Movie Recommendation System

Chase Thompson and Mclean Dries

Central Research Question: How effectively can a machine learning model predict movie ratings using user history and movie metadata?

Baseline Models:

Baseline 1: Global Mean Rating

Predict every unknown rating as the overall dataset mean (~3.5).

Baseline 2: User Mean Rating

Predict the user's average past rating.

Baseline 3: Item Mean Rating

Predict each movie's mean rating from all other users.

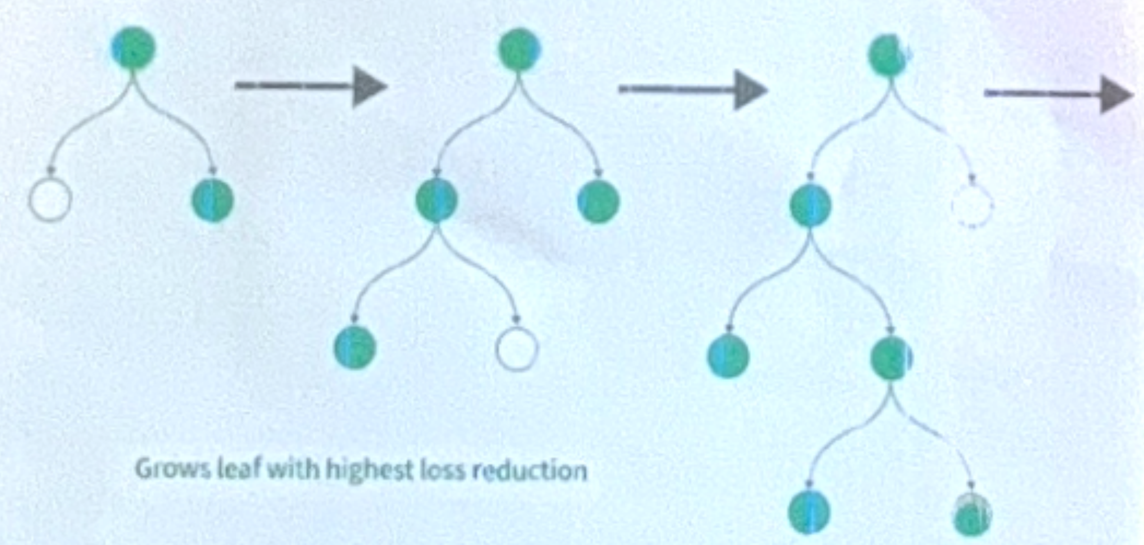
Examples of engineered features:

Feature	Description
user_avg_rating	Mean rating by the user
user_rating_count	How many movies the user has rated
movie_avg_rating	Mean rating for the movie
movie_rating_count	Popularity proxy
genre_vector	One-hot encoding of genres
rating_timestamp	Time-based ordering

Primary Contributions:

1. A cleaned and processed subset of the MovieLens dataset.
2. Implementation of three baselines (global mean, user mean, item mean).
3. A LightGBM model trained on user–movie interaction features.

How LightGBM Works



Grows leaf with highest loss reduction

Input-Output Example Table

Input Example	Output Example
user_avg_rating = 3.9, movie_avg_rating = 4.2, genres = (Drama, Crime), movie_rating_count = 12,000	Predicted rating: 4.4
user_avg_rating = 2.8, movie_avg_rating = 3.1, genres = (Comedy), movie_rating_count = 400	Predicted rating: 2.5

Conclusion & Future Work

We built a simplified but effective movie recommendation model using LightGBM and the MovieLens 25M dataset. The model significantly outperformed all baselines and demonstrated strong predictive power through engineered user–movie features.

If given more time, future work could include incorporating review text using NLP embeddings, addressing popularity bias with feature regularization, adding user demographic features, and/or deploying the model as a simple web app.

Movie Recommendation System

Chase Thompson and Mclean Dries

Central Research Question: How effectively can a machine learning model predict movie ratings using user history and movie metadata?

Primary Contributions:

1. A cleaned and processed subset of the MovieLens dataset.
2. Implementation of three baselines (global mean, user mean, item mean).
3. A LightGBM model trained on user–movie interaction features.

Baseline Models:

Baseline 1: Global Mean Rating

Predict every unknown rating as the overall dataset mean (~3.5).

Baseline 2: User Mean Rating

Predict the user's average past rating.

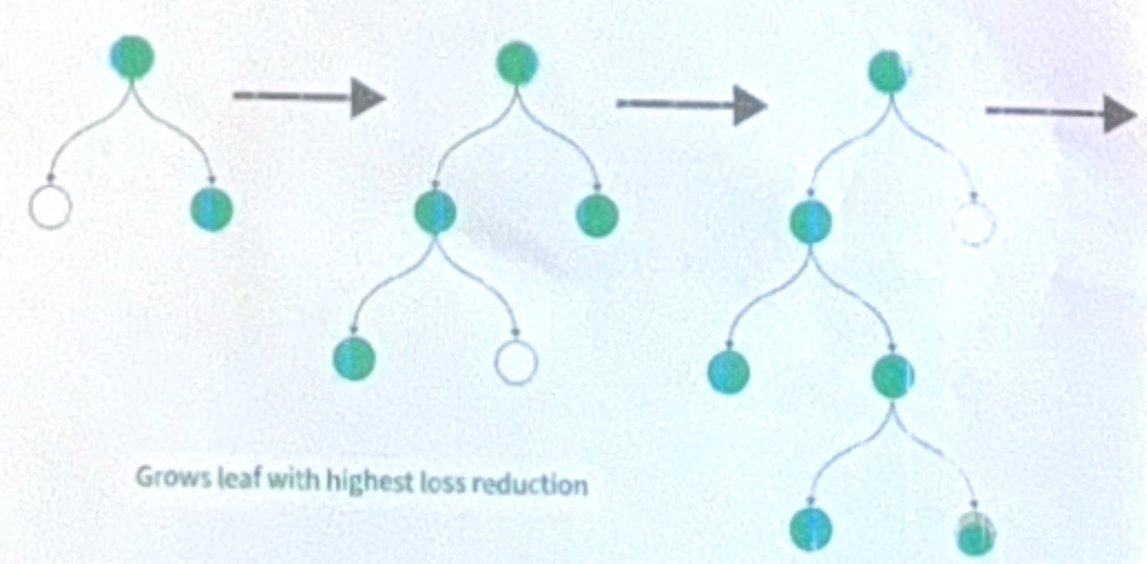
Baseline 3: Item Mean Rating

Predict each movie's mean rating from all other users.

Examples of engineered features.

Feature	Description
user_avg_rating	Mean rating by the user
user_rating_count	How many movies the user has rated
movie_avg_rating	Mean rating for the movie
movie_rating_count	Popularity proxy
genre_vector	One-hot encoding of genres
rating_timestamp	Time-based ordering

How LightGBM Works



Input–Output Example Table

Input Example	Output Example
user_avg_rating = 3.9, movie_avg_rating = 4.2, genres = {Drama, Crime}, movie_rating_count = 12,000	Predicted rating: 4.4
user_avg_rating = 2.8, movie_avg_rating = 3.1, genres = {Comedy}, movie_rating_count = 400	Predicted rating: 2.5

Conclusion & Future Work

We built a simplified but effective movie recommendation model using LightGBM and the MovieLens 25M dataset. The model significantly outperformed all baselines and demonstrated strong predictive power through engineered user–movie features.

If given more time, future work could include incorporating review text using NLP embeddings, addressing popularity bias with feature regularization, adding user demographic features, and/or deploying the model as a simple web app.