TEAM 2- BFS TRACK

FRAUD CLICK DETECTION WEBAPP

Rahul Tomar	Team Lead
Dipanshu Upreti	Python Track
Vaishnavi Subramanya Desai	Python Track
Girish Kumar Reddy Tokala	Python Track
Ragini Gaurav	UI Track

Presentation Outline

TOPICS FOR DISCUSSION

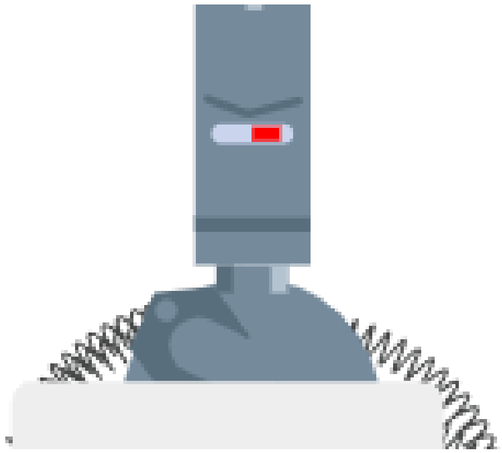
The Problem!
Approach
The Solution
Implementation

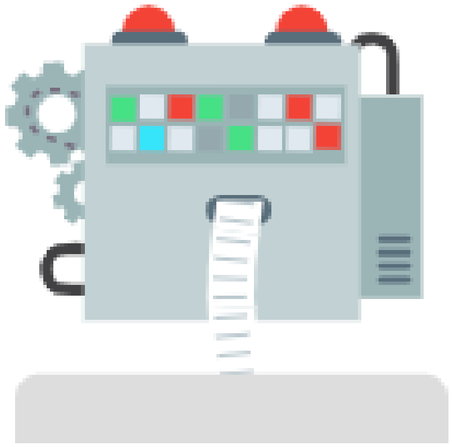
The Problem

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

Familarising 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 ateam and peer to peer meets for development

TIME MANAGEMENT

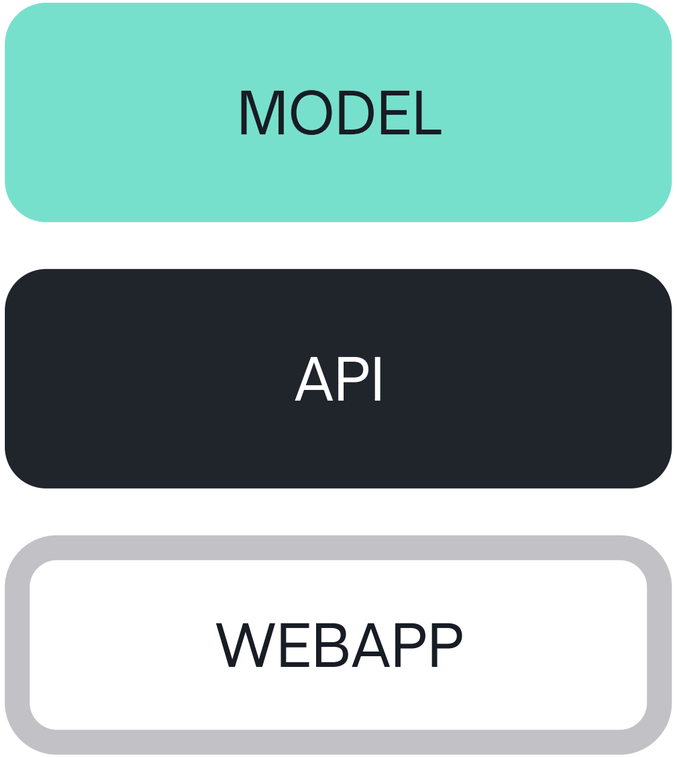
Team members were busy with their regular schedules and had to compenstae for different schedules

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 commnication

The Solution

Overall Project can be divided into 3 major parts-



Selecting the Model

STEPS IN MODEL DEVELOPMENT

Gettingt datasets and doing basic analysis- Kaggle

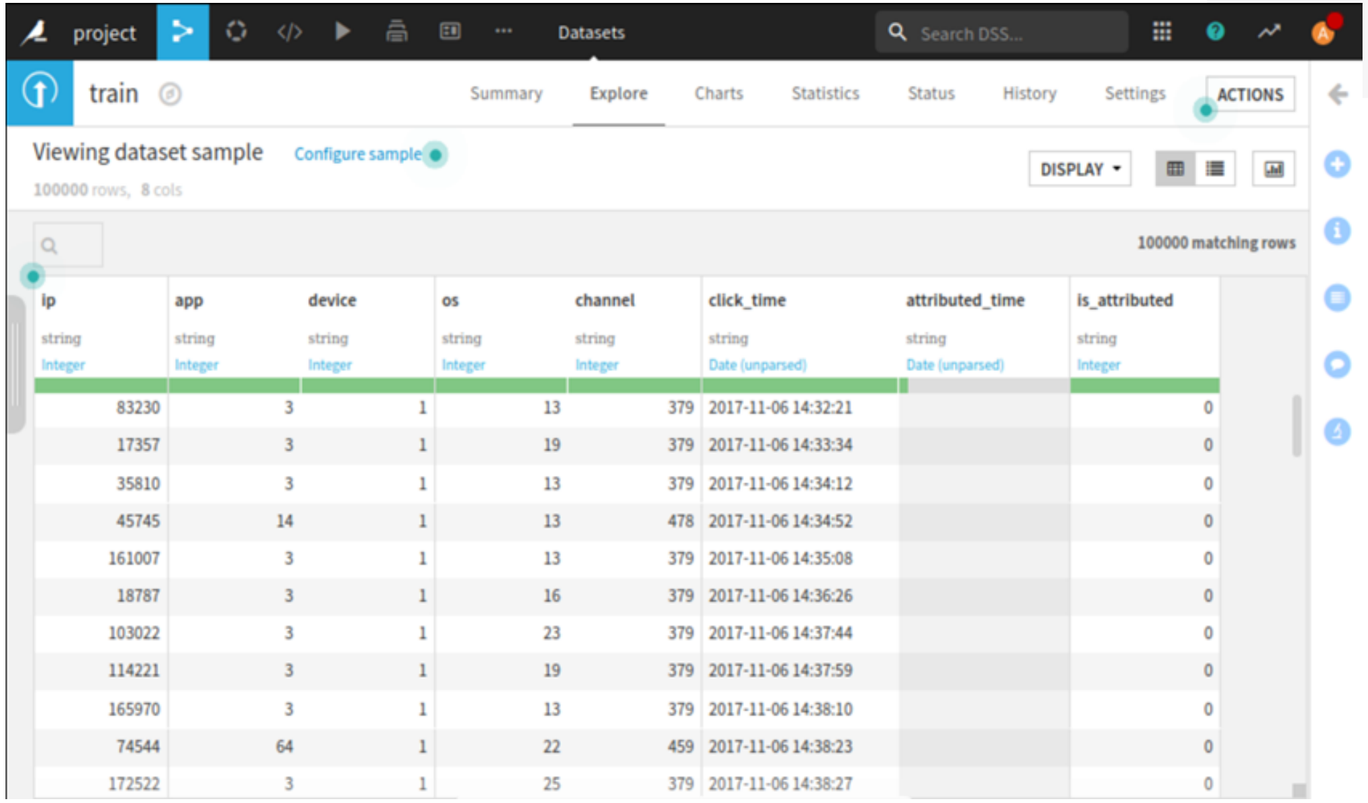
Data Cleaning and pre-processing - Using Dataiku

Research for possible models that can be used- Online

Monitoring performance for different models- Dataiku

Finalize the model - Random Forest

Building Model



ip	app	device	os	channel	click_time	attributed_time	is_attributed
string Integer	string Integer	string Integer	string Integer	string Integer	string Date (unparsed)	string Date (unparsed)	string Integer
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

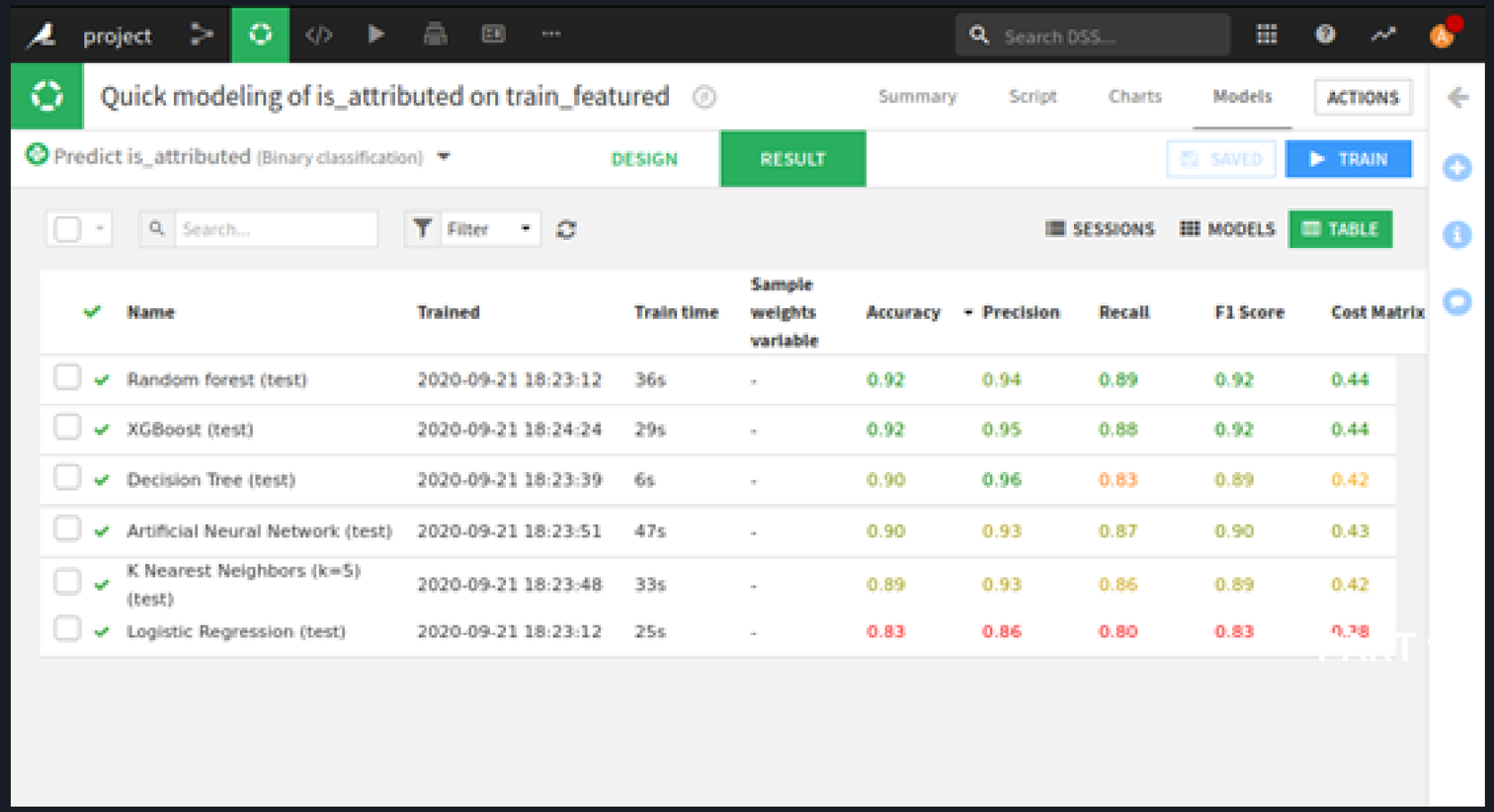
Mostly the data was clean. removed

FEATURE ENGINEERING

Dataset had some features through which certain properties could be extracted

DROPPING IRRELEVANT FIELDS

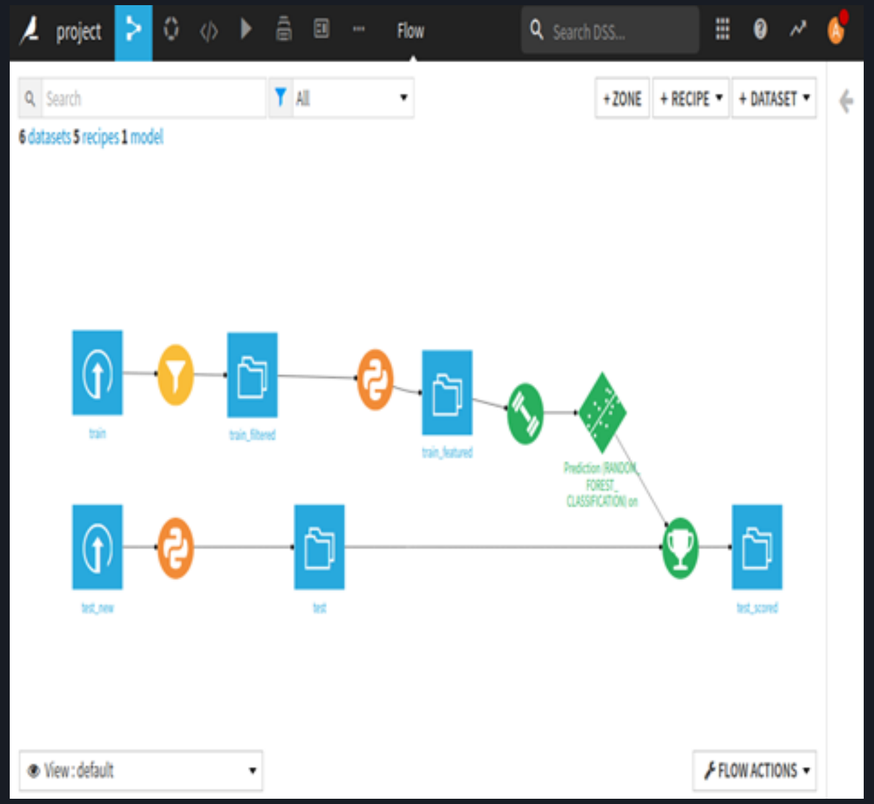
Dropping attributed time

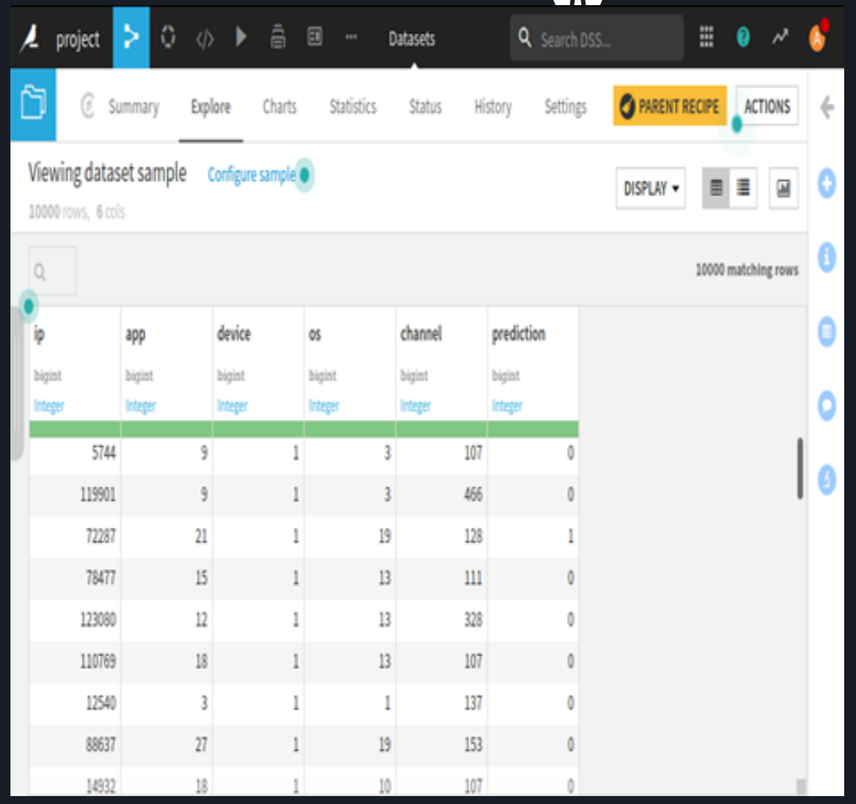


Choosing Model

RANDOM FOREST GAVE BET RESULT IN MULTIPLE TEST RUNS

Final Flow and Predictions in Dataiku





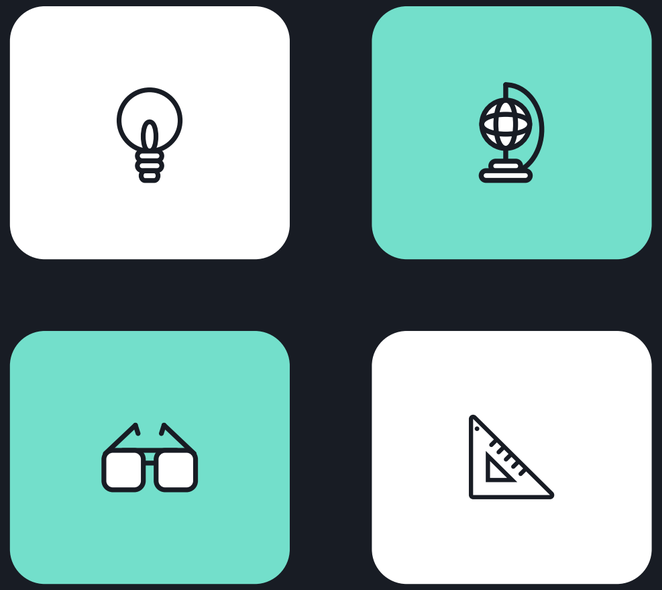
GETTIBNG OUR HANDS DIRTY

We faced our fiirst 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

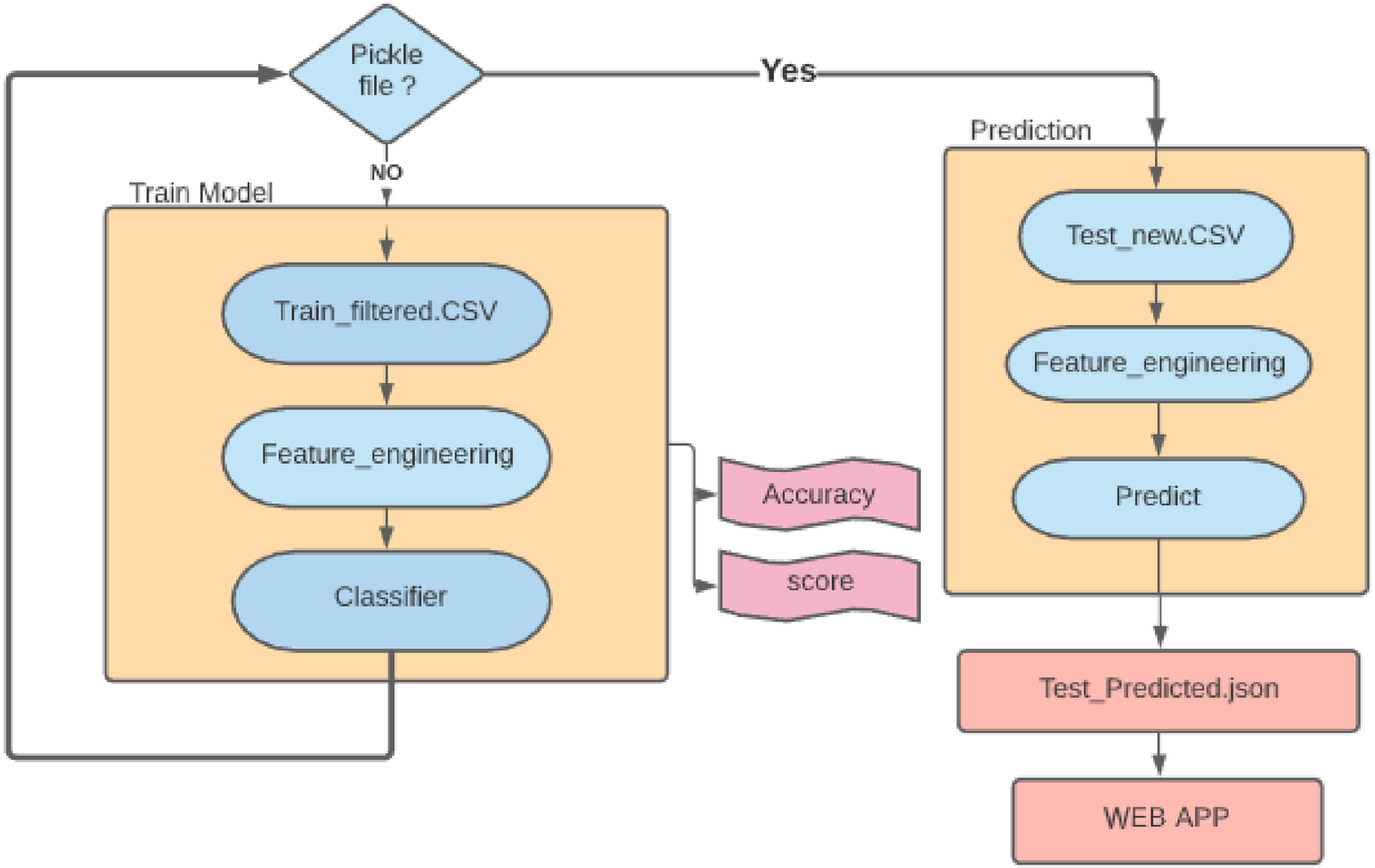


Building API

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

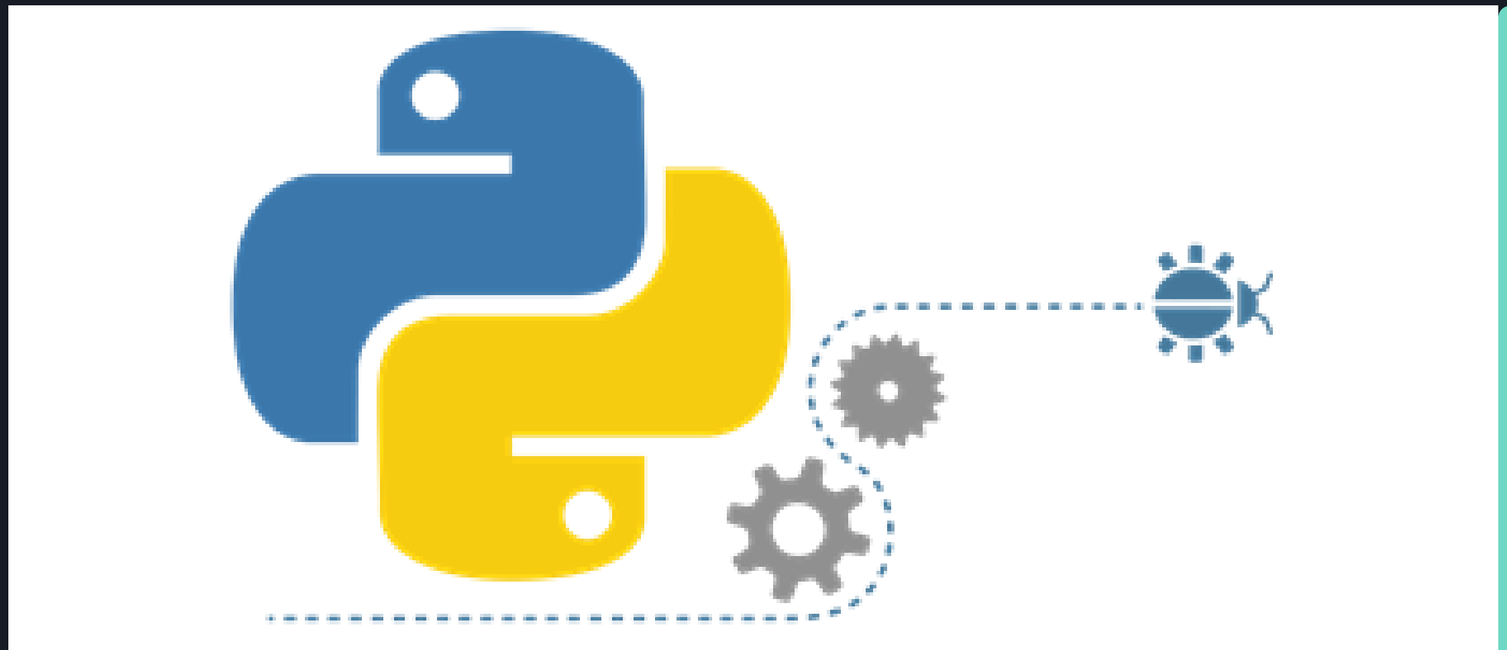
In this project Flask Restful API framework is used for deploying model.

Flowchart API



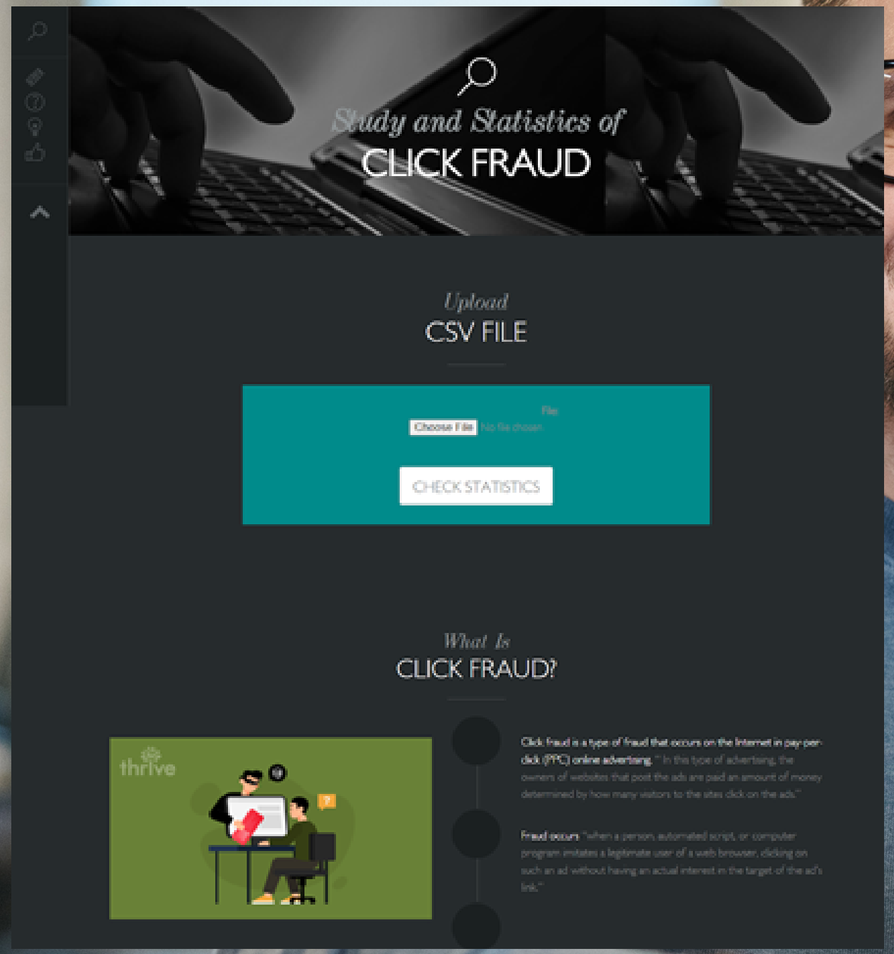
DJANGO WEBAPP

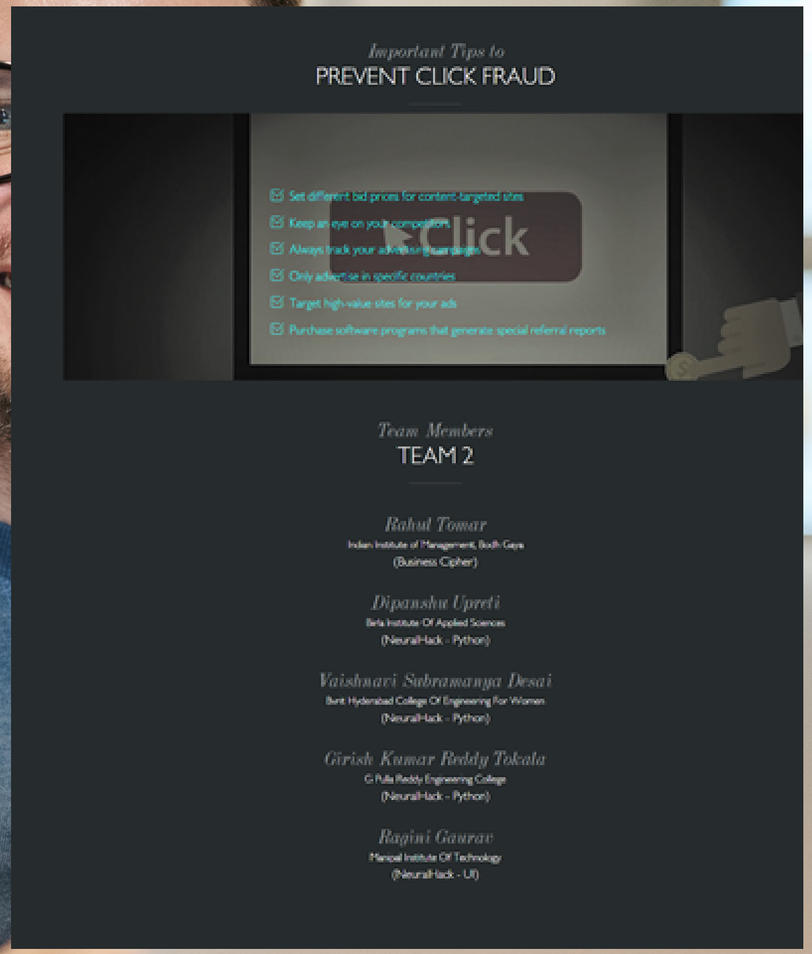
Django is a Python-based free and open-source web framework that follows the model-template-views architectural pattern.



First Page

Takes CSV as input

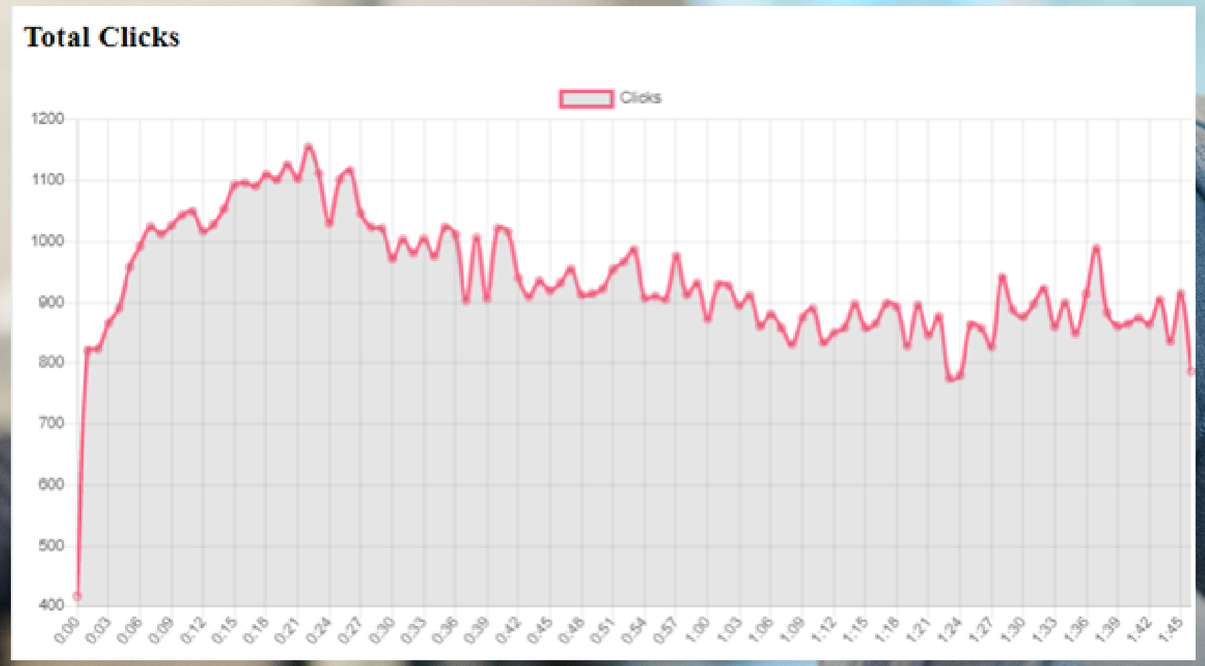




Dashboard

Training Prediction -

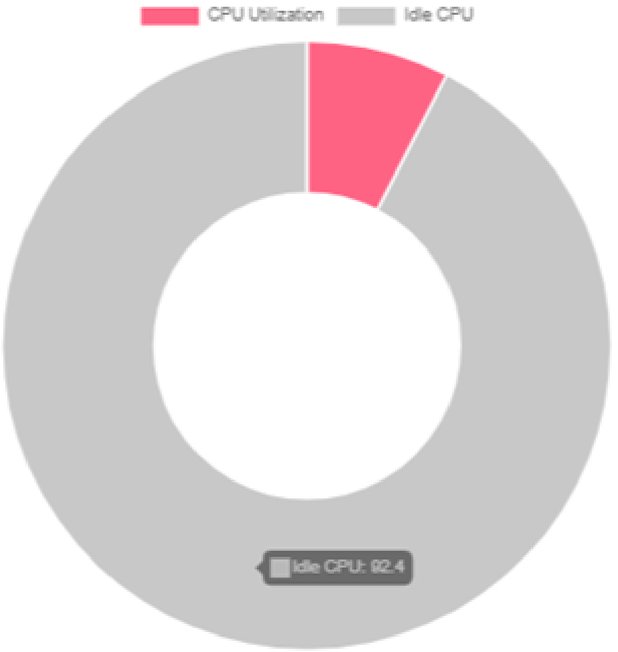
Training Time:- 0.5966494083404541 sec



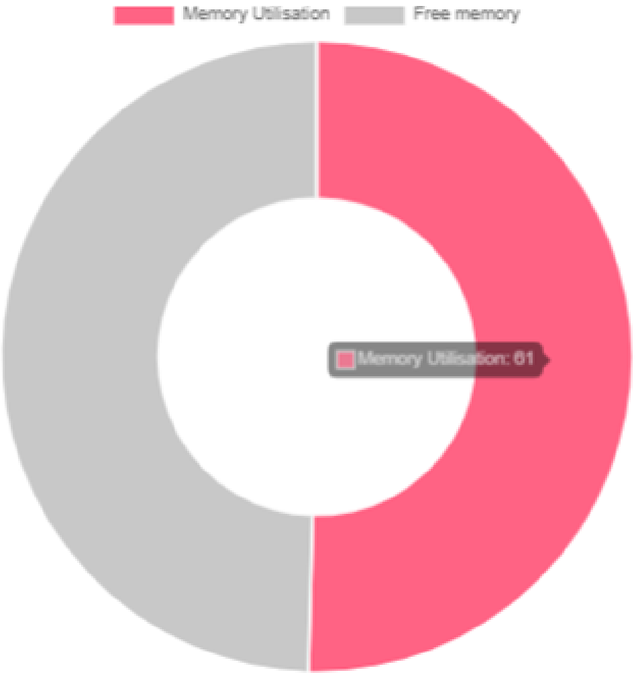
Second Page

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

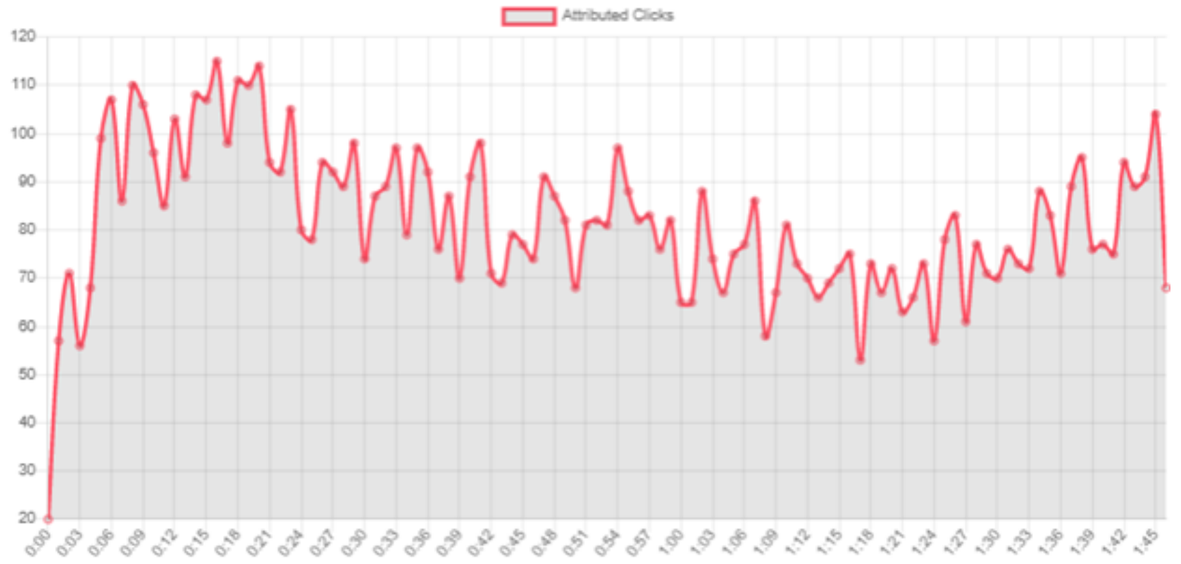
CPU Utilization



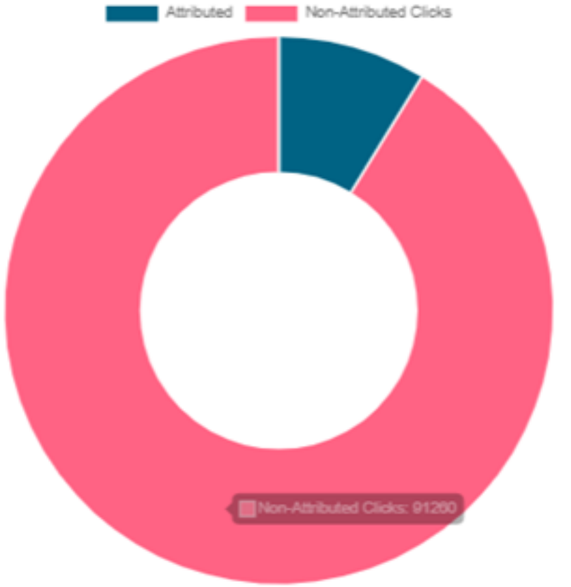
Memory Utilization



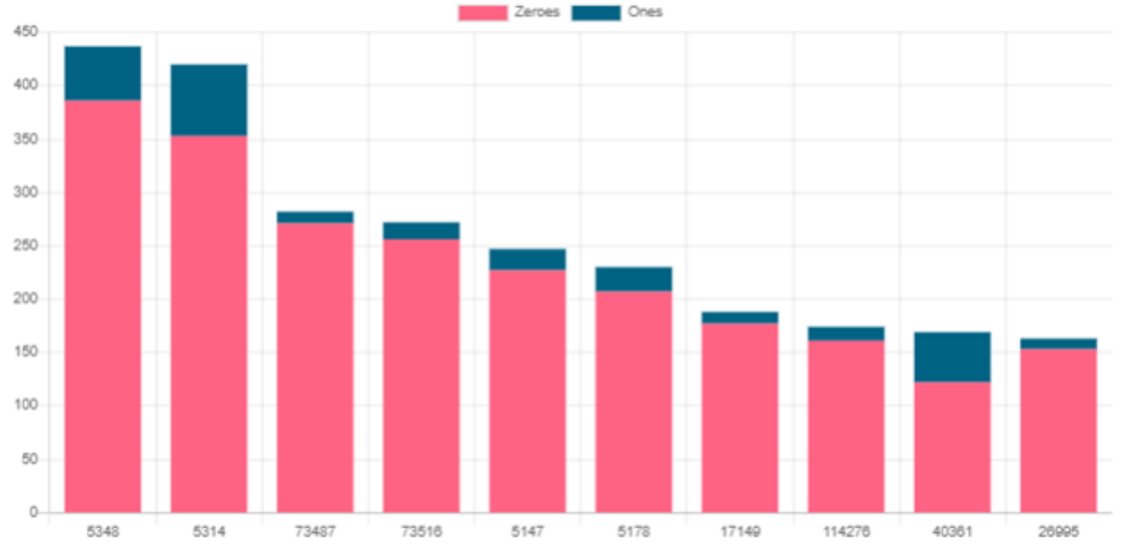
Predicted Attributed Clicks



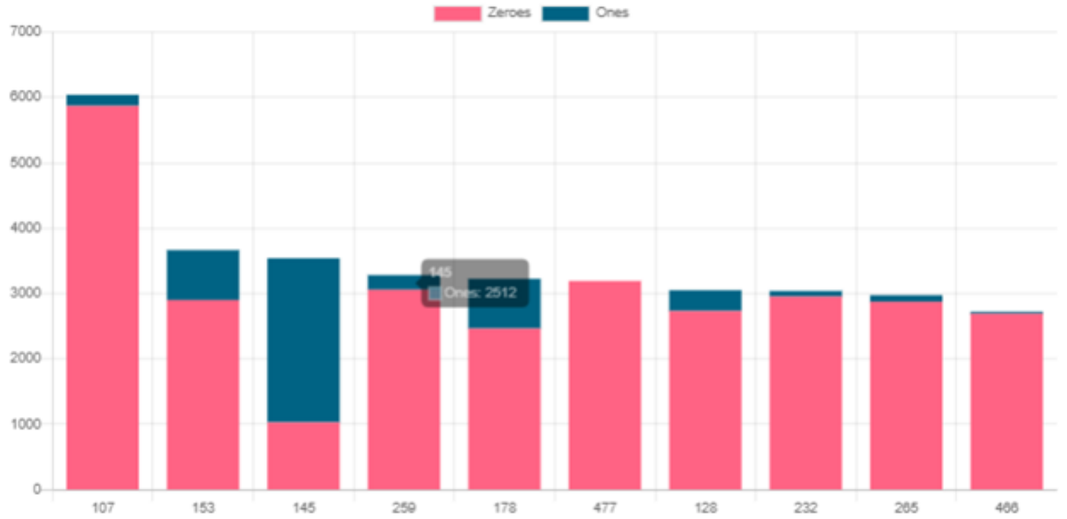
Atrubuted vs Non-attributed Clicks



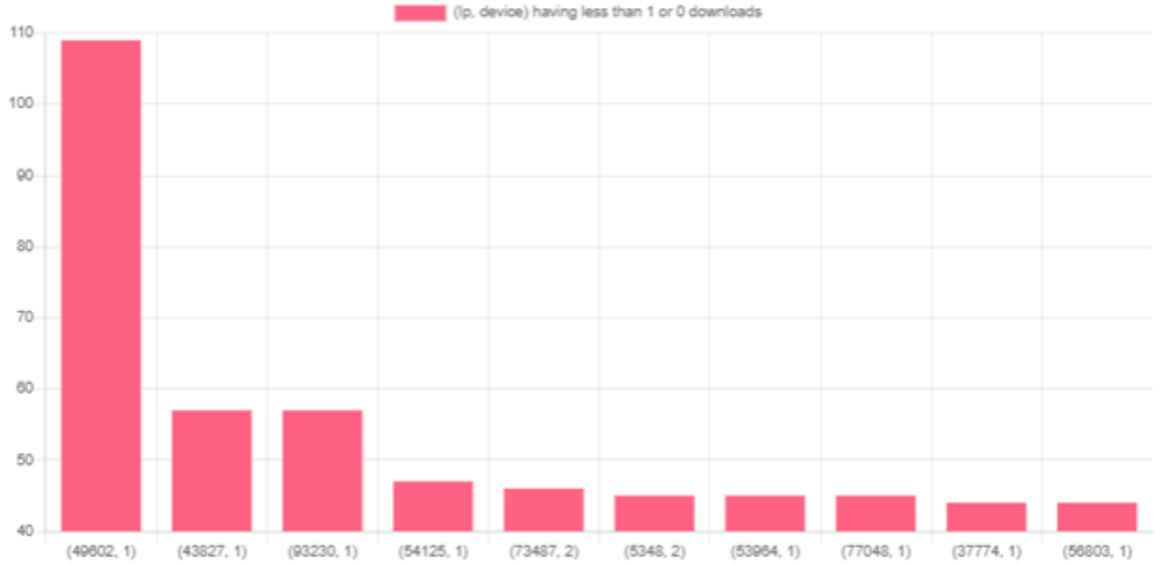
IPs with most clicks



Channel with most clicks



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



Thank You