Chongkyung Kim, Lyla Kiratiwudhikul, Mina Lee, Tom Zhang
AC209a

The Effects of Footwear Product Displays on Prices

Introduction

Nike and Adidas are two leading sportswear and athleisure companies in the world. While keeping excellence in products over years, both companies have expanded their markets through diverse advertising tactics such as adding new color schemes to Nike's Dunk Low and adding exclusive lines like Jordans, Air Max, NMD, Yeezy, and more. Both companies would take current market conditions into account when setting its product price points and ranges, but companies also would utilize the value-based pricing strategy which incorporates the consumers' willingness to pay (Entreprene, 2021). While many factors will play a role in altering the consumers' willingness to pay, we can observe companies' final analysis on those through the final price decision of each product along with the descriptions to persuade consumers on their websites. Therefore, using the premise that each company has decided the price and display of each product to target consumer's needs and willingness to purchase, it would be not surprising to investigate the relationship between the price and displays of footwear products on the companies' websites such as wording, colors, and styles. Therefore, our analysis will focus on finding features that are predictive of product prices. The results can potentially be used to predict the price of a footwear product given its description, number of styles available, type of shoes as well as other indicators.

Data Description

I. Data variables

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

- Product title and subtitle
- 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 product details such as materials made and product code

The total numbers of Nike and Adidas footwear products originally obtained are 1,368 and 2,390, respectively. The data include all men's, women's, kids' and infants' footwear products as well as their information that were available on the official websites at the time of data collection.

II. Data Cleaning:

There were some mismatches and missing entries in the raw data that were extracted from the original websites. Below briefly describes the cleaning process of the Nike and Adidas datasets.

There were originally 1,368 Nike products in the raw dataset. However, a few of them were gift card items instead of footwear products, and hence they were dropped from the final dataset. In addition, for products that have only one color choice, the *color* column entries were missing and needed to be filled in with the information in the description box of the product page.[1]

The Adidas dataset had 2,390 observations, but a substantial proportion of them had missing entries for the title, subtitle, number of colors, prices, and reduced prices, which were essential for analysis. This was due to the fact that Adidas html structures were not consistent throughout the website. Therefore, the missing information was re-collected from their individual product pages.

```
Int64Index: 1366 entries, 0 to 1367
Data columns (total 11 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   label          408 non-null    object
 1   title          1366 non-null   object
 2   subtitle       1366 non-null   object
 3   num_colors     1366 non-null   object
 4   price          1366 non-null   object
 5   reduced_price  579 non-null    object
 6   url            1366 non-null   object
 7   description    1338 non-null   object
 8   colors         1231 non-null   object
 9   n_reviews      1271 non-null   float64
 10  avg_stars      1271 non-null   float64
```

```
Int64Index: 2390 entries, 0 to 2389
Data columns (total 12 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   title          2389 non-null   object
 1   subtitle       2360 non-null   object
 2   num_colors     2390 non-null   object
 3   url            2390 non-null   object
 4   price          2386 non-null   object
 5   reduced_price  2246 non-null   object
 6   description    2368 non-null   object
 7   details        2248 non-null   object
 8   colors         1945 non-null   object
 9   n_reviews      2180 non-null   float64
 10  avg_stars      2180 non-null   float64
 11  product_code   2390 non-null   object
```

Figure 1: Feature Summary for Nike (left) and Adidas (right) datasets

Observe that some reduced prices are missing if the products were not on sale. In addition, there are some products that do not have reviews and average rating stars as they might not be so popular that a review had been given for them or they might be newly released.

For both Nike and Adidas, the *subtitle* features contain information that can be split into two categorical features: *category* and *purpose*, which refer to targeted consumer groups (Men's, Women's, Kid's, Infant) and types of shoes (e.g. Lifestyle, Running, etc), respectively. The values for each of the two features can be found in Table 1 under Appendix.

Data Analysis

For the price analysis, in order to consider the primary choice of a product's price, we will focus on the regular price of a product and not the sales price. Nike has 1,368 products with prices ranging from \$25.00 to \$350.00. The average price of Nike products is \$108.34. On average, each product has 3.53 styles of color choices. Compared to Nike's, Adidas' products have a wider range of prices, from \$25.00 to \$1,125.00, with a few outliers from the specialty products (e.g. the Balenciaga x Adidas Collaboration). The average price of products is also slightly higher at \$119.74. In addition, Adidas has more varieties for their customers in terms of color choices for each product, with an average of 5.69 styles.

---

[1] Note that there was a time gap between data scraping and data cleaning stages, which resulted in inconsistent numbers of observations. For example, products were sold out and were no longer available on the websites at the time of cleaning.

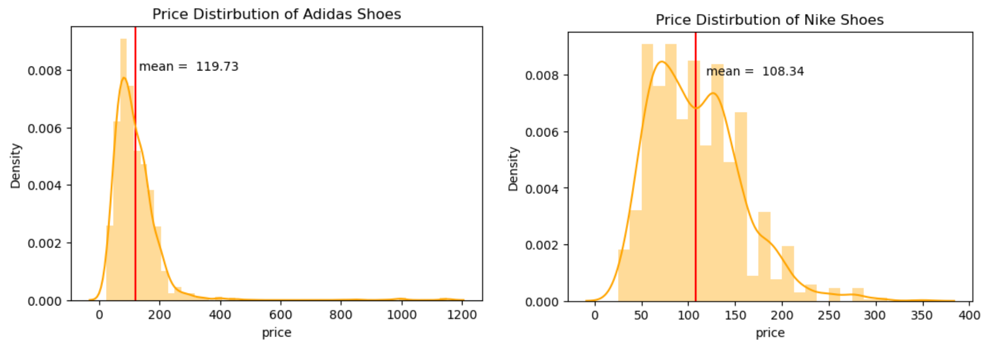


Figure 4: Price distributions of Adidas shoes (left) and of Nike shoes (right)

Focusing on the numerical variables of both Nike and Adidas, the scatterplots and the correlation maps below show no evident linear relationships between prices and the numerical features.

Figure 2: Scatter plots to visualize the the relationships between prices and numerical variables for Nike (top) and Adidas (bottom)

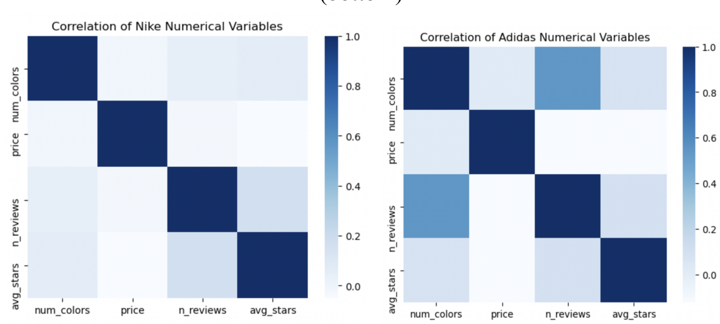


Figure 3: Correlations among price and numerical features for Nike (left) and Adidas (right)

To extend the analysis to non-numerical variables, the *subtitle* feature has been broken into two subcategories of *category* (Men's, Women's, Kid's, Infant) and *purpose* (e.g. Lifestyle, Running, etc) to categorize the products and compare. Figure 4 and 5 show that some types of footwear have higher prices than others. For instance, adults' price ranges are higher than those of kids' and infants'. In particular, men's shoes overall are slightly more expensive than women's for both Adidas and Nike. However, there is no clear evidence on the effect of product purposes on prices. One interesting observation is that products that are not designed specifically to a particular type of sports (i.e. those labeled as *lifestyle* or *sportswear*) tend to have the widest ranges of prices.

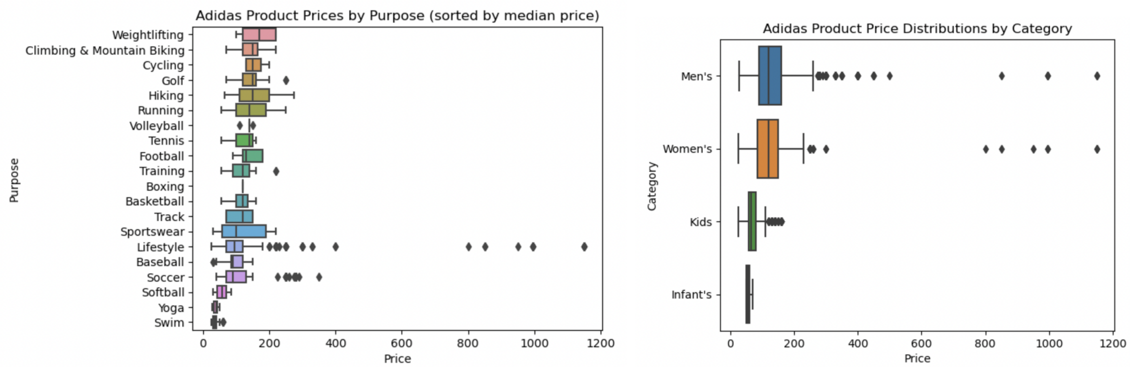


Figure 4: Adidas product prices by purpose (left) and by category (right)

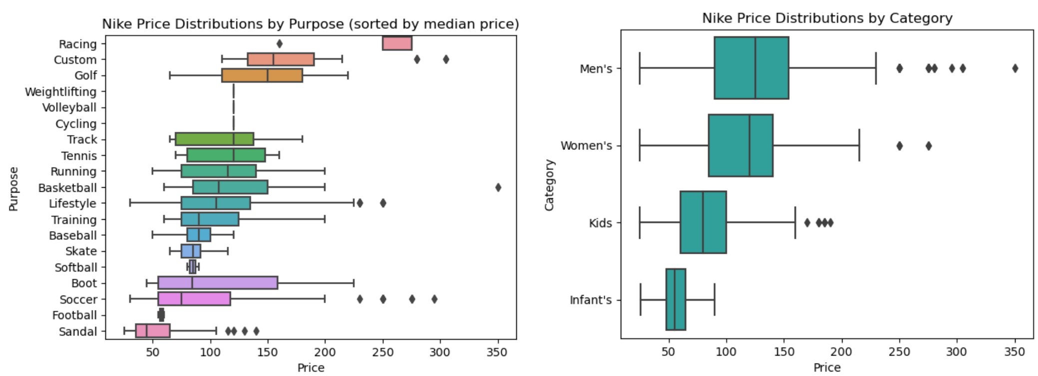


Figure 5: Nike product prices by purpose (left) and by category (right)

Next, we performed sentiment analysis on the data to contextually mine the text to identify and extract subjective descriptions of products. In this approach, the variable of interest is a text combination of *title, category, purpose,* and *description*. We adopted a normal classification approach to the variable description to analyze whether the description is positive, neutral, and negative. The processed data expectedly showed a bigger cluster in the positive score since the description itself is provided by the company.

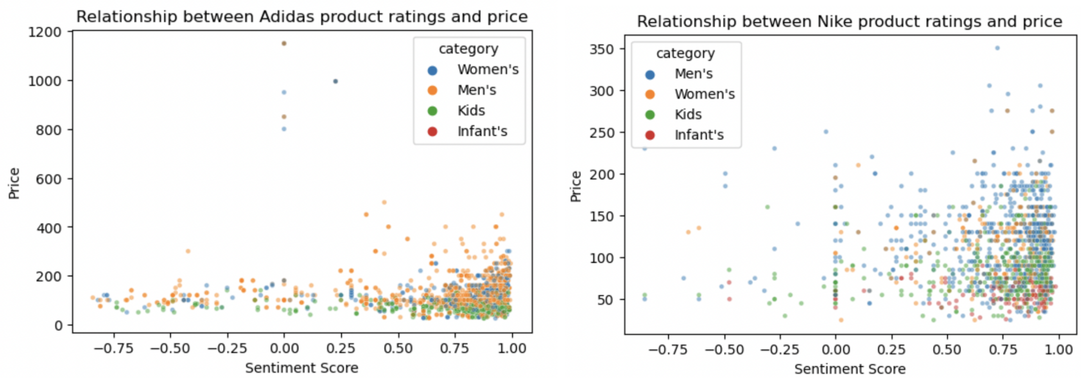


Figure 6: Sentiment analysis distribution for each category of Adidas (left) and of Nike (right)

Further, we also performed preliminary text analysis on the data. We adopted the standard preprocessing approach for converting the raw text combination mentioned above to bag-of-words (BOW) representation. The process involved: lowercasing, removing punctuation, tokenization, removing stop words, calculating word counts, and filtering out low-occurrence words. The resulting BOW representations then have columns as the vocabulary, and their occurrences are recorded for each row/product. 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 performs better in a classification setting than regression (Frank et al., 2000), we predict the target “price” that is grouped into five brackets. The resulting models have a test accuracy of about 77% and 31% for Adidas and Nike, respectively. Observe that the Nike models perform extremely poor on the BOW representations. We will discuss the poor model performance of Nike’s in detail later.

By extracting relevant attributes from the fitted Multinomial Naive Bayes classifier, we are able to extrapolate feature importances within each price bracket - that is, the most predictive words from the text combinations. The following table on the left of Figure 6 illustrates our result for Nike, 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”. From the Adidas table located on the right, we can see the words “boost” in the second highest bracket. 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

	(24.825, 68.75]	(68.75, 112.5]	(112.5, 156.25]	(156.25, 200.0]	above_200
0	product	product	upper	upper	product
1	upper	upper	product	product	upper
2	recycled	shoes	shoes	shoes	adidas
3	shoes	recycled	outsole	50	shoes
4	outsole	outsole	recycled	midsole	outsole
5	content	content	50	adidas	midsole
6	adidas	color	midsole	recycled	rubber
7	color	adidas	adidas	outsole	color
8	50	code	color	boost	imported
9	code	imported	content	rubber	code

Figure 7: Top 10 most predictive features for each price bracket in Nike (left) and Adidas (right)

In addition, we also calculated TF-IDF for all words on each price bracket, which measures how relevant each word is to a specific bracket. In our case, price brackets are equivalent to documents. Mathematically, it is the product of “term frequency” (of a word in a bracket) and “inverse document frequency” (that is, how rare a word is relative to all brackets). Hence, words with high TF-IDF scores are the most relevant words unique to a specific bracket. The following output illustrates the top 10 keywords in the lowest and highest price brackets for Nike.

```
guo         0.052319
ishod       0.042210
utility     0.042210
wair        0.042210
effort      0.042210
kidright    0.042210
gen         0.042210
lane        0.042210
burrow      0.042210
vapormax    0.042210
Name: (24.845, 63.75], dtype: float64
```

```
describe      0.04939
notice        0.04939
tailwind      0.04939
towards       0.04939
pays          0.04939
king          0.04939
embroidery    0.04939
beginnings    0.04939
national      0.04939
vapormax      0.04939
Name: above_180, dtype: float64
```

Figure 8: Top 10 keywords based on TF-IDF in the lowest and highest price brackets for Nike

Similarly, these are the keywords for Adidas:

```
tnd            0.040213
scurry         0.040213
vulcraid3r     0.040213
locker         0.040213
pool           0.040213
poolside       0.040213
seem           0.040213
children       0.040213
postworkout    0.040213
channel        0.040213
Name: (24.825, 68.75], dtype: float64
```

```
highcut         0.052141
dusk            0.052141
balenciaga      0.052141
qasa            0.052141
triplelayer     0.052141
carbitex        0.052141
gliding         0.052141
remaking        0.052141
outdoorboost    0.052141
ajatu           0.052141
Name: above_200, dtype: float64
```

Figure 9: Top 10 keywords based on TF-IDF in the lowest and highest price brackets for Adidas

Finally, we can incorporate these results into the modeling process. Specifically, the TF-IDF representation matrix is concatenated with the original features as extra columns and input into the models in the next section.

Modeling

I. Classification Models

First, we performed our analysis in the classification setting. Product prices of each brand are segmented into five brackets as follows:[2]

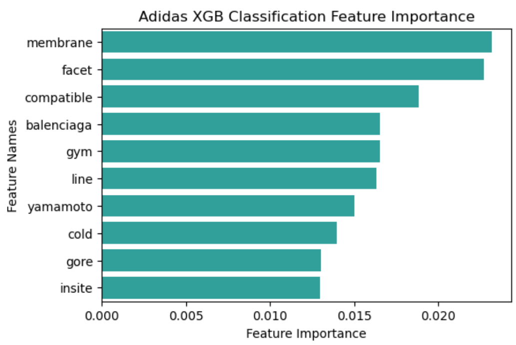
> Adidas: ($24.825, $68.75], ($68.75, $112.5], ($112.5, $156.25], ($156.25, $200.0], and above $200.
> Nike: ($24.845, $63.75], ($63.75, $102.5], ($102.5, $141.25], ($141.25, $180.0], and above $180.

We trained XGBoost and LightGBM classifiers with the price brackets as the target variable and investigated features or words that yield high predictive power based on the Mean Decrease in Impurity

[2] The highest price bracket for each brand is chosen manually so that it contains about 6-7% of products. The rest of the data are then discretized into four equal-width price brackets.

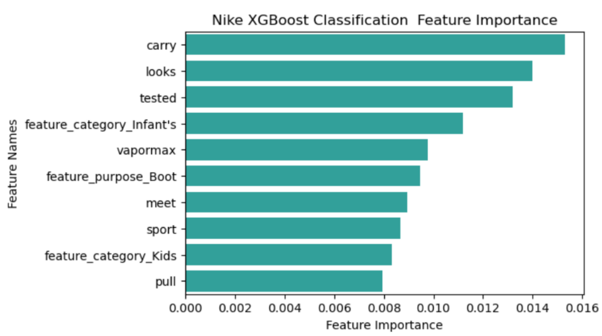
(MDI) measure.[3] The results are shown in Figure 10-13. Both XGBoost and LightGBM perform well in classifying the price brackets for Adidas products, with F-1 scores above 87% across all target classes. The top ten features based on their MDI-based importances suggest that Adidas' product prices are driven by the types of footwear (e.g. "facet", "gym"), high-end brand collaborations (such as "Balenciaga" and "Yamamoto"), and the materials (e.g. "membrane" and "synthetic").

However, the results for Nike products are less impressive with low F1-scores for most classes. One hypothesis for Adidas' better performance is that Adidas has certain product lines that are priced in unique, small ranges, making it easier to distinguish products and their corresponding prices into distinct groups. In addition, Nike data contains not only a smaller number of products but also relatively less information about the products themselves. Their descriptions are rather non-descriptive and do not contain much detail on the materials or technology used for the design.



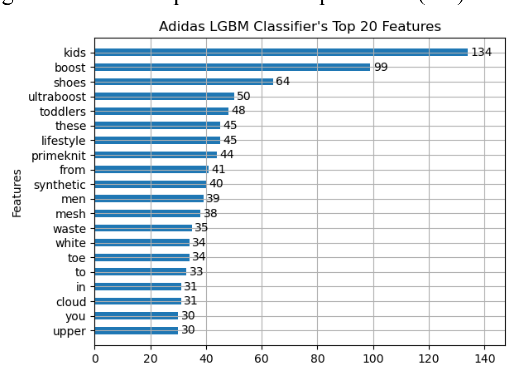
	precision	recall	f1-score	support
(24.825, 68.75]	0.89	0.90	0.90	156
(68.75, 112.5]	0.87	0.91	0.89	311
(112.5, 156.25]	0.83	0.84	0.84	171
(156.25, 200.0]	0.92	0.83	0.87	105
above_200	0.89	0.76	0.82	45
accuracy			0.87	788
macro avg	0.88	0.85	0.86	788
weighted avg	0.87	0.87	0.87	788

Figure 10: Adidas' top 10 feature importances (left) and performance metrics (right) from the XGBoost classifier



	precision	recall	f1-score	support
(24.845, 63.75]	0.71	0.55	0.62	85
(63.75, 102.5]	0.45	0.53	0.49	144
(102.5, 141.25]	0.41	0.57	0.48	129
(141.25, 180.0]	0.14	0.05	0.08	56
above_180	0.14	0.05	0.08	37
accuracy			0.45	451
macro avg	0.37	0.35	0.35	451
weighted avg	0.42	0.45	0.43	451

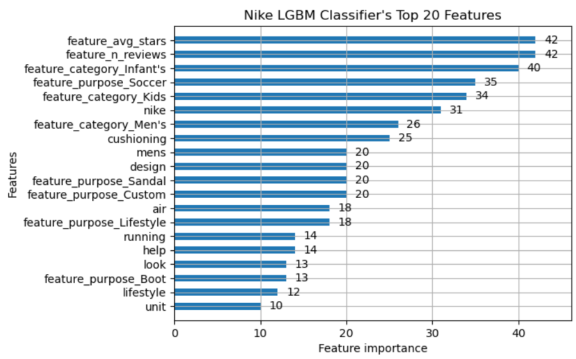
Figure 11: Nike's top 10 feature importances (left) and performance metrics (right) from the XGBoost classifier



	precision	recall	f1-score	support
0	0.92	0.91	0.92	156
1	0.88	0.91	0.89	311
2	0.82	0.86	0.84	171
3	0.92	0.85	0.88	105
4	0.97	0.76	0.85	45
accuracy			0.88	788
macro avg	0.90	0.86	0.88	788
weighted avg	0.88	0.88	0.88	788

Figure 12: Adidas' top 20 feature importances (left) and performance metrics (right) from the LightGBM classifier

[3] Note that permutation feature importance results are not available due to computational resource constraints.



	precision	recall	f1-score	support
0	0.73	0.62	0.67	85
1	0.46	0.61	0.52	144
2	0.44	0.56	0.49	129
3	0.23	0.05	0.09	56
4	0.29	0.05	0.09	37
accuracy			0.48	451
macro avg	0.43	0.38	0.37	451
weighted avg	0.46	0.48	0.45	451

Figure 13: Nike's top 20 feature importances (left) and performance metrics (right) from the LightGBM classifier

II. Regression Analysis

For the regression task, we instead used our original price data as the target variable and fit XBGboost and LightGBM regressors. The root mean squared errors (RMSE) on the test set for Adidas' XGBoost and LightGBM are at $101.16 and $100.59, respectively. The prediction plots on the test set shown in Figure 14 and 16 suggest that the models struggle to predict outlying prices of high-end products such as the Y-3 and Balenciaga Collections. The results from feature importance analysis are aligned with those from the classification task: Adidas prices are driven by footwear types, materials, and special product lines.

Despite the poor performance under the classification task, Nike's regression models are able to predict prices with test RMSE of approximately $37-$42. According to the feature importance plots, the models seem to pick up the categories whether they are men's, women's or children's as well as whether the products are sandals or custom. The evidence here confirms our hypothesis proposed in the previous section that Nike's textual displays of products are not descriptive and strongly predictive of prices.

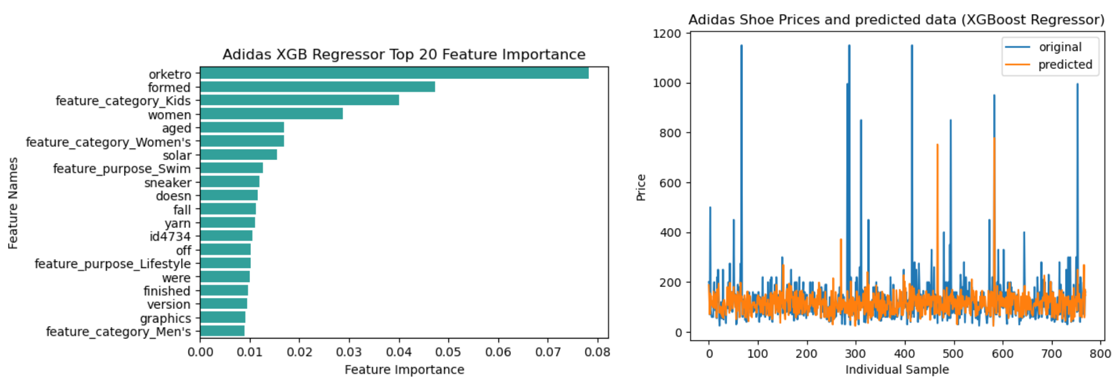


Figure 14: Adidas XGBRegressor feature importance (left) and regression plot (right)

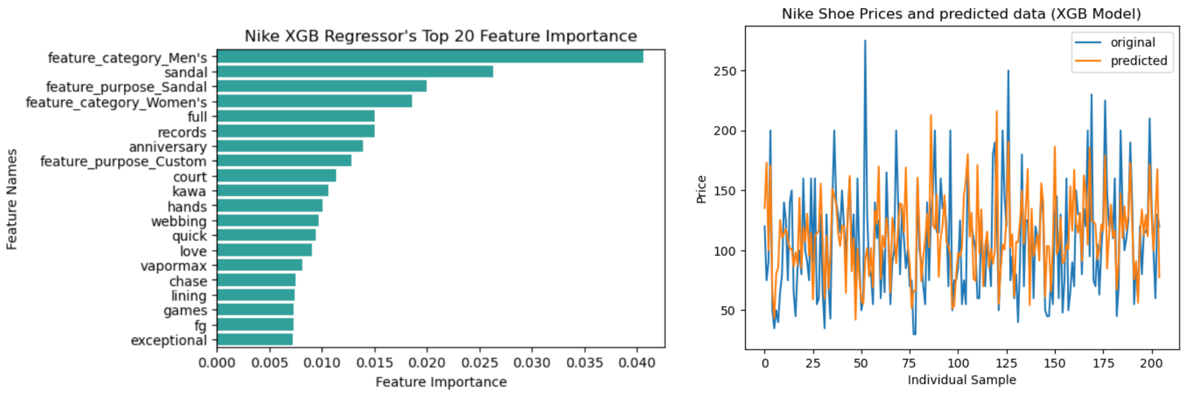


Figure 15: Nike XGBRegressor feature importance (left) and regression plot (right)

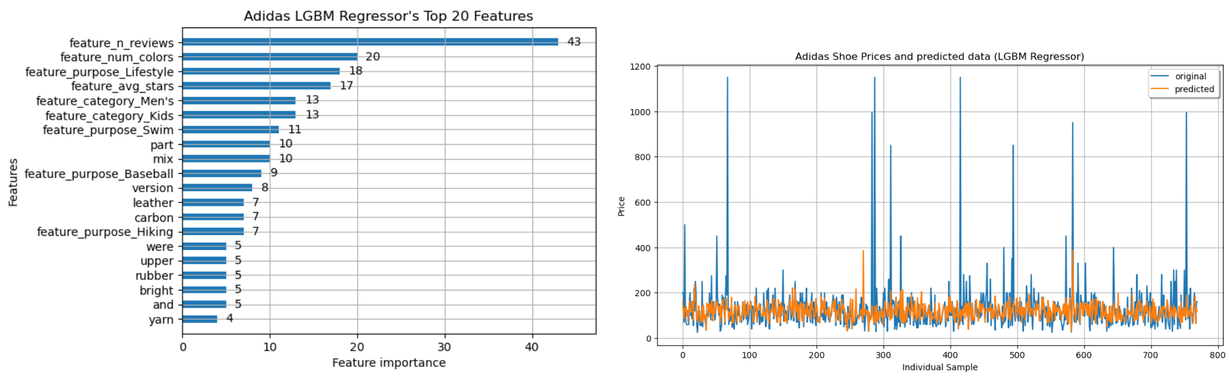


Figure 16: Adidas LGBMRegressor feature importance (left) and regression plot (right)

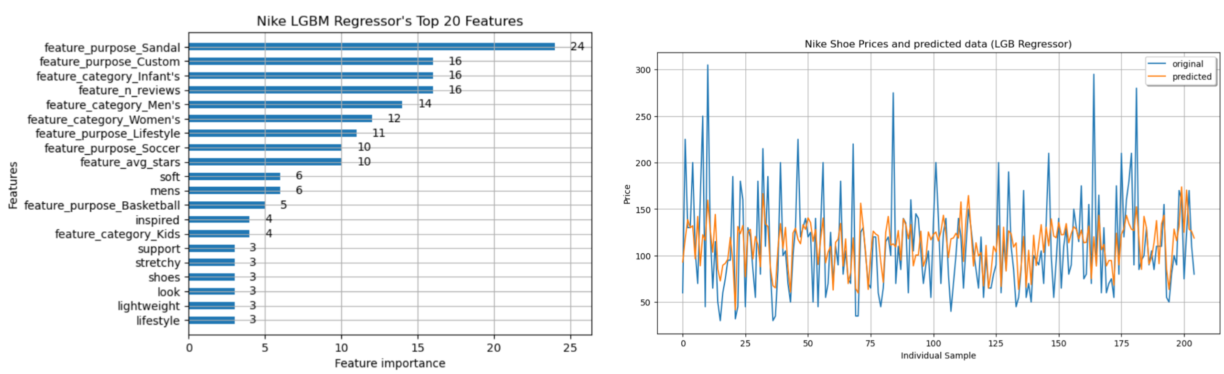


Figure 17: Nike LGBMRegressor feature importance (left) and regression plot (right)

Conclusion

The purpose of this analysis was to find the classification and regression models' top feature importances[4] on Nike and Adidas dataset and ultimately find the relationship between the price and displays of footwear products on the companies' websites such as wording, colors, and styles. With XGBoost and LightGBM models, we may conclude that Adidas' product prices are driven by the types of footwear, technology, high-end collaboration lines, and the materials. However, Nike's textual displays of products are not predictive of prices because descriptions of Nike products have limited information compared to Adidas. Therefore, it is not a good practice to predict price ranges of Nike products with textual displays.

[4] Feature importance refers to an indication of the relative importance of each feature when making a prediction

Nevertheless, there is still room for improvement. The current study has been limited to finding price predictive features from product's display on websites. If time permits, we can actually perform price optimization with information given by the price predictive features and we can also extend the study to be on the customer perspective by predicting prices using the sentiment analysis of customer's review for future research.

References

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

Frank, E., Trigg, L., Holmes, G. et al. Technical Note: Naive Bayes for Regression. Machine Learning 41, 5–25 (2000). https://doi.org/10.1023/A:1007670802811

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

Gupta, S. (2018, January 19). *Sentiment analysis: Concept, analysis and applications*. Medium. Retrieved December 9, 2022, from https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17

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

Appendix

Table 1: Values of *category* and *purpose* features for Adidas and Nike

Adidas' and Nike's Categories	Adidas' *purpose* values	Nike's *purpose* values
• Men's • Women's • Kids' • Infants'	• Lifestyle • Running • Soccer • Basketball • Sandal • Custom • Training • Boot • Golf • Track • Tennis • Baseball • Cycling • Softball • Weightlifting • Volleyball • Sportswear • Yoga • Hiking • Swim • Climbing & Mountain Biking • Boxing	• Lifestyle • Running • Soccer • Basketball • Sandal • Custom • Training • Boot • Golf • Skate • Track • Tennis • Baseball • Racing • Cycling • Football • Softball • Weightlifting • Volleyball