Chongkyung Kim, Lyla Kiratiwudhikul, Mina Lee, Tom Zhang
AC209

# Milestone 2

**Project Question:**

Nike and Adidas are two leading sportswear and athleisure companies in the world. While keeping excellence in products over years, both companies have extended their markets by adding exclusive lines like Jordans, Air Max, NMD, Yeezy, and more. Even though both companies consider current market conditions in setting its price points and price ranges of shoes, customers are unaware if the companies utilize the value-based pricing strategy which considers the consumers' willingness to pay or premium pricing strategy which is based on quality and value of products. Many factors can alter customers' willingness to pay for the company's products, but our analysis will focus on the displays of footwear products on the companies' websites such as wording, colors, and materials. From this analysis, we want to extend our market research practice and ultimately predict the price of a shoe using the name of products, number of styles available, materials as well as other indicators.

**Data Description:**

**I. Data variables**

Our footwear data is acquired directly from the Nike and Adidas websites. The following information are gathered:

- Product title and subtitle/category
- Product special label (e.g. *Best Seller*, *Sustainable Materials*, *Member Access*, etc.)
- Direct URL link to the product page
- Prices (original and discounted)
- Product description
- Color choices / number of color choices
- Number of reviews
- Average rating for the product based on the reviews
- Other details such as materials made

The total numbers of Nike and Adidas footwear products obtained are 1,368 and 2,390, respectively.

**II. Data Cleaning:**

```
label           959
title             0
subtitle          1
num_colors        0
price             2
reduced_price   789
url               0
description      28
colors          528
n_reviews        97
avg_stars        97
dtype: int64
```

Nike data:
There are 2 missing values for price;
1 missing value for subtitle;
789 missing values for reduced price;
28 missing values for description;
528 missing values for colors; and
97 missing values for both n_reviews and avg_stars.

Adidas data:
There are 1 missing value for title;
30 missing values for subtitle;
4 missing values for original price;

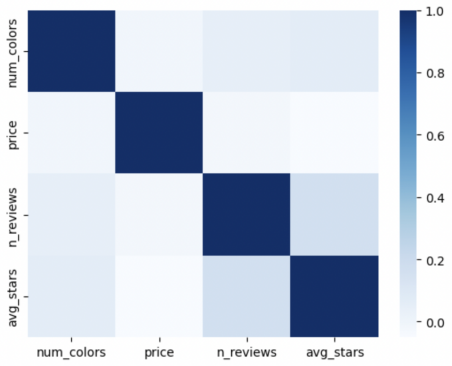
144 missing values for reduced price;
22 missing values for description;
142 missing values for details;
445 missing values for colors; and
210 missing values for both n_reviews and avg_stars.

Reduced prices could be missing if the products are not on sale.
Reviews and Stars could be missing if the products are not very popular that a review has been given for them.
Both description, details, and subtitles can be treated as accessory elements.
However, the missingness in *price* is due to data collection error. Therefore, we have imputed the value with median at the moment.

**Data Analysis:**

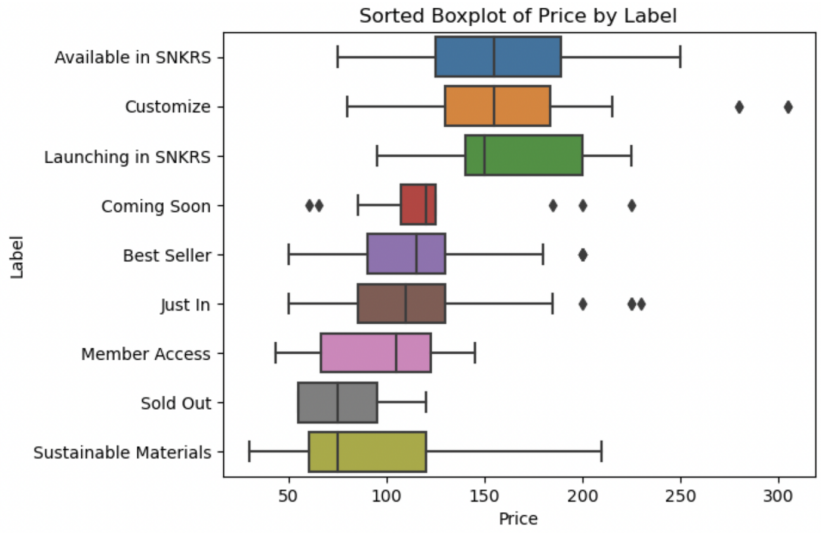
For our dataset, Nike has 1,368 products with prices ranging from $25.00 to $350.00. The average price of products is $108.34. However, in order to consider the primary choice of a product's price, we will focus on the regular price of a product and not consider the sales price. On average, each product has 3.53 style choices.

Focusing on the numerical variables, number of colors seem to be a better predictor to predict price compared to other two indicators. However, even though the correlation between numerical variables are relatively low, as we extend the research, we will attempt to categorize the products and compare.


Succeeding attempts are primarily focused on the wordings of products. Labels with “Available in SNKRS” and “Customize" have relatively higher price ranges than other labels. This is reasonable since the launchment on

SNKRS app is usually based on the products with higher demand which would align with Nike's value-based price decision tactics.




Further, we also performed preliminary text analysis on the data. The variables of interest are the two main text columns: title and description. Since there is much overlap between the two, we will be investigating “description” at the moment, which contains more text information.

We adopted the standard preprocessing approach for converting raw text to bag-of-words representation. The process involved: lowercasing, removing punctuation, tokenization, removing stop words, calculating word counts, and filtering out low-occurrence words. The resulting data frame should then have columns as the vocabulary, and their occurrences are recorded for each row. The processed data in this format enables us to fit a Naive Bayes model as it works well when the feature set is much larger than observations. Because Naive Bayes is a better classifier than regressor, we will predict on the target “price” grouped into five brackets based on percentiles. The resulting model has a test accuracy of about 72%.

By extracting relevant attributes from the fitted scikit-learn Multinomial Naive Bayes classifier, we are able to extrapolate feature importance within each class - that is, the most predictive words from “description” for each price bracket. The following data frame illustrates our result, and the output is quite intuitive. In the lowest price bracket, we see words such as “foam”, and in the second highest bracket, we see “elite”. These instances might give us some idea of the unique characteristics for Nike shoes within each price bracket.

| | (24.675, 90.0] | (90.0, 155.0] | (155.0, 220.0] | (220.0, 285.0] | (285.0, 350.0] |
|---|---|---|---|---|---|
| **0** | nike | air | air | air | air |
| **1** | design | nike | nike | nike | cleat |
| **2** | air | design | design | features | nike |
| **3** | little | comfort | max | phantom | zoom |
| **4** | comfort | cushioning | cushioning | design | like |
| **5** | foam | classic | feel | elite | soccer |
| **6** | classic | look | comfort | zoom | made |
| **7** | cushioning | upper | style | next | control |
| **8** | durable | feel | upper | provides | nods |
| **9** | made | max | look | control | world |

**Baseline Model:**

| Predictors | MSE |
|---|---|
| ['num_colors'] | 1974.42 |
| ['n_reviews'] | 1967.54 |
| ['avg_stars'] | 1976.13 |
| ['num_colors', 'n_reviews'] | 1968.04 |
| ['avg_stars', 'n_reviews'] | 1970.56 |
| ['num_colors', 'avg_stars'] | 1976.64 |
| ['num_colors', 'n_reviews', 'avg_stars'] | 1971.13 |

| Predictors | R2 score |
|---|---|
| ['num_colors'] | -0.01 |
| ['n_reviews'] | -0.0 |
| ['avg_stars'] | -0.01 |
| ['num_colors', 'n_reviews'] | -0.0 |
| ['avg_stars', 'n_reviews'] | -0.0 |
| ['num_colors', 'avg_stars'] | -0.01 |
| ['num_colors', 'n_reviews', 'avg_stars'] | -0.0 |

Since the main goal of the project is to predict the shoe price, our first choice model that we start with is the linear regression model. We started off with variables “number of colors”, “number of reviews”, and “average star ratings”. As of now, none of these variables or the combination of these variables seem to show a linear relationship with the shoe prices. To that end, we will have to explore ways to incorporate other variables into our model to find a linear relationship, if at all, between the available variables and and the shoe prices.

**Reference:**


Entreprene. (2021, December 19). *Pricing strategies of nike*. The Strategy Watch. Retrieved November 27, 2022, from https://www.thestrategywatch.com/pricing-strategies-of-nike/

Wondershare EdrawMind. *Adidas Marketing Mix (4PS) Analysis*. Retrieved December 2, 2022, from https://www.edrawmind.com/article/adidas-marketing-mix-analysis.html

Gema Mateos. *Nike's Pricing strategy within the extended marketing mix*. Medium. Retrieved December 2, 2022, from https://medium.com/@gemamateos.gmc/nikes-pricing-strategy-withing-the-extended-marketing-mix-52e644f67954