Multimodal Price Prediction Using Text and Image Fusion with LGBMRegressor

Arpan Pramanick, Rounak Koner, Unnati Mishra, Debargha Mitra Roy

---

1 Executive Summary

Our solution involves the use of a multimodal machine learning model, which combines the textual and visual data about products in order to precisely estimate the best prices. With a `SentenceTransformer`-based text embedding, `ResNet50` visual representations, and a `LightGBM` regression model, we are able to sum up the semantic and aesthetic indicators that affect price. Such a hybrid approach provides more interpretability, scalability and predictive accuracy across different product categories.

2 Methodology Overview

2.1 Problem Analysis

The problem aimed to predict optimal product prices using multimodal data, combining textual catalog descriptions and product images. The dataset contained missing text fields, which were filled with empty strings, and all text was standardized to lowercase. Exploratory analysis revealed that textual content encoded product category, brand, and quality cues, while visual data offered complementary insights like color, texture, and design. Together, these modalities provide a holistic representation of the product, making a multimodal strategy ideal.

Key Observations:

The fusion of textual and visual features significantly improved price prediction accuracy compared to single-modality models. Text embeddings captured semantic cues like brand and quality, while image features added visual context, resulting in a balanced and robust multimodal pricing model.

2.2 Solution Strategy

We are proposing to use a multimodal machine learning model, which will unite the textual and the visual data concerning products with the aim of accurately approximating the most appropriate prices. We can combine semantic and aesthetic indicators which influence price with a `SentenceTransformer`-based text embedding, `ResNet50` visual representations, and a `LightGBM` regression model. A hybrid strategy offers higher interpretation, scale and predictive validity between various product lines.

Approach Type: Hybrid Multimodal Model

Core Innovation: Fusion of `SentenceTransformer` text embeddings and `ResNet50` visual features, trained with a `LightGBM` Regressor for accurate price prediction.

Workflow Summary:

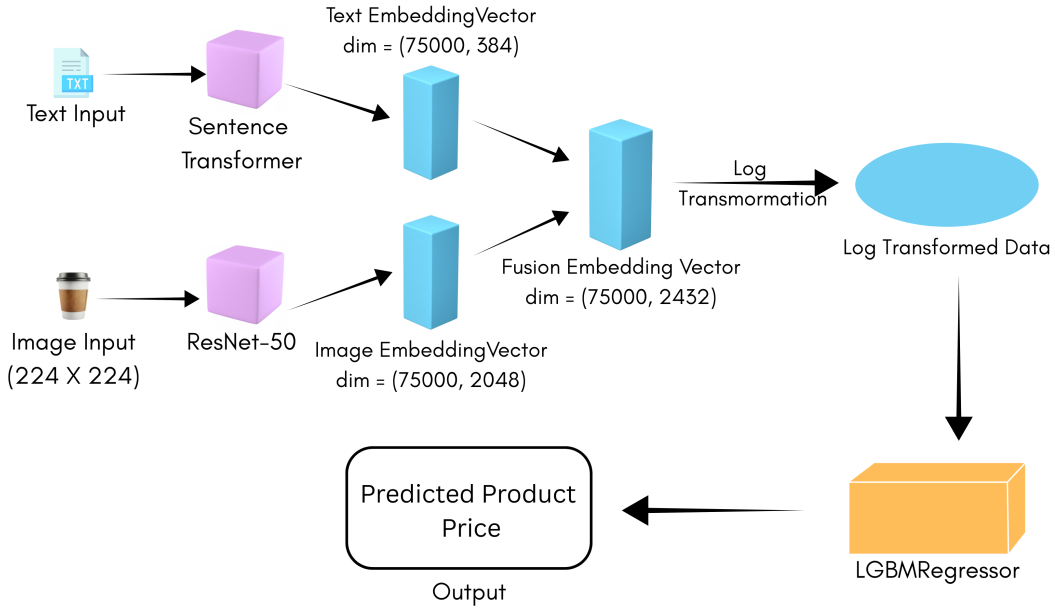
1. Text preprocessing and embedding generation using `SentenceTransformer`.
2. Image feature extraction using pre-trained `ResNet50`.
3. Fusion of multimodal embeddings.
4. Training `LightGBM` on combined features.
5. Model evaluation using $SMAPE$, MAE, $RMSE$, and R^2 metrics.

3 Model Architecture

3.1 Architecture Overview

The model comprises three major components:

- **Text Processing Pipeline:** Encodes product descriptions intosemantic vectors.
- **Image Processing Pipeline:** Extracts high-level visual embeddings.
- **Fusion + Regression Layer:** Concatenates both embeddings and passes them to `LightGBM` for final price prediction.



Model Architecture

3.2 Model Components

Text Processing Pipeline:

- Preprocessing steps: Lowercasing, missing value handling, cleaning.
- Model type: `SentenceTransformer(all-MiniLM-L6-v2)` (384-dimension embeddings).
- Key parameters: Dense vector per sentence capturing context and brand-specific meaning.

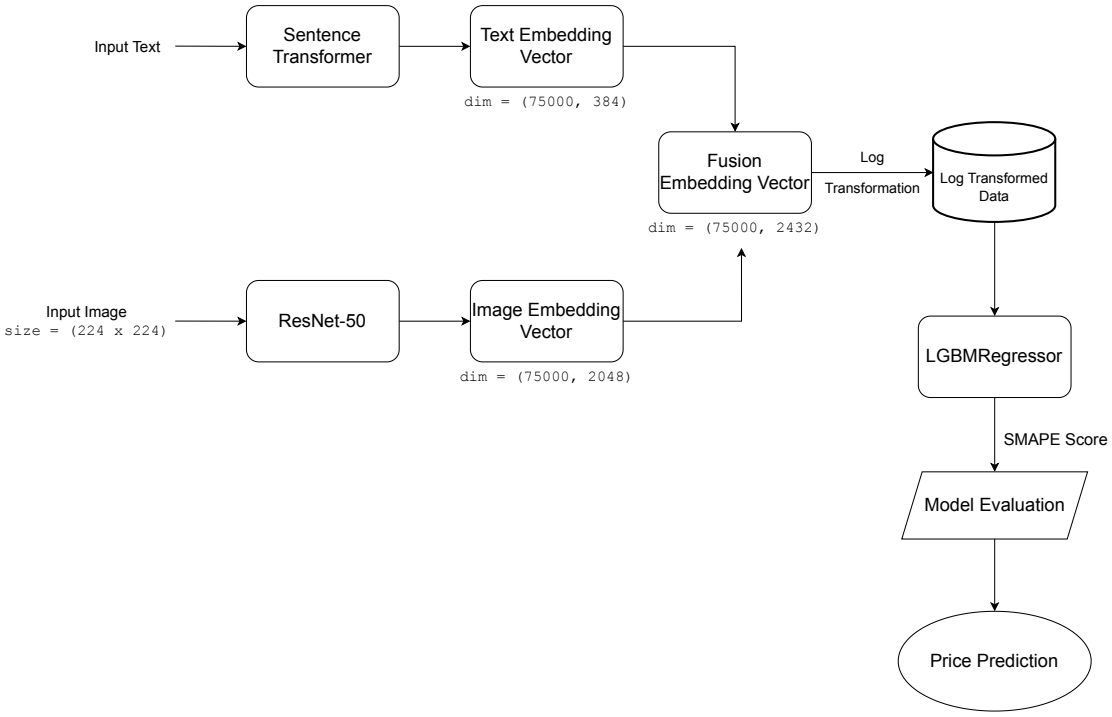
Image Processing Pipeline:

- Preprocessing steps: Resize image of size (224×224), normalization via `preprocess_input()`.

- Model type: Pre-trained `ResNet50` (without top layers).
- Key parameters: 2048-dimension feature vector from global average pooling.

Fusion + Regression:

- Features concatenated using `scipy.sparse.hstack()`.
- Model: `LightGBMRegressor` (n_estimators=1000, learning_rate=0.05).
- Objective: Regression task on price variable.



Model Architercture

4 Model Performance

4.1 Validation Results

- **SMAPE Score:** 22.85%
- **Other Metrics:** Calculated R^2, MAE, $RMSE$.
  - R^2: 0.58
  - MAE: 0.58
  - $RMSE$: 0.58

5 Conclusion

The proposed Smart Product Pricing System is an excellent blend between language and vision intelligence to make the right predictions regarding prices. The model is expected to have excellent

generalization and interpretability with the combination of `SentenceTransformer`-based textual understanding, `ResNet50`-based visual representation, and `LightGBM` regression.

This is a scalable and adaptable hybrid methodology that offers a strong structure of intelligent pricing automation across various e-commerce platforms.

6 Appendix

6.1 Code Artefacts

- Kaggle Dataset Download Link: Kaggle Link
- Uploaded Code Google Drive Link: Drive Link