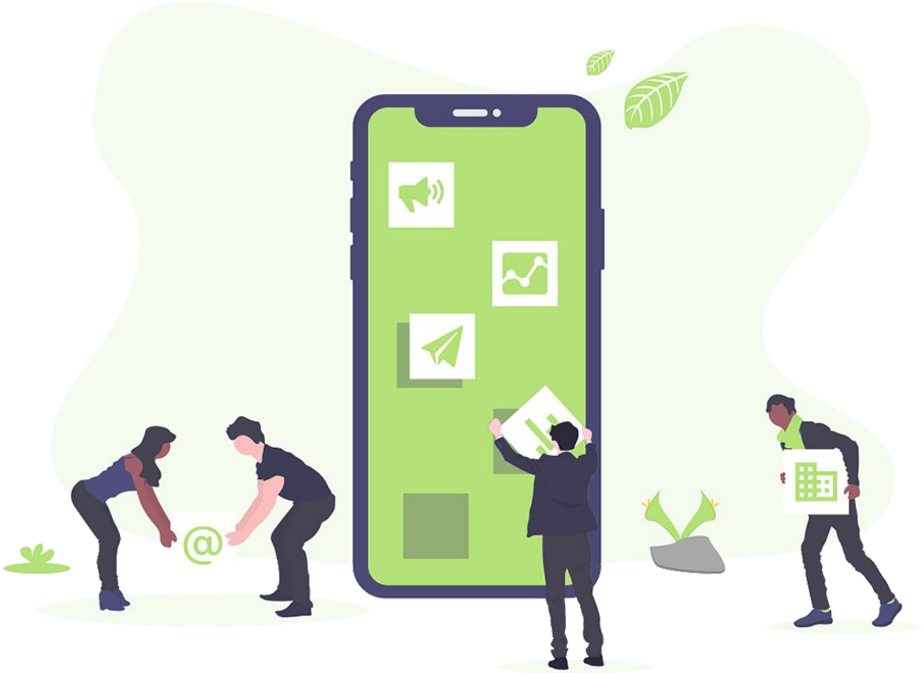
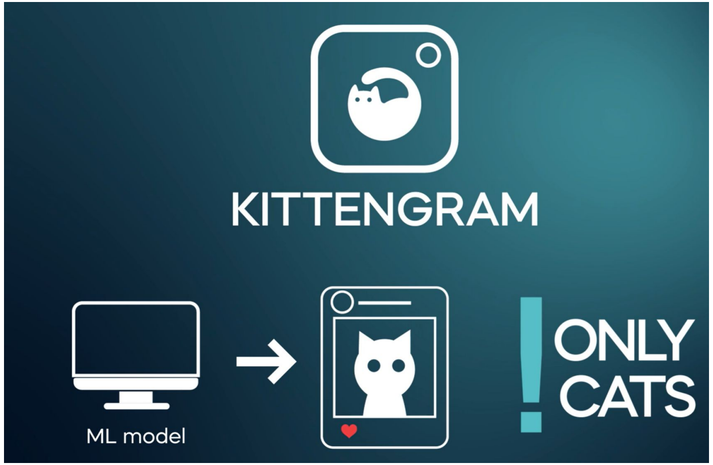
A/B testing in a digital product

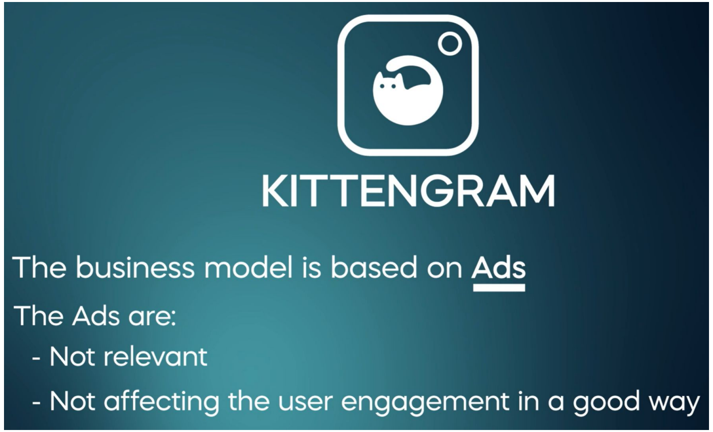


Aadhavan Alakan

Problem Definition

- Performing A/B testing on a frictional app (Kittengram).
- Kittengram is an app where users can share their cat pictures and users can only comment or like about other user’s pictures.
- They have an ML model which verifies the user’s pictures and makes sure that they are posting only cat pictures.
- The business model is based on Ads
- The company is still functioning on investors money because they are not making enough money on their own.
- The ads are not relevant to the user’s interest.
- They are not affecting the user engagement in a good way




KITTENGRAM

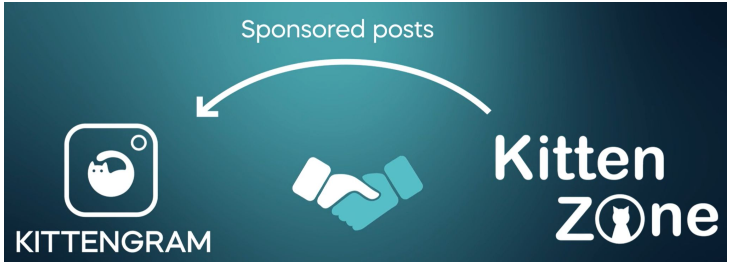
The business model is based on Ads

The Ads are:

- Not relevant
- Not affecting the user engagement in a good way

Problem Definition

- Company is planning to partner with a big retailer who sells cat products.
- They want to introduce sponsored post in the app so that the users can get relevant ads.
- Through affiliate links the company can also increase the revenue when an user engages with an ad and buys the product through it.
- The goal is to determine whether new feature (sponsored post) is liked by the users and the company's revenue increases.



Key Performance Indicators (KPIs)

- Daily Active Users (DAU)

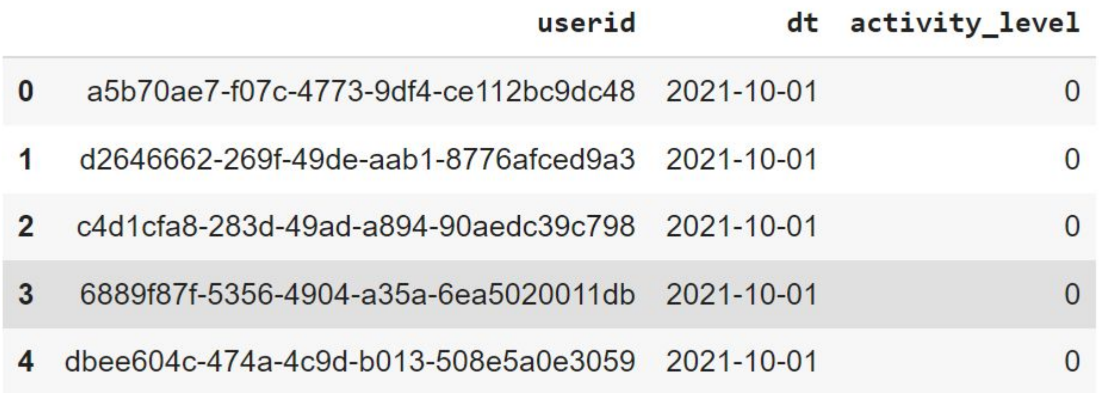
  To check if the users are coming back to the app on a daily basis.

- Click-Through Rate (CTR)

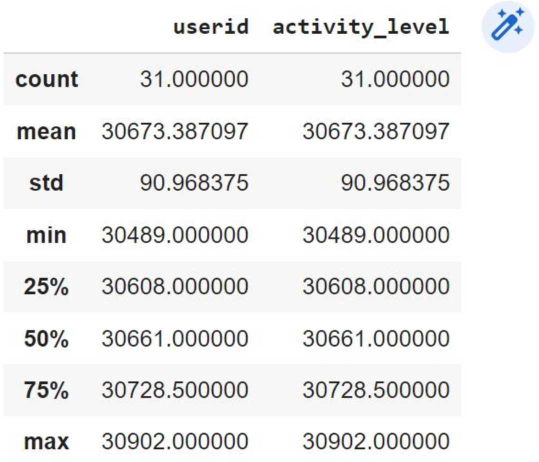
  Checking the CTR between the new type of sponsored posts vs. the old type of sponsored posts that were not tailored to our user base.

Analyzing the Dataset - DAU

- For every user registered in the app, for every day that the app worked, we calculate the activity level of that user.
- The **activity_level** is represented by an integer: the number of times the user opened the app during the day.
- **userid**: describes the unique user who is registered and using the app.
- **dt**: it refers to the date on which the activity was recorded.



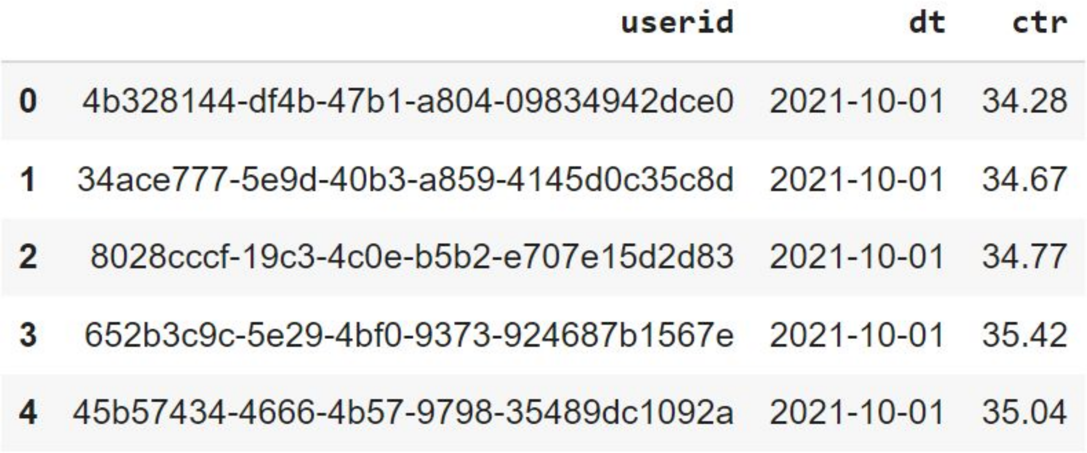
	userid	dt	activity_level
0	a5b70ae7-f07c-4773-9df4-ce112bc9dc48	2021-10-01	0
1	d2646662-269f-49de-aab1-8776afced9a3	2021-10-01	0
2	c4d1cfa8-283d-49ad-a894-90aedc39c798	2021-10-01	0
3	6889f87f-5356-4904-a35a-6ea5020011db	2021-10-01	0
4	dbee604c-474a-4c9d-b013-508e5a0e3059	2021-10-01	0



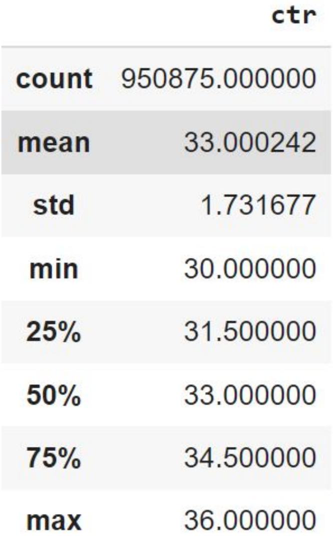
	userid	activity_level
count	31.000000	31.000000
mean	30673.387097	30673.387097
std	90.968375	90.968375
min	30489.000000	30489.000000
25%	30608.000000	30608.000000
50%	30661.000000	30661.000000
75%	30728.500000	30728.500000
max	30902.000000	30902.000000

Analyzing the Dataset – CTR

- For every user registered in the app, for every day that the app worked, we calculate the activity level of that user.
- The **ctr** is represented by a float: the number of ads user clicked on/total number of ads this user was exposed to during the day.
- **userid**: describes the unique user who is registered and using the app.
- **dt**: it refers to the date on which the activity was recorded.

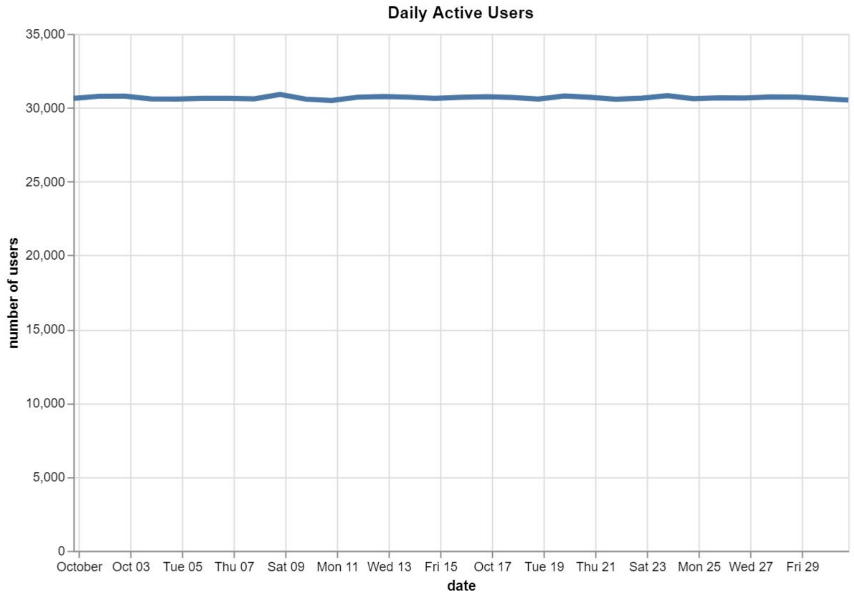


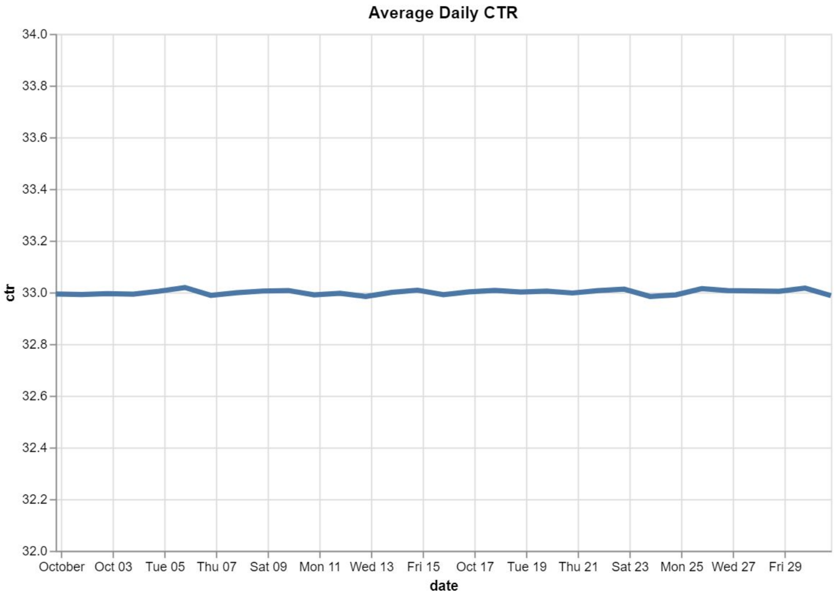
	userid	dt	ctr
0	4b328144-df4b-47b1-a804-09834942dce0	2021-10-01	34.28
1	34ace777-5e9d-40b3-a859-4145d0c35c8d	2021-10-01	34.67
2	8028cccf-19c3-4c0e-b5b2-e707e15d2d83	2021-10-01	34.77
3	652b3c9c-5e29-4bf0-9373-924687b1567e	2021-10-01	35.42
4	45b57434-4666-4b57-9798-35489dc1092a	2021-10-01	35.04



	ctr
count	950875.000000
mean	33.000242
std	1.731677
min	30.000000
25%	31.500000
50%	33.000000
75%	34.500000
max	36.000000

Checking the average DAU & CTR





- Both metrics look stable. This will be useful for us to perform the test and determine how long the test should run and how many people we should expose to this test to reach the statistical significance for the metrics (sample size).
- The above dataset runs between October 1st to October 31st.

Why are the average DAU & CTR stable?

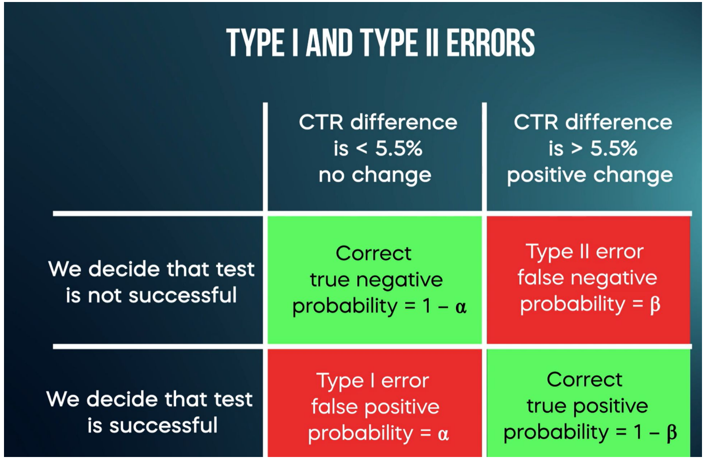
- The dataset is generated for practice purpose and it does not represent actual users.\
- Some of the potential reasonings might be if this was observed in real life would be fraud accounts, bots or some automated activities performed by the users etc.
- Maybe there are promotions in the app that made the users come back to the apps everyday.

Defining the Hypothesis

- H1 - with the introduction of the tailored ads, users are more likely to click on the ads, therefore the CTR will increase per user.
- H0 - Tailored ads will have no effect on the user engagement with the ads, and will not affect the CTR.
- Success metric for our case will be CTR, it is based on the hypothesis -it is the dependable variable.
- We want to control the test impact for the business-critical metrics. The Guardrail metrics for our case will be DAU (Daily active Users). It is important not to lose the users.
- We are setting alpha = 5% and beta = 20%

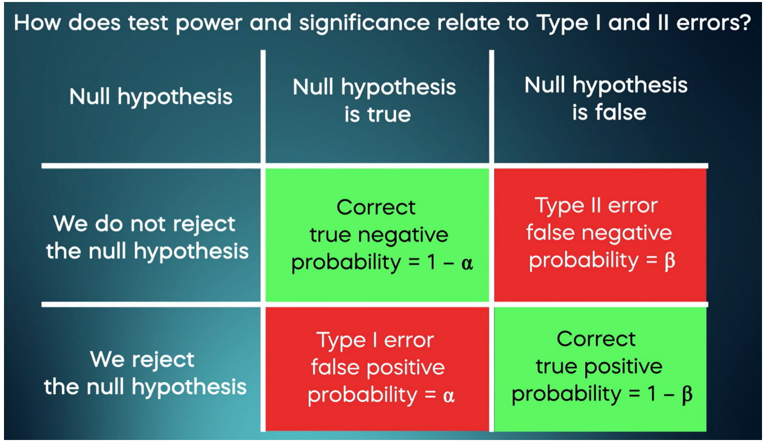
Defining the Minimum Detectable Effect

- By analyzing the mean and standard deviations in both the datasets we fix our MDE.
- We would like to determine at least 91 DAU difference between the test and control group. This equals to approximately 0.33% increase.
- We would like to determine at least 1.8 percentage points difference between test and control groups. This equals to approximately 5.5% increase.



TYPE I AND TYPE II ERRORS

	CTR difference is < 5.5% no change	CTR difference is > 5.5% positive change
We decide that test is not successful	Correct true negative probability = 1 – α	Type II error false negative probability = β
We decide that test is successful	Type I error false positive probability = α	Correct true positive probability = 1 – β



How does test power and significance relate to Type I and II errors?

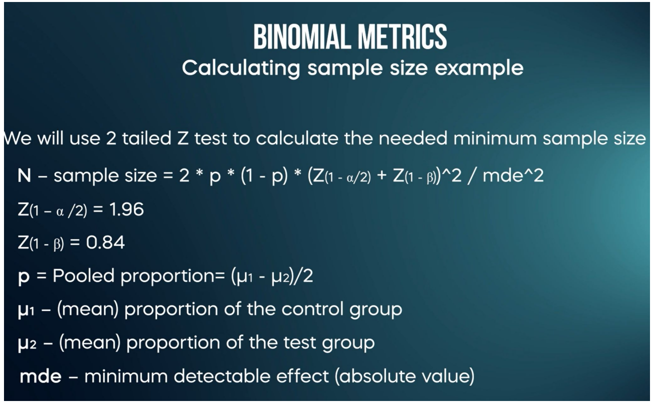
Null hypothesis	Null hypothesis is true	Null hypothesis is false
We do not reject the null hypothesis	Correct true negative probability = 1 – α	Type II error false negative probability = β
We reject the null hypothesis	Type I error false positive probability = α	Correct true positive probability = 1 – β

Calculating the sample size for Binomial metric (CTR)

- We will use 2 tailed Z test to calculate the needed minimum sample size.

- We found out that we need at least 8797 users in each group, provided the groups are equally split.

- For our case, we are taking around 60,000 users in total, and they are nearly split equally, thereby getting around 30,000 users between the test group and control group.



BINOMIAL METRICS

Calculating sample size example

We will use 2 tailed Z test to calculate the needed minimum sample size

N – sample size = 2 * p * (1 - p) * (Z$_{(1 - \alpha/2)}$ + Z$_{(1 - \beta)}$)^2 / mde^2

Z$_{(1 - \alpha /2)}$ = 1.96

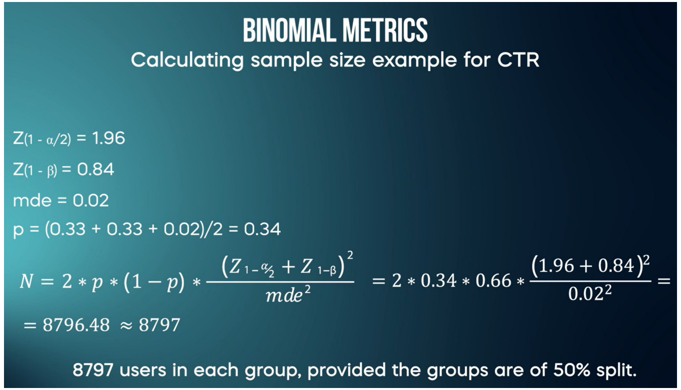
Z$_{(1 - \beta)}$ = 0.84

p = Pooled proportion= (μ_1 - μ_2)/2

μ_1 – (mean) proportion of the control group

μ_2 – (mean) proportion of the test group

mde – minimum detectable effect (absolute value)



BINOMIAL METRICS

Calculating sample size example for CTR

Z$_{(1 - \alpha/2)}$ = 1.96

Z$_{(1 - \beta)}$ = 0.84

mde = 0.02

p = (0.33 + 0.33 + 0.02)/2 = 0.34

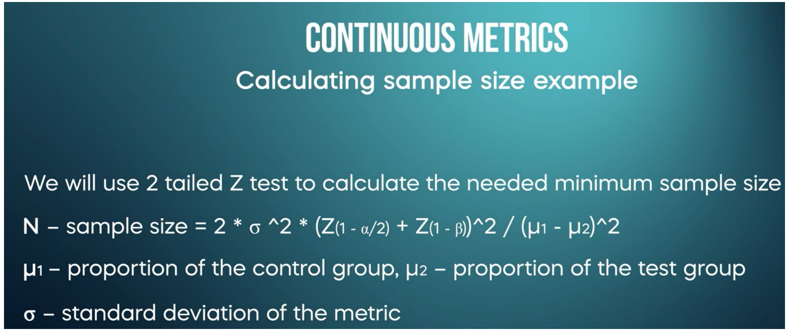
$$N = 2 * p * (1 - p) * \frac{\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2}{mde^2} = 2 * 0.34 * 0.66 * \frac{(1.96 + 0.84)^2}{0.02^2} =$$

$$= 8796.48 \approx 8797$$

8797 users in each group, provided the groups are of 50% split.

Calculating the sample size for Continuous metric (DAU)

- We will use 2 tailed Z test to calculate the needed minimum sample size.
- We found out that we need at least 13 days in each group, provided the groups are equally split.
- For our case, we are running the test for one full month (30 days).



CONTINUOUS METRICS

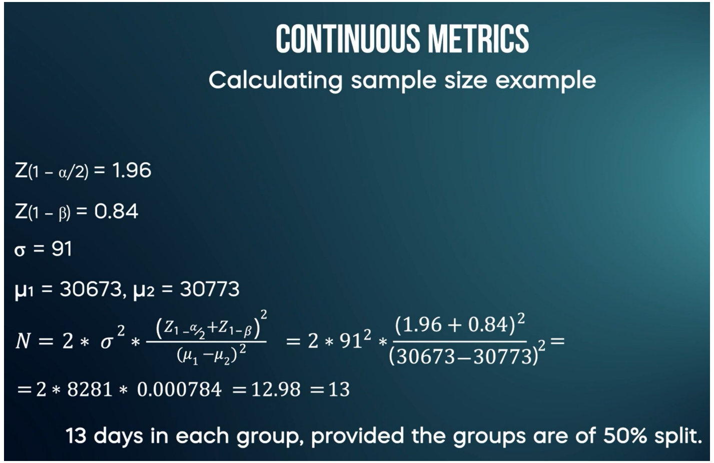
Calculating sample size example

We will use 2 tailed Z test to calculate the needed minimum sample size

N – sample size = 2 * σ ^2 * ($Z_{(1 - \alpha/2)}$ + $Z_{(1 - \beta)}$)^2 / (μ_1 - μ_2)^2

μ_1 – proportion of the control group, μ_2 – proportion of the test group

σ – standard deviation of the metric



CONTINUOUS METRICS

Calculating sample size example

$Z_{(1 - \alpha/2)} = 1.96$

$Z_{(1 - \beta)} = 0.84$

$\sigma = 91$

$\mu_1 = 30673$, $\mu_2 = 30773$

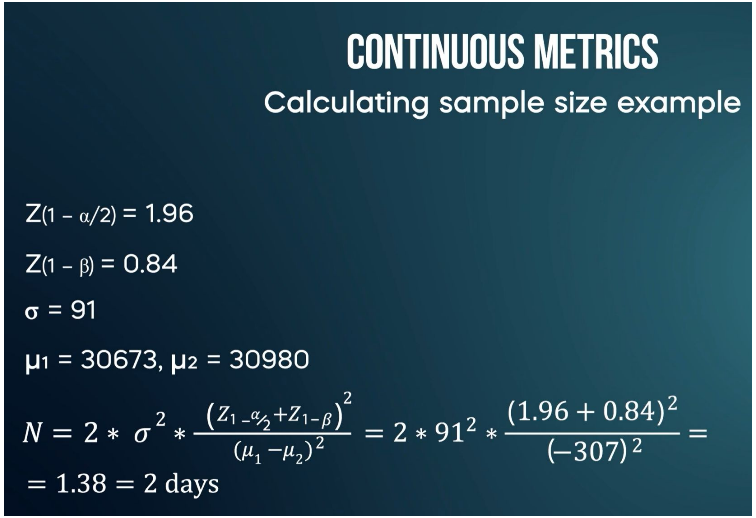
$$N = 2 * \sigma^2 * \frac{(Z_{1-\alpha/2} + Z_{1-\beta})^2}{(\mu_1 - \mu_2)^2} = 2 * 91^2 * \frac{(1.96 + 0.84)^2}{(30673 - 30773)^2} =$$

$$= 2 * 8281 * 0.000784 = 12.98 = 13$$

13 days in each group, provided the groups are of 50% split.

Changing the relative difference (DAU)

- Solving this issue for 100 users might be very tight and too much statistical significance for us. Also performing the test for 13 days is potentially very long.
- Hence we want to calculate the approximately 1% relative difference between the test and control group instead of 0.33%.
- This increases the daily active users (DAU) from 100 to 300.
- We found that we need at least two days in each group in the revised calculation, provided the groups are equally split.



CONTINUOUS METRICS

Calculating sample size example

$Z_{(1-\alpha/2)} = 1.96$

$Z_{(1-\beta)} = 0.84$

$\sigma = 91$

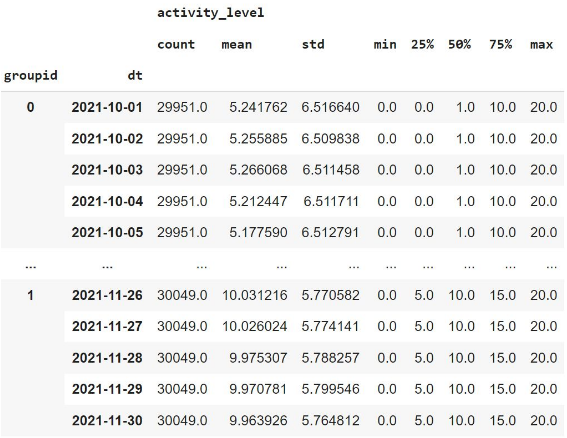
$\mu_1 = 30673,\ \mu_2 = 30980$

$$N = 2 * \sigma^2 * \frac{\left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2}{(\mu_1 - \mu_2)^2} = 2 * 91^2 * \frac{(1.96 + 0.84)^2}{(-307)^2} =$$

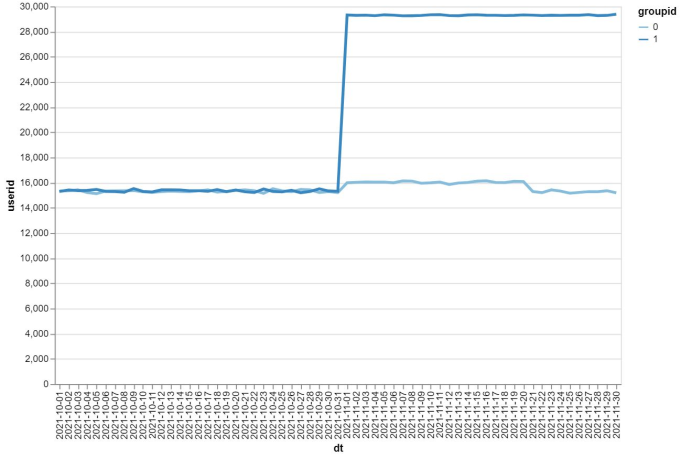
$$= 1.38 = 2 \text{ days}$$

Analyzing the A/B test - DAU

- This dataset analyzes the data from October 1st to November 30th. The test starts on November 1st.
- We can observe the difference between the groups by observing the median values in group - 0 and group -1.
- The control group has a median activity of 1 per day, while in test group, it has 5 activities.
- Also, by observing the graph, we see a spike in the number of users active in a day from November 1st, and this indicates that the test has made a positive impact.
- Finally, we performed a t-test and calculated the P-value between the groups. The P-value was very small; hence we rejected the Null hypothesis, which implies a big difference in the user activity between the groups, which is evident from the graph.

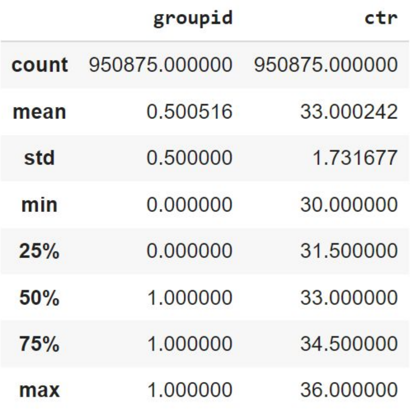


		activity_level							
		count	mean	std	min	25%	50%	75%	max
groupid	dt								
0	2021-10-01	29951.0	5.241762	6.516640	0.0	0.0	1.0	10.0	20.0
	2021-10-02	29951.0	5.255885	6.509838	0.0	0.0	1.0	10.0	20.0
	2021-10-03	29951.0	5.266068	6.511458	0.0	0.0	1.0	10.0	20.0
	2021-10-04	29951.0	5.212447	6.511711	0.0	0.0	1.0	10.0	20.0
	2021-10-05	29951.0	5.177590	6.512791	0.0	0.0	1.0	10.0	20.0
...	...	...	...	...	...	...	...	...	...
1	2021-11-26	30049.0	10.031216	5.770582	0.0	5.0	10.0	15.0	20.0
	2021-11-27	30049.0	10.026024	5.774141	0.0	5.0	10.0	15.0	20.0
	2021-11-28	30049.0	9.975307	5.788257	0.0	5.0	10.0	15.0	20.0
	2021-11-29	30049.0	9.970781	5.799546	0.0	5.0	10.0	15.0	20.0
	2021-11-30	30049.0	9.963926	5.764812	0.0	5.0	10.0	15.0	20.0

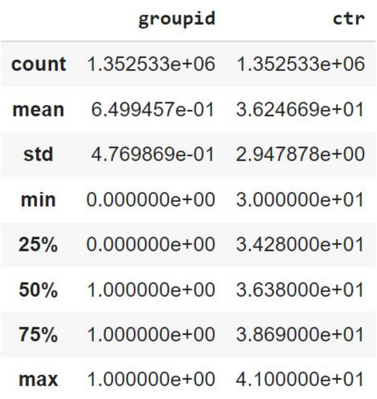


Analyzing the A/B test – CTR

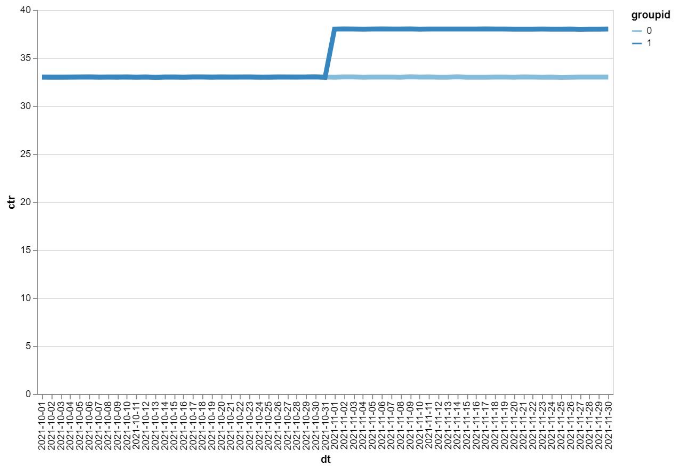
- We can see that the click-through rate between the groups has increased. The mean ctr has increased by 9.81 % and the user activity rate has increased by a whopping 42.24%.
- The above increase can be seen from the graph below.
- Finally, we performed a t-test and calculated the P-value between the groups. The P-value was very small; hence we rejected the Null hypothesis, which implies a difference in the click through rate between the groups, which is evident from the graph.



	groupid	ctr
count	950875.000000	950875.000000
mean	0.500516	33.000242
std	0.500000	1.731677
min	0.000000	30.000000
25%	0.000000	31.500000
50%	1.000000	33.000000
75%	1.000000	34.500000
max	1.000000	36.000000



	groupid	ctr
count	1.352533e+06	1.352533e+06
mean	6.499457e-01	3.624669e+01
std	4.769869e-01	2.947878e+00
min	0.000000e+00	3.000000e+01
25%	0.000000e+00	3.428000e+01
50%	1.000000e+00	3.638000e+01
75%	1.000000e+00	3.869000e+01
max	1.000000e+00	4.100000e+01



Conclusion

- H1 - with the introduction of the tailored ads, users are more likely to click on the ads, therefore the CTR difference > 5.5% per user.
- H0 - Tailored ads will have no effect on the user engagement with the ads, and CTR difference < 5.5% per user.
- We observed a mean ctr of 33 ad-clicks per user in the control group whereas in the test group we got 36.24 ad-clicks - an increase of 9.81% per user.
- The user activity also saw a massive 42.24% increase. This might be good news for the business, but we should calculate user retention since high user activity rates may often result in user burnout.
- We can conclude that the above A/B test has been successful because we rejected the Null hypothesis since the difference in CTR between the groups was well above 5.5%.