SALES PREDICTION MODEL FOR BIG MART

Nikita Malik[1]*, Karan Singh[2]
[1]Assistant Professor, MSI
[2]Student, MSI
Janakpuri, New Delhi
[1]*nikitamalik@msi-ggsip.org, 9971633991
[2]karan01921202017@msi-ggsip.org

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

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

I. Introduction

In today's modern world, huge shopping centers such as big malls and marts are recording data related to sales of items or products with their various dependent or independent factors as an important step to be helpful in prediction of future demands and inventory management. The dataset built with various dependent and independent variables is a composite form of item attributes, data gathered by means of customer, and also data related to inventory management in a data warehouse. The data is thereafter refined in order to get accurate predictions and gather new as well as interesting results that shed a new light on our knowledge with respect to the task's data. This can then further be used for forecasting future sales by means of employing machine learning algorithms such as the random forests and simple or multiple linear regression model.

1.1 Machine Learning

The data available is increasing day by day and such a huge amount of unprocessed data is needed to be analysed precisely, as it can give very informative and finely pure gradient results as per current standard requirements. It is not wrong to say as with the evolution of Artificial Intelligence (AI) over the past two decades, Machine Learning (ML) is also on a fast pace for its evolution. ML is an important mainstay of IT sector and with that, a rather central, albeit usually hidden, part of our life [1]. As the technology progresses, the analysis and understanding of data to give good results will also increase as the data is very useful in current aspects. In machine learning, one deals with both supervised and unsupervised types of tasks and generally a classification type problem accounts as a resource for knowledge discovery. It generates resources and employs regression to make precise predictions about future, the main emphasis being laid on making a system self-efficient, to be able to do computations and analysis to generate much accurate and precise results [2]. By using statistic and probabilistic tools, data can be converted into knowledge. The statistical inferencing uses sampling distributions as a conceptual key [11].

ML can appear in many guises. In this paper, firstly, various applications of ML and the types of data they deal with are discussed. Next, the problem statement addressed through this work is stated in a formalized way. This is followed by explaining the methodology ensued and the prediction results observed on implementation. Various machine learning algorithms include [3]:

- Linear Regression: It can be termed as a parametric technique which is used to predict a continuous or dependent variable on basis of a provided set of independent variables. This technique is said to be parametric as different assumptions are made on basis of data set.
- K-Nearest Neighbors (KNN): It is a learning algorithm which is based on instances and knowledge gained through them [4]. Unlike mining in data stream scenarios, cases where every sample can simultaneously belong to multiple classes in hierarchical multi-label classification problems, k-NN is being proposed to be applied to predict outputs in structured form [5].
- Decision tree: It is an intuitive model having low bias and it can be adopted to build a classification tree with root node being the first to be taken into account in a top-down manner. It is a classic model for machine learning [6].
- Naïve Bayes classifiers: These are based on Bayes theorem and a collection of classification algorithms where classification of every pair is independent of each other. Bayesian learning can provide predictions with readable reasons by generating an if-then form of list of rules [8].
- Random Tree: It is an efficient algorithm for achieving scalability and is used in identification problems for building approximate system. The decisions are taken considering the choices made on basis of possible consequences, the variables which are included, input factor. Other algorithms can include SVM, xgboost, logistic regression and so on [7].
- K-means clustering: This algorithm is used in unsupervised learning for creating clusters of related data based on their closeness to the centroid value [9].

1.2 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 Big Mart achieve this goal, a predictive model can be built to find out for every store, the key factors that can increase their sales and what changes could be made to the product or store's characteristics.

II. Methodology

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

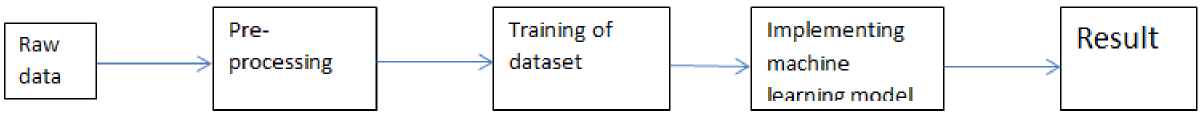

Fig1: Steps followed for obtaining results

2.1 Dataset and its Preprocessing

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

```
In [7]: df.head()
        #understanding rows and column
```

Out[7]:

	Item_Identifier	Item_Weight	Item_Fat_Content	Item_Visibility	Item_Type	Item_MRP	Outlet_Identifier	Outlet_Establishment_Year	Outlet_Size	Outlet_Locatic
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	

```
In [7]: df.head()
        #understanding rows and column
```

Out[7]:

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

Fig2: Screenshot of Dataset

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

```
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
```

Fig3: Various datatypes used in the Dataset

In the raw data, there can be various types of underlying patterns which also gives an in-depth knowledge about subject of interest and provides insights about the problem. But caution should be observed with respect to data as it may contain null values, or redundant values, or various types of ambiguity, which also demands for pre-processing of data. Dataset should therefore be explored as much as possible.

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

```
In [10]: df.describe()
Out[10]:
```

	Item_Weight	Item_Visibility	Item_MRP	Outlet_Establishment_Year	Item_Outlet_Sales
count	7060.000000	8523.000000	8523.000000	8523.000000	8523.000000
mean	12.857645	0.066132	140.992782	1997.831867	2181.288914
std	4.643456	0.051598	62.275067	8.371760	1706.499616
min	4.555000	0.000000	31.290000	1985.000000	33.290000
25%	8.773750	0.026989	93.826500	1987.000000	834.247400
50%	12.600000	0.053931	143.012800	1999.000000	1794.331000
75%	16.850000	0.094585	185.643700	2004.000000	3101.296400
max	21.350000	0.328391	266.888400	2009.000000	13086.964800

Fig4: Numerical variables of the Dataset

Preprocessing of this dataset includes doing analysis on the independent variables like checking for null values in each column and then replacing or filling them with supported appropriate data types, so that analysis and model fitting is not hindered from its way to accuracy. Shown above are some of the representations obtained by using Pandas tools which tells about variable count for numerical columns and 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 to be chosen at priority for further exploration tasks and analysis. Data types of different columns are used further in label processing and one-hot encoding scheme during model building.

2.2 Algorithms employed

Scikit-Learn can be used to track machine-learning system on wholesome basis [12]. Algorithms employed for predicting sales for this dataset are discussed as follows:

- Random Forest Algorithm

Random forest algorithm is a very accurate algorithm to be used for predicting sales. It is easy to use and understand for the purpose of predicting results of machine learning tasks. In sales prediction, random forest classifier is used because it has decision tree like hyperparameters. The tree model is same as decision tool. Fig.5 shows the relation between decision trees and random forest. To solve regression tasks of prediction by virtue of random forest, the *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.

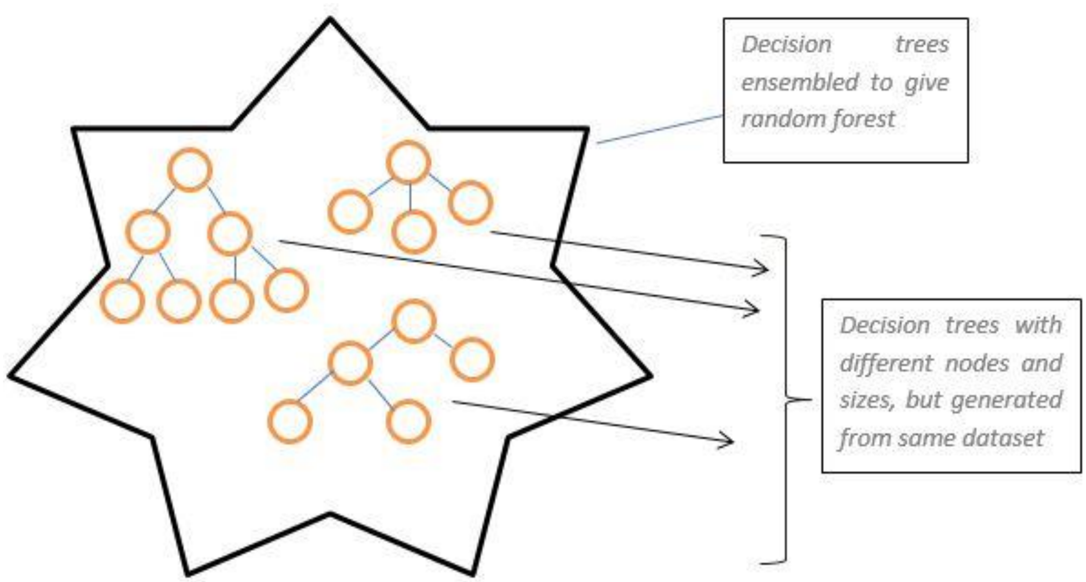

Fig5: Relation between Decision Trees and Random Forest

- Linear Regression Algorithm

Regression can be termed as a parametric technique which is used to predict a continuous or dependent variable on basis of a provided set of independent variables. This technique is said to be parametric as different assumptions are made on basis of data set.

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

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

Y - Variable to be predicted
X - Variable(s) used for making a prediction
βo - When X=0, it is termed as prediction value or can be referred to as intercept term
β1 - when there is a change in X by 1 unit it denotes change in Y. It can also be said as slope term
∈ -The difference between the predicted and actual values is represented by this parameter and also represents the residual value. However efficiently the model is trained, tested and validated, there is always a difference between actual and predicted values which is irreducible error thus we cannot rely completely on the predicted results by the learning algorithm. Alternative methods given by Dietterich can be used for comparing learning algorithms [10].

2.3 Metrics for Data Modelling

• The coefficient of determination R^2 (R-squared) is a statistic that measures the goodness of a model's fit i.e. how well the real data points are approximated by the predictions of regression. Higher values of R^2 suggest higher model accomplishments in terms of prediction along with accuracy, and the value 1 of R^2 is indicative of regression predictions perfectly fitting the real data points. For further better results, the use of adjusted R^2 measures works wonders. Taking logarithmic values of the target column in the dataset proves to be significant in the prediction process. So, it can be said that on taking adjustments of columns used in prediction, better results can be deduced. One way of incorporating adjustment could also have included taking square root of the column. It also provides better visualization of the dataset and target variable as the square root of target variable is inclined to be a normal distribution.

• The error measurement is an important metric in the estimation period. Root mean squared error (RMSE) and Mean Absolute Error (MAE) are generally used for continuous variable's accuracy measurement. It can be said that the average model prediction error can be expressed in units of the variable of interest by using both MAE and RMSE. MAE is the average over the test sample of the absolute differences between prediction and actual observation where all individual differences have equal weight. The square root of the average of squared differences between prediction and actual observation can be termed as RMSE. RMSE is an absolute measure of fit, whereas R^2 is a relative measure of fit. RMSE helps in measuring the variable's average error and it is also a quadratic scoring rule. Low RMSE values obtained for linear or multiple regression corresponds to better model fitting.

With respect to the results obtained in this work, it can be said that there is no big difference between our train and test sample since the metric RMSE ratio is calculated to be equal to the ratio between train and test sample. The results related to how accurately responses are predicted by our model can be inferred from RMSE as it is a good measure along with measuring precision and other required capabilities. A considerable improvement could be made by further data exploration incorporated with outlier detection and high leverage points. Another approach, which is conceptually easier, is to combine several sub-models which are low dimensional and easily verifiable by domain experts, i.e., ensemble learning can be exploited [9].

III. 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 language used extensively nowadays for solving domain problems instead of dealing with complexities of a system. It is also termed as the ‘batteries included language’ for programming. It has various libraries used for scientific purposes and inquiries along with number of third-party libraries for making problem solving efficient.

In this work, the Python libraries of Numpy, for scientific computation, and Matplotlib, for 2D plotting have been used. Along with this, Pandas tool of Python has been employed for carrying out data analysis. Random forest regressor is used to solve tasks by ensembling random forest method. As a development platform, Jupyter Notebook, which proves to work great due to its excellence in ‘literate programming’, where human friendly code is punctuated within code blocks, has been used.

3.2 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.

Considering the case of Item-Weight, the feature item weight is shown to have a low correlation with the target variable Item-Outlet-Sales in Fig.6.

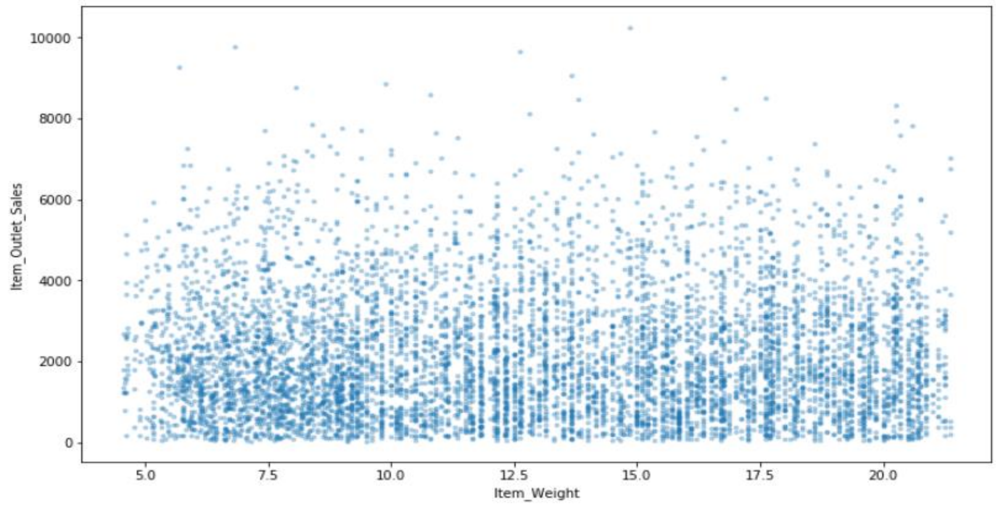

Fig6: Correlation between target variable and Item-weight variable

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

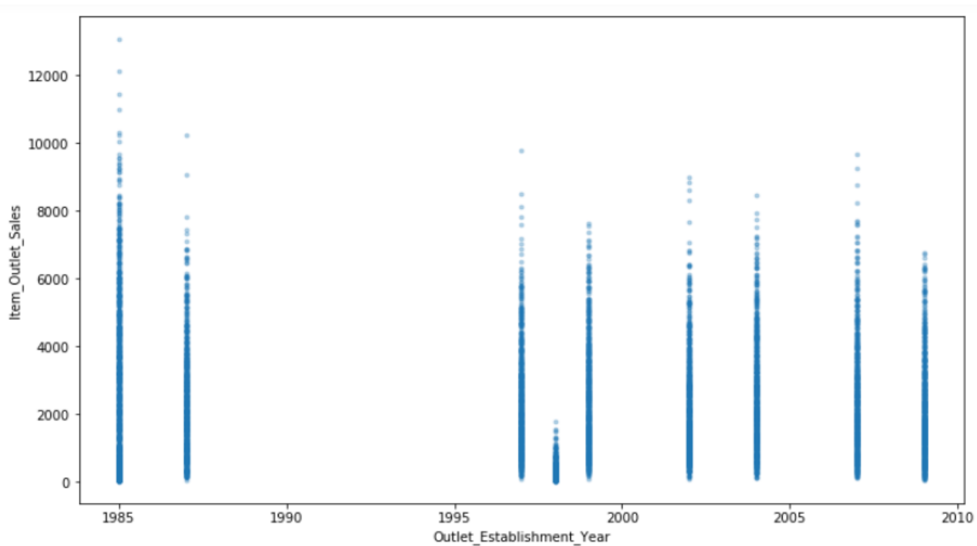

Fig7: Correlation between target variable and Outlet-establishment-year variable

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

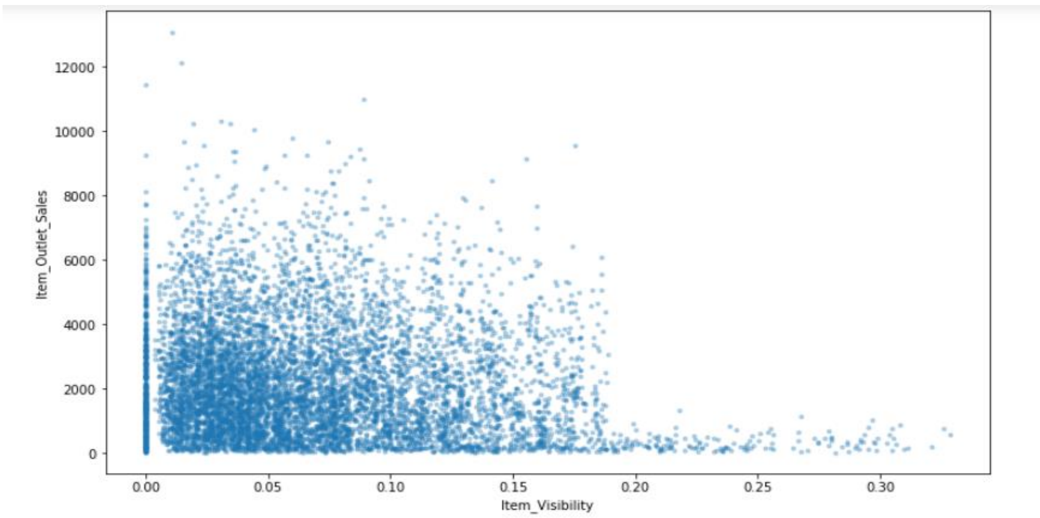

Fig8: Correlation between target variable and Item-visibility variable

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

```
In [16]: ct=[x for x in data.dtypes.index if data.dtypes[x]=='object']

In [17]: ct=[x for x in ct if x not in['Item_Identifier','Outlet_Identifier','source']]

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

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

Fig9: Different item categories in the dataset

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

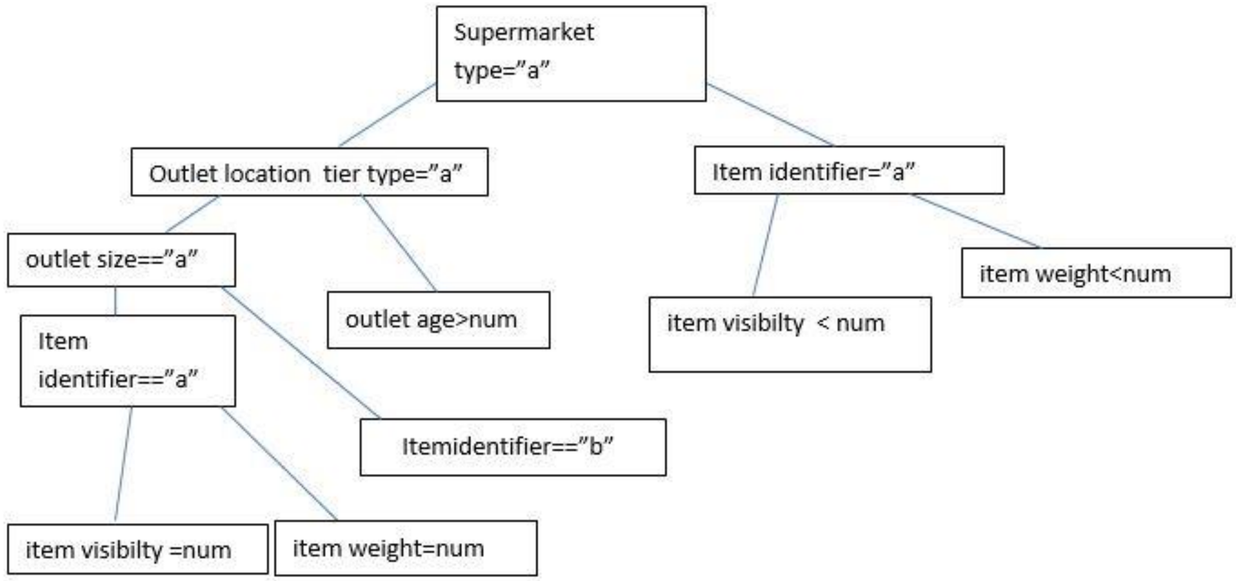

Fig 10: Flowchart for division of dataset on various factors (having proper leaves after pruning)

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

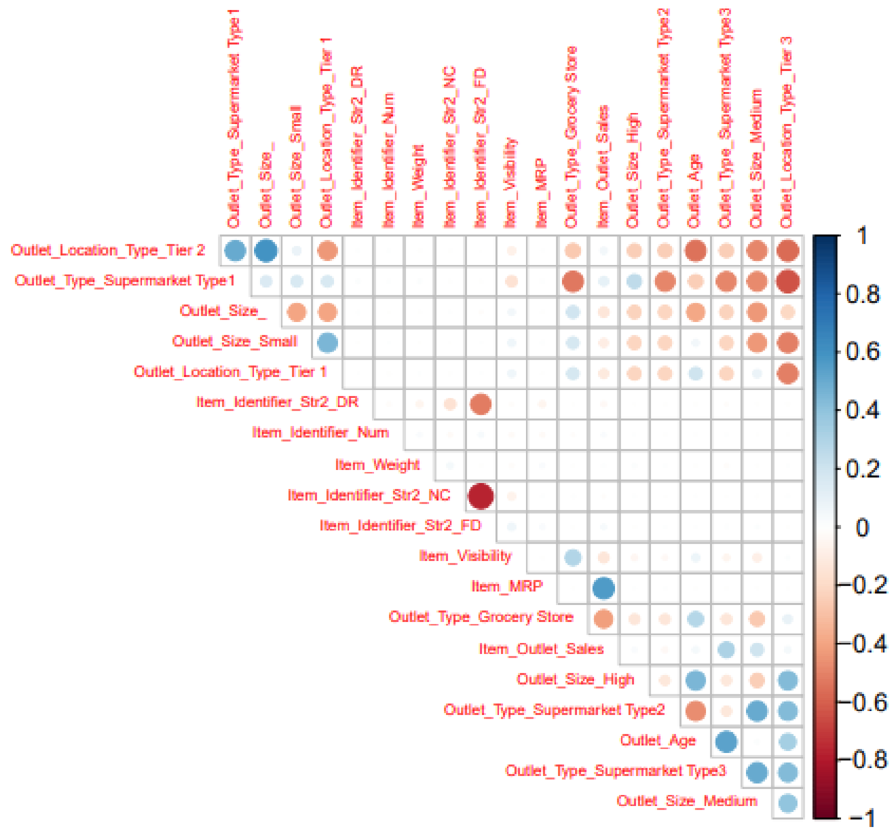

Fig11: Diagram showing correlation among different factors

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

- Item_visibility is having nearly zero correlation with our dependent variable item_outlet_sales and grocery store outlet_type. This means that the sales are not affected by visibility of item which is a contradiction to the general assumption of “more visibility thus, more sales”.
- Item_MRP (maximum retail price) is positively correlated with sales at an outlet, which indicates that the price quoted by an outlet plays an important factor in sales.
- Outlet situated in location with type tier 2 and size medium are also having high sales, which means that a one-stop-shopping-center situated in a town or city with populated area can have high sales.
- Variation in MRP quoted by various outlets depends on their individual sales.

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

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

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

OLS Regression Results

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

Fig12. Summary from linear regression model

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

3.3 Prediction results and Conclusion

- 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, as shown by code snippet in Fig.13.

```
from sklearn.ensemble import RandomForestRegressor
X_train = sd.drop(['Item_Outlet_Sales','Item_Identifier','Outlet_Identifier'],axis=1)
Y_train = sd['Item_Outlet_Sales']
X_test = ds.drop(['Item_Identifier','Outlet_Identifier'],axis=1).copy()
rf = RandomForestRegressor(n_estimators=400,max_depth=6, min_samples_leaf=100,n_jobs=4)
rf.fit(X_train,Y_train)
rf_pred = rf.predict(X_test)
rf_accuracy = round(rf.score(X_train,Y_train)*100,2)
print('accuracy of random forest is : %.4g' %rf_accuracy)

accuracy of random forest is : 60.8
```

Fig 13. Code showing model score of random forest

It can be concluded that more locations should be switched or shifted to Supermarket Type3 to increase the sales of products at Big Mart. 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.

4 Conclusion and Future Scope

In this paper, basics of machine learning and the associated data processing and modeling algorithms have been described, followed by their application for the task of sales prediction in Big Mart shopping centers at different locations. On implementation, the prediction results show the correlation among different attributes considered and how a particular location of medium size recorded the highest sales, suggesting that other shopping locations should follow similar patterns for improved sales.

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

References

[1] Smola, A., & Vishwanathan, S. V. N. (2008). Introduction to machine learning. *Cambridge University, UK*, *32*, 34.

[2] Saltz, J. S., & Stanton, J. M. (2017). *An introduction to data science*. Sage Publications.

[3] Shashua, A. (2009). Introduction to machine learning: Class notes 67577. *arXiv preprint arXiv:0904.3664*.

[4] MacKay, D. J., & Mac Kay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.

[5] Daumé III, H. (2012). A course in machine learning. *Publisher, ciml. info*, *5*, 69.

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

[7] Cerrada, M., & Aguilar, J. (2008). Reinforcement learning in system identification. In *Reinforcement Learning*. IntechOpen.

[8] Welling, M. (2011). A first encounter with Machine Learning. *Irvine, CA.: University of California*, *12*.

[9] Learning, M. (1994). Neural and Statistical Classification. *Editors D. Mitchie et. al*, 350.

[10] Mitchell, T. M. (1999). Machine learning and data mining. *Communications of the ACM*, *42*(11), 30-36.

[11] Downey, A. B. (2011). *Think stats*. " O'Reilly Media, Inc.".

[12] Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly Media.