Amazon ML Challenge 2025 – Technical Write-up

1. Introduction & Problem Statement
The objective of the Amazon ML Challenge was to develop a machine learning model capable of predicting the price of retail items based on multimodal inputs. The provided dataset consisted of approximately 75,000 samples, each containing a sample ID, a textual catalog description (including product name, size, unit, and packaging details), an image URL, and the target variable: price. The goal was to build a robust regression model that can accurately infer the product price using both textual and visual information.

2. Data Understanding & Preprocessing
The dataset included two main modalities: textual catalog content and product images. The catalog text contained rich descriptive attributes such as brand, unit size (e.g., “12 oz”, “pack of 6”), and product type. Image URLs were provided for most samples, although some were invalid or missing.

Text data was cleaned minimally since the pretrained transformer models handle tokenization and semantic normalization effectively. The images were pre-encoded into embeddings using a pretrained CLIP vision encoder, and missing images were masked using a corresponding *valid_mask* array. Numerical columns such as “value” and “unit” were implicitly encoded through the text embeddings. The dataset was split into training and validation subsets using an 80–20 ratio, ensuring no data leakage across splits.

3. Machine Learning Approach
A multimodal deep learning framework was employed, built upon the *openai/clip-vit-large-patch14* model from Hugging Face. The approach leveraged CLIP’s dual encoder architecture to project both image and text data into a shared embedding space. The extracted embeddings were then processed by a lightweight regression head trained to predict continuous price values.

The model architecture consisted of: Pretrained CLIP text and image encoders (frozen during initial epochs) Custom projection layers mapping embeddings to a unified latent space A fully connected regression head producing scalar price predictions The optimization was performed using the AdamW optimizer with weight decay (1e–5), and the learning rate was scheduled using OneCycleLR for stable convergence. The loss function used was Mean Squared Error (MSE), appropriate for continuous target prediction.

4. Experiments & Results
Several experiments were conducted to evaluate the effectiveness of unimodal and multimodal setups: **Text-only baseline:** Using CLIP text encoder embeddings with a simple regression head. Provided a solid baseline with reasonable generalization. **Image-only baseline:** Using CLIP image embeddings alone. Performance was comparatively weaker due to limited visual information about size and quantity. **Multimodal fusion model:** Combining both text and image embeddings, concatenated before the regression layer. This configuration achieved the best validation performance

and smoother convergence. The final model was trained for 10 epochs with a batch size of 64 on GPU, reaching stable loss values and reliable price predictions. Predictions for the test set were submitted in the required format with columns: *sample_id* and *price*.

5. Conclusion

The proposed CLIP-based multimodal regression model effectively learned the relationship between catalog descriptions, product imagery, and their corresponding prices. Incorporating both modalities led to noticeable improvement over single-modality baselines, validating the strength of pretrained contrastive models for cross-domain regression tasks. Future improvements could include fine-tuning the CLIP encoders jointly or incorporating metadata-based numerical features. Overall, the approach demonstrated that pretrained multimodal models can generalize effectively to structured e-commerce prediction tasks.