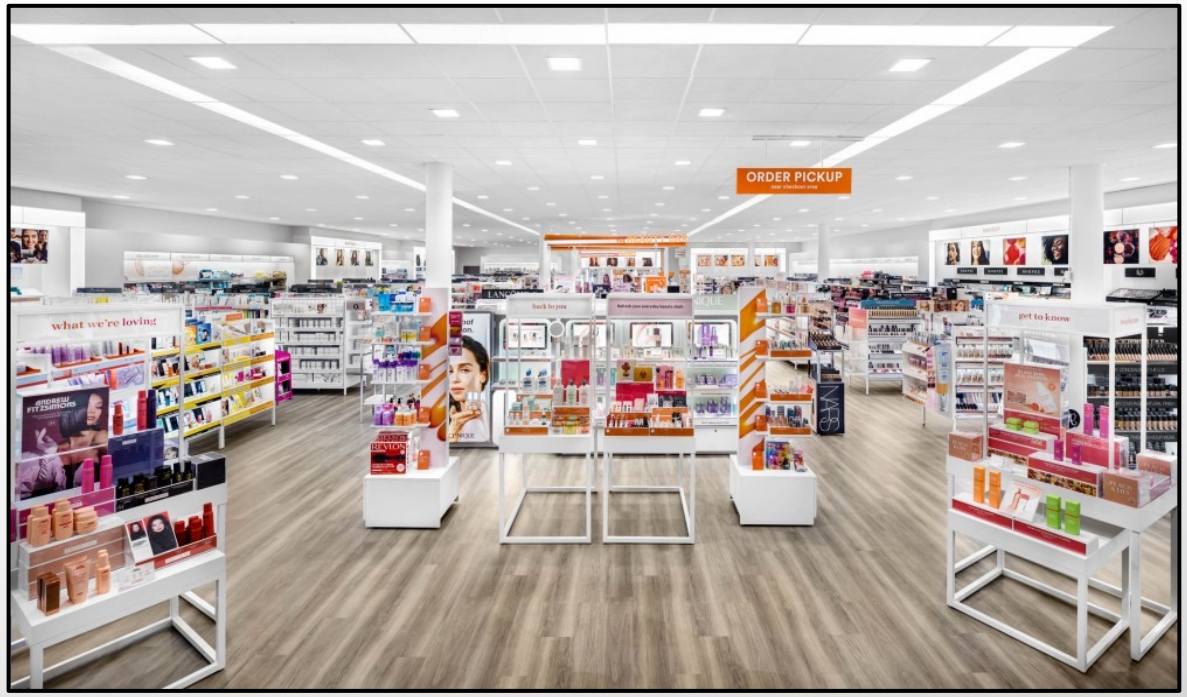
BEAUTY BEYOND RATINGS

UNVEILING THE SENTIMENT AND SECRETS OF BEAUTY PRODUCTS USING ULTA'S ONLINE TEXT DATA

Andrew Mashhadi – Stats 425

INTRODUCTION

- Ulta remains as one of the most popular chains of beauty stores in the United States.
- Carries both high-end and low-end cosmetics, fragrances, nail products, bath and body products, beauty tools, and haircare products.
- Contains over 600 brands and 120 product categories.
- Many brands are not provided with detailed performance metrics or sentiment of their corresponding products from Ulta.



INTRODUCTION

- Using modern text-mining methods with the reviews, ratings, and descriptions for each product, we can extract more information from the Ulta website than a traditional survey and would be able to save cost and time by leveraging the automation of our analyses.
- We aim to enhance our understanding of the Ulta dataset, unravel hidden patterns, and unlock invaluable insights for both consumers and businesses in the beauty industry by accomplishing the following distinctive goals:

Exploratory Approach

a) Uncover customer sentiment

b) Revealing word trends with term frequencies & TF-IDF

c) Discovering commonalities with topic modeling

d) Contrasting term frequencies

Modeling Approach

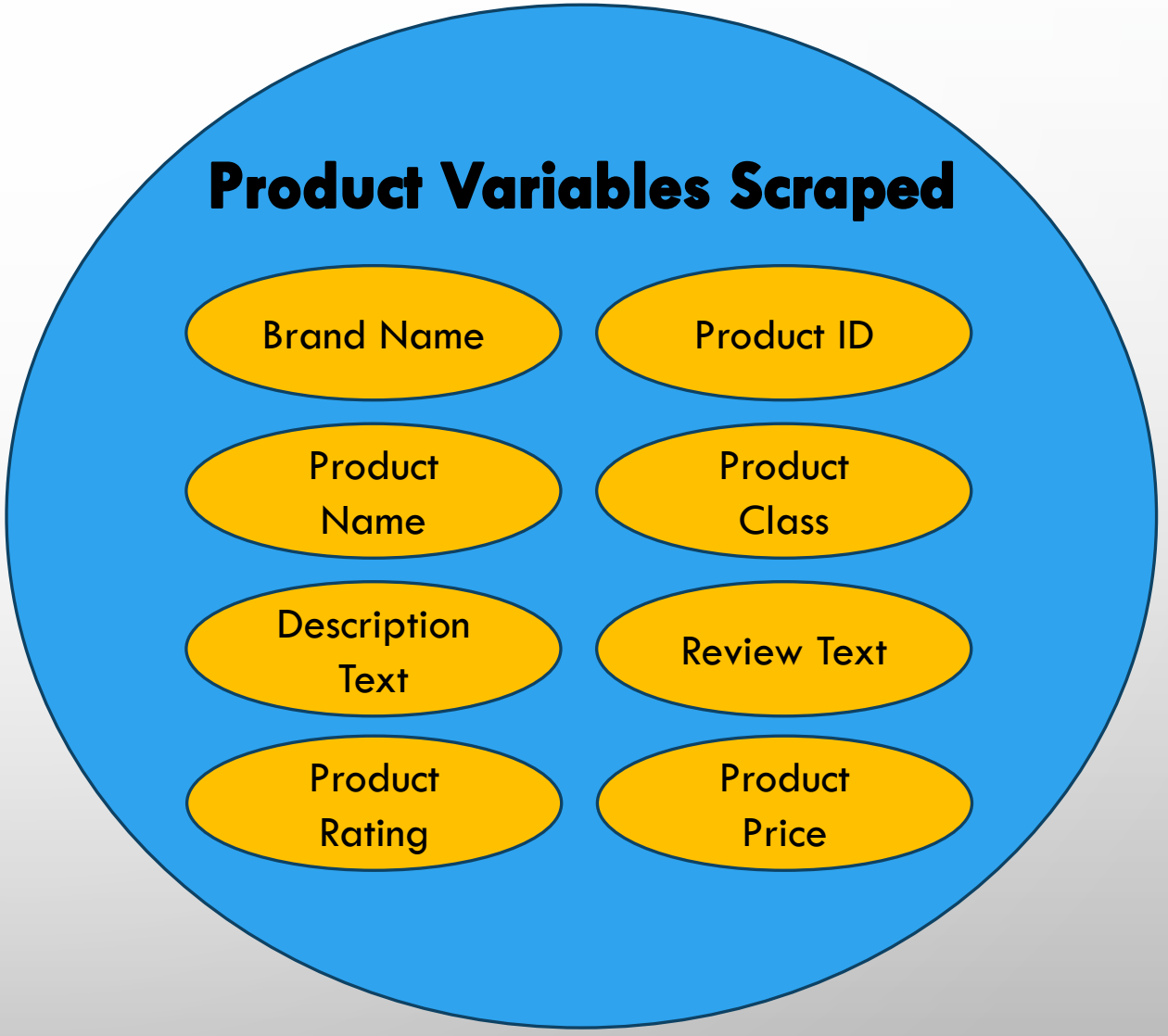
a) Predicting average star ratings with reviews

b) Unveiling the price-sentiment nexus

c) Leveraging product descriptions to predict rating

DATA COLLECTION

- Scraped Ulta's online text data from each brand and each product.
- Collected 15,247 unique products from the Ulta website.
- 5 reviews per product (to save storage space and processing time).
- Over 61,457 customer reviews were used in our study.



EXPLORATORY TEXT ANALYSIS

UNCOVERING CUSTOMER SENTIMENT

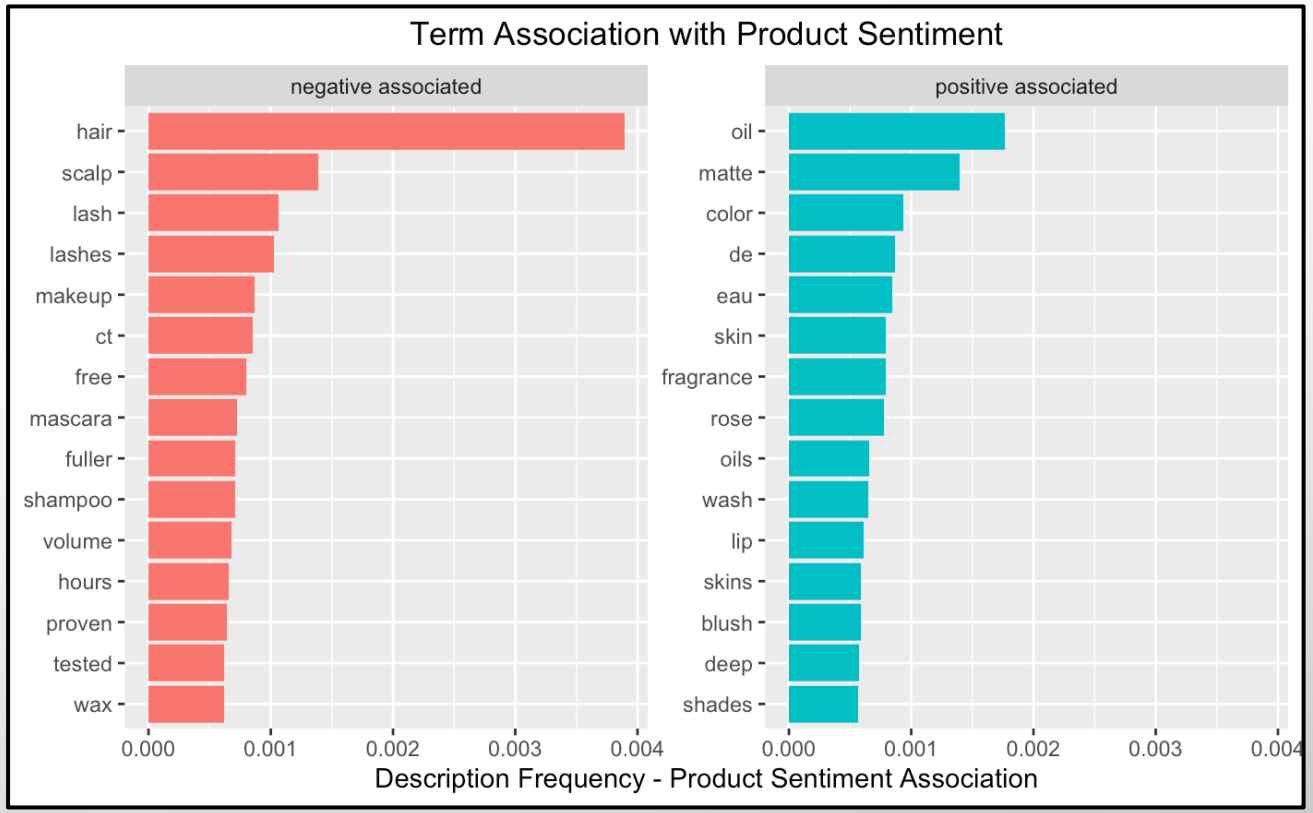
- Preprocessing: removed numbers, punctuation, and English stop-words, then eliminated duplicate reviews.
- Used *AFINN Sentiment Lexicon* to extract each review's overall score.
- Gave equal weight to each review (positive or negative), then used proportion of positives for each product as score.
- Used rating to validate predicted sentiment scores.
- Observed a significant positive relationship between star ratings and predicted sentiment, with a correlation estimate of 0.44.

(a) Rating < 1.5

(b) 4.5 < Rating

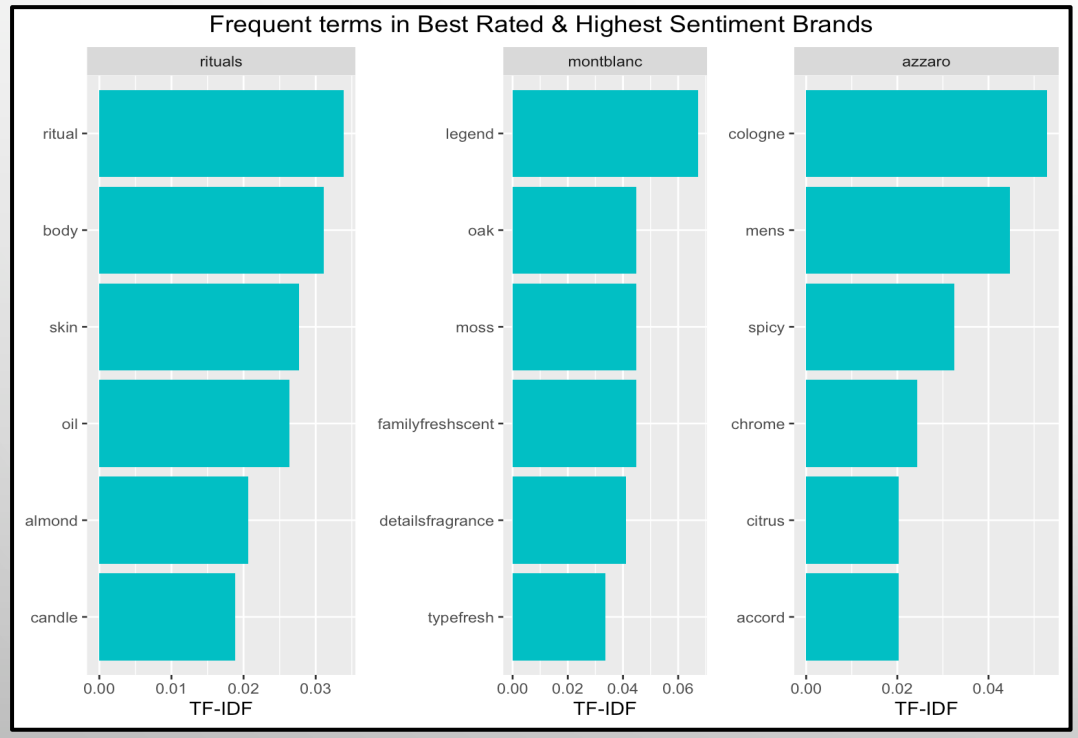
REVEALING WORD TRENDS WITH TERM FREQUENCIES & TF-IDF

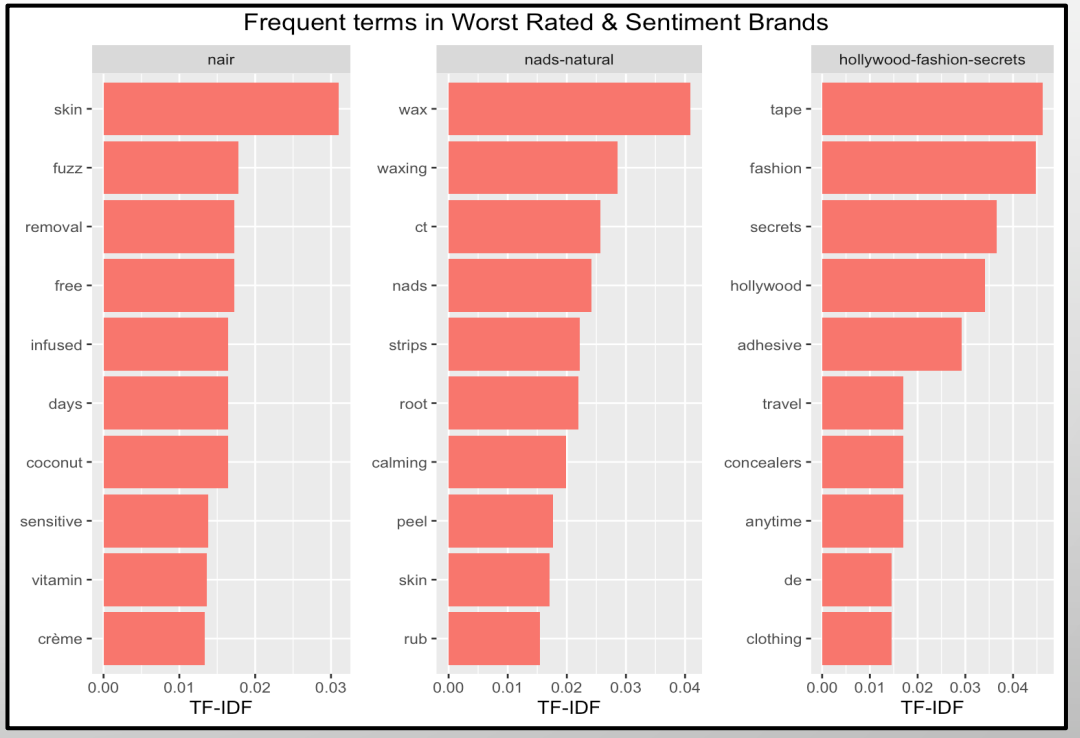
- We assessed word frequencies within product descriptions based on the rating of the individual products and the overall sentiment found in the previous section.
- Found the terms associated with the largest absolute difference in relative frequencies.
- Observed that negative sentiment was associated with products related to hair care and eye makeup, while positive sentiment was linked to products related to cleansers and lip & cheek makeups.



REVEALING WORD TRENDS WITH TERM FREQUENCIES & TF-IDF

- Examined words with the highest TF-IDF scores for brands or product categories with the best and worst combined customer sentiment and rating.
- Used weighted average of rating and predicted sentiment to extract more meaningful insights from the review text than from star ratings alone.
- Notice the cologne and fragrance (Montblanc and Azzaro) brands in the higher combined scores and the hair removal and body taping brands in the lower combined scores.





REVEALING WORD TRENDS WITH TERM FREQUENCIES & TF-IDF

- As a brand, consider asking the following questions:
    1. Which product categories exhibit the highest customer sentiment and ratings for our brand ?
    2. Which other brands receive the best customer sentiment and ratings within a specific product category ?
- We illustrate our approach to answering these questions with a few hand-selected brands and categories.
- Again, we used the combined rating to extract comprehensive insights from both customer reviews and star ratings.

Brand	Best Product Category	Worst Product Category
Peach & Lily	Face Peels & Exfoliators	Cleansing Balms & Oils
Colourpop	Lip Liner	Mascara

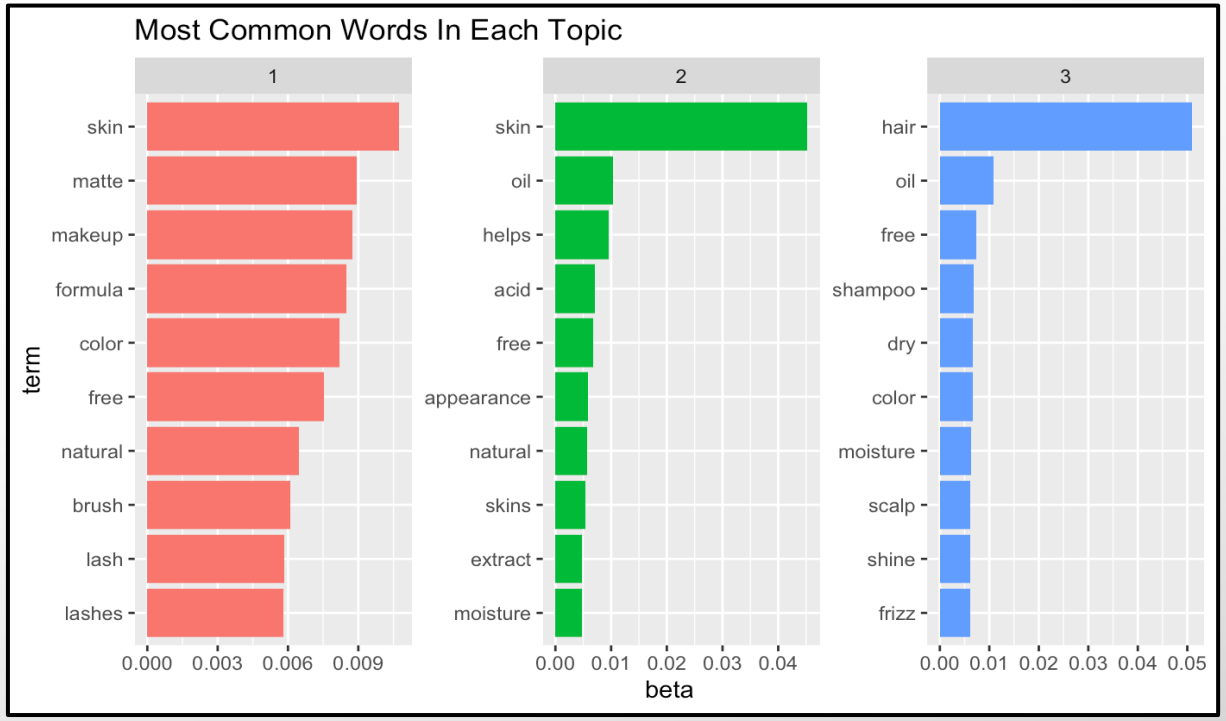
Table 1: Best and Worst Product Categories Associated with Each Brand

Product Category	Best Brand	Worst Brand
Shampoo	Blind Barber	Its A 10
Face Moisturizer	Black Opal Beauty	BYOMA
Mascara	Elf Cosmetics	Morphe Cosmetics

Table 2: Best and Worst Brands Associated with Each Category

DISCOVERING COMMONALITIES WITH TOPIC MODELING

- Utilized Latent Dirichlet Allocation (LDA) to build topic models on our text data.
- Goal was to identify underlying themes and commonalities across diverse brands, offering a holistic view of the beauty industry.
- Treated the set of product descriptions for each brand as a separate “document” to determine the most suitable mixture of topics that characterizes each brand.
- Found that k=3 topics provided the most effective separation of themes.



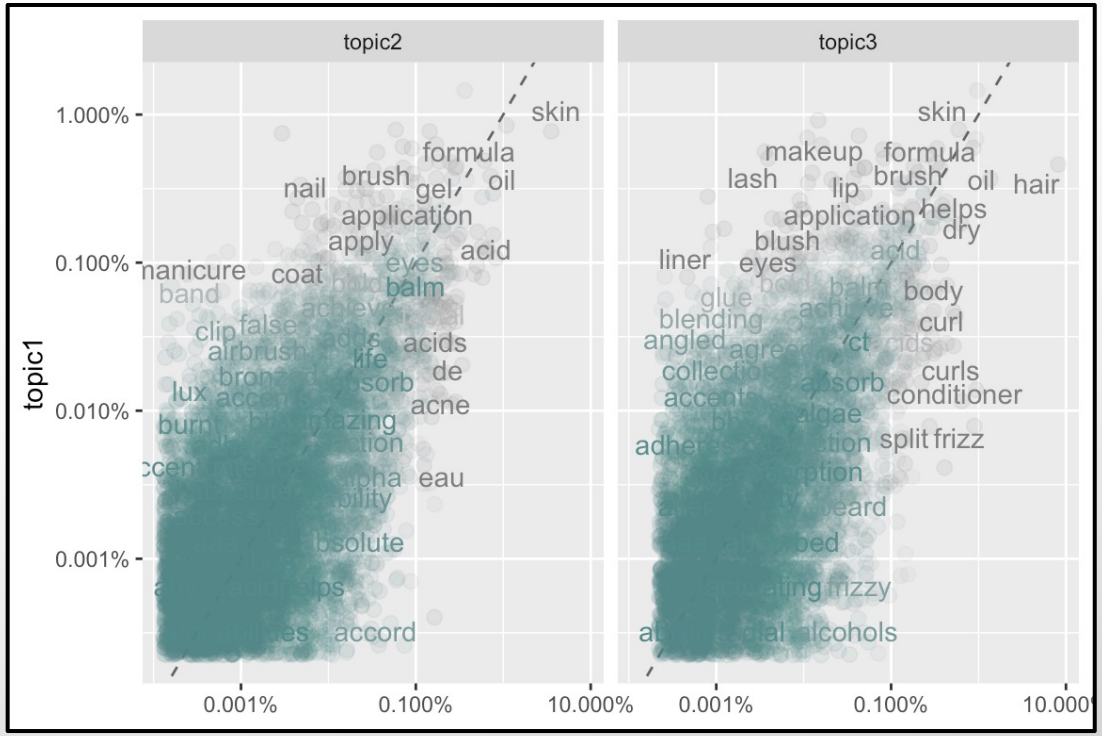
(1) Makeup / Lashes

(2) Face & Body Cleansers / Moisturizers

(3) Hair care

CONTRASTING TERM FREQUENCIES IN TOPICS & MAKEUP BRANDS

- First compared the term frequency proportions among the generated topics.
- Scatter plots were utilized to visually depict the correlation between topics and evaluate the words responsible for their associations.
- Pearson's correlation coefficients were then employed to quantify the level of correlation in term frequencies between topics.
- Finding indicates that word frequencies between makeup brands and skin care brands are closer related.
- This makes sense because makeup and skin care products are often applied to similar areas of the body.

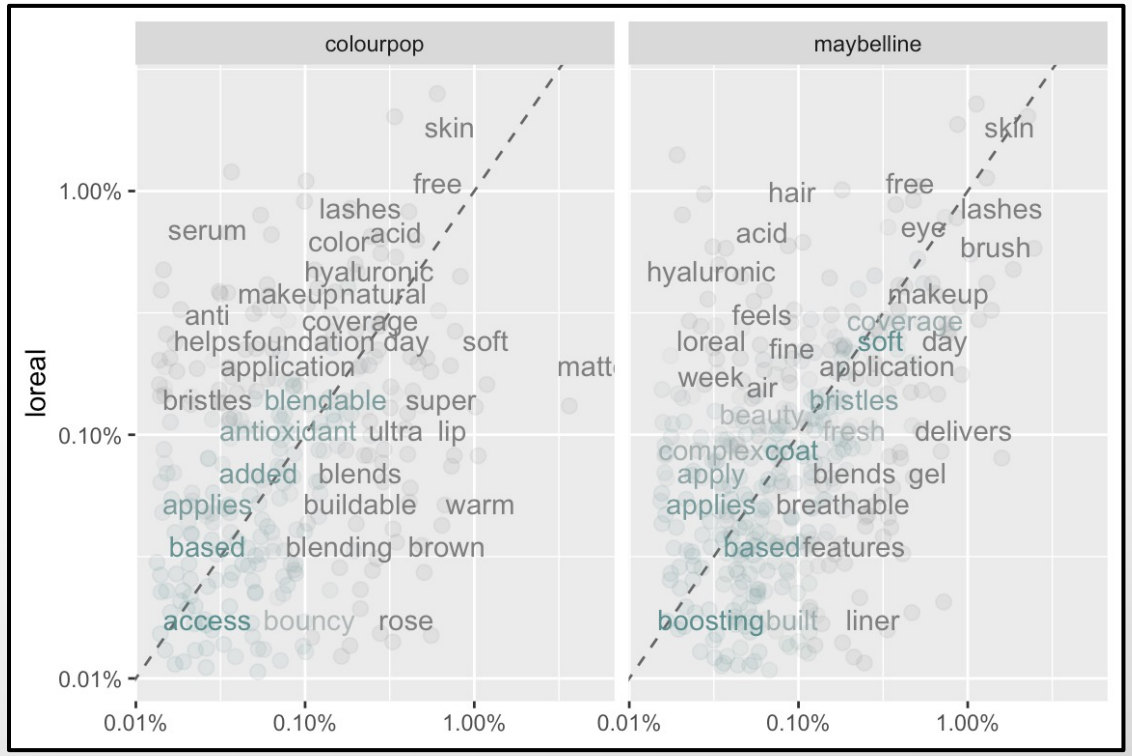


T1 vs. T2
Pearson's Corr. = 0.64

T1 vs. T3
Pearson's Corr. = 0.40

CONTRASTING TERM FREQUENCIES IN TOPICS & MAKEUP BRANDS

- Conducted a similar correlation analysis based on the word frequencies from three prominent makeup brands: L'Oreal, Colourpop, and Maybelline.
- Observed that the word frequencies between L'Oreal & Maybelline have stronger correlation than L'Oreal & Colourpop.
- Although all three brands exclusively focus on makeup, Colourpop stands out with their choice of descriptive words such as "blending", "brown", "warm", "matte", and "bouncy".



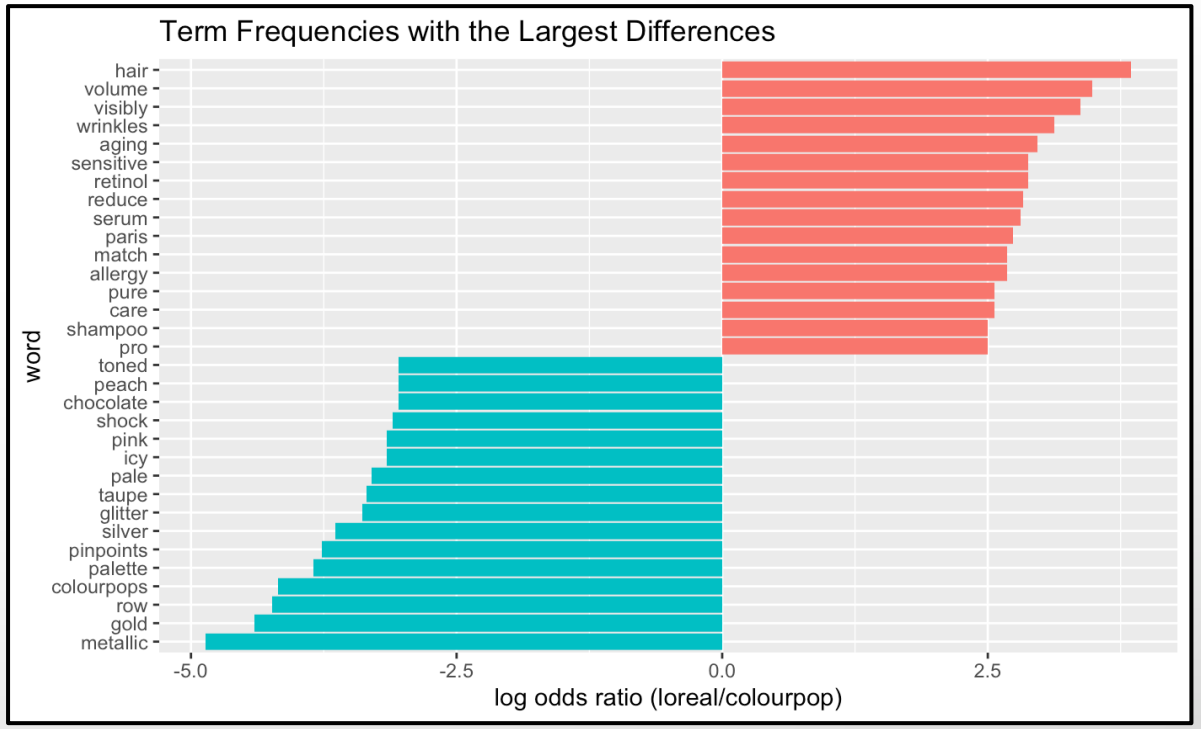
L'Oreal vs. Colorpop
Pearson's Corr. = 0.16

L'Oreal vs. Maybelline
Pearson's Corr. = 0.65

CONTRASTING TERM FREQUENCIES IN TOPICS & MAKEUP BRANDS

So, what distinguishes Colourpop from L'Oreal (and Maybelline) ?

- We utilized the log odds ratio to identify the top 15 most distinctive words for both L'Oreal and Colourpop.
- L'Oreal's distinctive words, such as "wrinkles", "aging"," and "retinol", are associated with slightly older customers seeking anti-aging products.
- Colourpop's distinct terms consist of more youthful and vibrant descriptors, including "peach", "chocolate", "pink", "icy", "glitter", "silver", "gold", and "metallic".
- L'Oreal (and Maybelline) has a long-standing history dating back to the early 1900s, whereas Colourpop was founded in 2014.

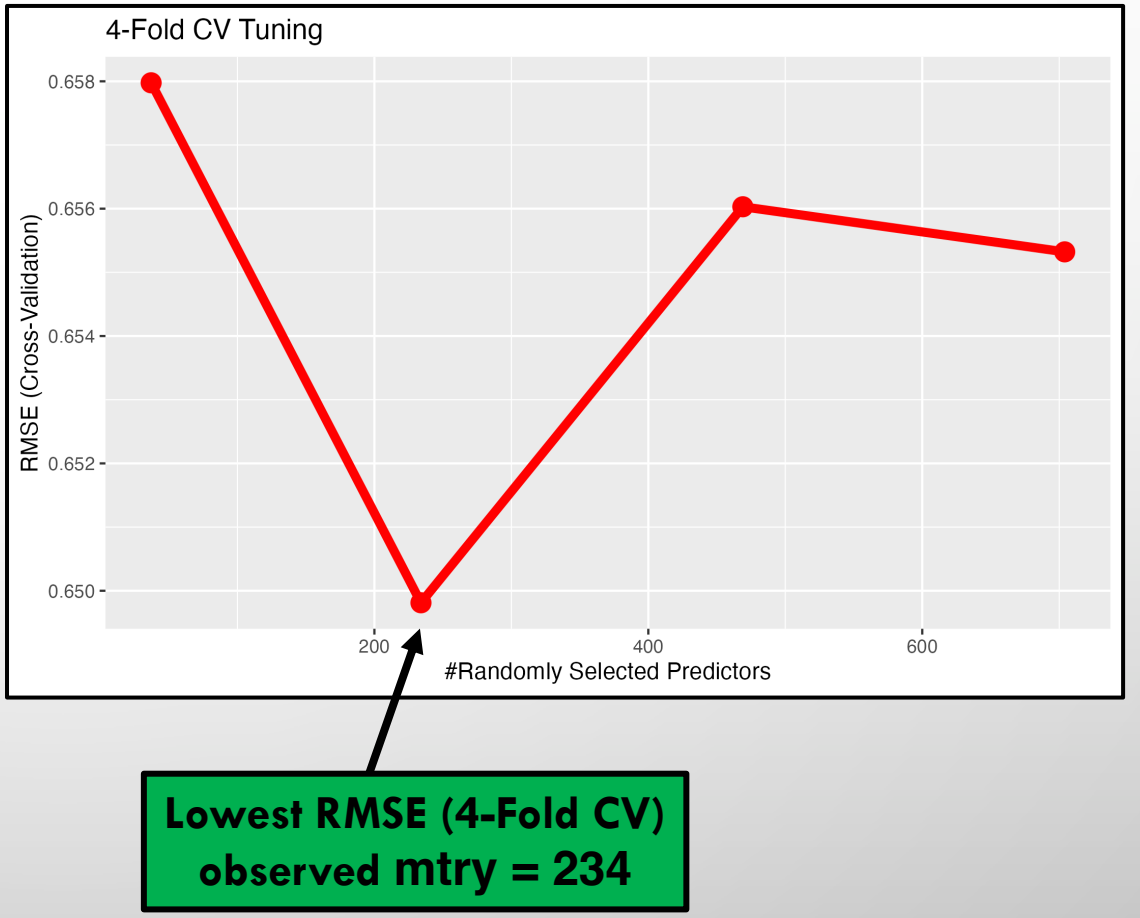


Colourpop's target customers are generally younger, possibly belonging to the Gen-Z demographic, while L'Oreal tends to cater to a slightly older customer base.

MODELING

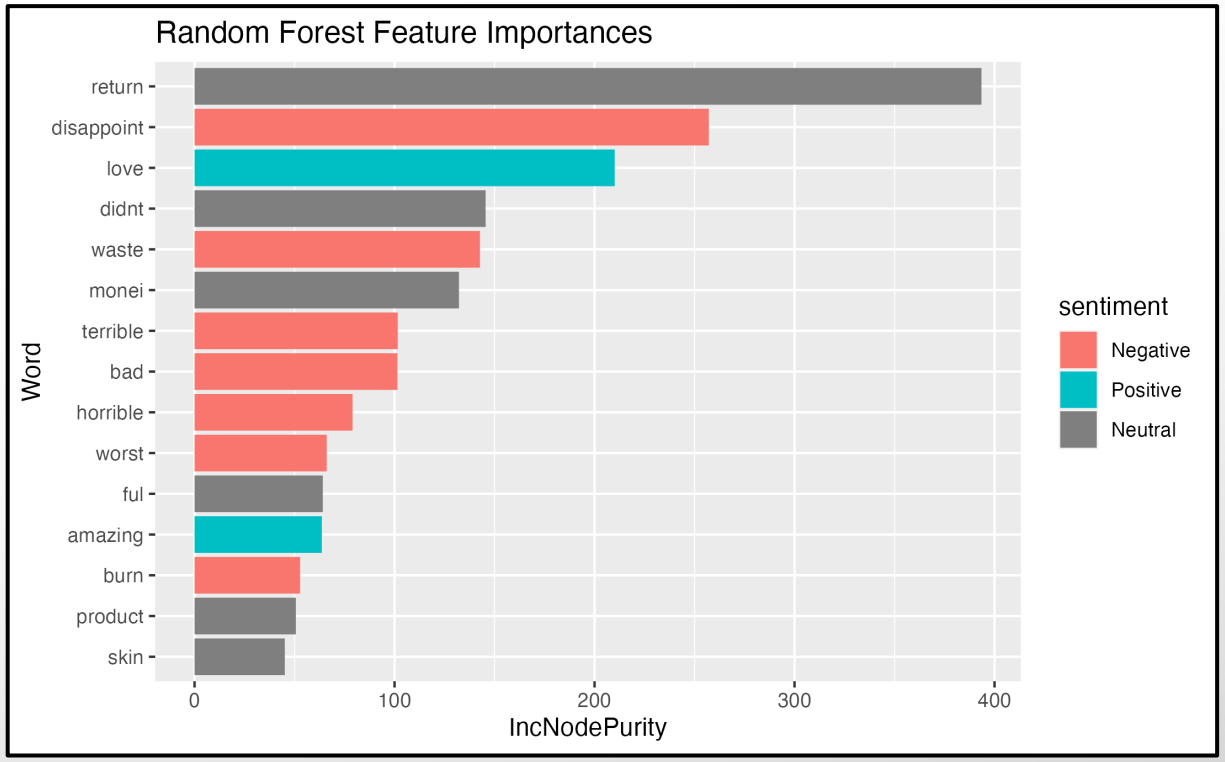
PREDICTING AVERAGE STAR RATINGS FROM REVIEWS

- Goal to assess how well the text in the customer reviews can be used to predict the average star rating of a product.
- Removed highly sparse terms (max sparse = 0.95) and performed stemming to reduce feature space and minimize confusion among words with similar meanings.
- Trained random forest regressor with 80% text data.
- Tuned model using grid search with 4-Fold CV and RMSE.
- Found that n=500 and mtry=234 were optimal.



PREDICTING AVERAGE STAR RATINGS FROM REVIEWS

- Final training took ~18min, with reported training RMSE of 0.2675 stars
- Using 20% unseen test data, observed an RMSE of 0.6176.
- Reported R-Squared of 0.427 for the test data, explaining a significant proportion of the variance in the star ratings.
- Feature importances show how much model error (RSS) increases when variable is randomly permuted.
- Bing sentiment shown where applicable, indicating only 2 positive words in the top 15 most important terms within the model.
- Importances align with intuition.



"I am *disappointed*."

"I *loved* this product"

"I *returned* this product"

UNVEILING THE PRICE-SENTIMENT NEXUS

- Unraveled the complex interplay between product price and customer sentiment by modeling their relationship.
- Used a binomial logistic regression model with product price as the predictor variable and sentiment as the response variable.
- Globally fitted model showed that an increase in price corresponds to a more positive sentiment across all product classes and brands.
- Analysis at the brand level revealed significant associations between price and sentiment for 7 brands.
- At the product category level, only 4 categories exhibited a significant relationship.

GLOBAL MODEL

Variable	Estimate	Std. Error	Z value	$\Pr(>\|Z\|)$
Intercept	2.474844	0.053189	46.529	≈ 0
price	0.005441	0.001489	3.654	≈ 0.0003

Table 4: Summary of Global Logistic Regression Model

Brand	Term	Estimate	Std. Error	Z value	$\Pr(>\|Z\|)$
Gimme Beauty	price	0.8005	0.3393	2.3591	0.0183
Ouidad	price	0.1336	0.0622	2.1460	0.0319
Invisibobble	price	0.4328	0.2079	2.0815	0.0374
Fourth Ray Beauty	price	0.6285	0.3052	2.0586	0.0395
Scunci	price	0.3547	0.1751	2.0262	0.0427
Burts Bees	price	-0.2047	0.1024	-1.998	0.0457
OGX	price	-0.3730	0.1870	-1.994	0.0461

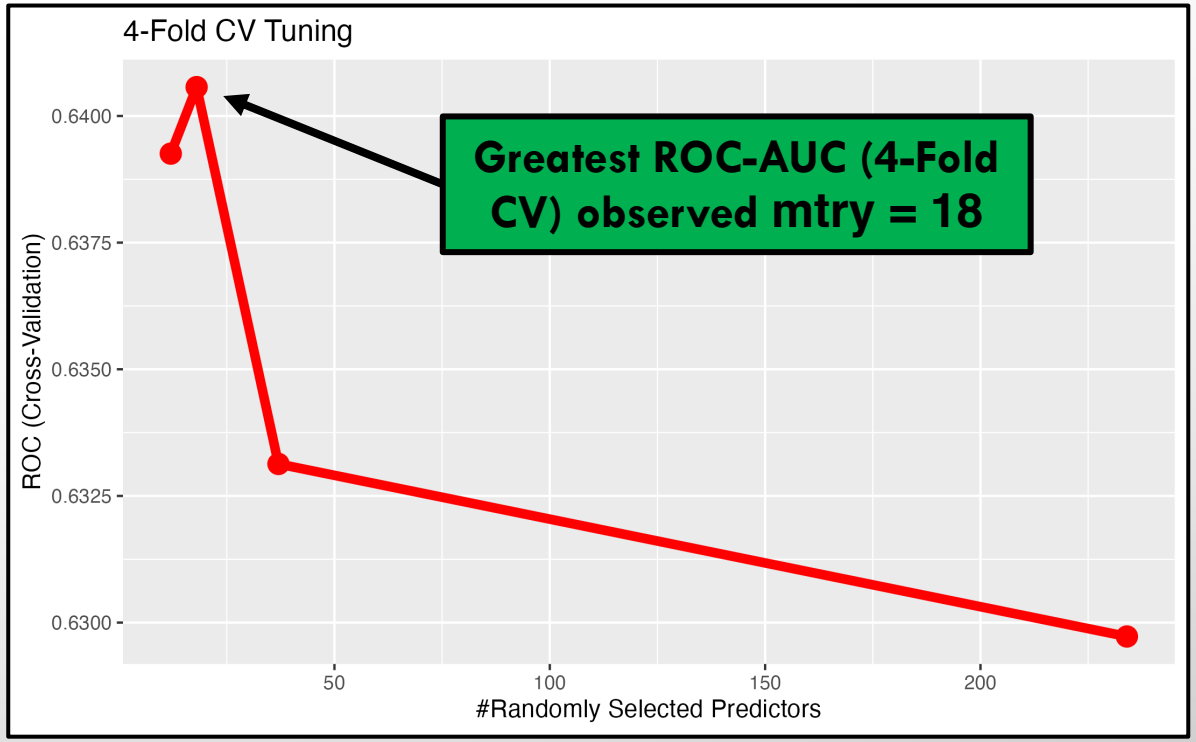
Table 5: Brands with Significant Relationships

Product Category	Term	Estimate	Std. Error	Z value	$\Pr(>\|Z\|)$
Anti-Aging	price	-0.0212	0.0092	-2.3084	0.0210
Wax & Pomade	price	0.1519	0.0705	2.1530	0.0313
Face Moisturizer	price	0.0193	0.0010	2.0182	0.0436
Masks	price	-0.0486	0.0246	-1.9727	0.0485

Table 6: Product Categories with Significant Relationships

LEVERAGING TEXTUAL DESCRIPTIONS

- Goal to predict the customer rating of a product based solely on the text found in the product descriptions.
- Considered rating >= 4 as “good” and rating < 4 as “bad”.
- Also removed sparse terms and stemmed all words.
- Trained random forest *classifier* with 80% text data.
- Tuned model using grid search with 4-Fold CV and area under the ROC curve.
- Found that n=500 and mtry=18 were optimal.

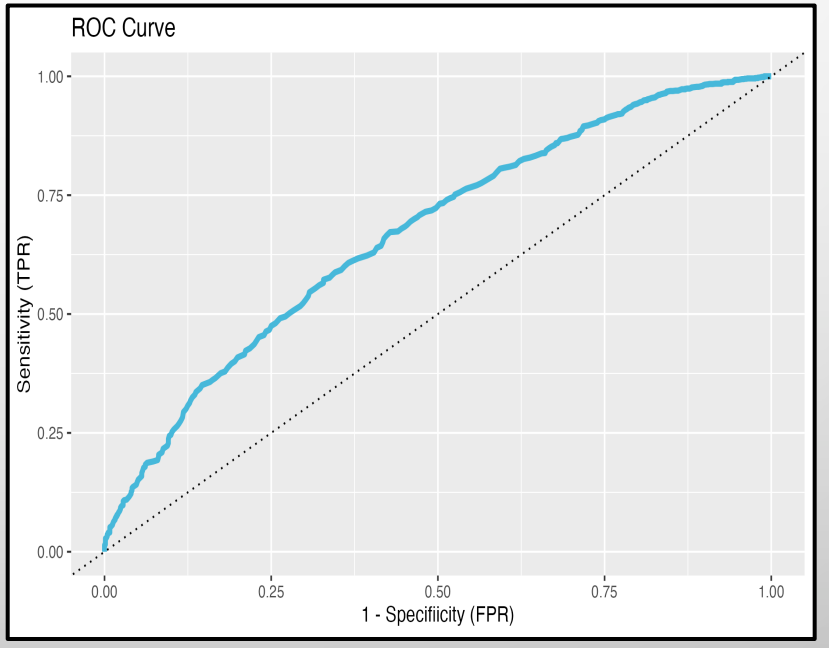


LEVERAGING TEXTUAL DESCRIPTIONS

- Final training took ~28min, with reported training accuracy of 98% (using default cut-point of 0.5).
- Used 5% of remaining unseen data as validation data to find optimal cut-point, prioritizing accuracy.
- Used optimal cut-point of t*=0.5 to test fully trained classifier with remaining 15% test set.
- Obtained test accuracy of **68%**, sensitivity of **94%**, and specificity of **20%**.
- Used ROC to assess performance without fixed cut-point.
- Area under the ROC curve was about 0.67, indicating an acceptable discrimination.

Cut-Point (t)	Sensitivity	Specificity	Balanced Accuracy	Accuracy
0.5	0.9214	0.1923	0.5568	0.656
0.55	0.8805	0.2582	0.5694	0.654
0.49	0.9214	0.1813	0.5514	0.652
0.51	0.9120	0.1978	0.5549	0.652
0.53	0.8931	0.2308	0.5619	0.652

Table 7: Random Forest Classifier Cut-Point Tuning with Validation Set



With t* = 0.5

		Predicted		
		Bad	Good	Total
Actual	Bad	151	601	752
	Good	78	1280	1358
	Total	229	1881	

CONCLUSION & FUTURE WORK

- Overall, this project significantly enhanced our understanding of the Ulta dataset, uncovering hidden patterns, and providing invaluable insights for consumers and businesses in the beauty industry.
- These findings demonstrated practical implications for brands to optimize their strategies, improve customer satisfaction, and foster a deeper understanding of customer preferences.
- Further research can build upon these findings with a larger sample size of reviews for each product to provide a more accurate measure of sentiment.