Machine Learning Analysis of Supermarket Product Sales

Name: BibeK Yadav

Course: CAI4002

Date: DEC10

1. Introduction

The goal of this project is to analyze supermarket product performance using classical machine learning techniques. Using the provided product_sales.csv dataset, we clean and preprocess the data, apply K-means clustering (implemented from scratch) to identify groups of similar products, and train regression models to predict monthly profit. Finally, we present the results through a simple Streamlit dashboard designed for non-technical supermarket managers.

2. Data Preprocessing

The dataset contains 200 products with variables including product_id, product_name, category, price, cost, units_sold, promotion_frequency, shelf_level, and profit. There were 4 missing values in product_name, which were imputed using neutral placeholders of the form Unknown_. No numeric variables had missing values, so no rows were dropped.

To handle outliers, the IQR rule was applied to price, cost, units_sold, promotion_frequency, shelf_level, and profit. Values beyond Q1 - 1.5×IQR and Q3 + 1.5×IQR were capped (winsorized) at these bounds. This preserved all 200 products while preventing a few extreme values from dominating clustering and regression.

For K-means clustering, numeric features used for clustering (price, cost, units_sold, promotion_frequency, and shelf_level) were standardized using Z-scores so that each feature has mean 0 and standard deviation 1. This is necessary because K-means relies on Euclidean distance and unscaled features with larger numeric ranges would otherwise dominate the distance calculations.

3. K-means Clustering

K-means clustering was implemented from scratch using NumPy, without scikit-learn's clustering utilities. The algorithm repeatedly assigns each product to the nearest centroid, then recomputes centroids as the mean of points in each cluster, until centroid movement falls below a tolerance or a maximum number of iterations is reached. To reduce sensitivity to initialization, K-means was run multiple times with different random seeds and the solution with the lowest within-cluster sum of squares (WCSS) was chosen.

To select the number of clusters, K was varied from 2 to 8 and WCSS was recorded. The elbow plot of WCSS against K shows a clear bend around K = 4, after which additional clusters provide diminishing returns. Therefore, K = 4 was chosen as the optimal number of clusters.

Cluster-level summaries reveal four intuitive groups: - Cluster 0: Promo-Driven Budget Best-Sellers – low price, very high units sold, high profit, frequent promotions, dominated by beverages and snacks. Cluster 1: Mid-Range Steady Performers – medium price and volume, strong and stable profit, a mix of bakery, dairy, and beverages with good shelf placement. - Cluster 2: Low-Visibility Value Items – moderate prices and volumes but fewer promotions and lower shelf levels, with many produce and snack items; these appear under-promoted and under-exposed. - Cluster 3: Premium Low-Volume

Specialties – high price, low volume, moderate profit; mostly meat and dairy products that offer high per-unit margins but sell less frequently.

4. Regression Analysis

The target variable for regression is monthly profit. Features include price, cost, units_sold, promotion_frequency, shelf_level, and one-hot encoded product category. The data was split into 80% training and 20% testing sets.

Two models were trained and compared: 1) Linear Regression, which assumes a linear relationship between the features and profit. 2) Polynomial Regression of degree 2, which uses polynomial features (squared terms and pairwise interactions) followed by linear regression.

Model performance was evaluated using Mean Squared Error (MSE) and Mean Absolute Error (MAE) on both training and test sets. The linear model achieved a test MAE of about $68, whereas the degree-2 polynomial model reduced the test MAE to approximately $23. This indicates that the polynomial model captures important nonlinear relationships and interactions between features.

The polynomial model does show some signs of overfitting, with training error much lower than test error, but it still generalizes significantly better than the simple linear model. Therefore, the polynomial regression model was selected as the preferred predictor of profit.

5. Visualization and User Interface

To present the results to non-technical users, a Streamlit dashboard was created. The dashboard includes: - A data overview section showing a sample of the dataset and explanations of cleaning steps. - An elbow plot to illustrate the choice of K = 4 clusters. - A scatter plot of price versus units sold, with colors indicating clusters and centroids marked for clarity. - A regression section comparing linear and polynomial models with a metrics table and an actual-versus-predicted profit plot. - An interactive "Profit What-If Calculator" where a user can adjust sliders for price, cost, units sold, promotion frequency, shelf level, and category and see the model’s predicted monthly profit.

This interface allows a supermarket manager to explore how similar products group together and how changes in pricing and promotion might impact expected profit.

6. Conclusions, Limitations, and Future Work

The analysis shows that supermarket products can be meaningfully segmented into four clusters with distinct pricing, promotion, and sales characteristics. A degree-2 polynomial regression model can predict monthly profit with reasonably low error, providing useful guidance for pricing and promotion decisions.

Key limitations include the use of only a single snapshot of monthly data, the absence of explicit business constraints in the model (for example, profit should be negative if price is below cost), and the lack of external factors such as competition or seasonality. Future work could incorporate time-series

data, regularized regression models (such as Ridge or Lasso) to control complexity, and more sophisticated clustering or dimensionality reduction methods.

In summary, this project demonstrates how classical machine learning techniques—implemented and interpreted carefully—can provide actionable insights into product performance and support data-driven retail decisions.

7. Use of AI Tools

A generative AI assistant was used to help brainstorm the project structure, suggest code snippets for data preprocessing, K-means implementation, regression modeling, and Streamlit UI components. All code was run, tested, and understood