Building a Movie Recommendation System to Decrease User Churn and Improve Engagement

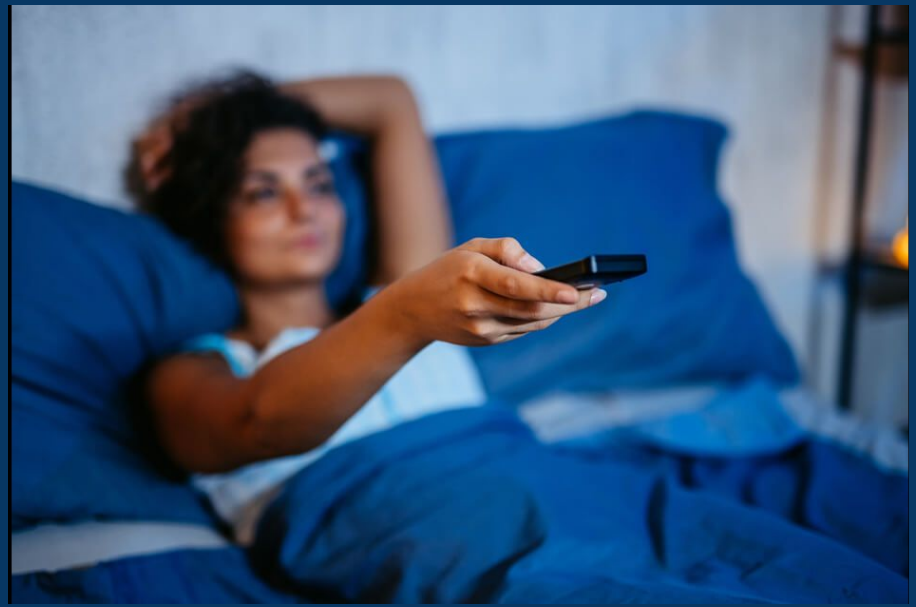
•••

Sarah Berkin

Data Science Capstone Project, Sept 2024 Cohort

The Problem

Over 110 hours per year scrolling through streaming platforms



52% of users say a streaming service's browsing experience significantly influences their decision to subscribe

56% cancel subscriptions after completing a single show

Data Information for this Analysis

Data Source

The dataset originates from the GroupLens research group at the Dept of Computer Sci. & Eng. at the University of Minnesota

Sample Population

1,000,209 anonymous movie ratings

6,040 MovieLens users who joined in 2000

3,900 movies

Variables

Target variable: User rating on scale of 0.5-5

Ind. variables include User features (gender, age group, occupation) & Movie features (genres, release year)

Data Cleaning

Mixed text in Title field

Extracted release year from movie title using regular expressions & created separate year feature

Genres field contained multiple genres in single string

Applied one-hot encoding to expand genres into individual binary columns

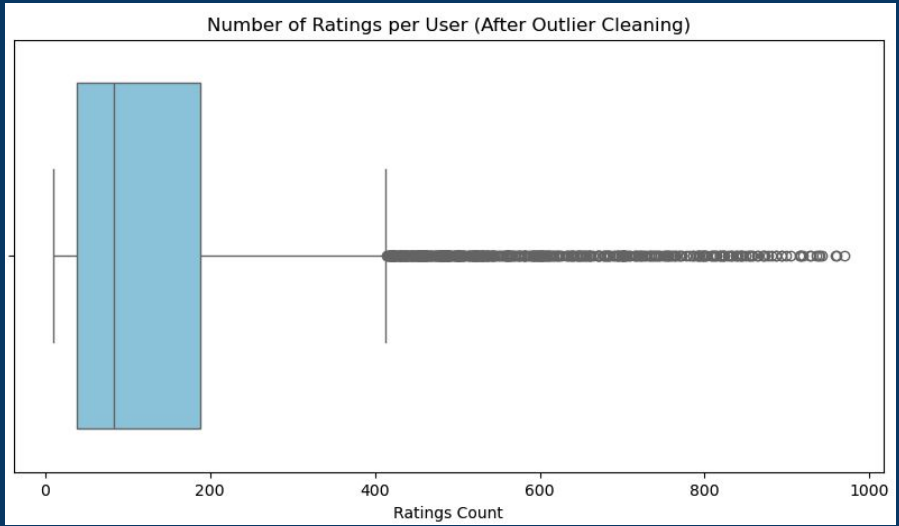
User demographics (gender, occupation) was categorical

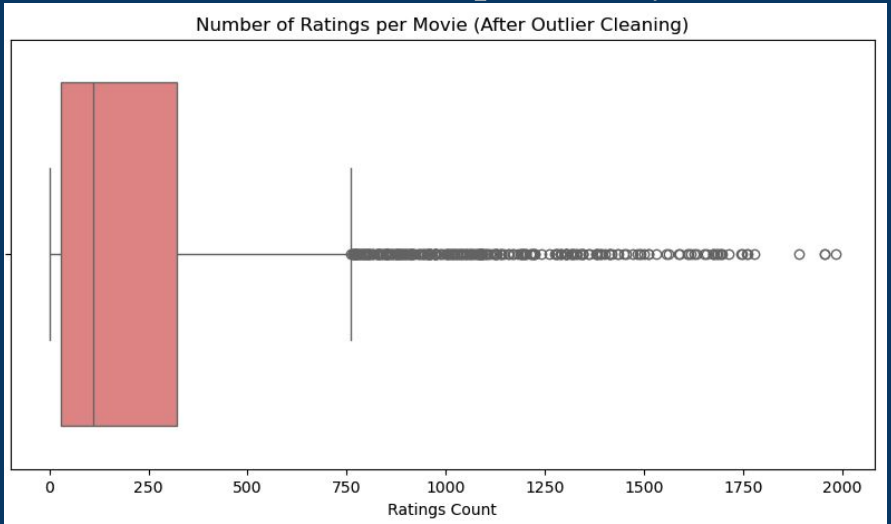
Applied one-hot encoding & created a bucketed age group feature

Data Cleaning

Boxplots detected rating count outliers (super-users & blockbuster movies)

Trimmed most extreme outliers based on IQR method & boxplot analysis

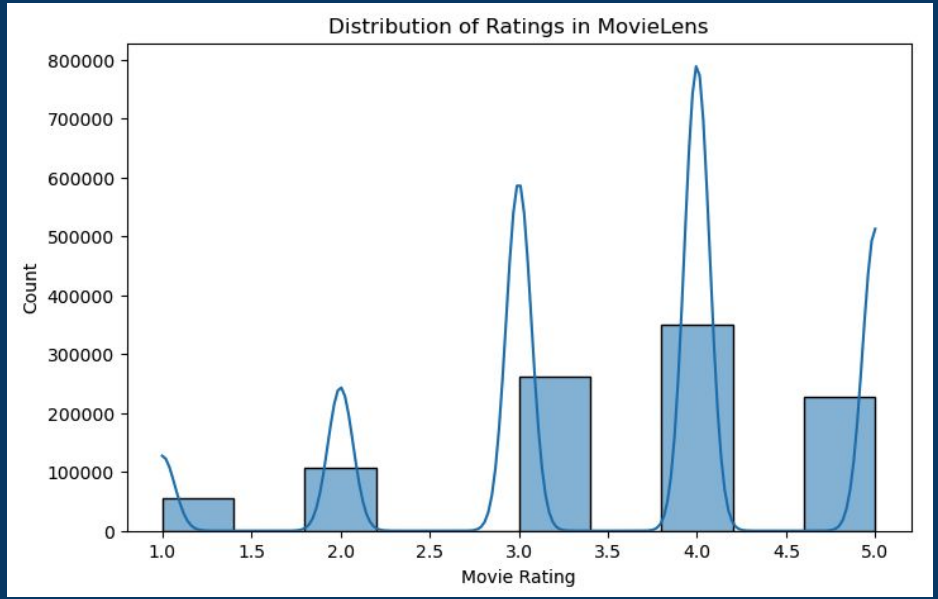




Exploratory Data Analysis

Visualizing Ratings Distribution

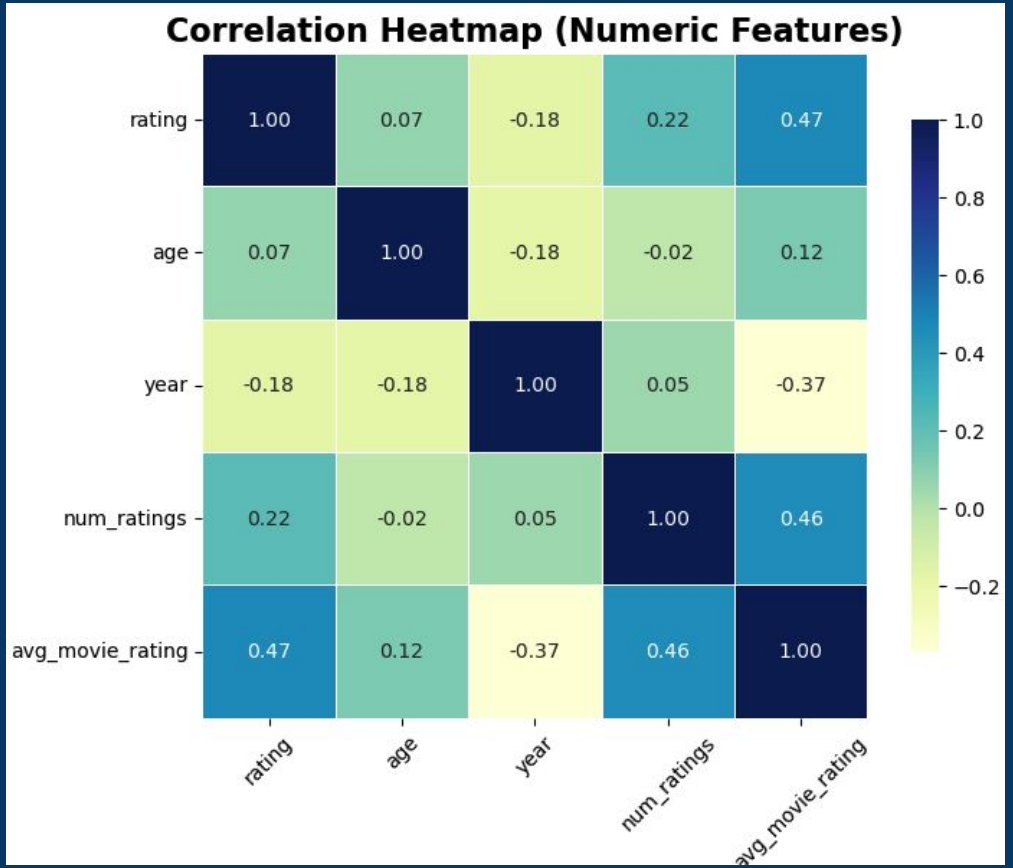
If there's an extreme on either side (ex. ~90% above or below 3.5), then the dataset may be imbalanced. Results showed 57% of ratings were 3.5 or higher.



Exploratory Data Analysis

Visualizing Correlations between the Variables

Heatmap identified weak relationships between age, movie release, year, & ratings.



Exploratory Data Analysis

Does it appear that low ratings are a proxy for scrolling frustration?

Checked how many low (1-2) scores existed. Behavior patterns suggested moderate scroll frustration in raw data.

16.37% Low Scores

Feature Engineering & Pre-Processing

One-Hot Encoding

Prepared User features for ML models by converting into binary

Data Splitting

Train Test Split method

80% Train / 20% Test

Feature Standardization

Applied StandardScaler to standardize numeric & dummy variables

Modeling

I used the Surprise, Scikit-learn, and Matplotlib libraries for training and visualizing my recommendation system

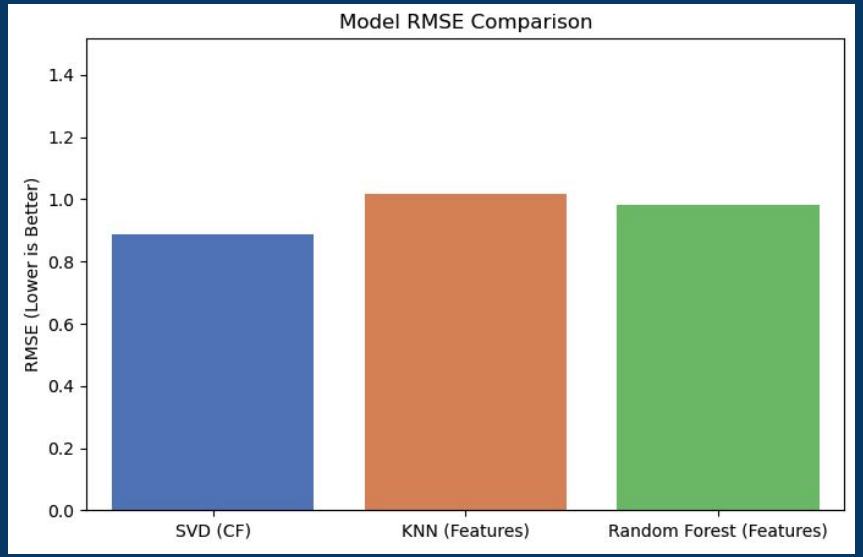
Model	Description
SVD (Collaborative Filtering)	Matrix factorization using userId and movieId
KNN Regressor	Content + demographic feature-based distance model
Random Forest Regressor	Ensemble model using full feature set
Train-test split: 80/20	

Modeling

Behavior simulation showed that without recommendations, users frequently rated multiple movies poorly before finding one they liked

With the SVD model, the top 10 recommendations included 10/10 highly-rated movies, showing significant improvement in browsing experience.

Model	RMSE (Lower = Better)
SVD (Collaborative Filtering)	0.8869 (Best)
KNN Regressor	1.0165
Random Forest Regressor	0.9813



Recommended Business Use Cases

Improved content discovery	Reduced subscription churn by personalizing homepage feed to keep users engaged
Personalized watchlists & Push notifications	Get in front of scrolling problem by serving up high-confidence recommendations
Targeted upsell based on User preferences	Upselling premium content aligned with user taste

Future Work

Quantify scroll time reduction	Scroll behavior tracking to more precisely measure impact on reducing user browsing
Explore hybrid models	Get in front of “cold start” problem for new users by integrating content-based recommendations
Incorporate Implicit Feedback	Add features like watch time partial watches, and browsing behavior
Expand user feature set	Integrate additional user demographic data like location or watch history patterns
Conduct A/B Testing	Evaluate real-world engagement improvements & quantify the model’s effect on churn

The Team

Project by:

Sarah Berkin

Data Science Trainee

Special thanks to:

Jaleed Khan, PhD

Data Science Mentor
Sr. Researcher
Honorary Research Fellow
Lecturer