Research Article

Machine Learning for Intelligent Sales Prediction

Md. Arafath Kafy [1]

[1] Daffodil International University, Dhaka, Bangladesh; arafath15-3946@diu.edu.bd
* Correspondence: arafath15-3946@diu.edu.bd

Abstract: — Business is an activity where the main goal is to make more profit. Decision-making is the key to achieving this objective, the result depends on which decision was taken in the beginning. That's why the decision which will give the best outcome gets more priority. Though for ages people have relied on their intelligence and experience to make decisions, in recent times they have depended on technology because it analyzes much more efficiently and shows an approximate outcome of that decision. The study focuses on predictive models relating to sales projections for businesses. The collaboration of traditional methods and modern technology can bring the best possible outcome for a business. The study verifies the potential impact of the selected strategies on sales results and the potential increase in revenue. Through this analysis, the paper highlights the strategic significance of incorporating technology-driven insights rather than relying solely on personal experience. By assessing how technology shapes decision-making processes in sales contexts, this research contributes to the growing understanding of business strategies in a rapidly digitizing world. The findings underscore the benefits of technology for increasing business profitability. Ultimately, this study provides insights for practitioners and stakeholders to harness the power of technological tools for more informed and effective decision-making, thereby improving business performance.

Keywords: Sales prediction; Linear Regression; Lasso Regression; Decision Tree; Random Forest; XGBRegressor

1. Introduction

Maximizing profit is the fundamental objective for business organizations. And to make this possible a strategic approach is the most important thing. In recent times, advanced data mining techniques and predictive modeling jointly brought changes into business operations to achieve this objective. The purpose of this research is to explore how various methodologies and models can be harnessed to optimize profitability within an organization. Making decisions based on data rather than insight is known as data-driven decision-making. (Provost and Fawcett, 2013). Companies in the top third of their industries are 5% more productive and 6% more profitable when they use data-driven decision-making. (McAfee and Brynjolfsson, 2012; Brynjolfsson et al., 2011). In acceptance studies of decision support systems (DSS), explanation capabilities have been shown to increase compliance with recommendations (Arnold et al., 2006; G¨on¨ul et al., 2006). In contrast, machine learning models such as random forests, boosting, support vector machines (SVMs), and neural networks perform significantly better than simple, interpretable models such as decision trees, naive Bayes, or decision rules. They are top-performing black box ML models (Caruana and Niculescu-Mizil, 2006).

The research methodology involves the application of different models at different points in time to get reliable and accurate results. The process may involve integrating results from multiple models to lay a solid foundation for achieving the overarching goal of maximum

profit. The objective is to prioritize outcomes that are not only reliable but also offer maximum returns, thus ensuring informed decision-making essential to sustainable success.

Several challenges may emerge during research, including issues such as data scarcity, incomplete datasets, and identification of outliers. An important consideration is the need for historical data analysis, a prerequisite for effective prediction and forecasting. This need is particularly pronounced when introducing a novel product to the market, as a lack of historical data can prevent accurate sales forecasting. As a result, strategizing to increase sales of such new products becomes a complex task. Addressing these challenges and achieving accurate predictions constitute the key challenges that this research addresses.

The author of this paper lives in Bangladesh. Which has a major developing market economy. This country is developing day by day but slowly. Also, many economic institutions are growing over time. So it is important to make the economy of this country stable to help it develop rapidly and stand side by side with the top-ranked developed country. This situation motivated the author to do something to push the economy by helping those economic institutions and making the country financially stable.

So, in this paper, we aimed to provide a comprehensive exploration of data mining techniques and predictive modeling to optimize profit margins for financial institutions. By perfectly analyzing the application of diverse models and addressing associated challenges, the study aims to find out the efficient decision-making landscape that maximizes sales for the organization.

2. Related Works

To enhance profitability, this study focuses on the strategic use of data mining techniques and predictive models. By using machine learning algorithms, prediction models and forecasting techniques organizations can uncover insights that empower informed decision-making. Many research has been done on this topic using various models previously.

'Walmart's Sales Data Analysis - A Big Data Analytics Perspective'. An analysis of Walmart's sales data is presented in this study. This study examines the data collected from a retail store in order to predict future strategies related to the management of the store. Various sequences of events, including weather conditions, holidays, etc., can actually alter the state of different departments, so it examines these effects as well as their influence on sales [1].

A study by Harsoor & Patil in 2015 [2] used Hadoop, MapReduce, and Hive to forecast sales of Walmart stores. This paper used 45 Walmart stores 3 years data for analysis, to forecast the sales for upcoming 39 weeks. Using HDFS and Map Reduce, they collected enormous sales data and transferred it to HDFS [3], but were later unable to draw conclusions due to the enormous data size. The average sales feature for all 45 stores and 99 departments was calculated using Hive processing. R programming was used to compute statistics using machine learning algorithms. Thus, Holt Winters [2] was used to train the dataset provided by Walmart and then to predict sales. An interactive data visualization tool called Tableau was then used to visualize the predicted sales.

Katal, Wazid, & Goudar [5] conducted a thorough study in 2013 regarding handling of Big Data; issues, challenges, tools, and best practices. Data quality, data scalability, fault tolerance,

and heterogeneous data processing were also mentioned as technical challenges. Distributed file systems, MapReduce [4] and Spark are proposed as good tools for handling big data [5] using Parallel Programming Models.

'Intelligent Sales Prediction Using Machine Learning Techniques'. This study explores the decisions to be taken based on experimental data and insights obtained through the visualization of data through Intelligent Sales Prediction Using Machine Learning Techniques. It has used data mining techniques. In terms of picturing future transactions, gradient boost has shown maximum accuracy [6].

'Utilization of artificial neural networks and GAs for constructing an intelligent sales prediction system'. In the study, deep neural network techniques were used to predict a company's sales strategy regarding electronic components in advance. To optimize the efficiency of the system, certain optimization algorithms are used, such as Genetic Algorithms [7].

'Applying machine learning algorithms in sales prediction' this thesis examines several distinct methods of implementing machine learning algorithms to achieve better, more optimal results, which are then analyzed for prediction purposes. Four algorithms were used, as well as an ensemble technique. Different tactics have also been used to select features [8].

'Research on retailer data clustering algorithm based on Spark' Forecasts of sales are essential inputs for many decision making activities at the organizational level, including operations, marketing, sales, production, and finance [9].

'Sales Prediction System Using Machine Learning'. In this paper, techniques such as Clustering Models and measures for sales prediction are used to obtain correct results in terms of predicting future sales or demands for a firm. Researchers are utilizing algorithmic methods in future research due to their high potential [10].

'Retail sales prediction and item recommendations using customer demographics at store level' This paper describes a system for predicting sales along with a system for recommending products to a group of retail stores. The sales of each individual have been carefully designed based on consumer demographic information [11].

'Bayesian learning for sales rate prediction for thousands of retailers' The paper shows that from the prediction of the single vendor's transaction rate, a large number of vendors would benefit from it, thereby enabling the construction of a set-up capable of estimating a large number of outputs. A neural network approach is used to make the prediction. Using Bayesian learning, they gained insights here [12].

'A Survey on the Clustering Algorithms in Sales Data Mining' It is difficult for the business sector to identify an accurate data mining technique and an effective prediction strategy [13].

'Combining Data Mining and Machine Learning for Effective User Profiling' In this research, an automatic prototype is employed in order to detect suspicious behavior. To create this appropriate prototype, several machine learning techniques have been used. To uncover the discrepancy in cell phone owners' conducts, data mining and constructive induction techniques are combined [14].

'BRAND LOYALTY ANALYSIS SYSTEM USING K-MEANS ALGORITHM' To be competent enough and generate higher revenue, business organizations continuously seek a better technique or model for data mining [15].

'Analysis & Prediction of Sales Data in SAP-ERP System using Clustering Algorithms' Due to

the enormous amount of data generated by e-commerce transactions. In terms of sales data analysis, there are a lot of issues, and the most important are identifying product attributes, fixing prices, realizing net sales and launching new products. The Expectation Maximization algorithm (EM) is discussed with various prediction methods, sales forecasting strategies, and sales forecasting strategies [16].

3. Experiments and Methods

The purpose of this study is to find the most effective model among some selected models to predict sales. This will improve the efficiency and reliability of forecasting sales.

A. Data Collection and Preparation

The dataset used for this research was collected from "Kaggle". In 2013 the data scientists at BigMart Collected sales data for 1559 products across 10 stores in different cities. The products and stores have also been defined according to certain attributes. Predicting the sales of each product in a specific store using a predictive model is the objective. The Dataset contains two CSV files named "Train" and "Test" both of the files contains similar types of data which includes Item Identifier, Item Weight, Item Fat Content, Item Visibility, Item Type, Item MRP, Outlet Identifier, Outlet Establishment Year, Outlet Size, Outlet Location Type, Outlet Type and one extra data in Train file- Item Outlet Sales. Because both the files contain similar types and values of data we merged both of them and created one dataset for our work.

B. Outlier Detection

In order to achieve optimal performance, outlier detection is an essential part of the preprocessing of data. It is possible to use the outlier detection process to deploy a model or to begin optimizing it further. In addition to displaying generic information without regard to the parameters of a model, it can be used to display more detailed information. Most importantly to maintain the quality of each data attribute.

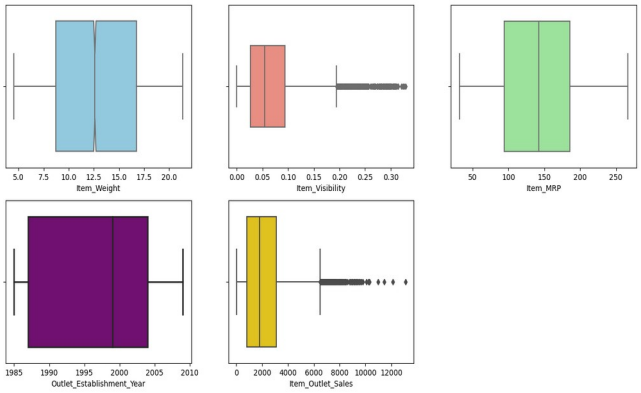


Figure 1: Boxplots for Outlier detecting

After completing the process of outlier detection, outliers were found in “Item Visibility” and “Item Outlet Sales” data columns as shown in figure 1.

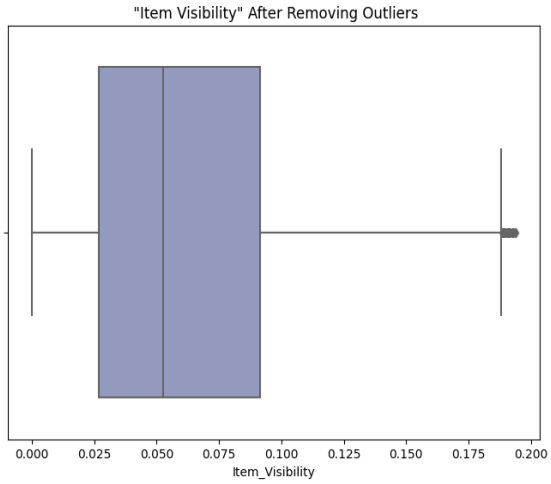


Figure 2: “Item Visibility” after removing outliers

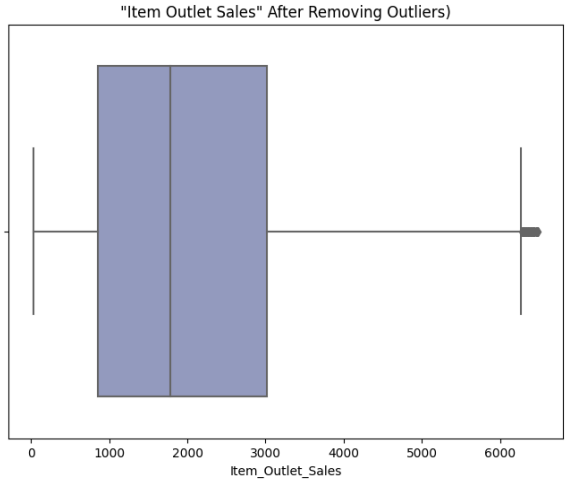


Figure 3: “Item Outlet Sales” after removing outliers

After removing those outliers as shown in Figure 2 and Figure 3 the dataset was updated.

C. Missing Value Handling

Because of corrupted data or failure to record data, the dataset contains missing values. These missing values can negatively affect the result so it is essential to remove or handle them. In our dataset among 14204 data of each column, “Item Weight” contains 2439 missing values, “Outlet Size'' contains 4016 missing values and “Item Outlet Sales” contains 5681 missing values. To handle those missing values, they were filled with their mode and mean value.

D. Data Visualization

To visualize the correlation between the dataset attributes we used heat map. Heat map is a two dimensional data visualization process which is an element of Seaborn library. It is a color encoded matrix which is used to display the correlation between target variable and other attributes. In a heat map the color intensity represents the dependency of the target variable on the respective attribute. The higher the intensity of the color of an attribute relative to the target variable, the dependency is lower of the target variable on the respective attribute.

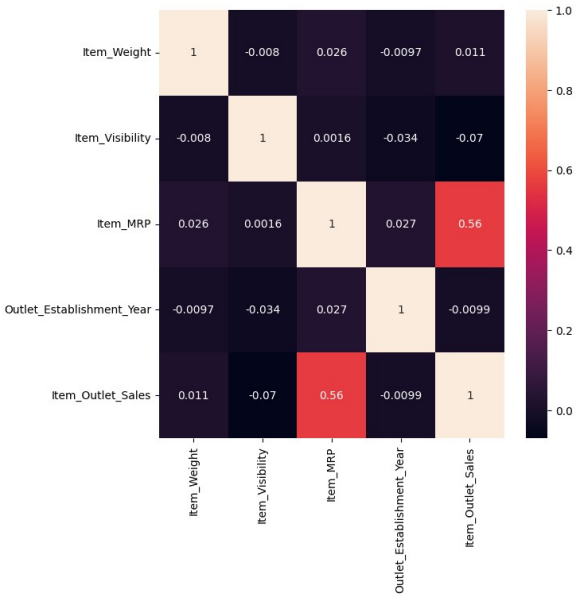


Figure 4: Heat map for Correlation between attributes

As shown in Figure-4 the target variable Item Outlet Sales is least dependent on item Visibility and most dependent on Item MRP. Therefore If the MRP increases the Item Outlet Sales decreases.

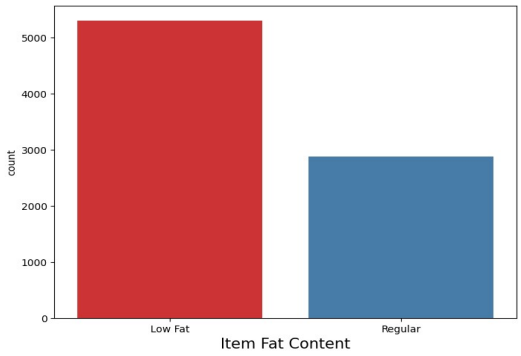


Figure 5: Fat content of Items

The Figure-5 is displaying the amount of fat in the items depending on two categories Low Fat and Regular Fat. It is observed that most of the items contain low fat.

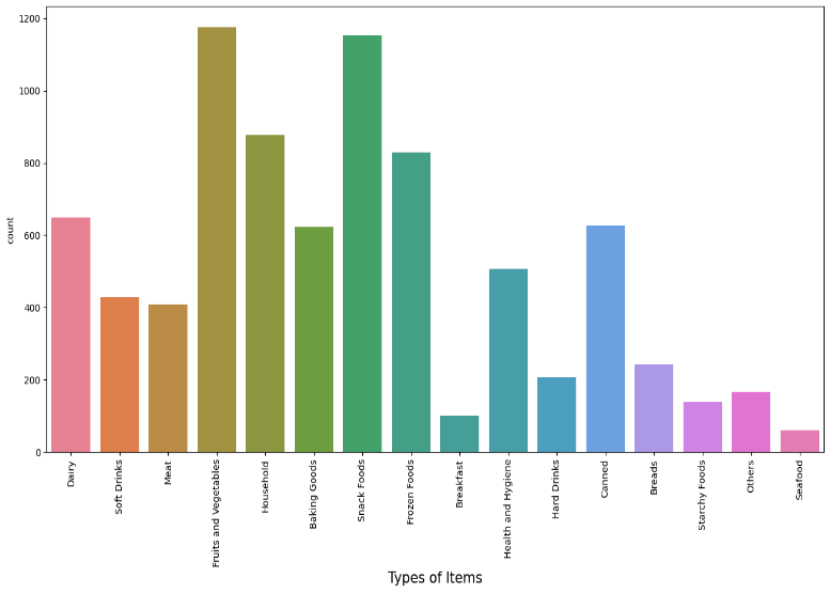


Figure 6: Amount of different types of items

Figure-6 displays which type of item has the highest amount. As we can see from the plot Fruits and Vegetables have the highest amount compared to any other item.

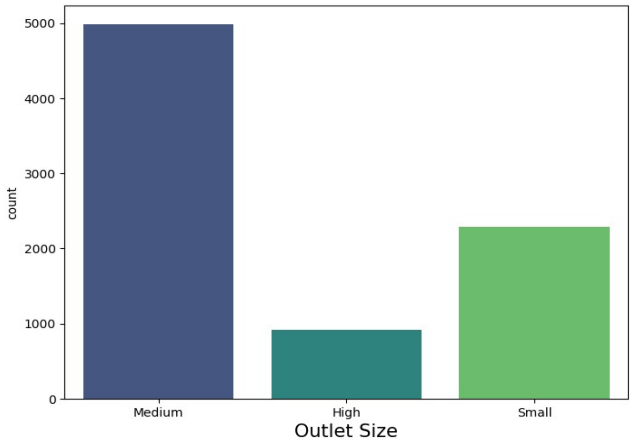


Figure 7: Size of different Outlets

From Figure-7 we can visualize that highest amount of outlet is medium in size. And high size outlets are in the lowest amount.

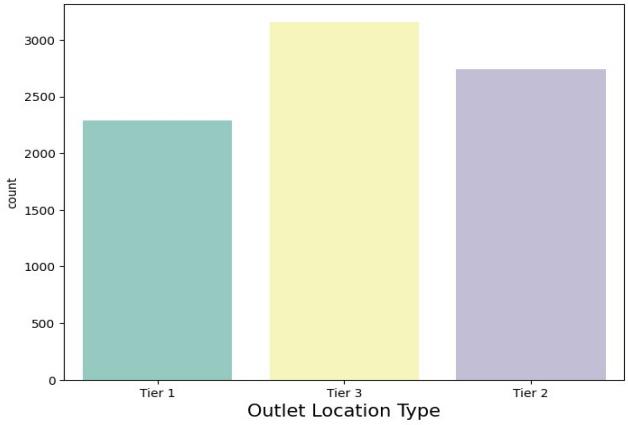


Figure 8: Number of outlet in different types of location

As shown in figure-9 among three different types of outlet location types 'Tier 3' type of location have the highest number of outlet.

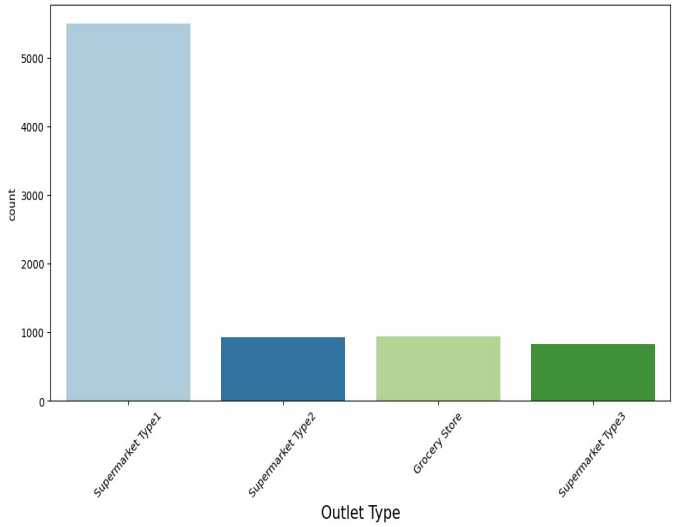


Figure 9: Number of each Outlet types

From Figure-9 it is observed that Supermarket Type 1 contains maximum number among various other Outlet types.

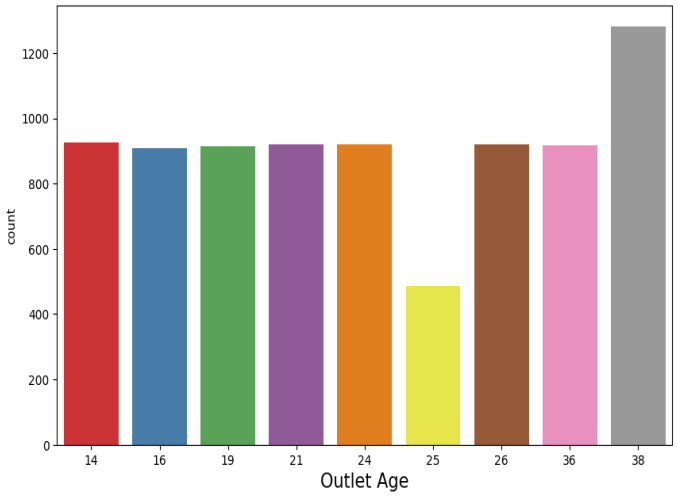


Figure 10: Age of different Outlet

In Figure-10 age of the Outlets are displayed. Where maximum number of outlets are 38 years old.

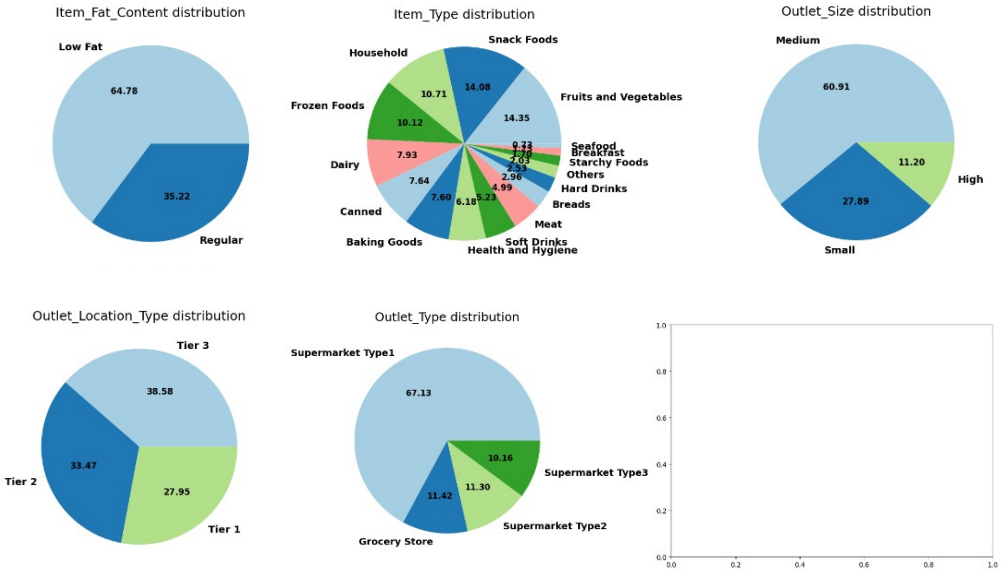
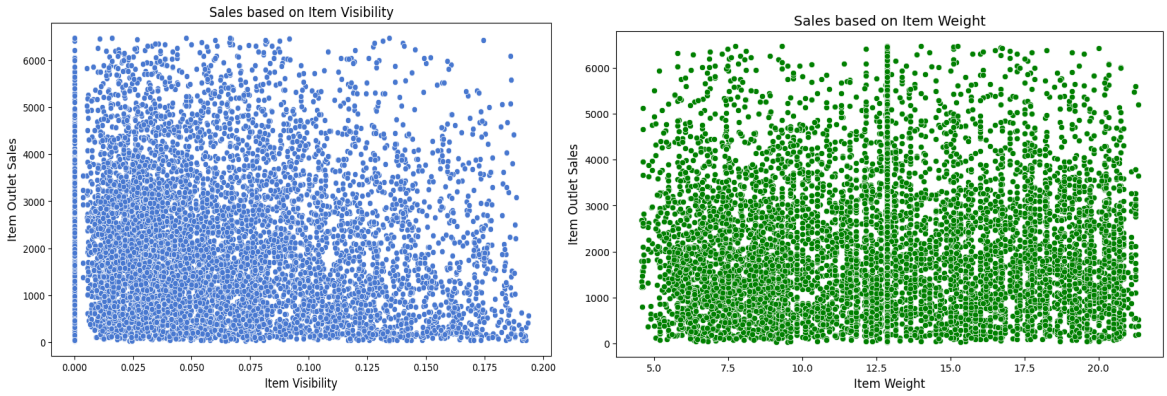
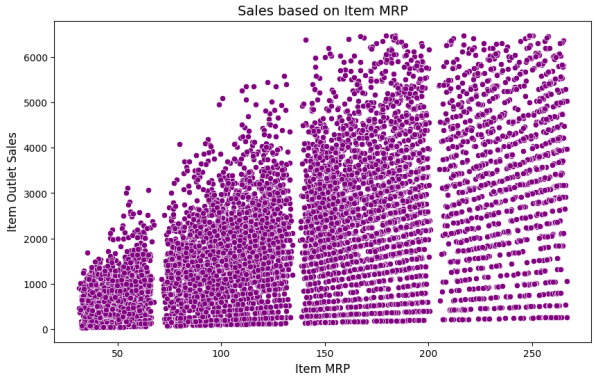


Figure 11: Distribution of all attributes

From Figure-11, a clear picture of how different attributes are distributed based on different categories can be seen. Using this diagram, all the data visualization of the dataset can be unified.



(a) Amount of Item Outlet Sales based on Item Visibility (b) Amount of Item Outlet Sales based on Item Weight



(c) Amount of Item Outlet Sales based on Item MRP

Figure 12: Amount of Item outlet sale based on different conditions

In the scatter diagrams of Figure-12 the amount of Item outlet sale based on different conditions is shown. In the regions where points are densely packed indicate a high rate of sales. And where the density is low the rate of sales is also low. In respect to different attributes the position of density is different.

E. Data Transformation

For some types of attributes, transforming the type of the data is important to implement algorithms and models on them. In our dataset attributes- Fat Content, Outlet Type, Outlet Size, Outlet Location Type, Item Type contains non-numeric data. To work with these data we transformed these data to numeric value we used Label Encoder for Fat Content, Outlet Type, Outlet Size, and Outlet Location Type. For Item Type we used One Hot Encoder and created separated attribute columns for each type of Item.

F. Implementing Machine-Learning Models

In our study, we utilized 5 types of models to find the best model for our dataset. By selecting Item outlet sales as the dependent variable and other attributes as independent variables we implemented those models. In data splitting part we employed train-test split for model evaluation where text size = 0.2 and random state = 2.

a) Linear Regression: One of the most common and widely used Machine Learning algorithms is linear regression. Linear regression shows the relation between the dependent and independent variable.
b) Lasso Regression: Lasso regression is also one type of regression model which is also one type of linear regression but it uses L1 regularization to penalize the absolute size of the regression coefficients. It shrinks the less important feature coefficients to zero, effectively performing feature selection.
c) Decision Tree: Decision tree is one type of non-linear tree based model. It recursively splits the data based on features to make predictions. They can be used for both classification and regression tasks.
d) Random Forest: It is an ensemble learning method which is also a tree based model. Random forest builds multiple decision trees and combines their outputs to make predictions. Just like Decision tree it is also effective for both classification and regression tasks.
e) XGBRegressor(XGBoost): XGBregressor is also an ensemble method and it is a Boosting model. XGBRegressor is a part of XGBoost library, which is an implementation of the gradient boosting algorithm. It is designed for regression tasks and builds a sequence of trees to improve model accuracy.

By evaluating these Machine learning models, we presented evaluation metrics for each model, including Mean Absolute Error (MAE), Mean Square Error (MSE) and R-squared scores. Cross-validation scores also provide insights into model performance across different subsets of the data.

4. Results

Both the Linear Regression and Lasso Regression models were employed for predicting sales figures. The models yielded comparable results with a Mean Absolute Error (MAE) of approximately 810-811 and a Squared Mean Squared Error (MSE) of around 1056-1057. The R-squared scores stood at about 0.4858, indicating a moderate fit to the data.

Table 1. Forecasting Errors for Linear and Lasso Regression.

Model	Mean Absolute Error	Mean Squared Error	R^2 score
Linear Regression	810.69124	1056.70978	0.48582475
Lasso Regression	809.86913	1055.44162	0.48582475

The similar performance of these models suggests that the regularization techniques applied in Lasso Regression did not significantly enhance predictive accuracy compared to the basic Linear Regression model.

The Decision Tree model demonstrated promising results, showcasing a lower MAE of around 705. However, it reported a considerably higher MSE of approximately 920897, suggesting potential overfitting issues. The R-squared scores were notably low, ranging around 0.15, which could indicate the model's limited ability to explain variance in the data. In contrast, the Random Forest model showed improvement with a slightly higher MAE of about 703, a reduced MSE of around 904560, and higher R-squared scores varying from 0.49 to 0.53. The ensemble-based nature of Random Forest led to enhanced predictive performance compared to the standalone Decision Tree.

Table 2. Forecasting Errors for Decision Tree, Random Forest.

Model	Mean Absolute Error	Mean Squared Error
Decision Tree	705.16565	920897.1807
Random Forest	702.77092	904559.9567

The XGBoost Regressor was utilized to predict sales figures. It provided an MAE of approximately 753 and an MSE of around 1080861. The R-squared scores fluctuated between 0.456 to 0.502, displaying a better performance compared to the Decision Tree but not surpassing the Random Forest.

Table 3. Forecasting Errors for XGBRegressor.

Model	Mean Absolute Error	Mean Squared Error
XGBRegressor	752.96553	1080861.683

By calculating the mean accuracy as the average of the cross validation scores obtained of each model we compared the accuracy of the models performance. The model with the highest accuracy is the best performer.

Table 4. Accuracy of trained Models

Model	Mean Absolute Error
Linear Regression	48.2%
Lasso Regression	48.2%
Decision Tree	14.6%
Random Forest	52%
XGBRegressor	48.6%

As shown in Table-4 out of the 5 models Random forest works best on the data we used with the accuracy of 52%.

5. Conclusions

Machine learning models play a crucial role in refining sales predictions. Random Forests appear to be the most effective solution, with an impressive 52% accuracy rate. Machine learning techniques are superior to traditional ones when it comes to providing reliable sales projections due to their robustness. To enhance decision-making, this study advocates integrating modern data mining approaches with traditional sales strategies. In addition to being able to overcome challenges such as data scarcity, predictive modeling is capable of driving business profitability and competitiveness. Predictive modeling is undeniably transformative, emphasizing the significance of informed decision-making in sales forecasting. Particularly, the role of Random Forest stands out as a driving force behind accurate and reliable sales predictions.

References

1. Singh Manpreet, Bhawick Ghutla, Reuben Lilo Jnr, Aesaan FS Mohammed, and Mahmood A. Rashid. "Walmart's Sales Data Analysis-A Big Data Analytics Perspective." In 2017 4th Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE), pp. 114-119. IEEE, 2017.
2. S. Harsoor and A. Patil, “Forecast of sales of walmart store using Big Data application,” International Journal of Research in Engineering and Technology, vol. 4, p. 6, June 2015.
3. M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. Mccauley, M. J. Franklin, S. Shenker, and I. Stoica, “Fast and interactive analytics over Hadoop data with Spark,” Usenix - The Advanced Computing Systems Association, 2012.
4. J. Dean and S. Ghemawat, MapReduce: simplified data processing on large clusters. Association for Computing Machinery, 2008.
5. Katal, M. Wazid, and R. H. Goudar, Big Data: Issues, Challenges, Tools and Good Practices, 2013.
6. Cheriyan, Sunitha, Shaniba Ibrahim, Saju Mohanan, and Susan Treesa. "Intelligent Sales Prediction Using Machine Learning Techniques." In 2018 International Conference on Computing, Electronics & Communications Engineering (iCCECE), pp. 53-58. IEEE, 2018.
7. Baba, Norio, and Hidetsugu Suto. "Utilization of artificial neural networks and GAs for constructing an intelligent sales prediction system." In Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium, vol. 6, pp. 565-570. IEEE, 2000.
8. Sekban, Judi. "Applying machine learning algorithms in sales prediction." (2019).
9. Huang, Q., & Zhou, F. (2017, March). “Research on retailer data clustering algorithm based on spark”. In AIP Conference Proceedings (Vol. 1820, No. 1, p. 080022). AIP Publishing.
10. Panjwani, Mansi, Rahul Ramrakhiani, Hitesh Jumnani, Krishna Zanwar, and Rupali Hande. “Sales Prediction System Using Machine Learning”. No. 3243. EasyChair, 2020.
11. Giering, Michael. "Retail sales prediction and item recommendations using customer demographics at store level." ACM SIGKDD Explorations Newsletter 10, no. 2 (2008): 84-89.
12. Ragg, Thomas, Wolfram Menzel, Walter Baum, and Michael Wigbers. "Bayesian learning for sales rate prediction for thousands of retailers." Neurocomputing 43, no. 1-4 (2002): 127-144.
13. Maingi, M. N. A Survey on the Clustering Algorithms in Sales Data Mining.
14. Fawcett, Tom, and Foster J. Provost. "Combining Data Mining and Machine Learning for Effective User Profiling." In KDD, pp. 8-13. 1996.
15. Saylı, A., Ozturk, I., & Ustunel, M. (2016). Brand loyalty analysis system using K-Means algorithm. Journal of Engineering Technology and Applied Sciences, 1(3).
16. Sastry, S. H., Babu, P., & Prasada, M. S. (2013). Analysis & Prediction of Sales Data in SAP-ERP System using Clustering Algorithms. arXiv preprint arXiv:1312.2678.