Helpfulness Recommendation of Amazon Reviews

Ricky Deng
University of California, San Diego
La Jolla, California, U.S.A.
rdeng@ucsd.edu

Ryan Rickey
University of California, San Diego
La Jolla, California, U.S.A.
rtrickey@ucsd.edu

ABSTRACT

In the age of online shopping, the reviews of an item can be one of the most valuable resources to consult before making the decision to purchase. Items with more reviews may appear more trustworthy to a potential customer, as more people have tried it out and spoken their mind. That being said, it is often the case that some reviews are more helpful than others. Maybe someone had a bad day and turned to the review section as a means to vent. Maybe someone has a lot of experience purchasing one type of good, and they are able to provide context on their insight about an item's quality. Which reviews should be shown to a user first? Can a recommender system be built to predict the helpfulness of an Amazon review before users react to it? Is a particular user more likely to leave a helpful review than another? Our model seeks to answer these at reasonable accuracy.

Keywords

Recommender Systems; Review Helpfulness; Video Game Reviews, Amazon Reviews.

1. INTRODUCTION

We are primarily interested in predicting the "helpfulness" of a given Amazon review. Amazon offers a built-in feature to allow "helpful votes" on reviews, however this feature relies on human selection that is inherently unreliable due to varying levels of exposure. Thus, we seek to model the relationship between various features of review data with its "helpfulness" to other users, an implicitly human feature. This predictive modeling can assist with recommending reviews that are very helpful– but not necessarily highly voted for– due to lack of voting data, that is, human assistance.

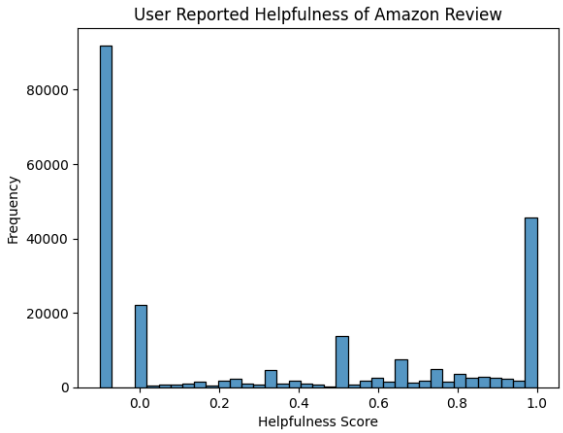

Figure 1: A plot of "helpfulness score" for each Amazon review, derived from user provided helpful votes over total votes

Please refer to Figure 1. We obtained our data as a specific subset of Amazon reviews from May 1996 - July 2014 by users and items with at least 5 reviews, specifically in the Video Game product category[1]. Our reasoning was as follows: although a more recent dataset extending to 2018 was listed, we discovered one major issue in that, the particular dataset only labeled reviews with "helpfulness votes" and not

[1]cseweb.ucsd.edu/~jmcauley/pdfs/emnlp19a.pdf

total votes. This would not show the total user interactions with the review, preventing us from scaling a helpfulness score. We also chose the Video Game category and selected only 5-core (ie, there are at least 5 entries of each type in the dataset) data because this allowed us to maintain our sufficiently large dataset while still ensuring we had high quality data from high quality users. In Figure 1 we have scaled user given helpfulness votes out of total votes, however we notice immediately that the majority of Amazon reviews do not fall on our scale because they have received no user interaction– that is, no voting. Our predictive model can help to remedy this situation by using inherent review features to predict a helpfulness score rather than unreliably expecting users to rate all reviews.

2. THE MODEL

Although removing these "null helpfulness" reviews may be an intuitive notion, we realized doing so cut off a large portion of our data we would need to train our data. Testing our baseline model on this approach would yield mediocre results and anything further would exponentially increase our mean squared error (from here on, MSE). Instead, we choose to map helpfulness votes of 0 to a scaled score of 0, regardless of total vote quantity. In practice this means 0/0 scores (untouched reviews) are labeled as a 0 scaled helpfulness score.

2.1 Identifying a Baseline

We begin with the classic univariate linear regression model with the review length as our feature, which I believe to positively correlate helpfulness. Defined simply by:

$$X\theta = y \quad (1)$$

$$\theta = (X^T X)^{-1} X^T y \quad (2)$$

I've taken the liberty of deriving (2), the formula for the coefficient, theta, from (1), represented in sklearn as .coef_. Before we make any adjustments to our baseline model, it is imperative we examine our baseline theta, however if you're following along, I'm sure you have reached a relatively simple predictor:

$$\hat{y} = \Theta_0 + \Theta_1 * L \quad (3)$$

Where, $\hat{y}$ is our predicted helpfulness score and L represents length of review text. Now we can comfortably inspect our bias term Θ_0 which our model returns as approximately 0.285. Although not high, it's certainly positive. This suggests that even for reviews of little to no text (zero in length), our model's baseline helpfulness score indicates that these will still tend to be perceived as some level of helpful. Similarly, Θ_1 is approximately 8×10^{-5}, suggesting that indeed there is a positive correlation between review text and helpfulness, as hypothesized. Our MSE is also an excellent .159 courtesy of our high quality data set and simple model. Of course these values will shift, but this assures us that we can move on from our baseline as everything is in order. I will additionally consider summary length and overall rating, but of course not the given helpful votes because that would simply be a theta of 1.

2.2 Further Models

Due to the nature of our relevant features: review length, summary length, and overall rating, we can actually be somewhat confident in our multi-featured linear regression model performing as well or better than most other more complex models.

$$y_i = \begin{cases} 1 & \text{if } X_i \cdot \theta > 0 \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

$$\sigma(t) = \frac{1}{1+e^{-t}} \quad (5)$$

Consider the logistic regression formula (4), even supposing we applied some sigmoid function (5), we would still be severely limited by the binary nature of classification based models. This effectively eliminates our ability to use any sort of classification such as Jaccard similarity, without drastically decreasing our

accuracy. Instead my options would be based primarily on text based approaches such as TF-IDF and cosine similarity, which we will explore in training. For reference, we know Cosine Similarity(A,B) is the dot product of word vectors A and B divided by the magnitude of the vectors multiplied. We can define a rudimentary pseudocode implementation for sets:

```
def cosine(set1,set2)
        numerator=length(set1.inter(set2))
        denominator=sqrt(length(set1))*
                    sqrt(length(set2))
        If denominator ==0:
            return 0
        return numerator/denominator
```

3. TRAINING

Features	Θ	MSE
Review length	[2.84825094e-01, 7.99856883e-05]	.159162
+summary length	[2.35666595e-01, 7.30167590e-05, 2.10386671e-03]	.1580303
+overall rating	[2.18219593e-01, 7.31545439e-05, 2.12197384e-03, 4.11112053e-03]	.158014

Figure 2. Linear Regression Model Optimization through Text and Rating

Revisiting our best performing model, linear regression, our theta is listed as bias, review length, summary length, and overall rating coefficients. As evidenced by extraordinary MSE even univariately, our linear regression outperformed even our best combination of TF-IDF and Cosine with all features– which we achieved approximately 0.48, still nearly half a point in scaled helpfulness. However, in our testing we identified that Θ_3, review rating was the greatest predictor and a positive predictor at that. Although summary length, which we expected to be our best predictor, performed great at 0.002, overall rating appears to have almost double the effect on helpfulness. Contrary to what one might expect, the higher the rating, the more helpful the review is perceived to be.

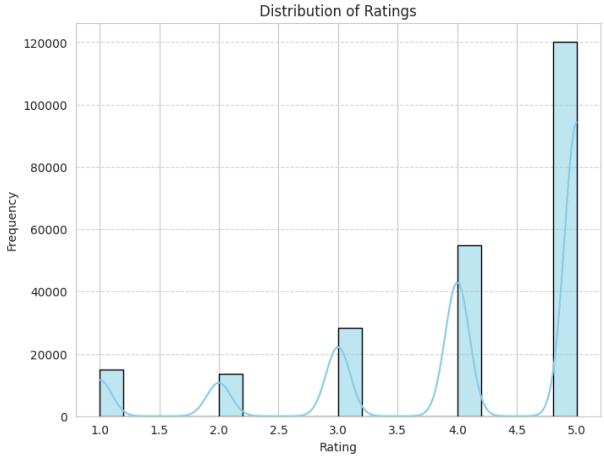

Figure 3. Distribution of Ratings across Data

We see in Figure 3 that a majority of our data on reviews actually sits at very high ratings. This could account for the positive correlation between rating and helpfulness. However, plenty of our data was marked as 0 because of the lack of votes, thus it is reasonable to assume that there were also a fair amount of the high scores that were mapped to 0 helpfulness that could have prevented a stronger correlation from forming.

Perhaps not regretfully, but our exploratory data analysis and high quality dataset has left us qualitative features unresponsive to classification but highly predictive.

4. EXPERIMENTS

With our modeling work done, recall the predictor equation (3) from 2.1 Identifying a Baseline:

$$\hat{y} = \Theta_0 + \Theta_1 * L \quad (3)$$

We have confirmed that this linear regression is the most optimal method of predicting the relationship between the features (review text length, review summary length, and overall review rating) and the review's helpfulness score, scaled to a unit of 1. Thus, we can apply our formula to all 3 features:

$$\hat{y} = \Theta_0 + \Theta_1 * L + \Theta_2 * S + \Theta_3 * R \quad (3)$$

We have thus mapped a clear relation for our predicted helpfulness based on our defined features. To put this into practice we can apply on some newer data from the Amazon reviews dataset.[2]

5. RELATED WORK

The task of predicting the helpfulness of reviews based on their length finds resonance with various research endeavors in recommender systems and user-generated content analysis, while differing from others.

On our dataset, we found quantitative factors to be the best measure of predicting review helpfulness. Others have shown that qualitative evidence, such as reviewer experience, reviewer impact, and reviewer cumulative helpfulness, are factors that may uncover new insights [1]. Nonlinear regression models seem to be optimal in some cases, for such helpfulness prediction on a dataset of movie reviews [2]. While including the length of the title (or 'summary' field) improved our model, others have shown such title features to be weak determinants of online review helpfulness [3].

Beyond rudimentary regression models, many services and approaches seek to leverage more nuanced machine-learning approaches to predict helpfulness and make recommendations. Jinni [4] leverages more tags in the metadata to provide content-based recommendations in entertainment. In improvements made to the Netflix Prize model, matrix factorization and temporal temporal elements of the data are included to build a predictor. [5, 6].

Many image-based recommendation approaches have been taken regarding fashion trends and styles and substitutes from this same dataset of Amazon reviews [7, 8]. Indeed, we ignored the image metadata for our analysis task under the assumption that it would not help predict helpfulness, but it is definitely useful for other tasks.

6. CONCLUSION

In this study, we delved into the realm of predicting review helpfulness within a repository of 5-core video game reviews from Amazon. Although limited to instances of games and users that have 5 reviews each, we still found many reviews that hadn't been rated as helpful. We ultimately assigned these reviews a 'helpful' score of 0 in the training of our linear regressor. The length of the review text and the length of the summary turned out to be the most powerful predictors on this dataset. While a quantitative approach has been shown to not offer the best performance on helpfulness prediction, we found it to fit our desires in searching for a reasonably good prediction. Granted, the specific limitations of our dataset may limit this approach in the wild, but it sheds light on a basic notion: assuming users aren't typing gibberish on items that already have reviews, a linear regression model does a decent job of predicting whether or not the review will be useful.

References

[1] Albert H. Huang, Kuanchin Chen, David C. Yen, Trang P. Tran, A study of factors that

[2] cseweb.ucsd.edu/~jmcauley/pdfs/emnlp19a.pdf

contribute to online review helpfulness, Computers in Human Behavior, Volume 48, 2015, Pages 17-27, ISSN 0747-5632, https://doi.org/10.1016/j.chb.2015.01.010. (https://www.sciencedirect.com/science/article/pii/S0747563215000229)

[2] Y. Liu, X. Huang, A. An and X. Yu, "Modeling and Predicting the Helpfulness of Online Reviews," 2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy, 2008, pp. 443-452, doi: 10.1109/ICDM.2008.94.

[3] Akbarabadi, M., & Hosseini, M. (2020). Predicting the helpfulness of online customer reviews: The role of title features. International Journal of Market Research, 62(3), 272-287. https://doi.org/10.1177/1470785318819979

[4] http://www.jinni.com/about.html

[5] R.M. Bell, Y. Koren, and C. Volinsky. The bellkor solution to the netflix prize, 2007.

[6] Koren, Y The BellKor Solution to the Netflix Grand Prize, 2009.

[7] Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering, R. He, J. McAuley WWW, 2016

[8] Image-based recommendations on styles and substitutes, J. McAuley, C. Targett, J. Shi, A. van den Hengel, SIGIR, 2015