




Storyline for 242 Presentation

Google Slides - create and edit presentations online, for free.

Create a new presentation and edit with others at the same time. Get stuff done with or without an internet connection. Use Slides to edit PowerPoint files. Free from Google.

 <https://docs.google.com/presentation/d/1XtuWunGpKp6ZsJuV7gl-Gcyi8i5P4VtqiljDBU3Vt1k/edit#slide=id.p>



Google Colaboratory

 https://colab.research.google.com/drive/1hpt0L824lhWyRV3vTFu6cKV113_zWKVg



Presentation Timeline

Introduction (30s) - Chloe

- Explain the topic, data
 - For CTR (Click-through Rate) prediction, a traditional industrial model would use standard classification algorithms to provide predictions on billions of events per day, using a correspondingly large feature space, and then learn from the resulting mass of data. Therefore, the goal of our group is to build a Machine Learning model that beats standard classification algorithms in real industrial settings by using 11 days worth of Avazu data. Our model may eventually be applied to the advertising industry as a better alternative for traditional models.

Data Analysis - Yunshun (1 min)

Feature Engineering - Alan (1.5 min)

Model and Training - Jo (1.5 min)

- Model Selection
- Pipeline

- Steps
 - Convert all data into Hash Int (if not originally int) - Light GBM?
- Results
 - Metrics

Conclusion and Future Work - Chloe (30s - 1 min)

- Future work

Original Proposal

Analytic Techniques

- Feature engineering: Generate new useful features for the models
- Logistics Regression: Create a baseline and measure the importance for every feature using the values and the weights
- Tree Models (Decision Tree, Random forest, Gradient Boosting/Adaboost): Use more complex methods to improve the model performance
- Model Integration: Integrate models and get our final results
- Visualization: Visualize data original distributions as well as the results

Goals/Evaluation Metrics

We'll use the following metrics to evaluate different machine learning models to measure accuracy and performance. Confusion Matrix will show the performance of our classification model on a set of test data (1 day of click-through data) by measuring recall, precision, accuracy, and AUC(Area Under The Curve) - ROC (Receiver Operating Characteristics) curve. AUC-ROC curve will give performance of a classification model at all thresholds where it plot true positive rate (TPR) vs false positive rate (FPR). Also, cross-validation will be used to assess the predictive performances of our models.

Impact

Predicting ad click-through rates (CTR) is central to the multi-billion dollar online advertising industry. As an essential part of the marketing funnel model, an increase in the prediction accuracy would empower advertisers to have better

estimate on the overall campaign performance, and therefore adjust their marketing plan to optimize their ROI (Return on Investment).

Introduction

Data

The data set is collected by AVAZU. The 11 days of click-through rate is collected.

Training set - 10 days of click-through data, ordered chronologically.

- Test set - 1 day of click-through data. This will be used to test the predicted model.
-

Data Exploration (Train)

General Meta Data

Data Size - 40 million

count	40428967
-------	----------

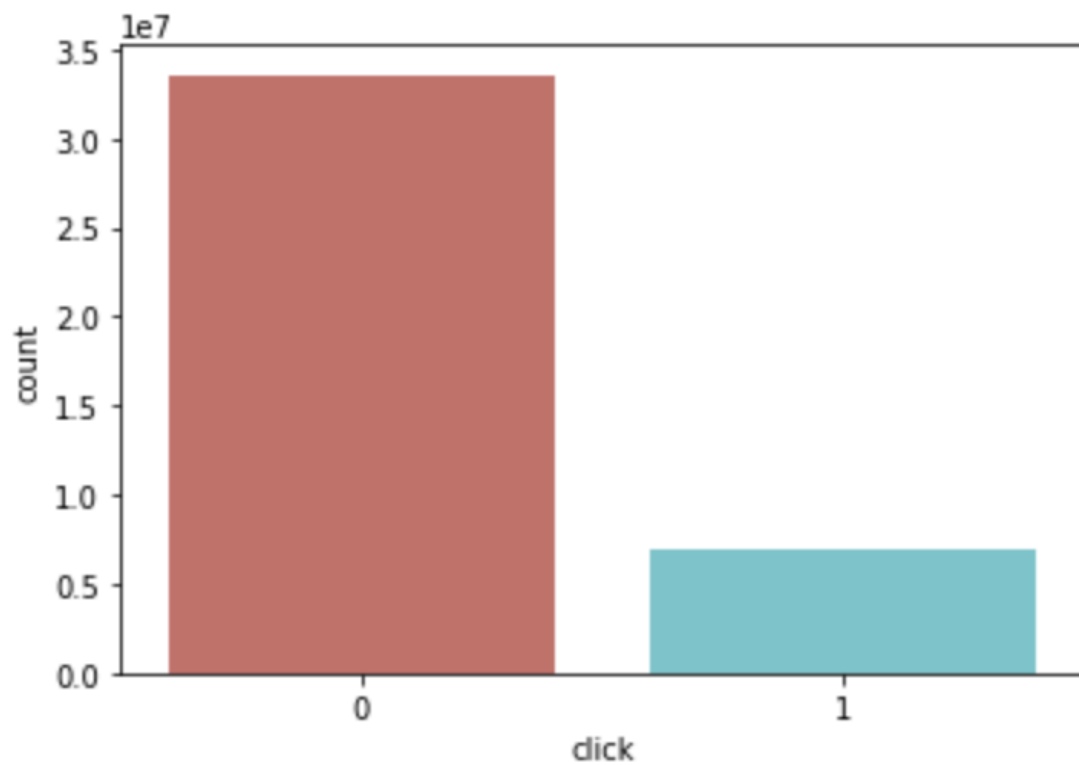
Feature Types

id	object
click	int64
hour	datetime64[ns]
C1	int64
banner_pos	int64
site_id	object
site_domain	object
site_category	object
app_id	object
app_domain	object
app_category	object
device_id	object
device_ip	object
device_model	object
device_type	int64
device_conn_type	int64
C14	int64
C15	int64
C16	int64
C17	int64
C18	int64
C19	int64

C20	int64
C21	int64

Distribution of Clicks

0	0.830194
1	0.169806



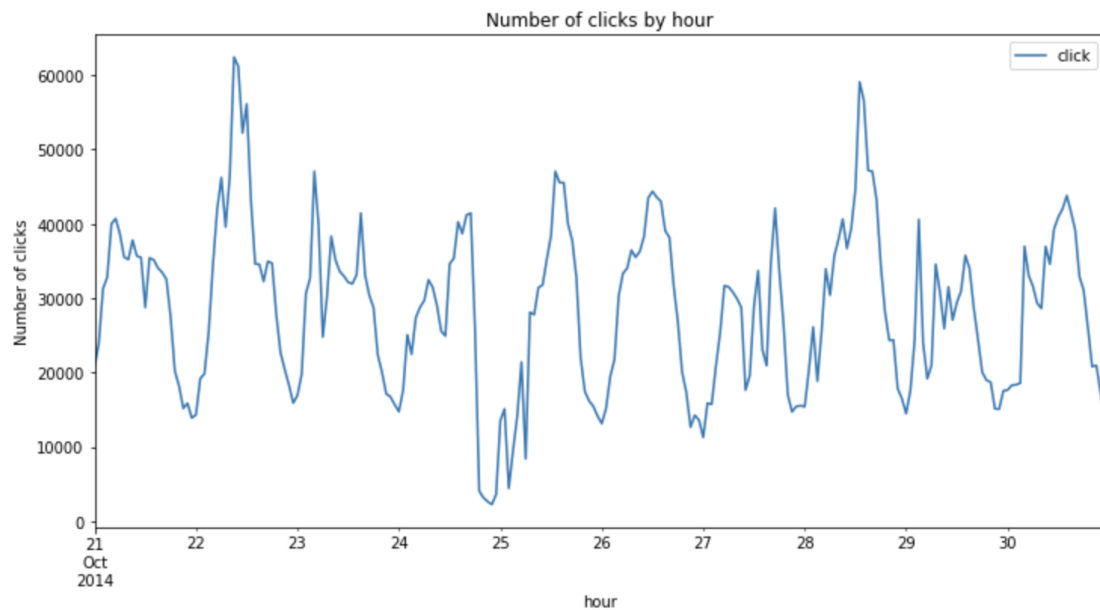
Distribution by Time

Hours

We have about 10 days of clickstream data

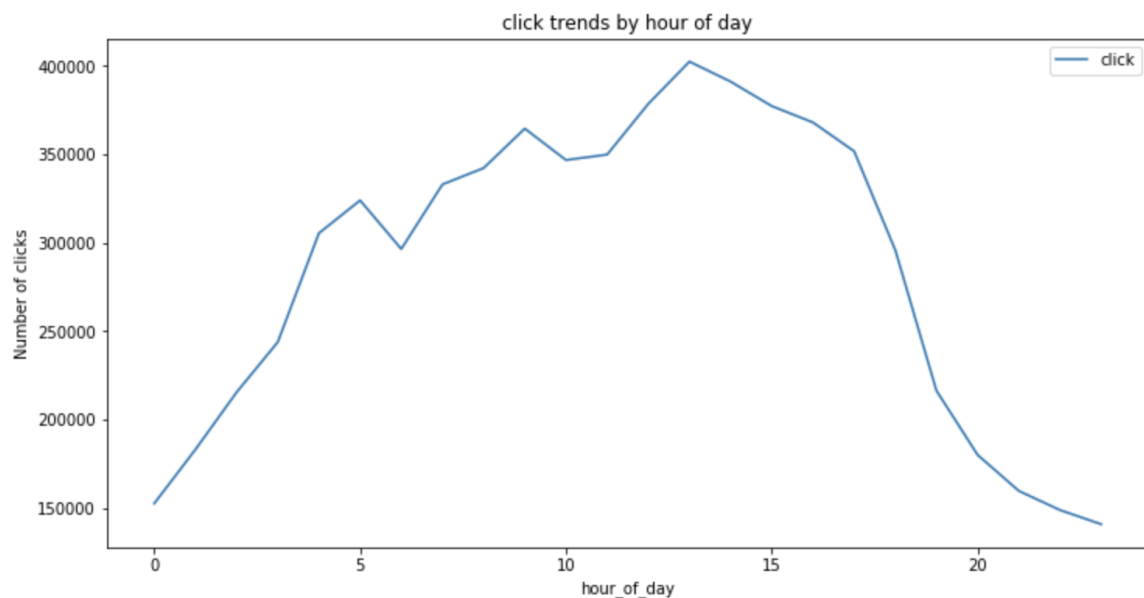
unique	240
top	2014-10-22 09:00:00
freq	447783
first	2014-10-21 00:00:00
last	2014-10-30 23:00:00

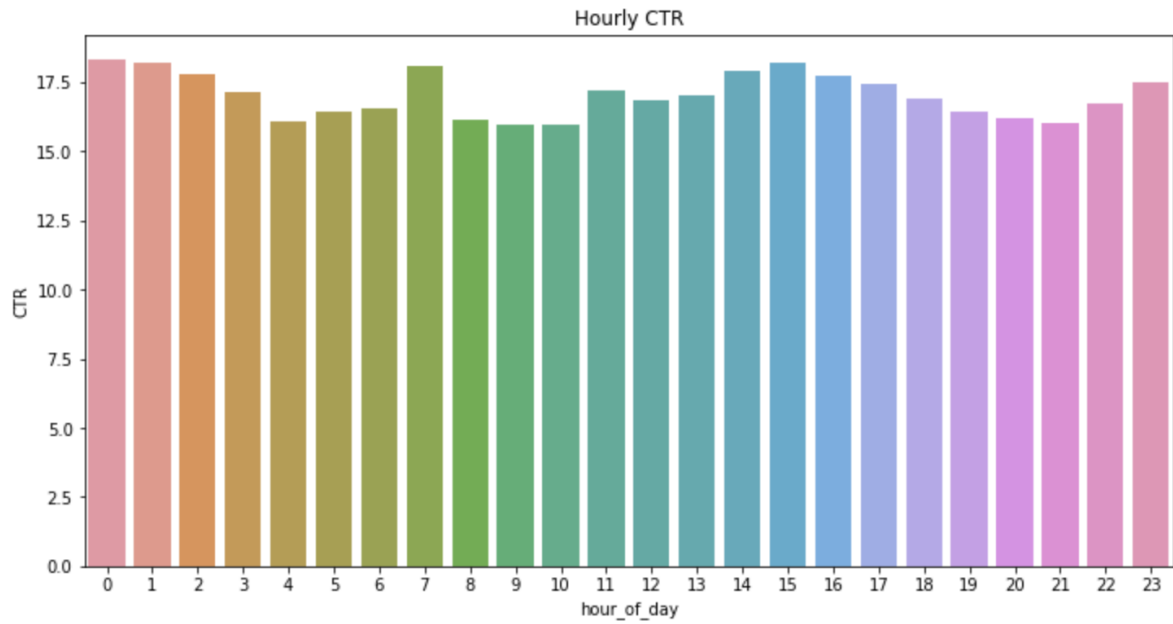
There seems to be strong cyclicity by **time of day** as well as **days of week**



Time of Day

Within a day, the number of impressions and clicks seem to peak in the middle of the day; however, CTR seems to fluctuate across different times of a day, with the change being mostly consistent

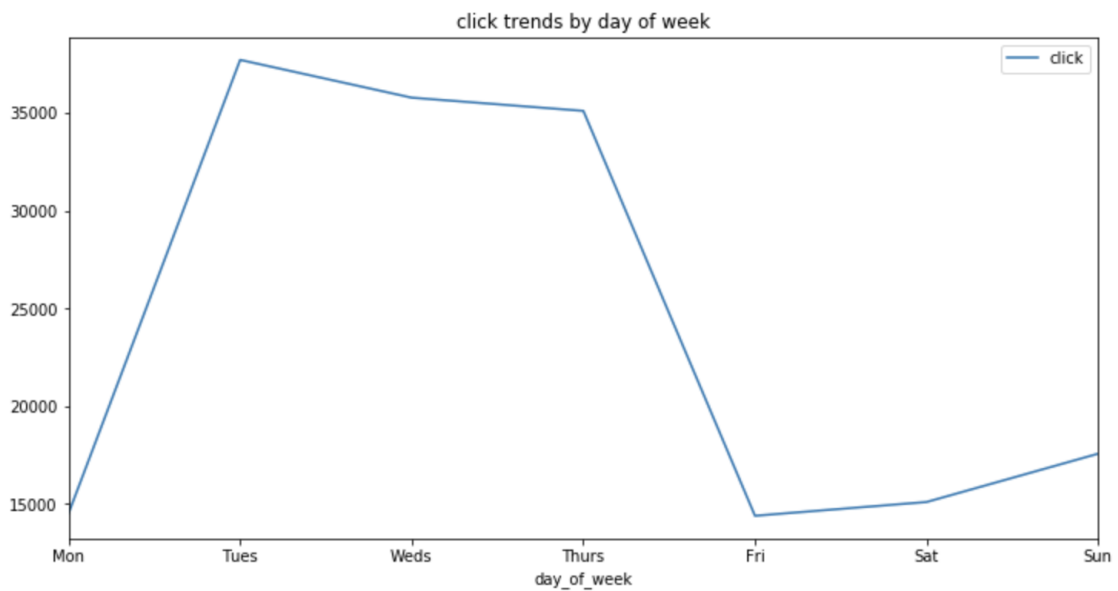


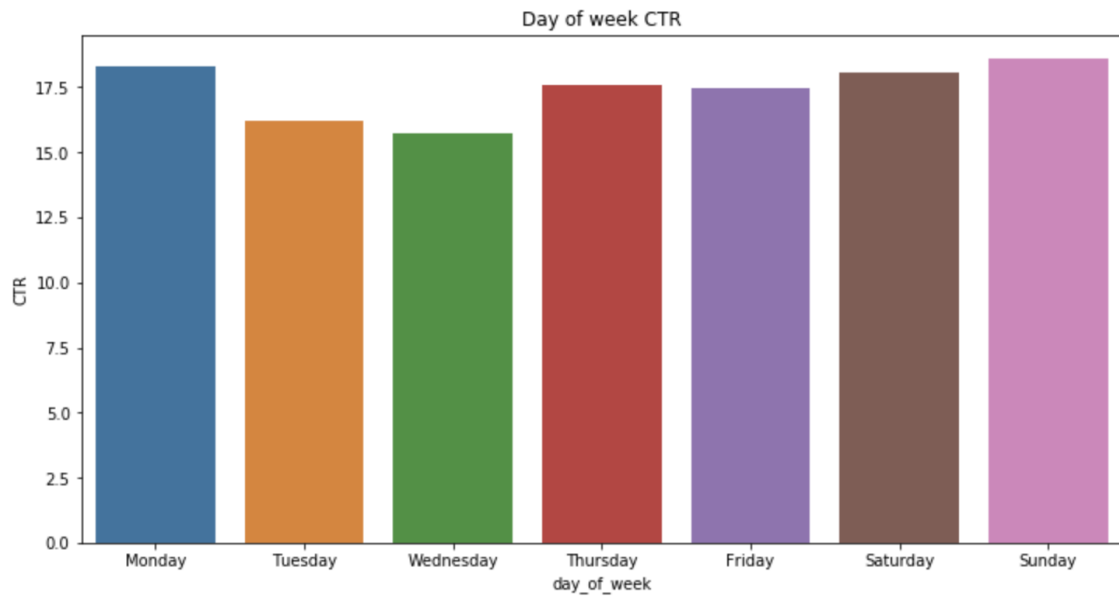


Day of Week

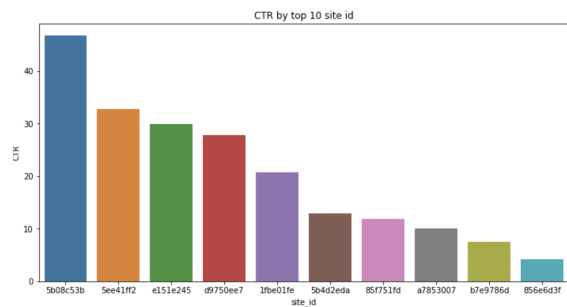
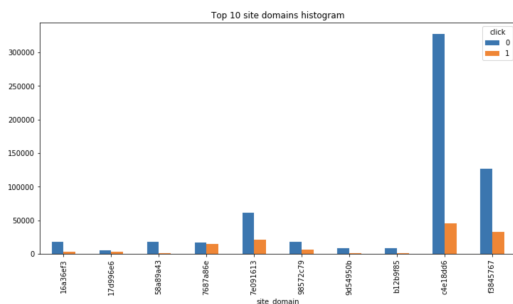
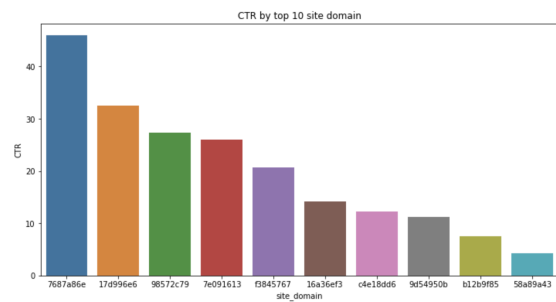
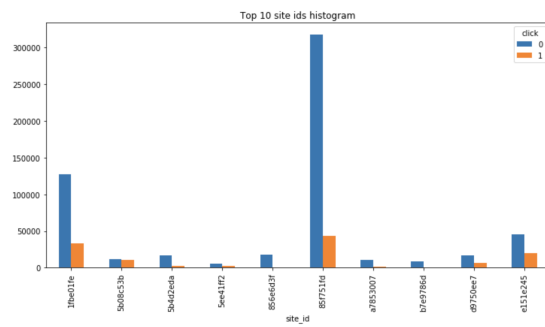
The number of clicks is obviously higher in Tues/Weds/Thurs and lower in Mon/Fri and weekends

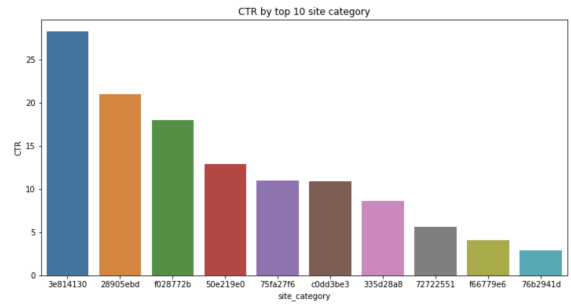
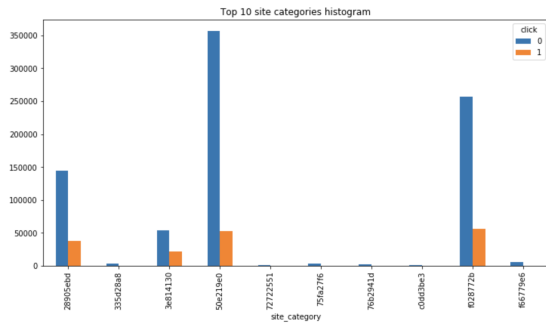
(would be great to get CTR data across days as well)



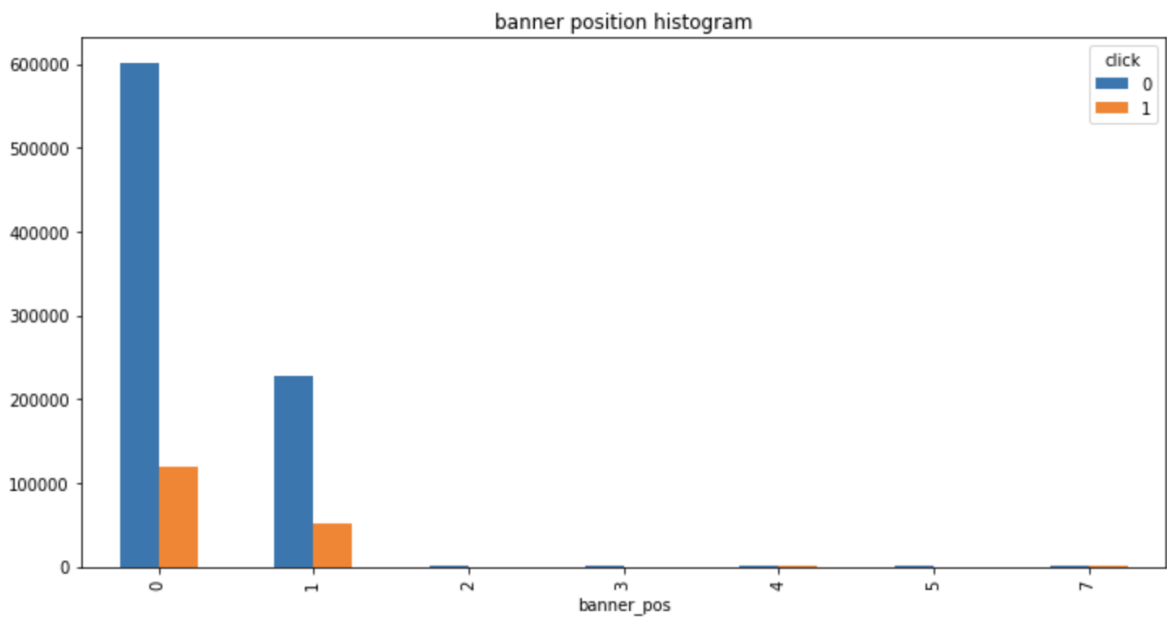


Distribution by SitesID

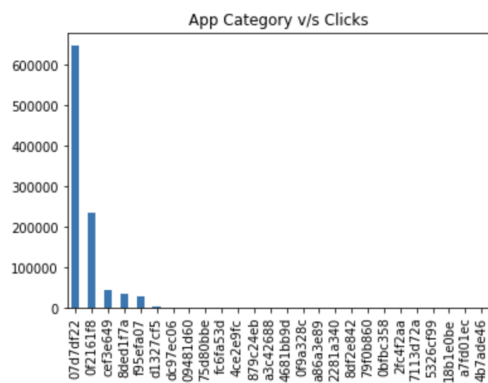


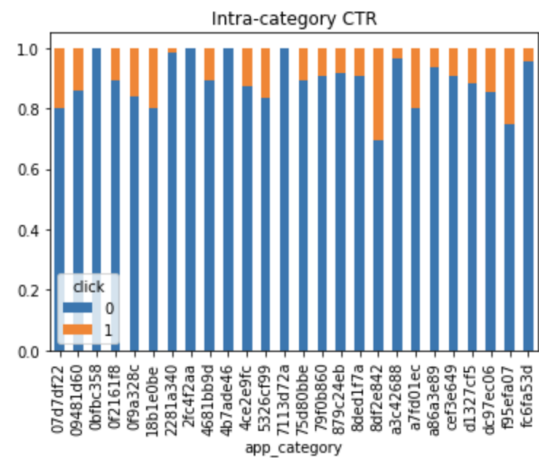


Distribution by Banner Position



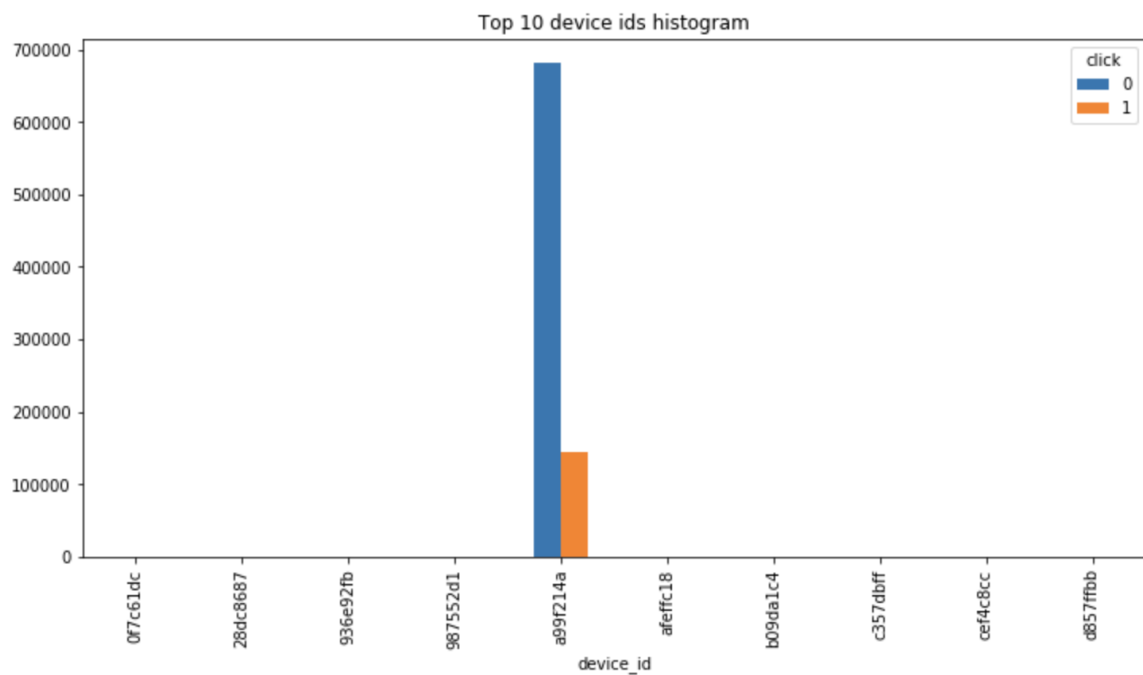
By App Categories



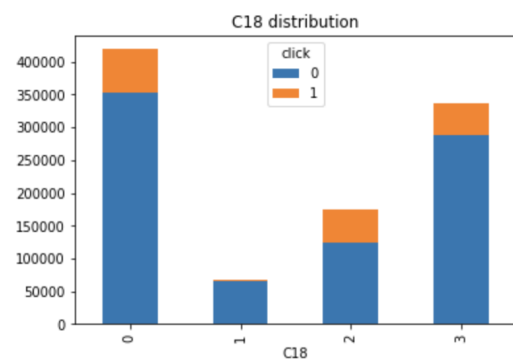
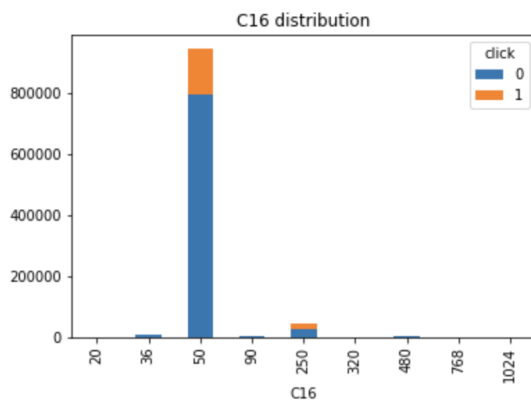
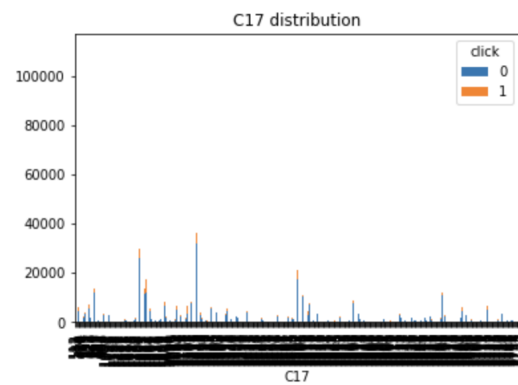
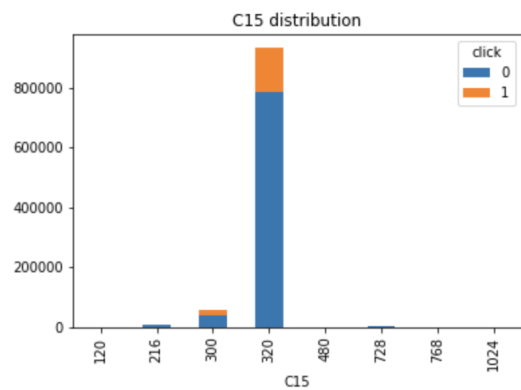
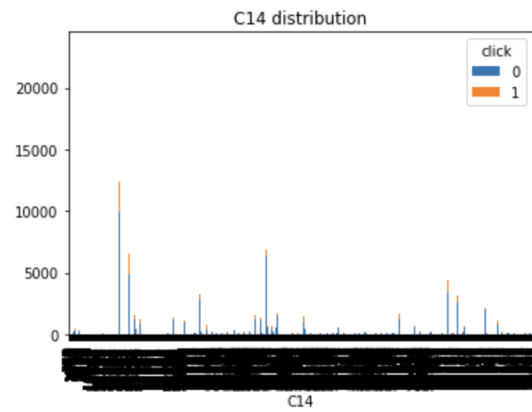
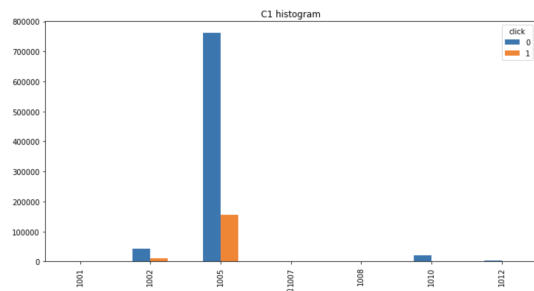


Other features

Device information is not as useful



Some of the unknown categorical data seem to be very skewed, not giving us too much information



Feature Engineering

Original Features

`click` - 0 or 1

`hour` - 14102108(originally a timestamp)

```
parse_date = lambda val : pd.datetime.strptime(val, '%y%m%d%H')
banner_pos
```

Site

site_id

site_domain

site_category

App

app_id

app_domain

app_category

Device

device_id

device_ip

device_model

device_type

device_conn_type

	id	click	hour	C1	banner_pos	site_id	site_domain	site_category	app_id	app_domain	app_category	device_id	device_ip	device_model	device_type	device_conn_type
0	10009635774586344851	0	2014-10-21	1005	0	543a539e	c7ca3108	3e814130	ecad2386	7801e8d9	07d7df22	a99f214a	37018b2d	24f6b932	1	0
1	10010452321736390000	1	2014-10-21	1005	0	1fbc01fe	f3845767	28905ebd	ecad2386	7801e8d9	07d7df22	a99f214a	cede6db1	a0f5f879	1	0
2	10014385711019128754	0	2014-10-21	1005	0	1fbc01fe	f3845767	28905ebd	ecad2386	7801e8d9	07d7df22	a99f214a	12c3d700	ef726eae	1	0
3	10015944270539844899	0	2014-10-21	1005	0	1fbc01fe	f3845767	28905ebd	ecad2386	7801e8d9	07d7df22	a99f214a	9cd175b0	ecb851b2	1	0
4	10016492574846482398	0	2014-10-21	1005	0	1fbc01fe	f3845767	28905ebd	ecad2386	7801e8d9	07d7df22	a99f214a	83b564b0	a5bce124	1	0

Time

hour → hour_of_day + day_of_week

Publisher

pub_id / pub_domain / pub_category

- By default, use site_id (domain, category)
- If site_id = 85f751fd, then use app_id
 - Did this because more than 90% of data appears in the same website or same app - use app_id and site_id to better characterize id into more evenly-distributed categories

▼ Code

```

if is_app(row):
    new_row['pub_id'] = row['app_id']
    new_row['pub_domain'] = row['app_domain']
    new_row['pub_category'] = row['app_category']
    writer_app.writerow(new_row)
else:
    new_row['pub_id'] = row['site_id']
    new_row['pub_domain'] = row['site_domain']
    new_row['pub_category'] = row['site_category']
    writer_site.writerow(new_row)

```

Device

`device_id_count` - the number of unique device id - corresponding to `device_id`

`device_ip_count` - the number of unique device ip - corresponding to `device_ip`

User

`user_count` : `device_id` + `device_ip` + `device_model`

- Definition: the number of unique user sessions

`smooth_user_hour_count` : `clicks` + `hour_of_day` + `user_id`

- Definition: number of impressions (the total of 0 and 1) for one user within a specific hour

`user_click_history` :

- Definition: a string that summarizes the latest four user actions in the previous hour

▼ Code

```

user, hour = def_user(row), row['hour']
new_row['user_count'] = user_cnt[user]
new_row['smooth_user_hour_count'] = str(user_hour_cnt[user+'-'+hour])

```

Hash Transformation

- Transform all non-int data into Hash int

Model and Training

Additional Visualization

- Old - remaining ones @Alan Zhang
- New - further exploration @Changjin Liu
- Model Selection - new features for presentation
 - Logistics Regression - baseline model? @Alan Zhang
 - Random Forest @Chloe Kim
 - Adaboosting @Jo Kim @Yunshun Zhong
 - LGBM (Light GBM) @Changjin Liu
- Evaluation metrics
 - Confusion Matrix (recall, precision, accuracy)
 - AUC/ROC

Future Work

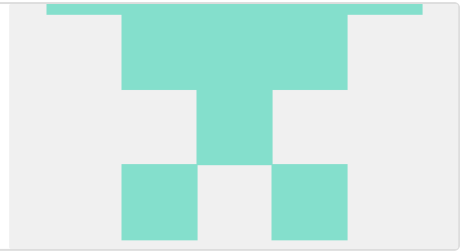
- Based on data analysis
 - Get rid of abnormal data in the higher range (depending on objectives)
 - Feature selection based on data analysis
 - Further feature engineering based on data analysis - User_click_history - merge data patterns (0010, 0100, 1000...)
- C-number → Hash values

References

ycjuan/kaggle-avazu


You can't perform that action at this time. You signed in with another tab or window. You signed out in another tab or window. Reload to refresh your session. Reload to refresh your session.

 <https://github.com/ycjuan/kaggle-avazu/tree/master/bag>



Code


In last posting, we have figured out how to import files from local hard drive. In this posting, I will delineate how to import files directly from Google Drive. As you know Colab is based on Google Drive, so it is convenient to import files from Google Drive once you know the drills.

 <https://buomsoo-kim.github.io/colab/2018/04/16/Importing-files-from-Google-Drive-in-Google-Colab.md/>

Google Drive

Meet Google Drive - One place for all your files

Google Drive is a free way to keep your files backed up and easy to reach from any phone, tablet, or computer. Start with 15GB of Google storage - free.


 <https://drive.google.com/drive/folders/1zvpKblOKNj9BfSSL9mmc-kDgQ3JN0EXm>



Project Proposal

Google Docs - create and edit documents online, for free.

Create a new document and edit with others at the same time -- from your computer, phone or tablet. Get stuff done with or without an internet connection. Use Docs to edit Word files. Free from

 <https://docs.google.com/document/u/1/d/13JOliB66SEl6cr2Q7MPECTq5BUNJIKrL5UdgwcbIqGc/edit>

