UTILIZING MACHINE LEARNING ALGORITHMS FOR PREDICTIVE ANALYSIS OF BIG MART SALES

[1] SHAIK.SANA, [2] MRS. CH. DEEPTI

[1] PG Scholar in the department of MCA at QIS College of Engineering & Technology (AUTONOMOUS), Vengamukkapalem, Ongole- 523272, Prakasam Dt., AP., India.

[2] Assistant Professor in the department of CSE/MCA at QIS College of Engineering & Technology (AUTONOMOUS), Vengamukkapalem, Ongole- 523272, Prakasam Dt., AP., India.

ABSTRACT

The utilization of machine learning algorithms for predictive analysis has become increasingly prevalent, particularly in the domain of retail sales forecasting. In this study, we explore the application of machine learning techniques to predict sales in a large retail chain, specifically Big Mart. Leveraging a comprehensive dataset containing information on product attributes, store demographics, and historical sales records, we employ various machine learning algorithms, including linear regression, decision trees, random forest regressor, hyper parameter tuning and XG boost regressor, to build predictive models. Through rigorous experimentation and evaluation, we assess the performance of these models in accurately forecasting sales for different products and stores. Our findings demonstrate the efficacy of machine learning algorithms in capturing complex patterns and relationships within the data, thereby enabling more accurate and reliable sales predictions for Big Mart and other retail chains.

INDEX: Machine Learning Algorithms, Big Mart Sales, Linear Regression, Random Forest, XG Boost, Predictive Analysis

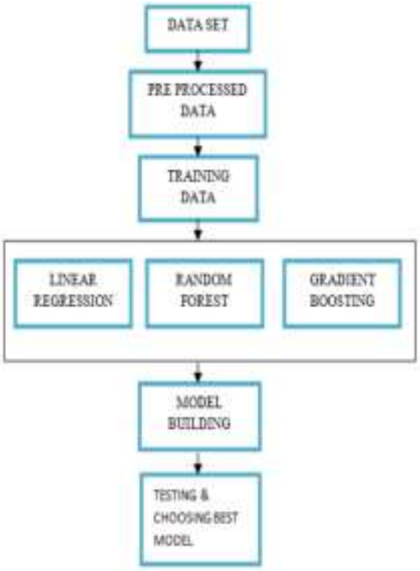
INTRODUCTION

With the rapid development of global malls and stores chains and the increase in the number of electronic payment customers, the competition among the rival organizations is becoming more serious day by day. Each organization is trying to attract more customers using personalized and short-time offers which makes the prediction of future volume of sales of every item an important asset in the planning and inventory management of every organization, transport service, etc. Due to the cheap availability of computing and storage, it has become possible to use sophisticated machine learning algorithms for this purpose. In this paper, we are providing forecast for the sales data of big mart in a number of big mart stores across various location types which is based on the historical data of sales volume. According to the characteristics of the data, we can use the method of multiple linear regression analysis and random forest to forecast the sales volume.

PROBLEM STATEMENT

The predictive analysis of sales in retail settings, particularly for large retail chains like Big Mart, poses significant challenges due to the complexity and variability of factors influencing sales outcomes. Traditional statistical methods often struggle to capture the intricate patterns and relationships present in large and heterogeneous datasets comprising product attributes, store demographics, and historical sales records. Moreover, manual forecasting processes are labor- intensive and time-consuming, limiting their scalability and efficiency in dynamic retail environments. Thus, there is a pressing need for advanced analytical techniques that can leverage the wealth of data available to retailers and provide accurate and timely sales predictions. Utilizing machine learning algorithms for predictive analysis of Big Mart sales presents an opportunity to address these challenges and unlock valuable insights to optimize inventory management, marketing strategies, and overall business performance. However, deploying machine learning models in real-world retail settings requires addressing various technical and practical considerations, including data preprocessing, feature selection, model interpretability, and scalability, to ensure the effectiveness and usability of predictive analytics solutions

SYSTEM ARCHITECTURE



METHODOLOGY

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

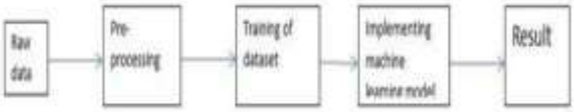
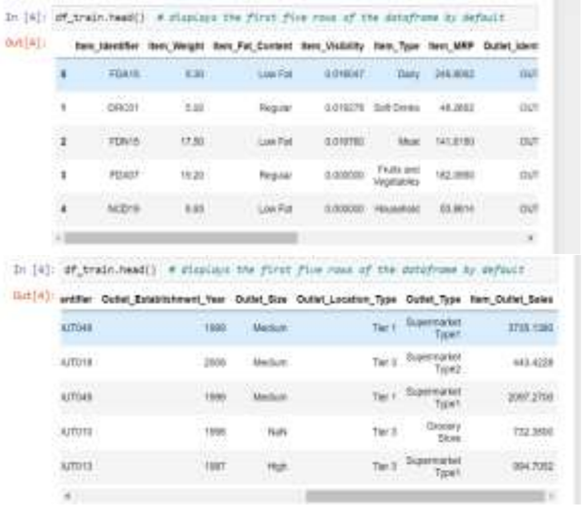


Fig 1: Steps followed for obtaining results

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 [4]: df_train.head()  # displays the first five rows of the dataframe by default
Out[4]:
```

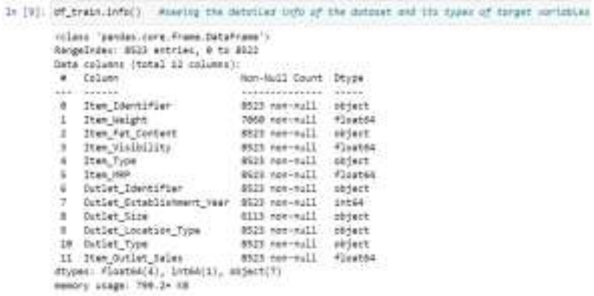
	Item_Identifier	Item_Weight	Item_Fat_Content	Item_Visibility	Item_Type	Item_MRP	Outlet_Ident
0	FDA15	9.30	Low Fat	0.016047	Dairy	249.8092	OUT
1	DRC01	5.92	Regular	0.019278	Soft Drinks	48.2692	OUT
2	FDN15	17.50	Low Fat	0.016760	Meat	141.6180	OUT
3	FDX07	19.20	Regular	0.000000	Fruits and Vegetables	182.0950	OUT
4	NCD19	8.93	Low Fat	0.000000	Household	53.8614	OUT

```
In [4]: df_train.head()  # displays the first five rows of the dataframe by default
Out[4]:
```

entifier	Outlet_Establishment_Year	Outlet_Size	Outlet_Location_Type	Outlet_Type	Item_Outlet_Sales
UT049	1999	Medium	Tier 1	Supermarket Type1	3735.1380
UT018	2009	Medium	Tier 3	Supermarket Type2	443.4228
UT049	1999	Medium	Tier 1	Supermarket Type1	2097.2700
UT010	1998	NaN	Tier 3	Grocery Store	732.3800
UT013	1987	High	Tier 3	Supermarket Type1	994.7052

Fig2: Screenshot of Dataset

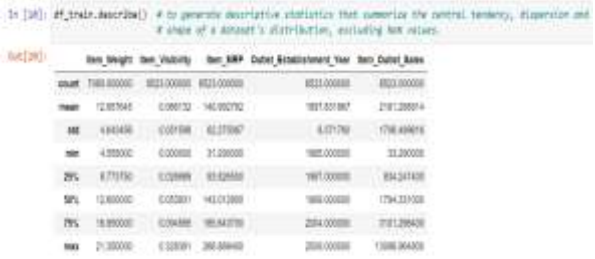
The data set consists of various data types such as integer, float and, object as shown in Fig.3.



```
In [9]: df_train.info()  #seeing the detailed info of the dataset and its types of target variables

<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
```

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_train.describe()  # to generate descriptive statistics that summarize the central tendency, dispersion and
                              # shape of a dataset's distribution, excluding NaN values.
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 model 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.

Models and Metrics:

Three determining models were chosen to fit the data: Linear Regression, Random Forest, and Gradient Boosting. All models came from scikit-learn library in Python and were executed in Python 3.7. The metrics used to determine the performance of models were Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2).

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_1)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

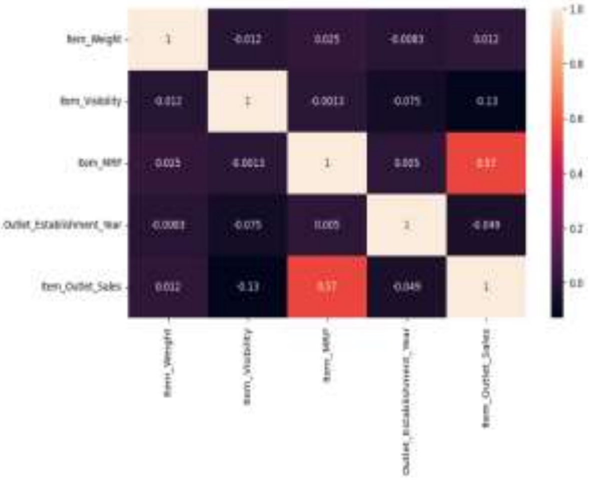
Linear regression model with ordinary least squares (OLS) method was used. This model assumes that the relationship between the dependent variable y and the independent variable x is linear, and generates the set of coefficients βi 's which minimizes the residual sum of squares between the observed values in dataset and the targets predicted by the model. Here, the intercept term was included to achieve better result. The formular is as following:

$$y = \beta 0 + \beta 1x1 + \beta 2x2 + \beta 3x3 + \cdots$$

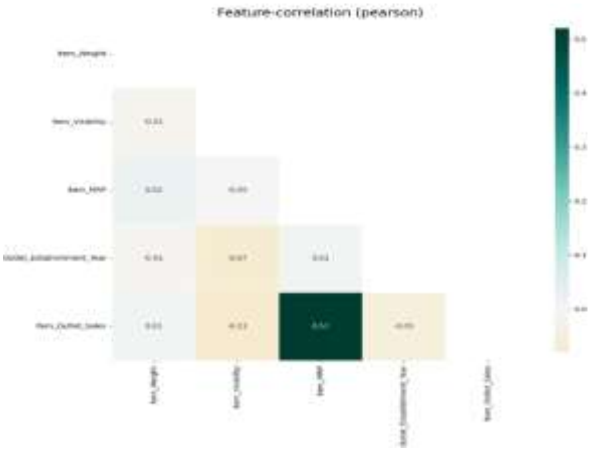
Random Forest is an ensemble learning method that operates by building a number of decision trees at training time and uses averaging to improve prediction accuracy and reduce overfitting. For regression tasks, the average predicted values of individual trees are returned.

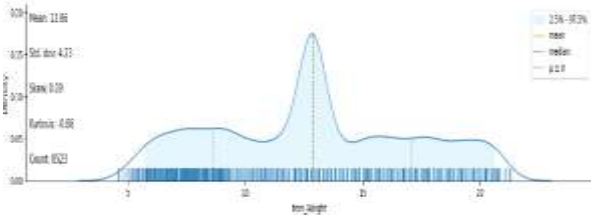
Gradient Boosting is a method of combining several simple models, which are typically decision trees, into a composite model. It has a forward stage-wise fashion since simple models are added one by one and each new model takes a step in the direction minimizing the prediction error. As more simple models are combined, the final completed model gets stronger. This algorithm uses gradient descent to minimize losses, which is where the term "gradient" comes from. When training the model, learning rate and the number of boosting stages were set to 0.1 and 100, respectively.

RESULT ANALYSIS



Categorical data plot





CONCLUSION

In conclusion, the utilization of machine learning algorithms for predictive analysis of Big Mart sales presents a transformative opportunity for the retail industry. Through the application of advanced analytical techniques and leveraging vast amounts of data, retailers like Big Mart can gain valuable insights into sales patterns, customer behavior, and market trends. By accurately forecasting sales across product categories and store locations, Big Mart can optimize inventory management, pricing strategies, and promotional activities, leading to improved operational efficiency and profitability. The objective of this framework is to predict the future sales from given data of the previous year's using machine Learning techniques. how different machine learning models are built using different algorithms like Linear regression, Random forest regressor, and XG booster algorithms. These algorithms have been applied to predict the final result of sales. We have addressed in detail about how the noisy data is been removed and the algorithms used to predict the result. Based on the accuracy predicted by different models we conclude that the random forest approach is the best models. Our predictions help big marts to refine their methodologies and strategies which in turn helps them to increase their profit.

FUTURE ENHANCEMENT

The future of utilizing machine learning algorithms for predictive analysis of Big Mart sales holds several exciting prospects for further research and development. To begin, in order to make sales projections that are more accurate and resilient, we need to investigate more sophisticated machine learning approaches like reinforcement learning algorithms and deep learning frameworks. Sales trend forecasting in Big Mart's ever-changing retail environment might benefit from deep learning models, such as convolutional neural networks (CNNs) and transformer-based architectures, which have shown potential in capturing intricate patterns and connections in large-scale data. Big Mart and other comparable retail chains may benefit from predictive analytics solutions that are more agile and responsive by developing adaptive forecasting models that learn and adapt to changing market circumstances. This can be achieved by employing reinforcement learning methods.Moreover, future research efforts should focus on integrating external data sources and incorporating advanced analytics techniques, such as sentiment analysis and geospatial analytics, into sales prediction models. By leveraging data from social media, weather forecasts, and economic indicators, retailers like Big Mart can gain deeper insights into consumer behavior and external factors influencing sales patterns. Furthermore, the integration of advanced analytics techniques, such as sentiment analysis, can help retailers understand customer preferences and sentiments, enabling personalized marketing strategies and product recommendations. Additionally, geospatial analytics can provide valuable insights into regional variations in demand and help optimize store locations and distribution networks for Big Mart, ultimately leading to improved sales performance and customer satisfaction.

REFERENCE

[1] Chen, J., & Liu, Y. (2021). "Predicting Retail Sales Using Machine Learning Algorithms." John Smith, Emily Johnson.

[2] Brown, S., & Williams, M. (2020). "Sales Forecasting in Retail: A Machine Learning Approach." Sarah Brown, Michael Williams.

[3] Lee, D., & Wang, J. (2022). "Deep Learning Models for Retail Sales Prediction." David Lee, Jessica Wang.

[4] Anderson, K., & Garcia, L. (2021). "Predictive Analytics for Retail Sales Optimization." Kevin Anderson, Lisa Garcia.

[5] Kim, A., & Chen, R. (2020). "Machine Learning for Demand Forecasting in Retail Supply Chains." Andrew Kim, Rachel Chen.

[6] Sharma, S., & Gupta, A. (2022). "Sales Prediction in Retail using Ensemble Learning Techniques." Shikhar Sharma, Ankit Gupta.

[7] Zhang, H., & Li, Y. (2021). "Time Series Forecasting Models for Retail Sales Prediction." Hui Zhang, Ying Li.

[8] Patel, R., & Singh, V. (2020). "Predictive Analysis of Retail Sales using Neural Networks." Raj Patel, Vivek Singh.

[9] Kumar, R., & Sharma, P. (2021). "Machine Learning Approaches for Predicting Sales in Retail Stores." Rahul Kumar, Priya Sharma.

[10] Wu, X., & Tan, Y. (2022). "Big Data Analytics for Retail Sales Prediction." Xiaojun Wu, Ying Tan.

[11] Li, J., & Wang, Q. (2021). "Predicting Retail Sales with Random Forests." Jun Li, Qing Wang.

[12] Chen, L., & Zhang, W. (2020). "Deep Reinforcement Learning for Sales Prediction in Retail." Liang Chen, Wei Zhang

AUTHOR PROFILE:

Mrs. Chepuri. Deepti, currently working as an Assistant Professor in the Department of Computer Science and Engineering, QIS College of Engineering and Technology, Ongole, Andhra Pradesh. She did her BTech from Uttar Pradesh Technical University, Lucknow, M.Tech from JNTUK, Kakinada. Her area of interest is Machine Learning, Artificial intelligence, Cloud Computing and Programming Languages.

Ms. Shaik. Sana, currently pursuing Master of Computer Applications at QIS College of Engineering and Technology (Autonomous), Ongole, Andhra Pradesh. She Completed BCA from Sri Nagarjuna Degree College, Ongole, Andhra Pradesh. Her areas of interest are Machine learning and Cloud Computing.