Course: CAI 4002 Artificial Intelligence

Professor: Xian Su

Authors:

- Eduardo Goncalvez (PID: 6526311)
- Alex Waisman (PID: 6529880)
- Ivan Salazar (PID: 6237206)

Repository Link: https://github.com/Edugre/ML_Assignment

Regression & K-Means Analysis Report

1. Introduction

This project analyzes a supermarket sales dataset to identify meaningful product clusters and develop predictive models for profit estimation. This work consists of three major components:

1. Data Pre-processing.
2. K-means clustering (Implemented from scratch).
3. Regression model development for profit prediction.

The goal is to improve data quality, segment products into actionable groups, and determine which modeling approach achieves the strongest predictive performance.

2. Dataset Description

High-quality preprocessing is essential for both clustering and regression. The following steps were applied to the raw dataset to ensure consistency and reliability.

2.1 Missing Value Handling

Missing values were first analyzed to determine their frequency across attributes. Two strategies were applied, depending on the extent of the missingness:

1. **Product Name Repair:** Missing product names were filled by referencing other rows with the same product_id. If no such match existed, a placeholder label was assigned.
2. **< 5% Missing:** Numerical columns with small amounts of missing data were imputed using the median. As this method is robust to skewed distributions.
3. **>= 5% Missing:** Columns exceeding 5% missing values were flagged, and all rows missing values in those columns were removed. This avoids introducing excessive bias through imputation.

This approach ensured minimal distortion while retaining the integrity of the dataset.

2.2 Outlier Detection and Treatment

Outliers were detected using the Interquartile Range (IQR) method:

- Calculated Q1 and Q3
- Computed IQR = Q3 – Q1
- Defined outlier bounds as:
  - Lower Bound = Q1 – 1.5 * IQR
  - Upper Bound = Q3 + 1.5 * IQR

Instead of removing outliers, they were capped to the nearest acceptable bound. This prevented extreme values from distorting model training while preserving overall dataset size and structure.

2.3 Normalization and Feature Scaling

Because K-means clustering relies on Euclidean distance, feature scale must be standardized. After outlier capping, all numerical features used for clustering were transformed using Z-score standardization:

$$z = \frac{x - \mu}{\sigma}$$

Standardizing the data ensured equal contribution from each feature and supports stable centroid convergence. Z-score was the more appropriate choice given the capped distributions.

3. K-Means Clustering

K-Means clustering was fully implemented from scratch following the algorithm's five core steps: Centroid initialization using k-means++, assignment, updated, convergence check, and final labeling.

3.1. Elbow Method and Optimal K selection

The algorithm was executed for K = 2 to 8, and the Within-Cluster Sum of Squares (WCSS) was recorded for each value. The elbow curve indicated a sharp drop in WCSS from k=2 to k=3, with diminishing results thereafter.

A reasonable elbow appeared between k = 3 and k = 4. To improve segmentation detail while avoiding overfitting, k = 4 was selected as the optimal number of clusters, which were used to analyze averaged attributes: price, units sold, cost, profit, and promotion frequency.

3.2. Visualization

Two key visualizations were produced:

- Elbow Curve: Shows diminishing WCSS improvement past K = 4.
- 2D Cluster Scatter Plot: Uses standardized price and units sold to illustrate cluster boundaries and centroid positions.

These visuals support the selection of K = 4 and offer intuitive insights into product groupings.

4. Regression Analysis

The goal of the regression portion was to build predictive models for profit, comparing multiple algorithms and evaluating their performance using MSE, MAE, and R^2.

4.1. Linear Regression

A basic linear model was trained as a baseline.

- Test R^2: ~0.49
- MAE: ~108.29

Performance was moderate, indicating nonlinear relationships between predictors and profit.

4.2. Polynomial Regression (Degrees 2 and 3)

To capture nonlinear interactions, polynomial features were generated:

Degree 2:

- Test R^2: 0.895
- Strong improvement over linear regression
- MAE: ~42.26

Degree 3:

- Test R^2: 0.932
- Lowest error metrics across all models.
- Best overall predictor of profit.

Polynomial Regression (Degree 3) was the top-performing model.

4.3. Ridge Regression

Ridge regression was applied to reduce variance and mitigate overfitting. While more stable than linear regression, it did not outperform polynomial regression.

4.4. Model Visualization

For each model, the following were visualized:

- Actual vs. Predicted Profit scatter plot
- Residual distribution (to assess model fit)

Polynomial degree 3 showed the tightest clustering around the ideal prediction line and the most symmetric residual distribution.

5. Conclusion

This project successfully applied preprocessing, clustering, and regression techniques to generate actionable insights from a supermarket sales dataset.

Key findings:

- Z-score standardization and IQR capping produced stable inputs for K-means.
- The elbow method supported selecting k=4, resulting in meaningful product segments.
- Cluster 2 (“Budget Best-Sellers”) delivered the highest sales volume, while clusters 0 and 3 offered premium-price opportunities.
- Polynomial Regression (degree 3) achieved the strongest profit predictions with $R^2 = 0.93$.

Overall:

The combined clustering and regression framework provides a strong foundation for retail decision-making, including pricing strategy, inventory management, and product portfolio optimization.

6. Use of AI Tools

AI tools were used for brainstorming, reviewing preprocessing logic, and refining the report narrative. All code, modeling decisions, interpretations, and final outputs reflect the student’s own analysis and implementation.