Recommendation System

Zining Wang, Jiayu Yao, Qiyang Li

1. Intro & Background

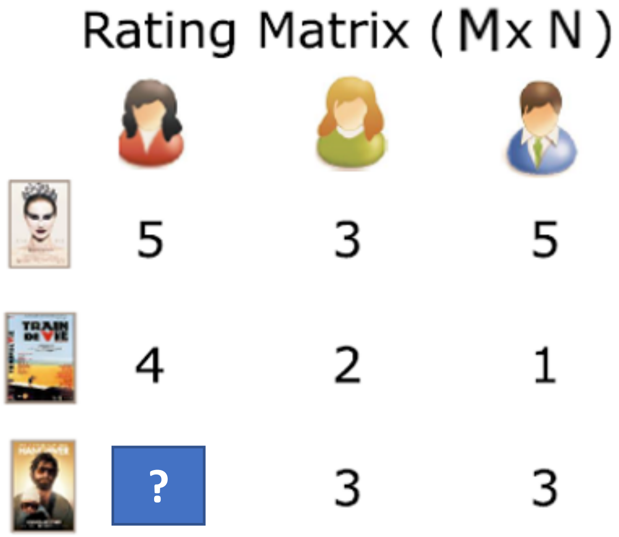
Recommendation System

Problem

- User, Movie → Rating(1-5)
- Given: History ratings dataset
- Given: User & Movie information
- Objective: Predict ratings and reduce error

Applications

- Recommend different movies to different users
- Find top 10 favorite movies
- Personalized recommendations



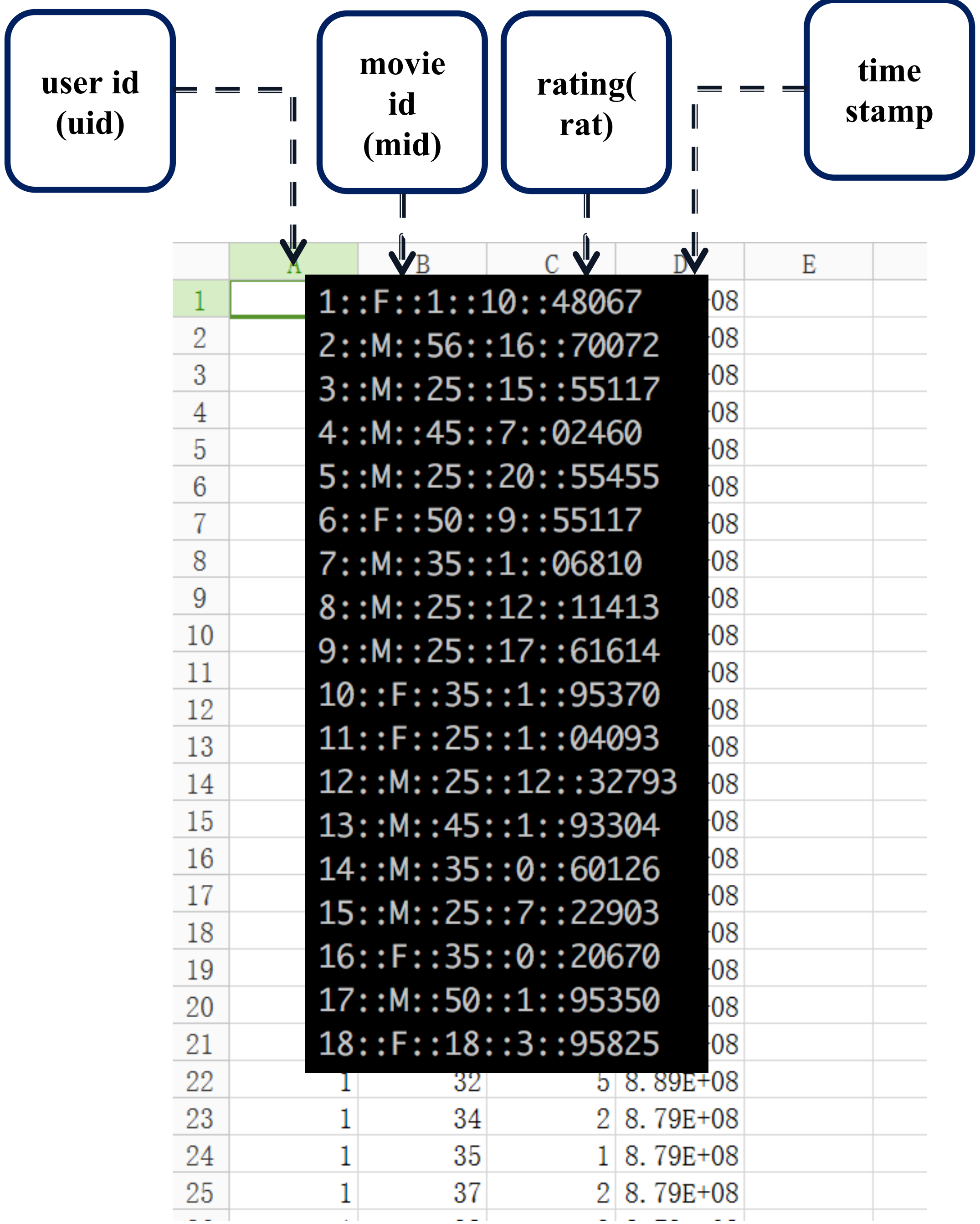
Rating Matrix (M x N)

5	3	5
4	2	1
?	3	3

Dataset

Dataset:

- 1 M dataset by Grouplens
- 6040 users, 3952 movies
- User id, movie id, ratings
- Additional user and movie information



user id (uid)

movie id (mid)

rating(rat)

time stamp

```
1::F::1::10::48067
2::M::56::16::70072
3::M::25::15::55117
4::M::45::7::02460
5::M::25::20::55455
6::F::50::9::55117
7::M::35::1::06810
8::M::25::12::11413
9::M::25::17::61614
10::F::35::1::95370
11::F::25::1::04093
12::M::25::12::32793
13::M::45::1::93304
14::M::35::0::60126
15::M::25::7::22903
16::F::35::0::20670
17::M::50::1::95350
18::F::18::3::95825
```

	A	B	C	D	E
22	1	32	5	8.89E+08	
23	1	34	2	8.79E+08	
24	1	35	1	8.79E+08	
25	1	37	2	8.79E+08	

2. Methods

Recommendation System

Methods:

- Global Average/Movie Average/User Average/Combined Average
- Weighted K Nearest Neighbor (KNN)
- Singular Vector Decomposition (SVD)
- XGBoost
- Ensemble/Hybrid Recommendation System Methodology
- Novel ideas..

Only Using Average...

- Predict every rating by global average/movie average/user average
- Baseline Model

KNN

Item-based KNN

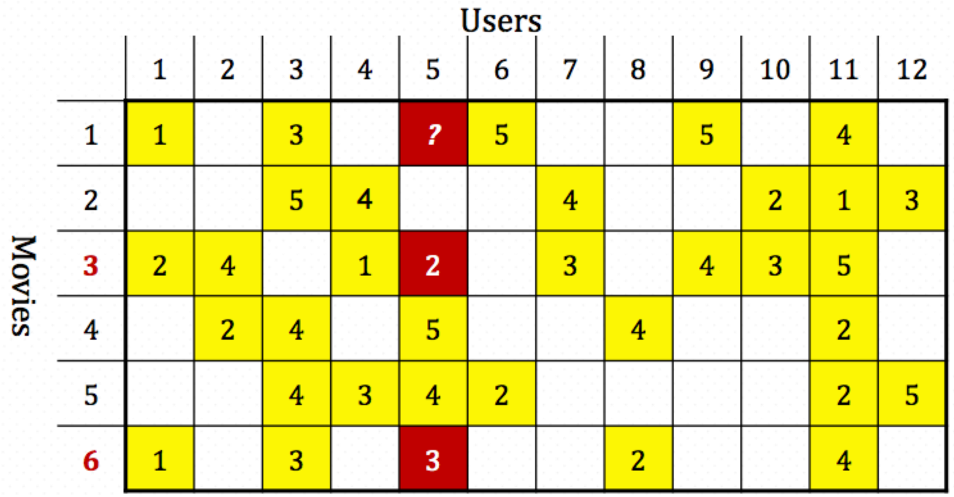
- •Intuition: each individual will rate similar movie similarly
- •Pearson Correlation

$$w_{i,j} = \frac{\sum_{u\in U}(r_{u,i}-\bar{r}_i)(r_{u,j}-\bar{r}_j)}{\sqrt{\sum_{u\in U}(r_{u,i}-\bar{r}_i)^2 \sum_{u\in U}(r_{u,j}-\bar{r}_j)^2}}$$

- Distance Metric - Pearson Distance d=1-w

Weighted KNN

- Take weighted average



Movies	Users												
	1	2	3	4	5	6	7	8	9	10	11	12	
1	1		3		?	5			5		4		k=2
2			5	4			4			2	1	3	
3	2	4		1	2		3		4	3	5		sim=0.4
4		2	4		5			4			2		
5			4	3	4	2					2	5	
6	1		3		3			2			4		sim=0.3

Prediction =(2*0.4+3*0.3)/(0.3+0.4)=2.43

SVD

Simple SVD

- Factorized the movie-user matrix
- Impute missing values with average

SVD + bias

- Incorporate user/movie bias
- Randomly initialize matrices
- Set default user/movie bias vectors
- Use gradient descent to find P,Q and bias
- Update for certain iterations

$$\min_{Q,P} \sum_{(x,i)\in R} \underbrace{\left(r_{xi} - (\mu + b_x + b_i + q_i\, p_x)\right)^2}_{\text{goodness of fit}}$$

$$+\underbrace{\left(\lambda_1 \sum_i \|q_i\|^2 + \lambda_2 \sum_x \|p_x\|^2 + \lambda_3 \sum_x \|b_x\|^2 + \lambda_4 \sum_i \|b_i\|^2\right)}_{\text{regularization}}$$

λ is selected via grid-search on a validation set

$$\min_{P,Q} \sum\nolimits_{(i,x)\in \mathrm{R}} \left(r_{xi} - q_i \cdot p_x\right)^2$$

$$r_{xi} = \underbrace{\mu}_{\text{Overall mean rating}} + \underbrace{b_x}_{\text{Bias for user } \boldsymbol{x}} + \underbrace{b_i}_{\text{Bias for movie } \boldsymbol{i}} + \underbrace{q_i \cdot p_x}_{\text{User-Movie interaction}}$$

Models Using Additional Information

- Additional Information
- User: Gender, Age, Occupation, Zip Code, Timestamp
- Movie: Year, Genre (Action, Drama, Horror, etc.), Number of Rating

KNN

- Distance Metric: Hamming (N_unequal(x, y) / N_tot)

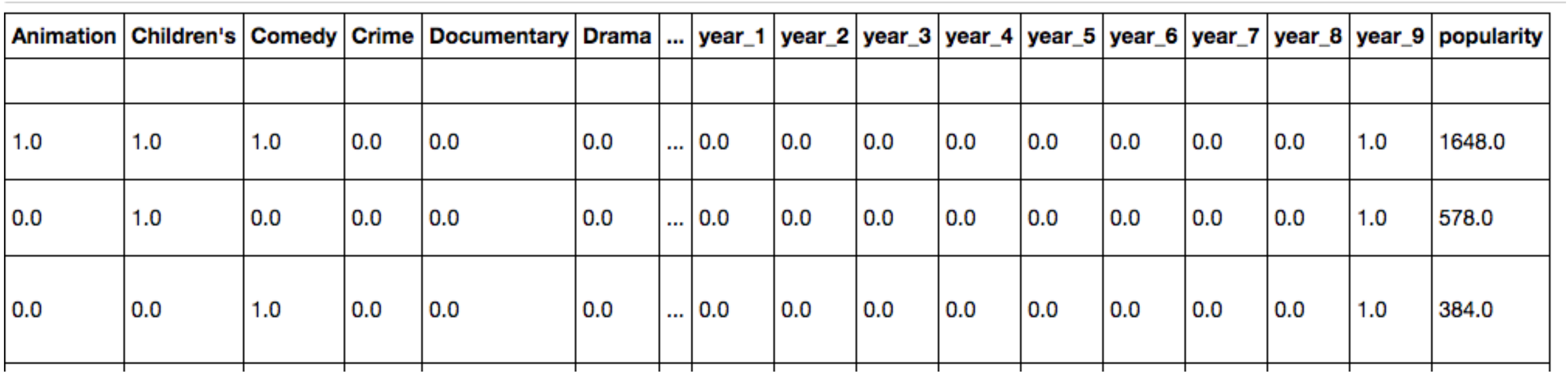
Tree Models

- XGBoost

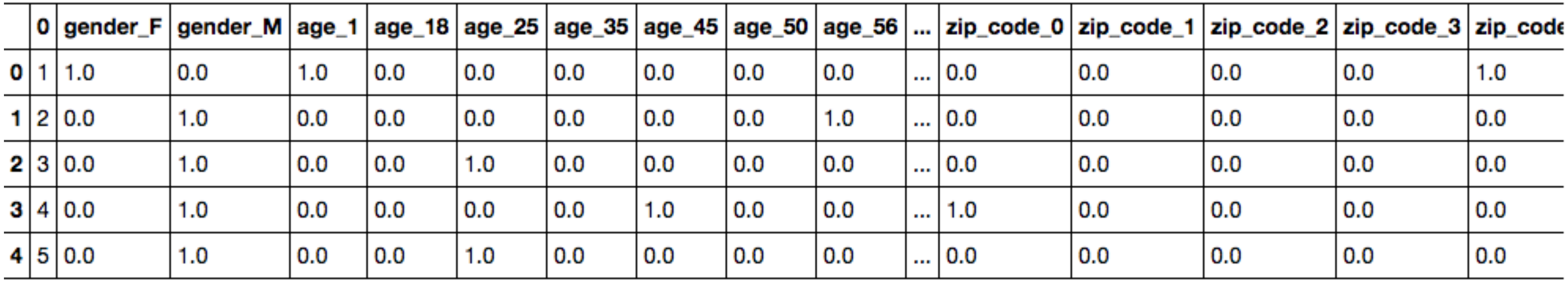
3.Experimental Results

Preprocessing Techniques

- Dummy Code
- Imputation



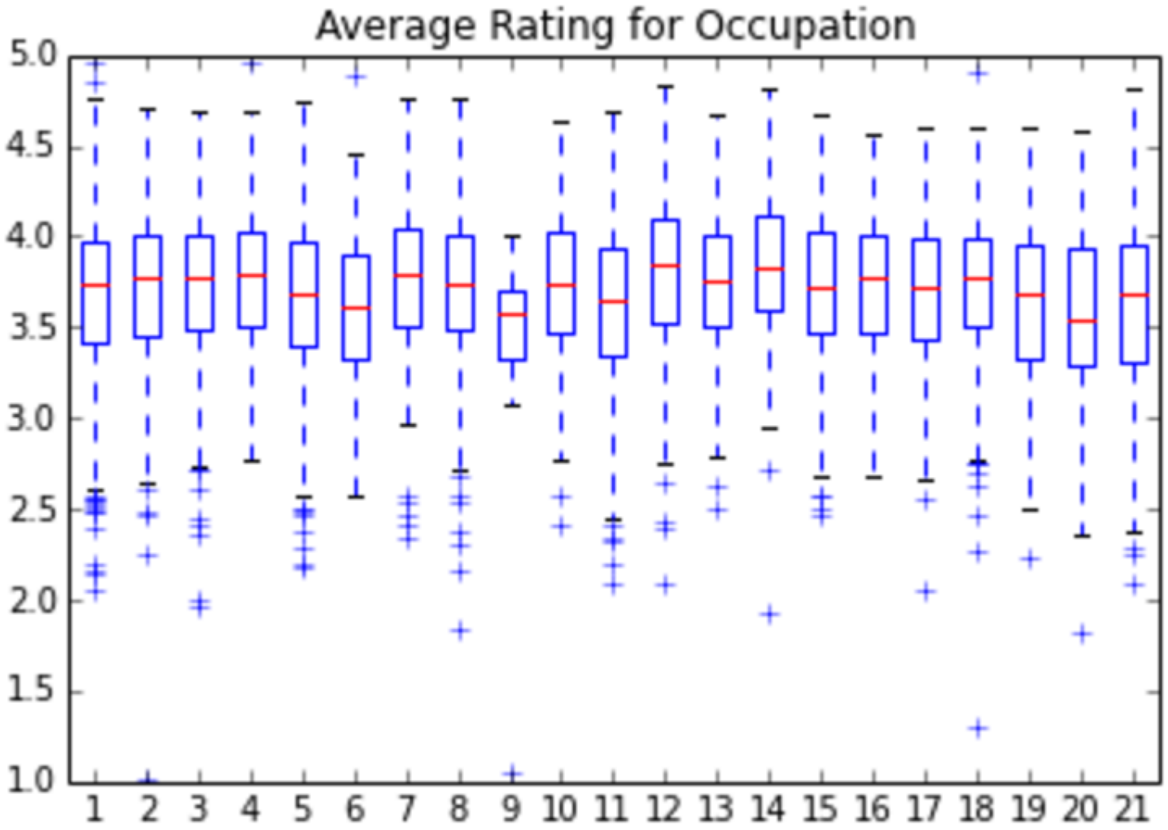
Animation	Children's	Comedy	Crime	Documentary	Drama	...	year_1	year_2	year_3	year_4	year_5	year_6	year_7	year_8	year_9	popularity
1.0	1.0	1.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1648.0
0.0	1.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	578.0
0.0	0.0	1.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	384.0



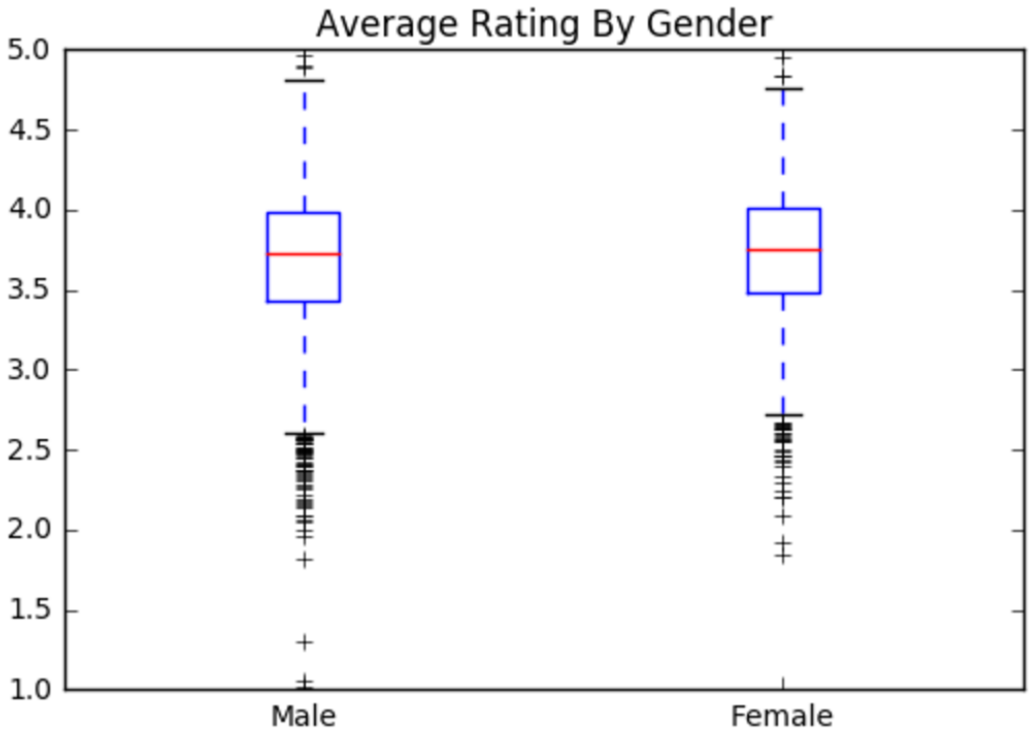
	0	gender_F	gender_M	age_1	age_18	age_25	age_35	age_45	age_50	age_56	...	zip_code_0	zip_code_1	zip_code_2	zip_code_3	zip_code
0	1	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	1.0
1	2	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	...	0.0	0.0	0.0	0.0	0.0
2	3	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0
3	4	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	...	1.0	0.0	0.0	0.0	0.0
4	5	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0

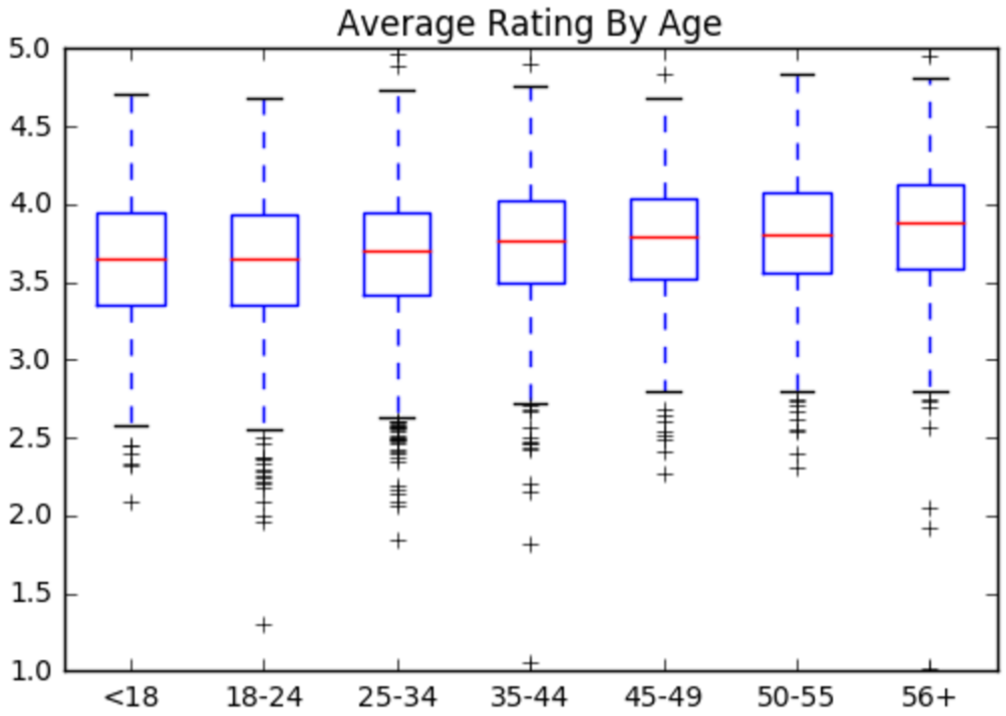
...

Plots


Average Rating for Occupation
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21

Plots


Average Rating By Gender
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
Male
Female


Average Rating By Age
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
<18
18-24
25-34
35-44
45-49
50-55
56+

Train/Test Split

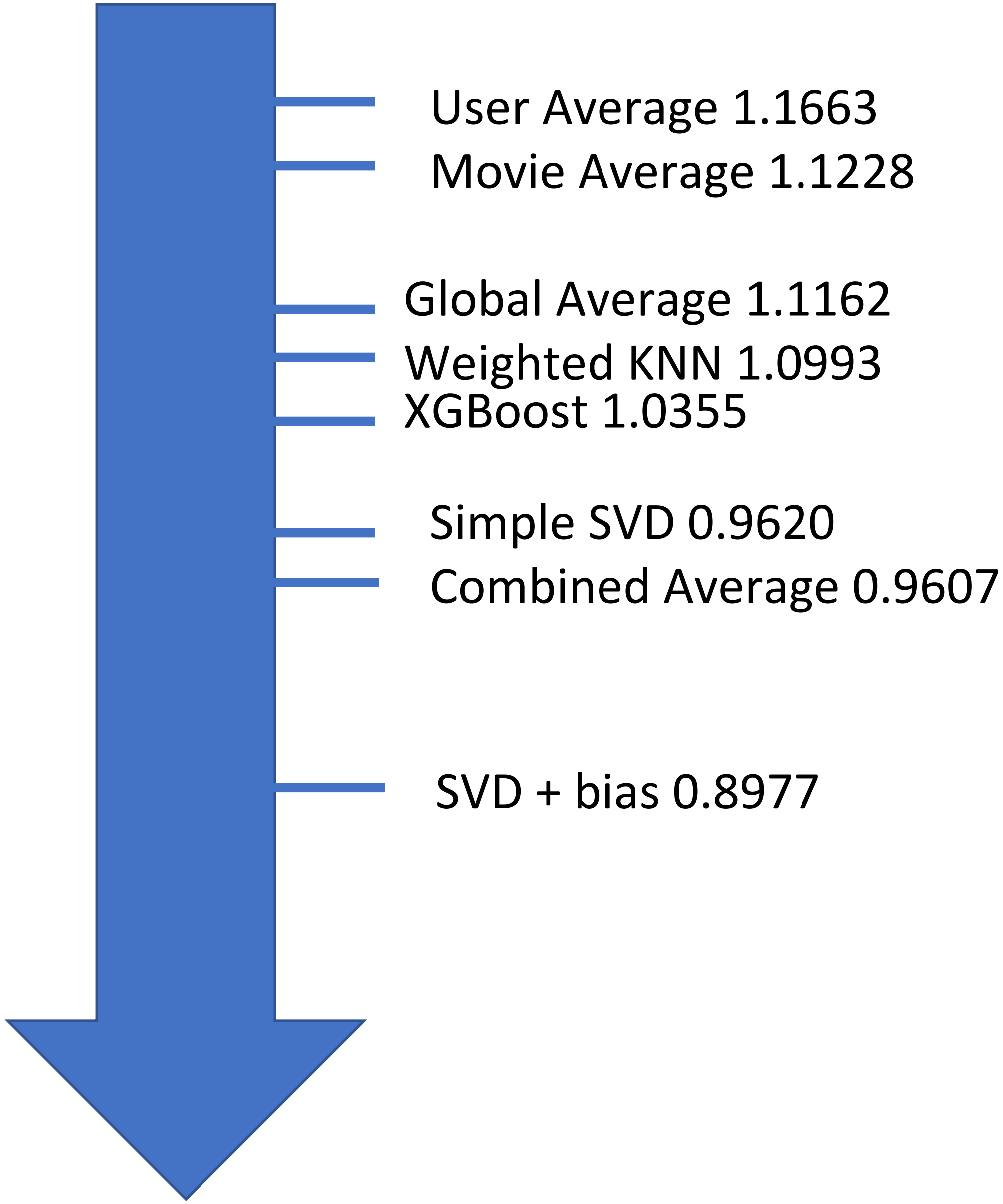
- 80% train (cross validation), 20% test
- Same train/test set for all methods
- Why?

Evaluation Metrics

- root mean square error (RMSE)
- Why?

Result Comparison

- Test RMSE



User Average 1.1663

Movie Average 1.1228

Global Average 1.1162

Weighted KNN 1.0993

XGBoost 1.0355

Simple SVD 0.9620

Combined Average 0.9607

SVD + bias 0.8977

4.Discussion

Key Challenges

- Preprocessing user/movie info
- Deal with missing values
- Defining distance metrics for KNN
- Running time and complexity

Problems

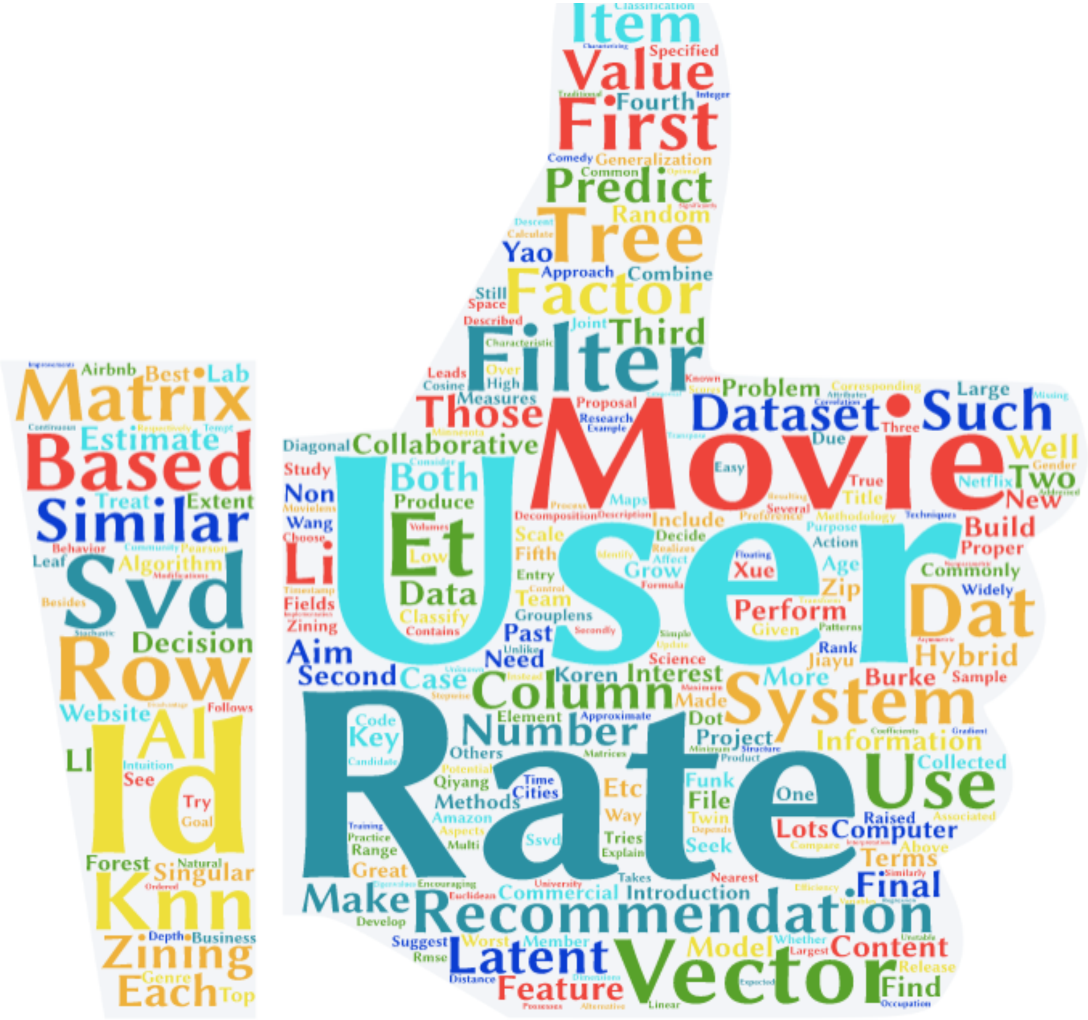
- How to get dummy variables for zip code?
- Imputation for missing values
- How to incorporate temporal information

Next step

- Explore smart ways to preprocess data
- Ensemble
- Try advanced models (if we have time..)

5.Q & A

Thank you!


User
Rate
Movie
Recommendation
Vector
System
Filter
Factor
Tree
First
Value
Item
Predict
Matrix
Based
Similar
Svd
Row
Id
Al
Knn
Zining
Each
Singular
Forest
Decision
Website
Estimate
Dataset
Such
Collaborative
Latent
Feature
Column
Number
Key
Make
Dat
Hybrid
Use
Information
Content
Final
Computer
Terms
Model
Find
Those
Problem
Both
Et
Li
Data
Aim
Second
Case
Interest
Koren
Build
Well
Two
New
Perform
Xue
Commonly
Burke
Third
Yao
Combine
Approach
Random
Generalization
Fourth
Specified
Algorithm
Extent
Treat
Best
Lab
Airbnb
Knowledge
Follows
Intuition
See
Try
Goal
Natural
Depth
Business
Genre
Top
Produce
Scale
Fifth
Low
Team
Classify
Past
Need
Grouplens
Entry
Project
Dot
Element
Funk
File
Seek
Twin
Lots
One
Raised
Collected
Commercial
Introduction
Great
Range
Practice
Amazon
Methods
Cities
Qiyang
Others
Etc
Way
Tries
Ssvd
Suggest
Rmse
Worst
Member
Largest
Release
Non
Wang
Study
Diagonal
Fields
Include
Decide
Action
Age
Zip
Jiayu
More
Purpose
Methodology
Title
True
Rank
Sample
Proper
Widely
Netflix
Large
Three
Due
Leads
High
Measures
Proposal
Research
Example
Still
Space
Described
Joint
Nearest
Above
Given
Science
Made