



Novelist

Find your next page turner

Meghan Thommes

Health Data Science Fellow

goodreads: A website for book lovers

The screenshot displays the Goodreads website interface. At the top, the navigation bar includes the Goodreads logo, links for Home, My Books, Browse, and Community, a search bar, and notification icons. The main content is organized into several sections:

- CURRENTLY READING:** A vertical list of books with progress indicators. Visible titles include "Goldenhand (#5)" by Garth Nix, "I Contain Multitudes: The Microbes Within..." by Ed Yong, and "Fall of Light (The Kharkanas Trilogy, #...)" by Steven Erikson.
- 2020 READING CHALLENGE:** A badge showing "3 books completed" and "You're on track!" with a progress bar at 3/50 (6%).
- WANT TO READ:** A section for books on the user's shelf, with "The Humboldt Current" by Sara Pennypacker visible.
- UPDATES:** A central feed of user activity. It shows "Anita wants to read" "Jane Anonymous" by Laurie Faria Stolarz (with a "Want to Read" button and a 5-star rating) and "Anita made progress on The Hand on the Wall (Truly Devious, #3)" by Maureen Johnson (with a progress bar at 43% and a "Want to Read" button).
- NEWS & INTERVIEWS:** A featured article titled "Exclusive Excerpt from the Upcoming YA Novel 'This Is my America'" with a book cover image.
- RECOMMENDATIONS:** A section titled "Because you enjoyed Abhorsen (Abhorsen, #3):" featuring "Year of the Griffin (Derkholm, #2)" by Diana Wynne Jones.
- GOODREADS CHOICE AWARDS:** A banner for the "Announcing the BEST BOOKS of 2019!" with the Goodreads Choice Awards logo.








Which book should a user read next?

Can sort by

- Average Rating
- Date Added
- Author
- Title
- *etc*

goodreads Home My Books Browse Community Search books

My Books: Want to Read (266)

#	cover	title	author	avg rating	rating	shelves	date read	date added	
3		Ringworld (Ringworld, #1)	Niven, Larry	3.96	★★★★★	to-read [edit]	not set [edit]	Feb 24, 2014	edit view ✕
4		Sun of Suns (Virga, #1)	Schroeder, Karl	3.74	★★★★★	to-read [edit]	not set [edit]	Feb 24, 2014	edit view ✕
5		Paleofantasy: What Evolution Really Tells Us about Sex, Diet, and How We Live	Zuk, Marlene	3.69	★★★★★	to-read [edit]	not set [edit]	Feb 24, 2014	edit view ✕
6		The Sun Also Rises	Hemingway, Ernest	3.82	★★★★★	to-read [edit]	not set [edit]	Feb 24, 2014	edit view ✕
7		The Little Prince	Saint-Exupéry, Antoine de	4.30	★★★★★	to-read [edit]	not set [edit]	Feb 24, 2014	edit view ✕
8		A Grief Observed	Lewis, C.S.	4.20	★★★★★	to-read [edit]	not set [edit]	Feb 24, 2014	edit view ✕
9		Blood Meridian, or the Evening Redness in the West	McCarthy, Cormac	4.17	★★★★★	to-read [edit]	not set [edit]	Feb 24, 2014	edit view ✕

30 of 266 loaded per page infinite scroll sort Date added

Which book should a user read next?

Can sort by

- Average Rating
- Date Added
- Author
- Title
- *etc*

The screenshot shows the Goodreads interface for a user's 'Want to Read' shelf. The page title is 'My Books: Want to Read (266)'. The table lists books with columns for #, cover, title, author, avg rating, rating, shelves, date read, date added, and edit/view options. The books are sorted by 'date added' in descending order. A red box highlights the header row of the table.

#	cover	title	author	avg rating	rating	shelves	date read	date added	edit	view
3		Ringworld (Ringworld, #1)	Niven, Larry	3.96	★★★★★	to-read [edit]	not set [edit]	Feb 24, 2014	edit	view
4		Sun of Suns (Virga, #1)	Schroeder, Karl	3.74	★★★★★	to-read [edit]	not set [edit]	Feb 24, 2014	edit	view
5		Paleofantasy: What Evolution Really Tells Us about Sex, Diet, and How We Live	Zuk, Marlene	3.69	★★★★★	to-read [edit]	not set	Feb 24, 2014	edit	view
6		The Sun Also Rises								
7		The Earth From Space								
8		A Grief Observed	Lewis, C.S.	4.20	★★★★★	to-read [edit]	not set [edit]	Feb 24, 2014	edit	view
9		Blood Meridian, or the Evening Redness in the West	McCarthy, Cormac	4.17	★★★★★	to-read [edit]	not set [edit]	Feb 24, 2014	edit	view

At the bottom of the page, there is a pagination bar showing '30 of 266 loaded per page', a dropdown menu for 'infinite scroll', a 'sort' dropdown menu, and a 'Date added' dropdown menu.

No personalized information

Novelist

<https://insight-novelist.herokuapp.com/>

Find your next page turner

This app ranks your "To-Read" books on [Goodreads](#)

Would you like to upload a CSV of your exported Goodreads library?

Upload your Goodreads data?

- Yes, upload my own Goodreads data
- No, use pre-loaded data

Upload your Goodreads library:

Upload your exported Goodreads Library CSV file

Drop files here to upload
or
[browse files](#)

Obtain data on how **users** rate **books**

Data cleaning

goodreads
UCSD Book Graph
15.7M reviews

{JSON}

16.7GB



2GB

UCSD Book Graph

Scraped Goodreads users' public shelves in late 2017

- 2.1M books
- 465k users

Most books are rated **3 stars** or more

Data cleaning

Feature extraction

goodreads

UCSD Book Graph

15.7M reviews

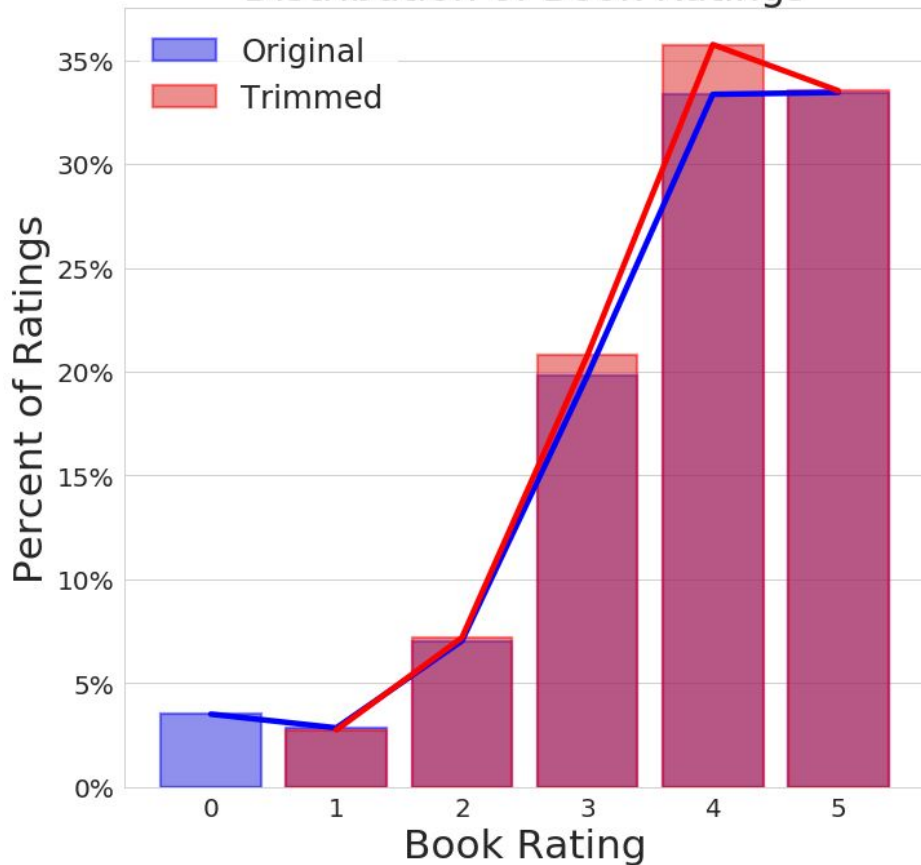


9.3M reviews



Book Ratings

Distribution of Book Ratings



Use **collaborative filtering** to predict book ratings

Data cleaning

Feature extraction

Model
tuning



Evaluation &
Validation

goodreads

UCSD Book Graph

15.7M reviews



9.3M reviews



Book Ratings



Collaborative-based

?

Sometimes the simplest method is the best

Model	Root Mean Square Error (# stars)
Baseline (Alternating Least Squares)	0.988
Baseline (Stochastic Gradient Descent)	0.990
Matrix Factorization (Singular Value Decomposition)	1.027

The error is not great...but neither was Netflix's

Netflix Prize

COMPLETED

[Home](#) [Rules](#) [Leaderboard](#) [Update](#) [Download](#)

Leaderboard

Showing Test Score. [Click here to show quiz score](#)

Display top leaders.

Best RMSE = 0.95 → 0.86

...but the additional accuracy gains did not justify the effort needed to bring them into production

Rank **Team Name** **Best Test Score** **% Improvement** **Best Submit Time**

Grand Prize - RMSE = 0.8567 - Winning Team: BellKor's Pragmatic Chaos

1	BellKor's Pragmatic Chaos	0.8567	10.06	2009-07-26 18:18:28
2	The Ensemble	0.8567	10.06	2009-07-26 18:38:22
3	Grand Prize Team	0.8582	9.90	2009-07-10 21:24:40

Use **collaborative filtering** to predict book ratings

Data cleaning

Feature extraction

Model tuning



Evaluation & Validation

Use model to predict ratings

goodreads
UCSD Book Graph

15.7M reviews



9.3M reviews



Book Ratings



Collaborative-based

Baseline
(ALS)

Baseline estimates for each
user and item

{JSON}



mongoDB®



PostgreSQL

Top 10
Ratings

Bottom 10
Ratings

NoveList gives users a **personalized** experience

User A

Your top 10 ranked books are:

	Title	Author
0	When Breath Becomes Air	Paul Kalanithi
1	Homo Deus: A History of Tomorrow	Yuval Noah Harari
2	Algorithms to Live By: The Computer Science of Human Decisions	Brian Christian
3	Never Let Me Go	Kazuo Ishiguro
4	Seveneves	Neal Stephenson
5	The Sixth Extinction: An Unnatural History	Elizabeth Kolbert
6	The Winds of Winter (A Song of Ice and Fire, #6)	George R.R. Martin
7	Anne of Green Gables (Anne of Green Gables, #1)	L.M. Montgomery
8	The Little Prince	Antoine de Saint-Exupéry
9	Animal Farm	George Orwell



User B

Your bottom 10 ranked books are:












These books may be ranked low because you have not read similar books

	Title	Author
0	The Relic Master	Christopher Buckley
1	You're Never Weird on the Internet (Almost)	Felicia Day
2	The Savage Detectives	Roberto Bolaño
3	The Stone Sky (The Broken Earth, #3)	N.K. Jemisin
4	Love, Nina: Despatches from Family Life	Nina Stibbe
5	A Face Like Glass	Frances Hardinge
6	Sacred Games (Sacred Games)	Vikram Chandra
7	American War	Omar El Akkad
8	Human Voices	Penelope Fitzgerald
9	Love & Gelato	Jenna Evans Welch



Additional Slides

Collaborative filtering to make personalized predictions

						
	5		4			4
		3		3		2
			5		5	5
	4		5		4	
	4				2	

Personalized predictions are **hard**

Every person is **unique** with a variety of interests

Data Sparsity: Have **large datasets**, but a small amount of **data per user**

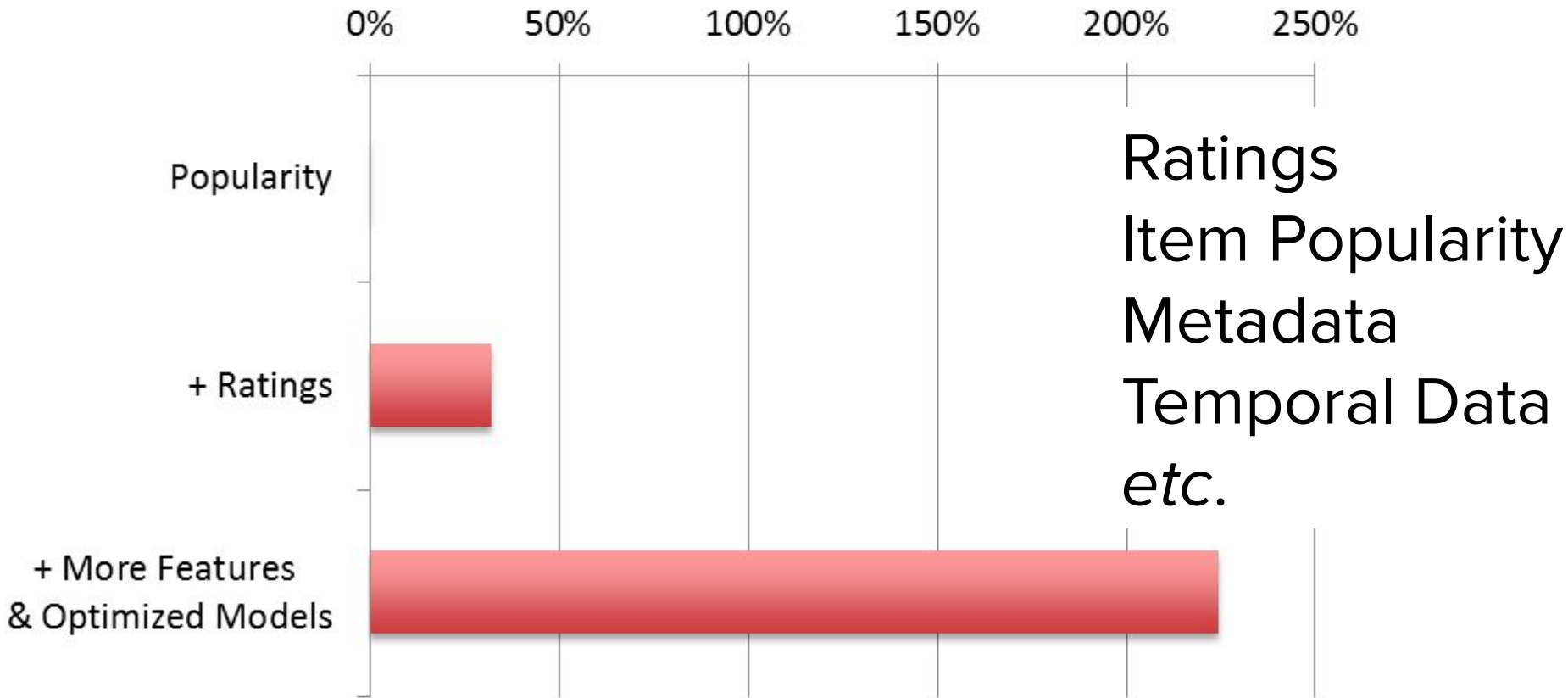
Cold-start: Cannot make a prediction if a user or book does not have **enough ratings**

Enjoyment depends on mood, context, ...

...or a if user just wants something new, fresh, etc

Adding **features** is the best way to improve rankings

Ranking improvement over baseline



Potential features for **content-based** filtering

goodreads

Author(s)

Genre(s)

Publisher

Year Published

Number of Pages

Duration (*how long it takes to read*)



Aggregates bibliographic information into **subject terminology schemas**

FAST subject headings

- Faceted Application of Subject Terminology

Recommender system to predict book ratings

Data cleaning

Feature extraction

Model tuning



Evaluation & Validation

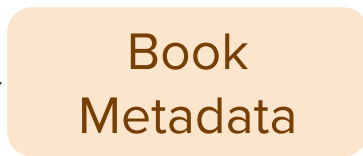
Use model to predict ratings

goodreads

15.7M reviews



9.3M reviews



Book Metadata

Content-based
(engineered features)

k Nearest
Neighbors

TF-IDF

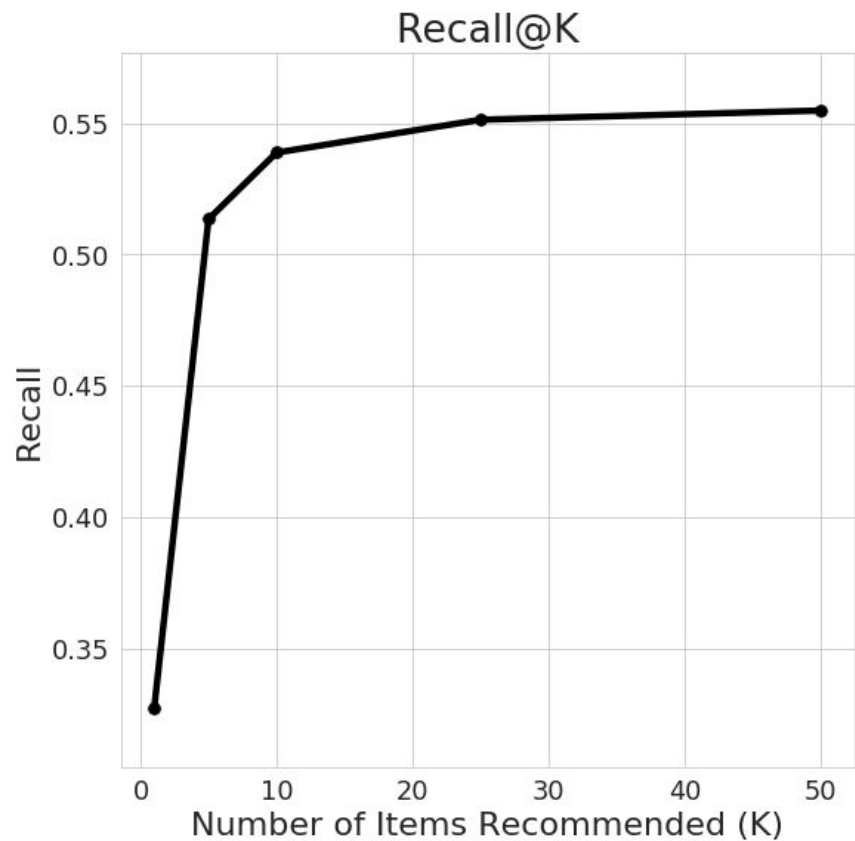
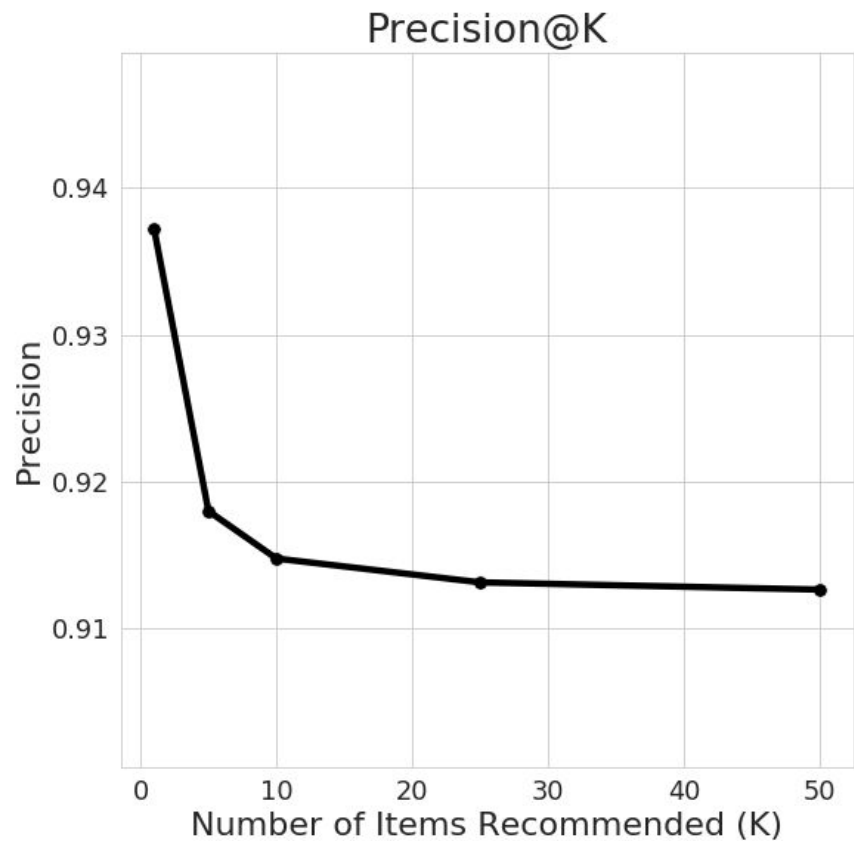
Collaborative-based

Baseline Only

Baseline estimates for each
(user, item) pair

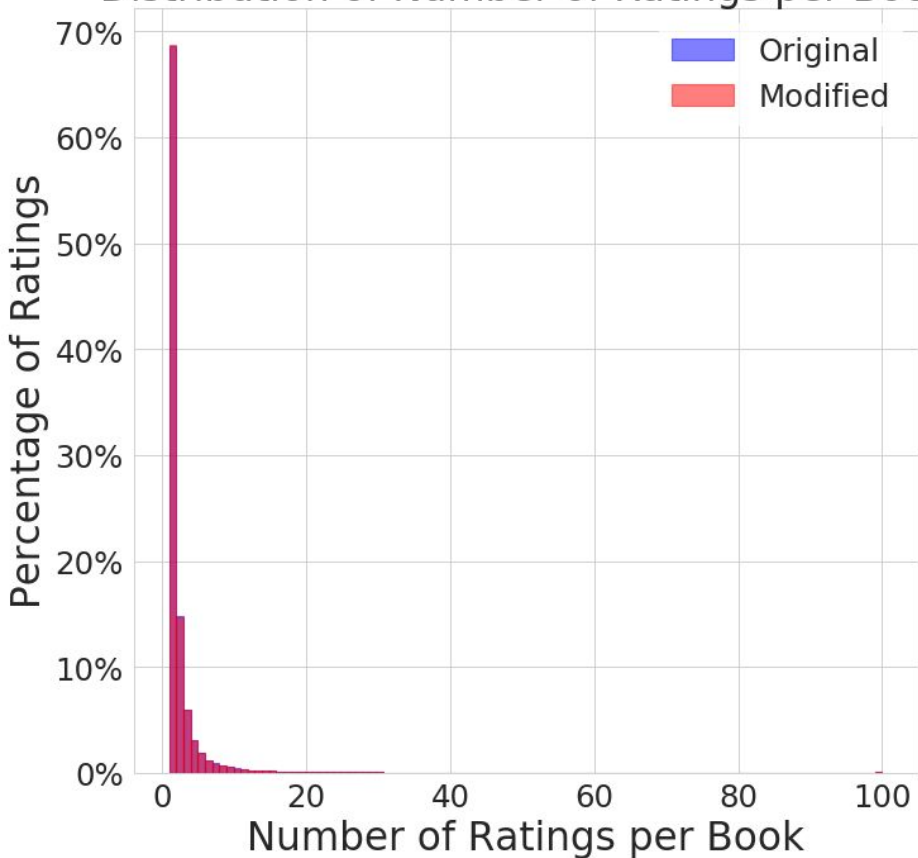
Rank
Ratings

Precision and recall at cutoff K

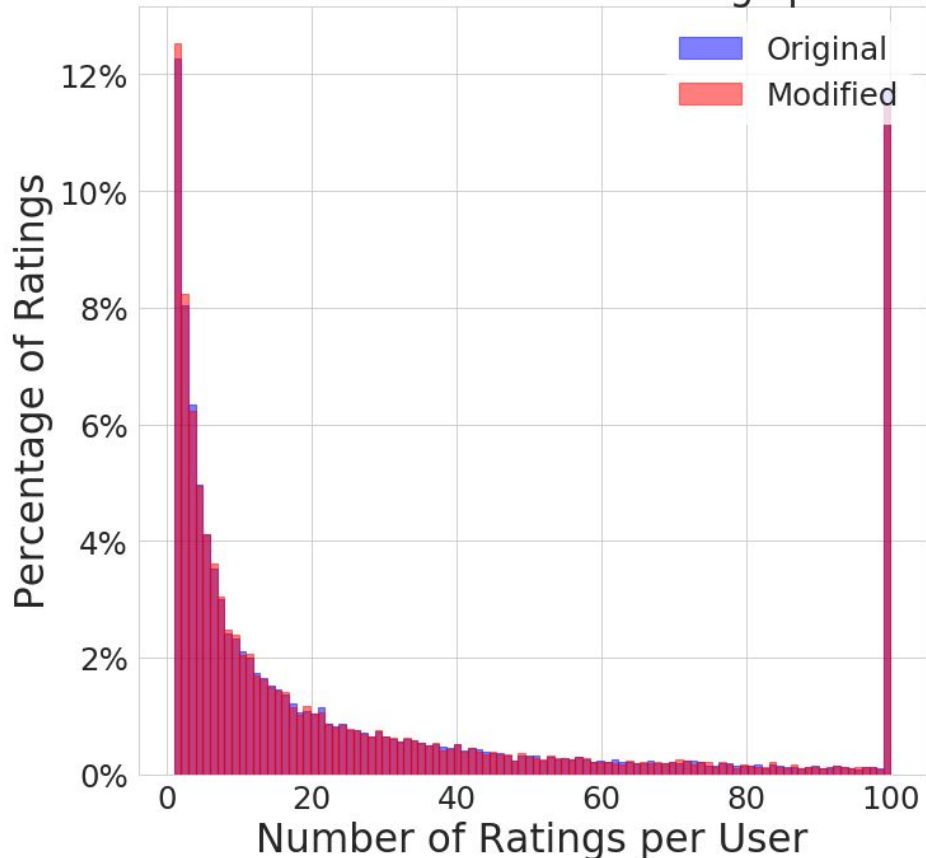


Histograms: Number of Ratings

Distribution of Number of Ratings per Book

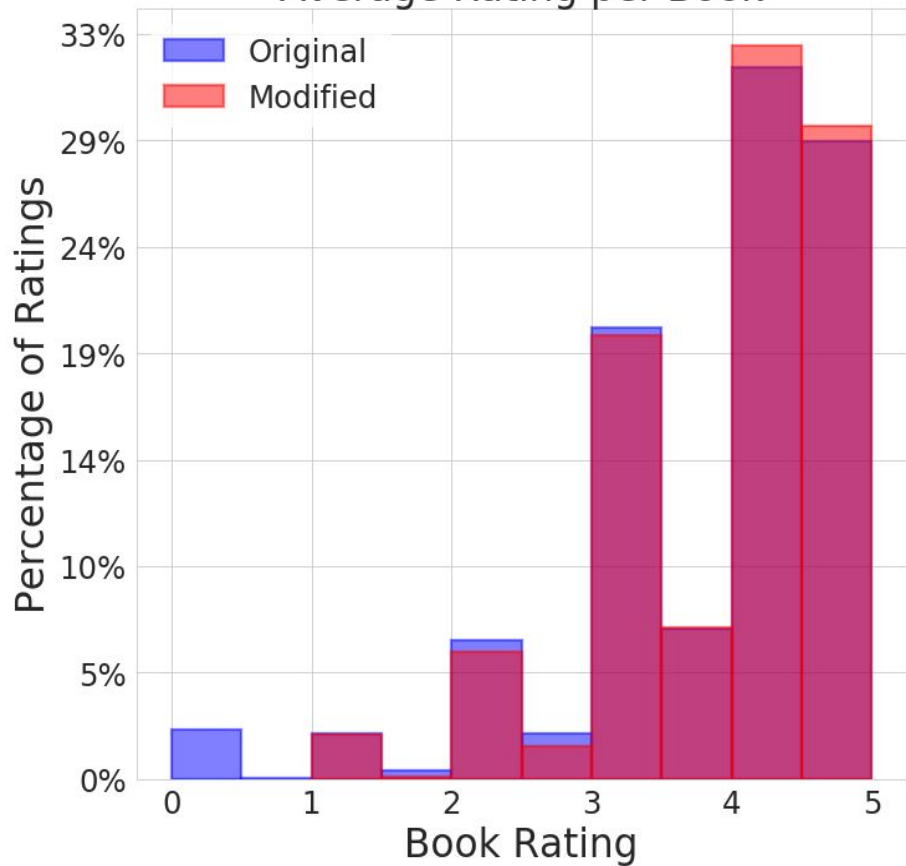


Distribution of Number of Ratings per User

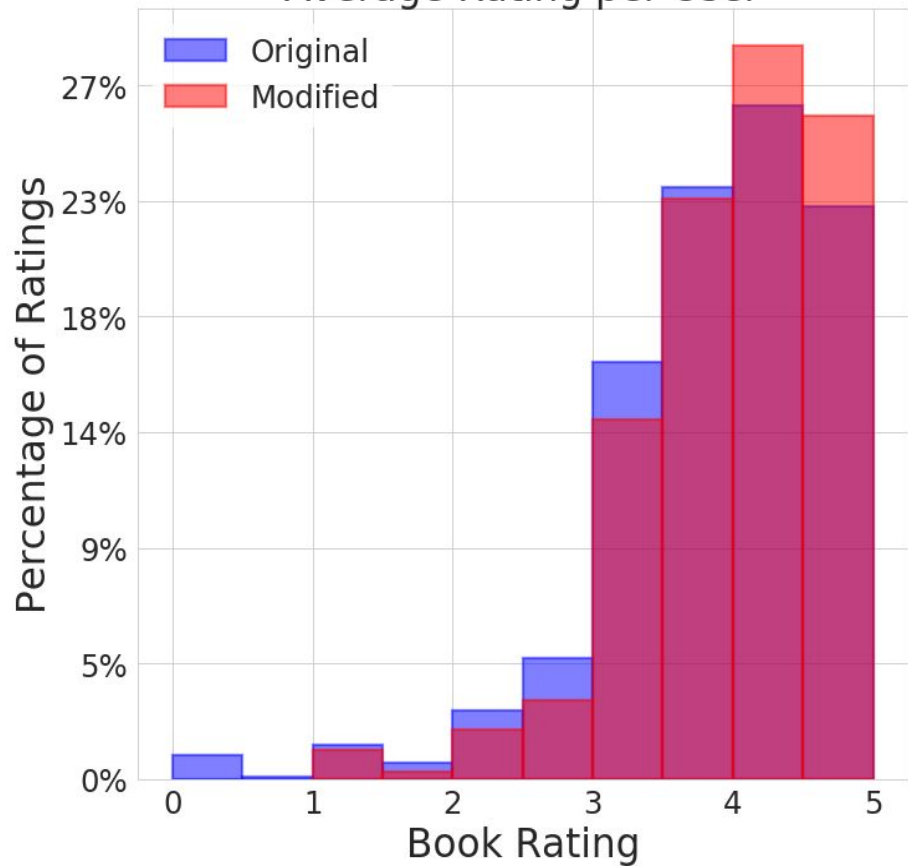


Histograms: Average Ratings

Average Rating per Book



Average Rating per User



Why not the Goodreads API?

- ✓ Rest API
- ✗ Opens up users to some security vulnerabilities
- ✗ Python wrappers are not well supported (developed as a side project by non-employees)
- ✗ OAuth integration is not well documented (and runs into errors)

Method Overview: Baseline

Rating is predicted based on the baseline estimate for each user and book:

$$\hat{r}_{user,book} = b_{user,book} = \mu + b_{user} + b_{book}$$

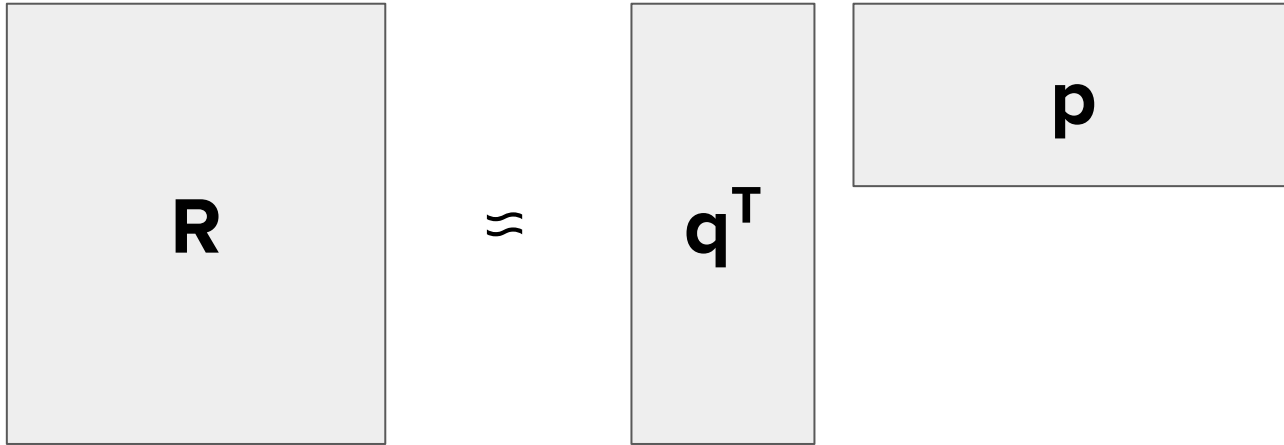
Minimizing the regularized square error:

$$\sum_{r_{user,book} \in R_{train}} (r_{user,book} - (\mu + b_{user} + b_{book}))^2 + \lambda_{user} b_{user}^2 + \lambda_{book} b_{book}^2$$

Can use stochastic gradient descent (SGD) or alternating least squares (ALS) to minimize the error

Method Overview: Matrix Factorization

Find **latent features**



Method Overview: Matrix Factorization

Rating is predicted based on the baseline estimate for each user and book and the latent book and user factors:

$$\hat{r}_{user,book} = b_{user,book} = \mu + b_{user} + b_{book} + q_{book}^T p_{user}$$

Minimizing the regularized square error:

$$\sum_{r_{user,book} \in R_{train}} (r_{user,book} - \hat{r}_{user,book})^2 + \lambda (b_{book}^2 + b_{user}^2 + \|q_{book}\|^2 + \|p_{user}\|^2)$$

Use stochastic gradient descent (SGD) to minimize the error

Method Overview: Matrix Factorization

Singular Value Decomposition (SVD)

Non-Negative Matrix Factorization (NMF)

- Factors are non-negative

$$p_{user} \leftarrow p_{user} + \gamma(e_{user,book} \cdot q_{book} - \lambda p_{user})$$

$$q_{book} \leftarrow q_{book} + \gamma(e_{user,book} \cdot p_{user} - \lambda q_{book})$$

$$p_{user,factor} \leftarrow p_{user,factor} \cdot \frac{\sum_{book \in B_{user}} q_{book,factor} \cdot r_{user,book}}{\sum_{book \in B_{user}} q_{book,factor} \cdot \hat{r}_{user,book} + \lambda_{user} |B_{user}| p_{user,factor}}$$

$$q_{book,factor} \leftarrow q_{book,factor} \cdot \frac{\sum_{user \in U_{book}} p_{user,factor} \cdot r_{user,book}}{\sum_{user \in U_{book}} p_{user,factor} \cdot \hat{r}_{user,book} + \lambda_{book} |U_{book}| q_{book,factor}}$$

Method Overview: Clustering (k -NN)

Use similarity between users or items

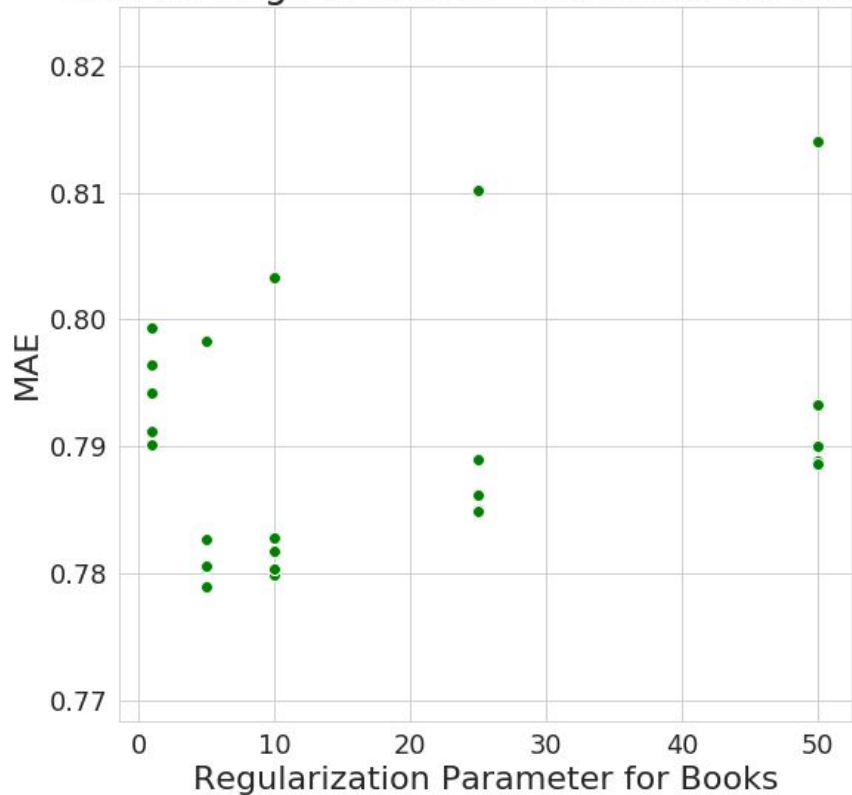
$$\text{User-centric: } \hat{r}_{user,book} = \frac{\sum_{v \in N_{book}^k(user)} \text{sim}(user, v) \cdot r_{v,book}}{\sum_{v \in N_{book}^k(user)} \text{sim}(user, v)}$$

$$\text{Item-centric: } \hat{r}_{user,book} = \frac{\sum_{j \in N_{user}^k(book)} \text{sim}(book, j) \cdot r_{user,j}}{\sum_{j \in N_{user}^k(book)} \text{sim}(book, j)}$$

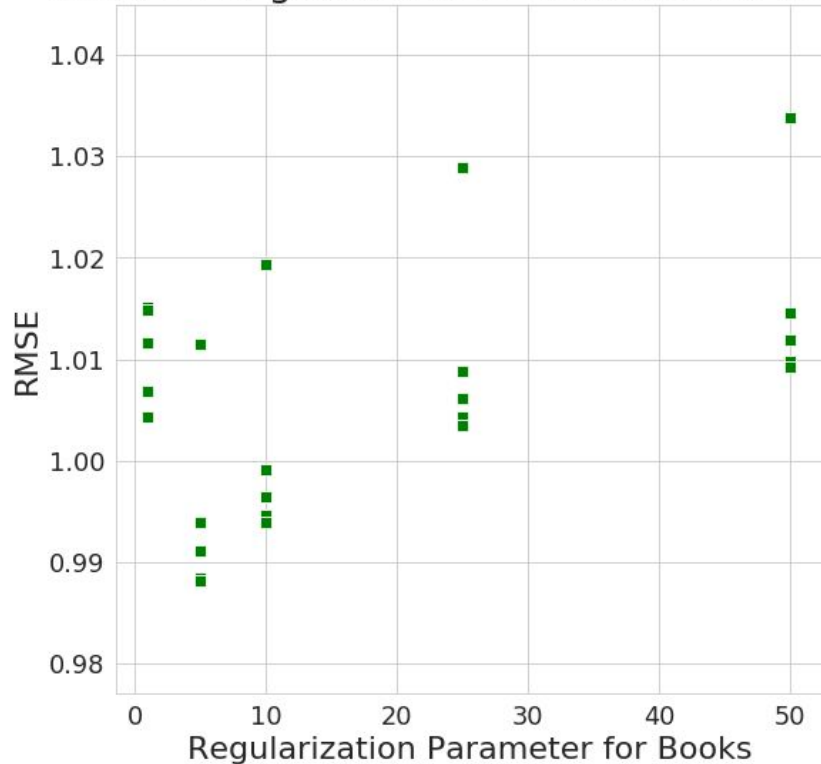
Can also offset by mean or baseline

Parameter Tuning: Baseline (λ_{books})

MAE vs Regularization Parameter for Books

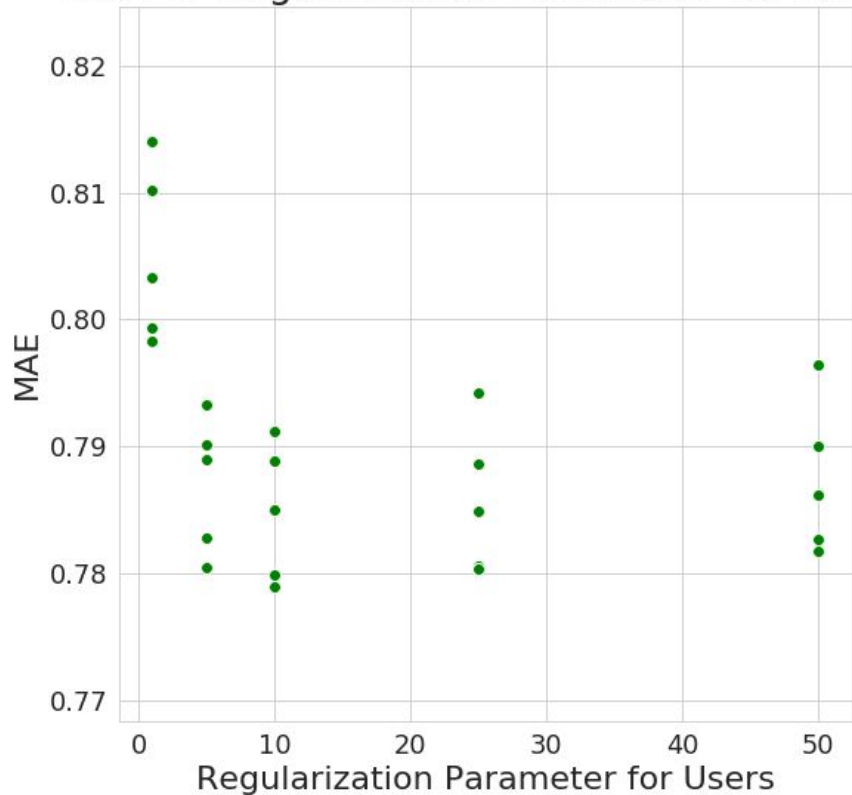


RMSE vs Regularization Parameter for Books

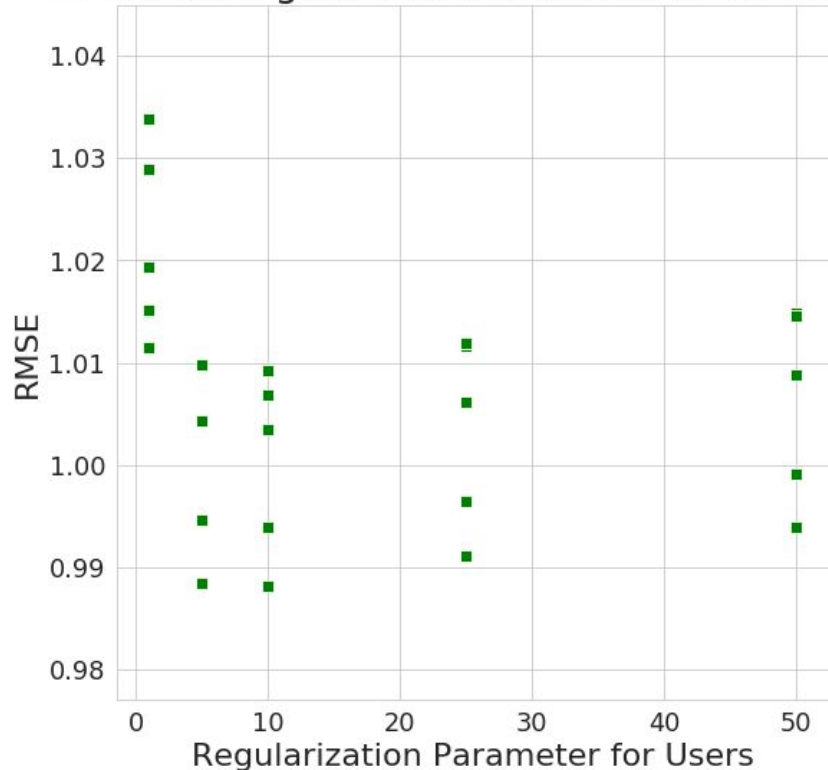


Parameter Tuning: Baseline (λ_{users})

MAE vs Regularization Parameter for Users

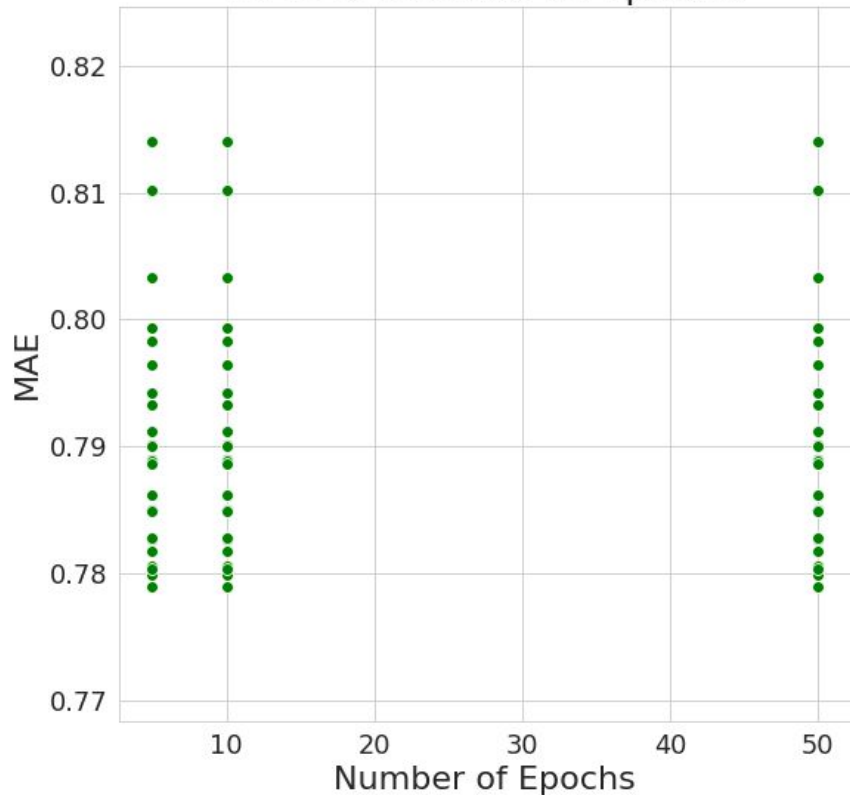


RMSE vs Regularization Parameter for Users

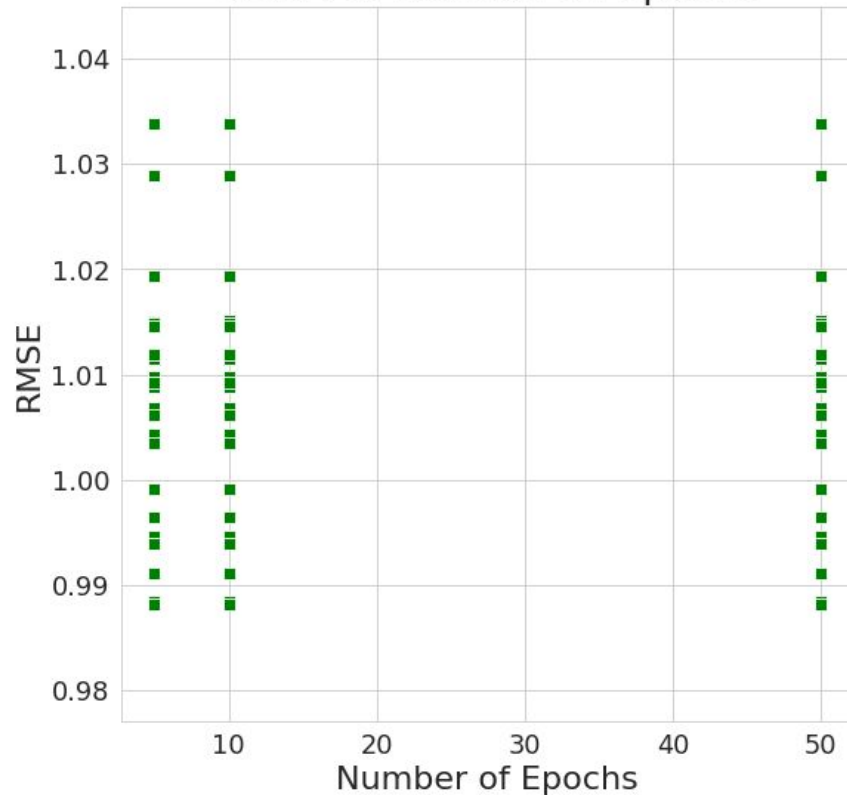


Parameter Tuning: Baseline (n_{epochs})

MAE of Number of Epochs



RMSE vs Number of Epochs



Assessing Models

