**Project Report: A Hybrid Recommendation System for Post Engagement**

**Collab link:**
[https://colab.research.google.com/drive/1Qc1o98ioqXcFVVzlPt6a9CMcxrMsS2VC?usp=sharing](https://colab.research.google.com/drive/1Qc1o98ioqXcFVVzlPt6a9CMcxrMsS2VC?usp=sharing)

It was a wonderful opportunity and a genuine pleasure to work on the problem statement your company is solving. The challenge of connecting users with meaningful content is a fascinating one, and I thoroughly enjoyed applying my knowledge of machine learning to build a system that recommends the top 3 most relevant posts to each user. This project was not only a great application of existing skills but also a valuable learning experience.

---

**Project Goal**

The primary objective was to develop a recommendation model that takes user profile data, post attributes, and historical engagement patterns to generate a personalized list of the top 3 posts for every user, aiming to maximize engagement.

---

**My Approach: A Two-Pronged Hybrid System**

To tackle this, I developed a hybrid recommender system that combines the strengths of two distinct modeling techniques. This ensures the recommendations are both relevant (based on content) and personalized (based on behavior).

**1. Content-Based Model (LightGBM)**

This model acts like a feature-based matchmaker. It uses a **LightGBM classifier**, a lightweight and powerful gradient-boosting model, to learn the relationship between user/post attributes and engagement.

- **User Features**: age, gender, top_3_interests, past_engagement_score.
- **Post Features**: content_type, tags, and an engineered popularity_score.

The goal of this model is to answer: "Based on a user's interests and a post's characteristics, how likely is an engagement?"

**2. Collaborative Filtering Model (SVD)**

This model works on the principle of "users who liked this also liked...". I used **Singular Value Decomposition (SVD)**, a classic and efficient matrix factorization technique.

- It processes the user-post engagement matrix to uncover latent (hidden) patterns in user tastes.

- It excels at providing personalized recommendations based on the behavior of similar users, even if the content features aren't descriptive.

---

**A Key Challenge & Learning: The "Noisy Zero" Problem**

One of the most interesting challenges I faced was designing a robust evaluation strategy. In recommendation systems, we only have data on what users *did* engage with (the '1's). A lack of engagement (a '0') doesn't necessarily mean dislike; it could simply mean the user never saw the post. This is the **"Noisy Zero" problem**.

To overcome this for evaluation, I implemented the following strategy:

- **Ground Truth**: For a user in the validation set, I took the posts they *actually engaged with* (their positive interactions from Engagements.csv) and held them out as the "ground truth." These are the items we hope our model successfully recommends.
- **Candidate Pool**: The model's task is then to rank this small set of ground-truth posts against a much larger pool of posts the user has never seen (the assumed '0's).

This method tests the model's ability to make positive items stand out from the noise, which is a much more realistic measure of success than simple classification accuracy.

---

**Evaluation Metrics**

To measure the quality of the ranked recommendations, I used two industry-standard metrics:

- **AUC (Area Under the Curve)**: Measures whether the model is generally better at giving higher scores to positive items than to negative ones across the entire candidate pool.
- **nDCG (Normalized Discounted Cumulative Gain)**: A top-heavy metric that evaluates if the *most relevant* items appear at the very top of the recommendation list.

---

**Results & Key Insights**

I implemented a **grid search** to perform hyperparameter tuning and find the optimal blend between the content-based and collaborative models. The results were clear and insightful:

**For this specific dataset, collaborative patterns are a much stronger and more reliable signal than the available content features.**

The optimal configuration heavily favored the SVD model (content_weight of 0.1). The content model, especially with its generic popularity feature, was acting more like a noisy distraction than a helpful signal. By giving it a very low weight, we effectively **benched the noisy player and let our star player—the personalized SVD model—run the game.** This highlights the immense value of user interaction data.

---

### Conclusion

I am confident that this hybrid framework provides a robust and well-reasoned solution to the recommendation problem. The project reinforced the importance of a rigorous evaluation setup and demonstrated that sometimes, the simplest, most direct signals—in this case, user behavior—are the most powerful. Thank you again for the fantastic opportunity to work on this challenge.