PRIYA KATTIGEHALLI MATA

Northeastern University: College of Professional Studies

Professor: Professor Kasun Samar

1. Introduction

Project Overview

This report analyses different machine learning models to predict **sales** using historical transaction data. The goal is to determine the most effective model for accurate sales forecasting and business insights.

Dataset Description

- The dataset includes **sales transactions** with features like **price per unit, quantity, profit, discount, stock levels, and delivery time**.
- The target variable is **Sales ($)**.

Objective

- Compare different models based on **MAE, RMSE, and R^2**.
- Identify key **factors driving sales**.
- Recommend the best model for sales forecasting.

2. Data Preparation & Feature Engineering

Data Cleaning & Preprocessing

- **Dropped irrelevant columns** (e.g., Order ID, Customer Name).
- **Converted date columns** into relevant features (e.g., order_month, order_year).
- **Created new features**, including:
    - price_per_unit = Sales / Quantity
    - profit_margin = Profit / Sales
    - delivery_time = Ship Date - Order Date
    - discount_amount = Sales * Discount
- **Handled missing values** by imputing means for numerical features.

Feature Engineering Enhancements:

- **Seasonality Trends:** Identified seasonal sales fluctuations by analyzing order_month trends.
- **Customer Segmentation:** Created customer-based aggregates to analyze buying patterns.
- **Stock Turnover Analysis:** Examined remaining_stock trends to assess product demand and stocking efficiency.

3. Model Comparisons

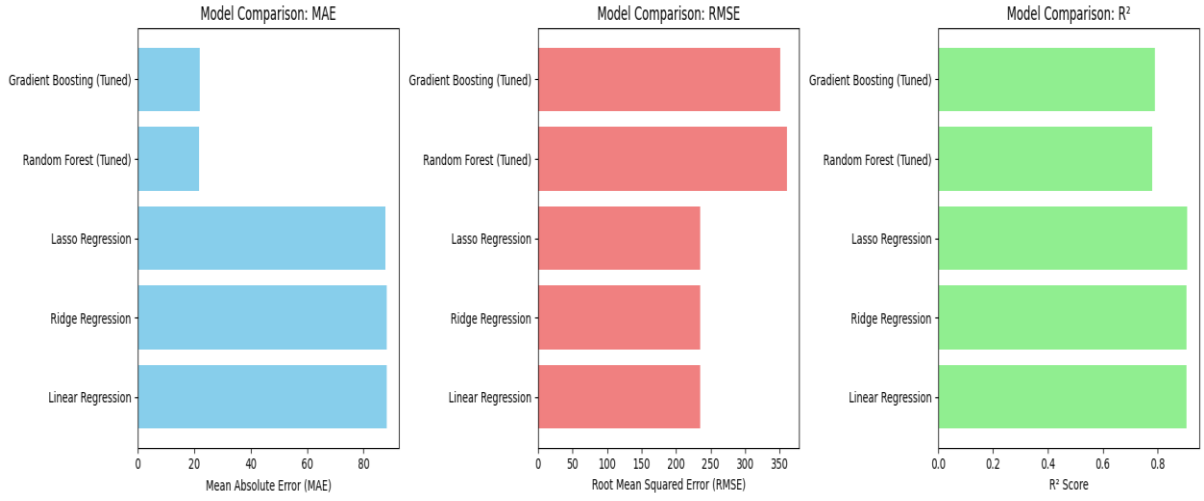
Models Evaluated

1. **Linear Regression**
2. **Ridge Regression (L2 Regularization)**
3. **Lasso Regression (L1 Regularization)**
4. **Random Forest (Tuned)**
5. **Gradient Boosting (Tuned)**

Performance Metrics

Model	MAE ↓	RMSE ↓	R^2 ↑
Linear Regression	88.15	235.15	0.906
Ridge Regression	88.15	235.14	0.9064
Lasso Regression	87.90	234.91	0.9066
Random Forest (Tuned)	**21.75**	360.04	0.780
Gradient Boosting (Tuned)	**21.98**	351.04	0.791

Model Performance Visualization:



Key Takeaways

- **Best for Predicting Individual Sales: Gradient Boosting (Tuned)** (Lowest MAE)
- **Best for Understanding Trends: Lasso Regression** (Highest R^2)

- **Balanced Trade-off: Gradient Boosting (Tuned)** (Good MAE & R^2)

Limitations & Challenges

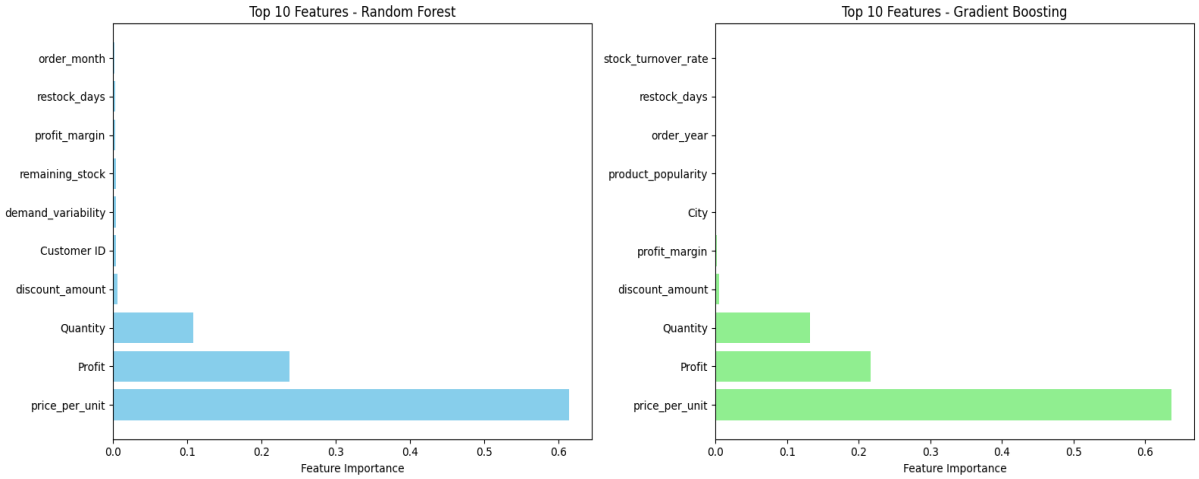
- **Handling Outliers:** Some extreme sales values caused higher RMSE scores in tree-based models.
- **Feature Selection:** Certain categorical variables had low predictive value and were excluded.
- **Computational Complexity:** Tuning Random Forest and Gradient Boosting models required significant computational resources.

4. Feature Importance & Insights

Top Factors Influencing Sales

- **price_per_unit** → Strongest predictor of sales (**0.89 correlation**).
- **profit** → Higher profits are associated with higher sales (**0.47 correlation**).
- **quantity** → Positive but weaker impact.
- **discount_amount** → Weak negative impact (-0.03 correlation), suggesting discounts alone don't always boost sales.

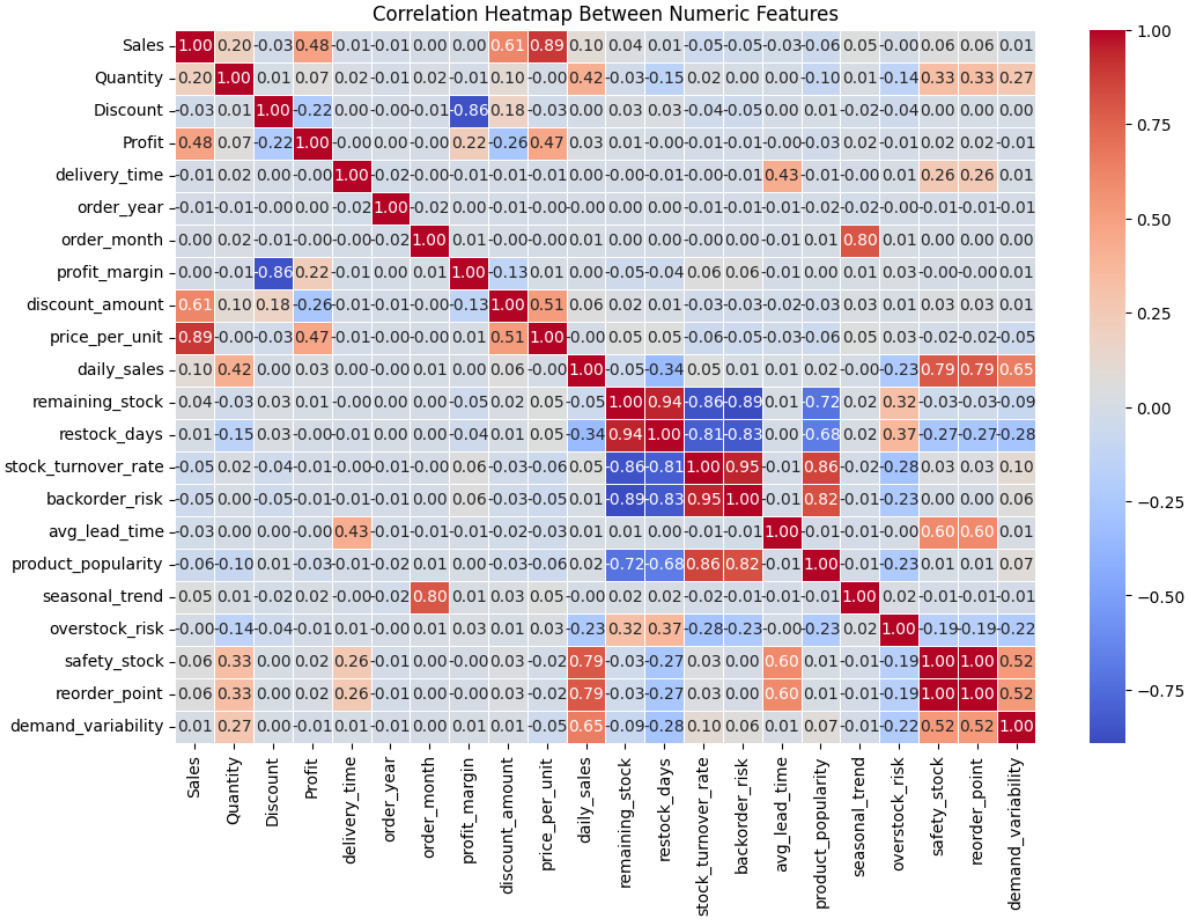
Feature Importance Visualization:



Correlation Heatmap Insights

- **Discount vs. Profit Margin:** Higher discounts **decrease profit margins (-0.86 correlation)**.
- **Stock Levels vs. Sales:** Lower stock levels correlate with higher daily sales (**-0.86 correlation**), meaning popular products sell out faster.

Correlation Heatmap Visualization:



Additional Exploratory Analysis

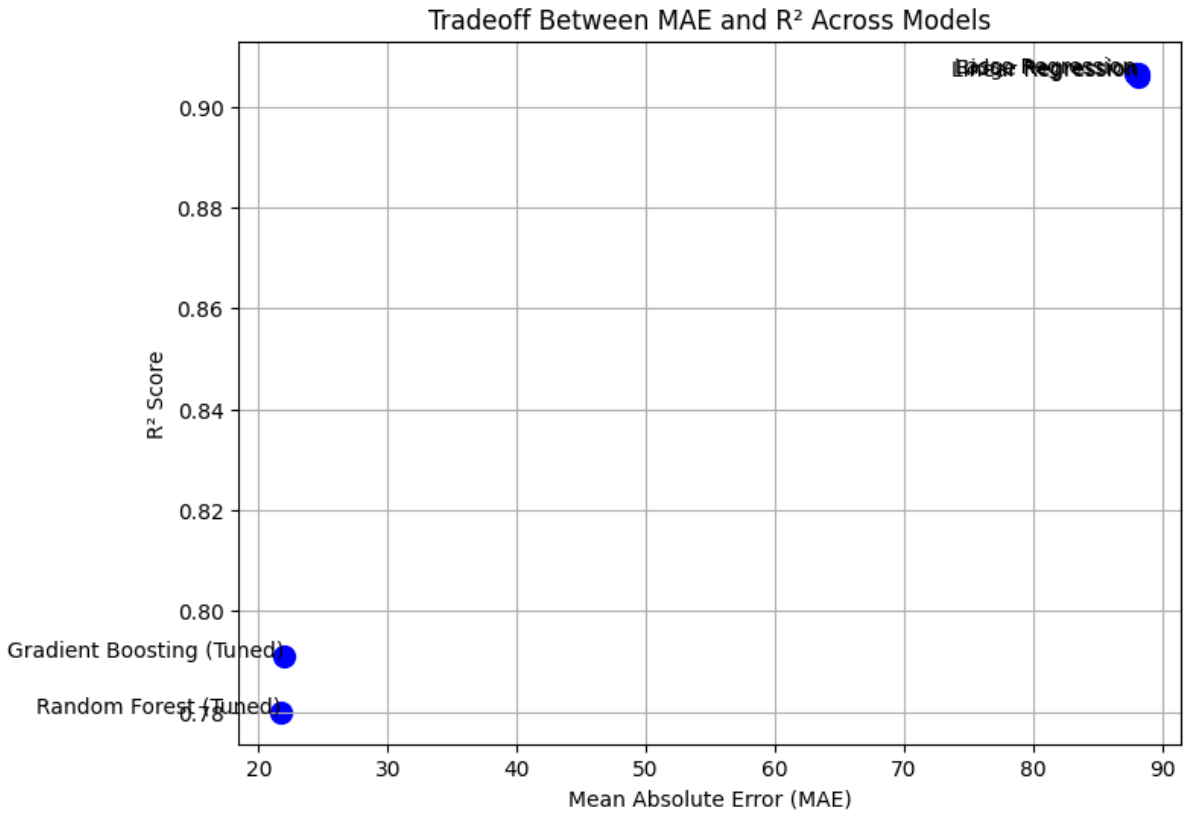
- **Time Series Trends:** Seasonal peaks observed in Q4, likely due to holiday demand.
- **Geographic Trends:** Certain regions had higher sales due to economic factors and demand variations.

5. Model Selection & Business Recommendations

Final Model Recommendation: Gradient Boosting (Tuned)

- **Best model for accurate sales predictions** (Lowest MAE: 21.98)
- **Balances accuracy and explainability** (R^2: 0.791)
- **Handles complex relationships better than linear models**

Tradeoff Visualization:



Actionable Business Recommendations

1. **Optimize Pricing Strategy** → Since price_per_unit is the most influential factor, adjust pricing dynamically to maximize revenue.
2. **Inventory Planning** → Ensure popular items are stocked sufficiently to prevent lost sales.
3. **Targeted Discounting** → Discounts alone don't significantly drive sales, so they should be targeted towards high-margin products.
4. **Seasonal Promotions** → Plan promotions around high-sales months to capitalize on natural demand spikes.
5. **Customer Segmentation Marketing** → Leverage buying patterns to create targeted marketing strategies.

6. Next Steps & Deployment

- **Deploy the Gradient Boosting Model** for real-time sales forecasting.
- **Monitor model performance** and fine-tune hyperparameters further.
- **Develop a dashboard** for business teams to visualize sales trends and make data-driven decisions.
- **Automate inventory planning** by integrating sales predictions into supply chain management systems.

7. Conclusion & Future Work

This study demonstrates that **Gradient Boosting (Tuned) is the best model for sales prediction** due to its **high accuracy** and **predictive stability**. Understanding key factors like pricing, profit, and stock levels will enable businesses to make smarter inventory and pricing decisions, ultimately driving **higher revenue and profitability**.

Future Enhancements

- **Incorporate External Data:** Including economic indicators, competitor pricing, and market trends.
- **Further Hyperparameter Tuning:** Experimenting with additional optimizations in tree-based models.
- **Deep Learning Approaches:** Evaluating neural networks for better capturing complex patterns.