A/B Hypothesis Testing: Ad campaign performance

INTRODUCTION

- SmartAd is a mobile first advertiser agency. It designs Intuitive touch-enabled advertising.

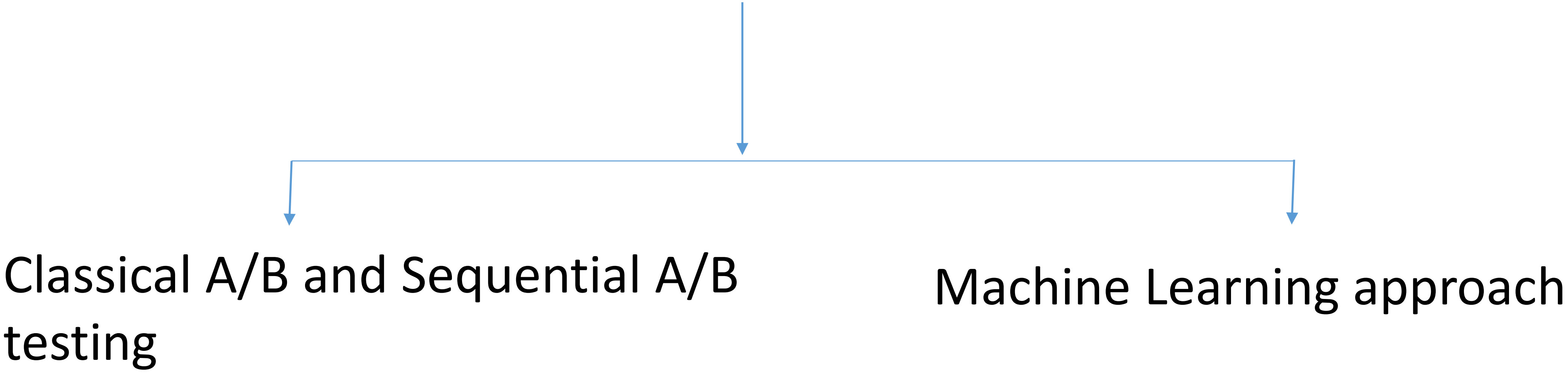
- SmartAd provides an additional service called Brand Impact Optimiser (BIO), a lightweight questionnaire, served with every campaign to determine the impact of the creative, the ad they design, on various upper funnel metrics, including memorability and brand sentiment.

- An hypothesis as formulated to investigate the influence of a new design

- The main objective is to validate the hypothesis algorithm you built

METHODOLOGY

- The method of testing the algorithm are divided into two methods



Classical A/B and Sequential A/B testing

Machine Learning approach

METHODS: CLASSICAL A/B TESTING

- Classical A/B testing involves comparing two versions of an idea to see which performs better.
- These variations, known as A and B, are presented randomly to users. A portion of them will be directed to the first version, and the rest to the second.
- A statistical analysis of the results then determines which version, A or B, performed better, according to certain predefined indicators such as conversion rate.
- The classic A/B test presents users with two variations of the proposed idea at similar conditions. That way, you can compare two or several variations of the same element. The Z-test is used to investigate the significant difference between the A and B.

METHOD: SEQUENTIAL A/B TESTING

- A common issue with classical A/B-tests, especially when you want to be able to detect small differences, is that the sample size needed can be prohibitively large.
- In many cases it can take several weeks, months or even years to collect enough data to conclude a test.
- Start your experiment with choosing sample size, let’s call it N; randomly assign variations under test to the treatment and control, with 50% probability each.
- Track the number of incoming successes for both variations. Let’s refer to the conversion rate of treatment variation as T, and CR of control as C.
- It’s necessary to finish the test when T−C reaches √2N and declare the treatment variation to be the winner of your A/B test. It’s necessary to finish the test when T+C reaches N. In such case, declare that the experiment had no winner.

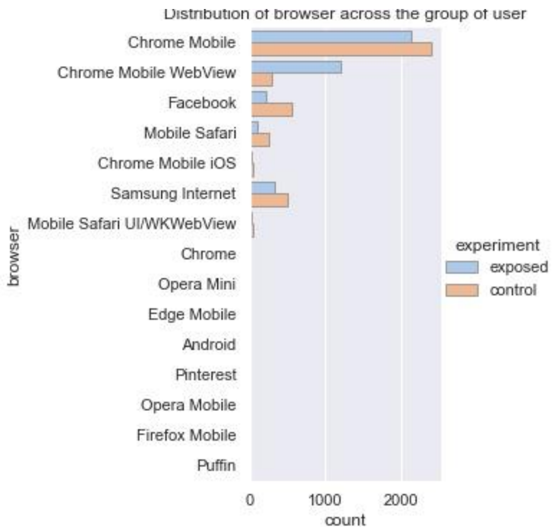
METHOD: MACHINE LEARNING

- Machine learning algorithm are used for investigating the hypothesis
- The algorithm employed are decision tree, xgboost, logistic Regression
- A 5 folds cross validation was used in grid search to simultaneously obtain a good fit and tune the hyper parameters
- Machine learning algorithms are used to optimize the investigation process.
- Machine learning algorithms can also take into consideration other features of the users.
- The algorithms are used to train a sub data. These models are then used to predict the interactions of the rest of the data, these predictions are compared with the actual interactions using certain metrics.
- The metrics gives the level of significance of the experimental feature

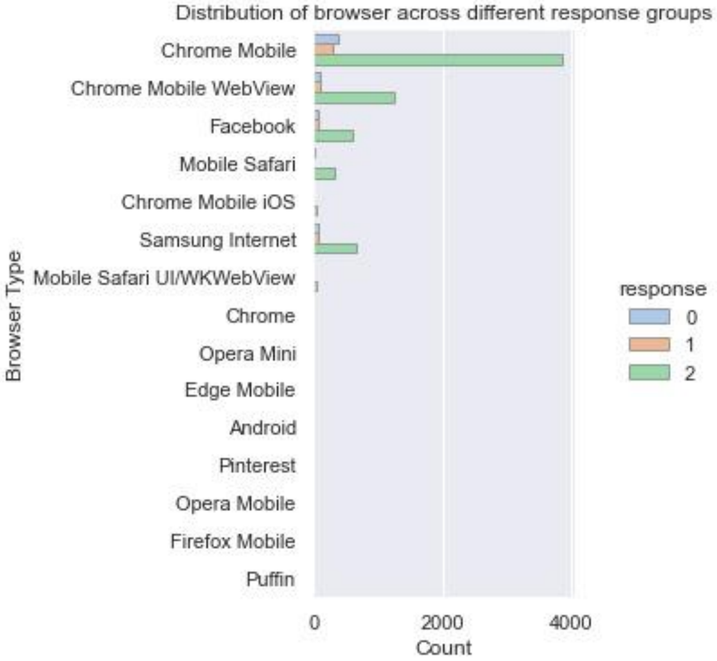
DATA

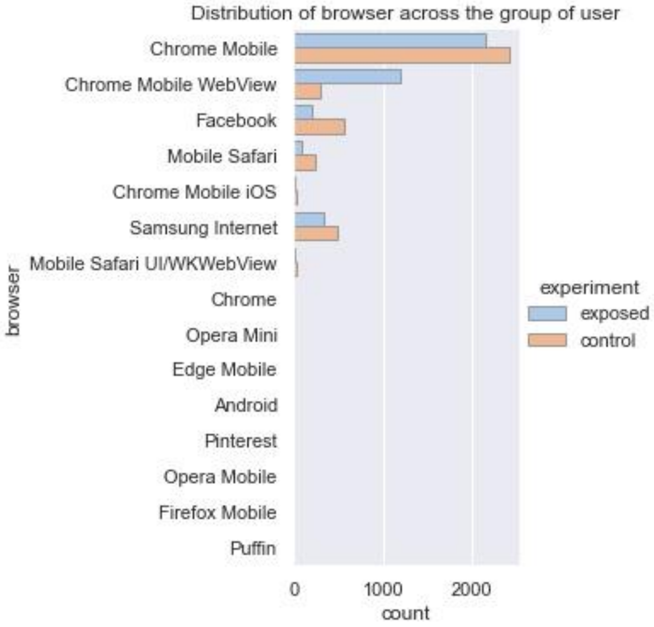
- The data collected for this challenge has the following columns
- **auction_id:** the unique id of the online user who has been presented the BIO. In standard terminologies this is called an impression id. The user may see the BIO questionnaire but choose not to respond. In that case both the yes and no columns are zero.
- **experiment** : which group the user belongs to - control or exposed.
- **date** : the date in YYYY-MM-DD format. The date range from 3 to 10 July 2020
- **hour** : the hour of the day in HH format.
- **device_make** : the name of the type of device the user has e.g. Samsung
- **platform_os:** the id of the OS the user has.
- **browser** : the name of the browser the user uses to see the BIO questionnaire.
- **yes** : 1 if the user chooses the “Yes” radio button for the BIO questionnaire.
- **no** : 1 if the user chooses the “No” radio button for the BIO questionnaire.
- **Response**: This is a derived feature from the yes and no feature

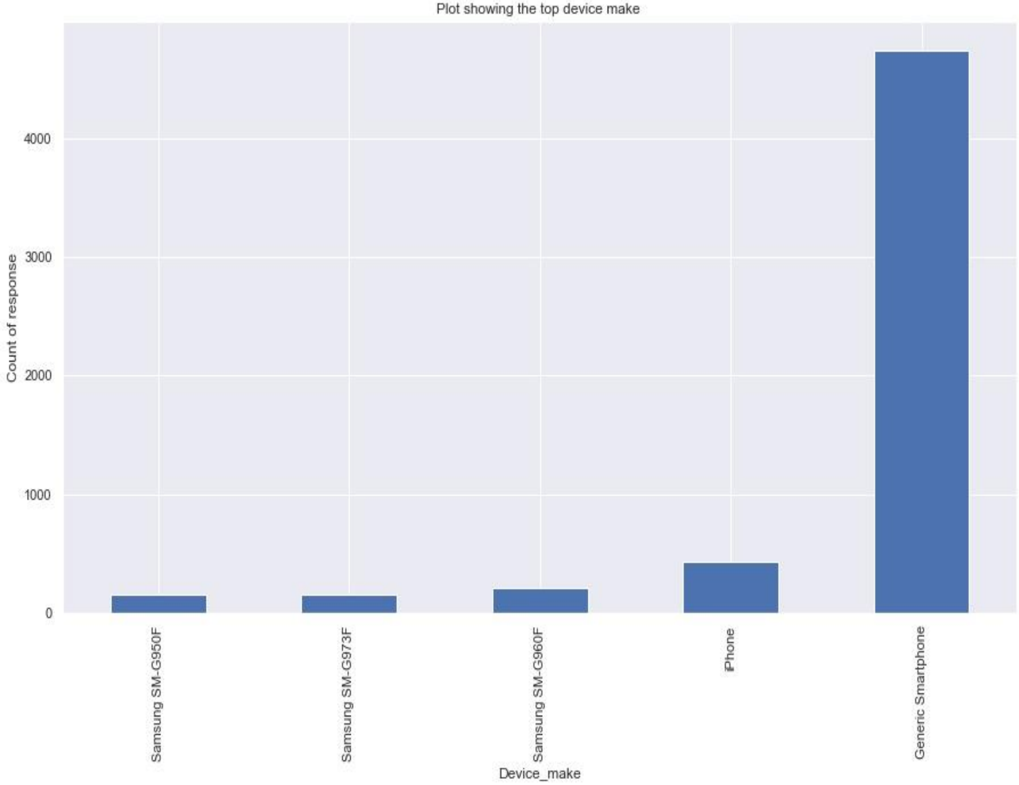
DATA



The plots shows the distribution of the data across features, it can be inferred that a lot of the response were from users that didn’t answer the questionnaire







RESULT: CLASSICAL A/B TESTING

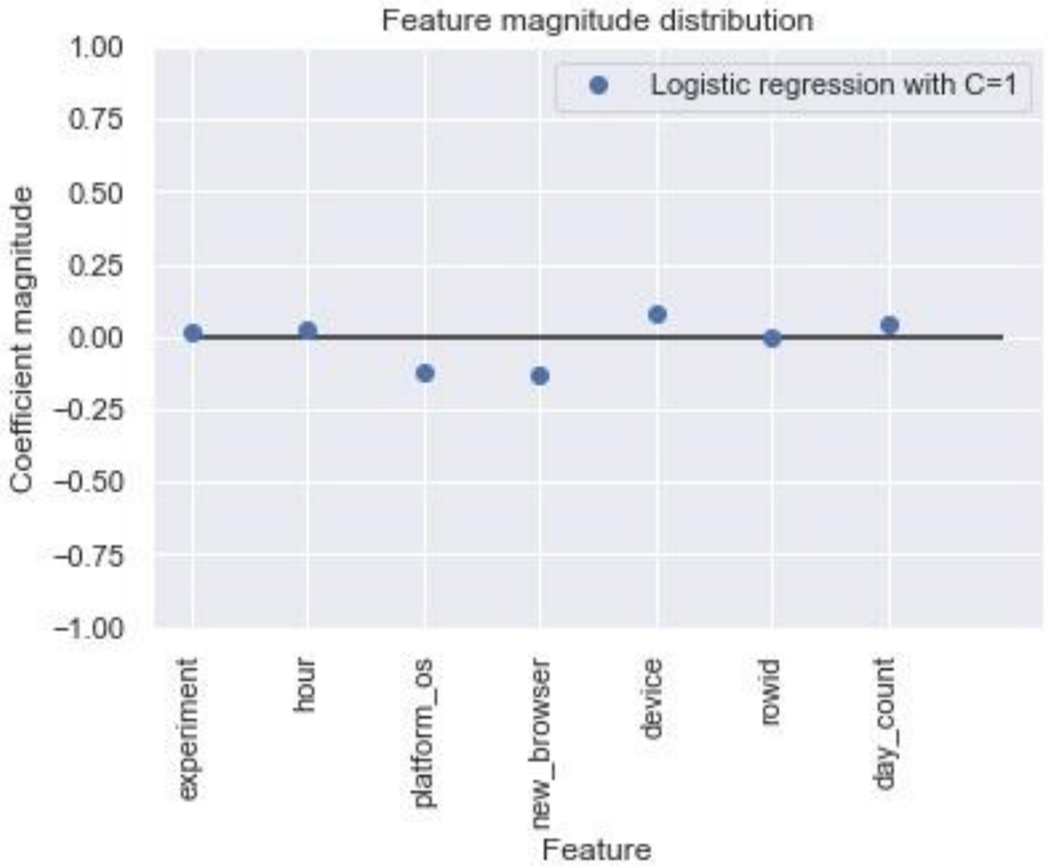
- Based on the model used, the required sample size per class is 38932
- Split of control users who saw dummy vs treatment users who saw new design: 47.14 % 52.86 %
- Number of control users: 586
- Percentage of control users who responded: 45.05 %
- Number of treatment users: 657
- Percentage of treatment users who responded: 46.88 %
- The number of data in the groups is below the optimum size

RESULT: CLASSICAL A/B TESTING

- The lower bound of the confidence interval is -3.72 %
- The upper bound of the confidence interval is 7.38 %
- Investigation shows we should not reject the null hypothesis

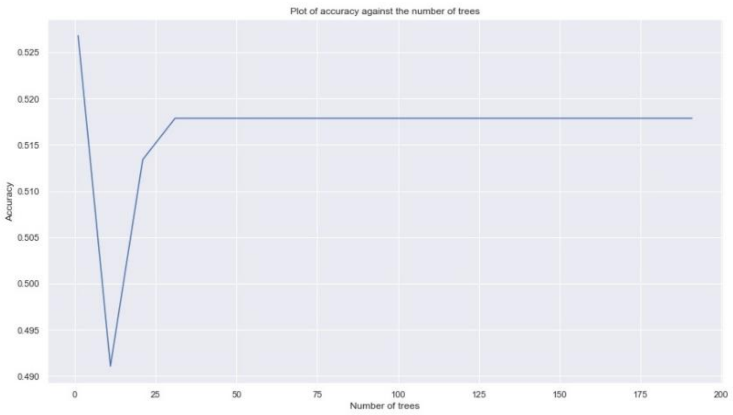
RESULTS: MACHINE LEARNING(LOGISTIC REGRESSION)

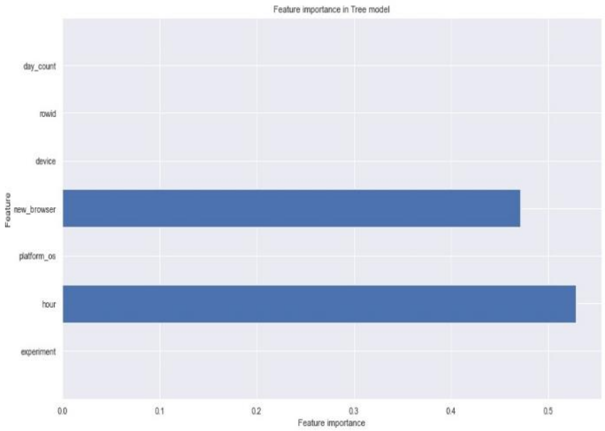
- This model gives an accuracy of about 51.3%
- The model plot shows that device, hour and experiment are important features

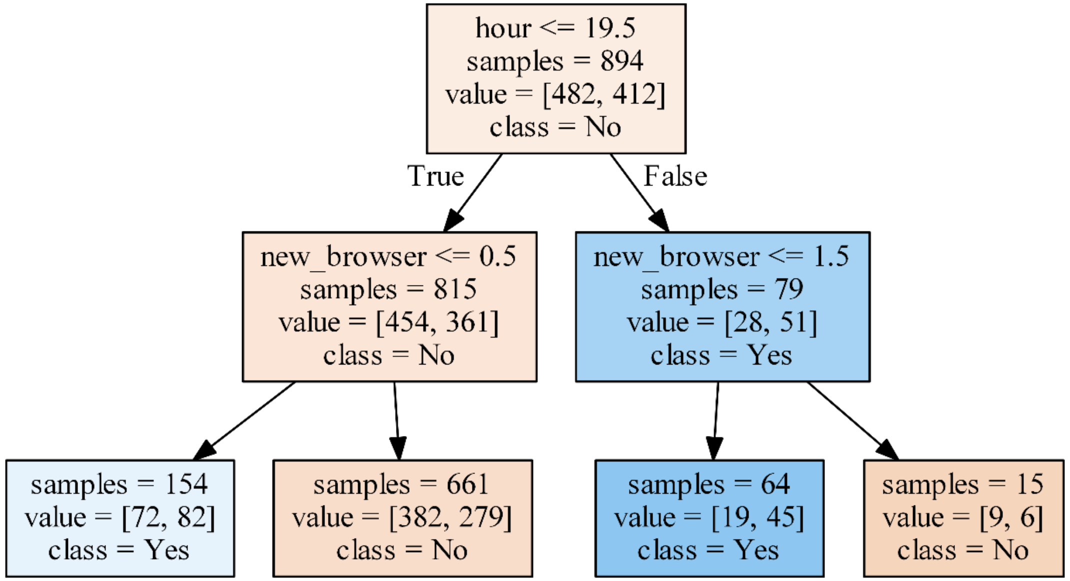


RESULTS: MACHINE LEARNING(DECISION

- This model has a similar accuracy (52.2 percent) to logistic regression model
- The first plot shows the maxim accuracy at a tree_depth of 2.
- The second and third plot emphazize on the importance of some features in separating the classes
- The feature importance highlight browser type and the hour as the important feature in this model







hour <= 19.5
samples = 894
value = [482, 412]
class = No

True

False

new_browser <= 0.5
samples = 815
value = [454, 361]
class = No

new_browser <= 1.5
samples = 79
value = [28, 51]
class = Yes

samples = 154
value = [72, 82]
class = Yes

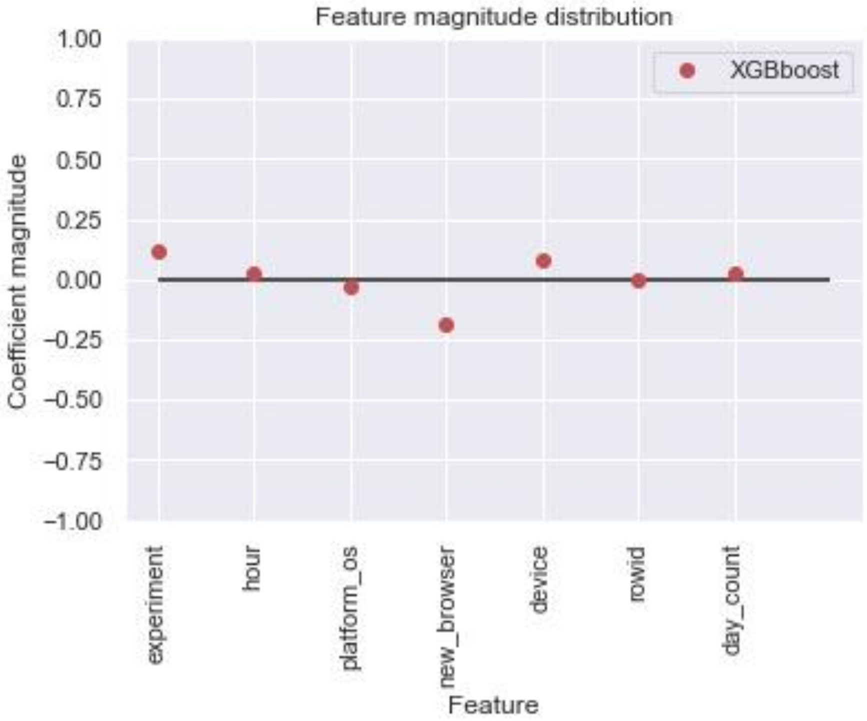
samples = 661
value = [382, 279]
class = No

samples = 64
value = [19, 45]
class = Yes

samples = 15
value = [9, 6]
class = No

RESULTS: MACHINE LEARNING(XGBOOST)

- The accuracy for this model is about 51.3%
- It uses the experiment and device efficiently in modelling the data



COMPARISON OF THE TWO METHODS

- The classical A/B method was able to predict a significant different between the exposed and control groups, but this could be due to chance, because the optimal number of sample size was not obtained
- The Machine Learning approach gives a more detailed explanation of the features that are defines each groups
- The algorithms was able to separate the two groups and classified them based on response

CLASSIFICATION REPORT

- The tables shows the classification report on the model evaluation on the test set
- The Xgboost model had the best performance in terms of accuracy, precision and recall

Logistic Regression report

	no	yes	accuracy	macro avg	**weighted avg**
precision	0.519231	0.142857	0.456	0.331044	**0.359648**
recall	0.75	0.056604	0.456	0.403302	**0.456**
f1-score	0.613636	0.081081	0.456	0.347359	**0.387833**
support	72	53	0.456	125	**125**

Decision Tree report

	no	yes	accuracy	macro avg	**weighted avg**
precision	0.535354	0.269231	0.48	0.402292	**0.422517**
recall	0.736111	0.132075	0.48	0.434093	**0.48**
f1-score	0.619883	0.177215	0.48	0.398549	**0.432192**
support	72	53	0.48	125	**125**

Xgboost report

	no	yes	accuracy	macro avg	**weighted avg**
precision	0.54902	0.304348	0.504	0.426684	**0.445279**
recall	0.777778	0.132075	0.504	0.454927	**0.504**
f1-score	0.643678	0.184211	0.504	0.413944	**0.448864**
support	72	53	0.504	125	**125**

RECOMMENDATION AND OUTCOMES

- It can be concluded that machine learning approach was better suited for the investigation
- For better performance with the classical A/B testing approach, more data need to be obtained
- There is a significant difference between the exposed and control users
- The investigation was successful, although there is need for more data gathering
- Conclusions about the effectiveness of the ads can only be reliably obtained from the machine learning approach due to the limitation of data sample size and duration of training.

LIMITATIONS OF THE ANALYSIS

- The data was AGGREGATED - To truly understand customer behavior, we should run the analysis on unaggregated data to determine probability of an individual customer enrolling.
- There are no features related to the Customer in the data set - The customer journey and their characteristics are incredibly important to understanding complex purchasing behavior. Including good features is the best way to improving model performance, and thus insights into customer behavior.
- This study do not reveal if the product needs overhauling or a total change
- This study only reveals which is better, it does not inform on how to improve the product
- The methods solely depends on the user responses, humans change!

REFERENCES

- https://splitmetrics.com/blog/splitmetrics-sequential-ab-testing/
- https://www.business-science.io/business/2019/03/11/ab-testing-machine-learning.html