Data Preprocessing and Feature Engineering

Before model training, extensive preprocessing and feature engineering steps were conducted to prepare the dataset for analysis. The original dataset comprised a mix of textual, numerical, and boolean features.

**3.1 Data Cleaning**

Initial exploratory data analysis (EDA) was performed to understand the structure and quality of the data. Redundant columns that did not contribute meaningful information were removed. Additionally, all rows containing missing values were dropped to ensure the dataset remained clean and consistent for downstream modeling.

**3.2 Target Variable Transformation**

A histogram of the target variable, price, revealed a right-skewed distribution, indicating a concentration of low-price items and a long tail of high-price items. To address this skewness and stabilize variance, a log-transformation was applied to the price column. This transformation enhances model performance by bringing the distribution closer to normality.

**3.3 Feature Engineering**

To improve the predictive capacity of the models, several new features were engineered:

**Color Features**: The color column, which contained textual descriptions, was used to derive three binary indicators:

* + plain\_color: equal to 1 if the item was a solid color.
  + pattern: equal to 1 if the item featured any kind of pattern.
  + print: equal to 1 if the item had printed designs.

These distinctions were made using regular expressions to parse the text content from mainCatCode column and assign the appropriate flags.

**Category-Based Features**: The mainCatCode column was parsed using regular expressions to extract multiple layers of categorical information:

* 1. **Premium Indicator**: A binary premium flag was created, with value 1 for items belonging to premium product lines.
  2. **Gender Tags**: Three binary gender flags—men, ladies, and neutral—were generated to classify the target audience.
  3. **Item Category Flags**: Additional binary features were created for specific product categories such as shoes, bottoms, jeans, tops, socks, outerwear, accessories, underwear, dresses, sportswear, and loungewear, which were further divided into even smaller categories. These categories were inferred from textual patterns.

**Material Tier Features**: The materials column, containing composition text, was parsed to extract dominant materials using regex, filtering only those with ≥50% presence. Each item’s key materials were then matched against three predefined quality tiers:

* **High-tier** (e.g., leather, silk, cashmere)
* **Mid-tier** (e.g., cotton, linen, modal)
* **Low-tier** (e.g., polyester, acrylic, plastic)

Binary flags (high\_tier, mid\_tier, low\_tier) were assigned based on material presence, allowing multiple tiers per item. For instance, a product made of 70% of leather, 60% of cotton and 10% of polyester would have high\_tier = 1, mid\_tier = 1, low\_tier = 0 for the threshold set at 0.5. This structured encoding captures the value impact of materials for price prediction.

**Textual Embedding**: To utilize unstructured textual information, the details column underwent TF-IDF (Term Frequency–Inverse Document Frequency) vectorization. This technique transforms textual content into numerical features based on the relevance and uniqueness of words, improving the model's ability to capture semantic cues from product descriptions.

**Feature Utility**

All engineered features were developed with the goal of enhancing the model’s ability to accurately predict item price. By isolating key product attributes—such as category, composition, visual design, and description—this enriched feature space provides meaningful signals that support more robust learning and generalization.