Rashmi Amardeep, 2024, 12:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology

An Open Access Journal

Machine Learning-Based Sales Forecasting and Inventory Optimization

Associate Professor Rashmi Amardeep, Prakhar Anant, Priya B Gunali, Riya Raikar, S Deepak Dhore Reddy
Department of Information Science and Engineering,
Dayananda Sagar Academy of Technology and Management, Bengaluru, India

Abstract- This project introduces a novel approach to sales prediction utilizing machine learning algorithms, aiming to revolutionize sales forecasting and enhance business decision-making processes. By leveraging advanced machine learning techniques, the project analyzes historical sales data to predict future sales trends accurately. The methodology involves the implementation of three prominent machine learning algorithms—Long Short-Term Memory, ARIMA and XG Boost to identify patterns and correlations within the dataset. Key features such as temporal factors, promotional activities, and market dynamics are integrated into the analysis to capture the intricacies of sales behavior the system offers an intuitive interface for entering fresh data and generating sales predictions in real time. Performance metrics assessment underscores the effectiveness of each algorithm in precisely forecasting sales figures. The project's predictive analytics capabilities empower businesses to optimize inventory management, devise targeted marketing strategies, and allocate resources effectively. By leveraging the capabilities of machine learning, companies can acquire valuable insights into consumer behavior and market trends. This empowers proactive decision-making and fosters sustainable growth within the competitive business environment.

Keywords- Sales forecasting, Machine learning algorithms, Long Short-Term Memory (LSTM), Autoregressive Integrated Moving Average (ARIMA), Extreme Gradient Boosting (XGBoost), Historical sales data, Temporal factors, Promotional activities, Inventory management, Predictive analytics

I. INTRODUCTION

In response to the evolving landscape of the retail sector in India and the imperative for enhanced operational efficiency, our project endeavors to address the challenges and opportunities inherent in sales prediction through the application of machine learning (ML) algorithms. Our endeavor seeks to revolutionize traditional sales forecasting methodologies by integrating advanced ML techniques to offer more accurate, timely, and scalable predictions. The retail industry in India is witnessing a paradigm shift, driven by the proliferation of digital technologies and the imperative for data- driven decision-making. Traditional sales forecasting methods, relying on intuition and historical data alone, fall short in capturing the dynamic nature of consumer behavior and market dynamics. Hence, there is an urgent need for innovative solutions that can leverage the vast amounts of sales data generated by retailers to provide actionable insights and anticipate future trends. Our project seeks to fill this void by utilizing three robust ML algorithms: Autoregressive Integrated Moving Average (ARIMA), eXtreme

Gradient Boosting (XGBoost), and Long Short-Term Memory (LSTM) networks..These algorithms offer distinct advantages in modeling time-series data, capturing complex relationships, and handling nonlinear patterns, respectively. By integrating these algorithms into our sales prediction framework, we aim to enhance predictive

By ensuring precision, scalability, and adaptability, this empowers retailers to make well-informed decisions, staying ahead of competitors.

In this paper, we present a comprehensive analysis of our sales prediction system, detailing the methodology, implementation approach, and experimental results.

We demonstrate the effectiveness of each algorithm individually and explore ensemble techniques to further improve prediction performance. Additionally, we discuss the implications of our findings for the retail industry in India, highlighting the potential for increased operational efficiency, customer satisfaction, and profitability through the adoption of advanced ML-based sales forecasting solutions.

II. LITERATURE SURVEY

1. Comparison Study: Product Demand Forecasting with Machine Learning for Shop

This research evaluates various demand prediction models, such as KNN, Random Forest, FNN, ANN, and Holt-Winters, using Mean Absolute Percentage Error (MAPE). Our approach introduces a methodology that utilizes K Nearest Neighbor, Gaussian Naive Bayes, and Decision Tree Classifier.

It takes into account factors like customer behavior, seasonal weather patterns, time, occasions, months, and product categories. In our local market, the Gaussian Naive Bayes algorithm demonstrates the best accuracy at 58.92%. Future implementation in the market will further refine the model, exploring additional decision-making attributes and incorporating diverse datasets from different city segments.

2. A Hybrid Machine Learning Model for Sales Prediction

This paper introduces a novel approach for predicting Wal-Mart's sales by combining XGBoost and LightGBM frameworks, emphasizing feature engineering for data processing. The models built on these frameworks outperform traditional methods like Logistic Regression and SVM in experiments. The integrated model, merging XGBoost and LightGBM, demonstrates superior predictive capabilities, showcasing its potential as a robust solution for sales prediction in retail contexts.

3. Sales Prediction Model for Big Mart

This research delves into the utilization of machine learning algorithms for sales prediction at Big Mart. Highlighting the evolving significance of machine learning in managing large datasets, the authors introduce several algorithms, such as Linear Regression, K-Nearest Neighbors, Decision Tree, Naïve Bayes Classifiers, Random Tree, and K-means Clustering. The study aims to discern the influence of item properties on sales at Big Mart. formulating a predictive model for each store. The methodology involves preprocessing a dataset from 10 stores and 1559 products, utilizing algorithms like Random Forest and linear regression for accurate sales predictions. Overall, the paper provides a concise overview of using machine learning for sales prediction, specifically in the context of a comprehensive dataset from a large shopping center like Big Mart.

4. Sales Forecasting with Machine Learning Algorithms

This paper explores machine learning algorithms for sales prediction, utilizing four algorithms, including an ensemble technique. It aims to predict future sales using clustering models and measures. The research involves decisions from experimental data and insights from data visualization, employing data mining techniques. It outlines a sales prediction system and product recommendation system for retail, leveraging consumer demographics. Deep neural networks are applied for understanding electronic component sales strategy. Bayesian learning, coupled with neural

networks, predicts individual retailer sales rates, benefiting a large number of outputs. The study detects suspicious behavior using an automatic prototype, merging various machine learning methodologies for effective user profiling.

5. A Predictive Analysis of Retail Sales Forecasting using Machine Learning Techniques

The paper emphasizes sales forecasting's critical role in information technology chain stores, Influencing inventory management, marketing strategies, customer service, and financial planning, this study underscores challenges such as over and under-forecasting.

It proposes a combination of human expertise and statistical models to ensure accuracy. By employing LSTM, ARIMA, Linear Regression, Random Forest, and XGBoost models, the research identifies XGBoost as the most fitting for the Citadel POS dataset. Future directions entail exploring deep learning methodologies with expanded datasets to improve precision in retail sales forecasting.

6. Predicting Big Mart Sales with Various Machine Learning

Techniques This paper underscores the critical role of accurate sales predictions for businesses in maintaining standards and enhancing profitability. The study formulates hypotheses pertaining to store and product attributes influencing sales, conducts comprehensive data exploration, and employs Utilizing machine learning models such as linear regression, ridge regression, decision tree, and random forest to forecast store sales.

Findings show enhancements compared to a baseline model, with ridge regression exhibiting superior performance in the analysis. public leaderboard. The paper identifies limitations, such as the exclusion of disaster-related factors, and suggests future work, including incorporating these elements and developing an online app for customer reviews and rankings. Overall, the study contributes to the field by proposing a robust framework for mart sales prediction and evaluating various machine learning models' performance.

III. METHODOLOGY

Initially, we conducted a literature review to gather relevant research findings, which serve as foundational input for In our examination of retail sales through machine learning approaches, our main aim is to evaluate the efficacy of LSTM (Long Short- Term Memory), ARIMA (Autoregressive Integrated Moving Average), and XGBoost (Extreme Gradient Boosting) models applied to point-of-sale sales data. Figure 3.1 outlines the detailed methodology of our proposed solution..

1. Data Collection

We collected sales data from various sources, including internal databases, POS systems, and online platforms (Kaggle). This data encompassed crucial features such as date, product ID, quantity sold, price, customer information, location, and brand. We ensured data integrity by verifying its accuracy and completeness, while also adhering to privacy regulations to safeguard sensitive information.

2. Data Preprocessing

Our data cleaning process involved identifying and removing duplicate records to eliminate redundancy and maintain data consistency. We addressed missing values (is null) using appropriate strategies such as imputation or removal, ensuring that our analysis was based on complete and reliable data. Data validation techniques were employed to detect and rectify inconsistencies or errors in the dataset, ensuring its quality and reliability for further analysis. The clean data was analyzed using the following methods:

Data Visualization

To gain insights into sales trends and patterns, we visualized the data using various charts and graphs. Key metrics such as sales by brand, location, reorder quantity, and approximate total value were represented visually using tools like bar plots, line charts, scatter plots, and histograms. Visualizations helped us understand the distribution of sales across different categories and identify trends or anomalies that could inform decision-making.

Correlation Matrix

We calculated the correlation between variables such as reorder quantity and approximate total value to assess the strength and direction of relationships. A correlation matrix or heatmap was generated to visualize these relationships, providing insights into how different variables interacted with each other.

Location-Based Filtering

By filtering sales data based on specific locations or regions, we analyzed regional sales trends and identified geographical areas with the highest sales performance. Through this analysis, we were able to efficiently allocate resources and customize marketing strategies to target particular regions.

Brand-Based Filtering

Similarly, filtering sales data based on brands or product categories allowed us to assess brand performance and market share. By comparing sales metrics across different brands, we identified top-performing brands and areas for potential improvement.

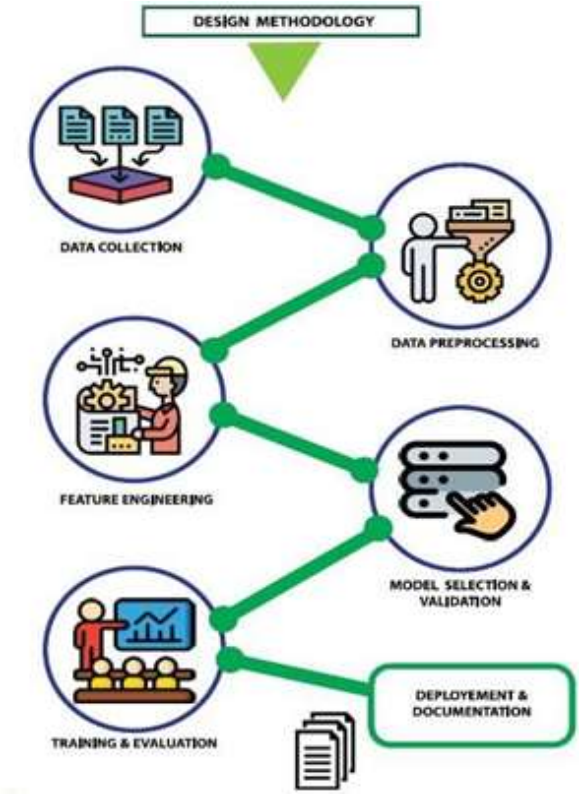

Fig 1: Design Methodology

3. Feature Engineering

Feature engineering proved pivotal in our project, refining our dataset to enable optimal modeling. We not only created new features and transformed existing ones but also ensured compatibility for machine learning algorithms. Handling missing values, encoding categorical variables, and incorporating time-related features were key steps in this process. Additionally, we introduced novel features, which provided deeper insights into consumer preferences and purchasing behavior. These efforts collectively bolstered the predictive power of our model, enabling more accurate sales predictions.

4. Model Selection and Validation

XGBoost: XGBoost, a gradient boosting algorithm, was utilized for sales prediction based on historical data and pertinent features. Training the model involved historical sales data, and hyper parameters were adjusted to enhance its performance and precision. Cross-validation techniques were employed to evaluate the model's performance and ensure its robustness.

ARIMA (AutoRegressive Integrated Moving Average): Employing ARIMA as a time series forecasting technique, we utilized it to capture temporal patterns within the sales data. Parameters like the order of differencing (d), autoregressive (p), and moving average (q) were chosen based on the stationarity of the data. The ARIMA model was trained on historical sales data and validated using out-of-sample testing to assess its predictive capability.

LSTM (Long Short-Term Memory): LSTM, belonging to the category of recurrent neural networks (RNN), was employed for forecasting sequential data. The sequential sales data underwent preprocessing to form input-output pairs conducive for LSTM training. The model was trained using historical sales sequences, and parameters were optimized to improve its forecasting accuracy.

5. Training and Evaluation

In the training phase of our model using XGBoost, ARIMA, and LSTM algorithms, we meticulously fine-

tuned each model to optimize performance and accuracy. With XGBoost, we employed gradient boosting techniques to iteratively improve the model's predictive capability, adjusting hyperparameters through cross-validation to achieve optimal results For ARIMA, meticulous parameter selection was conducted, encompassing factors like the order of differencing and the autoregressive and moving average components. This ensured the model adeptly captured temporal patterns. With LSTM, a recurrent neural network architecture, we trained the model on historical sales sequences, optimizing parameters to capture complex temporal dependencies in the data. Through these training processes, we aimed to develop robust models capable of making accurate sales predictions.

In evaluating the performance of our predictive models, several key metrics provide insights into their accuracy and effectiveness. Mean Squared Error (MSE) acts as a foundational metric, measuring the average squared deviation between predicted and actual values, thereby indicating the overall accuracy of our predictions. Additionally, Mean Absolute Error (MAE) provides a straightforward evaluation by calculating the average absolute difference between predicted and actual values, offering a clear insight into the magnitude of prediction errors. Furthermore, R-squared (R^2) elucidates the goodness of fit of our models, offers a more intuitive interpretation, calculating the square root of the MSE and providing a measure of prediction error in the original units of the data.

6. Deployment and Documentation

A user-friendly web interface was developed to provide stakeholders with easy access to sales predictions and insights. Frontend components were designed using HTML, CSS, and JavaScript frameworks to ensure a seamless user experience. Backend functionality was integrated into the web interface to fetch and display sales forecasts, reports, and visualizations. The interface was optimized for responsiveness and cross-browser compatibility to accommodate users accessing the system from different devices and browsers.

Forecasts generated by the XGBoost, ARIMA, and LSTM models were combined to provide comprehensive sales predictions. Detailed reports summarizing sales forecasts, trends, and insights were generated to facilitate decision-making by stakeholders. Visualizations such as charts, graphs, and tables were included in the reports to present findings effectively and aid in data interpretation.

IV. RESULT & DISCUSSION

The results and discussions section provides a comprehensive analysis of the performance of various predictive models employed in the project. It presents detailed insights into the effectiveness of each model in forecasting sales accurately, highlighting their respective strengths and limitations.

By comparing performance metrics like Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and coefficient of determination (r-squared), this section provides valuable insights into the predictive capabilities of each model. Additionally, the discussions delve into the implications of the findings, emphasizing the significance of accurate sales forecasting in optimizing business operations, enhancing decision-making processes, and achieving organizational objectives.

1. Predictive Analysis

XGBoost Model

XGBoost, incorporating three essential components—extreme, gradient, and boosting—employs boosting, a crucial concept in systematic ensemble methods, to enhance weak learners such as regression trees.

Through iterative refinement, new models are added sequentially to correct errors from previous iterations, creating a more robust predictive model. By combining multiple decision trees and fitting new ones to residuals of preceding ones, XGBoost effectively minimizes prediction errors, making it a powerful tool for enhancing predictive accuracy in handling complex data relationships.

Table 4.1: XGBoost Model Performance Results

Index	Score
RMSE	0.28742880
MSE	0.08261531
MAE	0.24680418
R-squared(R2)	0.75009001

Table 4.1 showcases The Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) achieved via Gradient Boosting Regression on the validation test were examined. RMSE signifies the standard deviation of the prediction error, demonstrating the extent to which data points diverge from the regression line (0.28742). In contrast, MAE assesses the average magnitude of the error. Irrespective of its direction (0.246804).

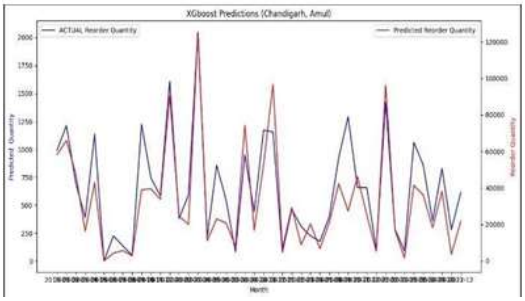

Fig 2: XGBoost Model Forecasting Performance

Figure 4.1 The visualization portrays the actual and predicted sales of the target variant, generated from the Gradient Boosting regression model. In this depiction, the blue line represents the actual sales values, whereas the red line indicates the forecasted sales of the targeted variant.

2. LSTM Model

LSTM models have been utilized on datasets to predict sales using historical retail sales data, commonly applied to detect patterns in financial market datasets such as stock market data.

Table 2: LSTM Model Performance Results

Index	Score
MAE	1035.65554
MSE	2067311.01
RMSE	1437.81466
MAP	4.89249106

Table 4.2 illustrates the outcomes of Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) from LSTM Regression on the validation set. RMSE serves as a metric for the standard deviation of prediction errors, indicating the extent of deviation of data points from the regression line, with a value of 1437.81466. Meanwhile, MAE quantifies the average error magnitude, independent of the direction between actual and predicted observations, with a value of 1035.6555.

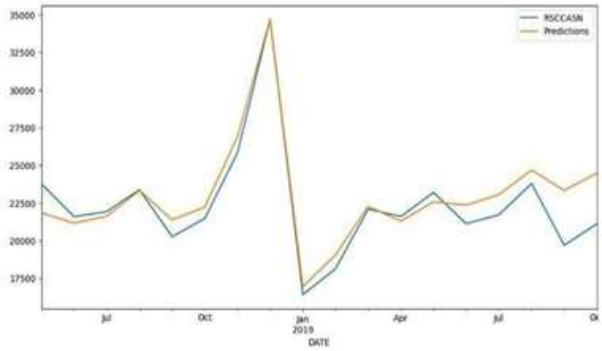

Fig 3: LSTM Model Prediction Performance

Figure 4.2 illustrates The visualization shows the actual and predicted sales of the target variant, obtained from the LSTM regression model. In this representation, the blue line denotes the actual sales values, while the red line represents the forecasted sales of the targeted variant.

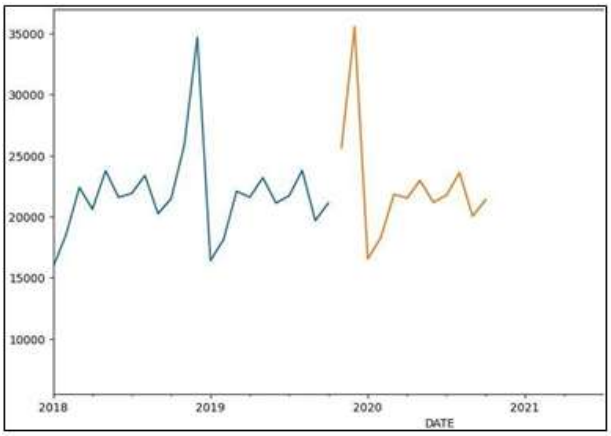

Fig 4: LSTM Model Forecasting Performance

Figure 4.3 depicts a forecast for brand AMUL in Chandigarh for the year 2020 based on learning from previous year data. The blue line represents the original data, whereas the red line depicts the

forecasted values. By comparing these lines, the accuracy of the model be evaluation and identification of any discrepancies for potential model refinement.

3. ARIMA Model

The ARIMA model is instrumental in forecasting sales trends, employing statistical methods tailored for time series analysis. It encompasses several parameters:

- P: Represents the autoregression order for trend analysis.
- D: Denotes the order of trend differentiation.
- Q: Specifies the moving average order for trend analysis.

Moreover, the ARIMA model does not include seasonal differences, which are treated independently using the SARIMA model, identified as SARIMA (p, d, q) (P, D, Q) m.

Table 3: ARIMA Model Performance Results

Index	Score
MSE	66375431.14
MAE	7027.662325
R2	0.375734632

Table 4.3 displays the performance metrics of the ARIMA model on the validation test dataset. The Mean Squared Error (MSE), which represents the standard deviation of prediction errors, is computed at 66375431.14.

Fig 5: ARIMA Model Forecasting Performance

Additionally, the Mean Absolute Error (MAE), reflecting the average magnitude of prediction errors, registers at 7027.662325. Additionally, the coefficient of determination (r-squared), representing the proportion of the variance in the dependent variable that is predictable from the independent variable, is reported at 0.37573463.

Figure 4.4 The visualization illustrates the actual and predicted sales of the target variant, as determined by the ARIMA regression model. Here, the blue line represents the actual sales values, while the red line indicates the forecasted sales of the targeted variant.

The comparative analysis of XGBoost, LSTM, and ARIMA models for sales prediction underscores the importance of selecting the appropriate forecasting method for optimal results. XGBoost emerges as the preferred model due to its superior performance in minimizing prediction errors, as evidenced by the lowest RMSE and MAE values among the three models. This indicates its effectiveness in handling complex data relationships and generating more accurate sales forecasts. In contrast, LSTM exhibits significantly higher errors, indicating less precise predictions compared to XGBoost. Additionally, ARIMA demonstrates the highest errors, suggesting relatively poorer performance in sales prediction compared to XGBoost and LSTM. While ARIMA yields a moderate coefficient of determination (r-squared), highlighting its moderate level of predictability, XGBoost' s superior accuracy underscores its effectiveness in meeting the demands of sales forecasting tasks. The findings emphasize the importance of leveraging advanced machine learning techniques, such as XGBoost, to enhance sales prediction accuracy and inform strategic decision- making processes in inventory management, marketing, and financial planning. Moving forward, further research may explore the integration of deep learning approaches and the expansion of dataset size to further improve forecasting accuracy, particularly in the context of large retail sales datasets. Ultimately, accurate sales forecasting is pivotal for achieving customer satisfaction, optimizing channel relationships, and

realizing significant cost savings in business operations.

V. FUTURE ENHANCEMENTS

These future improvements are poised to further elevate the project's impact and unlock new opportunities for businesses to leverage predictive analytics for strategic decision-making and operational optimization. The future enhancements to the system could include:

1. Integration of External Data Sources
Integrating supplementary external data sources like weather forecasts, social media trends, and economic indicators could enhance the accuracy and reliability of sales forecasts even further.

2. Advanced Feature Engineering
Investigating advanced feature engineering techniques to extract deeper insights from the data and capture intricate relationships between variables could augment the predictive capability of the system.

3. Ensemble Learning
Implementing ensemble learning techniques, such as model stacking or blending, could help combine the strengths of multiple machine learning algorithms and improve overall prediction performance.

4. Dynamic Model Updating
Creating mechanisms for dynamically updating predictive models in response to shifts in market conditions, customer behavior, and other pertinent factors could guarantee that forecasts remain accurate and up-to-date over time.

5. Customer Segmentation
Segmenting customers according to their purchasing behavior, demographics, and preferences could facilitate personalized sales forecasts and tailored marketing strategies.

Overall, the project lays the foundation for leveraging advanced data analytics techniques to drive business growth, improve operational efficiency, and gain a competitive edge in today's dynamic marketplace. By embracing innovation and continuously refining the predictive models, businesses can stay ahead of the curve and thrive in an increasingly data- driven world.

VI. CONCLUSION

When assessing the performance of XGBoost, LSTM, and ARIMA models for sales prediction, Table 4.1 demonstrates that XGBoost achieves the lowest RMSE (0.28742) and MAE (0.246804) values among the three models, indicating superior predictive accuracy. Conversely, LSTM exhibits a considerably higher RMSE (1437.81466) and MAE (1035.6555) compared to XGBoost, suggesting less accurate predictions. Additionally, ARIMA demonstrates the highest errors, with an MSE of 66375431.14 and MAE of 7027.662325, highlighting its relatively poorer performance compared to XGBoost and LSTM. While ARIMA yields a coefficient of determination (r-squared) of 0.37573463, indicating a moderate level of predictability, XGBoost' s superior accuracy, particularly in minimizing prediction errors, underscores its effectiveness in handling complex data relationships. Therefore, based on the comparative analysis, XGBoost emerges as the preferred model for sales prediction due to its superior performance and predictive accuracy.

In conclusion, our project presents an innovative solution to the challenges encountered by businesses in accurately forecasting sales By leveraging advanced machine learning algorithms like ARIMA, LSTM, and XGBoost, the project endeavors to provide highly accurate and reliable sales forecasts. Through the integration of these robust algorithms, the project addresses the shortcomings of traditional forecasting methods, offering enhanced predictive accuracy, streamlined processes, and the ability to capture complex data relationships. The proposed system boasts numerous advantages, including accuracy, adaptability, comprehensive insights, timeliness, usability, and scalability. By furnishing businesses with actionable insights into their sales performance, the system enables informed

decision-making, optimized inventory management, and enhanced operational efficiency. Moreover, with its real-time data analysis capabilities and user-friendly interface, the system is accessible to businesses of all sizes and technical expertise levels.

REFERENCES

1. Md. Ariful Islam Arif, Saiful Islam Sany, Faiza Islam Nahin and A.K.M. Shahariar Azad Rabby, "Comparison Study: Product Demand Forecasting with Machine Learning for Shop", Proceedings of the SMART–2019, IEEE Conference ID: 468668th International Conference on System Modeling & Advancement in Research Trends, pp. 171-176, 22nd–23rd November, 2019.
2. Muhammad Sajawal, Sardar Usman, Hamed Sanad Alshaikh, Asad Hayat and M. Usman Ashraf, "A Predictive Analysis of Retail Sales Forecasting using Machine Learning Techniques", LGU Research Journal of Computer Science & IT, Vol. 06 Issue 04, pp. 33-43, October – December 2022, doi:10.54692/lgurjcsit.2022.0604399.
3. Purvika Bajaj, Renesa Ray, Shivani Shedge, Shravani Vidhate and Prof.Dr. Nikhilkumar Shardoor, "Sales Prediction using Machine Learning Algorithm", International Research Journal Of Engineering & Tech, Vol. 07 Issue 06, pp. 3619-3625, June 2020.
4. Nikita Malik and Karan Singh, "SALES PREDICTION MODEL FOR BIG MART", MSI Janakpuri, New Delhi, Vol. 03, pp. 22-32, July 2020.
5. Rao Faizan Ali, Amgad Muneer, Ahmed Almaghthawi, Amal Alghamdi, Suliman Mohamed Fati and Ebrahim Abdulwasea Abdullah Ghaleb,"BMSP-ML: big mart sales prediction using different machine learning techniques", IAES International Journal of Artificial Intelligence (IJ-AI), Vol. 12, No. 2, pp. 874-883, June 2023, doi:10.11591/ijaiv12.i2.
6. Jingru Wang,"A hybrid machine learning model for sales prediction", International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI), pp. 363-366, 2020, doi:10.1109/ICHCI51889.2020.00083.
7. Michael Giering. "Retail Sales Prediction and Item Recommendations Using Customer Demographics at Store Level", ACM SIGKDD Explorations Newsletter, Volume 10, Issue 2 , pp. 84–89, December 2008.
8. Ruiyun Kang,"Sales Prediction of Big Mart Based on Linear Regression, Random Forest, and Gradient Boosting",Proceedings of the 2nd International Conference on Business and Policy Studies Vol. 17,pp. 200- 207, 13 September 2023.
9. Sai Nikhil Boyapati and Ramesh Mummidi, "Predicting sales using Machine Learning Techniques", DV1478 Bachelor Thesis in Computer Science, this thesis is submitted to the Faculty of Computing at Blekinge Institute of Technology in 2020.
10. Aakanksha Ramesh Jadhav and Dr Ramesh D Jadhav, "Machine Learning for Sales Prediction in Big Mart". Sinhgad Institutes - Sinhgad Institute of Technology and Science, August 3, 2023.