Predicting Amazon Ratings Based on Review Text

Adithyan Subramanian
Mentor: Mukesh MIthrakumar

Background and Goals

- Amazon
- Sells over 100 million products
- 4000 orders placed every minute
- Based on the text of the reviews can we predict the rating of the review?

Data

- Data from data.world
- Product reviews of AmazonBasics products.
- 28,332 rows

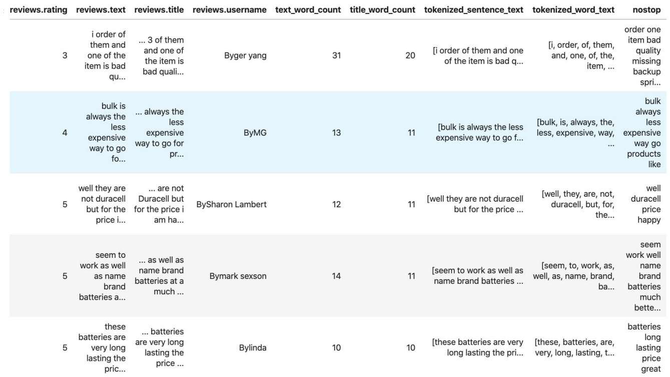
Data Wrangling and Cleaning

Cleaning

- Dropped irrevelant columns such as URLs
- No Nan values in relevant columns

Processing

- Used NLTK library
- Deleted numerical values, as well as periods and commas from text data
    - Chose to keep exclamation points and question marks
- Tokenized by sentence and words
- Removed stop words

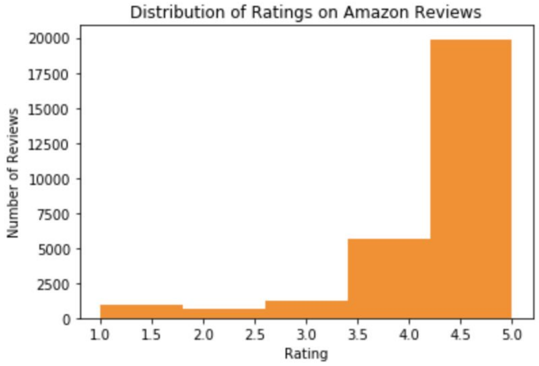


reviews.rating	reviews.text	reviews.title	reviews.username	text_word_count	title_word_count	tokenized_sentence_text	tokenized_word_text	nostop
3	i order of them and one of the item is bad qu...	... 3 of them and one of the item is bad quali...	Byger yang	31	20	[i order of them and one of the item is bad q...	[i, order, of, them, and, one, of, the, item, ...	order one item bad quality missing backup spri...
4	bulk is always the less expensive way to go fo...	... always the less expensive way to go for pr...	ByMG	13	11	[bulk is always the less expensive way to go f...	[bulk, is, always, the, less, expensive, way, ...	bulk always less expensive way go products like
5	well they are not duracell but for the price i...	... are not Duracell but for the price i am ha...	BySharon Lambert	12	11	[well they are not duracell but for the price ...	[well, they, are, not, duracell, but, for, the...	well duracell price happy
5	seem to work as well as name brand batteries a...	... as well as name brand batteries at a much ...	Bymark sexson	14	11	[seem to work as well as name brand batteries ...	[seem, to, work, as, well, as, name, brand, ba...	seem work well name brand batteries much bette...
5	these batteries are very long lasting the pric...	... batteries are very long lasting the price ...	Bylinda	10	10	[these batteries are very long lasting the pri...	[these, batteries, are, very, long, lasting, t...	batteries long lasting price great

Exploratory Data Analysis

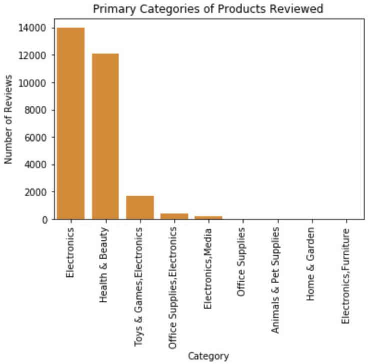
Distribution of ratings

- Data highly skewed to 5 star reviews
  - Roughly 20,000 reviews of 28,332 reviews were 5 star reviews
- 4 star ratings had the next highest number of reviews
  - 6,000 reviews
- Chose to assume that this distribution was indicative of the population



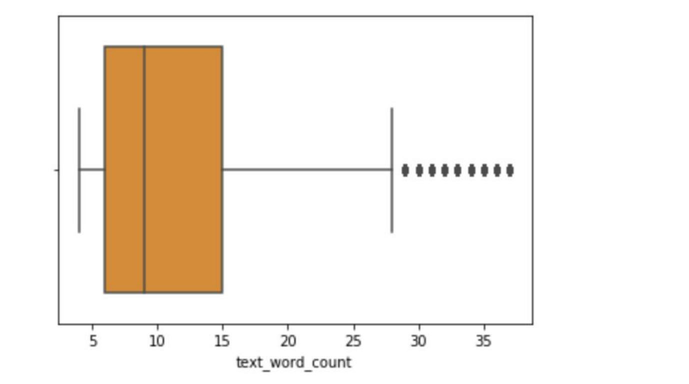
Categories of Products

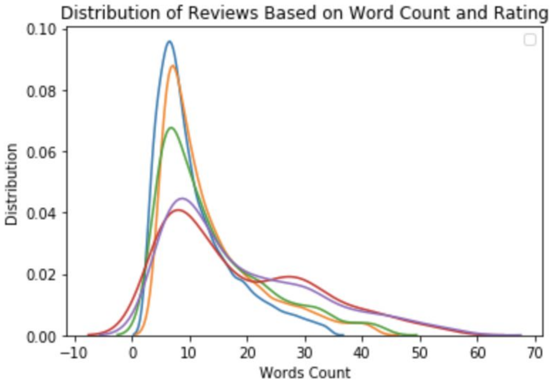
- Most products were electronics
- Health and beauty products also had a significant percentage



Word count

- Maximum word count was 719
- Average was close to eight
- There was no significant difference in word counts between each of the 5 ratings





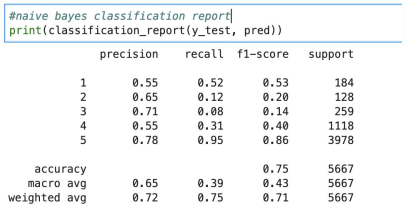
Machine Learning

Initial Processing

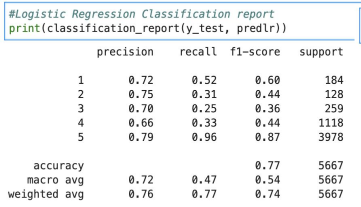
- Initialized countvectorizer and tfidfvectorizer
- Used vectorizers to create test and train sets (80/20 split)
- Converted sets to dataframes

Initial modeling

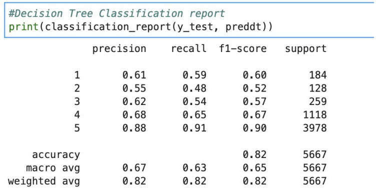
- Chose three initial models
  - Naive bayes
    - 0.75 accuracy
  - Logistic regression
    - 0.77 accuracy
  - Decision tree
    - 0.82 accuracy
- Chose to use decision tree



```
#naive bayes classification report
print(classification_report(y_test, pred))
              precision    recall  f1-score   support

           1       0.55      0.52      0.53       184
           2       0.65      0.12      0.20       128
           3       0.71      0.08      0.14       259
           4       0.55      0.31      0.40      1118
           5       0.78      0.95      0.86      3978

    accuracy                           0.75      5667
   macro avg       0.65      0.39      0.43      5667
weighted avg       0.72      0.75      0.71      5667
```



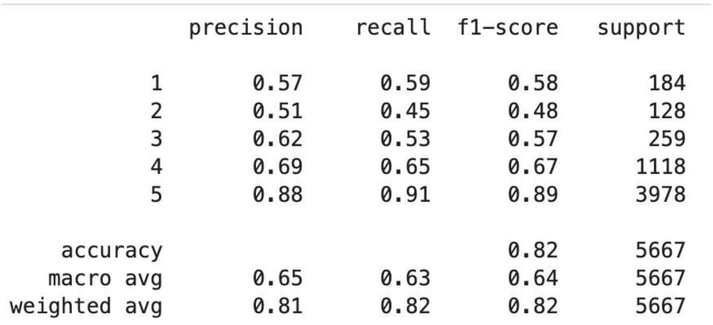
```
#Logistic Regression Classification report
print(classification_report(y_test, predlr))
              precision    recall  f1-score   support

           1       0.72      0.52      0.60       184
           2       0.75      0.31      0.44       128
           3       0.70      0.25      0.36       259
           4       0.66      0.33      0.44      1118
           5       0.79      0.96      0.87      3978

    accuracy                           0.77      5667
   macro avg       0.72      0.47      0.54      5667
weighted avg       0.76      0.77      0.74      5667
```



```
#Decision Tree Classification report
print(classification_report(y_test, preddt))
              precision    recall  f1-score   support

           1       0.61      0.59      0.60       184
           2       0.55      0.48      0.52       128
           3       0.62      0.54      0.57       259
           4       0.68      0.65      0.67      1118
           5       0.88      0.91      0.90      3978

    accuracy                           0.82      5667
   macro avg       0.67      0.63      0.65      5667
weighted avg       0.82      0.82      0.82      5667
```

Model optimization

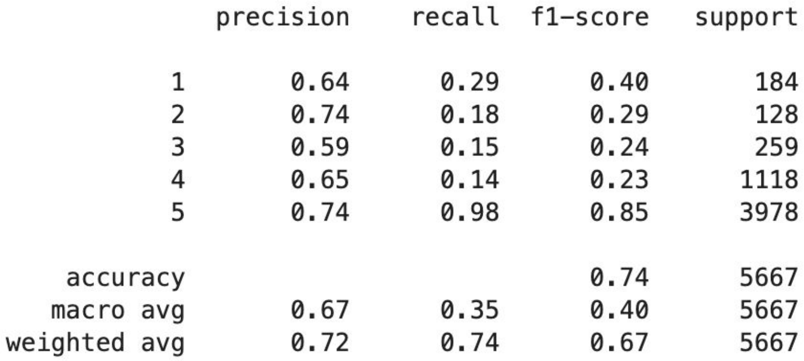
- Used gridsearchCV to determine best parameters
    - Max depth of 5000, and minimum samples split of 2
    - Training accuracy = 0.984
    - Testing accuracy = 0.818
- Concerns of overfitting



```
              precision    recall  f1-score   support

           1       0.57      0.59      0.58       184
           2       0.51      0.45      0.48       128
           3       0.62      0.53      0.57       259
           4       0.69      0.65      0.67      1118
           5       0.88      0.91      0.89      3978

    accuracy                           0.82      5667
   macro avg       0.65      0.63      0.64      5667
weighted avg       0.81      0.82      0.82      5667
```

Compensating for overfitting

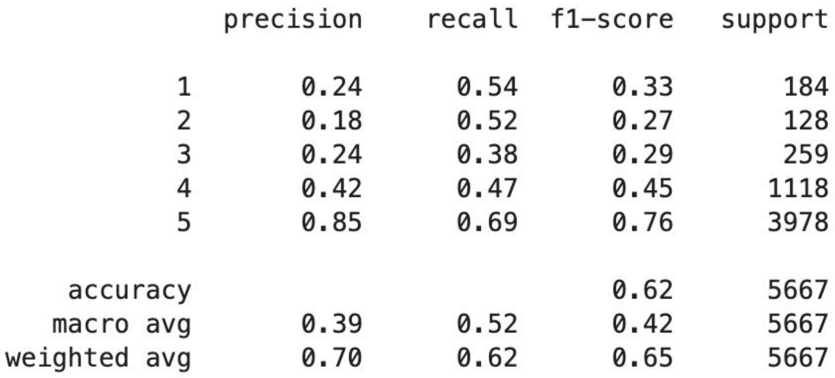
- Eventually settled on hyperparameters
  - Max depth = 20
  - Min samples split=2
- Training set accuracy = 0.775
- Testing set accuracy = 0.736
- Overfitting largely solved, but lower accuracy



```
              precision    recall  f1-score   support

           1       0.64      0.29      0.40       184
           2       0.74      0.18      0.29       128
           3       0.59      0.15      0.24       259
           4       0.65      0.14      0.23      1118
           5       0.74      0.98      0.85      3978

    accuracy                           0.74      5667
   macro avg       0.67      0.35      0.40      5667
weighted avg       0.72      0.74      0.67      5667
```

Oversampling

- Used SMOTE to oversample reviews with lower than a 5-star rating
- Training set accuracy = 0.627
- Testing set accuracy = 0.624
- Overfitting completely solved but accuracy extremely low
- Chose to recommend compensated decision tree model with no oversampling



```
              precision    recall  f1-score   support

           1       0.24      0.54      0.33       184
           2       0.18      0.52      0.27       128
           3       0.24      0.38      0.29       259
           4       0.42      0.47      0.45      1118
           5       0.85      0.69      0.76      3978

    accuracy                           0.62      5667
   macro avg       0.39      0.52      0.42      5667
weighted avg       0.70      0.62      0.65      5667
```

Conclusion

Possible next steps

- Get more data on 1-star and 2-star reviews
- Feature and parameter optimization

Final thoughts

- Final model has a testing set accuracy of 0.736
- Far better than chance (0.20)
- This is even better when focusing only on 5-star reviews (0.85)
- Other 4 classifications were above chance as well