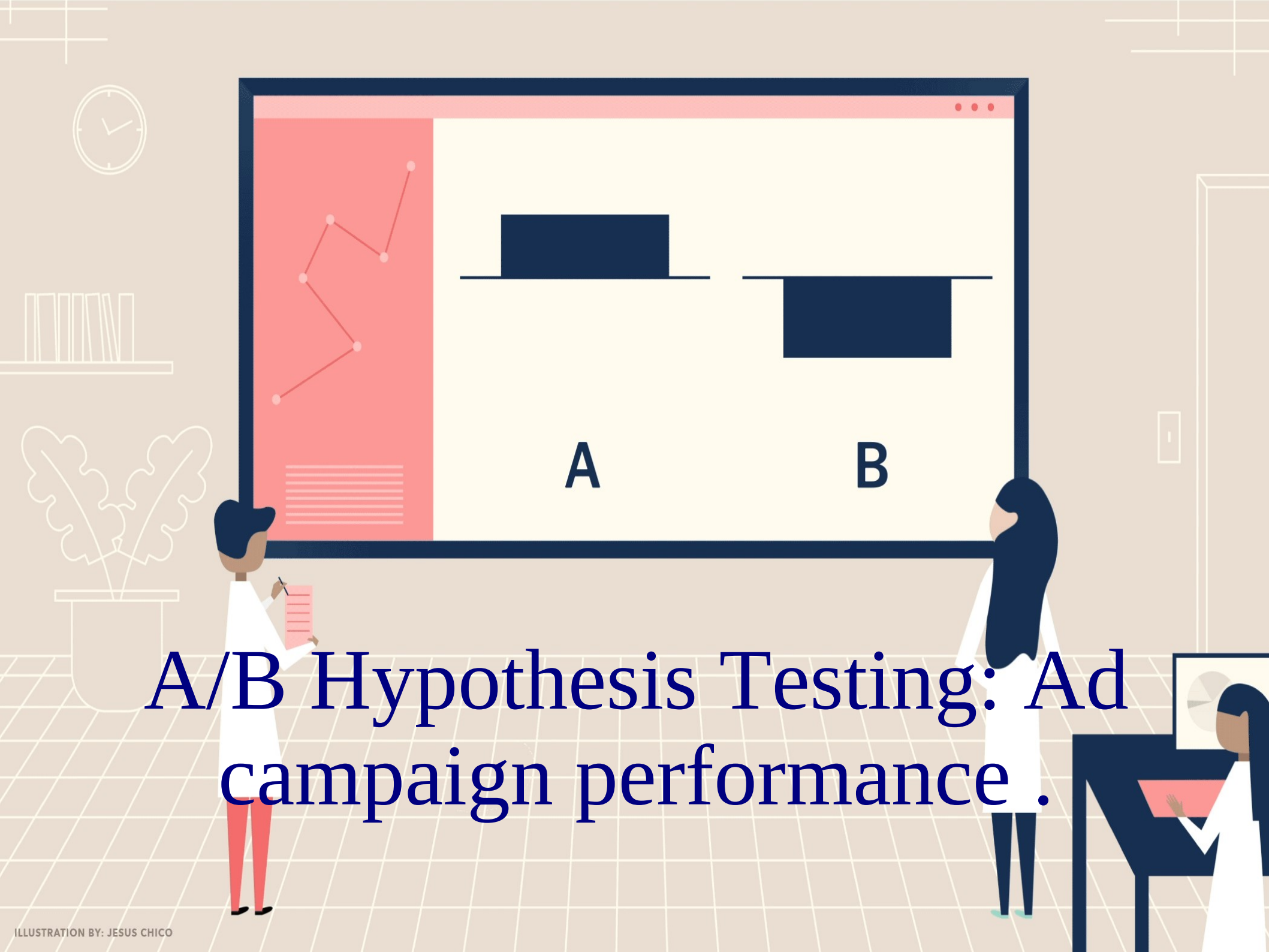


A/B Hypothesis Testing: Ad campaign performance .

ILLUSTRATION BY: JESUS CHICO

Introduction and Objectives

Case Overview:

SmartAd is a mobile first advertiser agency. It is running an online ad for a client with the intention of increasing brand awareness.

The company provides an additional service called Brand Impact Optimizer (BIO), a lightweight questionnaire, served with every campaign to determine the impact of the ad they design.

Objectives:

The task at hand is to design a reliable hypothesis testing algorithm for the BIO service and determine whether the recent advertising campaign resulted in significant lift in brand awareness.

Methods

- A/B testing is a user experience research methodology that that consists of a randomized experiment with 2 variants. Statistical hypothesis testing is applied to compare the 2 variants and determine which is more effective.
- The following techniques of A/B testing is be conducted in this analysis:
  - **Sequential A/B testing:** Involves conducting the test on 2 versions of a single variable at a time. It goes with the belief that not more than one factor should be varied at the same time.
  - **Classic A/B testing:** Unlike sequential, classic involves conducting the test on the variants at a go. It allows checking of results at the very end of the test.
  - **A/B testing with Machine Learning:** The ML approach allows modeling of complex systems unlike the statistical inference approach. Feature significance in ML models is what provides insights towards wether the experiment had impact or not. It also outlines the contributions of other features towards the viewed outcome.

Data: Features

- The BIO data for this project is a “Yes” and “No” response of online users to the following question: ***`Q: Do you know the brand SmartAd?`***
- The data has the following features:
  - *auction_id*: the unique id of the online user who has been presented the BIO.
  - *experiment*: which group the user belongs to - control or exposed.
  - *Date* and *Hour*: the date and hour the response was recorded.
  - *device_make*: the name of the type of device the user has.
  - *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.

Data: Univariate Analysis

.

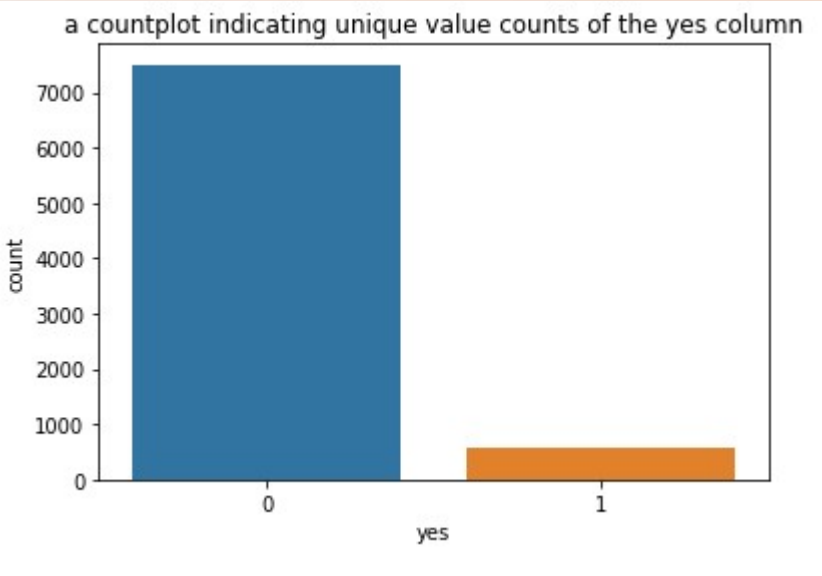


figure 1.

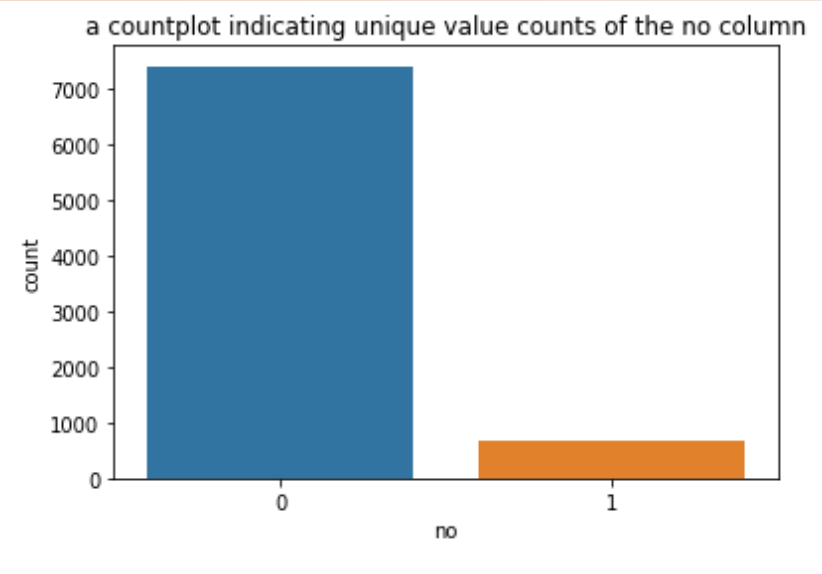


figure 2.

- In figure 1, the 0 entries have a higher count cause it represents users who either answered no or didn't respond to the questionnaire. The count of 1 entries also indicates the number of people who recall the ad.
- In figure 2, the 0 entries have a higher count cause it represents users who either answered yes or didn't respond to the questionnaire.
- In summary, the number of 1 entries are low. This means a small percentage of the users responded to the questionnaire.

Data: Univariate Analysis

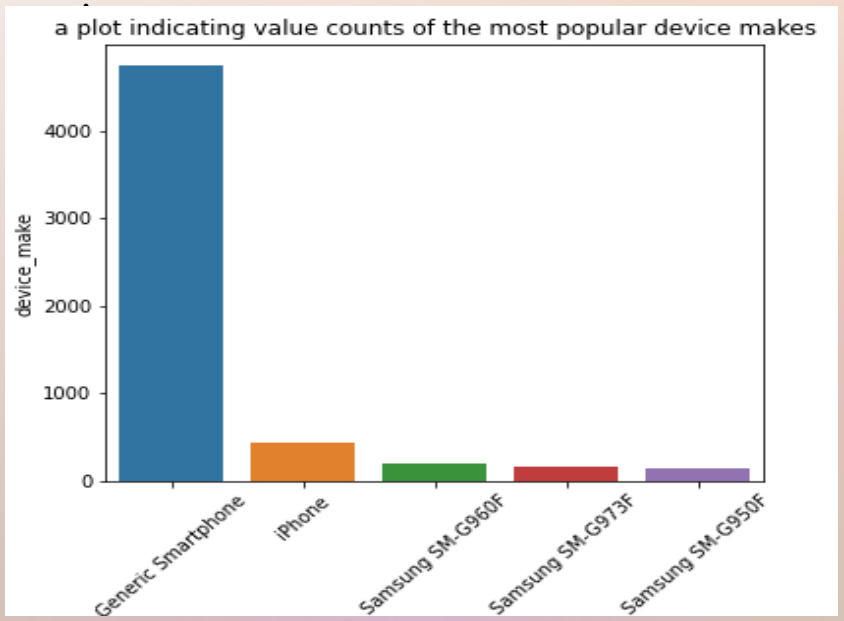


Figure 3.

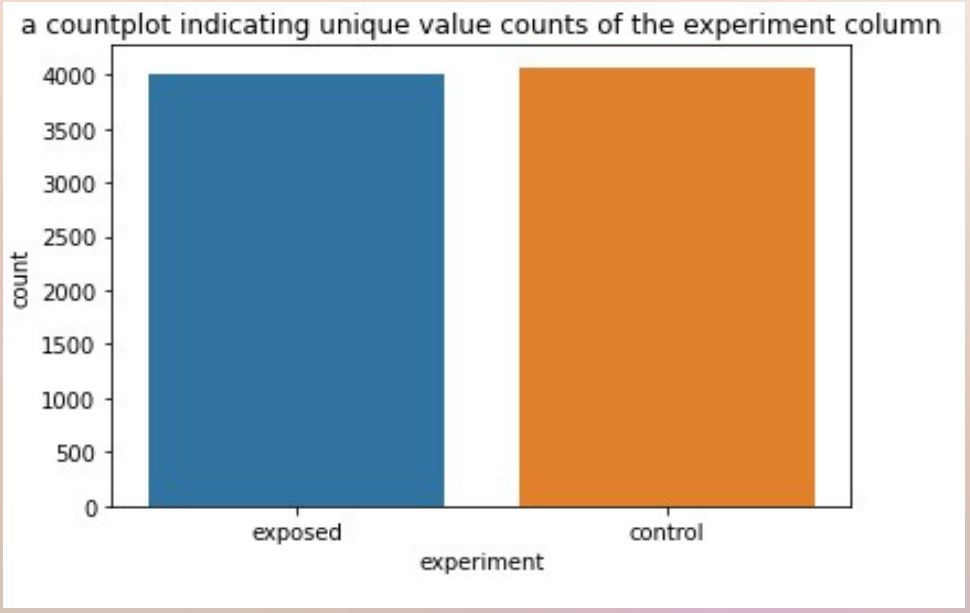


Figure 4.

- Figure 3 indicates top 5 most used device makes. Samsung devices make up 60% of these.
- Figure 4 indicates the number of users in the 2 groups: *exposed* (shown the SmartAd ad) and the *control* (shown a dummy ad). There appears to be a balance in the numbers. This is good for the test since the difference between the two is required to be statistically insignificant in order to render the test valid.

Data: Multivariate Analysis

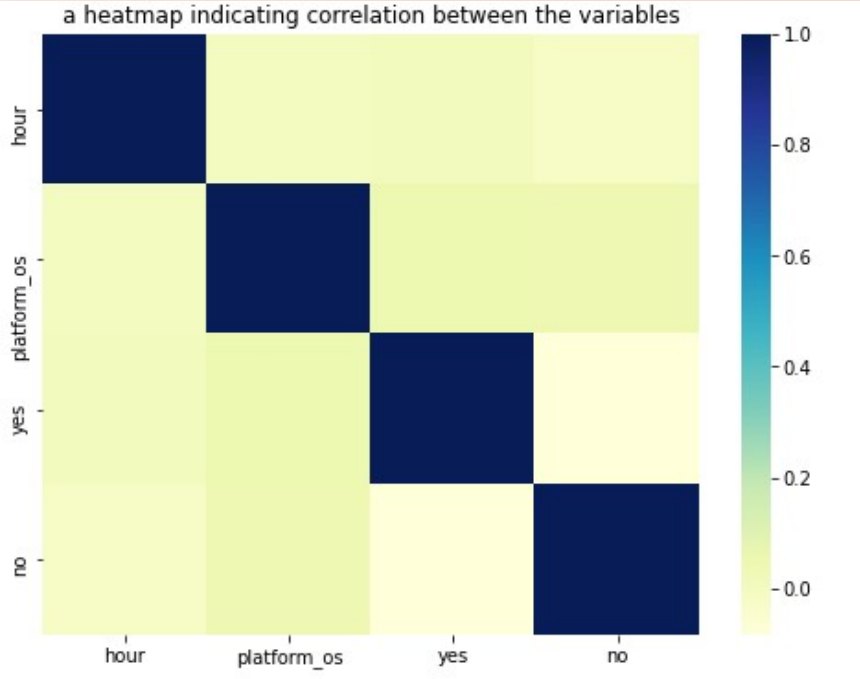
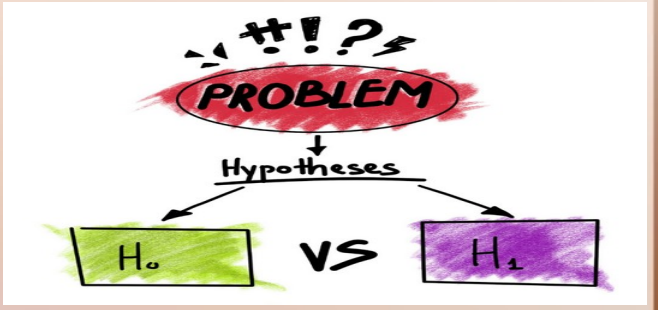


Figure 5.

- Correlation is a statistical technique that can show whether and how strongly pairs of variables are related.
- The heatmap in figure 5 show little to no correlation between the variables.

Results: Classic A/B Testing

Hypothesis



- H0: There's no difference in brand awareness between the 2 groups.
- H1: There's a difference in brand awareness between the 2 groups.

Metrics

- *Invariate metrics* are sanity checks that make sure the experiment is not inherently wrong. They include:
  - Total Number of users in each experiment group.
  - Average Number of online users per day.
  - Device Make unique counts in each group.
  - Browser type unique counts in each group.
  - Platform OS unique counts in each group.

- *Evaluation Metrics* are the metrics in which we expect to see a change, and are relevant to our goal. It includes:
  - Net Conversion: Proportion of users who recall the SmartAd ad to the total number of users aggregated on a daily basis.

Results: Classic A/B Testing

Sample Size

- The minimum sample size is calculated so that the experiment can have enough significance as well as statistical power.
- I6,162 users is the value obtained. Assuming we take 90% of daily users (1010 users), the data collection period for the experiment should be about 2 weeks, 3 days. The data available is for 8 days, thus we're one week less on the optimal sample size.

Results

- Using the collected data as it is:
- All the invariant metrics are upheld apart from the unique device counts across the groups. Their difference is statistically significant.
- The test shows a positive change in brand awareness in the exposed group by 1.2% which is higher by 0.2% than the minimum detectable change set of 1%. In conclusion, we reject the null hypothesis. Thus, there is a variation in brand awareness between the 2 groups.

Results: Sequential A/B Testing

Hypothesis

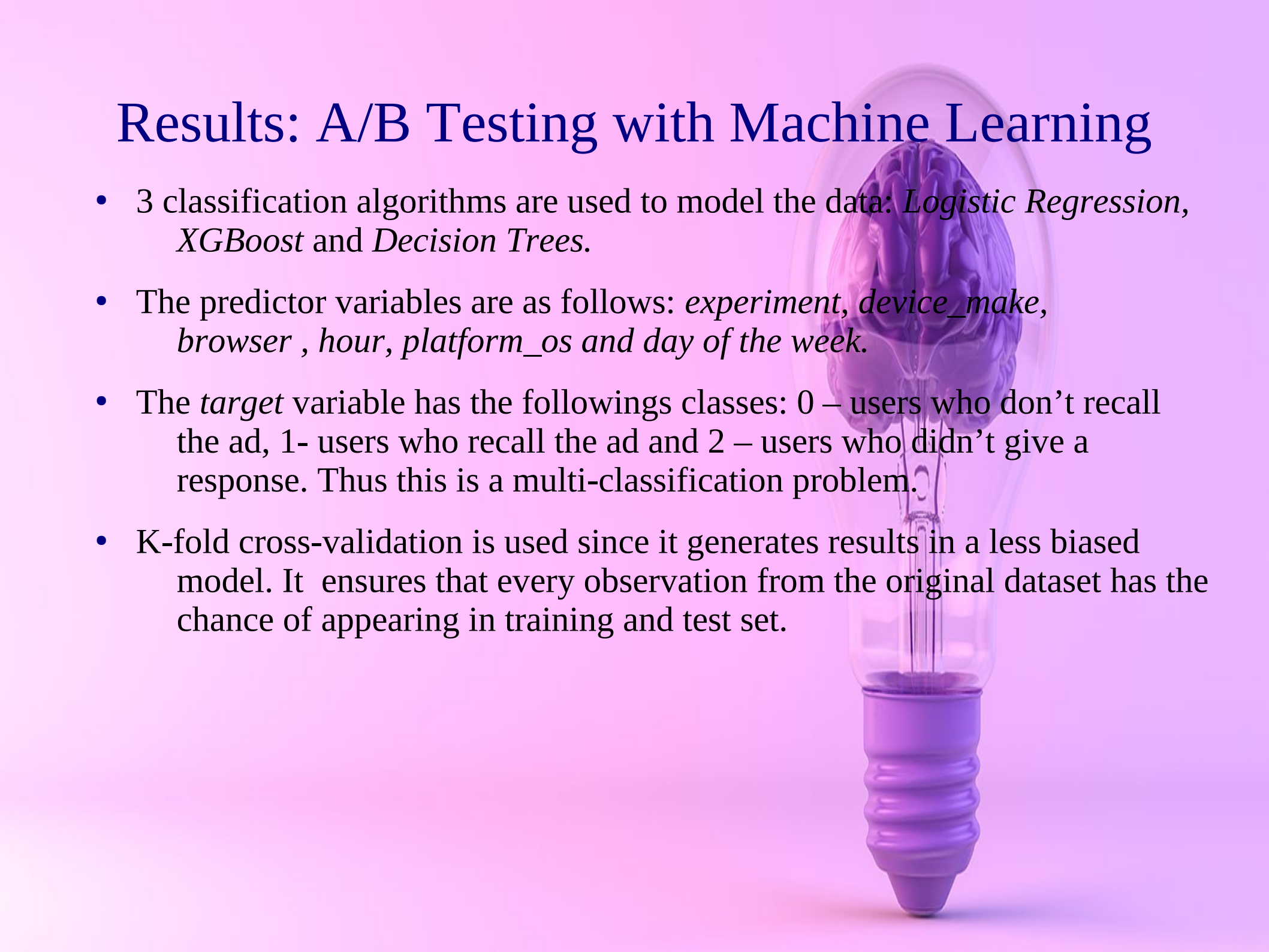
- H0: There's no difference in brand awareness between the 2 groups.
- H1: There's a difference in brand awareness between the 2 groups.

Outcome:

'Unable to conclude. Needs more sample.'

The quoted sentence is the outcome of the test. This matches the sample size limitation highlighted in the classic test. Therefore we cannot make a conclusion on the hypothesis stated.

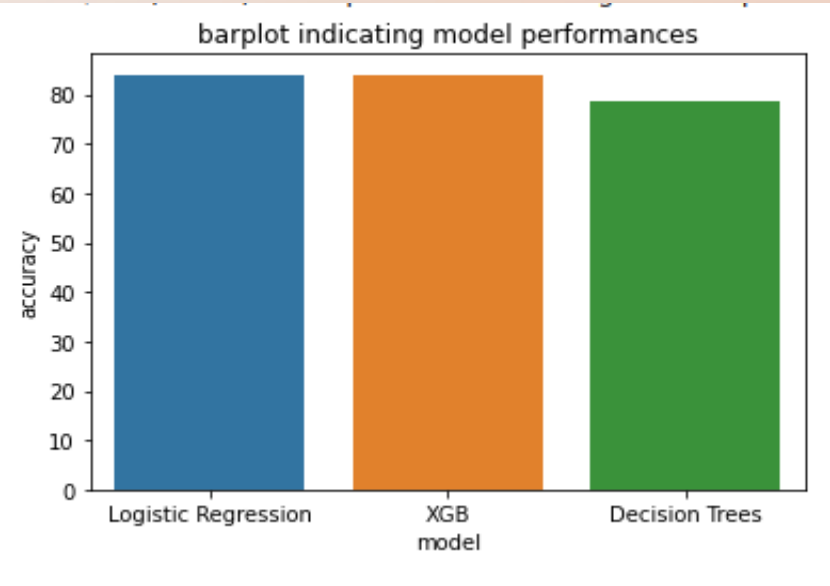
Results: A/B Testing with Machine Learning



- 3 classification algorithms are used to model the data: *Logistic Regression*, *XGBoost* and *Decision Trees.*
- The predictor variables are as follows: *experiment*, *device_make*, *browser* , *hour*, *platform_os and day of the week.*
- The *target* variable has the followings classes: 0 – users who don't recall the ad, 1- users who recall the ad and 2 – users who didn't give a response. Thus this is a multi-classification problem.
- K-fold cross-validation is used since it generates results in a less biased model. It ensures that every observation from the original dataset has the chance of appearing in training and test set.

Results: A/B Testing with Machine Learning

Model Performance and Loss Functions



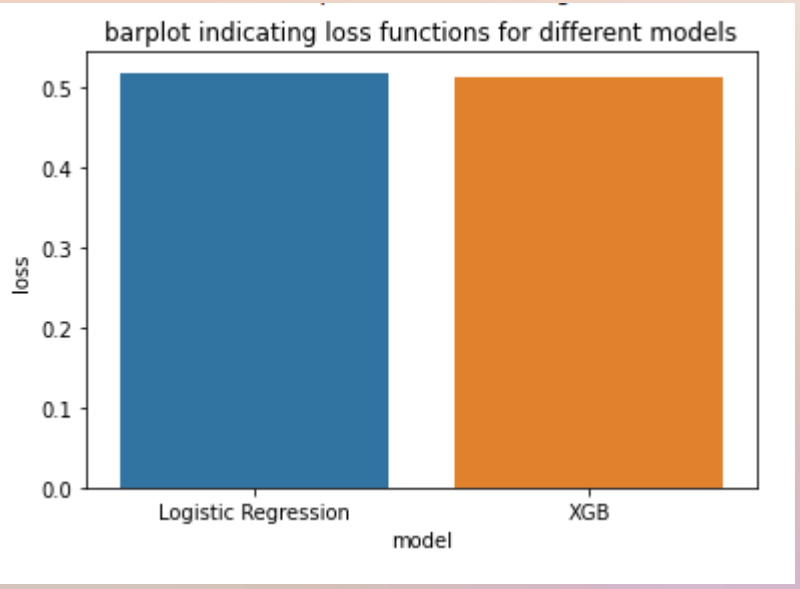


Figure 6 .

Figure 7.

- Figure 6 indicates the accuracies of the 3 models. Decision Trees Classifier records the lowest accuracy.
- Figure 7 indicates the loss functions of the 2 most accurate models. XGB's loss is slightly lower than Logistic Regression

Results: A/B Testing with Machine Learning

Feature Importance

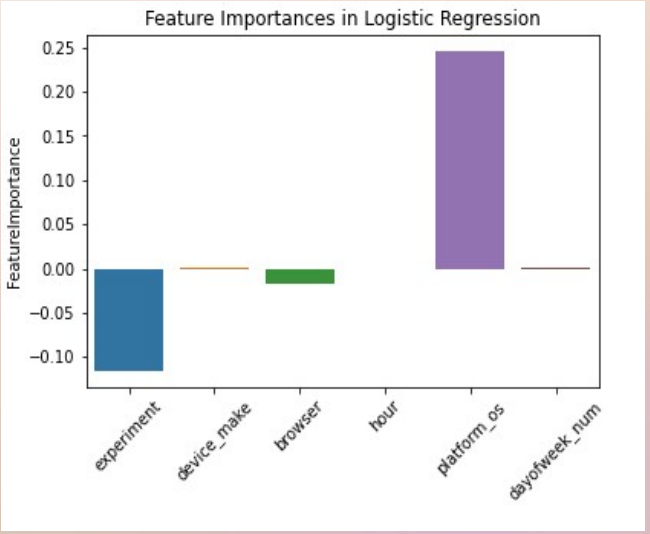


Figure 8.

Figure 9.

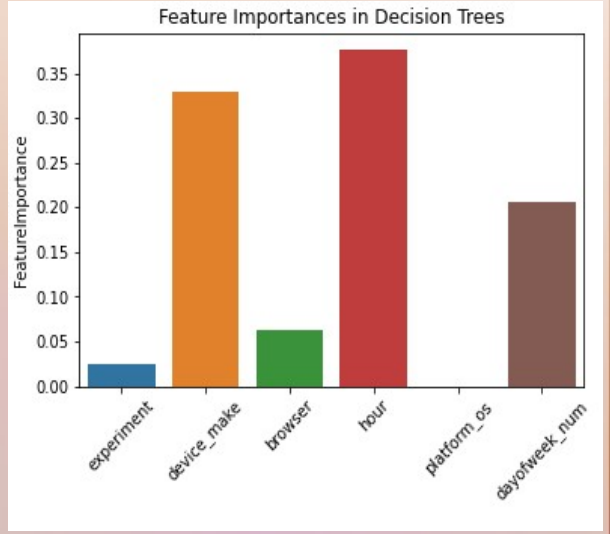


Figure 10.

- *Experiment* is significant in the prediction according to all the 3 models.
- The *experiment* coefficient in logistic regression is interpreted as the expected change in log odds for a one-unit increase in the experiment group.
- Other top significant features are: *device_make*, *day of_the week*, *browser* and *hour*.

Overall Results

- According to the results of classic A/B testing, there's an increase in brand awareness in the exposed group.
- According to the results of Machine Learning, *experience* proves to be a significant feature in predicting whether a user recalls the ad thus the experiment was worth it.
- In conclusion, there's difference in brand awareness between the 2 groups. It is higher in the exposed group.

Recommendations

- The advertisement campaign resulted in a significant lift in brand awareness thus the ad should be used in place of other alternatives.
- The metrics used in designing the ad should be extended to future ad designs.
- SmartAd can increase the client charge since the ad has proven useful.

Limitations

- Insufficient sample size. This is concluded based on the minimum sample size calculated from the classic test and the outcome of the sequential test.
- Majority of the users didn’t respond to the questionnaire, thus the ML models had to be run on 3 classes since dropping them would result to poor performance.

References

1) William Q. Meeker, Jr, ‘A Conditional Sequential Test for the Equality of Two Binomial Proportions’, Journal of the Royal Statistical Society. Series C (Applied Statistics), # Vol. 30, No. 2 (1981), pp. 109-115.
2) https://www.kaggle.com/tammyrotem/ab-tests-with-python/notebook
3) https://www.business-science.io/business/2019/03/11/ab-testing-machine-learning.html
4) https://medium.com/analytics-vidhya/a-b-testing-clearly-explained-56488430156
5) https://github.com/Testispuncher/Sequential-Probability-Ratio-Test