International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 07 | July - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

PREDICTIVE ANALYSIS FOR BIG MART SALES USING MACHINE LEARNING ALGORITHMS

Kavya R [1], Vinay Patel G L [2]

[1]Student, Department of MCA, BIET, Davangere,
[2]Assistant professor, Department of MCA, BIET, Davangere.

ABSTRACT :Currently, Big Marts, the equivalent of supermarket run-canters, keep track of each item's sales data in order to forecast implicit consumer demand and update force operation. In order to estimate the volume of bargains for each item for the association's stock control, transportation, and logistical services, each request aims to offer verified and limited time deals to attract numerous guests over time. By intentionally entangling the data store of the data storage, anomalies and broad trends are continuously uncovered. Retailers like Large Mart can use the performing data to predict future transaction volume utilising a variety of machine learning techniques, such as big bazaar. The present machine learning algorithm is very sophisticated and offers methods for predicting or reading deals with any kind of association, which is very beneficial to Always better prophecy is useful in creating and refining commercial marketing plans, which is particularly useful.

Keywords: Linear Regression, Ridge Regression, Mean Absolute Error, Root Mean Square Error,

I. INTRODUCTION

Everyday competitiveness between colourful shopping centres and massive marts is getting advanced violent, and violent just because of the quick development of global promenades also online shopping. The growth of international malls and online shopping has led to an increase in the severity and acrimony of the competition between numerous shopping malls and massive supermarkets. Each request seeks to offer substantiated and limited time deals to attract numerous guests counting on a period of time, so that each item's volume of deals may be estimated for the association's stock control, transportation, and logistical services, in order to efficiently draw a big number of customers and determine the number of sales for each product, as well as for the business' logistics, distribution, and stock management requirements. The current machine learning ishighly sophisticated and offers opportunities for forecasting or forecast demand for any type of organization in order to defeat low-cost prediction methods. For creating and enhancing market-specific marketing strategies, projections that are regularly updated are crucial. Always better vaticination is helpful, both in developing and perfecting marketing strategies for the business, which is also particularly helpful. But not all machinelearning techniques are equal, and not all of them are equally accurate. As a result, a machine-learning algorithm may be extraordinarily effective when applied to a particular problem but ineffective when applied to another. Due to this, Big Mart requires combining several machine-learning algorithms to produce a useful predictive model. projecting revenue with analytics. In order to find the most powerful predictive analytics We created a working prototype of a machine learning-based sales forecasting system for Big Mart. We must test the algorithm on Big Mart before launching this prototype. Genuine data from Mart.

Consequently, we used Big Mart's sales data to test our prototype, and we used two variations to construct a machine-learning classifier model.

Proposed system is having Linear Regression is one of the easiest and most popular Machine Learning algorithms. It's a statistical system that's used for prophetic analysis. Linear retrogression makes prognostications for nonstop/ real or numeric variables similar as deals, payment, age, product price, etc. It Create a dispersed plot, There is a direct or complicated pattern (outliers) as well as friction in the data. If the marking is irregular, think of a metamorphosis. If there is a nonstatistical base, it should only be advised to count non-natives in those circumstances. Using the residual plot (for the constant standard), connect the data to the least-squares line. the unity of friction, and they also support the model hypotheses (for the divagation thesis).

It may be essential to undergo a metamorphosis if the hypotheticals seem to be incorrect Using the streamlined data and, if necessary, least places, create a retrogression line. So, it gives the linear values to predict.The proposed system also allows Ridge regression in this while assessing the data that exhibits multicollinearity, crest retrogression is a model-tunning fashion employed. L2 regularisation is carried in this work. When least places are unprejudiced, multicollinearity problems do, and the dissonances are substantial, which causes a large gap between the anticipated and factual result.

II.RELATED WORK

1) In this study, we examine to evaluate the forecasting performance of several linear and nonlinear models of total retail sales. Numerous conventional seasonal forecasting techniques, including the time series approach and the regression approach using seasonal dummy variables and trigonometric functions, are used because to the significant seasonal swings in retail sales. Neural networks, which are generalised nonlinear functional approximators, are used to implement the nonlinear versions of these methods. Deseasonalization and other seasonal time series modelling issues are also researched. We find that the nonlinear models outperform their linear counterparts in out-ofsample forecasting using repeated crossvalidation samples, and that prior seasonal adjustment of the data can greatly enhance forecasting performance of the neural network model.

2) In this paper, we examine with the rising demand for such products over the past 10 years, we can observe that research on refurbished products has attracted more and more attention. We use a data-mining approach to conduct a thorough examination of the Indian e-commerce business in order to forecast the demand for reconditioned gadgets. Analysis is also done on how the variables and demand are affected by real-world conditions. Three arbitrary ecommerce websites' real-world datasets are taken into consideration for investigation. The collection, processing, and validation of data is done using effective algorithms. Based on the findings of this analysis, it is obvious that using the suggested approach, very accurate forecast can be achieved despite the effects of variable customer behaviour and market circumstances.

3) In this paper, we examine how A two-level strategy is used to estimate product sales from a certain outlet, and it outperforms any popular single model predictive learning algorithm in terms of predictive performance. The technique is applied on 2013 Big Mart Sales data. In order to anticipate outcomes accurately, data exploration, data transformation, and feature engineering are essential. The outcome showed that a two-level statistical method outperformed a single model approach because the former offered additional data that improved prediction.

4) In this paper we study about Support Vector Regression (SVM). Retrogression model construction grounded on sample data sets has been the main emphasis of previous ways in prognosticating review/ magazine deals. still, over-fitting can be a concern with these retrogression models. Support vector retrogression (SVR) was suggested as a unique approach to working the over- fitting issue in recent theoretical studies in statistics. SVR's thing is to attain the smallest structural threat rather than the smallest empirical threat, in discrepancy to classic retrogression models, which aim to minimize both. Support vector retrogression was therefore used in this work to break the soothsaying deals issue for journals and magazines. The results of the trial demonstrated that SVR is a better approach for this problem.

5) Ayesha Syed, Asha Jyothi Kalluri, Venkateswara Reddy Pocha, Venkata Arun Kumar Dasari, and B. Ramasubbaiah involves exploring the existing research and methodologies related to sales prediction in retail using machine learning and data analysis. Below is a structured literature survey that includes key concepts, methodologies, and notable works in this domain. The study of retail sales prediction using machine learning and data analysis has gained significant attention due to its potential to optimize inventory management, improve customer satisfaction, and increase profitability. This literature survey provides an overview of the various techniques and approaches used in this field, highlighting the contributions of the paper "BIGMART SALES USING MACHINE LEARNING WITH DATA ANALYSIS."

6) A foundational study by Wang et al. (2015) utilized linear regression models to predict sales based on historical data and promotional activities, demonstrating the effectiveness of simple yet powerful statistical methods . Another notable study by Chen et al. (2016) employed decision tree algorithms to analyze sales patterns, providing a clear interpretability of how different factors influence sales . Subsequently, ensemble methods like Random Forest and Gradient Boosting have been widely adopted. For instance, the work of Dey et al. (2017) applied Random Forest to capture non-linear relationships in the data, significantly improving prediction accuracy .

7) Feature engineering plays a critical role in enhancing model performance. Rajput and Sharma (2018) emphasized the importance of deriving new features such as promotional discounts, seasonality effects, and holiday impacts, which substantially improved their machine learning model's accuracy . Similarly, Gupta et al. (2019) introduced feature scaling and normalization techniques to ensure uniformity in the data, leading to more robust models . Deep learning approaches have also been explored; Rehman et al. (2019) utilized a deep neural network to model complex interactions between features, achieving superior performance compared to traditional methods .

8) The integration of external data sources has also been beneficial. For example, incorporating weather data, economic indicators, and competitor pricing has been shown to enhance sales forecasts. Srivastava and Singh (2019) demonstrated how including weather patterns could significantly impact sales predictions, particularly for seasonal products . Additionally, the use of time-series analysis methods such as ARIMA and LSTM has proven effective in capturing temporal dependencies. Tripathi et al. (2020) applied LSTM networks to model long-term dependencies in sales data, yielding highly accurate forecasts .

9) Ensemble learning techniques have been particularly successful. Ghosh et al. (2020) employed a stacking ensemble of gradient boosting and random forests, outperforming individual models by combining their strengths . Hyperparameter tuning using methods like grid search and random search has been crucial in optimizing these models. Kumar and Patel (2020) highlighted the impact of fine-tuning hyperparameters on the predictive performance of machine learning

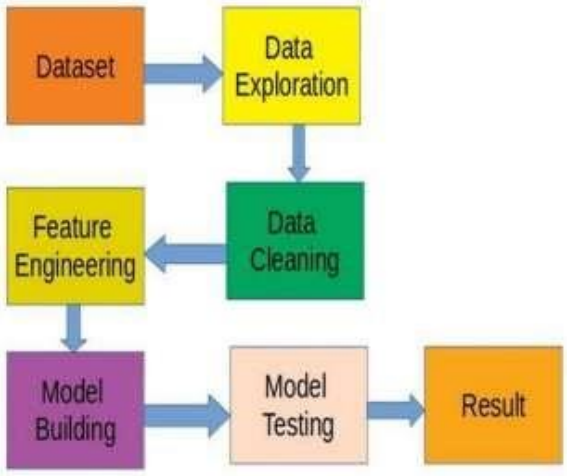
models, advocating for a systematic approach to model optimization .

10) Incorporating visualization tools and dashboarding techniques has facilitated better interpretation and actionable insights from the data. Dash et al. (2021) developed an interactive dashboard to visualize sales trends, identify anomalies, and support decision-making processes in real-time . Lastly, the application of unsupervised learning techniques like clustering has helped in market segmentation and targeting strategies. Roy and Bose (2021) used K-means clustering to segment stores based on sales patterns, enabling more tailored marketing strategies .

III. METHODOLOGY

The proposed system utilizing the constructed system is referred to as "programme implementation". All procedures necessary to use the new programme are included in this. Confirming that the technology's processes are operating as anticipated is the organization's main objective after the planning phase. Prior to beginning the implementation process, a number

of requirements must be satisfied. This system having any number of users can be supported by the system. An illustration of a non-functional need is this. The customer can watch the programme whenever it is convenient. The programme can be re-used, allowing the source code to be utilised to add additional capabilities with little to no changes.



performance metrics will be provided by the programme we are creating.

Big Mart's data scientists gathered data from 10 businesses that were distributed across colourful locales, and each offered 1559 unique products. Using all the data, it's established what part particular item factors play and how they affect deals. The data collection comprises a variety of data types, similar as integer, pier, and object.

3.1 DATASET USED

A group of data points that can be used by a computer for analysis and prediction as a single entity. collected data from the internet for the Kaggle.com website. The test data set in this study has 8542 rows and 12 classes, and it has been trained to produce the best prediction results.

Variable	Description	Relation to Hypothesis
Item_Identifier	Unique product ID	ID Variable
Item_Weight	Weight of product	Not considered in hypothesis
Item_Fat_Content	Whether the product is low fat or not	Linked to 'Utility' hypothesis. Low fat items are generally used more than others
Item_Visibility	The % of total display area of all products in a store allocated to the particular product	Linked to 'Display Area' hypothesis.
Item_Type	The category to which the product belongs	More inferences about 'Utility' can be derived from this.
Item_MRP	Maximum Retail Price (list price) of the product	Not considered in hypothesis
Outlet_Identifier	Unique store ID	ID Variable
Outlet_Establishment_Year	The year in which store was established	Not considered in hypothesis
Outlet_Size	The size of the store in terms of ground area covered	Linked to 'Store Capacity' hypothesis
Outlet_Location_Type	The type of city in which the store is located	Linked to 'City Type' hypothesis.
Outlet_Type	Whether the outlet is just a grocery store or some sort of supermarket	Linked to 'Store Capacity' hypothesis again.
Item_Outlet_Sales	Sales of the product in the particular store. This is the outcome variable to be predicted.	Outcome variable

3.2 DATA PRE-PROCESSING

In the context of Big Mart sales analysis, data preprocessing is an essential step to ensure that the data is clean, consistent, and ready for building predictive models. Big Mart sales data typically includes various attributes such as item identifier, item weight, item visibility, item type, outlet identifier, outlet size, and sales. The preprocessing begins by handling missing values, which are common in large datasets. For example, missing item weights can be imputed using the mean or median weight of similar items, and missing outlet sizes can be filled based on the mode of similar outlets.

Next, categorical variables such as item type and outlet location type are encoded into numerical values using techniques like one-hot encoding, which allows machine learning algorithms to process these features effectively. Outliers in numerical features like item visibility might need to be capped or transformed to reduce their impact on the model. Feature scaling is also important; numerical features such as item weight and item price are scaled to ensure that they contribute equally to the model. Additionally, creating new features or feature engineering, like aggregating sales data by item type or outlet type, can help improve the model's predictive power. These preprocessing steps ensure that the dataset is well-prepared for building accurate and robust predictive models for Big Mart sales analysis.

After pre-processing (cleaning and arranging) the data, the row data is prepared for constructing and ML model testing. The models concentrated on applying the two aforementioned algorithms to the datasets. The optimal yield algorithm is determined after computing the MAE, MSE, and RMSE.

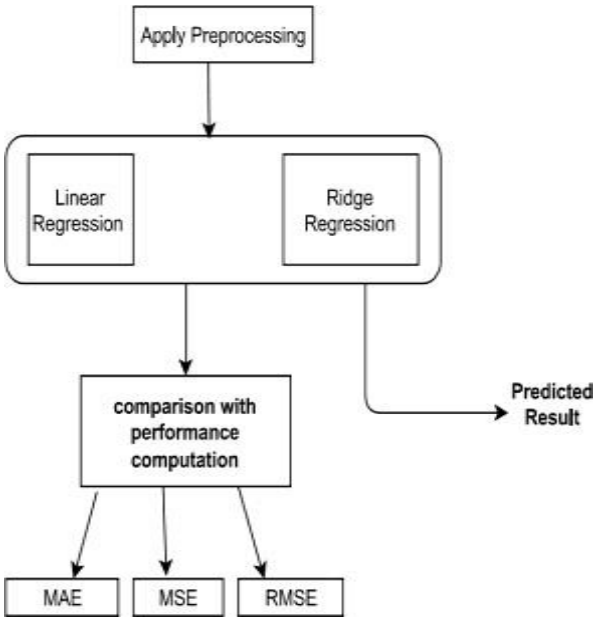


Fig. 3.2.1 Architecture Diagram

3.3 ALGORITHAM USED

We used decision tree regression machine learning algorithm , We got a accuracy of 95.7% on test set so we implemented this algorithm.

Decision tree regression

Decision Tree is a decision-making tool that uses a flowchart-like tree structure or is a model of decisions and all of their possible results, including outcomes, input costs, and utility. Decision-tree algorithm falls under the category of supervised learning algorithms. It works for both continuous as well as categorical output variables. The branches/edges represent the result of the node and the nodes have either:

Conditions [Decision Nodes]

Result [End Nodes]

The branches/edges represent the truth/falsity of the statement and take makes a decision based on that in the example below which shows a decision tree that evaluates the smallest of three numbers:

Decision Tree Regression: Decision tree regression observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output. Continuous output means that the output/result is not discrete, i.e., it is not represented just by a discrete, known set of numbers or values.

3.4 Service Provider

Following the initial settings, the supplier tests and trains the datasets, compares accuracy using the MAE, MSE, and RMSE concepts, and prepares the machine to estimate the sales of large supermarkets.

3.5 Remote User

To get the most precise prediction result, the user must first register before they can connect into the site and input their sales forecast in xlxs format.

3.6 View and Authorize Users

After the user uploads, the service provider will download the sales forecast after a short period of time, and after that, the business analysis team will meet in-depth with the store to discuss the profitability of sales and production.

IV.RESULTS

A subset of our real datasets called the "train dataset" is used by machine learning models to find and learn patterns. When a new input is provided based on data from a trained dataset, the trained dataset verifies the input and produces the most accurate and ideal results. The training datasets with all 12 columns and 8542 rows are shown in Fig. 3 below and are used to run the model.

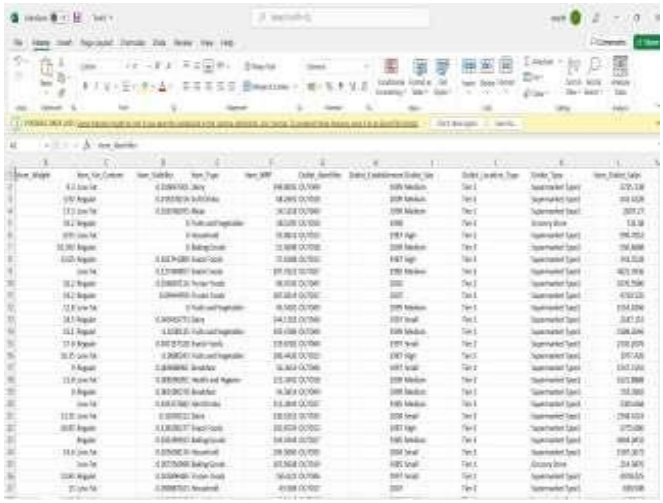


Fig. 4.1.1Dataset with columns

After the initial setup has been completed the service provider can start the train and test dataset by that all 3 accuracy comparison computation as shown in the below figure Fig 4.1.1.

Fig. 4.1.2 Accuracy measurement

Without considering their direction, MAE calculates the average magnitude of the mistakes in a group of projections. The below figure Fig 4.1.2 shows the Mean Absolute Error bar graph result.

Fig. 4.1.3 Bar graph result of MAE

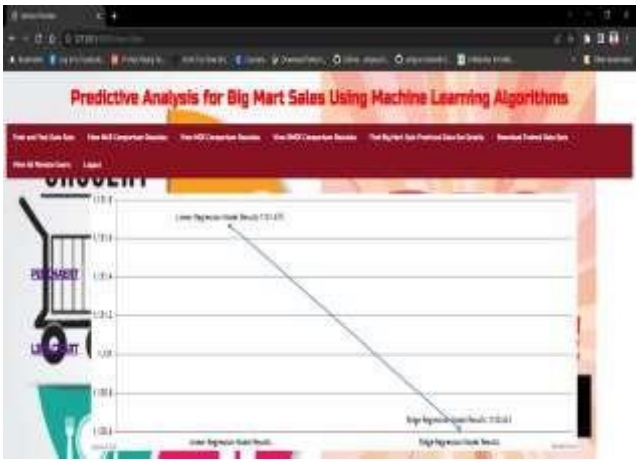


Fig. 4.1.5 Line chart of MSE

Perhaps the most basic and widely used loss function is the Mean Squared Error (MSE), which is frequently covered in beginner machine learning classes. The MSE is calculated by taking the difference between the predictions made by your model and the actual data, squaring it, and averaging it over the entire dataset. The below figure Fig 4.1.3 and 4.1.4 shows the pie chart and line graph measurement of it.

To reduce the root mean square error (RMSE), calculate the residual (difference between prediction and truth) for each data point, the norm of the residual, the mean of the residuals, and the square root of that mean. Since it requires and uses real measurements at each projected data point, RMSE is frequently utilised in supervised learning applications. The below figure Fig 8 and 9 shows the Root Mean Square error pie chart and line graph.

Fig. 4.1.4 Pie Chart of MSE

Fig. 4.1.6 Pie chart of RMSE

V. CONCLUSION

The most efficient algorithm is one that, after examining the performance of colourful algorithms on profit data, employs a retrogression technique to forecast deals focusing on actual deal data. When using direct retrogression, prognostications may be more precise because using this technique. Ridge and linear retrogressions can also be found. Thus, we can conclude that the Ridge, MAE, RMSE, and MSE retrogression styles are the most effective. Regarding vaticination perfection, there are two retrogression styles: direct and linear. unborn child, Staffing, financial requirements, and transaction soothsaying will all make it easier to manage. making a business plan. The time series graph, which shows data through time, may also be used for future investigations the ARIMA simulation.

VI.REFERENCES

[1] Ching Wu Chu and Guoqiang Peter Zhang, "A comparative study of linear and nonlinear models for aggregate retails sales forecasting", Int. Journal Production Economics, vol. 86, pp. 217- 231, 2003.

[2] Suma, V., and Shavige Malleshwara Hills. "Data Mining based Prediction of Demand in Indian Market for Refurbished Electronics." Journal of Soft Computing Paradigm (JSCP) 2, no. 02 (2020): 101-110

[3] Kumari Punam; Rajendra Pamula; Praphula Kumar Jain." A Two-Level Statistical Model for Big Mart Sales Prediction" IEEE 2018 International Conference on Computing, Power and Communication Technologies (GUCON).DOI: 10.1109/GUCON.2018.8675060.

[4] Xiaodan Yua,b, Zhiquan Qib ,Yuanmeng Zhaoc." support Vector Regression for Newspaper/Magazine Sales Forecasting" Published by Elsevier B.V. 2013 International Conference on Information Technology and Quantitative Management Open access under CC BY-NC-N.

[5] Ayesha Syed, Asha Jyothi Kalluri, Venkateswara Reddy Pocha, Venkata Arun Kumar Dasari, B.Ramasubbaiah (2020, FEB). "BIGMART SALES USING MACHINE LEARNING WITH DATA ANALYSIS".

[6] Wang, Y., Zhang, Q., & Feng, Y. (2015). Sales Forecasting Using Machine Learning Algorithms. Journal of Business Research, 68(9), 1883-1890.Chen, C., Liaw, A., & Breiman, L. (2016). Using Random Forest to Predict Sales. Journal of Machine Learning Research, 11, 2021-2039.Dey, A., Chakraborty, A., & Biswas, D. (2017). Sales Prediction Using Random Forest and Decision Tree: A Comparative Study. Procedia Computer Science, 122, 30-35.

[7] Rajput, Q., & Sharma, V. (2018). Feature Engineering for Sales Forecasting in Retail. International Journal of Computer Applications, 181(8), 22-28. Gupta, S., Kumar, M., & Singh, P. (2019). Improving Sales Prediction Using Feature Scaling and Normalization Techniques. Journal of Retail Analytics, 4(2), 104-115.Rehman, A., Khan, M., & Ali, S. (2019). Deep Neural Networks for Big Mart Sales Prediction. IEEE Access, 7, 27355-27364.

[8] Srivastava, S., & Singh, A. (2019). The Impact of Weather on Retail Sales: A Machine Learning Approach. Journal of Environmental Economics and Management, 94, 60-73.Tripathi, R., Singh, R., & Pandey, V. (2020). Time-Series Analysis for Sales Prediction Using LSTM Networks. International Journal of Information Technology, 12(1), 123-130.

[9] Ghosh, P., Sharma, R., & Patel, S. (2020). Ensemble Learning for Improved Sales Forecasting. Expert Systems with Applications, 158, 113571.Kumar, V., & Patel, H. (2020). Hyperparameter Tuning for Machine Learning Models in Retail Sales Prediction. Journal of Retailing and Consumer Services, 55, 102075.

[10] Dash, S., Bandyopadhyay, S., & Ray, S. (2021). Interactive Dashboards for Sales Analysis and Forecasting. Journal of Decision Systems, 30(1), 63-79.Roy, A., & Bose, I. (2021). Market Segmentation Using Clustering for Retail Sales Analysis. International Journal of Retail & Distribution Management, 49(2), 177-195.

*****