**Shopper Categorization for RetailerData Insights**

**Project Overview**

The aim of the RetailerData Insights Shopper Categorization project is to classify shopper behavioral data into meaningful categories, specifically leveraging product names and search terms from Amazon, Target, and Walmart applications. The project’s goal is to deliver actionable insights by building a robust categorization model complemented with descriptive analytics, focusing initially on the fashion product category.

Note : Because of confidentiality requirements, I anonymized the project name and dashboard data

**Approach and Planning:**

We organized our project into three parallel workstreams: Descriptive Analysis, Fashion Data Filtering using R, and Model Building. These tasks were designed to progress simultaneously to maximize efficiency.

**A screenshot of a diagram

Description automatically generated**

1. **Descriptive Analysis**

* We began with exploratory data analysis (EDA) to understand the structure and shape of the dataset.
* Key steps included identifying missing values, inconsistent entries, and any messy data that required cleaning.
* We performed essential calculations needed for analysis such as conversion rates, funnel analysis, and other engagement metrics.
* A general-purpose dashboard was developed using the entire dataset, without category-specific filtering initially.
* Once the first level of categorization (fashion vs. non-fashion) was ready, we applied that filter to the dashboard, effectively converting the general dashboard into a fashion dashboard for more targeted insights.

1. **Fashion Data Filtering using R**

* The filtering of fashion products from the whole dataset was done using R code. This was performed by our team member Eva Halsne.
* The primary focus was to identify as many fashion-related products as possible, creating a labeled dataset used for model training.
* The labeled dataset, combined with the dataset provided by RetailerData, served as the foundation for the first phase of model building. The dataset generated from our labeling process was used for training, testing, and validation. After final inference, the original dataset provided by RetailerData, which appears to have been pre-verified with accurate labels, was used for cross-verification of the model’s performance.

1. **Model Building**

* For modeling, we used a pre-trained model based on BERT (Google’s original BERT base model, loaded via Hugging Face library) to classify products as fashion or non-fashion.
* The modeling approach is planned to be performed in phases:
  + We built a separate model for the first level of taxonomy classification rather than attempting to classify the entire taxonomy in a single model.
  + This staged approach will allow for better traceability of model errors and easier performance tuning.
* After successfully building and fine-tuning the initial fashion vs. non-fashion classification model, we planned the following future steps:
  + We will explore clustering techniques to identify natural groupings within the fashion category.
  + If clustering does not produce meaningful separations, we will continue using supervised modeling (e.g., further fine-tuning BERT models) to classify into deeper categories and simultaneously explore other techniques such as semi-supervised learning, active learning, or weak supervision approaches.
* As more categories are identified, these categorizations will be used to further slice the data within the dashboard and enhance analytical capabilities.
* Initially, we aim to build only three levels of taxonomy classification to assess the value for insights generation. Deeper hierarchies may fragment the analysis excessively and reduce interpretability.

**Additional Considerations**

* The process for classifying search terms will be treated separately.
  + Search terms often involve challenges like fuzzy matching, spelling errors, and incomplete entries, requiring a specialized cleaning process.
  + Due to missing values and different characteristics compared to product names, we determined that a different model or method might be necessary for search terms.

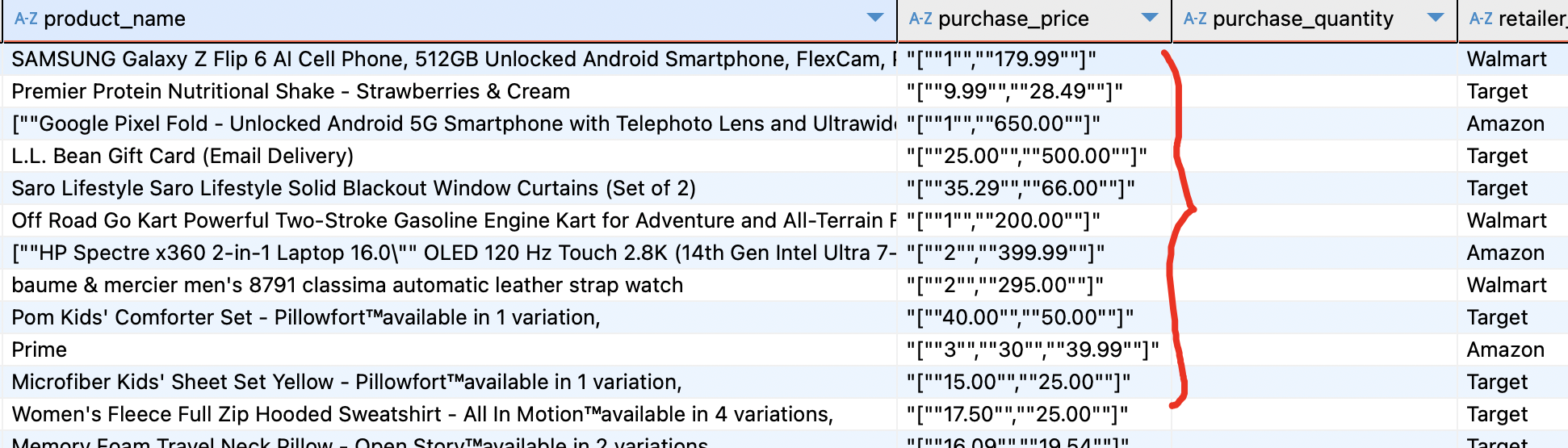
Each of the following sections provides more detailed explanations of the points outlined in the approach above.

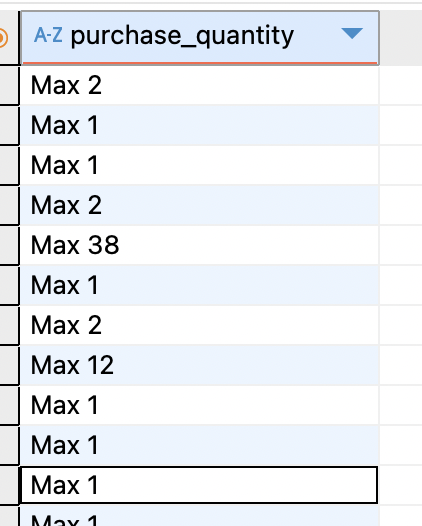
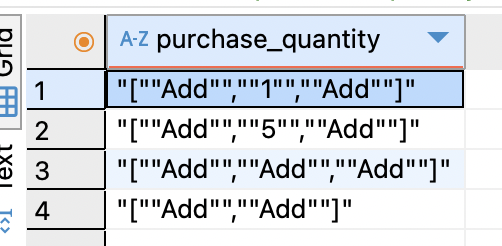
* + 1. **Descriptive Analysis**

**Exploratory data analysis**

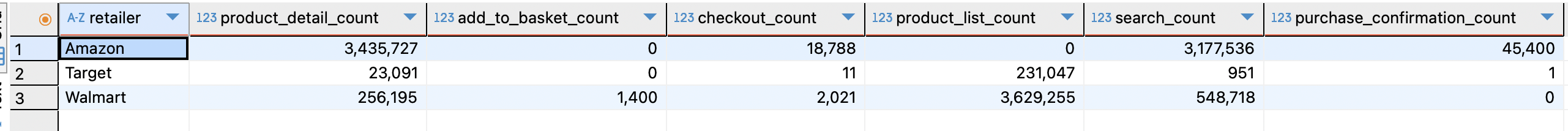
Initial data exploration revealed several findings as follows

* A significant number of purchase confirmation events were found to be missing product names. Since we could not determine the category for these purchase confirmations, they could not be included in our calculations.
* Due to the high number of missing product names in purchase confirmations, we decided not to use purchase confirmation as our primary metric for analysis. Instead, we used the next best available metric — checkouts — which serve as a strong indicator of user interest.
* There were several entries in the purchase price field that contained multiple values. These rows lacked a consistent schema, with some entries having two values and others having up to twelve. Since there was no standardized structure for interpreting these multi-valued fields, and they represented only about 0.5 percent of the overall dataset, we decided to ignore these rows for the current phase of analysis.

****

**** ****

* Due to missing values in the ‘add to basket’ events for two retailers and missing product names in all ‘Purchase Confirmation’ events, we decided to base our dashboards primarily on checkout events, supported by a summary of event counts.



All calculations were performed at the event level, and this approach was confirmed with RetailerData.

These findings can be useful for testing and verifying the data as the project progresses.

**Product Distribution Analysis**

To quantify potential insights before model development, a preliminary analysis was conducted using keyword-based categorization (e.g., keywords like “bag,” “dress,” etc.) to estimate the distribution of fashion-related products. This analysis established an initial baseline for subsequent modeling phases and sampling strategies.

**General Dashboard Development**

An initial dashboard was constructed, incorporating key metrics such as:

* Conversion rates based on checkout events (chosen for their stability and reliability)
* Top products based on engagement metrics
* Demographic analysis based on gender, income, and DMA (Designated Market Area)

This dashboard initially provided overall insights without categorical granularity, but it will be expanded as new product categories are identified through the modeling process.

* + 1. **Fashion Data Filtering using R**
* The initial data cleaning was performed, which involved removing duplicate entries, filtering out unnecessary columns, combining rows,cleaning up text fields (product names, product IDs, and search terms), removing punctuation, and converting text to lowercase for consistency.
* Following this, the data filtering process was carried out. The filtering involved identifying fashion-related products through multiple layers: matching brand names, department names (e.g., men, women, kids), and apparel-related keywords (e.g., tops, dresses, shoes). Additional logic was applied to remove false positives such as unrelated products containing misleading terms. After applying these filters, a random sampling check confirmed approximately 94-96 percent filtering accuracy across different trials. Finally, department labels were extracted from product names, and the filtered dataset was saved for further use in modeling and dashboard development
  + 1. **Categorization Model Development**

**Choice of Model and Initial Phase**

Given the availability of labeled data and recommendations from professors and the teaching assistant, we decided to start with BERT. However, we will ultimately use the technique that produces the best results and are open to adjusting our approach based on model performance.

**Why BERT?**

BERT was selected because of its proven effectiveness in handling complex textual data, its deep contextual understanding of language, and its strong performance when fine-tuned on labeled datasets for classification tasks.

**Modeling Strategy and Phases**

The categorization modeling was planned in multiple phases to systematically address complexities inherent in hierarchical taxonomy classification:

* **Phase 1: Binary Classification (Fashion vs. Non-Fashion)**
  + Implemented and fine-tuned a BERT-based binary classifier.
  + Achieved measurable accuracy, validating the effectiveness of the approach.

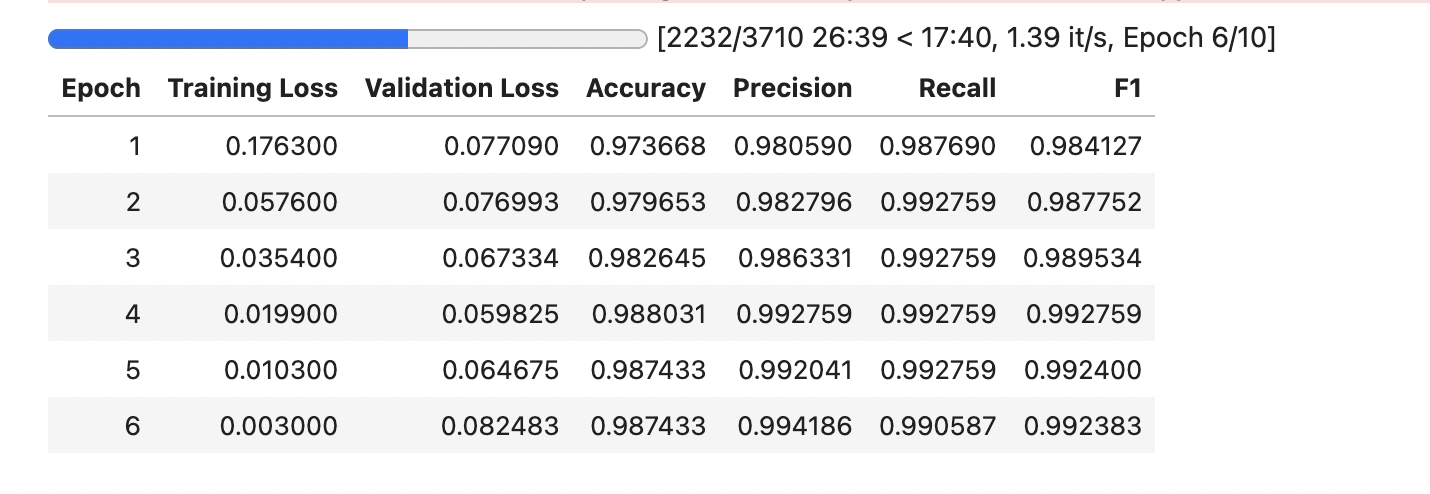
For each iteration, training, testing, and validation datasets are tracked to prevent any reuse. We currently use the R-filtered labeled data for modeling and cross-verify results with the dataset provided by RetailerData.

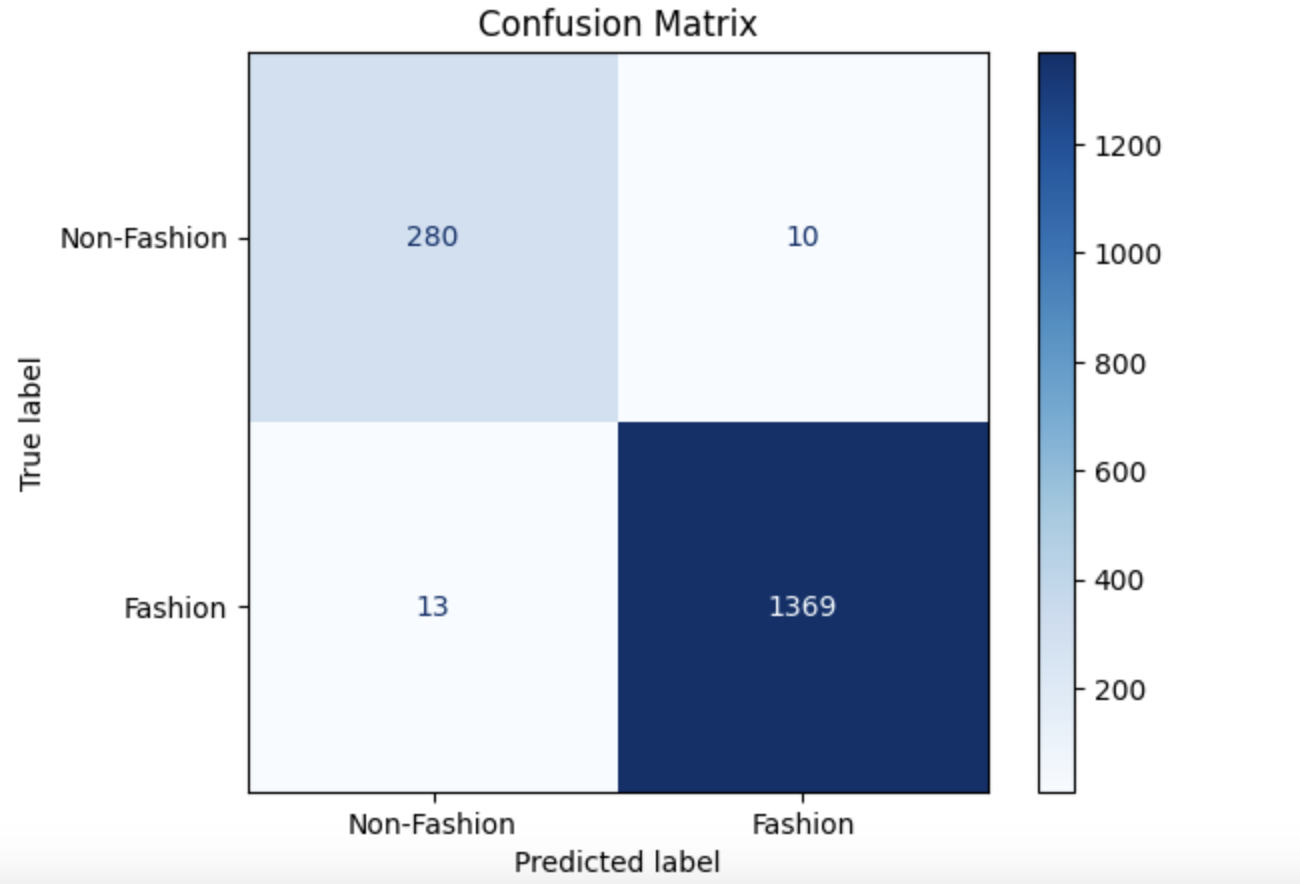
Early Model iteration results :

Results : Iteration 1

Sample unbalanced – 1671 row dataset – more fashion than non-fashion

Note : early\_stopping\_patience=2



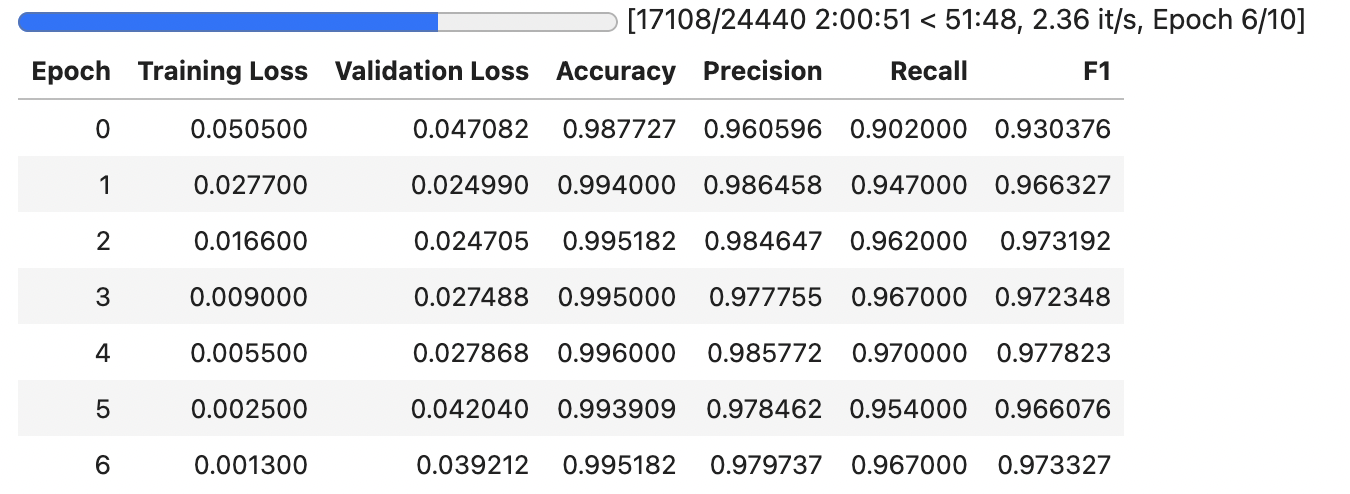


There were many misclassifications in this iteration, including several obvious non-fashion items.

Iteration 2 : early\_stopping\_patience=2 to avoid overfitting

Much larger dataset

10k fashion , 100k non fashion – sample ratio maintained similar to original







A blue squares with white text

Description automatically generated

Current prediction performance is not optimal, with approximately 20 out of 100 items being misclassified. However, further fine-tuning is planned to improve the model and evaluate its performance more accurately.

We have implemented an initial method for tracking and preventing data reuse across training, validation, and testing by writing data back to a separate database table. However, optimizing and standardizing this process for greater efficiency is a priority.

* **Phase 2 and Beyond: Detailed Taxonomy**
  + Planned to iteratively develop subsequent classifiers for more granular categories (Men’s, Women’s, Kids).
  + Each level uses the output of the previous model as input, allowing clear traceability and easier debugging.

**4. Model Refinement and Future Methods**

Post-validation of the initial BERT model, several analytical methods were identified for subsequent phases:

* Further Fine-tuning: We plan to refine the current BERT model using extended, higher-quality labeled datasets to improve classification accuracy.
* Clustering Techniques: We will apply clustering methods to identify natural groupings within product categories, particularly where labeled data is sparse or unavailable.
* Hybrid Approaches: Based on model performance and clustering results, we will explore combining supervised and unsupervised techniques to enhance categorization accuracy and robustness.

**5. Integration of Model Output into Analytical Framework**

Upon successful development of hierarchical categorization:

* The categorized product names will be integrated back into the main dataset.
* The general dashboard initially developed will be refined by adding categorical filters
* This integration will enable nuanced analyses such as:
  + Category-specific conversion rates.
  + Identification of shopping patterns.
  + Loyalty metrics and basket composition.
  + Path-to-purchase analytics for fashion products.

**Challenges**

* Session-based analysis and basket analysis are still pending. Identifying user sessions has proven to be challenging, as we have found only a handful of examples with a clear path to purchase. As a result, we are not yet confident in producing descriptive insights from session analysis.
* Basket analysis has not been attempted so far but will be considered in future phases as we refine the data and modeling approach.

**Final Approach**

We are using a phased approach, starting with fashion vs. non-fashion classification using a fine-tuned BERT model. Since the R-based filtering, which used hardcoded brand lists and regular expressions, achieved approximately 94 percent accuracy, we consider 94 percent a reasonable performance target for the model as well. For deeper levels of categorization, we will explore clustering, hybrid methods, and other machine learning techniques. Once satisfactory accuracy is achieved, we will join the categorized data back to the main dataset and expand descriptive analysis. Future work includes refining session-based and basket analysis and improving data handling processes.