Zee Recommender Systems: Personalized Movie Recommendations

❖ Topic: Building a Personalized Recommender System for Movies
❖ Duration: 1 week

Why this case study?

From the company's perspective:

- Zee Recommender Systems represents an ambitious venture by Zee to enhance user experience through personalized movie recommendations.
- The focus is on leveraging user ratings and similarities among users to create a robust, personalized movie recommender system.
- Utilizing a comprehensive dataset of movie ratings, user demographics, and movie details, Zee aims to develop a system that can accurately predict user preferences and suggest movies accordingly.
- The insights gained from this system are expected to drive user engagement, increase satisfaction, and foster a more intuitive user experience.

From the learner's perspective:

- This case study offers an opportunity to delve into the complex and highly relevant field of recommender systems, a cornerstone of modern AI-driven platforms.
- Learners will explore collaborative filtering techniques, both item-based and user-based, and understand their applications in real-world scenarios.
- The project will cover a range of concepts including Pearson Correlation, Cosine Similarity, and Matrix Factorization, providing a comprehensive learning experience in building recommender systems.

- Participants will gain hands-on experience in not just building but also evaluating these systems, honing skills that are highly sought after in the field of data science and machine learning.

Dataset Explanation: Zee Recommender Systems Data

The dataset for this project is composed of three primary files, each contributing essential information for building the recommender system:

1. Ratings File (ratings.dat):

    - Format: UserID::MovieID::Rating::Timestamp
    - Contains user ratings for movies on a 5-star scale.
    - Includes a timestamp representing when the rating was given.
    - Each user has rated at least 20 movies.

2. Users File (users.dat):

    - Format: UserID::Gender::Age::Occupation::Zip-code
    - Provides demographic information about the users, including gender, age group, occupation, and zip code.
    - Demographic data is voluntarily provided by users and varies in accuracy and completeness.

3. Movies File (movies.dat):

    - Format: MovieID::Title::Genres
    - Lists movie titles alongside their respective genres.
    - Genres are categorized into multiple types like Action, Comedy, Drama, etc., and are pipe-separated.

Key Points to Note:

1. UserID and MovieID serve as unique identifiers for users and movies, respectively.

2. Ratings reflect user preferences and are crucial for understanding individual and collective tastes.
3. User Demographics (gender, age, occupation) can provide insights into user preferences and behavior patterns.
4. Movie Details (title, genres) are essential for categorizing movies and understanding their appeal to different user segments.

What is Expected?

Assuming you're a data scientist at Zee, your responsibility involves creating a personalized movie recommender system. Your primary goals are:

- To analyze user ratings, demographic data, and movie characteristics to understand viewing preferences.
- To apply collaborative filtering, Pearson Correlation, Cosine Similarity, and Matrix Factorization techniques to build an effective recommender system.
- To evaluate the system's performance and refine it for accuracy and user relevance.

Submission Process:

Upon completing the Zee Recommender Systems project...

- Document Your Findings: Compile your methodologies, analysis, and insights in a Jupyter Notebook.
    - Ensure your notebook includes:
        - Demonstrated Python code for data processing, building the recommender system, and its evaluation.
        - Visualizations that support your analysis, such as user-rating distribution charts, similarity matrices, etc.
        - Final insights and actionable recommendations based on your analysis to enhance Zee's user experience.

- Convert to PDF: Turn your Jupyter Notebook into a PDF (using the Chrome browser's Print function).

- Follow Submission Guidelines: Adhere to the prescribed submission process and upload your PDF on the designated platform.

- Note on Revisions: Once submitted, revisions to your work will not be possible.

General Guidelines:

- Approach as a Real-World Challenge: Embrace this project as a typical task you would encounter as a data scientist in the entertainment industry.
- Navigating Through Challenges:
    - Frequently revisit the problem statement to ensure alignment with the objectives.
    - Break complex tasks into smaller, more manageable steps.
    - Utilize online resources, forums, or documentation for overcoming coding challenges or conceptual doubts.
- Collaboration and Discussion: Engage with peers in discussion forums for a broader perspective and collaborative problem-solving.
- Seeking Clarity and Knowledge: Revisit educational materials or seek external resources for a better understanding of complex concepts.
- Instructor Assistance: Reach out to your instructor for clarifications or if facing significant challenges.
- Adopt a Growth Mindset: View every challenge as a learning opportunity. Tackle the project with enthusiasm, dedication, and a willingness to learn.

What does 'good' look like?

1. Define the Problem Statement and perform Exploratory Data Analysis

	Hint	Approach
a. Definition of problem	Clearly articulate the objective of creating a recommender system.	a. Understand the significance of personalized movie recommendations in enhancing user experience. b. Identify the key elements a successful recommender system should address, such as accuracy, relevance, and user engagement.
b. Exploratory Data Analysis (EDA)	Dive deep into the dataset to uncover underlying patterns and insights.	a. Analyze the distribution of movie ratings, user demographics, and movie genres. b. Utilize functions like data.describe(), data.info(), and visual tools to explore the data. c. Investigate the range of ratings and the frequency of ratings per movie and per user. d. Examine user demographics and their potential influence on movie preferences.
c. Scope for Exploration	Visualize the data in various forms (histograms, bar charts, scatter plots) to get a comprehensive view of the distribution and relationships.	a. Explore relationships between different variables, such as age groups and genre preferences. b. Look for trends or anomalies in the data, such as unusually high or low ratings for certain movies.
d. Initial Insights	Document your observations and initial hypotheses based on your EDA.	a. Note any surprising findings, such as specific genres being more popular in certain age groups. b. Comment on the rating distribution and its potential impact on the recommender system's performance.

		c. Provide a preliminary assessment of the data quality and any potential challenges in building the recommender system.

2. Data Preprocessing

	Hint	Approach
a. Data Cleaning and Formatting	Ensure the dataset is clean and formatted correctly for analysis.	a. Check for and handle missing values or anomalies in the dataset. b. Format the data correctly, especially the timestamps and categorical data like genres and occupations. c. Normalize or standardize the data if necessary, particularly for any numerical fields that might be used in the analysis.
b. Data Transformation	Transform the data into a format suitable for building a recommender system.	a. Encode categorical data, like movie genres and user occupations, using methods such as one-hot encoding or label encoding, making them suitable for algorithmic processing. b. Create a matrix or a table that aligns users with their movie ratings, which is essential for collaborative filtering.
c. Feature Engineering	Derive new features that could enhance the recommender system's performance.	Possible Steps: a. Consider creating features like average rating per user, average rating per movie, total number of ratings per movie, etc. b. Explore creating user profiles based on their rating patterns and demographic information. Scope for Exploration: a. Investigate whether incorporating the time factor (when the rating was given)

		could provide insights into user preferences. b. Extract and utilize features like “Release Year”, “Year”, “Month”, “Day”, “DayOfWeek”, “isWeekend”, “isHoliday”, and “State_US”. b. Experiment with different ways of representing movie genres, such as one-hot encoding or multi-label encoding.
d. Handling Sparse Data	Address the issue of sparsity in the user-item interaction matrix.	a. Evaluate the level of sparsity in the dataset and its potential impact on the model. b. Consider techniques like matrix factorization or imputation methods to handle sparse data effectively.

3. Model building

	Hint	Approach
a. Collaborative Filtering with Pearson Correlation	Implement an item-based collaborative filtering approach using Pearson Correlation.	a. Create a pivot table with movie titles and user IDs, imputing NaN values appropriately. b. For a given movie, calculate the Pearson Correlation with other movies to find the most similar ones. c. Recommend movies to a user based on the highest correlations to the movies they have already rated.
b. Collaborative Filtering with Cosine Similarity	Use Cosine Similarity for measuring item-item and user-user similarities.	a. Generate similarity matrices for items (movies) and users based on their rating patterns. b. Employ the Nearest Neighbors algorithm with Cosine Similarity to find and recommend similar movies.

c. Matrix Factorization Techniques	Apply Matrix Factorization for a more sophisticated approach to recommendations.	a. Utilize libraries like ‘cmfrec’ or ‘Surprise’ to perform Matrix Factorization on the rating data. b. Explore different factorization dimensions (e.g., d=4) and observe their impact on the recommendations. c. Evaluate the model using metrics like RMSE and MAPE to gauge its predictive accuracy.
d. Model Evaluation and Tuning	Evaluate and refine your models for the best performance.	a. Assess the models based on their ability to accurately predict user preferences. b. Consider performing a train-test split specifically for matrix factorization and evaluate the performance on unseen data.
e. Advanced Collaborative Filtering Techniques	Experiment with embedding-based approaches and visualization techniques.	a. Use matrix factorization to derive embeddings for items and users, and explore item-item and user-user similarities based on these embeddings. b. For a deeper analysis, visualize the embeddings in two dimensions and interpret the resulting clusters and relationships.

4. Results Interpretation & Stakeholder Presentation

	Hint	**Approach**
a. Understand the Business Context	Interpret the results in the context of Zee’s business goals and user experience.	a. Discuss how the recommender system can enhance user engagement and satisfaction. b. Highlight how personalized recommendations can potentially increase user retention and platform usage.

b. Presentation of Findings	Present your findings in a manner that is accessible and compelling to stakeholders.	a. Use clear and informative visualizations to demonstrate the effectiveness of the recommender system. b. Prepare a presentation or report that includes key metrics, insights, and the business implications of your findings.
c. Discussion of Model Performance	Provide a detailed analysis of the models' performance.	a. Discuss the accuracy of the recommendations and any patterns or trends identified in the user preferences. b. Compare the performance of different models (Pearson Correlation, Cosine Similarity, Matrix Factorization) and highlight the strengths and limitations of each.
d. Recommendations for Improvement	Suggest actionable improvements based on your analysis.	a. Propose strategies for further refining the recommender system, such as incorporating additional user feedback or exploring more advanced algorithms. b. Consider recommending methods for ongoing monitoring and updating of the system to adapt to changing user behaviors and preferences.
e. Future Considerations	Look ahead to future enhancements and potential developments.	a. Suggest areas for further research or additional data that could be leveraged to improve the system. b. Discuss the scalability of the recommender system and its adaptability to other types of content or user segments.

Questionnaire (Answers should be presented in the text editor along with insights):

1. Users of which age group have watched and rated the most number of movies?

2. Users belonging to which profession have watched and rated the most movies?

3. Most of the users in our dataset who've rated the movies are Male. (T/F)

4. Most of the movies present in our dataset were released in which decade?
    - 70s b. 90s c. 50s d.80s

5. The movie with the maximum no. of ratings is ___.

6. Name the top 3 movies similar to 'Liar Liar' on the item-based approach.

7. On the basis of approach, Collaborative Filtering methods can be classified into ___-based and ___-based.

8. Pearson Correlation ranges between ___ to ___ whereas, Cosine Similarity belongs to the interval between ___ to ___.

9. Mention the RMSE and MAPE that you got while evaluating the Matrix Factorization model.

10. Give the sparse 'row' matrix representation for the following dense matrix -
    [[1 0]
    [3 7]]