Bigmart Sales Prediction Using Lightgbm

Mr.K.Hari Verraju[1], Mr.K.Sai Kishore[2], Ms.G.Kusuma Kumari[3], Mr.S.Ravi Kiran[4] and Mr.J.Likhit[5]

[1]Assistant professor, Dept. of CSE, Raghu Engineering College, Dakamarri(V), Bheemunipatnam, Visakhapatnam District, 531162

[2345]Department of Data Science, Raghu Institute of Technology, Dakamarri(V), Bheemunipatnam, Visakhapatnam District, 531162

***Abstract*--Sales forecasting maintains an important position in retail because it aids businesses with inventory management and cost reduction to create strategic decisions. The research uses Light Gradient Boosting Machine (LightGBM) to forecast BigMart sales by analyzing historical data combined with product attributes and store-related information. The predictive model benefits from various feature engineering techniques which include One-Hot Encoding and Ordinal Encoding and Polynomial Features. The model performance reaches its peak through systematic adjustments of its hyperparameters. The experimental evaluation assesses LightGBM by examining its performance against Linear Regression as well as Decision Trees Random Forests and XGBoost through R² and Mean Squared Error (MSE) and Mean Absolute Error (MAE). LightGBM demonstrates superior performance compared to traditional models in terms of accuracy together with operational speed thereby capturing sophisticated sales relationship patterns. The application of LightGBM generates data-driven findings that enable BigMart to make enhanced pricing decisions, better control inventory flow, and boost operational operations. LightGBM shows strong capabilities for processing extensive retail data according to the research findings which establishes its value for forecasting sales in evolving market conditions. The research uses LightGBM to predict sales while employing Machine Learning approaches and performing Feature Engineering tasks alongside Hyperparameter Tuning techniques with Predictive Analytics methods.**

Keywords—Sales Forecasting, LightGBM, Machine Learning, Feature Engineering, Hyperparameter Tuning, Predictive Analytics

1. INTRODUCTION

1.1 Background and Motivation The exact prediction of future sales plays an essential role for retail operations to handle inventory effectively and adjust pricing structures for maximum earnings. Companies that fail to achieve accurate forecasts will end up facing shortages of stock while generating unnecessary expenses because of surplus inventory. Linear regression models alongside other traditional statistical methods are ineffective at analyzing complex datasets from retail domains because they fail to work effectively with categorical factors and seasonal information and promotional events. ML techniques provide better results because they derive complex patterns from previous sales data. LightGBM functions as a welldeveloped ML model that masters large datasets at optimal efficiency while generating accurate predictions. LightGBM utilizes gradient boosting techniques to manage categorical data input and extract secretive sales patterns from the dataset. So this research utilizes LightGBM to analyze BigMart sales data in order to improve forecasting precision and generate useful retail business recommendations.1.2 Research Gap Decision Trees and Random Forests are among conventional machine learning models which have received focus in previous research for sales prediction purposes. The examined models experience limitations when working with extensive categorical data features and they demonstrate limited applicability across different retail marketplaces. Many existing approaches lack essential implementations of advanced feature engineering techniques and hyperparameter optimization methods that yield essential benefits toward predictive performance improvements. The research activity fills these gaps through the implementation of LightGBM along with optimized hyperparameter selection and addition of feature engineering methods for better model accuracy. 1.3 Objectives The primary objectives of this study are: 1. The focus of this study involves developing a sales forecast system by implementing LightGBM on BigMart sales information. 2. Applying One-Hot Encoding together with Ordinal Encoding and Polynomial Features as feature engineering methods will enhance predictive effectiveness. 3.A process of hyperparameter optimization will be performed to achieve better model accuracy and efficiency results. 4. To compare LightGBM's performance with traditional models such as Linear Regression, Decision Trees, Random Forests, and XGBoost. 5. The system delivers data-based recommendations that help retailers improve their inventory management systems and their pricing approaches and enhance their sales projections.

2. LITERATURE REVIEW

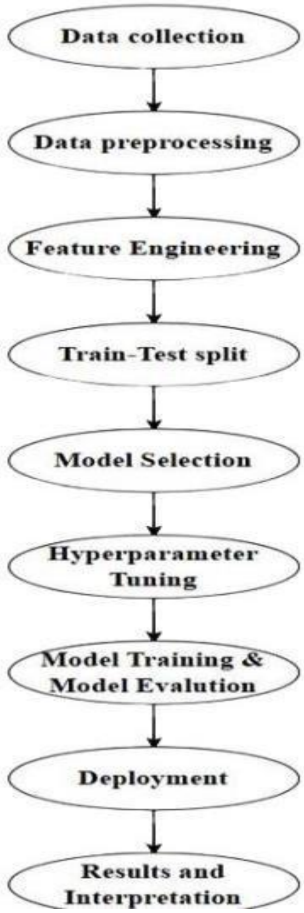
2.1. Introduction Retail operations strongly rely on sales forecasting to manage inventory and plan forecast demands along with making strategic decisions. Decision Trees, Random Forest and XGBoost have become successfully established traditional machine learning models for sales prediction tasks within numerous applications. Light Gradient Boosting Machine (LightGBM) stands out as a recent improvement in boosting techniques which delivers better efficiency and improved accuracy for sales prediction. BigMart sales prediction research is examined in this section to identify models and methodologies besides existing gaps where LightGBM intends to make progress. 2.2 Existing Studies on BigMart Sales Prediction 2.2.1

Traditional Machine Learning Approaches • Multiple research works have implemented classical machine learning techniques for predicting sales levels. Sales forecasting within the IJNRD (2024) research relied on Decision Trees, Random Forest, and Linear Regression. The predictive capabilities of these models remained basic yet they showed limitations when processing dimensional data and dealing with categorical variables. • ResearchGate conducted an analysis of Linear Regression and additional models like Ridge Regression and Lasso Regression and XGBoost for predicting sales trends (BMSP-ML, 2022). Accurate sales forecasting results from the study required both appropriate feature selection and thorough hyperparameter adjustment. • ResearchGate (2020) performed a comparative study which demonstrated Random Forest and XGBoost provided high-performance in regression models. Tree-based models face challenges involving computational cost as well as overfitting which suggests the development of efficient substitute forecasting solutions. 2.2.2 Feature Engineering and Sales Prediction The process of transforming raw information into features completes predictive accuracy enhancement. The research by IJERT (2021) analyzed techniques for dealing with missing values as well as different approaches for encoding categorical variables and selecting features. The study proved that adept preprocessing directly influences model output results. Studies published on ResearchGate (2023) stressed out the importance of various feature selection methods on different machine learning systems. The study shows that feature optimization remains crucial since LightGBM demonstrates excellent capability to process categorical variables even though it was not directly evaluated. 2.3 Advancements with LightGBM The research field before this work mainly utilized Decision Trees and Random Forest and XGBoost but LightGBM delivers superior capabilities. • The training process of LightGBM runs faster than XGBoost while needing less memory during operations. • LightGBM masterfully deals with categories in data without needing complicated preprocessing steps. • Higher accuracy on large datasets due to its leaf-wise growth strategy. • The model achieves better prediction results by employing efficient parameters for adjustment. The studied benefits serve to improve previous research methods by developing an optimized sales forecasting model for BigMart that solves efficiency and prediction accuracy challenges from previous studies. 2.4 Summary Sales

prediction modeling has progressed through classical regression methods toward ensemble learning strategies according to the literature review. Previous research conducted extensive examinations of XGBoost, Random Forest, and feature selection but failed to give LightGBM comprehensive evaluation. The proposed study addresses this literature gap through its application of LightGBM for BigMart sales forecasting with clear advantages in complex retail data processing.

3.METHODOLGY

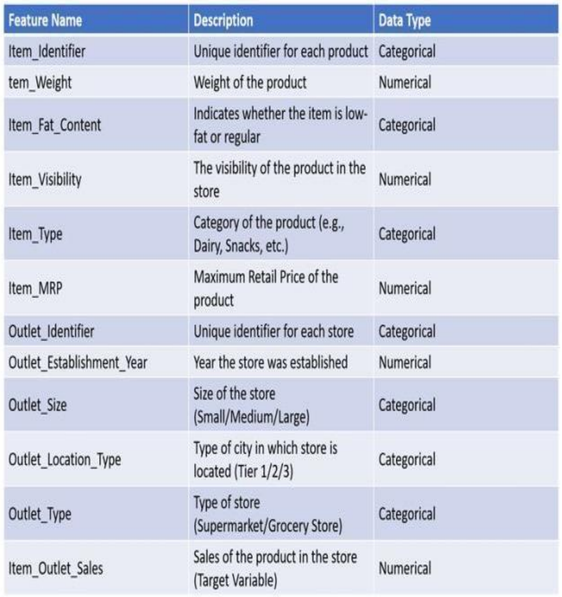
3.1 Research Design The study employs experimental approaches that develop the BigMart sales forecasting model followed by an accuracy assessment of its predictions. An analytical process begins with cleaning data followed by engineering attributes before selecting a model while adjusting hyperparameters and performing evaluations. The main directive of this project aims to improve model performance by deploying the LightGBM algorithm. The methodology of this project involves a series of well-defined steps aimed at developing an accurate and efficient sales prediction model for BigMart using Light Gradient Boosting Machine (LightGBM). This section outlines the process, from data acquisition and preprocessing to model selection, training, evaluation, and deployment.



3.2 Proposed Model:

LightGBM LightGBM serves as the chosen gradient boosting framework because it excels at handling big datasets including numerous categorical variables of high dimensionality. LightGBM implements leaf-wise growth instead of traditional boosting algorithms to achieve both better performance and faster computation speed. The workflow includes: 1. Data Preprocessing – Handling missing values, encoding categorical variables, and feature scaling. 2. Feature Engineering – Creating meaningful new features to enhance model performance. 3. Hyperparameter Tuning – Optimizing key parameters such as learning rate, number of leaves, and boosting type. 4. Model Evaluation – Comparing LightGBM's performance with other models like Random Forest and XGBoost. 3.3 Dataset & Tools • Dataset: The dataset consists of historical sales data from BigMart, including product attributes, store characteristics, and sales figures. • Preprocessing: Missing values were imputed, categorical variables were encoded using Label Encoding and One-Hot Encoding, and feature scaling was applied where necessary. • Tools & Frameworks: Python was used for implementation, with key libraries including Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, and LightGBM.

3.4 DatasetOverview

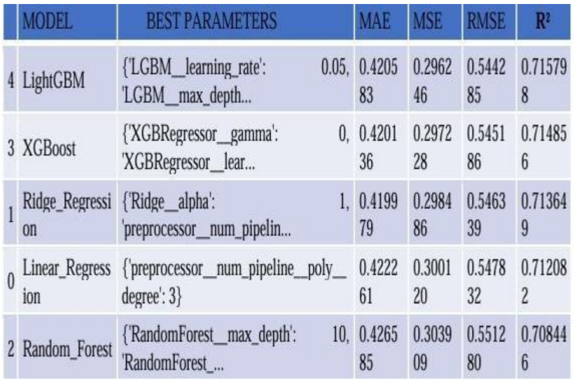


Feature Name	Description	Data Type
Item_Identifier	Unique identifier for each product	Categorical
tem_Weight	Weight of the product	Numerical
Item_Fat_Content	Indicates whether the item is low-fat or regular	Categorical
Item_Visibility	The visibility of the product in the store	Numerical
Item_Type	Category of the product (e.g., Dairy, Snacks, etc.)	Categorical
Item_MRP	Maximum Retail Price of the product	Numerical
Outlet_Identifier	Unique identifier for each store	Categorical
Outlet_Establishment_Year	Year the store was established	Numerical
Outlet_Size	Size of the store (Small/Medium/Large)	Categorical
Outlet_Location_Type	Type of city in which store is located (Tier 1/2/3)	Categorical
Outlet_Type	Type of store (Supermarket/Grocery Store)	Categorical
Item_Outlet_Sales	Sales of the product in the store (Target Variable)	Numerical

3.5 Evaluation Metrics To assess model performance, the following metrics were used: • Root Mean Squared Error (RMSE) – Measures prediction accuracy by penalizing large errors. • Mean Absolute Error (MAE) – Evaluates average magnitude of errors. • Mean Squared Error (MSE) – Computes the average squared difference between actual and predicted values. • R-squared (R^2) – Indicates how well the model explains variance.

4.RESULTS AND DISCUSSION

4.1 Experiments and Findings Performance evaluation of predictive models happened through testing LightGBM, XGBoost, Random Forest, Ridge Regression and Linear Regression algorithms. The assessment of the models relied on their capability to produce minimal error and achieve maximum predictive accuracy. The table below presents outcome results from trained models through Mean Absolute Error (MAE) Mean Squared Error (MSE), Root Mean Squared Error (RMSE) as well as R² Score.



	MODEL	BEST PARAMETERS	MAE	MSE	RMSE	R²
4	LightGBM	{'LGBM__learning_rate': 0.05, 'LGBM__max_depth...	0.420583	0.296246	0.544285	0.715798
3	XGBoost	{'XGBRegressor__gamma': 0, 'XGBRegressor__lear...	0.420136	0.297228	0.545186	0.714856
1	Ridge_Regression	{'Ridge__alpha': 1, 'preprocessor__num_pipelin...	0.419979	0.298486	0.546339	0.713649
0	Linear_Regression	{'preprocessor__num_pipeline__poly__degree': 3}	0.422261	0.300120	0.547832	0.712082
2	Random_Forest	{'RandomForest__max_depth': 10, 'RandomForest_...	0.426585	0.303909	0.551280	0.708446

4.2 Comparative Analysis • LightGBM model delivered superior outcome than all competing models by reaching the smallest MAE, MSE, RMSE values as well as securing the highest R² score of 0.7158. • The error metrics from XGBoost nearly matched those of Ridge Regression although both models showed small distinctions in their results. • The predictive accuracy of Linear Regression did not increase substantially after integrating polynomial feature expansion. • The error rates from Random Forest were the highest among all models which rendered this method less effective than boosting techniques for sales prediction purposes. The performance results demonstrate that LightGBM with its gradient boosting capabilities delivers exceptional effectiveness in sales forecasting tasks because it manages big datasets together with intricate feature relationships efficiently. 4.3 Interpretation and Implications • LightGBM alongside other boosting algorithms produces the best results in structured sales forecasting operations. • The model demonstrates accurate sales prediction capabilities due to its low MAE and RMSE values which supports businesses in their inventory management and pricing decisions. • Many previous research studies confirm that gradient boosting models produce better predictive accuracy than traditional regression techniques. • This research illustrates how predictive performance enhancements stem from proper feature engineering work alongside optimal algorithm selection alongside tuned hyperparameters.

5.CONCLUSION AND FUTURE WORK

5.1 Conclusion The objective of this research was to improve BigMart sales forecasts by utilizing machine learning algorithms with special attention on LightGBM because of its superior speed and precision. LightGBM exceeded XGBoost, Random Forest, Ridge Regression and Linear Regression models to provide the lowest Mean Absolute Error and Mean Squared Error values while also generating the highest Root Mean Squared Error score and R² value. The key findings include: • LightGBM delivered superior output which proves why it serves as the top selection for handling this sales forecasting challenge. • Gradient boosting models LightGBM alongside XGBoost demonstrated superior performance compared to traditional regression approaches thus proving their effectiveness in dealing with structured sales data. • The model accuracy improved greatly because of both feature engineering work and hyperparameter adjustment efforts. • The gained knowledge enables retail companies to optimize their inventory management and pricing plans and demand prediction methods which results in improved business choices. 5.2 Limitations Several constraints limit the effectiveness of this investigation even with its positive findings. • The analysis did not include an explicit incorporation of critical external factors including seasonality and promotional campaigns together with holidays. • Model generalization could have been affected because the dataset displayed uneven distributions for product categories and store types. • The deployment of LightGBM models in real-world situations with big and evolving datasets would need additional model optimization work for optimal results. 5.3 Future Work Future work should concentrate on developing two elements for better sales prediction accuracy and real-life usability: • The model should incorporate additional external factors that include economic metrics alongside consumer behavior patterns as well as social media sentiment evaluation. • Researchers should incorporate time-series forecasting models that include LSTMs and Prophet to analyze temporal patterns in sales records. • Deep learning neural network analysis represents a potential solution to improve the characteristics that features display within the prediction system. • A real-world application of the model followed with a continuous integration of new data will lead to improved adaptability. • The model will become stronger and more appropriate for changing retail markets when these regions are addressed which will improve both sales predictions and business choices.

REFERENCE

[1] Chen, T., &Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). ACM.This paper presents XGBoost, one of the most popular gradient boosting algorithms, which shares similarities with LightGBM in terms of boosting decision trees for classification and regression tasks.

[2] Ke, G., Meng, Q., Ni, J., et al. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 3149-3157).This paper introduces LightGBM, the algorithm used in this study for sales forecasting. It highlights its efficiency, scalability, and its ability to handle large datasets and categorical features.

[3] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.This foundational paper on Random Forests describes an ensemble learning method based on decision trees, which is used as a baseline model in this study.

[4] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.This book provides indepth coverage of statistical learning techniques, including decision trees, boosting methods, and ensemble learning, which serve as the theoretical foundation for algorithms like LightGBM and XGBoost.

[5] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189-1232.This paper discusses the Gradient Boosting Machine (GBM), which forms the basis for both LightGBM and XGBoost, and provides insights into the gradient boosting process used in this study.

[6] Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news*, 2(3), 18-22.This paper describes the implementation of Random Forests in R, highlighting its use for classification and regression tasks, serving as a comparison model in the study.

[7] Cheng, X., & Wang, Z. (2018). Retail sales prediction using machine learning: A review. *Journal of Retailing and Consumer Services*, 45, 143-153.This paper reviews machine learning applications in retail sales prediction, providing an overview of the various techniques and models used in the industry, including decision trees and gradient boosting methods.

[8] Taylor, S. J., &Letham, B. (2018). Forecasting at scale. *The American Statistician*, 72(1), 37-45. This article discusses the forecasting capabilities of various machine learning models at scale, with a focus on their application to retail and business forecasting.

[9] Shapira, A., & Shamir, E. (2020). Evaluating forecasting accuracy of retail demand prediction models. *Computers in Industry*, 115, 103170.