Netflix Recommendation System Analysis

Project Overview

This project focuses on understanding **how recommendation systems work in real-world platforms like Netflix** by analyzing large-scale user–movie rating data.
The main objective is to **predict user preferences** and recommend relevant movies based on historical behavior.

By studying millions of user interactions, the project demonstrates how data-driven personalization improves user engagement, content discovery, and customer retention.

Dataset Summary

The dataset represents **real-world Netflix-style rating data** collected at scale.

- **Total ratings:** 24M+
- **Users:** Millions of unique customers
- **Movies:** Thousands

Key information included in the dataset:

- **User data:** customer ID representing individual users
- **Movie data:** movie ID representing unique content
- **Interaction data:** explicit ratings (1–5 scale)
- **Behavioral insight:** historical user preferences inferred from ratings

Due to the scale of the dataset, efficient memory handling and preprocessing were essential before analysis and modeling.

Data Preparation & Exploratory Analysis Using Python

The analysis began in **Python**, where the dataset was loaded and reviewed to understand its structure, data types, and distribution.

Key steps performed:

- Converted raw data into a structured tabular format
- Standardized column names for consistency
- Optimized data types to reduce memory usage
- Checked for missing or invalid ratings

- Analyzed rating distribution and user activity patterns

Exploratory analysis revealed:

- High sparsity in user–movie interactions
- Most users rated only a limited number of movies
- Certain movies received significantly more ratings than others

These characteristics confirmed that **collaborative filtering** was the most suitable approach.

Data Preparation for Machine Learning

Before model training, the cleaned dataset was transformed into a format compatible with machine learning algorithms.

Steps included:

- Defining the rating scale (1–5)
- Converting data into Surprise library format
- Splitting data into training and test sets
- Ensuring scalability for large-volume learning

This step ensured that the model could efficiently learn from historical user behavior.

Recommendation Model Development

A **Collaborative Filtering model using Matrix Factorization (SVD)** was implemented.

Why this approach?

- Effectively handles sparse datasets
- Learns hidden (latent) user preferences
- Widely used in real-world recommender systems

The model learns:

- Which users have similar tastes
- Which movies share similar audience preferences
- How strongly a user is likely to rate an unseen movie

Model Training & Evaluation

The SVD model was trained on millions of ratings to learn user–movie interaction patterns.

Evaluation:

- Model performance was evaluated using **RMSE (Root Mean Squared Error)**
- Predicted ratings were compared with actual user ratings
- Stable prediction accuracy confirmed the model's reliability

This step ensured that recommendations were data-driven and statistically meaningful.

Recommendation Generation

Once trained, the model was used to:

- Predict ratings for movies a user had not watched
- Rank movies based on predicted preference score
- Generate **personalized top-N recommendations**

The final output simulates how Netflix presents tailored movie suggestions to each user.

BELOW ARE THE TWO USERS TOP 5 AND BOTTOM RECOMMENDATION

```
user_2031561.head(5)
```

	Movie_Id	title	genres	Estimated_Score
27	28	Persuasion (1995)	Drama|Romance	3.978231
17	18	Four Rooms (1995)	Comedy	3.888712
24	25	Leaving Las Vegas (1995)	Drama|Romance	3.878116
4	5	Father of the Bride Part II (1995)	Comedy	3.851733
2	3	Grumpier Old Men (1995)	Comedy|Romance	3.728791

```
user_2031561.tail(4)
```

	Movie_Id	title	genres
27274	131256	Feuer, Eis & Dosenbier (2002)	Comedy
27275	131258	The Pirates (2014)	Adventure
27276	131260	Rentun Ruusu (2001)	(no genres listed)
27277	131262	Innocence (2014)	Adventure|Fantasy|Horror

```
user_823628.head(5)
```

	Movie_Id	title	genres	Estimated_Score
17	18	Four Rooms (1995)	Comedy	4.061956
24	25	Leaving Las Vegas (1995)	Drama|Romance	4.041365
27	28	Persuasion (1995)	Drama|Romance	3.922843
7	8	Tom and Huck (1995)	Adventure|Children	3.826373
4	5	Father of the Bride Part II (1995)	Comedy	3.677769

```
user_823628.tail(5)
```

	Movie_Id	title	genres	Estimated_Score
5	6	Heat (1995)	Action|Crime|Thriller	3.513983
23	24	Powder (1995)	Drama|Sci-Fi	3.500218
15	16	Casino (1995)	Crime|Drama	3.351807
16	17	Sense and Sensibility (1995)	Drama|Romance	3.337929
25	26	Othello (1995)	Drama	3.246668

Key Insights

- User behavior patterns are strong predictors of future preferences
- Similar users tend to consume similar content over time
- Collaborative filtering scales efficiently for large platforms
- Accurate recommendations do not require detailed movie metadata
- Data preprocessing and optimization are critical at scale

Business Impact

If deployed in a real-world environment, this system could:

- Increase user engagement and watch time
- Improve recommendation relevance
- Enhance customer satisfaction
- Reduce churn through personalization

Business Recommendations

Based on the project findings, the following recommendations can be made:

- **Invest in Personalization:**
  Recommendation engines significantly enhance user experience and retention.
- **Leverage User Behavior Data:**
  Historical interaction data is more valuable than explicit content descriptions.
- **Optimize Data Pipelines:**
  Large-scale systems require efficient preprocessing and memory management.
- **Continuously Improve Models:**
  Regular retraining helps adapt to changing user preferences.

Tools & Technologies Used

Python, Pandas, NumPy, scikit-surprise, Jupyter Notebook

Conclusion

This project demonstrates an **end-to-end recommendation system** built on real-world scale data, covering:

- Business understanding
- Data preprocessing
- Machine learning modeling
- Evaluation and recommendation generation

It showcases how **raw user data is transformed into intelligent, personalized decisions** — the foundation of modern digital platforms.