Detecting Fraud in Online Ad Clicks

Exploratory Data Analysis and Visualization

Yu-Chieh Chen	yc4015
Kevia Qu	kq2153
Sarosh Sopariwalla	sjs2303
Jace Yang	jy3174
Yunzhe Zhang	yz4197

Overview of Dataset

Logs of users' clicks collected from mobile devices by TalkingData

- Data represents 4 days' worth of click-traffic for mobile app ads in China
- Target column ("is_attributed") indicates if the click led to an app download
- Three key characteristics define this dataset:

1. Extremely large
   - Order of 180 million rows
2. Lack of descriptive features
   - 7 features available
3. Target class imbalance
   - Minority class (app is downloaded) ~ 0.24% of data

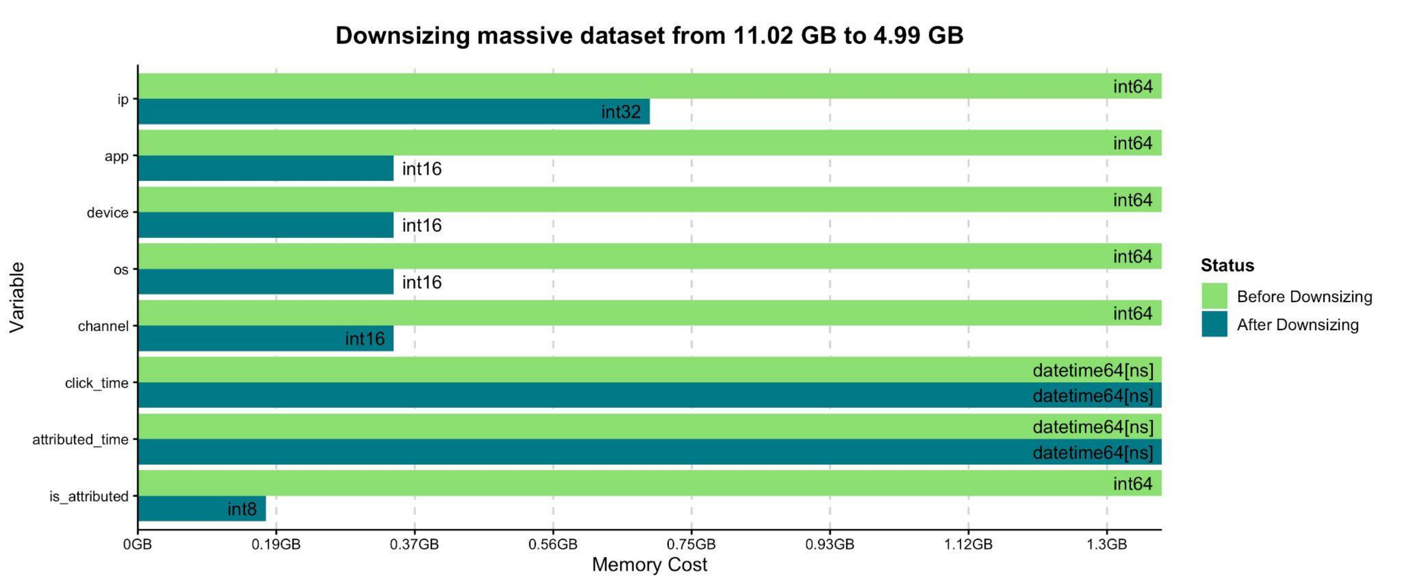
	ip	app	device	os	channel	click_time	attributed_time	is_attributed
0	83230	3	1	13	379	2017-11-06 14:32:21	NaT	0
1	17357	3	1	19	379	2017-11-06 14:33:34	NaT	0
2	35810	3	1	13	379	2017-11-06 14:34:12	NaT	0
3	45745	14	1	13	478	2017-11-06 14:34:52	NaT	0
4	161007	3	1	13	379	2017-11-06 14:35:08	NaT	0
...	...	...	...	...	...	...	...	...
184903885	121312	12	1	10	340	2017-11-09 16:00:00	NaT	0
184903886	46894	3	1	19	211	2017-11-09 16:00:00	NaT	0
184903887	320126	1	1	13	274	2017-11-09 16:00:00	NaT	0
184903888	189286	12	1	37	259	2017-11-09 16:00:00	NaT	0
184903889	106485	11	1	19	137	2017-11-09 16:00:00	NaT	0

184903890 rows × 8 columns

Raw Data Cleaning

Large file size problematic for model training

- Raw dataset is 11 Gb, which causes memory overload during training
    - ~185 million rows (7 features, 1 target)
- Memory reduction techniques were used to downsize the dataset
    - Features recast to lower-memory types



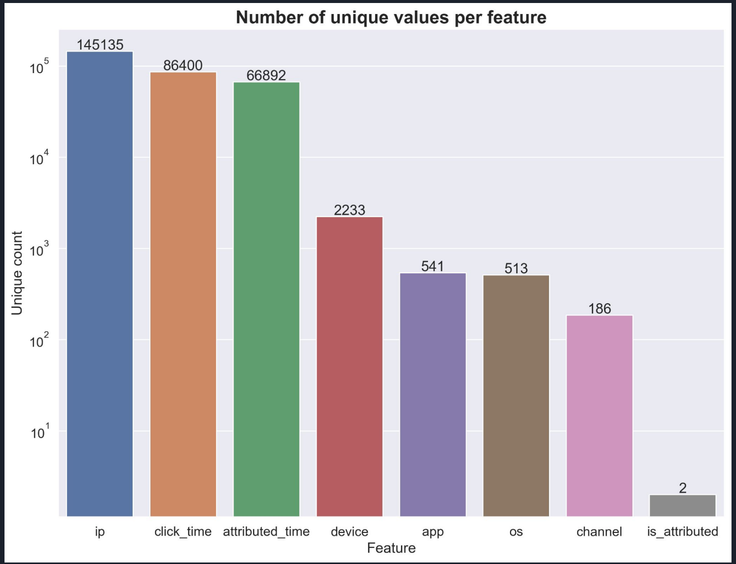
Feature Exploration

Most features are randomized discrete IDs

- 5/7 features are discrete, other 2 are datetime
- Target class for prediction is 'is_attributed', a binary (0,1) indicator

Feature Name	Description	Variable Type
ip	ip address of click	Categorical, unordered
app	ID of app clicked on	Categorical, unordered
device	User's phone type	Categorical, unordered
os	the os version of user's phone	Categorical, unordered
channel	id of mobile ad publisher	Categorical, unordered
click_time	the timestamp of the click in UTC	datetime
attributed_time	the timestamp when app is downloaded after clicking an ad.	datetime

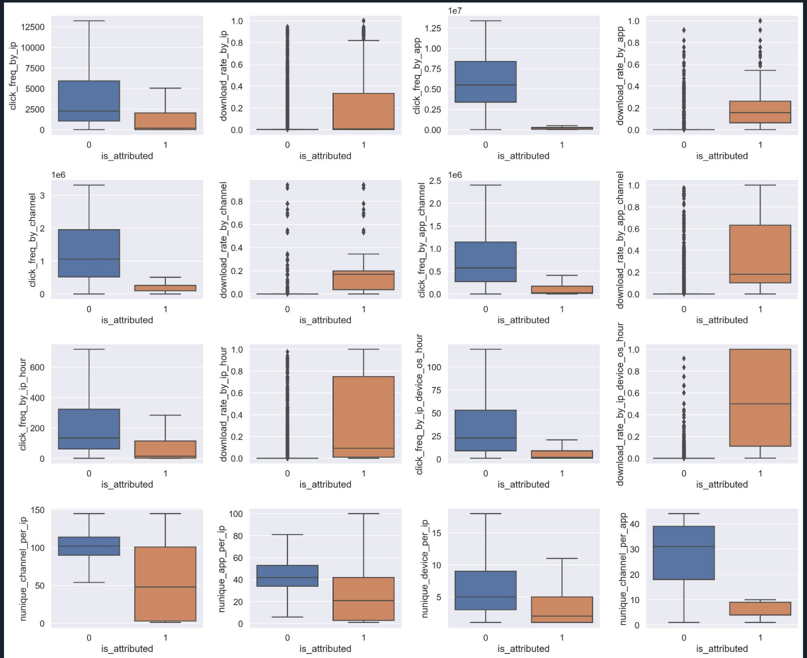
- Categorical features have high cardinality
  - One-hot encoding will expand data too much
- Missingness only appears in 'attributed_time'
  - Only null when 'is_attributed' =0 (no download)
  - Missingness can be one-hot encoded



Data Preprocessing

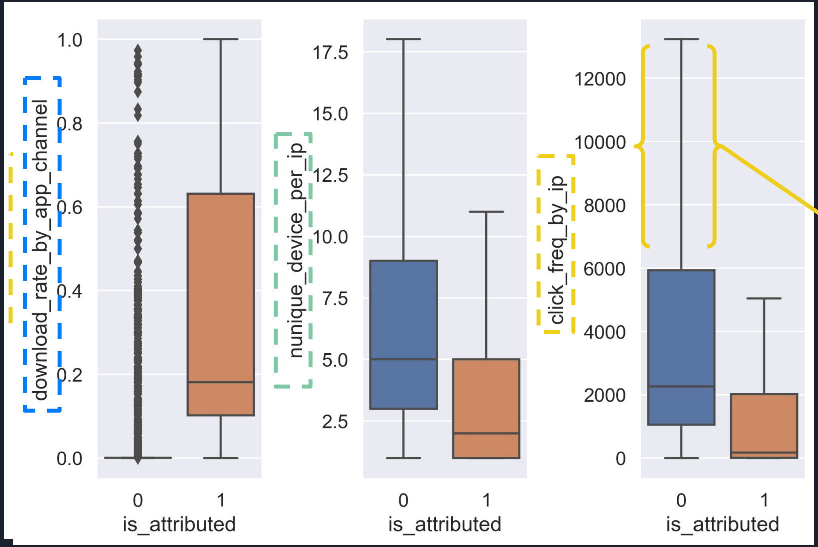
Apply Categorical Encoding to new features

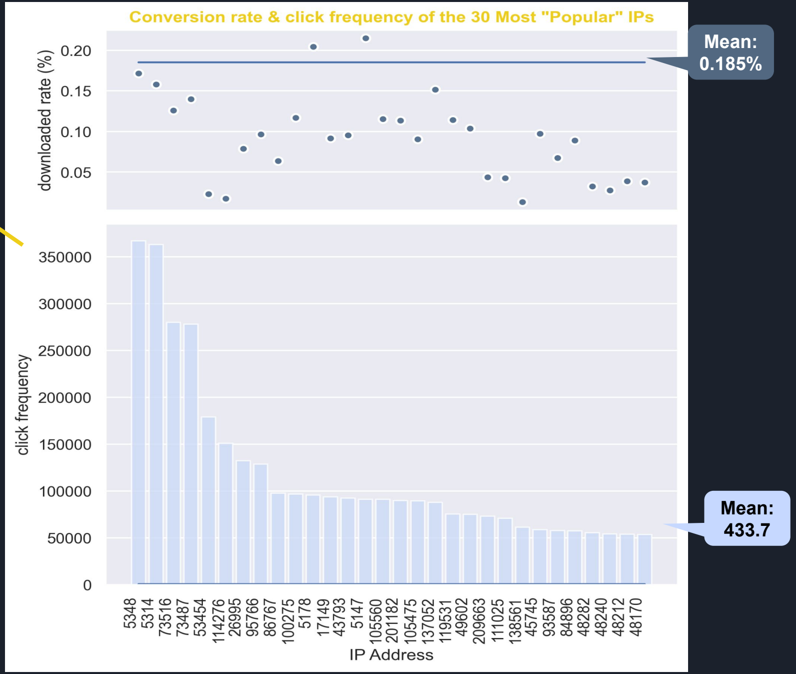
- Categorical encoding was applied by grouping data by the different ID fields to generate numerical features, e.g
    - *click_freq_by_ip*: total clicks by an ip address
    - *download_rate_by_ip_device_hour*: conversion rate for a ip in certain hour. (target encoding)
    - *nunique_device_per_ip*: how many device types one ip have
- New features give more descriptive measures of each ID as related to the target class label
- Does not expand dataset with sparse features like one-hot encoding



Data Preprocessing

New features offer insights into IDs' relationships with target class





💡 Clicks that result in no download tend to come from ip addresses with more clicks in its history, app channels with lower conversion rates, and devices that download one app through multiple channels.

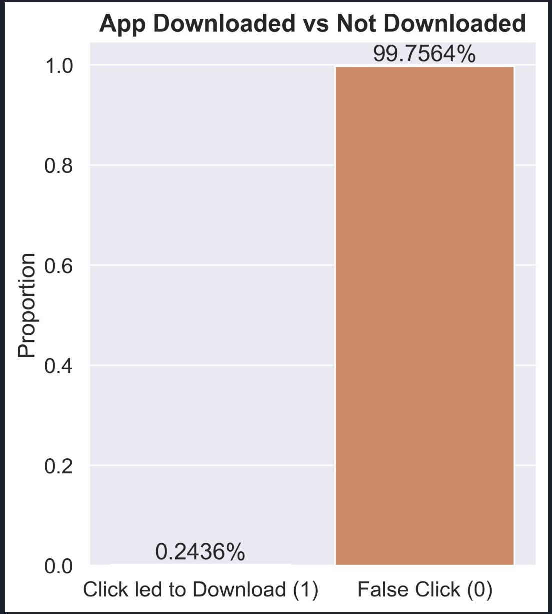
Top 30 IP addresses with the most clicks tend to have download rates < dataset average

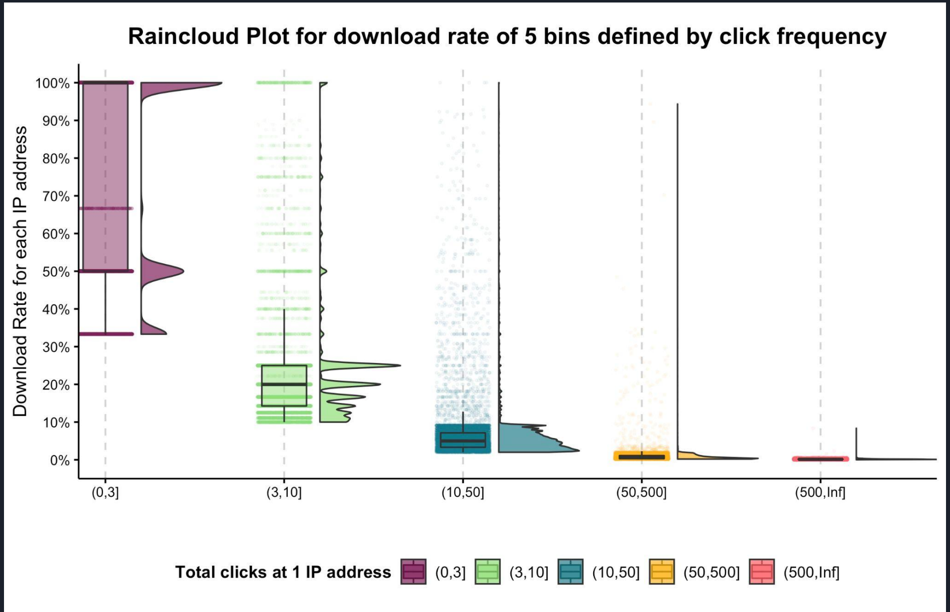
Target Label Exploration
Class labels are extremely imbalanced

- Only 0.24% of data is of positive class (click resulted in a download) among 180 million+ clicks

2 observations that likely cause this:

1. People rarely download apps even after clicking the ad
2. “Click-bots” easily generate millions of clicks with few downloads, and they do this throughout the whole day (next slide)

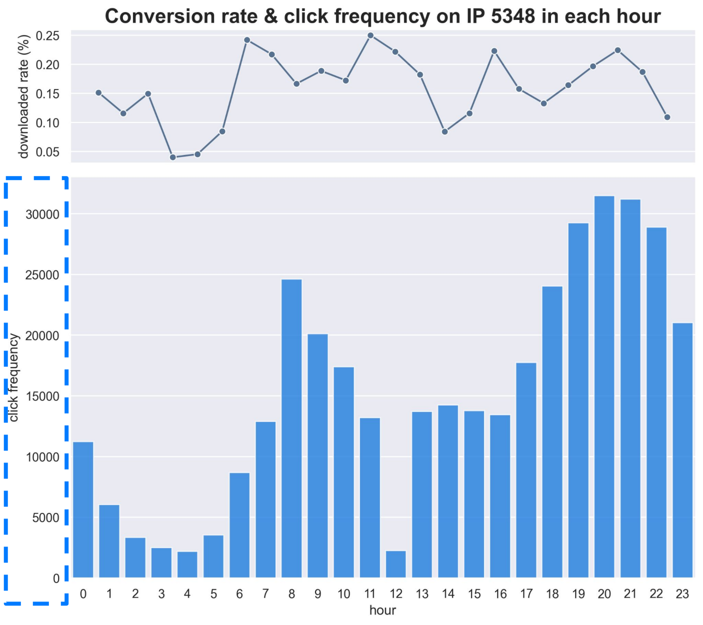


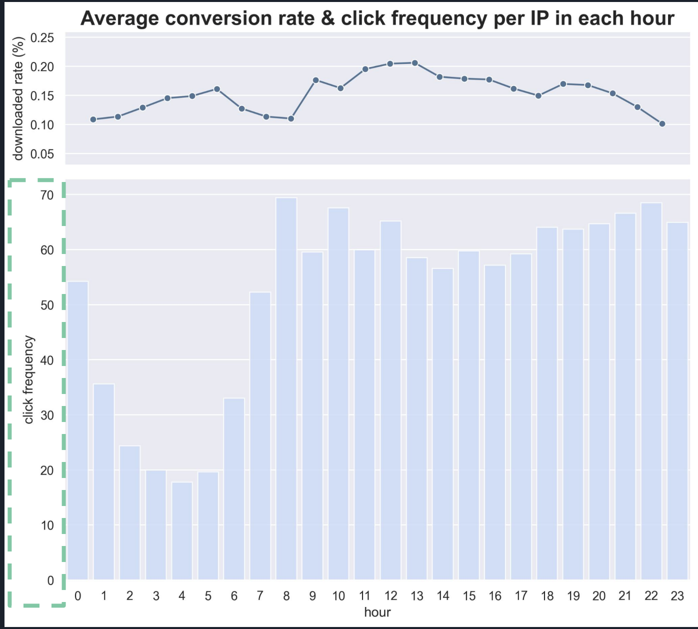


Target Label Exploration

Labels are extremely imbalanced

- “Click-bots”, ip addresses such as ip address 5348, which recorded the highest number of clicks, generate many more clicks across all hours of the day compared to the dataset average.





Sampling and Modelling Techniques

Resampling and balanced modelling will be combined to account for class imbalance

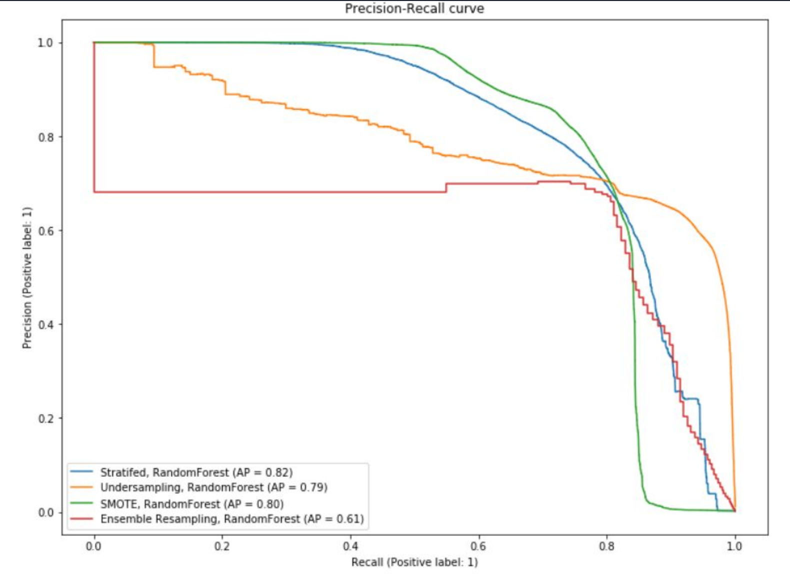
Resampling techniques alone to reduce the impact of target class imbalance (without expanding existing dataset) may not be sufficient:

- Undersampling
- Ensemble Resampling
- SMOTE (on a subsample of data)

Resampling may allow for a model that prioritizes higher recall without large trade-offs in precision

Sampling methods will be combined with adjusting precision/recall thresholds and weighting loss functions for the following modelling methods:

- Logistic Regression
- Random Forest
- Gradient Boosting
- XGBoost
- LightGBM
- Kernel SVMs



Experiment on a small Random Forest (50 classifiers):
Sampling techniques improve recall/precision trade-off over simple stratified training