Click Through Rate Prediction

Classification Assignment

2023/08/15

Topic

CONTENTS

Problem Statement

Most of the websites you visit include ads. The online advertising industry is huge, and players such as Google, Amazon, and Facebook generate billions of dollars by targeting the correct audiences with relevant ads. Most of the decisions about ads are data-driven solutions such as the following:

- How do you know which ad to use and who to target?
- Many companies advertise products in the same category, so how do you decide whose ad to display?
- Which ad should be placed on which part of the web page?
- Should a particular ad be pushed on a mobile device or remain on a desktop or laptop?

These decisions depend on numerous factors, including when the ad is placed, the site it is placed on, the characteristics of the people who will see the ad, the overall demographics, and more.

An important exercise marketing companies need to do before making any of the above decisions is a click-through rate (CTR) prediction. 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.

PART 01

Data Analysis and Preparation

Flow of the assignment

Raw Data	EDA	Data cleansing	Model Building	Results
Extracting data	Understanding the data	Removing unnecessary columns and rows	Building a machine learning predictive model	Finalizing the model with best accuracy metric value

EDA

Checking for Null Values

```
*******Check count of Null Values*
click                0
C1                   0
banner_pos           0
site_id              0
site_domain          0
site_category        0
app_id               0
app_domain           0
app_category         0
device_id            0
device_ip            0
device_model         0
device_type          0
device_conn_type     0
C14                  0
C15                  0
C16                  0
C17                  0
C18                  0
C19                  0
C20                  0
C21                  0
month                0
dayofweek            0
day                  0
hour                 0
y                    0
dtype: int64
```

Duplicate record Check

Checking duplicate rows

```
data[data.duplicated()].shape
```

```
(658, 27)
```

Shape of data after dropping duplicates

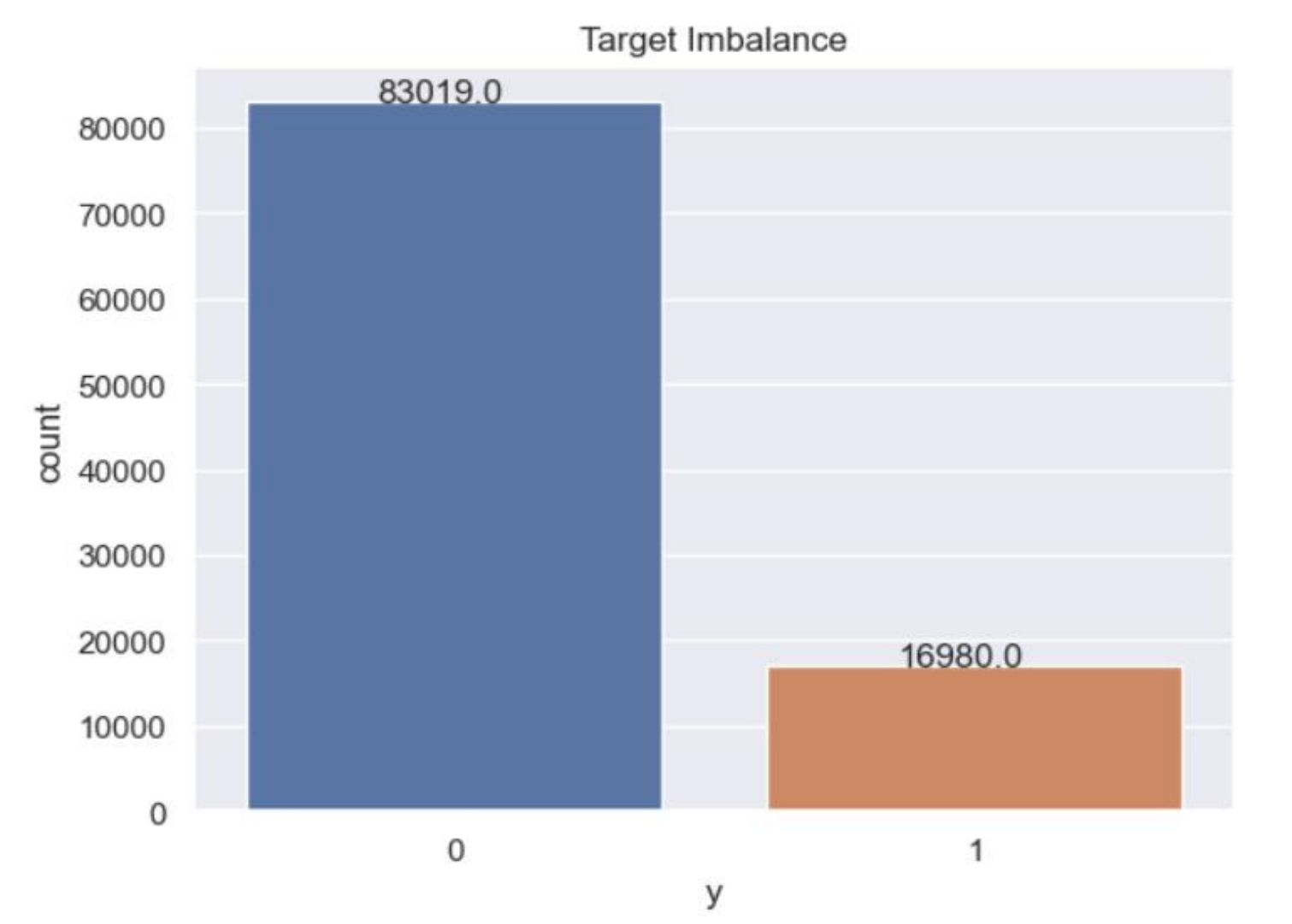
```
data[data.duplicated()].shape
```

```
(0, 27)
```

Unique values count

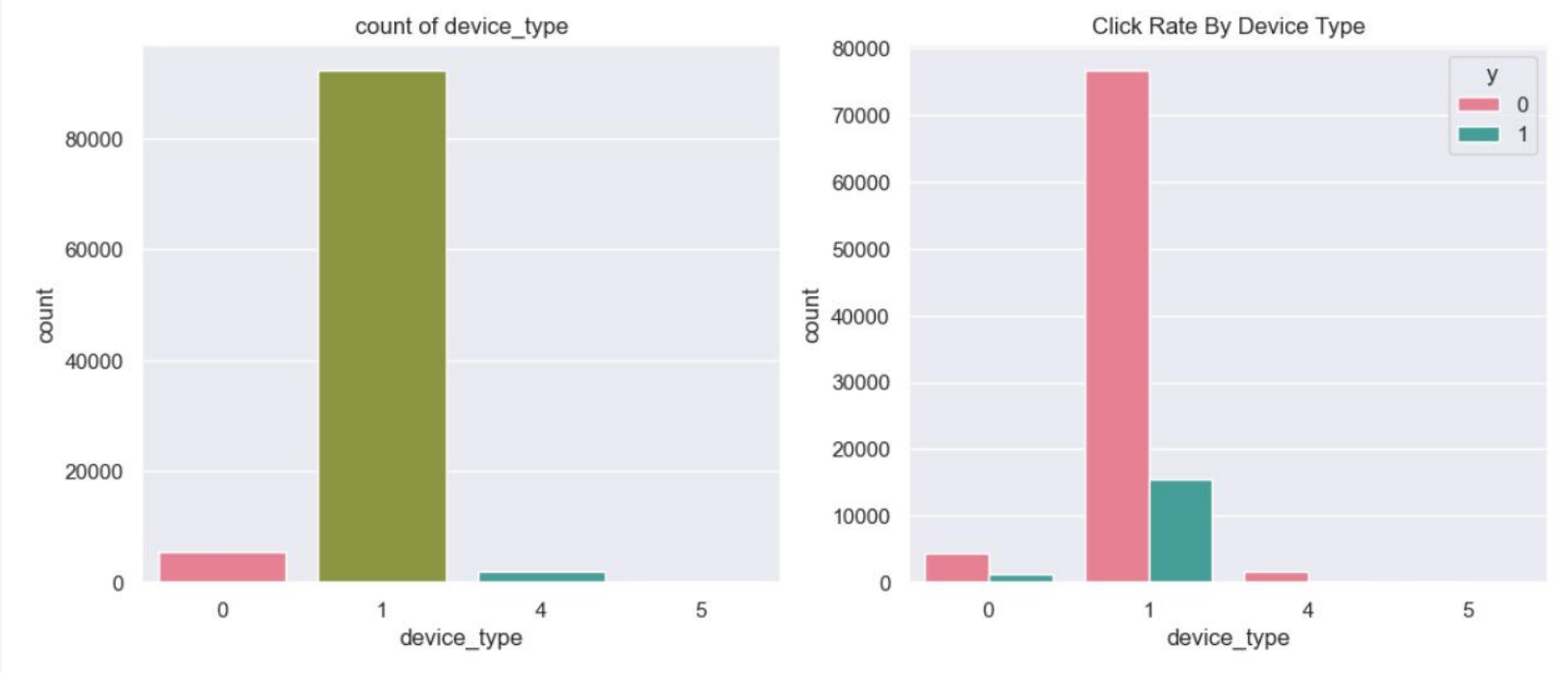
```
*******Check count of Unique Values*****
click                    2
C1                       7
banner_pos               7
site_id               1485
site_domain           1331
site_category           19
app_id                1354
app_domain              96
app_category            21
device_id            16801
device_ip            78013
device_model          3145
device_type              4
device_conn_type         4
C14                   1722
C15                      8
C16                      9
C17                    399
C18                      4
C19                     64
C20                    154
C21                     60
month                    1
dayofweek                7
day                     10
hour                    24
y                        2
dtype: int64
```

Checking for Target Data Imbalance

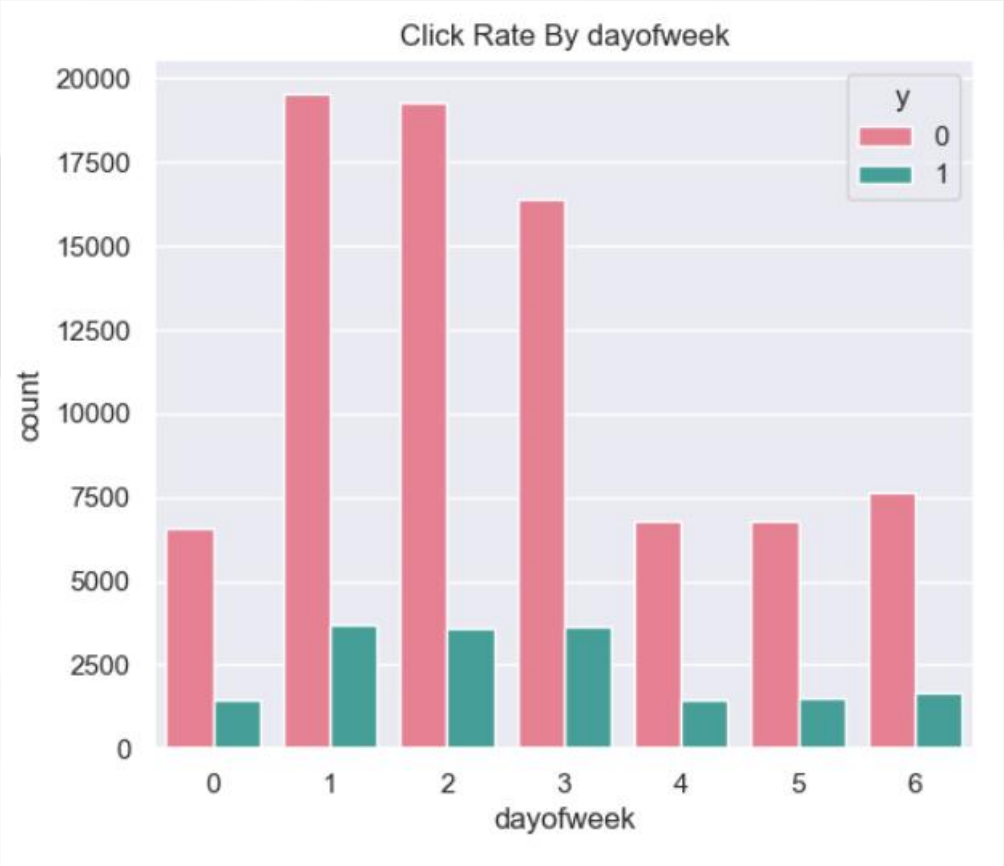


Understanding the click rate based on the different banner positions

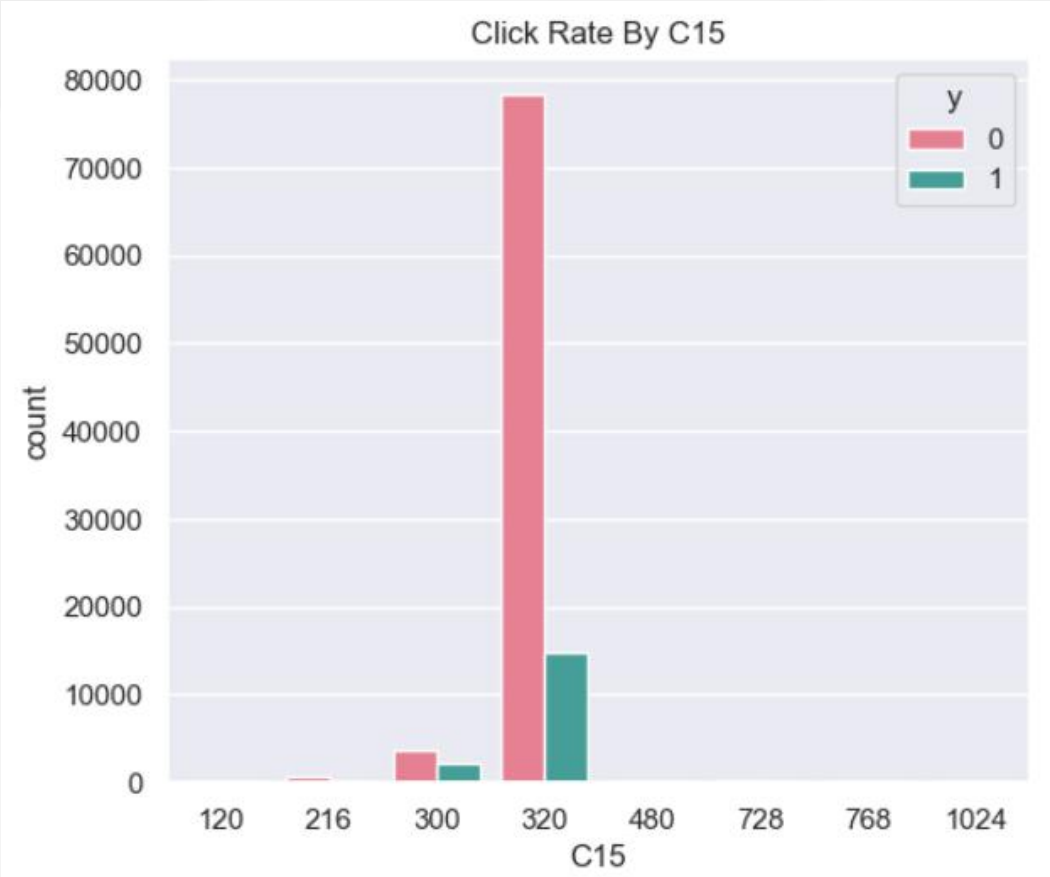
Getting insights on click rate based on the device type of the customer



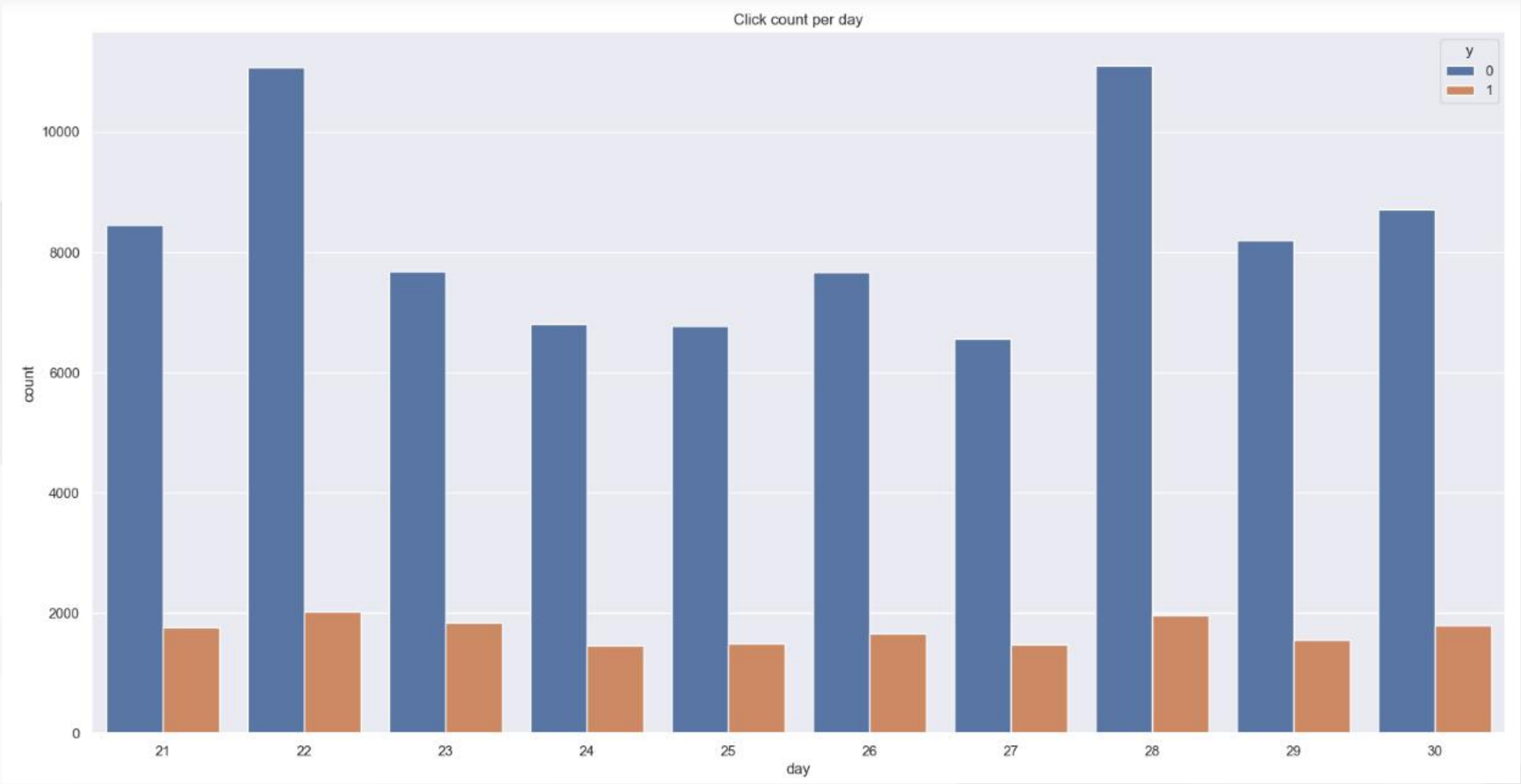
Click Rate by dayofweek



Click Rate by C15

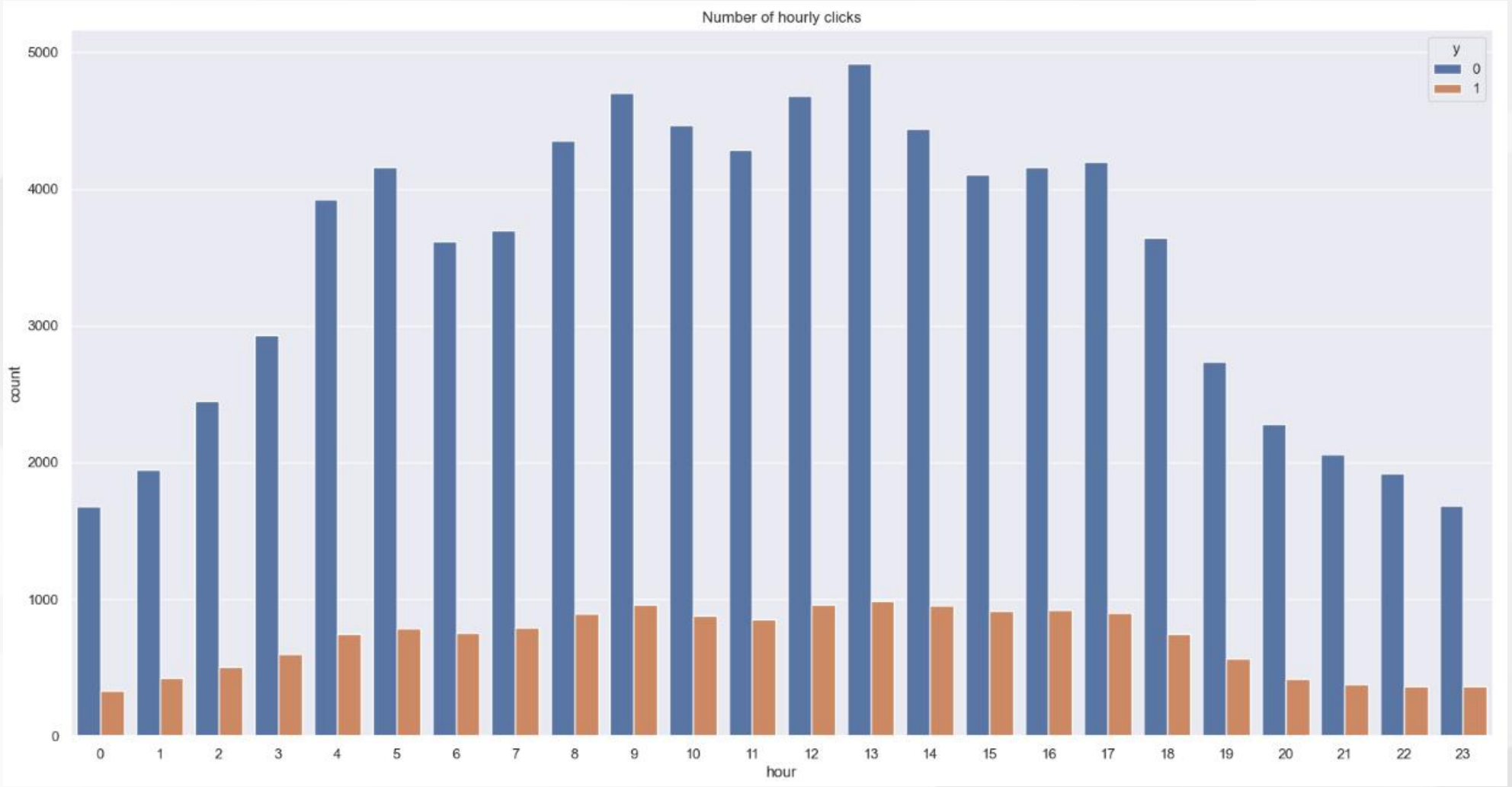


Click Count per day



This graph gives better understanding of clicks per day

Hourly Click Rate

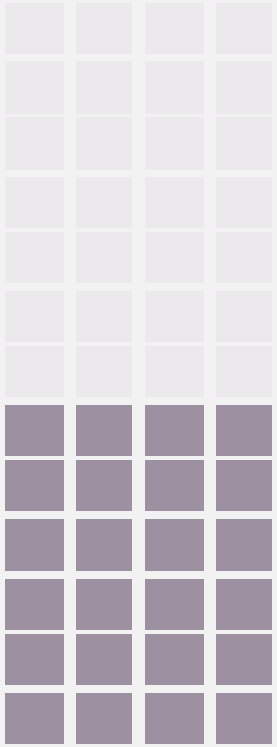


Deep understanding of click throughout the day for better understanding

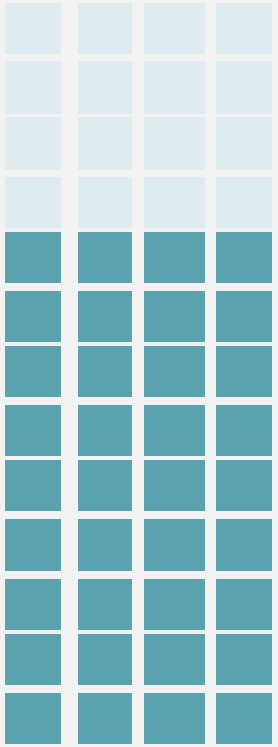
PART 02

Model Building

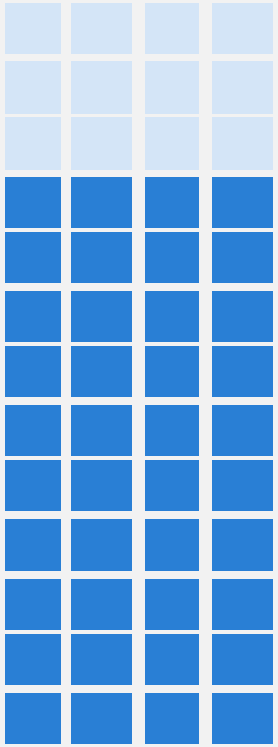
Machine Learning Models used



Logistic Regression



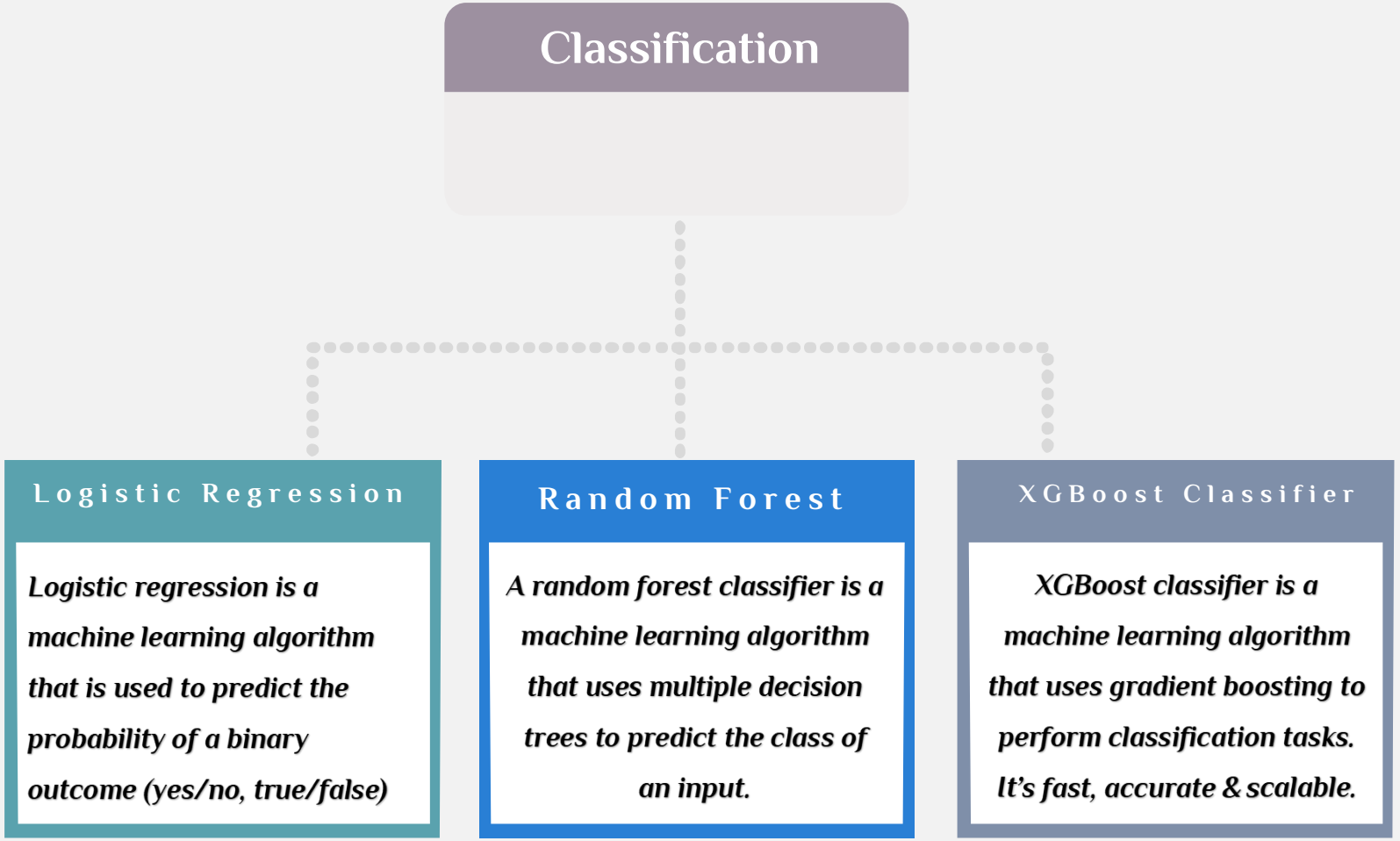
Decision Tree Classifier



Random Forest Classifier

Why Those 3 models

As our dataset has a categorical target dataset hence a classification based



PART 03

Conclusion

Comparing with the results of the model

	Accuracy	Recall	Precision	f1_score	ROC_AUC
Logistic regression	0.830051	0.003579	0.253521	0.007057	0.542474
DecisionTree Classifier	0.830051	0.003579	0.253521	0.007057	0.542474
RandomforestClassifier	0.796866	0.184095	0.321975	0.234252	0.584781
LogisticRegression_after_feature_selection	0.830084	0.003777	0.263889	0.007448	0.547672
DecisionTree_after_feature_selection	0.768782	0.207555	0.264371	0.232543	0.555099
RandomForest_after_feature_selection	0.795792	0.182704	0.317554	0.231954	0.582390
LogisticRegression_after_resampling	0.593299	0.601020	0.595328	0.598161	0.593264
DecisionTree_after_resampling	0.815892	0.888634	0.777579	0.829405	0.822714
RandomForest_after_resampling	0.832119	0.910168	0.788933	0.845225	0.840472

As per the results we can certainly say that after resampling models are performing better and Random forest Classifier is producing better results after resampling.

THANK YOU

Reporter: Kaustubh Nitin Patil