FilmFinder: Simplifying Choices with Data-Centric Recommendations

Ali Nikan
Kimia Naghavi

Introduction:

In this project, we propose to develop a movie recommendation system leveraging machine learning techniques, with a strong emphasis on conducting a data-centric exploration. Inspired by the ideas presented in the DataPerf reading, this project will not only improve the system's accuracy but also guarantee the usefulness of suggestions by using data valuation strategies and closely studying different dataset selection decisions. Delivering a customized recommendation engine that responds to each user's unique preferences and enhances their movie-watching experience is the ultimate goal in mind.

Objective:

The primary objective of this project is to design and implement a movie recommendation system that provides users with personalized movie suggestions based on their viewing history, preferences, and demographic information. Understanding how different data selection methods and valuation techniques affect the system's overall performance is our main objective, since this implies that our suggestions are not only reliable but also highly personalized and contextually relevant. We will explore various personalization techniques, including behavioral and contextual personalization, to provide a clear picture of how we intend to achieve high levels of personalization.

- **Behavioral Personalization:** To provide suggestions, this approach monitors users' past activities. A user's viewings, ratings, and search histories are examples of their activities. They may also include their interactions with the movies they have seen and rated (e.g., watch time, ratings given, reviews made). The recommendation system can determine what additional movies the user would enjoy by analyzing these behaviours and activities. For example, the system might suggest more animated films if a user usually watched and gave them good ratings.
- **Contextual Personalization:** Contextual personalization considers the user's interaction context with the system. The time of day, the device being used, the place, the user's present feelings as assumed by their interactions, and even the weather may all be included in this. For example, if the user has preferences for certain types of movies at certain times, the system may suggest lighter comedy on a rainy day or action-packed movies on a weekend evening.

Methodology:

- **Data Collection and Preprocessing:**
  From credible sources, we will collect a comprehensive and varied dataset containing movie details, user ratings, genres, and demographic data. The preprocessing stage will include feature engineering to gather important characteristics that can greatly impact recommendation outcomes, as well as careful data cleaning to handle missing values and outliers. We will also address potential challenges such as handling biased data and propose preliminary solutions to ensure the quality and fairness of our dataset.
- **Model Development:**

We will use a combination of content-based and collaborative filtering techniques. For example, we will look into neighborhood-based strategies and matrix factorization techniques for collaborative filtering; deep learning models will also be taken into consideration if more complex procedures are needed along the way. To further customize the suggestions for content-based filtering, we will look at user preferences and movie metadata. We will detail the criteria for choosing between different ML models and explain when we might opt for deep learning approaches over traditional matrix factorization, ensuring the best fit for our system's needs.

- **Evaluation and Validation:**
  We intend to employ a split dataset that includes training, validation, and testing to evaluate the system's performance fully. To accurately assess the level of quality and usability of the suggestions, we will use several kinds of metrics, such as precision, recall, F1-score, and other advanced metrics that relate to recommendation systems. The best models and data procedures will be found by A/B testing.

Data Exploration Component:

This important phase includes an extensive review of the dataset to identify the essential characteristics that have an important effect on the quality of recommendations. We will evaluate the inherent worth of distinct data segments using data valuation approaches, and we will run some experiments with different dataset types to see how they affect model performance. In addition to identifying the features of the data, this examination will guide the best choices for datasets and preprocessing techniques.

Reporting and Documentation:

A thorough report will cover all aspects of the project, from methods to learned lessons. It will highlight the results of the data-centric exploration findings, offering an overview of how data selection and valuation decisions affect how successful movie recommendation systems are.

Expected Outcome:

By the end of the project, we expect to have developed a sophisticated movie recommendation system with a focus on data-centric exploration. The project will provide valuable insights into the factors influencing recommendation accuracy and the effectiveness of different data selection strategies. Additionally, the findings will contribute to advancing the field of recommendation systems and personalized user experiences.

Conclusion:

This project, which hopes to fully improve customized entertainment experiences, represents a unique progress of data-centric exploration and machine learning innovation. Our goal is to raise the bar for recommendation systems' efficiency and sophistication in the constantly evolving industry of entertainment by using advanced techniques and insights from the DataPerf reading.