Predicting price of skincare products use linear regression model

By: Jing Lee

Abstract

In this project, I am attempting to predict the price of skincare products using a linear regression model. What are the most important factors that influence the price of skincare products? I used collected data on skincare products using webs craping. Then, I use linear regression to model the relationship. Additionally, regression with regularization (lasso, ridge, and elastic net) was modeled. Data set was split into test and training for proper evaluation. The model scores the same on the test and train set using regression with and without regularization. I selected the regression model without regularization since it is less complex and model was not overfitting based on scores.

Design

The clients will be both businesses that sell beauty products and consumers. Businesses can use the model to understand how products are priced when they are presenting a new product. Consumers can use this information to decide which product to purchase based on their budget and needs.

Data

The data set was 1847 skincare products data obtained from www.ulta.com. The target was price. The features of interest were price (target), volume (oz), number of reviews, product name, product description, benefits, ingredients list, clinical benefits, and clean product type.

Algorithms

Feature engineering

1. Creating binary features of the following:
    a. ingredients list: to account whether an active ingredient is present or not
    b. brand name based on list of brands
    c. presence of products to a list of clean ingredients obtained from ulta.com site.
    d. product type: moisturizer vs serum
2. Converting rating to rating^2 based on pair plot where rating has a quadratic shape.

Models

Linear regression and Linear regression with regularization: Lasso, Ridge, and Elastic Net.

Model evaluation and selection

The entire training set of 1501 data was split into 80% train and 20% test. The validation method used was cross validation ideal or a small/medium dataset.

Linear regression without regularization results

- Train R^2 [0.39054358, 0.41968104, 0.37822006, 0.52167151, 0.33326757]
- Mean train R^2 0.409 +/- 0.063
- Test R^2: 0.413
- MAE +- $: 13.66

Tools
Web scraping: BeautifulSoup
Modeling: Sklearn
Visualizations: pandas, matplotlib, seaborn

Communication
Slides and jupyter notebook (code) are available on GitHub:
https://github.com/lee-jin81/metis_project_2_regression