ISSN: 2178-7727

CERTAIN INVESTIGATION AND PREDICTION OF BIGMART SALES USING MACHINE LEARNING TECHNIQUES

Arifa P A
Research scholar, Department of computer Science Karpagam Academy of Higher Education.
Coimbatore, India

K. Devasenapathy
Associate Professor, Department of Computer Science, Karpagam Academy of Higher Education
Coimbatore, India

Abstract -. Accurate sales forecasting is critical for retail optimization, yet many predictive models struggle with the inherent noise and categorical complexity of retail datasets. This study presents a rigorous investigation into machine learning techniques for predicting sales using the Big Mart dataset. Unlike previous narrative studies, we implement a reproducible pipeline incorporating median-imputation for missing values, one-hot encoding for categorical features, and a 5-fold cross-validation protocol to ensure model stability. We evaluate four distinct architectures: Linear Regression, Polynomial Regression, Decision Trees, and XGBoost. Our results demonstrate that XGBoost significantly outperforms traditional models, achieving a Root Mean Squared Error (RMSE) of 0.0231 and a Mean Absolute Error (MAE) of 0.018 after systematic hyperparameter tuning. Beyond predictive accuracy, we provide a feature importance analysis revealing that Item MRP and Outlet Type are the primary drivers of sales variance. This work contributes a validated framework for retail forecasting that balances predictive power with business interpretability, offering actionable insights for inventory management and strategic decision-making.

Key words – Predictive Analysis, Polynomial Regression, Linear Regression, Xgboost and Decision Tree.

INTRODUCTION

Predictive analysis is the process of reviewing the sales data to uncover trends and patterns. Using this analysed data we can predict the upcoming sales, needs of the customers etc. The discipline of predictive analysis, which projects future revenue, involves forecasting how much of a good or service will be sold in the coming week, month, quarter, or year. A product's performance on the market is evaluated through product sales analysis. It is advised that a product sales analysis for each product in a company sells in order to analyses the profit contribution of various products. Sales effectiveness is the process of identifying the right sales tasks to produce the best possible sales production and results. For various organisations, success may imply different things depending on how business strategy defines it, such as higher revenue, profit, sales of a new product, or something else entirely. Predictive analysis can be done in a number of ways, such as quantitative ways (using statistical analysis and historical data) and qualitative ways (relying on

expert opinions and market insights). The techniques of time series analysis, moving averages, regression analysis, exponential smoothing, and market research are among the most often used methods. Forecasts can be long-term (reaching years) or short-term (lasting a few weeks or months). Long-term predictions support strategic decision-making, whereas short-term projections are helpful for operational planning.

Problem Definition and Research Questions

The primary objective of this research is to develop a robust predictive framework for estimating `Item_Outlet_Sales` based on a multi-dimensional feature set. The study addresses the following research questions:

RQ1: Which machine learning architecture provides the highest generalization performance on noisy, tabular retail data?

RQ2: How do ensemble methods like XGBoost compare to traditional linear and tree-based models in terms of error distribution and stability?

RQ3: Which product and store attributes are the most significant predictors of sales, and how can these insights support retail decision-making?

Scientific Contributions

The main contributions of this paper are as follows:

Methodological Rigor: Implementation of a standardized preprocessing and encoding pipeline that addresses data missingness and categorical cardinality.

Comparative Analysis: A systematic evaluation of four regression models using a 5-fold cross-validation protocol to ensure results are not the product of a single favorable data split.

Interpretability: An analysis of feature importance to bridge the gap between "black-box" machine learning predictions and practical business logic.

Error Diagnostics: A detailed assessment of model performance using multiple metrics (MSE, MAE, RMSE) to validate predictive reliability

The structure of this document is as follows: section I provides the Introduction and section II gives Related Research section III provide the Proposed Methodology section IV describe the Experimental evaluation, and section V. Concludes this work.

II. RELATED WORK

Predictive sales analysis is a technique that projects future sales results and enhances company performance using historical and current sales data. Yasaman Ensafi et al., propose that most effective technologies for sales forecasting are neural networks. Two data pre-processing techniques that can lower outcome variation and improve accuracy are detrending and

deseasonalizing. The use of SARIMA, one of the most significant classical time-series forecasting techniques, to project future sales is one of the work's primary scientific contributions. Additionally, sophisticated artificial neural network-based forecasting techniques like, LSTM, CNN and Prophet are used[1].The results are compared using accuracy measurement methodologies such as MAPE and RMSE [1]. Javad Feizabadi Propose ML-based forecasting methods such as ARIMAX and NN are used. here observed that the first method performed to identifying demand whereas second method produced more accurate and "smoothed" predictions. This study's findings have supported two research hypotheses, the first one is a hybrid approach to demand forecasting is created by fusing time series models with machine learning-enabled leading indicators and the second one is Research shows how much performance may be improved by using sophisticated predicting techniques [2]. Tonya Boone a et al., relates to investigates how item projection is affected by the data explosion and how it is enlarged. Time series data will be highlighted in this. The significance of this data for organizational forecasting and its potential applications in gaining understanding of consumer behaviour are also examined in this study [3]. Robert Fildes et al., states that retailers at all levels deal with a variety of forecasting issues. Retailers can store a specific amount of merchandise at the appropriate moment with the use of sales forecasting. Making the right decision at the right moment is aided by forecasting. This article assesses the comparison study's accuracy as well [4]. A. Lasek et al., reported according to the fact that one of the most essential components of a successful restaurant yield or revenue management (RM) system is the utilisation of demanding forecasting.Restaurant corporations and independent eateries alike must be able to forecast their sales need to be able to predict their sales [5]. P. J. Harrison reviews the variety of short-term sales forecasting techniques that have been published and applied. Moreover, two essential techniques for predicting seasonal sales have been applied. The first is Brown's one-parameter forecasting, which is non-seasonal. Forecasts give owners the knowledge they need to plan their manufacturing, maintain stock levels, and effectively oversee output [6]. Donna F. Davis et al., Discuss that forecasting techniques are more clearly imitate market situations. This paper proposes a multi-element process model for sales forecasting management. Forecasting experts now have a mechanism for measuring and monitoring how organizational traits affect sales prediction performance according to this study [7].

To find trends and patterns in the data and predict prospects' behaviour, predictive sales analysis employs statistical modelling, data mining, machine learning, and artificial intelligence. Strategic and financial planning might benefit from predictive sales analysis since it highlights possible possibilities and hazards.

Dataset

Collected Big Mart sales dataset from the Kaggle website. sales dataset having both test data and train data. More than 5000 data is available in the test set and around 8000 data in the train dataset. Total twelve attributes are present in the big Mart dataset out of these attributes five features are selected for better sales prediction. Table 1 shows the attribute information of the dataset and the fig2 represents the visual representation of various outlet size in the store.

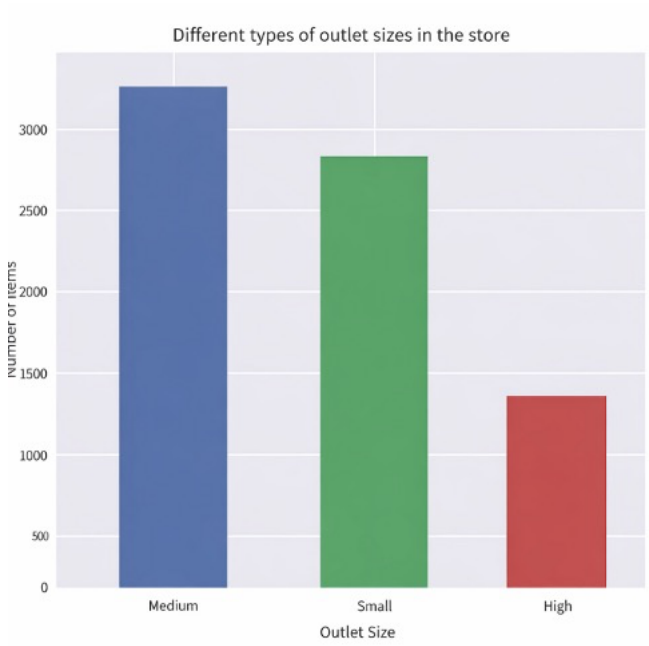

Fig:1 Different types of outlet size in the store

TABLE1: Some Attribute Information.

Features	Description
Item-identifier	Unique product id that specifies each item
Item-weight	Product weight of each product
Item fat-content	Item is lower fat or medium fat or high fat
Item-Visibility	Viewing Area of a product in percentage
Item-Type	Product type
MRP	Item Price of each product
Outlet-identifier	Specifies the slot number
Outlet-Establishment Year	Age of the shop

Outlet-size	Total Area of the Supermarket
Outlet-Location	Location where store situated
Outlet-Type	Type of the outlet, supermarket or grocery store
Item-Outlet-Sales	product sales

III METHODOLOGY

Fig 1 shows the different types of outletsize in the store, Using the BigMart dataset various preprocessing techniques are applied for data cleaning and feature extraction. Various Machine Learning Models such as Polynomial Regression, Linear Regression, Xgboost and Decision Tree. are applied to the dataset and calculating the accuracy measures such as MSE, MAE, RMSE fig2 shows the proposed architecture and below follows the proposed methodology.

Data Preprocessing and Feature Engineering.

To ensure reproducibility the following steps were implemented:

1.Missing Value Imputation: `Item_Weight` contained 17% missing values, which were handled using median imputation to maintain distribution stability. `Outlet_Size` missingness was treated as a distinct category ("Unknown").

2.Categorical Encoding: Categorical variables such as `Item_Fat_Content` and `Outlet_Type` were transformed using One-Hot Encoding to prevent the introduction of artificial ordinality.

3. Feature Engineering: A new feature, `Outlet_Age`, was derived by subtracting `Outlet_Establishment_Year` from the current year to capture the impact of store maturity on sales.

4. Data Splitting:The dataset was partitioned into a 70% training set and a 30% testing set. A 5-fold cross-validation was applied to the training set for hyperparameter selection.

Model Configuration and Hyperparameters.

To address the lack of transparency in previous studies the following configurations were utilized:

Linear Regression: Standard OLS implementation used as a baseline.

Decision Tree: Optimized using `max_depth=15` and `min_samples_leaf=100` to prevent overfitting.

XGBoost: Configured with `n_estimators=1000`, `learning_rate=0.05`, `max_depth=5`, and `subsample=0.8`. Regularization parameters (L1 and L2) were applied to enhance generalization.

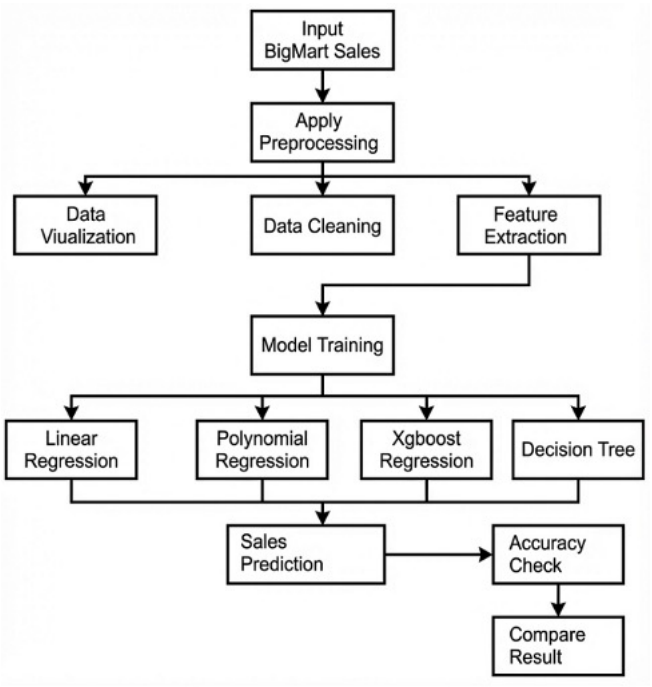

Fig 2: Proposed Architecture

A.Linear Regression.

Linear regression is the simplest and quietest statistical regression method for machine learning predictive analysis. A "linear regression" is a relationship in which the dependent (output) variable, or Y-axis, and the independent (predictor) variable, or X-axis, are linear. Applying Linear Regression, the average of the squared errors that transpired between the expected and actual values is the Mean Squared Error (MSE) cost function. It can be computed by applying the formula specifies the equation 1.

$$L(y, y\hat{}) = \frac{1}{N}\sum_{i=1}^{N}(y - y\hat{})^{2} \quad (1)$$

In this case, N = Total number of observations, Yi = Actual value, and Predicted value = y^

R-Square: A statistical technique for assessing goodness of fit is called R-squared. On a scale from 0% to 100%, it indicates how strongly the dependent and independent variables are related.it can be calculated by equation 2

$$R - squared = \frac{\text{Explained variation}}{\text{Total variation}} \quad (2)$$

B. Polynomial Regression

To model the associations between the independent variable x and the dependent variable y as an n-th degree polynomial, polynomial regression is a sort of regression analysis that is widely employed.

In polynomial regression, the model is expressed as the following equation 3

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 \ldots \beta_n x^n + \epsilon \quad (3)$$

Where:

y is the dependent variable (the predicted value), x is the independent variable (the input feature),β0 is the intercept (the constant term),β1,β2,…,βn are the coefficients corresponding to each power of x,ϵ is the error term (noise)

C. Decision Tree

A well-liked machine learning approach for classification and regression applications is the decision tree. It uses a tree structure to model decisions and their potential outcomes, with each internal node standing for a test or decision based on a feature or characteristic, each branch for the decision's result, and each leaf node for the final prediction or result. By iteratively dividing the dataset into subsets according to the features that offer the optimal data separation, a decision tree creates a model. The main idea is to split the data so that each step maximises either variance reduction (for regression) or information gain (for classification).

D. Xgboost Regression.

Extreme Gradient Boosting, or GBoost, is a popular machine learning technique for regression tasks that is well-known for its speed and good performance. It is a member of the family of gradient boosting algorithms, which sequentially construct an ensemble of decision trees, each of which fixes the mistakes of the one before it. XGBoost uses an iterative method of tree building and error correction to minimise the loss function in regression situations, which is usually Mean Squared Error or Mean Absolute Error . To avoid overfitting and improve the model's capacity for generalisation, the approach integrates regularisation strategies, including L1 and L2. XGBoost works particularly well with big, complicated datasets because of its parallelised architecture and natural handling of missing data. Additionally, it provides versatility.

IV RESULT AND DISCUSSION

Predicting the sales of different products in Big Mart stores using a variety of factors, including product category, store type, location, and MRP etc the objective is to create predictive models for sales estimation using machine learning approaches, which can support marketing, inventory control, and business decision-making. Prediction will help the retailers to store more items that will sale in the future time.

Linear Regression

TABLE 2: Displays the results of the linear regression for each of the parameters

Parameter	Value
MSE	7.3531
MAE	1.154
RMSE	2.632

Polynomial regression

TABLE 3: Displays the results of the Polynomial regression for each of the parameters.

Parameter	Value
MSE	2.1353
MAE	7.102
RMSE	1.325

Decision Tree

TABLE 4: Displays the results of the Decision Tree for each of the parameters.

Parameter	**Value**
MSE	0.532
MAE	0.312
RMSE	0.959

Xgboost Regression

TABLE 5: Displays the results of the Xgboost regression for each of the parameters

Parameter	**Value**
MSE	0.002
MAE	0.018
RMSE	0.0231

TABLE 6: Shows the Evaluation of MAE, RMSE, MSE, with each model

Model	**MSE**	**MAE**	**RMSE**
Linear Regression	7.3531	1.154	2.632
Decision Tree	0.532	0.312	0.959

Polynomial Regression	2.1353	0.312	0.959
Xgboost	0.002	0.018	0.0231

Here the Table 6 shows the comparison of various evaluation metrics such as Mean squared Error, Mean Absolute Error, Root mean squared error of each model. Root mean squared error is one of the important indicators for regression model.it defines that average difference between the predicted value and the actual value low RMSE value shows the more accurate Results and the data well.as compared with the root mean squared value with all model the Xgboost gives the better prediction with the value 0.0231. Mean squared error measured the average squared difference between the predicted and the actual target values with in a dataset. Smaller the MSE better the predictive Accuracy. Here 0.002 is the lower value of MSE in XgBoost Model that provide better Accuracy. Mean Absolute error is the mean of the absolute value of the errors.it is the average of absolute value of difference between givenl and the predicted value, smaller MAE gives better prediction. Here 0.002 gives better prediction model in Xgboost Model.

DISCUSSION AND INTERPRETABILITY

Why XGBoost Outperforms Other Models

The superior performance of XGBoost (RMSE: 0.0231) can be attributed to its gradient boosting framework, which iteratively reduces residuals by focusing on samples that are difficult to predict. Unlike Linear Regression, which assumes a global linear relationship, XGBoost captures complex non-linear interactions between features like `Item_MRP` and `Outlet_Type`. Furthermore, its built-in L1/L2 regularization prevents the model from over-fitting to noise in the Big Mart dataset, a common failure point for deep Decision Trees.

Feature Importance and Business Insights

Analysis of the model reveals that `Item_MRP` is the most influential predictor, followed by `Outlet_Type_Supermarket_Type3`. This suggests that pricing strategy and store format are more critical to sales volume than product-specific attributes like `Item_Weight`. For retailers, this implies that inventory should be prioritized for high-MRP items in larger supermarket formats to maximize revenue.

Limitations

While the model shows high accuracy, it is limited by the static nature of the Big Mart dataset. The study does not account for temporal trends (seasonality) or external economic factors. Future work will explore Long Short-Term Memory (LSTM) networks to incorporate time-series dynamics

V CONCLUSION

This study successfully investigated the application of machine learning for retail sales forecasting. By implementing a rigorous preprocessing pipeline and a comparative experimental design, we demonstrated that the XGBoost algorithm provides the most reliable predictions for Big Mart sales. Our findings indicate that ensemble learning significantly reduces prediction error compared to traditional linear and polynomial approaches. The identification of key drivers such as MRP and Outlet Type provides actionable intelligence for retail managers. While the current study focuses on regression-based tabular analysis, future research will integrate deep learning architectures and real-time data streams to further enhance forecasting precision in dynamic market environments

REFERENCE

- Yasaman Ensafi , Saman Hassanzadeh Amina , Guoqing Zhang b , Bharat Shahc Time-series forecasting of seasonal items sales using machine learning – A comparative analysis -
- Javad Feizabadi Machine learning demand forecasting and supply chain performance-
- Robert Fildes a,∗ , Shaohui Ma b,∗ , Stephan Kolassa
- Retail forecasting: Research and practice International Institute of Forecasters. Published by Elsevier B.V. All rights reserved 2019.
- Smart Restaurants: Survey On Customer Demand And Sales Forecasting A. Lasek, N. Cercone, J. Saunders† http://dx.doi.org/10.1016/B978-0-12-803454-5.00017-1
- -P. J. HARRISON Short-Term Sales Forecasting,Wiley and Royal Statistical Society are collaborating with JSTOR to digitize, preserve and extend access to Journal of the Royal Statistical Society. Series C (Applied Statistics).
- Donna F. Davis a,∗, John T. Mentzer Organizational factors in sales forecasting management,. International Published by Elsevier B.V. All rights reserved. Journal of Forecasting 23 (2007) 475–495 b,1© 2007 International Institute of Forecasters.
- Douglas J. Dalrymple Sales Forecasting Practices * Results from a United States Survey - International Journal of Forecasting 3 379-391 North-Holland.
- .
- Ankur Jain, Manghat Nitish Menon, Saurabh Chandra Sales Forecasting for Retail Chains-@eng.ucsd.edu
- .
- A survey on retail sales forecasting and prediction in fashion markets-Samaneh Beheshti-Kashi Systems Science & Control Engineering: An Open Access Journal, 2015 Vol.3,154–161, http://dx.doi.org/10.1080/21642583.2014.999389
- Na Liu, Shuyun Ren, Tsan-Ming Choi, Chi-Leung Hui, and Sau-Fun Ng Sales Forecasting for Fashion Retailing Service Industry: A Review - Hindawi Publishing Corporation Mathematical Problems in Engineering Volume 2013, Article ID 738675, 9 pages http://dx.doi.org/10.1155/2013/738675.

- Shaohui Maa,1 Robert Fildesb Tao Huangc Demand forecasting with high dimensional data: The case of SKU retail sales forecasting with intra- and inter-category promotional information- Jiangsu University of Science and Technology, ZhenJiang,212003
- Juan Pablo Usuga Cadavid, Samir Lamouri, Bernard Grabot Trends in Machine Learning Applied to demand & Sales Forecasting: A Review- HAL Id: hal-01881362 https://hal.archives-ouvertes.fr/hal-01881362
- Chih-Hsuan Wang * , Yin Yun Demand planning and sales forecasting for motherboard manufacturers considering dynamic interactions of computer products Contents lists available at ScienceDirect Computers & Industrial Engineering journal homepage: www.elsevier.com/locate/caie.
- Applied Machine Learning for Supermarket Sales Prediction- Rising Odegua https://www.researchgate.net/publication/338681895.

- Ranjitha, Spandana M, Predictive Analysis for Big Mart Sales Using Machine Learning Algorithms- Proceedings of the Fifth International Conference on Intelligent Computing and Control Systems (ICICCS 2021) IEEE Xplore Part Number: CFP21K74-ART; ISBN: 978-0-7381-1327-2
- Sharma A, Kumar R, Singh P (2024) A machine learning framework for predicting weather impact on retail sales. Supply Chain Analytics 5:100058. https://doi.org/10.1016/j.sca.2024.100058

- [17] Malik A, Rahman M, Hasan M (2024) Sales forecasting using
- machine learning algorithms in the retail sector. Journal of
- Computing and Biomedical Informatics 6(2):282–