Click Through Rate Prediction

Manisha Paliwal (MSDS: University of Arizona, June 2022)

Problem Statement

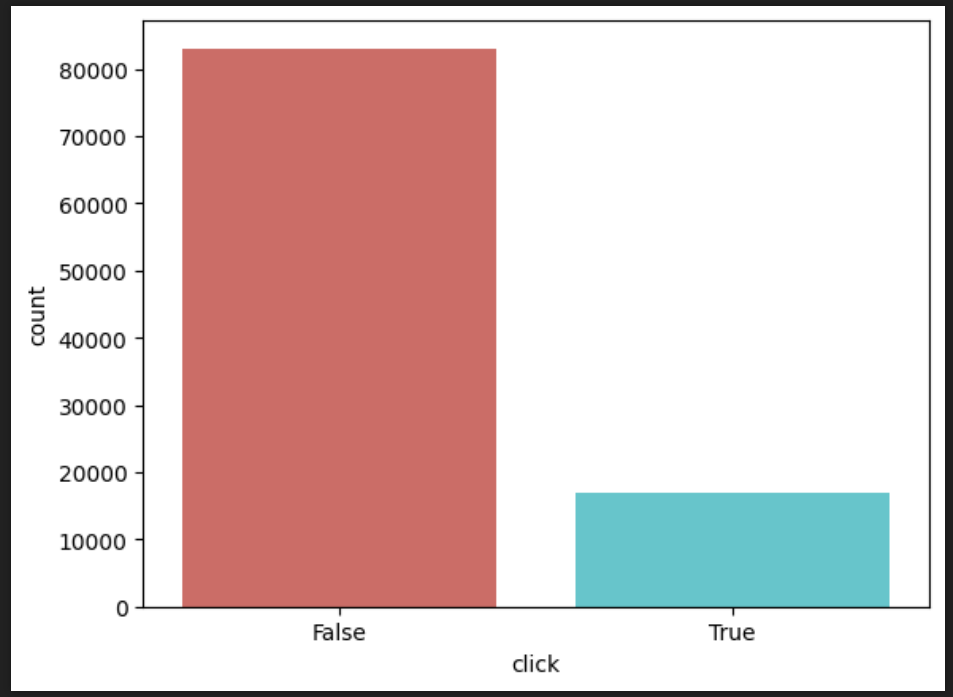
- CTR is an important exercise for marketing companies.
- The objective is to predict whether the audience will click on an ad or not and thus help the marketing team answer ad placement-related questions

Data Attributes

click:	0/1 for non-click/click
hour: format is YYMMDDHH, so 14091123 means 23:00 on Sept. 11, 2014 UTC.	
C1	anonymized categorical variable
banner_pos	position of the ad/banner on the page
site_id	unique id of the site on which the ad is shown
site_domain	unique domain of the site on which the ad is shown
site_category	category of the site on which the ad is shown
app_id	app id of the site on which the ad is shown
app_domain	app category of the site on which the ad is shown
app_category	category id of the site on which the ad is shown
device_id	device id on which the add was shown
device_ip	ip address of the device on which the ad was shown
device_model	model type of the device on which the ad was shown
C14 - C21	anonymized categorical variable

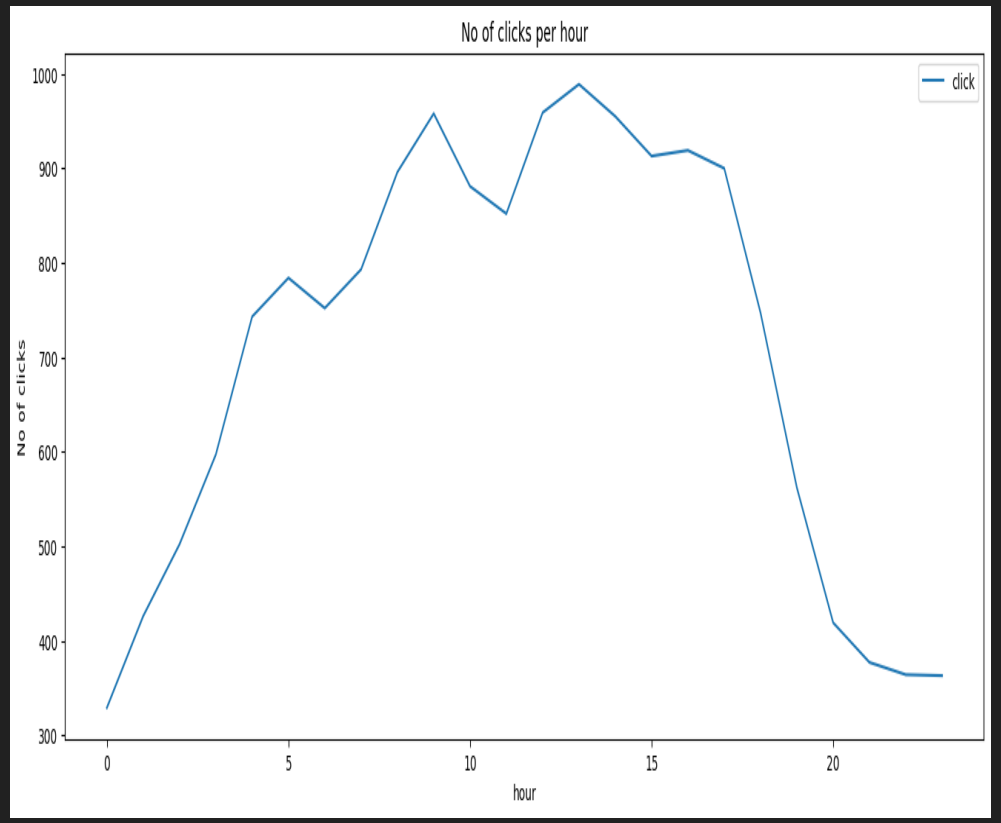
Analysis: CTR

- From this analysis, we can say that approx. 83% did not click and 17% clicked



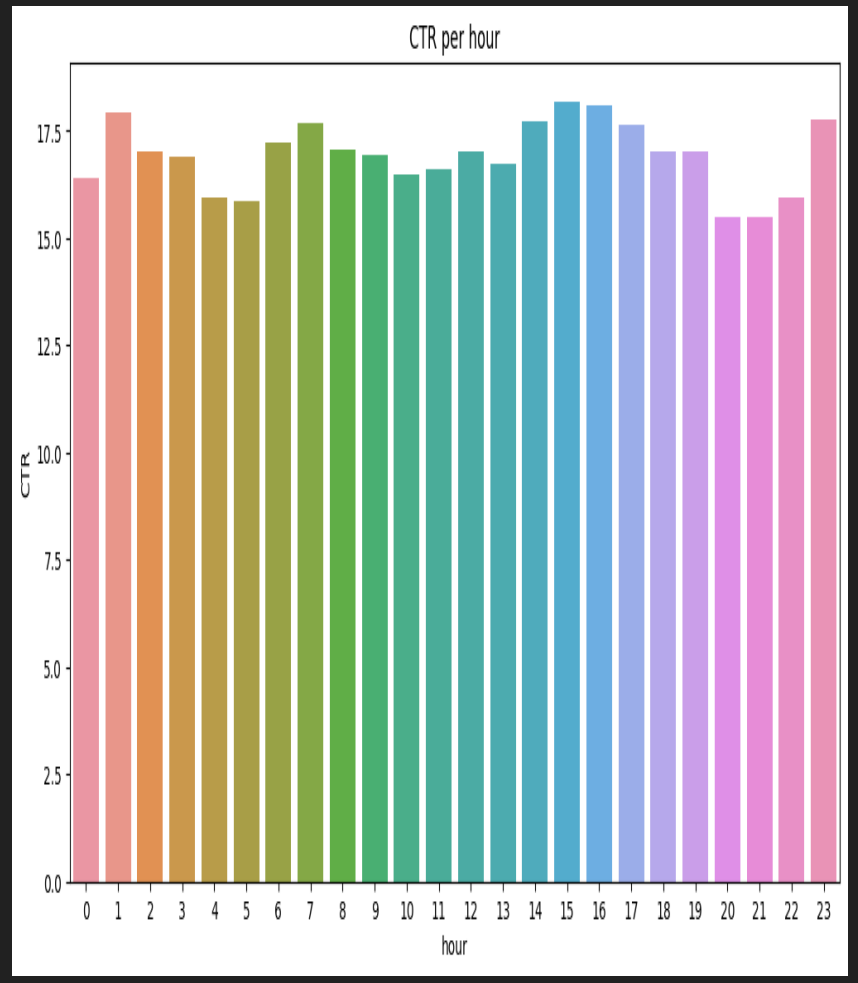
Analysis: Number of clicks per hour

- It can be seen that peak hours are at 9 and 13



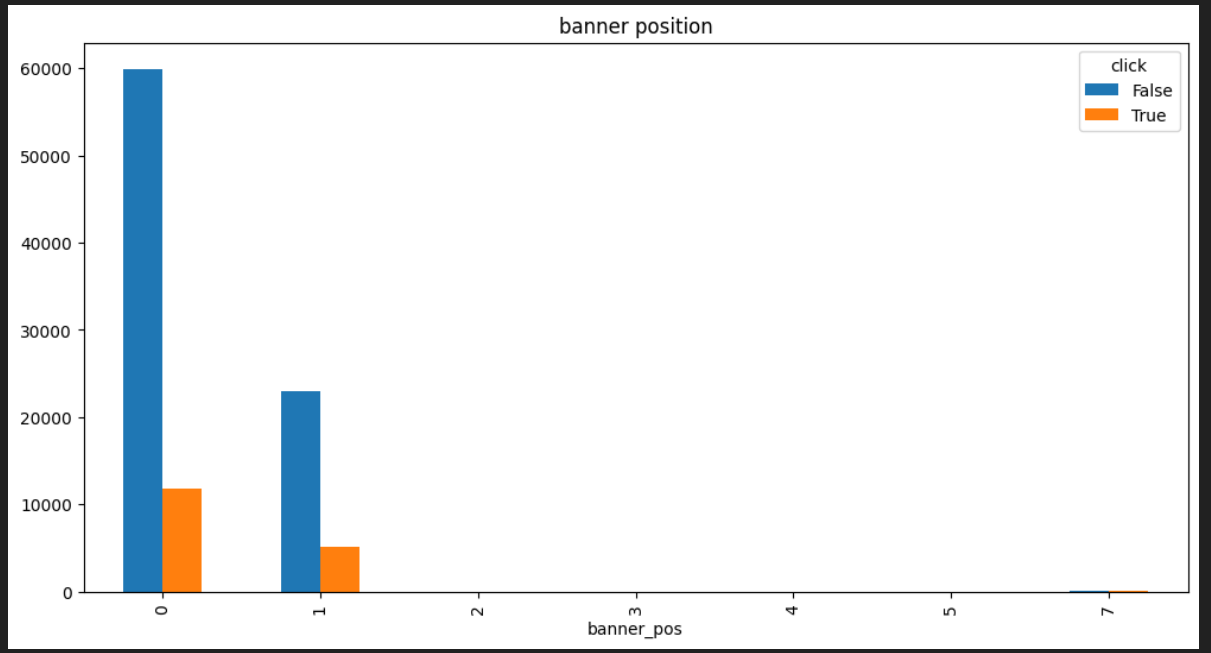
Analysis: CTR per hour

- From this analysis, we can say that CTR is higher at mid night hours.



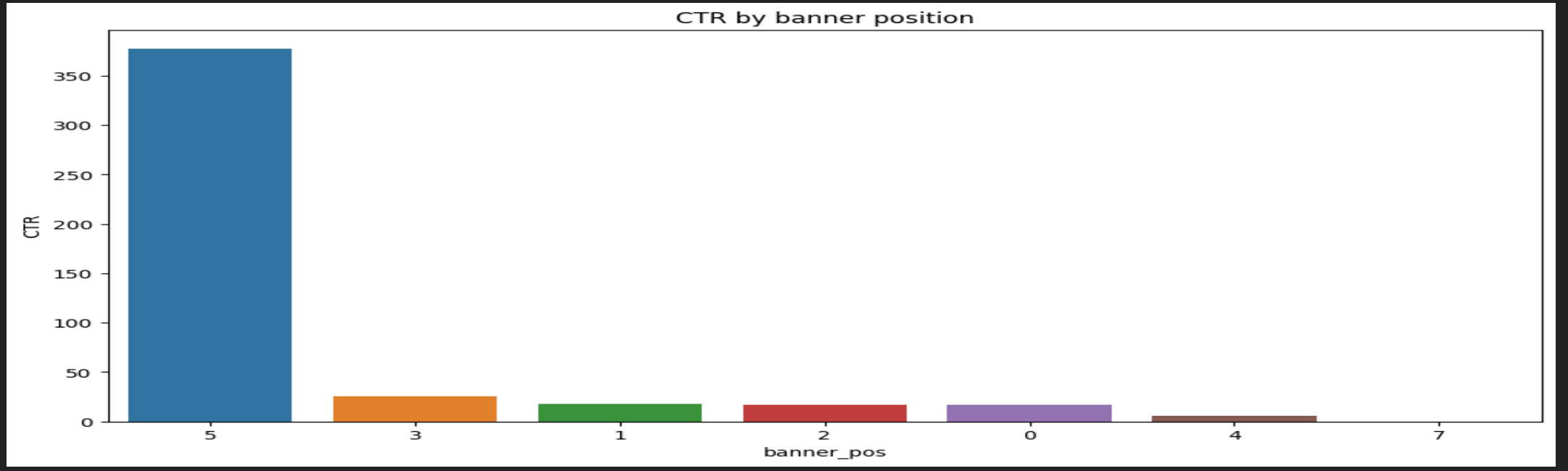
Analysis: Banner position

- We see that only two banner position i.e. 0 and 1 are significant



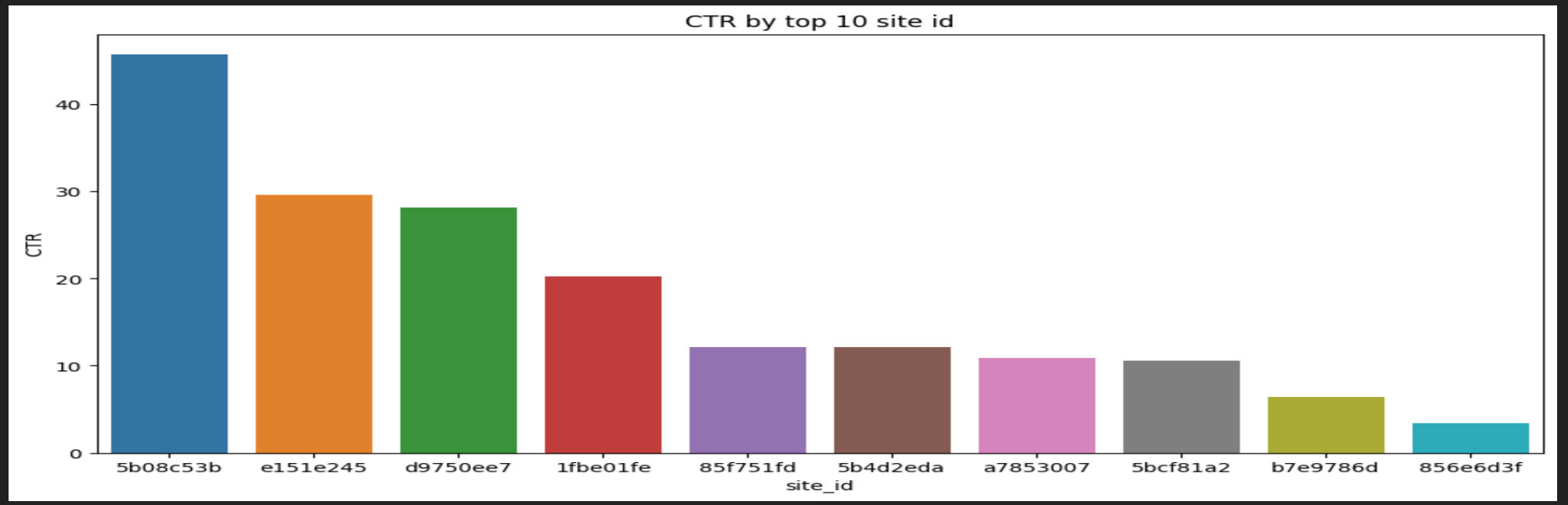
Analysis: CTR by banner position

- We see that position 5 has highest CTR



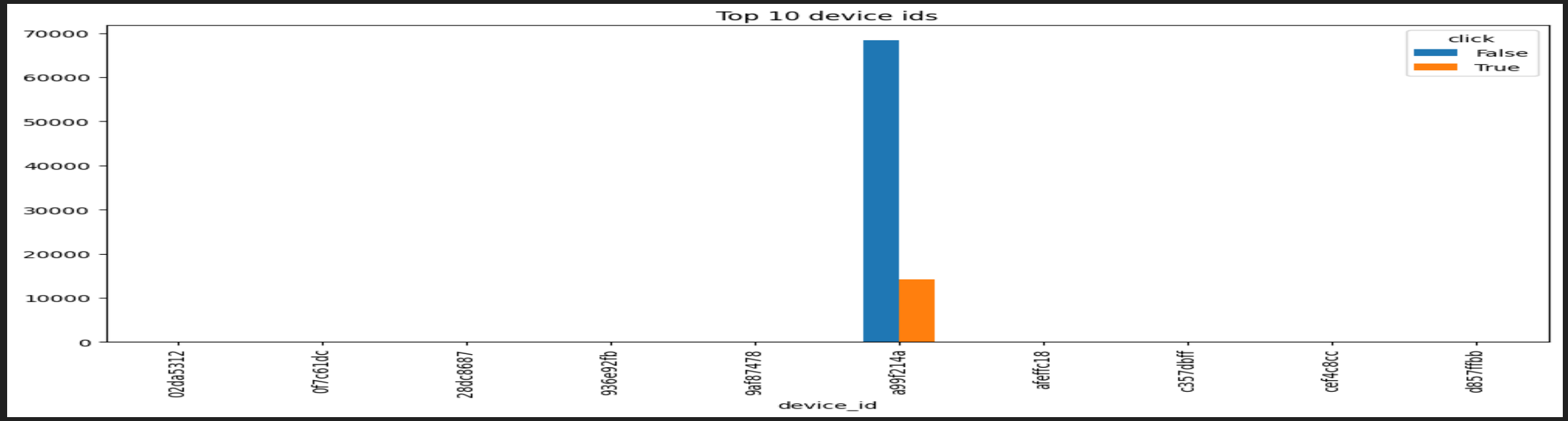
Analysis: CTR by top 10 site ids

- Analysis done by site id



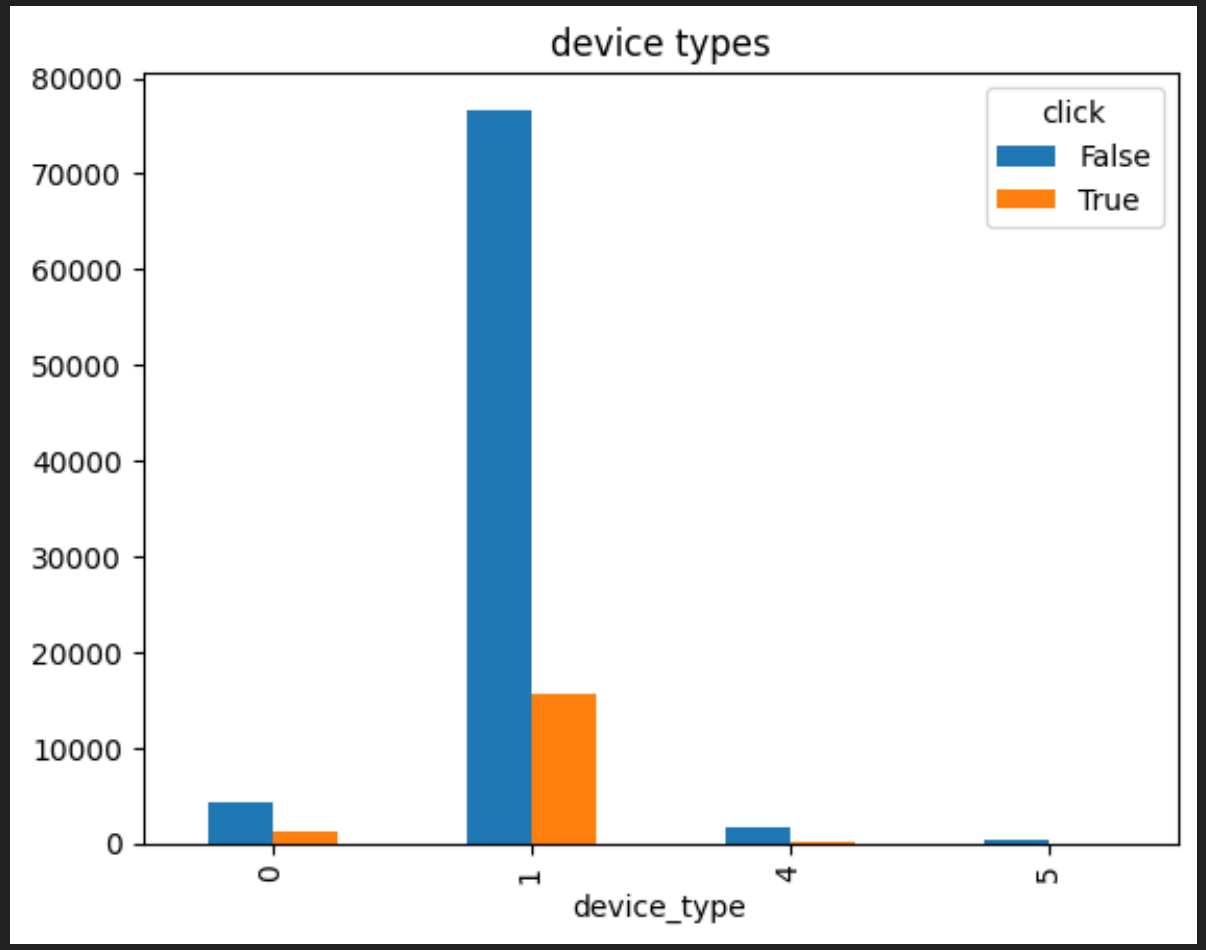
Analysis: Top 10 device id

- Analysis done by device ids



Analysis: Device type

- Analysis done by device type



Choosing Classification Method

- For this analysis, we have chosen below three methods –
  - Logistics Regression: It is used to find the probability of success and failure when the dependent variable is binary. It is very much easier to set up and train
  - Random Forest: Used due to its readability and interpretability. It has less chances of overfitting
  - Boosting: It reduces variance and bias. Algorithm work on combining weak learners into strong learners

Model Evaluation

- Logistics Regression has an AUC score 0.5 which is less
- Random Forest has recall value 1 which means 100% true positives are discovered
- Boosting also has recall value 1 which means Random Forest and Boosting performed better than Logistics Regression

Associated Risks

- Logistics Regression
  - It assumes linearity between dependent and independent variables
  - May not be accurate for small sample size
- Random Forest
  - Large number of trees can make algorithm very slow and ineffective
  - Not suitable for linear methods with a lot of sparse features
- Boosting:
  - Sensitive to outliers
  - It is almost impossible to scale up