Predicting Sephora Product Ratings

Madison Martin

2023-07-27

Introduction

This project aims to predict the user ratings for beauty products sold on the popular website for beauty retailer Sephora. Specifically, it aims to identify the role product attributes like brand, price, online popularity, limited edition offerings, and exclusivity have on customer satisfaction with the products.

Research Questions:

What product characteristics (if any) play a role in predicting the customer satisfaction with the product, as evidenced by the product's rating? Do limited edition products tend to have higher or lower ratings? Does interest in the product seem to correlate with higher ratings?

Sourcing the Data

This data set was sourced from the Kaggle website and uploaded by user Raghad Alharbi, which you can find at this link here. It was collected utilizing web scraping methods during the month of April 2020 from the Sepohra US website.

Download the Data

```
# download the product data from sephora website
# Load the necessary libraries
library(readr)
library(tidyr)
library(ggplot2)
library(tidyverse)
library(dplyr)

# read in csv
sephora_df <- read_csv(file = "sephora_website_dataset.csv")
head(sephora_df)
```

```
## # A tibble: 6 x 21
##        id brand        category name  size  rating number_of_reviews  love price
##     <dbl> <chr>        <chr>    <chr> <chr>  <dbl>             <dbl> <dbl> <dbl>
## 1 2218774 Acqua Di Pa~ Fragran~ Blu ~ 5 x ~    4                 4  3002    66
## 2 2044816 Acqua Di Pa~ Cologne  Colo~ 0.7 ~    4.5              76  2700    66
```

```
## 3 1417567 Acqua Di Pa~ Perfume  Aran~ 5 oz~    4.5                26  2600   180
## 4 1417617 Acqua Di Pa~ Perfume  Mirt~ 2.5 ~    4.5                23  2900   120
## 5 2218766 Acqua Di Pa~ Fragran~ Colo~ 5 x ~    3.5                 2   943    72
## 6 1417609 Acqua Di Pa~ Perfume  Fico~ 5 oz~    4.5                79  2600   180
## # i 12 more variables: value_price <dbl>, URL <chr>, MarketingFlags <lgl>,
## #   MarketingFlags_content <chr>, options <chr>, details <chr>,
## #   how_to_use <chr>, ingredients <chr>, online_only <dbl>, exclusive <dbl>,
## #   limited_edition <dbl>, limited_time_offer <dbl>
```

Precleaning for Analysis

```
# take out the vars we don't think we will use in the analysis
# going to remove heavy text ones: url, instructions, ingredients
product_df <- sephora_df %>%
  select(- URL, - how_to_use, -ingredients)
# view(product_df)
```

I'm already seeing that some products have a different price vs. value_price. I'm guessing these are mostly sets/kits (as you can see if you head the df), but the data documentation does not state this explicitly. I am going to make a new column for these products specifically and keep it in mind as I go through my exploratory analyses.

```
# add column to identify products who have a different price from the "value price"
product_df <- product_df %>%
  mutate(deal = ifelse(value_price - price > 0, 1, 0))
```

Exploratory Data Analysis