Enhancing E-commerce Recommendations with Shopping Session

Team Member: Minwoo Sohn, Jiayu Shi

Exploratory Data Analysis (EDA)

This session presents a comprehensive Exploratory Data Analysis (EDA) of 'product_train' and 'session_train' datasets, focusing on key attributes like price, brand, locale, color, size, and session length. The goal is to identify patterns that will inform the development of a recommendation system.

<Basic Analysis>
The dataset comprises 1,551,057 entries across 11 attributes: ID, Locale, Title, Price, Brand, Color, Size, Model, Material, Author, and Description. Essential fields such as ID, Locale, Title, and Price exhibit no missing values, indicating complete product listings. Conversely, Color, Size, Model, Author, and Description show a significant number of missing values (19,371 to 789,922), likely reflecting non-applicability to certain products.

<Price Analysis>
Price analysis reveals most products priced around 70€, predominantly under 20€. For Japanese products, prices generally fall below 45€ after currency conversion, utilizing outlier boundaries defined by traditional IQR methods.

<Brand Analysis>
Top brands in descending order include Amazon Basics, LEGO, APPLE, Independently Published, Generic, Samsung, and BOSH, highlighting a diverse range of popular products.

<Locale Analysis>
The distribution of product listings by locale is as follows: DE (518,327), UK (500,180), JP (395,009), IT (50,461), FR (44,577), and ES (42,503). The UK locale was selected for initial recommendation system analysis due to the team's proficiency in English and number of samples.

<Color & Size>
Globally, the most popular colors are Black, White, Multicolor, Blue, and Gray, indicating a preference for standard color schemes.

"One Size" or "Free Size" dominates across all locales, suggesting a trend towards products that do not require precise sizing. Medium, Large, and Small sizes are consistently present, indicating traditional clothing demand. Single-unit sales prevail, hinting at a dataset leaning towards individual rather than bulk purchases.

<Session Train Dataset Analysis>

The dataset captures session data, including previous items (prev_items) and the next purchased item (next_item), along with the locale of the Amazon store. The UK, DE, and JP locales exhibit the highest session counts. Session length analysis reveals a concentration of sessions with fewer than four interactions, suggesting a focus on shorter sessions for computational efficiency.

Modeling

Our team will construct a graph-based model for recommending products on Amazon. The graph's nodes represent products, characterized by attributes such as id (unique Amazon Standard Item Number, ASIN), locale, title, price, brand, color, size, model, material, author, and a description of key features (desc). Attributes will be processed to balance numerical and textual data: prices will remain numerical, while textual attributes will be transformed into numerical vectors. We plan to merge all textual information per product into a comprehensive sentence and utilize BERT to obtain a 512-dimensional vector representation.

Edges in the graph symbolize the sequence of items in the session train dataset, indicating the transition from one product to another within a user's session. Directed edges will represent these transitions, with their frequency serving as the weight, enhancing the model's ability to learn from repeated patterns.

Focusing on UK locale products will streamline our dataset, reducing the nodes' size by approximately 80%. Further refinement includes eliminating products appearing less than three times and limiting the analysis to the five most recent items in a customer's purchase history to maintain relevance and manageability.

The model's architecture comprises an adjacency matrix (A) and a feature matrix (X), the latter incorporating both the price and the BERT-derived vectors for a comprehensive 513-dimensional representation. By multiplying A with X, we integrate neighboring node information, subsequently concatenating the original and integrated feature matrices to enrich our dataset (X'').

For prediction, we'll employ LSTM or RNN, leveraging their sequential data handling capabilities to forecast the next likely purchase. Optimization will focus on improving metrics such as Mean Reciprocal Rank (MRR) or using cross-entropy as the loss function.

This approach aims to create a nuanced and dynamically updating model capable of offering personalized recommendations based on a user's browsing and purchasing history, potentially increasing customer satisfaction and engagement on the Amazon platform.

Weekly timeline

Week 2 (April 1 - April 7): Establishing a Baseline

- Baseline Model Development: Implement a baseline model to establish initial performance benchmarks for the project. This model will serve as a comparison point for future improvements.
- LSTM model with price
- Initial Evaluation: Assess the baseline model's performance using predefined metrics, highlighting areas for immediate enhancement.

Week 3 (April 8 - April 12): Optimization and Exploration

- Parameter Tuning: Refine the baseline model through extensive parameter optimization to improve accuracy and efficiency.
- Exploratory Modelling: Investigate various model architectures and methodologies that could lead to better performance, including transfer learning and deep learning techniques.

Week 4 (April 15 - April 19): Finalization and Preparation

- Model Finalization: Conclude the experimental phase by finalizing the model with the best performance.
- Presentation Preparation: Start preparing for the final presentation. This includes creating slides, visualizations, and any other necessary materials to effectively communicate the project's findings and methodology.

Week 5 (April 22 - April 25): Conclusion and Cleanup

- Code Cleanup: Clean the project code to ensure readability, efficiency, and reproducibility.
- Project Finalization: Wrap up the project by finalizing all documentation, including a comprehensive report and a detailed explanation of the methodology, findings, and conclusions.
- Presentation and Submission: Deliver the final presentation, showcasing the project's achievements, insights, and potential impact on e-commerce personalization.