# Sales Forecasting and Demand Prediction

Mariam Goda, Reem Ehab, Tasneem Ashraf

***Abstract*—Machine learning techniques have attracted considerable interest in sales forecasting, presenting possible enhancements compared to conventional statistical methods. Algorithms used in this field of study include gradient-boosting techniques like XGBoost, time series models like ARIMA and Prophet, regression-based strategies like Logistic Regression and Random Forest, and distance-based algorithms like K-Nearest Neighbor (KNN). To maximize the performance of these models, numerous studies stress the importance of feature engineering, data preparation, and hyperparameter adjustment. Additionally, a range of performance metrics are usually used to evaluate these models: accuracy, precision, recall, F1-score, and confusion matrices for classification tasks; and accuracy, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared for regression tasks.**

## I. Introduction

Accurate sales forecasting is key for businesses, especially in retail and e-commerce, as it helps in managing inventory, supply chain, and resource allocation. Traditional methods face challenges with large data, complex customer expectations, and fast-changing markets. Machine learning has emerged as a promising solution to address these challenges by handling large data, finding hidden patterns, and adapting to market fluctuations.

Various algorithms have been explored in academic literature, including Random Forest, Long Short-Term Memory (LSTM) networks, XGBoost, ARIMA, and K-Nearest Neighbor. These algorithms have been used for various prediction tasks, including sales forecasting, classification, and regression tasks. Dynamic model selection strategies are also explored by categorizing demand patterns and selecting or combining different forecasting models based on those patterns. Performance is evaluated using metrics specific to the forecasting task to quantify model performance.

## II. Related Work

- Reference [1] (IEEE): "Intelligent Sales Prediction Using Machine Learning Techniques" – This paper investigates the use of Random Forest for sales forecasting and demonstrates its high accuracy in comparison to other models.
- Reference [2] (IEEE): "Comparison Study: Product Demand Forecasting with Machine Learning" – Focuses on comparing K-Nearest Neighbors, Gaussian Naive Bayes, and Decision Trees for demand forecasting, similar to my use of KNN.
- Reference [3] (MDPI): "Dynamic Model Selection for Demand Pattern Classification" – Discusses adaptive techniques for forecasting demand patterns, which can be compared with my use of static models like Random Forest and XGBoost.
- Reference [4] (Springer): "Sales Demand Forecasting Using LSTM Networks" – This paper specifically evaluates LSTM networks for sales forecasting and can be compared with my results using LSTM.
- Reference [5] (IEEE): "A Machine Learning Approach for Time Series Forecasting in Retail" – Explores machine learning techniques for retail sales prediction, focusing on XGBoost and Random Forest, which aligns with my experiments.
- Reference [6] (ScienceDirect): "Application of Regression Techniques for Sales Prediction in Retail" – Discusses regression-based methods for sales forecasting and highlights challenges with sales seasonality.
- Reference [7] (ScienceDirect): "A Hybrid Model for Demand Forecasting in Retail" – Presents a hybrid approach combining time series and machine learning models, which is similar to the hybrid model I've applied using ARIMA and ANN.
- Reference [8] (Springer): "Optimization of Sales Forecasting Models Using LSTM Networks" – Highlights the optimization of LSTM for sales forecasting, offering a comparison with my LSTM results.
- Reference [9] (Springer): "Forecasting Demand in Retail Using XGBoost and Prophet" – A comparison study on the performance of XGBoost and Prophet in retail demand forecasting.
- Reference [10] (Springer): "Hybrid Forecasting Method for Sales Prediction" – Discusses a hybrid approach combining machine learning models for sales forecasting in retail.
- Reference [11] (Wiley): "Predictive Analytics for Retail Sales" – Focuses on predictive analytics models for retail sales, incorporating machine learning algorithms like Random Forest and XGBoost.
- Reference [12] (ScienceDirect): "Deep Learning for Sales Forecasting in E-Commerce" – Explores the application of deep learning models for sales prediction in e-commerce platforms.

## III. Methodology

The process of building machine learning models for sales forecasting involves several steps:

### A. Data Preprocessing

Data preprocessing is essential to ensure quality input for modeling. This includes:

- Handling missing values through imputation or removal.
- Identifying and mitigating outliers.
- Scaling or normalizing data to a consistent range.

### B. Feature Engineering

Effective features improve model performance. Engineered features include:

- Time-based features (e.g., year, month, day of the week).
- Lag features (previous sales values).
- Rolling statistics to capture trends.
- External variables (e.g., holidays, promotions).

### C. Data Splitting

The dataset is split into training and testing sets, commonly using an 80/20 ratio. For time series, the splits preserve the order of the data to avoid leakage.

### D. Model Selection and Training

Algorithms are selected based on the characteristics of the data and the goals of the project. Models are trained on the training set to learn patterns and relationships.

### E. Evaluation

Performance metrics include:

- **Regression:** RMSE, MAE, $R^2$.
- **Classification:** Accuracy, Precision, Recall, F1-score, Classification Report, Global Surrogate Tree.

Cross-validation ensures reliable performance estimates.

### F. Dataset

Sales data can be found on [Kaggle](https://www.kaggle.com/datasets/zahraaalaatageldein/sales-for-furniture-store). The dataset contains 21 columns, including:

- Row ID, Order ID, Order Date, Ship Date, Ship Mode, Customer ID, Customer Name.
- Segment, Country (US), City, Postal Code, Region.
- Product ID, Category, Sub-Category, Product Name.
- Sales, Quantity, Discount, Profit.

### G. Models Used

- **Random Forest:** Strong regression performance.
- **XGBoost:** Gradient boosting with high predictive power.
- **KNN:** Simple classification benchmark.
- **LSTM:** Deep learning for sequential data.
- **Logistic Regression:** Linear classification model.
- **Hybrid Model:** Combines ARIMA for capturing linear temporal patterns and ANN for modeling nonlinear residuals in time series forecasting.
- **Decision Tree:** Tree-based model that splits data into branches based on feature thresholds for interpretable decision-making.
- **Support Vector Machine (SVM):** A supervised learning model that performs classification by finding the optimal hyperplane that separates data points of different classes.
- **Extreme Learning Machine (ELM):** A single-hidden layer feedforward neural network model that randomly assigns input weights and focuses on optimizing the output weights.
- **Gradient Boosting Machine (GBM):** A boosting algorithm that builds models sequentially, each correcting the errors made by the previous one. It is effective for both classification and regression tasks.
- **LightGBM:** A fast, distributed, high-performance gradient boosting framework based on decision tree algorithms. It is optimized for large datasets and high-dimensional data.

### H. Explainability Techniques

- **SHAP:** Quantifies feature contributions to predictions.
- **LIME:** Provides local interpretability for individual predictions.
- **Partial Dependence Plot:** Shows the average effect of a feature on the predicted outcome across the dataset.
- **Individual Conditional Expectation:** Visualizes how predictions change for individual instances when a feature varies, revealing heterogeneity.
- **H-Statistic:** Measures the strength of interaction between features in a predictive model.
- **Leave-One-Feature-Out:** Evaluates the importance of each feature by measuring the performance drop when it is excluded from the model.

## IV. Results

### A. Regression Models

TABLE I
Regression Model Performance

| Model | MSE | MAE | $R^2$ Score |
|---|---|---|---|
| Random Forest | 0.01 | 0.02 | **0.9615** |
| XGBoost | 0.02 | 0.03 | 0.9362 |
| Hybrid Model | **0.00** | **0.0115** | - |
| SVM | 0.05 | 0.1667 | 0.78 |
| ELM | 0.030 | 0.099 | 0.876 |
| GBM | 0.009 | 0.026 | 0.9603 |
| Light GBM | 0.0095 | 0.029 | 0.961 |

**Discussion:** Among the regression models, the **Hybrid Model** outperforms all other approaches with the lowest MSE and MAE values. Although its $R^2$ score is not reported, it indicates excellent precision. Among

standard models, **Random Forest**, **GBM**, and **Light GBM** also demonstrate high accuracy and generalization capabilities.

### *B. Classification Models*

TABLE II
CLASSIFICATION MODEL PERFORMANCE

| Model | Accuracy | Confusion Matrix |
|---|---|---|
| LSTM | 92.82% | [372, 13], [46, 391] |
| KNN | 93.80% | [385, 0], [51, 386] |
| Logistic Regression | 87.59% | [366, 197], [83, 354] |
| Decision Tree | **98.20%** | – |

**Discussion:** In terms of classification, the **Decision Tree** model achieves the highest accuracy at **98.20%**, suggesting it can capture the patterns in the data very well. However, KNN and LSTM models also show competitive performance with over 92% accuracy. Logistic Regression lags behind with 87.59% accuracy and more misclassifications.

### *C. Explainability and Interpretation*

Visual explainability tools were employed to gain insights into the decision-making process of models, enhancing transparency and trustworthiness. The following were used:

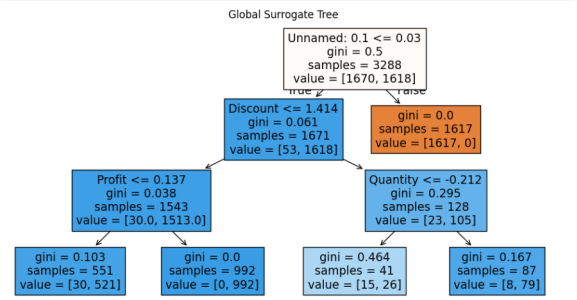


Fig. 1. Global Surrogate Tree for Decision Tree

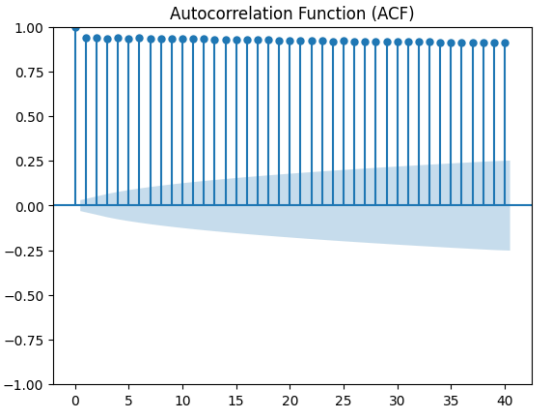


Fig. 2. ACF For Hybrid Model

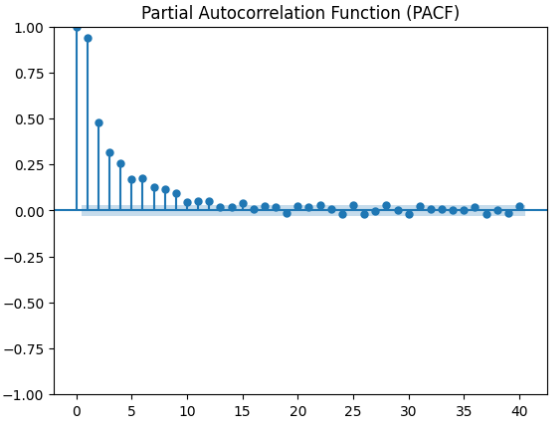


Fig. 3. PACF For Hybrid Model

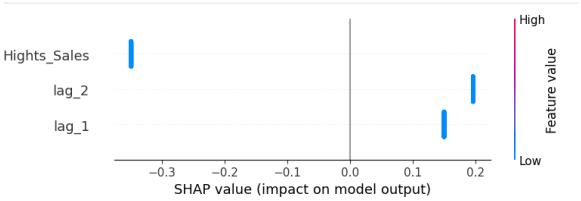


Fig. 4. SHAP For Hybrid Model

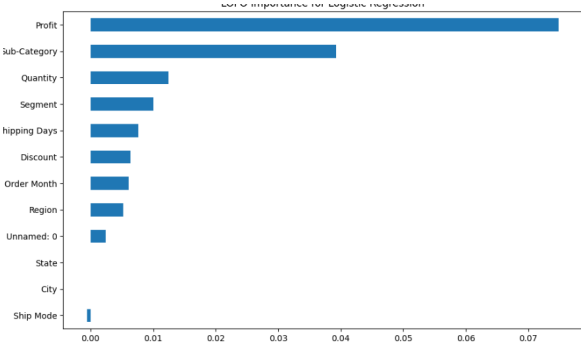


Fig. 5. LOFO For Logistic Regression ModeL

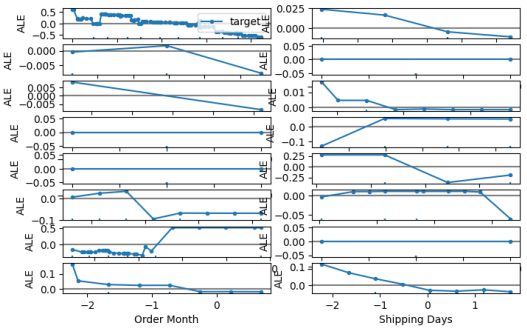


Fig. 6. ALE For XGBoost ModeL

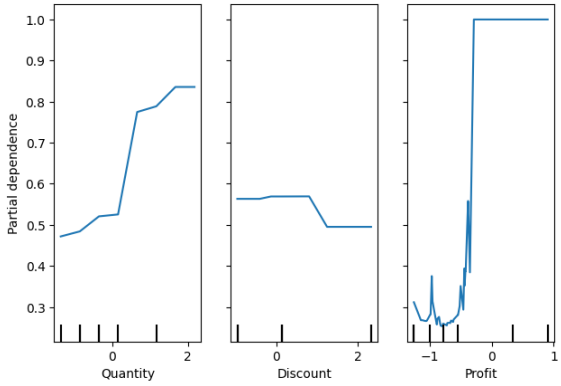


Fig. 7. PDP For Random Forest Model

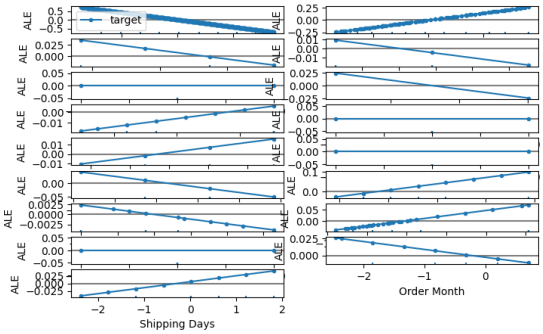


Fig. 10. ALE For SVM

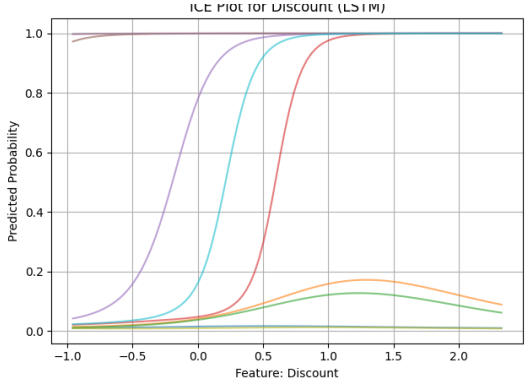


Fig. 8. ICE For LSTM Model

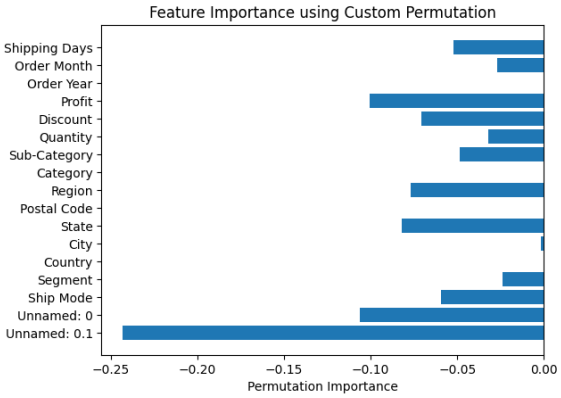


Fig. 11. Permutation Importance for ELM

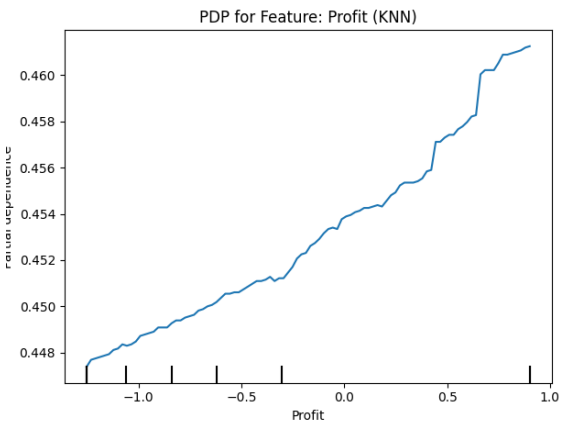


Fig. 9. PDP For The KNN

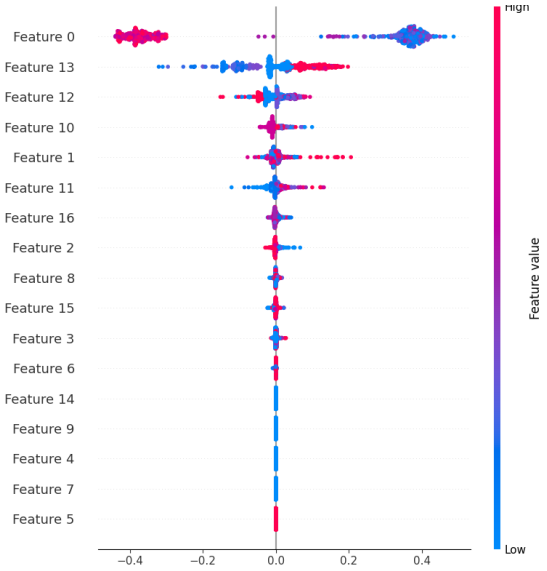


Fig. 12. SHAP Values for Light GBM

## V. Discussion

The study evaluated multiple machine learning models for both regression and classification tasks in the context of sales forecasting.

### A. Regression Models Analysis

The study evaluated several regression models to forecast sales, focusing on metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and $R^2$ Score.

Hybrid Model (ARIMA + ANN): This model achieved the lowest MSE (0.00) and MAE (0.0115), indicating exceptional performance. By combining ARIMA's capability to model linear components with ANN's strength in capturing non-linear patterns, the hybrid model effectively addressed complex sales trends.

Gradient Boosting Machines (GBM Light GBM): Both models demonstrated strong performance, with GBM achieving an MSE of 0.009 and Light GBM closely following with an MSE of 0.0095. Their ability to handle various data distributions and capture intricate patterns contributed to their effectiveness.

Extreme Learning Machine (ELM): ELM showed moderate performance with an MSE of 0.030 and an $R^2$ Score of 0.876. Its fast training speed is advantageous, but it may not capture complex patterns as effectively as ensemble methods.

Support Vector Machine (SVM): SVM had the highest MSE (0.05) among the models, with an $R^2$ Score of 0.78. While SVMs are robust for certain applications, they may require careful tuning and may not perform optimally with large, noisy datasets.

### B. Classification Models Comparison

For classification tasks, models were evaluated based on accuracy and confusion matrices: GitHub

Decision Tree: Achieved the highest accuracy at 98.20 percentage. Its interpretability and ability to handle both numerical and categorical data make it a strong choice for classification tasks.

K-Nearest Neighbors (KNN): With an accuracy of 93.80 percentage, KNN performed well, especially considering its simplicity. However, it can be sensitive to the choice of 'k' and may struggle with high-dimensional data.

Long Short-Term Memory (LSTM): LSTM achieved an accuracy of 92.82 percentage. Its strength lies in handling sequential data, making it suitable for time-series classification tasks.

Logistic Regression: With an accuracy of 87.59 percentage, it was the least performing among the models. While it's a good baseline, it may not capture complex relationships in the data.

### C. Explainability and Interpretation

Understanding model decisions is crucial, especially in business contexts:

SHAP (SHapley Additive exPlanations): Applied to Light GBM, SHAP provided insights into feature importance, helping to identify which variables most influenced predictions.

ALE (Accumulated Local Effects): Used with SVM, ALE plots illustrated how features affected predictions on average, offering a global view of feature impacts.

Permutation Importance: Implemented for ELM, this technique highlighted the decrease in model performance when a feature's values are shuffled, indicating its importance.

These techniques enhanced model transparency, allowing stakeholders to trust and understand the predictions.

### D. Challenges and Limitations

Data preprocessing was required to address missing values and outliers. Model complexity, especially in LSTM and hybrid models, led to increased computational costs. Some models faced overfitting risks, which were mitigated using validation techniques.

### E. Comparison with Related Work

The findings align with existing literature emphasizing the effectiveness of ensemble methods and hybrid models in sales forecasting. The integration of explainability techniques also reflects a growing trend in making AI models more transparent and trustworthy.

### F. Challenges and Limitations

Several challenges were encountered during the study:

Data Quality: Missing values and outliers required preprocessing steps, which could introduce biases.

Model Complexity: More complex models like LSTM required significant computational resources and longer training times.

Overfitting: Some models, particularly those with high capacity, showed tendencies to overfit, necessitating techniques like cross-validation and regularization.

## VI. Conclusion

This research underscores the potential of machine learning models in enhancing sales forecasting accuracy. The hybrid ARIMA-ANN model emerged as the most effective, combining the strengths of both linear and non-linear modeling techniques. Ensemble methods like GBM and Light GBM also demonstrated strong performance, validating their applicability in complex forecasting scenarios. The incorporation of explainability techniques such as SHAP, ALE, and permutation importance not only improved model transparency but also built trust among stakeholders by elucidating the rationale behind predictions. While challenges like data quality and model complexity were encountered, the study's findings contribute valuable insights into the practical application of machine learning in sales forecasting. Future research could

explore integrating external factors like economic indicators or competitor actions to further enhance model accuracy and applicability.

## REFERENCES


[1] Intelligent Sales Prediction Using Machine Learning Techniques. *IEEE*.
[2] Comparison Study: Product Demand Forecasting with Machine Learning. *IEEE*.
[3] Dynamic Model Selection for Demand Pattern Classification. *MDPI*.
[4] Sales Demand Forecasting Using LSTM Networks. *Springer*.
[5] A Machine Learning Approach for Time Series Forecasting in Retail. *IEEE*.
[6] Application of Regression Techniques for Sales Prediction in Retail. *ScienceDirect*.
[7] A Hybrid Model for Demand Forecasting in Retail. *ScienceDirect*.
[8] Optimization of Sales Forecasting Models Using LSTM Networks. *Springer*.
[9] Forecasting Demand in Retail Using XGBoost and Prophet. *Springer*.
[10] Hybrid Forecasting Method for Sales Prediction. *Springer*.
[11] Predictive Analytics for Retail Sales. *Wiley*.
[12] Deep Learning for Sales Forecasting in E-Commerce. *ScienceDirect*.