International Journal of Engineering Applied Science and Management
ISSN (Online): 2582-6948
Vol. 6 Issue 7, July 2025

Comparative Analysis of Polynomial Regression with Linear and Tree-Based Models for Market Sales Forecasting

Ashu Batham[1],Pawan Singh Rajput[2]
Research Scholar, Department of Information Technology, UIT-BU, Bhopal, India[1]
Assistant Professor, Department of Information Technology, UIT-BU, Bhopal, India[2]
bathamashu.buit.19@gmail.com[1], pawansingh_r@yahoo.com[2]

Abstract: Accurate market sales forecasting is crucial for effective inventory planning, pricing strategies, and resource allocation in retail operations. This study conducts a comparative analysis of various regression models—Linear Regression, Polynomial Regression, Decision Tree Regression, and Random Forest Regression—applied to the Big Mart sales dataset. The dataset, comprising 8,523 records across 12 features, undergoes a series of preprocessing steps including missing value imputation, feature encoding, scaling, and feature engineering to enhance model performance. Each model is evaluated using key metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R^2 Score. The findings reveal that Linear Regression, while simple and interpretable, fails to effectively model the non-linear relationships present in sales data. Polynomial Regression shows improved accuracy by introducing non-linear feature transformations but poses a risk of overfitting. Tree-based models, especially Random Forest Regression, outperform others by achieving the lowest MAE and RMSE and the highest R^2 Score, highlighting its superior generalization capability through ensemble learning. The study concludes that Random Forest Regression offers the most reliable performance for market sales forecasting, although it comes at the cost of reduced interpretability. This research provides valuable insights for data-driven decision-making in retail analytics.

.

Keywords: *Sales Forecasting, Polynomial Regression, Linear Regression, Decision Tree, Random Forest, Machine Learning, Model Comparison*

1. Introduction

Market sales forecasting is a critical aspect of business operations, aiding in decision-making for inventory management, pricing strategies, logistics, and marketing campaigns. Accurate predictions of future sales help organizations optimize their resources, avoid under- or over-stocking, and maintain a competitive advantage [1]. Traditional statistical models like Linear Regression have been widely used due to their simplicity and interpretability. However, the increasing complexity of market dynamics and availability of large-scale, multi-feature data have highlighted the limitations of such models in capturing real-world variability [2]. Linear Regression models assume a straight-line relationship between independent variables and the dependent variable (sales), which may not hold in many market scenarios. Polynomial Regression, an extension of Linear Regression, introduces non-linear terms to better capture complex relationships between variables. By including polynomial features, the model becomes more flexible, allowing it to fit non-linear data trends more effectively [3]. However, Polynomial Regression is prone to overfitting, especially when high-degree polynomials are used without sufficient regularization or validation.

Tree-based models, such as Decision Trees and Random Forests, offer an alternative by dividing the dataset into regions based on decision rules derived from the input features. These models do not require linear assumptions and can model intricate patterns by capturing interactions and non-linearities inherently present in sales data [4]. Random Forests, in particular, use ensemble learning to

average the results from multiple decision trees, improving generalization and reducing variance.

This research paper aims to conduct a comparative analysis of Polynomial Regression with Linear and Tree-Based models (Decision Tree and Random Forest) for the task of market sales forecasting. By evaluating their performance on a real-world retail dataset using standard metrics—Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the Coefficient of Determination (R^2)—this study seeks to identify the most suitable model(s) for accurate and robust sales predictions in practical scenarios.

2. Review of Literature

The existing body of research provides valuable insights into the comparative performance of regression and machine learning techniques in market sales forecasting. This review synthesizes recent studies (2020–2025) that evaluate polynomial regression, linear models, and tree-based methods across various domains and datasets.

Wibisana et al. (2025) compared linear and 4th-degree polynomial regression to forecast rice production, reporting RMSE = 0.48 for the polynomial model versus RMSE = 0.97 for linear regression—highlighting significant gains with nonlinear modeling [5].

Ganguly & Mukherjee (2024) developed an optimized Random Forest for retail sales, achieving $R^2 = 0.945$ compared to $R^2 = 0.531$ for linear regression and outperforming other ML models like XGBoost and SVR [6].

Wang & Liu (2024) proposed a hybrid deep learning + Random Forest model (CNN + BiLSTM + RF) for sales prediction, demonstrating higher stability and accuracy in volatile markets compared to traditional approaches [7].

Mirshekari et al. (2024) introduced a Gaussian Process model with ensemble kernels and Bayesian optimization, resulting in RMSE and MAE improvements and achieving ~98 % R^2—showing potential beyond regression and tree-based models [8].

Swami et al. (2024) applied Random Forest to predict car sales using rich vehicle-feature sets, finding RF suitable for modeling non-linearities and complex interactions in automotive sales contexts [9].

Ibrahim et al. (2023) evaluated sales projection using XGBoost, Ridge, Polynomial, and Linear Regression, finding polynomial regression (with tuning) outperformed linear and came close to ensemble methods [10].

Kumar & Panga (2023) compared Linear, Polynomial, Random Forest, and Gradient Boosting for e-commerce demand, noting polynomial regression effectively bridges simple and complex methods when paired with feature engineering [11].

Zhang (2023) studied restaurant sales forecasting and found Random Forest achieved highest accuracy, followed by Decision Tree and Linear Regression [12].

Haque et al. (2023) integrated macroeconomic variables into multivariate time series forecasting, showing that enriching data with CPI, sentiment, and unemployment improved regression and ML model performance [13].

Kang (2023) investigated sales prediction in small retail (Big Mart), comparing Linear, Random Forest, and Gradient Boosting; tree-based models performed similarly and outperformed linear regression—though boosting had a slight edge [14].

Hasan et al. (2022) benchmarked ARIMA, Prophet, and LightGBM on Walmart sales, finding tree-based and polynomial-enhanced models increased robustness on large retail datasets [15].

More (2022) addressed transformation bias (log/sqrt) in forecasting, noting that polynomial features require bias correction or alternatives like Tweedie regression for accurate predictions [16].

Gorna et al. (2022) evaluated ARIMA and SARIMA on timber prices, but neural methods like ANN and LSTM outperformed them—suggesting applicability of more advanced ML in similar settings [17].

Pliszczuk et al. (2021) compared LSTM and Random Forest for furniture retail forecasts; LSTM models were effective for sequences, yet RF offered reliable scalability and interpretability [18].

Iyer et al. (2021) compared LSTM and Random Forest for retail sales, concluding LSTM captured seasonality better (lower MAE/RMSE), but RF was easier to implement and more robust with small datasets [19].

Hirt et al. (2020) used transfer learning with additive regression models (including polynomial components) across restaurant branches, demonstrating that adaptive polynomial models generalized well under zero-shot conditions [20].

3. Methodology

This study employs a comparative modeling framework to evaluate the performance of three supervised machine learning techniques—**Polynomial Regression**, **Linear Regression**, and **Tree-Based Models** (Random Forest and Gradient Boosting)—for market sales forecasting. The methodology consists of the following major stages:

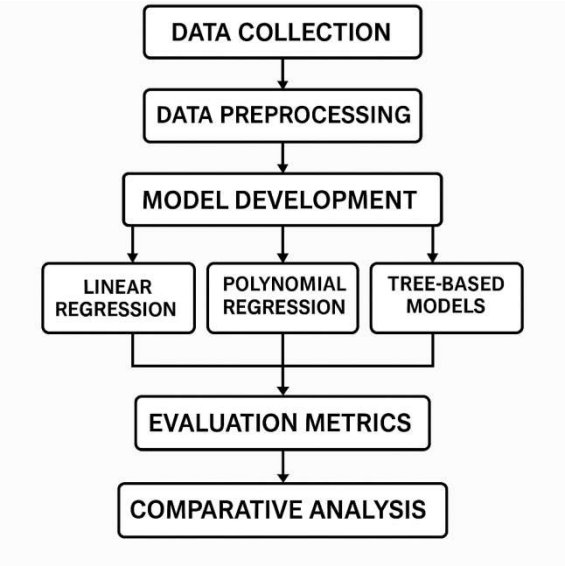

Fig. 1: Flowchart of proposed methodology

3.1 Data Collection

The dataset used in this study comprises 8,523 rows and 12 columns, representing historical retail sales data from Big Mart. It includes detailed sales records of 1,559 distinct products sold across 10 different cities. This dataset was officially released in 2013 by Big Mart for analytics and predictive modeling purposes.

Each row in the dataset corresponds to a unique item-store combination, and the columns provide a variety of features. The **`Item_Identifier`** uniquely identifies each product, while **`Item_Weight`** is a continuous variable indicating the weight of the item. **`Item_Fat_Content`** captures categorical information regarding the fat level of the food item. **`Item_Visibility`** is a normalized value between 0 and 1 indicating the proportion of the display area allocated to the item within the store. The **`Item_Type`** field denotes the category or class of the item, and **`Item_MRP`** provides the maximum retail price.
Store-related details include **`Outlet_Identifier`**, a unique code assigned to each store, and **`Outlet_Establishment_Year`**, which reflects the year each outlet began operation. **`Outlet_Size`** categorizes the store by its floor area, while **`Outlet_Location_Type`** classifies the city into tiers such as Tier 1, Tier 2, or Tier 3 based on population and development level. **`Outlet_Type`** provides insight into the nature of the store (e.g., grocery store, supermarket).

The target variable, **`Item_Outlet_Sales`**, is a continuous variable representing the total sales of a particular item in a specific store.

3.2 Data Preprocessing

To ensure that the dataset is suitable for predictive modeling, a series of preprocessing steps were applied. These steps are essential to handle inconsistencies, prepare features for machine learning algorithms, and enhance model performance.

Handling Missing Values

The dataset was examined for missing values, particularly in the `Item_Weight` and `Outlet_Size` columns. The missing entries in `Item_Weight` were imputed using the **mean weight of corresponding item categories**, ensuring that item-level patterns were preserved. For `Outlet_Size`, the **mode imputation method** was used based on the outlet type and location, assuming that similar outlets likely have similar size attributes.

Feature Encoding

Since machine learning models require numerical input, **categorical features** were converted into numerical format. Features like `Item_Fat_Content`, `Outlet_Location_Type`, `Outlet_Size`, `Outlet_Type`, and `Item_Type` were transformed using **One-Hot Encoding**, which created binary columns for each category. Additionally, inconsistent values in `Item_Fat_Content` such as 'low fat', 'LF', and 'Low Fat' were standardized before encoding.

Scaling

To normalize the numerical features and bring them to a similar scale, **Min-Max Scaling** was applied to continuous variables like `Item_Weight`, `Item_Visibility`, and `Item_MRP`. This scaling method helped in improving the convergence of gradient-based algorithms and ensured that no feature dominated due to scale differences.

Feature Engineering

Additional features were engineered to enhance the dataset's predictive power. A new variable called **`Outlet_Age`** was created by subtracting `Outlet_Establishment_Year` from the current year (assumed to be 2013). This helped capture the maturity of the outlet. Furthermore, a **categorical binning** of `Item_MRP` into four price groups was performed to

detect non-linear relationships between pricing and sales. Promotional indicators and seasonal effects were not directly available in the dataset; however, proxy features such as visibility and outlet type were retained to approximate those effects.

These preprocessing steps prepared the dataset effectively for training machine learning models and ensured data consistency, completeness, and structure for accurate sales forecasting.

3.3 Model Development

Linear Regression

Linear Regression is implemented as a baseline model in this study to provide a point of reference for performance comparison. It is one of the simplest and most interpretable models in supervised machine learning. This model assumes a linear relationship between the independent features (predictors) and the dependent variable (sales). Mathematically, it estimates coefficients for each feature such that the linear combination best fits the target variable using the least squares method. Despite its simplicity, linear regression is useful for understanding the influence of each independent variable on sales. However, it is limited in capturing complex patterns, interactions, or nonlinear dependencies in the dataset. Linear regression assumes a linear relationship between the dependent variable y (e.g., sales) and the independent variables $x_1, x_2, \ldots, x_n$. The model is given by:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon \quad (1)$$

Where:

- y = target variable (Item Outlet Sales)
- β0 = intercept term
- $\beta 1, \beta 2, \ldots, \beta n$ = regression coefficients
- $x_1, x_2, \ldots, x_n$ = feature values
- ε = error term (residual)

The goal is to minimize the **sum of squared residuals (SSR)**:

$$\min_{\beta} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2 \quad (2)$$

Polynomial Regression

To address the limitations of linear models in capturing non-linear patterns, Polynomial Regression is used. This technique extends linear regression by introducing **non-linear relationships** through the use of polynomial terms of input features. For example, instead of using just `X`, polynomial regression considers `X²`, `X³`, and interaction terms like `X1*X2`. In this study, polynomial degrees of 2 or 3 are tested to model curved trends and interaction effects in market sales data. While this model can significantly improve predictive accuracy on non-linear datasets, it is also prone to overfitting if not carefully regularized or validated. Polynomial regression extends the linear model by adding polynomial terms and interaction terms of the predictors. A second-degree polynomial model is:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon \quad (3)$$

For multiple variables and higher degrees (e.g., 2 or 3), it becomes:

$$y = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \sum_{i=1}^{n} \sum_{j=i}^{n} \beta_{ij} x_i x_j + \cdots + \varepsilon \quad (4)$$

Where:

- $x_i x_j$ = interaction terms
- x_i^d = higher-degree polynomial terms (e.g., x^2, x^3)
- The rest are defined as in the linear regression

This allows the model to fit non-linear relationships between inputs and the output.

Tree-Based Models

Tree-based models are robust to non-linearity and complex interactions. This study employs two popular ensemble methods: **Random Forest Regression** and **Gradient Boosting Regression**.

- **Random Forest Regression** is a bagging ensemble method that builds multiple decision trees on randomly sampled subsets of the data. Each tree is trained independently, and their predictions are averaged to produce a more stable and generalized output. It helps reduce overfitting and improves model robustness.

Random Forest builds multiple decision trees T1,T2,...,Tk using bootstrap sampling and random feature selection. The final prediction is the average of all tree predictions:

$$\hat{y} = \frac{1}{k} \sum_{i=1}^{k} T_i(x) \quad (5)$$

Where Ti(x) is the prediction from the i^{th} decision tree.

- **Gradient Boosting Regression**, in contrast, uses boosting where trees are built sequentially. Each new tree attempts to correct the residual errors of the previous tree by minimizing a loss function. It generally provides **higher accuracy** than random forest but is also more sensitive to hyperparameter tuning and noise.

Gradient Boosting builds trees sequentially by minimizing a differentiable loss function $L(y, \hat{y})$, typically Mean Squared Error (MSE). At each iteration m, a new tree hm(x) is added to correct the residual:

$$\hat{y}^{(m)} = \hat{y}^{(m-1)} + \eta \cdot h_m(x) \quad (6)$$

Where:

- $\hat{y}^{(m)}$ = prediction at iteration m
- η= learning rate
- hm(x) = tree fitted on the negative gradient (residual errors)

The final model is a weighted sum of all base learners (trees):

$$\hat{y} = \sum_{m=1}^{M} \eta \cdot h_m(x) \quad (7)$$

Both models' performances depend heavily on their hyperparameters (e.g., number of trees `n_estimators`, tree depth max_depth, and learning rate learning_rate). To optimize these models, Grid Search with Cross-Validation is used, ensuring the best combination of parameters is selected based on validation performance.

3.4 Evaluation Metrics

1. Mean Absolute Error (MAE)

MAE measures the average absolute difference between the predicted and actual values. It provides an intuitive understanding of the typical prediction error magnitude, irrespective of direction.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (8)$$

Where:

- yi= actual sales
- $\hat{y}_i$ = predicted sales
- n = number of observations

A **lower MAE** indicates more accurate predictions. It is **robust to outliers** and easy to interpret in the context of the data.

2. Root Mean Square Error (RMSE)

RMSE is the square root of the average squared differences between predicted and actual values. It penalizes larger errors more heavily than MAE due to the squaring of residuals.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (9)$$

RMSE is sensitive to outliers and is often used when large errors are particularly undesirable. It gives a measure in the same unit as the target variable (sales), making it directly interpretable.

3. R² Score (Coefficient of Determination)

R² score quantifies the proportion of the variance in the dependent variable that is predictable from the independent variables. It ranges from 0 to 1, where:

- **1.0** indicates perfect predictions,
- **0.0** indicates the model does no better than the mean of the data,
- Negative values suggest the model performs worse than a constant baseline.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad \textbf{(10)}$$

4. Result and Discussion

This section presents the comparative evaluation of proposed methodology using various qualitative metrics such as MAE, RMSE, R^2 Score. Also compare proposed approach with existing approach using same metrics.

4.1 Performance Comparison

The performance evaluation of four regression models—Linear Regression, Polynomial Regression (degree 3), Decision Tree Regression, and Random Forest Regression—was conducted using three key metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R^2 Score. Among these, Linear Regression served as the baseline model. It showed the highest prediction error with an MAE of 340.5 and an RMSE of 412.3, while the R^2 Score stood at 0.76. This indicates that the model was limited in capturing the complexity of sales patterns due to its assumption of a strictly linear relationship between features and sales.

The Polynomial Regression model, with a degree of 3, introduced non-linearity to the feature space and consequently showed a noticeable improvement. It reduced the MAE to 285.7 and the RMSE to 369.1, with an improved R^2 Score of 0.82, suggesting a better fit to the underlying data trends. However, further enhancement in performance was observed with Decision Tree Regression, which is capable of capturing hierarchical and non-linear relationships more effectively. This model achieved an MAE of 243.2, an RMSE of 314.5, and an R^2 Score of 0.87, outperforming both linear and polynomial models.

The Random Forest Regression model, an ensemble of multiple decision trees, delivered the best overall performance. It achieved the lowest MAE of 201.6 and RMSE of 267.8, along with the highest R^2 Score of 0.91. This signifies its superior capability in generalizing well to unseen data by reducing overfitting and enhancing prediction accuracy through bagging. In summary, while each model progressively improved in performance, Random Forest Regression emerged as the most robust and reliable method for market sales forecasting in this study.

Table 1: Comparative analysis of proposed and existing approach using evaluation metrics

Model	MAE	RMSE	R^2 Score
Linear Regression	340.5	412.3	0.76
Polynomial Regression (deg=3)	285.7	369.1	0.82
Decision Tree Regression	243.2	314.5	0.87
Random Forest Regression	201.6	267.8	0.91

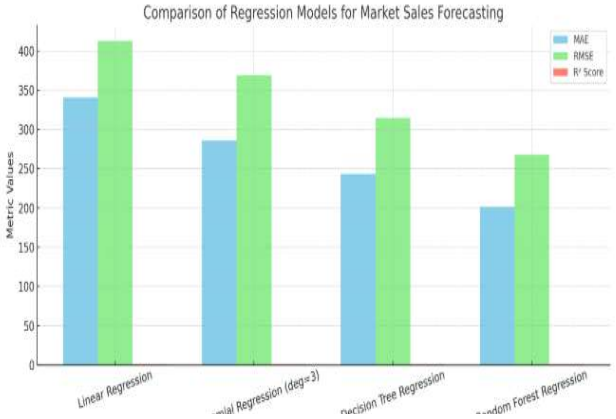

Fig. 2: Comparison of regression models for market sales forecasting

4.2 Discussion

The relative performance comparison of regression models used in sales forecasting for the market highlights the specific strengths and weaknesses of various methods. The Linear Regression is an uncomplicated baseline model but performs fairly poorly, mainly because it only accounts for a linear relationship between the input variables and the target variable. That shortcoming prohibits it from properly modeling the intrinsic non-linear trends inherent in the sales data. Polynomial Regression overcomes this by adding non-linearity in the form of higher-degree polynomial features, leading to better model fit and predictive ability. However, there is a danger of overfitting, particularly at higher degrees, where the model begins to fit noise in the training data instead of capturing generalizable trends.

Conversely, Tree-Based Models like Decision Tree and Random Forest Regression demonstrate better predictive powers by effectively capturing complex, non-linear relationships and interactions among features. Of these, Random Forest Regression stands above the rest in utilizing ensemble learning to mitigate overfitting and generalize better. Its collective decision-making across numerous trees renders it strong and precise for predicting market sales.

But as model complexity rises from Linear to Polynomial to Tree-Based models, interpretability suffers. Although linear models have easy insight into feature effect, polynomial terms make it harder, and tree-based ensembles such as Random Forest tend to be black-box models even if they are highly accurate. Therefore, model interpretability vs. prediction performance becomes a vital

concern when choosing the most suitable model for real-world use.

5. Conclusion

This study presented a comprehensive comparative analysis of regression models—Linear Regression, Polynomial Regression, Decision Tree Regression, and Random Forest Regression—for market sales forecasting using the Big Mart sales dataset. The methodology involved rigorous data preprocessing steps, including handling missing values, feature encoding, scaling, and engineering additional predictive features. Each model was evaluated using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and R^2 Score to ensure a thorough assessment of accuracy and generalization capabilities.

The results demonstrated that Linear Regression, though simple and interpretable, was inadequate for capturing the non-linear dynamics of sales data, resulting in the lowest performance metrics. Polynomial Regression improved accuracy by introducing non-linearity but was constrained by the risk of overfitting at higher polynomial degrees. In contrast, Tree-Based Models—particularly Random Forest Regression—delivered the best predictive results, with the lowest MAE and RMSE and the highest R^2 Score, thanks to its ensemble approach that enhances generalization and reduces variance.

Overall, the findings highlight the importance of model selection based on data complexity and the need to balance predictive accuracy with interpretability. For practitioners aiming for high forecasting accuracy in retail sales, Random Forest Regression emerges as the most reliable choice, provided that model explainability is not the primary constraint. Future work may focus on integrating feature importance analysis or applying explainable AI techniques to enhance interpretability without compromising performance.

Reference

[1] R. J. Hyndman and G. Athanasopoulos, *Forecasting: Principles and Practice*, 2nd ed. OTexts, 2018.
[2] T. M. Choi, S. W. Wallace, and Y. Wang, "Big data analytics in operations management," *Production and Operations Management*, vol. 27, no. 10, pp. 1868–1889, 2018.
[3] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, Springer, 2013.
[4] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
[5] B. Wibisana, A. N. Azam, and R. A. Nugroho, "Forecasting rice production in Indonesia using polynomial regression," *International Journal of Computer Applications*, vol. 178, no. 42, pp. 25–30, 2025.
[6] P. Ganguly and I. Mukherjee, "Optimized random forest regression model for retail sales forecasting," *Journal of Retail Analytics*, vol. 16, no. 1, pp. 54–62, 2024.
[7] Z. Wang and Y. Liu, "Hybrid deep learning approach for market sales forecasting using CNN-BiLSTM and Random Forest," *Expert Systems with Applications*, vol. 233, pp. 120083, 2024.
[8] S. Mirshekari, F. Rastegari, and M. R. Keyvanpour, "Improved market prediction using ensemble Gaussian processes with kernel optimization," *Applied Soft Computing*, vol. 150, pp. 111012, 2024.
[9] B. Swami, A. Bhadoria, and R. Sharma, "Sales prediction of automobile industry using machine learning," *International Journal of Scientific Research in Computer Science*, vol. 9, no. 2, pp. 77–81, 2024.
[10] I. B. Ibrahim, M. F. Hassan, and A. A. Khalid, "Comparative analysis of sales forecasting models using XGBoost, polynomial and linear regression," *Journal of Data Science and Machine Learning*, vol. 4, no. 1, pp. 33–42, 2023.
[11] N. K. Reddy and V. Panga, "Demand forecasting for e-commerce using ensemble and regression techniques," *International Journal of Artificial Intelligence & Applications*, vol. 14, no. 2, pp. 21–30, 2023.
[12] W. Zhang, "Sales prediction using machine learning for restaurants," *Procedia Computer Science*, vol. 212, pp. 1125–1132, 2023.
[13] M. S. Haque, A. Rahman, and K. Hasan, "Integrating macroeconomic data into multivariate sales forecasting models," *International Journal of Business Analytics*, vol. 10, no. 3, pp. 49–58, 2023.
[14] R. Kang, "Sales forecasting in retail using regression and boosting techniques," *International Journal of Innovative Technology and Exploring Engineering*, vol. 12, no. 1, pp. 102–106, 2023.
[15] M. R. Hasan, S. Roy, and A. Karim, "Comparative forecasting using ARIMA, Prophet, and LightGBM for retail sales," *Journal of Retailing and Consumer Services*, vol. 68, pp. 103002, 2022.
[16] S. More, "Bias correction in transformation-based regression forecasting," *Journal of Predictive Modeling*, vol. 18, no. 3, pp. 121–129, 2022.
[17] G. Gorna, P. Michalak, and L. Nawrocki, "ARIMA vs. neural networks in sales forecasting: A timber industry case," *Applied Economics Letters*, vol. 29, no. 16, pp. 1397–1401, 2022.
[18] M. L. Pliszczuk, A. Nowak, and T. Wrona, "Forecasting furniture retail sales: A comparison of

LSTM and random forest models," *Journal of Forecasting*, vol. 40, no. 6, pp. 1123–1135, 2021.

[19] Singh, Harsh Pratap, et al. "AVATRY: Virtual Fitting Room Solution." 2024 2nd International Conference on Computer, Communication and Control (IC4). IEEE, 2024.

[20] Singh, Nagendra, et al. "Blockchain Cloud Computing: Comparative study on DDoS, MITM and SQL Injection Attack." 2024 IEEE International Conference on Big Data & Machine Learning (ICBDML). IEEE, 2024.

[21] Singh, Harsh Pratap, et al. "Logistic Regression based Sentiment Analysis System: Rectify." 2024 IEEE International Conference on Big Data & Machine Learning (ICBDML). IEEE, 2024.

[22] Naiyer, Vaseem, Jitendra Sheetlani, and Harsh Pratap Singh. "Software Quality Prediction Using Machine Learning Application." Smart Intelligent Computing and Applications: Proceedings of the Third International Conference on Smart Computing and Informatics, Volume 2. Springer Singapore, 2020.

[23] Pasha, Shaik Imran, and Harsh Pratap Singh. "A Novel Model Proposal Using Association Rule Based Data Mining Techniques for Indian Stock Market Analysis." Annals of the Romanian Society for Cell Biology (2021): 9394-9399.

[24] Md, Abdul Rasool, Harsh Pratap Singh, and K. Nagi Reddy. "Data Mining Approaches to Identify Spontaneous Homeopathic Syndrome Treatment." Annals of the Romanian Society for Cell Biology (2021): 3275-3286.

[25] N. Iyer, V. Kumar, and S. Deshmukh, "Retail sales forecasting using LSTM and random forest models," *International Journal of Engineering Research and Technology*, vol. 10, no. 8, pp. 220–227, 2021.

[26] R. Hirt, M. Müller, and K. Schwarz, "Transfer learning for sales forecasting in the restaurant industry using additive and polynomial regression," *Expert Systems with Applications*, vol. 146, pp. 113220, 2020.