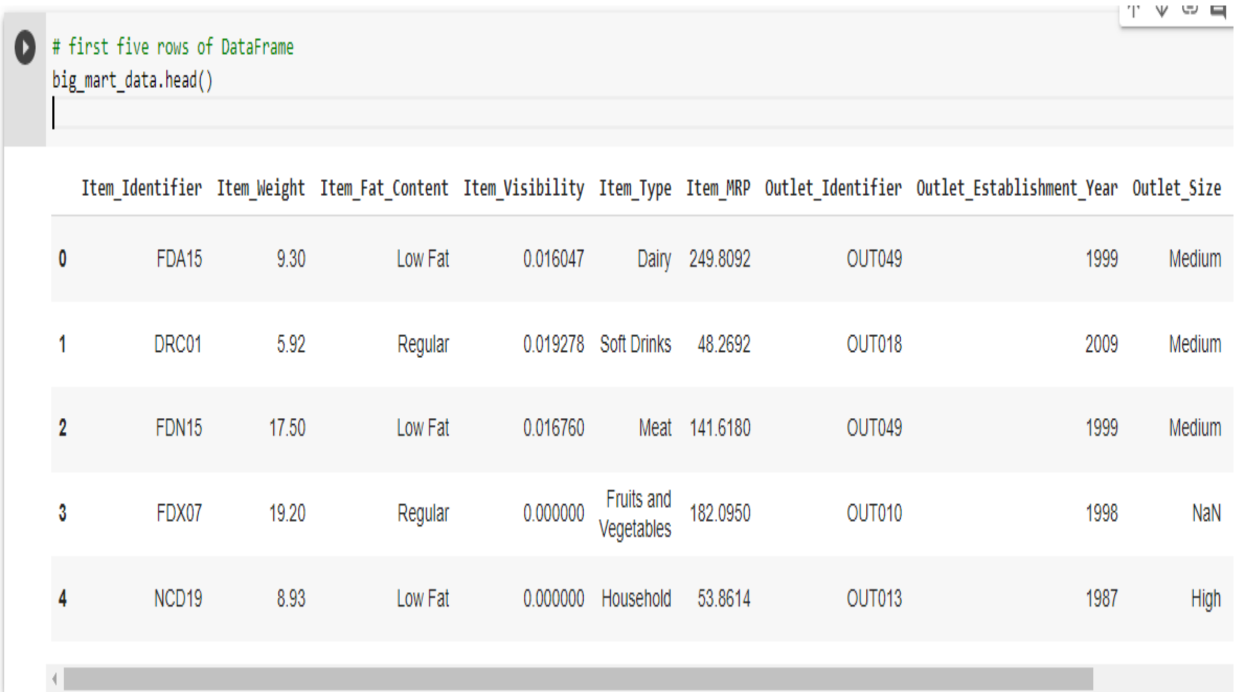


Fig. 1. Steps followed for obtaining results

V. DATASET AND ITS PREPROCESSING

BigMart‟s data scientists collected sales data of their 10 stores situated at different locations with each store having 1559different products as per 2013 data collection.Using all the observations it is inferred what role certain properties of an item play and how they affect their sales.The dataset looks like shown in Fig.2 on using head() function on the datasetvariable.



```
# first five rows of DataFrame
big_mart_data.head()
```

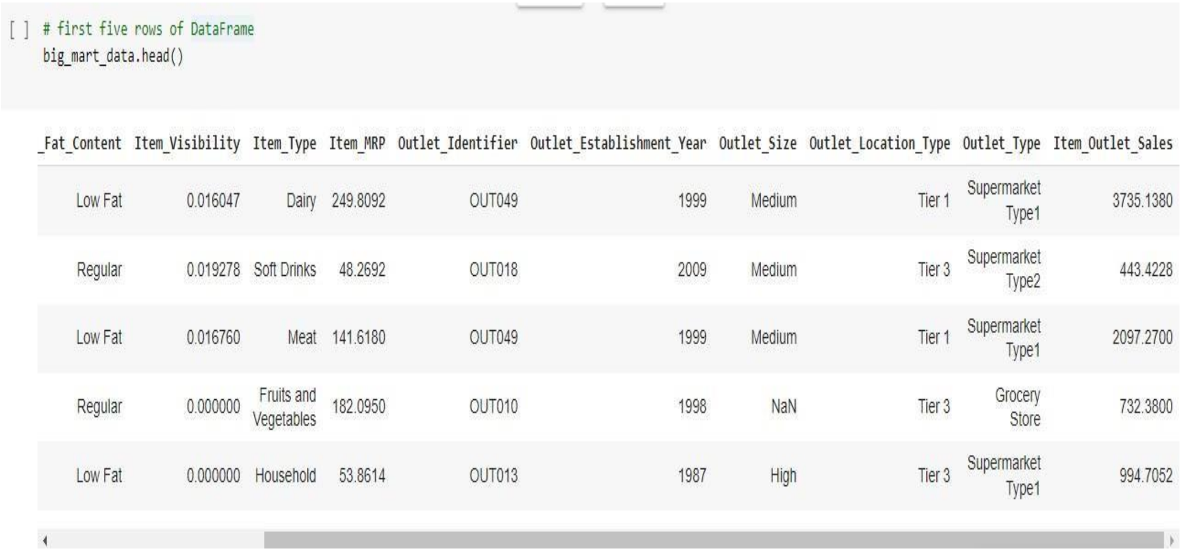
	Item_Identifier	Item_Weight	Item_Fat_Content	Item_Visibility	Item_Type	Item_MRP	Outlet_Identifier	Outlet_Establishment_Year	Outlet_Size
0	FDA15	9.30	Low Fat	0.016047	Dairy	249.8092	OUT049	1999	Medium
1	DRC01	5.92	Regular	0.019278	Soft Drinks	48.2692	OUT018	2009	Medium
2	FDN15	17.50	Low Fat	0.016760	Meat	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

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⋞ Impact Factor 7.39 ⋞ Vol. 11, Issue 5, May 2022

DOI: 10.17148/IJARCCE.2022.115106

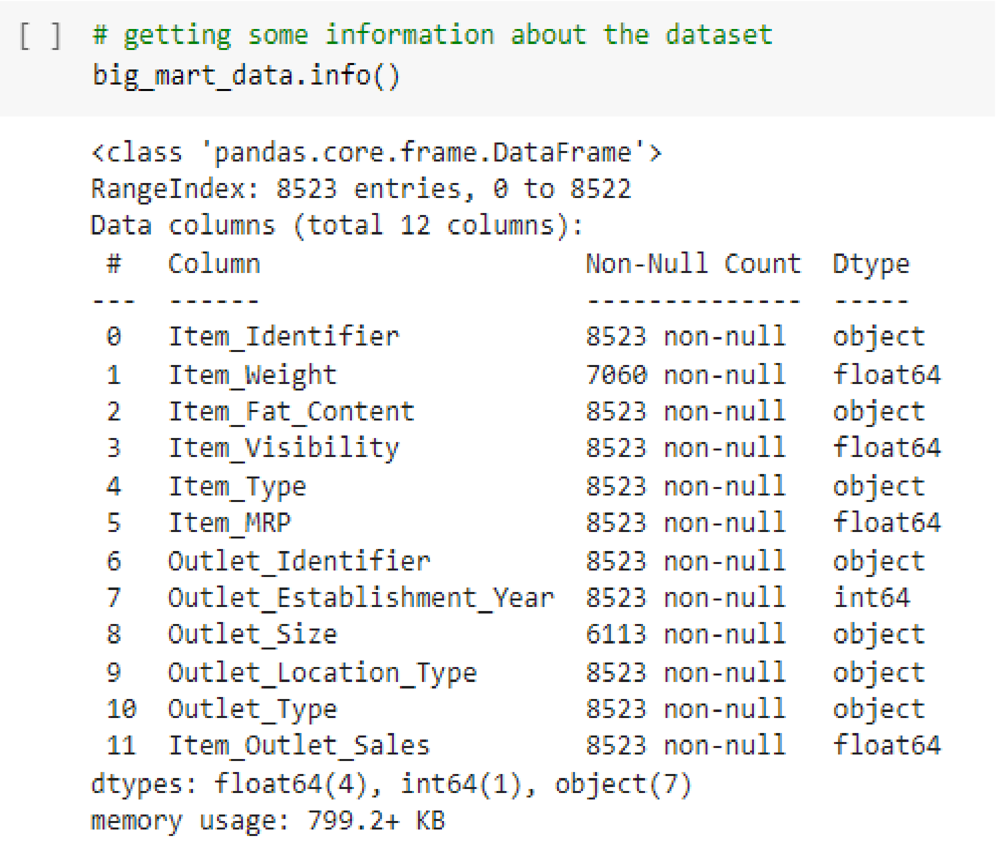


```
[ ] # first five rows of DataFrame
    big_mart_data.head()
```

_Fat_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. 2. Screenshot of Dataset

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

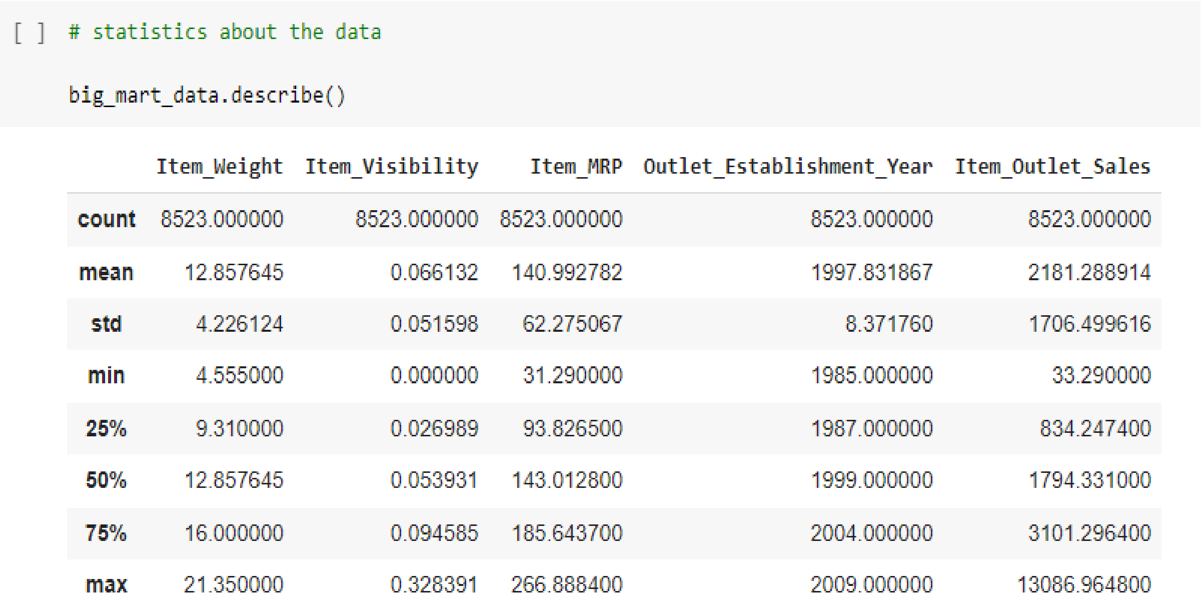


```
[ ] # getting some information about the dataset
    big_mart_data.info()

    <class 'pandas.core.frame.DataFrame'>
    RangeIndex: 8523 entries, 0 to 8522
    Data columns (total 12 columns):
     #   Column                     Non-Null Count  Dtype
    ---  ------                     --------------  -----
     0   Item_Identifier            8523 non-null   object
     1   Item_Weight                7060 non-null   float64
     2   Item_Fat_Content           8523 non-null   object
     3   Item_Visibility            8523 non-null   float64
     4   Item_Type                  8523 non-null   object
     5   Item_MRP                   8523 non-null   float64
     6   Outlet_Identifier          8523 non-null   object
     7   Outlet_Establishment_Year  8523 non-null   int64
     8   Outlet_Size                6113 non-null   object
     9   Outlet_Location_Type       8523 non-null   object
     10  Outlet_Type                8523 non-null   object
     11  Item_Outlet_Sales          8523 non-null   float64
    dtypes: float64(4), int64(1), object(7)
    memory usage: 799.2+ KB
```

Fig. 3. Various data types used in the Dataset

In the raw data, there can be various types of underlying patterns which also gives an in-depth knowledge about the subject of interest and provides insights into the problem. But caution should be observed concerning data as it may contain null values, or redundant values, or various types of ambiguity, which also demands pre-processing of data. The dataset should therefore be explored as much as possible. Various factors important by statistical means like mean, standard deviation,median, count of values and maximum value, etc. are shown in Fig.4 for numerical variables of ourdataset.



```
[ ]  # statistics about the data

     big_mart_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. 4. Numerical variables of the Dataset

Preprocessing of this data set includes analyzing the independent variables like checking for null values in each column and then replacing or filling them with supported appropriate data types so that analysis and model fitting is not hindered from their way to accuracy. Shown above are some of the representations obtained by using Pandas tools which tell 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 the median, play an important factor in deciding which value to be chosen at priority for further exploration tasks and analysis. Data types of different columns are used further in label processing during model building.

VI. ALGORITHM EMPLOYED

□ **Linear Regression**

Linear Regression falls under the category of Supervised Learning. It is a linear model that shows a linear relationship between a dependent (Y) and independent (X) variables i.e predicting output (Y)on input values (X). When there is a single input variable (x), then it is termed simple linear regression. Ex:- Predicting Price of the house by analyzing Area of house. And, when there are multiple input variables, it is termed as multiple linear regression. Ex:- Predicting the Price of the house by analyzing different features like Area, Locality, No. of rooms, etc.

□ **Random Forest**

The RF (Breiman 2001) is a supervised classification technique based on classification trees (Breiman et al. 1984) and the algorithm consists in learning a set of weak learners, in the case of Random Forest, weak learners are decision trees[7]. A generalized model has low bias and low variance but in the case of Decision Tree, it provides an overfitted model i.e a model with high variance, so to tackle this problem, Random Forest Regression is used.

□ **LASSO Regression**

LASSO regression aims to identify the variables and corresponding regression coefficients that leadto a model that minimizes the prediction error. This is achieved by imposing a constraint on the model parameters, which ‘shrinks’ the regression coefficients towards zero, that is by forcing the sum of the absolute value of the regression coefficients to be less than a fixed value (λ).

1. Uses L1 Regularization.
2. Penalty Term:- constant * magnitude of the slope.
3. It shrinks the less important features, so it can be used in feature engineering when we have alargenumber of features.

□ **XGBOOST**

Extreme Gradient Boosting (XGBoost) is an open-source library that provides an efficient and effective

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 5, May 2022

DOI: 10.17148/IJARCCE.2022.115106

implementation of the gradient boosting algorithm. Shortly after its development and initial release, XGBoost became the go-to method and often the key component in winning solutions for a range of problems in machine learning competitions. Regression predictive modeling problems involve predicting a numerical value such as a dollar amount or a height. **XGBoost** can be used directly for **regression predictive modeling**.

VII. IMPLEMENTATION AND RESULTS

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

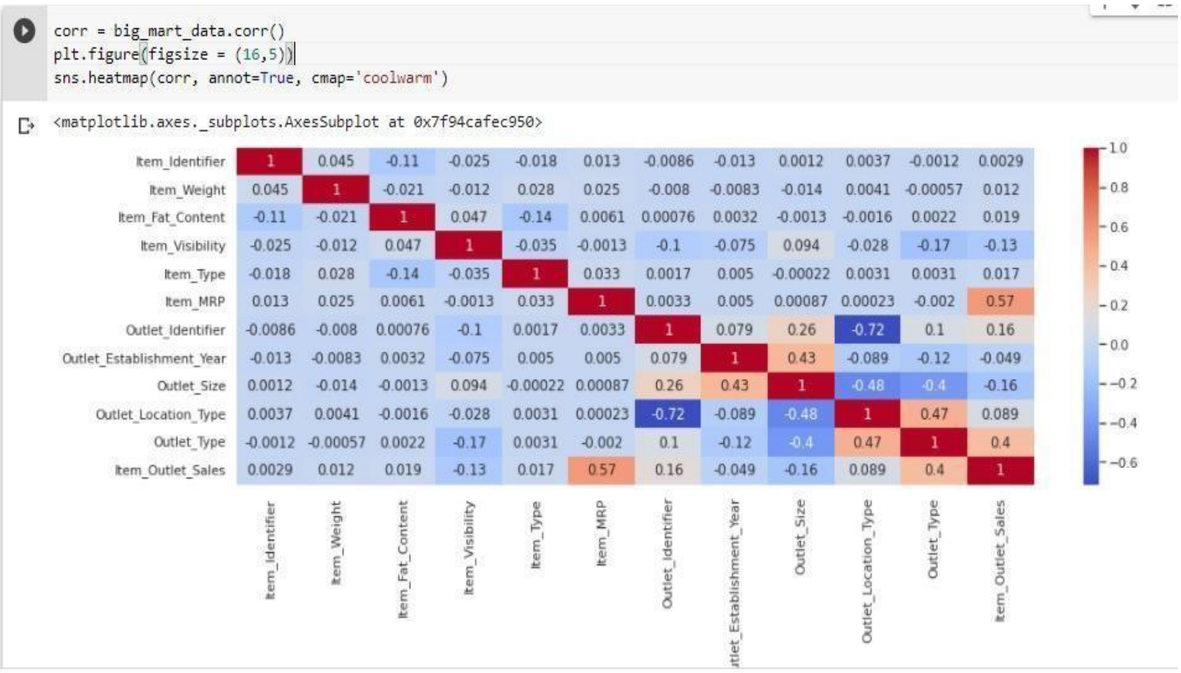
Implementation Platform and Language

Python is a general-purpose, interpreted-high-level language used extensively nowadays for solving domain problems instead of dealing with the complexities of a system. It is also termed as the „batteries included language" for programming. It has various libraries used for scientific purposes and inquiries along with several third-party libraries for making problem-solving efficient. In this work, the Python libraries of Num py, for scientific computation, and Matplotlib, for 2D plotting have been used. Along with this, Pandas tool of Python has been employed for carrying out data analysis.

XGBoost regressor is used for prediction. As a development platform, Jupyter Notebook, which proves to work great due toits excellence in „literate programming", where human-friendly code is punctuated within code blocks, has been used.

Data Modeling and Observations

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



```
corr = big_mart_data.corr()
plt.figure(figsize = (16,5))
sns.heatmap(corr, annot=True, cmap='coolwarm')
```

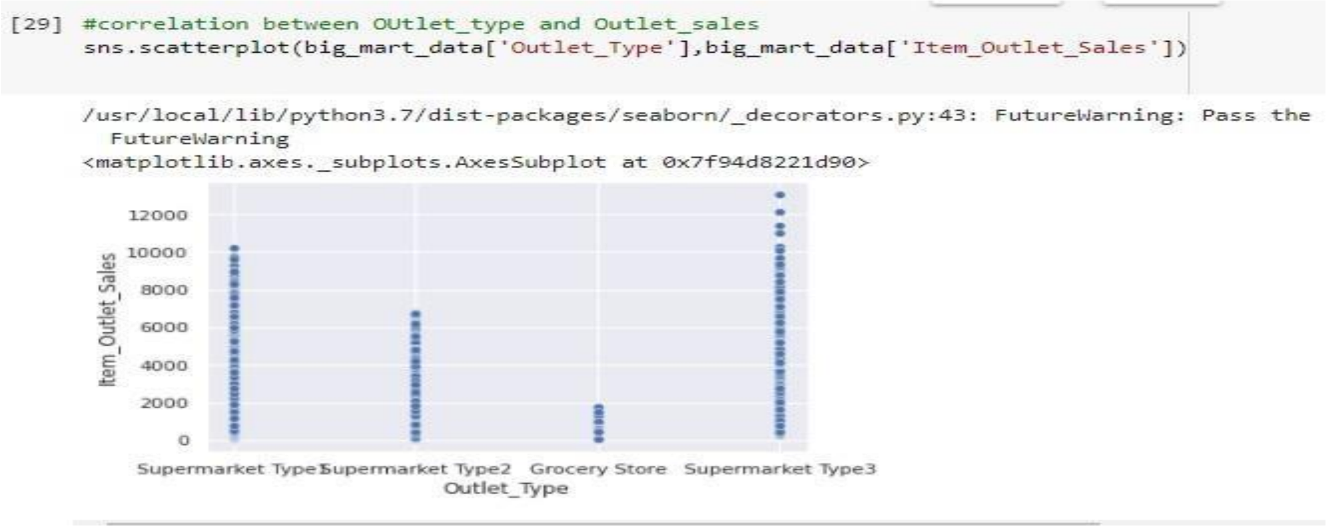
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f94cafec950>
```

	Item_Identifier	Item_Weight	Item_Fat_Content	Item_Visibility	Item_Type	Item_MRP	Outlet_Identifier	tlet_Establishment_Year	Outlet_Size	Outlet_Location_Type	Outlet_Type	Item_Outlet_Sales
Item_Identifier	1	0.045	-0.11	-0.025	-0.018	0.013	-0.0086	-0.013	0.0012	0.0037	-0.0012	0.0029
Item_Weight	0.045	1	-0.021	-0.012	0.028	0.025	-0.008	-0.0083	-0.014	0.0041	-0.00057	0.012
Item_Fat_Content	-0.11	-0.021	1	0.047	-0.14	0.0061	0.00076	0.0032	-0.0013	-0.0016	0.0022	0.019
Item_Visibility	-0.025	-0.012	0.047	1	-0.035	-0.0013	-0.1	-0.075	0.094	-0.028	-0.17	-0.13
Item_Type	-0.018	0.028	-0.14	-0.035	1	0.033	0.0017	0.005	-0.00022	0.0031	0.0031	0.017
Item_MRP	0.013	0.025	0.0061	-0.0013	0.033	1	0.0033	0.005	0.00087	0.00023	-0.002	0.57
Outlet_Identifier	-0.0086	-0.008	0.00076	-0.1	0.0017	0.0033	1	0.079	0.26	-0.72	0.1	0.16
Outlet_Establishment_Year	-0.013	-0.0083	0.0032	-0.075	0.005	0.005	0.079	1	0.43	-0.089	-0.12	-0.049
Outlet_Size	0.0012	-0.014	-0.0013	0.094	-0.00022	0.00087	0.26	0.43	1	-0.48	-0.4	-0.16
Outlet_Location_Type	0.0037	0.0041	-0.0016	-0.028	0.0031	0.00023	-0.72	-0.089	-0.48	1	0.47	0.089
Outlet_Type	-0.0012	-0.00057	0.0022	-0.17	0.0031	-0.002	0.1	-0.12	-0.4	0.47	1	0.4
Item_Outlet_Sales	0.0029	0.012	0.019	-0.13	0.017	0.57	0.16	-0.049	-0.16	0.089	0.4	1

Fig. 5. Diagram showing correlation among different factors

From Fig.5, the correlation among various dependent and independent variables is explored to be able to decide onthe further steps that are to be taken. Variables used are obtained after data pre-processing, and followingare some of the important observations about some of the used variables:

DOI: 10.17148/IJARCCE.2022.115106



```
[29] #correlation between OUtlet_type and Outlet_sales
     sns.scatterplot(big_mart_data['Outlet_Type'],big_mart_data['Item_Outlet_Sales'])
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning: Pass the
  FutureWarning
<matplotlib.axes._subplots.AxesSubplot at 0x7f94d8221d90>
```

Fig. 6. Correlation between target variable and Item-visibility variable

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

VIII. RESULT AND CONCLUSION

Today, every shopping mall wants to know what their customers want so they can earn profit. In order to getan idea about customers' needs, they can keep an eye on their past sales, as sales is directly proportional to profit. As the profit made by a company is directly proportional to the accurate predictions of sales, the Big marts are desiring more accurate prediction algorithm so that the company will not suffer any losses. In this research work, we have designed a predictive model by modifying Gradient boosting machines as Xgboost technique and experimented it on the 2013 Big Mart dataset for predicting sales of the product from a particular outlet. Experiments support that our technique produce more accurate prediction compared to than other available techniques like decision trees, ridge regression etc**.**

Table 5.1: Final Result

S.no	**Name of Model**	**MAE**	**R^2**
1.	LASSO	895.14	0.484
2.	Linear Regression	886.90	0.4965
3.	Random Forest	785.12	0.5455
4.	Random ForestHyper Tuning	783.37	0.5597
5.	XGBOOST	742.8	0.5977

Multiple instances parameters and various factors can be used to make this sales prediction more innovative and successful. Accuracy, which plays a key role in prediction-based systems, can be significantly increased as the number of parameters used are increased. Also, a look into how the sub-models work can lead to increase in productivity of system.

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

DOI: 10.17148/IJARCCE.2022.115106

correctly.

REFERENCES

[1] Mahesh, Batta. "Machine Learning Algorithms-A Review." International Journal of Science and Research (IJSR).[Internet] 9 (2020): 381-386.

[2] 1. Makridakis, S., Wheelwright, S.C., Hyndman, R.J.: Forecasting methods and applications. John wiley & sons (2008)

[3] Meenakshi, Meenakshi, Machine Learning Algorithms and their Real-life Applications: A Survey (May 7, 2020). Proceedings of the International Conference on Innovative Computing & Communications (ICICC) 2020, Available at SSRN: https://ssrn.com/abstract=3595299 or http://dx.doi.org/10.2139/ssrn.3595299

[4] https://ijcrt.org/papers/IJCRT2106802.pdf

[5] Kadam, H., Shevade, R., Ketkar, P. and Rajguru.: "A Forecast for Big Mart Sales Based on Random Forests and Multiple Linear Regression." (2018).

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

[7] Mauricio A Valle and Gonzalo A Ruz. "Turnover prediction in a call center: Behavioral evidenceof loss aversion using random forest and naıve bayes algorithms". In: Applied Artificial Intelli- gence 29.9 (2015), pp. 923–942.

[8] Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y. and Cho, H., 2015. Xgboost: extreme gradient boosting. R package version 0.4-2, 1(4), pp.1-4.

[9] T. Alexander and D. Christopher, quot;An Ensemble Based Predictive Modeling in Forecasting Sales of Big Martquot;, International Journal of Scientific Research, vol. 5, no. 5, pp. 1- 4, 2016.