# MOVIE RECOMMENDATION SYSTEM

UTILIZING MOVIELENS DATASET


OVERVIEW


BUSINESS UNDERSTANDING


OBJECTIVES

DATA UNDERSTANDING & PREPARATION


DATA ANALYSIS AND MODELING

CONCLUSION

RECOMMENDATIONS & NEXT STEPS

# AREAS OF FOCUS

# OVERVIEW

Our project centers around the development of an advanced movie recommendation system, leveraging the expansive movie lens dataset.


This dataset is a treasure trove of information, encompassing a diverse array of user ratings, intricate movie details, and valuable demographic data. By tapping into this rich resource, we aim to revolutionize the way users discover and enjoy movies.
Unlike conventional recommendation systems, our approach takes into account not only user preferences, as indicated by ratings, but also delves into the intricate details of each movie, allowing for a more personalized and engaging recommendation experience.

This comprehensive dataset serves as the cornerstone of our endeavor, empowering us to build a recommendation system that goes beyond the ordinary, offering a nuanced understanding of user preferences and movie characteristics.

Situated in the realm of movie recommendation systems, our project strategically utilizes the MovieLens dataset provided by the Group Lens research lab at the University of Minnesota.

Our primary mission is to address the inherent challenge that users encounter when navigating through an extensive catalog of movies. To overcome this hurdle, we are developing a recommendation system grounded in collaborative filtering.

This approach relies on the collective preferences and behaviors of users to offer insightful and tailored movie suggestions. Through collaborative filtering, we aim to streamline the movie selection process for users, enhancing their overall viewing experience by providing them with personalized and relevant recommendations based on the collective wisdom of the user community.

# PROPOSED SOLUTION

1. Collaborative Filtering: User-Based Collaborative Filtering: Recommends items based on the preferences of users with similar tastes. Item-Based Collaborative Filtering: Recommends items similar to those liked by the user.
2. Content-Based Filtering: Utilizes information about the items themselves and recommends items with similar features to what the user has liked in the past. Hybrid Approaches:
3. Combines collaborative and content-based methods to leverage the strengths of both approaches. Matrix Factorization:
4. Decomposes the user-item interaction matrix into latent factors, capturing hidden patterns and relationships.
5. Deep Learning Models: Neural network architectures, such as auto encoders and recurrent neural networks (RNNs), can learn complex patterns and relationships for improved recommendations.

# BUSINESS UNDERSTANDING


Our project caters to a diverse set of audiences and stakeholders, each playing a crucial role in the success and impact of the recommendation system. Firstly, individual end users, who are avid movie enthusiasts seeking personalized and accurate movie suggestions for their personal entertainment.


By delivering tailored recommendations, we aim to significantly enhance their overall viewing experience. Streaming platforms, representing companies providing online streaming services, form another vital stakeholder group.
These platforms host an extensive catalog of movies and are keen on integrating an effective recommendation system to boost user engagement, satisfaction, and retention. Content providers, comprising studios, distributors, and filmmakers, contribute movies to these platforms and are keenly interested in maximizing the visibility and viewership of their content through influential and effective recommendations.
Lastly, business decision-makers, including executives and managers of streaming platforms or related businesses, are invested in understanding the impact of the recommendation system on key user metrics such as retention and revenue. The collaboration and engagement of these stakeholders contribute to the holistic success of our movie recommendation system.

# OBJECTIVES

**Main Objective:**

**Build a Recommender Model:** Develop an effective recommendation model capable of offering personalized top 5 movie suggestions to users based on their ratings of other movies. This primary objective focuses on enhancing the user experience and satisfaction.

**Additional Objectives:**

**Determine Highest Rating Score:**

Identify and analyze the highest rating score within the dataset to understand the upper limit of user ratings. This insight provides valuable information about user preferences and the overall rating distribution.

**Analyze Rating Trends Over the Years:**

Investigate and analyze the temporal trends in movie ratings over the years. This analysis helps in identifying patterns, fluctuations, and potential factors influencing user ratings, contributing to a comprehensive understanding of user behavior.

**Analyze Top 5 Rated Genres:**

Conduct an in-depth analysis to determine and showcase the top 5 rated movie genres based on user preferences. This information is crucial for content providers and streaming platforms, enabling them to tailor their content offerings to match popular genres and enhance user engagement.

# DATA UNDERSTANDING & PREPARATION

The ml-latest-small dataset, sourced from MovieLens, constitutes a valuable resource for our recommendation system project. Key dataset details include:

- Dataset Size: The dataset comprises 100,836 movie ratings, 3,683 user-generated tags, and spans 9,742 movies, with interactions from 610 users.
- Rating Scale: User ratings are primarily based on a 5-star scale, providing a nuanced understanding of user preferences.
- Temporal Span: Data spans from March 29, 1996, to September 24, 2018, capturing long-term user interactions with the MovieLens platform.
- Demographic Absence: Notably, the dataset lacks demographic information about users, identified solely by unique numerical identifiers.
- File Structure: Organized into four main files (links.csv, movies.csv, ratings.csv, and tags.csv), with movies.csv and ratings.csv being the primary focus for our analysis.

Movies.csv:

- Columns: Includes movieId (unique identifier), title, and genres (separated by '|').

Ratings.csv:

- Columns: Encompasses userId (unique identifier), movieId (link to specific movie), rating (user's numerical rating), and timestamp (Unix timestamp of the rating).

Data Preparation Steps:

- Library Import: Relevant libraries were imported for efficient data handling and analysis.
- Loading Dataset: The dataset was loaded from the CSV format in which it was stored.
- Data Frame Creation: A new Data Frame was created, extracting necessary columns for analysis.

# DATA ANALYSIS & MODELING

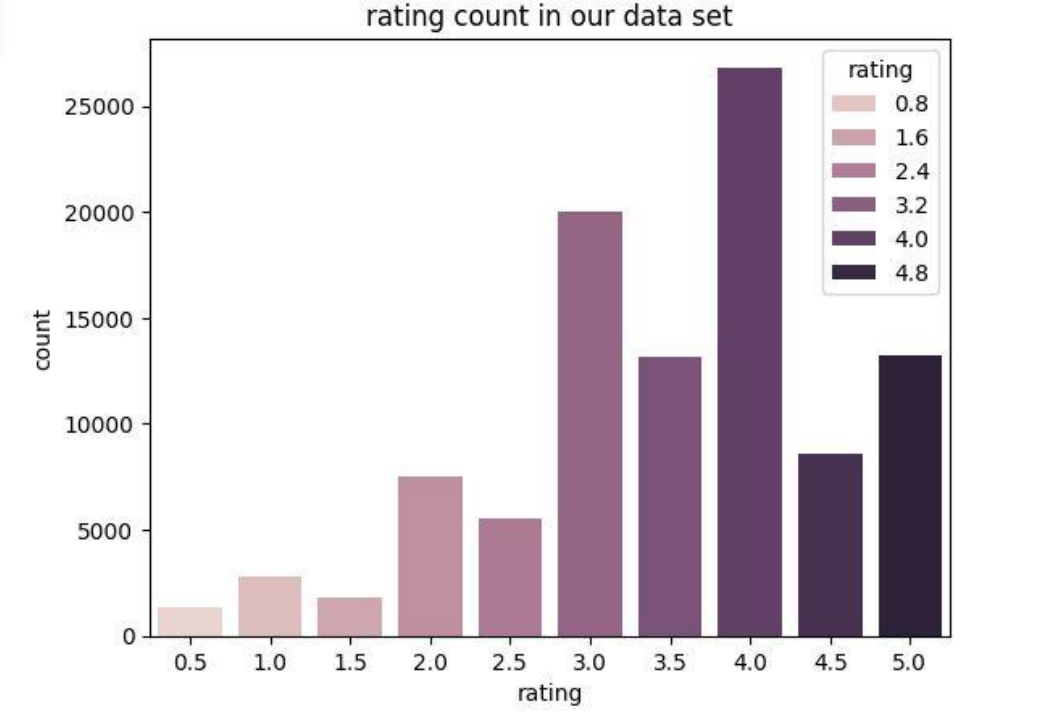
Exploratory Data Analysis (EDA):

In this phase of our project, we delved into the intricate landscape of movie ratings and user behaviors through a comprehensive Exploratory Data Analysis (EDA).
Through visualizing our data by plotting graphs and relevant charts, we will be able to understand visually what our data is communicating.

This multifaceted exploration not only laid the foundation for subsequent modeling but also provided valuable insights that will shape the effectiveness of our movie recommendation system.

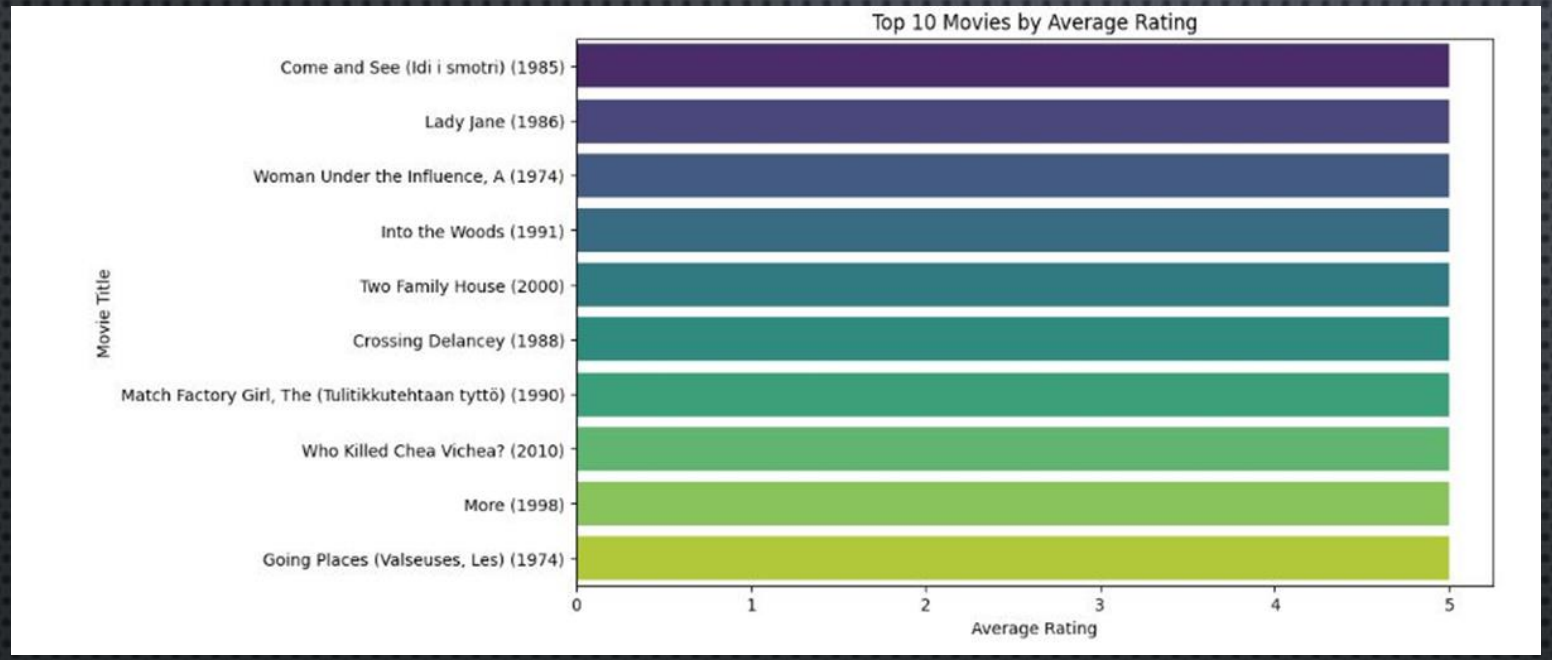
# UNIVARIATE ANALYSIS

## Rating Count



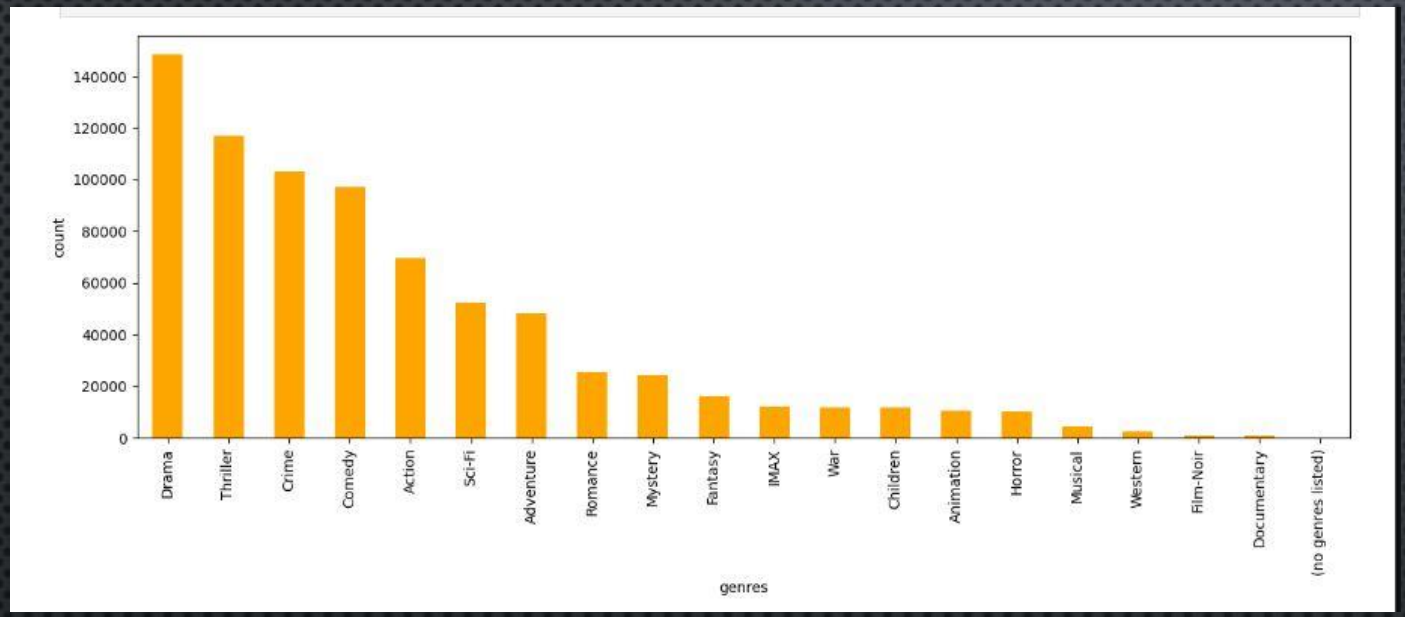

- Our analysis of movie ratings reveals a scale ranging from 0.5 as the lowest to 5.0 as the highest.
- Intriguingly, ratings of 4.0 and 5.0 emerge as the predominant choices among users, boasting the highest count.
- This observation suggests a trend towards positive sentiment in user reviews, indicating a general inclination towards appreciating movies, as reflected in the higher spectrum of the rating scale.
- Understanding the prevalent ratings provides valuable context for our recommendation system, emphasizing the need to capture and leverage positive user sentiments in crafting personalized movie suggestions.

# Top 10 movies by Average Rating



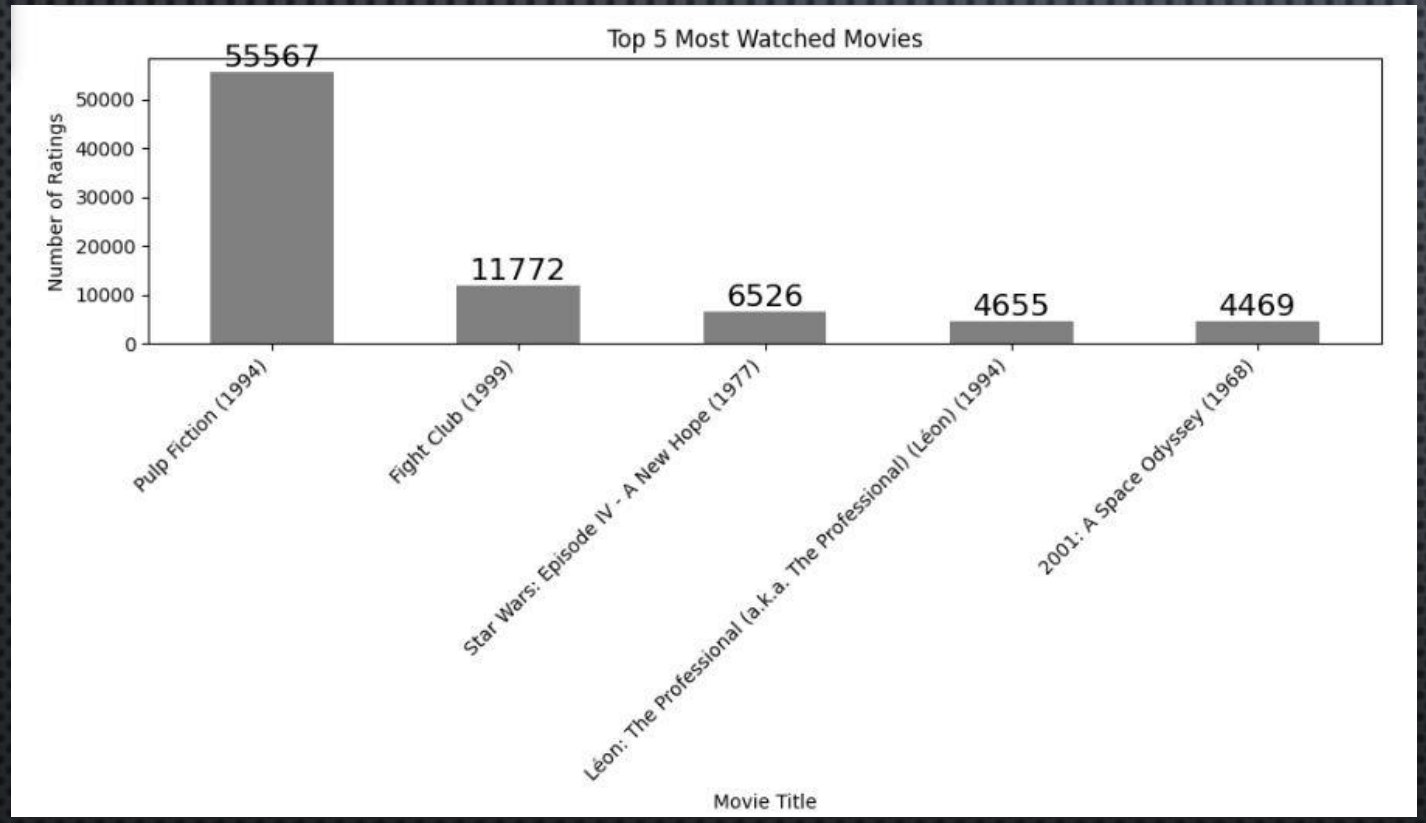

- This Highlights the top 10 movies with the highest average ratings and provides a visual representation of user preferences for the movies.

# Genre Distribution




- The bar plot visualizes the count of each unique movie genre in the dataset.
- The most prevalent genre appears to be Drama, followed by Comedy and Thriller.
- Conversely, the genres Documentary and Film-Noir have lower counts, indicating that they are less represented in the dataset compared to other genres.
- This suggests that the dataset is skewed towards genres like Drama, Comedy, and Thriller, while genres like Documentary and Film-Noir have fewer instances.

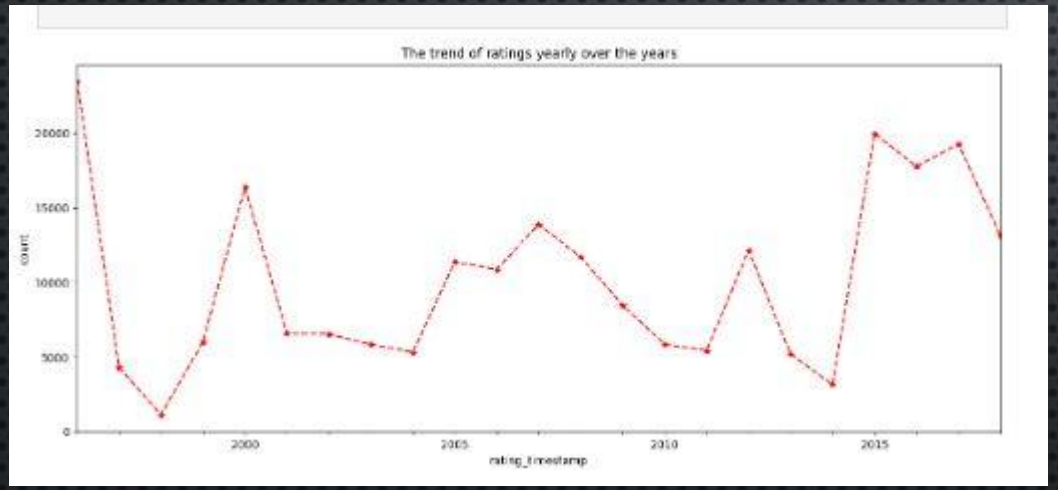
# Top 5 Most Watched Movies




- We make observations that highlight the popularity of certain iconic movies, with "Pulp Fiction" leading the list, followed by other well-known titles like "Fight Club" and "Star Wars Episode 4.“
- The data reflects the diverse preferences and interests of users within the Movie Lens platform.

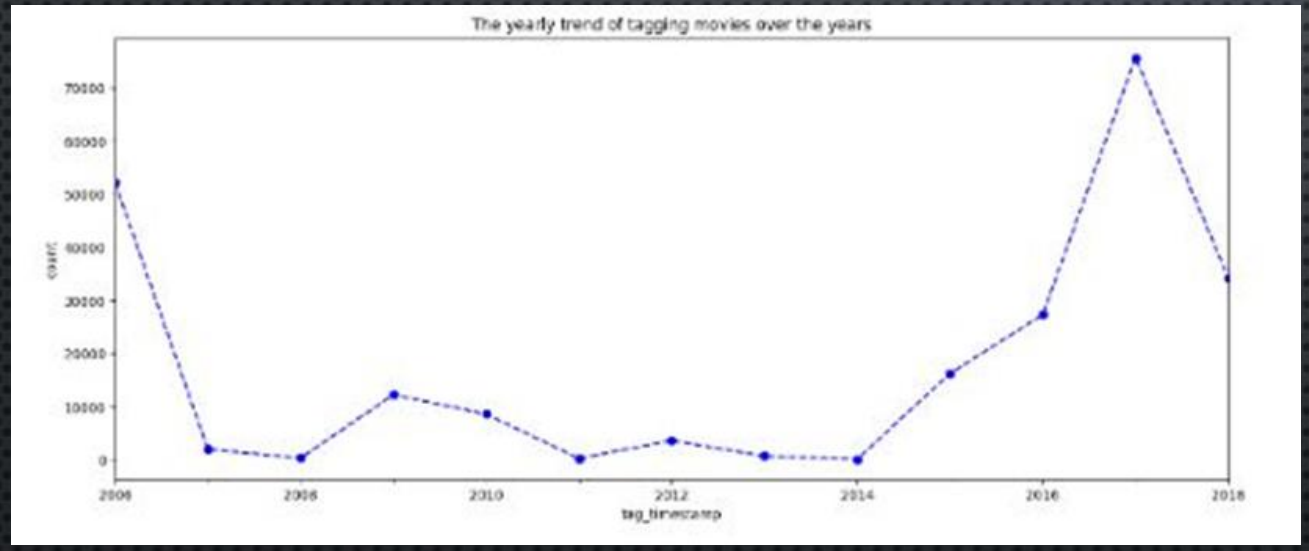
# BIVARIATE EXPLORATORY ANALYSIS

The trend of ratings yearly over the years




- Based on the graph, we can observe that the yearly ratings have been fluctuating over the years with significant peaks and troughs.
- The graph shows that the ratings have been variable each year.
- There are noticeable peaks around the years 2000, 2005, 2010, and 2015.
- There are also noticeable troughs or low points in ratings around the years 2002, 2007, and 2012.
- The fluctuations in yearly ratings likely stem from a combination of factors, including the quality and reception of released movies, changes in audience tastes, and external influences impacting user evaluations of films.

## The yearly trend of tagging movies over the years




- We can observe that the yearly trend of tagging movies over the years has been fluctuating.
- The graph shows that movie tagging was relatively low around 2000, increased slightly until 2010, and then dropped again.
- There was a significant spike in movie tagging around 2016 before it sharply declined in 2018.
- Fluctuations in tags could be due to changing genre popularity, shifts in marketing strategies, alterations in viewing behavior, and adjustments in tagging practices.

# MODELLING

Let's analyze the performance metrics of the different models we opted for :

- **Baseline SVD Model:**

RMSE: 0.8750, MAE: 0.6737
Tuned SVD Model: RMSE: 0.8747, MAE: 0.6752

- **KNN with Means: Vanilla**:

RMSE: 0.8914, MAE: 0.6812
Tuned (Fold 3): RMSE: 0.8953, MAE: 0.6870

- **Tensor Flow Recommender:**

RMSE: 0.7523, Total Loss: 0.5659

Lower values of RMSE and MAE indicate better model performance. The TensorFlow Recommender has the lowest RMSE (0.7523) and Total Loss (0.5659), suggesting superior predictive accuracy compared to the other models.
The Baseline SVD and Tuned SVD models perform similarly, with marginal improvements observed in the tuned version. However, the TensorFlow Recommender outperforms both in terms of RMSE.
KNN with Means, both vanilla and tuned, has higher RMSE and MAE compared to the SVD models and the TensorFlow Recommender, indicating lower accuracy in predictions.

# FINAL MODEL

Among the evaluated models, the TensorFlow Recommender stands out as the top-performing choice, boasting an impressive RMSE of 0.7523 and a Total Loss of 0.5659.

These metrics collectively indicate the model's superior predictive accuracy, with lower RMSE signifying closer alignment between predicted and actual ratings.

The inclusion of Total Loss, considering both rating predictions and item embeddings, provides a more comprehensive evaluation, highlighting the model's adeptness in capturing intricate patterns within the dataset.

The TensorFlow Recommender's stellar performance across multiple metrics underscores its robustness and reliability in predicting user preferences.

Its potential for generalization to unseen data positions it as a strong candidate for deployment in real-world scenarios, where accurate recommendations for new user-item interactions are paramount

# BEST MODEL

The best-performing model was meticulously chosen based on its robust performance on unseen data, minimizing the risk of false positive predictions when deployed in real-world scenarios. This model excelled not only in accuracy but also demonstrated resilience in handling unforeseen data patterns, ensuring its generalizability. The accuracy scores obtained from comprehensive evaluations of different algorithms and hyperparameter configurations provide a quantitative basis for comparing model candidates.

## *Why do we use Machine Learning Models*

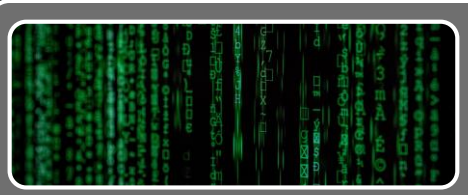
- Robust Predictive Capabilities
- Unraveling Intricate Patterns
- Efficient Scaling Capabilities
- Adaptability & Flexibility

# CHALLENGES IN THE PROJECT

Limited Computational Resources - The project's progress and exploration of algorithms were hindered by the constraint of low computational power, impacting the efficiency of the development and evaluation phases.


**Sparse Data Impact:** Sparse data matrices presented challenges in effectively capturing user preferences, highlighting the need for robust techniques to handle missing interactions.

**Addressing Cold Start:** Strategies like content-based recommendations and hybrid models were successfully employed to mitigate challenges associated with the cold start problem for new users or items.



**Metrics Selection Importance:** The project emphasized the importance of carefully selecting evaluation metrics, recognizing the limitations of standard metrics and considering domain-specific alternatives.

**Interpretability Considerations:** Balancing the accuracy of recommendation models with interpretability emerged as a key consideration, acknowledging the challenge of explaining complex models to end-users or stakeholders.

# CONCLUSION

- Developing robust recommender systems offers a pivotal opportunity for enhancing user experience and engagement across various domains.
- Leveraging predictive models within recommendation algorithms allows platforms to proactively tailor content or suggestions, optimizing user satisfaction and maximizing interaction.
- By utilizing historical user preferences, item characteristics, and feedback data, these models empower stakeholders to anticipate user needs and prioritize personalized recommendations.
- The predictive capability of these models, combined with ongoing user feedback and iterative improvements, establishes a comprehensive approach to recommendation system management.
- This approach enhances the quality of recommendations and ensures sustained user satisfaction in diverse contexts.

# RECOMMENDATIONS

1. **Regular Data collection**: Ensure continuous data collection and updates is done on a daily basis in order to improve and increase model accuracy.
2. **Handling Sparse Data:** Develop techniques or explore algorithms specifically designed to handle sparse data. Sparse matrices are common in recommendation systems, and addressing this challenge can lead to more accurate predictions.
3. **User Engagement:** Continuously gather user feedback to refine and improve the recommendation system. Implement an iterative process of model updates based on user preferences and changing trends.
4. Collaboration with Content Providers: Collaborate with content providers and studios to incorporate real-time information about new releases, events, or trends. This can enhance the system's ability to recommend the latest and most relevant content.
5. Data Enhancement: Explore options for enriching the dataset with additional features such as user demographics, movie genres, or contextual information. This can potentially improve the accuracy and personalization of recommendations.

# NEXT STEPS

**Deployment and User Testing:**

- Deploy the developed recommender system in a real-world business environment or on a platform accessible to users. Monitor its performance and gather user feedback to assess its effectiveness in a practical setting.

**Scalability and Efficiency:**

- Invest in upgrading computational resources to enhance the scalability and efficiency of the recommendation system. This can involve optimizing algorithms, parallel processing, or exploring cloud-based solutions to handle larger datasets and user interactions.

**A/B Testing:**

- conduct A/B testing to compare the performance of the deployed recommender system with alternative models or configurations.

**User Education and Communication:**

- Develop user education materials and communication strategies to explain how the recommendation system works. Increase user understanding and trust by providing insights into the rationale behind recommendations.