Website P.K. App

Data Analysis on Avazu Mobile Ads Data

Nicole (Linna) Li

Dataset Background

- **Resource:** Kaggle - Click-Through Rate Prediction
  (https://www.kaggle.com/c/avazu-ctr-prediction)

- **Sponsor:** Avazu is a leading multinational corporation in the digital marketing industry, specializing in cross-device advertising and mobile game publishing.

- **Mobile Ads:**
  - Mobile Web
  - Mobile Application
  - Text Message (SMS)
  - Multimedia messaging service (MMS)
  - Mobile Video and TV

  **from Wikipedia*

Avazu
A LEADING ADVERTISING PLATFORM
WITH GLOBAL COVERAGE

Research Objective

- Explore possible methods to discover patterns from a CA data
- Determine the important predictors
- Compare different mobile ads channels
- Draw conclusions based on the research

Tools

- Python (Jupyter)
- MS Excel
- Gephi
- R

Data Preparation

- **Original dataset -** 40,428,967 records in total

- Data Balance and Data Reduction
- Missing Values Detection and Replacement (Missing Value Code → 0)
- Generate derived attibutes on time/date
- Feature/Instance Reduction
- Data Splitting

- **Final dataset**
  - 300,000 records and 24 attributes
  - 80% for training set, 20% for test set

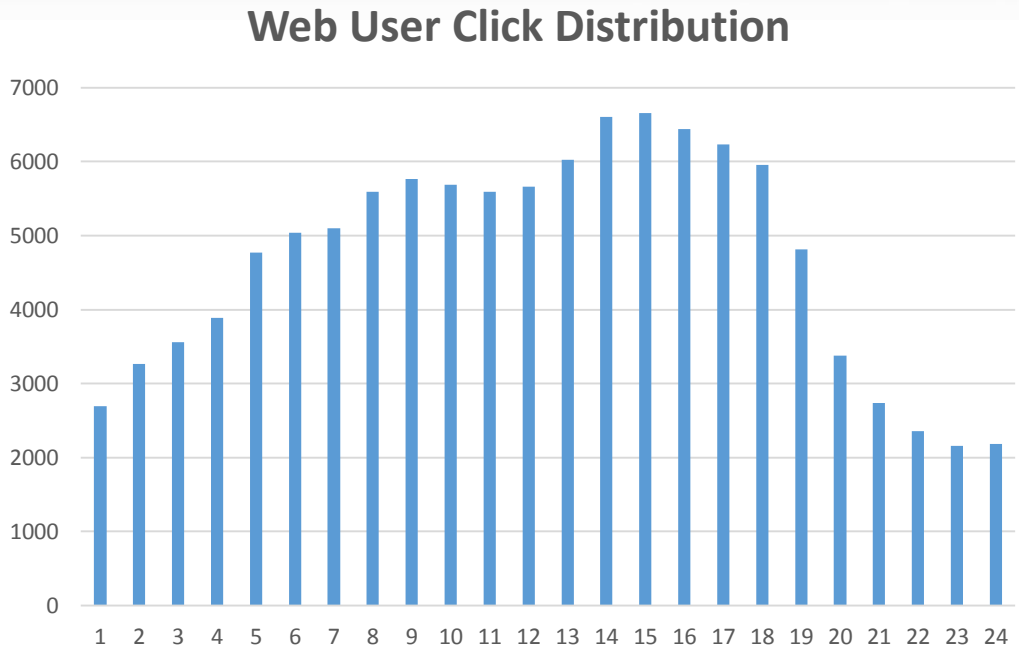
Data Description

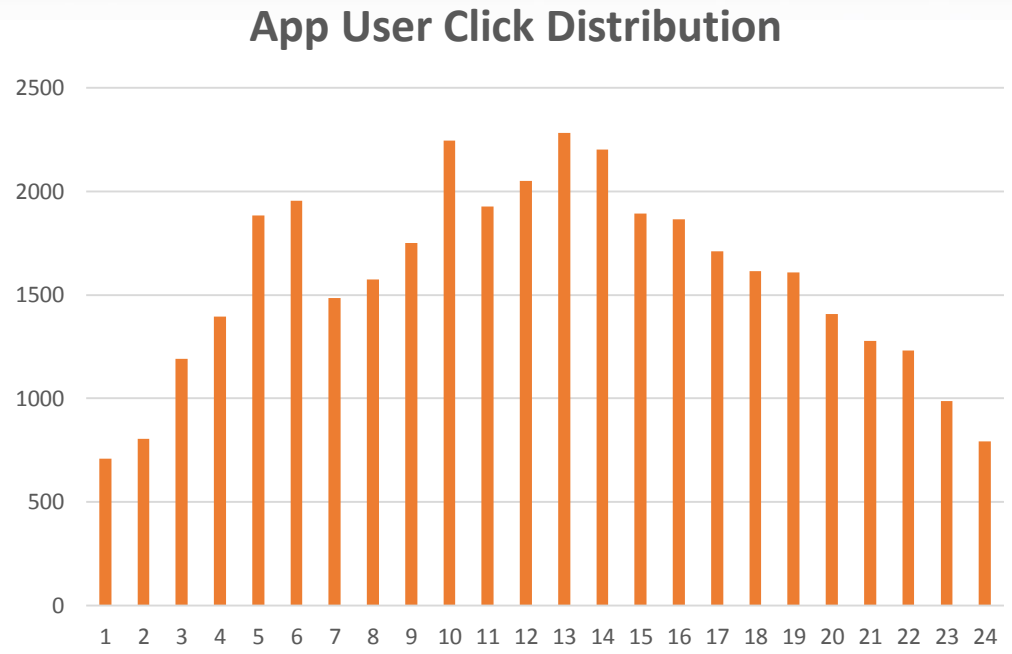
- **Channel:** Website (ID, Domain, Category), App (ID, Domain, Category)
- **Device:** ID, IP, Model, Type, Connect Type
- **Anonymized Categorical Variable:** C1, C14-C21
- **HOUR:** YYMMDDHH → DayOfWeek, HourOfDay
- **Date range:** 2014/10/21 (Mon) – 2014/10/30 (Thu)

Rows

Columns	**299991**	**299992**	**299993**	**299994**	**299995**	**299996**	**299997**	**299998**	**299999**	**300000**
site	0	d9750ee7	0	0	8ab5bdcf	0	0	0	0	1fbe01fe
	0	98572c79	0	0	db11867b	0	0	0	0	f3845767
	0	f028772b	0	0	c0dd3be3	0	0	0	0	28905ebd
app	6efb59d5	0	e2fcccd2	53de0284	0	f888bf4c	f0d41ff1	9c13b419	e2fcccd2	0
	d9b5648e	0	5c5a694b	d9b5648e	0	5b9c592b	2347f47a	2347f47a	5c5a694b	0
	0f2161f8	0	0f2161f8	0f2161f8	0	0f2161f8	0f2161f8	f95efa07	0f2161f8	0

Click Distribution - Channel





- The Website users are more likely (x2.5) to click on ads than App users.
- Different peak periods.
- Web-users want more ads in the afternoon, but the number drops dramatically after 8pm.

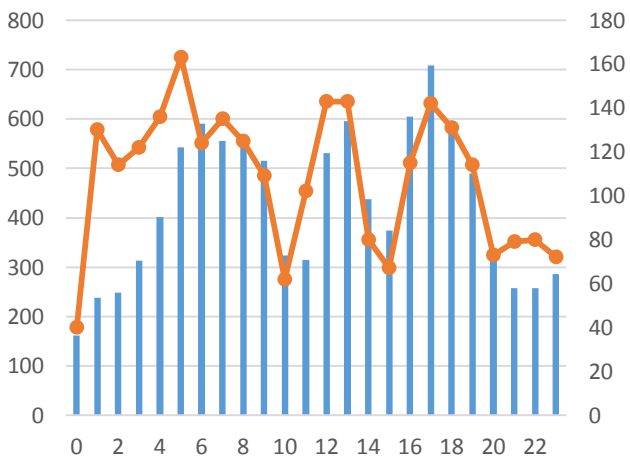
Click Distribution – Day of Week

Web-click

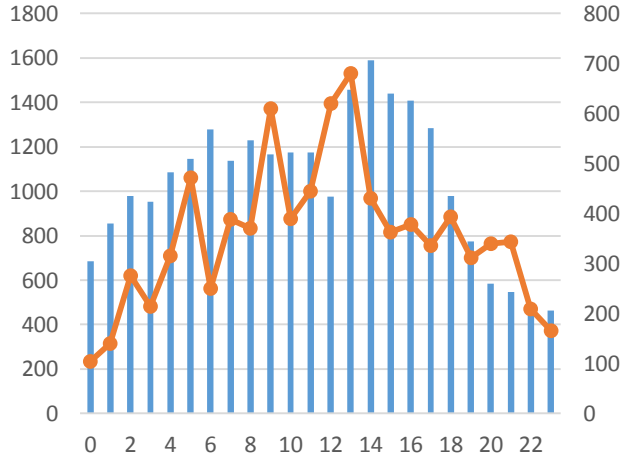
App-click

Date range: 2014/10/21 (Mon) – 2014/10/30 (Thu)

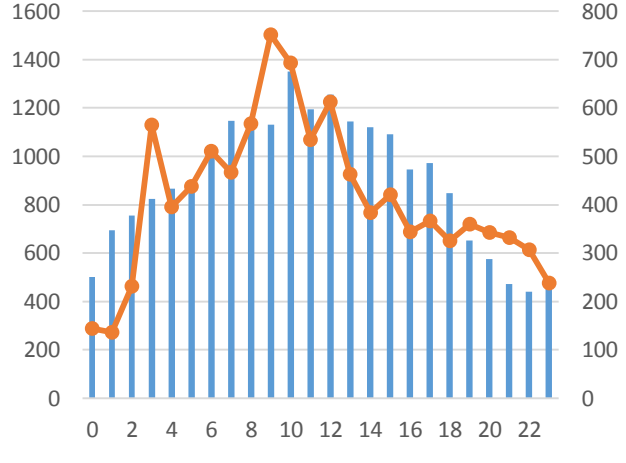
Monday



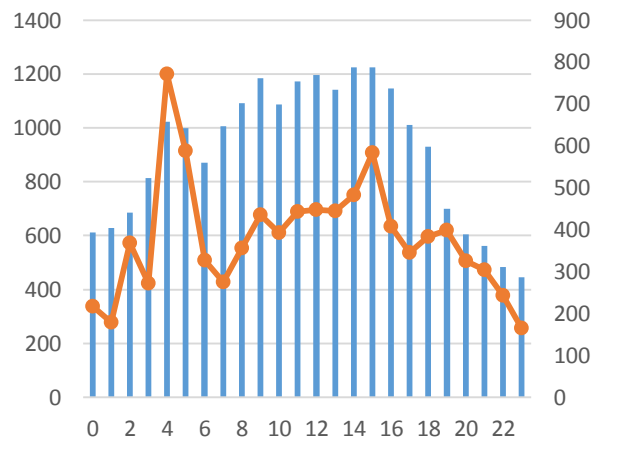
Tuesday



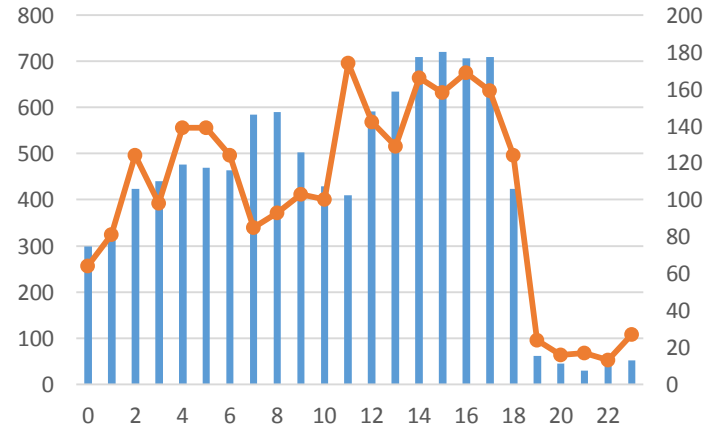
Wednesday



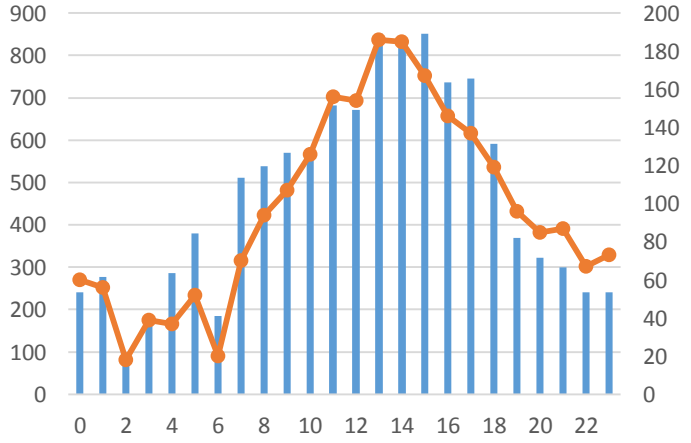
Thursday



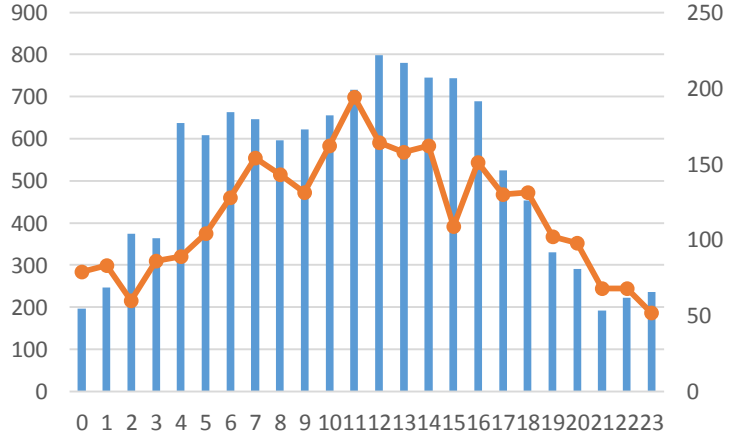
Friday



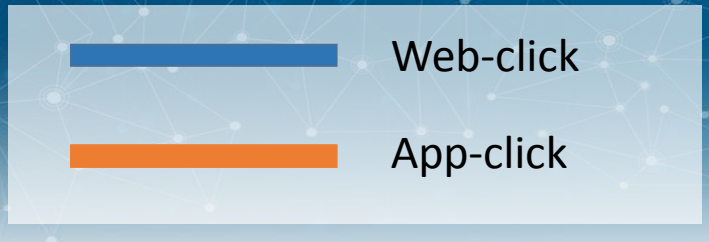
Saturday



Sunday

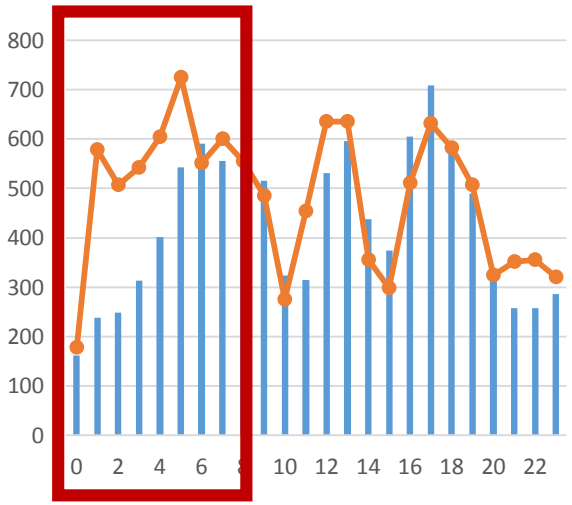


Click Distribution – Day of Week

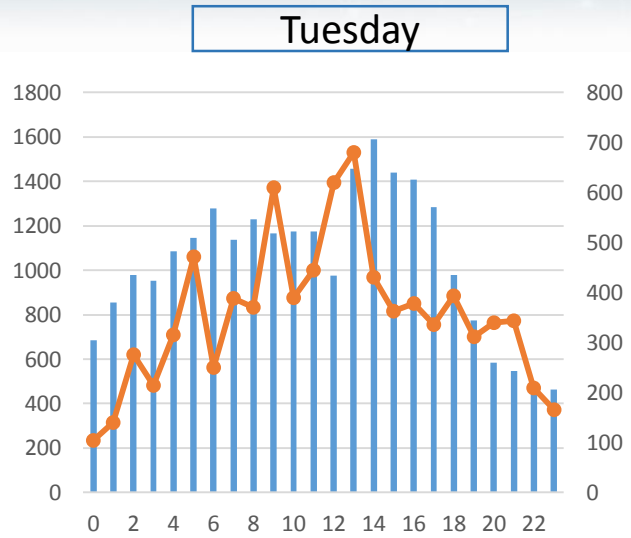


Date range: 2014/10/21 (Mon) – 2014/10/30 (Thu)

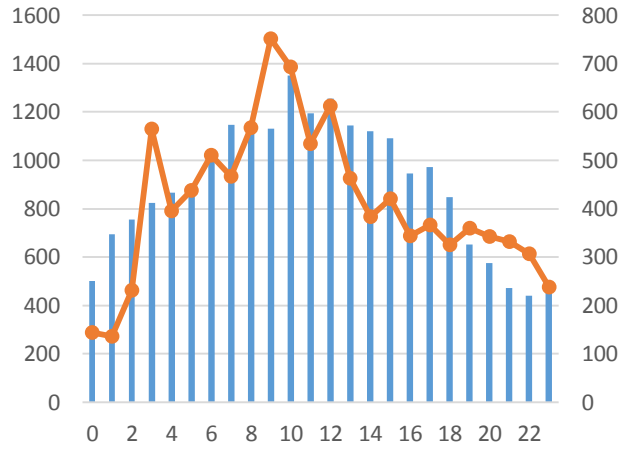
Monday



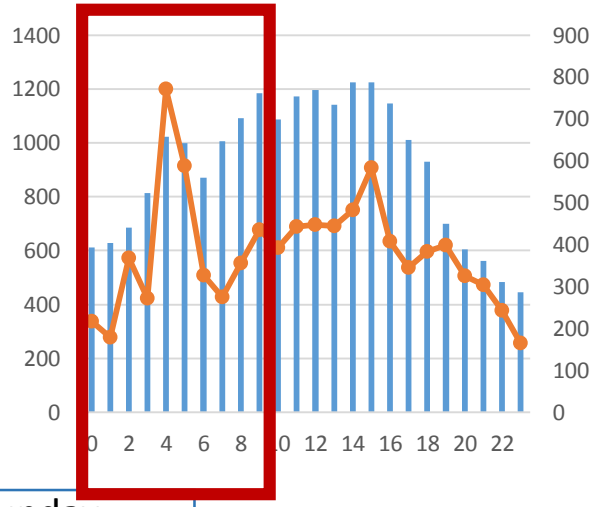
Tuesday



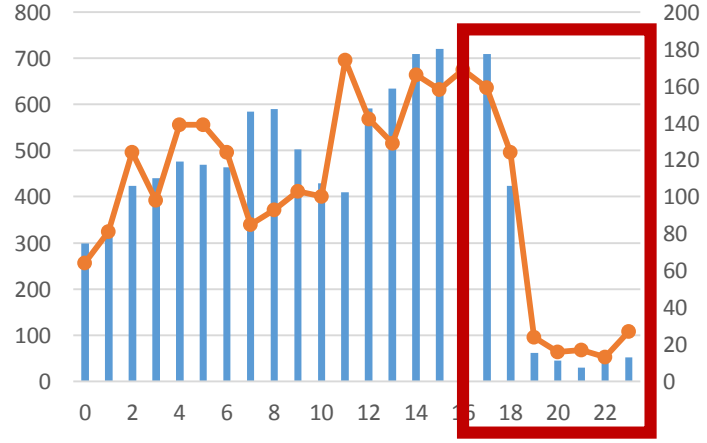
Wednesday



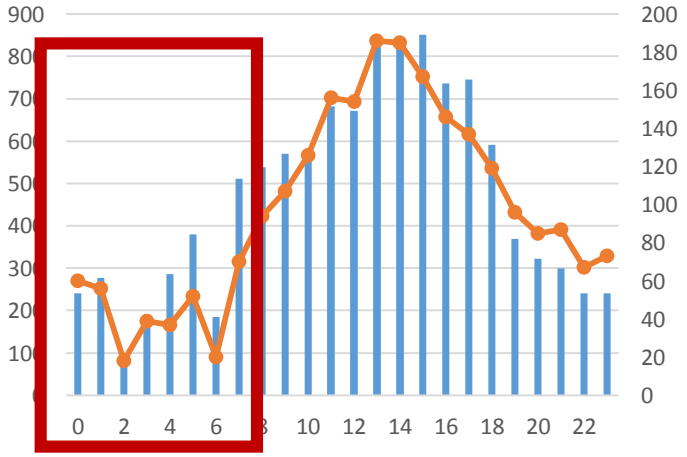
Thursday



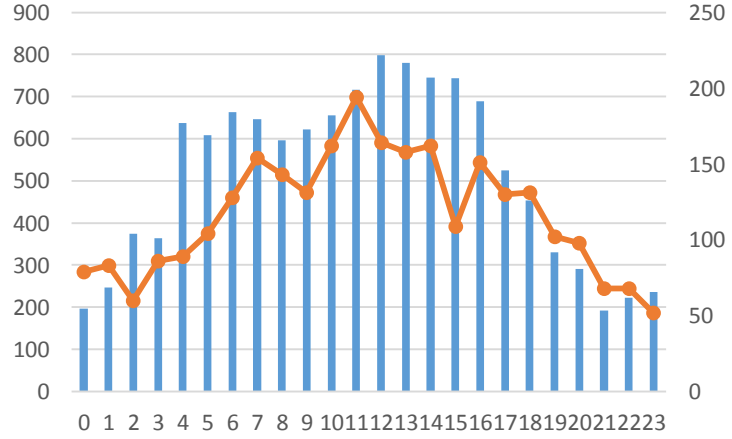
Friday



Saturday



Sunday



Click Distribution

Findings:

- Click happens on Website more than on App.
- The hourly patterns of Web-click and App-click are not same.
- During the weekdays, App-click tends to be high in the early morning and evening.
- During the weekdays, 12pm to 16pm can be an active period for Ads clicking.
- People don’t like clicking on Ads on Friday after work.

Simulated Click-Through Rate

Purpose: Find the easily-clicked Ads and Ads Category.

Original definition

- The percentage of people visiting a web page who access a hypertext link to a particular advertisement.
- Typically, CTR is less than 1%.

Re-definition for this research

- The data set contains much more “Clicked” data.
- With the grouped sample data, the Click Through Rate would be the rate between “Clicked” and the total number of cases in the group.

- Group by: Ad Group ID, Ad ID
- If the ‘appearance’ of a ID is less than the Threshold, this ID will be kicked off. Otherwise, it’s easy to get 100%.
- Threshold: 5 for Ad Group ID, 15 for Ad ID

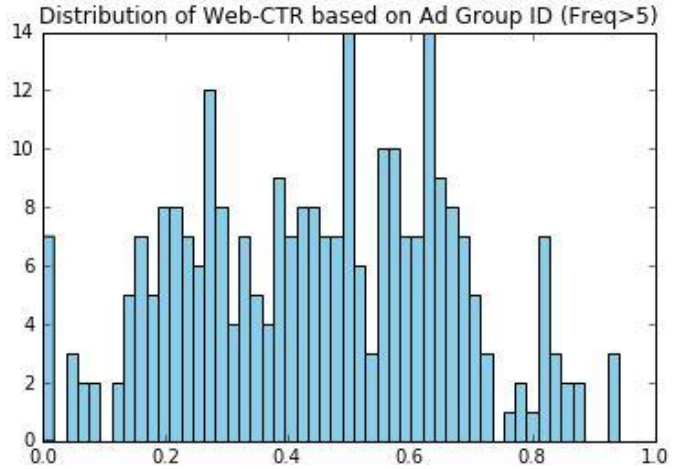
The target is not only a particular Ad, but also a particular category of Ads.

Click-Through Rate - Based on Ad Group ID (C17)

Number of unique Ad Group for Web-click: 297

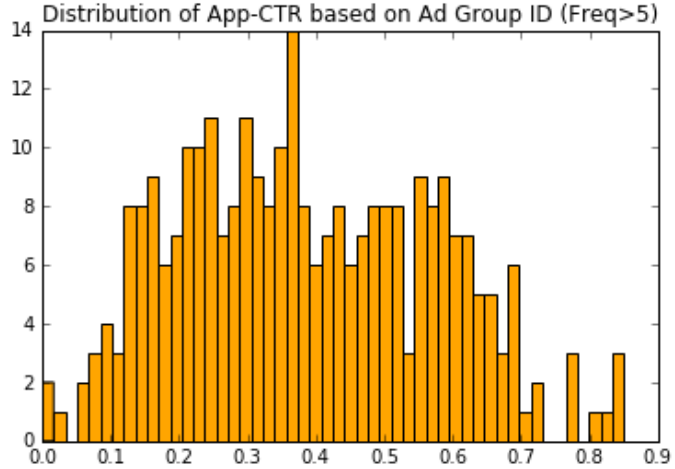
Number of unique Ad Group for App-click: 302

Website Click



webGroup	Total	Click	CTR
1694	101	95	0.940594
2518	45	42	0.933333
2662	26	24	0.923077
2286	1008	886	0.878968
2295	2850	2474	0.86807
827	262	225	0.858779
2101	7	6	0.857143
2569	400	338	0.845
1903	24	20	0.833333
2659	6	5	0.833333

App Click



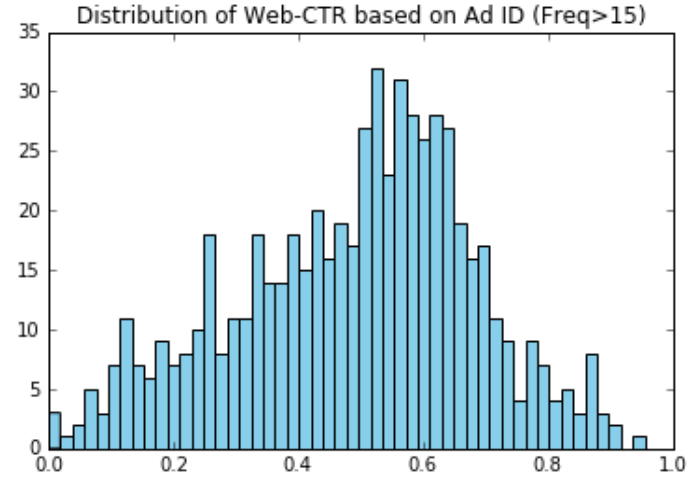
appGroup	Total	Click	CTR
2659	20	17	0.85
1926	126	106	0.84127
2295	119	100	0.840336
2702	11	9	0.818182
2438	10	8	0.8
2638	18	14	0.777778
2421	31	24	0.774194
2331	190	146	0.768421
1272	152	109	0.717105
2510	7	5	0.714286

Click-Through Rate - Based on Ad ID (C14)

Number of unique Ad ID for Web-click: 1416

Number of unique Ad ID for App-click: 1483

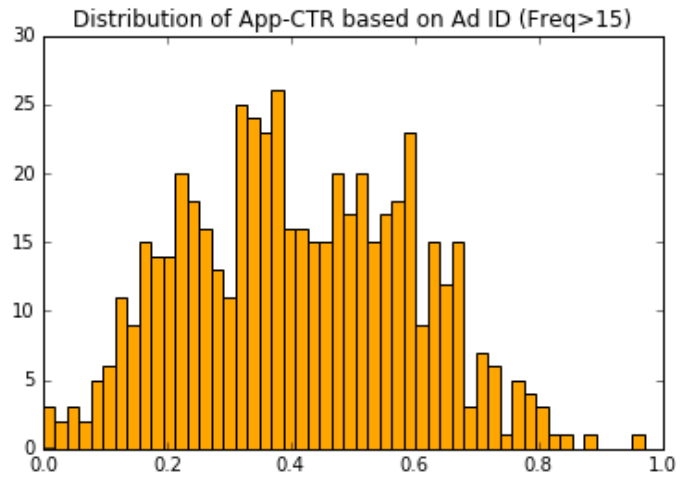
Website Click



webAds	Total	Click	CTR
21814	22	21	0.9545
21812	23	21	0.9130
17759	20	18	0.9000
20018	503	446	0.8867
22272	171	151	0.8830
20635	305	269	0.8820
21277	301	264	0.8771
21914	16	14	0.8750
8996	80	70	0.8750
22270	70	61	0.8714

webAds	Total	Click	CTR
20019	505	440	0.8713
20093	2850	2474	0.8681
23087	22	19	0.8636
21918	22	19	0.8636
8994	85	73	0.8588
23376	33	28	0.8485
8995	97	82	0.8454
22804	24	20	0.8333
20015	64	53	0.8281
20346	1410	1166	0.8270

App Click



appAds	Total	Click	CTR
16989	33	32	0.9697
18936	27	24	0.8889
20093	119	100	0.8403
20274	29	24	0.8276
21590	16	13	0.8125
20215	26	21	0.8077
21791	254	205	0.8071
18925	29	23	0.7931
20009	37	29	0.7838
22756	18	14	0.7778

appAds	Total	Click	CTR
23553	18	14	0.7778
12472	31	24	0.7742
20345	100	77	0.7700
21677	39	30	0.7692
20346	90	69	0.7667
23216	37	28	0.7568
21595	20	15	0.7500
20271	38	28	0.7368
18934	37	27	0.7297
21273	92	67	0.7283

Click Prediction Analysis – Overall

Data Modification

- Create a new binary attribute to define if a case is about Web-click or App-click
- If the case is Web-click, fill “1”. Otherwise, fill “0”.
- For categorical data, only include the ones with 35 or less levels

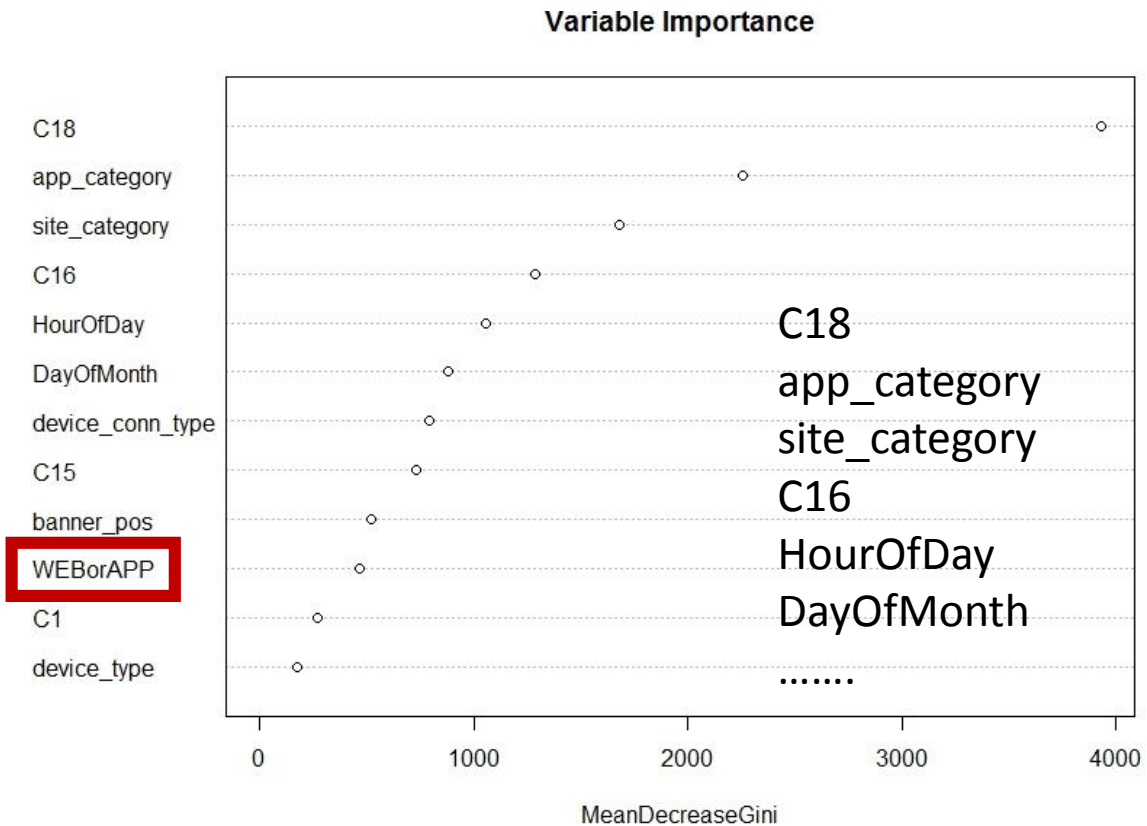
Methodology

- Decision Tree

Measure	Value
Accuracy	0.6098
Sensitivity	0.6064
Specificity	0.6081

- Random Forest
  - Prune the Random Forest (ntree=300)

Measure	Value
Accuracy	0.6509
Sensitivity	0.6419
Specificity	0.6591



Click Prediction Analysis – Web-Click

Data Modification

- Extract the web-click data from the full-data set.
- Keep numerical data and categorical data that with less than 50 levels.

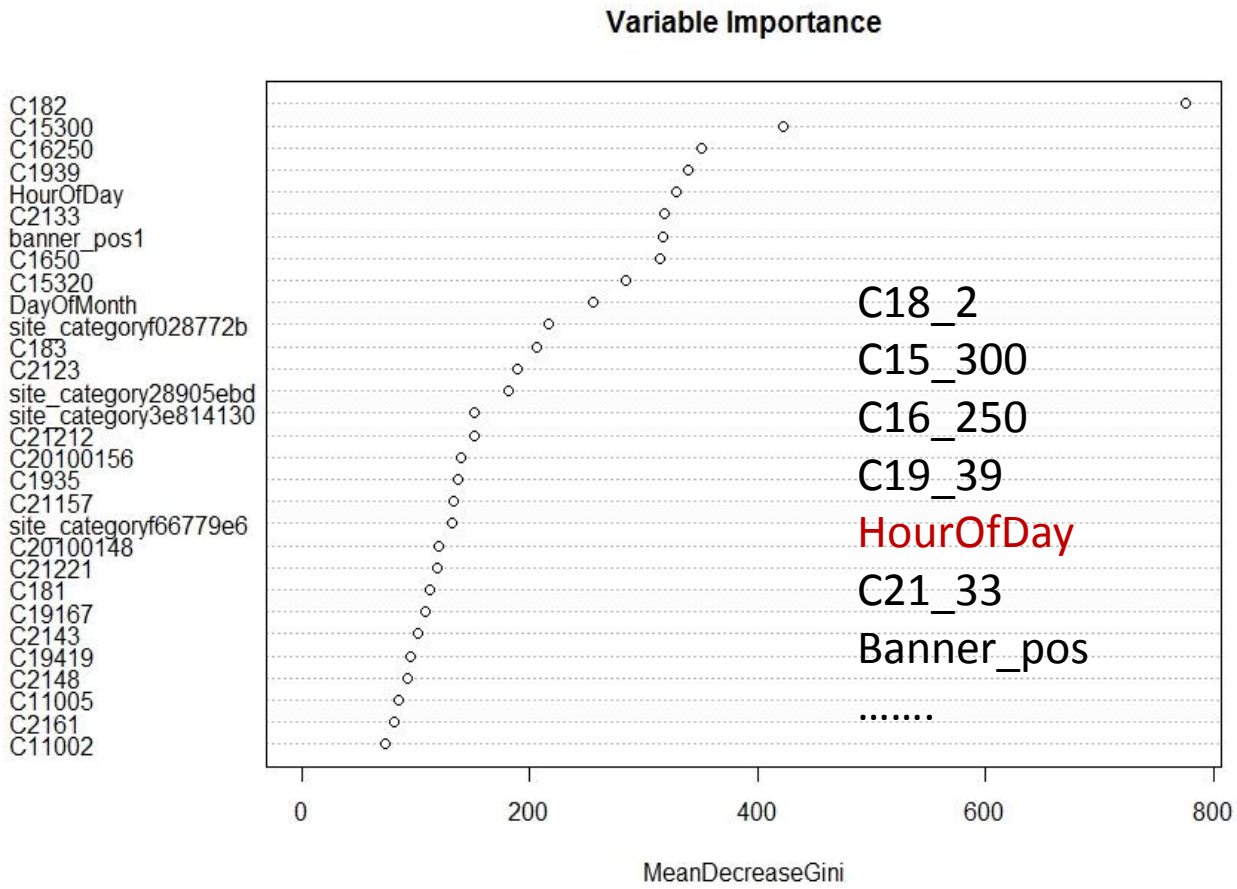
Methodology

- Decision Tree

Measure	Value
Accuracy	0.6105
Sensitivity	0.5932
Specificity	0.6233

- Random Forest
  - Prune the Random Forest (ntree=150)

Measure	Value
Accuracy	0.6614
Sensitivity	0.6650
Specificity	0.6546



Click Prediction Analysis – App-Click

Data Modification

- Extract the app-click data from the full-data set.
- Keep numerical data and categorical data that with less than 50 levels.

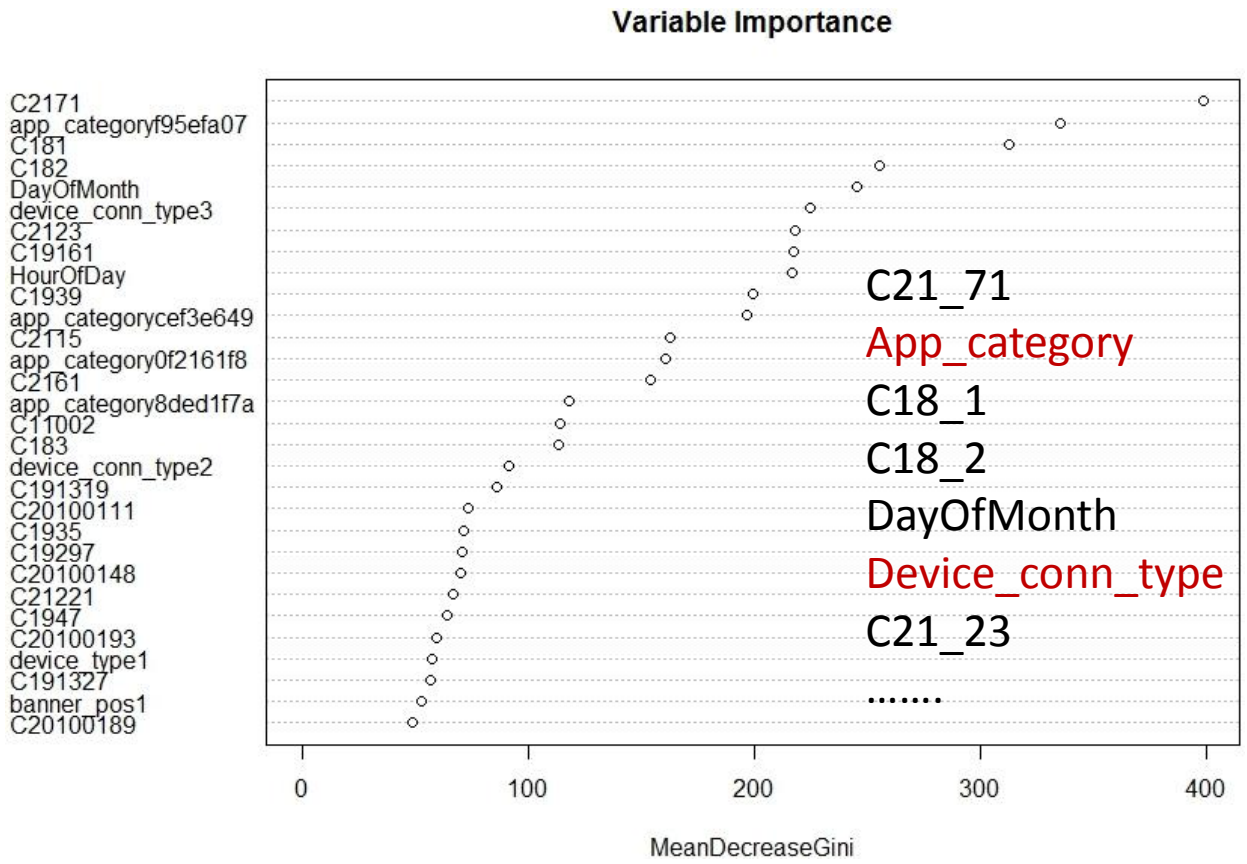
Methodology

- Decision Tree

Measure	Value
Accuracy	0.6720
Sensitivity	0.6960
Specificity	0.6072

- Random Forest
  - Prune the Random Forest (ntree=150)

Measure	Value
Accuracy	0.6614
Sensitivity	0.6650
Specificity	0.6546

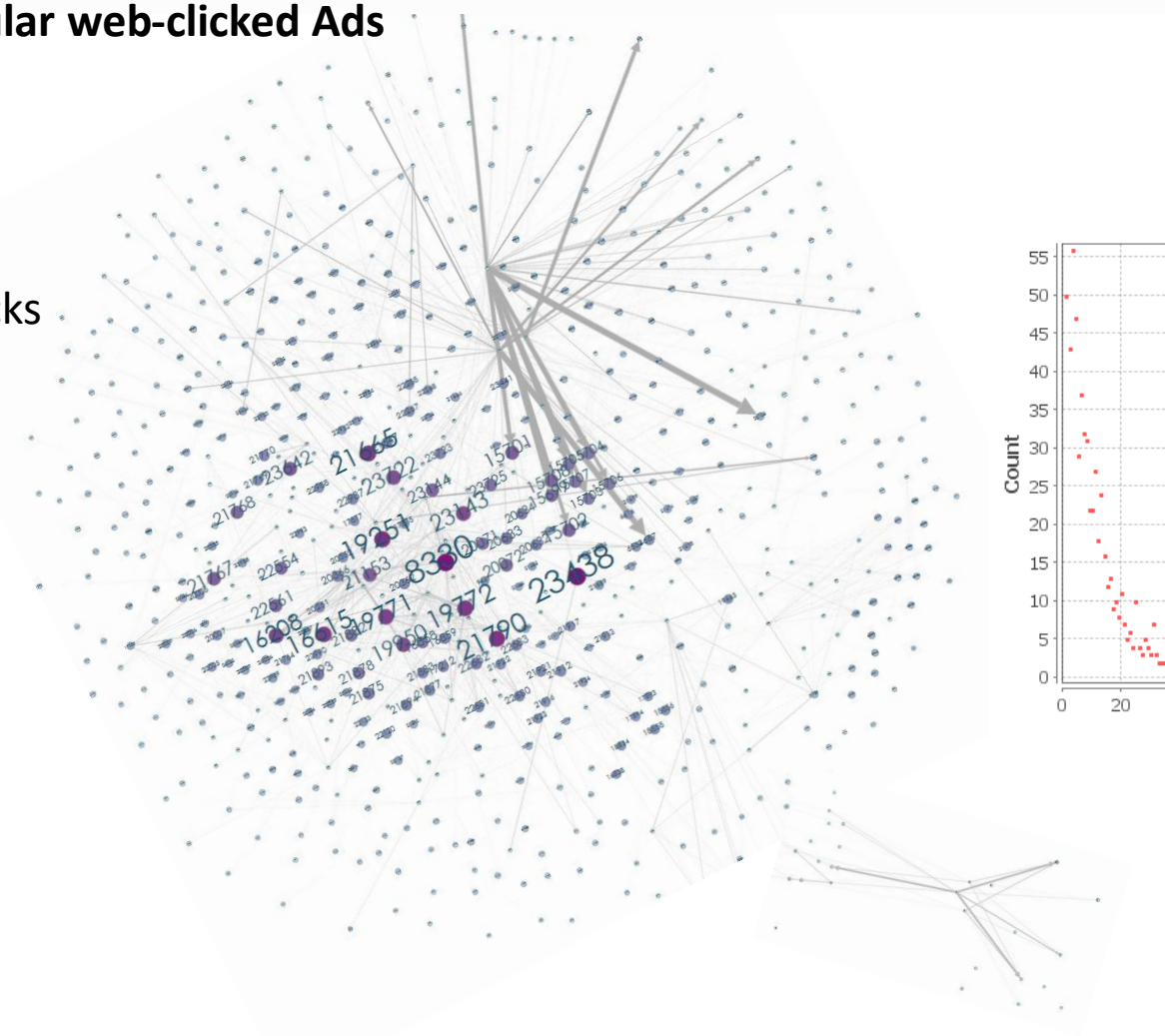


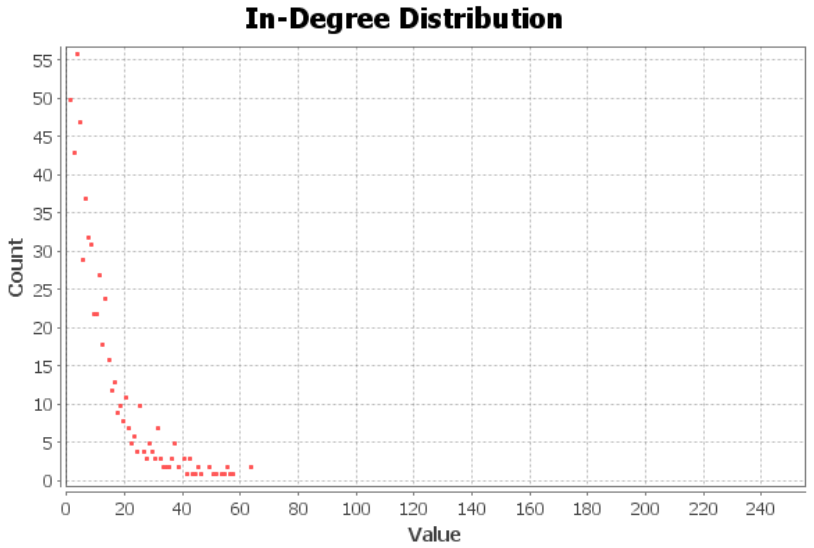
Social Network Analysis on Web Ads

Objective: Find the most popular web-clicked Ads

Node size: In-degree
Node color: In-degree
Label size: In-degree
Edge thickness: Number of clicks

Site ID → Ads



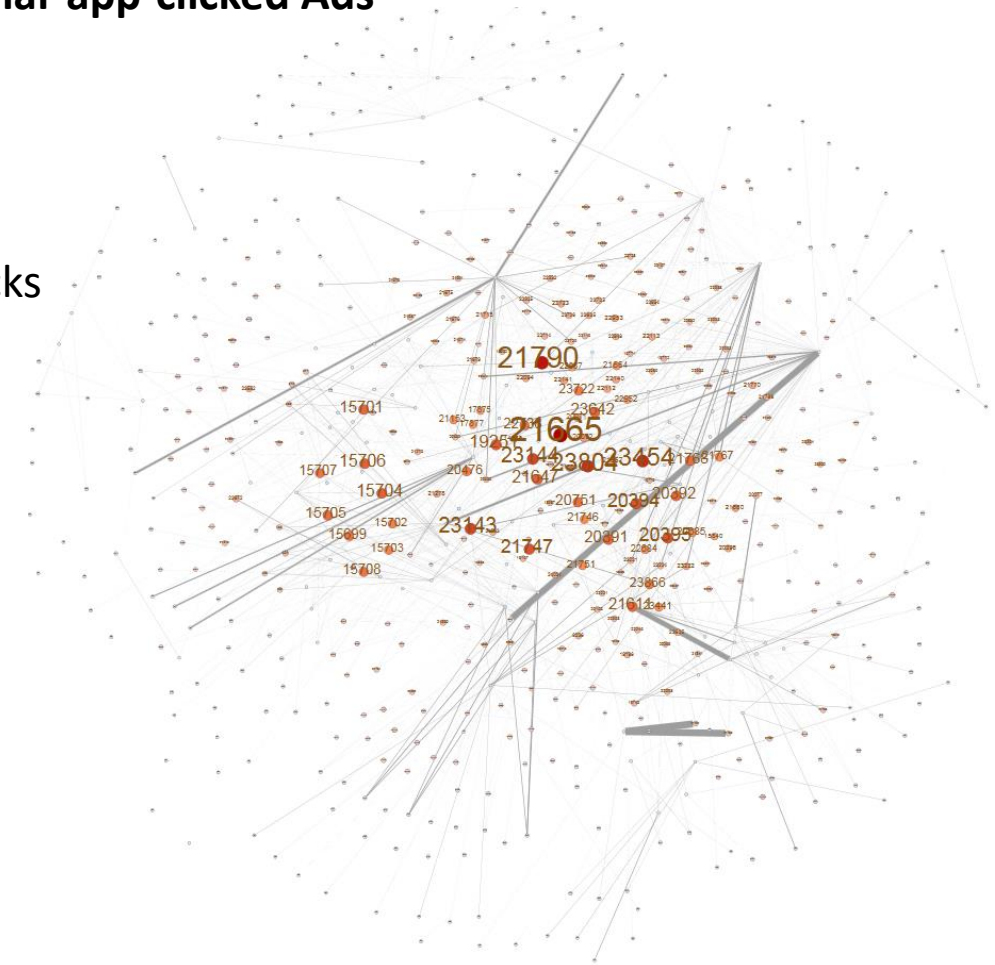


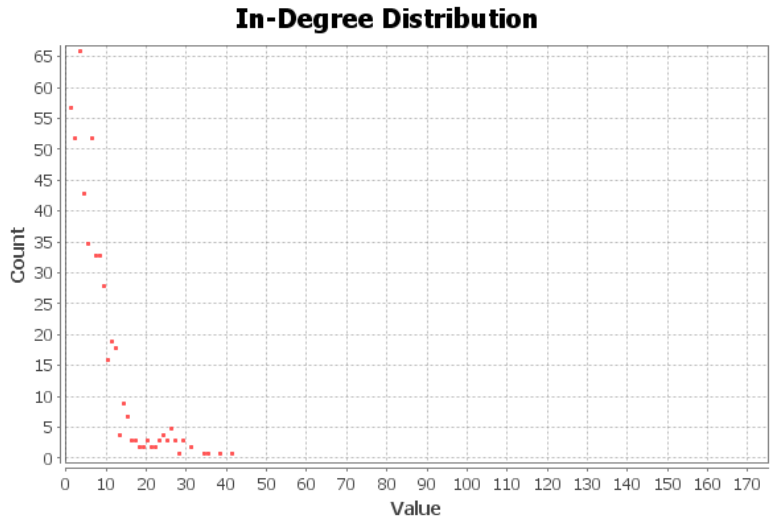
Social Network Analysis on App Ads

Objective: Find the most popular app-clicked Ads

Node size: In-degree
Node color: In-degree
Label size: In-degree
Edge thickness: Number of clicks

App ID → Ads





Conclusion – Model Results

- Website Ads are more likely to be clicked.
- Important predictors:
  - Common: Good timing can bring more clicks on Ads.
  - Web-click: The characteristic of Ads itself is more important
  - App-click: External influences, such as App category and Device Connection Type
- Random Forest performs better than Decision Tree.
- Social Network is a good way to visualize the click pattern.

Conclusion – Difficulties

- Data set from Computational Advertising is always very large.
- Data set is unbalanced.
- ID or IP information matters.
- Lack of information.

Future Research

- All the researches are based on the sampled data. Include more in future.
- Try more classification algorithms for high-level data.
- Deal with unbalanced data.
- Depending on current work, add the cost of misclassification.
- Association analysis on IP and Ads.

Reference

- Powers, D. & Xie, Y. (1999), Statistical Methods for Categorical Data Analysis
- Trofimov, I. & Kornetova, A. & Topinskiy, V., Using boosted trees for click-through rate prediction for sponsored search
- Berrar, D. (2012), Random forests for the detection of click fraud in online mobile advertising

Thank you