Comparative Assessment of Predictive Sales Models Using Big Data and Visual Data Exploration

Bipul Chakraborty[1], Yashwanth Gowda B[2]
[1]M.Tech Research Scholar, [2]Assistant Professor
Department of Computing and Engineering, Sri Guru Granth Sahib World University, Fatehgarh Sahib, Punjab, India

Abstract:

Effective sales forecasting is essential for businesses across various sectors, including retail, manufacturing, logistics, and marketing, to streamline operations, optimize inventory, and enhance revenue planning. Conventional forecasting methods often fall short in modeling the complex and nonlinear dynamics embedded in modern sales data, which restricts their predictive reliability. To address this challenge, this study investigates the use of advanced machine learning algorithms on the Big Mart dataset, utilizing PySpark's distributed computing platform to efficiently process large volumes of data. Several algorithms—such as K-Nearest Neighbor, Decision Trees, XGBoost, Random Forest, and traditional regression models—were developed and evaluated. Key sales-driving factors like product characteristics, promotional campaigns, seasonal fluctuations, and outlet features were integrated as model inputs. Performance was rigorously measured using error metrics including RMSE, MAE, and R-squared. Results reveal that ensemble-based models, particularly XGBoost and Random Forest, outperform conventional statistical models by a significant margin, with XGBoost achieving notable prediction accuracy due to its ability to model complex relationships. Additionally, data visualization with Power BI supported actionable insights for strategic decision-making. The research also investigates a hybrid forecasting method combining moving averages with neural networks, showcasing promising results. This work highlights the worth of machine learning in attractive sales forecast reliability and paves the way for incorporating deep learning in future studies.

Keywords — **Sales forecasting, machine learning, XGBoost, PySpark, Random Forest, K-Nearest Neighbor, Big Mart dataset, inventory optimization, predictive modeling, data visualization.**

1. INTRODUCTION

Sales forecasting plays a crucial role across sectors like retail, logistics, manufacturing, marketing, and wholesale. Reliable demand predictions empower organizations to maintain optimal inventory, enhance supply chain efficiency, and ensure effective resource allocation and formulate effective marketing strategies, all of which contribute to maximizing revenue and long-term growth. Conventional forecasting techniques, which primarily rely on statistical and mathematical models, frequently face limitations in recognizing and adapting to the intricate, nonlinear patterns present in real-world

data exhibited by modern sales data, especially in large-scale and dynamic retail environments. The advent of big data and advancements in computational power have paved the way for Machine learning methods have gained prominence as robust alternatives, capable of managing large-scale datasets and uncovering valuable patterns that enhance forecasting precision. This research seeks to investigate the effectiveness of different machine learning approaches applied to the Big Mart sales dataset and evaluate their predictive performance relative to conventional models. The goal is to identify models that can provide reliable sales forecasts to support decision-making in retail management.

2. LITERATURE SURVEY

Over the past decade, numerous approaches to sales forecasting have been explored, ranging from classical time series analysis to modern machine learning methods. Traditional models such as Moving Average, Exponential Smoothing, and Autoregressive Integrated Moving Average (ARIMA) have been widely used due to their interpretability and simplicity. However, these methods often fail to capture complex seasonal trends and nonlinear relationships in sales data. More recent studies have focused on applying This study considers a series of machine learning representations—such as Result Trees, Random Forestry, Support Vector Machines (SVM), Incline Enhancing, and Neural Links—that validate enhanced suppleness and advanced predictive accurateness. For example, Random Forestry and XGBoost have demonstrated robustness in handling high-dimensional data with multiple interacting features. Cross models merging statistical and machine learning techniques, such as integrating Moving Average with Artificial Neural Networks (ANN), have also shown promise. Despite these advances, challenges remain in managing large datasets efficiently, selecting optimal features, and balancing model complexity with interpretability.

3. PROPOSED WORK

This research proposes a comprehensive sales forecasting framework that leverages the Big Mart dataset, a rich source of retail sales data containing product details, store characteristics, and promotional information. The study focuses on developing predictive models using machine learning algorithms implemented in PySpark, a distributed computing framework suitable for processing large-scale datasets. Key features such as product weight, visibility, category, outlet type, and promotional indicators are used as inputs. The framework evaluates multiple models, including K-Nearest Neighbor (K-NN), Decision Trees, Linear, Polynomial and Ridge Regression, Random Forest, ARIMA, and XGBoost. The objective is to benchmark the predictive accuracy of these models through rigorous evaluation

metrics and identify the best-performing algorithm for retail sales forecasting.

4. METHODOLOGY

The methodology involves several phases: data preprocessing, model development, training, evaluation, and visualization. First, the Big Mart dataset undergoes cleaning to handle missing values and outliers, followed by feature engineering to derive relevant attributes such as promotional flags, seasonal indicators, and product categorization. Information is then divided into training and testing subsections to authorize typical presentation. Machine learning models are trained using PySpark to exploit parallel processing, enabling faster computation on the large dataset. Hyperparameter tuning is performed to optimize each model's predictive capacity. To evaluate model performance and predictive reliability, metrics such as Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2) are employed.

5. IMPLEMENTATION

The experimental environment consists of a PySpark cluster configured for distributed data processing, ensuring scalability and efficiency. Models are implemented using PySpark's MLlib library and The process is further supported by Python libraries where required, with the workflow encompassing data loading, preprocessing, and model training using cross-validation techniques and generating predictions on the test set. Additionally, Power BI is utilized to create dynamic dashboards that visualize forecasting results and key sales trends, enabling business users to derive actionable insights. MATLAB is employed for implementing a hybrid forecasting approach combining moving average and Artificial Neural Networks (ANN), providing an additional comparative perspective.

6. DATASET

The Big Mart dataset contains over 8,000 records encompassing diverse product categories and multiple outlet types. Key variables include Item Identifier, Weight, Visibility, Category, MRP (Maximum Retail Price), Outlet Identifier, Outlet Type, Location Type, and Outlet Size. The dataset includes sales figures, which serve as the target variable for prediction. Missing values, particularly in item weight and outlet size, are imputed using mean substitution and domain knowledge-based rules. Exploratory data analysis reveals seasonal fluctuations and the impact of promotions on sales volume, which justify their inclusion as model features.

7. MODELS USED FOR FORECASTING

A range of forecasting models is implemented:

- **K-Nearest Neighbor (K-NN):** Uses similarity between data points for

prediction, simple yet effective in certain scenarios.

- **Linear, Polynomial, Ridge Regression:** Serve as baseline statistical models capturing linear and nonlinear relationships.
- **Decision Trees and Random Forest:** Tree-based models that handle nonlinearity and interactions effectively.
- **XGBoost:** An ensemble gradient boosting technique known for its accuracy and speed.
- **ARIMA:** A classical time series model for comparison.
- **Hybrid Model:** Combines Moving Average and ANN to leverage strengths of both statistical and deep learning methods.

8.CRITERIA FOR COMPARATIVE ANALYSIS

Models are evaluated primarily on three metrics:

- **RMSE (Root Mean Square Error):** Captures the typical magnitude of prediction errors, with larger errors receiving greater penalty.
- **MAE (Mean Absolute Error):** Reflects the mean of absolute differences between observed and predicted outcomes.
- **R-squared (R^2):** Represents the share of variance in the target variable accounted for by the model, serving as an indicator of overall fit.

These metrics provide a balanced view of accuracy and reliability, crucial for practical forecasting applications.

9. EXPERIMENTAL RESULTS

The findings reveal that ensemble-based machine learning techniques, especially XGBoost and Random Forest, deliver superior performance compared to conventional regression and time-series approaches. XGBoost achieved the lowest RMSE (~1081) and highest R^2 (~0.59), demonstrating its ability to capture complex feature interactions and nonlinear patterns. Random Forest showed competitive performance with about a 20% reduction in prediction error compared to classical methods. K-NN and linear regression models lagged behind, confirming limitations in handling high-dimensional and nonlinear data. The hybrid moving average-ANN approach implemented in MATLAB showed promising accuracy but requires further tuning. Visualizations via Power BI revealed seasonal sales peaks and promotion-driven spikes, providing practical insights for inventory management.

10. CONCLUSION

This study confirms that machine learning techniques significantly enhance sales forecasting accuracy over traditional methods, especially when leveraging rich feature sets and large datasets. Among tested models, XGBoost stands

out for its superior predictive power and robustness. The integration of PySpark enables scalable processing of big retail data, while visualization tools like Power BI facilitate informed business decision-making. The exploration of hybrid models combining statistical and neural network approaches offers an exciting avenue for future research. Organizations seeking to improve resource allocation, inventory control, and revenue management can benefit greatly from adopting these advanced analytical methods. Future work may include incorporating deep learning architectures and real-time forecasting to further boost accuracy and responsiveness.

11. REFERENCE

[1] Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and Machine Learning forecasting methods: Concerns and ways forward. PLOS ONE, 13(3), e0194889.

[2] Carbonneau, R., Laframboise, K., & Vahidov, R. (2008). Application of machine learning techniques for supply chain demand forecasting. European Journal of Operational Research, 184(3), 1140–1154.

[3] Zhang, G., Eddy Patuwo, B., & Hu, M. Y. (2003). Forecasting with artificial neural networks: The state of the art. International Journal of Forecasting, 14(1), 35–62.

[4] Lemke, C., Gabrys, B., & Buhmann, J. M. (2009). Automatic selection of spectral channels using random forests. Pattern Recognition Letters, 30(9), 879–886.

[5] Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32.

[6] Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice (2nd ed.). O Texts.

[7] Bandara, K., Bergmeir, C., & Smyl, S. (2020). Forecasting across time series databases using recurrent neural networks on groups of similar series: A clustering approach. Expert Systems with Applications, 140, 112896.

[8] Crime Prediction using K Nearest Neighbour Algorithm by Akash Kumar, Aniket Verma,Gandhali Shinde,YashSukhdeve,Nidhi Lal published by 2020 International Conference on Emerging Trends in Information Technology and Engineering.

[9] House Price Prediction Using Regressio Techniques: A Comparative Study by CH.Raga Madhuri, Anuradha G, M.VaniPujitha published by IEEE 6th International Conference on smart structures and systems 2019.

[10] Deepa Rani Gopagoni, P V Lakshmi and Ankur Chaudhary, “Evaluatıng Machıne Learnıng Algorıthms For Marketıng Data Analysıs -Predıctıng Grocery Store Sales” https://www.researchgate.net/publication/344508907-2019.

[11] Rising Odegua, “Applied Machine Learning for Supermarket Sales Prediction”,

https://www.researchgate.net/publication/338681895, 2020.

[12] Grigorios Tsoumakas, "A survey of machine learning techniques for food sales prediction", Artif Intell Rev, https://doi.org/10.1007/s10462-018-9637-z-Springer-2018.

[13] Yuta Kaneko and Katsutoshi Yada, "A Deep Learning Approach for the Prediction of Retail Store Sales", IEEE 16th International Conference on Data Mining Workshops-2016.

[14] Purvika Bajaj, Renesa Ray2, Shivani Shedge, Shravani Vidhate, Prof. Dr. Nikhilkumar Shardoor, "sales prediction using machine learning algorithms", International Research Journal of Engineering and Technology (IRJET), Volume: 07 Issue: 06 | June 2020

[15] Bohanec, M., Robnik-Šikonja, M. and Borštnar, M.K., 2017. Organizational learning supported by machine learning models coupled with general explanation methods: A Case of B2B sales forecasting. Organizacija, 50(3), pp.217-233.

[16] Cheriyan, S., Ibrahim, S., Mohanan, S. and Treesa, S., 2018, August. Intelligent sales prediction using machine learning techniques. In 2018 International Conference on Computing, Electronics & Communications Engineering (iCCECE) (pp. 53-58). IEEE.

[17] Hofmann, E. and Rutschmann, E., 2018. Big data analytics and demand forecasting in supply chains: a conceptual analysis. The international journal of logistics management, 29(2), pp.739-766.

[18] Sivarajah, U., Kamal, M.M., Irani, Z. and Weerakkody, V., 2017. Critical analysis of Big Data challenges and analytical methods. Journal of business research, 70, pp.263-286.

[19] Raj Theeng Tamang, M., Sharif, M.S., Al-Bayatti, A.H., Alfakeeh, A.S. and Omar Alsayed, A., 2020. A machine-learning-based approach to predict the health impacts of commuting in large cities: Case study of London. Symmetry, 12(5), p.866.

[20] Barbierato, E. and Gatti, A., 2024. The challenges of machine learning: A critical review. Electronics, 13(2), p.416.

[21] Tsoumakas, G., 2019. A survey of machine learning techniques for food sales prediction. Artificial Intelligence Review, 52(1), pp.441- 447.