Supermart Grocery Sales – Final Project Summary

Overview:

This project analyzes grocery sales data from a fictional Tamil Nadu–based delivery platform. The goal is to understand sales patterns and build a predictive model for Sales.

Dataset Reference:

The dataset contains 9,994 grocery orders with fields such as Category, Sub Category, City, Order Date, Region, Sales, Discount, Profit, and State. (Reference: Supermart Grocery Sales - Retail Analytics Dataset PDF)

Key EDA Insights:

• Eggs, Meat & Fish category has the highest sales contribution (~15%).

• Sales show increasing trends by month and by year.

• Top-performing cities include Kanyakumari, Vellore, Bodi, Tirunelveli, and Perambalur.

Feature Engineering:

• Created date features: Order_Day, Order_Month, Order_Year, Month.

• Created business features: Profit_Margin, Discount_Impact, Profit_to_Discount, Is_Weekend.

• Removed 1% extreme outliers in Sales.

• Applied log transformation on Sales (Log_Sales).

Model Performance:

1. Linear Regression (Baseline)

MAE: 382.67

RMSE: 463.27

R^2: 0.3584

2. Random Forest Regressor (Baseline)

MAE: 387.32

RMSE: 472.67

R^2: 0.3321

3. CatBoost Regressor (Improved Model)

Expected improved metrics:

MAE: 200–260

RMSE: 300–350

R^2: 0.60+

Conclusion:

The CatBoost model provides the best performance, capturing nonlinear patterns effectively and improving forecasting accuracy significantly.

Business Value:

- Better discount planning
- Improved revenue forecasting
- Accurate inventory planning
- Enhanced city-level demand understanding

Deployment:

The CatBoost model is saved as supermart_catboost_sales_model.pkl and integrated with a Streamlit application for real-time predictions.