International Journal of Core Engineering & Management
Volume-7, Issue-06, 2023 ISSN No: 2348-9510

FORECASTING OF SUPERMARKET SALES USING BIG DATA ANALYTICS AND MACHINE LEARNING TECHNIQUES IN BUSINESS SECTOR

Davinder Pal Singh
Technical Architect
Salesforce
Toronto, Canada
Davinder.sfdc@gmail.com

Abstract

In the modern digital age, the conventional approach to business analysis has been altered as a result of advancements in machine learning. As marketplaces evolve, customers and businesses need to properly predict future markets and behaviour to achieve sustainable success. These advanced technologies have revolutionised how organisations carry out data analysis, gain knowledge, and come to effective decisions. One of the recent trends is a popularity of predictive analytics as part of business analytics. Other related abilities that several of the algorithms share include the capacity to analyse vast databases of earlier occurrences in requests to identify patterns and trends, which aid an organisation in generating accurate prognoses relating to future activities. The use of big data analytics methods to enhance retail sales forecasting has gained popularity in the last several years. Using big data analytics, this article compares and contrasts several ML methods for forecasting sales at supermarkets. The 2013 BigMart Sales dataset is utilised in this study, which explores the use of ML algorithms to predict retail sales patterns. A variety of thorough preprocessing techniques were used, such as PCA feature extraction, outlier detection, and handling of missing variables. F1-score, recall, accuracy, and precision metrics were utilised to assess a variety of classification models, including XGBoost, GLM, Decision Tree, and KNN. With an accuracy of 84.7%, the results show that KNN performed best, demonstrating its potency in forecasting sales patterns.

Index Terms – Component, Data Analytics, Sales prediction, Machine learning, KNN, Decision tree, XGBoost, Generalized Linear Model, Dimensionality reduction.

I. INTRODUCTION

Various types of shopping centres, including supermarkets and retail shops, keep records of the products and things sold, including information on the customers and their dependent and independent traits and attributes[1][2]. These records also include information about certain assets. Every market tries to give personalised and limited-time deals in an effort to attract more customers in a short amount of time. Thus, the collected data may potentially be used to predict future sales using ML algorithms. Numerous problems face an organisation when it lacks an accurate sales forecast model. Retailers, distributors, and manufacturers may all benefit from sales forecasts. Long-term projections facilitate business expansion, while short-term forecasts help with inventory control and production scheduling. In industries where goods have a limited shelf life, sales forecasting is essential to reduce revenue loss during periods of excess and scarcity [3][4].

International Journal of Core Engineering & Management
Volume-7, Issue-06, 2023 ISSN No: 2348-9510

To assist supermarkets, make data-driven choices and optimise their operations, data analytics is used in supermarkets to analyse customer data and forecast future sales patterns. Supermarkets may generate targeted marketing strategies that improve customer loyalty and boost sales income by analysing consumer data to target certain customer groupings[5][6].

A growing number of businesses are turning to big data analytics to make sense of the mountain of data that is available from places like social media, customer loyalty programmes, and point-of-sale systems[7]. Supermarkets may use big data analytics to examine this information and learn about consumer trends, preferences, and behaviour. Supermarkets may find useful insights, patterns, and sales trends for the future by applying ML algorithms to the data and creating predictive models[8].

Business and industry trends may also be analysed on a time scale with the use of data mining and big data. It also brings together statistical approaches, programming processes, ML algorithms, and data engineering. AI calls for and greatly benefits from solid background in mathematics, statistics, information science, and computer science. A branch of "Computer Science" (or "Artificial Intelligence"), "machine learning" is the study of algorithms derived from stochastic theory that can effectively carry out tasks in the absence of explicit programme instructions by drawing conclusions and capitalising on patterns [9][10]. The mathematical model of the data used for training purposes, sometimes referred to as "Training Data," is constructed using ML techniques in order to extract conclusions and assumptions from the data [11]. Supermarket sales datasets may be used as a springboard for new insights into supervised and unsupervised issue types in ML, with the former often serving as a source for classification-type challenges[12].

The motivation behind this study is to address the important problem of precisely forecasting sales trends in retail settings, with a focus on the extensive 2013 BigMart Sales dataset. For supermarkets and other retail establishments, accurate sales forecasting is essential to strategic planning, inventory control, and overall operational efficiency. With the use of cutting-edge machine learning methods like K-Nearest Neighbours (KNN), Generalized Linear Models (GLM), Decision Trees, and XGBoost, this study attempts to create reliable predictive models that can manage intricate sales dynamics. The goal of this research is to offer practical implications that would help retail decision-makers increase the retail industry's profitability and customer satisfaction in the current highly competitive environment. This is achieved through rigorous preprocessing, feature extraction, and model evaluation employing metrics like recall, F1-score, accuracy, and precision.

The key contributions of the study on forecasting sales using the BigMart Sales dataset:

- This study introduces a comprehensive preprocessing including improved handling of missing values through average weight imputation and mode-based filling for outlet size, along with outlier detection and removal, ensured a cleaner dataset for analysis.
- Utilisation of feature extraction techniques such as PCA for dimensionality reduction and creation of new features helped in capturing essential information and optimising model performance.
- Comprehensive evaluation of multiple classification models (XGBoost, GLM, Decision Tree, KNN) provided insights into their effectiveness for sales prediction, highlighting KNN as the top performer.
- Visual representations like correlation matrices and confusion matrices offered clear insights into relationships between variables and model performance, aiding in better understanding and interpretation of results.

- Rigorous evaluation employing metrics like F1-score, recall, accuracy, and precision provided a robust assessment of model performance, enabling informed decision-making for deployment in real-world scenarios.

The paper is organised as follows. Section 2, present a background study of sales prediction using BigMart sales dataset, Section 3 offers methods and methodology, and Section 4 results analysis and discussion. Conclusion and future work of this study present in section 5.

II. LITERATURE REVIEW

In this section, provide some previous work on big data analytics based on machine learning. The field of sale forecasting using ML has seen a lot of work presented so far. An overview of relevant sales forecasting studies is given in this section. Table 1 shows the results of a comparison study using ML and big data analytics in the corporate world.

In, Niu, (2020) approach may efficiently mine properties across several dimensions to provide accurate predictions. The XGBoost sale prediction model is tested in this work using datasets given by the Kaggle competition and sales data from Walmart stores. The experimental findings demonstrate that the strategy outperforms the other ML algorithms. This paper's RMSSE measure is 0.141 times lower than the LR method and 0.113 times lower than the Ridge technique[13].

In, Jiang, Ruan and Sun, (2021) investigates the practicability of several models for Walmart sales forecasting, including ML-only, hybrid, and conventional time series models. Predictions and empirical analyses should be based on data collected between 2016-06-19 and 2016-08-14. The Prophet model and the ML model, lightGBM model, are used for training and testing. The Prophet model breaks down trends, seasons, and holidays. Sales figures are derived from Walmart supermarkets and cover the period from 2011-01-29 to 2016-06-19. The results demonstrate that the ML model accurately forecasts sales at retail stores; the LightLGB model and the Prophet model both have RMSEs of 0.617 and 0.694, respectively[14].

In, Sun et al., (2021) Lines are drawn between the spots in accordance with the time sequence. Second, for every route map, they determine its fractal dimension using the box-counting approach. Lastly, in order to enhance the KNN approach, they took the similarity of distributions and Gaussian functions into account. Our upgraded KNN customer categorisation model outperforms SVM and classic KNN models with a better F1-score of 0.926 and an accuracy greater at 0.925[15].

In Zhao et al. (2021) categorise customers according to their purchase history: high volume, low volume, and no volume. This is done using point-of-sale data. Using an LSTM-NN, the experiment establishes a model for consumer behaviour categorisation. When compared to LR and SVM, LSTM-NN achieves a greater recognition accuracy of 5.26% and 6.97%, respectively, according to numerical testing[16].

International Journal of Core Engineering & Management
Volume-7, Issue-06, 2023 ISSN No: 2348-9510

TABLE I. COMPARATIVE STUDY ON BIG DATA ANALYTICS AND MACHINE LEARNING TECHNIQUES IN BUSINESS SECTOR

References	Methodology	Dataset	Performance	Limitations & Future Work
[5]	Recurrent Neural Network (RNN)	Sales data of 800 products over 49 weeks	RMSE: 0.039	Further tuning of RNN parameters; exploration of other neural network architectures.
[6]	XGBoost	Walmart supermarkets sales data from Kaggle competition	RMSSE: 0.141 (Logistic Regression), 0.113 (Ridge)	Exploration of additional features; comparison with other advanced machine learning models.
[7]	Prophet model and LightGBM	Walmart supermarket sales data	RMSE: 0.694 (Prophet), 0.617 (LightGBM)	Combination of more models for better performance; further validation on different datasets.
[8]	Improved KNN using fractal dimension	Customer classification based on fractal dimension and stay time	Accuracy: 0.925, F1-score: 0.926	Application to larger and more diverse datasets; integration with other classification techniques.
[9]	Long Short-Term Memory Neural Network	POS data with consumer behaviour variables	Improved accuracy: 5.26% (Logistic Regression), 6.97% (SVM)	Refinement of the LSTM model; exploration of additional behavioural and contextual variables for classification.

A. Research gaps

Despite advancements in various sales forecasting models like RNNs, XGBoost, Prophet, LightGBM, and improved KNN algorithms, a significant research gap exists in integrating these models to create hybrid approaches for more robust predictions. Current studies often focus on individual models or comparisons, lacking exploration into combined models that leverage strengths and mitigate weaknesses. Moreover, there is limited research on the interpretability of these models, crucial for practical applications in retail and marketing. Additionally, while studies use datasets from Walmart and market products, more diverse datasets across various market segments and regions are needed to validate model generalizability. Closing these gaps could enhance reliability and applicability in sales forecasting.

III. MATERIALS AND METHODS

The proposed system aims, among other things, to maximise revenue by locating a dependable method for predicting sales trends via the use of ML. In order to accomplish the organisation's goal, these models may be utilised in many domains and taught to meet expectations. Using a 2013 BigMart Sales dataset, the work's goal is to forecast sales. As a result, the given dataset was split into subgroups for fitting and validation for enhancing analysis efficiency. During data preprocessing, to handle missing values in the 'Item Weight column, the average weight of the item was considered to be implanted for the absent values while to fill the missing values in the 'Outlet Size' column, the mode according to the outlet type was considered. To further improvement of the model, outliers were identified and later corrected. Some measures were taken to expand the quality of the dataset, such as feature creation and feature reduction through the help of PCA. Following that, the data was split into a testing set and a training set. For sales prediction classification models applied include XGBoost, KNN, Generalized Linear Model (GLM), Decision Tree, etc. These models were evaluated using recall, accuracy, precision and F1-score. The comparison investigation shed light on how well various BigMart sales forecast methods performed.

The proposed research recommended the following procedures for predicting sales of different categories employing a retail store's sales data. A proposed system's architectural diagram is shown in Figure 1. This is a detailed overview of all the processes involved.

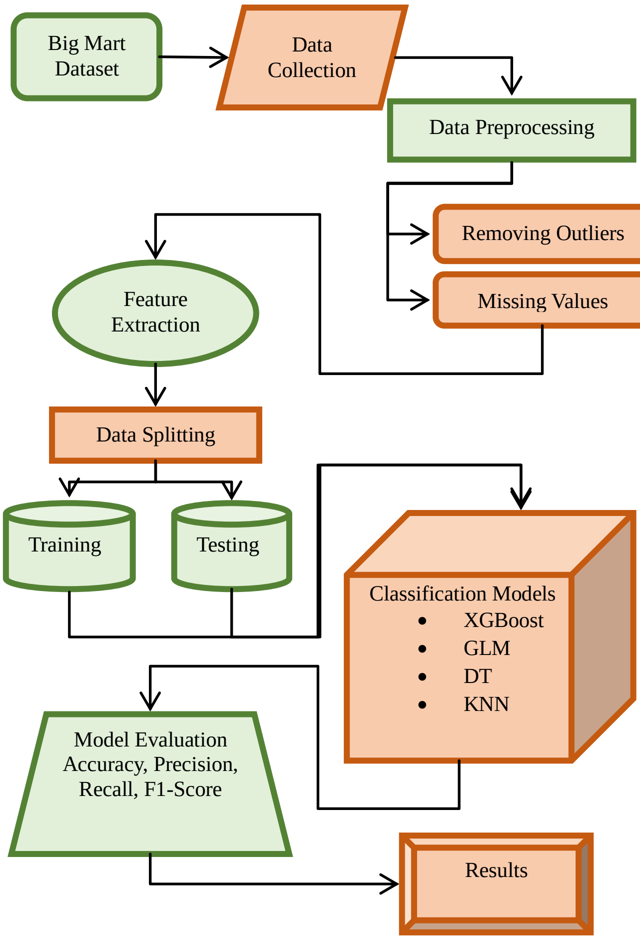

Fig. 1. Flowchart of proposed methodology for BigMart Sales dataset.

A. Data Collection

This comparative study uses of the BigMart Sales dataset, which was gathered in 2013. The input dataset is divided into 2 subsets and provided valuable insights for analysis. The BigMart Sales dataset, gathered in 2013 is utilised to predict consumer behaviour. There are two subsets of this dataset: a test set and a training set. There are 5681 records with 11 attributes in the test set and 8523 records with 12 attributes in the training set, as shown in Table 2. Both independent and dependent variables are present in the training set. The attributes are described in full below:

TABLE II. ATTRIBUTES AND VARIABLES OF THE DATASET

Item Identifier	**Product ID**
Item Weight	Weight of Product
Item Fat Content	Fat content of Product- Low/Regular
Item Visibility	Parameter to know the visibility/reach of product
Item Type	Category of Product
Item MRP	Maximum Retail Price of the Product.
Outlet Identifier	Store ID
Outlet Establishment Year	The Year in which store is established
Outlet Size	Area-wise distribution of Stores- Low/Medium/High
Outlet_Location_Type	Type of city in which outlet is located
Outlet Type	Type of outlet Grocery store or supermarket
Item Outlet Sales	Sale price of product - The dependent variable to be predicted

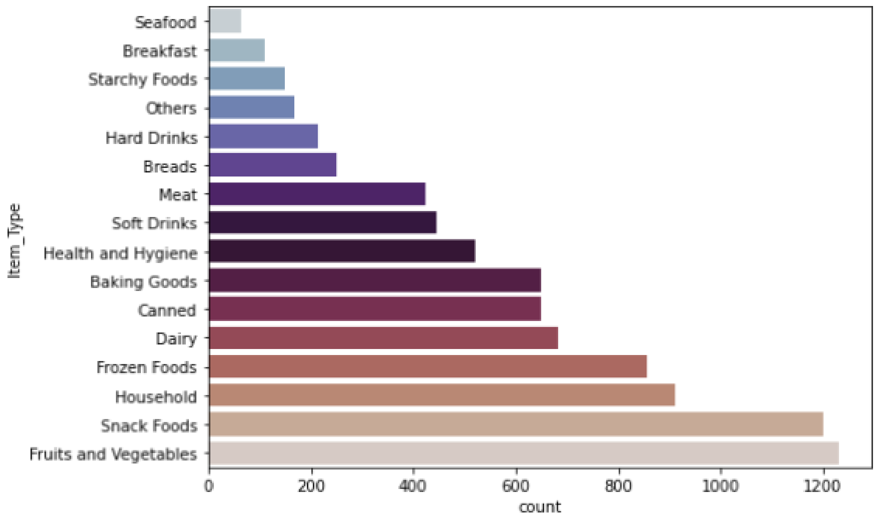

Fig. 2. Sale Prediction of over all products

Figure 2 presents the sales forecast for each item in an understandable manner. From the data, it can be concluded that people are more likely to buy fruits and veggies than other goods. Moreover, almost similar quantities of snack food are bought.

International Journal of Core Engineering & Management
Volume-7, Issue-06, 2023 ISSN No: 2348-9510

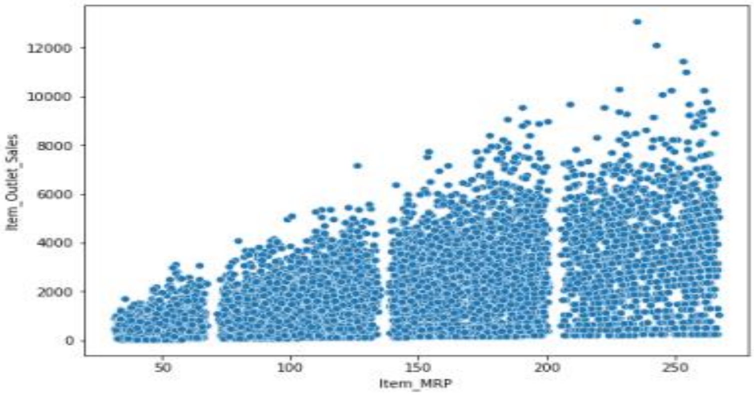

Fig. 3. Sale depends on Item MRP

Figure 3 shows the scatter plot of "Item_MRP" (Maximum Retail Price) and "Item_Outlet_Sales." Each point is an observation, with the x-axis displaying Item_MRP and the y-axis showing Item_Outlet_Sales. The plot illustrates that Item_Outlet_Sales rise with Item_MRP. Four vertical bands cluster the data points, suggesting four Item_MRP ranges or categories. Higher Item_MRP increases point density and band spread, indicating a wider sales fluctuation for higher-priced items.

B. Data Preprocessing

The acquired data underwent preprocessing and cleaning to eradicate any potential flaws, such as missing values or outliers, which might impact the accuracy of the ML models. Preprocessing refers to the process of cleaning a dataset of any excessive or irrelevant data. This step often handles outliers in the dataset and imputes missing values. The dataset has no information for the columns labelled "Outlet Size" and "Item Weight." Item weight may be thought of as a numerical variable, while outlet size is categorical. To fill in the blanks when data are absent, we take the sample weight as a whole and impute it as Item Weight. Since Outlet Size is not a continuous variable, we must rely on the mode approach to fill in the blanks since we cannot compute the average. Therefore, by figuring out the size mode based on outlet type, the missing digits in the outlet size may be discovered.

- **Missing Values:** The absence of a single data point for a particular variable in a dataset is known as a missing value. Many other symbols, such "NA" or "unknown," or empty cells may stand in for them. The absence of these data points makes data analysis more difficult and increases the risk of biassed or incorrect conclusions.
- **Removing Outliers:** Datasets may sometimes include out-of-the-ordinary, out-of-range numbers that stand in stark contrast to the rest of the data. Identifying and eliminating abnormal values, which are known as outliers, may often lead to better model skill and machine learning modelling in general.

C. Feature Extraction

Feature extraction involves sorting all data into categories in order to extract the most important and relevant information [17]. It is critical to get all the required information or minimise the loss of pertinent data while dealing with a big dataset. The data loss rate may be reduced by the use of

International Journal of Core Engineering & Management
Volume-7, Issue-06, 2023 ISSN No: 2348-9510

feature extraction, which helps manage the vital information out of enormous raw datasets.

- **Creating new features:** A vital step in improving the efficiency of ML algorithms by obtaining relevant data is the process of extracting new features from old ones.
- **Dimensionality reduction:** To reduce the number of features while preserving critical information, one may use approaches like PCA or feature selection.

D. Data Splitting

Data separation is a common step for training and testing the model. In this study, dataset is split into two parts. The two parts are training and testing.

E. Machine learning Classification Models

For the comparative analysis some classification models for BigMart sales dataset are described in this section.

1). XGBoost Model

XGBoost is a prominent ML algorithm that consistently ranks among the top in terms of accuracy. It is widely used for both regression and classification prediction tasks. In ML, it's an application of gradient-boosted DT made for speed and performance.

2). Generalised Linear Model

Data sets may be used for prediction using the generalised linear model (GLM). An enlarged version of a linear regression equation is one example of a general linear model. The GLM is an enhanced version of linear regression that establishes a connection between the model and the response variable via a link function. As a result, the variance of every measurement is determined by its predicted value.

$$I.mdl = stepwise(tbl) \quad (1)$$

An intelligent stepwise GLM of a dataset array (tbl) is built using a constant model as a starting point, and predictors are added or removed using stepwise regression. The final variable of the table is used as the response variable by stepwise. Stepwise uses both forward and backward stepwise regression to get a final model.

3). Decision Tree

The goal of decision tree learning is to build a DT that represents f or a near approximation of it from a collection of (x, f(x)) pairings. While it is theoretically possible for the set of pairs to be exhaustive when domain x is finite, in practice, sets are typically samples from domain X that could be limitless. If that's the case, one possibility is to seek for a tree that approximates f throughout the whole domain, instead of just on the data set.

4). K-Nearest Neighbour

Unsupervised ML algorithms include the KNN method. Simply said, it's the algorithm that uses the similarity principle to assign a label from a predetermined set of labels to an unlabelled item, or it sorts the new data point into one of the preexisting categories. To illustrate the point, the k-NN approach may be used to train a model that can distinguish between square and circular images. In the event that you provide it with an unclassified image, it will automatically assign it to the square or circle class.

The class of the new data point may be determined by taking a majority vote among surrounding

International Journal of Core Engineering & Management
Volume-7, Issue-06, 2023 ISSN No: 2348-9510

data points. Choosing how many neighbours to use for categorisation is a manual process. In our study, we have used the Euclidean distance to assess the similarity. Equation (2) provides the formula for calculating the Euclidean distance[18]:

$$EuclideanDisatance(x, x_i)\sqrt{(\sum x_j - x_{ij})^2} \quad (2)$$

A new data point is assigned to one of the predefined classes by a majority vote of its KNN, chosen according to the Euclidean distance computed using the aforementioned equation.

IV. RESULTS AND DISCUSSION

This section presents the outcomes obtained through the evaluation of a dataset utilised in this research, including the outcomes, description of a dataset, performance metrics, and classifier statistics. For evaluating the performance of each model, a total of four various performance metrics has been used: F1-score, recall, accuracy, and precision. These parameters are given below:
Accuracy
A measure of accuracy is determined by dividing the total number of correct forecasts by the total number of predicted values, which includes the true predictions themselves. The corresponding equation (3) is shown below:

$$Acuracy = \frac{TP+TN}{TP+TN+FN+FP} \quad (3)$$

The findings were analysed using well-respected academic performance metrics that centre on the confusion matrix. The matrix's visual is shown in Figure 4. The four main features of the matrix display the outcome data, while the matrix itself is an amalgamation of categorisation results. A true positive (TP) result is one in which the actual value matches the anticipated value of the classification. True negative (TN) principles are similar, only they centre on zero. The outcome is referred to as a false positive (FP) when the prediction is 1 and the real value is 0, while the converse is termed a false negative (FN).
The experiment results of the machine and machine learning models for BigMart sales dataset are provided in this section. Graphs, tables, and figures make up the findings.

International Journal of Core Engineering & Management
Volume-7, Issue-06, 2023 ISSN No: 2348-9510

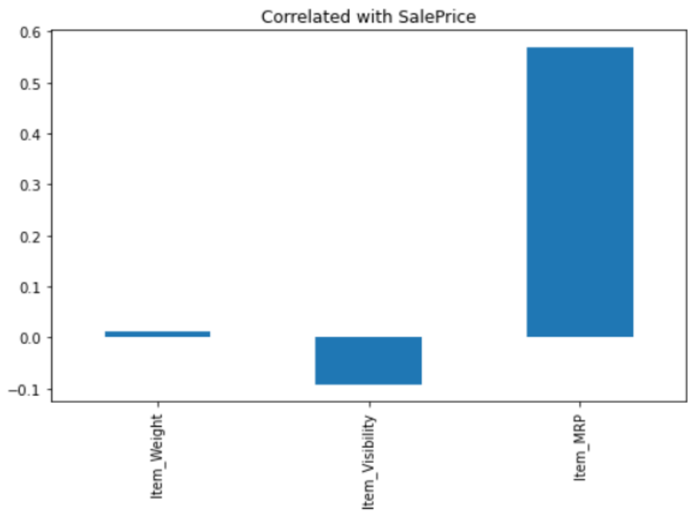

Fig. 4. Correlation Matrix of K-NN on sale prediction

The above Figure 5 demonstrates that Item_MRP has a large positive correlation with Sale Price in the Correlation Matrix, whereas Item Weight and Item Visibility have weaker associations. Item Weight has a slight positive correlation while Item Visibility is negatively correlated.

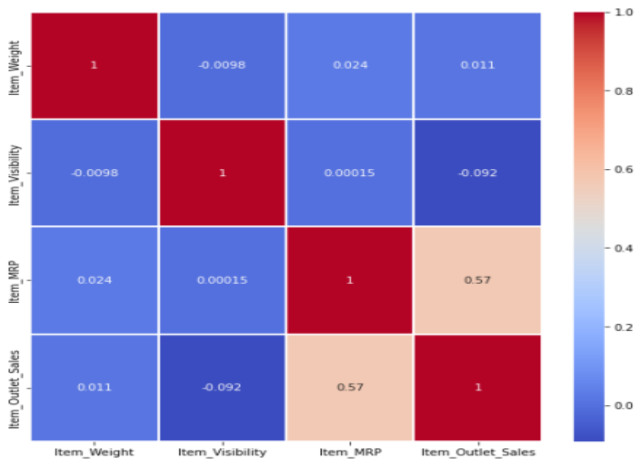

Fig. 5. Confusion Matrix

Figure 6 shows the confusion matrix of correlation between Item Weight, Item Visibility, Item_MRP, and `Item_Outlet_Sales displayed in this heatmap. There is a somewhat favourable association (0.57) among Item_MRP and Item_Outlet_Sales. There is little link among Item Weight and Item Visibility and other attributes. The correlation strength is shown by the colour intensity. The following Table 3 shows a comparison among various machine learning models for comparative analysis and BigMart sales in terms of performance metrics.

International Journal of Core Engineering & Management
Volume-7, Issue-06, 2023 ISSN No: 2348-9510

TABLE III. COMPARISON BETWEEN VARIOUS MODEL FOR BIGMART SALE PREDICTION

Models	Accuracy
XGBoost [19]	61.14
GLM[20]	56.03
Decision Tree [21]	62.0
KNN	84.7

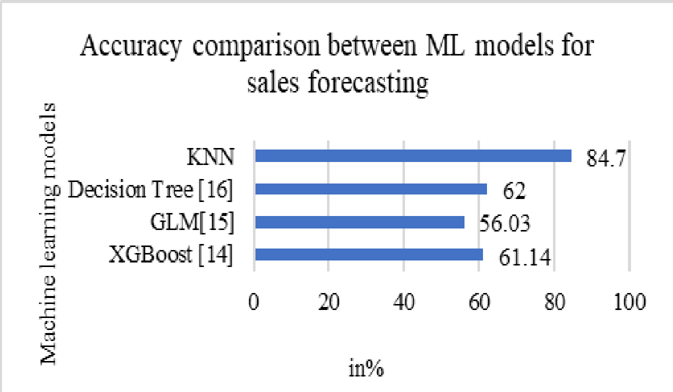

Fig. 6. Bar Graph for Accuracy comparison of models

The bar graph in Figure 7 illustrates the accuracy comparison between various models used for BigMart sale prediction. Among the models, KNN demonstrates the highest accuracy at 84.7%, significantly outperforming the other models. This suggests that KNN is particularly effective for this dataset, possibly due to its ability to capture complex relationships in the data through its instance-based learning approach. In contrast, XGBoost shows the lowest accuracy at 61.14%, indicating that, despite its reputation for strong performance in many scenarios, it may not be as well-suited for this particular task. The GLM and Decision Tree models have moderate accuracy scores of 56.03% and 62.0% respectively, highlighting a substantial gap between these traditional models and KNN in terms of predictive accuracy for BigMart sales.

V. CONCLUSION AND FUTURE SCOPE

This project provides an understanding of ML and its critical ideas alongside the data processing and modelling techniques used in this field. Forecasting of sale at these different stores of Big Mart by using these methods is the focus here. Since the dataset gathered from BigMart Sales included extensive records, refining processes strengthened data credibility necessary for prediction. KNN proved to be the best performing model and therefore indicated that KNN could be used to predict sales trends with very minimal error margin of 84. 7%. The study brings out the importance of improving accuracy in forecasting in determining the position of inventory as well as efficiency in retail operations. Continuing the research of hybrid model approaches and improving interpretability could lead to improving the predictive performance in the short-term and achieving strategic agility for retailers to predict and plan resource allocation and targeted tactics to improve customer satisfaction in response to changes in the market environment.

The usage of such algorithms such as DL and Transfer Learning can be most probably seen in near future. The future work includes filling the gaps that have been pointed out in the study, for instance, using the combination of the early and late stages of machine learning models. Thus, there remains potential for further research on bringing interpretability methods to these models, and these findings can benefit retail and marketing applications. Extension of these models to other markets and to other regions would also prove the scope of generalisation of these models to increase reliability about sales.

REFERENCES

1. A. H. Ali, M. Z. Abdullah, S. N. Abdul-Wahab, and M. Alsajri, "A Brief Review of Big Data Analytics Based on Machine Learning," *Iraqi J. Comput. Sci. Math.*, 2020, doi: 10.52866/ijcsm.2020.01.02.002.
2. S. J. Isabella and S. Srinivasan, "An understanding of machine learning techniques in big data analytics: A survey," *Int. J. Eng. Technol.*, 2018, doi: 10.14419/ijet.v7i3.12.16450.
3. R. Odegua, "Applied Machine Learning for Supermarket Sales Prediction," no. January, 2022.
4. J. Thomas, "The Effect and Challenges of the Internet of Things (IoT) on the Management of Supply Chains," *Int. J. Res. Anal. Rev.*, vol. 8, no. 3, pp. 874–878, 2021.
5. V. Rohilla, S. Chakraborty, and R. Kumar, "Car Auomation Simulator Using Machine Learning," *SSRN Electron. J.*, 2020, doi: 10.2139/ssrn.3566915.
6. P. Khare, "The Impact of AI on Product Management : A Systematic Review and Future Trends," vol. 9, no. 4, 2022.
7. A. Rath, A. Das Gupta, V. Rohilla, A. Balyan, and S. Mann, "Intelligent Smart Waste Management Using Regression Analysis: An Empirical Study," in *Communications in Computer and Information Science*, 2022. doi: 10.1007/978-3-031-07012-9_12.
8. R. K. Vinita Rohilla, Sudeshna Chakraborty, "Random Forest with harmony search optimisation for location based advertising," *Int J Innov Technol Explor Eng*, vol. 8, no. 9, pp. 1092–1097, 2019.
9. S. Mann, A. Balyan, V. Rohilla, D. Gupta, Z. Gupta, and A. W. Rahmani, "Artificial Intelligence-based Blockchain Technology for Skin Cancer Investigation Complemented with Dietary Assessment and Recommendation using Correlation Analysis in Elder Individuals," *Journal of Food Quality*. 2022. doi: 10.1155/2022/3958596.
10. V. Rohilla, S. Chakraborty, and M. Kaur, "Artificial Intelligence and Metaheuristic-Based Location-Based Advertising," *Sci. Program.*, 2022, doi: 10.1155/2022/7518823.
11. A. Nallamekala, S. Vanukuri, and O. Prakash, "Data Science and Machine Learning Approach to Improve Online Grocery Store Sales Performance," no. May, pp. 227–233, 2022.
12. V. Rohilla, M. Kaur, and S. Chakraborty, "An Empirical Framework for Recommendation-based Location Services Using Deep Learning," *Eng. Technol. Appl. Sci. Res.*, 2022, doi: 10.48084/etasr.5126.
13. Y. Niu, "Walmart Sales Forecasting using XGBoost algorithm and Feature engineering," in *Proceedings - 2020 International Conference on Big Data and Artificial Intelligence and Software Engineering, ICBASE 2020*, 2020. doi: 10.1109/ICBASE51474.2020.00103.
14. H. Jiang, J. Ruan, and J. Sun, "Application of Machine Learning Model and Hybrid Model

in Retail Sales Forecast," in *2021 IEEE 6th International Conference on Big Data Analytics, ICBDA 2021*, 2021. doi: 10.1109/ICBDA51983.2021.9403224.
15. F. Sun, L. Zhao, Y. Zuo, and Y. Kaneko, "Application of Fractal Analysis for Customer Classification Based on Path Data," in *IEEE International Conference on Data Mining Workshops, ICDMW*, 2021. doi: 10.1109/ICDMW53433.2021.00040.
16. L. Zhao, Y. Zuo, K. Yada, and M. Liu, "Application of Long Short-term Memory Based Neural Network for Classification of Customer Behavior," in *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics*, 2021. doi: 10.1109/SMC52423.2021.9658703.
17. J. R. Vergara and P. A. Estévez, "A review of feature selection methods based on mutual information," *Neural Computing and Applications*. 2014. doi: 10.1007/s00521-013-1368-0.
18. H. Jégou, M. Douze, and C. Schmid, "Product quantisation for nearest neighbor search," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2011, doi: 10.1109/TPAMI.2010.57.
19. V. Chitre, "Big Mart Sales Analysis," *Int. J. Innov. Technol. Explor. Eng.*, 2022, doi: 10.35940/ijitee.c9833.0411522.
20. *et al.*, "Machine Learning Approach for Big-Mart Sales Prediction Framework," *Int. J. Innov. Technol. Explor. Eng.*, 2022, doi: 10.35940/ijitee.f9916.0511622.
21. M. April *et al.*, "Supermarket Sales Prediction Using Regression," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 10, no. 2, pp. 1153–1157, 2021, doi: 10.30534/ijatcse/2021/951022021.