Predicting ratings of Moisturizers

Mariam Yousuf

Overview

- Goal is to predict ratings of moisturizers listed on Sephora

- Reviews are important for companies as consumers base buying decision on reviews

- If a company can identify the features that make a highly rated product, they can create similar products

Intro to Data

- Scraped data from Sephora's moisturizer section

- Collected the pink underlined information

- Collected info on 399 products

- 129 brands

- 23 different formulations

- 4 different ingredient callouts

Distribution of Ratings

- Ratings between 0-5 stars, but not evenly distributed

- No ratings between 1.5-2

- Mean of 4.18

Histogram of Ratings
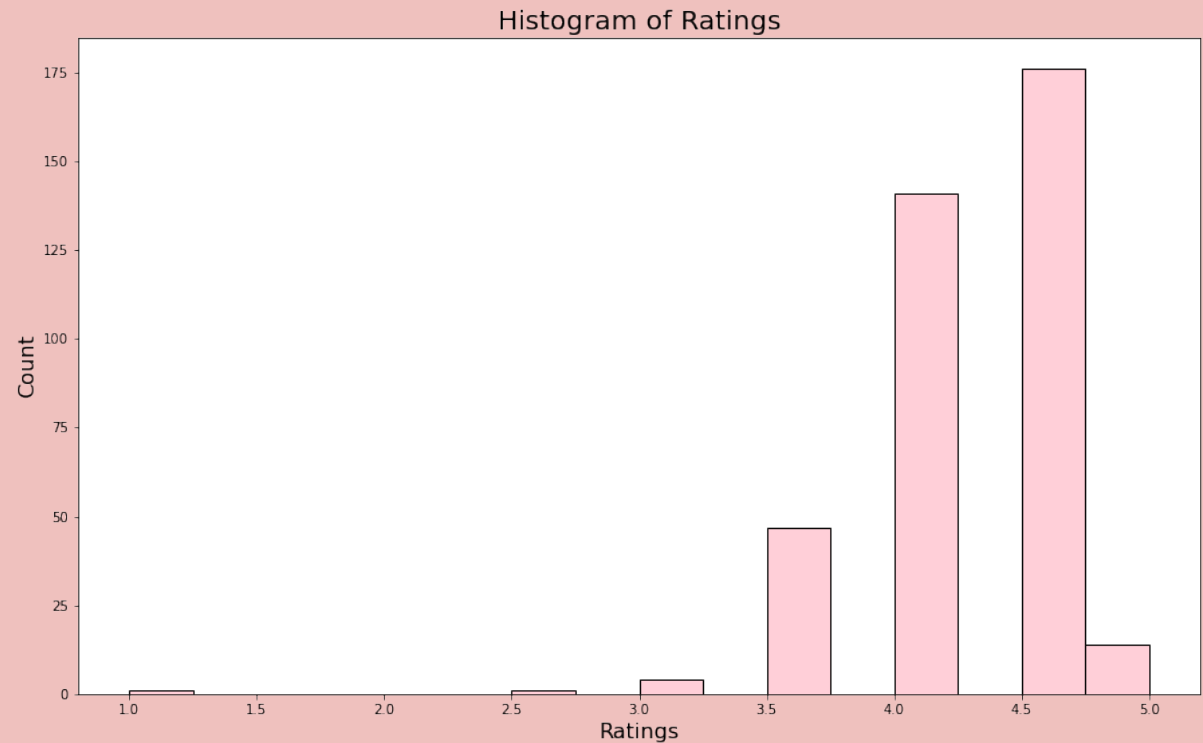
Results of Basic Models

- Basic model (1) includes just price per ounce

- Basic model (2) includes price per ounce and all the dummies for brand, formulation, skin type, ingredient call-outs, and clean at Sephora

- Overfitting clearly a problem

R^2	Training Set	Testing Set
Model 1	0.00326	-0.0419
Model 2	0.570	-4.103E+21

Results of LASSO with Cross Validation

- Ran LASSO with CV on basic model (2) includes price per ounce and all the dummies for brand, formulation, skin type, ingredient call-outs, and clean at Sephora

- 36 variables remained, down from 165

- Overfitting still a problem

- Can't make model less complex, so need more data

R^2	Training Set	Testing Set
Model 2	0.342	-0.541

Future Work

- Add more data

- Suggestions?