WHITE PAPER

By Renee De Wolf,
MS MBA CFE,
Sr. Data Scientist,
Appriss Retail

Consumer Shopping Behaviors: Online Reviews and Sentiments Analysis

Consumer online shopping behaviors can be predicted. In a recent analysis we used a women's apparel ecommerce review dataset to predict whether an item would be "recommended" by the consumer based on sentiments analysis of the review's text and attributes of the item. Intuitively, the use of positive words in reviews increases the likelihood of a positive recommendation, but we also saw that consumers found the most useful reviews sometimes warned against a purchase vs. encouraging them towards a purchase. Can this be applied beyond purchases? Yes, this process could be deployed in the future to predict behaviors regarding the relationship between negative product recommendations and returns to help prevent returns proactively.

In this analysis, we review how a machine learning model can be used to predict whether an item would be "recommended" by the consumer, based on sentiments analysis of the review's text and attributes of the item.

Data & Methodology

For this analysis, we used a women's ecommerce review dataset that was de-identified and contributed by a retailer to the public domain for research. The scrubbed dataset included 22,628 observations and 10 variables/features:

- **Clothing ID**: Integer Categorical variable that refers to the specific piece being reviewed.
- **Age**: Positive Integer variable of the reviewer's age.
- **Title**: String variable for the title of the review.
- **Review Text**: String variable for the review body.
- **Rating**: Positive Ordinal Integer variable for the product score granted by the customer from 1 to 5, Worst to Best.
- **Recommended IND**: Binary variable stating where the customer recommends the product where 1 is recommended, 0 is not recommended.
- **Positive Feedback Count**: Positive Integer documenting the number of other customers who found this review positive.
- **Division Name**: Categorical name of the product high level division.
- **Department Name**: Categorical name of the product department name.
- **Class Name**: Categorical name of the product class name.

The goal of the analysis is to try and predict a Recommended Indicator (RI); i.e., would the consumer recommend the item being reviewed. 18,527 (81.88%) of all reviews has a recommendation indicator =1/Yes.

General Analysis

As illustrated in Figure 1, most ratings 3.5 and above, 18,527 (x81.88%), resulted in a RI =1. As expected, the most mixed recommendations came from a Rating of 3 while Ratings of 4 or 5 solidly corresponded to a positive recommendation while Ratings of 1 or 2 consistently corresponded to a negative one. Distributions for other categorical variables such as Positive Feedback Count were also visualized to look for obvious disproportions that might indicate a strong relationship to the recommendation indicator for specific categories.

Few disproportions were observed visually so mathematical tests for both the categorical and quantitative variables were executed to support feature selection. Figure 2 illustrates the split of the recommendation indicator across the distribution of how many other customers found this review positive/useful. The distribution of Age was also visualized, and percentile statistics were calculated for use in created even width Age Group bins for modeling. The median age of reviewers in this dataset was 41 and 75% of reviewers were 52 or younger. Figure 3 shows the distribution of Age split by Recommendation Indicator.

Figure 1: Recommended Indicator vs Consumer Rating

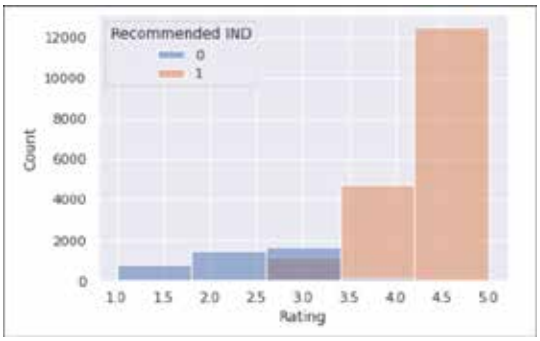


Figure 2: Recommended Indicator vs Positive Feedback Count

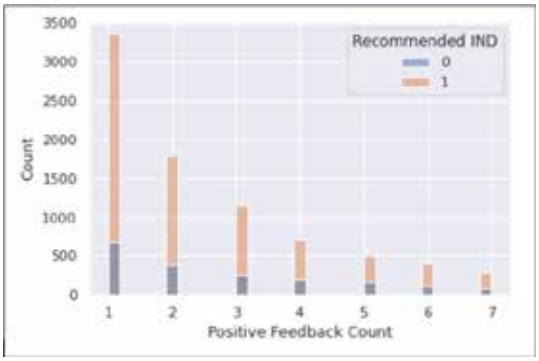
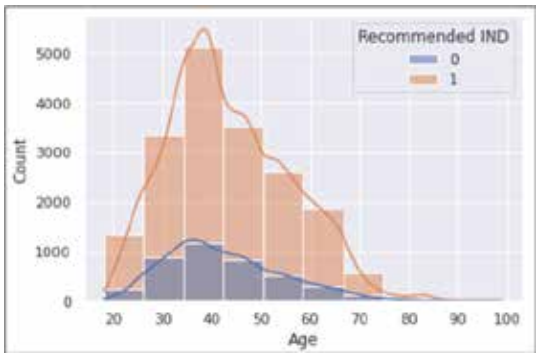


Figure 3: Recommended Indicator vs Age



APPRISS® RETAIL

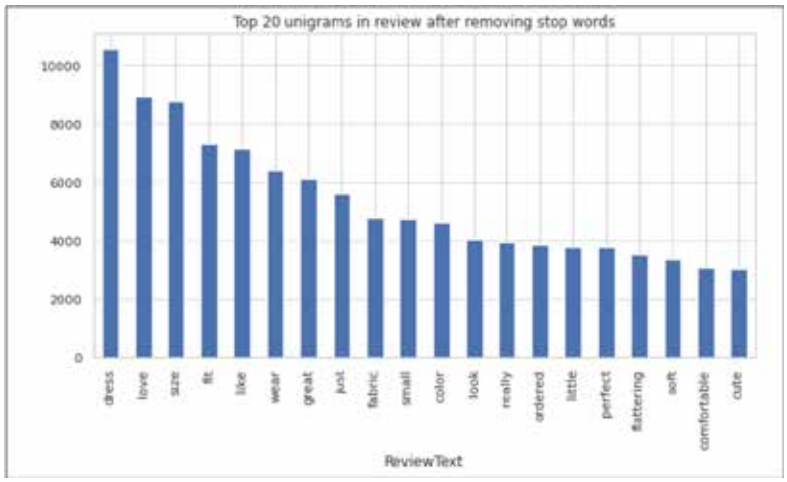
Additional exploratory analysis showed that the item's SKU hierarchy (division, class, department) were also relatively evenly distributed across the Recommendation Indicator based on appearance and therefore less likely to be strong predictors.

Natural Language Processing Analysis

Natural Language Processing (NLP) is part of Artificial Intelligence (AI) concerned with how humans and computers interact using natural language. Specifically, it studies how to program computers to read, understand, and analyze language. In our analysis, NLP was conducted on the text of the item review itself. As part of the analysis, we excluded words that are considered to be stop words (such as "and", "or", "by"). For the purposes of NLP stop words are frequently excluded because they are words that, while very common in written language, don't tend to add value to text analysis. For the purposes of this analysis, we did not employ lemmatization or stemming (reducing them to their root word).

Reduction of words to their root can often be useful in text analysis. In this case however we determined that it was preferable to omit this step to improve the single word sentiment extraction. A consumer for example, may say that something "fits", but would not typically say that some something does NOT "fits". As a result, rooting both "fits" and "fit" down to "fit" by removing the "s" from the former would potentially erode some insight around whether the sentiment of a review was positive or negative based on its inclusion of that word. Illustrated in Figure 4 are the top 20 frequency words across all reviews, good and bad.

Figure 4: Top Frequency Words in the Review Text



As the next step in the analysis, words (unigrams) were identified as "positive", based on their appearance in reviews associated with ratings of 4 or 5, or "negative", based on their appearance in reviews associated with ratings of 1 or 2. Words that appeared in both "positive" and "negative" groups were excluded from both. We created two variables using the count of positive and negative exclusive words in a review.

Modeling

Prior to modeling, we cleaned and scaled all of the variables and corrected the imbalance of the dependent variable (RI). Imbalance corrections are used to avoid the model missing rare events by preferencing the higher frequency event (in this case a recommendation of yes).

Two tree-based models were attempted, and the data was split into test and train sets for modeling. Shapely values were also calculated to assist in visualizing feature importance. Shapely values are an artifact of game theory wherein each prediction row represents a "game", each variable represents a "player" in the game and the contribution of each player to the game is represented by a Shapely value.

The mean accuracy of the models was 70%. More importantly, however, the mean Recall for RI(1) was 74%. Overall the models performed reasonably well at identifying items that would be recommended by a consumer.

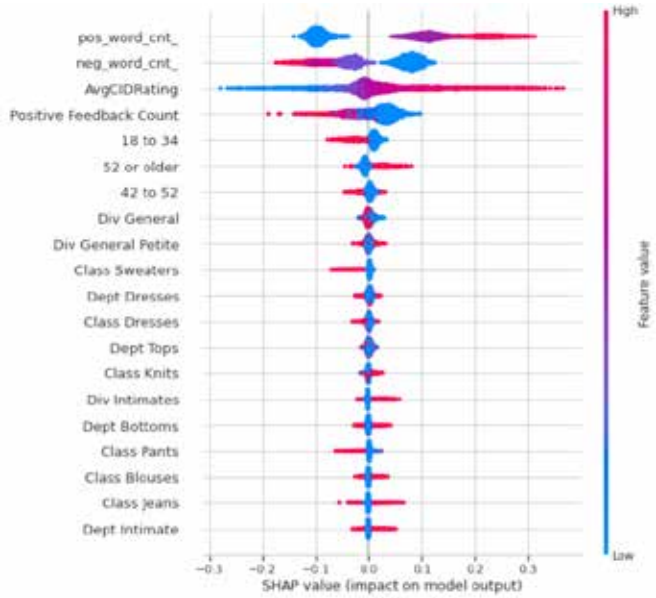
Given that it so cleanly divided the data in 4 of the 5 groups, the Rating variable was not entered into to the modeling directly. This was done to determine how well the other features would predict the recommendation outcome when Rating was unknown. As previously outlined, the Rating variable was used indirectly to support classifying the reviews as positive or negative as part of the NLP process to create the positive and negative sentiment word count variables that did become part of the model.

Those sentiment features were the two most high impact features in both models. The count of consumers indicating that review provided positive/useful feedback, the historical review rating of the Clothing ID being reviewed, and whether the consumer was in the 18-34-year-old age group rounds out the top 5 variables that models deemed most important to the prediction.

Figure 5. illustrates the visualization of the shapely values of the model variables. The features at the top hold the most weight in the model decision. Each dot represents a review with its shapely value for each variable. Blue dots indicate that the likelihood of a positive recommendation was reduced by that variable, while a red dot indicates that the likelihood of a positive recommendation was increased. Intuitively we can see that as the number of positive words increases so does the likelihood that the recommendation is positive with the reverse being true for negative words. Notably, the Positive Feedback Count had a slightly negative relationship with recommendations. This might indicate that reviews that consumers found the most useful leaned somewhat towards the ones that warned against a purchase vs. encouraging them towards one.

Reviews that consumers found the most useful leaned somewhat towards the ones that warned against a purchase vs. encouraging them towards one.

Figure 5: Confusion Matrix of Random Forest Model



Limitations

This analysis was limited specifically to women's apparel. It's possible that the outcomes may not generalize to other verticals or segments of products. In our analysis we limited the natural language process to the parsing of unigrams and didn't use optimizers like lemmatization, stemming or custom stop words. While we still achieve some level of feature importance it is likely that this could be improved upon through optimization or expansion to bi-gram or tri-gram phrase analysis. We only attempted one type of target rebalancing here. Other under-sampling techniques such as ENN may be useful as well.

Summary and Future Work

Future work here would be to examine the relationship between negative recommendations and returns to determine whether real-time recommendation sentiment might be directly useful in understanding consumer issues with certain products as they purchase to prevent returns proactively.

Americas +1 949 262 5100
Europe/Middle East/Africa +44 (0)20 7430 0715
Asia/Pacific +1 949 262 5100

APPRISS® RETAIL
apprissretail.com