Big Mart Sales Prediction and Analysis

Shantanu Choudhary[1], Utkarsh Singh[2], Nikhil Saxena[3], Sameer Jain[4]
[1,2,3,4] *Computer Engineering, Raj Kumar Goel Institute of Technology*

Abstract— **Machine Learning is a technology that allows machines to become more accurate in predicting outcomes without being explicitly programmed for it. The basic premise of machine learning is to build models and deploy algorithms that can receive input data and use statistical analysis to predict an output while modifying outputs as the new data becomes available. These models can be used in different areas and trained to match the expectations so that accurate steps can be taken to achieve the organization's target. In this paper, the case of Big Mart Shopping Centre has been discussed to predict the sales of different types of items and for understanding the effects of different factors on the sales of different items. Taking various features 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 used to take decisions to improve sales.**

Index Terms: **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 task's 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.

1.1 Machine Learning

The volume of data is increasing day by day and such unprocessed data is needed to be analysed more precisely, as it could give very informative and accurate gradient results as per current standard requirements. Artificial Intelligence (AI) is evolved over the past two decades, Machine Learning (ML) is also on a fast pace of evolution. Machine Learning is an important pillar of IT sector and with that it becomes a rather important part of our life [1]. As the technology progresses further, 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 problems and generally a classification type of problem accounts as a source for knowledge discovery. It generates resources and deploys regression to make accurate predictions about the future, the main emphasis being laid on making the system self-efficient, to be able to do analysis and computations to generate more accurate and precise results [2]. By using statistics and probabilistic tools, data can be converted into knowledge.

In this paper, we have discussed various applications of Machine Learning and the types of data it could deal with. Now, the problem statement addressed through this work is stated in a formalized manner. This is followed by explaining the methodologies ensued and the prediction results observed on implementation. Various machine learning algorithms include [3]:

- Linear Regression: It can be referred to as a parametric ML technique which is used to predict a continuous or dependent variable since a dataset of independent variables is provided. This technique is said to be parametric as different predictions are made based on the dataset.
- K-Nearest Neighbors (KNN): It is a machine learning algorithm which is based on instances and knowledge gained through them [4]. Unlike data mining in streams, cases where every sample could simultaneously belong to multiple classes in hierarchical multi-label classification

problems, k-NN can produce results in a structured form [5].

- Naïve Bayes classifiers: Naïve Bayes Classifiers are based on Bayes theorem and a collection of classification algorithms where classification of every pair is independent of each other. Bayesian learning is capable of provide predictions with readable reasons by generating an if-then form of list of rules [8].
- Random Tree: It is an efficient machine learning algorithm for achieving scalability and is used in identification problems for building approximation systems. The decisions are taken considering the choices made based on the possible consequences, the variables which are included, input factor. Other algorithms that could be employed are SVM, XG-Boost, logistic regression and so on [7].
- K-means clustering: This algorithm is used in unsupervised learning for creating clusters of related data based on their closeness to the centroid value [9].

1.2 Problem Statement

"To build a framework that is able to predict future sales of Big Mart from given data using the Machine Learning Algorithms".

II.METHODOLOGY

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

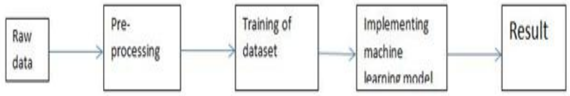
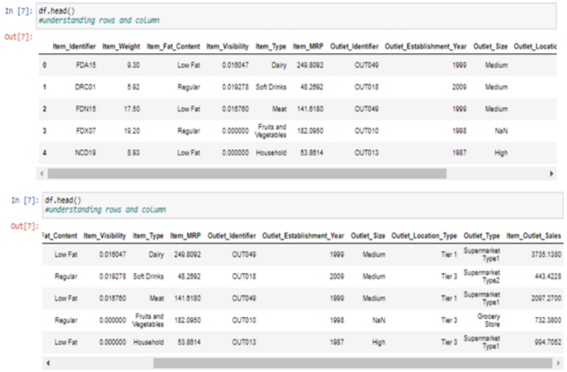


Fig1: Steps followed for obtaining results

2.1 Pre-processing of Dataset

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



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

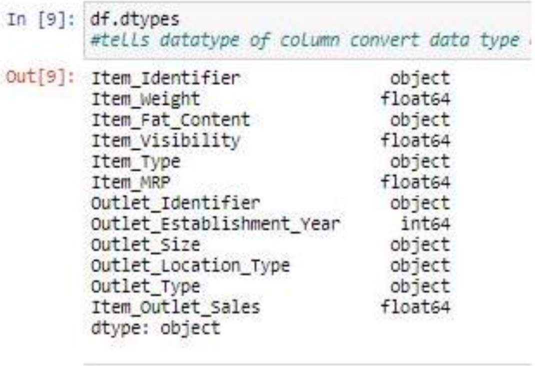
	Item_Identifier	Item_Weight	Item_Fat_Content	Item_Visibility	Item_Type	Item_MRP	Outlet_Identifier	Outlet_Establishment_Year	Outlet_Size	Outlet_Locatic
0	FDA15	9.30	Low Fat	0.016047	Dairy	249.8092	OUT049	1999	Medium	
1	DRC01	5.92	Regular	0.019278	Soft Drinks	48.2692	OUT018	2009	Medium	
2	FDN15	17.50	Low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

at_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 such as integer, float and, object as shown in Fig.3.



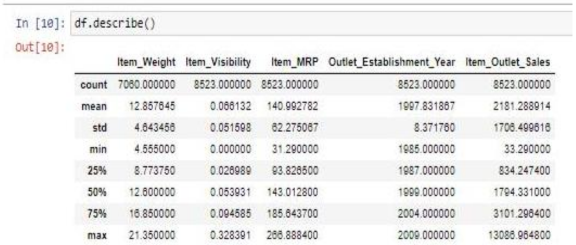
```
In [9]: df.dtypes
        #tells datatype of column convert data type

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 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.

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.



```
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

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.

2.2 Algorithms employed

Scikit-Learn can be used to track ML system on wholesome basis [12]. Algorithms deployed for predicting sales for this Big Mart dataset are discussed as follows:

Random Forest Algorithm

Random forest algorithm is an accurate machine learning algorithm which will 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 similar hyper parameters like decision tree. The tree model is same as decision tool. Fig.5 shows the relationship between random forest and decision tree. To solve regression tasks of prediction by virtue of random forest, the sklearn.ensemble library's random forest regressor class is used. The important role is played by the parameter labled as n_estimators which also comes under random forest regressor. Random forest could be referred to as a meta-estimator that is used to fit upon numerous decision trees by taking the different sub-samples of the dataset.

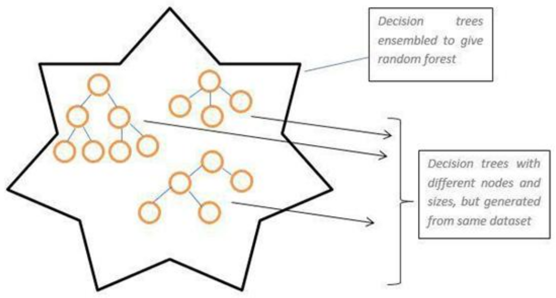

Decision trees ensembled to give random forest
Decision trees with different nodes and sizes, but generated from same dataset

Fig5: Relation between Decision Trees and Random Forest

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.

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

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

Y – Dependent Variable

X – Independent Variable

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

$\in$ -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].

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

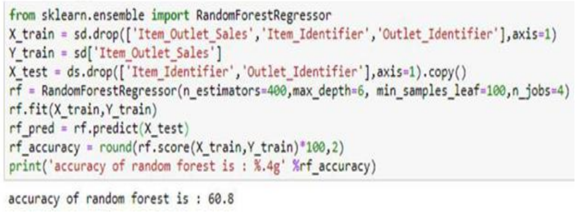
Python is a general purpose, interpreted-high level programming language that is used extensively nowadays for solving domain problems instead of dealing with complexities of a system. It is also known as the 'batteries included language' for programming. It has various libraries that could be used for 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.

Random forest regressor is used to solve problems by using random forest method. As a development platform, Jupyter Notebook, which has proven to work great due to its excellence in ‘literate programming’ scheme, where human-friendly code is punctuated within code blocks to make it more presentable and readable.

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.
- 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.
- Also, model accuracy and score of regression model can reach nearly 61% if it is built with more hypothesis consideration and analysis, as represented 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

From this it can 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 centres like Big Mart could benefit from this model by using it to predict the sales of its products on different locations.

IV.CONCLUSION AND FUTURE SCOPE

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 show s 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 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.

REFERENCES

[1] Kadam, H., Shevade, R., Katekar, P. and Rajguru (2018). A Forecast of Big Mart Sales based on Random Forests and Multiple Linear Regression

[2] Makridakis, S., Wheelwrigh.S.C., Hyndman. R.J (2008). Forecasting methods and applications

[3] C.M. Wu, P.Patil and S. Gunaseelan (2018). Comparison of different machine learning

algorithms for Multiple Regression on Black Friday Sales Data

[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.