Executive Summary

Abigail Ajamian

Course D214

11/1/24

Problem and Hypothesis

This study dives into Sephora's e-commerce website to gain insight into customer satisfaction with products and allow the implementation of data-driven decisions for marketing strategies. The dataset holds 8,494 observations, each of which is a product found on Sephora's website. Remember that although the dataset exceeds the required 7,000 observations, this is still a small sample of products on the website. The dataset also has 27 columns that have data on each observation.

The research question presented is, "Can an ordinal logistic regression be performed to determine product characteristics that influence the rating of a product?" The null hypothesis is that an ordinal logistic regression model cannot be made using this data to determine product characteristics that influence the rating of a product. The alternative hypothesis is that an ordinal logistic regression model can be created to determine product characteristics that influence the rating of a product with an accuracy> 70%.

Data Analysis Process

Data Preparation

Cleaning the data of missing values, duplicates, and outliers was the first step to the data for analysis. This dataset had no duplicates but had missing values and outliers that needed treatment. The columns with a high percentage of missing values were dropped entirely.

The rating column had less than 3% missing values. Therefore, the observations with the missing values were dropped. This also removed all missing values in the reviews column. For any other columns that contained missing values, the values were imputed. Any outliers detected were imputed with the mean or mode, dependent on the variable data type. Next, categorical variables were encoded to be represented numerically to prepare the data.

The dependent variable needed to be grouped due to the ordinal regression analysis. This was done by implementing KMeans clustering with the number of clusters being 5, standing for ratings 1-5, and stored in a new column rating_clusters. The last step in data preparation was removing all unnecessary columns.

Analysis

Exploratory Data Analysis (EDA) was completed to extract univariate and bivariate statistics of the data. A few patterns were noticeable; this included higher-rated products tended to have more reviews. The dataset was then split into training and testing sets using a 70-30 split.

After EDA, the first ordinal logistic model was created using mord.LogisticIT function and fit to the x_train and y_train datasets. This was used to make predictions and output classification_report that showed the accuracy of this model as 39%. Since this is not the wanted accuracy, feature elimination was implemented via SequentialFeatureSelector. This returned the

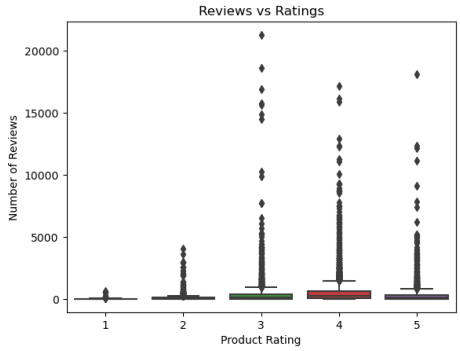
k best (4) variables based on accuracy score. These features were then used to create a reduced model, which decreased the model's accuracy by 1%.

Since this model shows a poor fit, a multinomial logistic regression was also completed to decide if the data satisfies the ordinal logistic regression assumption of proportional odds. This was done using the LogisticRegression function with the multiclass attribute set to multinomial. This model's accuracy was calculated using the accuracy_score function on the prediction and actual y_test values. The output was an accuracy of 39%, showing no improvement. Therefore, it is concluded that the proportional odds assumption does not cause a poor fit of the ordinal regression model.

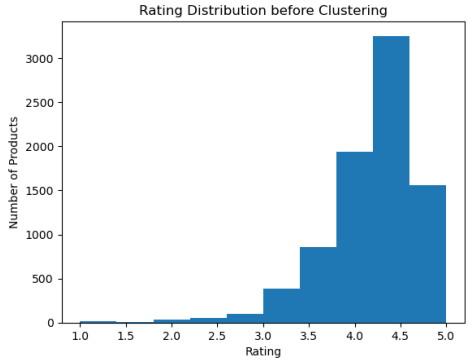
Findings

Univariate and Bivariate Analysis

1. Products with higher ratings often have a higher number of reviews.



2. Product rating distribution is unimodal, with the peak around the ratings 4-4.5.



Ordinal Regression

We failed to reject the null hypothesis, which indicates that the analysis could not identify predictor variables that had statistical significance to the rating of a product.

Limitations

1. An ordinal logistic model assumes proportional odds (Lee, 2019). A Brant-Wald test is used to identify if the dataset satisfies this assumption. Since Python does not have this ability, it needs to be concluded that proportional odds do not impact the fit of the ordinal regression model by comparing the accuracy score to a multinomial regression model's accuracy.
2. Although the dataset meets the requirement of 7,000 observations, 8,494 observations is still a small sample size for the number of products on the Sephora website. This can create a challenge when detecting statistical significance and relationships (Analytics, 2024).

Proposed Actions

There are two courses of action that are recommended for further study. The first is to gather more data on Sephora products that can be joined with this dataset and then complete the ordinal logistic regression again. The second course of action is to use the written reviews of the products to create a sentiment analysis. Sentiment analysis will classify the review's sentiment "as positive, negative, or neutral" (Gupta, 2018). With this, Sephora could still gain insight into customer satisfaction.

Benefits of Study

This study's expected benefits include informing Sephora of the relationship between the independent variables used in the ordinal regression and product ratings. Because we failed to reject the null hypothesis, there is no correlation between product ratings and the 11 independent variables used in this sample set. If the sample data is assumed to have the same relationships as the population data, this result will be helpful for Sephora stakeholders. Due to the results from this analysis, Sephora can implement that any of these 11 independent variables should not be utilized in business decisions to achieve a higher-rated product catalog. This helps Sephora stakeholders and marketers remove these 11 independent variables from consideration, allowing them to investigate other product variables that may correlate with product rating.

Sources

Analytics, E. I. (2024, January 2). *Decoding data size: Pros and cons of working with small data sets*. Medium. https://medium.com/@analyticsemergingindia/decoding-data-size-pros-and-cons-of-working-with-small-data-sets-bc1ea0792da6

Gupta, S. (2018, January 19). *Sentiment analysis: Concept, analysis and applications*. Medium. https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17

Lee, E. (2019, May 29). *Ordinal logistic regression on World happiness report*. Medium. https://medium.com/evangelinelee/ordinal-logistic-regression-on-world-happiness-report-221372709095