Predicting the Difference in Goodreads Ratings and Amazon Ratings

By: Tien Nguyen & Tu Lam

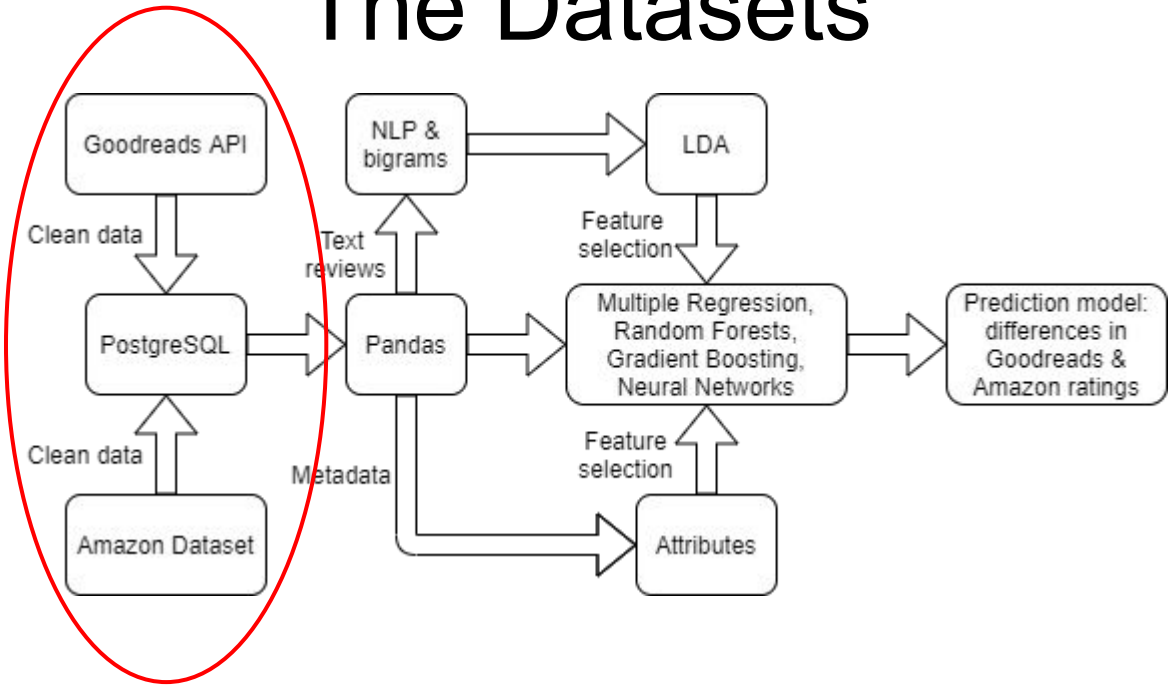
Project Objective

- Motivation: Goodreads and Amazon are both well-established sources of book reviews. Should a user look at Goodreads or Amazon for book recommendations?
- Problem: Goodreads and Amazon have different user base and different star rating interpretation
- Goal: Predict differences in ratings between Goodreads and Amazon so that users can have a better idea of a book's ratings when comparing it between the two platforms

Star Rating	Amazon Interpretation	Goodreads Interpretation
5	I love it	It was amazing
4	I like it	Really like it
3	It's okay	Liked it
2	I don't like it	It was ok
1	I hate it	Did not like it

Differences in star rating interpretation between Amazon and Goodreads

The Datasets



Amazon Datasets

Amazon reviews dataset :
+ 5,683,680 reviews (rows)
+ 6 features

Amazon metadata dataset :
+ 37,233 books
+ 10 features

overall	reviewTime	asin	count_before	count_after	reviewText	cleaned_text
5.0	08 12 2005	1713353	23	7	This book is a winner with both of my boys. T...	book winner boys enjoy picture story classic
5.0	03 30 2005	1713353	170	81	The King, the Mice and the Cheese by Nancy Gur...	king mouse cheese nancy gurney excellent child...
5.0	04 4 2004	1713353	55	27	My daughter got her first copy from her great-...	daughter get first copy greatgrandmother fathe...
5.0	02 21 2004	1713353	80	32	I remember this book from when I was a child a...	remember book child year ago remember wonderfu...
5.0	10 3 2016	1713353	31	13	Just as I remembered it, one of my favorites f...	remember one favorites childhood great conditi...

Amazon reviews dataset

asin	average	rating_count	text_reviews_count	genres	rank	verifiedTrue_count	Format	am_countText_before	am_countText_after
1713353	4.83	54	54	Childrens Books, Literature & Fiction	1461315	36	Paperback, Hardcover	2362	1037
1061240	4.87	45	45	Childrens Books, Literature & Fiction	321557	30	Hardcover	3085	1326
1711296	4.44	107	107	Literature & Fiction	2884610	69	Library Binding, VHS Tape, Paperback, Hard...	5667	2574
2007649	3.37	19	19	Science & Math, Chemistry	9799524	3	Kindle Edition, Paperback, Hardcover	5668	2810
1716069	4.61	59	59	Literature & Fiction, Poetry	3841172	44	Kindle Edition, Paperback, Hardcover	3081	1457

Amazon metadata dataset

Goodreads Datasets

Goodreads reviews dataset:

+ 906,876 reviews
+ 6 features

Goodreads metadata dataset:

+ 37,233 books
+ 21 features

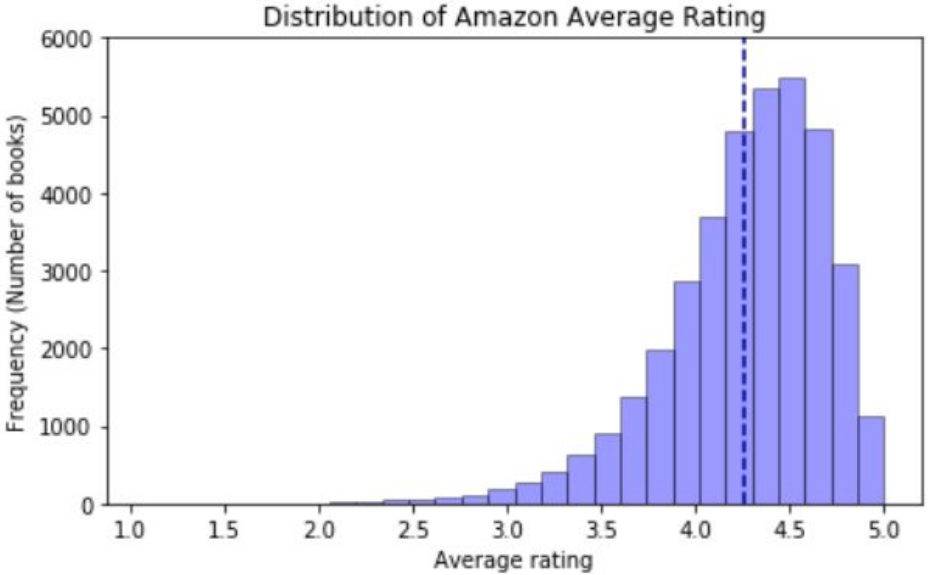
overall	reviewTime	asin	count_before	count_after	reviewText	cleaned_text
4	Dec 14 2016	0307408868	462	206	Another hard to put down nonfiction book from ...	another hard put nonfiction book eric arson en...
5	Dec 21 2016	0062273205	1328	619	I haven't read many (any?) books that are writ...	read many book write leos leo ceo aspire ceo m...
0	Mar 20 2014	006073731X	4	3	Sacca and Nate recommend	sicca name recommend
5	Dec 21 2016	0071424911	943	397	A truly inspirational book by a truly inspirat...	truly inspirational book truly inspirational m...
3	Aug 05 2012	0062041266	347	152	A fun dark slightly comical western about tw...	fun dark slightly comical western two killer c...

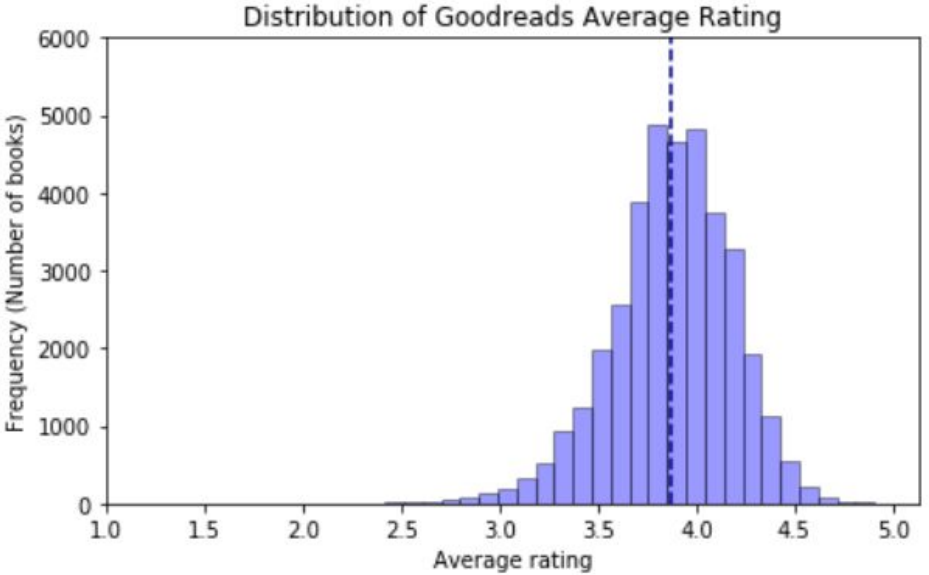
Goodreads reviews dataset

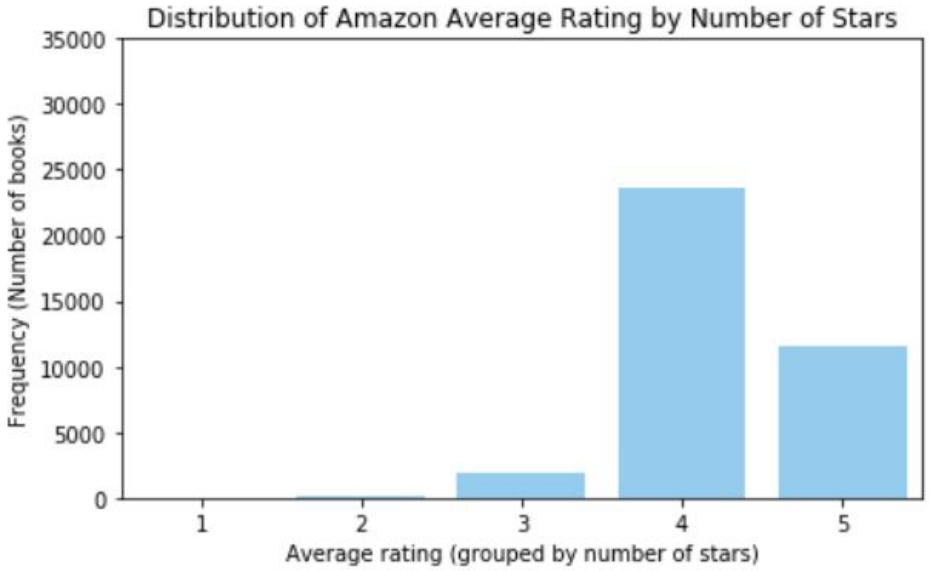
asin	average_	ratings_	reviews_	text_reviews_	total_ratings_	total_reviews_	total_text_	publication_ye	publication_mc	publication_da	publisher	num_pages	forma	descr	cleaned_desc	gr_countD	gr_countDes	cleaned_genres	gr_countText_	gr_countText
00010003	4.23	2E+05	163625	5535	220088	196528	8847	2010	1	1	Rupa & Co	127	Paper	Kahli	tahsil vibrant	106	66	poetry, fiction, n	42320	17834
1053655	4.08	16	33	6	676	1552	85	1997			HarperColl	268	Hardcover					history, historica	158	75
1061240	4.62	10	22	2	221	603	36	1959	12	1	Western P	324	Hardcover					poetry, children	49	18
00016110	3.86	33	74	4	2929	5786	75					190		The s	snobby girl fa	47	25	children, fiction,	130	61
1711296	4.29	604	1319	48	738	1564	65				Random H	63						children, fiction,	257	117
1712691	4.22	40	101	2	2387	3130	62	1982	4	1	HarperColl	28	Paper	A you	young bear y	34	19	children, fiction	83	36
1712713	4.14	50	105	5	2883	4522	141	1991	8	8	Beginner B	32	Paper	Illus.	illus full color	33	18	children, fiction	47	28
1712764	4.15	72	89	4	14445	19835	203	2007	9	25	Random H	36	Hard	Wher	dr sus take u	58	8	children, fiction,	49	22

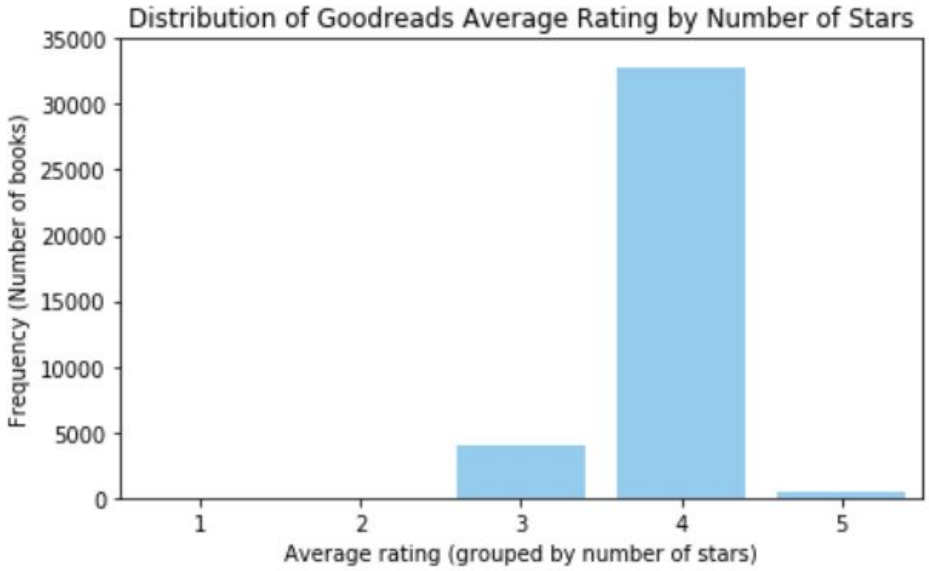
Goodreads metadata dataset

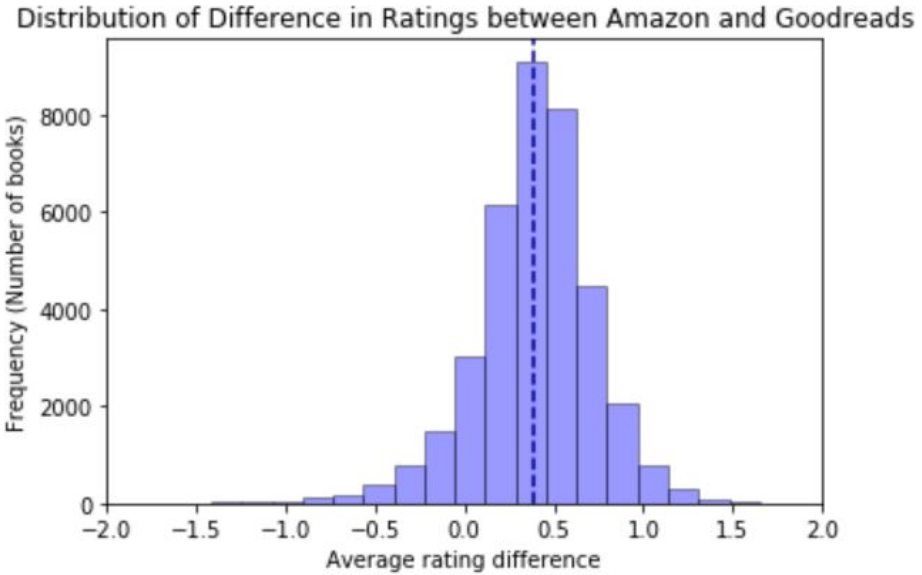
Exploratory Data Analysis (EDA)

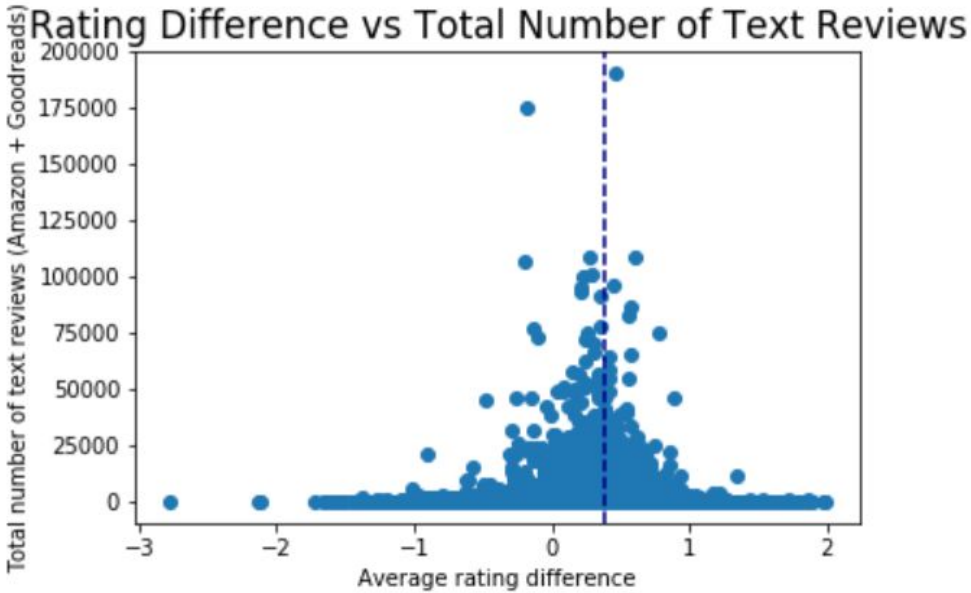

Distribution of Amazon Average Rating
Frequency (Number of books)
Average rating
0
1000
2000
3000
4000
5000
6000
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0


Distribution of Goodreads Average Rating
Frequency (Number of books)
Average rating
0
1000
2000
3000
4000
5000
6000
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0


Distribution of Amazon Average Rating by Number of Stars
Frequency (Number of books)
35000
30000
25000
20000
15000
10000
5000
0
1
2
3
4
5
Average rating (grouped by number of stars)


Distribution of Goodreads Average Rating by Number of Stars
Frequency (Number of books)
35000
30000
25000
20000
15000
10000
5000
0
1
2
3
4
5
Average rating (grouped by number of stars)


Distribution of Difference in Ratings between Amazon and Goodreads
Frequency (Number of books)
0
2000
4000
6000
8000
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
Average rating difference

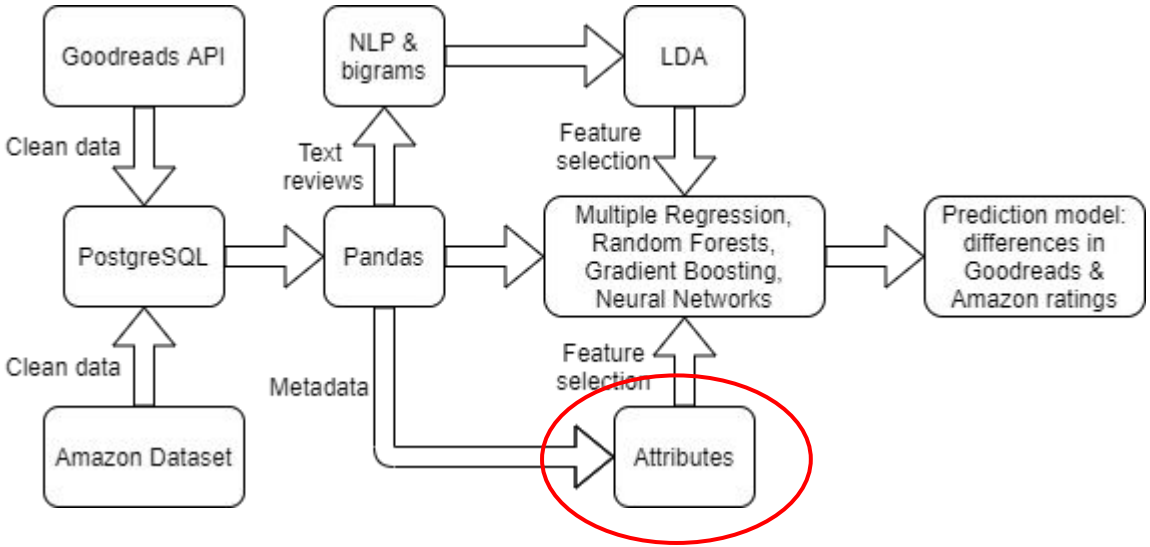

Rating Difference vs Total Number of Text Reviews
Total number of text reviews (Amazon + Goodreads)
0
25000
50000
75000
100000
125000
150000
175000
200000
−3
−2
−1
0
1
2
Average rating difference

Data Modeling

Machine Learning Algorithms

- Numeric and categorical data to predict
- Multiple Regression
  - Assumptions:
    - Linearity ❌
    - Normality of the Error Terms ❌
    - No Autocorrelation of the Error Terms ✅
    - Homoscedasticity ✅
    - No Multicollinearity among Predictors ✅
- Random Forest
  - Can handle many predictor variables
  - Can handle skewed and multi-modal data
  - Can handle categorical data (e.g., one-hot encoding)
  - No formal assumptions
- XgBoost
  - Can handle missing values
  - Can handle categorical data (e.g., one-hot encoding)
  - Accurate and fast (parallelization)
- Neural Network
  - Doesn't have assumptions about normality, linearity, variable independence, etc.
  - Can handle categorical data (e.g., one-hot encoding)
  - Can capture complex patterns in data
- Evaluation：MAE, MSE, RMSE, and R2

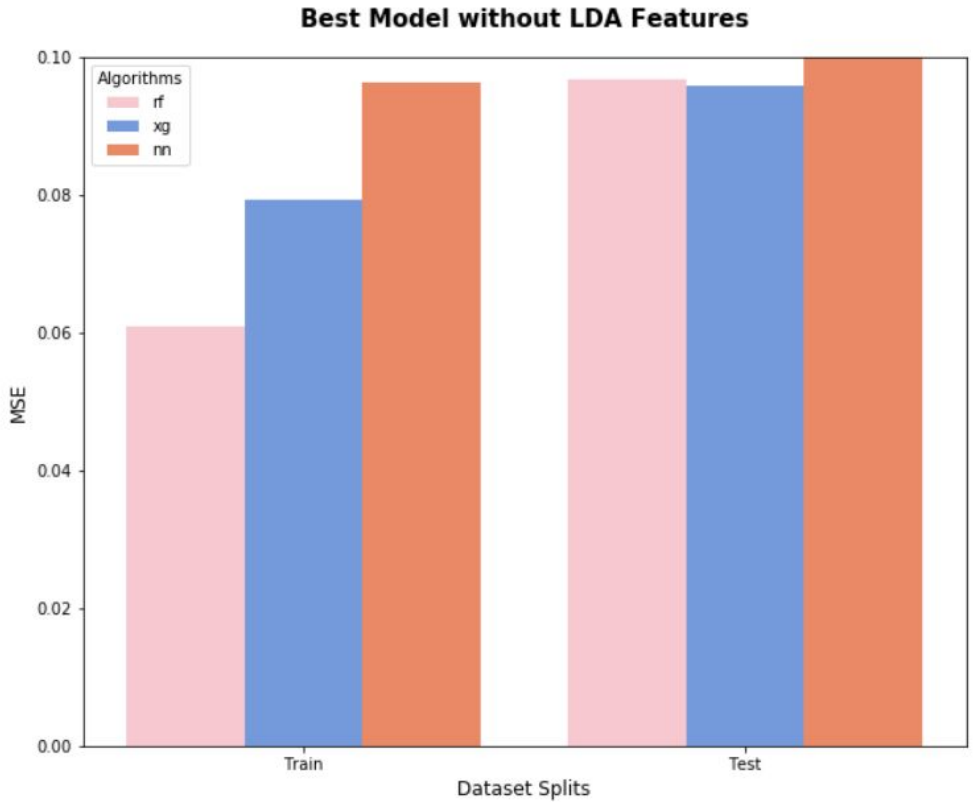
Data Modeling (Basic Features)



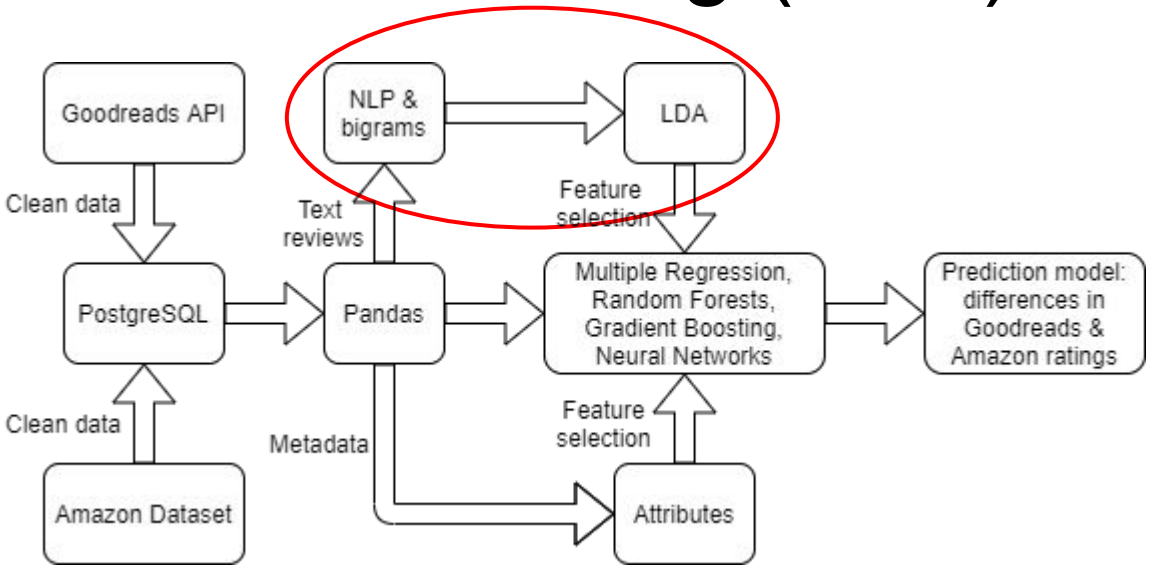
Data Modeling (Basic Features)

- Features selection
  - Numerical features
    - Ratings count, reviews count, rank, reviews word counts before/after cleaning, etc.
    - Data cleaning: combining columns
    - Null values: delete rows, replace w/ mean or median, impute?
  - Categorical features
    - Format, publisher, genres
    - One-hot encoding
- Hyperparameters tuning using Grid Search Cross Validation

Data Modeling



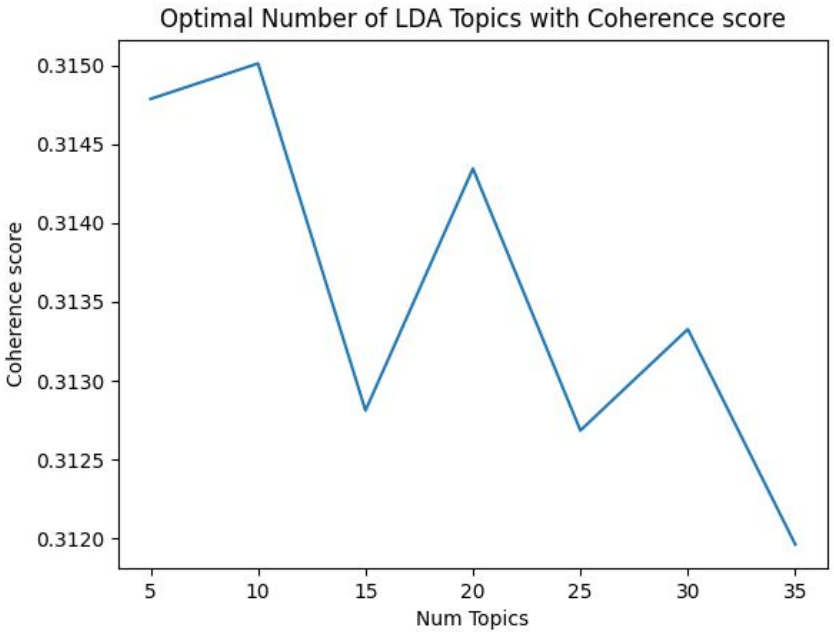
Data Modeling (LDA)



Topic Modeling : Latent Dirichlet Allocation (LDA)

- Gensim package
- Mallet's implementation (via Gensim)
- Input : review texts of each book
- Output : number of the proportion under each topic
- Evaluate Topic Models : CV Coherence Score
  - Coherence Score : similarity between semantic meaning and statistically derived weights of the highest score words.
  - CV coherence score has the highest correlation between human ranked topic and its ranking.

Optimal Number of Topics for LDA



Gensim's inbuilt version of the LDA algorithm

- Number of Topic : 10
- Coherence Score : 0.315

Mallet's LDA from within Gensim itself

- Coherence Score : 0.337

Topic Modeling (LDA) Continued

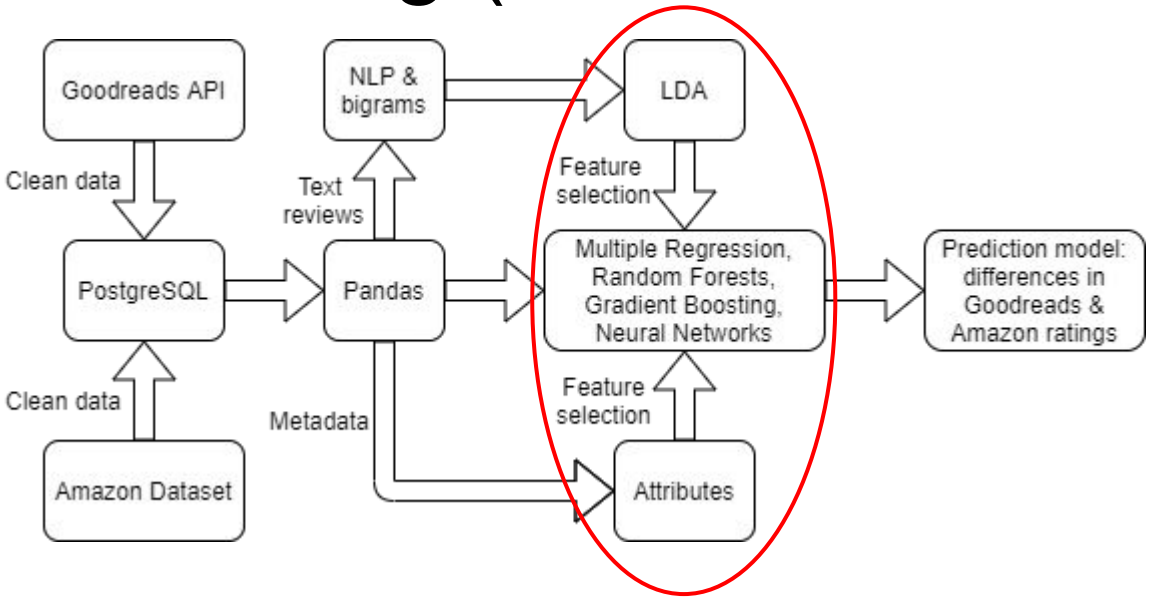
```
(0, '0.027*"book" + 0.023*"vampire" + 0.016*"love" + 0.009*"read" + 0.008*"adam" + 0.007*"anna" + 0.007*"make" + 0.007*"story"')
(1, '0.033*"book" + 0.017*"read" + 0.010*"people" + 0.009*"make" + 0.008*"write" + 0.006*"bad" + 0.006*"case" + 0.006*"story"')
(2, '0.023*"book" + 0.012*"fan" + 0.012*"life" + 0.011*"great" + 0.011*"music" + 0.009*"read" + 0.008*"write" + 0.007*"time"')
(3, '0.033*"read" + 0.020*"classic" + 0.018*"book" + 0.015*"great" + 0.013*"movie" + 0.013*"time" + 0.010*"version" + 0.010*"story"')
(4, '0.096*"series" + 0.091*"book" + 0.035*"read" + 0.028*"love" + 0.017*"character" + 0.017*"great" + 0.016*"good" + 0.014*"wait"')
(5, '0.038*"god" + 0.021*"book" + 0.020*"christian" + 0.014*"church" + 0.014*"bible" + 0.014*"jesus" + 0.013*"faith" + 0.010*"religion"')
(6, '0.032*"book" + 0.013*"business" + 0.013*"read" + 0.012*"make" + 0.012*"work" + 0.009*"people" + 0.008*"good" + 0.008*"company"')
(7, '0.073*"book" + 0.032*"read" + 0.031*"life" + 0.011*"great" + 0.010*"make" + 0.010*"time" + 0.010*"give" + 0.010*"good"')
(8, '0.015*"write" + 0.011*"life" + 0.008*"time" + 0.008*"work" + 0.008*"character" + 0.007*"reader" + 0.007*"writer" + 0.007*"story"')
(9, '0.041*"book" + 0.018*"read" + 0.015*"world" + 0.010*"game" + 0.009*"character" + 0.009*"make" + 0.009*"story" + 0.008*"end"')
```

Dot product of weight and frequency words

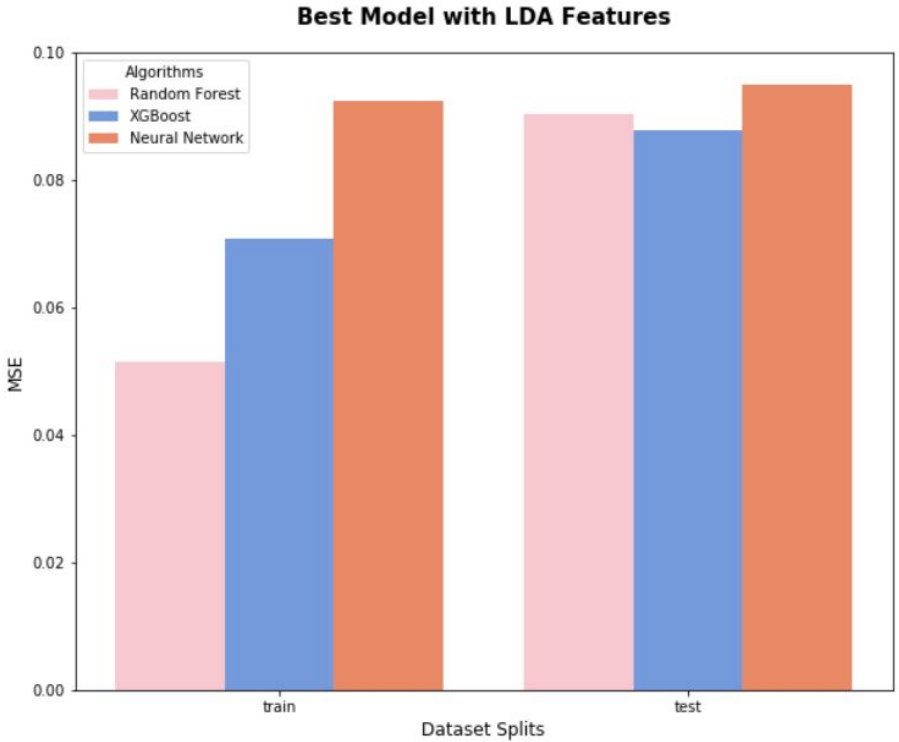
asin	prop_topic_0	prop_topic_1	prop_topic_2	prop_topic_3	prop_topic_4	prop_topic_5	prop_topic_6	prop_topic_7	prop_topic_8	prop_topic_9
0001713353	0.002965	0.006142	0.003064	0.009717	0.007532	0.002567	0.012497	0.037718	0.004156	0.007830
0001061240	0.002599	0.004051	0.002987	0.092998	0.006955	0.003470	0.003761	0.022731	0.036765	0.003858
0001711296	0.002122	0.004148	0.004668	0.009395	0.006226	0.003888	0.003317	0.032458	0.008616	0.005083
0002007649	0.001322	0.006238	0.002341	0.003492	0.004599	0.004865	0.019789	0.005529	0.002828	0.002164
0001716069	0.003595	0.008401	0.004867	0.016812	0.004443	0.005715	0.011723	0.022608	0.010097	0.004231

Combining Basic Features with LDA Features

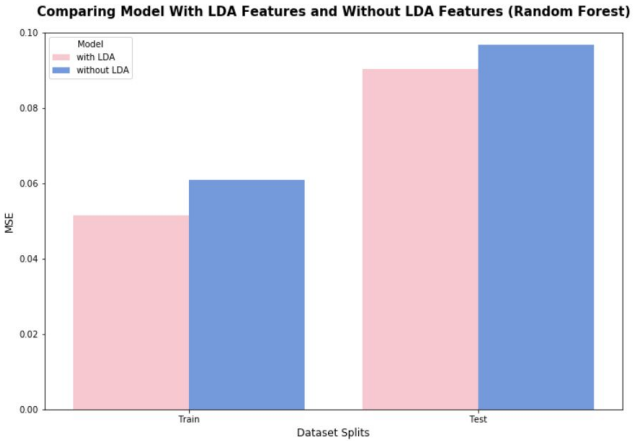
Data Modeling (Prediction Models)

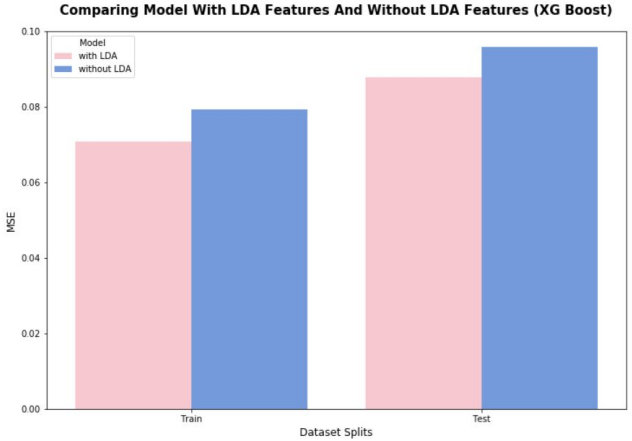


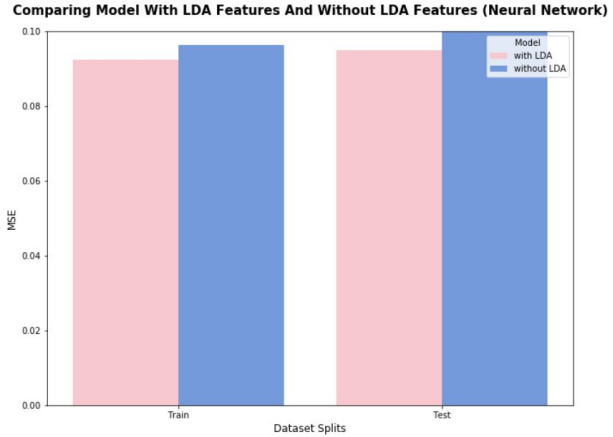
Data Modeling



Data Modeling







Conclusion

- The rating difference between Amazon and Goodreads, on average: 0.4
- Our errors are small because the range of our prediction--difference in ratings between Goodreads and Amazon--is small (-4, 4)
- Text review features play an important role.
- Best models overall: Xgboost & Neural Network

Future Work

- Combining Random Forest, Xgboost, and Neural Network algorithms into an ensemble to get better errors
- Predict Amazon ratings based on features from Goodreads
  - LDA on Goodreads description
- Predict Goodreads ratings based on features from Amazon
- Compare various ratings difference prediction models (model using both Amazon and Goodread features, just Goodread features, or just Amazon features) and find the best one

QUESTIONS?

SUGGESTIONS?