# Project Title:

# Building a Recommendation System Using Collaborative Filtering and Content-Based

# Outline of the Presentation

- Business Understanding
    1. Overview of the Project
    2. Problem statement
    3. Objectives of the project
    4. Business challenge and proposed solution
- Data Understanding
- Data Preparation- Setting up, Importing Libraries and Utility Functions and Loading, data Cleaning
- Data Exploration
- Building a Recommendation System Model:
    - ✓ Collaborative Filtering Recommender
    - ✓ Content-Based Recommender
- Comparison Analysis
- Recommendations and Conclusion
- Next Steps

# Problem Statement:

- With the increasing competition in the streaming industry and the exponential growth of video-on-demand platforms across Africa, Showmax is looking to enhance user engagement and retention by offering a personalized movie recommendation system. Showmax currently caters to a diverse audience across 44 African countries with a wide range of local and international content, including series, movies, and documentaries. However, as the content library continues to grow, users often face challenges in discovering relevant content tailored to their unique preferences, resulting in a suboptimal user experience.

- Showmax Films faces the challenge of keeping users engaged on the platform for longer periods. Many users may experience decision fatigue when faced with an overwhelming number of movie options. This can lead to dissatisfaction and even subscription cancellations. Showmax Films needs advanced recommender system to provide personalized, relevant movie suggestions based on users' past ratings and preferences, thus enhancing the overall user experience and reducing defection.

# Stakeholder Understanding

- The Key Stakeholders interest in this project include:
- **1. Showmax Executive Team (Business Leaders & Decision Makers)**

  **- Interest:** Improving customer retention, increasing revenue, and staying competitive in the streaming market.

  **- Expectation:** A recommendation system that enhances user experience, boosts watch time, and maximizes subscription renewals.

- **2. Marketing & Content Strategy Team**

  **-Interest:** Understanding user preferences to tailor content promotion and licensing decisions.

  **- Expectation:** Insights from the recommendation system to guide content acquisition, targeted promotions, and personalized marketing campaigns.

- **3. Showmax Users (Subscribers)**

  **- Interest:** Finding movies they enjoy quickly without spending too much time searching.

  **- Expectation:** A seamless and personalized recommendation experience that enhances their viewing satisfaction.

# Data Understanding: Data Insights

- The MovieLens dataset used in building this movie recommender system was sourced from GroupLens Research Lab at the University of Minnesota being data collected over various period of time.
- The MovieLens Zipped Dataset has Small: 100,000 ratings and 3,600 tag applications applied to 9,000 movies by 600 users. Last updated 9/2018.
- The MovieLens dataset has to csv files of interest i.e "Movies.csv" and "ratings.csv"

✓ Movies.csv file:

- Contains 9742 rows and 3 columns
- Columns include: movieId, title and genres

✓ Ratings.csv file:

- Contains 100836 rows and 4 columns
- Columns include: UserId, movieId, rating and timestamp

# Data Preparation:

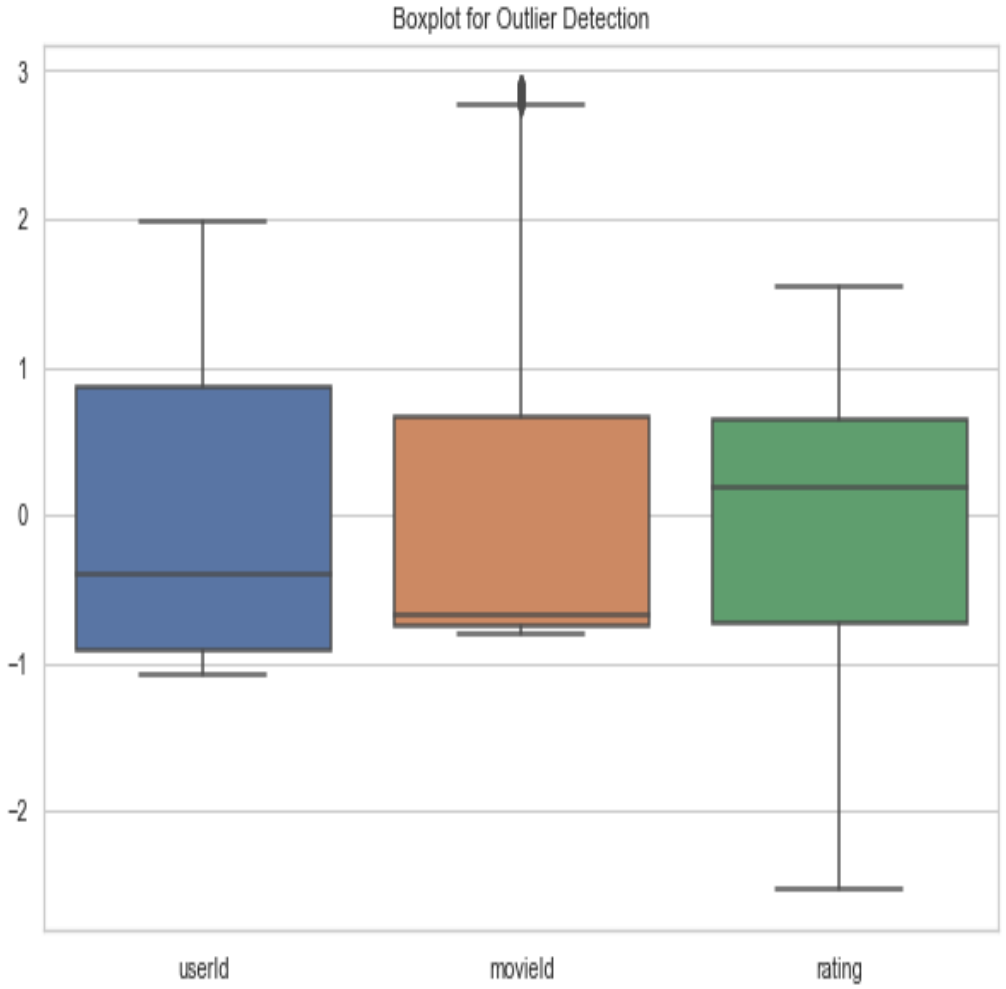
- Except for ratings which is a **float** dtype all other features of the dataset are **int** or **object** dtypes

- No missing values in both "Movie.csv" and "ratings.csv"

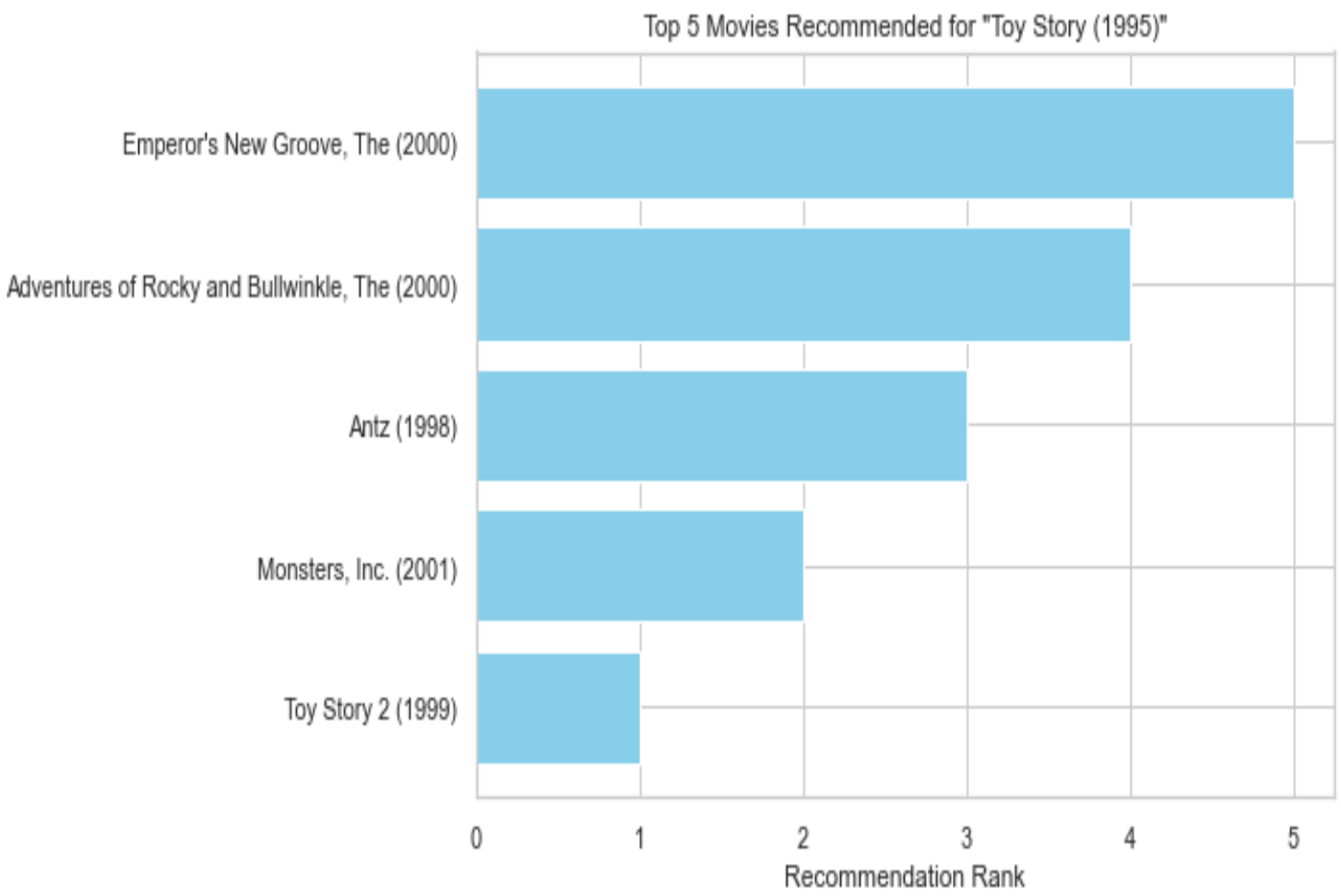
- However, 91112 duplicate values were identified and further dropped
  - The duplicate values were dropped improve model performance since presence of duplicate values reduces model accuracy by making predictions on unseen data in overfitting in machine learning
- To build an effective recommendation system, we need to combine both user interactions (ratings) and movie details (titles, genres

# Data Visualization: Checking For Outliers

- **Explanation of the Boxplot Visualization:**
  - ❑ There are few outliers in the rating columns (indacting potential of very few anomalies) compared to the general distribution of ratings.
  - ❑ While “userId” and “movieId” column shows possibility of many outliers in the upper whisker of the box plot



# Data Visualization: Top 5 Recommended Movies Using Recommender System




- **Explanation of Horizontal Bar chart Visualization:**
  - ❑ Using the recommendation system Model developed:
    - ✓ Emperor's New Groove,The (2000) shows the highest rating of 5, follow by Adventues of Rocky and Bullwinkle, The(2000)

# Modelling:

- Given the project was to build a Movie Recommendation System For ShowMax Films ussing MovieLens Dataset, the following steps were followed:
    1. Building a Preprocessing Pipeline, Training a Model using Gradient Boosting Regression Technique and Further Tuning of the model using XGBoost
    2. Installing the additional Surprise library, Loading and preprocessing the data set
    3. Collaborative Filtering: Build a model using matrix factorization-based techniques.
    4. Content-Based Filtering: Handle the cold-start problem using movie features.
    5. Model Performance Evaluation and Interpretation

# Model Evaluation

# Model Performance Evaluation : XGBoost Model

- **Explanation of the XGBoost Model Performance Metrics:-**
  - **Mean Squared Error = 0.83**: On average, the squared prediction error is 0.83 units, suggesting that there is room for improvement in the model's predictions.
  - **Mean Absolute Error = 0.70** : The average magnitude of prediction errors is 0.70 units. This is the average distance between predicted and actual values.
  - **$R^2$ Score = 0.26:** The model explains 26% of the variance in the target variable, indicating that it captures some patterns but still misses a significant portion of variability.
    - **Observations:**
      - There is an improvement in $R^2$ Score from 0.17 to 0.26

# Model Performance Evaluation: Collaborative Filtering

- **Evaluating RMSE, MAE of algorithm SVD on 5 split(s).**

| | Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 | Mean | Std |
|---|---|---|---|---|---|---|---|
| RMSE (testset) | 0.9521 | 0.9255 | 0.9517 | 0.9377 | 0.9751 | 0.9484 | 0.0166 |
| MAE (testset) | 0.7386 | 0.7258 | 0.7444 | 0.7400 | 0.7660 | 0.7430 | 0.0131 |
| Fit time | 1.18 | 1.21 | 1.20 | 0.76 | 1.20 | 1.11 | 0.18 |
| Test time | 0.02 | 0.02 | 0.02 | 0.02 | 0.00 | 0.02 | 0.01 |

Average RMSE: 0.95

Average MAE: 0.74

- **Root Mean Squared Error (RMSE):**
  - Measures the average squared difference between predicted and actual ratings.
  - Smaller RMSE values indicate more accurate predictions.
  - RMSE = 0.95 means the average prediction error is 0.95 stars.
- **Mean Absolute Error (MAE):**
  - Measures the average magnitude of prediction errors.
  - MAE is less sensitive to outliers than RMSE.
  - MAE = 0.74 means that, on average, the model's predictions deviate by 0.74 star

# Recommendation :

- **Collaborative Filtering would be fit for Existing Users:**
  - Collaborative filtering works best when a large amount of user-item interaction data is available. It effectively captures user preferences and offers personalized recommendations

- **Content-Based Filtering would be fit for New Users or Items:**
  - Content-based filtering is crucial for addressing the cold-start problem, as it can recommend items based on their features (e.g., genres, keywords) without needing historical ratings.

- **Hybrid Approach for Robustness:**
  - Consider combining both approaches to build a hybrid recommender system:
    - Use collaborative filtering when sufficient interaction data is available.
    - Fallback to content-based filtering for new users or items.

# Next Steps:

- The organization can further improve the model performance using both approaches into a hybrid recommender system, then Deploy the Recommendation System as an API to further improve on recommendations

# Thank You