IS 733 DATA MINING

PROJECT MILESTONE - 1

GROUP - 11

MOVIE REVIEW

By

Asmita Deshpande
Gayathri Gurram
Ravi Sharma
Tarunsingh Jodha

INDEX

Use Case Development

1) What is the context for this proposed project? Why is it important?

The TMDB 5000 Movie Dataset is a large dataset of movie metadata, including information such as title, release date, genre, budget, revenue, and cast. The growth of OTT and VOD continue to reach millions of viewers as online media takes control. The digital media platforms are diving in for applications that read user minds and facilitate a recommended list of movies and shows that they are more likely to get drawn to due to a rapid shift in consumer behavior and perceptions. And the internet entertainment sector is subject to the same rule. Calling recommendations includes a wide range of strategies to boost OTT growth and revenue. Leading OTT and VOD services use recommendation systems to entice users and are backed by subscription-based, transactional, and ad-based business models. Recommendation engines today do more than just promote tailored content. The hybrid recommendation system increases user stickiness and loyalty while customizing the online streaming experience for each individual user.

2) Who will benefit from this proposed solution?

There are multiple industries who can benefit from the proposed solution like the stakeholders which include are listed below:

1. **Users**: The movie recommendation system primarily benefits users who seek personalized movie recommendations based on their viewing history, preferences and ratings. The system can help users find new movies that match their interests and save time and effort searching for movies to watch. Movie fans can use the database to discover new movies and learn more about the movies they love.
2. **Movie streaming platforms**: Companies like Netflix, Amazon Prime and Hulu can use this system to improve their user experience and retention rates. By offering personalized recommendations, these companies can keep their users engaged and subscribe to their services for longer. Film studios and production companies.
3. **Movie studios and production companies**:The information about audience preferences and trends that a recommendation system would give to film studios and production businesses would also be advantageous. They can utilize this information to better inform their choices on movie production, release, and promotion. The dataset can be used by movie studios to track the success of their productions and spot patterns in the movie business.
4. **Advertisers**: a recommendation system can help advertisers target their ads more effectively to users who are most likely to be interested in their products or services. By providing personalized recommendations, the system can improve ad clicks and results. Film distributors can use the dataset to identify potential markets for their films and target marketing campaigns.

3) Describe a use case of this data science system (e.g. how the user might use this tool?)

One potential use case for the TMDB 5000 Movie Dataset is to develop a movie recommendation engine. A movie recommendation engine is a system that can suggest movies to users based on their past viewing history and preferences. The TMDB 5000 Movie Dataset can be used to train a movie recommendation engine by providing it with information about the movies that users have watched and enjoyed. Once the movie recommendation engine is trained, it can be used to suggest movies to users who are looking for something new to watch.

Another potential use case for the TMDB 5000 Movie Dataset is to develop a movie search engine. Using their keywords or search terms, people can find movies with the use of a movie search engine. The movie data may be easily searched for by indexing it using the TMDB 5000 Movie Dataset. After it has been indexed, the movie search engine can be used by users to find movies by title, genre, actor, director, or any other search term.

Finally, tools for film analysis can be created using the TMDB 5000 Movie Dataset. A system that can improve user comprehension of movies is known as a movie analysis tool. Users can access details about the movies they are watching, including cast and crew, budget, box office results, and reviews, by using the TMDB 5000 Movie Dataset. The dataset can be used to develop a variety of data science systems that can help people find, watch, and understand movies.

Datasource Description

1) A high-level description of the data source

The TMDB 5000 Movie Dataset is a large dataset of movie metadata, including information such as title, release date, genre, budget, revenue, and cast. The dataset was created by scraping the website The Movie Database (TMDb).

2) What kind of work already exists for this data source?

There is a lot of work that has already been done with the TMDB 5000 Movie Dataset. Some of the work that has been done includes:

- Sentiment Analysis: This usually involves analyzing the sentiment expressed in movie reviews, generally categorized on a scale of positive,negative or neutral.The goal is to extract insights from the reviews to understand strengths and weaknesses of a movie and to predict how well it is to be received by the audience.
- Genre classification: Genre classification involves predicting the genre of a movie based on its features such as plot, characters, and setting. This can help in organizing and categorizing movies for better search and browsing experience for users.
- Topic modeling: Topic modeling involves extracting topics from movie reviews using techniques such as Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF). This

can help in identifying themes and trends in the reviews and can be used to improve the recommendation system.

- Opinion mining: Opinion mining involves extracting and summarizing opinions from movie reviews using Natural Language Processing (NLP) techniques. This can help in identifying the most common opinions about a movie and can be used to generate a summary of the reviews.
- Recommender systems: Recommender systems analyze user behavior data and provide personalized movie recommendations to users based on their past viewing history, ratings, and reviews. These systems use various techniques such as collaborative filtering, content-based filtering, and hybrid filtering to generate personalized recommendations.

3) Link to the data source

The data source can be found here:
https://www.kaggle.com/datasets/tmdb/tmdb-movie-metadata?select=tmdb_5000_movies.csv

Model Development

A high-level idea of the model

The model will be a machine learning model that can be used to predict the rating of a movie based on its metadata. The model will be trained on a dataset of movies that have already been rated by users. Once the model is trained, it can be used to predict the ratings of new movies.

1) **What type of ML problem is it? Supervised/unsupervised; if supervised, is it classification or regression?**

The model is a supervised learning model. This means that the model is trained on a dataset of labeled data. In this case, the data is labeled with the ratings of the movies.Overall, the choice of algorithm depends on the specific requirements of the recommender system. Factors such as scalability, interpretability, and accuracy should be considered when choosing a supervised machine learning algorithm for building a movie recommender system.

2) **What does training data look like?**

The training data will consist of a dataset of movies that have already been rated by users. The dataset will include the following information about each movie:

- Title
- Release date
- Genre
- Budget
- Revenue
- Cast

- Crew
- User ratings

Unit of analysis - what each row in the training data table represents

Each row in the training data table represents a single movie. The rows are ordered by the release date of the movies.Since out project is based on recommender systems, typical features include,user ID, the movie ID, and the rating given by the user.

What are the input variables?

The input variables to the model are the following:

- Title: A unique identifier for the movie title in the TMDB database
- Release date: The date at which the movie has been released
- Genre: Indicating the genre of the movie
- Budget: The budget of the movie (US dollars)
- Revenue: The box office revenue (US dollars)
- Cast/Crew: The cast of the movie
- Runtime: The runtime of the movie, in minutes.
- spoken_languages: A list of the languages spoken in the movie.
- status: The status of the movie (e.g., released, Postproduction, rumored, etc.).
- tagline: The movie's tagline or marketing slogan
- ID: A unique identifier for the movie in the TMDb database.
- homepage: The URL of the movie's official website.
- Keywords: A list of keywords associated with the movie.
- original_language: The original language of the movie.
- original_title: The original title of the movie.
- overview: A summary or description of the movie.

What are the output/labels? If applicable

The possible outputs or labels for a movie reviews project depend on the specific goals of the project.For out project we have decided to consider rating as out variable to be predicted.Rating label indicate the numerical rating given to a movie by the reviewer, usually on a scale of 1-5 or 1-10. This is a common label used in movie recommendation systems.

Model Evaluation

1) **Which performance metrics do you plan to use to evaluate your model performance?**

I plan to use the following performance metrics to evaluate my model performance:

- Accuracy_Score is the percentage of predictions that the model gets correct.
- Precision is the percentage of positive predictions that are actually positive.
- Recall is the percentage of positive instances that are predicted as positive.

- F1 score is a weighted average of precision and recall.

- K Fold cross validation is splitting the data into k parts and evaluate the model's ability with the new data.

2) How do you go about gathering those metrics?

I will gather those metrics by splitting the dataset into a training set and a test set. The training set will be used to train the model, and the test set will be used to evaluate the model's performance. I will then use the following steps to gather the metrics:

a.Calculate the accuracy of the model on the test set.

b.Calculate the precision of the model on the test set.

c.Calculate the recall of the model on the test set.

d.Calculate the F1 score of the model on the test set.