International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048
Volume 9, Issue 3, pp 326-331, March-2025
https://doi.org/10.47001/IRJIET/2025.903047

Predictive Analysis for Big Mart Sales Using Machine Learning Algorithms

[1]Alphan Shaikh, [2]Mohd Khalid Shaikh, [3]Ayaan Sayyed, [4]Prof. A.N.Gedam

[1,2,3]Student, Computer Engineering, AISSMS College of Polytechnic, Pune, Maharashtra, India
[4]Asst. Professor, Computer Engineering, AISSMS College of Polytechnic, Pune, Maharashtra, India

***Abstract* - Currently, supermarket run-centres, Big Marts keep track of each individual item's sales data in order to anticipate potential consumer demand and update inventory management. Anomalies and general trends are often discovered by mining the data warehouse's data store. For retailers like Big Mart, the resulting data can be used to forecast future sales volume using various machine learning techniques like big mart. A predictive model was developed using Decision Tree Regression for forecasting the sales of a business such as Big -Mart, and it was discovered that the model outperforms existing models.**

Keywords: Inventory Management, Anomalies, Big Mart, Machine Learning, Decision Tree Regression.

I. INTRODUCTION

The daily competition between different malls as well as big malls is becoming more and more intense because of the rapid rise of international supermarkets and online shopping's. Every mall or mart tries to provide personal and short-term donations or benefits to attract more and more customers on a daily basis, such as the sales price of everything which is usually predicted to be managed through different ways such as corporate asset management, logistics, and transportation service, etc. Current machine learning algorithms that are very complex and provide strategies for predicting or predicting long-term demand for a company's sales, which now also help in overcoming budget and computer programs.

In this paper, we basically discuss the subject of specifying a large mart sale or predicting an item for a customer's future need in a few supermarkets in various locations and products that support the previous record. Various ML algorithms such as linear regression, random forest, etc. are used to predict sales volume. As we know, good marketing is probably the life blood of all organizations, so sales forecasting now plays an important role in any shopping mall. It is always helpful to predict the best, and develop business strategies about useful markets and to improve market knowledge. Regular sales forecasting research can help in-depth analysis of pre-existing conditions and conditions and then, assumptions are often used in terms of customer acquisition, lack of funding, and strength before setting budgets and marketing plans for the coming year.

In other words, sales forecasts are predicted on existing services of the past. In-depth knowledge of the past is required to develop and enhance market opportunities no matter what the circumstances, especially the external environment, which allows to prepare for the future needs of the business. Extensive research is ongoing in the retailer's domain to predict long term sales demand. An important and effective method used to predict the sale of a mathematical method, also called the conventional method, but these methods take more time to predict sales. And these methods could not manage indirect data so to overcome these problems in traditional methods the machine learning techniques used. ML methods can handle not only indirect data but also large data sets well.

II. LITERATURE REVIEW

1) A comparative study of linear and nonlinear models for aggregate retails sales forecasting.
AUTHORS: Ching Wu Chu and Guoqiang Peter Zhang.

The purpose of this paper is to compare the accuracy of various linear and nonlinear models for forecasting aggregate retail sales. Because of the strong seasonal fluctuations observed in the retail sales, several traditional seasonal forecasting methods such as the time series approach and the regression approach with seasonal dummy variables and trigonometric functions are employed. The nonlinear versions of these methods are implemented via neural networks that are generalized nonlinear functional approximators. Issues of seasonal time series modeling such as deseasonalization are also investigated. Using multiple cross-validation samples, we find that the nonlinear models are able to outperform their linear counterparts in out-of-sample forecasting, and prior seasonal adjustment of the data can significantly improve forecasting performance of the neural network model. The overall best model is the neural network built on deseasonalized time series data. While seasonal dummy variables can be useful in developing effective regression models for predicting retail sales, the performance of dummy regression models may not be robust. Furthermore,

ISSN (online): 2581-3048
Volume 9, Issue 3, pp 326-331, March-2025
https://doi.org/10.47001/IRJIET/2025.903047

trigonometric models are not useful in aggregate retail sales forecasting.

2) Sustainable development and management in consumer electronics using soft computation.

AUTHORS: Wang, Haoxiang.

Combination of Green supply chain management, Green product deletion decision and green cradle-to-cradle performance evaluation with Adaptive-Neuro-Fuzzy Inference System (ANFIS) to create a green system. Several factors like design process, client specification, computational intelligence and soft computing are analysed and emphasis is given on solving problems of real domain. In this paper, the consumer electronics and smart systems that produce nonlinear outputs are considered. ANFIS is used for handling these nonlinear outputs and offer sustainable development and management. This system offers decision making considering multiple objectives and optimizing multiple outputs. The system also provides efficient control performance and faster data transfer.

3) Data Mining based Prediction of Demand in Indian Market for Refurbished Electronics.

AUTHORS: Suma, V., and Shavige Malleshwara Hills.

There has been an increasing demand in the e-commerce market for refurbished products across India during the last decade. Despite these demands, there has been very little research done in this domain. The real-world business environment, market factors and varying customer behavior of the online market are often ignored in the conventional statistical models evaluated by existing research work. In this paper, we do an extensive analysis of the Indian e-commerce market using data-mining approach for prediction of demand of refurbished electronics. The impact of the real-world factors on the demand and the variables are also analyzed. Real-world datasets from three random e-commerce websites are considered for analysis. Data accumulation, processing and validation is carried out by means of efficient algorithms. Based on the results of this analysis, it is evident that highly accurate prediction can be made with the proposed approach despite the impacts of varying customer behavior and market factors. The results of analysis are represented graphically and can be used for further analysis of the market and launch of new products.

4) Forecasting Monthly Sales Retail Time Series: A Case Study.

AUTHORS: Giuseppe Nunnari, Valeria Nunnari.

This paper presents a case study concerning the forecasting of monthly retail time series recorded by the US Census Bureau from 1992 to 2016. The modeling problem is tackled in two steps. First, original time series are de-trended by using a moving windows averaging approach. Subsequently, the residual time series are modeled by Non-linear Auto-Regressive (NAR) models, by using both Neuro-Fuzzy and Feed-Forward Neural Networks approaches. The goodness of the forecasting models, is objectively assessed by calculating the bias, the mae and the rmse errors. Finally, the model skill index is calculated considering the traditional persistent model as reference. Results show that there is a convenience in using the proposed approaches, compared to the reference one.

5) Multiple Linear Regression Analysis of the Overlay Accuracy Model Zone.

AUTHORS: Zone-Ching Lin, Wen-Jang Wu.

The multiple linear regression method was used to analyze the overlay accuracy model and study the feasibility of using linear methods to solve parameters of nonlinear overlay equations. The methods of analysis include changing the number of sample points to derive the least sample number required for solving the accurate estimated parameter values. Besides, different high-order lens distortion parameters were ignored, and only the various modes of low-order parameters were regressed to compare their effects on the overlay analysis results. The findings indicate that given a sufficient number of sample points, the usage of multiple linear regression analysis to solve the high-order nonlinear overlay accuracy model containing seventh-order lens distortion parameters is feasible. When the estimated values of low-order overlay distortion parameters are far greater than those of high-order lens distortion parameters, excellent overlay improvement can still be obtained even if the high-order lens distortion parameters are ignored. When the overlay at the four corners of image field obviously exceeds that near the center of image field, it is found, through simulation, that the seventh-order parameters overlay model established in this paper has to be corrected by high-order lens distortion parameters to significantly improve the overlay accuracy.

III. SYSTEM DESIGN

For building a model to predict accurate results the dataset of Big Mart sales undergoes several sequence of steps as mentioned in Figure 1 and in this work we propose a model using Xgboost technique. Every step plays a vital role for building the proposed model. After preprocessing and filling missing values, we used ensemble classifier using Decision trees, Linear regression, Ridge regression, Random forest and Xgboost. Both MAE and RSME are used as accuracy metrics for predicting the sales in Big Mart. From the accuracy metrics it was found that the model will predict best using minimum MAE and RSME.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 9, Issue 3, pp 326-331, March-2025

https://doi.org/10.47001/IRJIET/2025.903047

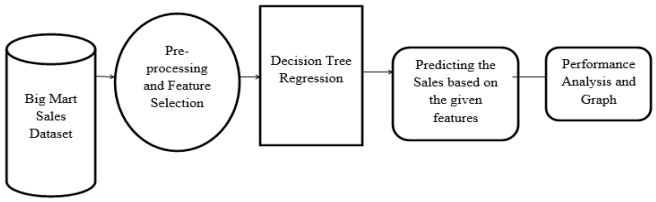

Figure 1: System Architecture

IV. IMPLEMENTATION

Modules:

- Data Collection
- Dataset
- Data Preparation
- Model Selection
- Analyze and Prediction
- Accuracy on test set
- Saving the Trained Model

Modules description:

1) Data Collection:

- This is the first real step towards the real development of a machine learning model, collecting data. This is a critical step that will cascade in how good the model will be, the more and better data that we get, the better our model will perform.
- There are several techniques to collect the data, like web scraping, manual interventions and etc.
- Predictive Analysis for Big Mart Sales Using Machine Learning Algorithms
- Data set Link: https://www.kaggle.com/shivan118/big-mart-sales-prediction-datasets

2) Dataset:

The dataset consists of 8523 individual data. There are 12 columns in the dataset, which are described below.

1. Item*Identifier* ---- Unique product ID
2. Item*Weight* ---- Weight of product
3. Item*FatContent* ---- Whether the product is low fat or not
4. Item*Visibility* ---- The % of the total display area of all products in a store allocated to the particular product
5. Item*Type* ---- The category to which the product belongs
6. Item*MRP* ---- Maximum Retail Price (list price) of the product
7. Outlet*Identifier* ---- Unique store ID
8. Outlet*EstablishmentYear* ---- The year in which the store was established
9. Outlet*Size* ---- The size of the store in terms of ground area covered
10. Outlet*LocationType* ---- The type of city in which the store is located
11. *Outlet*Type* ---- Whether the outlet is just a grocery store or some sort of supermarket
12. Item*OutletSales* ---- sales of the product in t particular store. This is the outcome variable to be predicted.

Data Preparation:

- Wrangle data and prepare it for training. Clean that which may require it (remove duplicates, correct errors, deal with missing values, normalization, data type conversions, etc.)
- Randomize data, which erases the effects of the particular order in which we collected and/or otherwise prepared our data
- Visualize data to help detect relevant relationships between variables or class imbalances (bias alert!), or perform other exploratory analysis
- Split into training and evaluation sets

Model Selection:

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.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 9, Issue 3, pp 326-331, March-2025

https://doi.org/10.47001/IRJIET/2025.903047

Analyze and Prediction:

- In the actual dataset, we chose only 9 features:
  1. Item*Weight* ---- Weight of product
  2. Item*FatContent* ---- Whether the product is low fat or not
  3. Item*Visibility* ---- The % of the total display area of all products in a store allocated to the particular product
  4. Item*Type* ---- The category to which the product belongs
  5. Item*MRP* ---- Maximum Retail Price (list price) of the product
  6. Outlet*EstablishmentYear* ---- The year in which the store was established
  7. Outlet*Size* ---- The size of the store in terms of ground area covered
  8. Outlet*LocationType* ---- The type of city in which the store is located
  9. *Outlet*Type* ---- Whether the outlet is just a grocery store or some sort of supermarket

Accuracy on test set:

- We got an accuracy of 95.80% on test set.

Saving the Trained Model:

- Once you're confident enough to take your trained and tested model into the production-ready environment, the first step is to save it into a .h5 (or) .pkl file using a library like pickle.
- Make sure you have pickle installed in your environment.
- Next, let's import the module and dump the model into .pkl file

V. RESULTS

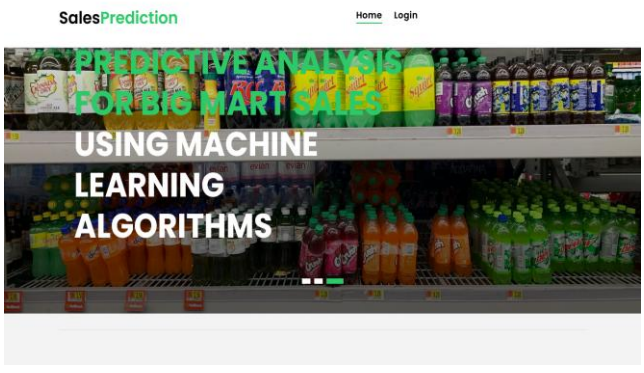

Figure 2: Home Page

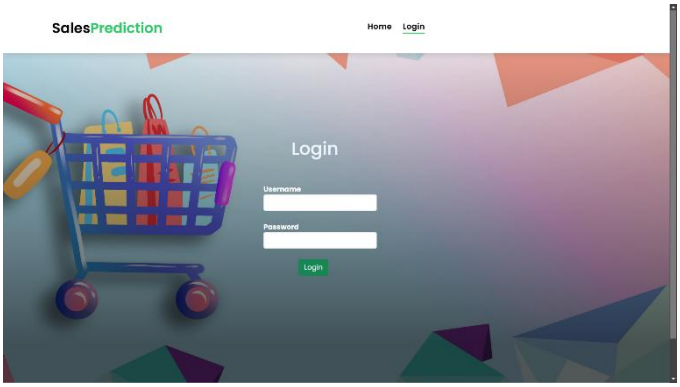

Figure 3: Login Page

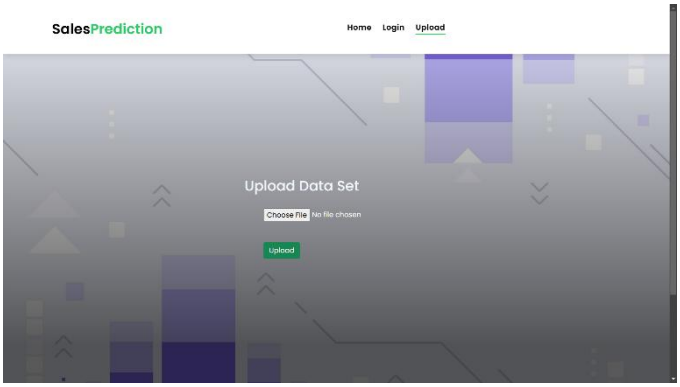

Figure 4: Upload Dataset Page

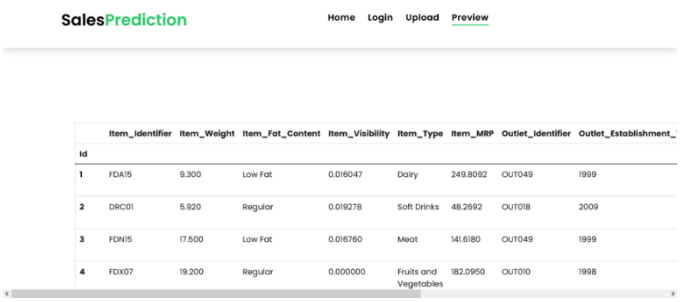

Figure 5: Preview Dataset

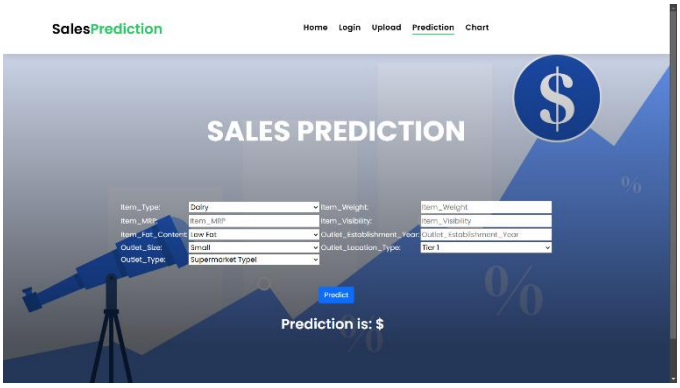

Figure 6: Prediction Page

ISSN (online): 2581-3048
Volume 9, Issue 3, pp 326-331, March-2025
https://doi.org/10.47001/IRJIET/2025.903047

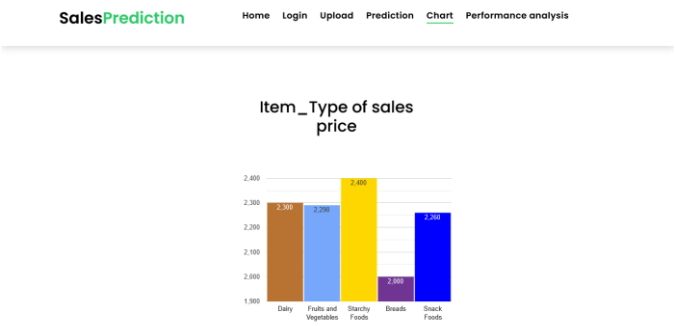

Figure 7: Chart

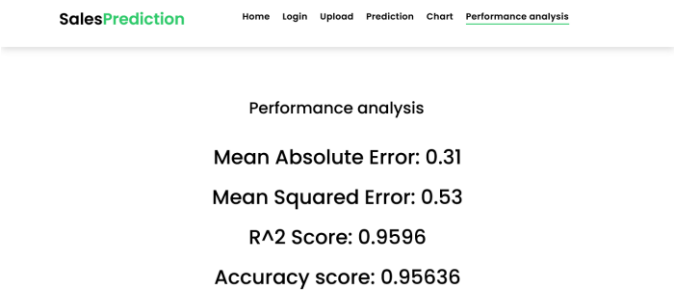

Figure 8: Performance Analysis

VI. CONCLUSION

In this work, the effectiveness of Decision Tree Regression on the data on revenue and review of, best performance-algorithm, here propose software to using regression approach for predicting the sales centered on sales data from the past the accuracy of linear regression prediction can be enhanced with this method, and Decision Tree Regression can be determined. So, we can conclude Decision Tree Regression gives the better prediction with respect to Accuracy.

In future, the forecasting sales and building a sales plan can help to avoid unforeseen cash flow and manage production, staff and financing needs more effectively. In future work we can also consider with the ARIMA model which shows the time series graph.

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] Wang, Haoxiang. "Sustainable development and management in consumer electronics using soft computation." Journal of Soft Computing Paradigm (JSCP) 1, no. 01 (2019): 56.- 2.

[3] 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

[4] Giuseppe Nunnari, Valeria Nunnari, "Forecasting Monthly Sales Retail Time Series: A Case Study", Proc. of IEEE Conf. on Business Informatics (CBI), July 2017.

[5] O. Ajao Isaac, A. Abdullahi Adedeji, I. Raji Ismail, "Polynomial Regression Model of Making Cost Prediction In Mixed Cost Analysis", Int. Journal on Mathematical Theory and Modeling, vol. 2, no. 2, pp. 14 – 23, 2012.

[6] Xinqing Shu, Pan Wang, "An Improved Adaboost Algorithm based on Uncertain Functions", Proc. of Int. Conf. on Industrial Informatics – Computing Technology, Intelligent Technology, Industrial Information Integration, Dec. 2015.

[7] A.S. Weigend and N. A. Gershenfeld, "Time series prediction: Forecasting the future and understanding the past", Addison-Wesley, 1994.

[8] N. S. Arunraj, D. Ahrens, A hybrid seasonal autoregressive integrated moving average and quantile regression for daily food sales forecasting, Int. J. Production Economics 170 (2015) 321-335P

[9] D. Fantazzini, Z. Toktamysova, Forecasting German car sales using Google data and multivariate models, Int. J. Production Economics 170 (2015) 97-135.

[10] X. Yua, Z. Qi, Y. Zhao, Support Vector Regression for Newspaper/Magazine Sales Forecasting, Procedia Computer Science 17 (2013) 1055–1062.

[11] E. Hadavandi, H. Shavandi, A. Ghanbari, An improved sales forecasting approach by the integration of genetic fuzzy systems and data clustering: a Case study of the printed circuit board, Expert Systems with Applications 38 (2011) 9392–9399.

[12] P. A. Castillo, A. Mora, H. Faris, J.J. Merelo, P. Garcia Sanchez, A.J. Fernandez-Ares, P. De las Cuevas, M.I. Garcia-Arenas, Applying computational intelligence methods for predicting the sales of newly published books in a real editorial business management environment, Knowledge-Based Systems 115 (2017) 133-151.

[13] R. J. Kuo, Tung Lai HU and Zhen Yao Chen "application of radial basis function neural networks for sales forecasting", Proc. Of Int. Asian Conference on Informatics in control, automation, and robotics, pp. 325- 328, 2009.

[14] Suresh K and Praveen O, "Extracting of Patterns Using Mining Methods Over Damped Window," 2020 Second International Conference on Inventive Research

ISSN (online): 2581-3048
Volume 9, Issue 3, pp 326-331, March-2025
https://doi.org/10.47001/IRJIET/2025.903047

in Computing Applications (ICIRCA), Coimbatore, India, 2020, pp. 235-241, DOI:10.1109/ICIRCA48905.2020.9182893.

[15] Shobha Rani, N., Kavyashree, S., & Harshitha, R. (2020). Object Detection in Natural Scene Images Using Thresholding Techniques. Proceedings of the International Conference on Intelligent Computing and Control Systems, ICICCS 2020, Iciccs, 509–515.

[16] https://www.kaggle.com/brijbhushannanda1979/bigmartsalesdata.

AUTHORS BIOGRAPHY

Alphan Shaikh,
Student, Computer Engineering, AISSMS College of Polytechnic, Pune, Maharashtra, India.

Mohd Khalid Shaikh,
Student, Computer Engineering, AISSMS College of Polytechnic, Pune, Maharashtra, India.

Ayaan Sayyed,
Student, Computer Engineering, AISSMS College of Polytechnic, Pune, Maharashtra, India.

Prof. A.N.Gedam,
Asst. Professor, Computer Engineering, AISSMS College of Polytechnic, Pune, Maharashtra, India.

Citation of this Article:

Alphan Shaikh, Mohd Khalid Shaikh, Ayaan Sayyed, & Prof. A.N.Gedam. (2025). Predictive Analysis for Big Mart Sales Using Machine Learning Algorithms. *International Research Journal of Innovations in Engineering and Technology - IRJIET*, 9(3), 326-331. Article DOI https://doi.org/10.47001/IRJIET/2025.903047

****

