

# EVALUATION TASK REPORT

This is the dataset link--→ [Walmart Sales](#)

The goal of this task was to build a sales prediction model using the Walmart dataset.

We implemented two models:

1. Linear Regression (from scratch using gradient descent).
2. Random Forest (using scikit-learn).

The task also required data cleaning, feature engineering, and performance evaluation.

## APPROACH

### Part A: Python Programming

#### 1. Statistical Functions

- Implemented functions to calculate mean, median, mode, variance, and standard deviation.
- First implemented using pure Python (loops, conditionals, basic math).
- Then re-implemented using NumPy, which gave concise and efficient solutions.

#### 2. Product Billing

- Implemented a function that accepts a dictionary of products ({product: (price, quantity)}).
- Computed:
  - Total bill amount.
  - Product with highest contribution (price × quantity).

This part demonstrated ability to write clean Python functions without relying on advanced libraries, and then optimize using NumPy.

## **Part B: Machine Learning Task**

### **DATA CLEANING & EDA**

- Loaded dataset with Pandas.
- Converted `Date` to datetime format and extracted `Year`, `Month`, `Day`, `Week`.
- Handled missing values and checked sales distribution.
- Plotted sales trends for exploratory analysis.

### **FEATURE ENGINEERING**

- Features used: `Store`, `Holiday\_Flag`, `Temperature`, `Fuel\_Price`, `CPI`, `Unemployment`, plus date features.
- Scaled features (MinMaxScaler) for Linear Regression to improve gradient descent convergence.
- Random Forest used raw values (scaling not required).

### **MODEL IMPLEMENTATION**

- Linear Regression (Scratch)
  - Implemented cost function, gradient descent, and weight updates manually.
  - Predictions made using learned weights and bias.
- Random Forest (Sklearn)
  - Used `RandomForestRegressor` with 100 trees.
  - Leveraged scikit-learn's efficient tree-based implementation.

### **EVALUATION**

- Metrics: RMSE, MAE,  $R^2$
- Added error handling: evaluation only happens if training is complete.
- Plotted actual vs predicted sales for visual comparison.

## RESULTS

MODEL	RMSE	MAE	R <sup>2</sup>
Linear Regression	533,199.35	440,893.64	0.1175
Random Forest	106,771.64	54,029.61	0.9646

- Linear Regression explained only ~12% of variance, struggling with complex patterns.
- Random Forest explained ~96% variance, showing excellent predictive performance.

## LEARNINGS

- Linear Regression: good for demonstrating fundamentals but limited for real-world, non-linear data.
- Random Forest: captures complex relationships effectively.
- Feature Engineering: extracting date features significantly improved performance.
- Error Handling: ensured robustness by checking training completion before evaluation.

## CONCLUSION

- Random Forest is clearly the better model for this dataset.
- However, the exercise of building Linear Regression from scratch demonstrated understanding of ML fundamentals.
- The project highlights the importance of feature engineering and model choice in predictive tasks.