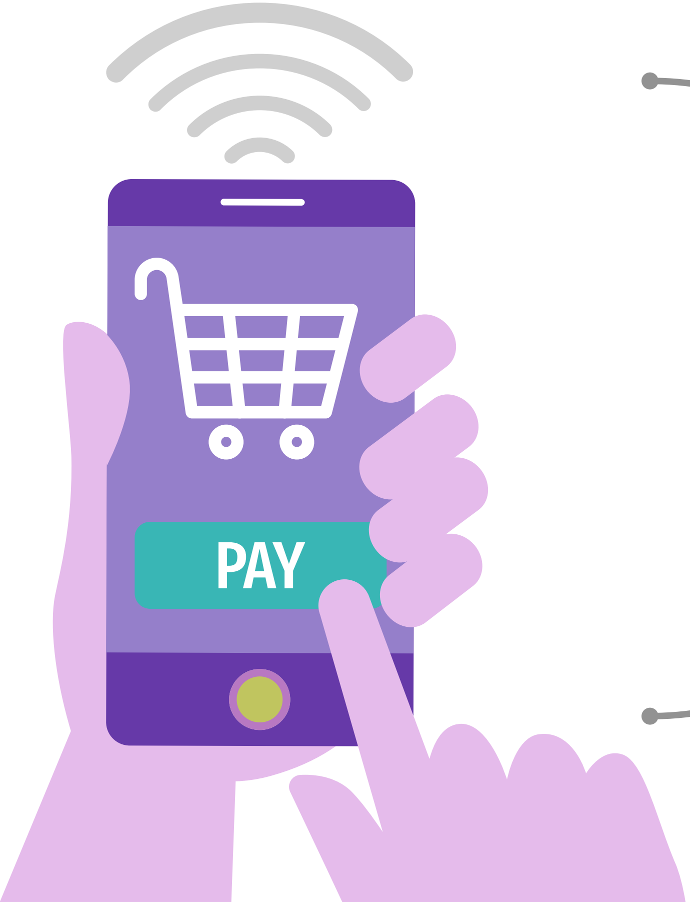
Amazon Next Product Recommendation

Minwoo & Jiayu

Table of Contents



Introduction

- **High-level**: Our project focuses on enhancing e-commerce personalization through advanced session-based product recommendations. This approach seeks to predict what products customers will engage with next, tailored to their past session activities
- **Significance**: In the competitive e-commerce landscape, accurately predicting customer behavior in real-time sessions is crucial. It allows platforms to offer highly relevant product suggestions, improving user experience and potentially increasing sales.

Problem

01 Number of Sessions and Products

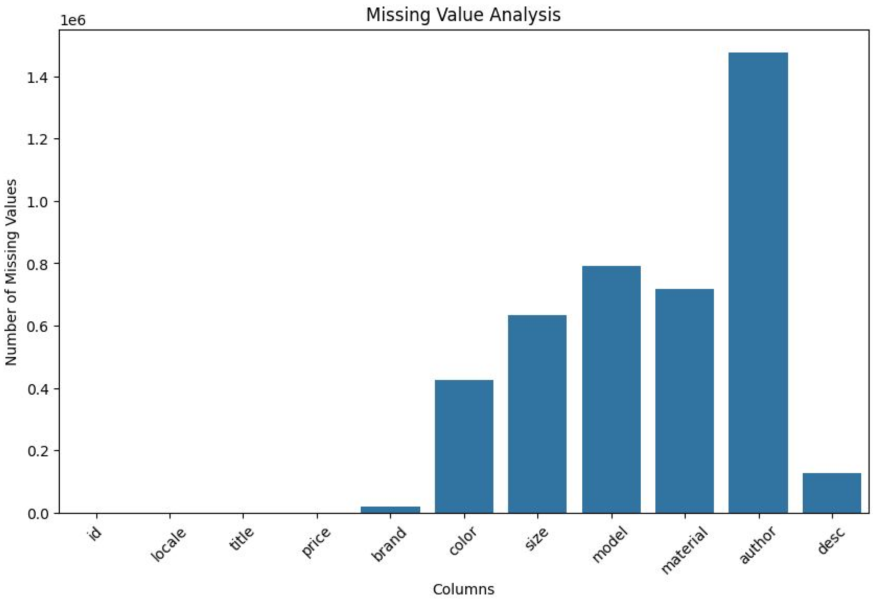
02 Data Complexity

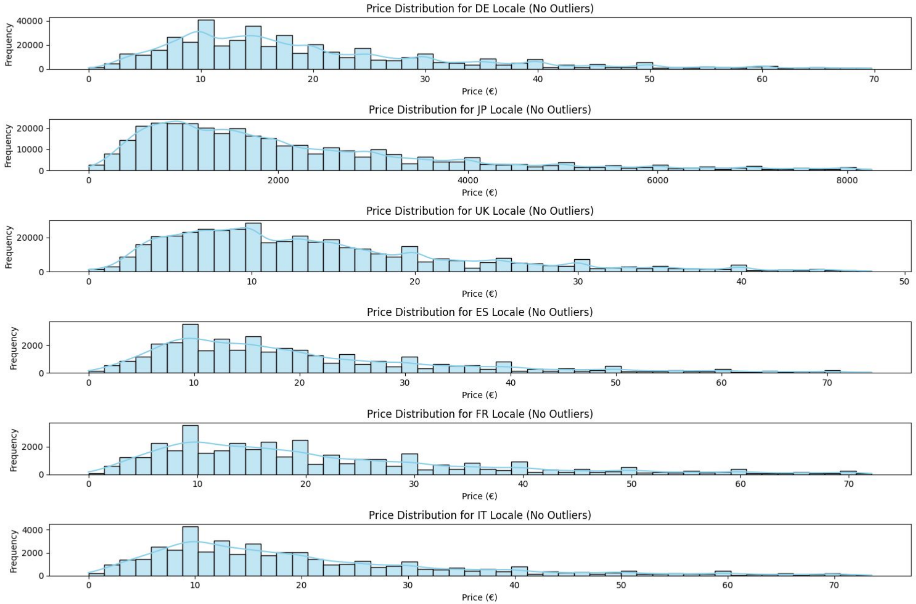
03 Context-aware Personalization and Recommendation

04 Different Languages

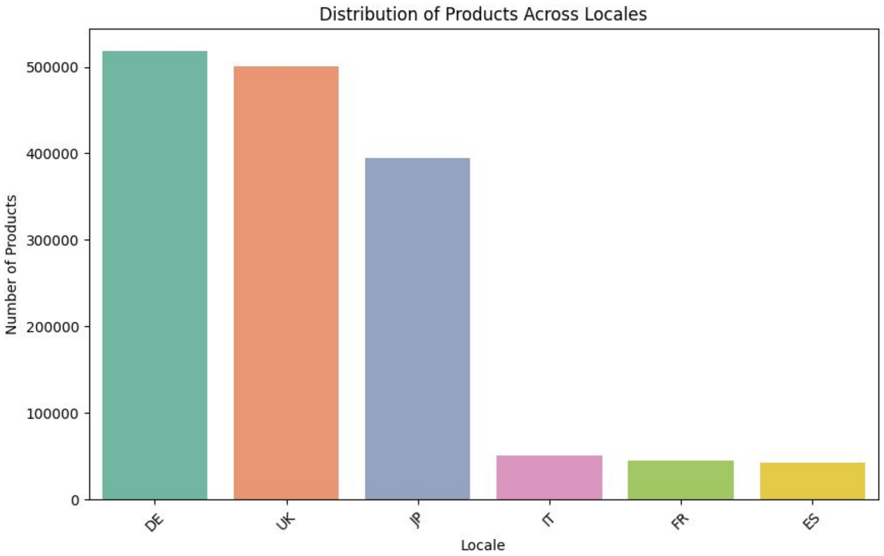
SHOP

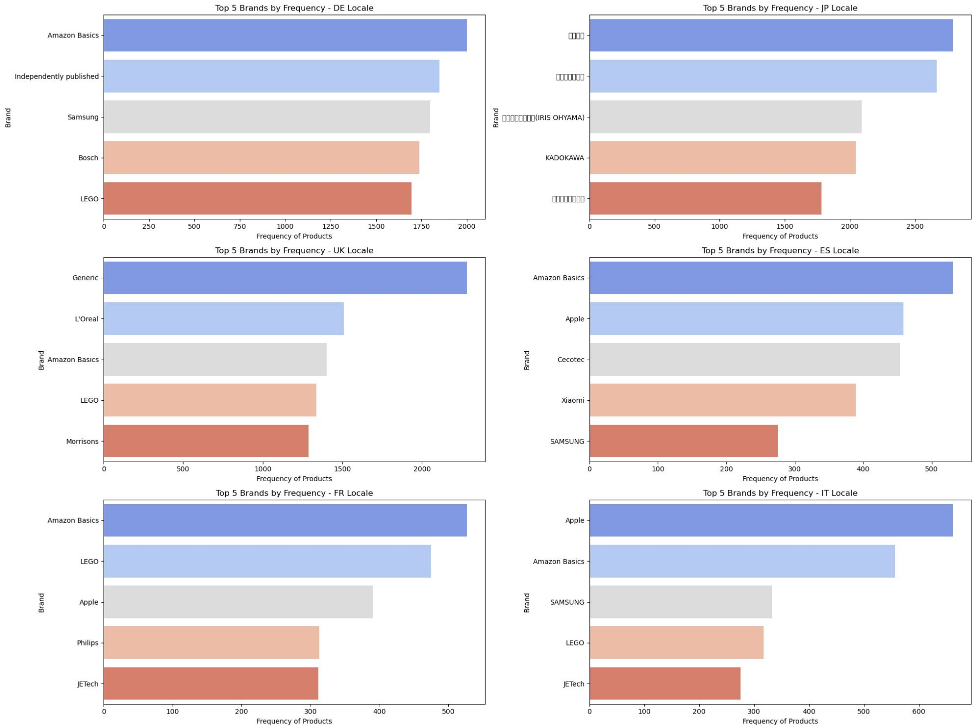
Exploratory Data Analysis



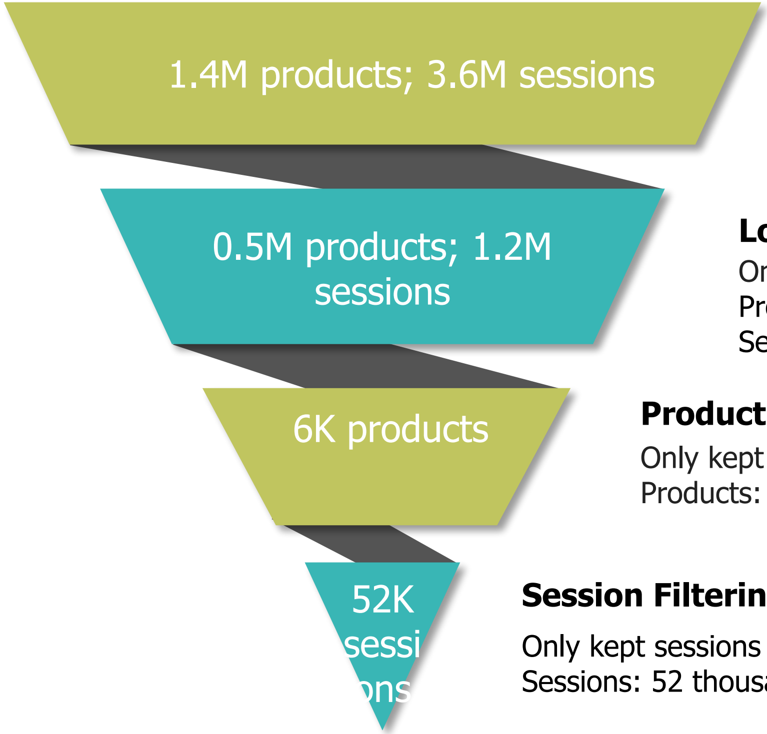


Exploratory Data Analysis





Challenge



Complete Dataset

Products: 1.4 million products
Sessions: 3.6 million sessions

Locale Filtering

Only kept products in UK
Products: 0.5 million products
Sessions: 1.2 million sessions

Product Filtering

Only kept products appear > 100 times in UK sessions
Products: 6 thousand

Session Filtering

Only kept sessions having >=2 previous items
Sessions: 52 thousand

Feature Engineering

Original Columns

price, title, brand, color, size, material, author, desc

Price

Keep price as is

Representation

Combined all other columns into one paragraph, then pass to BERT

Final Features *X*

1. Price
2. Representation

Modeling (Baseline)

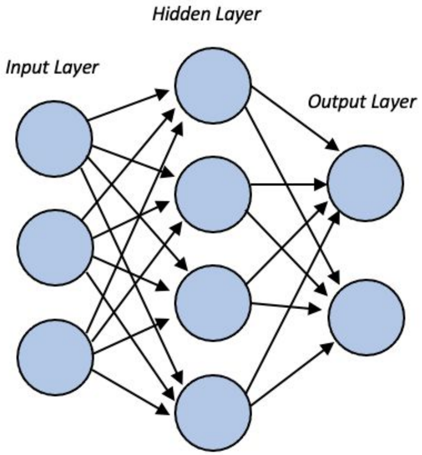
- Accuracy Metric: Checks if the exact next product purchased appears in the top 100 recommended products.

- Baseline Model: Randomly selects 100 products from a total of 6K available products.
- Performance: The accuracy stands at 1.49%.
- Explanation: This model is the most intuitive and naive approach, serving as a basic benchmark for comparison.

Modeling (NN)

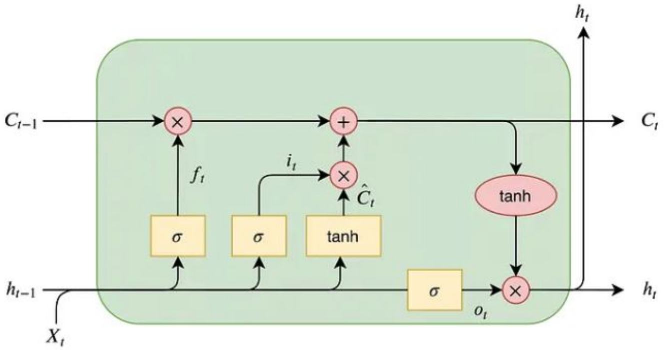
- Split Sessions: Divide the sessions into an 80:20 ratio for training and testing datasets.
- Adjacency Matrix: Construct the adjacency matrix ***A*** using only the links from the training dataset.
- Feature Matrix: Multiply the feature matrix ***X*** with ***A*** to derive the aggregated feature matrix ***X'***.
- Final Features: Concatenate ***X*** and ***X'*** to form the final features ***[X X']***, which are inputted into a 5-layer neural network for predicting target product features.
- Product Matching: Utilize cosine similarity to identify the top 100 products with the highest similarity scores.
- **Accuracy Metrics:**

- Using only price as ***X***: Achieves an accuracy of 2.5%.
- Using both price and additional representations as ***X***:
  Results in an accuracy of 2.26%.



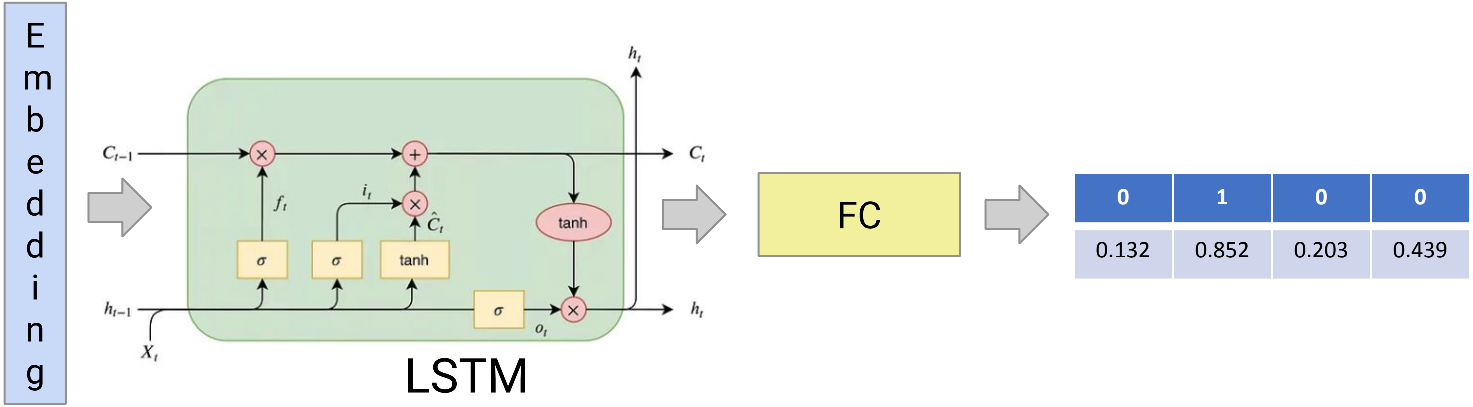
Modeling (LSTM)

- An LSTM (Long Short-Term Memory) model is a type of recurrent neural network (RNN) that is well-suited to making predictions based on time series data or sequence data
- **Memory Cells**: LSTMs have a special architecture that includes memory cells that can maintain information in memory for long periods. Each cell has gates that manage the flow of information into and out of the cell, which helps it to remember or forget information dynamically.
- **Handling Long Dependencies**: One of the primary advantages of LSTMs is their ability to connect previous information to the present task, which is particularly useful in tasks where context from far back in the sequence is crucial to understand the current input.



Modeling (LSTM)

- Tried with Session Sequence only vs Session Sequence + Price
- Accuracy: 0.0095% (Sequence only)
- Accuracy: 60.9% (Sequence + Price)



Conclusion and Future work

Methods to improve Model Performance:

- Adjacency Matrix Improvement: Retain all links within each training session in the adjacency matrix ***A***, rather than limiting to links between the last two products or between the last product and the next.
- LSTM Model Enhancement: Incorporate all available features in the LSTM model to capture a richer set of data inputs.
- Expanding Product Inclusion: If computational resources permit, include all products in the graph, not just those with a frequency above 100.

Thank You!

TOP SALE E-COMMERCE ITEMS

Q & A