Amazon Review Sentiment Analysis

Marissa, Cory, Mike, & Jordan

Agenda

- Problem & Data
- Exploratory Data Analysis
- Data Preprocessing
- Models & Results
- Conclusions & Future Research

Executive Summary

Research Question:

Can we predict whether Amazon reviews are positive or negative based on the review text?

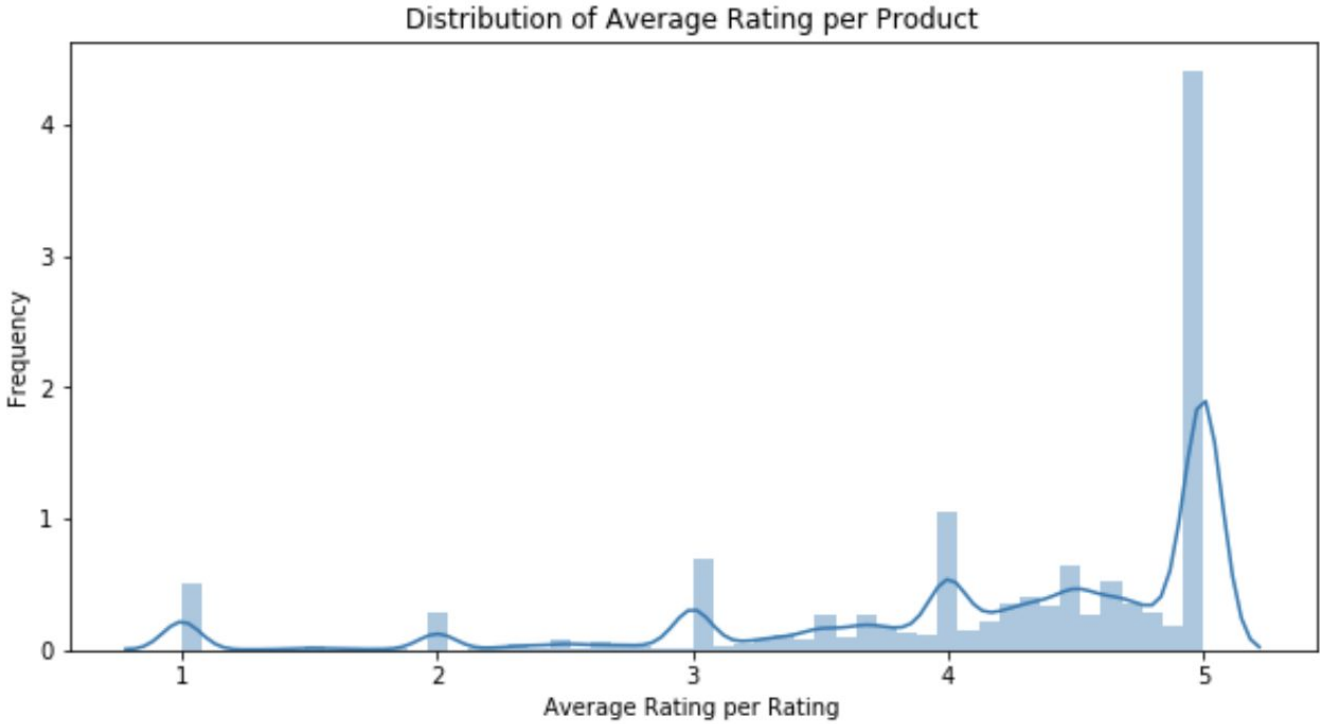
Approach:

- Convert rating (number of stars 1-5) to binary positive or negative
  - 3 stars and below are negative
  - Over 3 stars are positive
- Use sentiment analysis on review text with binary label

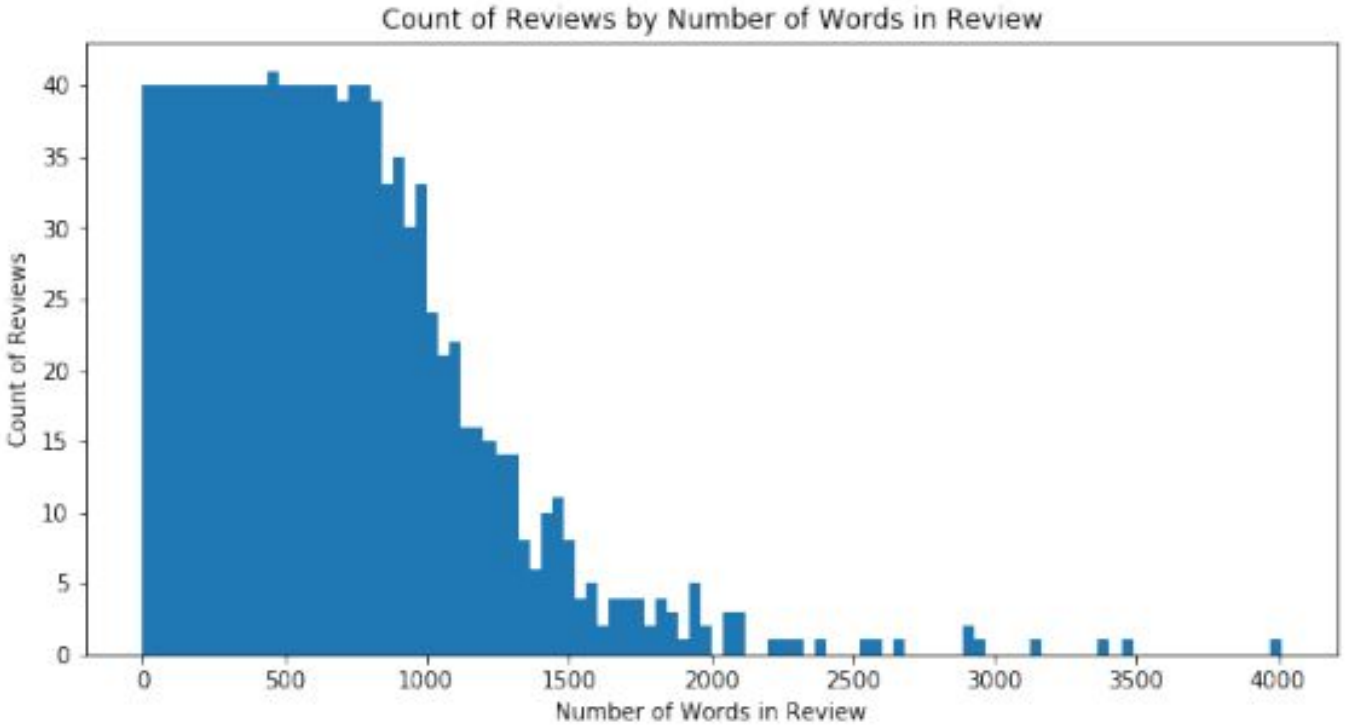
Data

- Sourced by Julian McAuley, UCSD
- Full Data set 142.8 million reviews spanning May 1996 - July 2014
- Project subset Grocery and Gourmet Food 5,074,160 reviews
- Data 2.05 GB
- Data Link

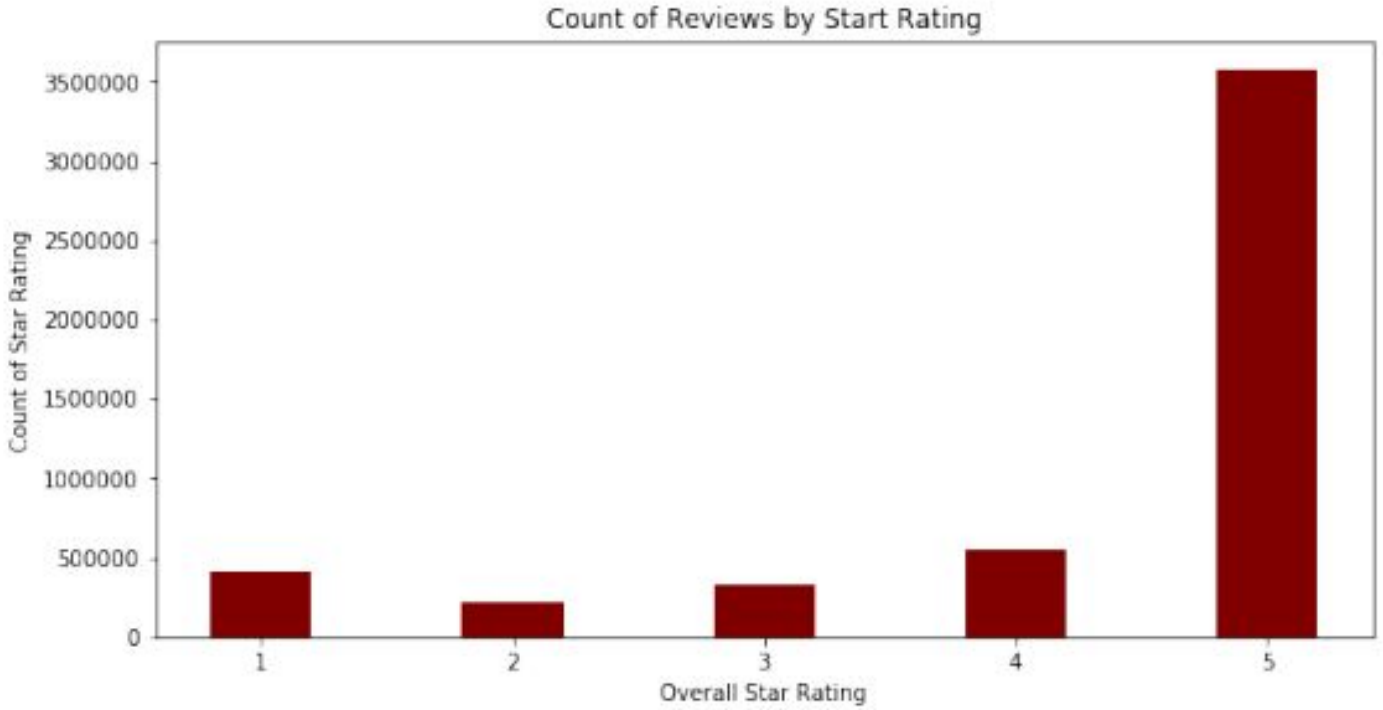
Exploratory Data Analysis



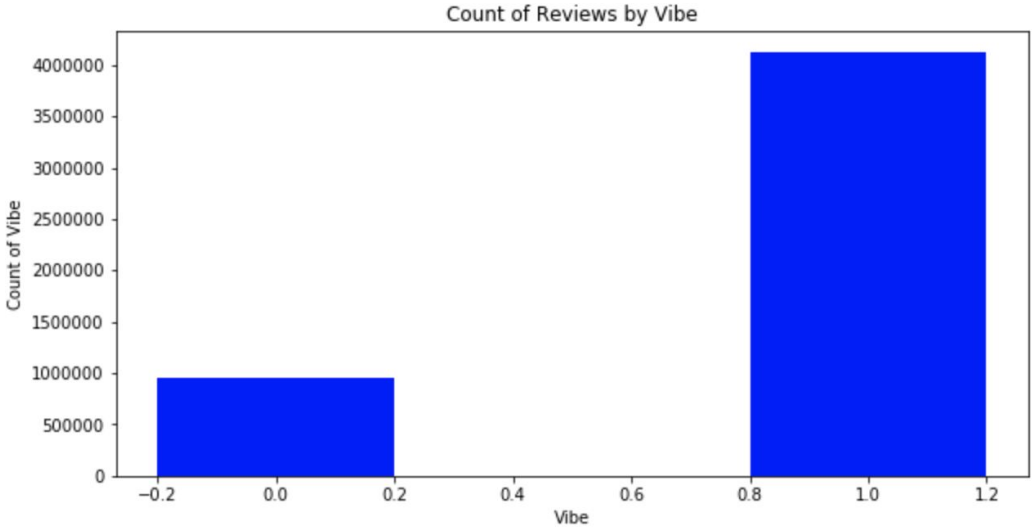
Exploratory Data Analysis

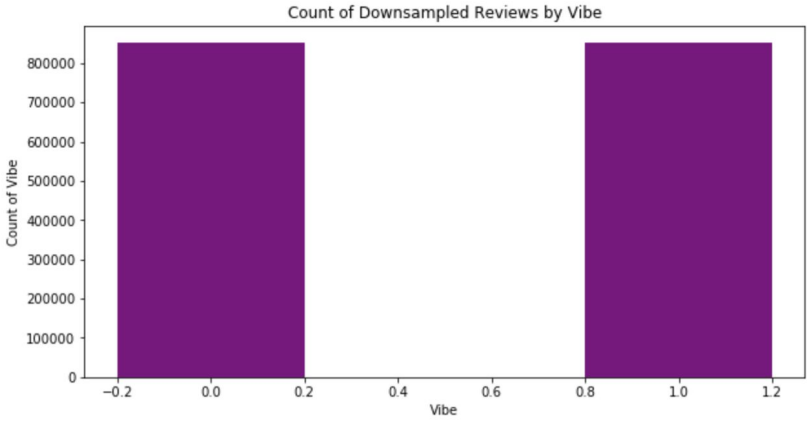


Exploratory Data Analysis



Exploratory Data Analysis





Data Processing

- The review text as input
- Positive or negative review label
  - Positive defined as 4 or 5 stars
- Down Sampled to balance the data
- Cleaned the review text
  - Set to lowercase, removed punctuation and stop words
- Hashing Term Frequency
- Inverse Document Frequency
  - Used as the model features

Models

1. Logistic Regression
    - Max Iterations = 100
    - Regularization Parameter = 0.1
    - Threshold = 0.4
2. Random Forest
    - Number of Trees = 30
    - Max Depth = 10
    - Max Bins = 32
3. Gradient Boosted Random Forest
    - Max Iterations = 20
    - Max Depth = 5
    - Max Bins = 32

Model Results - Metrics

	Accuracy	Recall	Precision	F1
Logistic Regression	0.899	0.943	0.933	0.938
Random Forest	0.800	0.841	0.906	0.872
Gradient Boosted Trees	0.619	0.561	0.949	0.705

Model Results - Confusion Matrices

Logistic Regression

67,156	27,843
23,309	388,258

Random Forest

52,251	35,748
65,401	346,166

Gradient Boosted Trees

82,694	12,305
180,484	231,083

TN	FN
FP	TP

Conclusions and Future Research

Conclusions

- Our model utilizing logistic regression is able to accurately predict review sentiment
- Logistic regression outperformed tree based models for this task

Future Research

- Threshold can be changed to flag very negative reviews
- Identify positive and negative aspects of the reviews
  - To help customers understand pros and cons of the product
  - To help sellers understand where they need to improve

Questions