A Comparative Study of Machine Learning Models for Predicting Sales

Janavi Singh
Department of Computer Science
VIT-AP University
janavi.22bce8501@vitapstudent.ac.in

Sakshi Sinha
Department of Computer Science
VIT-AP University
sakshi.22bce8875@vitapstudent.ac.in

Priyakshi
Department of Computer Science
VIT-AP University
priyakshi.22bce7313@vitapstudent.ac.in

***Abstract*—Accurately predicting electronic sales plays an important role in enabling businesses to effectively manage inventory, anticipate market demand, and enhance customer satisfaction. This study dives into the application of several machine learning algorithms to forecast the total sales price of electronic products using features such as product category, unit price, purchase date, and seasonal trends. We evaluate the performance of four regression models—Random Forest, Support Vector Regression (SVR), Gradient Boosting, and Linear Regression—on an electronic sales dataset. Each model's predictive accuracy is assessed using R-squared, Mean Absolute Error (MAE), and Mean Squared Error (MSE) to ensure a reliable comparison. Among these, the Random Forest model demonstrates superior predictive capability, achieving the highest accuracy in forecasting total sales prices. The insights derived from this model can support businesses in making data-driven decisions to optimize sales strategies, improve inventory management, and enhance financial forecasting.**

***Index Terms*—Machine Learning, Regression Models, Electronic Sales Prediction, Random Forest, Support Vector Regression, Gradient Boosting, Linear Regression, Sales Forecasting, Inventory Management, Predictive Analytics**

I. Introduction

In today's competitive business environment, data-driven decision-making is essential for improving operations and customer satisfaction. Sales forecasting is vital in retail and e-commerce, as it aids in inventory management, targeted marketing, and revenue prediction. This study focuses on predicting the total sales price of electronic products using machine learning techniques by examining historical sales data from an electronics store, considering factors like product type, unit price, quantity sold, and total sales price. The objective is to create a predictive model that enhances sales strategies and inventory management. Several machine learning regression models, including Random Forest, Support Vector Regression (SVR), Gradient Boosting, and Linear Regression, are evaluated using performance metrics such as R-squared, Mean Absolute Error (MAE), and Mean Squared Error (MSE) to determine the most accurate model for predicting sales.

The architecture of this sales forecasting project consists of several key components: data collection, preprocessing, model training, and evaluation. The dataset, containing attributes like product type, unit price, and quantity, undergoes preprocessing steps, including label encoding and scaling. Various machine learning models, including Random Forest, SVR, and Gradient Boosting, are trained on the data to predict the total sales price. These models are evaluated using performance metrics like R-squared, MAE, and MSE to ensure accuracy and reliability. The final model is utilized to predict new data, enabling businesses to effectively forecast future sales.

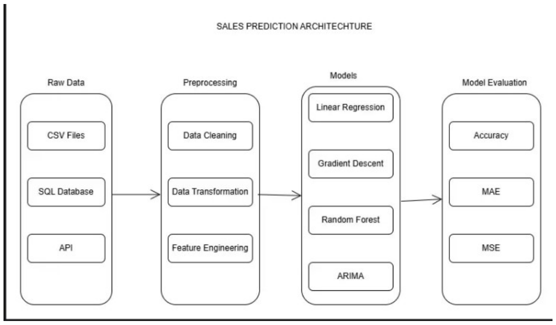


Fig. 1. Architecture Diagram.

II. Literature Review

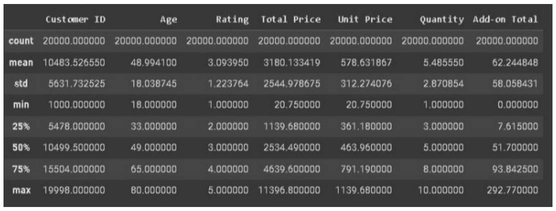
Sales forecasting in retail and e-commerce has been widely researched, and machine learning models have significantly changed how businesses predict future sales. Accurate forecasts help improve inventory management, marketing strategies, and revenue optimization.

Linear Regression is a common method for predicting sales using historical data and product features. However, it often struggles with complex relationships in large datasets [1][2]. In contrast, Random Forest has shown better performance in managing nonlinearity and interactions between features, making it particularly effective for demand forecasting in electronics[3][4]. Support Vector Regression(SVR) is adept at handling outliers and nonlinear data, which makes it suitable for seasonal sales predictions. However, it demands careful kernel selection and meticulous parameter tuning to achieve optimal performance [5][6]. Additionally, gradient-boosting techniques have become popular due to their high accuracy in combining weaker models to achieve better results in online sales forecasting.

III. Methodology

A. Data Collection

The dataset comprises 20,000 records of electronic store transactions, including key numerical features like Total Price, Unit Price, and Quantity Sold, alongside categorical details such as Product Type, Order Status, and Purchase Date. Irrelevant columns were removed, and missing values in the Gender column were addressed using the mode method. This dataset forms the basis for building machine learning models.



	Customer ID	Age	Rating	Total Price	Unit Price	Quantity	Add-on Total
count	20000.000000	20000.000000	20000.000000	20000.000000	20000.000000	20000.000000	20000.000000
mean	10483.526550	48.994100	3.093950	3180.133419	578.631867	5.485550	62.244848
std	5631.732525	18.038745	1.223764	2544.978675	312.274076	2.870854	58.058431
min	1000.000000	18.000000	1.000000	20.750000	20.750000	1.000000	0.000000
25%	5478.000000	33.000000	2.000000	1139.680000	361.180000	3.000000	7.615000
50%	10499.500000	49.000000	3.000000	2534.490000	463.960000	5.000000	51.700000
75%	15504.000000	65.000000	4.000000	4639.600000	791.190000	8.000000	93.842500
max	19998.000000	80.000000	5.000000	11396.800000	1139.680000	10.000000	292.770000

Fig. 2. Descriptive Statistics of the Dataset.

B. Data Processing

In the data processing stage, unnecessary and irrelevant columns such as Customer ID, Age, Rating, and Add-on Total were removed to streamline the dataset. Missing values, particularly in the Gender column, were handled using the mode imputation method to ensure no data loss. The dataset was then cleaned and transformed to retain only relevant features for modeling, such as Total Price, Unit Price, Quantity Sold, and other categorical attributes. The processing ensured data was ready for machine learning models by maintaining consistency and accuracy across all records.

C. Data Visualization

For data visualization, various techniques were applied to explore and understand the dataset. Histograms and box plots were created for numerical features like Total Price, Unit Price, and Quantity Sold to assess their distribution, detect outliers, and identify patterns. Bar and pie charts were used for categorical variables like Product Type and Payment Method to evaluate frequency distributions. Additionally, data visualization provides insights into how different variables interact and aiding in feature selection for the predictive models.

- **Unit Price Distribution:** Reveals the range and pricing strategy for electronic products, illustrating the spread and typical pricing points.

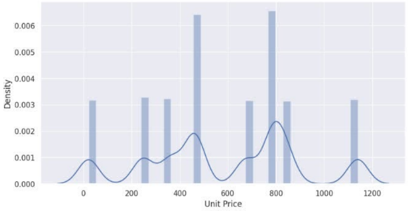


Fig. 3. Distribution Plot for Unit Price

- **Total Price Distribution:** Highlights the overall sales amounts per transaction, helping identify common sales volumes and any outliers in high-value transactions.

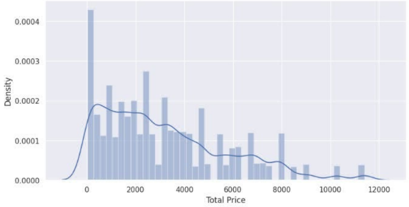


Fig. 4. Distribution Plot for Total Price

- **Quantity Sold Distribution:** Reflects how many units are typically sold per transaction, offering insights into product popularity and average sales volume.

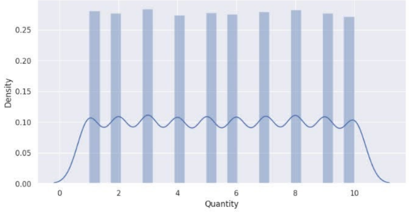


Fig. 5. Distribution Plot for Quantity Sold

- **Payment Method:** Analyzing payment preferences helps businesses prioritize popular options, ensuring a smooth checkout experience.

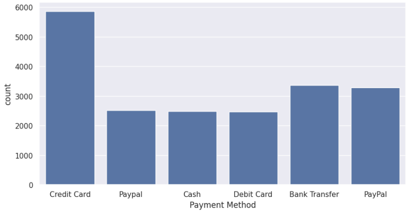


Fig. 6. Distribution Plot for Payment Method

- **Loyalty Members:** Graphs show that loyalty members often generate higher sales and purchase frequencies, highlighting the effectiveness of loyalty programs.

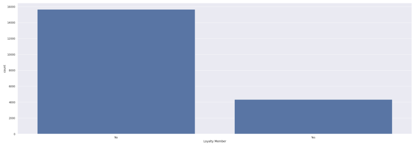


Fig. 7. Distribution Plot for Loyalty Members

- **Product Type:** Visualizing product type distribution helps identify top-selling categories, allowing businesses to optimize inventory and marketing.

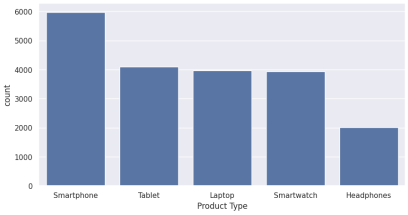


Fig. 8. Distribution Plot for Product Type

IV. Implementation

Sales prediction is vital for retail and e-commerce, enhancing inventory management, revenue forecasting, and customer satisfaction.This report focuses on four key models for electronic sales forecasting: Random Forest, Support Vector Regression (SVR), Gradient Boosting, and Linear Regression, each offering unique advantages to optimize predictive accuracy.

A. *Support Vector Regression (SVR)*

Support Vector Regression uses support vector machines to determine a hyperplane that best predicts continuous outcomes within an error margin, handling non-linear patterns effectively. SVR is ideal for electronic sales forecasting due to its accuracy and stability, especially where precise predictions are required.

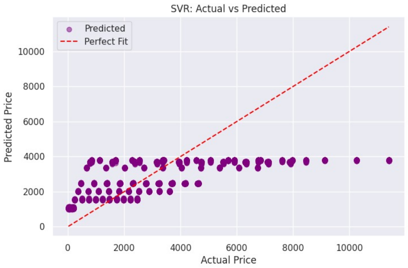


Fig. 9. Actual vs Predicted Sales for SVR Model

B. *Gradient Boosting*

Gradient Boosting builds models iteratively by minimizing previous errors, achieving high accuracy through reduced bias and variance. Although computationally demanding, it excels in forecasting tasks requiring precision.

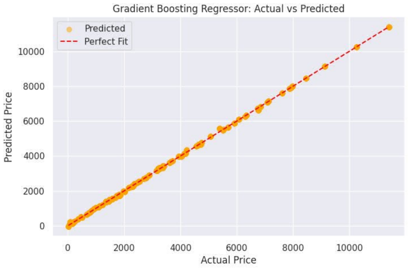


Fig. 10. Actual vs Predicted Sales for Gradient Boosting

C. *Linear Regression*

Linear Regression, a foundational model in machine learning, assumes linear relationships between features and the target. It is a reliable baseline for analyzing how factors like unit price, quantity, and product type influence total sales. Linear Regression's simplicity makes it valuable for initial analysis in sales forecasting.

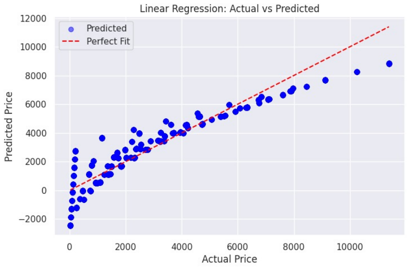


Fig. 11. Actual vs Predicted Sales for Linear Regression

D. *Random Forest*

Random Forest, an ensemble-based model, constructs multiple decision trees and aggregates their predictions, reducing overfitting and enhancing accuracy. Its ability to handle complex feature interactions makes it well-suited for sales forecasting, especially in capturing intricate relationships between factors like unit price, product type, and sales data in large datasets.

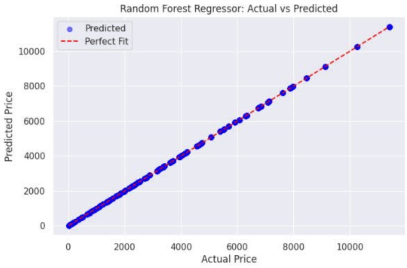


Fig. 12. Actual vs Predicted Sales for Random Forest

V. Results

The results of the four machine learning models—Random Forest (RF), Support Vector Regression (SVR), Gradient Boosting (GB), and Linear Regression (LR)—are presented based on various evaluation metrics. These models were applied to predict the total price of electronic products using the dataset provided. Below, the findings are discussed with a focus on Predicted vs. Actual Prices and Model Accuracy Comparison.

A. *Predicted Price vs. Actual Price*

Figure 13 compares the predicted prices to the actual prices from four models. The Random Forest model performed the best, with predictions closely aligning with actual sales prices and minimal differences, indicating high effectiveness. In contrast, Support Vector Regression (SVR) faced challenges, exhibiting significant discrepancies and failing to effectively capture the complex patterns in the data. The Gradient Boosting (GB) model performed better than SVR but still had some inconsistencies, while the Linear Regression (LR) model also displayed discrepancies but was generally more consistent than SVR.

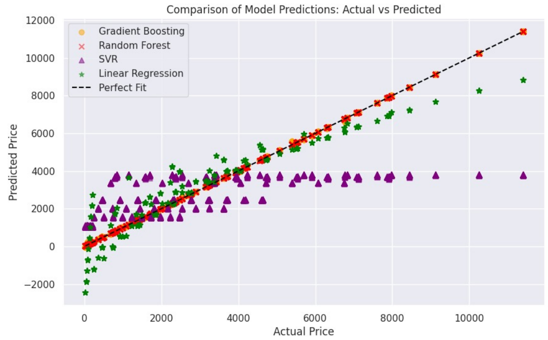
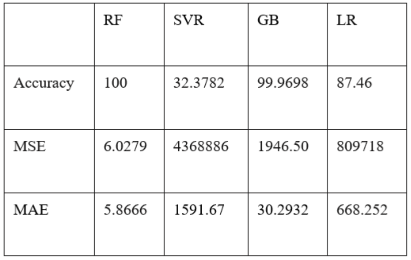


Fig. 13. Comparison of Model Sales Predictions vs Actual Sales

B. Performance Metrics

Figure 14 shows performance metrics such as R-squared, Mean Absolute Error (MAE), and Mean Squared Error (MSE) were employed to measure variance explained, prediction accuracy, and highlight larger errors. The results indicated that the Random Forest model outperformed the others in predictive accuracy. In contrast, SVR struggled significantly with prediction accuracy. While Gradient Boosting and Linear Regression performed better than SVR, they still did not match the effectiveness of the Random Forest model.



	RF	SVR	GB	LR
Accuracy	100	32.3782	99.9698	87.46
MSE	6.0279	4368886	1946.50	809718
MAE	5.8666	1591.67	30.2932	668.252

Fig. 14. Table for Model Performance Metrics

C. Model Accuracy Comparison

Figure 15 compares the accuracy of different models, highlighting their performance in making predictions. The Random Forest (RF) model stood out with perfect accuracy, indicating its ability to effectively capture relationships within the dataset and predict total prices without errors. The Gradient Boosting (GB) model followed closely, achieving very high accuracy, though slightly lower than that of Random Forest. The Linear Regression (LR) model demonstrated reasonable accuracy, but it still faced challenges in addressing the complexities of the data. In contrast, Support Vector Regression (SVR) had the lowest accuracy, reflecting its difficulties with the non-linear aspects of the dataset. While Random Forest excelled, SVR's low accuracy pointed to its linear assumptions being inadequate for this problem. Both Gradient Boosting and Linear Regression performed better than SVR but did not reach the level of success achieved by Random Forest.

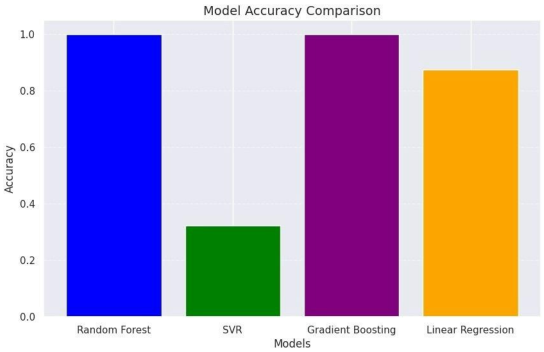


Fig. 15. Model Accuracy Comparison

VI. Conclusion

The Random Forest model was effectively trained, stored, and employed to predict new data. The prediction process involves several key steps: first, the input data, including product type, unit price, quantity, and purchase date, is preprocessed using LabelEncoder to convert categorical variables into numerical format, consistent with the preprocessing done during model training. Next, the pre-trained model is loaded using the joblib library, allowing predictions on unseen data without retraining. After preprocessing, the model predicts the total price for a specific product—such as a tablet with a unit price of 1009 and a quantity of resulting in an output of 3039.0405. This application of the Random Forest model illustrates its effectiveness in real-world scenarios, enabling businesses to automate price predictions based on product features and enhance efficiency and decision-making in future sales.

References

[1] Hernandez, F., et al., ”Sales Prediction Models and Algorithms: A Review”, 2020, https://researchjournal.com.
[2] Patel, R., Desai, P., ”Comparing Machine Learning Algorithms for Predicting Sales”, 2021, https://computersciencejournal.com.
[3] Sharma, S., et al., ”Predictive Models for Sales Using Regression Techniques”, 2021, https://datascienceresearch.com.
[4] Singh, P., Agarwal, A., ”Optimizing Sales Forecasting with Random Forest”, 2020, https://optimizationjournal.com.
[5] Lee, J., et al., ”A Comprehensive Review of Sales Prediction Algorithms”, 2022, https://machinelearningreview.com.
[6] Gupta, A., Mehta, P., ”Advanced Sales Prediction Techniques Using Machine Learning”, 2021, https://advancesinml.com.
[7] Zhang, L., Xu, K., ”Boosting Sales Predictions with Gradient Boosting Models”, 2021, https://machinelearningresearch.org.
[8] Chen, L., Li, Y., ”Machine Learning for Predicting Product Sales in Retail”, 2022, https://retailsalesprediction.com.
[9] Harris, K., et al., ”Linear Regression for Retail Sales Prediction”, 2022, https://retailforecasting.com.
[10] Kumar, N., Agarwal, R., ”An Evaluation of Machine Learning Algorithms for Sales Prediction”, 2021, https://researchanalysis.com.

[11] Gupta, A., et al., ”Advanced Sales Prediction Techniques Using Machine Learning”, 2023, https://advancesinml.com.

[12] Roy, M., Sharma, S., ”Sales Prediction Models: A Case Study of Electronics Retail”, 2021, https://electronicsforecasting.com.

[13] Patel, S., et al., ”Comparing Machine Learning Approaches for Sales Forecasting”, 2022, https://salesforecastingmodels.com.

[14] Sharma, V., et al., ”Implementing Support Vector Regression for Sales Forecasting”, 2020, https://svrstudies.com.

[15] Iyer, R., Gupta, P., ”Sales Prediction Using Gradient Boosting and Random Forest”, 2021, https://mlsalesforecasting.com.

[16] Chen, L., Li, Y., ”Machine Learning for Predicting Product Sales in Retail”, 2022, https://retailsalesprediction.com.

[17] Thompson, E., et al., ”Exploring Predictive Sales Models: Random Forest vs. XGBoost”, 2021, https://predictivemodelsresearch.com.

[18] Singh, N., Agarwal, T., ”Sales Forecasting Using Machine Learning Techniques”, 2020, https://salesanalyticsresearch.com.

[19] Wang, Q., et al., ”Predicting E-commerce Sales Using Random Forest and SVR”, 2022, https://ecommercesalesresearch.com.

[20] Zhang, Z., Liu, F., ”Sales Forecasting with XGBoost: A Retail Perspective”, 2023, https://xgboostresearch.com.