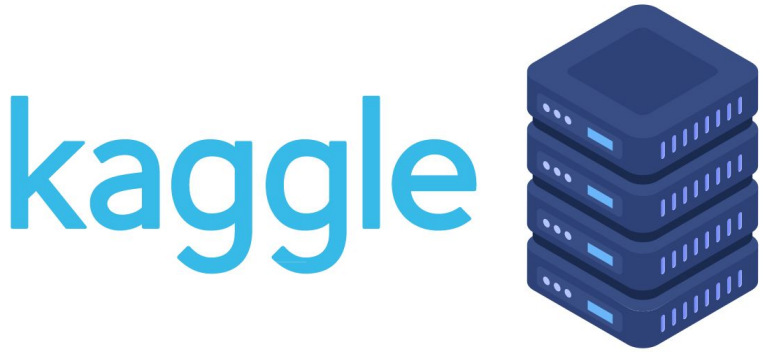
Click Fraud Detection

Frederikke Noerager Julie Ngo
Maggie Lin Maitry Mistry Rick Miao

Detection of Click Fraud

- Volume of click fraud poses a real threat to the survival of companies
- Increases costs
- On average, companies incur 14% of invalid clicks
- Focus on mobile devices and apps



Data Description

- Accessed through Kaggle
- Random-selected observations
- Provided by TalkingData
  - China’s largest independent big data service platform
  - Data was collected from devices in China
    - Results may not generalize outside Chinese market

Data Quality

- Consolidate the number of clicks per IP address over a period of time
    - Convert date format and split into day, month, year, hour
- Predict whether they download or not
- Assumption: 25% of the data is potentially fraudulent, top 25% of most frequent clicks that don’t have downloads are fraudulent

	ip	app	device	os	channel	click_time	attributed_time	is_attributed	click_year	click_day	click_hour	click_month
1	87540	12	1	13	497	2017-11-07 09:30:38	NA	0	2017	Tuesday	9	11
2	105560	25	1	17	259	2017-11-07 13:40:27	NA	0	2017	Tuesday	13	11
3	101424	12	1	19	212	2017-11-07 18:05:24	NA	0	2017	Tuesday	18	11
4	94584	13	1	13	477	2017-11-07 04:58:08	NA	0	2017	Tuesday	4	11
5	68413	12	1	1	178	2017-11-09 09:00:09	NA	0	2017	Thursday	9	11
6	93663	3	1	17	115	2017-11-09 01:22:13	NA	0	2017	Thursday	1	11
7	17059	1	1	17	135	2017-11-09 01:17:58	NA	0	2017	Thursday	1	11
8	121505	9	1	25	442	2017-11-07 10:01:53	NA	0	2017	Tuesday	10	11
9	192967	2	2	22	364	2017-11-08 09:35:17	NA	0	2017	Wednesday	9	11
10	143636	3	1	19	135	2017-11-08 12:35:26	NA	0	2017	Wednesday	12	11
11	73839	3	1	22	489	2017-11-08 08:14:37	NA	0	2017	Wednesday	8	11
12	34812	3	1	13	489	2017-11-07 05:03:14	NA	0	2017	Tuesday	5	11
13	114809	3	1	22	205	2017-11-09 10:24:23	NA	0	2017	Thursday	10	11

```
> colSums(is.na(training))
                ip                app             device                 os 
                 0                  0                  0                  0 
         channel         click_time     attributed_time      is_attributed 
                 0                  0                  0                  0 
```

Proposed Analysis

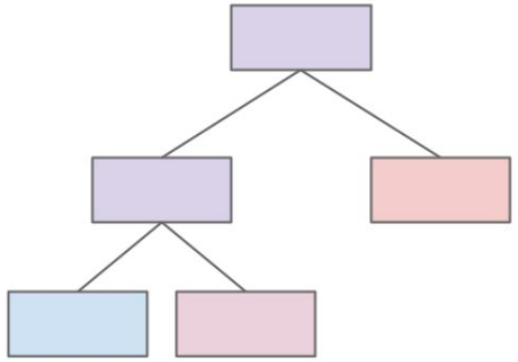
Input variables

- IP address
- Application
- Device
- Operating system
- Timestamp of clicks

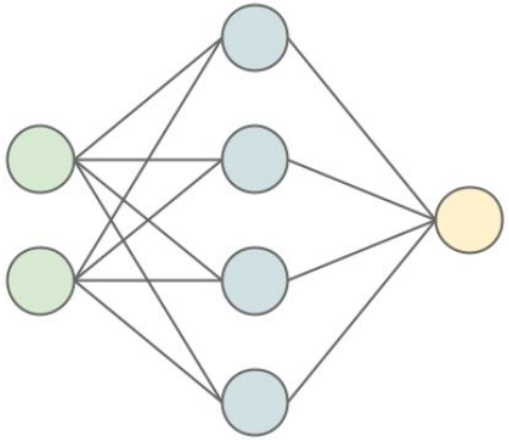
Target variable

- Whether user will install application

Classification models

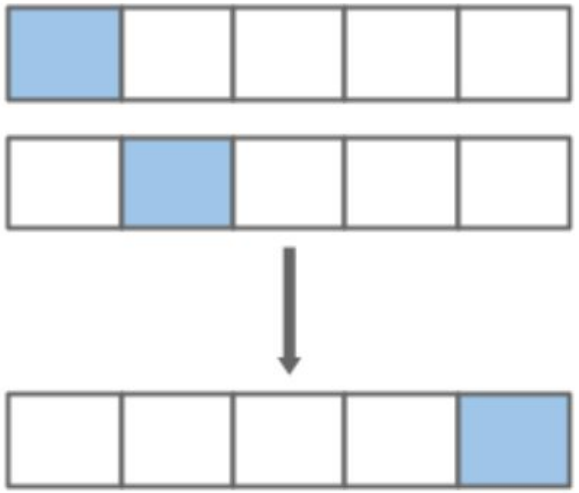


Neural networks

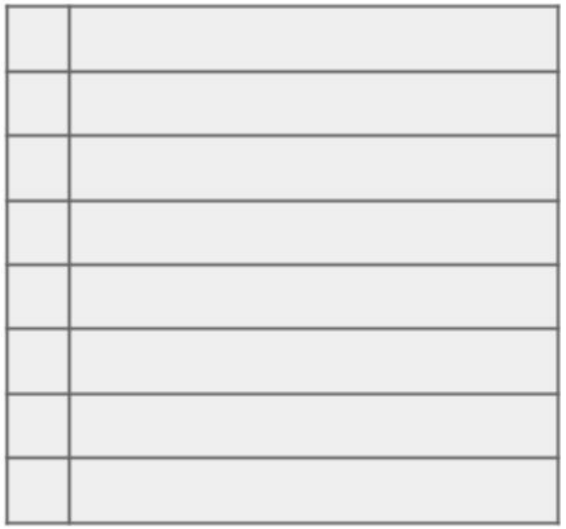


Proposed Analysis

Cross validation



List of fraudulent IP addresses



Expected Results

Conversion rate of 2% ~ 2,000 downloads total

Fraudulent rate of 25% ~ 25,000 fraudulent clicks total

Create an IP blacklist

Low test/train error

Risks

Not enough features/attributes	• **Consequence:** Inaccurate predictions for target variable • **Solution:** Create new features such as combining collinear variables (x_1x_2) or using higher order variables (x^2)
Assumption that the top 90% of clicks are fraudulent	• **Consequence:** High False Positive (FP) error rate • **Solution:** Test different thresholds (i.e. 60%, 70%, 80%) for deviation in FP rates and deviation in frequency of clicks
Data is not scaled appropriately	• **Consequence:** Unfair weights given to certain attributes • **Solution:** Scale dataset so that equal weight is given to attributes

Thank you!