Click-Through-Rate Prediction Capstone

By: Amanda Ahn

Springboard

CTR Prediction

Online advertising revenue equates to more than hundreds of billions of dollars for companies. As society continues to modernize, this source of revenue will only increase.

Accurately predicting CTR can help increase total company profits by modifying select features of advertisements.

The Problem

An average of

60-80%

of company revenues in the U.S. comes from advertisements and ad placements play a major role in increase CTR.

The goal of this project is to identify the significant features of advertisements to improve in order to increase the CTR by about 20% through the use of machine learning techniques.

Data

Data Source: Kaggle, Avazu CTR Prediction Dataset

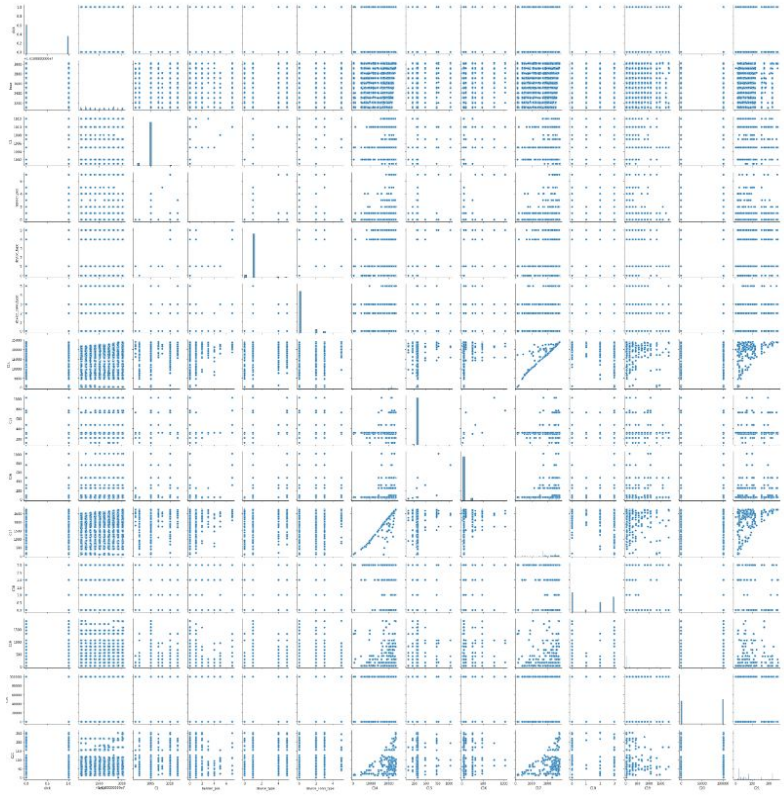
Number of records: 40,428,967 rows and 24 columns

Column names: id, click, hour, C1, banner_pos, site_id, site_domain, site_category, app_id, app_domain, app_category, deivce_id, device_model, device_type, device_conn_type, C14, C15, C16, C17, C18, C19, C20, C21.

Data Types: floats, integers, and objects/strings

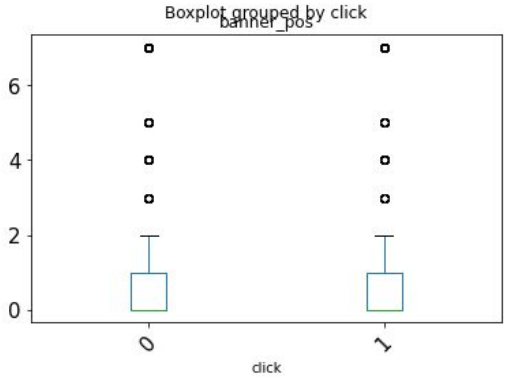
Cleaning: No null or duplicate values

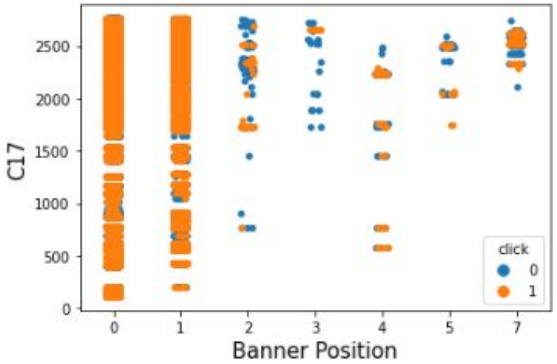
Exploratory Data Analysis

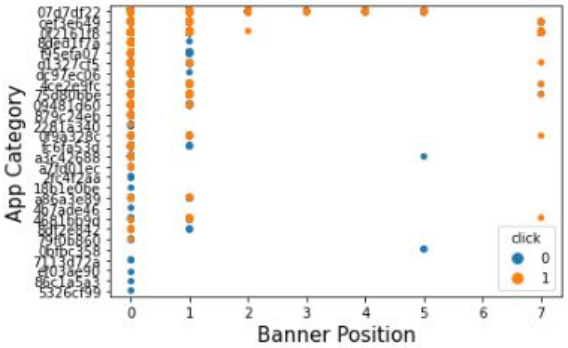


Seaborn pairplot for visualization of each feature with each other

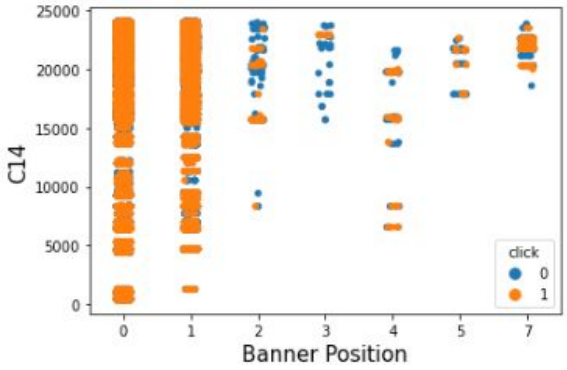
Feature C14 and C17 have positive correlation with each other

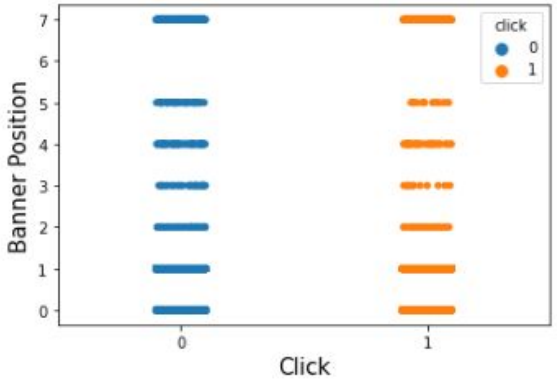






Boxplot of click vs banner_pos columns (self hypothesized important features)





Specific comparison stripplots of suspected important individual features deduced from pairplot

Modeling

Logistic Regression

	precision	recall	f1-score	support
0	0.58	0.64	0.61	67128
1	0.60	0.53	0.56	67128
accuracy			0.59	134256
macro avg	0.59	0.59	0.59	134256
weighted avg	0.59	0.59	0.59	134256

Accuracy Score: 0.588115242521749 5

[0.56904185 0.57170646 0.56874493 0.56117495 0.56779267]
Mean cross validation test score: 0.5676921711582379
Mean cross validation train score: 0.5684506178500486
Standard deviation in cv test scores: 0.0035079554870482425

3

Random Forest

	precision	recall	f1-score	support
0	0.68	0.63	0.65	67128
1	0.66	0.71	0.68	67128
accuracy			0.67	134256
macro avg	0.67	0.67	0.67	134256
weighted avg	0.67	0.67	0.67	134256

Accuracy Score: 0.6680073888690263

[0.63983914 0.63203599 0.64046306 0.63785728 0.64052348]
Mean cross validation test score: 0.6381437886637685
Mean cross validation train score: 0.6528238289912107
Standard deviation in cv test scores: 0.0032032318344256297

1

Gradient Boosting

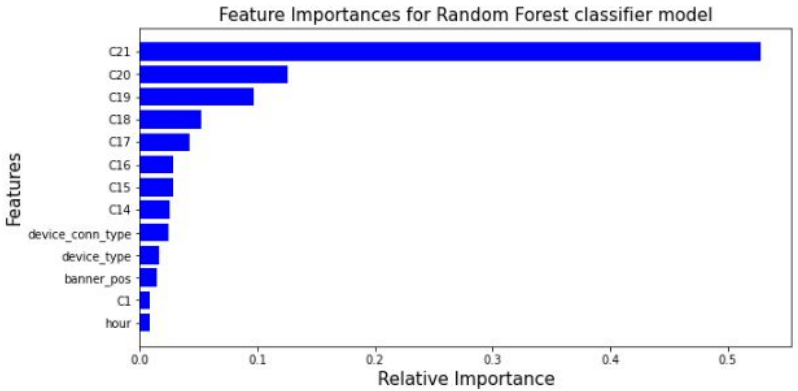
	precision	recall	f1-score	support
0	0.64	0.57	0.60	67128
1	0.61	0.67	0.64	67128
accuracy			0.62	134256
macro avg	0.62	0.62	0.62	134256
weighted avg	0.62	0.62	0.62	134256

0.6224228339887975

[0.67245329 0.66695963 0.67128813 0.67956636 0.67452751]
Mean cross validation test score: 0.6729589833945869
Mean cross validation train score: 0.6681433491516522
Standard deviation in cv test scores: 0.004126920509444569

2

Feature Importance



	Features	Importance scores
0	hour	0.008086
1	C1	0.008771
2	banner_pos	0.014281
3	device_type	0.016927
4	device_conn_type	0.024372
5	C14	0.025311
6	C15	0.028052
7	C16	0.028483
8	C17	0.042065
9	C18	0.052380
10	C19	0.097249
11	C20	0.125926
12	C21	0.528097

Feature C21

Conclusion

Random Forest Model with GridSearchCV:

Accuracy score: 0.67

ROC-AUC score: 0.72

When trying to improve click through rate for advertisements, companies should focus on what feature C21 is.

Stakeholders: Marketing, Advertising, Design Directors