SALES FORECASTING USING MACHINE LEARNING

Ms.NANDHINI K, Ms.NANDHINI K, Mr.SACHIN M, Mr. ROHIT.R

Mrs.Revathi M (MENTOR)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SRI SHAKTHI INSTITUTE OF ENGINEERING AND TECHNOLOGY

---***---

Abstract - Sales forecasting remains a critical component in the strategic planning and operational efficiency of retail enterprises. This study presents a data-driven forecasting model tailored for retail sales prediction, using a real-world dataset comprising historical sales records, outlet characteristics, and product features. The methodology emphasizes robust data preprocessing, including handling of missing values, outlier treatment, and categorical encoding, followed by the application of ensemble-based regression. Among various algorithms evaluated, the Extreme Gradient Boosting Random Forest Regressor (XGBRFRegressor) demonstrated superior predictive performance, achieving consistent accuracy across cross-validation folds. To enhance practical applicability, the forecasting system was integrated into an interactive interface, enabling real-time prediction based on user-defined input parameters. The proposed approach offers a scalable and reliable framework for informed decision-making in retail operations, particularly in inventory management, demand planning, and revenue optimization.

Key Words: predicting, forecasting, demand planning,sales

1.INTRODUCTION

Sales forecasting plays a vital role in retail operations,enabling businesses to anticipate demand, manage inventory, and plan strategically. Traditional forecasting methods often fall short in capturing complex patterns within retail data, particularly when dealing with multiple interdependent variables. This project proposes a machine learning-based approach to forecast sales using the BigMart Sales dataset, which includes diverse features such as item attributes, outlet characteristics, and establishment timelines. Advanced regression techniques, especially the XGBRFRegressor, are utilized to model the nonlinear relationships present in the data and enhance prediction.

2.LITERATURE REVIEW

A. Traditional Methods in Sales Forecasting

Sales forecasting traditionally used methods like Linear Regression and Random Forest. While Linear Regression is simple, it struggles with complex data, whereas Random Forest improves accuracy by handling complex patterns but still has limitations with very large datasets.

B. Advanced Machine Learning Techniques

More advanced methods like XGBoost have emerged, as they can capture complex relationships in large datasets. XGBoost is more accurate than simpler models, especially when tuned properly and with feature selection.

C. Model Deployment

Modern retail forecasting emphasizes real-time predictions. By deploying models through user-friendly interfaces, businesses can receive instant sales forecasts based on product details, enhancing decision-making in real-time.

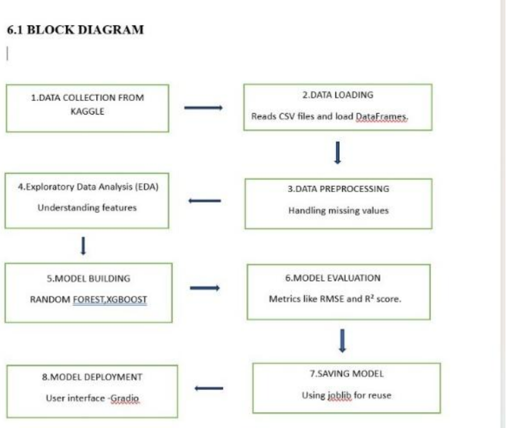
3.METHODOLOGY

Data Collection: The dataset used in this project contains historical sales data, including product details, store characteristics, and sales figures. Additional information, such as the maximum retail price (MRP), store size, location, and outlet type, was collected to enhance predictive modeling.

Feature Engineering: Relevant features were extracted from the dataset to enhance the model's predictive power. New variables, such as month, year, and holiday information, were derived to capture seasonality effects. Additionally, feature selection was performed using model-driven importance metrics to eliminate irrelevant or redundant features, ensuring the model remains efficient and interpretable.

Model Selection: Multiple machine learning algorithms were evaluated for sales forecasting, including Random Forest and XGBoost. Among these, the XGBRFRegressor was selected for its ability to capture complex, nonlinear relationships and handle feature interactions effectively.

Model Training: The model was trained on the preprocessed data using XGBoost and Random Forest regressors. The dataset was split into training and testing sets, with cross-validation employed to ensure generalization and prevent overfitting.

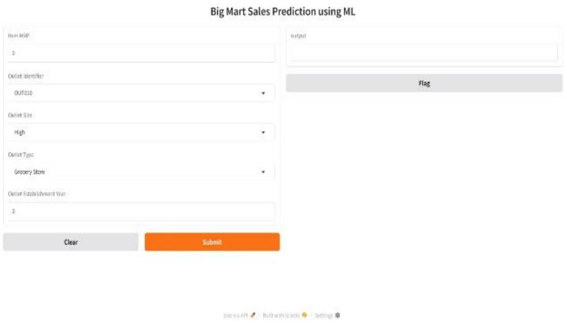


Model Evaluation: The model's performance was evaluated using metrics like R², Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). The XGBRF Regressor achieved an R² score of approximately 0.595 and a MAE of 714.42 on the test data, indicating reliable predictive accuracy.

Model Deployment: To facilitate real-time sales predictions, a lightweight, interactive application was developed using Gradio. This application allows users to input features such as item MRP, outlet size, and location type, and obtain instant sales forecasts. The model deployment ensures practical use in retail decision support systems, helping businesses optimize inventory and plan marketing strategies.

4.RESULT: The XGBRF Regressor model achieved a 0.595 R² score and 714.42 MAE, indicating reliable sales predictions. XGBoost and Random Forest effectively captured sales patterns. A Gradio app was created for real-time sales forecasting.

Output:


Big Mart Sales Prediction using ML

Item MRP
0

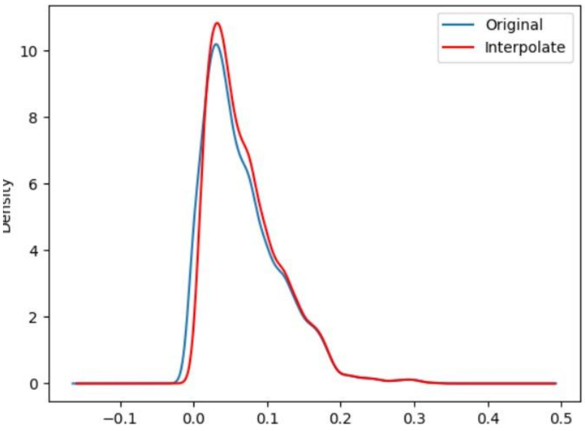
Outlet Identifier
OUT010

Outlet Size
High

Outlet Type
Grocery Store

Outlet Establishment Year
0

output



5.CONCLUSION

The sales forecasting project highlights the effectiveness of XGBoost and Random Forest in predicting sales using historical data and relevant features. The XGBRFRegressor achieved a strong R² score and low mean absolute error, demonstrating its predictive accuracy. The model was deployed through a Gradio interface for real-time sales predictions, aiding businesses in inventory planning, marketing, and decision-making. This project provides a scalable, data-driven solution for modern retail forecasting challenges.

ACKNOWLEDGEMENT

We express our gratitude to our mentors and teammates for their support throughout this project. Thanks to the availability of the dataset, which was key to building the model. We also appreciate the developers of **XGBoost** and **Random Forest** for providing essential tools for our forecasting solution.

REFERENCES

[1] Deepak, R., Sathyanarayanan, R., & Arunnehru, J. (2025). Demand and Sales Forecasting Using Random Forest and Linear Regression. In Smart Computing Paradigms: Advanced Data Mining and Analytics (pp. 525–540). Springer.

[2] Ganguly, P., & Mukherjee, I. (2024). Enhancing Retail Sales Forecasting with Optimized Machine Learning Models. arXiv preprint arXiv:2410.13773.

[3] Wang, M., Liu, Y., Li, G., Payne, T. R., Yue, Y., & Man, K. L. (2024). Unlocking Your Sales Insights: Advanced XGBoost Forecasting Models for Amazon Products. arXiv preprint arXiv:2411.00460.

[4] Swami, D., Shah, A. D., & Ray, S. K. B. (2020). Predicting Future Sales of Retail Products using Machine Learning. arXiv preprint arXiv:2008.07779.

[5] Zhang, Y., Wu, X., Gu, C., & Xie, Y. (2019). Predict Future Sales using Ensembled Random Forests. arXiv preprint arXiv:1904.09031.

[6] Kumar, M.S., Raut, D.R.D., Narwane, D.V.S., Narkhede, D.B.E., 2020. Applications of industry 4.0 to overcome the COVID-19 operational challenges. Diabetes Metab. Syndr. Clin. Res. Rev. 14,12831289. https://doi.org/10.1016/j.dsx.2020.07.010.

[7] De Gooijer, J.G., Hyndman, R.J., 2006. 25 years of time series forecasting. Int. J. Forecast., Twenty five Years Forecast. 22, 443–473. https://doi.org/10.1016/j.ijforecast.2006.01.001.

[8] Aderonke Anthonia Kayode, Noah Oluwatobi Akande, Adekanmi Adeyinka Adegun, Marion Olubunmi Adebiyi (2019), ―An automated mammogram classification system using modified support vector machine‖, Medical Devices: Evidence and Re-search, 12, 275—284.

[9] Kayode Anthonia Aderonke, Akande Noah Oluwatobi, Saheed O Jabaru, Oladele O Tinuke (2020), ―An Empirical Investigation of the Prevalence of Osteoarthritis in South West Nigeria: A Population-Based Study‖, International Journal of Online and Biomedical Engineering (iJOE), 16(1), 100-114.

[10] Oluwatobi Noah Akande, Oluwakemi Christiana Abikoye, Aderonke Anthonia Kayode, and Yema Lamari (2020), ―Implementation of a Framework for Healthy and Diabetic Retinopathy Retinal Image Recognition‖, Scientifica, Volume 2020, Article ID 4972527, pp. 1-14.

[11] Karmy, J.P., Maldonado, S., 2019. Hierarchical time series forecasting via Support Vector Regression in the European Travel Retail Industry. Expert Syst. Appl. 137, 59–73. https://doi.org/10.1016/j.eswa.2019.06.060.

[12] Wilson, J. H., Dingus, R., & Hoyle, J. (2020). Women count : Perceptions of forecasting in sales. Business Horizons, 63(5),637–646. https://doi.org/10.1016/j.bushor.2020.06.00

[13] Ahmed N. K., Atiya A. F., Gayar N. E., and El-Shishiny H., An empirical comparison of machine learning models for time series forecasting, *Econometric Reviews*. (2010) **29**, no. 5-6, 594621, https://doi.org/10.1080/07474938.2010.481556, 2-s2.0-77956724793.

[14] Box G. E. P. and Jenkins G. M., Time series analysis: forecasting and control, *Journal of Time*. (2010) **31**, no. 4, 303.