Recommender System

2024

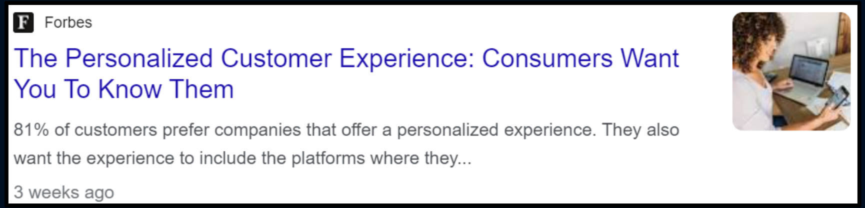
Agenda

- **What & Why**
- **Exploratory Data Analysis**
  - **Popularity & Self-Selection Biases**
- **Matrix Completion**
- **Model Testing, Model Evaluation & Model Building**
  - **Train vs Validation Performance**
  - **Test Performance**

Recommender System (RecSys)

WHAT

RecSys is an area of Machine Learning that analyzes **user behavior** and **item characteristics** to predict and present the most relevant items to individual users.



Forbes

The Personalized Customer Experience: Consumers Want You To Know Them

81% of customers prefer companies that offer a personalized experience. They also want the experience to include the platforms where they...

3 weeks ago

WHY

Businesses use RecSys to achieve **personalization at scale**, enhancing CX by accurately suggesting products/services, which **increases loyalty and improves conversion rates**.

Business Insider

Netflix recommendation engine worth $1 billion per year

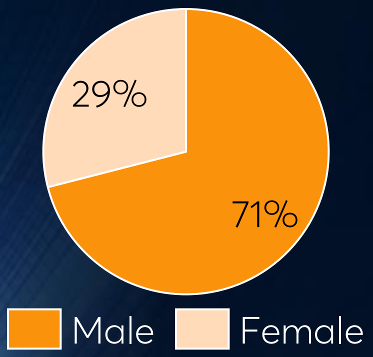
After a long refinement process, Netflix finally released its first "global" recommendation engine in December.

Jun 14, 2016

Exploratory Data Analysis

943 users

Gender Distribution

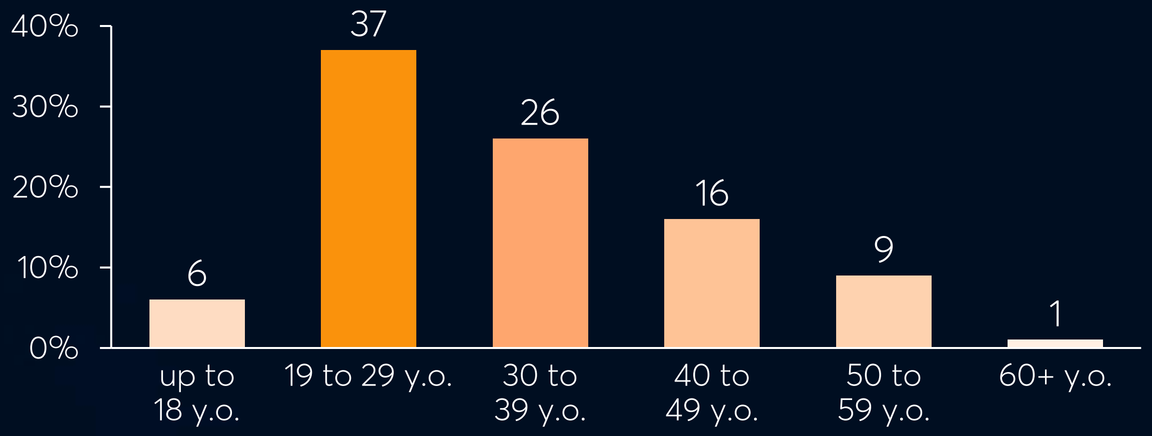


Occupations Top-5

Occupation	Share
Student	21%
Educator	10%
Administrator	8%
Engineer	7%
Programmer	7%

Note : Other is an occupation with an 11% share, but it was decided to not include it due to its descriptive effect.

Age Distribution



Exploratory Data Analysis

1682 movies

Movie Genres Top-5

Genre	Share
Drama	43%
Comedy	30%
Action	15%
Thriller	15%
Romance	15%

Note : A movie can belong to multiple genres.

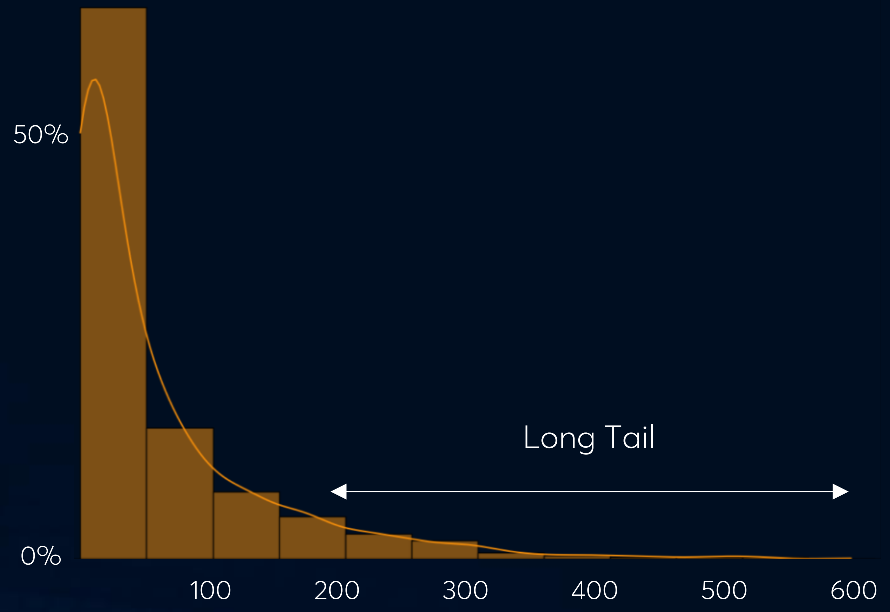
Ratings Summary

The distribution of ratings among items often satisfies a property in real-world settings, which is referred to as the **long-tail** property.

- More than 50% of the movies were rated up to 50 times.
- On average, each movie was rated 59 times.

According to this property, only a small fraction of the items are rated frequently – popular items.
This results in a **highly skewed** distribution of the underlying ratings.

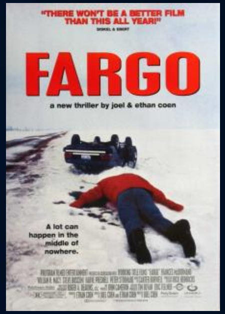
Ratings Distribution across Movies



Exploratory Data Analysis

Most Popular Movies Top-3

Star Wars (1977)

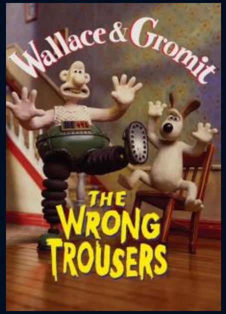


Fargo (1996)

Contact (1997)

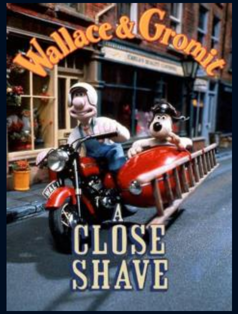
views	Star Wars (1977)	Fargo (1996)	Contact (1997)
	583	509	508

Hottest Movies Top-3



The Wrong Trousers (1997)

Schindler's List (1993)



A Close Save (1995)

★★☆ average rating	The Wrong Trousers (1997)	Schindler's List (1993)	A Close Save (1995)
	4.49	4.46	4.46

Note : The top-3 hottest movies were selected according to a threshold based on the number of views(/ratings). This prevents that movies with a couple of views and high ratings are not eligible for this list.

Exploratory Data Analysis

Self Selection Bias

Not only RecSys are prone to popularity bias, as shown before, but also **selection bias**. This translates as

Users generally rate movies they have chosen to watch, and often, they select movies they anticipate they will enjoy based on genres, actors, directors, or past experiences.

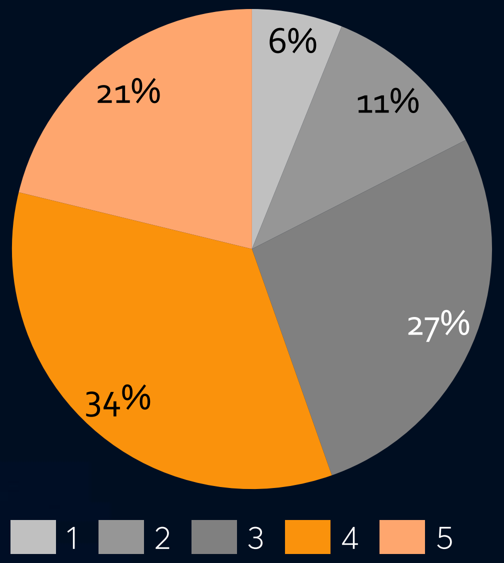
The **average rating**, of the 100k ratings, is **3.52** stars. Additionally

+3.2 X

- **55%** of the ratings are classified as **4 or 5** stars.
- **17%** of the ratings are classified as **1 or 2 stars**.
- **68%** of users classified ≥50% of their ratings as 4 or 5 stars.

★★☆ average rating **3.52**

Ratings Distribution



Matrix Completion

Given m users and n items, the data can be represented by an incomplete mn matrix. Each entry, r_{ui}, represents the unknown or given rating by user u to the item i.

The goal is to predict the missing values in this user-item interaction matrix, by leveraging the observed entries in the matrix.

By completing the matrix, RecSys can suggest items that a user might like but has not yet interacted with.

CHALLENGES

SPARSITY Sparsely populated matrix w/ a lot of missing values

SCALABILITY As the number of users and items frows, the size of matrix increases, posing computational challenges

DYNAMIC DATA User preferences and items properties change over time, requiring RecSys to adapt and update dynamically

	item A	item B	item C	item D	item E
user 1	5	3		1	
user 2	4			3	2
user 3			5	4	1
user 4	1	2			4
user 5		4	3	2	

Model Testing and Model Building

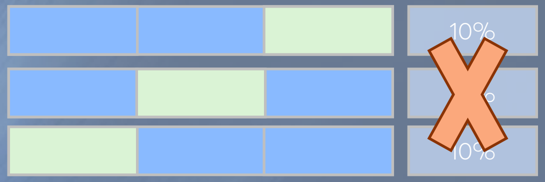
There are **3 main processes** to select and evaluate the methodology, and finally generate results.

Model Testing

GOAL

The first step is to choose which methodology will generate the movie recommendations. During this phase different algorithms with different hyper-parameters are tested.

DATA



~90% of the data is used to test different models. For example, cross-validation with 3 folds is tested.

Model Evaluation

GOAL

Once the methodology is chosen, we then want to estimate its error. This will allow us to understand how the results of the predictive model will generalize to an independent data set.

DATA

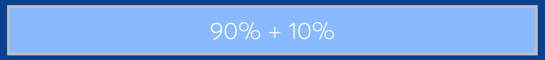
The best model is trained on 90% of data and tested on the remaining 10%. This way, we ensure the model was not optimized for this 10%, giving us then a realistic view of its performance.

Model Building

GOAL

Once we have tested every possible combination of models and evaluated the best one, it's time to generate results. To do this, we will leverage the full potential of the data.

DATA



The model is trained is on the whole dataset and predictions are generated.

Model Testing

In the process of testing different models, it was applied the following algorithms with their default parameters. Matrix Factorization stood out as the best algorithm.

Baseline

Performance

TEST RMSE	1.02
TRAIN RMSE	0.92

Description

The baseline estimator only considers the average rating, the user bias, and the item bias when filling up the rating matrix. Tends to minimize the self selection bias effect.

KNN

Performance

TEST RMSE	1.09
TRAIN RMSE	0.77

Description

Similar users have similar ratings on the same item. The predicted ratings of user *A* are computed as the weighted average ratings of these "peer group".

MF

Performance

TEST RMSE	1.02
TRAIN RMSE	0.67

Description

Dimensionality reduction techniques are used to create a new fully specified representation of the incomplete dataset. A low-rank matrix can capture redundancies in the data.

NMF

Performance

TEST RMSE	1.08
TRAIN RMSE	1.06

Description

The advantage of Non-negative matrix factorization is not necessarily one of accuracy, but one of interpretability which provides an understanding of the user-item interactions.

CoClustering

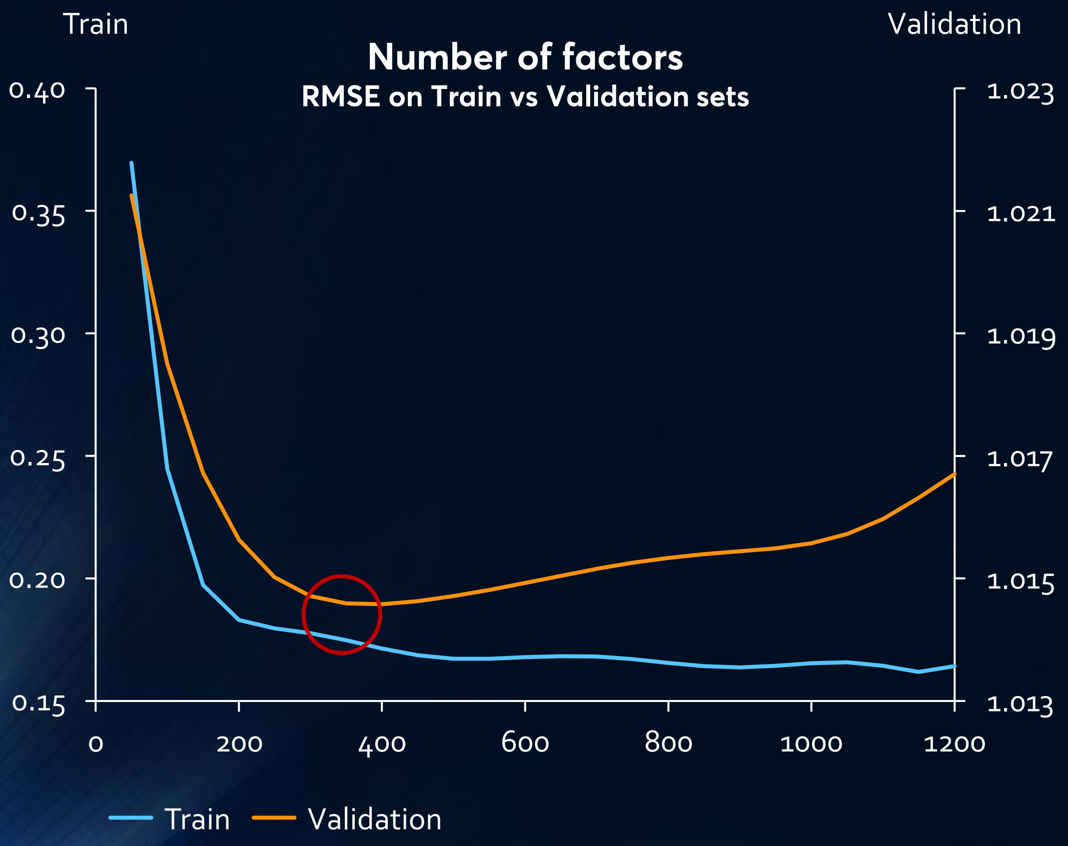
Performance

TEST RMSE	1.09
TRAIN RMSE	0.90

Description

Simultaneously creates user and item groups, unlike KNN-based algorithms. It's prone to struggle with scalability and sparsity issues compared to matrix factorization algorithms.

Train vs Validation Performance



IDEAL NUMBER OF FACTORS

~350

BIAS vs VARIANCE TRADE-OFF

As the complexity of the model increases, the training error decreases – decreasing the bias component of the error.

A model with high variance pays a lot of attention to training data and does not generalize on the data that it hasn't seen before.
To avoid overfitting, the number of factors should be chosen when the validation error starts to Increase on the validation set.