

Predicting User Ratings from Review Text

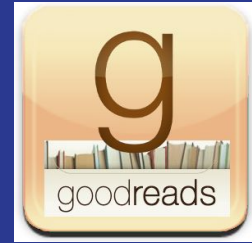
A Natural Language Processing & Machine Learning Case Study

Springboard Career Track Capstone 1
Heather A. Passmore

Problem

Customer ratings influence sales. Rating errors have big effects.

What companies care?



Problem

Customer ratings influence sales. Rating errors have big effects.

- On average customers read more than two Yelp reviews before deciding to use a business.
- A one-star increase in ratings on Yelp leads to a 5-9% increase in revenue for a business.

User Reviews and Ratings



Ratings matter

Consumers & vendors depend on ratings for:

- Purchase decisions
- Predicting sales

High and low ratings affect sales directly.

Predict rating

Benefits of predicting ratings from text:

- Error catching
- Suggest rating
- Improve consistency
- Flag problematic ratings

Motivation

Build a system to identify positive and negative feedback from customers to give businesses the power to intercede and to improve customer engagement and satisfaction.

Top customer reviews



triple0triple1



Get the first edition.

August 29, 2013

Format: Hardcover | **Verified Purchase**

I have just bought this brand new 3rd edition and compared it with my 1st edition. They are almost exactly the same, even down to the page numbers. This is an old scam by textbook publishers, but usually they will at least change around the table of contents and put in new diagrams, photos and tables. Not so with this title.

And if the 1st and 3rd editions are almost exact copies, then I'm sure the 2nd edition would be fine too.

Also, do not think that you need to buy the "Custom Edition" for your school, it's just the campus bookstore's and publisher's way of keeping you off Amazon and charging you double for an already overpriced book.

[Comment](#)

181 people found this helpful. Was this review helpful to you?

Yes

No

[Report abuse](#)



J.T884



Beware! Ebook has false advertising about including Mastering Chemistry

November 7, 2016

Format: Kindle Edition | **Verified Purchase**

[Chemistry: A Molecular Approach](#)

Even though it says twice in the description that the ebook comes with Mastering Chemistry, and it is the first result with an ISBN search that includes Mastering Chemistry, it actually doesn't contain Mastering Chemistry. False advertising at its finest. Chatted with support and all they would do is refund my money so I can order a hardcover. I can't go without a text while in a class. Very bad customer service.

[Comment](#)

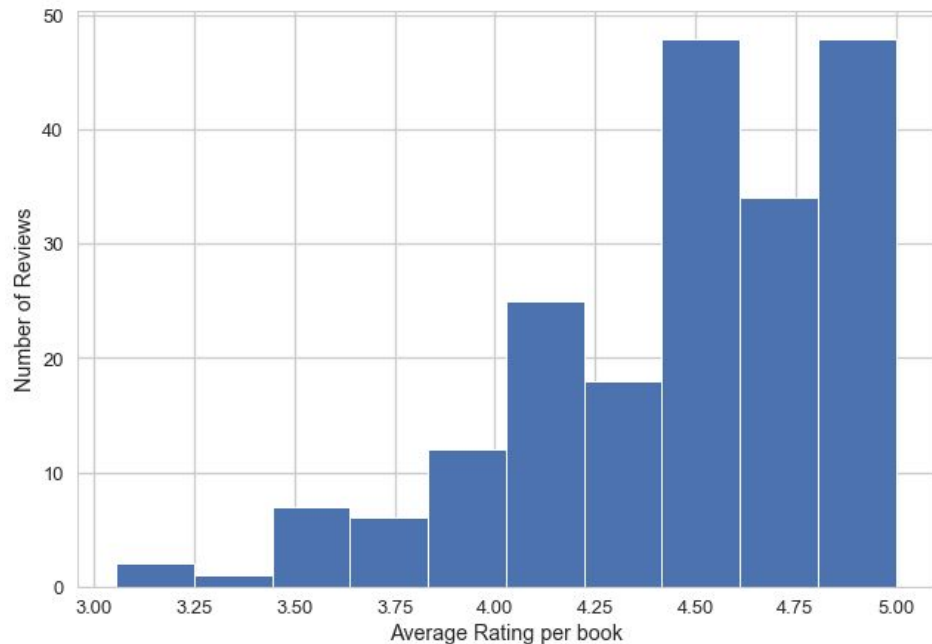
7 people found this helpful. Was this review helpful to you?

Yes

No

[Report abuse](#)

Reviewers love books! (most ratings are 'high')



- Majority of books reviewed by more than 10 reviewers have mean ratings > 4 stars.
- This Science Textbook genre subset of reviews includes reviews for 729 different books in 11546 reviews.
- Binary categories: 'low' ratings ($n = 3884$) and 'high' ratings ($n = 7662$).
- Word counts are significantly higher for reviews with 'low' ratings.

Machine Learning Highlights

Preprocessing

Specialized stopwords

Remove 'english' stopwords except:

- Leave 'not' in corpus for bigrams
- Remove 'book', books'

Vectorization

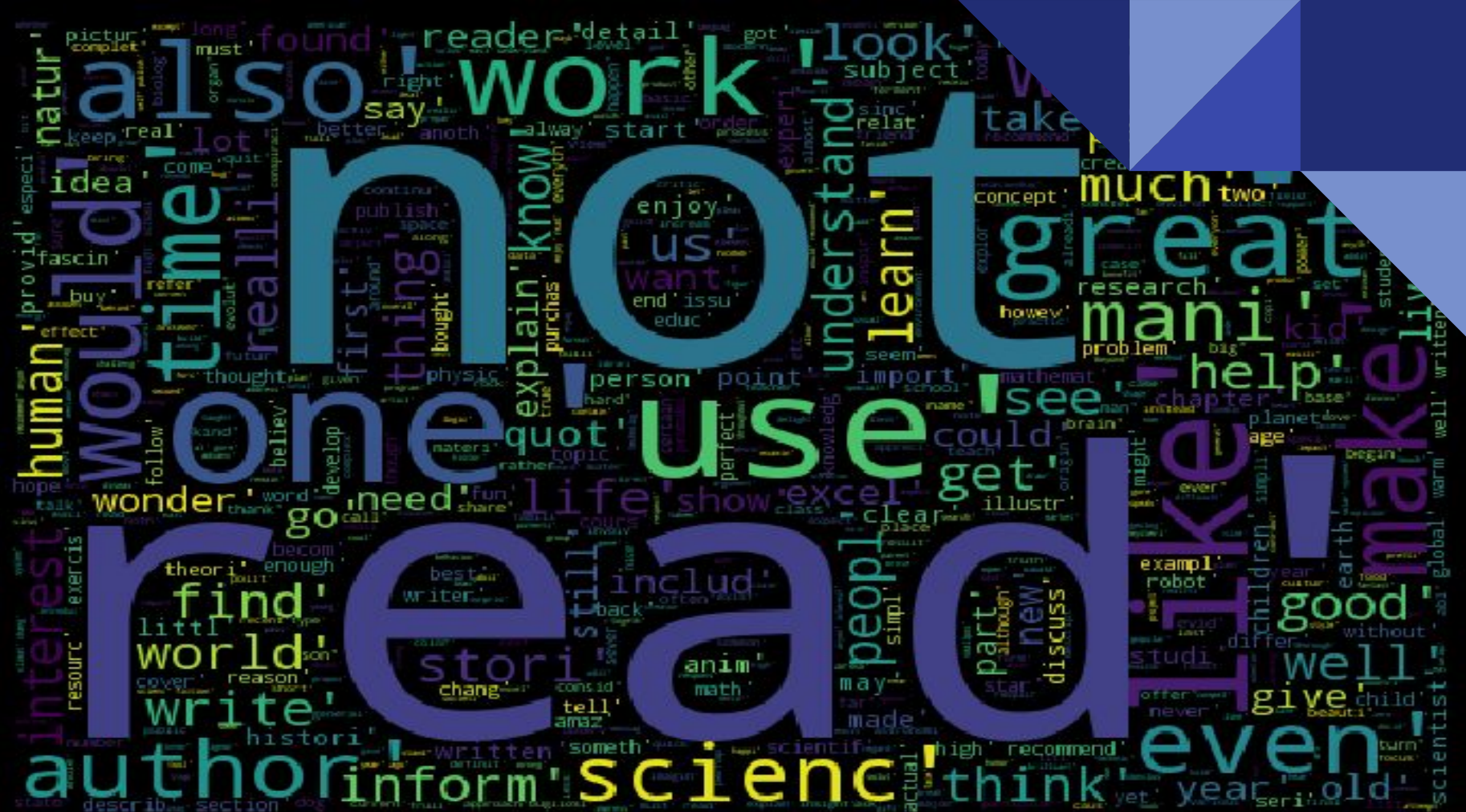
TF-IDF Vectorizer

- Features from single words and bigrams
- Baseline ROC-AUC score = 0.78 with Multinomial Naive Bayes classifier.

Model Tuning

Important Parameters

- Set class_weight to 'balanced' since target classes imbalanced.
- Tune with GridSearchCV.
- Compare ROC-AUC scores.



Classifier Comparisons

Classifier:	ROC-AUC:	GridSearchCV best params:
MultinomialNB	0.78	alpha=1, fit_prior=True
RandomForestClassifier	0.76	max_features=750, min_samples_leaf=8
LogisticRegressionCV	0.81	Cs=10, max_iter=100, tol=0.0001
SGDClassifier	0.75	alpha=0.1

Best Classifier

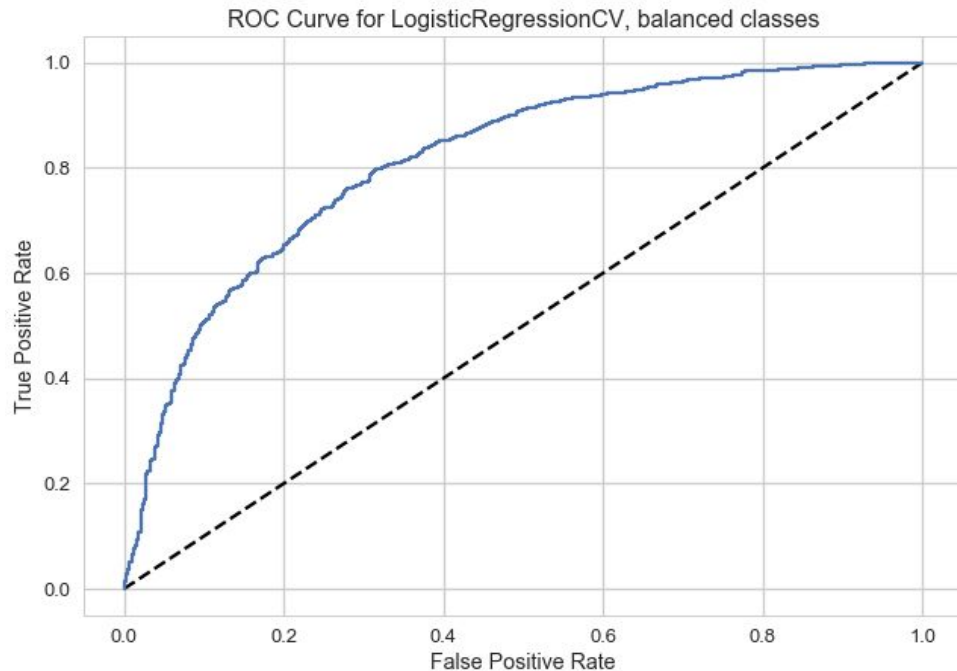
Logistic Regression with Cross
Validation

y=1 top features

Weight?	Feature
+5.262	excel
+4.659	wonder
+4.364	love
+4.188	great
+3.765	perfect
+3.193	everi
+3.129	conspiraci
+3.125	high
+3.076	high recommend
... 19630 more positive ...	
... 17289 more negative ...	
-3.064	lack
-3.121	page
-3.123	ok
-3.196	bore
-3.240	good
-3.807	unfortun
-3.847	would
-3.860	seem
-3.992	howev
-5.268	disappoint
-5.320	not

Best Classifier

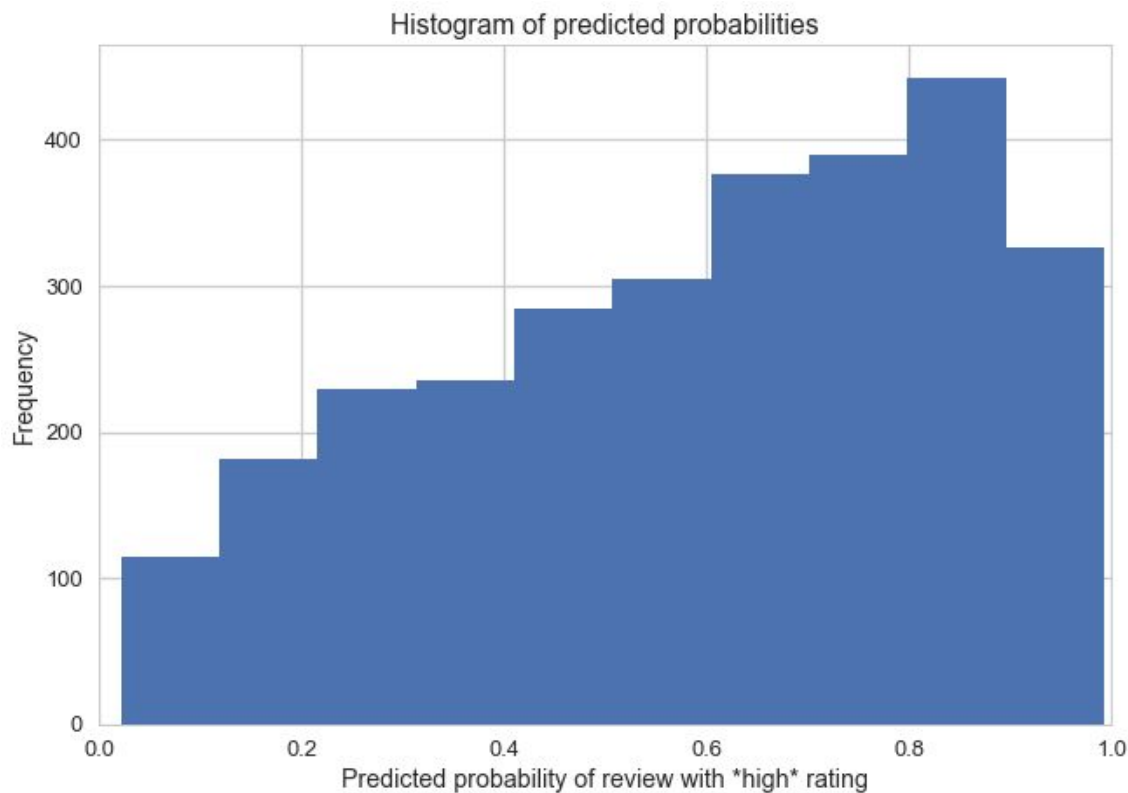
Logistic Regression with Cross
Validation

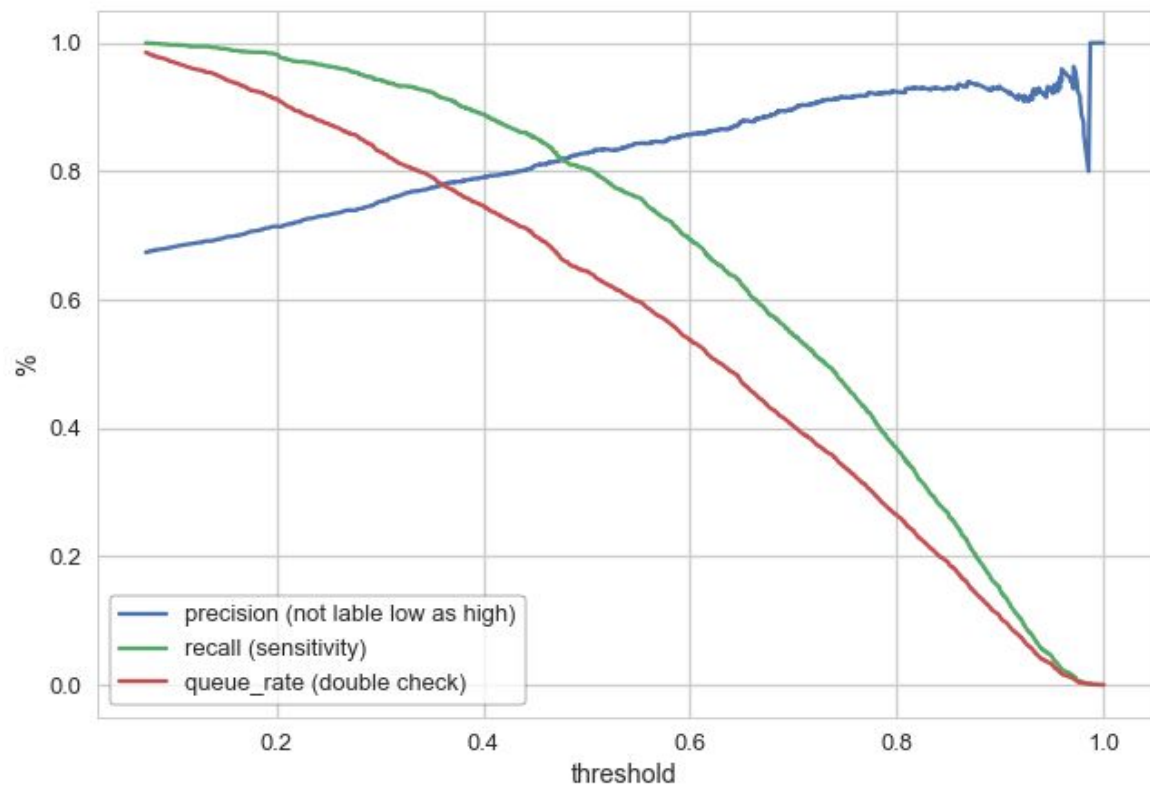




Improve classification

Distribution of predicted probabilities





[illegible]