Predicting SHEIN Product Ratings Using Similar Items' Reviews and Features

Ofir Herrera-318971694

August 2025

git clone link : `https://github.com/Ofir-Herrera/Fashion-project-Statistical-Analysis.git`

Abstract

This project investigates whether it is possible to predict whether a clothing item from the SHEIN website will be rated as "good" or "bad" based on data from similar items that already have customer reviews. A custom web crawler was developed to collect product data, including attributes, prices, and detailed ratings. The dataset underwent preprocessing steps including handling missing values, removing duplicates, and cleaning categorical and numerical features. Exploratory data analysis (EDA) was conducted to examine relationships between features using Spearman correlations and Chi-Square tests, followed by the construction of a binary target variable that incorporated both average ratings and the number of ratings per product. Logistic regression was employed as the predictive model, with model evaluation carried out using ROC and Precision–Recall curves, as well as F1-score. After removing features causing data leakage, the model achieved high performance (AUC $\approx$ 0.989, AP $\approx$ 0.976), demonstrating that customer feedback patterns from similar products can be effectively leveraged to predict the quality classification of unrated clothing items. The findings highlight the potential for leveraging historical review data to improve recommendation systems and assist consumers in decision-making.

1 Introduction

The rapid growth of e-commerce has transformed the way consumers evaluate and purchase products, with online reviews becoming a primary source of information for decision-making. Websites such as SHEIN, a major global fast-fashion platform, offer millions of clothing items, many of which lack sufficient customer feedback at the time of listing. This creates a challenge for both consumers, who face uncertainty about product quality, and retailers, who aim to improve recommendation systems and inventory decisions.

This project aims to address this challenge by predicting whether a newly listed clothing item on SHEIN will be classified as "good" or "bad" based on the reviews and attributes of similar items that already have feedback. By combining product metadata (such as price, material, and style) with aggregated customer rating information, a machine learning model is developed to provide accurate quality predictions. This approach can help consumers make informed purchasing decisions while supporting businesses in optimizing product visibility and marketing strategies.

2 Methods

The methodology consisted of five main stages: data collection, data preprocessing, exploratory data analysis (EDA), target variable creation, and model development and evaluation.

2.1 Data Collection

A custom web crawler was implemented to extract product information from the SHEIN website. For each clothing item, the following data was collected: product identifier (SKU), categorical attributes (e.g., material, pattern type, sleeve length), numerical attributes (e.g., price, points earned), and detailed ratings for both textual and image reviews (from 1-star to 5-star).

2.2 Data Preprocessing

The dataset initially contained **15,792** records and **57** features. After removing duplicates and filling missing values in categorical columns with the mode, the dataset was reduced to **13,796** records and **27** features, as shown in Table 1. This cleaned dataset was used for all subsequent analyses.

Table 1: Dataset size before and after data cleaning

	Rows	**Columns**
Before cleaning	15,792	57
After cleaning	13,796	27

2.3 Exploratory Data Analysis (EDA)

EDA was conducted to identify patterns, assess feature relationships, and inform model design. Both numerical and categorical variables were examined.

Continuous Features. Figure 1 shows histograms for `Points Earning` and `Price`. Both distributions are right-skewed, with most values clustered at the lower end. The strong correlation ($\rho \approx 0.99$) between these variables suggests redundancy, meaning only one may be necessary for predictive modeling.

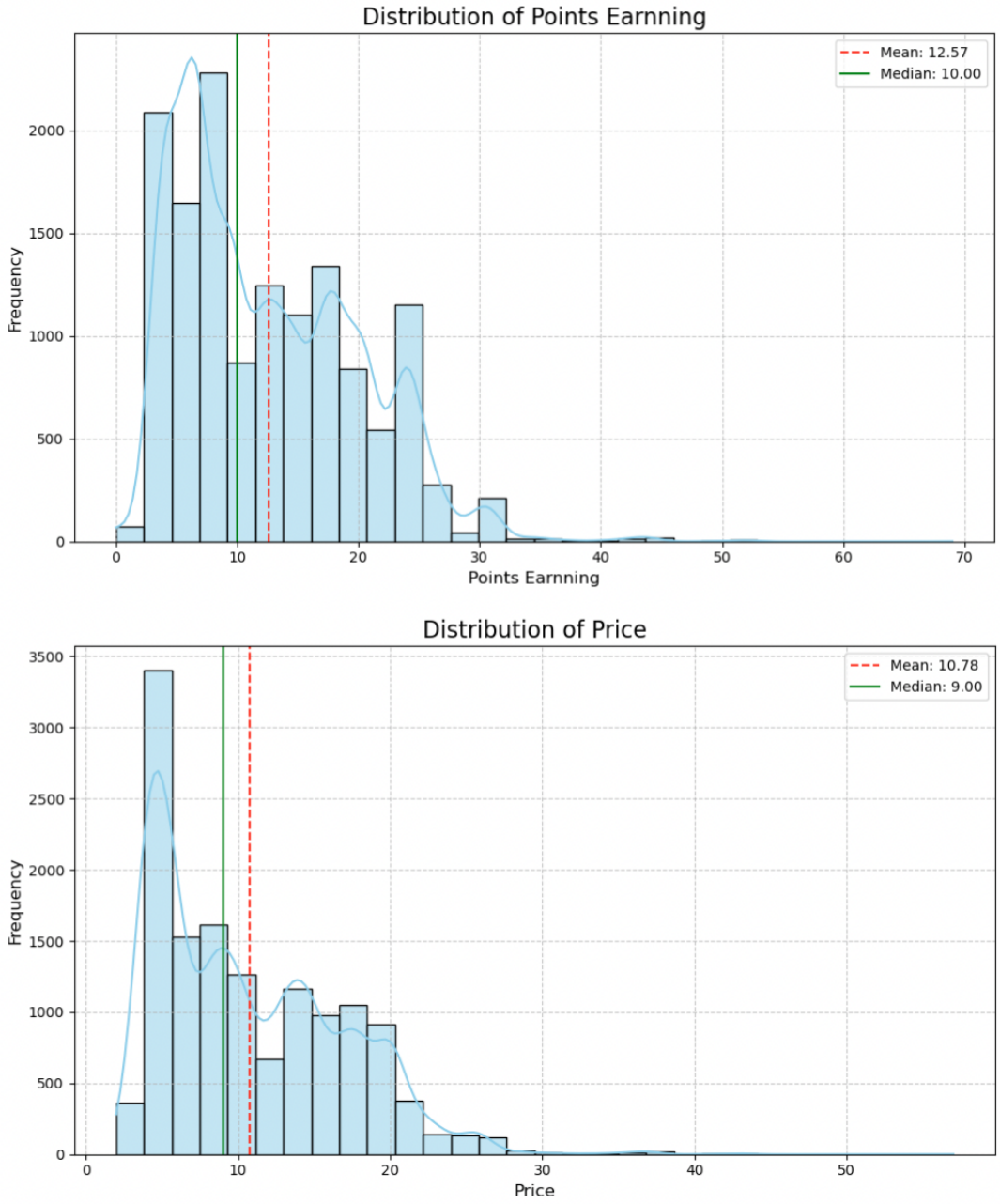


Figure 1: Distribution of Points Earning (top) and Price (bottom) with mean and median markers.

Categorical Features. Figure 2 shows the most common categories for `Style`, `Fabric`, `Season`, and `Fit Type`. For example, "Casual" dominates Style, while "Slight Stretch" is the most common fabric. These patterns suggest which product attributes are prevalent in the dataset, which can influence the model's bias.

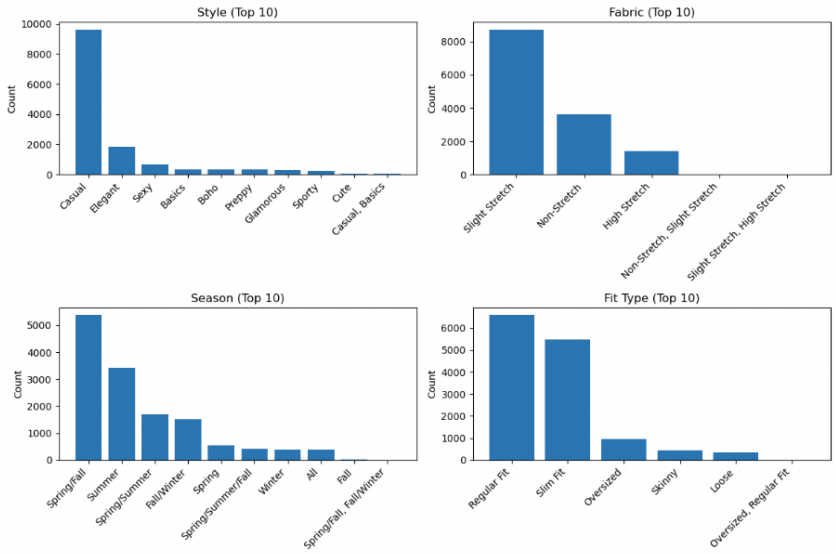


Figure 2: Top 10 categories for Style, Fabric, Season, and Fit Type.

Target Variable. The binary target variable (`Classification`) was created during feature engineering using a two-step process. First, the *average rating* for each product was computed from both textual and image reviews. Second, the *total number of ratings* was incorporated to ensure label reliability, so that items with very few reviews would not be labeled as "good" solely due to a high but noisy average. An initial threshold based on the first quartile (Q1) of the weighted average rating produced biased data and low model accuracy (approximately 60%). To address this, the threshold was redefined using the **median** of the weighted average rating, ensuring a balanced split between classes, together with a minimum number of ratings requirement and a minimum average rating of 4.0 out of 5. Items meeting these conditions were labeled as "good" (1), and all others as "bad" (0). As shown in Figure 3, this adjustment resulted in a perfectly balanced dataset (50% good, 50% bad), which is ideal for unbiased model training.

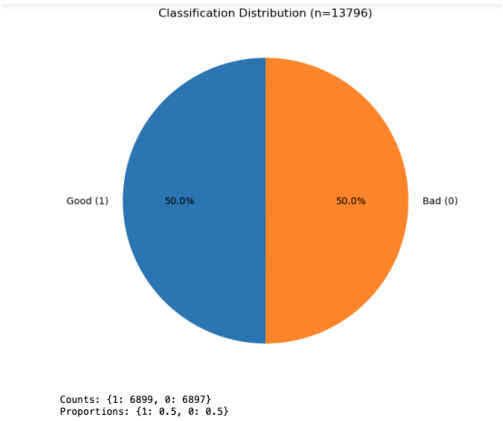


Figure 3: Distribution of the binary target variable (Classification).

Correlation Analysis. Figure 4 presents two correlation heatmaps:

(Left) Spearman Correlation: This heatmap shows monotonic relationships between continuous features and the target variable. Dark red tones along the `Classification` row/column indicate strong positive correlations, particularly for rating-based features such as `num_review_ratings`, `Total Ratings`, and `Weighted Combined Score` ($\rho \approx 0.85$–0.87). The high intercorrelation ($\rho \geq 0.97$) among these features

confirms the presence of multicollinearity. This insight guided the decision to remove features that risked leaking target information while retaining one representative rating-based predictor.

(Right) Chi-Square p-values: This heatmap visualizes the statistical association between categorical features and the target variable, where darker colors (low p-values) indicate stronger associations. Most categorical variables (e.g., `Style`, `Season`, `Fabric`) exhibit statistically significant relationships with the target ($p < 0.05$), meaning their distribution differs substantially between "good" and "bad" items. These findings supported the inclusion of categorical attributes in the model, as they provided complementary predictive power to the numerical features.

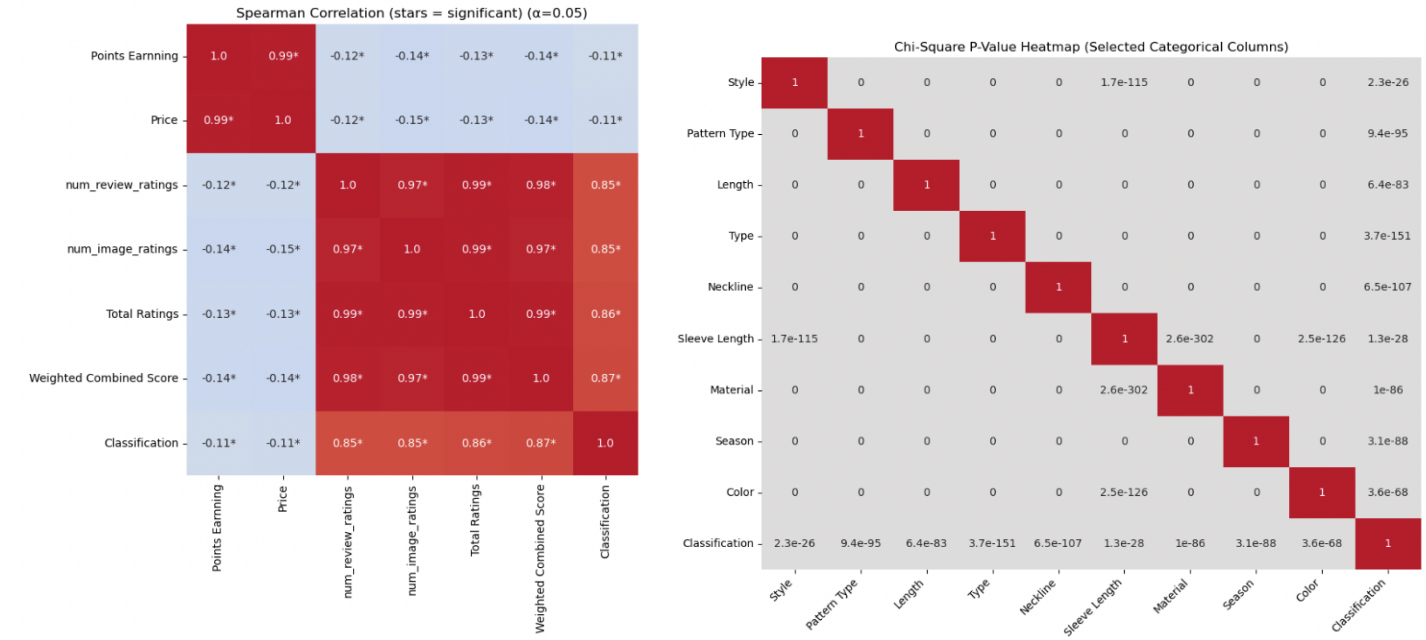


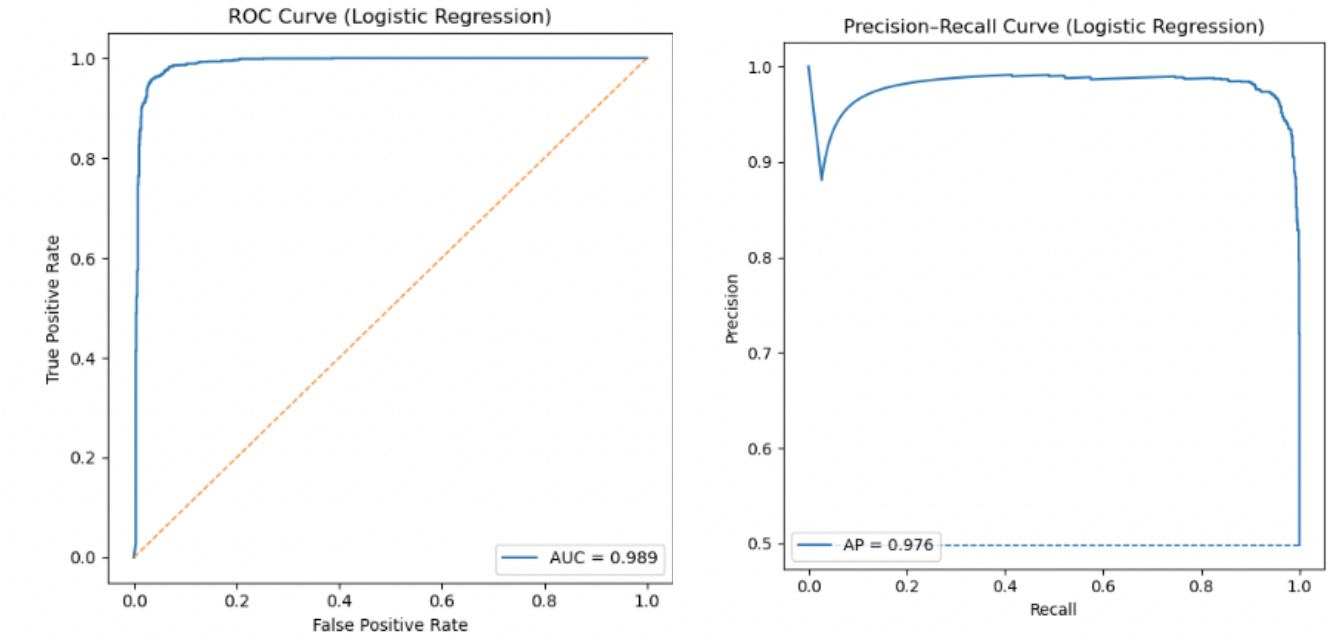
Figure 4: Left: Spearman correlation heatmap of continuous variables and the target. Right: Chi-Square p-value heatmap of categorical variables.

2.4 Feature Engineering and Selection

A Chi-Square test revealed statistically significant associations among most categorical features. Leakage-causing variables (`SKU`, `num_review_ratings`, `Total Ratings`, `Weighted Combined Score`) were dropped. The remaining features were retained for maximum predictive power.

3 Results

The logistic regression model was evaluated using ROC and Precision–Recall curves (Figure 5) and a confusion matrix (Figure 6). Performance metrics are shown in Table 2.



(a) ROC Curve (AUC = 0.989).

(b) Precision–Recall Curve (AP = 0.976).

Figure 5: Performance curves for Logistic Regression.

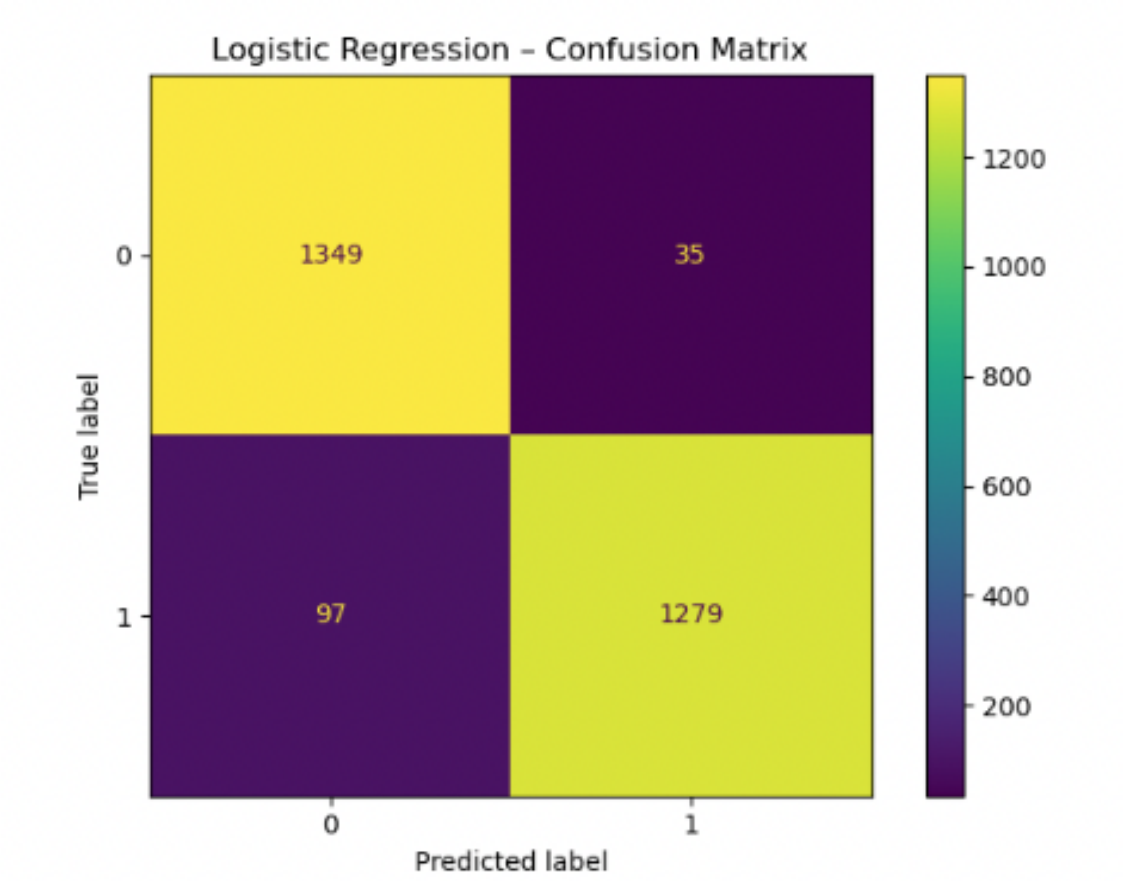


Figure 6: Confusion Matrix (Logistic Regression). TP = 1279, TN = 1349, FP = 35, FN = 97.

Table 2: Logistic Regression performance on the test set

Metric	AUC	AP	Accuracy	Precision	Recall	F1
Value	0.989	0.976	0.952	0.973	0.930	0.951

Interpretation. The model achieved near-perfect class separability (AUC = 0.989) and high precision even at high recall (AP = 0.976). The confusion matrix confirms a low error rate, with few false positives and negatives. These results were obtained *after* removing leakage-prone features, confirming that performance is not inflated by label proxies.

4 Discussion

One of the key challenges in this project was the construction of the binary target variable. An initial threshold based on the first quartile (Q1) of the weighted average rating produced a highly imbalanced dataset and limited the logistic regression model to approximately 60% accuracy with poor recall for the minority class. By redefining the threshold to the median of the weighted average rating and requiring a minimum number of ratings (and a minimum average rating of 4.0), the dataset achieved perfect class balance, and performance improved substantially (AUC = 0.989, AP = 0.976).

Overall, the analysis confirms that logistic regression can accurately predict clothing quality on SHEIN using product attributes and historical review data. Rating-based features were the strongest predictors but required careful handling to avoid leakage. While multicollinearity was present, it did not harm predictive

accuracy. Limitations include platform-specific data and dependency on customer engagement. Future work could explore additional algorithms and external datasets for generalization.