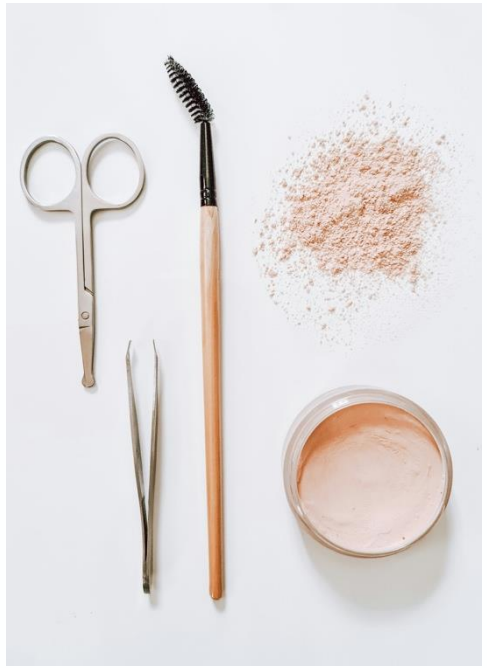
The Link Between Customer Reviews and Beauty Awards: An Analysis

Rachel Lewis





Background Information

Customer Reviews

- Defined as feedback about a company or product written by a customer on a commercial website; used to help inform purchases of other potential customers.[1]
- A 2023 survey of over 26,000 US consumers found that **88% of consumers consider ratings and reviews** when making a beauty purchase decision. [2]

Allure Best of Beauty Awards

- Allure Magazine is the leading publication in beauty journalism. [3]
- Since 1996, Allure Magazine has released an annual list of the top products in select beauty categories. [3]
- Winning an Allure Best of Beauty award can increase brand visibility, credibility, and sales. According to a report by The NPD Group, the sales of the Fenty Beauty Pro Filt'r Soft Matte Longwear Foundation **increased by 132% in the week following its win of the Allure Best of Beauty Award** in 2018. [4]

Sephora

- Sephora is a leading multinational retailer of personal care and beauty products, with over 500 stores in the United States. [5]
- The Cosmetify 2023 Index found that Sephora had the highest organic search traffic of beauty retailers at 16,100,000 and the highest number of Instagram followers at 22.1 million. [6]

Allure Magazine 2024 Best of Beauty Awards

Research Overview

This study compares Allure's Best of Beauty award winners for 2024 with customer sentiment, using Sephora reviews to address three key research questions:

Sentiment Comparison

How does customer sentiment differ between Best of Beauty award winners and non-award winners?

Predictive Analysis

How accurately can we predict which products will be Best of Beauty winners based on customer ratings and sentiment scores?

Keyword Analysis

Are there specific keywords in customer reviews that distinguish Best of Beauty winners from non-winners, and can these keywords help predict future award winners?

Data Sets

Allure Best of Beauty 2024 Winners at Sephora

- Collected **manually** by searching for all 357 Allure Best of Beauty Winners on Sephora's website and manually gathering the Product IDs from the product URL.
- Contains Category, Brand, Product Name, and Product ID for 85 Allure Best of Beauty 2024 award winners sold at Sephora.

Sephora Product IDs

- Collected from Sephora's website using a Python script to collect data from this API: https://www.sephora.com/api/v3/users/profiles/current/product/
- The Python script looped numerically in increments of 1, passing each number to the API as a possible Product ID.
- If the API returned a valid product name for the ID, then the script stored the Product ID in a Snowflake table. If not, the loop continued to the following number.
- Contains Product Name, Brand Name, and Product ID for over 7,700 products sold at Sephora in the same categories as the Allure Best of Beauty award winners.

Sephora Customer Review Data

- Used a Python script to pass the Product IDs stored in the other two tables to this API: https://api.bazaarvoice.com/data/reviews.json
- The script captured all relevant customer review data and stored it in a Snowflake table for further analysis.
- Contains almost 780,000 records of customer review data, including Product ID, Review Test, Rating (1-5), Is Recommended (Y/N), Submission Time, Skin Tone, Eye Color, Skin Type, Hair Color

Data Set Samples

Allure Best of Beauty 2024 Winners at Sephora

Category	Brand	Product Name	Product ID
Hair	K18	Damage Shield pH Protective Shampoo	P509691
Body	Kate McLeod	The Pebble Solid Bath & Shower Oil	P508743
Skin	Kate Somerville	KateCeuticals SuperCell Rejuvenation Serum	P510352
Scent	Maison Margiela Fragrances	Replica From the Garden Eau de Toilette	P507949

Sephora Product IDs

Product ID	Brand Name	Product Name
P481969	DIOR	Dior Addict Shine Lipstick
P202633	Anastasia Beverly Hills	Brow Wiz® Ultra-Slim Precision Eyebrow Pencil
P387589	Too Faced	Hangover Replenishing Face Primer

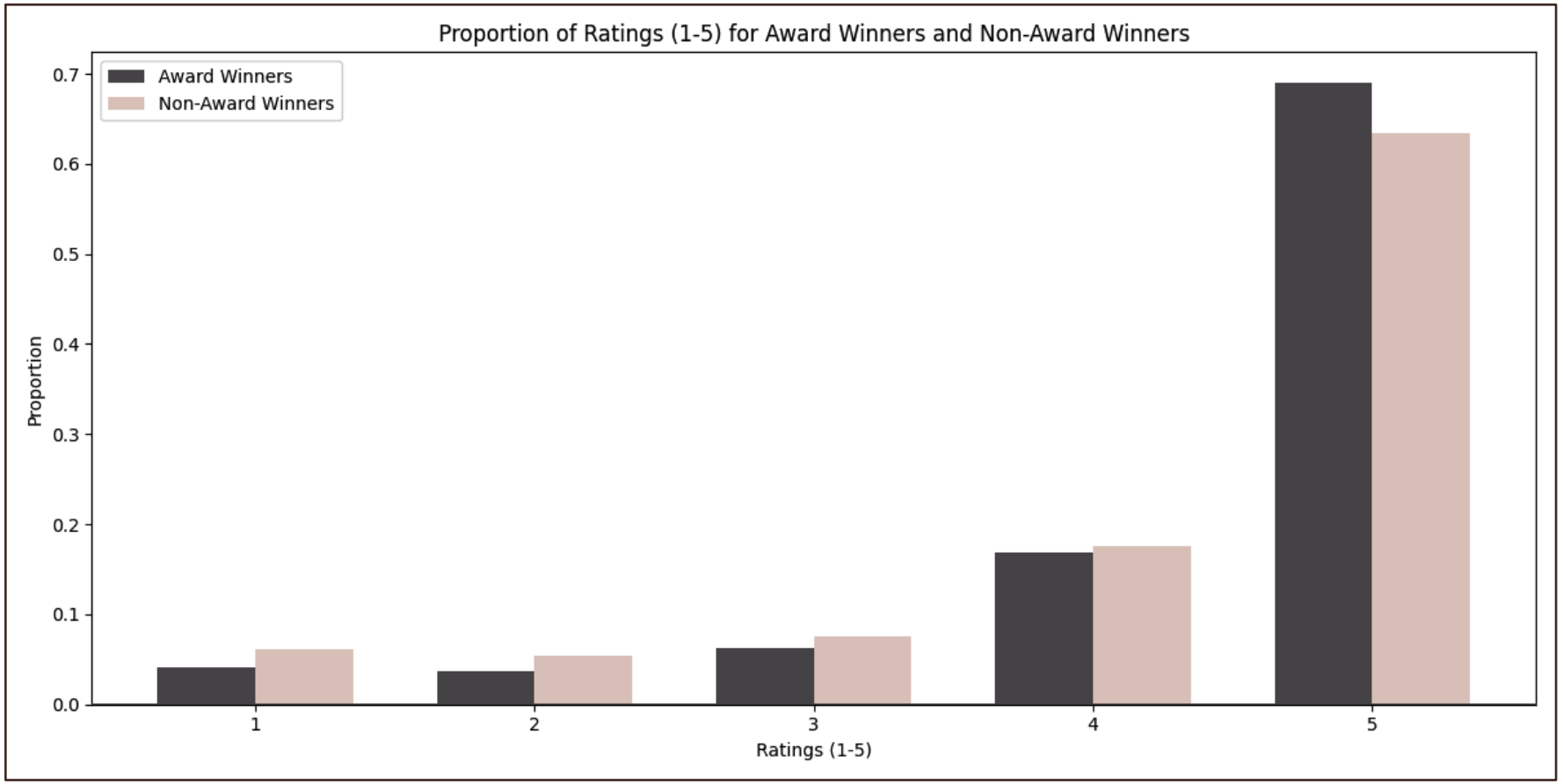
Data Set Samples

Sephora Customer Review Data

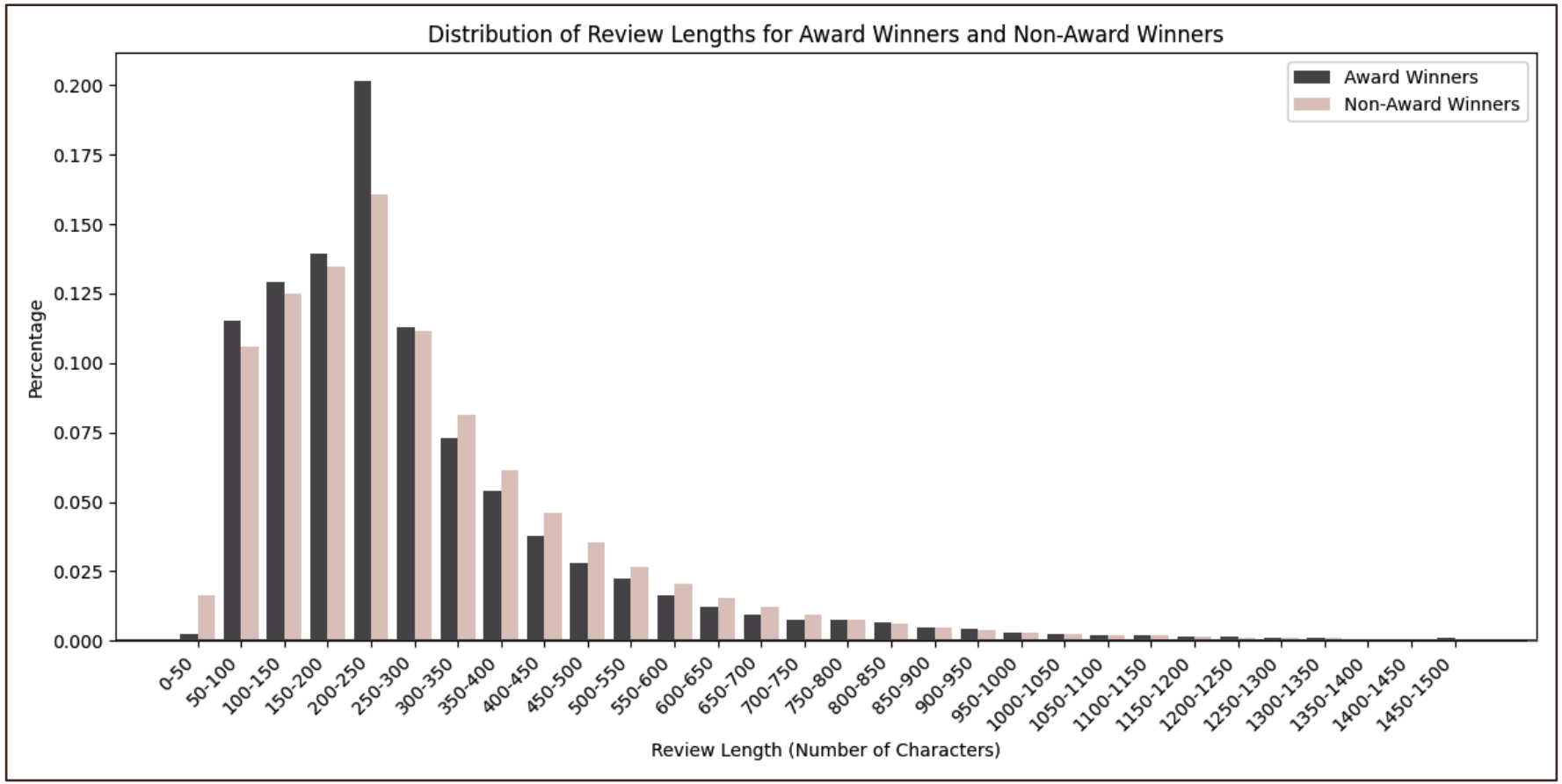
Product ID	Review Text	Rating	Is Recommended (T/F)	Submission Time	Skin Tone	Eye Color	Skin Tyoe	Hair Color
P421998	The store gave me the sample and I used it. I fall in love with it. My skin is very sensitive, but this oil works great on my skin.	5	TRUE	2024-11-03 01:32:43.000	medium	hazel	combination	black
P506273	It's a rich complex smell but doesn't last at all. Also, it's not fresh. It has kind of a chemical smell, like a room spray.	3	FALSE	2024-06-19 04:28:42.000	light	brown	combination	
P506273	I love the smell of this stuff and I really enjoy a different take on a clean scent. But the fragrance had an actual hair IN it. So gross	1	FALSE	2024-07-02 04:48:45.000	fairLight	brown	normal	brown

Exploratory Data Analysis

- Comparison of Customer Ratings for Winners and Non-Winners
- Comparison of Review Text Length
- Relationship Between Review Length and Rating

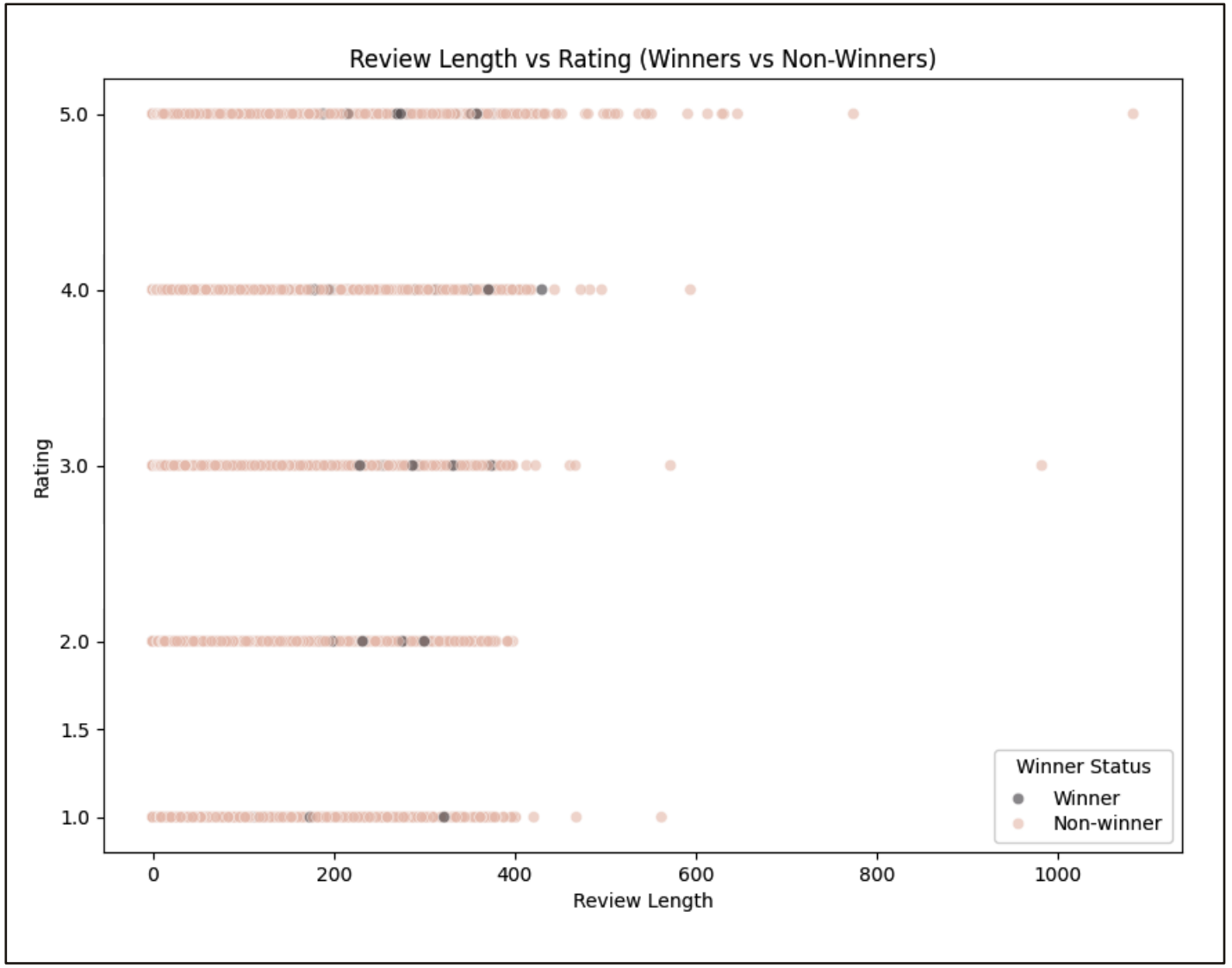


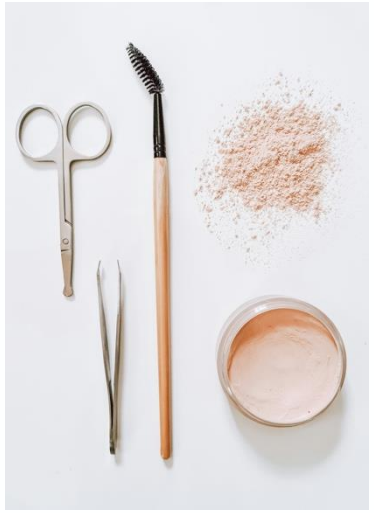
A comparison of customer ratings (1-5), where 1 is the lowest score and 5 is the highest, shows that products that won an Allure Best of Beauty 2024 award have a slightly higher proportion of 5-star reviews and a lower proportion of 1, 2, and 3-star reviews than non-winners.

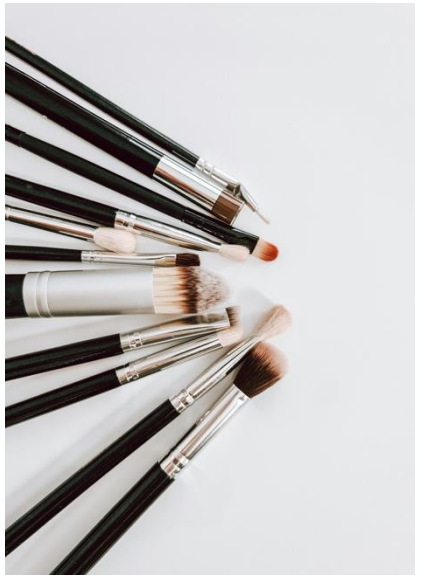


A comparison of the lengths of customer reviews, as calculated by character count, shows that the award winners have a higher percentage of reviews with 50-300 characters, with non-winners having a slightly higher percentage with more than 300 characters.

A comparison of review length and ratings shows that five and four-star ratings have slightly longer reviews, and products that won an Allure Best of Beauty award also tend to have longer reviews than non-winners.







TF-IDF

Using term frequency-inverse document frequency to quantify the importance of terms in Sephora customer review data.

TF-IDF

What:
Term Frequency-Inverse Document Frequency (TF-IDF) is a method for measuring the importance of a word in a document based on its frequency in that document (TF) and its rarity across a collection of documents (IDF). [9]

Why:
TF-IDF highlights important, distinctive terms while downplaying common words, which can improve search relevance, content analysis, and text classification. [9]

When:
TF-IDF is commonly used to analyze and rank textual data in document categorization, information retrieval, and keyword extraction. [9]

Vectorization:
Machine learning algorithms work with numerical data, so text data must be converted into numbers through vectorization. TF-IDF vectorization calculates the importance of each word in a document and converts that information into a numerical vector. [9]

Scikit-learn:
For this project, I used the TfidfVectorizer class from the scikit-learn library to convert customer review text into a matrix of TF-IDF features. By calculating the TF-IDF scores for the customer review text, I was able to represent text data in a numerical form that machine learning models can work with. [10]

TF-IDF Features

FEATURE_NAME
mattifying
balance
redness
consistency
legit
highlight
daily
improvement
wow
choice
combo
size
complimentary

A sample of the features identified in the TF-IDF metadata.

A sample of partial TF-IDF data for a single customer review. Each review has 8,388,608 characters of data.

REVIEW_ID	TFIDF_DATA
262888147	504B0304140000000800000021005 1A7535E85620000D03301000B001400696E64696365732E6E707901001000D033010000000000856200000000000009D9D09DC6763F9FFEF738E4496242485990943F635FB1A12224B25EB189264AC09957DC956764AB26F455124CA32288C6CD99992A52469F9F951A1FA5FEFE75C1FF775EEE77CBF8FDFDFABD3F96EF39C7BB9D6CFB5DCA77D62EB4D37DFA64A07A4AF4CDC65D7FD26EF3B71D5711357DF7D85894B8C9BF8B9BD27A5F44EBB2F62F7B9EC3AD35...

Top 50 Words in Customer Reviews for Best of Beauty **Winners**

1.	maybe	14.	side	27.	rather	40.	helped
2.	yellow	15.	along	28.	fav	41.	dark
3.	strikes	16.	experienced	29.	mask	42.	slight
4.	purchased	17.	refund	30.	gave	43.	wore
5.	main	18.	reccomend	31.	gets	44.	bit
6.	how	19.	kinda	32.	practical	45.	happy
7.	compare	20.	dyson	33.	game	46.	looking
8.	found	21.	lot	34.	giving	47.	try
9.	as	22.	flake	35.	skincare	48.	washing
10.	sunscreen	23.	setting	36.	only	49.	Instead
11.	perfectly	24.	of	37.	innovative	50.	causing
12.	two	25.	lip	38.	dispensing		
13.	soft	26.	helping	39.	tzone		

Top 50 Words in Customer Reviews for **Non-Winners**

1.	cause	14.	practical	27.	match	40.	smoothed
2.	hydrating	15.	ways	28.	top	41.	strikes
3.	maybe	16.	bronze	29.	those	42.	sprayed
4.	price	17.	found	30.	two	43.	quite
5.	back	18.	person	31.	they	44.	soothing
6.	chunky	19.	now	32.	detects	45.	free
7.	dark	20.	sunscreen	33.	device	46.	did
8.	setting	21.	nicely	34.	clean	47.	brand
9.	before	22.	behind	35.	difference	48.	seen
10.	plumping	23.	despite	36.	into	49.	expectations
11.	kinda	24.	however	37.	bit	50.	attachments
12.	soft	25.	deeply	38.	looking		
13.	difficult	26.	perfectly	39.	lot		




Models

- Sentiment Analysis (VADER)
- Logistic Regression
- Random Forest
- XG Boost

Sentiment Analysis (VADER)

Sentiment Comparison: How does customer sentiment differ between Best of Beauty award winners and non-award winners?

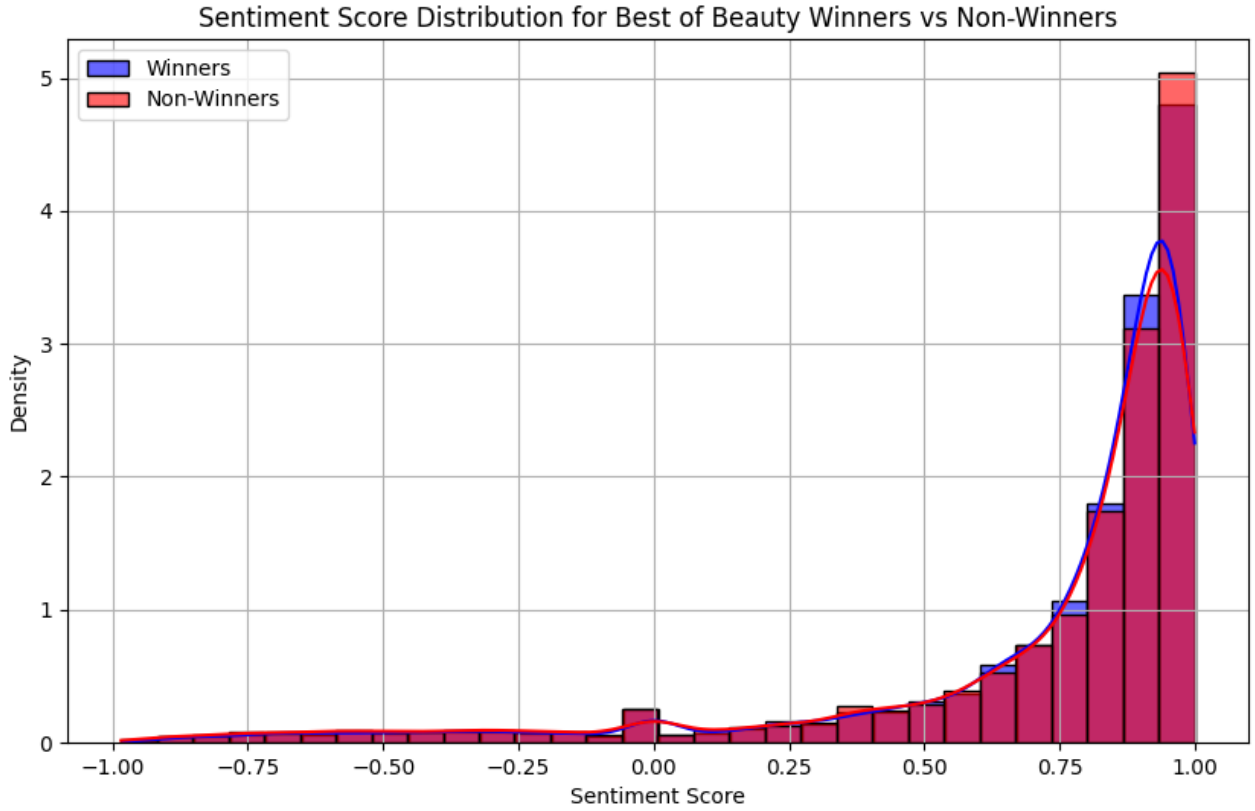
VADER (Valence Aware Dictionary and sEntiment Reasoner)

- A sentiment analysis tool that uses a pre-built lexicon to quantify text sentiment, ideal for reviews and social media. [11]
- Produces a **compound score** from -1 (most negative) to +1 (most positive). [11]
- **Benefit**: Efficiently captures emotional tone, providing quantitative sentiment scores to compare award winners with non-winners. [11]

Key Findings:

- **T-statistic (4.2714)** indicates a significant difference between the two groups.
- **P-value (0.0000)** confirms this difference is statistically significant.
- The density plot shows sentiment is generally higher for winners, consistent with statistical results.

Interpretation: Sentiment is higher for winners, and this difference is unlikely due to chance.



Predictive Analysis

Classification Models: How accurately can we predict which products will be Best of Beauty winners based on customer ratings and sentiment scores?

Logistic Regression

Objective: Since VADER sentiment analysis showed that sentiment scores are statistically significant when identifying Best of Beauty winners from non-winners, I took the analysis a step further.

Method: Train a logistic regression model on sentiment scores (as determined by VADER) and customer review ratings (1-5). The approach is to determine whether these two predicting variables allow us to accurately predict whether a product will be a Best of Beauty award winner.

Why Logistic Regression: The data set used in this analysis contains close to 800,000 records of customer review data. The largest table in the database used for the analysis is 19.5 GB. Processing time and capability are a concern in this analysis, and Logistic Regression is a simple classification model that is computationally efficient with large data sets. I am using it in this analysis for its scalability and efficiency. [12]

Results (First Model): After training the first Logistic Regression model, the test data set showed an **accuracy of 58.5%** along with significant class imbalance issues: it only identified 3% of non-winners and heavily favored winners, accurately identifying 98%.

Results (Second Model): After adjusting for class imbalance, the second Logistic Regression model showed a prediction **accuracy of 56%**. It improved on correctly predicting non-winners (up from 3% to 38%), but at the expense of accurately identifying winners, which dropped to 69% accuracy.

Next Steps: Train Random Forest and XG Boost models to see if either can handle class imbalance better than the Logistic Regression model.

Predictive Analysis (Continued)

Random Forest

```
Random Forest Accuracy: 0.5252019669827889
Random Forest Confusion Matrix:
 [[2254 2505]
 [2902 3727]]
Random Forest Classification Report:
              precision    recall  f1-score   support

           0       0.44      0.47      0.45      4759
           1       0.60      0.56      0.58      6629

    accuracy                           0.53     11388
   macro avg       0.52      0.52      0.52     11388
weighted avg       0.53      0.53      0.53     11388
```

After training and testing, the Random Forest model performed with an **overall accuracy of 52.5%**. The results were more balanced than the Logistic Regression model, accurately predicting non-winners at 47%. This confirms that the Random Forest model handles the class imbalance better, but still performed with an overall low accuracy.

XGBoost

```
XGBoost Accuracy: 0.5298559887600983
XGBoost Confusion Matrix:
 [[2267 2492]
 [2862 3767]]
XGBoost Classification Report:
              precision    recall  f1-score   support

           0       0.44      0.48      0.46      4759
           1       0.60      0.57      0.58      6629

    accuracy                           0.53     11388
   macro avg       0.52      0.52      0.52     11388
weighted avg       0.54      0.53      0.53     11388
```

The XGBoost model predicted winners and non-winners with **52.9% accuracy**, showing the best recall for non-winners of all the classification models (48%). Overall performance was comparable to the other classification models.

Keyword Analysis

Classification Models: Are there specific keywords in customer reviews that distinguish Best of Beauty winners from non-winners, and can these keywords help predict future award winners?

XGBoost

```
Accuracy: 0.568184534480698
Confusion Matrix:
 [[ 909 6410]
 [ 816 8599]]
Classification Report:
              precision    recall  f1-score   support

           0       0.53      0.12      0.20      7319
           1       0.57      0.91      0.70      9415

    accuracy                           0.57     16734
   macro avg       0.55      0.52      0.45     16734
weighted avg       0.55      0.57      0.48     16734
```

Since the XGBoost model showed the highest prediction accuracy while balancing the classes, I chose to expand on it further by incorporating features, or the top keywords in customer review text as determined by TF_IDF, as a predicting variable.

Overall model accuracy was 56.8%, but with high class imbalance again.

Conclusions

Classification models: Despite training and testing a range of classification models, including logistic regression, Random Forest, and XGBoost, modifying the parameters to accommodate class imbalance, and adding additional predicting variables, the features I trained the models on—sentiment scores, ratings, and keyword analysis—were not effective at accurately predicting Best of Beauty winners. The models consistently struggled with classifying non-winners, highlighting a strong bias toward predicting the majority class (winners) and achieving only moderate overall performance.

Sentiment analysis: The VADER analysis indicated a significant difference in customer sentiment between winning and non-winning products. However, the statistical significance does not imply that sentiment alone is a sufficient predictor for determining which products are winners.

Significant overlap between the two groups: Even though winners tend to have higher sentiment scores and customer ratings on average, there is still a large overlap in sentiment scores between winners and non-winners, which makes it difficult for models to separate the two groups reliably.

Conclusion: In summary, this analysis demonstrates that customer ratings, review text, and sentiment analysis, when considered in isolation, are not sufficient for predicting Allure Best of Beauty award winners. Generally, customer reviews skew towards positive sentiment and higher ratings. Still, there is a statistically significant increase in positive customer sentiment towards award-winning products versus non-winners.

Future Analysis

Ideas for building on this analysis in the future:

- Using customer review data from additional popular beauty retailers (e.g., Ulta). Some Best of Beauty winners are only sold on the brand's website, so future analyses could also incorporate brand data. However, precautions would need to be implemented to ensure that incorporating reviews from a brand's website does not lead to positively skewed review data.
- Incorporating additional product information into the classification models, such as product price.
- Implementing Time-Series analysis to analyze how customer sentiment evolves over time, particularly before and after a product wins a Best of Beauty award.
- Further advanced Natural Language Processing models may lead to more accurate winner vs. non-winner classifications based on sentiment analysis.

References

1. Cambridge Dictionary. (n.d.). *Product review*. In Cambridge Dictionary. https://dictionary.cambridge.org/us/dictionary/english/product-review
2. PowerReviews. (2023). *Health & beauty shopping trends 2023*. https://www.powerreviews.com/research/health-beauty-shopping-trends-2023/
3. Allure. (n.d.). *About Allure*. https://www.allure.com/info/about-allure
4. FasterCapital. (n.d.). *Beauty award nomination: Behind the scenes—How beauty brands secure award nominations*. https://fastercapital.com/content/Beauty-award-nomination--Behind-the-Scenes--How-Beauty-Brands-Secure-Award-Nominations.html
5. Inside Sephora. (n.d.). *Life at Sephora*. https://www.inside-sephora.com/en/usa/life-at-sephora
6. Cosmetify. (n.d.). *The Cosmetify Index*. https://www.cosmetify.com/us/the-cosmetify-index/
7. Allure. (2021). *How Allure's Best of Beauty judging process works*. https://www.allure.com/story/how-allure-best-of-beauty-judging-process-works
8. Allure. (2023). *Best of beauty 2024 winners*. https://www.allure.com/best-of-beauty-2024-winners
9. Capital One. (n.d.). *Understanding TF-IDF*. Capital One Tech. https://www.capitalone.com/tech/machine-learning/understanding-tf-idf/
10. Scikit-learn developers. (2024). *sklearn.feature_extraction.text.TfidfVectorizer — Scikit-learn 1.5.0 documentation*. https://scikit-learn.org/1.5/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

References (Continued)

11. Slavanya, R. (n.d.). *VADER: A comprehensive guide to sentiment analysis in Python*. Medium. Retrieved from https://medium.com/@rslavanyageetha/vader-a-comprehensive-guide-to-sentiment-analysis-in-python-c4f1868b0d2e
12. LinkedIn. (n.d.). *Which machine learning algorithms provide the fastest analysis?* Retrieved from https://www.linkedin.com/advice/3/which-machine-learning-algorithms-provide-fastest-nsq5f#:~:text=Linear%20algorithms%20such%20as%20Linear,significantly%20speeds%20up%20the%20analysis.