

# Amazon Product Review Project

Capstone II Project

Michael Trent

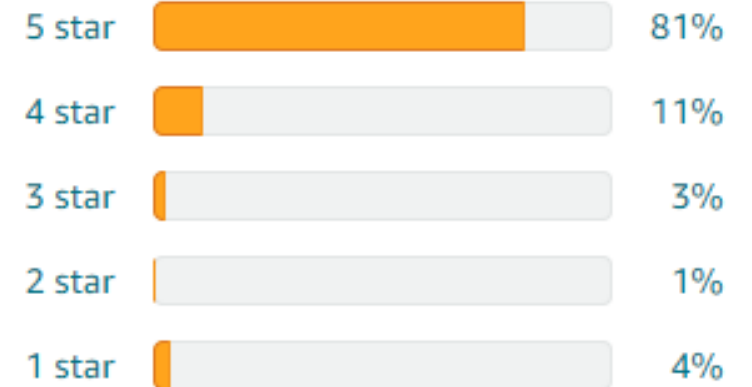
# Problem Statement

# Problem Statement

## Customer reviews


★★★★☆ 4.6 out of 5

8,820 global ratings



▼ How are ratings calculated?

# Problem Statement

 Gregg Mar

★★★★★

Great Product

Reviewed in the United States on February 27, 2018

Verified Purchase

I purchased this for my son-in-law because his car would occasionally not start. With 2 young kids, I felt that it was important that you would not get stranded somewhere. I received the unit and was surprised how well built it was. The case has a rubberized material instead of just a hard plastic. Everything was in a nice small zip up case which would easily fit in the spare tire area. The unit was almost fully charged when it arrived. My son-in-law went on vacation and I had his car in my garage while they were gone. I eventually tried to start it and the battery was drained. Although I had a battery charger/jumper on hand, it was a good time to test this unit. I hooked it up and it started right away. Pretty cool!

This is a lot better than carrying jumper cables around which my son-in-law did. Don't have to find someone to jump off of. Personally I am reluctant to offer a jump to anyone. There are a lot of electronics on a cars these days and there is a potential to damage them when jumping another car. I am thinking about purchasing one to carry around in my car.

283 people found this helpful

Helpful


|

▼ 9 comments

|

Report abuse

# Problem Statement

 Gregg Mar

★★★★★ **Great Product**

Reviewed in the United States on February 27, 2018

**Verified Purchase**

I purchased this for my son-in-law because his car would occasionally not start. With 2 young kids, I felt that it was important that you would not get stranded somewhere. I received the unit and was surprised how well built it was. The case has a rubberized material instead of just a hard plastic. Everything was in a nice small zip up case which would easily fit in the spare tire area. The unit was almost fully charged when it arrived. My son-in-law went on vacation and I had his car in my garage while they were gone. I eventually tried to start it and the battery was drained. Although I had a battery charger/jumper on hand, it was a good time to test this unit. I hooked it up and it started right away. Pretty cool!

This is a lot better than carrying jumper cables around which my son-in-law did. Don't have to find someone to jump off of. Personally I am reluctant to offer a jump to anyone. There are a lot of electronics on a cars these days and there is a potential to damage them when jumping another car. I am thinking about purchasing one to carry around in my car.

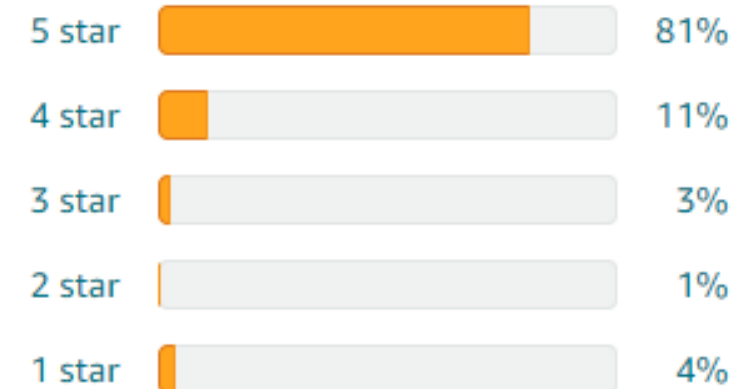
283 people found this helpful

Helpful | ▼ 9 comments | Report abuse

## Customer reviews


★★★★★ 4.6 out of 5

8,820 global ratings



▼ How are ratings calculated?

# Problem Statement

 Gregg Mar

★★★★★ **Great Product**

Reviewed in the United States on February 27, 2018

**Verified Purchase**

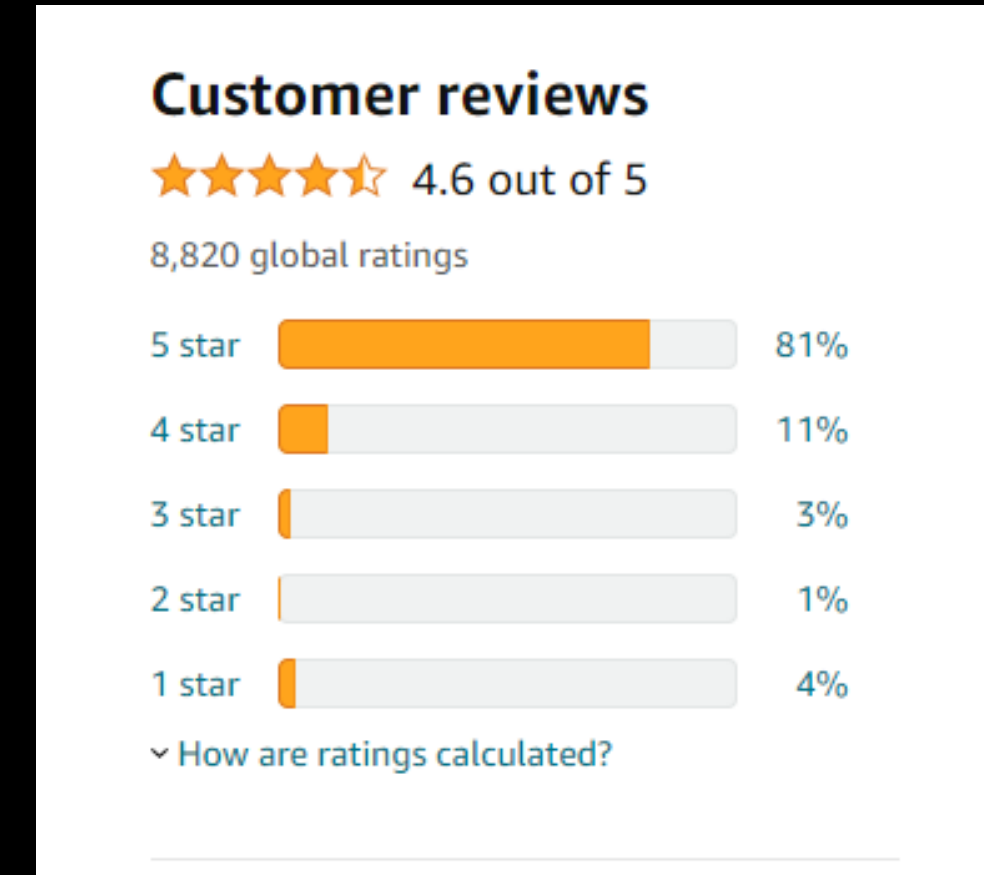
I purchased this for my son-in-law because his car would occasionally not start. With 2 young kids, I felt that it was important that you would not get stranded somewhere. I received the unit and was surprised how well built it was. The case has a rubberized material instead of just a hard plastic. Everything was in a nice small zip up case which would easily fit in the spare tire area. The unit was almost fully charged when it arrived. My son-in-law went on vacation and I had his car in my garage while they were gone. I eventually tried to start it and the battery was drained. Although I had a battery charger/jumper on hand, it was a good time to test this unit. I hooked it up and it started right away. Pretty cool!

This is a lot better than carrying jumper cables around which my son-in-law did. Don't have to find someone to jump off of. Personally I am reluctant to offer a jump to anyone. There are a lot of electronics on a cars these days and there is a potential to damage them when jumping another car. I am thinking about purchasing one to carry around in my car.

283 people found this helpful

[Helpful](#) | [▼ 9 comments](#) | [Report abuse](#)

?



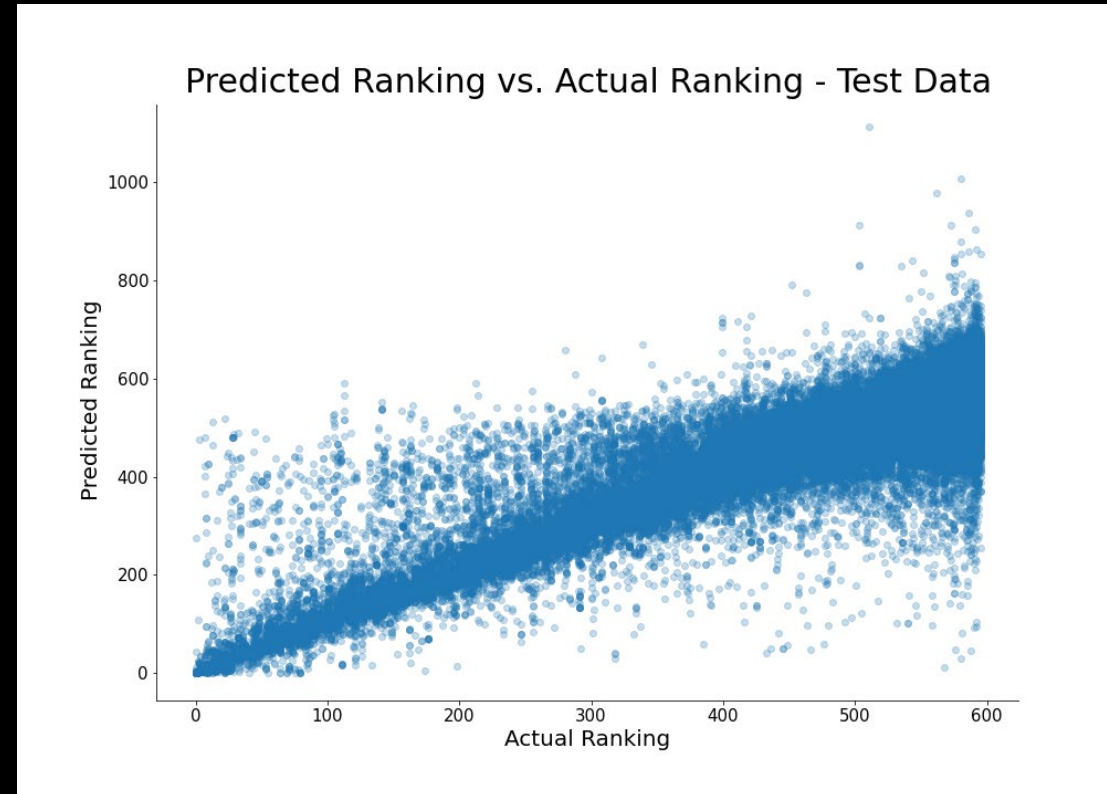


How well will your new product sell based on reviews and ratings?

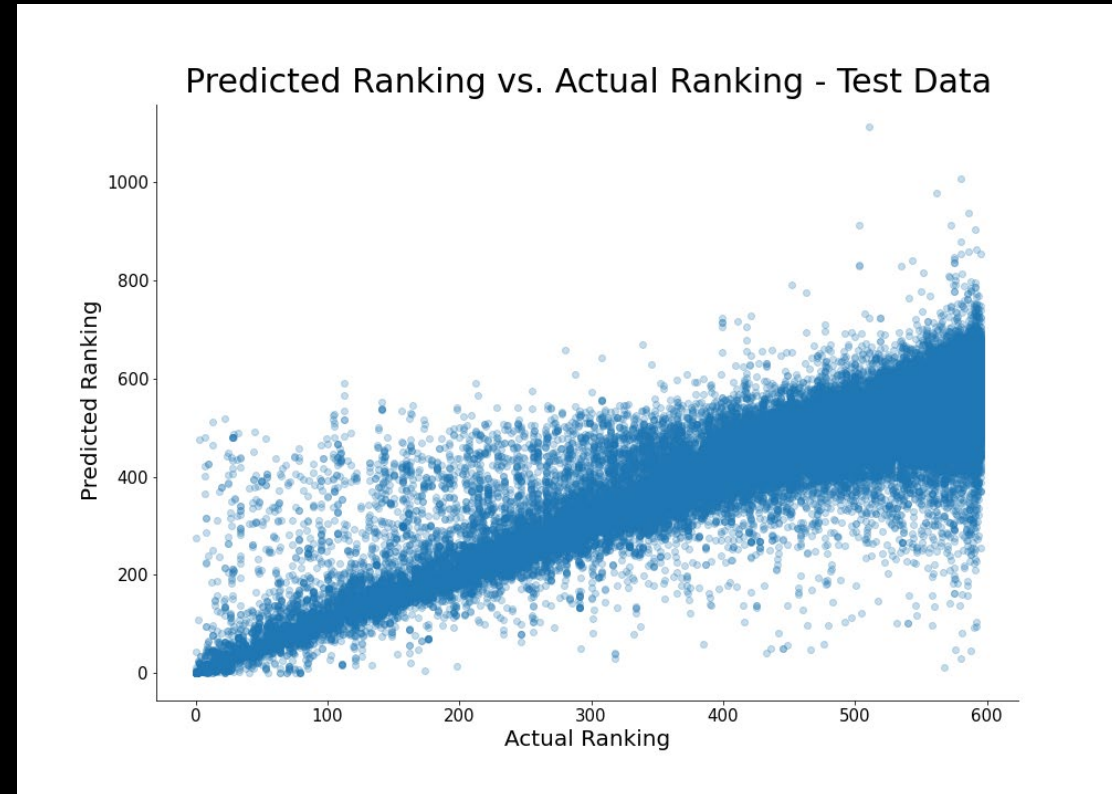




# How well will your new product sell based on reviews and ratings?







# How well will your new product sell based on reviews and ratings?



Our model predicts ~93% of the variability in the data

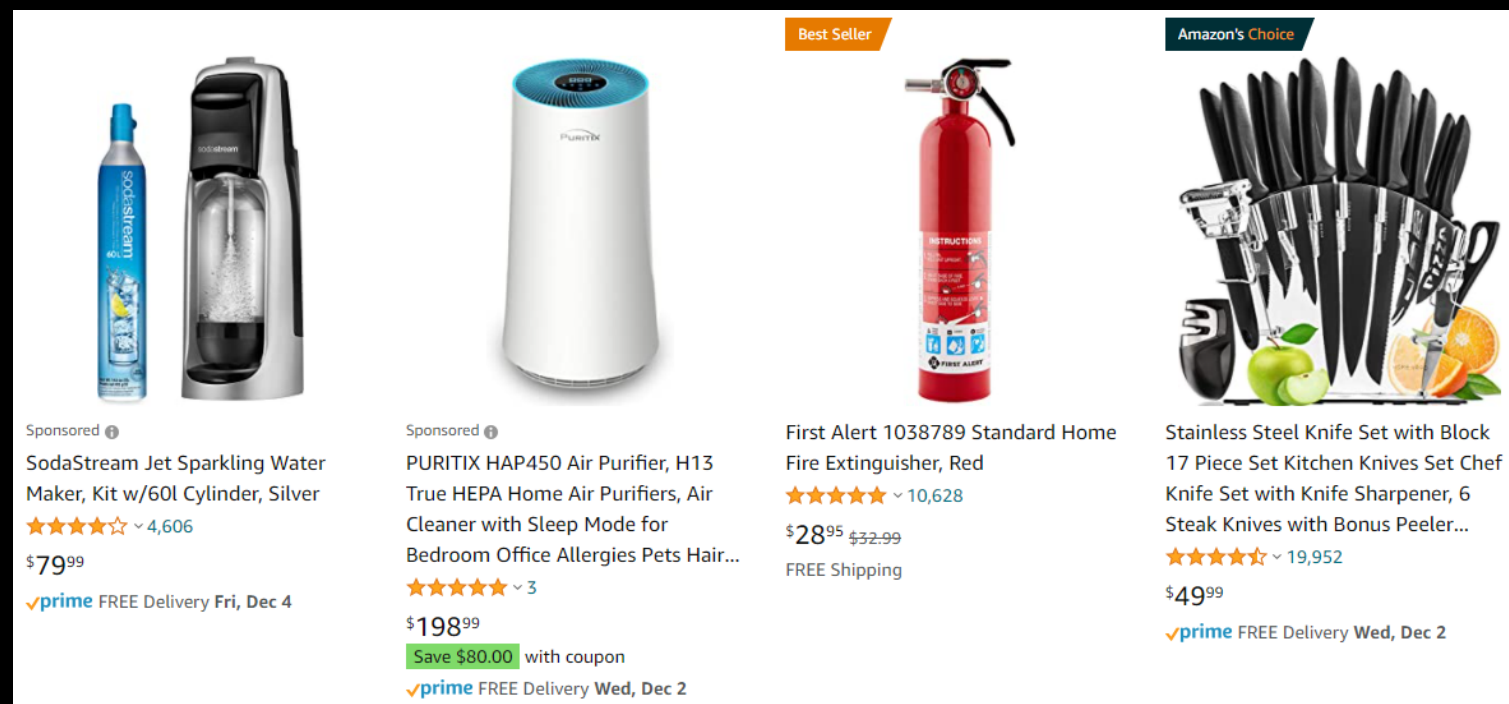
# Data

- ~ 7 million reviews
- ~ 1.5 million products

			
<p>Sponsored ⓘ</p> <p>SodaStream Jet Sparkling Water Maker, Kit w/60l Cylinder, Silver</p> <p>★★★★☆ ~ 4,606</p> <p>\$79<sup>99</sup></p> <p>✓prime FREE Delivery Fri, Dec 4</p>	<p>Sponsored ⓘ</p> <p>PURITIX HAP450 Air Purifier, H13 True HEPA Home Air Purifiers, Air Cleaner with Sleep Mode for Bedroom Office Allergies Pets Hair...</p> <p>★★★★★ ~ 3</p> <p>\$198<sup>99</sup></p> <p>Save \$80.00 with coupon</p> <p>✓prime FREE Delivery Wed, Dec 2</p>	<p>Best Seller</p> <p>First Alert 1038789 Standard Home Fire Extinguisher, Red</p> <p>★★★★★ ~ 10,628</p> <p>\$28<sup>95</sup> \$32.99</p> <p>FREE Shipping</p>	<p>Amazon's Choice</p> <p>Stainless Steel Knife Set with Block 17 Piece Set Kitchen Knives Set Chef Knife Set with Knife Sharpener, 6 Steak Knives with Bonus Peeler...</p> <p>★★★★☆ ~ 19,952</p> <p>\$49<sup>99</sup></p> <p>✓prime FREE Delivery Wed, Dec 2</p>

# Data

- ~ 7 million reviews
- ~ 1.2 million products



The screenshot displays four Amazon product listings side-by-side. Each listing includes a product image, a title, a star rating with the number of reviews, a price, and shipping information. The first listing is for a SodaStream Jet Sparkling Water Maker, which is sponsored and has a 4.5-star rating from 4,606 reviews, priced at \$79.99 with free delivery on Friday, December 4. The second listing is for a Puritix HAP450 Air Purifier, also sponsored, with a 4.5-star rating from 3 reviews, priced at \$198.99 (reduced from \$322.99 with a \$80.00 coupon), and free delivery on Wednesday, December 2. The third listing is for a First Alert 1038789 Standard Home Fire Extinguisher, marked as a 'Best Seller' with a 4.5-star rating from 10,628 reviews, priced at \$28.95 (reduced from \$32.99), and free shipping. The fourth listing is for a Stainless Steel Knife Set with a Block, marked as 'Amazon's Choice' with a 4.5-star rating from 19,952 reviews, priced at \$49.99, and free delivery on Wednesday, December 2.

Product	Rating	Reviews	Price	Shipping	Delivery
SodaStream Jet Sparkling Water Maker, Kit w/60l Cylinder, Silver	4.5 stars	~ 4,606	\$79.99	FREE	Fri, Dec 4
PURITIX HAP450 Air Purifier, H13 True HEPA Home Air Purifiers, Air Cleaner with Sleep Mode for Bedroom Office Allergies Pets Hair...	4.5 stars	~ 3	\$198.99 (Save \$80.00 with coupon)	FREE	Wed, Dec 2
First Alert 1038789 Standard Home Fire Extinguisher, Red	4.5 stars	~ 10,628	\$28.95 (was \$32.99)	FREE	Shipping
Stainless Steel Knife Set with Block 17 Piece Set Kitchen Knives Set Chef Knife Set with Knife Sharpener, 6 Steak Knives with Bonus Peeler...	4.5 stars	~ 19,952	\$49.99	FREE	Wed, Dec 2

## Citation:

Justifying recommendations using distantly-labeled reviews and fined-grained aspects. Jianmo Ni, Jiacheng Li, Julian McAuley. Empirical Methods in Natural Language Processing (EMNLP), 2019.

# Data Wrangling

- Determining Sales Ranking

# Data Wrangling

- Determining Sales Ranking

- ['>#9,714 in Kitchen & Dining (See Top 100 in Kitchen & Dining)', '>#29 in Kitchen & Dining > Bakeware > Baking Tools & Accessories > Baking Cups', '>#3,224 in Kitchen & Dining > Kitchen Utensils & Gadgets']

# Data Wrangling

- Determining Sales Ranking

- The overall ranking within Kitchen & Dining is meaningless.

- ['>#9,714 in Kitchen & Dining (See Top 100 in Kitchen & Dining)', '>#29 in Kitchen & Dining > Bakeware > Baking Tools & Accessories > Baking Cups', '>#3,224 in Kitchen & Dining > Kitchen Utensils & Gadgets']

# Data Wrangling

- Determining Sales Ranking

- The overall ranking within Kitchen & Dining is meaningless.
- We want the ranking within the primary (main) category.

- ['>#9,714 in Kitchen & Dining (See Top 100 in Kitchen & Dining)', '>#29 in Kitchen & Dining > Bakeware > Baking Tools & Accessories > Baking Cups', '>#3,224 in Kitchen & Dining > Kitchen Utensils & Gadgets']



# Data Wrangling

- Determining Sales Ranking

- The overall ranking within Kitchen & Dining is meaningless.
- We want the ranking within the primary (main) category.

- ['>#9,714 in Kitchen & Dining (See Top 100 in Kitchen & Dining)', '>#29 in Kitchen & Dining > Bakeware > Baking Tools & Accessories > Baking Cups', '>#3,224 in Kitchen & Dining > Kitchen Utensils & Gadgets']

# Data Wrangling

- Determining Sales Ranking
- Identify similar products

# Data Wrangling

- Determining Sales Ranking
- Identify similar products
- Identical products may have different names and product IDs

# Data Wrangling

- Determining Sales Ranking
- Identify similar products
  - Use fuzzy logic
- Identical products may have different names and product IDs

# Data Wrangling

- Determining Sales Ranking
- Identify similar products
  - Use fuzzy logic
- Identical products may have different names and product IDs
- Fuzzy Wuzzy



# Data Wrangling

- Determining Sales Ranking
- Identify similar products
  - Use fuzzy logic
  - $> 85\%$  set ratio &  $> 85\%$  sort ratio  
→ Same product
- Identical products may have different names and product IDs
- Fuzzy Wuzzy



# Data Wrangling

- Determining Sales Ranking
- Identify similar products
  - Use fuzzy logic
  - > 85% set ratio & > 85% sort ratio
    - Same product
- Identical products may have different names and product IDs
- Fuzzy Wuzzy
  - ['Hamilton Beach 31103A Countertop Oven with Convection and Rotisserie',
  - 'Hamilton Beach Countertop Oven with Convection and Rotisserie', 100, 95]

# Data Wrangling

- Determining Sales Ranking
- Identify similar products
  - Use fuzzy logic
  - > 85% set ratio & > 85% sort ratio
    - Same product
- Identical products may have different names and product IDs
- Fuzzy Wuzzy
  - ['Hamilton Beach 31103A Countertop Oven with Convection and Rotisserie',
  - 'Hamilton Beach Countertop Oven with Convection and Rotisserie', 100, 95]

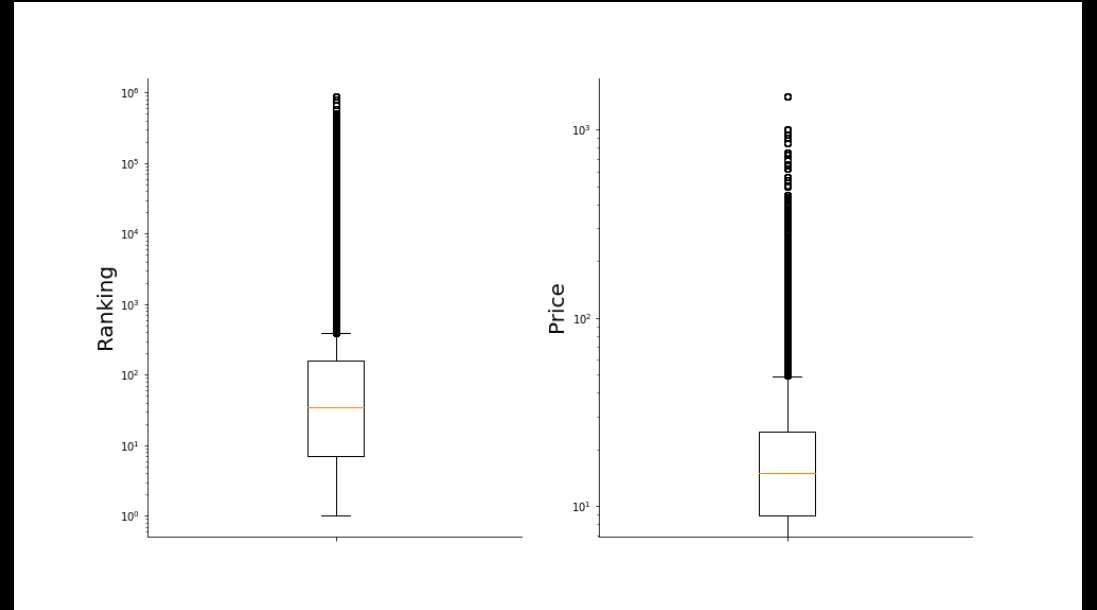


# Data Wrangling

- Determining Sales Ranking
- Identify similar products
- Drop unknowable missing values

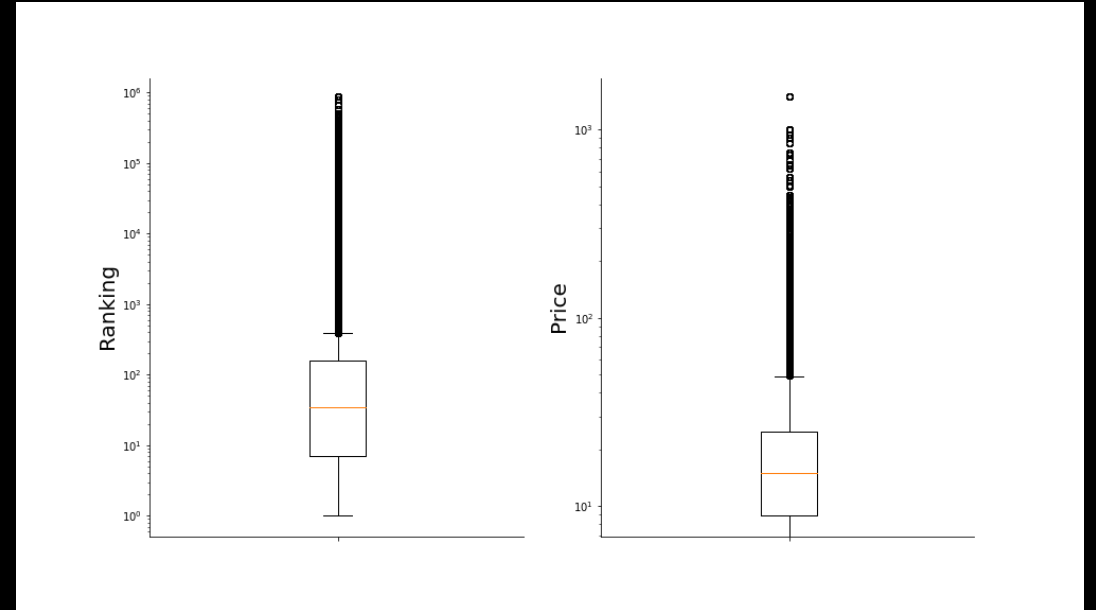
# EDA

- Determining Sales Ranking
- Identify similar products
- Drop unknowable missing values
- Drop outliers



# EDA

- Determining Sales Ranking
- Identify similar products
- Drop unknowable missing values
- Drop outliers
  - $> 4 \times \text{IQR}$  or  $< 4 \times \text{IQR}$



# Feature Engineering

- Determining Sales Ranking
  - Identify similar products
  - Drop unknowable missing values
  - Drop outliers
  - Binning categories and brands
- Infrequent categories (< 500 reviews)
  - Infrequent brands (< 300 reviews)  
→ Miscellaneous

# Feature Engineering

- Determining Sales Ranking
- Identify similar products
- Drop unknowable missing values
- Drop outliers
- Binning categories and brands
- Sentiment Analysis

- Nltk VADER



# Feature Engineering

- Determining Sales Ranking
- Identify similar products
- Drop unknowable missing values
- Drop outliers
- Binning categories and brands
- Sentiment Analysis

- Nltk VADER
  - Compound polarity score
    - $[-1, 1]$



# Feature Engineering

- Determining Sales Ranking
- Identify similar products
- Drop unknowable missing values
- Drop outliers
- Binning categories and brands
- Sentiment Analysis

- Nltk VADER
  - Compound polarity score
    - $[-1, 1]$
  - Calculate score for each sentence



# Feature Engineering

- Determining Sales Ranking
- Identify similar products
- Drop unknowable missing values
- Drop outliers
- Binning categories and brands
- Sentiment Analysis

- Nltk VADER
  - Compound polarity score
    - $[-1, 1]$
  - Calculate score for each sentence
  - Average scores across review





# Feature Engineering

- Determining Sales Ranking
- Identify similar products
- Drop unknowable missing values
- Drop outliers
- Binning categories and brands
- Sentiment Analysis

- Nltk VADER
  - Compound polarity score
    - $[-1, 1]$
  - Calculate score for each sentence
  - Average scores across review
  - Blank reviews scored as 0



# Feature Engineering

- Determining Sales Ranking
- Identify similar products
- Drop unknowable missing values
- Drop outliers
- Binning categories and brands
- Sentiment Analysis

- Nltk VADER
  - Compound polarity score
    - $[-1, 1]$
  - Calculate score for each sentence
  - Average scores across review
  - Blank reviews scored as 0
  - Numerical score for modeling!



# Feature Engineering

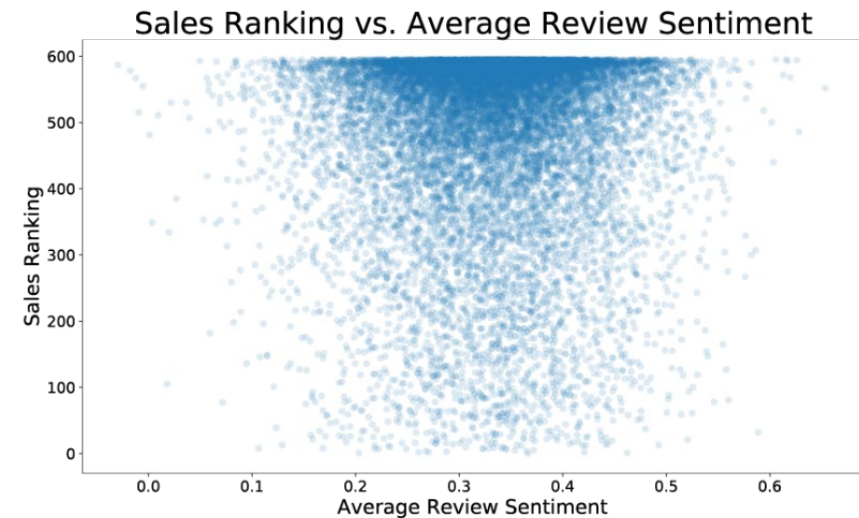
- Determining Sales Ranking
- Identify similar products
- Drop unknowable missing values
- Drop outliers
- Binning categories and brands
- Sentiment Analysis

- Nltk VADER
  - Compound polarity score
    - $[-1, 1]$
  - Calculate score for each sentence
  - Average scores across review
  - Blank reviews scored as 0
  - Numerical score for modeling!
  - Duplicate process for summaries



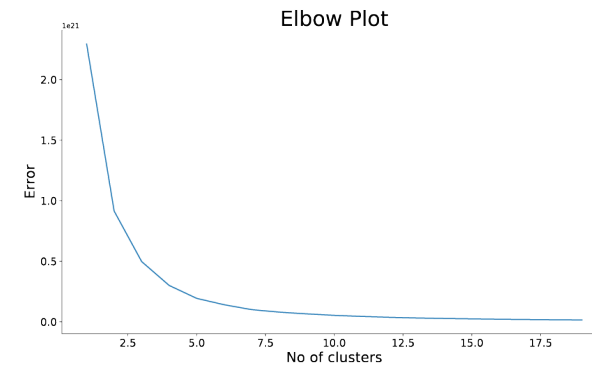
# Feature Engineering

- Determining Sales Ranking
- Identify similar products
- Drop unknowable missing values
- Drop outliers
- Binning categories and brands
- Sentiment Analysis
  - Non-linear relationship with ranking...



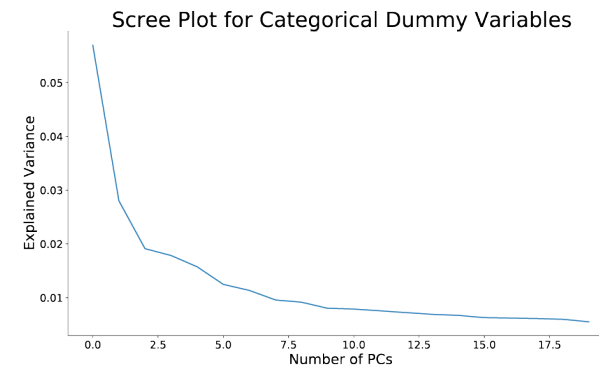
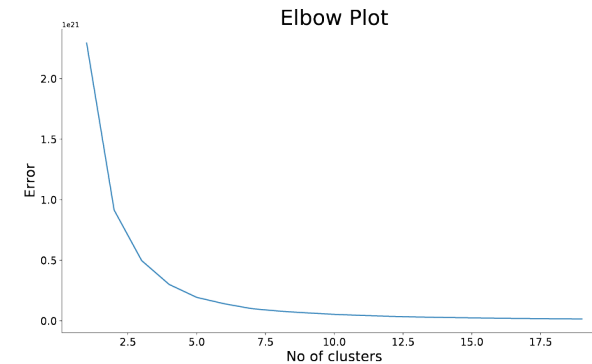
# Feature Engineering

- Determining Sales Ranking
- Identify similar products
- Drop unknowable missing values
- Drop outliers
- Binning categories and brands
- Sentiment Analysis
- Kmeans and PCA



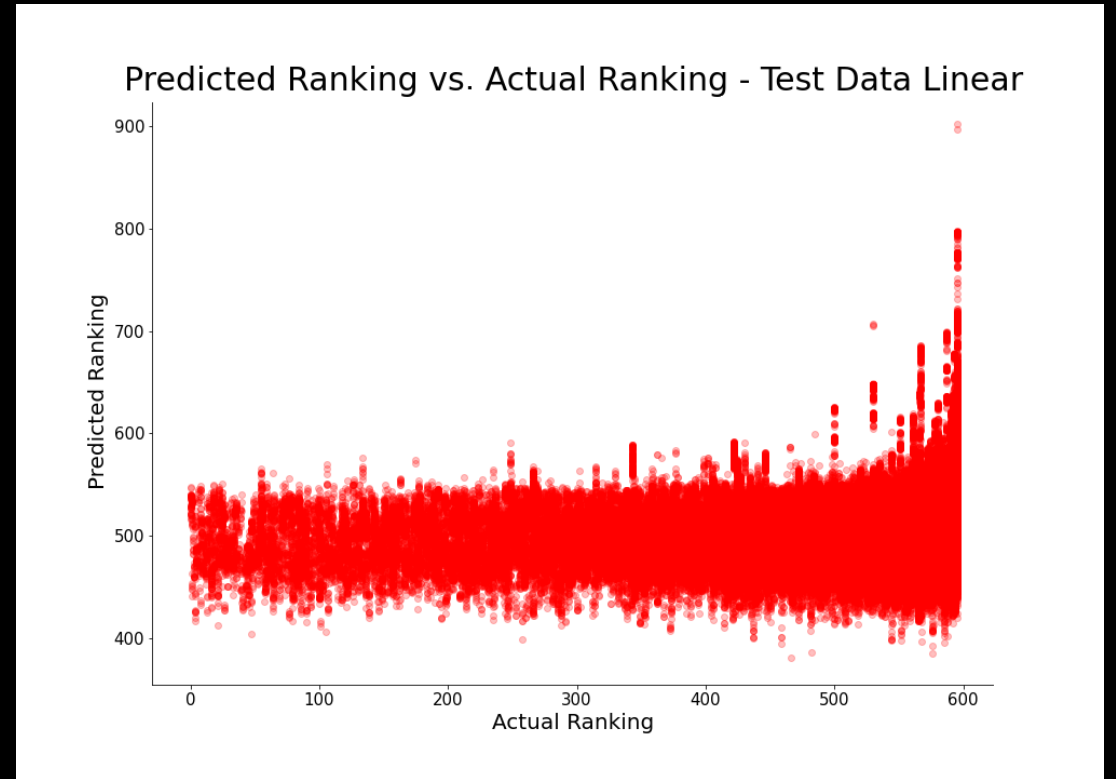
# Feature Engineering

- Determining Sales Ranking
- Identify similar products
- Drop unknowable missing values
- Drop outliers
- Binning categories and brands
- Sentiment Analysis
- Kmeans and PCA



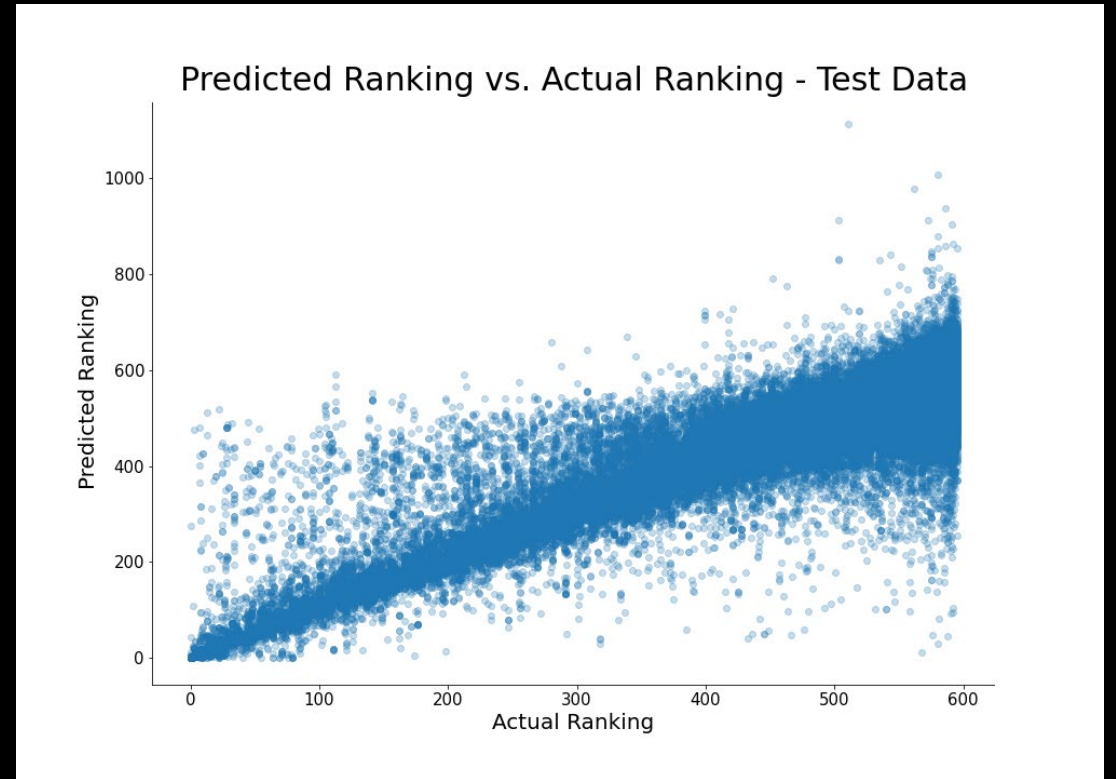
# Modeling

- Determining Sales Ranking
- Identify similar products
- Drop unknowable missing values
- Drop outliers
- Binning categories and brands
- Sentiment Analysis
- Kmeans and PCA
- Linear Modeling
  - $R^2 = 5.8\%$



# Modeling

- Determining Sales Ranking
- Identify similar products
- Drop unknowable missing values
- Drop outliers
- Binning categories and brands
- Sentiment Analysis
- Kmeans and PCA
- Linear Modeling
- XGBoost

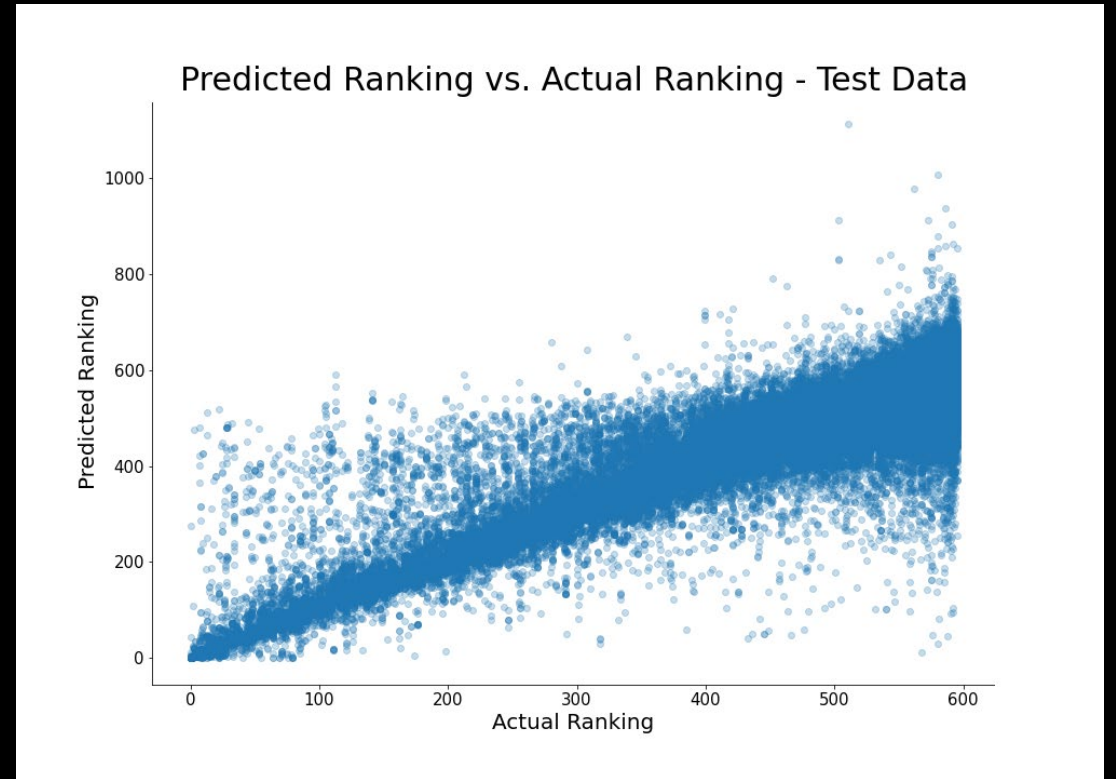


Bayesian Optimization for hyperparameter tuning



# Modeling

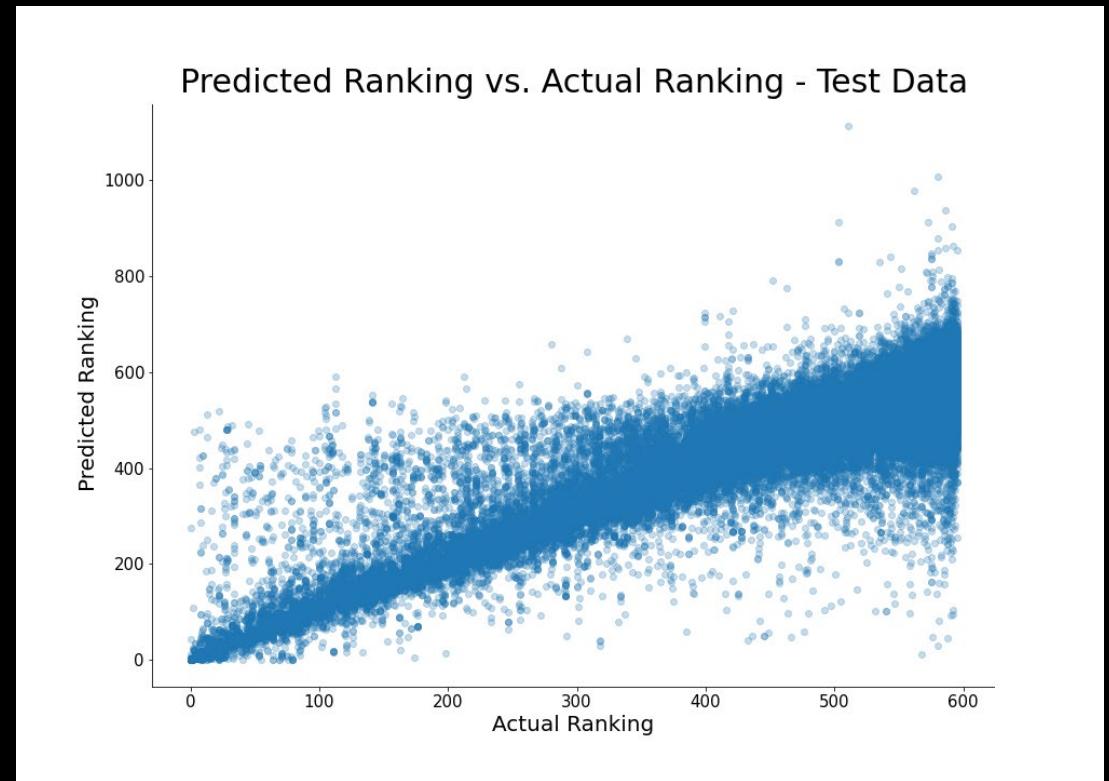
- Determining Sales Ranking
- Identify similar products
- Drop unknowable missing values
- Drop outliers
- Binning categories and brands
- Sentiment Analysis
- Kmeans and PCA
- Linear Modeling
- XGBoost
  - $R^2 = 92.745\%$



Bayesian Optimization for hyperparameter tuning

# Modeling

- Determining Sales Ranking
- Identify similar products
- Drop unknowable missing values
- Drop outliers
- Binning categories and brands
- Sentiment Analysis
- Kmeans and PCA
- Linear Modeling
- XGBoost
  - $R^2 = 92.745\%$  - Thanks optimizer!



Bayesian Optimization for hyperparameter tuning

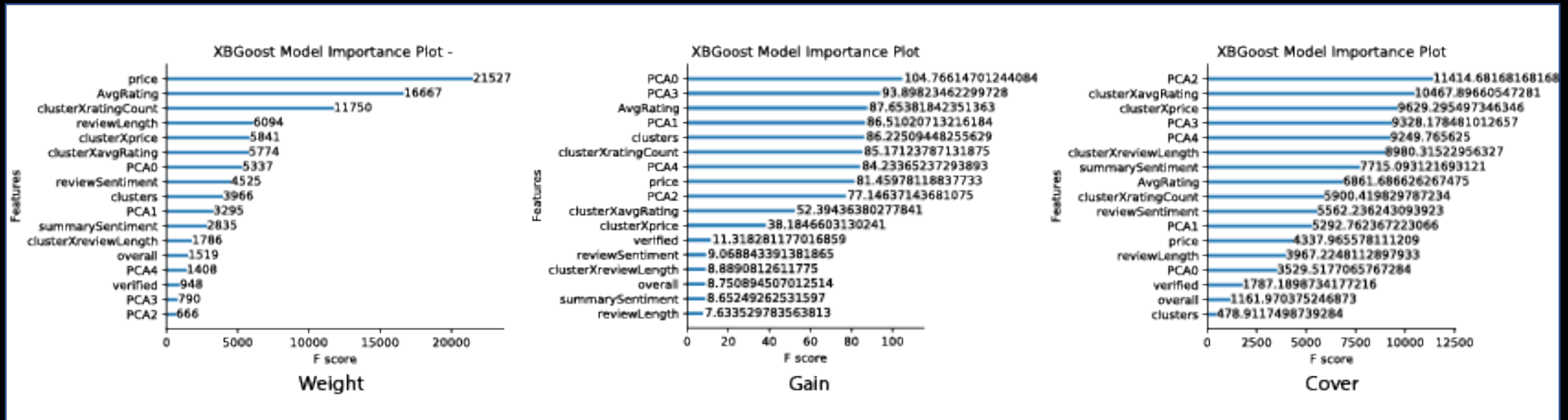
# Results

# Results

- Feature Importance

# Results

- Feature Importance - XGBoost



# Results

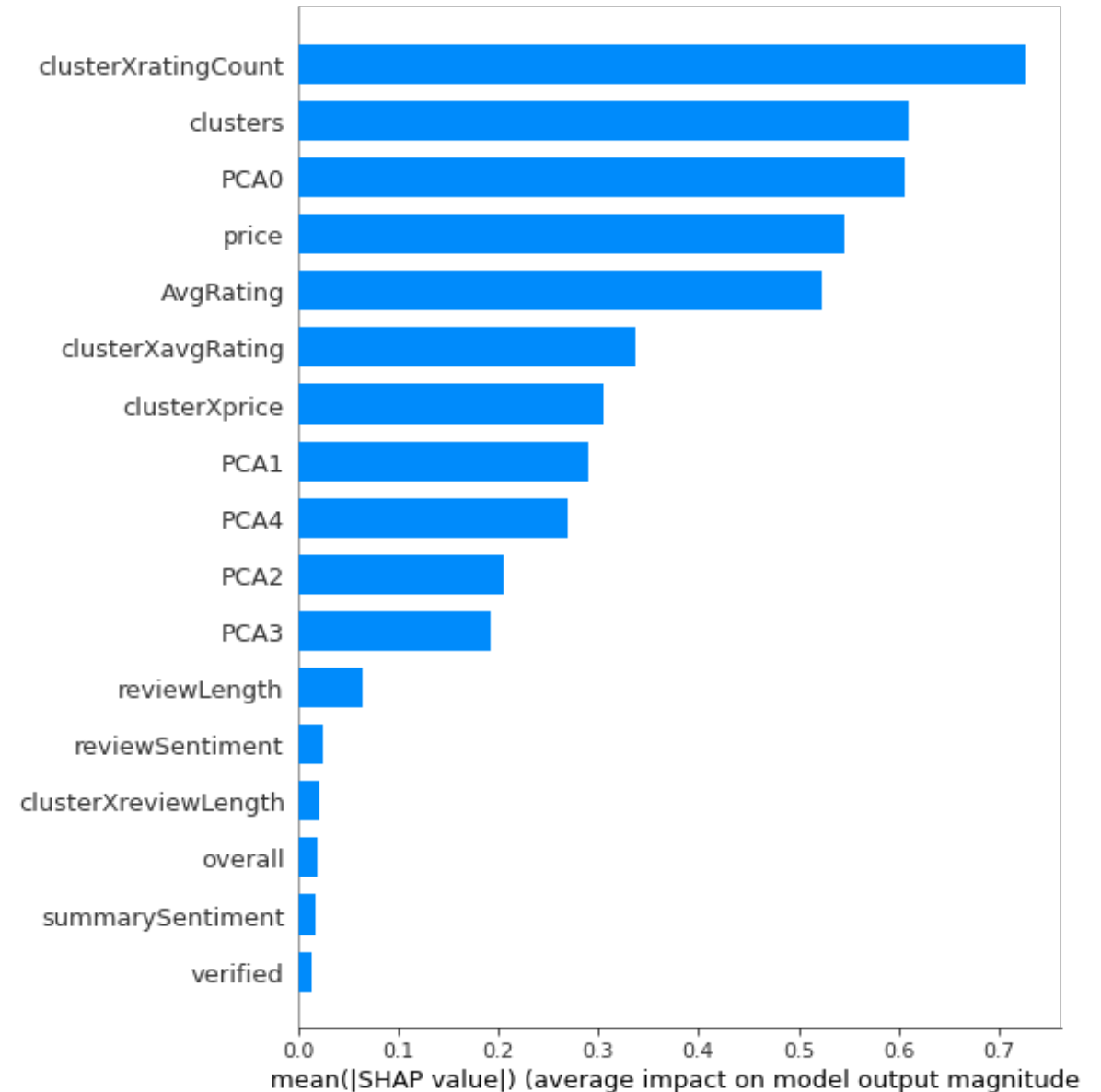
- Feature Importance - Shap

# Results

- Feature Importance – Shap
  - Uses the Shapley value to calculate each features contribution to the model

# Results

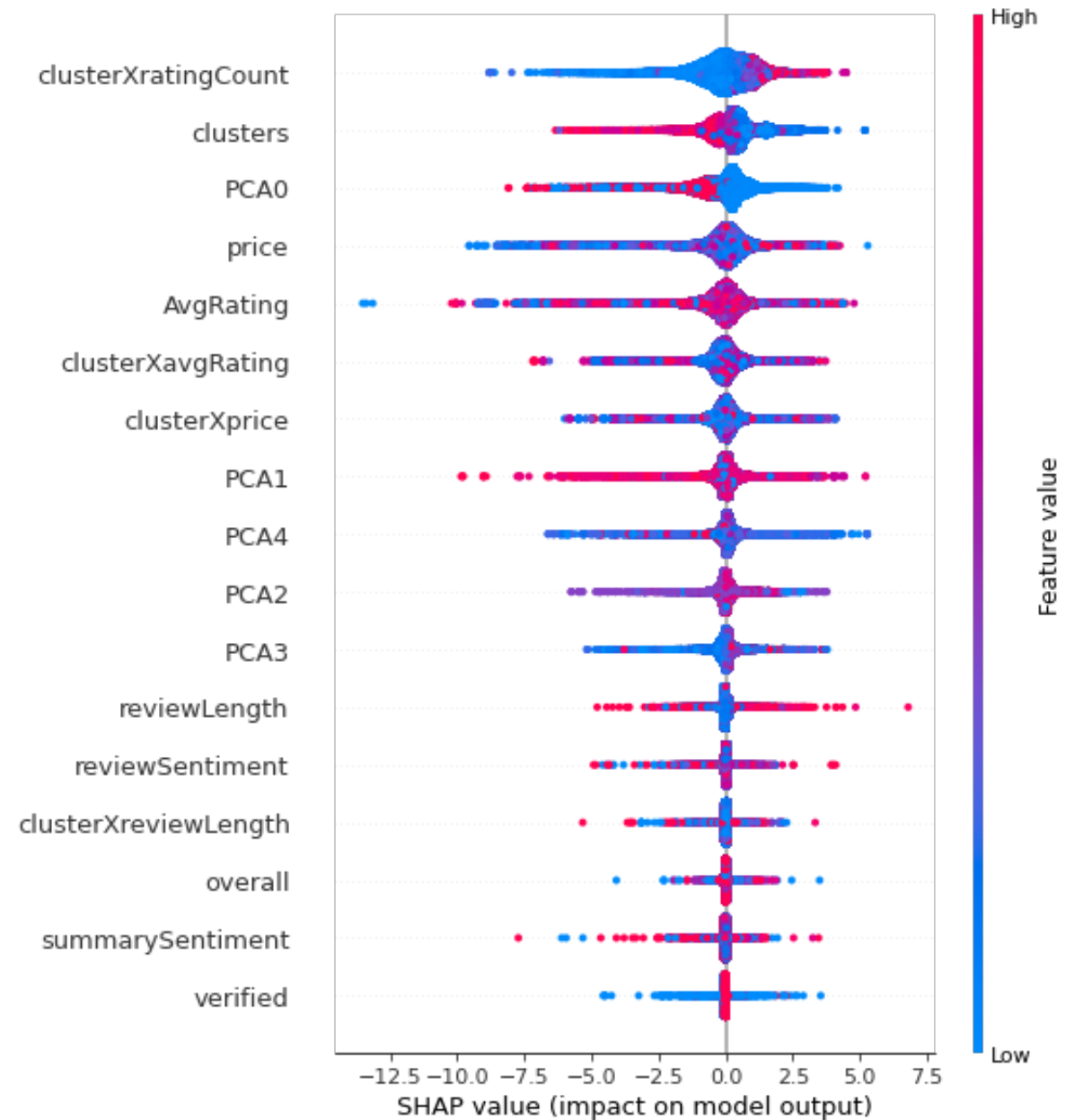
- Feature Importance – Shap
  - Price and average rating are important interpretable features
  - Review length and sentiment are not as important





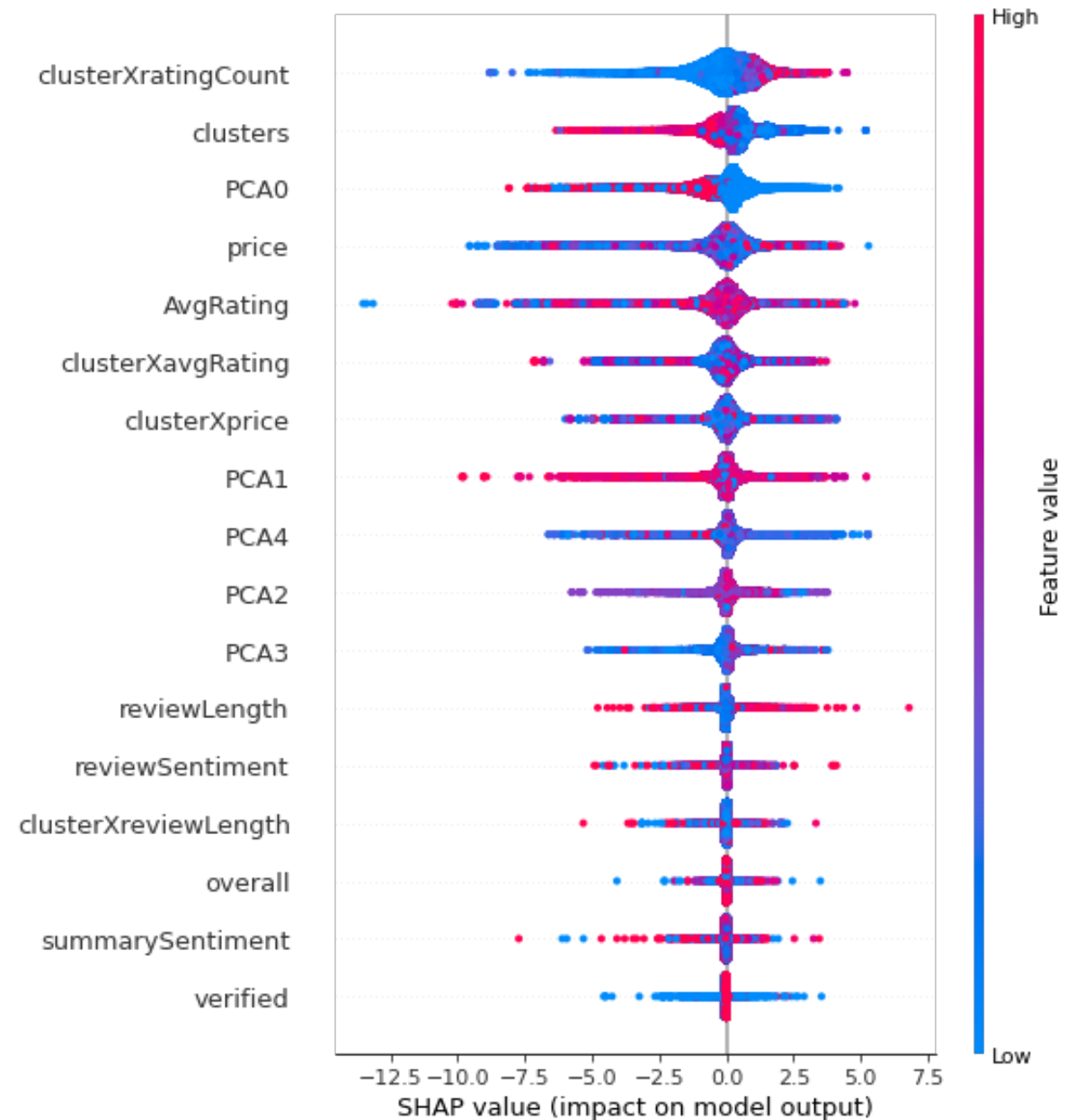
# Results

- Feature Importance – Shap



# Results

- Feature Importance – Shap
  - Possible correlation between price and Shap value
  - Possible correlation between average rating and Shap value



# Future directions

# Future directions

- Larger data set

# Future directions

- Larger data set
- More categories

# Future directions

- Larger data set
- More categories
- Mixed categories