

TEAM 2- BFS TRACK



FRAUD CLICK DETECTION WEBAPP

Rahul Tomar

Dipanshu Upreti

Vaishnavi Subramanya Desai

Girish Kumar Reddy Tokala

Ragini Gaurav

Team Lead

Python Track

Python Track

Python Track

UI Track

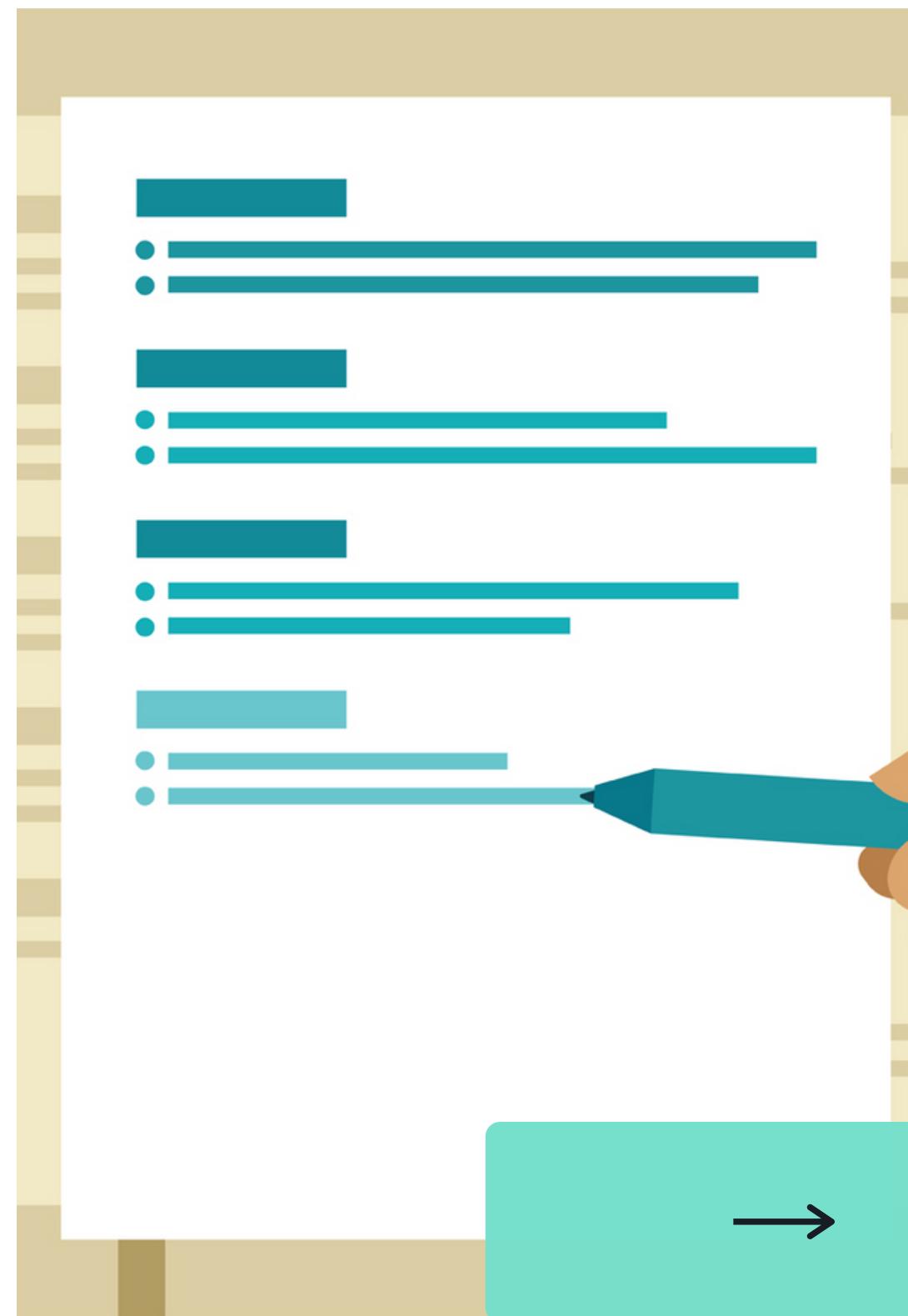
CLICK!
CLICK!
CLICK!



Presentation Outline

TOPICS FOR DISCUSSION

- The Problem!
- Approach
- The Solution
- Implementation

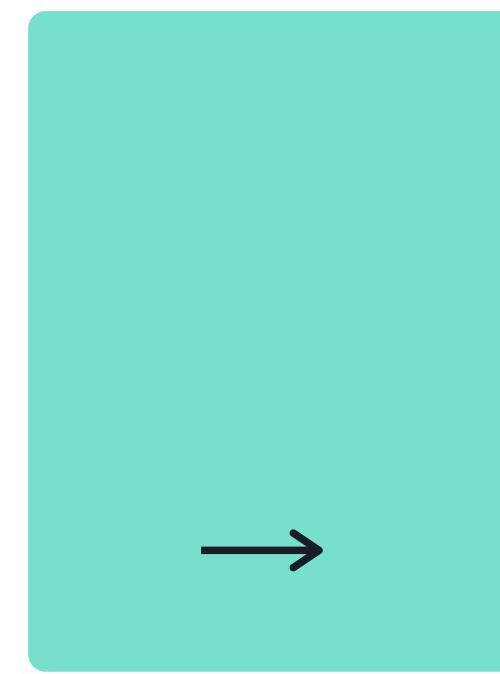
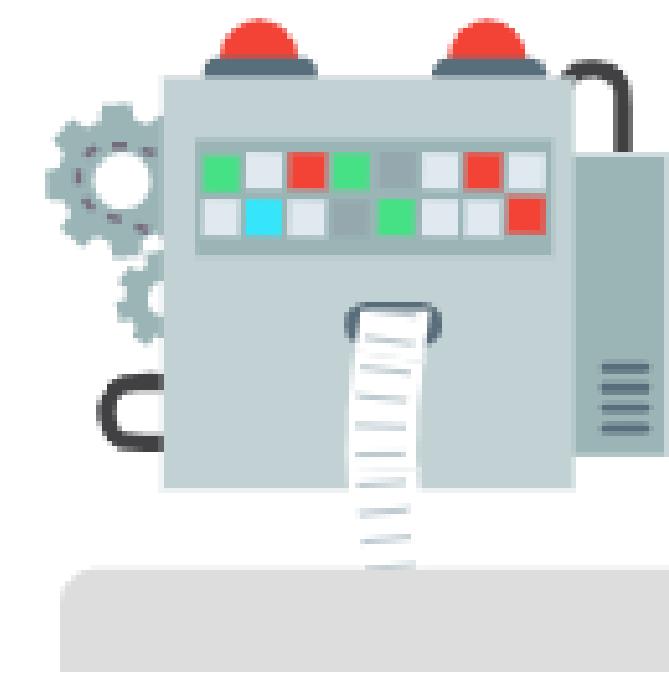
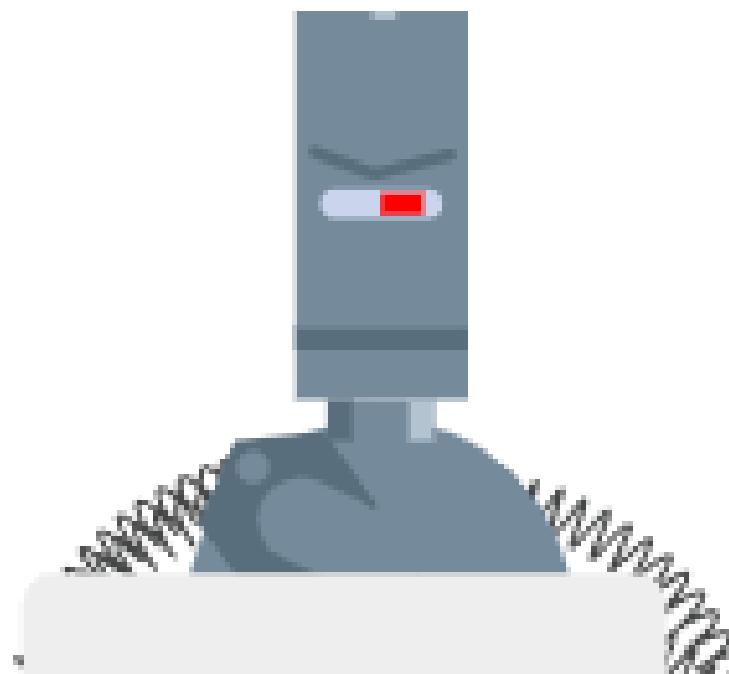
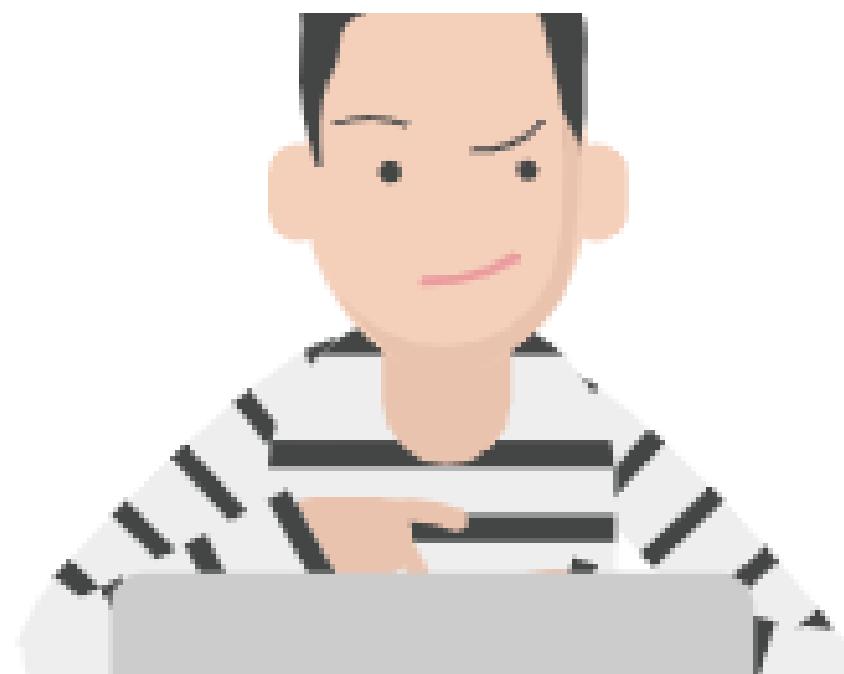


The Problem

WHAT IS CLICK FRAUD

Ad click fraud is prevalent in PPC advertising space to extrapolate the number of clicks for generating more revenue.

Although Google Adsense has checks in place it is still advisable to monitor your traffic details and analyzing them from time to time





“

According to AppsFlyer's annual fraud report, "The State of Mobile Ad Fraud 2020 Edition", global financial exposure to ad fraud in the first-half of 2020 was US\$1.6 billion.





Our Approach to the Solution

PLAN PLAN PLAN

Familiarising the team
with business
problem

Chalk out a logical
plan for approach

Plan for
implementation

DEVELOP SKILLS

Team members were
assigned specific
technology to
effectively manage time

TRIAL AND ERROR

Implement the final
build and improve upon
it

Agile and spiral model
of development



Challenges

NEW PLATFORMS

Most of the team were unaware of the platforms. We took up the challenge and developed the needed skills

COMMUNICATION

Communication was key for the development and being online was a bit tough. We overcame it by having regular meets as a team and peer to peer meets for development

TIME MANAGEMENT

Team members were busy with their regular schedules and had to compensate for different schedules

SEPARATE DEVELOPMENT

The code mostly worked fine but being in different platforms used to create bugs and irregular behaviour. It had to be smoothed out through one to one communication

The Solution

Overall Project can be divided into 3 major parts-

MODEL

API

WEBAPP





Selecting the Model

STEPS IN MODEL DEVELOPMENT



Gettingt datasets and
doing basic analysis-
Kaggle



Data Cleaning and pre-
processing - Using
Dataiku



Research for possible
models that can be
used- Online



Monitoring performance
for different models-
Dataiku



Finalize the model -
Random Forest



Building Model

The screenshot shows a dataset named "train" in a data science environment. The interface includes a top navigation bar with "project", "Datasets", and search functions. Below the navigation is a toolbar with icons for "Summary", "Explore", "Charts", "Statistics", "Status", "History", "Settings", and "ACTIONS". The main area displays a table titled "Viewing dataset sample" with 100,000 rows and 8 columns. The columns are labeled: ip, app, device, os, channel, click_time, attributed_time, and is_attributed. The table shows various numerical values and dates. A vertical sidebar on the right contains icons for "ADD", "INFO", "REFRESH", "SEARCH", and "EDIT".

ip	app	device	os	channel	click_time	attributed_time	is_attributed
83230	3	1	13	379	2017-11-06 14:32:21		0
17357	3	1	19	379	2017-11-06 14:33:34		0
35810	3	1	13	379	2017-11-06 14:34:12		0
45745	14	1	13	478	2017-11-06 14:34:52		0
161007	3	1	13	379	2017-11-06 14:35:08		0
18787	3	1	16	379	2017-11-06 14:36:26		0
103022	3	1	23	379	2017-11-06 14:37:44		0
114221	3	1	19	379	2017-11-06 14:37:59		0
165970	3	1	13	379	2017-11-06 14:38:10		0
74544	64	1	22	459	2017-11-06 14:38:23		0
172522	3	1	25	379	2017-11-06 14:38:27		0

DATASET

Each row of training data contains a click record, with the following features-

- ip: ip address of click.
- app: app id for marketing.
- device: device type id of user mobile phone (e.g., iphone 6 plus, iphone 7, huawei mate 7, etc.)
- os: os version id of user mobile phone
- channel: channel id of mobile ad publisher
- click_time: timestamp of click (UTC)
- attributed_time: if user download the app for after clicking an ad, this is the time of the app download
- is_attributed: the target that is to be predicted, indicating the app was downloaded



Basic Exploratory Data Analysis:-

LARGE SIZE OF TRAINING DATA

The actual train dataset was very large in size. It was around 7GB in size containing 180 million entries. With our personal system it was not possible to work in such big dataset

UNBALANCED DATA

- Our dataset was very highly unbalanced. In 100k entries, only 169 entries were actual click and rest all were fraud clicks.

SOME ATTRIBUTES WERE LESS RELEVANT

- Analysis of attributed_time showed that it contains mostly null or no values and hence was not helpful for prediction.

Pre Processing

DATA BALANCING

Resampling was done using Dataiku sample recipe after which we had a fairly balanced dataset.

DATA CLEANING

Mostly the data was clean. removed

FEATURE ENGINEERING

Dataset had some features through which certain properties could be extracted

DROPPING IRRELEVANT FIELDS

Dropping attributed time





The screenshot shows the Dataiku DSS interface with the following details:

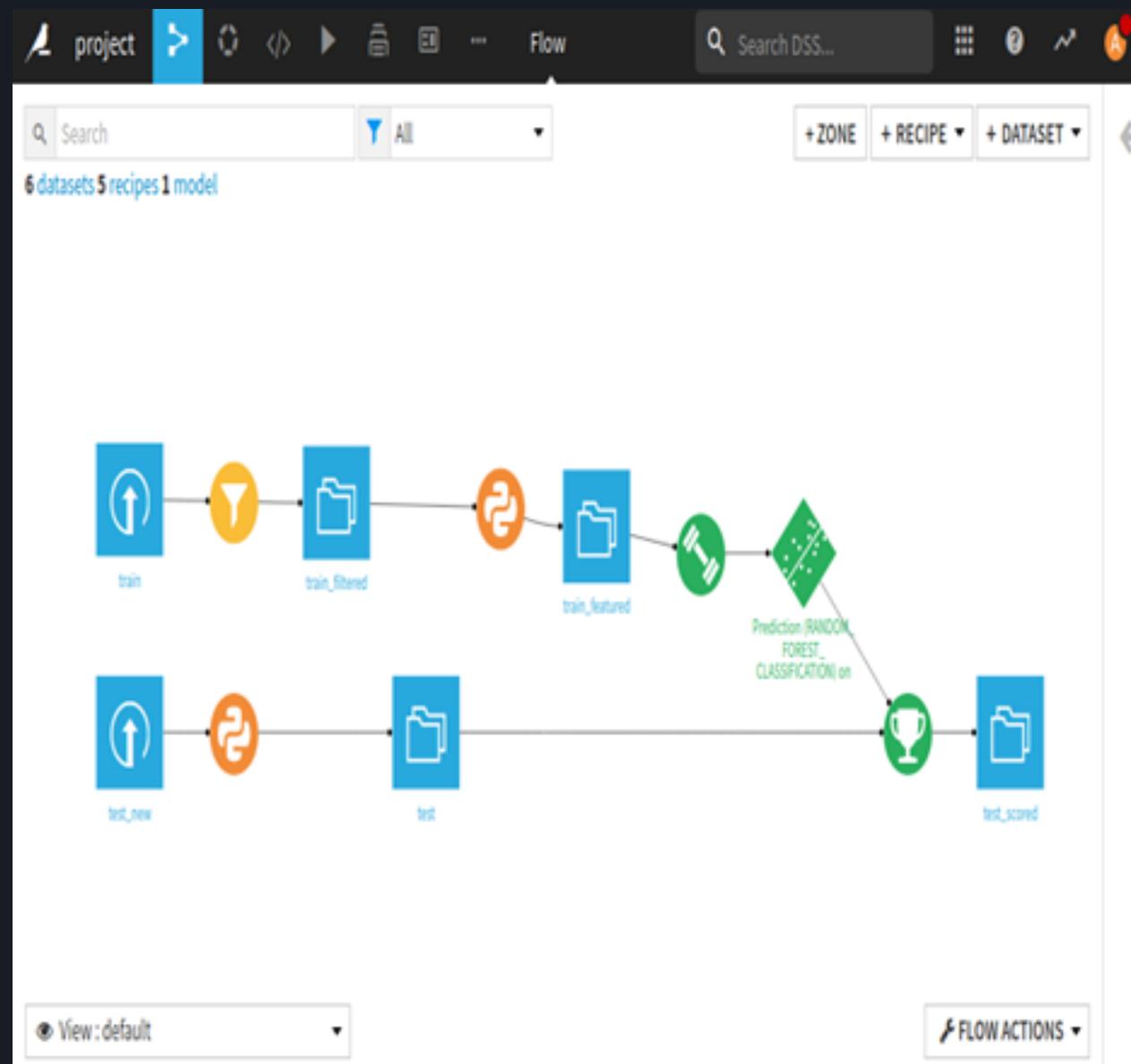
- Title:** Quick modeling of is_attributed on train_featured
- Section:** RESULT
- Model Type:** Predict is_attributed (Binary classification)
- Actions:** SAVED, TRAIN
- Table Headers:** Name, Trained, Train time, Sample weights variable, Accuracy, Precision, Recall, F1 Score, Cost Matrix
- Table Data:**

Name	Trained	Train time	Sample weights variable	Accuracy	Precision	Recall	F1 Score	Cost Matrix
Random forest (test)	2020-09-21 18:23:12	36s	-	0.92	0.94	0.89	0.92	0.44
XGBoost (test)	2020-09-21 18:24:24	29s	-	0.92	0.95	0.88	0.92	0.44
Decision Tree (test)	2020-09-21 18:23:39	6s	-	0.90	0.96	0.83	0.89	0.42
Artificial Neural Network (test)	2020-09-21 18:23:51	47s	-	0.90	0.93	0.87	0.90	0.43
K Nearest Neighbors (k=5) (test)	2020-09-21 18:23:48	33s	-	0.89	0.93	0.86	0.89	0.42
Logistic Regression (test)	2020-09-21 18:23:12	25s	-	0.83	0.86	0.80	0.83	0.38

Choosing Model

RANDOM FOREST GAVE BET RESULT IN MULTIPLE TEST RUNS

Final Flow and Predictions in Dataiku



The screenshot shows the Dataiku Datasets interface. At the top, there are tabs for 'project', 'Datasets', and 'Datasets'. Below the tabs, a search bar says 'Search DSS...' and a filter dropdown says 'All'. There are buttons for 'Summary', 'Explore', 'Charts', 'Statistics', 'Status', 'History', 'Settings', 'PARENT RECIPE', and 'ACTIONS'. The main area displays a table titled 'Viewing dataset sample' with 'Configure sample' and 'DISPLAY' buttons. It shows a sample of 10000 rows from a dataset with 6 columns: ip, app, device, os, channel, and prediction. The table has a header row with types: bigint, bigint, bigint, bigint, bigint, bigint. The data rows show various integer values. On the right side, there is a vertical toolbar with icons for 'i', 'm', 's', and 'a'.

ip	app	device	os	channel	prediction
5744	9	1	3	107	0
119901	9	1	3	466	0
72287	21	1	19	128	1
78477	15	1	13	111	0
123080	12	1	13	328	0
110769	18	1	13	107	0
12540	3	1	1	137	0
88637	27	1	19	153	0
14932	18	1	10	107	0





GETTING OUR HANDS DIRTY

We faced our first hiccup that we were no able to get a working api from Dataiku

Our mentor sir came to our rescue and helped us by suggesting alternative methods to build api



CODING THE MODEL

Model development was done and tested

Building API

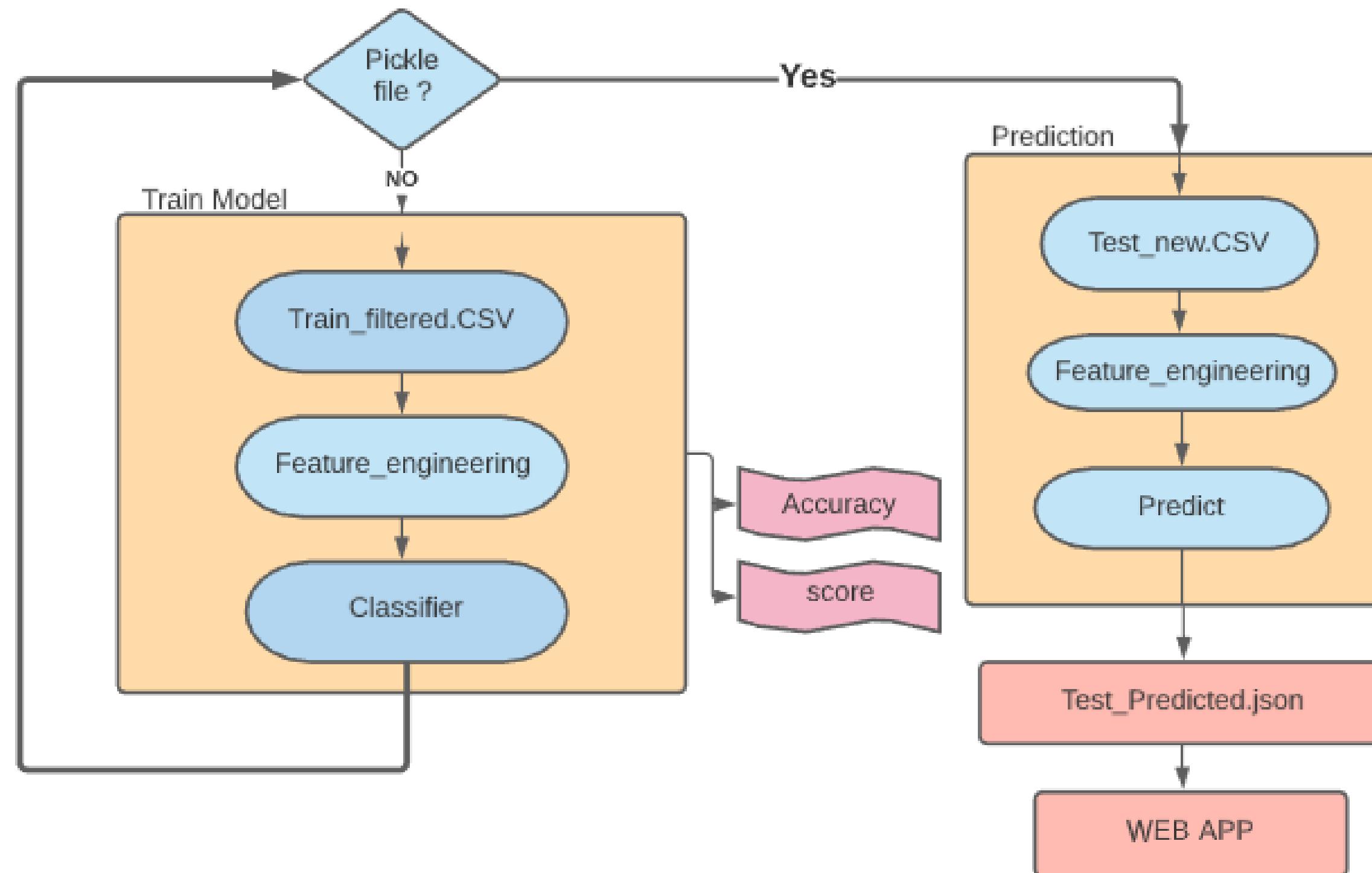


An application-programming interface (API) is a set of programming instructions and standards for accessing a Web-based software application or Web tool.

In this project Flask Restful API framework is used for deploying model.

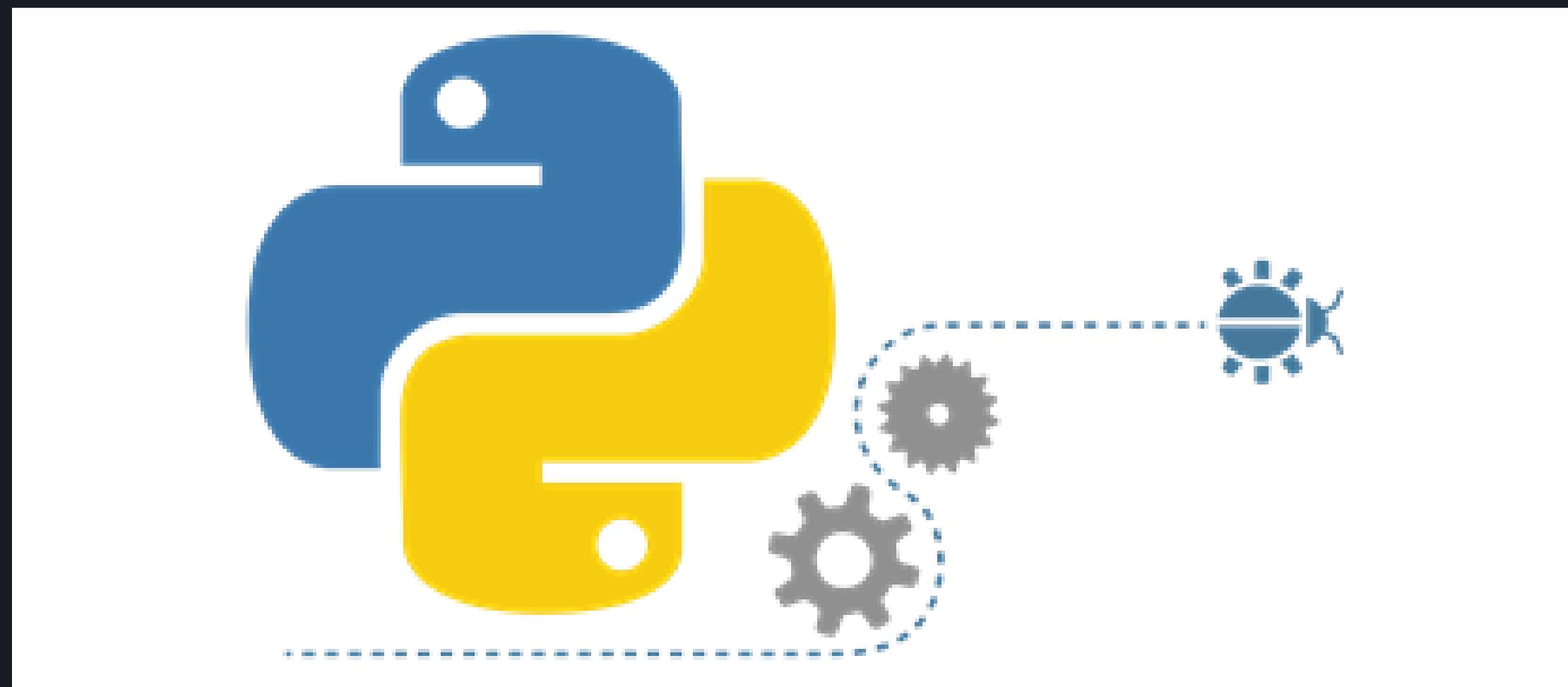


Flowchart API



DJANGO WEBAPP

Django is a Python-based free and open-source web framework that follows the model-template-views architectural pattern.





First Page

Takes CSV as input

Study and Statistics of
CLICK FRAUD

Upload
CSV FILE

Choose File No file chosen.

CHECK STATISTICS

What Is
CLICK FRAUD?

Click Fraud is a type of fraud that occurs on the Internet in pay per click (PPC) online advertising. "In the type of advertising, the owners of websites that post the ads are paid an amount of money determined by how many visitors to the sites click on the ads."

Fraud occurs "when a person, automated script, or computer program initiates a legitimate user of a web browser, clicking on such an ad without having an actual interest in the target of the ad's link."

Important Tips to PREVENT CLICK FRAUD

- Set different bid prices for content-targeted sites
- Keep an eye on your competition
- Always track your advertising campaigns
- Only advertise in specific countries
- Target High-value sites for your ads
- Purchase software programs that generate special referral reports

Team Members TEAM 2

Rahul Tomar
Indian Institute of Management, Rohtak
(Business Cipher)

Dipanshu Upreti
Beta Institute Of Applied Sciences
(NeuralHack - Python)

Vaishnavi Subramanya Desai
Birla Hyderabad College Of Engineering For Women
(NeuralHack - Python)

Girish Kumar Reddy Tokala
G R Rama Reddy Engineering College
(NeuralHack - Python)

Ragini Gaurav
Hercos Institute Of Technology
(NeuralHack - UI)

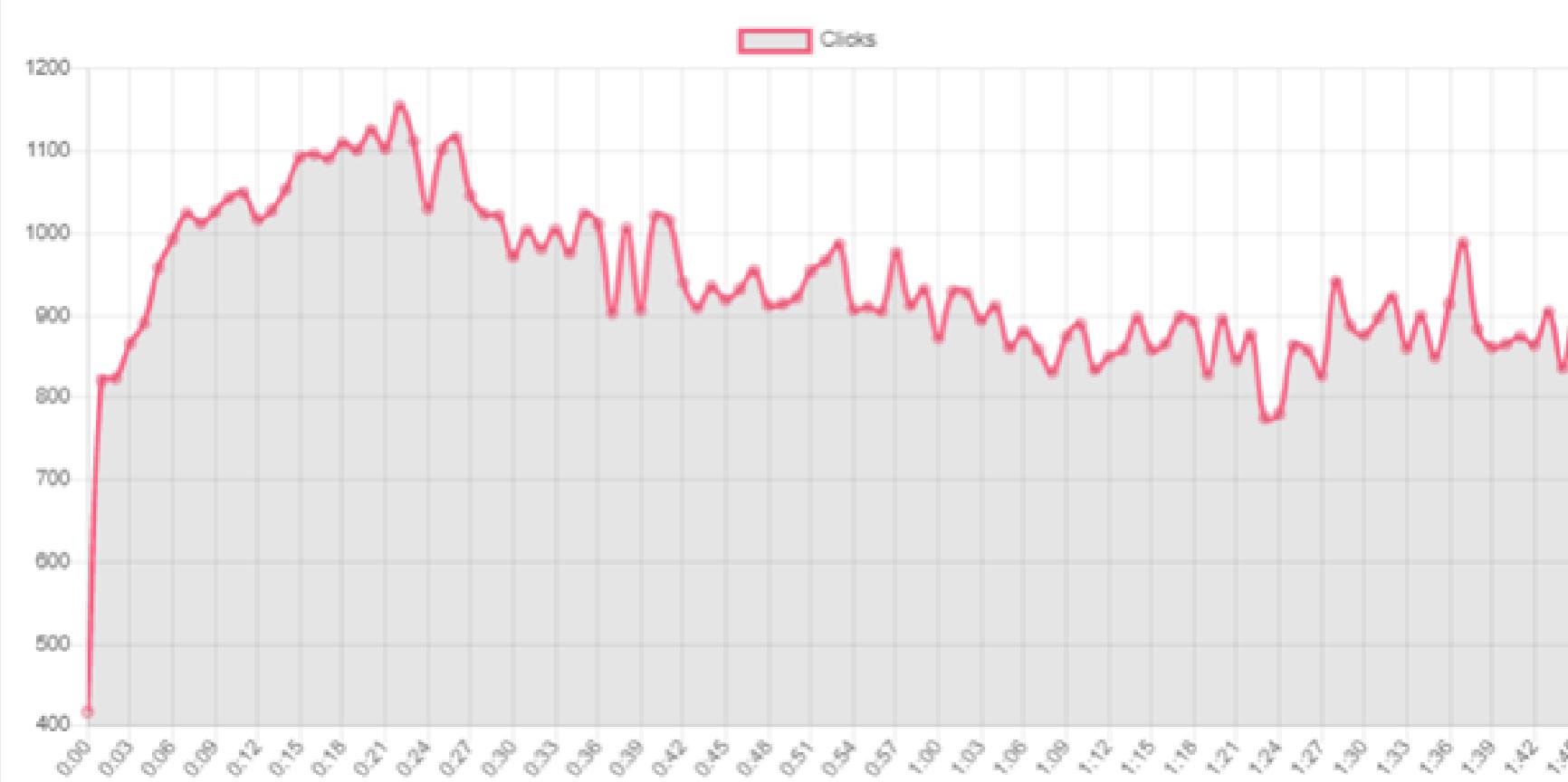
eClick

Dashboard

Training Prediction -

Training Time:- 0.5966494083404541 sec

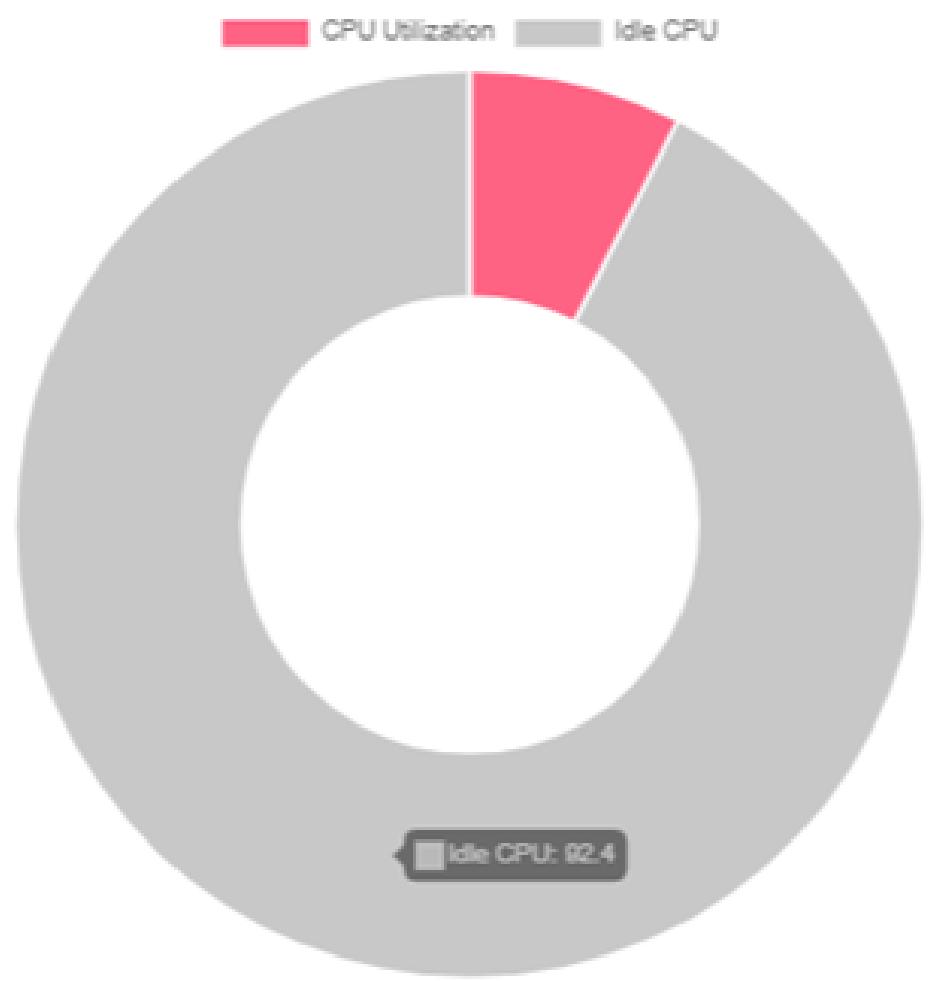
Total Clicks



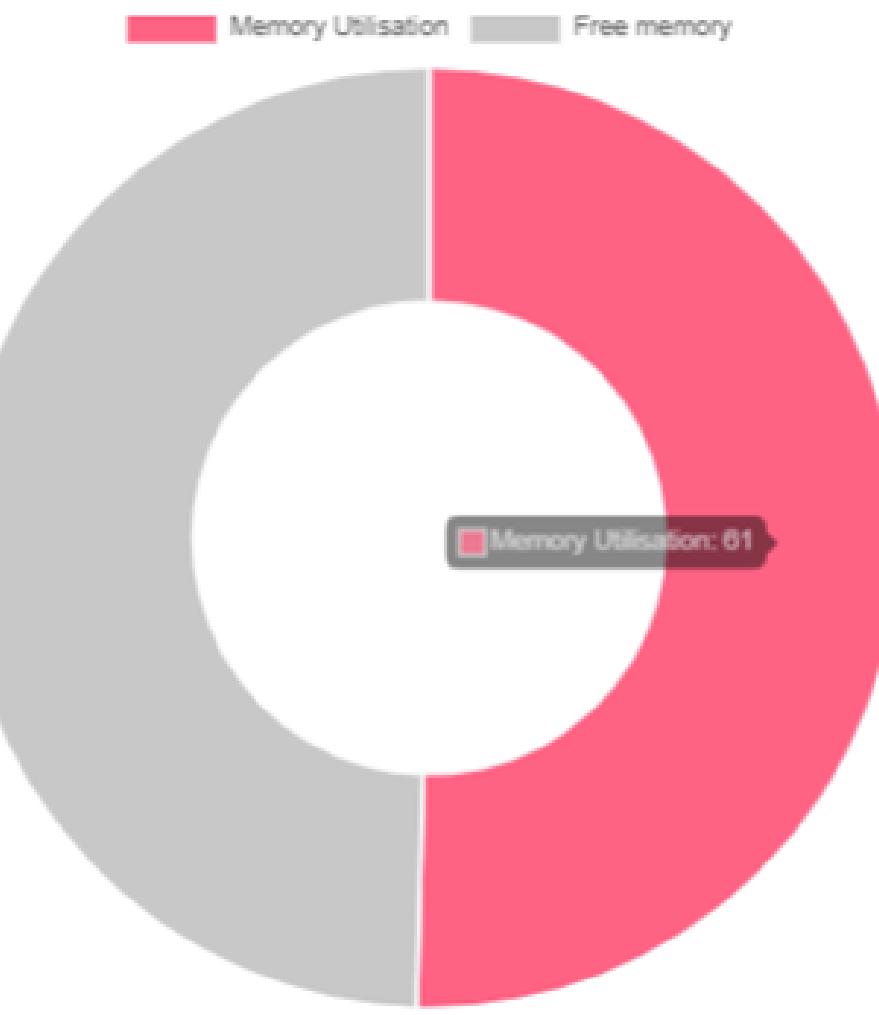
Second Page

- Visualizing the resources and giving results of prediction performed by API

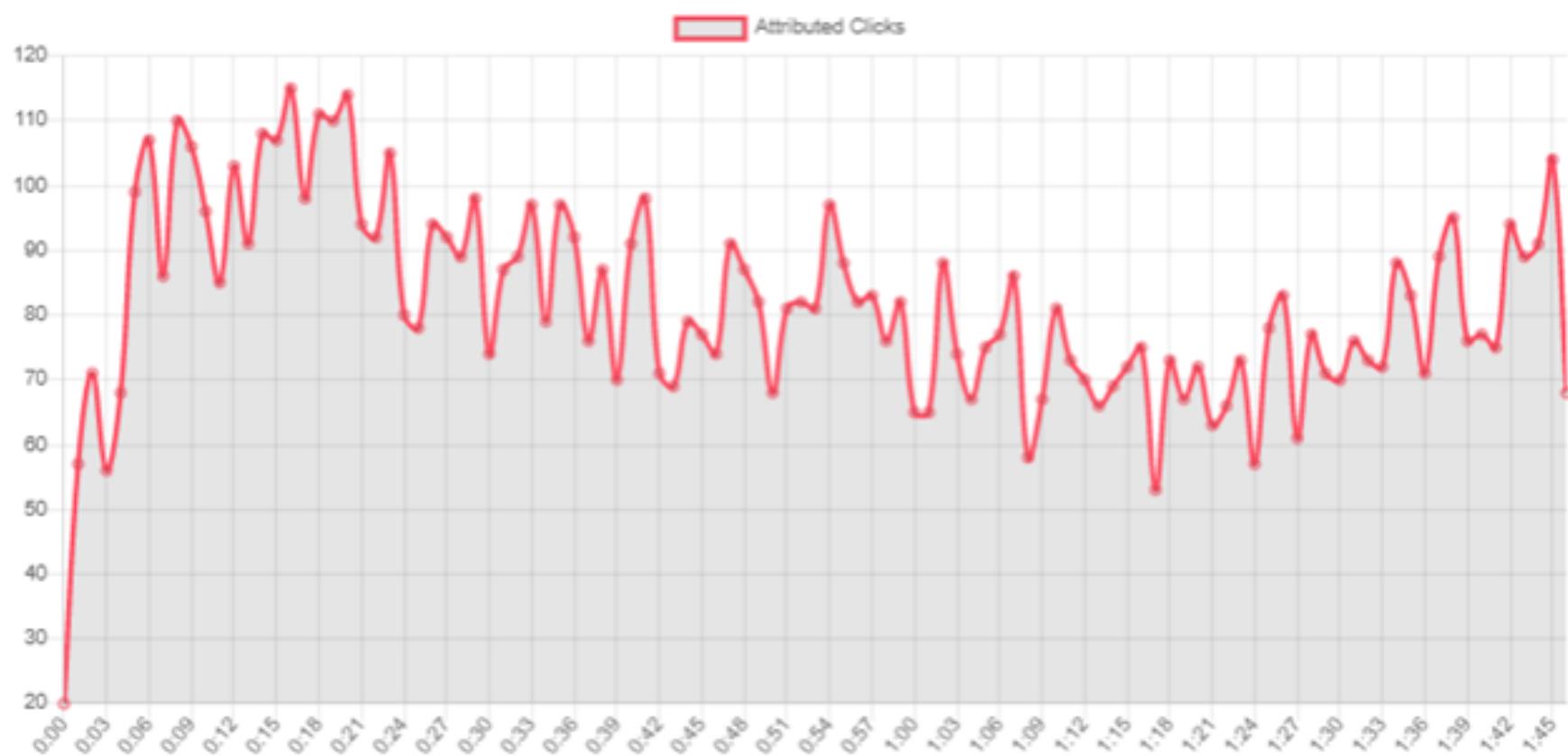
CPU Utilization



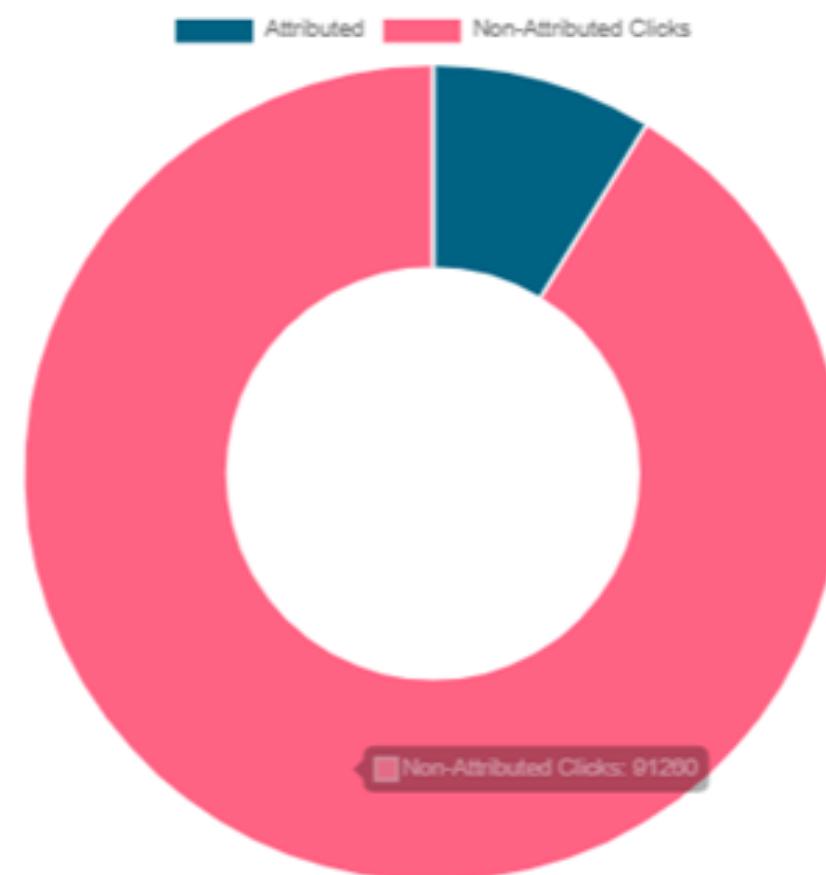
Memory Utilization



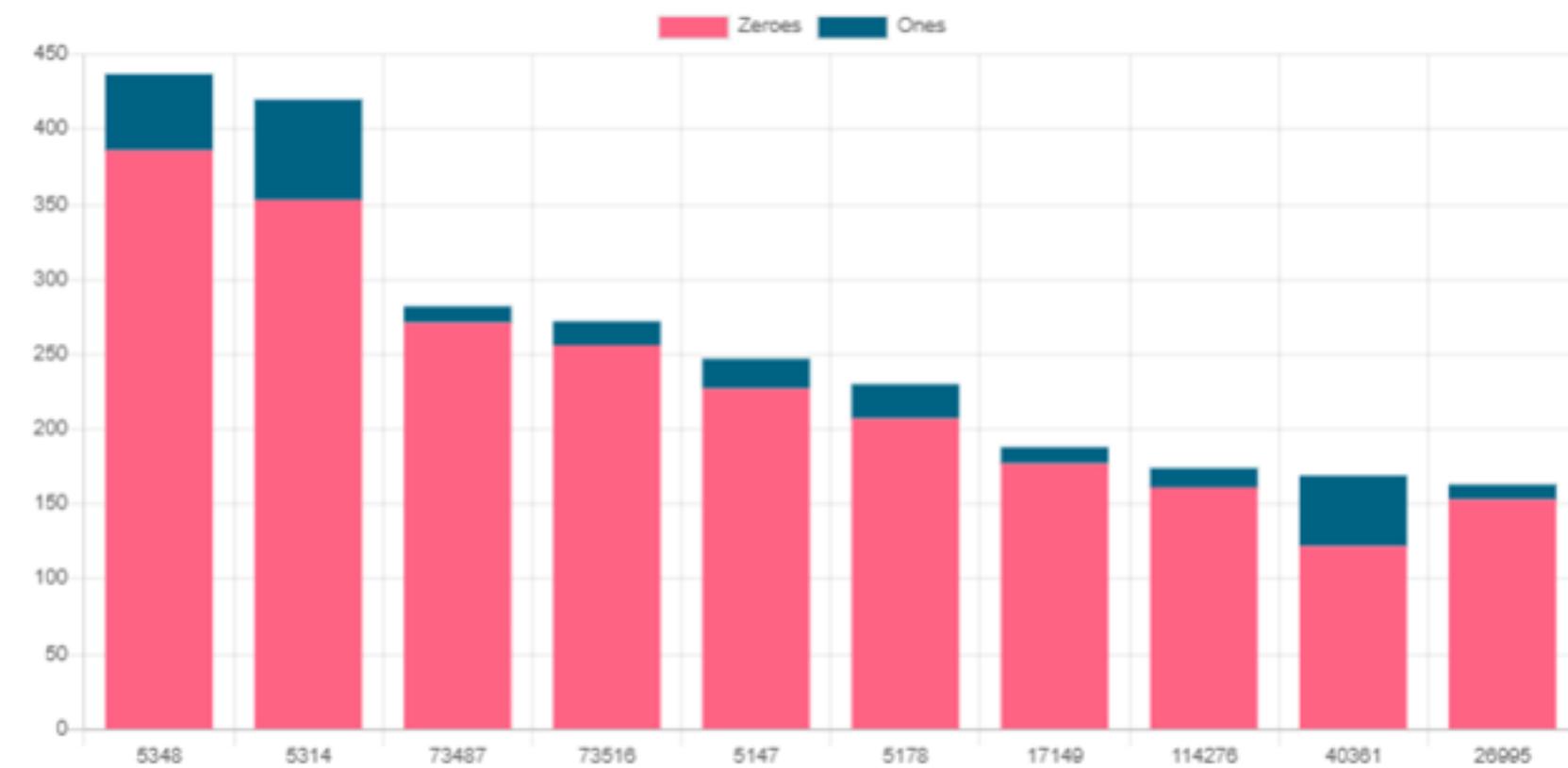
Predicted Attributed Clicks



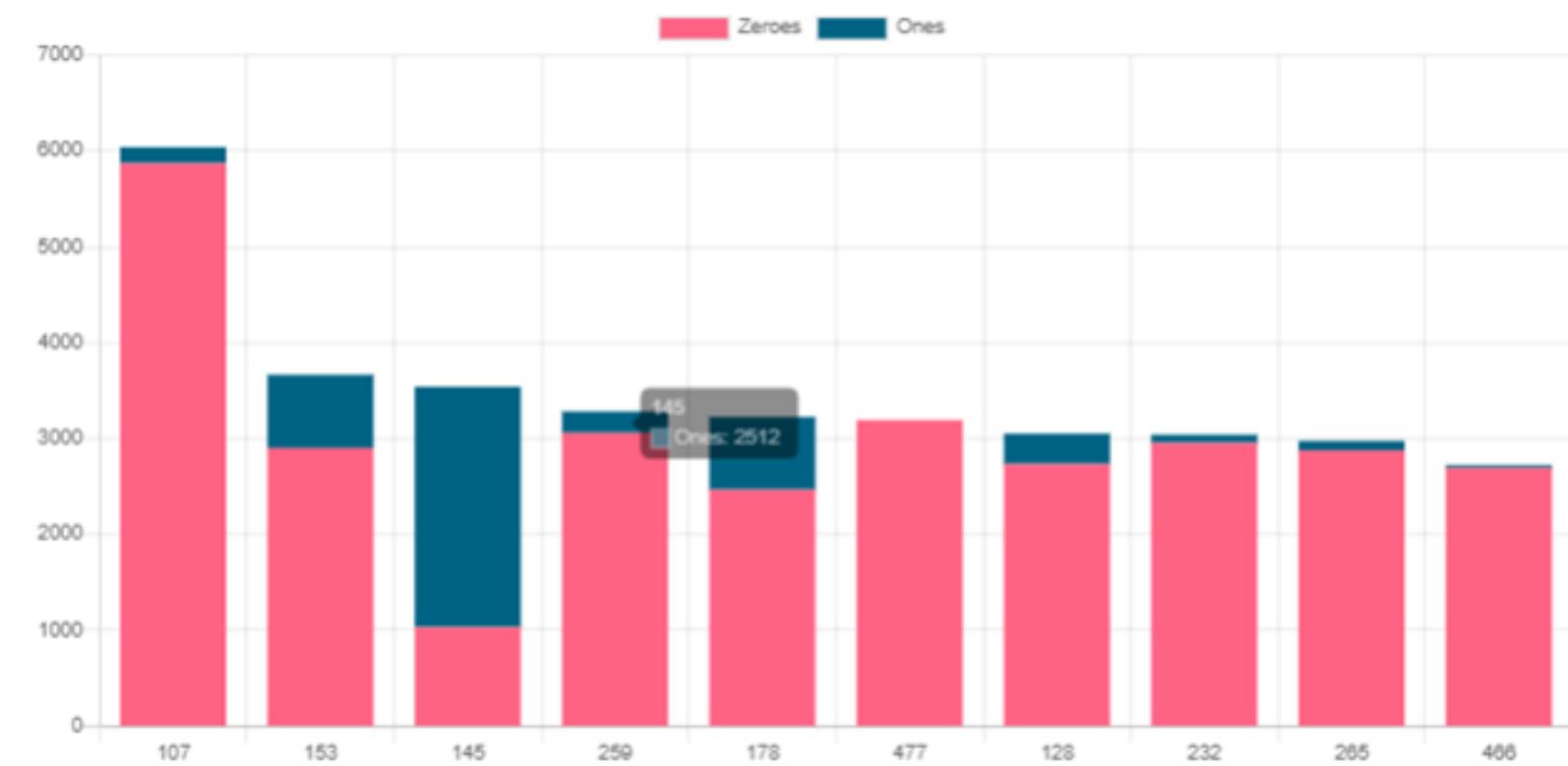
Attributed vs Non-attributed Clicks



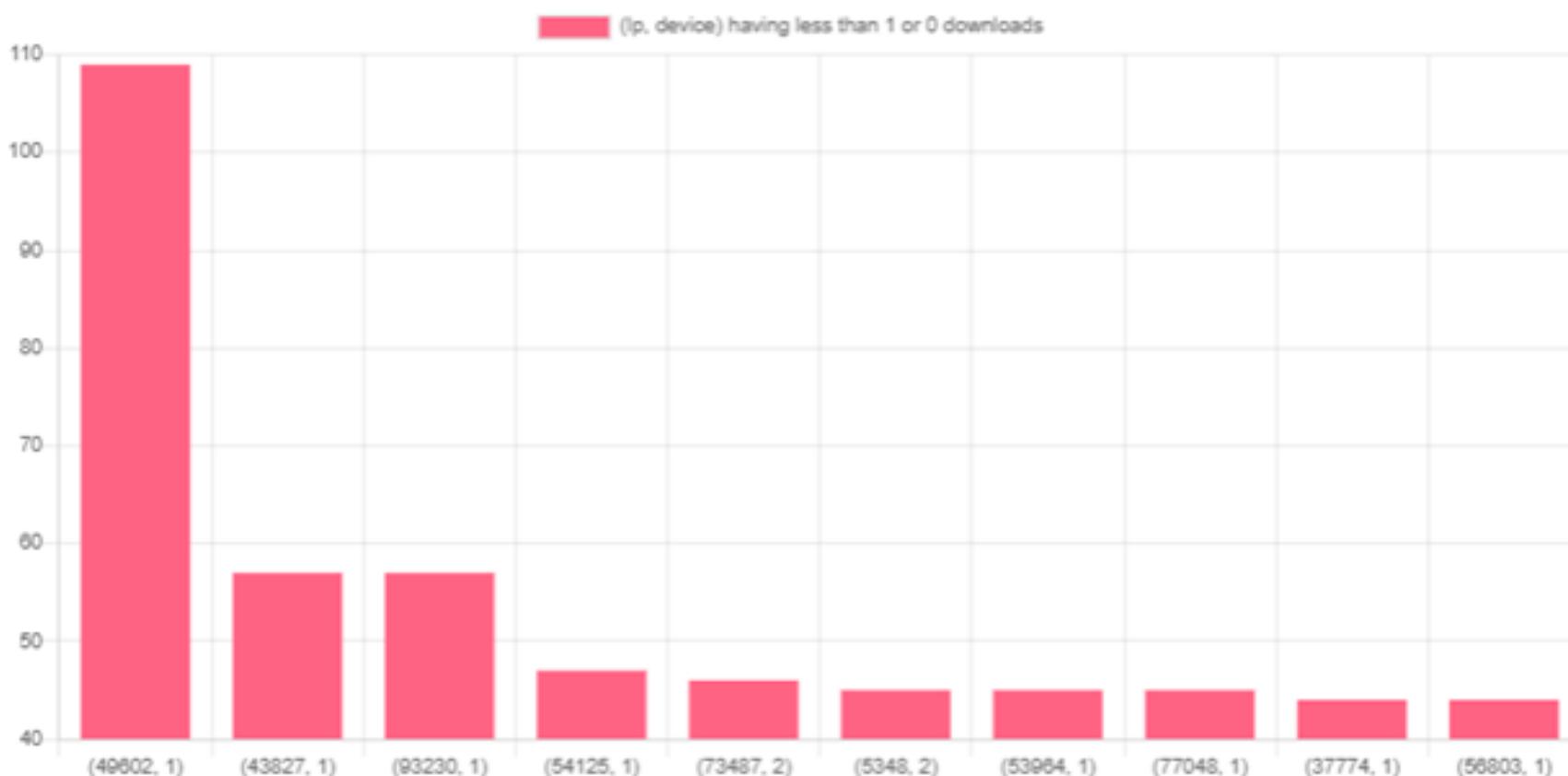
IPs with most clicks



Channel with most clicks



Suspected frauds (IP and Device combination with 0 or no attributed clicks)





Thank You