Building a Recommendation Engine for Learning Engagement

A Data-Driven Approach to Personalized Recommendations

Key points

- **Project Goal:** Develop a recommendation system that delivers engaging, relevant learning content tailored to each person's needs.
- **Core Datasets:**
  - **Users** – person profiles and learning preferences
  - **Content** – structured learning materials with metadata (topics, difficulty, type)
  - **Engagements** – interaction history (views, likes, completions) capturing learning behavior
- **Hybrid Approach:**
  - **Content-Based Filtering** to leverage metadata for matching content with student profiles and styles
  - **Collaborative Filtering** to identify patterns from similar learners' interaction behaviors
  - Combined to maximize personalization, coverage, and recommendation accuracy

Users Schema Design

- **Fields:**
  - **user_id** – Unique identifier for each user; primary key for linking with other datasets.
  - **title** – Role or designation of the user (e.g., Student, Instructor, Intern).
  - **department** – Academic or organizational unit the user belongs to (e.g., Computer Science, HR).
  - **seniority_level** – Experience or academic stage (e.g., Beginner, Intermediate, Advanced).
  - **learning_style** – Preferred mode of learning (e.g., Visual, Auditory, Kinesthetic)
- **Explanation:**
  - The schema goes beyond basic demographics by embedding **role, organizational context, and learning preferences**.
  - **user_id** serves as the linking key across engagements and content interactions.
  - This structure enables **fine-grained segmentation**, ensuring recommendations are not only content-relevant but also aligned with the learner's style and seniority.
  - By combining these attributes with interaction data, the hybrid model can deliver **personalized and context-aware learning recommendations**.

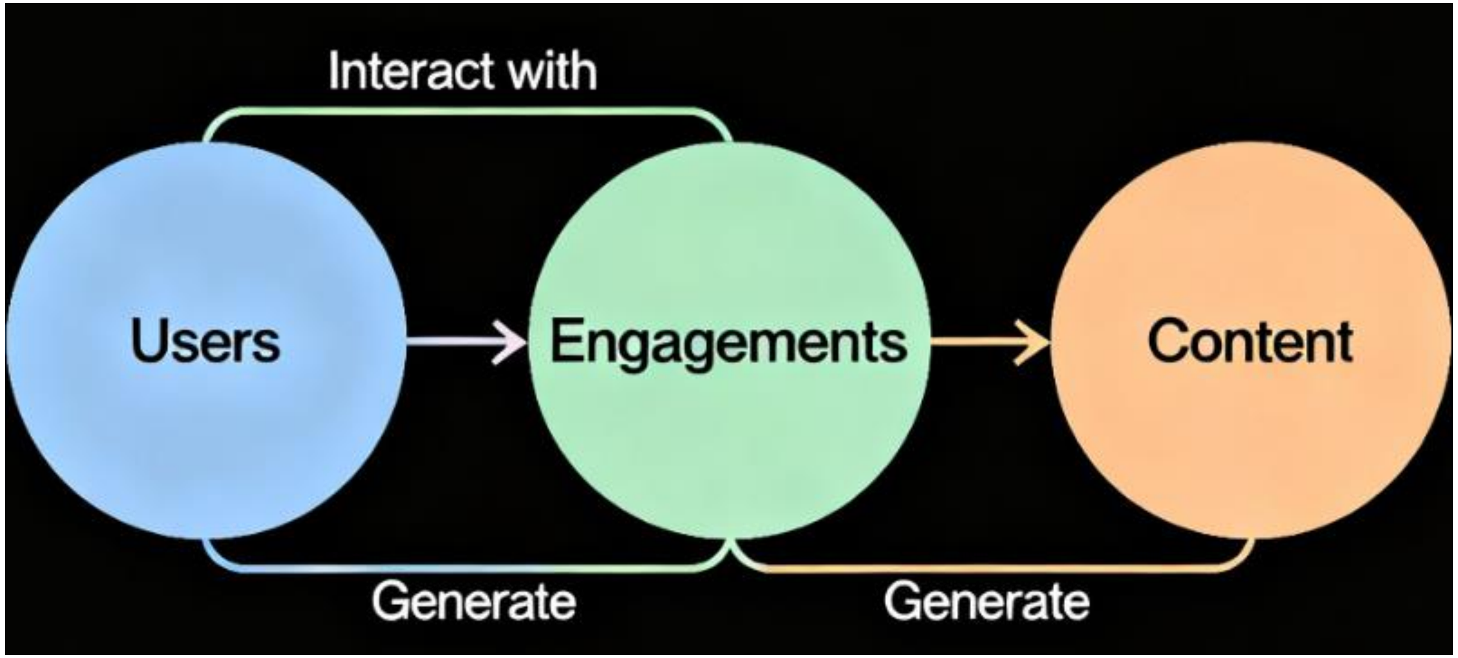
Content Schema Design

- **Fields:**
  - **content_id** – Unique identifier for each learning material; primary key for tracking.
  - **title** – Name or short description of the content (e.g., Introduction *to Statistics*).
  - **domain** – Broad subject area or discipline (e.g., Data Science, Finance, HR).
  - **subtopic** – Specific focus within the domain (e.g., Regression, Investments).
  - **difficulty_level** – Categorization of complexity (e.g., Beginner, Intermediate, Advanced).
  - **content_type** – Format of material (e.g., Video, Quiz, Article, Case Study).
- **Explanation:**
  - The schema captures **what the content is**, **where it belongs**, and **who it is suitable for**.
  - Metadata such as **domain, subtopic, and difficulty_level** are crucial for **content-based filtering**, ensuring recommendations match the learner’s knowledge level and interests.
  - **content_type** allows personalization by aligning recommendations with the learner’s preferred format (e.g., videos for visual learners).
  - Together, these fields enrich the recommendation engine, enabling precise and engaging suggestions.

Engagement Schema Design

- **Fields:**
  - **user_id** – Identifier linking the engagement to a specific user (foreign key from *Users* table).
  - **content_id** – Identifier linking the engagement to a specific lesson or material (foreign key from *Content* table).
  - **timestamp** – Date and time of the interaction, capturing learning sequence and recency.
  - **duration_seconds** – Time spent engaging with the content, useful for measuring depth of interaction.
  - **liked** – Boolean/flag indicating if the user liked the content (*nullable* to allow missing reactions).
  - **engagement_type** – Type of interaction (e.g., Completed, Bookmarked, Shared, Viewed).
- **Explanation:**
  - Captures the **behavioral footprint** of each learner with respect to the content.
  - Links **Users** and **Content** tables, forming the foundation of collaborative filtering.
  - Features like **duration_seconds** and **liked** enrich engagement signals, enabling the recommendation engine to distinguish between **casual interactions** and **meaningful learning activity**.
  - Provides a time-based perspective (**timestamp**) to model evolving learning preferences.
  - **engagement_type** records the **nature of interaction**, offering deeper insights into user intent and content utility.

Diagram showing the connection between Users, Engagements, and Content



Influential Attributes: The Title

- **Why It’s Influential:**
  - Serves as the **most direct representation** of the content’s subject matter.
  - Provides **textual features** that can be transformed into meaningful numerical representations.
  - Applying **TF-IDF (Term Frequency–Inverse Document Frequency)** converts titles into **vectorized features**, capturing both the importance of words and their uniqueness across content.
  - Forms the **foundation of the content-based filtering model**, enabling similarity matching between content items.
  - Enhances recommendations by ensuring that learners are suggested materials closely aligned with their **interests and topical relevance**.

Influential Attributes: The Domain and Subtopic

- **Why They're Influential:**
  - Provide **categorical tags** that add semantic context to each piece of content.
  - Enable the system to make **generalized recommendations** beyond exact keyword matches.
  - Example: A learner engaging with *Artificial Intelligence* content may also be recommended other topics within the broader *Technology* domain.
  - Improve **coverage and diversity** by grouping related materials under shared domains and subtopics.
  - Help mitigate the **cold-start problem** by leveraging domain/subtopic metadata for recommending new or less-interacted content.

Influential Attributes: The Timestamp

- **Why It's Influential:**
  - Adds a **temporal dimension** to user engagement, enriching behavioral analysis.
  - Enables creation of **recency features**, prioritizing newer interactions to reflect current learner interests.
  - Helps capture **seasonality patterns**, such as daily, weekly, or semester-based learning habits.
  - Supports **trend analysis**, revealing how preferences evolve over time (e.g., shifting from beginner to advanced topics).
  - Strengthens the **hybrid recommendation model** by combining static user/content attributes with **dynamic temporal signals**.

Simulating Engagements for Training

- **Strategy:**
  - Simulation data is stored in **engagements.csv**, with each row capturing a **user–content interaction** at a specific time.
  - **engagement_type** is the key simulated signal, designed to represent varying levels of user interest.
  - Interactions are encoded on a **0–10 scale**, combining both **implicit** and **explicit** feedback:
    - **0–5 points** from **duration_seconds** (time spent engaging).
    - **+5 points** if the user explicitly **liked** the content.
  - This design creates a **graded signal** that captures both **depth of engagement** and **sentiment**.
  - Provides a **controlled dataset** for training and testing the **SVD collaborative filtering model** before applying it to real-world data.
  - Enables experimentation with **different scoring distributions**, helping tune the recommendation engine for **accuracy, robustness, and realism**.

Simulating engagement_type

- The **engagement_type** column acts as a **proxy for user feedback**, representing the nature of each interaction.
- Different interaction types capture varying levels of **interest and intent**:
  - **Viewed** → Light engagement, passive interest
  - **Bookmarked** → Intent to revisit or learn later
  - **Completed** → Strong signal of commitment and interest
  - **Shared** → Very strong signal, indicating both interest and endorsement
- These categorical values serve as the **input for training the SVD collaborative filtering model**, allowing it to learn user preferences from diverse behaviors.
- By modeling engagement beyond simple likes/dislikes, the system gains a **richer understanding of user intent** and produces **more nuanced recommendations**.

Simulating User and Content Identity

- **Strategy:**
  - The **simulated user_id and content_id pairs** form the backbone of the dataset.
  - Each pair represents a **unique interaction** between a learner and a piece of content.
  - These pairs are organized into a **user–item interaction matrix**, where:
    - Rows = Users
    - Columns = Content items
    - Values = engagement signals (e.g., engagement_type score)
  - This matrix is the **foundation of collaborative filtering**, enabling the model to:
    - Detect **similarities across users** (user-based filtering).
    - Identify **similarities across content** (item-based filtering).
    - Learn latent patterns that power **personalized recommendations**.
  - Without this matrix, the system cannot generalize user preferences or make predictions for unseen items.

Data Insights - Key Features & Content Observations

- **Feature Redundancy:**
  - *Title* and *Department* strongly correlated → avoid using both in models.
- **Seniority & Domain Preference:**
  - Senior/Lead → Data/Product roles
  - Junior → Marketing roles
- **Learning Style:**
  - Consistent across users → better suited as a personalized filter, not for clustering.
- **Content Characteristics:**
  - Skewed toward beginners & video lessons → handle imbalance for recommendations.
  - Advanced learners have fewer lessons.
  - Limited interactive content → can be highlighted as high-value content.
  - Subtle domain patterns: more quizzes in Data Science, more advanced content in Software Engineering.

Data Insights - User Engagement Insights

- **User Activity:**
  - Majority are casual users (1–5 engagements), few power users (100+) → focus on retention strategies.
- **Content Interaction:**
  - Most content gets 100–1000 interactions; top content drives trends.
- **Daily Engagement:**
  - Stable ~8k/day → reliable platform, growth flat.
- **Engagement Types:**
  - Viewed dominates (~70%), other actions indicate higher commitment.
- **Likes & Session Duration:**
  - ~30% of interactions liked; liked sessions slightly longer (~16 min vs 14.5 min).
  - Stable across departments, domains, seniority.
  - Slightly higher likes in Marketing/Engineering; longer sessions for Leads → guide targeted engagement.
- **Overall Insight:**
  - User engagement is consistent → content appeals broadly across roles/domains.