Goodreads Books

Adam Gonzalez, Erin Lee, Jon Le,
Debbie Lu, and Corinne Smith

Motivation

Book reviews generally provide the following information:

- reader opinions or editorial reviews about a book
- simple ratings
- similar reads or recommendations

Source Image:
https://mixkit.co/free-stock-art/woman-lying-in-the-grass-cuddling-a-dog-with-an-22/

Motivation

But there is a gap....

- Book reviews are heavily qualitative as opposed to quantitative.
- What predicts the ratings?
- What are the strongest predictors of a higher or lower rating?

Research question: What are the factors that affect the average book rating?

Overview of Data Set

Data collection:

- The two datasets we used were both scraped from the GoodReads API.
- “Goodreads-books” from Kaggle
  - https://www.kaggle.com/jealousleopard/goodreadsbooks
- “GoodReads Authors” from Kaggle
  - https://www.kaggle.com/choobani/goodread-authors

Source Image:
https://goodereader.com/blog/electronic-readers/amazon-owns-goodreads-what-now

Data Cleaning

About the data:

- The first dataset contains information primarily on the books themselves.
- The second contains information primarily on the authors (hence the names).
- We combined both datasets into one **to analyze how average book ratings are affected by a number of different factors.**

Manipulation:

- Lots of character data types
- Added new feature:
  - Sentiment score

Cleaning Code

```
# 4) DATA CLEANING

```{r}
# mutate to correct column data types
books_1 <- books_sa %>% mutate(num_pages = as.numeric(num_pages),
                                avg_book_rating = as.numeric(avg_book_rating),
                               text_reviews_count = as.numeric(text_reviews_count),
                               publication_date = as.Date(publication_date,
format="%m/%d/%Y"),
                               born = as.Date(born, format="%m/%d/%Y"),
                               died = as.Date(died, format="%m/%d/%Y"),
                               gender = as.factor(gender)
                               )

# eliminate NAs, using " filter(!is.na()) " from Problem Set 5 cleaning code
books_total <- books_1 %>%
  filter(
    (!is.na(avg_book_rating)), (!is.na(book_ratings_count)), (!is.na(text_reviews_count)),
(!is.na(publication_date)),
    (!duplicated(title)),
    (avg_book_rating != 0),
    (author != "NOT A BOOK"))
```
```

More Cleaning Code

```
#Eliminate useless columns: sd(standard deviation of words in title), author ID, image_URL,
about, influence, website, twitter, original hometown, country, latitude, longitude

books_corti <- books_total %>% select(-isbn13,
                                      -sd,
                                      -authorid,
                                      -image_url,
                                      -about,
                                      -influence,
                                      -website,
                                      -twitter,
                                      -original_hometown,
                                      -country,
                                      -latitude,
                                      -longitude) %>% rename(
                                                        title_sentiment_avg = ave_sentiment,
                                                        title_word_count = word_count
                                      )
```

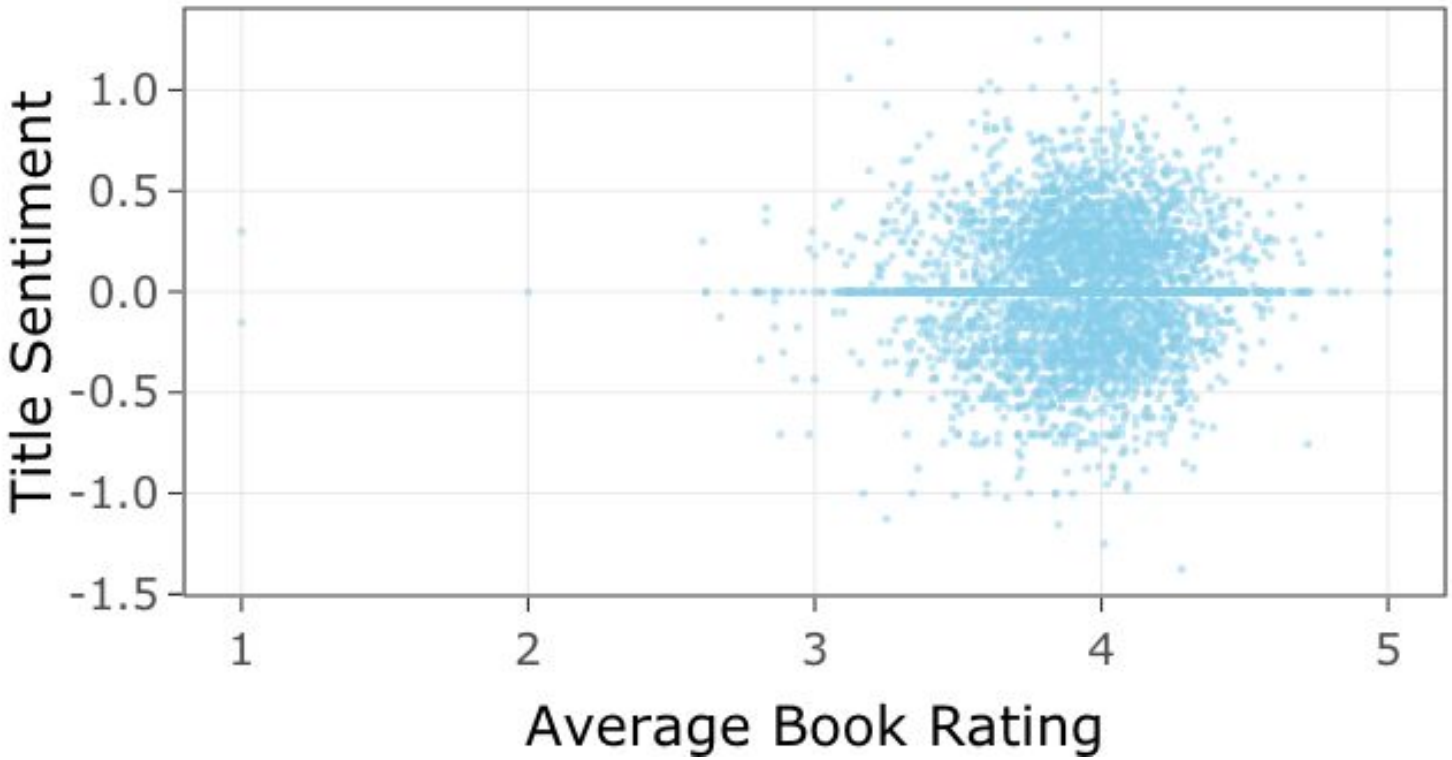
Data Exploration

Chosen variables:

- Outcome variable: average book rating
  - (“average_rating”)
- Predictors: 9 total
  - “num_pages”, “book_ratings_count”, “text_reviews_count”, “title_sentiment_avg”, “authorworkcount”, “author_fans”, “author_ratings_count, “author_review_count”, “gender”

Source Image:
https://mixkit.co/free-stock-art/person-using-binoculars-peering-out-from-behind-a-bush-24/

Exploring Data: Visualization 1



Exploring the Data: Visualization 2

Model 1: Linear Regression

Why a linear regression?

Training Metrics:

- RMSE: 0.282
- MAE: 0.214
- RSQ: 0.0606

Testing Metrics:

- RMSE: 0.254
- MAE: 0.180
- RSQ: 0.089

Pictured right:

- Linear Regression Output in a table

	avg_book_rating		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	3.82	3.80 – 3.84	**<0.001**
num_pages	0.00	0.00 – 0.00	**<0.001**
book_ratings_count	-0.00	-0.00 – 0.00	0.582
text_reviews_count	0.00	-0.00 – 0.00	0.221
title_sentiment_avg	0.05	0.02 – 0.08	**0.001**
authorworkcount	0.00	-0.00 – 0.00	0.051
author_fans	0.00	-0.00 – 0.00	0.124
author_ratings_count	0.00	0.00 – 0.00	**<0.001**
author_review_count	-0.00	-0.00 – -0.00	**<0.001**
gender [male]	0.03	0.01 – 0.05	**<0.001**
gender [unknown]	0.00	-0.02 – 0.03	0.773
Observations	4767		
R^2 / R^2 adjusted	0.061 / 0.059		

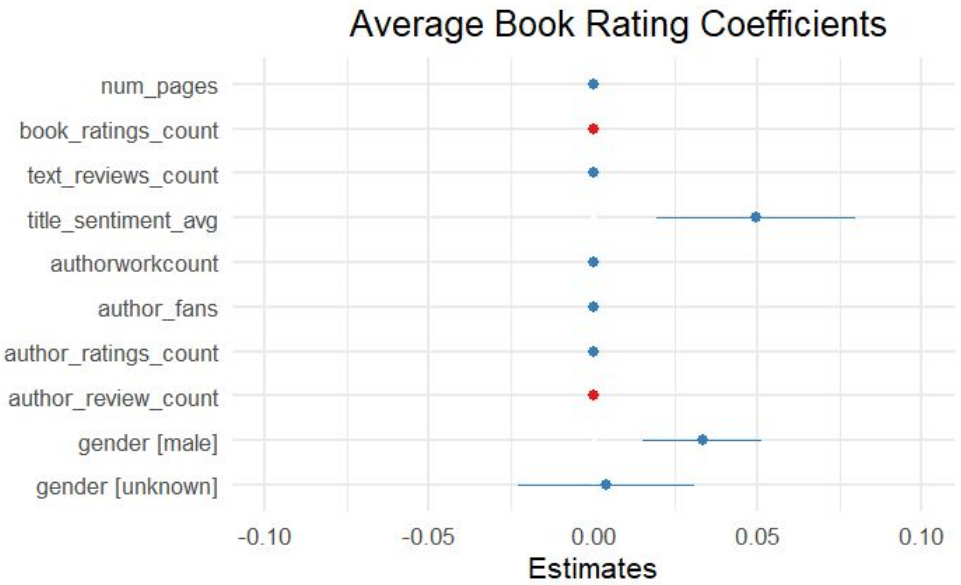
Model 1: Analysis

Assumption:

- Linear relationship between variables

Significant Predictors (based on p-value):

- Number of pages
- Average sentiment score of title
- Number of times author was rated
- Number of times author was reviewed
- Author identified as male

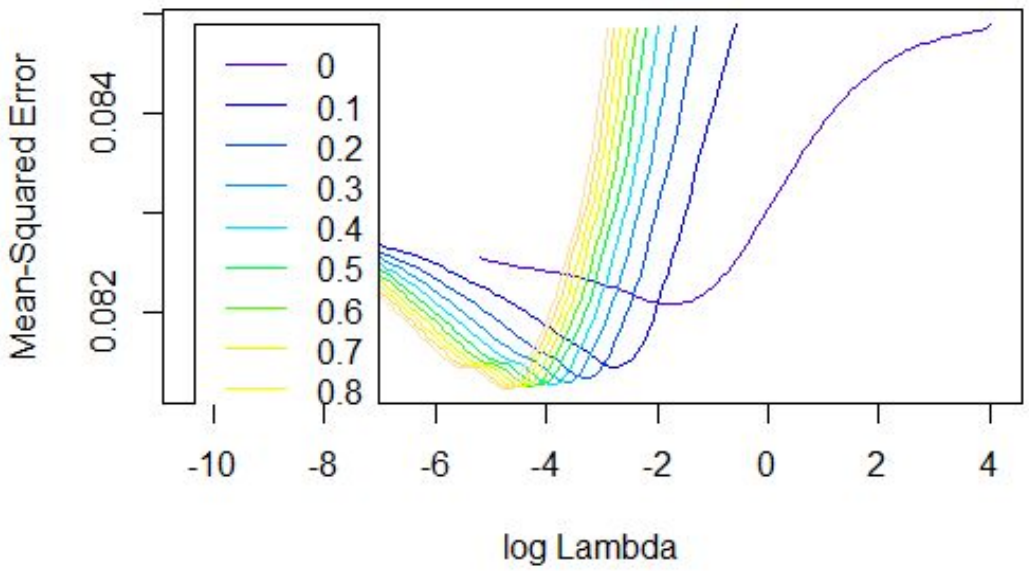


Model 2: Elastic Net

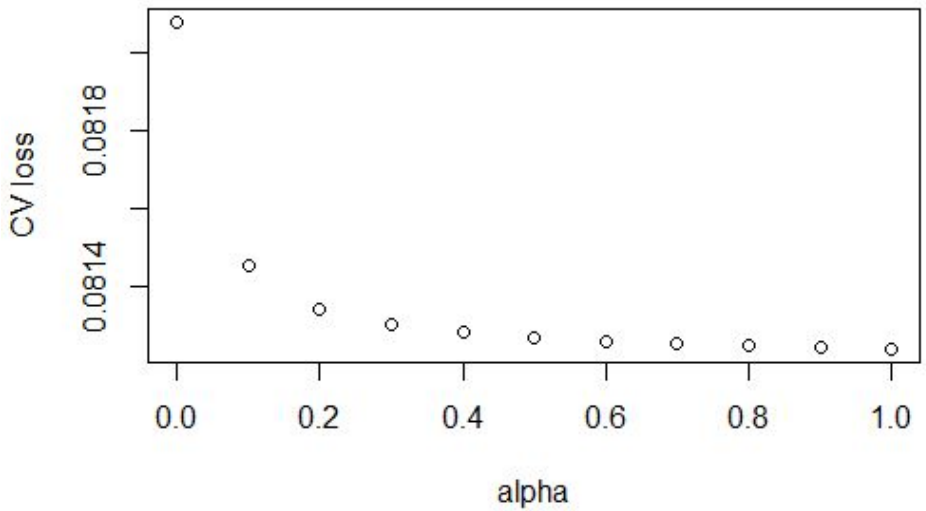
Why an elastic net model?

Best Model Parameters (based on Elastic Net Model):

- Alpha: 1
- Lambda.min: 0.00861
- Lambda.1se: 0.0419



Model 2: Analysis



alpha <dbl>	lambdaMin <dbl>	lambdaSE <dbl>	error <dbl>
1	0.008608862	0.04186147	0.08124101

Model 3: Lasso

Why use lasso?

Training Metrics:

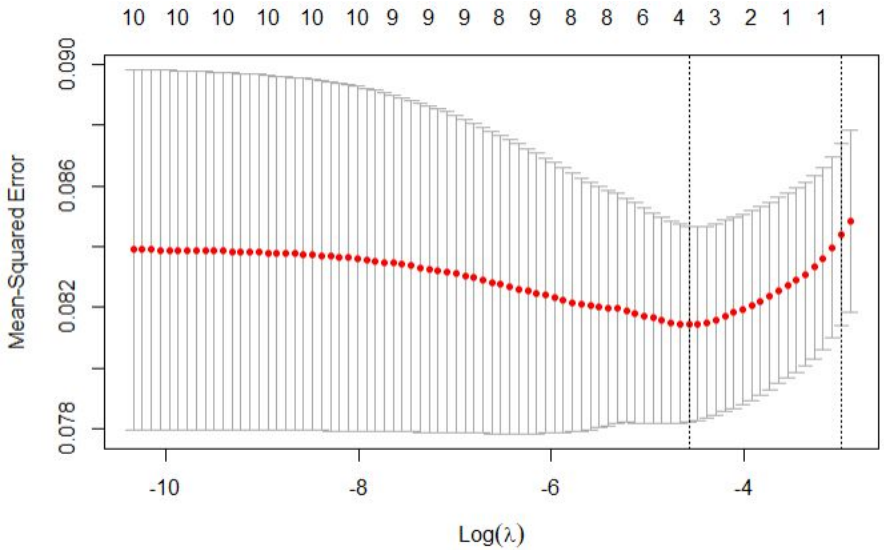
- RMSE: 0.285
- MAE: 0.216
- RSQ: 0.047

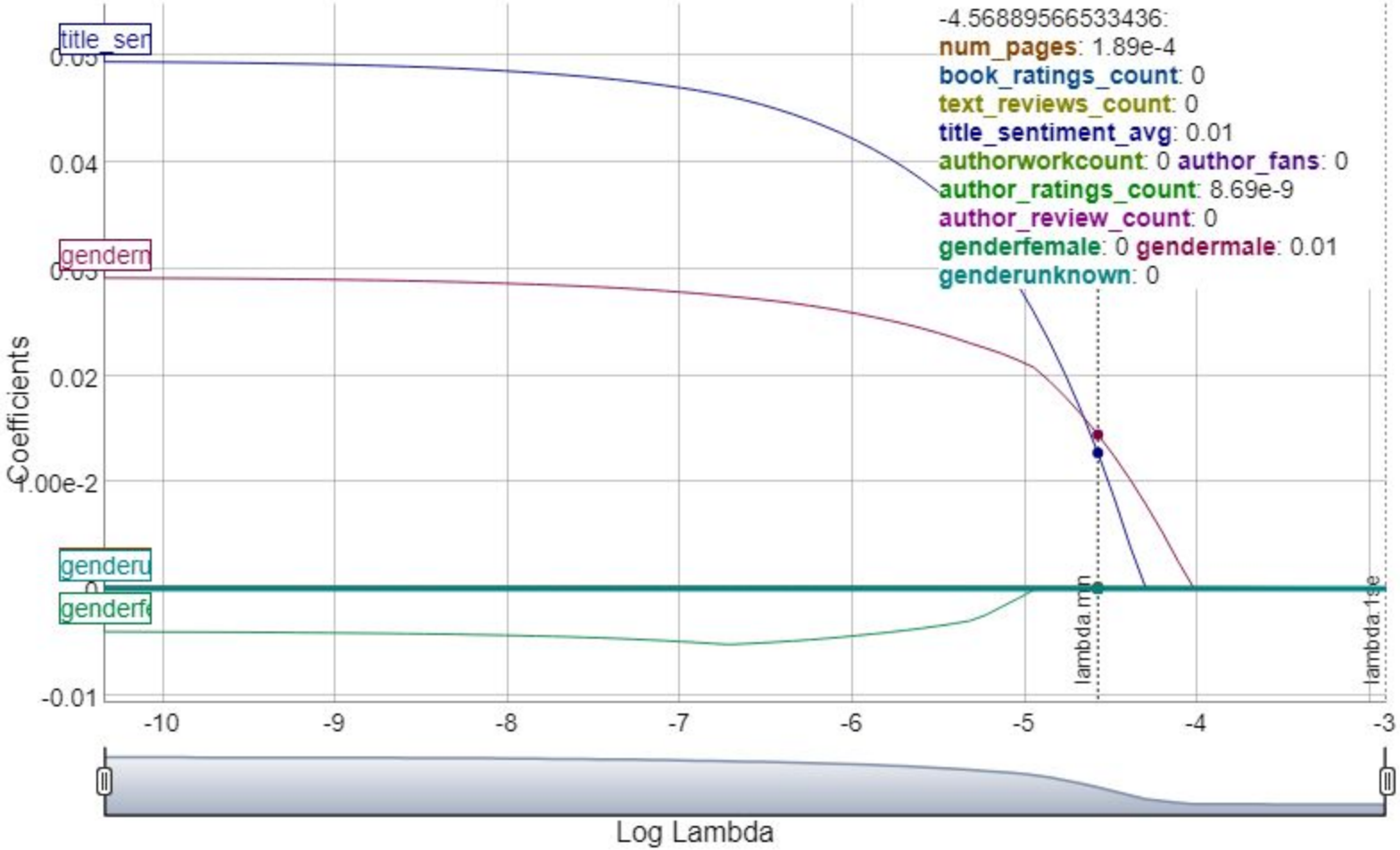
Testing Metrics:

- RMSE: 0.257
- MAE: 0.198
- RSQ: 0.079

Pictured left:

- Lasso plot with lambda.min and lambda.1se value




-4.56889566533436:
num_pages: 1.89e-4
book_ratings_count: 0
text_reviews_count: 0
title_sentiment_avg: 0.01
authorworkcount: 0 author_fans: 0
author_ratings_count: 8.69e-9
author_review_count: 0
genderfemale: 0 gendermale: 0.01
genderunknown: 0
title_ser
gendern
genderu
genderfe
0.04
0.02
4.00e-2
-0.01
Coefficients
-10
-9
-8
-7
-6
-5
-4
-3
lambda.min
lambda.1se
Log Lambda

Model 4: Random Forest

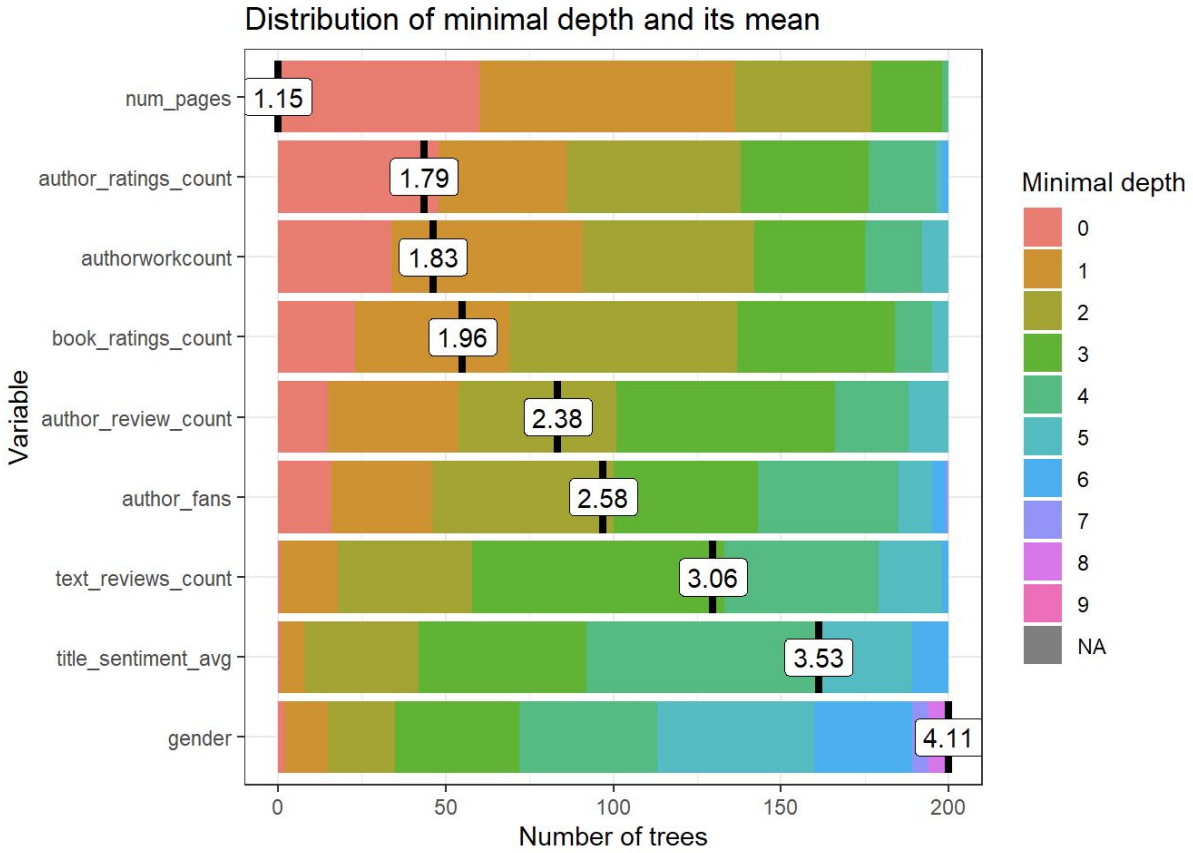
Why a random forest model?

Metrics:

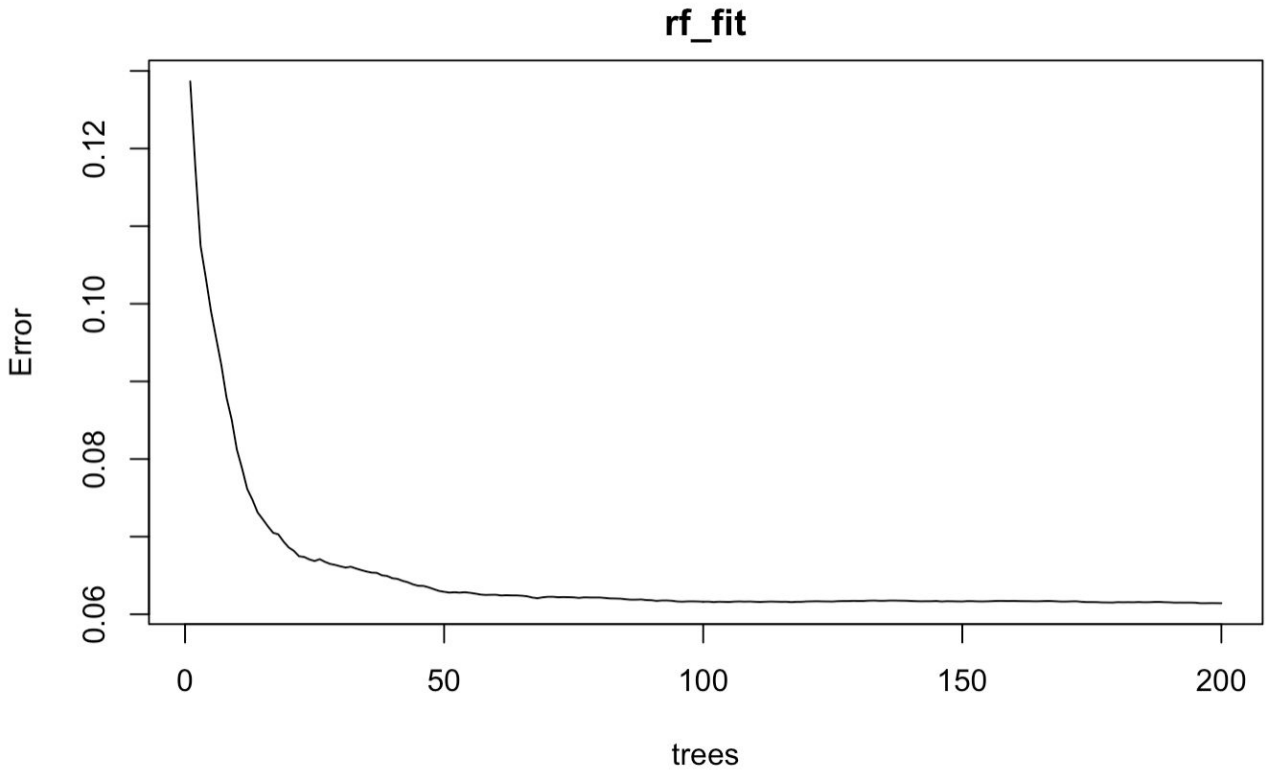
- RMSE: 0.249
- MAE: 0.180
- RSQ: 0.252

Pictured left:

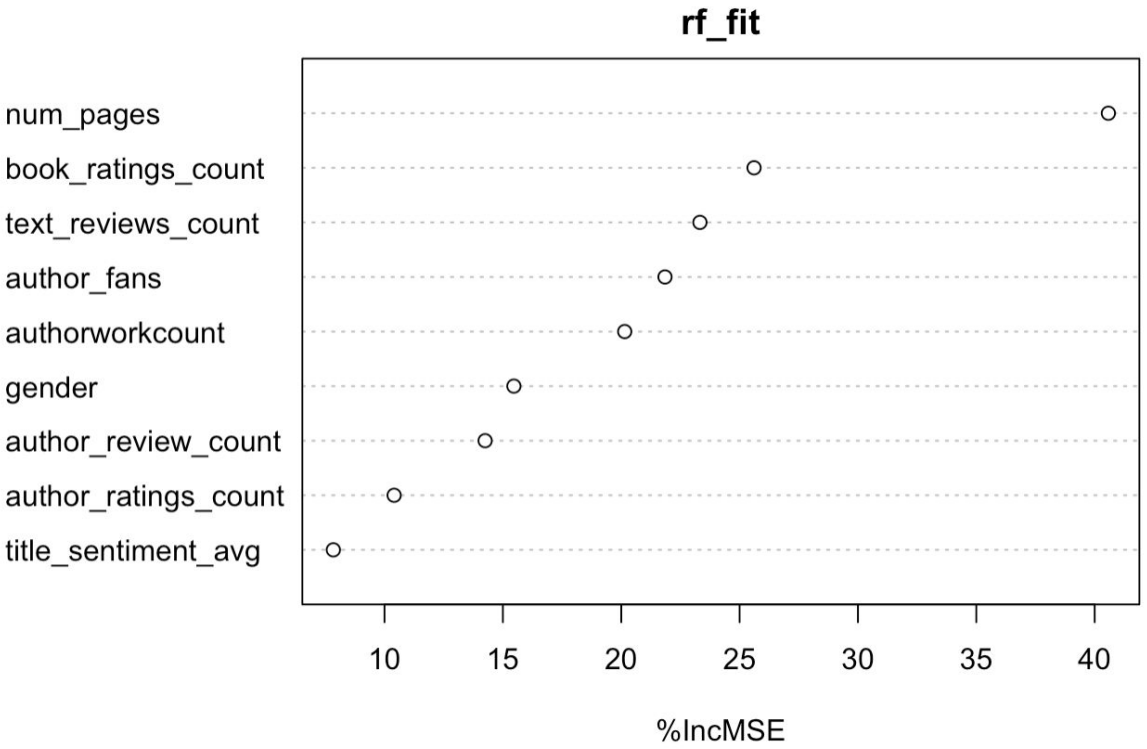
- Random Forest Explainer package



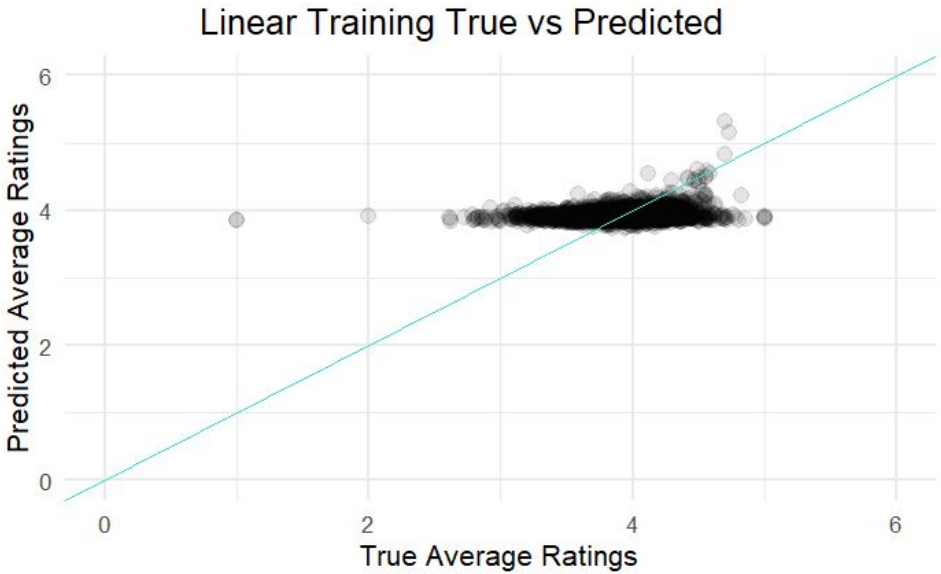
Model 4: Analysis

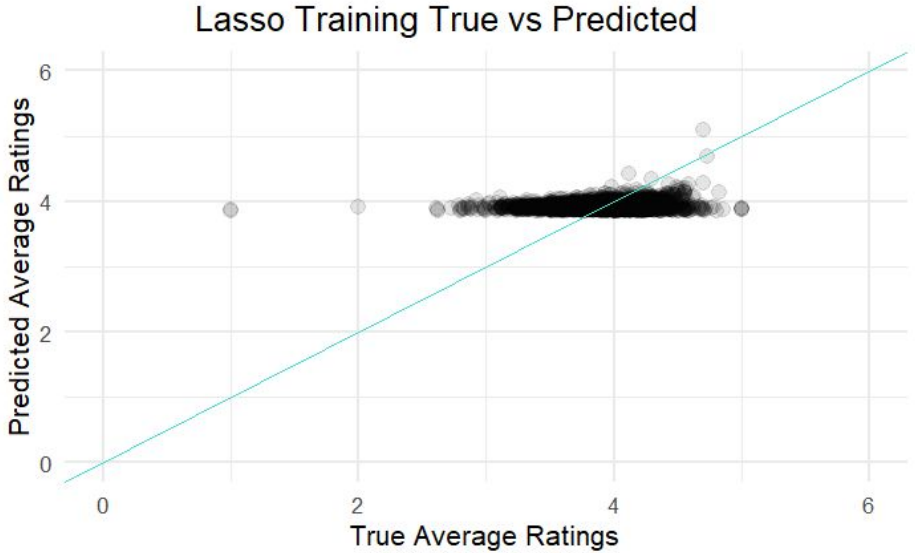


Model 4: Analysis

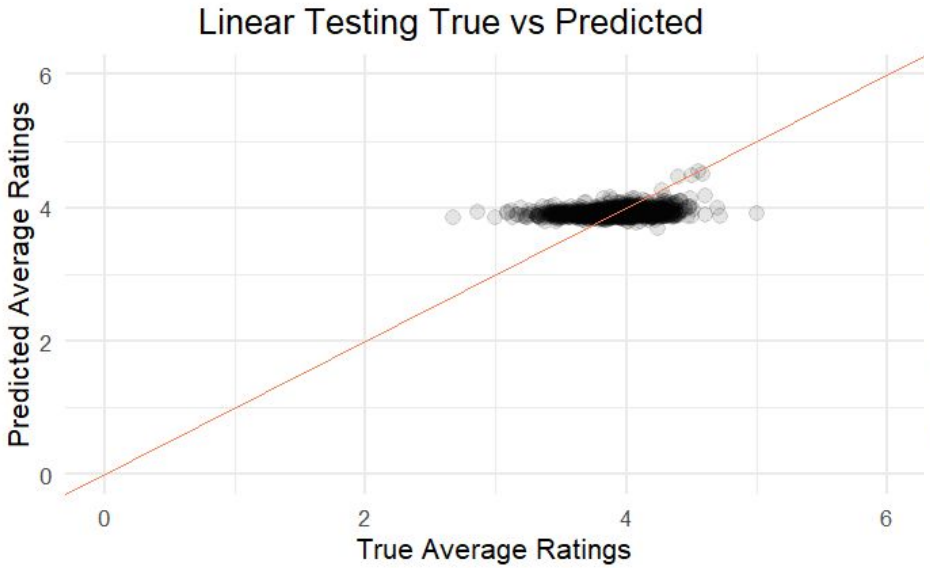


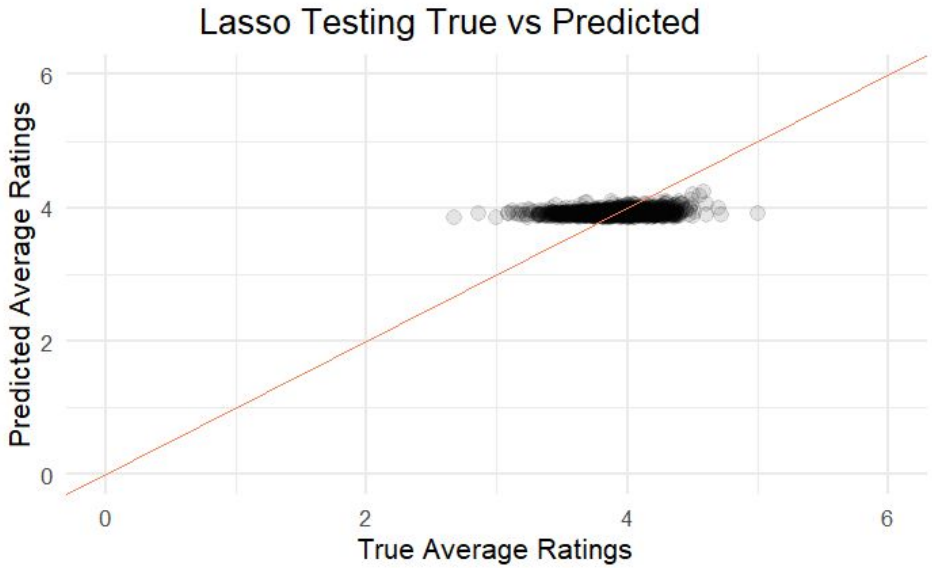
Training Comparisons





Testing Comparisons





Performance Comparisons

Metrics	Random Forest	Lasso	Linear	Best Value
rmse	0.2478224	0.25836801	0.2536445	0.2478224
rsq	0.2515691	0.07359474	0.0889949	0.2515691
mae	0.1797522	0.19916348	0.1950015	0.1797522

What went wrong, and why...

What?

- We find, at the end of our analysis, very low scores on just about all metrics
- The data is not compatible with our model types
- Essentially, minimal, if any, relationships are visible between our predictors

Why?

- We suspect this is due to the outcome variable being an average
- Because of this, all average data points become similar
- Thus, minimal relationships are visible, if any

From a Business Perspective

- Models are **not ready** for production.
- Data set is limited in its scope.
- Predictive inferences are missing sales statistics.
- Variables do not account for **when** consumers leave a rating.

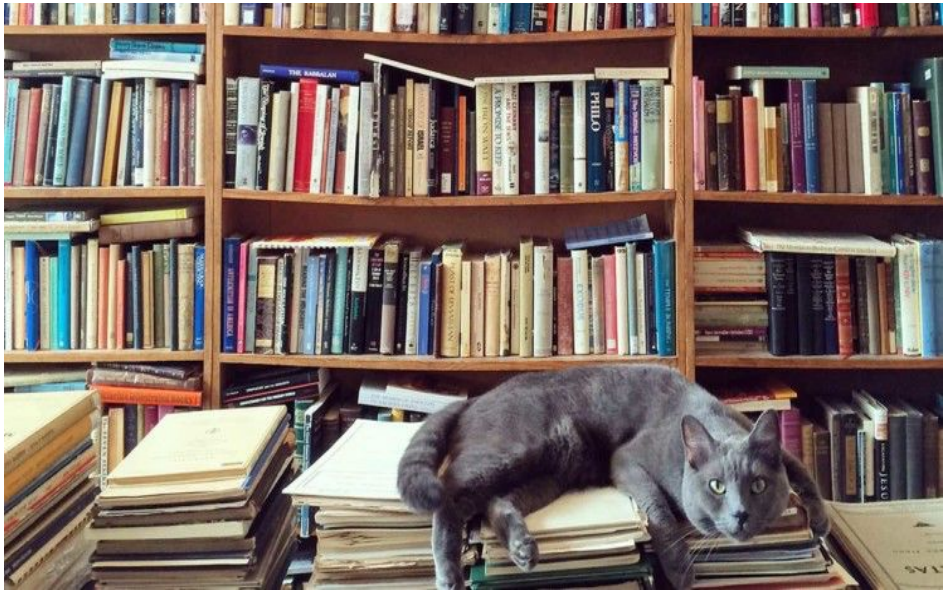
What Could Improve the Data?

This data set is limited:

- Includes only information on authors & publishers

Potential improvements:

- A different data set
  - (more information about consumers)
- Adding more variables to our existing data set



Questions?