Skincare Product Prediction using Regression

By: Crystal Han

Abstract

The goal of this project was to use data collected from the Sephora website to analyze and predict the price of skincare products. Exploring with several regression models, consumers, distributors, and resellers like Sephora can provide a reasonable price for skincare products that they will make or buy and sell. Products are scraped from the sephora website and gathers data on the ingredients, about, skin type, reviews, price, number of reviews, number of likes, brand, product name, and product type. The product types selected were determined whether the about part of the product could provide us any insight on its suitability of different skin types (Oily, Combination, Normal, Dry, Sensitive). Other product types may include certain ingredients or keywords that were a part of the search and were thus included in the analysis as these products are also used on the human body. After running through with the regression models, I provide the features to consider when buying or selling skincare products and visualize this through plots.

Design

The data for this project is scraped from the Sephora website which presents products that are commonly used on the human body, skin. Comparing reviews and price we see that it is not necessarily the price that affects reviews but the price itself can still affect the number of reviews. Sephora can allocate resources on providing skincare products that are more for combination skin type and a price around the average of 16-20$ to allow consumers to provide more reviews to the product.This data for this project is provided by Metropolitan Transport Authority (MTA).

Data

The dataset was scraped with the following products in mind: toners, cleansers, cushions (foundation), sunscreens, serums, face masks, face oils, face mists, and face creams. The data scraped is categorized by website, brand, product name, product type, about, ingredients, number of likes, and number of reviews. New features are then created based on the information collected such as ordinal encoding of skin types and ingredients.

Algorithms

Feature Engineering

1. Continuous Variables: Reviews, Number of Reviews, Number of Likes, Price
2. Categorical Variables: Brand, Product Name, Product Type, Ingredients, Skin type (normal, oily, combination, sensitive, dry)
3. Combining several hot encoded ingredient features into 3 dimensions using t-distributed stochastic neighbor embedding to simplify modeling
4. Ordinal encoding of brands and product types using price mean
5. Selecting regression model based on performance, in this case Lasso Regression

Models

Linear Regression, Lasso Regression, Ridge Regression, Polynomial Feature Regression, Random Forest Regressor, and Gradient Boosted Regression were used before settling on Lasso regression which has comparable cross-validation performance as the other regression models. Random Forest and Gradient Boosted Regression were dropped from consideration as their performance was not as good as the other regression models. Lasso Regression was used to guide the choice and order of variables to be included.

Model Evaluation and Selection

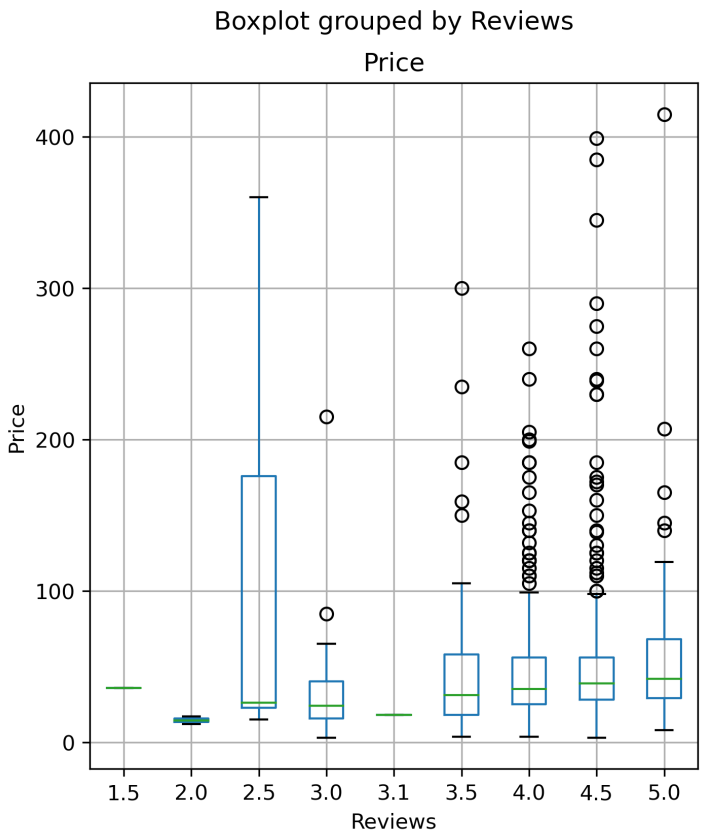
The entire training dataset of 1003 rows of data was split into 67/33 train vs. holdout. All scores reported were calculated with 5-fold cross validation on the testing portion to see how well we fared. Training and testing data were also scored to determine the performance of models.

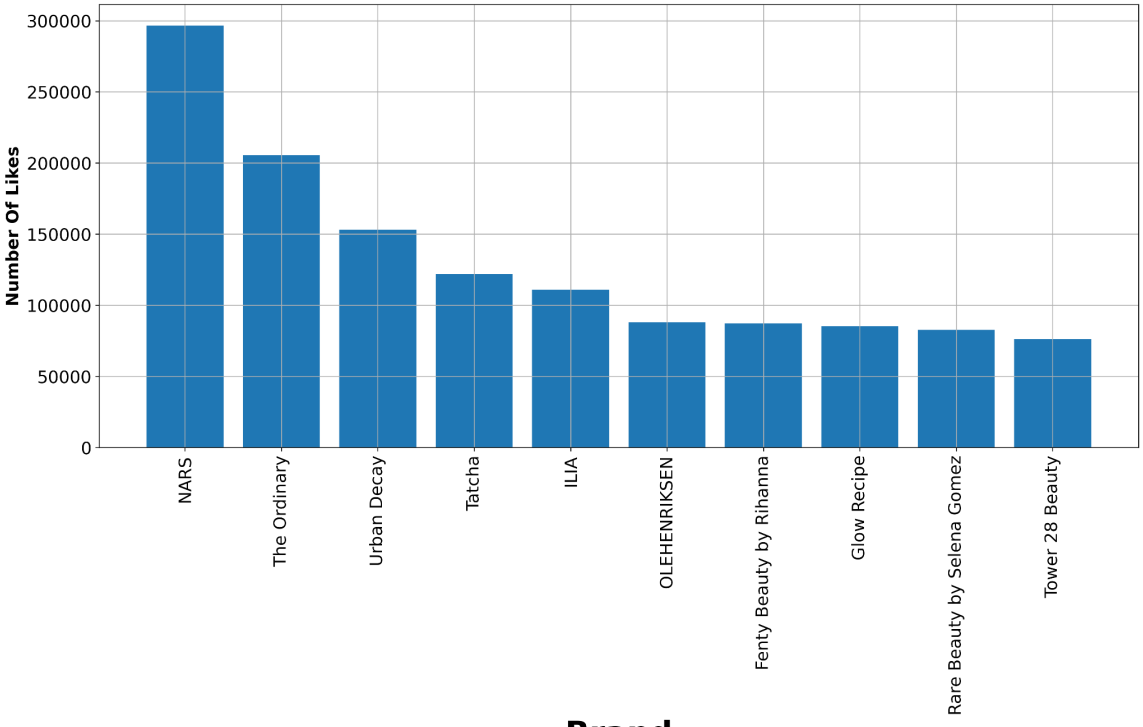
Tools

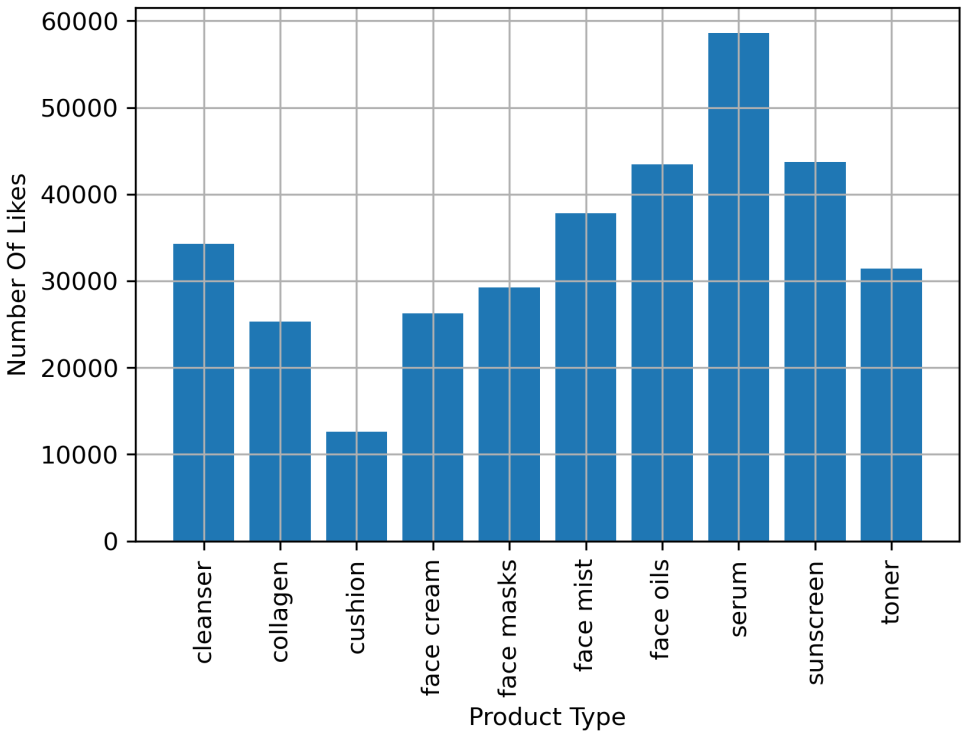
- NumPy and Pandas for data manipulation
- Matplotlib and Seaborn for plotting and visualizations
- Selenium and Beautiful soup for scraping the website
- Scikit-learn for modeling

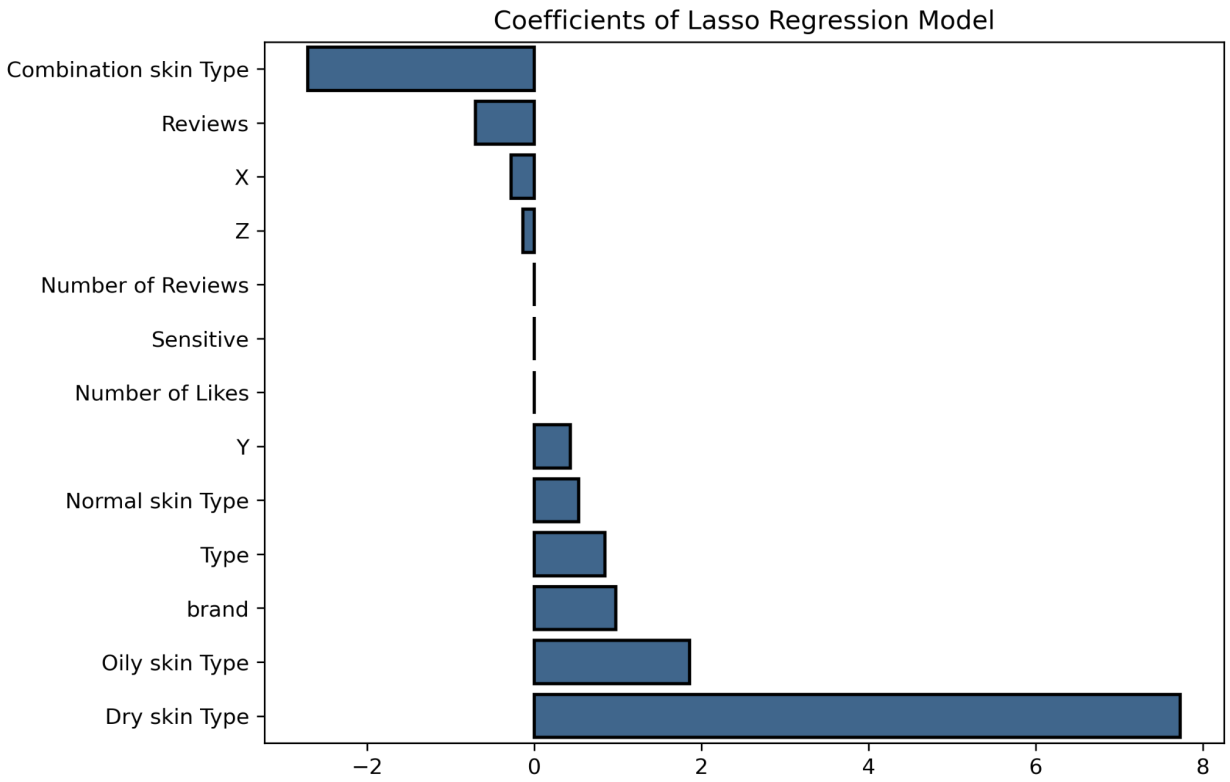
Communication

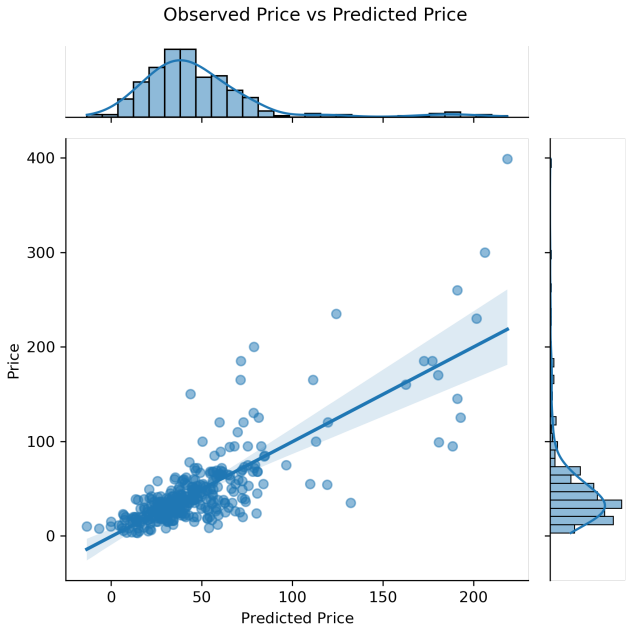
Train/Validation/Test	R^2	Mean Absolute Error	Cross Validation/Test	R^2
Linear Regression	0.6362	16.3508	Linear Regression	0.6024
Ridge Regression	0.6362	16.3413	Ridge Regression	0.6030
Lasso Regularization	0.6363	16.3398	Lasso Regularization	0.6024
Polynomial Regression	0.6362	16.2660	Polynomial Regression	0.6020

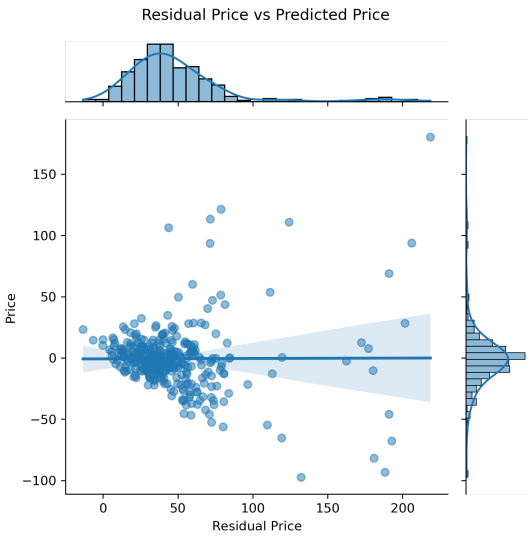

Boxplot grouped by Reviews
Price
Price
400
300
200
100
0
1.5
2.0
2.5
3.0
3.1
3.5
4.0
4.5
5.0
Reviews


Number Of Likes
300000
250000
200000
150000
100000
50000
0
NARS
The Ordinary
Urban Decay
Tatcha
ILIA
OLEHENRIKSEN
Fenty Beauty by Rihanna
Glow Recipe
Rare Beauty by Selena Gomez
Tower 28 Beauty
Brand


60000
50000
40000
30000
20000
10000
0
Number Of Likes
cleanser
collagen
cushion
face cream
face masks
face mist
face oils
serum
sunscreen
toner
Product Type


Coefficients of Lasso Regression Model
Combination skin Type
Reviews
X
Z
Number of Reviews
Sensitive
Number of Likes
Y
Normal skin Type
Type
brand
Oily skin Type
Dry skin Type
−2
0
2
4
6
8


Observed Price vs Predicted Price
Price
Predicted Price
0
100
200
300
400
0
50
100
150
200


Residual Price vs Predicted Price
Price
Residual Price
−100
−50
0
50
100
150
0
50
100
150
200