Amazon Alexa Review Classification

Catherine Somers

The Problem

- Customers are going to usually find things they will like and not like about an Amazon Alexa product
- **Potential Issue:** A product that is missing additional features or creates ease of use problems
- **Aim for this project:**
- **Audience:** Product Management and Sales side
- How this can help
    - Improving future product versions
    - Positively drive product sales

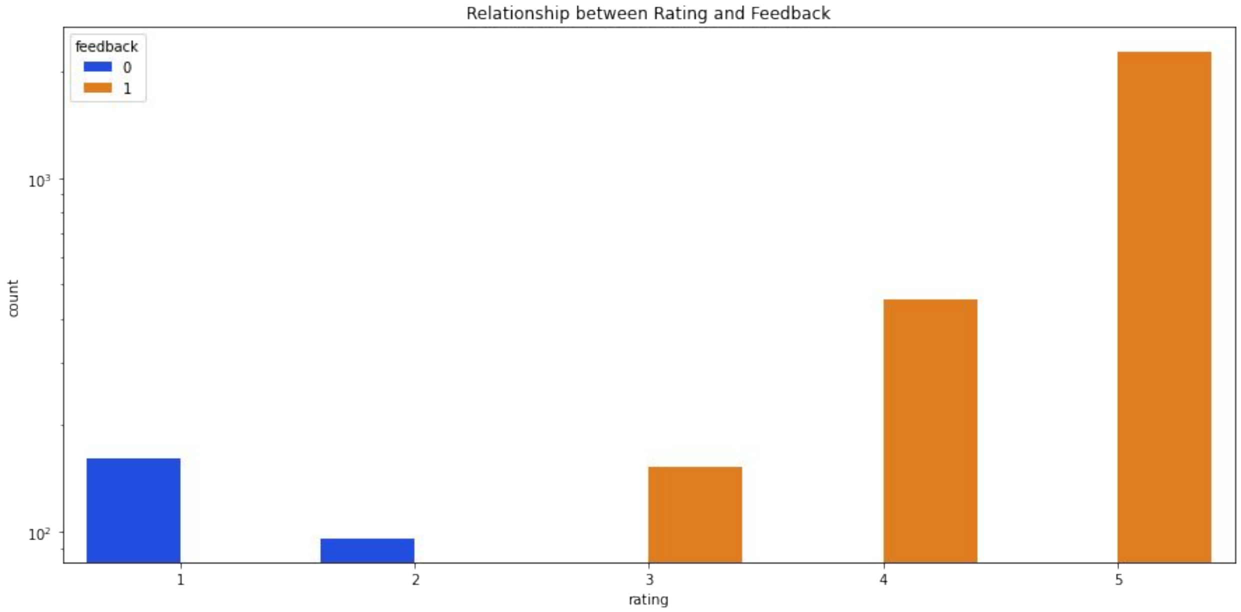
Data and Method Used

- Obtain Amazon Alexa Review dataset from Kaggle
- Took [Amazon Alexa Review data] for analysis
- Raw data: Total of 3151 values for feedback
- For each review there are 5 columns: rating, date, variation, verified_reviews, and feedback
- Rating, verified_reviews, and feedback relevant for analysis
- Rating has product rating on scale of 1-5, verified_reviews is composed of review text, and feedback denotes whether feedback is positive or negative

Data Wrangling

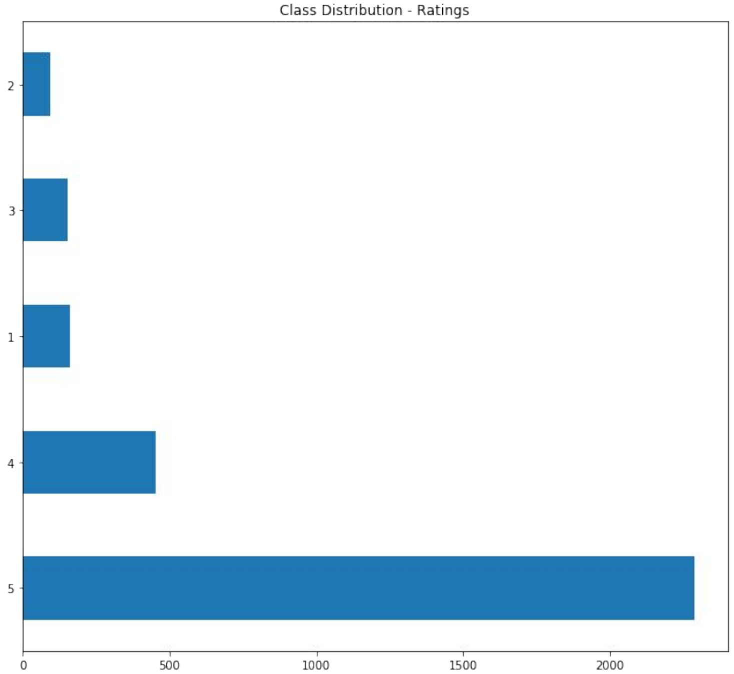
- Imported the .tsv file and formatted it into a Pandas dataframe.
- **Cleaning**:
    - Retain relevant columns and rows
    - Extract year, month, day of the week, and review length into separate columns
        - Estimating review length is an important feature for text classification in Natural Language Processing (NLP)

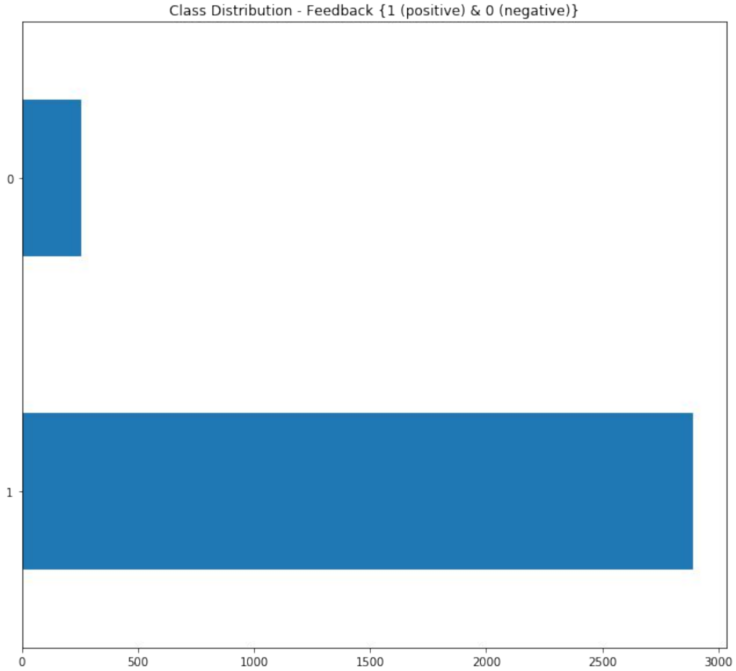
Exploratory Data Analysis (EDA)



- Plotting Rating and Feedback on a bar graph shows that in order for feedback to be considered positive, its rating has to be 3 or greater

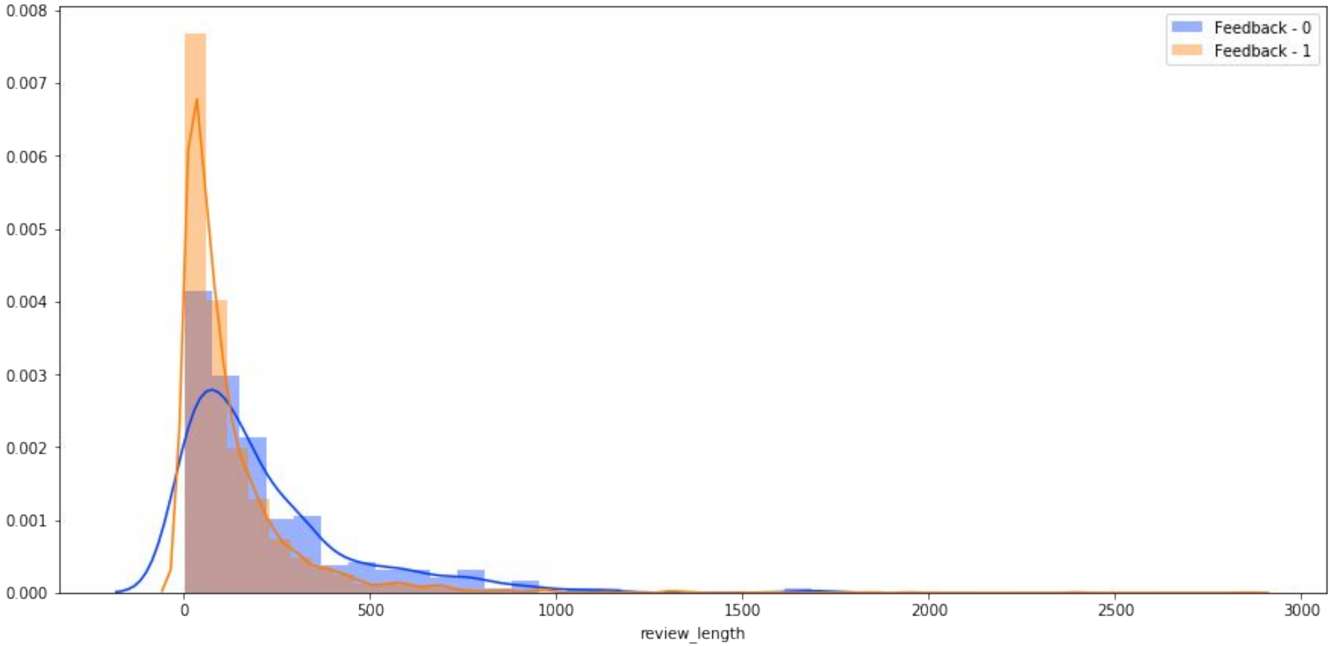
EDA





- In both cases, we see distribution among feedback and ratings is highly skewed on the positive side. For the most part, products have been well received by the customers.

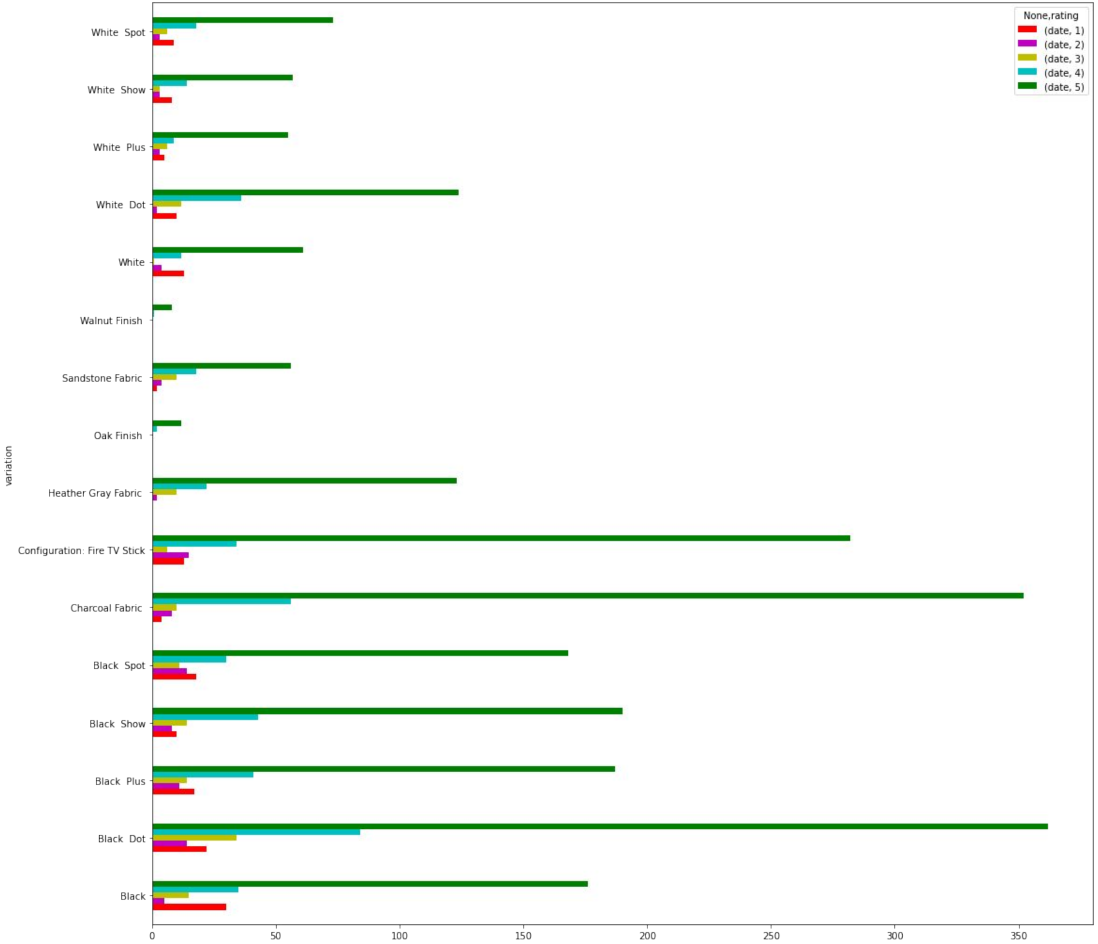
Length of Reviews and Feedback Type



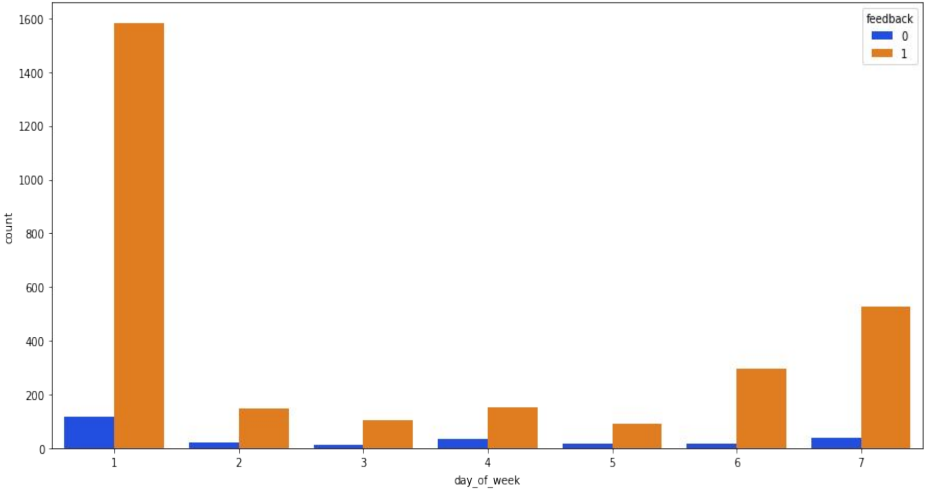
- Customers with negative reviews have a tendency to write a longer review

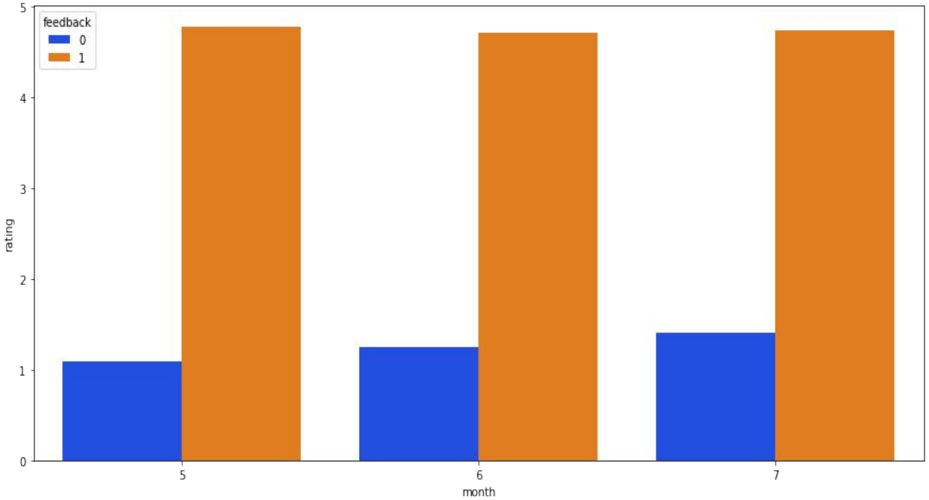
Product Variation and Rating

- The Black Dot variant of Amazon Alexa products has the most ratings of 5 among the Amazon Alexa products in the data



Day of Week, Month and Feedback Type



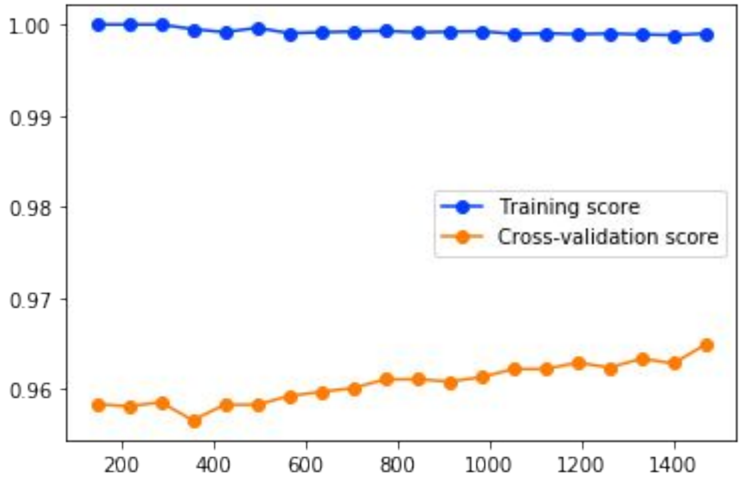


- Customers most likely to write their feedback on Monday and during the month of July

EDA: Key Findings

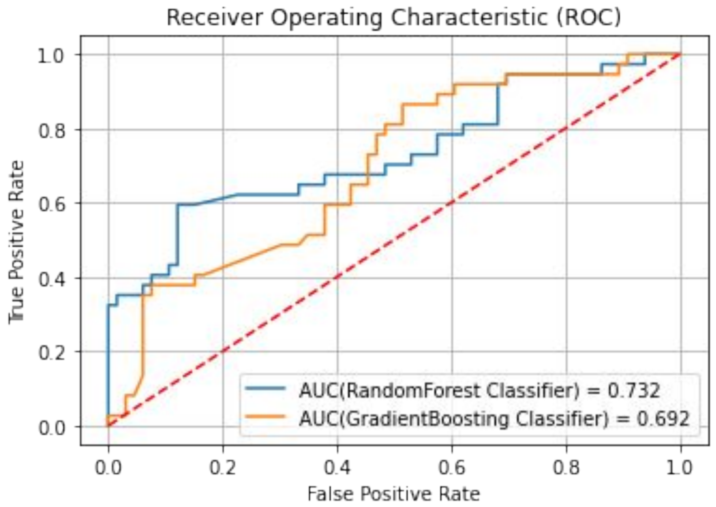
- Ratings need to be greater than 3 to be considered positive
- Distribution among feedback and ratings is highly skewed on the positive side
- What may help in the future is to stratify the data to avoid a class imbalance
- Customers were mostly likely to write reviews on Monday and gave feedback mostly in July

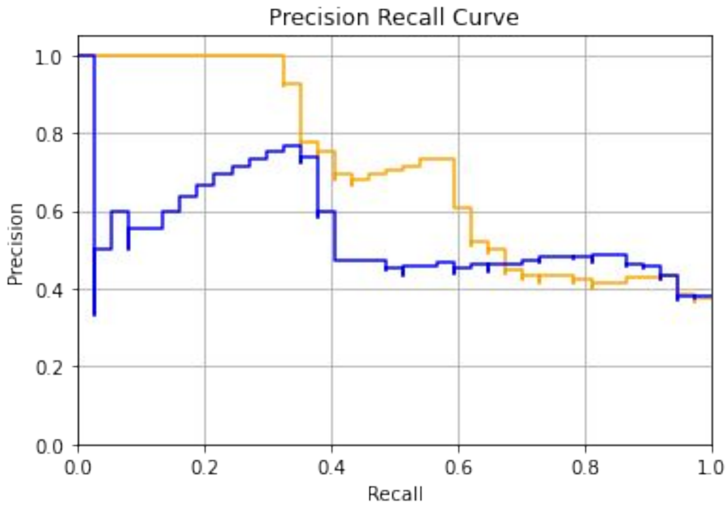
Predictive Modeling



- Trained and tested using Random Forest and Gradient Boosting
- In each scoring criteria, each model performed well and scored high
- Metrics used: Precision, Recall, F1 Score, and Confusion Matrix

Model Performance: Visualizations





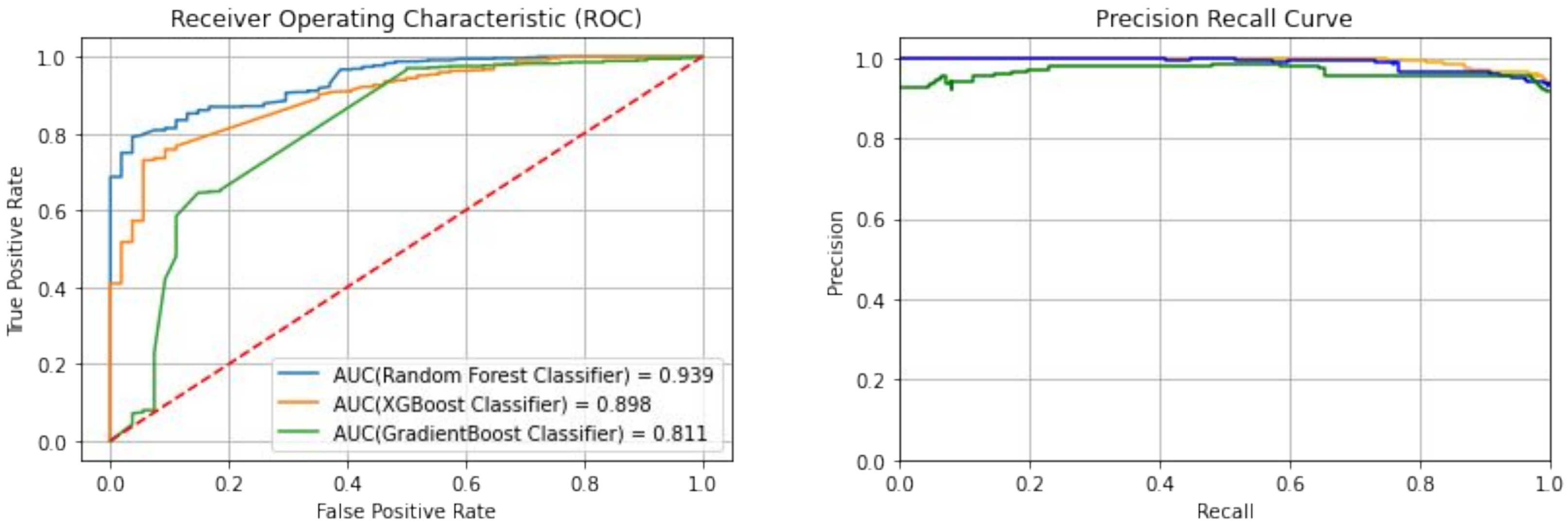
- Looks at least a 0.6 threshold for both ROC Curve and Precision Recall Curve looking at bad reviews

Evaluation of Model

- The targeted variable was feedback
- Table below summarizes performance on negative reviews

MODEL	ACCURACY	PRECISION	RECALL	F1 SCORE
Random Forest	73%	71%	41%	51%
Gradient Boosting	69%	61%	38%	47%

Sentiment Analysis with BERT



- Collectively, the three models had high performance as seen in the graphs above.

Sentiment Analysis with BERT Performance

- For each rating class:

RATING	ACCURACY
5	96.2%
4	24%
3	0%
2	0%
1	79.1%

Model Performance

MODEL	ACCURACY	PRECISION	RECALL
Random Forest	94%	95%	99.4%
XGBoost	94%	93.2%	100%
Gradient Boosting	93%	95.2%	97%

Conclusion and Future Work

- **Takeaway:** Predicting what determines negative feedback is more than just looking at the rating.
- If we look at the text a customer used in their written review, it helps gives us more insight why they chose certain rating
- Based on performance, the model I would pick would be XGBoost since it had highest rate of prediction
- Use a bigger dataset when doing multi-class classification