Name: Sejal Patil | 240516 | CSE IV

DAP Project Report: Product Price Classification Using Logistic Regression

1. Overview

The aim of the project is to create a full data analytics workflow that will classify clothing products into categories of prices. Instead of predicting price as a continuous numerical value, the project reframes the task as a binary classification problem: Standard-Priced versus High-Priced. This allows for more practical insights into understanding pricing patterns and informing decision-making in retail contexts.

2. Data Source

The analysis utilized the “Clothes Price Prediction” dataset, publicly available on Kaggle. This dataset includes features such as product category, brand, season, and user ratings, providing a basis for investigating the drivers of clothing prices.

Dataset Link: Clothes Price Prediction on Kaggle

3. Methodology

A multi-stage methodology was employed, moving from data preparation and exploration to statistical validation and predictive modeling.

a) Data Preparation & Feature Engineering
The dataset was cleaned and processed. The core feature engineering step was the creation of the binary target variable, Price_Category, by binning the Price column:

- 1 → High-Priced ((Price >= $108.00))
- 0 → Standard-Priced ((Price ≤ $108.00))
  This approach ensured the classes were balanced, creating a suitable foundation for a classification model.

b) Exploratory Data Analysis (EDA)
EDA was conducted to identify visual patterns that might separate the two price categories. Visualizations explored the relationships between Price_Category and features like Category, Brand, and Season to form initial hypotheses.

c) Statistical Validation: One-Way ANOVA
To rigorously test the visual patterns from EDA, a One-Way ANOVA test was performed to determine if there was a statistically significant difference in mean prices across the various product categories.

Metric	Value
F-statistic	1.3891
p-value	0.2257

Interpretation: With a p-value of 0.2257 (which is > 0.05), we fail to reject the null hypothesis.

Confidence Interval Confirms the ANOVA Results:

Observation: The confidence intervals for all categories show **significant overlap**. For instance, the price range for a 'Dress' ($106.67 - $122.34) overlaps almost completely with that of a 'Sweater' ($98.70 - $114.86). This overlap confirms that we cannot be confident that any one category is truly priced differently from another, supporting the ANOVA conclusion.

d) Predictive Modeling: Logistic Regression
A Logistic Regression classifier was trained on the prepared features to predict the Price_Category. This model was chosen as a baseline due to its interpretability.

e) Model Evaluation
The performance of the model was measured using:

- ➢ Confusion Matrix
- ➢ Accuracy Score
- ➢ Precision & Recall for each class
- ➢ Classification Report

The results were:

- ➢ **Precision:** 49% for Standard-Priced and 47% for High-Priced.
- ➢ **Recall:** 47% for Standard-Priced and 49% for High-Priced.
- ➢ **Accuracy: ~0.48**

This accuracy score is no better than random guessing, confirming that the model was unable to find any meaningful, predictive patterns in the data.

4. Key Takeaways

a. Product Category Does Not Strongly Influence Price
ANOVA results and visual patterns indicate that although the mean prices differ slightly across product categories, the differences are not statistically significant.

b. Median-Based Binning Creates Useful Price Groups
By categorizing price into two groups using the median, the dataset becomes balanced.
This approach:

- ➢ Reduces model training time
- ➢ Improves interpretability
- ➢ Enables classification with a binary target

c. Model Failure Confirms Weakness of Features
The Logistic Regression model's accuracy of ~48% is a direct consequence of the feature weakness identified by our statistical tests. The model's inability to perform better than a random guess serves as the final confirmation that:

- ➢ The available features lack the necessary signal to classify price.
- ➢ The data does not contain consistent, learnable patterns for this task.

d. The Workflow Demonstrates End-to-End Analytics
The project includes:

- ➢ Data preprocessing
- ➢ Exploratory Data Analysis (EDA)
- ➢ Statistical validation (ANOVA, correlation analysis)
- ➢ Model building
- ➢ Performance assessment

5. Conclusion

In this project, logistic regression was used to classify clothing items into price categories through a full data analytics workflow.

Although the final predictive performance of the model was poor, the analysis provided clear insights into the lack of strong relationships between the available features and price variation.

Both the ANOVA test and correlation analysis indicate that the features have minimal predictive power. Categories differ slightly in mean price, but not significantly, and all features show near-zero correlation with price. As a result, the logistic regression model could not learn meaningful patterns, explaining its poor performance (~48% accuracy).

Future improvements should not focus on more complex models, as they would likely fail for the same reasons. The only viable path to building a successful model is to **acquire a richer dataset**. Success would require incorporating more granular features-such as *style attributes, quality metrics, and seasonality*-which are the true drivers of complex pricing decisions in the fashion industry.