Final Exploratory Analysis Classification Report

Lori Fang, Trevor Kwan

This report is based on removing all sample-takers that did not convert into paying customers in the most recent 46 days of data. This is because there exists a time period between the time sample-takers take the sample and make their first purchase, so our model would be inaccurate if we evaluated whether or not a sample-taker would purchase without giving them a reasonable time period to try the sample and make a purchase. We selected 46 days because that is how long it takes for sample-takers to become paying customers on average.

After this, our data consisted of 593,404 sample-takers.

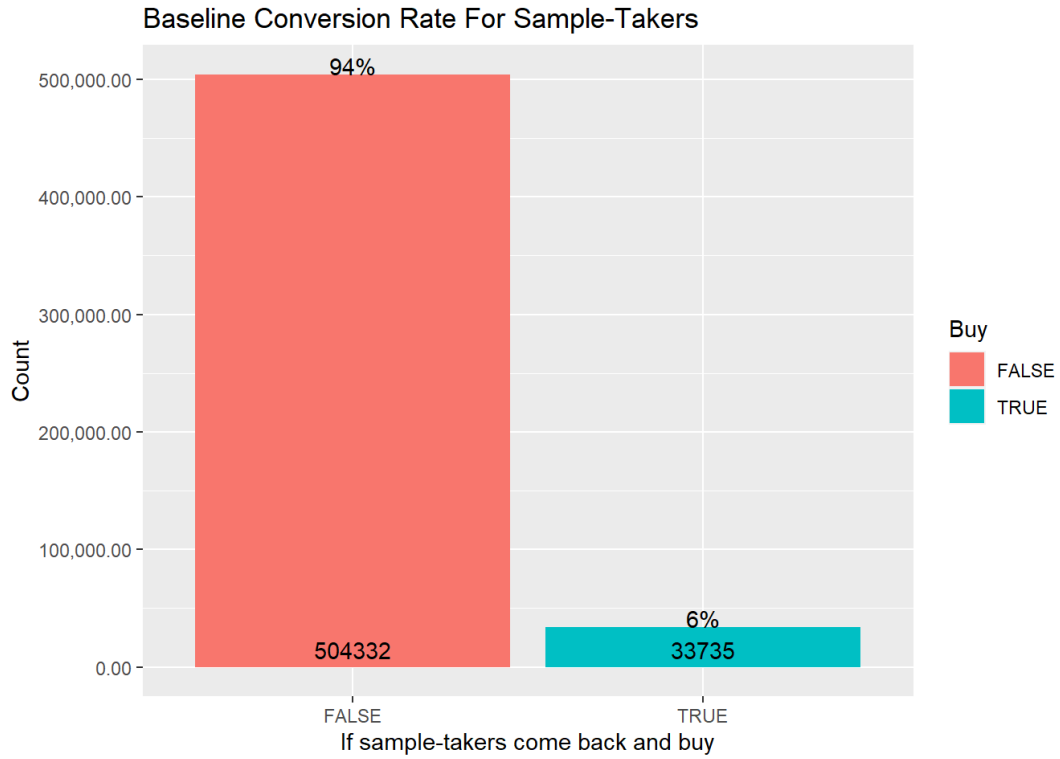
Data overview

customer_id	accepts_marketing	ordered_month	location	gender	free_shipping	product_type	skin_type	fv_site	buy
510336833	TRUE	4	ONTARIO, CANADA	unknown	TRUE	Anti-Aging	Normal to Dry	other	FALSE
544818881	FALSE	6	BRITISH COLUMBIA, CANADA	unknown	FALSE	Other	Normal to Oily	other	TRUE
577330433	FALSE	5	OREGON, UNITED STATES	female	FALSE	Other	Normal to Dry	other	TRUE
584222401	FALSE	5	NEW JERSEY, UNITED STATES	unknown	FALSE	Other	Normal to Oily	other	FALSE
593077569	FALSE	5	ONTARIO, CANADA	unknown	FALSE	Other	Normal to Dry	other	FALSE
599767169	FALSE	5	QUEENSLAND, AUSTRALIA	female	FALSE	Other	Normal to Dry	other	FALSE

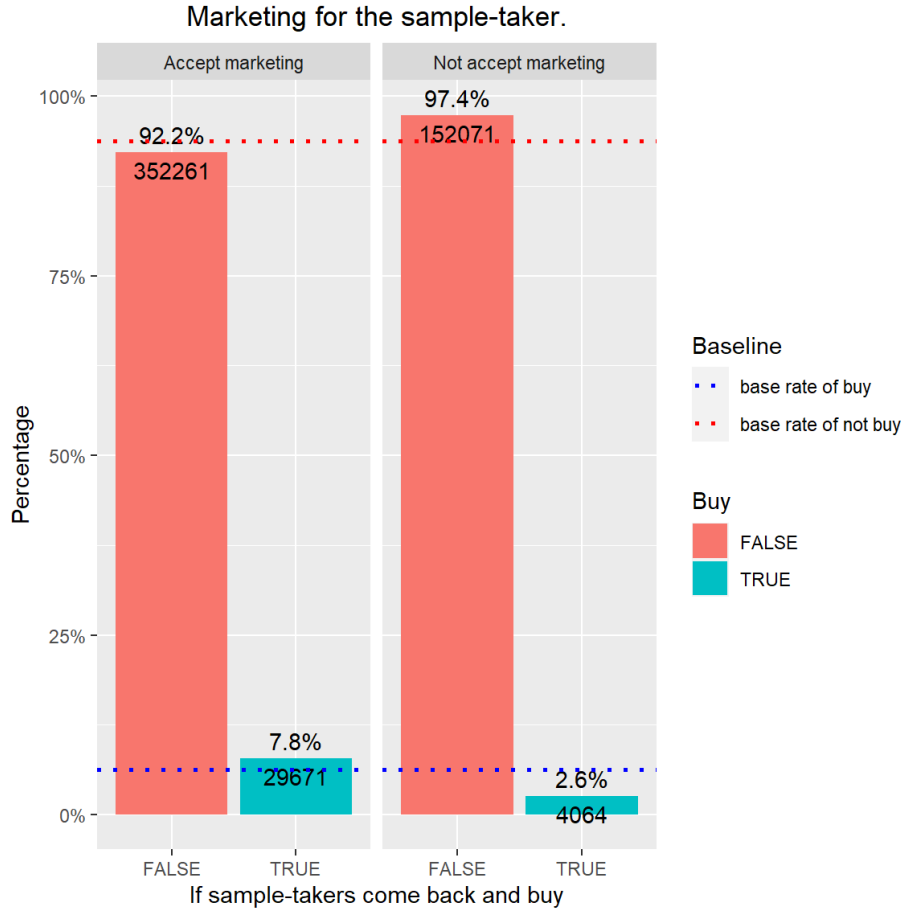
Exploratory Feature Analysis

Which features are important/contribute the most to whether or not a sample-taker will buy?

- Null hypothesis(H0): There is no difference in proportion of buyers(returned customers).
- Alternative hypothesis(Ha): There is a significant difference in proportion of buyers(returned customers).

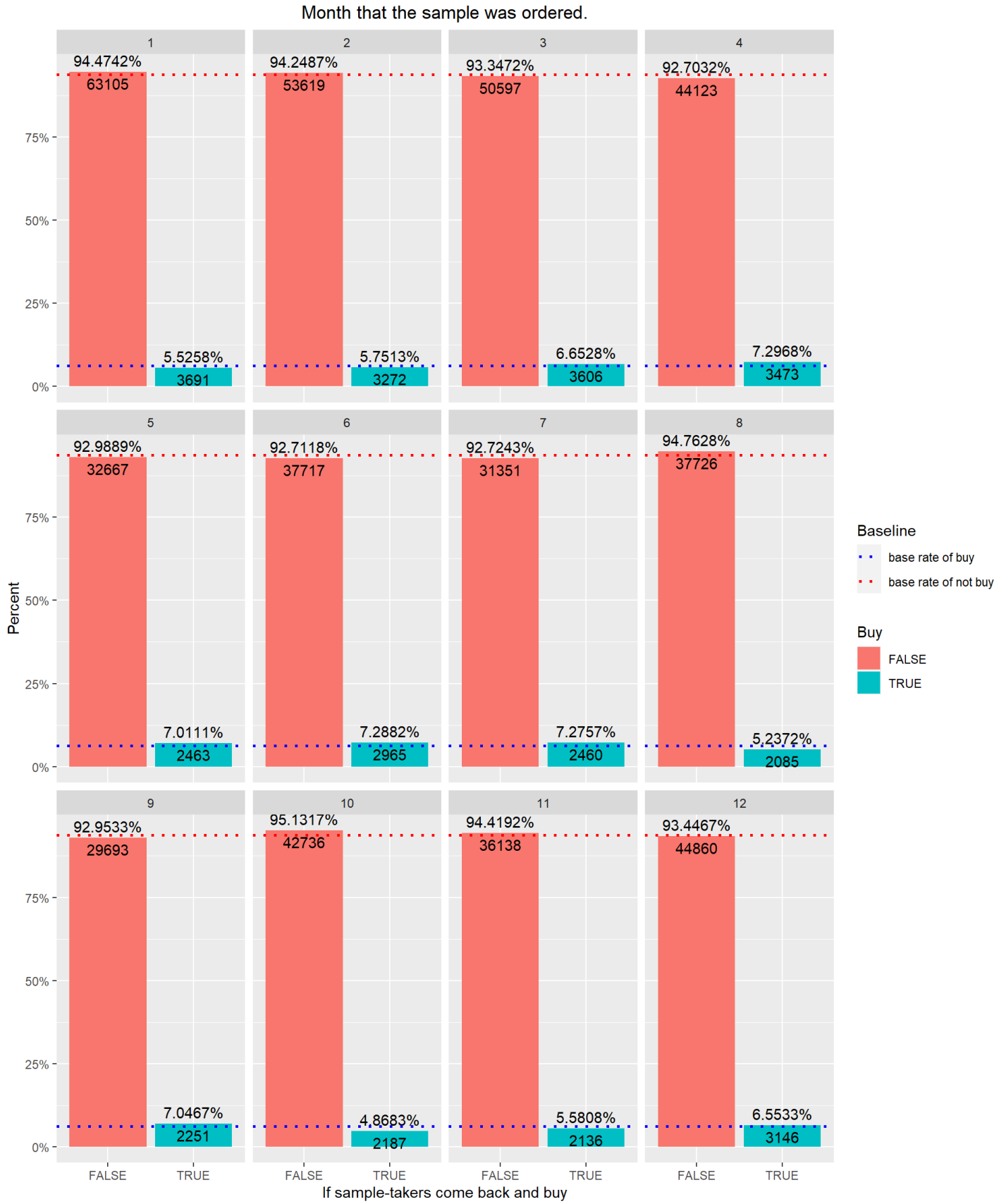


1. accepts_marketing vs buy



Sample-takers that accepted marketing had higher conversion rates than those that did not. And 72% of all sample-takers accepted marketing while the other 28% did not. With a p-value of 0, there is a significant difference in conversion rate between sample-takers that accept marketing and sample-takers that don't.

2. ordered_month vs buy

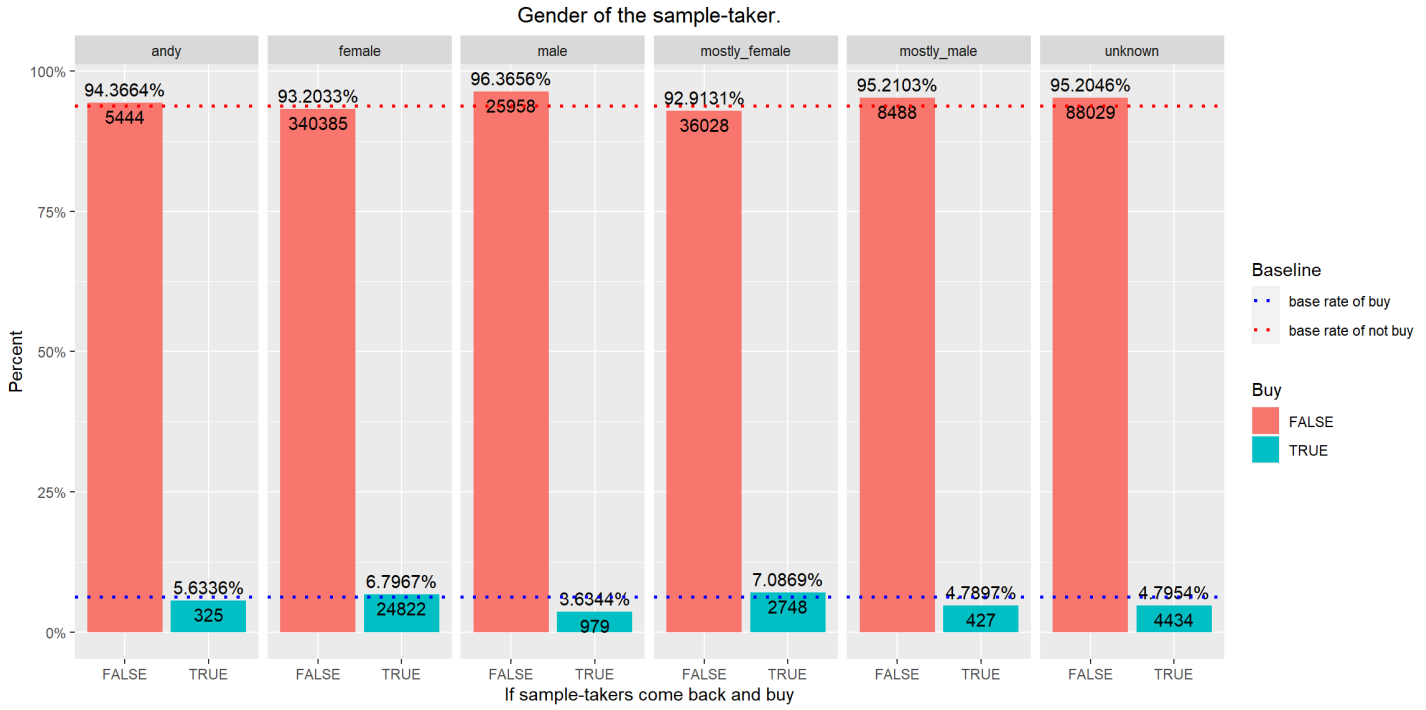


	1	2	3	4	5	6	7	8	9	10	11
2	1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
3	**0**	**0.00**	NA	NA	NA	NA	NA	NA	NA	NA	NA

	1	2	3	4	5	6	7	8	9	10	11
4	**0**	**0.00**	**0.00**	NA	NA	NA	NA	NA	NA	NA	NA
5	**0**	**0.00**	1.00	1	NA	NA	NA	NA	NA	NA	NA
6	**0**	**0.00**	**0.01**	1	1.00	NA	NA	NA	NA	NA	NA
7	**0**	**0.00**	**0.03**	1	1.00	1	NA	NA	NA	NA	NA
8	1	**0.04**	**0.00**	**0**	**0.00**	**0**	**0**	NA	NA	NA	NA
9	**0**	**0.00**	1.00	1	1.00	1	1	**0.00**	NA	NA	NA
10	**0**	**0.00**	**0.00**	**0**	**0.00**	**0**	**0**	0.99	**0.00**	NA	NA
11	1	1.00	**0.00**	**0**	**0.00**	**0**	**0**	1.00	**0.00**	**0**	NA
12	**0**	**0.00**	1.00	**0**	0.64	**0**	**0**	**0.00**	0.45	**0**	**0**

Sample-takers that ordered in October have the lowest conversion rate, and 8% of all sample-takers ordered in October.

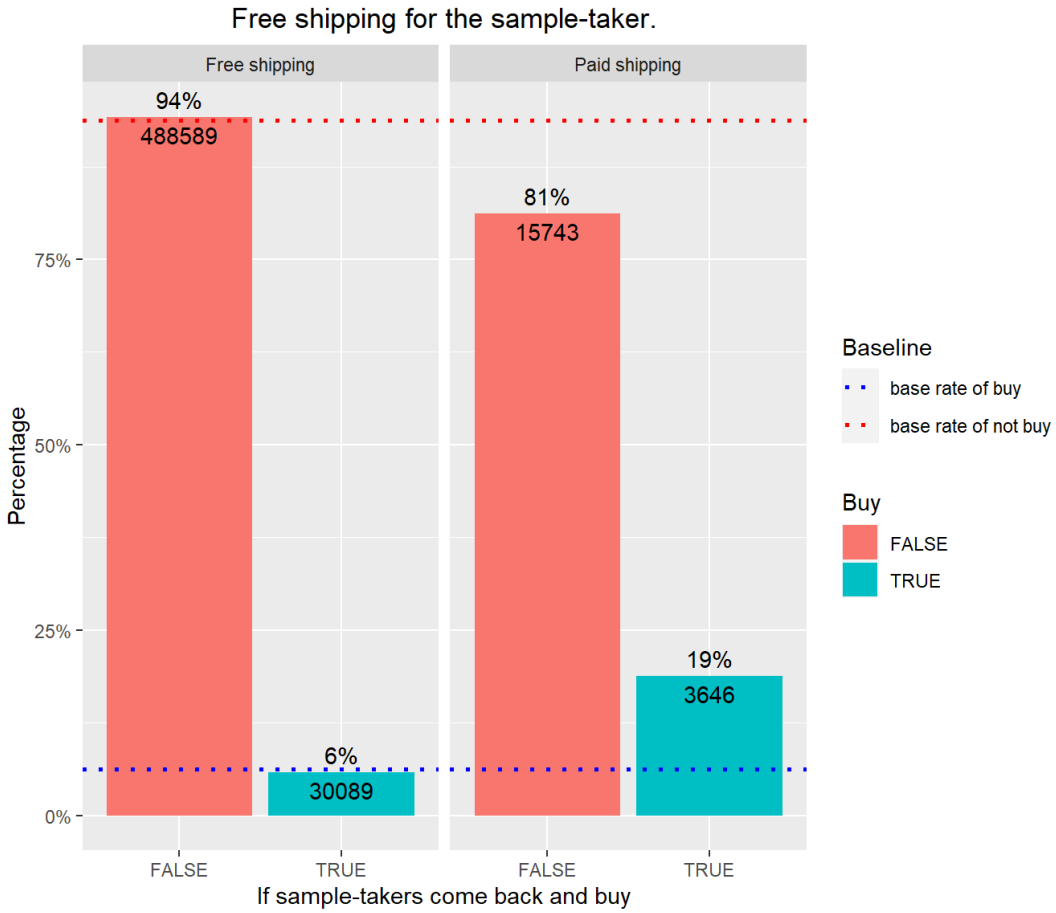
3. gender vs buy



	andy	female	male	mostly_female	mostly_male
female	**0.01**	NA	NA	NA	NA
male	**0.00**	**0.00**	NA	NA	NA
mostly_female	**0.00**	0.48	**0**	NA	NA
mostly_male	0.39	**0.00**	**0**	**0**	NA
unknown	0.07	**0.00**	**0**	**0**	1

“Andy” represents androgynous gender names that could be classified as either male or female. Female have higher conversion rates than males, and 74% of all sample-takers had female names.

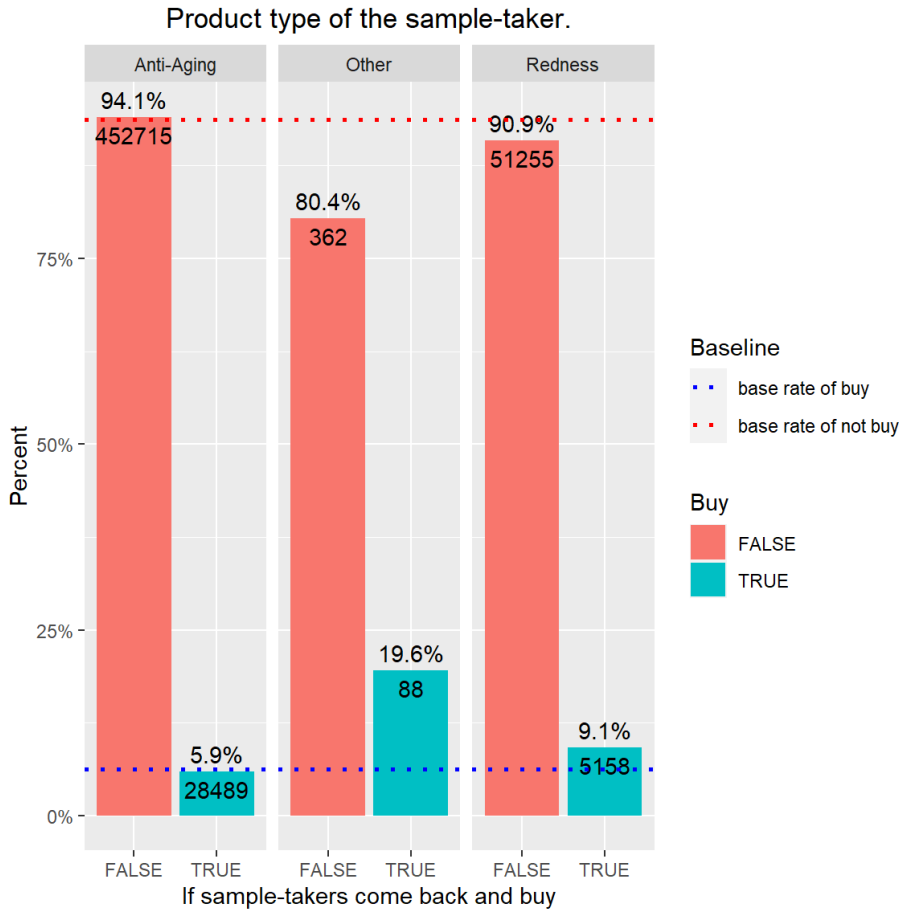
4. free_shipping vs buy



Sample-takers that paid for shipping had higher conversion rates than sample-takers that had free shipping. Sample-takers that paid for shipping made up of 3.6% of all sample-takers, while sample-takers that had free shipping made up the other 96.4% of sample-takers.

With a p-value of 0, there is a significant difference in conversion rate between sample-takers that had free shipping and sample-takers that don't.

5. product_type vs buy

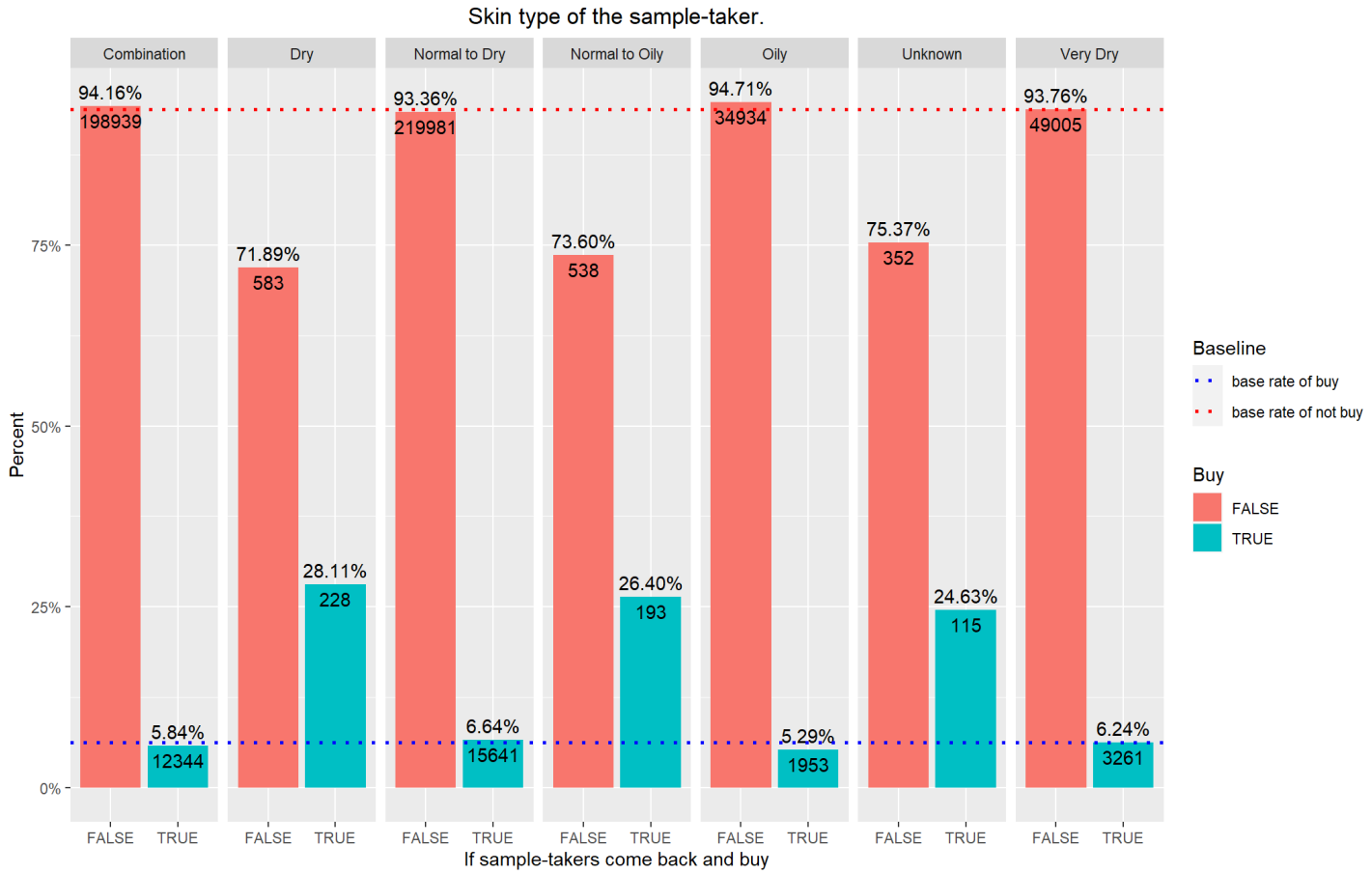


	Anti-Aging	Other
Other	**0**	NA
Redness	**0**	**0**

Sample-takers who ordered redness samples had higher conversion rates than sample-takers who ordered anti-aging samples. Anti-aging samples made up of 89% of all samples ordered and redness samples made up of 10% of all samples ordered.

There is a significant difference in conversion rate between sample-takers that ordered redness samples and sample-takers that ordered anti-aging samples.

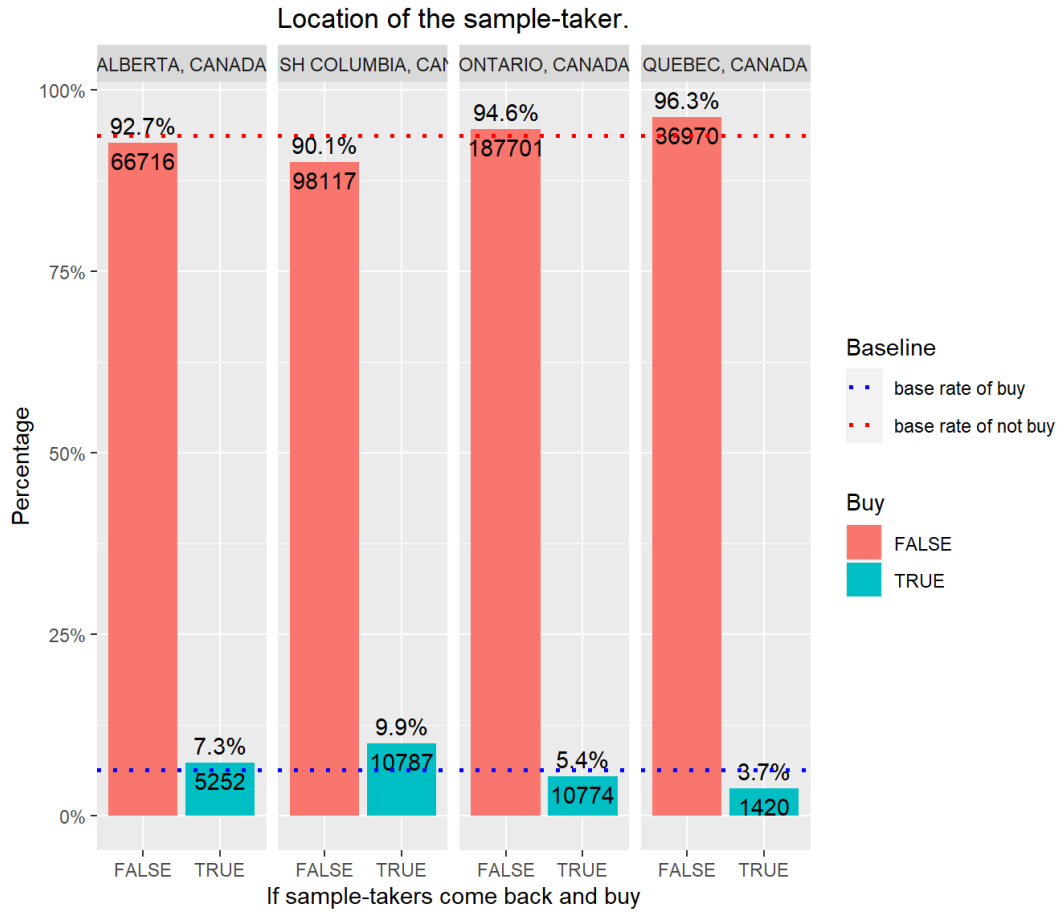
6. skin_type vs buy



	Combination	Dry	Normal to Dry	Normal to Oily	Oily	Unknown
Dry	**0.00**	NA	NA	NA	NA	NA
Normal to Dry	**0.00**	**0**	NA	NA	NA	NA
Normal to Oily	**0.00**	1	**0.00**	NA	NA	NA
Oily	**0.00**	**0**	**0.00**	**0**	NA	NA
Unknown	**0.00**	1	**0.00**	1	**0**	NA
Very Dry	**0.01**	**0**	**0.02**	**0**	**0**	**0**

Sample-takers with “dry” or “normal to oily” skin types had higher conversion rates than sample-takers with other skin types, but this could be due to chance since majority of sample-takers had a “normal to dry” or a “combination” skin type, with the total proportion of sample-takers making up “dry” and “normal to oily” skin types being 0.29% of all sample-takers.

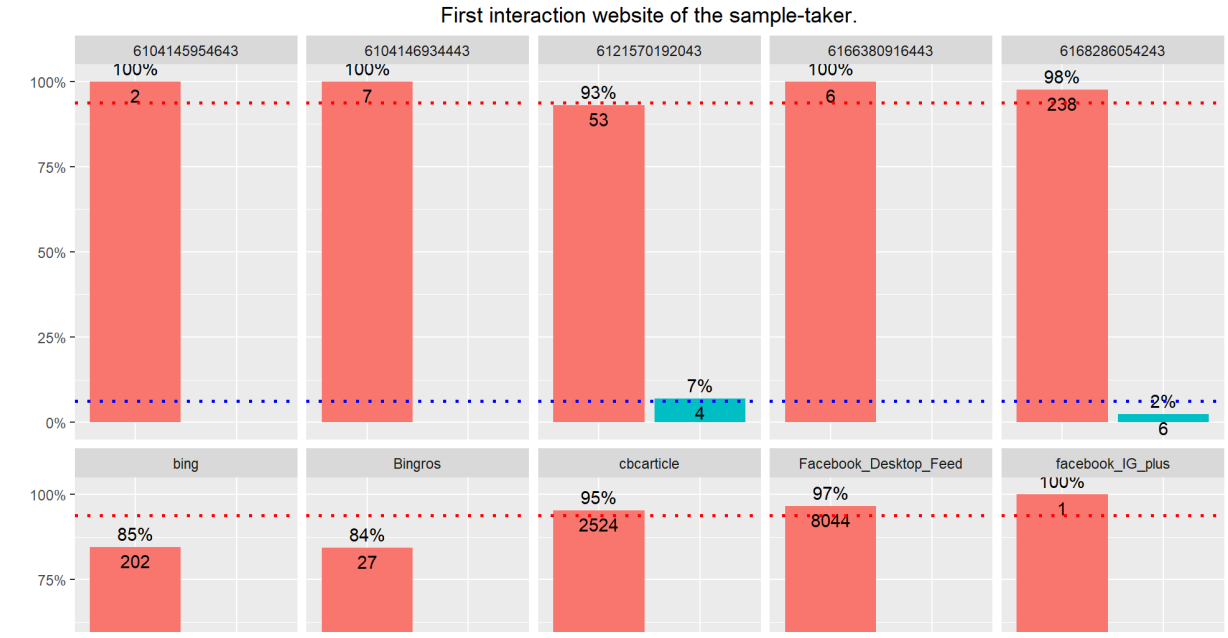
7. location vs buy



	ALBERTA, CANADA	BRITISH COLUMBIA, CANADA	ONTARIO, CANADA
BRITISH COLUMBIA, CANADA	**0**	NA	NA
ONTARIO, CANADA	**0**	**0**	NA
QUEBEC, CANADA	**0**	**0**	**0**

Sample-takers were most likely to come from British Columbia making up 20% of all sample-takers, and sample-takers from there had relatively high conversion rates.

8. fv_site vs buy



"Bing", "Bingros", and "googleshopping" also displayed higher conversion rates, and 260 sample-takers came from "bing", 38 sample-takers came from "Bingros", and 1315 sample-takers came from "googleshopping".

Ideal/Unideal Customer Features

The following analysis is of the top 4 locations, the 4 locations containing the largest amount of sample-takers.

"Baseline" represents the average conversion from the top 4 locations: British Columbia, Alberta, Ontario, and Quebec, regardless of whether or not sample-takers accepted marketing or had free shipping.

Ideal Customer From Top 4 Locations

Ideal Customer - conversion rate of sample-takers who accepted marketing and did not have free shipping.

In top 4 locations, especially British Columbia, sample-takers that accepted marketing and didn't have free shipping displayed higher conversion rates.

However, only 1.1% of sample-takers from top 4 locations accepted marketing and don't have free shipping.

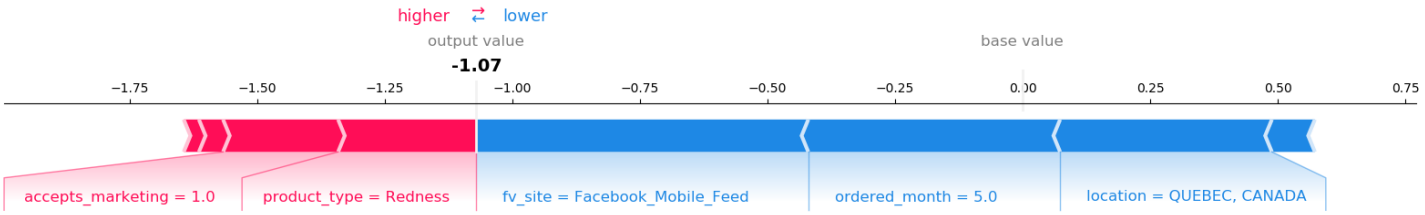
Unideal Customer From Top 4 Locations

Unideal Customer - conversion rate of sample-takers who did not accept marketing and had free shipping.

In top 4 locations, sample-takers that did not accept marketing and did not have free shipping displayed lower conversion rates.

And 26.7% sample-takers from top 4 locations did not accept marketing and did not have free shipping.

Overall Important Customer Features



Features that influenced whether or not a sample-taker will buy.

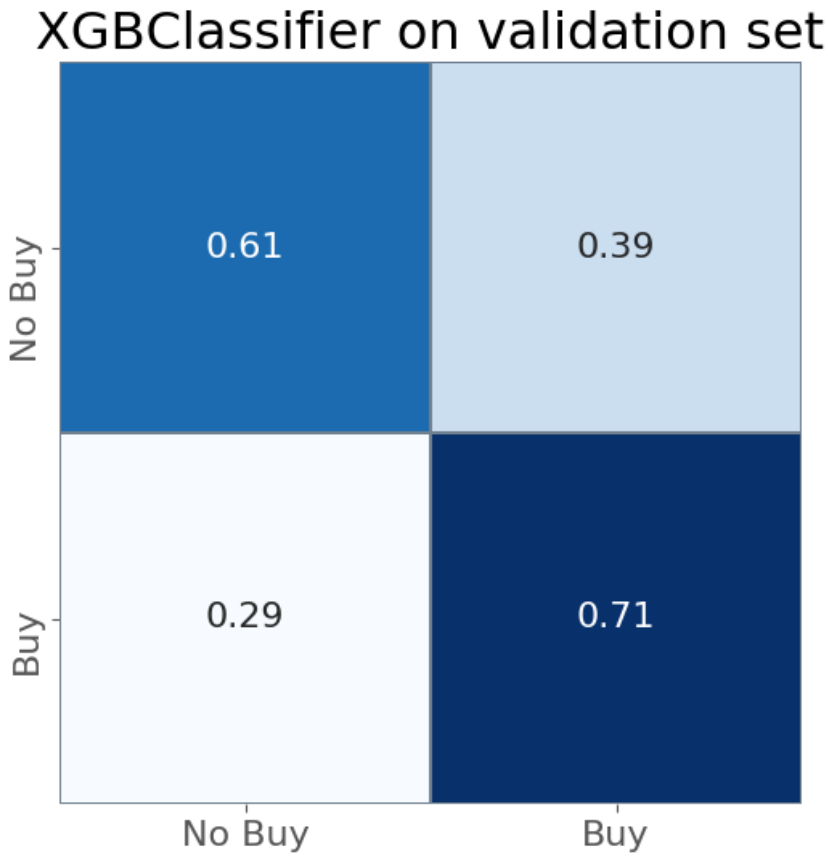
Certain features push the model to predict a higher output value, meaning a sample-taker with those features would be more likley to buy, and certain features push the model to predict a lower output value, meaning the sample-taker with those features would be less likely to buy.

Here, SHAP analysis suggests that sample-takers that accepted marketing and ordered a sample type “Redness” were more likely to purchase, and sample-takers that had the first interaction site of “Facebook_Mobile_Feed”, ordered their sample in the month of May, and ordered from the location Quebec were less likely to purchase.

Classification Model Results

Eight features were selected to optimize model prediction. Our model provides users with the option to tune the parameter “days”, which represents the number of days counting back from the most recent date and drops all sample-takers that haven’t purchased yet and fall in that “days” range.

The following plots show results where “days” = 46, which drops sample-takers that have not purchased in the most recent 46 days.



Confusion matrix - x-axis is what the model predicted the sample-taker would do, y-axis is what the sample-taker actually did.

Label	Precision	Recall	F1-score	Support
not_buy	0.9684593	0.6116720	0.7497851	161840
buy	0.1100302	0.7067491	0.1904155	10994

Our classification model yields a 11% precision score and a 70% recall score. A 11% precision implies that out of all sample-takers that our model *predicts* will “buy”, 11% of them will actually “buy”. A 70% recall implies that out of all sample-takers that *do* actually “buy”, 70% of them will be accurately detected by our model.

Our model is doing an acceptable job, but is very limited by the quality and number of features provided.

Limitations of the Classification Model

To achieve higher accuracy, precision, and recall of the model in the future, the quality of features and the number of features provided to the model could be improved. Some features to be added to improve model predictions include a sample-taker’s age or income. Features such as gender, skin_type, and product_type were extracted from the given data and thus were of lower quality. The quality of these features could be improved for better model performance.