Workshop No. 1 — Kaggle Systems Engineering Analysis

Melisa Maldonado Melenge – 20231020110
Jean Pierre Mora Cepeda – 20231020105
Juan Diego Martínez Beltrán – 20231020131
Luis Felipe Suárez Sánchez – 20231020033

Universidad Distrital Francisco José de Caldas
Computer Engineering Program — School of Engineering
Systems Analysis & Design
Eng. Carlos Andrés Sierra, M.Sc.
2025

1 Competition Overview

Goal

The competition aims to predict the **product category** that a Best Buy customer will ultimately choose based on their online search queries and browsing behavior. The objective is to build models that can accurately map user interactions to the most relevant product categories.

Dataset Structure

- **Search Logs:** Records of customer search queries, capturing the text typed by users.
- **Session Data:** Clickstream information over two years, including browsing events and interaction sequences.
- **Product Metadata:** Descriptions, categories, and identifiers of Best Buy products linked to customer activity.
- **Target Labels:** The final product category chosen by the customer at the end of each session.

Significant Constraints

- **Big Data Scale:** The dataset is approximately 7 GB, requiring efficient storage, preprocessing, and modeling techniques.
- **Noisy/Sparse Queries:** Many queries may be short, ambiguous, or contain little information, complicating accurate predictions.
- **Sequential Dependencies:** User behavior unfolds over sessions, making temporal order critical for capturing intent.
- **Category Imbalance:** Some product categories are much more frequent than others, which can bias models if not addressed.

2 System Objective

To predict which BestBuy product will most interest a mobile visitor based on their search query or behavior recorded over a two-year period.

3 Systemic Analysis

Inputs

The system is fed by data generated by user interaction with the mobile platform, including:

- Search queries (text queries).
- Browsing events (clickstream): product views, clicks, time on page, scrolling.
- Session metadata: device, timestamp, approximate location, and referral source.
- Purchase and conversion history.
- Product catalog: attributes such as ID, category, price, and specifications.
- Target variable: product of interest (click or purchase).

Processing

This is the core of the system, where a machine learning model transforms input data into predictions. Its key phases are:

1. **Ingestion and Preprocessing:** Reading and partitioning the dataset (7 GB), cleaning, deduplication, adjusting time zones, and handling missing values.
2. **Enrichment and Feature Engineering:** Tokenization and embeddings of queries, extraction of predictive variables, mapping between queries and products/categories, and handling class imbalance.
3. **Predictive Modeling:** Training classification and ranking algorithms (base models, text models, embeddings with nearest neighbors).
4. **Evaluation:** Use of ranking metrics: top-k accuracy, MAP, NDCG, recall@k.
5. **Deployment and Monitoring:** Implementation of prediction services and continuous monitoring.

Outputs

- Main prediction: top-N list of products sorted according to the user's probability of interest.
- UI signals: featured products and personalized suggestions.
- Prediction records: stored for feedback and continuous improvement.

Actors/Stakeholders

- **Data Science/ML Team:** Designs and builds the predictive model.
- **Product/UX Team:** Defines how recommendations will be integrated into the UI.
- **Engineering Team:** Develops and maintains data pipelines and infrastructure.
- **Business/Marketing Team:** Defines KPIs and uses the system for campaigns.
- **End Users:** Provide behavioral data and receive recommendations.

Relationships and Interactions

The relationship begins linearly: the user generates input data through browsing, which feeds into the model's processing to produce an output. In practice, this output is shown back to the user, influencing their future behavior and creating a feedback loop.

4 Complexity and Sensitivity

Constraints

- Computational resources (RAM/CPU/GPU).
- Optimization to the specific competition metric.

Conflicts and Points of Variability

- Generalization vs. overfitting.
- Cold start problem for new users/products.
- Temporal variability (seasonality, trends).
- Class imbalance and query noise.
- Multi-device sessions fragmenting history.
- Feedback loops reinforcing biases.
- Concept drift requiring retraining.

Chaos and Randomness

- Sensitivity to initial conditions (e.g., a single click or typo altering recommendations).
- Feedback loops forming filter bubbles.
- Random user intent (gifts, curiosity, price comparison).
- Nonlinear interactions (e.g., product position affecting CTR).

5 Visual Representation

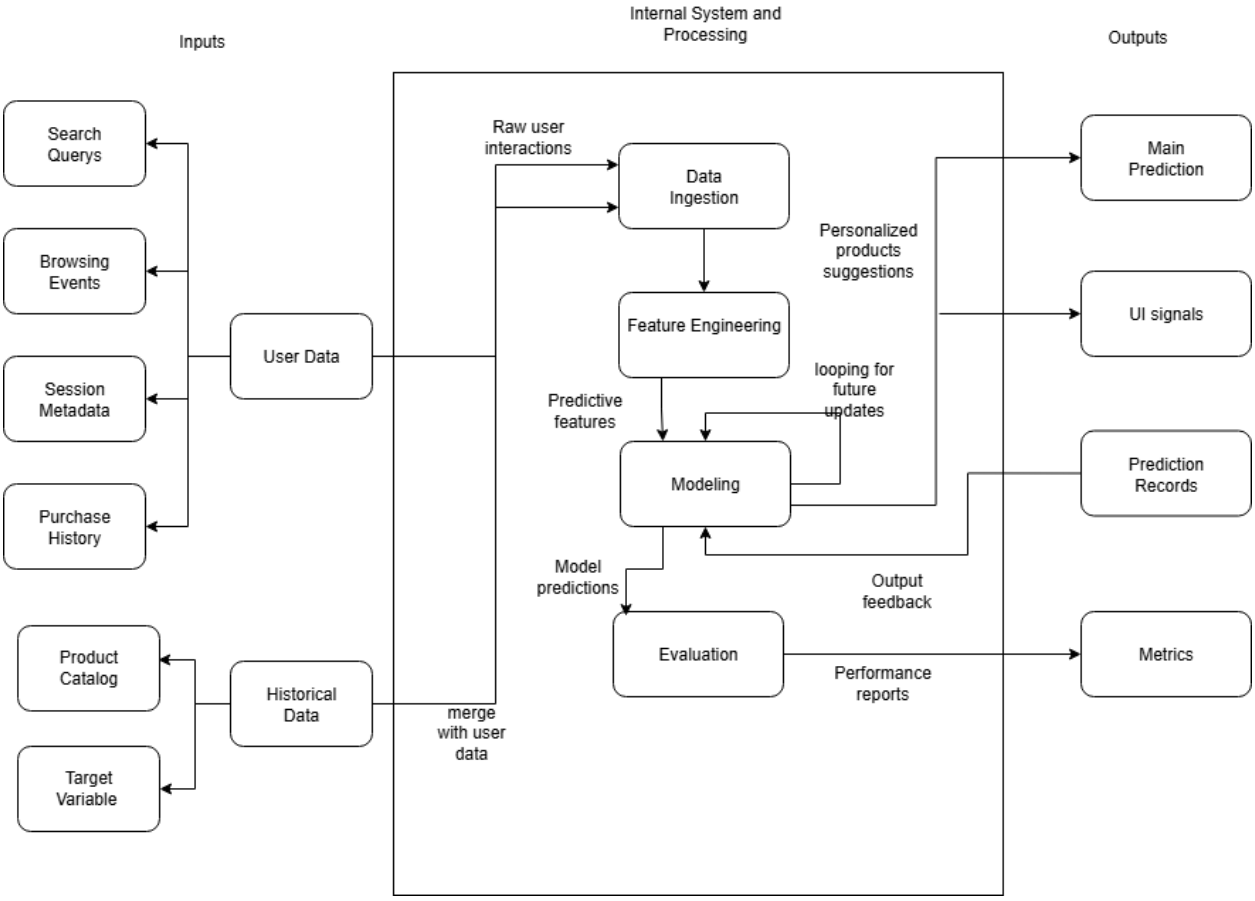

Figure 1: Diagram of the system analysis.

6 Conclusion

The analysis reveals a system with high potential but also inherent weaknesses.

Strengths

- Robust dataset: 2-year history and 7 GB of data.
- High business impact: potential to increase conversion and satisfaction.

Weaknesses

- Static dataset with no adaptation to new trends.
- Sensitivity to data quality (garbage in, garbage out).
- Risk of bias from historical promotions.

Key Findings

- System is multidimensional: NLP + clickstream + metadata.
- Query and session variables are most predictive; history aids personalization.
- Risks include drift and bias toward popular products.

Potential Strengths

- Large dataset allows seasonality and behavior capture.

Weaknesses / Points to Mitigate

- Imbalance and noise in labels.
- Cold-start problems.
- High infrastructure requirements.