Sephora

What Drives Product Recommendations?

IRONHACK FINAL PROJECT

Elham Allahbakhshi

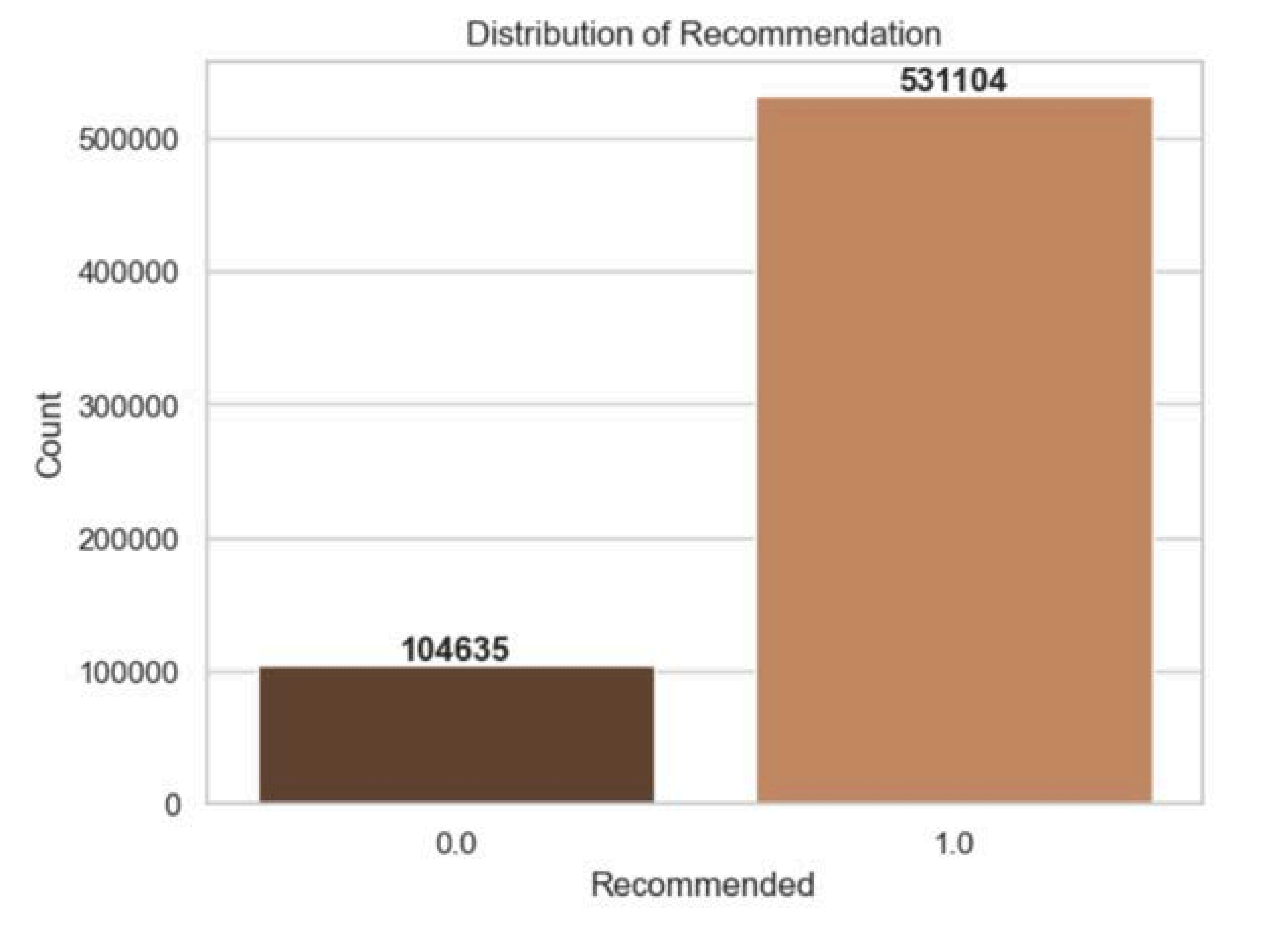
Project Overview

- Product name
- Brand name
- Price
- Discount price
- Sephora edition
- Rating
- Review text
- Review title
- Skin_tone
- Eye_color
- Skin_type
- Category
- Is_recommended

Project objective: Identifying factors that influenced consumers recommending a product

Distribution of Recommendation

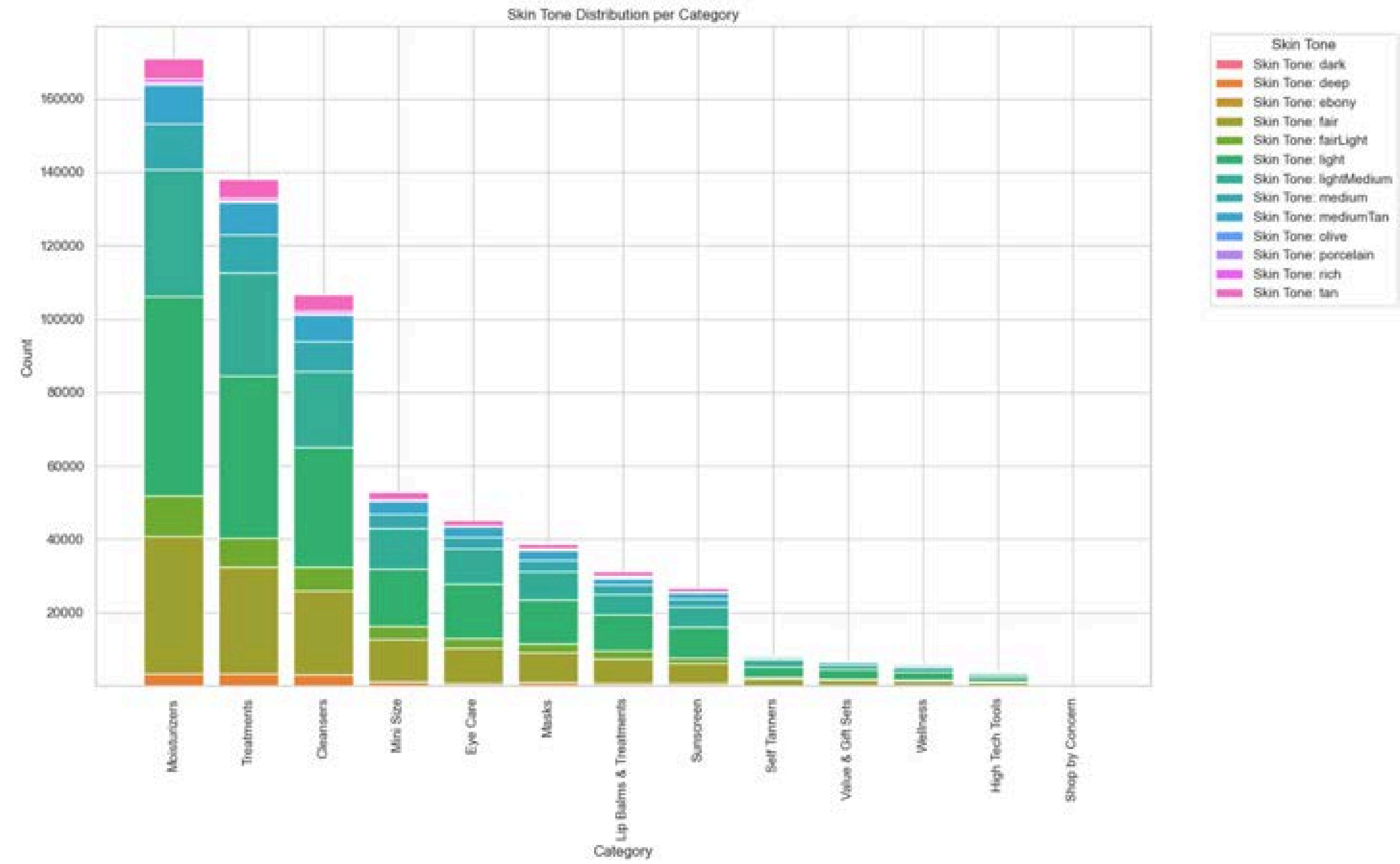
- Out of 635739 reviews
- 531104 recommended
- 104635 not recommened



Hypothesis Testing

- H0: There is no significant association between category and skin_tone.
- H1:There is significant association between categroy and skin_tone.

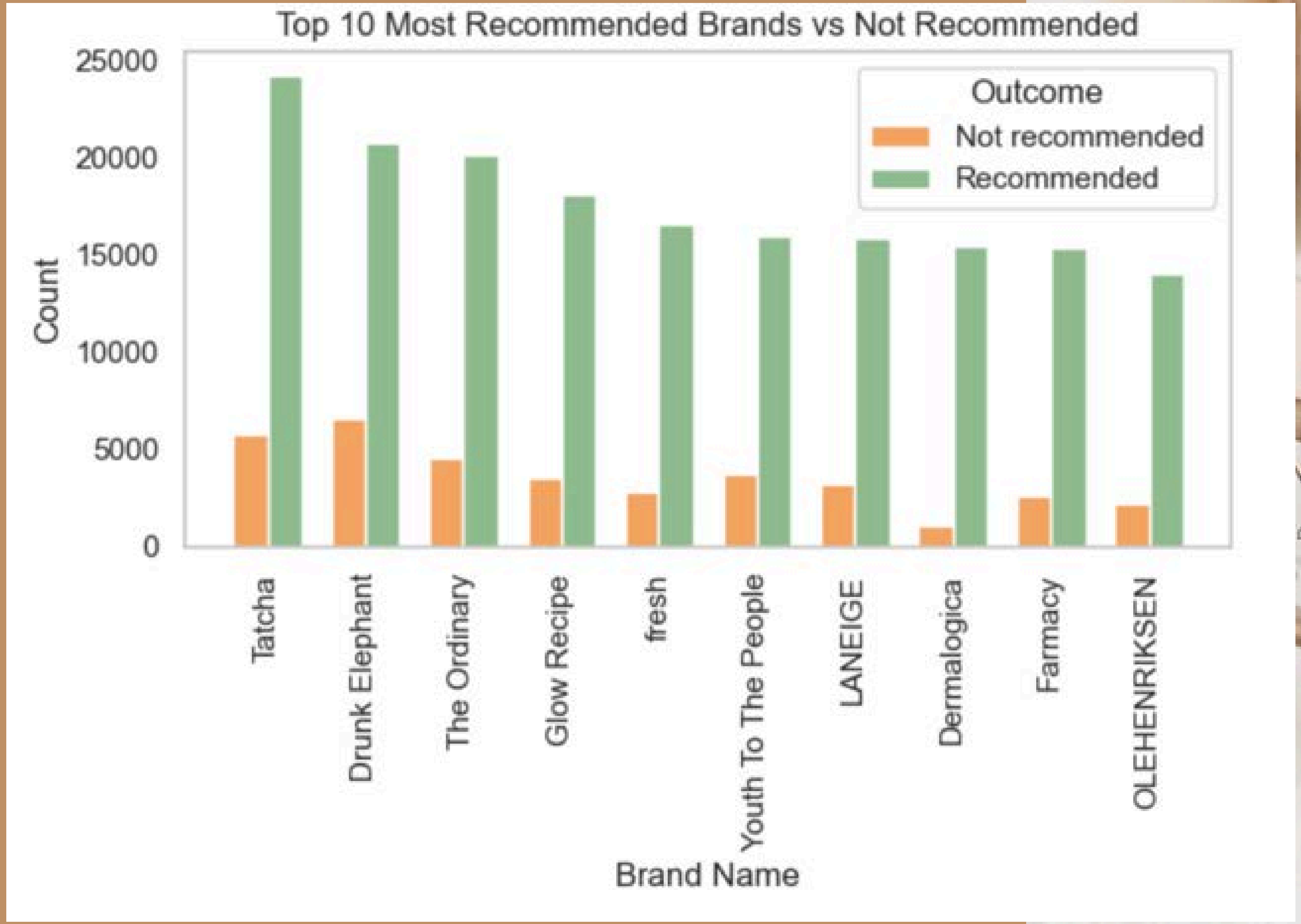
- Chi Squared test
- Reject null hypothesis
- Small Cremre's V 0.02



Hypothesis Testing

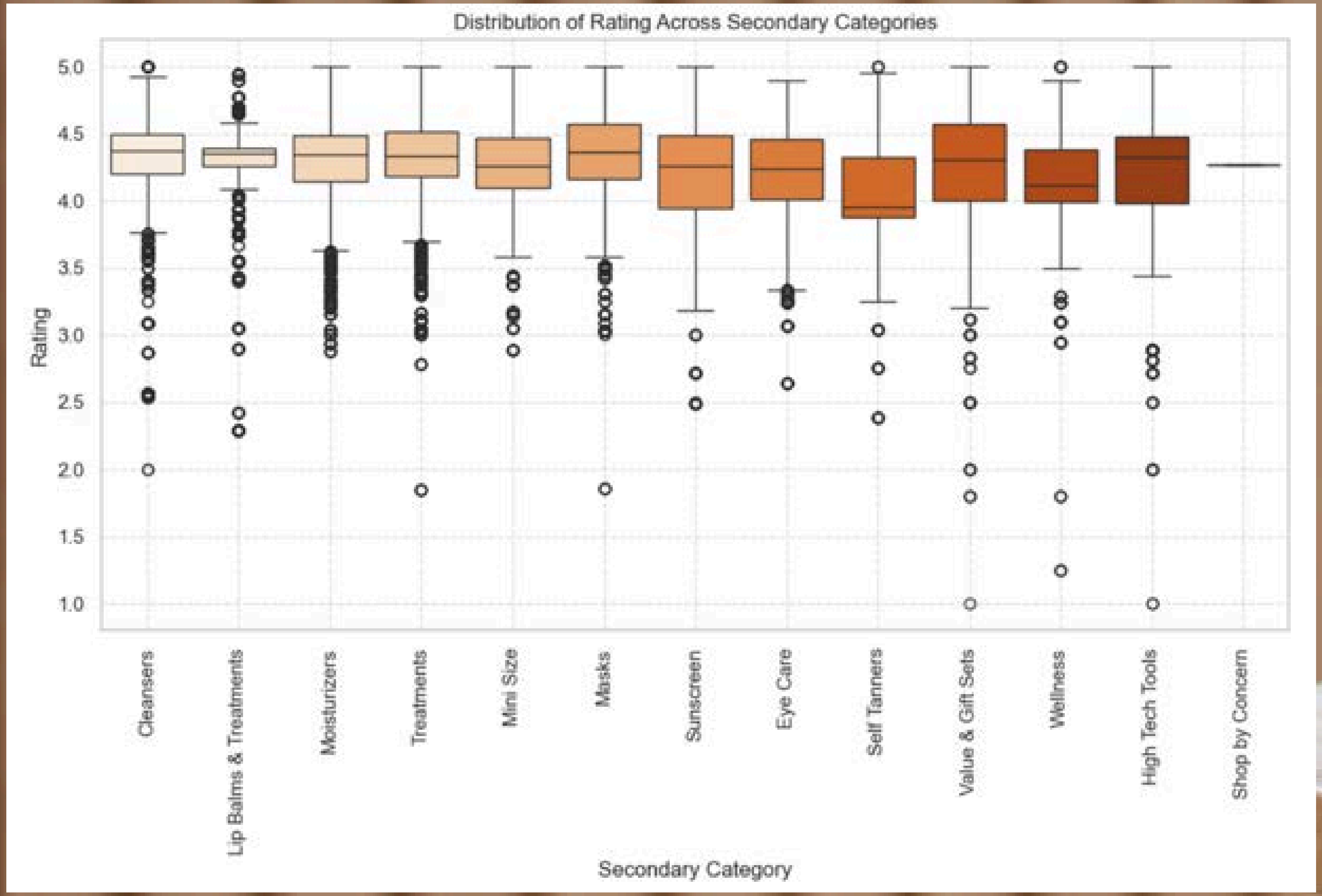
- H0: There is no significant association between is recommended and brand name.
- H1: There is a significant association between is recommended and brand name.

- Chi - squared test
- Reject Null Hypothesis
- Small Cremer's V 0.13



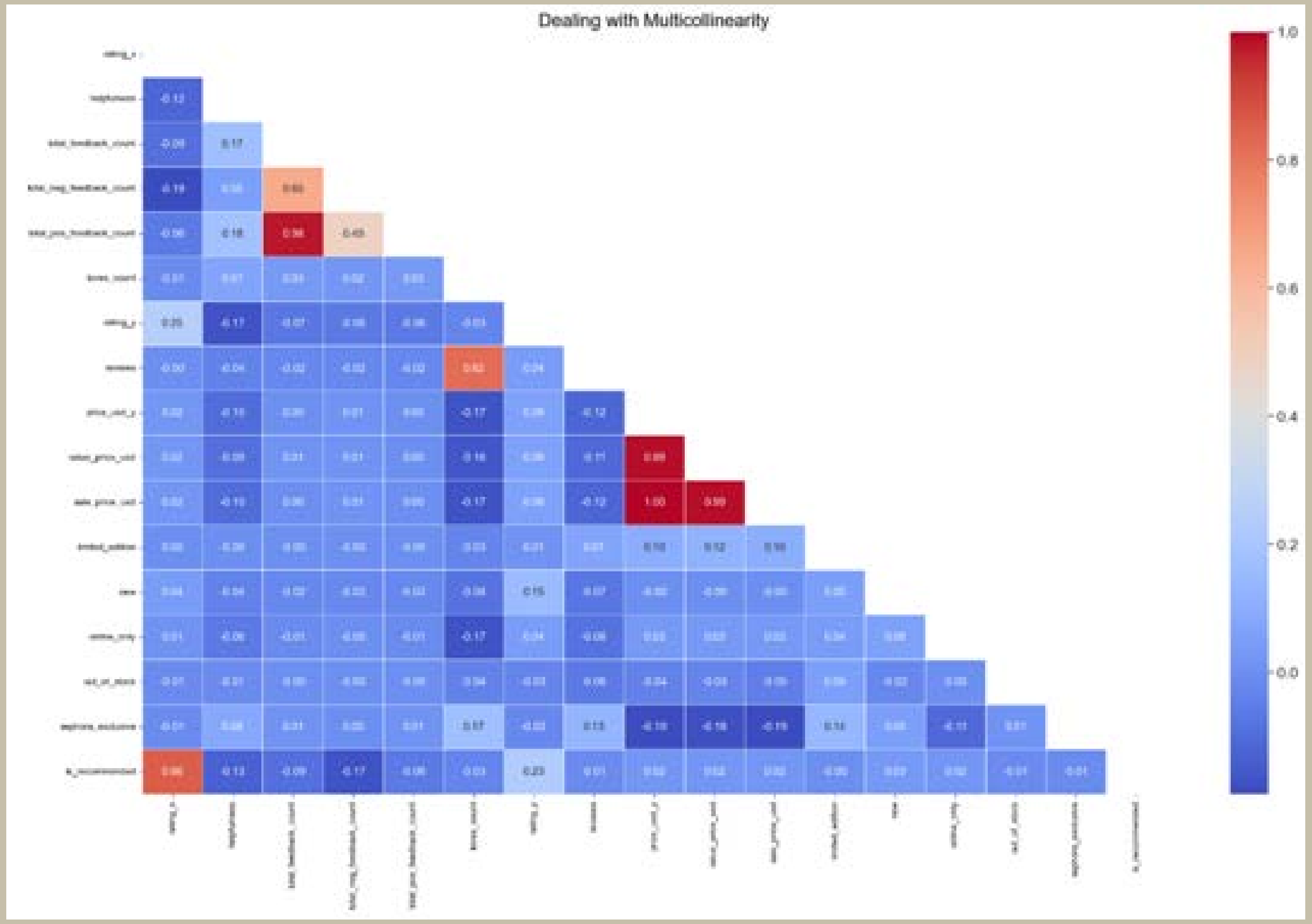
Hypothesis Testing

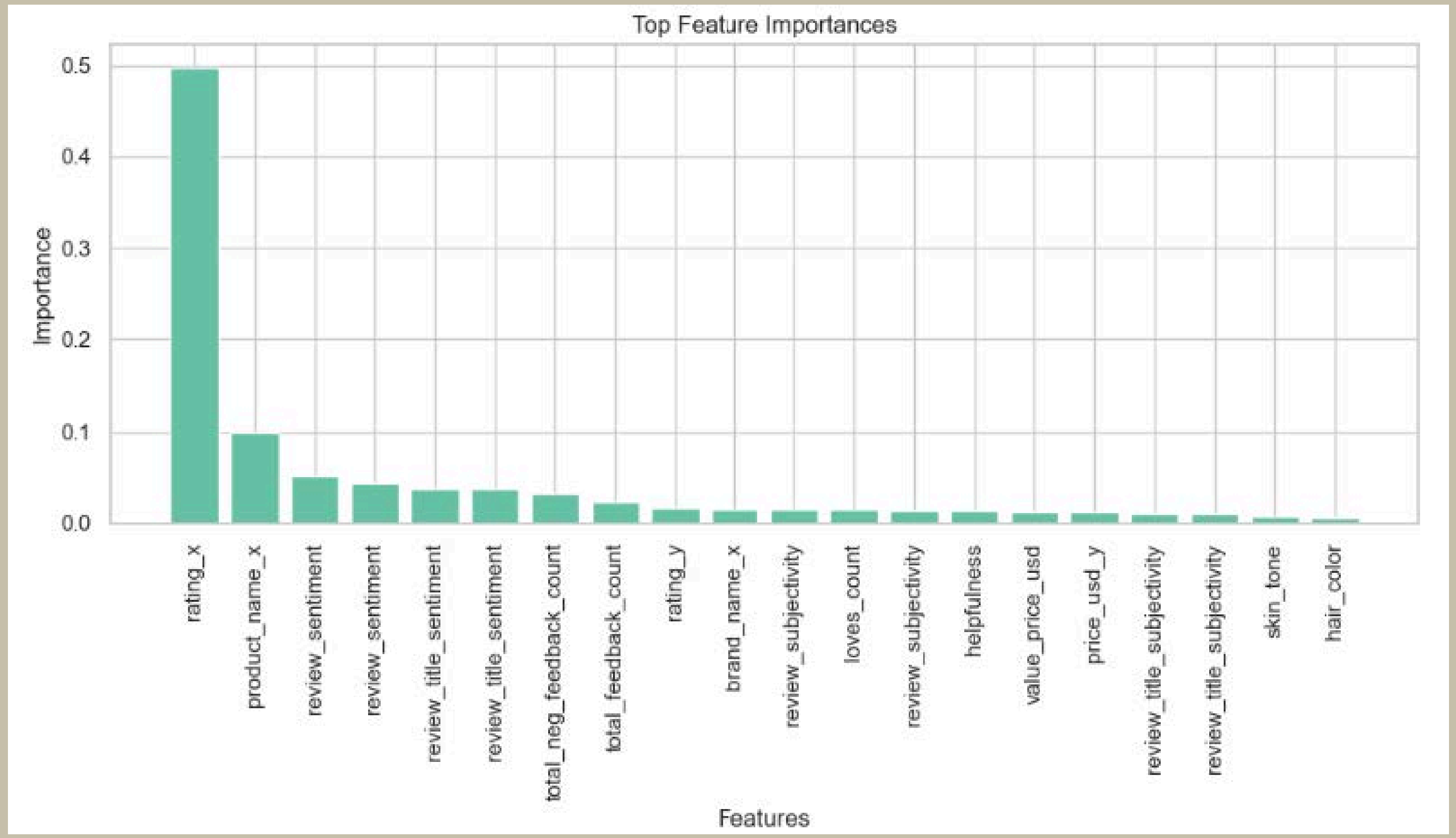
- H0: The mean rating is the same across all categories.
- H1: At least one category has a significantly different mean rating.



- Anova test
- Small P value- Reject null hypothesis
- Effect size (Eta squared) medium 0.0296

Machine Learning
Feature selection





Selected features:

rating, skin_tone, total_pos_feedback_count, loves_count, hair_color, skin_type, eye_color, reviews, sales_price, brand_name, category, online_only, review_title, review_text, etc

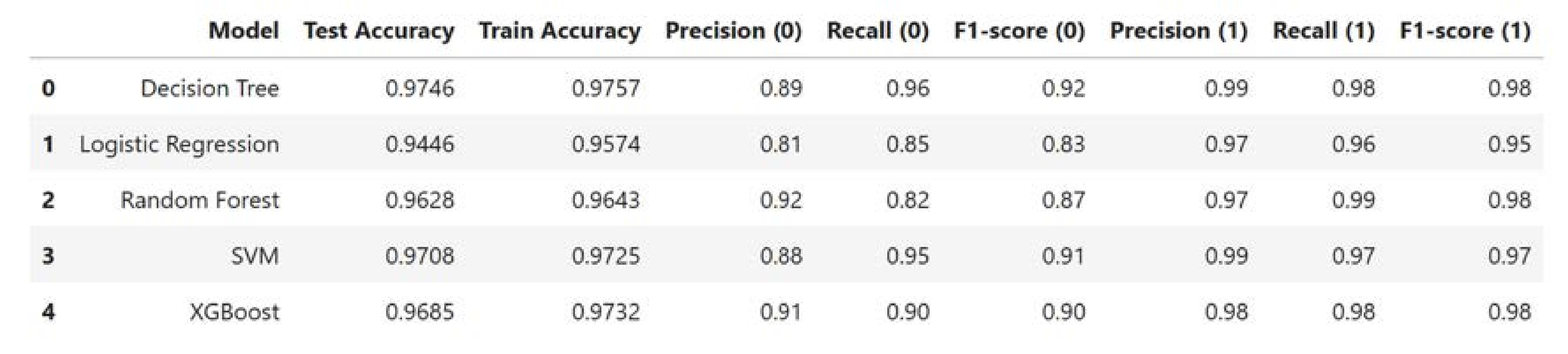
Target:

Is_recommended

Machine Learning Classification

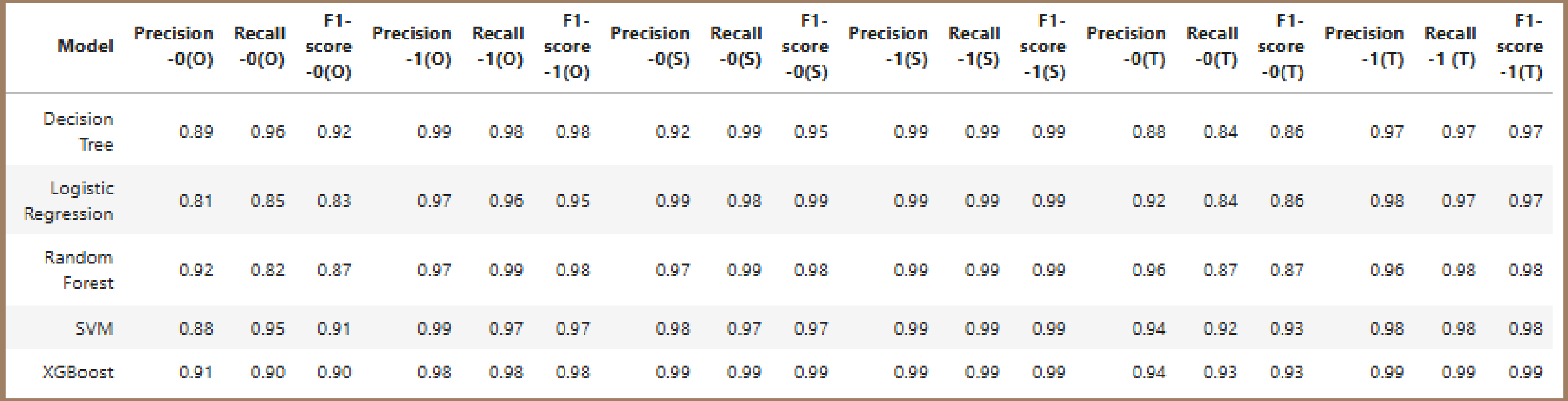
Models:

- Decision Tree
- Logistic Regression
- Random Forest
- Suport Vectore Model
- XGBoost



	Model	Test Accuracy	Train Accuracy	Precision (0)	Recall (0)	F1-score (0)	Precision (1)	Recall (1)	F1-score (1)
0	Decision Tree	0.9746	0.9757	0.89	0.96	0.92	0.99	0.98	0.98
1	Logistic Regression	0.9446	0.9574	0.81	0.85	0.83	0.97	0.96	0.95
2	Random Forest	0.9628	0.9643	0.92	0.82	0.87	0.97	0.99	0.98
3	SVM	0.9708	0.9725	0.88	0.95	0.91	0.99	0.97	0.97
4	XGBoost	0.9685	0.9732	0.91	0.90	0.90	0.98	0.98	0.98

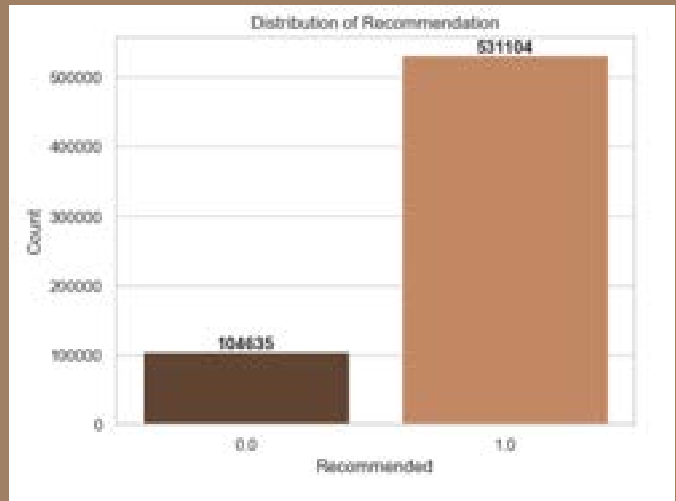
Machine Learning Imbalanced Data

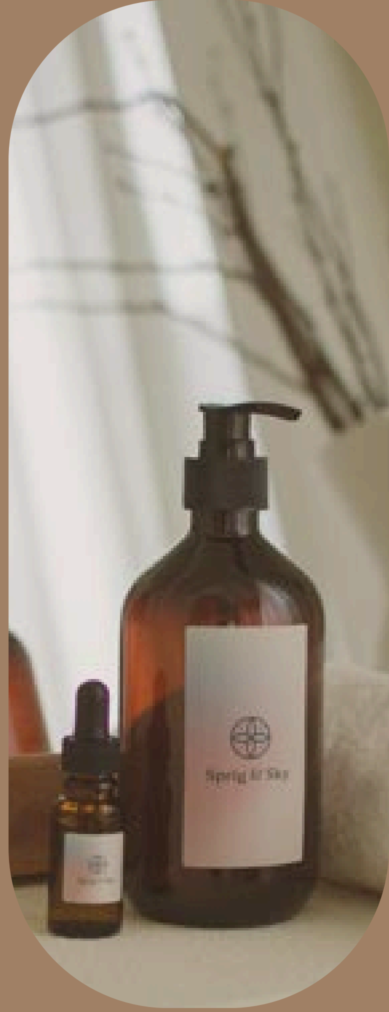


Model	Precision -0(O)	Recall -0(O)	F1-score -0(O)	Precision -1(O)	Recall -1(O)	F1-score -1(O)	Precision -0(S)	Recall -0(S)	F1-score -0(S)	Precision -1(S)	Recall -1(S)	F1-score -1(S)	Precision -0(T)	Recall -0(T)	F1-score -0(T)	Precision -1(T)	Recall -1 (T)	F1-score -1(T)
Decision Tree	0.89	0.96	0.92	0.99	0.98	0.98	0.92	0.99	0.95	0.99	0.99	0.99	0.88	0.84	0.86	0.97	0.97	0.97
Logistic Regression	0.81	0.85	0.83	0.97	0.96	0.95	0.99	0.98	0.99	0.99	0.99	0.99	0.92	0.84	0.86	0.98	0.97	0.97
Random Forest	0.92	0.82	0.87	0.97	0.99	0.98	0.97	0.99	0.98	0.99	0.99	0.99	0.96	0.87	0.87	0.96	0.98	0.98
SVM	0.88	0.95	0.91	0.99	0.97	0.97	0.98	0.97	0.97	0.99	0.99	0.99	0.94	0.92	0.93	0.98	0.98	0.98
XGBoost	0.91	0.90	0.90	0.98	0.98	0.98	0.99	0.99	0.99	0.99	0.99	0.99	0.94	0.93	0.93	0.99	0.99	0.99

Models:

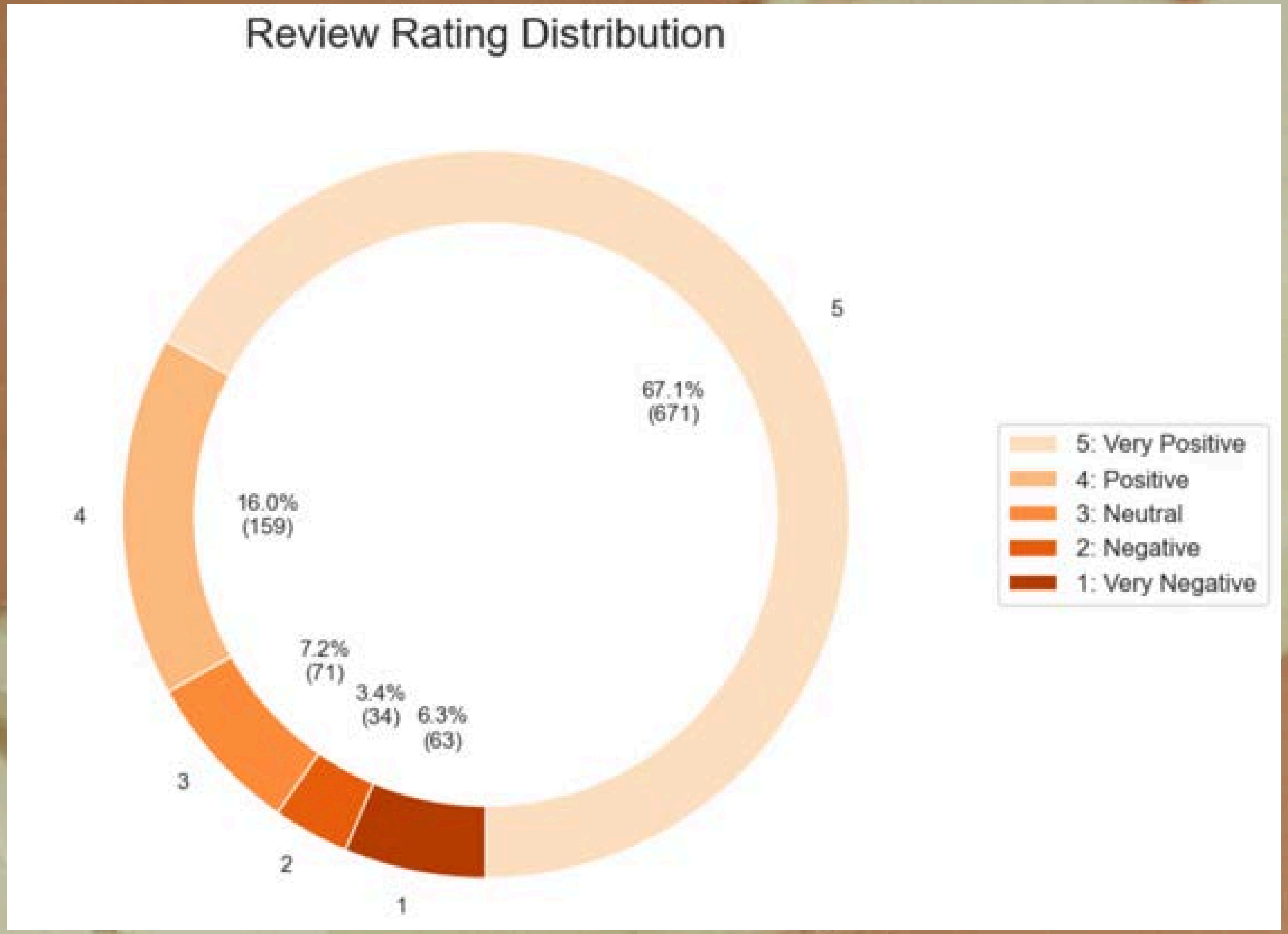
- Decision Tree
- Logistic Regression
- Random Forest: Good performance without Resampling
- SVM: Best Performer with Tomek
- XGBoost: Best Performer overll with SMOTE





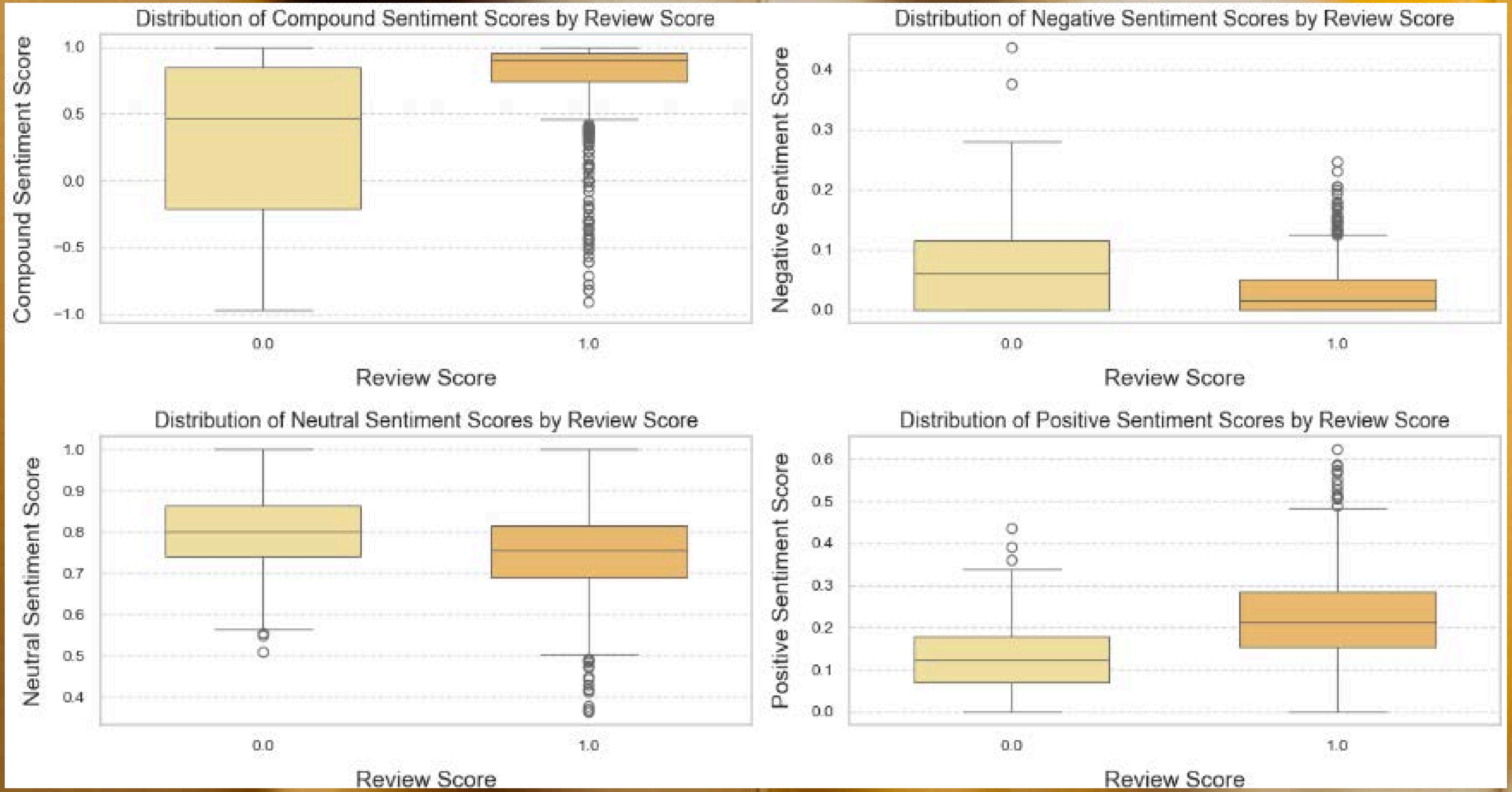
Review Rating Distrbution

- Majority are positive 83%
- 9.7% Negative
- Smallest fraction of 7.2% neutrals



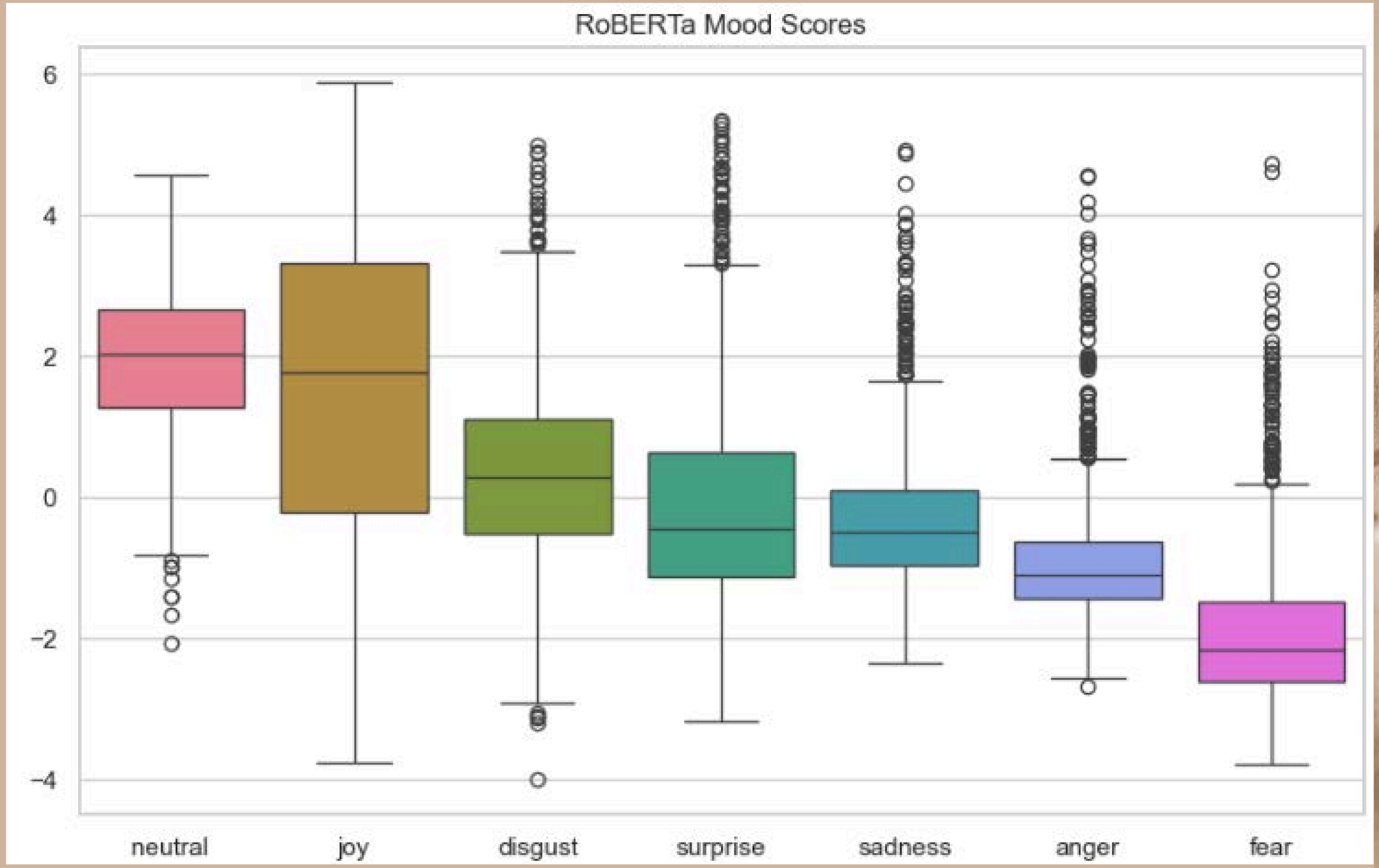
Distribution of sentiment scores by is_recommended

- Overall higher score for recommendation (Compound & Positive)
- Higher score in Non_recommended (Negative & Neutral)



Sentiment Analysis RoBERTa

- Identifying specific consumers' emotion and mood detection
- Neutral and Joy highest median illustrating customer satisfaction



Challenges & Solutions

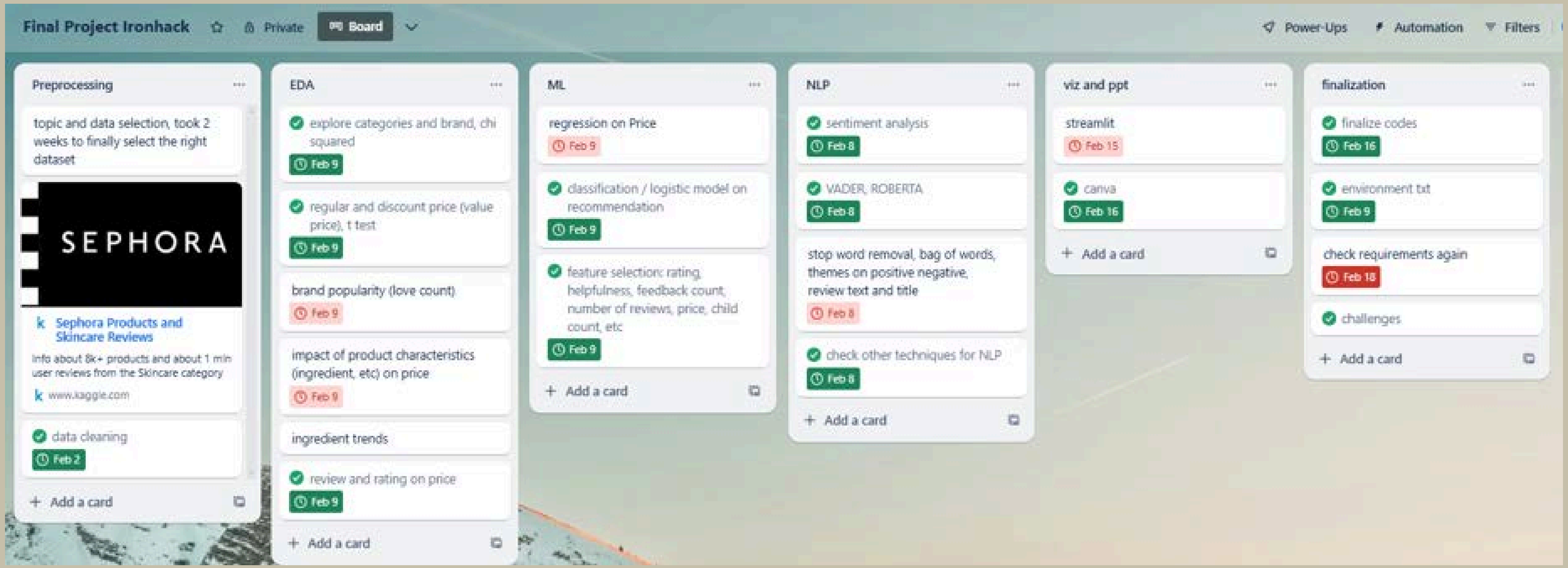
Challenges

- Large datasize long processing time
- Memory error
- High-dimensionality

Solution

- Smaller sample size
- Textblob for encoding

Planning



Final Project Ironhack

Preprocessing

- topic and data selection, took 2 weeks to finally select the right dataset
- SEPHORA
Sephora Products and Skincare Reviews
Info about 8k+ products and about 1 mln user reviews from the Skincare category
www.kaggle.com
- ✓ data cleaning — Feb 2

EDA

- ✓ explore categories and brand, chi squared — Feb 9
- ✓ regular and discount price (value price), t test — Feb 9
- brand popularity (love count) — Feb 9
- impact of product characteristics (ingredient, etc) on price — Feb 9
- ingredient trends
- ✓ review and rating on price — Feb 9

ML

- regression on Price — Feb 9
- ✓ classification / logistic model on recommendation — Feb 9
- ✓ feature selection: rating, helpfulness, feedback count, number of reviews, price, child count, etc — Feb 9

NLP

- ✓ sentiment analysis — Feb 8
- ✓ VADER, ROBERTA — Feb 8
- stop word removal, bag of words, themes on positive negative, review text and title — Feb 8
- ✓ check other techniques for NLP — Feb 8

viz and ppt

- streamlit — Feb 15
- ✓ canva — Feb 16

finalization

- ✓ finalize codes — Feb 16
- ✓ environment txt — Feb 9
- check requirements again — Feb 18
- ✓ challenges

Thank you