INSTITUTE FOR APPLIED
COMPUTATIONAL SCIENCE
AT HARVARD UNIVERSITY

Product Review Analysis With pattern

Milestone 1

Team: Sehaj Chawla, Taro Spirit, Lotus Xia, Heather Liu

Agenda

1. Motivation and Problem Statement
2. Background Knowledge
3. EDA
4. Baseline Models

Motivation

pattern is an ecommerce accelerator. It helps businesses grow faster and sell globally on ecommerce marketplaces.

How? For example, use AI-supported insights and reporting to help businesses

1) predict sales volumes in the future,
2) adapt their marketing strategy or even their products.

That’s where we come in!

pattern mostly sells their customers’ products on Amazon and would like us to predict how well their products will do in the future on this platform and to understand why.

But, what is a **metric of success** on Amazon?

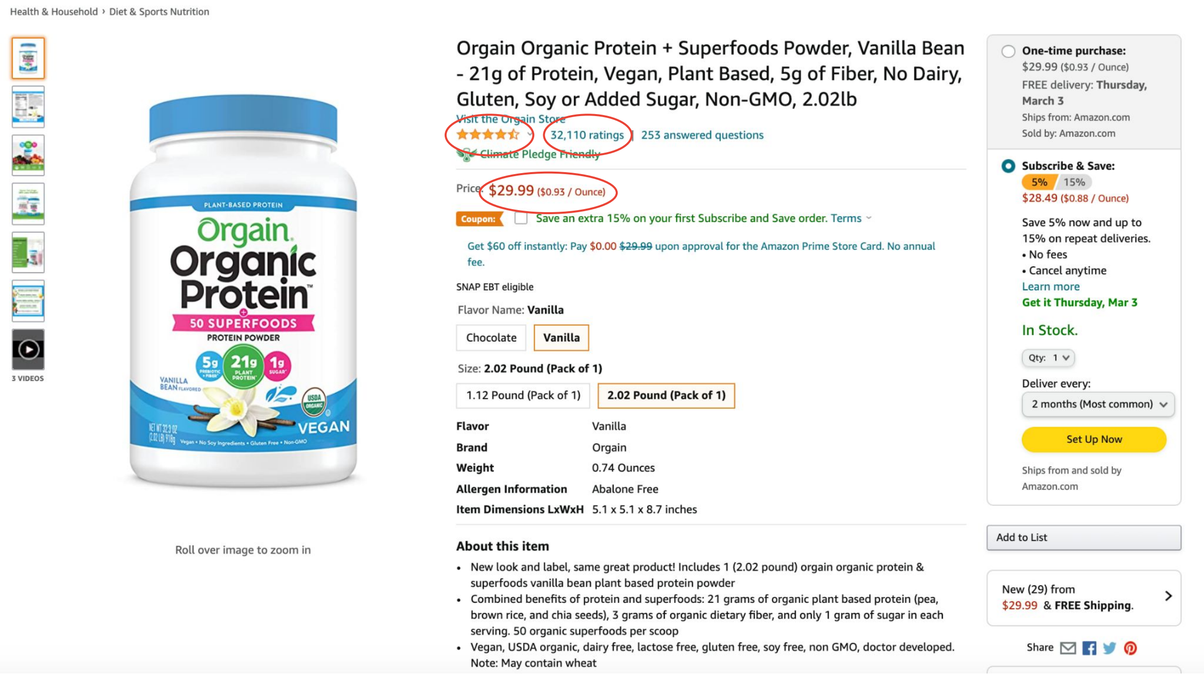
Use rank, strongly correlated with current sales volumes of a product. The smaller the rank the better a product is doing.

And **what kind of data** is predictive of rank?

Need all of the data that customers use to decide which product to buy.

Amazon

What are key informations that customers see on Amazon?



Health & Household › Diet & Sports Nutrition

Roll over image to zoom in

Orgain Organic Protein + Superfoods Powder, Vanilla Bean - 21g of Protein, Vegan, Plant Based, 5g of Fiber, No Dairy, Gluten, Soy or Added Sugar, Non-GMO, 2.02lb

Visit the Orgain Store

★★★★☆ 32,110 ratings | 253 answered questions

Climate Pledge Friendly

Price: $29.99 ($0.93 / Ounce)

Coupon: □ Save an extra 15% on your first Subscribe and Save order. Terms

Get $60 off instantly: Pay $0.00 ~~$29.99~~ upon approval for the Amazon Prime Store Card. No annual fee.

SNAP EBT eligible

Flavor Name: **Vanilla**

Chocolate | **Vanilla**

Size: **2.02 Pound (Pack of 1)**

1.12 Pound (Pack of 1) | **2.02 Pound (Pack of 1)**

Flavor	Vanilla
Brand	Orgain
Weight	0.74 Ounces
Allergen Information	Abalone Free
Item Dimensions LxWxH	5.1 x 5.1 x 8.7 inches

About this item

- New look and label, same great product! Includes 1 (2.02 pound) orgain organic protein & superfoods vanilla bean plant based protein powder
- Combined benefits of protein and superfoods: 21 grams of organic plant based protein (pea, brown rice, and chia seeds), 3 grams of organic dietary fiber, and only 1 gram of sugar in each serving. 50 organic superfoods per scoop
- Vegan, USDA organic, dairy free, lactose free, gluten free, soy free, non GMO, doctor developed. Note: May contain wheat

○ **One-time purchase:**
$29.99 ($0.93 / Ounce)
FREE delivery: **Thursday, March 3**
Ships from: Amazon.com
Sold by: Amazon.com

● **Subscribe & Save:**
5% 15%
$28.49 ($0.88 / Ounce)

Save 5% now and up to 15% on repeat deliveries.
- No fees
- Cancel anytime

Learn more

Get it Thursday, Mar 3

In Stock.

Qty: 1

Deliver every:
2 months (Most common)

Set Up Now

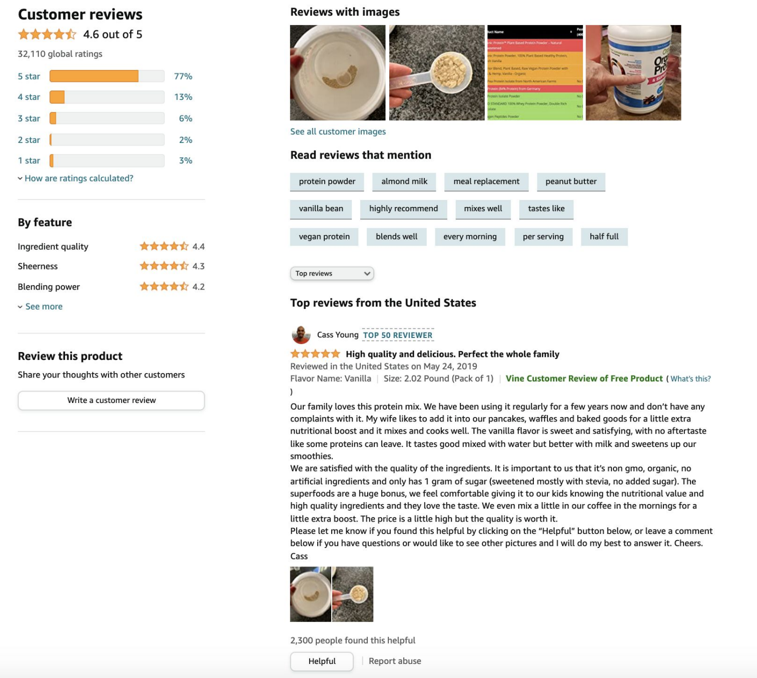
Ships from and sold by Amazon.com

Add to List

New (29) from $29.99 & **FREE Shipping**.

Share

Amazon



Customer reviews

4.6 out of 5

32,110 global ratings

5 star	77%
4 star	13%
3 star	6%
2 star	2%
1 star	3%

How are ratings calculated?

By feature

Ingredient quality	4.4
Sheerness	4.3
Blending power	4.2

See more

Review this product

Share your thoughts with other customers

Write a customer review

Reviews with images

See all customer images

Read reviews that mention

protein powder | almond milk | meal replacement | peanut butter | vanilla bean | highly recommend | mixes well | tastes like | vegan protein | blends well | every morning | per serving | half full

Top reviews

Top reviews from the United States

Cass Young TOP 50 REVIEWER

High quality and delicious. Perfect the whole family

Reviewed in the United States on May 24, 2019

Flavor Name: Vanilla | Size: 2.02 Pound (Pack of 1) | Vine Customer Review of Free Product (What's this?
)

Our family loves this protein mix. We have been using it regularly for a few years now and don't have any complaints with it. My wife likes to add it into our pancakes, waffles and baked goods for a little extra nutritional boost and it mixes and cooks well. The vanilla flavor is sweet and satisfying, with no aftertaste like some proteins can leave. It tastes good mixed with water but better with milk and sweetens up our smoothies.

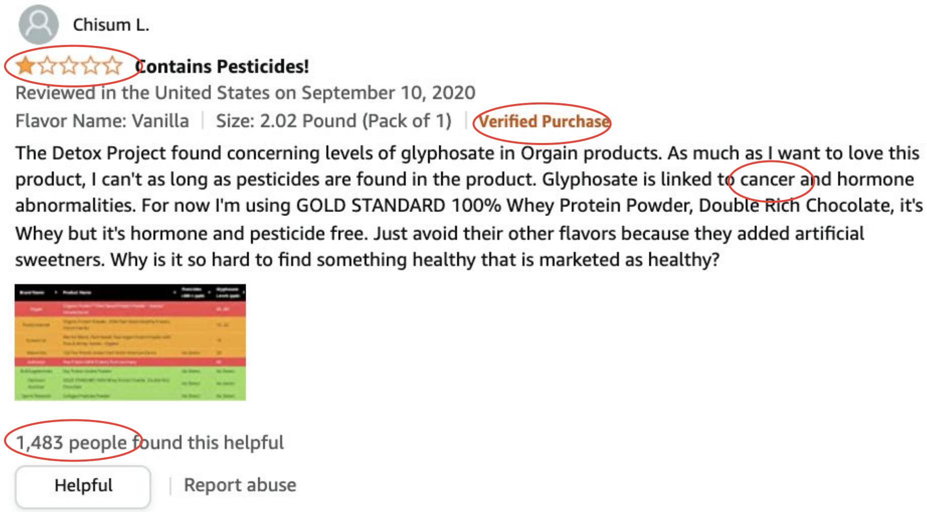
We are satisfied with the quality of the ingredients. It is important to us that it's non gmo, organic, no artificial ingredients and only has 1 gram of sugar (sweetened mostly with stevia, no added sugar). The superfoods are a huge bonus, we feel comfortable giving it to our kids knowing the nutritional value and high quality ingredients and they love the taste. We even mix a little in our coffee in the mornings for a little extra boost. The price is a little high but the quality is worth it.

Please let me know if you found this helpful by clicking on the "Helpful" button below, or leave a comment below if you have questions or would like to see other pictures and I will do my best to answer it. Cheers.

Cass

2,300 people found this helpful

Helpful | Report abuse



Chisum L.

Contains Pesticides!

Reviewed in the United States on September 10, 2020

Flavor Name: Vanilla | Size: 2.02 Pound (Pack of 1) | Verified Purchase

The Detox Project found concerning levels of glyphosate in Orgain products. As much as I want to love this product, I can't as long as pesticides are found in the product. Glyphosate is linked to cancer and hormone abnormalities. For now I'm using GOLD STANDARD 100% Whey Protein Powder, Double Rich Chocolate, it's Whey but it's hormone and pesticide free. Just avoid their other flavors because they added artificial sweetners. Why is it so hard to find something healthy that is marketed as healthy?

1,483 people found this helpful

Helpful | Report abuse

Problem Statement

★ Predict product rank using review data

★ Extract themes from review texts to gain insights on what keywords or topics are predictive of rank

Scope of Work

Project stages and focus:

- ★ Create several models to predict future rank including NLP and linear regression.
- ★ Focus on extraction of themes with the help of the prediction models.

Final Product:

- ★ Python notebooks which will be used by Pattern for future work/implementation

Team and Collaboration Infrastructure

Team collaboration:

- ★ Communication: slack channel
- ★ Working: google colab, github
- ★ Soon split our team into two subteams to work separately on the two different problems

Collaboration with Pattern:

- ★ Communication: email
- ★ Meeting: weekly

Learning Goals

Goal 1: Data handling

★ How to collect data

★ How to handle huge and noisy industry data

Goal 2: Modeling

★ How to adapt the models to real-world problems

★ How to select evaluation criteria

★ How to train models using clean data without losing the ability to generalize

Goal 3: Collaboration

★ How to clearly propose needs and delivery results to our partners

Literature Review

Literature on using NLP to analyze customers’ reviews

- Using NLP to extract quick and valuable insights from your customers’ reviews
- Sentiment Analysis of Movie Reviews with Google’s BERT
- Predicting Sales from the Language of Product Descriptions
- Amazon Product review Sentiment Analysis using BERT

Some popular models:

- BERT
- GPT-3
- XGBoost

Project Ideas

1. Analyze users’ sentiments towards a product
2. Predict product rankings/future sales using review data
3. Extract themes from review texts to gain insights on what keywords or topics are predictive of increasing/decreasing sales volumes

EDA: Data Description

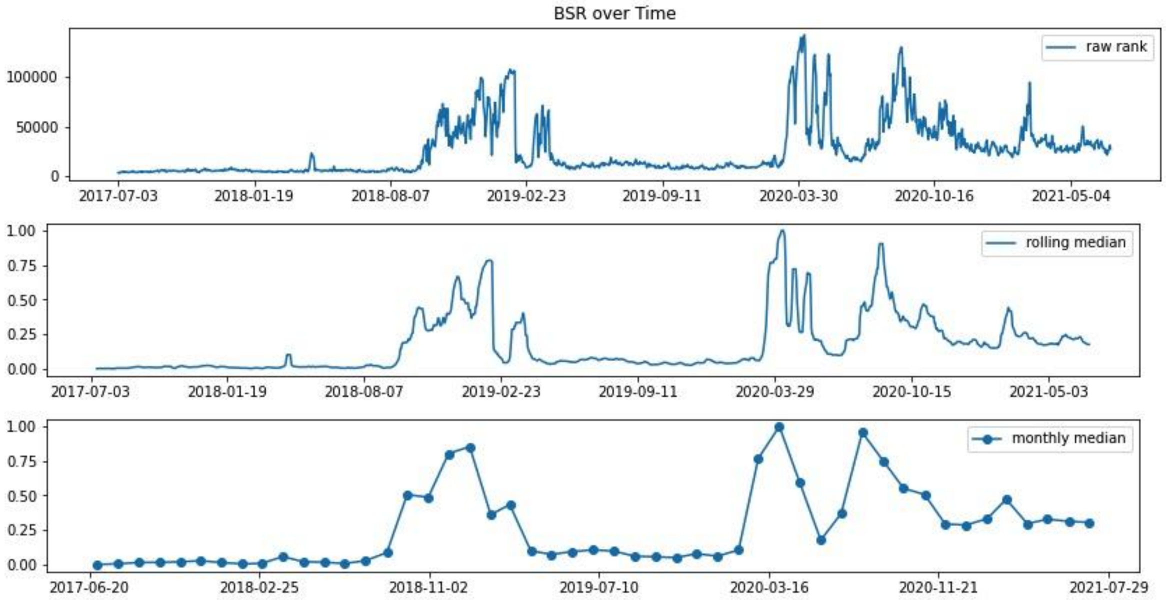
~9000 products in Amazon's Vitamins and Dietary Supplements category

Time range: 2017-07 to 2021-07

- ★ **Target variable: Amazon best seller rank data (BSR)**
  - ○ Daily rank of each product based on 1) current sales and 2) sales history
  - ○ Lower rank means better sales performance
- ★ **Predictors: Review data**
  - ○ All reviews under a product at the time of scraping
  - ○ Review title and review text
  - ○ Metadata such as review dates, review ratings, verified purchase, etc

EDA: Best Seller Rank

- ★ Rank data has high volatility
- ★ We aggregate rank by month after smoothing and normalizing daily ranks



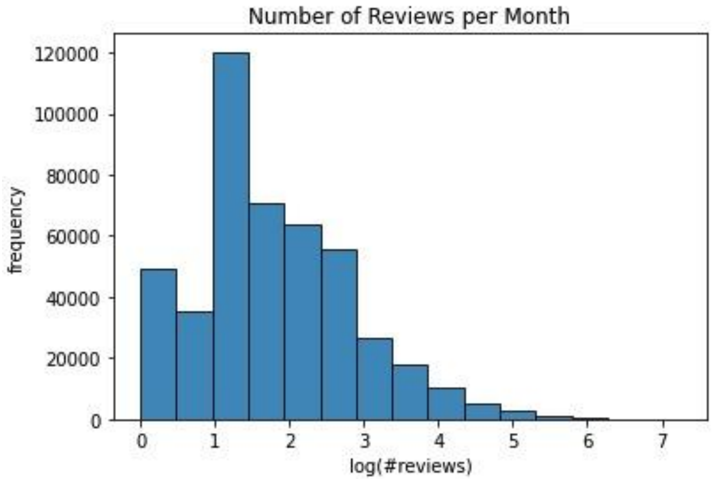
EDA: Review Data

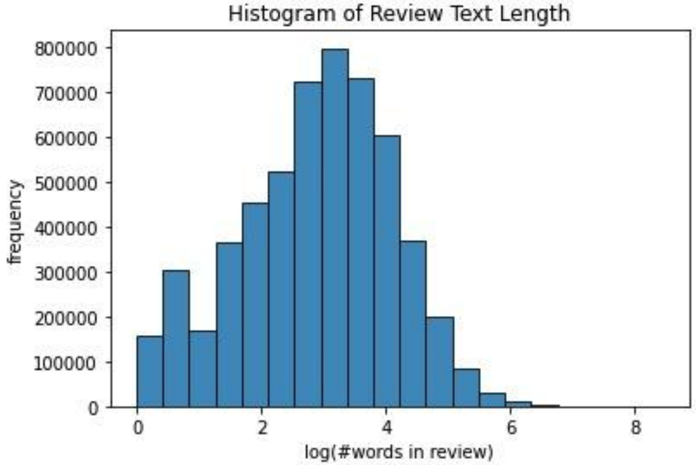
- ★ Varying number of reviews per month
  - ○ max #reviews per month = 1392
- ★ Varying length of reviews
  - ○ max #words per review = 4643
- ★ > 660,000 unique words

Example reviews:

“Horrible product, my mother in law ended up in the hospital with a severe allergic reaction. She had to be in the ICU for a couple of days. Please be careful with this product.”

“good.”



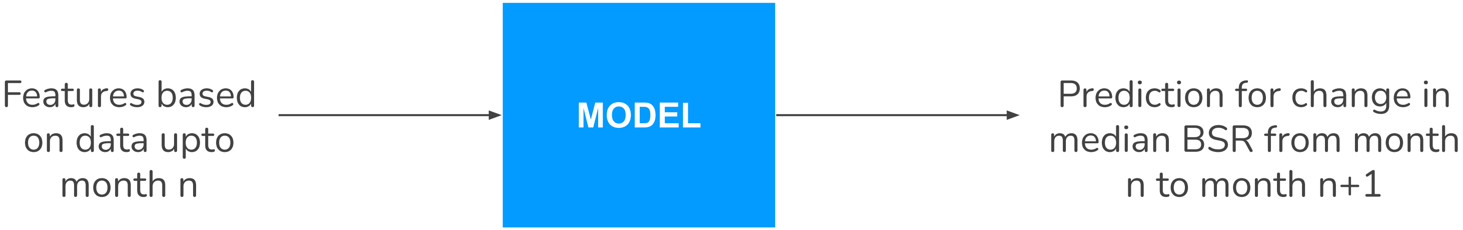


Modelling

Our plan is to ensemble two models

★ A non-text model that considers all the numeric data

★ A text-based model that is mainly influenced by the review texts



Features based on data upto month n

MODEL

Prediction for change in median BSR from month n to month n+1

Non-Text Regression Model

- ★ **Predicting BSR for the next month with no text in the input**
- ★ **Features in the input:**
  - ○ Daily median BSR over a rolling period of 10 days for the 30 days in the previous month
  - ○ Review ratings and verified status, fed in as weighted averages, weighted by the number of upvotes a review received.
- ★ **Models employed and OOD performance (MSE and R^2 score):**

Model	**MSE**	**R^2**
Linear Regression	0.016	0.222
XGBoost	0.014	0.311
Random Forests	0.013	0.331

Non-Text Regression Model

- **Future Work**
    - Train and test across whole data set (will be done on remote servers, once we have access to cloud services)
    - Hyper-parameter optimisation
    - Creation and inclusion of more features such as number of reviews, ratings of the 5 most popular reviews, and time series price data
    - Feature importance analysis

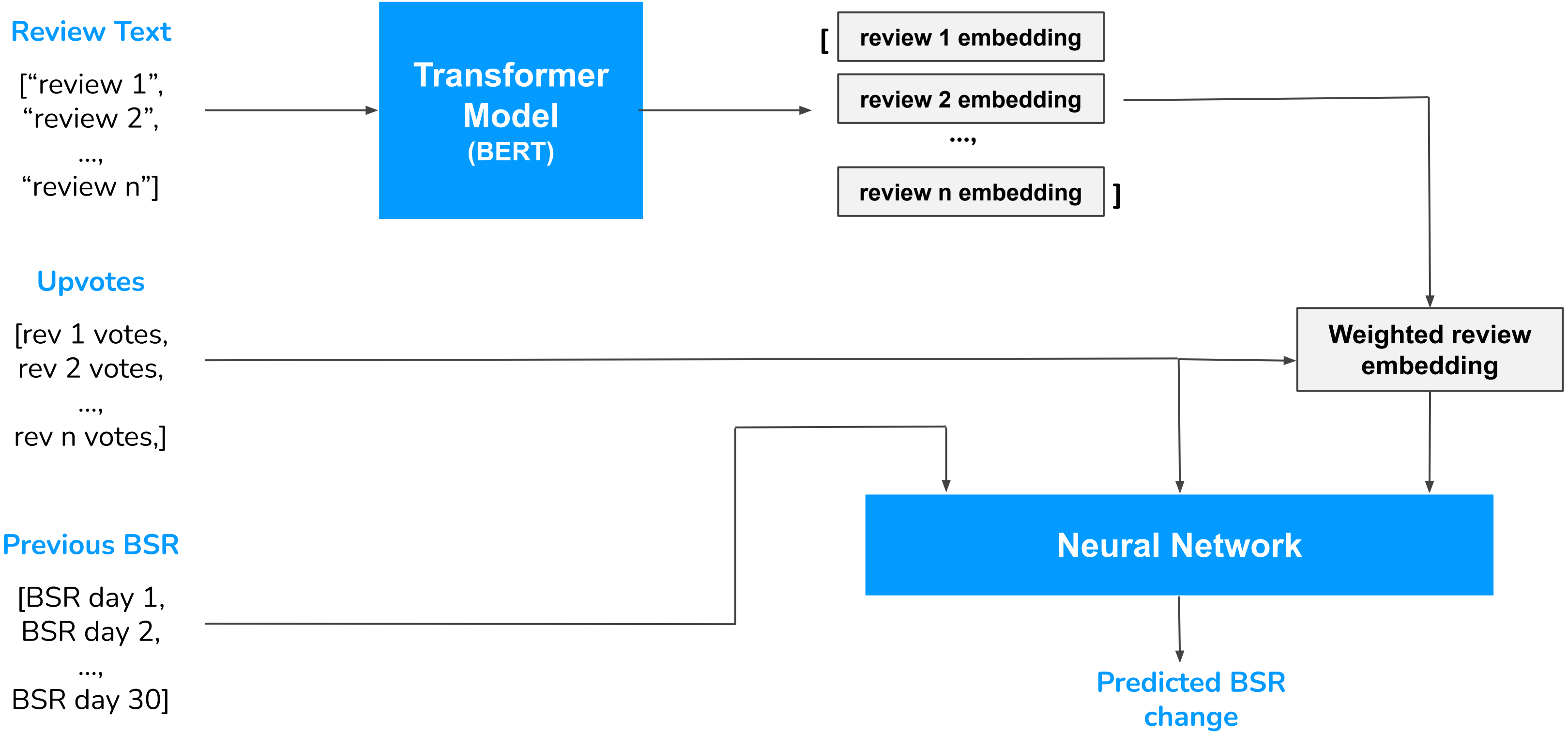
      (i) How important is each feature overall

      (ii) How does feature importance change with changing feature values

Text-Based Regression Model

- ★ **Predicting BSR for the next month with review text in the input**
- ★ **Features in the input:**
  - ○ Transformer (BERT uncased) embeddings for each of the reviews so far, weighted by the number of votes the review received
  - ○ Moving median BSR for the previous month and total upvotes for the product
- ★ **Model Performance after very limited training**
  - ○ MSE: 0.0504

Text-Based Regression Model



Review Text

["review 1",
"review 2",
…,
"review n"]

Transformer Model
(BERT)

[review 1 embedding
review 2 embedding
…,
review n embedding]

Upvotes

[rev 1 votes,
rev 2 votes,
…,
rev n votes,]

Weighted review embedding

Previous BSR

[BSR day 1,
BSR day 2,
…,
BSR day 30]

Neural Network

Predicted BSR change

Text-Based Regression Model

★ **Future Work**

  ○ Using just the current month’s reviews instead of cumulative reviews

  ○ Exploring other transformer models, especially tiny BERT

  ○ Removing upvote count as a feature and exploring impact of that (since there are concerns of data leakage)

  ○ Train and test on whole dataset once we have more computing resources

  ○ Training a baseline bag of words model for comparison with transformers

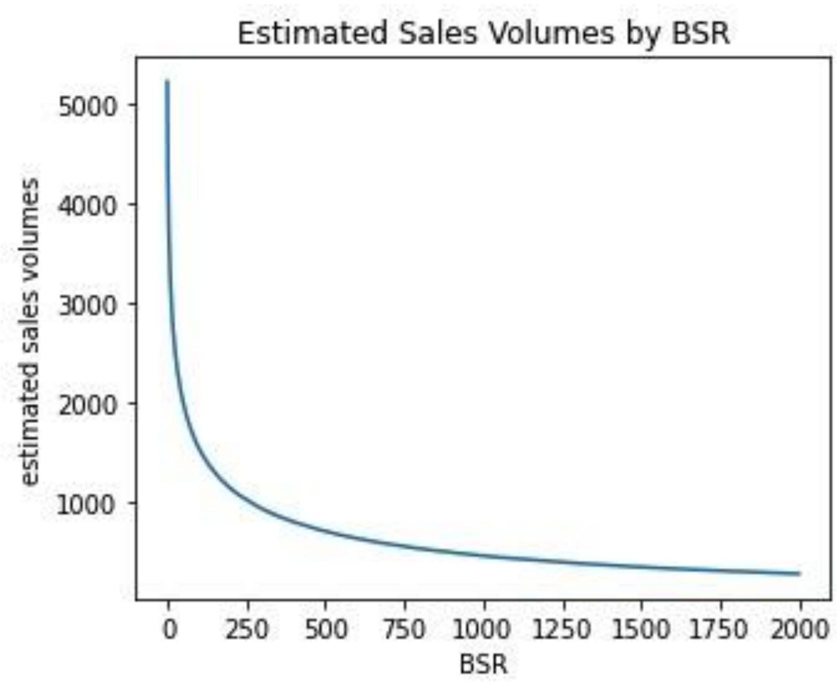
References

- “**Using NLP to extract quick and valuable insights from your customers’ reviews**,” https://medium.com/artefact-engineering-and-data-science/customer-reviews-use-nlp-to-gain-insights-from-your-data-4629519b518e
- “**Predicting Sales from the Language of Product Descriptions**,” https://nlp.stanford.edu/pubs/pryzant2017sigir.pdf
- “**Amazon Product review Sentiment Analysis using BERT**,” https://www.analyticsvidhya.com/blog/2021/06/amazon-product-review-sentiment-analysis-using-bert/
- “**Sentiment Analysis of Movie Reviews with Google’s BERT**,” https://medium.com/mlearning-ai/sentiment-analysis-of-movie-reviews-with-googles-bert-c2b97f4217f
- “**Amazon Best Seller Rank**,” https://www.sellerapp.com/amazon-best-seller-rank.html

Question?

Appendix

- ★ Pattern has provided **estimated sales volumes** corresponding to each rank


Estimated Sales Volumes by BSR
estimated sales volumes
BSR