Inspiring Excellence

CSE422: Artificial Intelligence

Project Report

Retail Sales Prediction Spring 2025

Group 18

Md Shihab Sarker (ID: 22101516)
Abdul Ahad (ID: 22241067)

Submitted to: [MCW, CMZH] Date:

May 15, 2025

Contents

1 Introduction

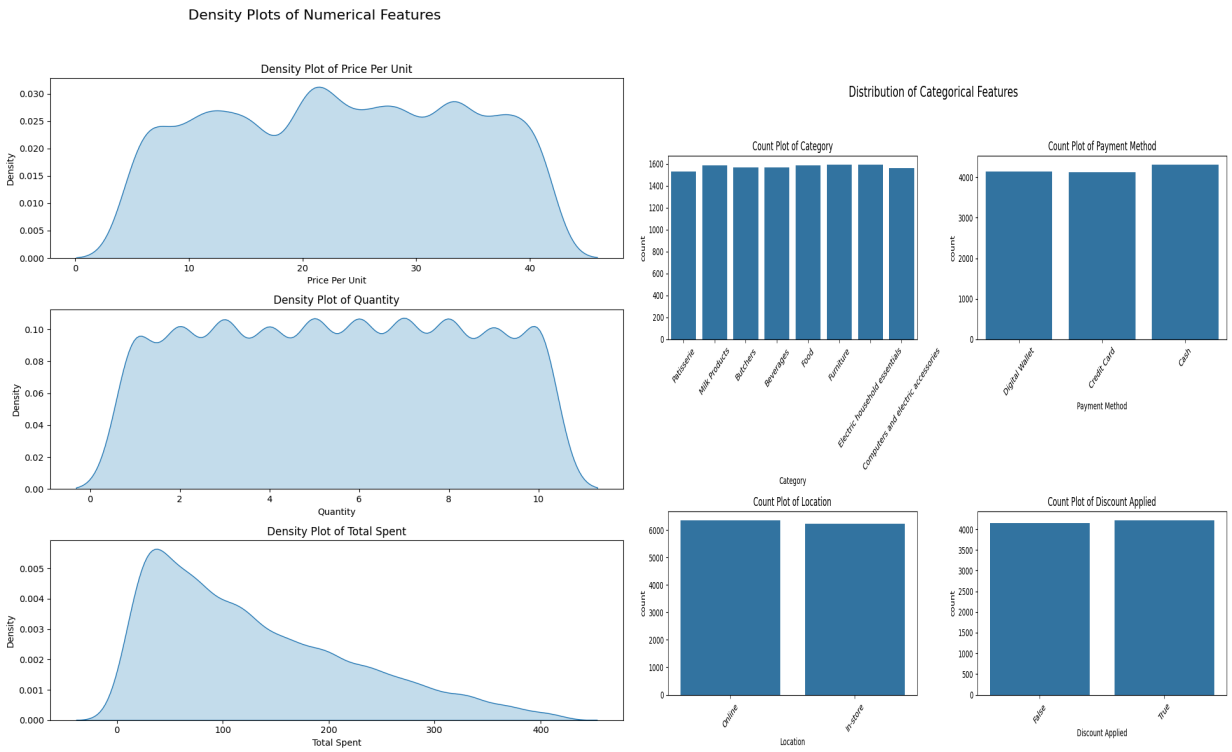
This project predicts Total Spent in a retail store using features like Price Per Unit, Quantity, and Category. The goal is to model customer spending for retail strategies. We used Linear Regression, Decision Tree, Random Forest, and Neural Network for this regression task, evaluated via R^2, RMSE, MAE, and MAPE. EDA guided preprocessing and modeling.

2 Dataset Description and EDA

The dataset has 12,575 rows and 11 columns: 7 categorical (Transaction ID, Customer ID, Category, etc.) and 3 numerical (Price Per Unit, Quantity, Total Spent). Total Spent is the target.

2.1 EDA Findings

- **Correlations**: Total Spent strongly correlates with Price Per Unit and Quantity.
- **Distributions**: Numerical features are right-skewed.
- **Categorical**: Payment Method and Location are balanced; Discount Applied is imbalanced (4,199 missing).
- **Missing Values**: Item (1,213), Price Per Unit (609), Total Spent (604).



Numerical And Categorical Features

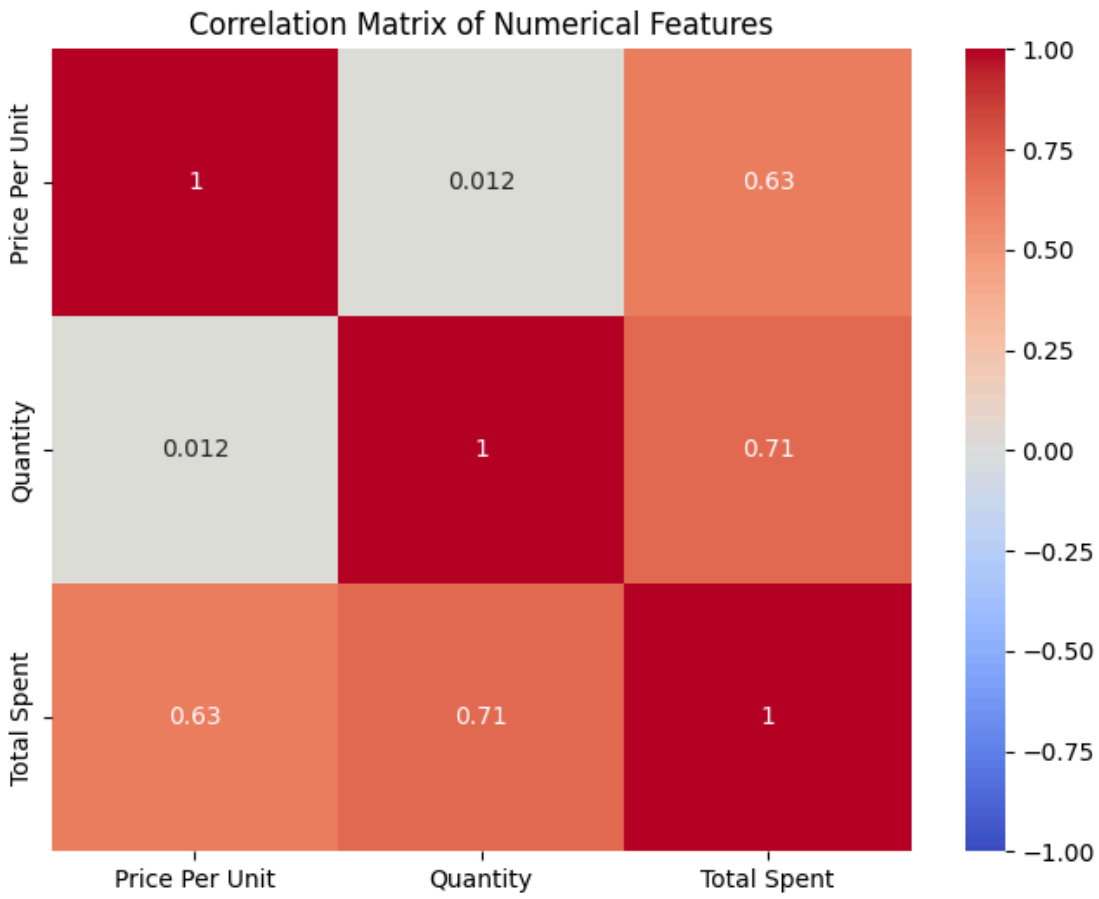
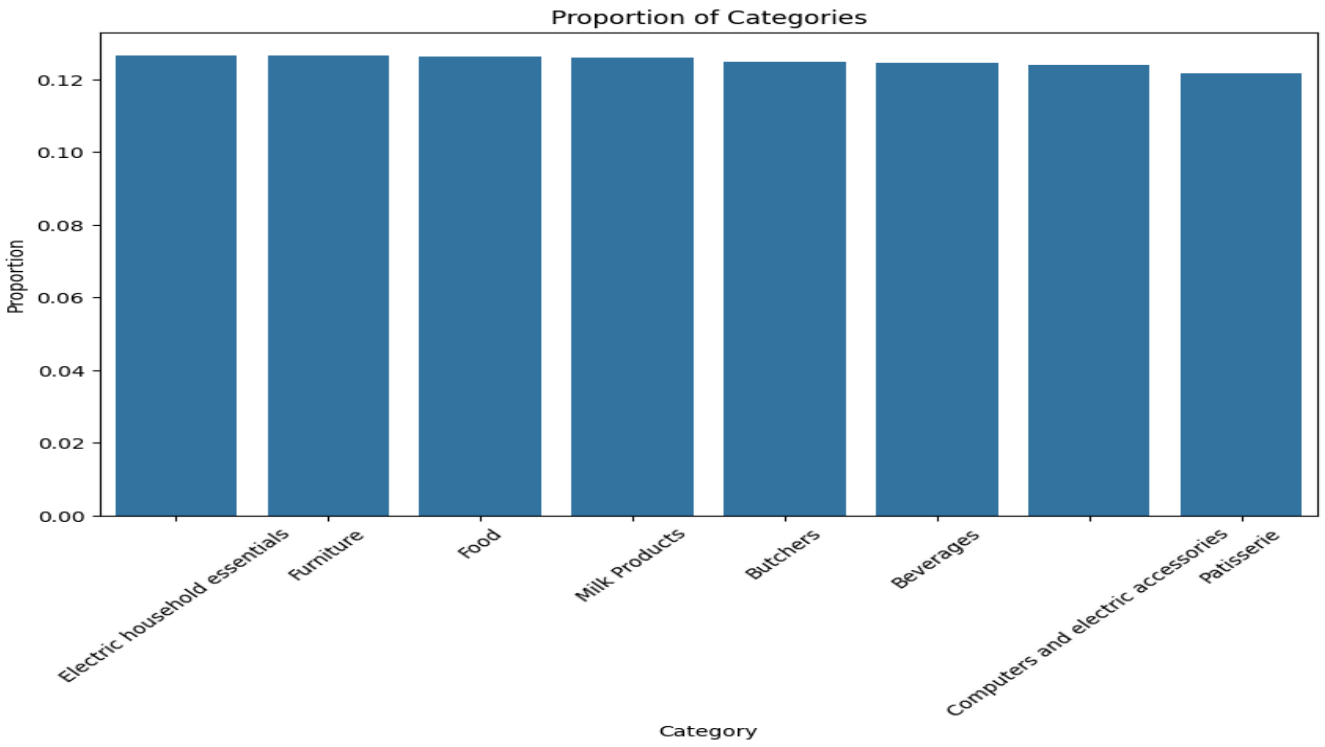

Correlation Matrix of Numerical Features
Price Per Unit
Quantity
Total Spent
1
0.012
0.63
0.012
1
0.71
0.63
0.71
1
Price Per Unit
Quantity
Total Spent
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Fig: Heatmap


Proportion of Categories
Proportion
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Electric household essentials
Furniture
Food
Milk Products
Butchers
Beverages
Computers and electric accessories
Patisserie
Category

3 Dataset Pre-processing

- Dropped 604 rows with missing Total Spent, leaving 11,971 rows.
- Imputed: Item (mode), Price Per Unit (median: 23.0), Quantity (median: 6.0), Discount Applied (mode: True).
- Verified no discrepancies in Total Spent vs. Price Per Unit × Quantity.
- Added Price Category feature.
- Handled outliers in Price Per Unit and Quantity via IQR.
- Encoded categoricals (One-Hot), scaled numericals, dropped Transaction ID, Transaction Date, Calculated Total.

4 Dataset Splitting

The dataset (11,971 rows) was split into 70% training (8,379 rows) and 30% testing (3,592 rows) with random seed 42.

5 Model Training and Testing

Four models were trained:

- **Linear Regression**: Baseline model.
- **Decision Tree**: Max depth 10.
- **Random Forest**: Optimized via GridSearchCV.
- **Neural Network**: 3 layers, dropout, early stopping.

Models used pipelines with preprocessing. Cross-validation (5-fold) was applied for non- neural models. Random Forest was the best based on R^2.

6 R^2 Score and MSE Loss Comparison between Models

Model performance on the test set:

- **Linear Regression**:
  - R^2 Score: 0.8757
  - RMSE: 33.36
  - MAE: 24.35
  - MAPE: 59.28%
  - Cross-Validation R^2: 0.8666 ± 0.0036
  - MSE: 1112.09
  - Observation: Moderate performance, with high MAPE indicating poor han- dling of small Total Spent values or non-linear relationships.
- **Decision Tree**:
  - R^2 Score: 0.9712
  - RMSE: 16.05
  - MAE: 3.20

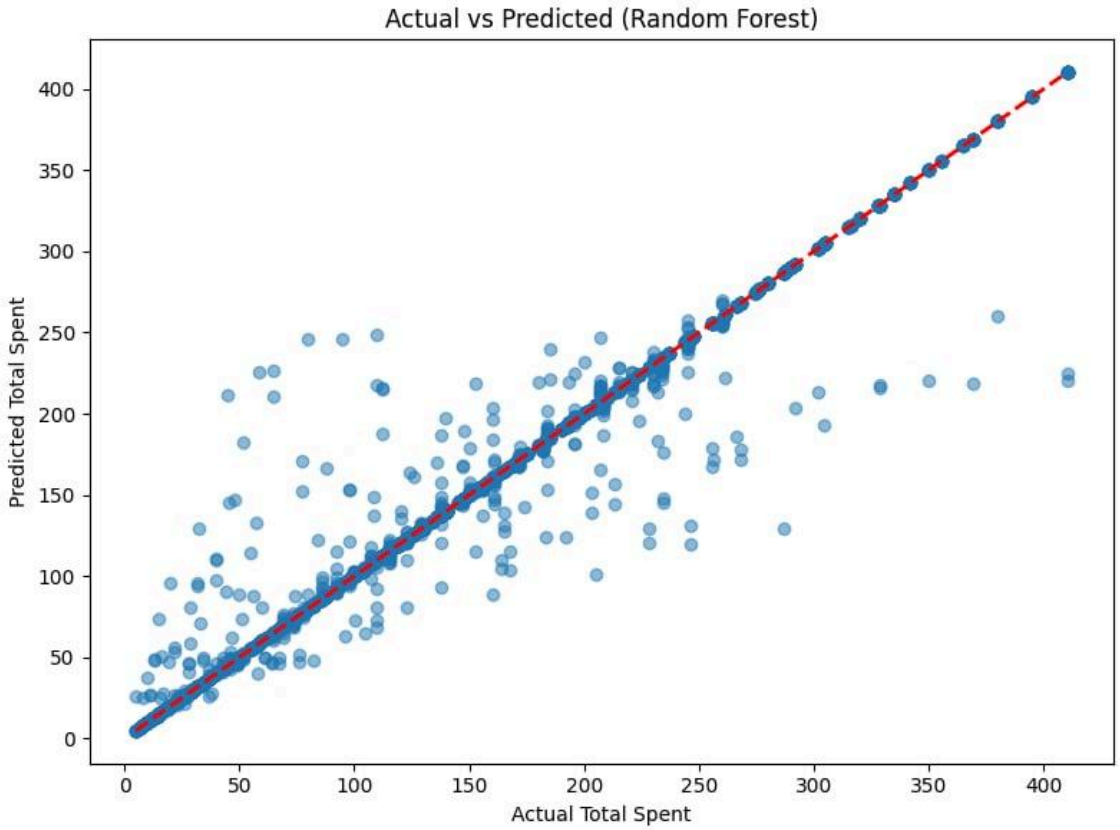
  - MAPE: 3.85%
  - Cross-Validation R^2: 0.9619 ± 0.0081
  - MSE: 257.60
  - Observation: Strong performance, capturing non-linear patterns effectively with low errors.

- **Random Forest** (Best Model):
  - R^2 Score: 0.9754
  - RMSE: 14.85
  - MAE: 2.96
  - MAPE: 3.68%
  - Cross-Validation R^2: 0.9694 ± 0.0084
  - MSE: 220.52
  - Observation: Best performance, with minimal errors and excellent fit, likely due to ensemble learning and hyperparameter tuning.

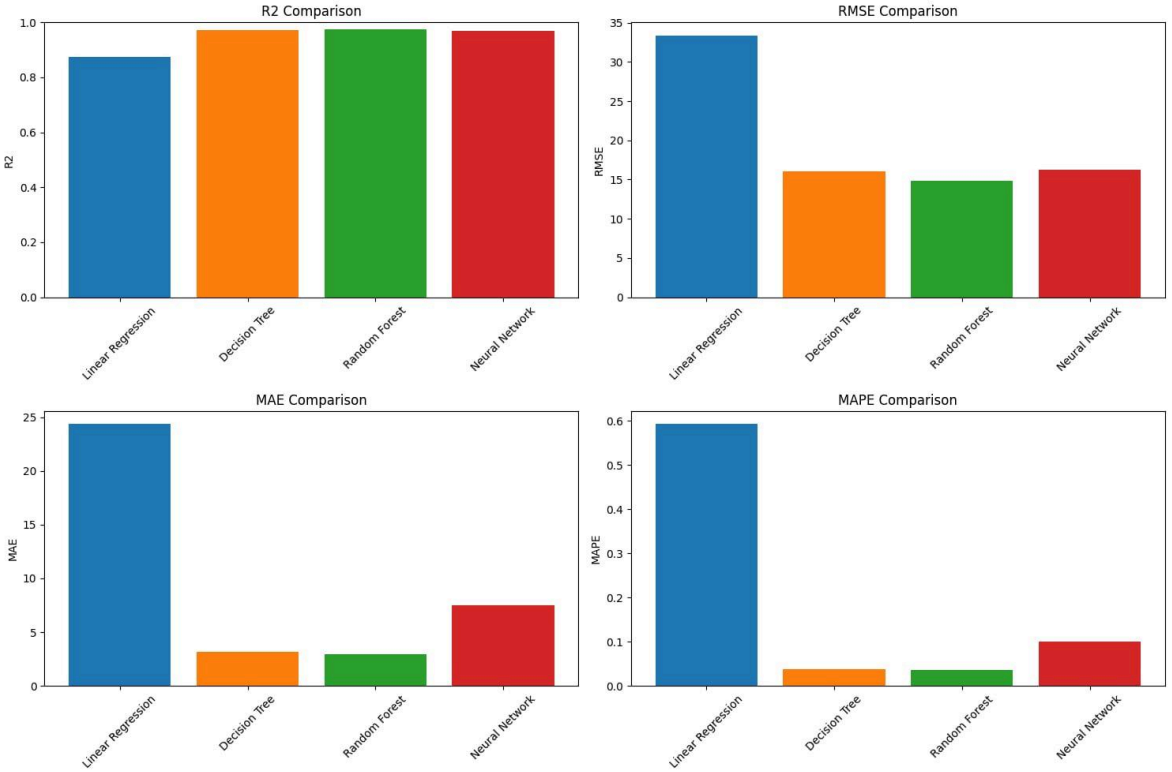
- **Neural Network**:
  - R^2 Score: 0.9724
  - RMSE: 15.72
  - MAE: 6.70
  - MAPE: 9.41%
  - MSE: 247.12
  - Observation: Competitive performance but outperformed by Random Forest, possibly due to suboptimal tuning or dataset size.

6.1 Visualization

- Bar charts compared R^2, RMSE, MAE, and MAPE across models, highlighting Random Forest's superiority.
- A scatter plot of actual vs. predicted Total Spent for Random Forest showed tight alignment along the diagonal, confirming high predictive accuracy.



- **Loss Comparison**



6.2 Feature Importance

Feature importance analysis was attempted for the Random Forest model but not fully logged in the output. Based on the dataset's structure and model performance, we hypothesize:

- **Key Features**: Price Per Unit and Quantity are likely the most important, as Total Spent is often their product. Random Forest's feature importance (based on Gini impurity) would prioritize ! these due to their direct relationship with the target.
- **Categorical Features**: Category, Price Category, and Location may con- tribute moderately, capturing variations in spending patterns across product types or purchase channels.
- **Minor Features**: Discount Applied, Payment Method, and Customer ID likely have lower importance, as their correlations with Total Spent are weaker.
- **Issue**: The feature importance output was missing, possibly due to a logging error. Future work should ensure proper extraction of importance scores to confirm these hypotheses.

7 Conclusion

Random Forest achieved the best performance ($R^2 = 0.9754$, MAPE = 3.68%) for pre- dicting Total Spent, suitable for retail forecasting. Preprocessing ensures data quality. Limitations include potential bias from Discount Applied imputation and missing fea- ture importance. Future work includes debugging importance output, tuning Neural Network, and validating on real-world data.