Data Mining Techniques for Click-Fraud Detection

Michael Albers

Utica University

DSC-680: Capstone Presentation

August 2023

What is Click Fraud?

Act of clicking on digital online ads that brings no benefit to the advertiser.

Who?	*How?*
Fraudulent Publisher websites	Crowdsourcing
Click Farms	Hit Inflation
Teams of Fraudsters	Botnets
Competitors	Publisher clicks on hosted ads

Detection Strategies

- Measure time duration between clicks
- Number of clicks in time bucket windows
- Click-Through Rate
- Total clicks by user and app
- Click Time Range

Exploratory Data Analysis

Predictor Variables

Target Variable

TalkingData - Data Set

- Provided by TalkingData Inc.
- Over 184 million click records
- Attributes: *click_time, app, channel, device, os, is_attributed*
- Target variable: is_attributed

  1 = Mobile App Downloaded

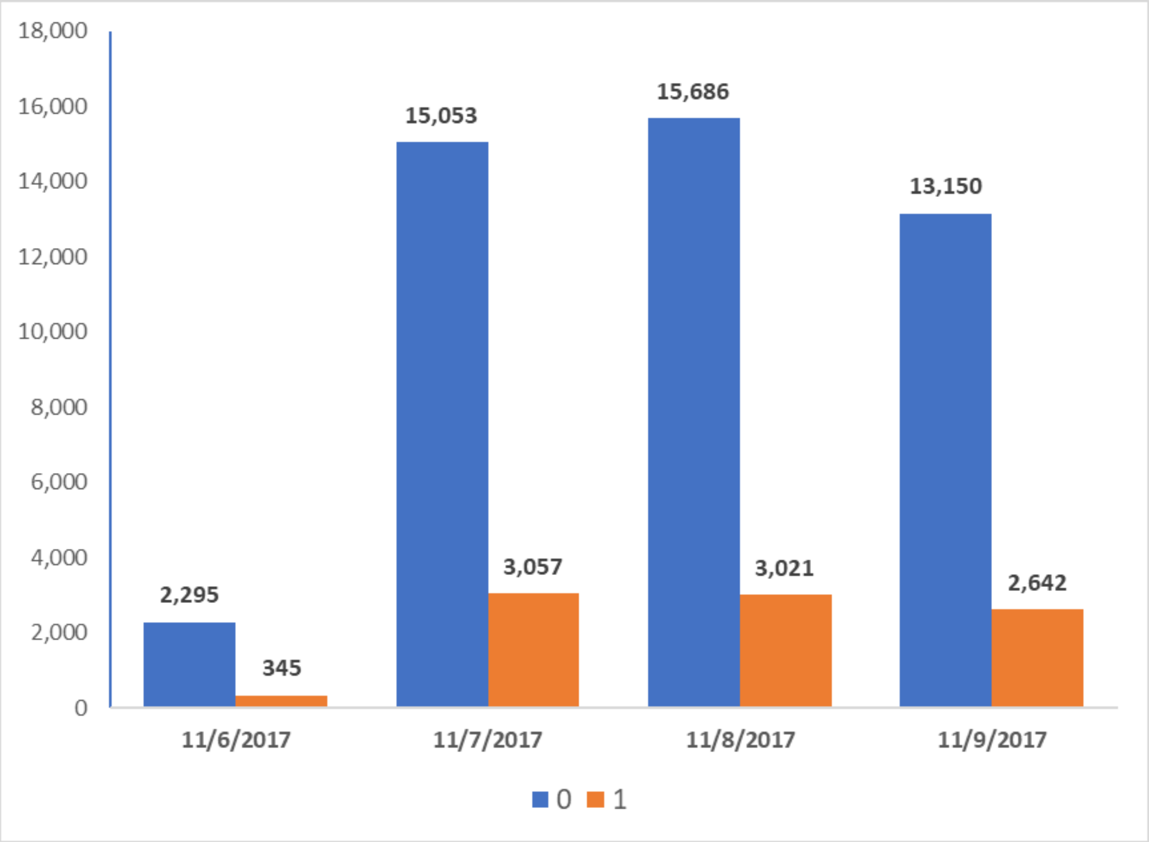
  0 = Mobile App Not Downloaded

Analytics Tools

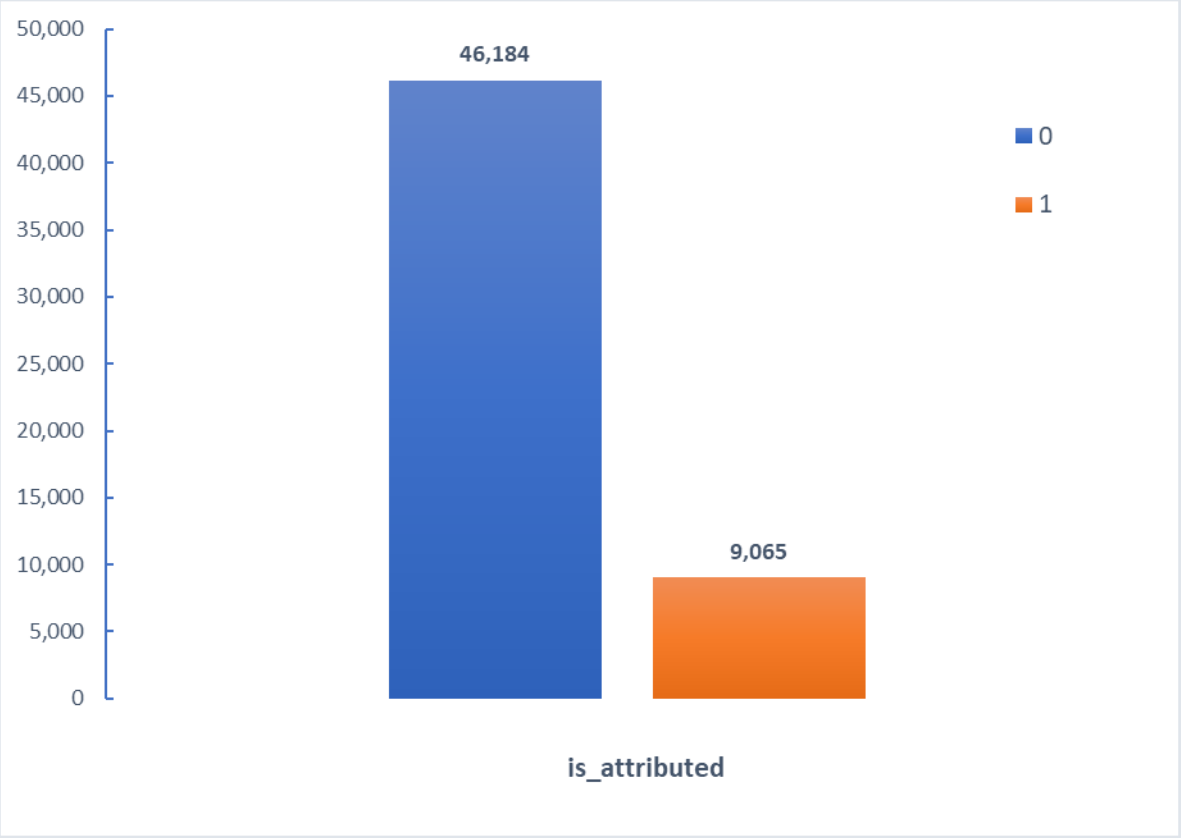
- Apache PySpark
- Google Storage, BigQuery
- Jupyter Notebooks on Google Colab
- Pandas
- Scikit-Learn
- Matplotlib and Seaborn Visualization Libraries

Stratified Sample

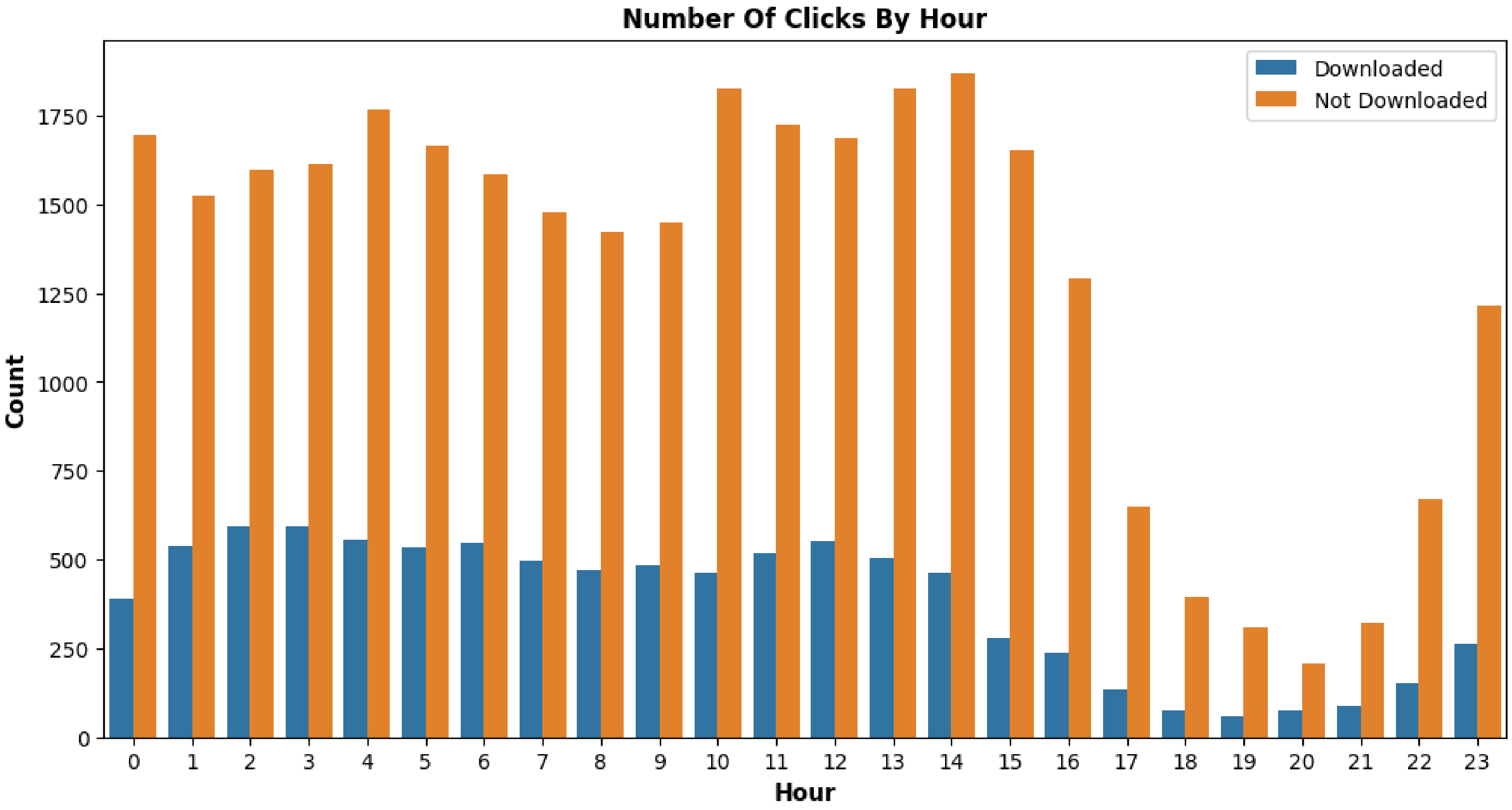
Total of 55,249 click observations



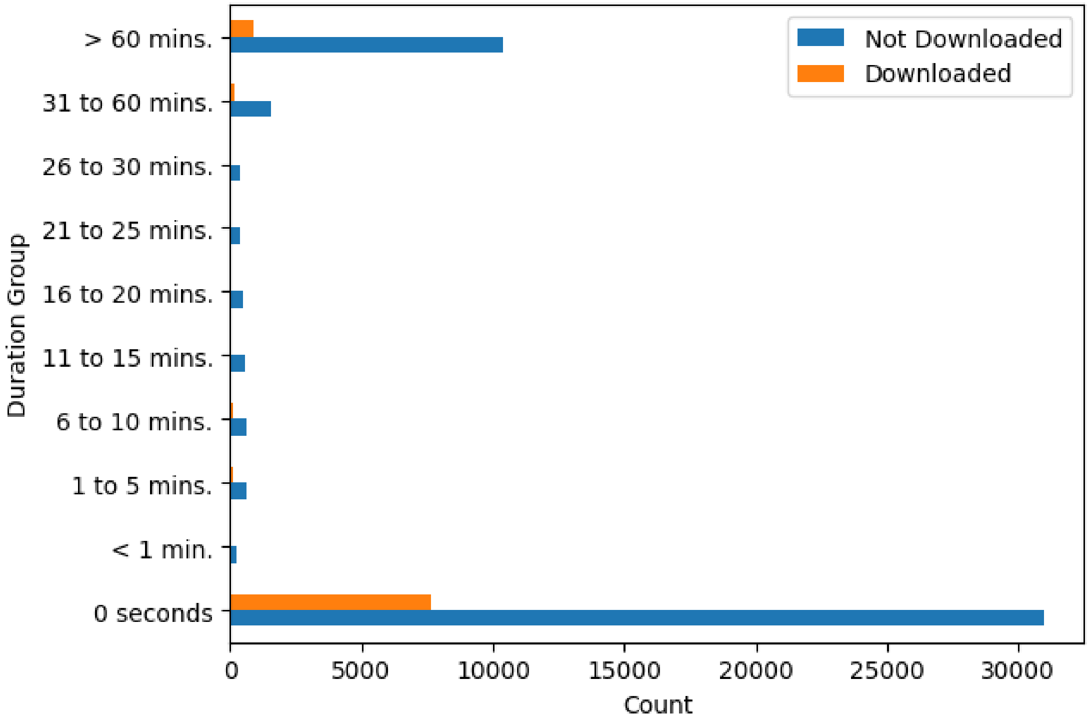
Target Class

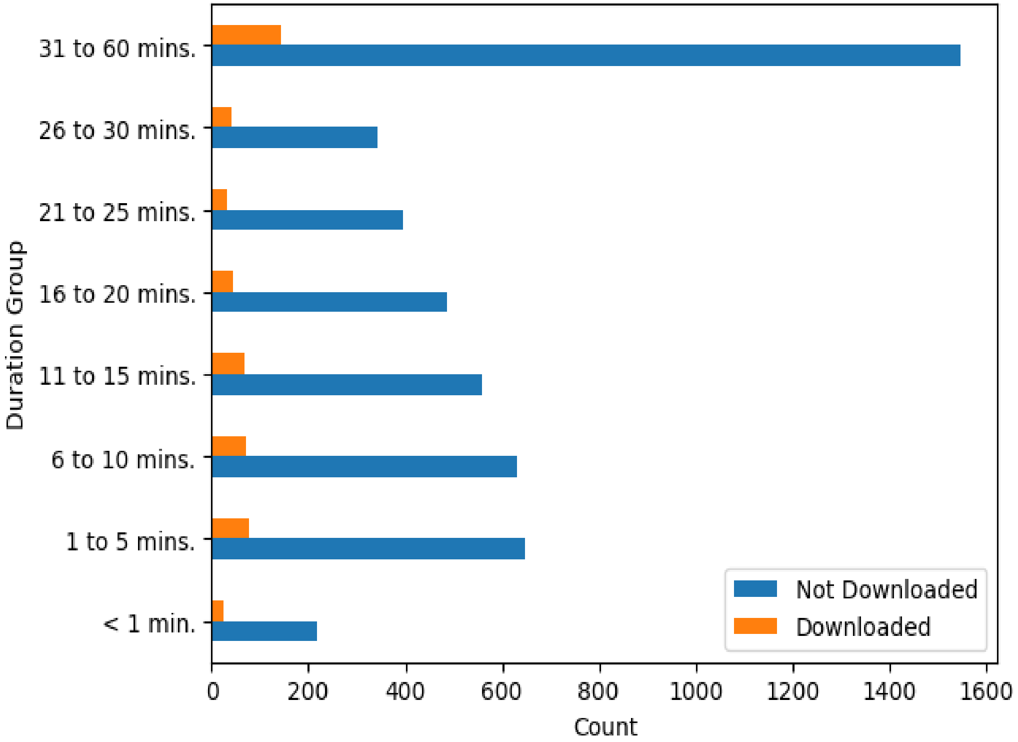


Hourly Clicks



Click Duration

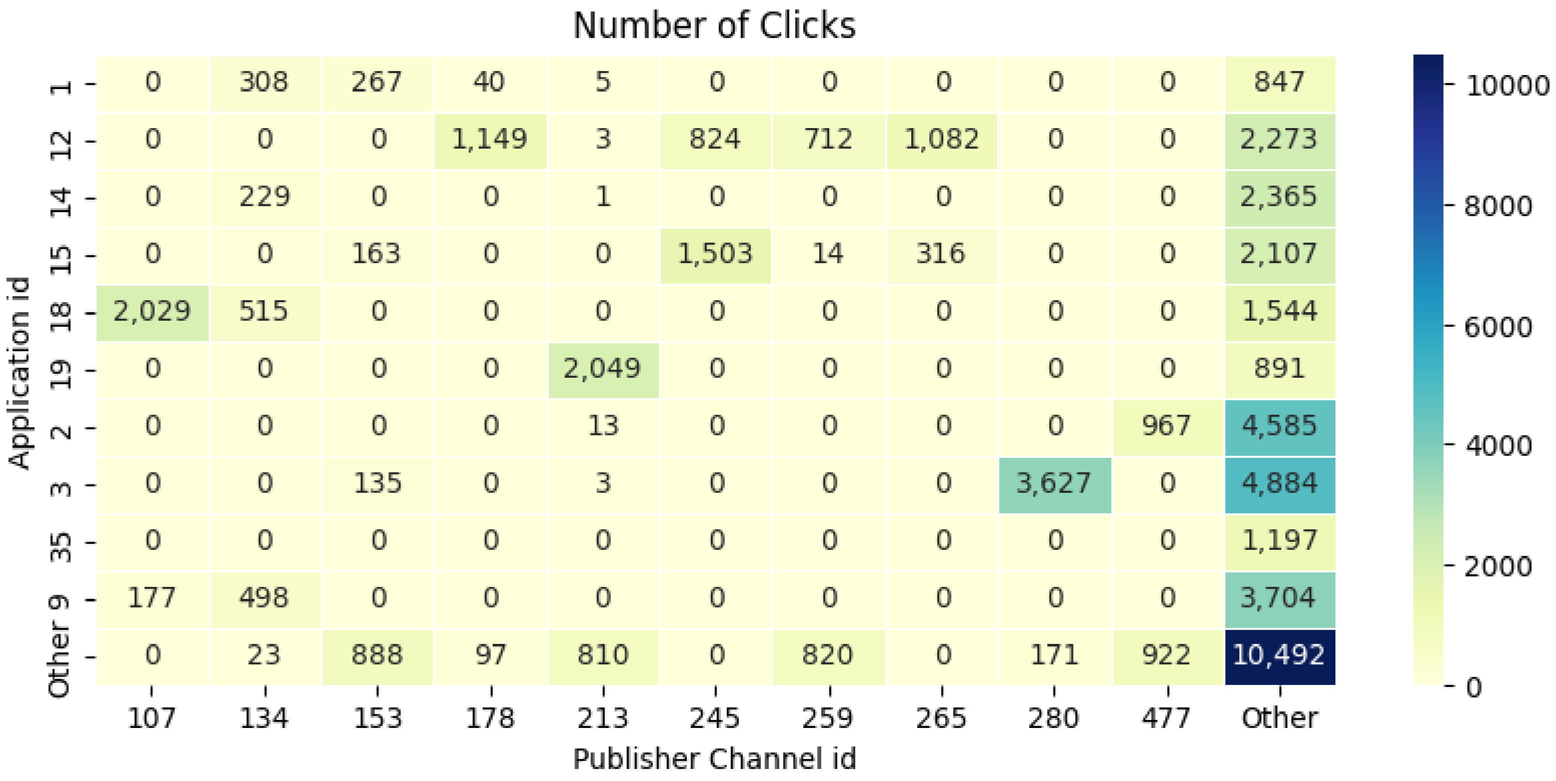




Categorical Variables

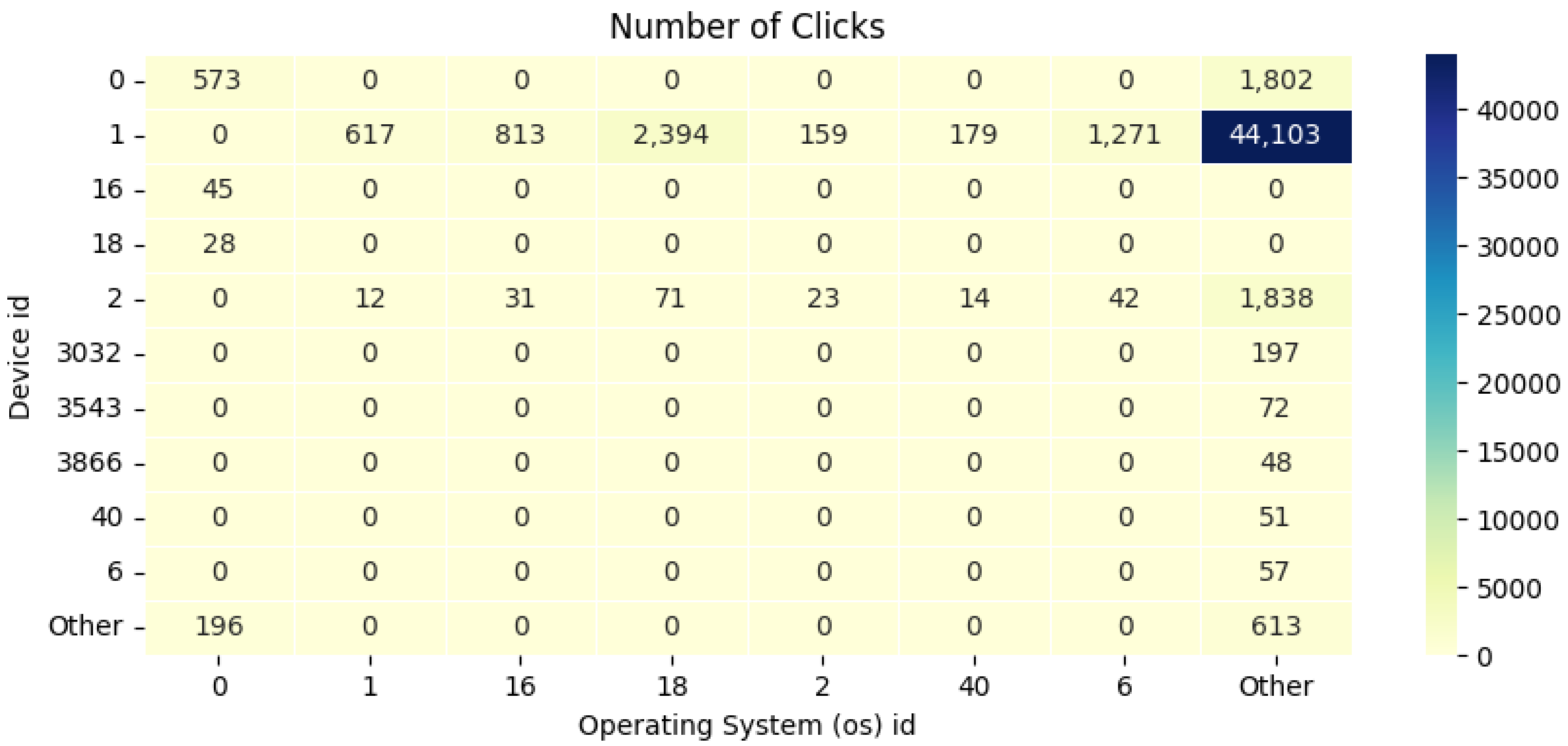
App, Publisher Channel, Device, Operating System

Clicks by App and Channel


Number of Clicks

Application id \ Publisher Channel id	107	134	153	178	213	245	259	265	280	477	Other
1	0	308	267	40	5	0	0	0	0	0	847
12	0	0	0	1,149	3	824	712	1,082	0	0	2,273
14	0	229	0	0	1	0	0	0	0	0	2,365
15	0	0	163	0	0	1,503	14	316	0	0	2,107
18	2,029	515	0	0	0	0	0	0	0	0	1,544
19	0	0	0	0	2,049	0	0	0	0	0	891
2	0	0	0	0	13	0	0	0	0	967	4,585
3	0	0	135	0	3	0	0	0	3,627	0	4,884
35	0	0	0	0	0	0	0	0	0	0	1,197
9	177	498	0	0	0	0	0	0	0	0	3,704
Other	0	23	888	97	810	0	820	0	171	922	10,492

Clicks by Device and Operating System

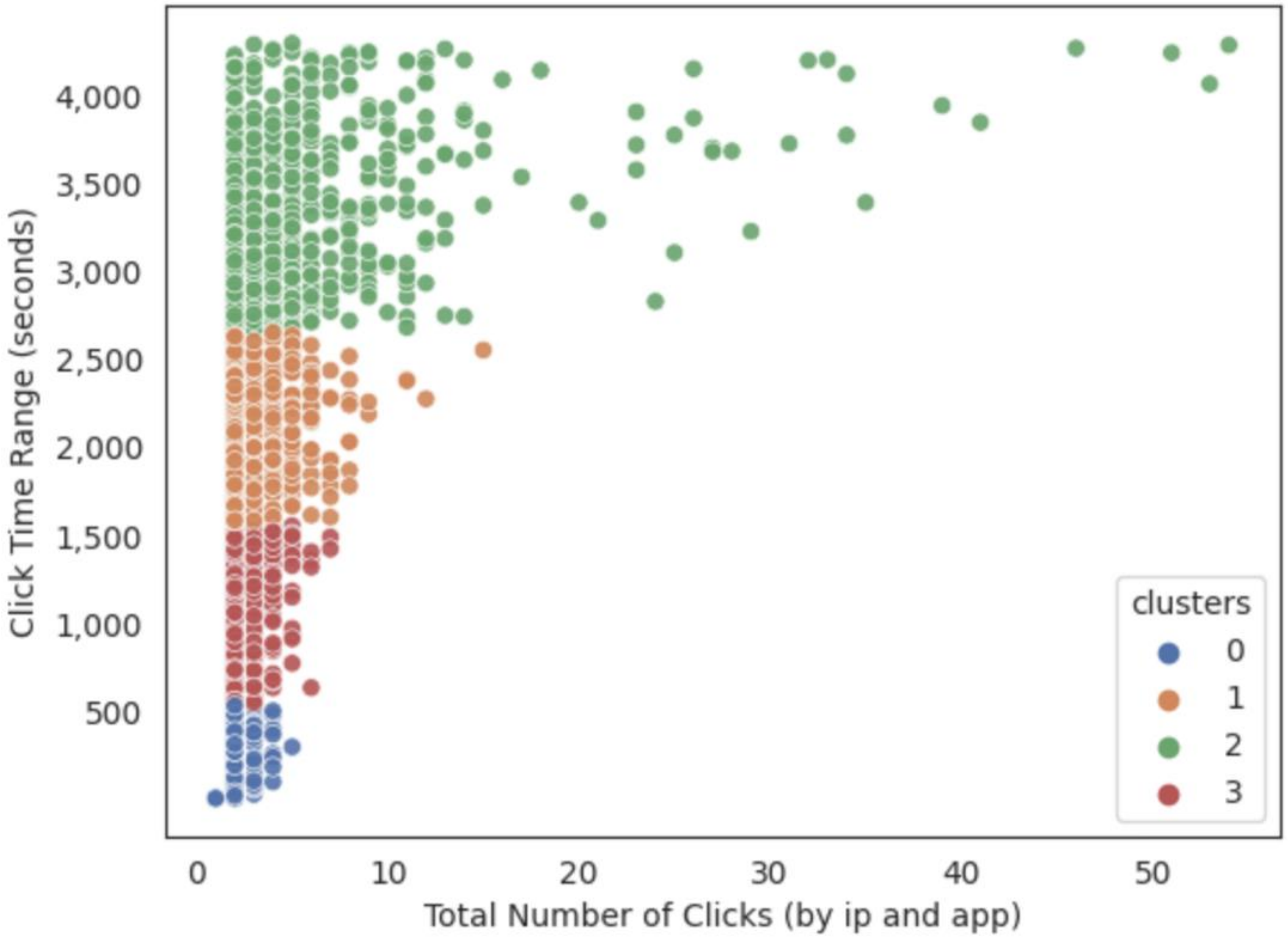


Device id	0	1	16	18	2	40	6	Other
0	573	0	0	0	0	0	0	1,802
1	0	617	813	2,394	159	179	1,271	44,103
16	45	0	0	0	0	0	0	0
18	28	0	0	0	0	0	0	0
2	0	12	31	71	23	14	42	1,838
3032	0	0	0	0	0	0	0	197
3543	0	0	0	0	0	0	0	72
3866	0	0	0	0	0	0	0	48
40	0	0	0	0	0	0	0	51
6	0	0	0	0	0	0	0	57
Other	196	0	0	0	0	0	0	613

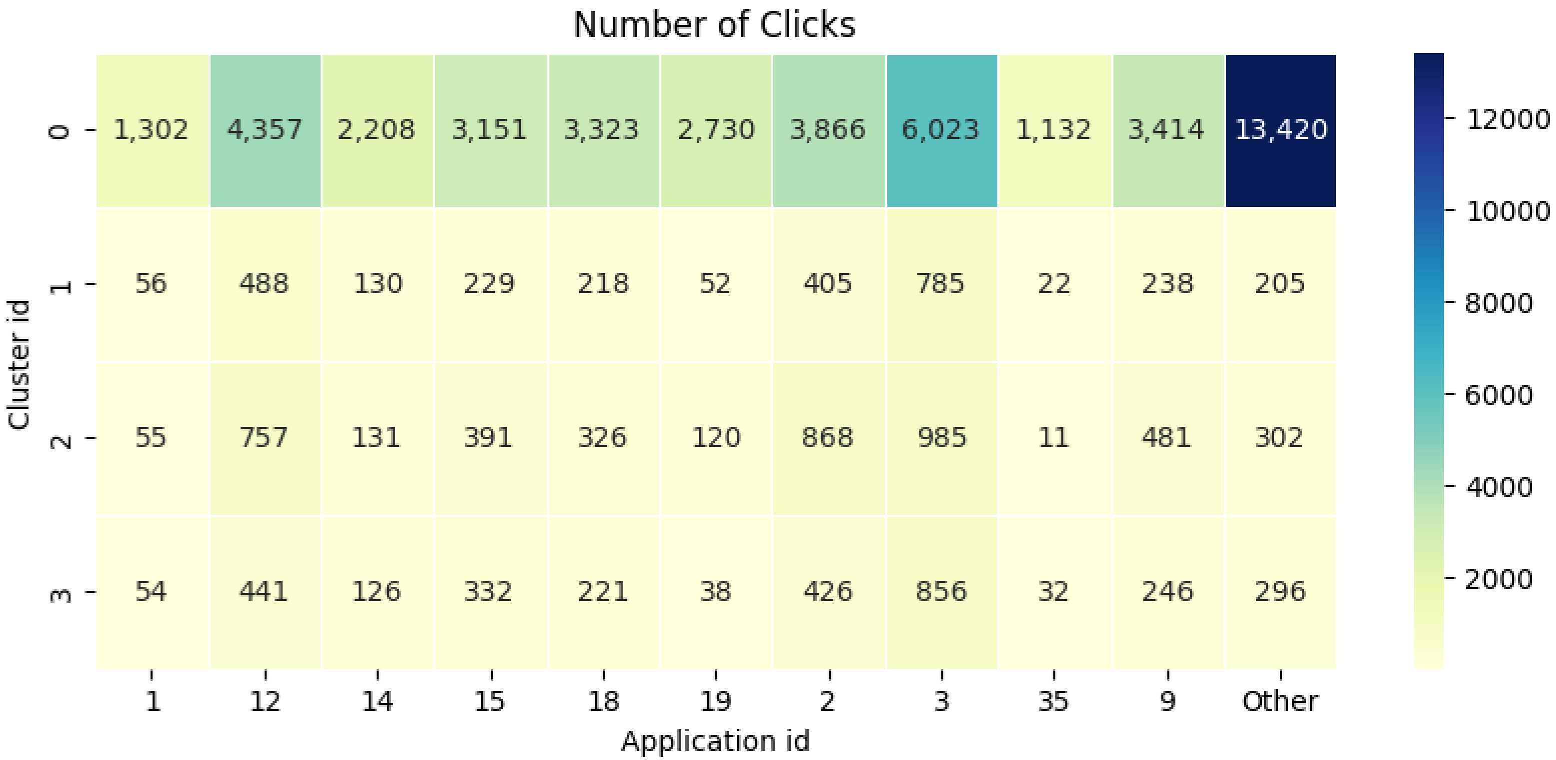
Clustering

K-Means Clustering with four clusters

Clusters Plot

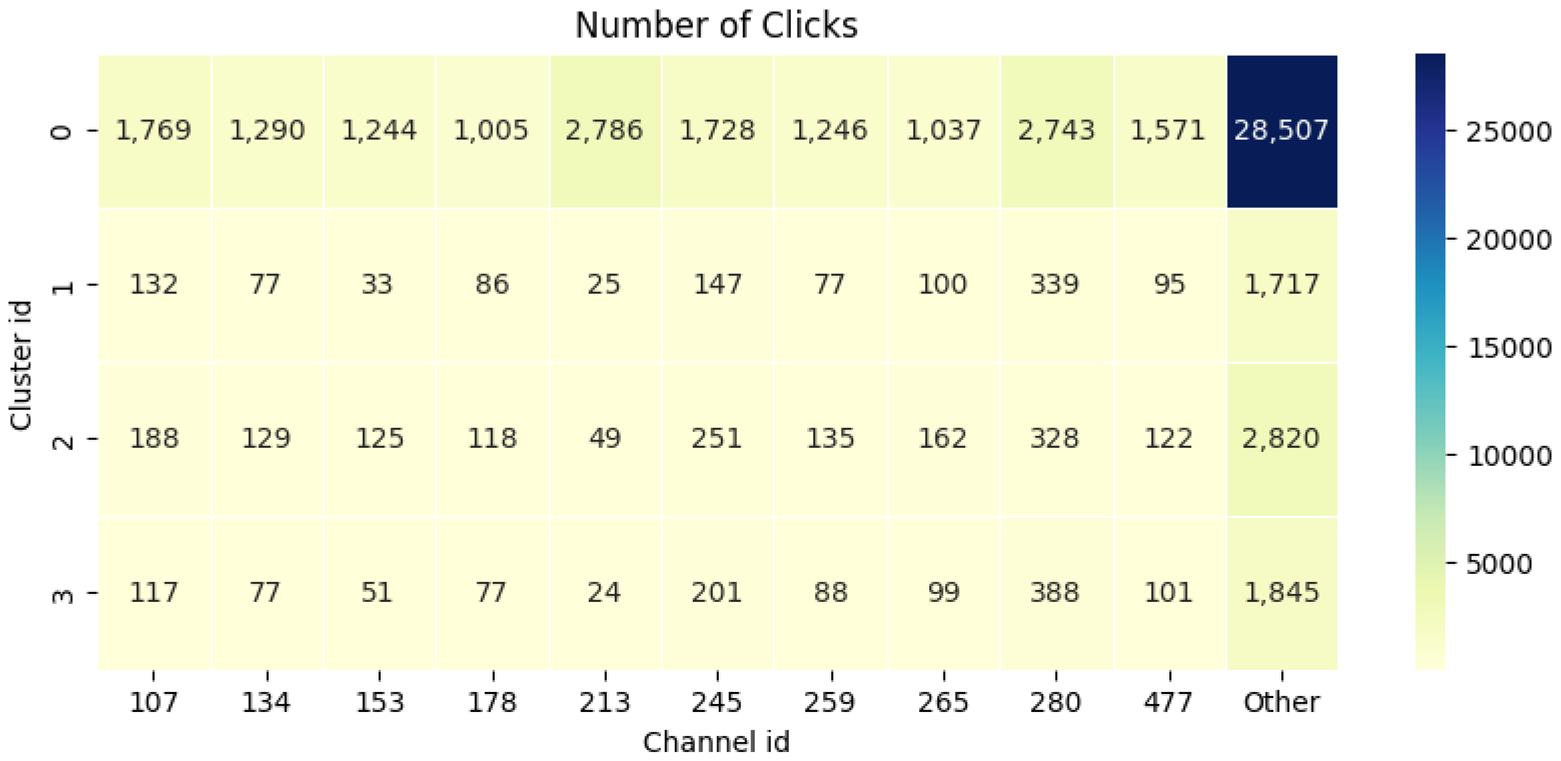


Heatmap - Clusters and App



Cluster id	1	12	14	15	18	19	2	3	35	9	Other
0	1,302	4,357	2,208	3,151	3,323	2,730	3,866	6,023	1,132	3,414	13,420
1	56	488	130	229	218	52	405	785	22	238	205
2	55	757	131	391	326	120	868	985	11	481	302
3	54	441	126	332	221	38	426	856	32	246	296

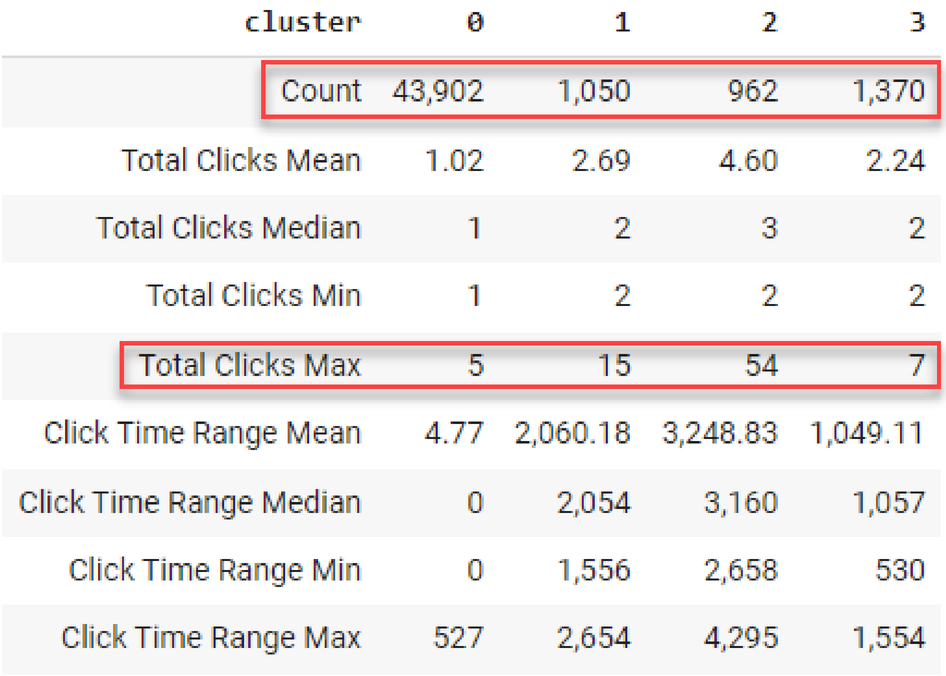
Heatmap – Clusters and Channel



Number of Clicks

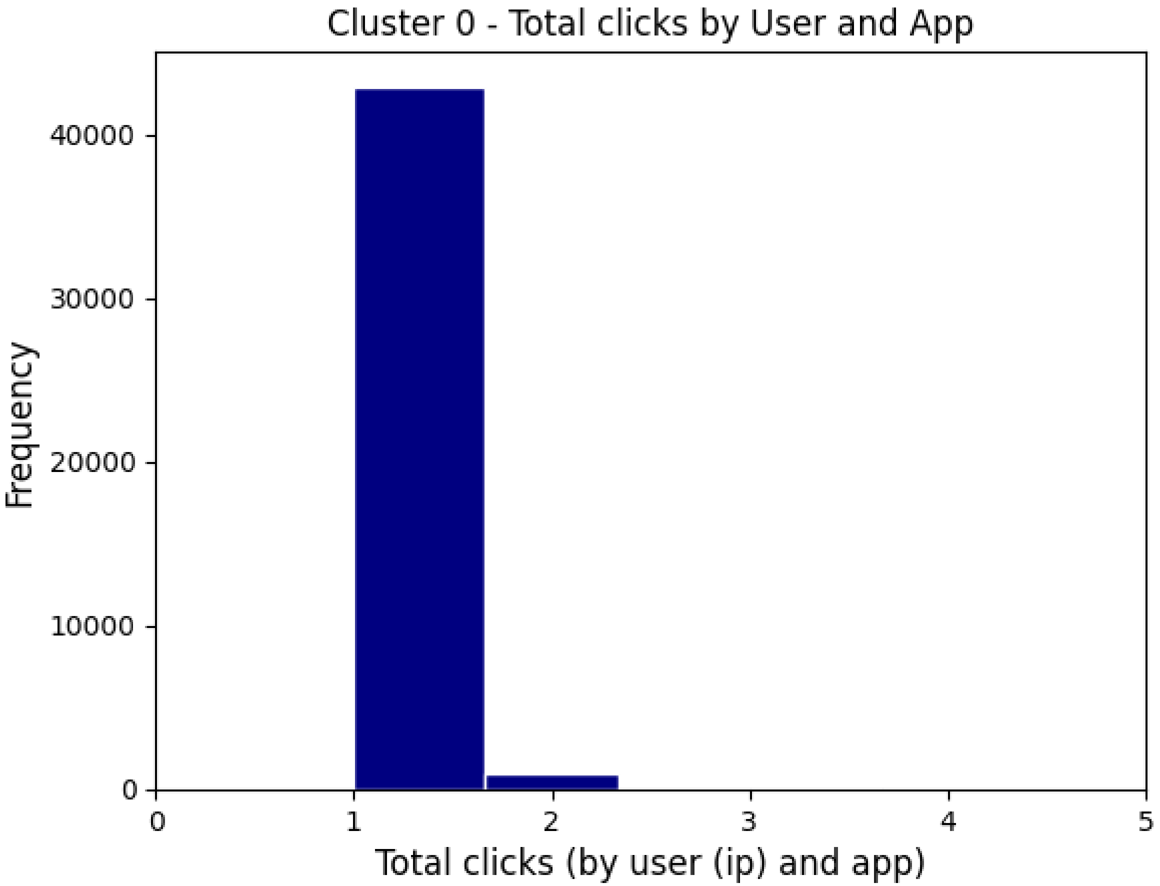
Cluster id	107	134	153	178	213	245	259	265	280	477	Other
0	1,769	1,290	1,244	1,005	2,786	1,728	1,246	1,037	2,743	1,571	28,507
1	132	77	33	86	25	147	77	100	339	95	1,717
2	188	129	125	118	49	251	135	162	328	122	2,820
3	117	77	51	77	24	201	88	99	388	101	1,845

K-Means Results

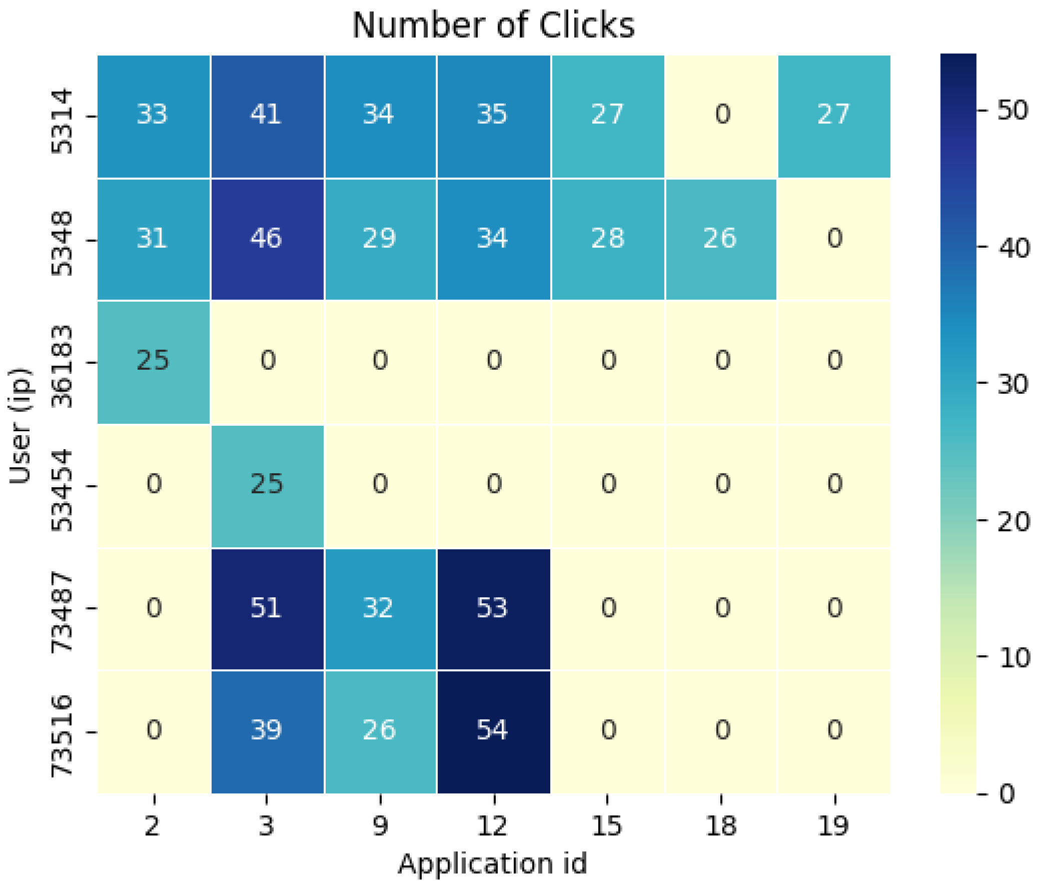


cluster	0	1	2	3
Count	43,902	1,050	962	1,370
Total Clicks Mean	1.02	2.69	4.60	2.24
Total Clicks Median	1	2	3	2
Total Clicks Min	1	2	2	2
Total Clicks Max	5	15	54	7
Click Time Range Mean	4.77	2,060.18	3,248.83	1,049.11
Click Time Range Median	0	2,054	3,160	1,057
Click Time Range Min	0	1,556	2,658	530
Click Time Range Max	527	2,654	4,295	1,554

Cluster 0 - “One-Time Clicks”



Cluster 2 - “Outliers”



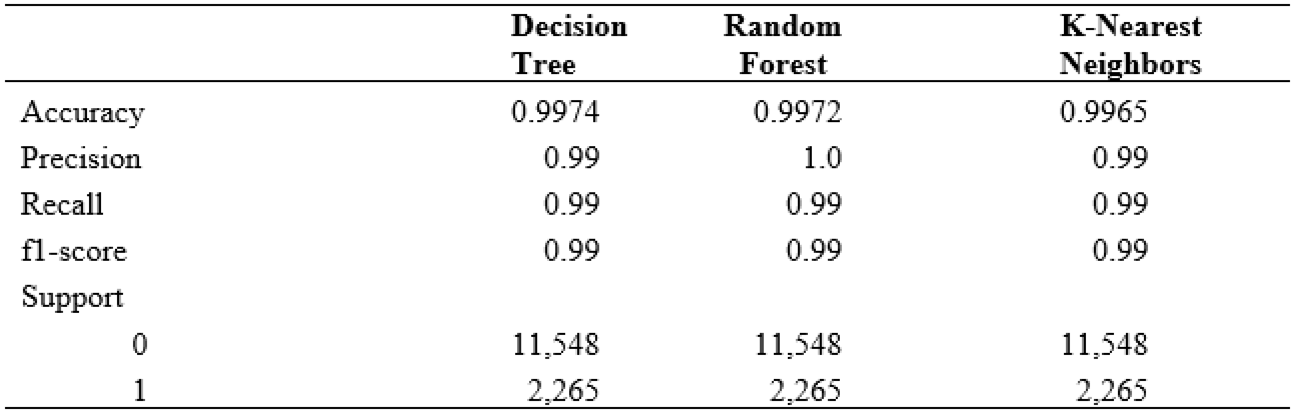
Number of Clicks

User (ip)	2	3	9	12	15	18	19
5314	33	41	34	35	27	0	27
5348	31	46	29	34	28	26	0
36183	25	0	0	0	0	0	0
53454	0	25	0	0	0	0	0
73487	0	51	32	53	0	0	0
73516	0	39	26	54	0	0	0

Classification Results

Decision Tree, Random Forest, K-Nearest Neighbors

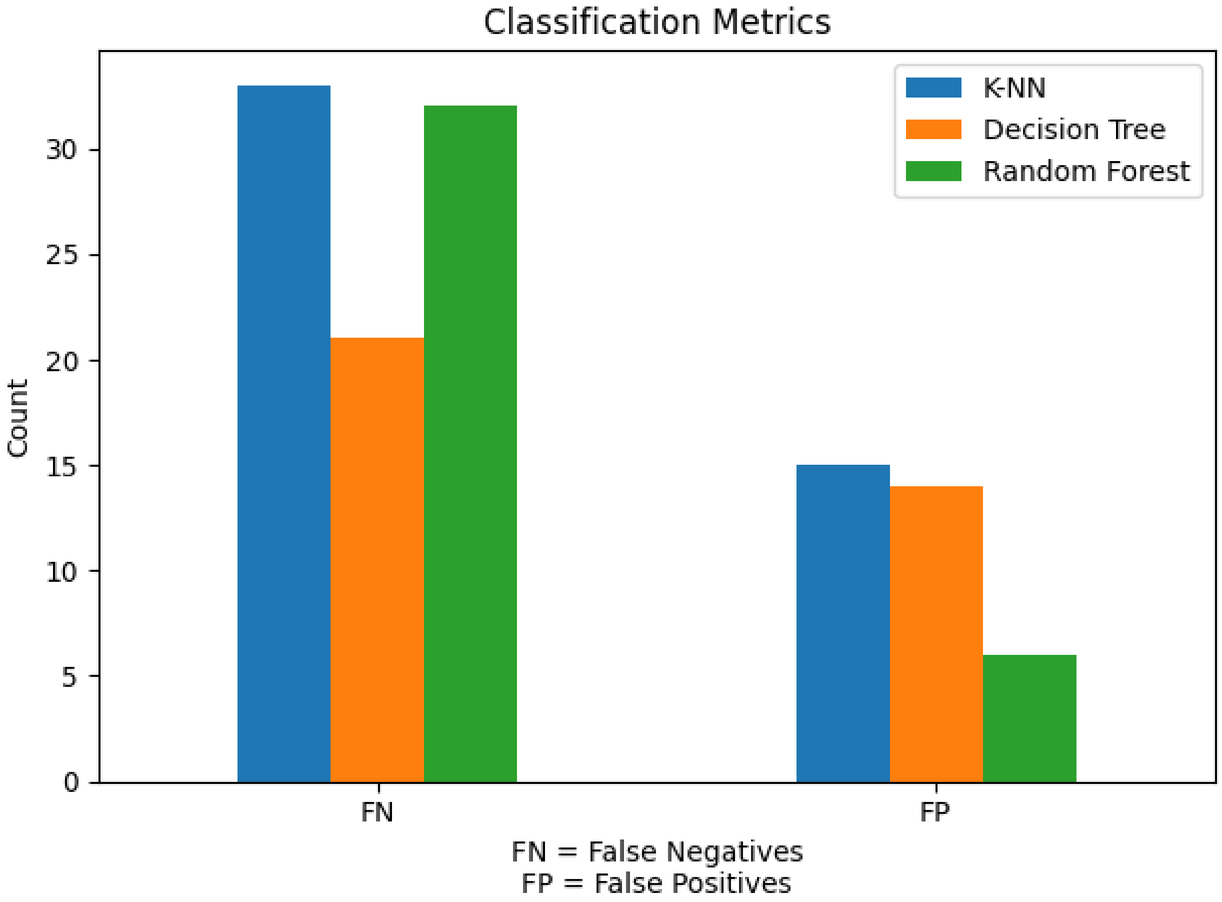
Classification Metrics



	Decision Tree	**Random Forest**	**K-Nearest Neighbors**
Accuracy	0.9974	0.9972	0.9965
Precision	0.99	1.0	0.99
Recall	0.99	0.99	0.99
f1-score	0.99	0.99	0.99
Support			
0	11,548	11,548	11,548
1	2,265	2,265	2,265

Note: Includes cluster variables

Misclassification Counts



Further Research

- Implement repeated sampling of raw data set
- Increase clusters count for K-Means Clustering
- Use additional features for K-Means Clustering
- Use Over and Under-sampling techniques to balance target classes

References

Yin, A., Kleinman, J., Elliott, J. & Yan, T. TalkingData AdTracking Fraud Detection Challenge. Kaggle. 2018. Retrieved from https://kaggle.com/competitions/talkingdata-adtracking-fraud-detection.