CS37300 Kaggle Competition Report
Ethan Fuller|0034368365|Los Pollos Hermanos (Gus Fring account)

1. **Task Specification**
The model is designed to predict the top 12 products a customer is likely to purchase within a week using transactional data from the past year.

Input Space:
- Historical transaction data containing customer IDs, product IDs, and purchase timestamps
- Training data containing lists of products purchased in a day by a customer spanning a year long time frame for a small subset of the overall customer population.
- Validation data containing the same customers over the span of one week after the training data.
- Different validation and training sets were used for training, with the combined predictions utilized for testing.

Output Space:
- For each customer in the prediction set, a list of 12 product IDs representing the most likely products they will purchase

Performance Metrics:
- Mean Average Precision (MAP@12)
- Additional validation metrics: Accuracy, Precision, Recall were used to develop improvements

2. **Data Representation**
Raw Data Format:
- Tabular data in CSV format
- Contains columns: customer_id, product_id, time_date
- Transactions in the training and validation sets have products in list_product_id format

Preprocessing Steps:
- Transaction expansion: Converts list_product_id entries into individual customer-product interactions, using the labels from the transactions dataset
- Time conversion: Converted time_date strings into datetime format
- ID mapping: Created numerical mappings for customer and product IDs
- Data filtering: Filtered newly created test and k-fold sections such that the training data matched the validation data customer wise.

3. **Knowledge Representation**
Model Description:
The model is a hybrid recommender system that combines collaborative filtering (cosine similarity scoring) with popularity-based recommendations and performs k-fold cross validation with random sampling to optimize hyperparameters.

Model Space:
- Customer-Product Interaction Matrix (sparse matrix)
- Product Popularity Scores (based on last 30 days)
- For each product, a score is generated based on this equation
  - y = C*w1 + P*w2
- Parameters:
  - C: Cosine similarity score calculated using customer product interaction matrix.
  - P: Popularity score taken via ranking every product based on its purchase frequency over 30 days from the prediction week.
 - Hyperparameters:
  - w1: Weight for Cosine Similarity scoring
  - w2: Weight for popularity score
- If there is not enough purchase data for a customer, fill the rest of a customer's purchase history with the top 12 most popular products.

4. **Learning Method**
a. Score Function:
- Primary: Mean Average Precision (MAP@12)
- The score function evaluates how well the model ranks the actual products a customer will purchase within the top 12 recommendations

b. Search:
- Optimization Algorithm: Two-stage approach
- Random sampling of weight combinations
- Hyperparameter Selection:
  - Number of K-Fold splits: 5
  - Number of random samples: 50
  - Weight constraints: Must sum to 1
  - Maximum iterations: 20

5. **Evaluation**
The model's evaluation strategy is comprehensive:
- Uses K-Fold cross-validation (5 folds) during weight optimization
- Evaluates on a separate validation dataset
- Calculates multiple metrics:
  - MAP@12 (primary metric)
  - Overall accuracy
  - Precision
  - Recall
- Performs final evaluation for the testing set (purchases after the validation set week) after training on combined training and validation data
- Uses progress tracking during evaluation to monitor performance for debugging