CLICK-THROUGH RATE PREDICTION CAPSTONE

By Amanda Ahn

01

CONTEXT

Online advertising revenue equates to more than hundreds of billions of dollars for companies. An average of 60-80% of companies' revenue in the U.S. comes from advertisements.

02

MOTIVE

As a society that continues to modernize and digitalize, online advertisements, as source of revenue, will only increase. Accurately predicting CTR can help increase total company profits by modifying select features of advertisements to increase click rate, which increases likelihood of customers buying the product.

03

PROBLEM STATEMENT

The goal of this project is to identify the significant features of advertisements to improve in order to boost CTR by increasing the base model by at least 5% through the use of machine learning techniques.

DATA

Source: Kaggle, Avazu CTR Prediction Dataset

NUMBER OF RECORDS:

40,428,967 rows & 24 columns

DATA TYPES:

- Floats
- Integers
- objects/strings

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 CLEANING:

No null or duplicate values

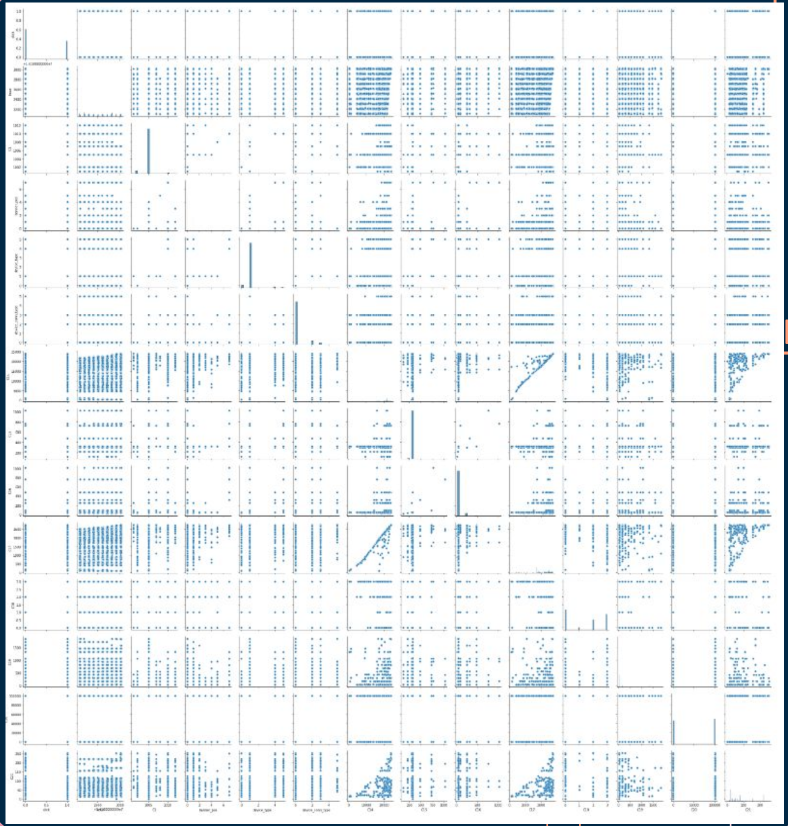
EXPLORATORY DATA ANALYSIS

Seaborn pairplot for visualization of each feature with each other

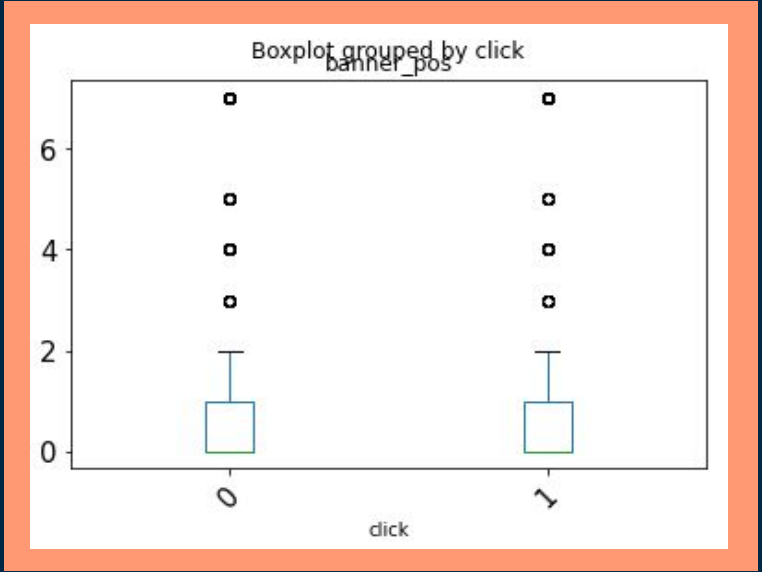
↓

Feature C14 and C17 have positive correlation with each other

- Keep these features in mind



FEATURE EXPLORATION



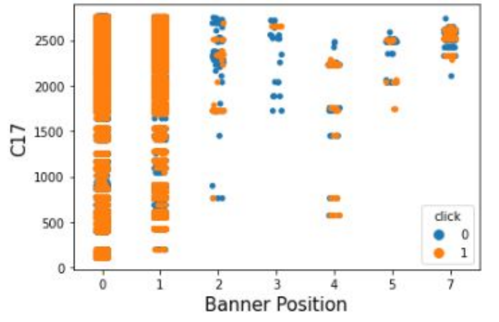
Boxplot of click vs banner_pos columns

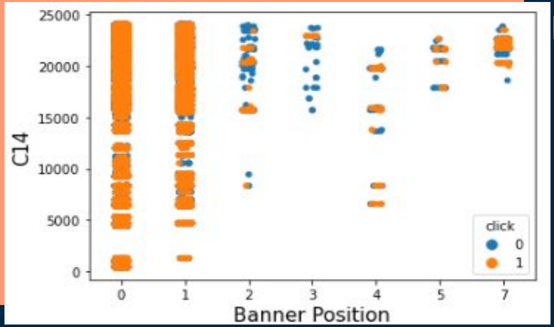
- Self hypothesized important features
- Background in UI/UX → advertisement position on app/site is very important
- No deducible conclusion from boxplot

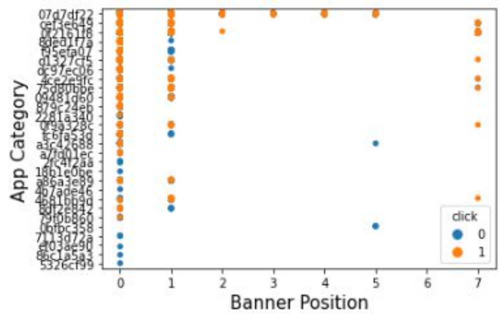
STRIP-PLOTS

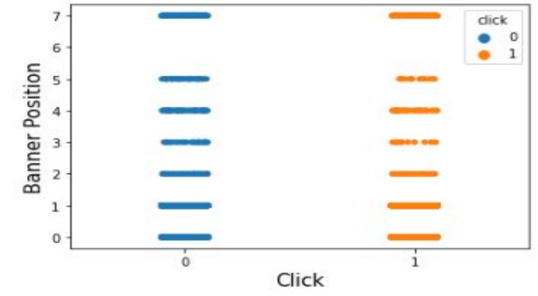
Specific comparison plots of suspected important individual features

No deducible conclusions again → move to modeling









MODELING PROCESS

FIRST MODEL

Choose first model:
Logistic Regression

most basic model to assess how it performs

2ND & 3RD MODELS

2nd Model: Random Forest
need greater accuracy than LR Model

3rd Model: Gradient Boosting
since RF Model > LR Model, use Gradient instead of Decision Tree

MODIFY DATA & REFIT MODELS

Use RandomOverSampler to balance data → had way more values for non-click data group than click data group

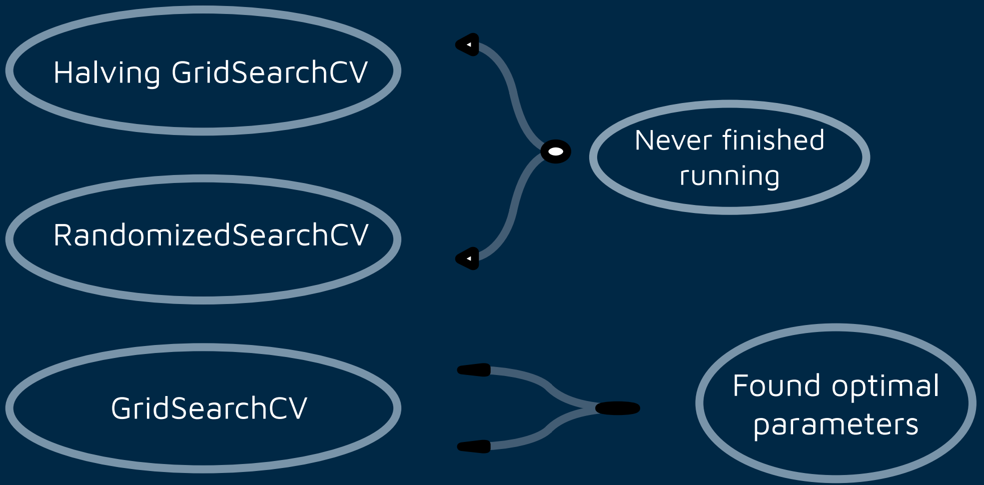
Refit all models to new datasets and determine best one

PARAMETRIC TUNING

Best model: Random Forest

Hyperparameter tuning using GridSearchCV to find optimal parameters

HYPERPARAMETER TUNING




```
rf = RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
            max_depth=None, max_features='auto', max_leaf_nodes=None,
            min_impurity_decrease=0.0,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, n_estimators=600, n_jobs=-1,
            verbose=1, random_state=None, warm_start=False, oob_score=False)
rf.fit(X_new_train, y_new_train)
y_new_pred = rf.predict(X_new_test)
acc = accuracy_score(y_new_test, y_new_pred)
cv = cross_val_score(rf, X_new_test, y_new_test,scoring='roc_auc').mean()
print(acc)
print(cv)
```

BEST MODEL

3

	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: 0.5881152425217495

LOGISTIC REGRESSION

1

	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

RANDOM FOREST

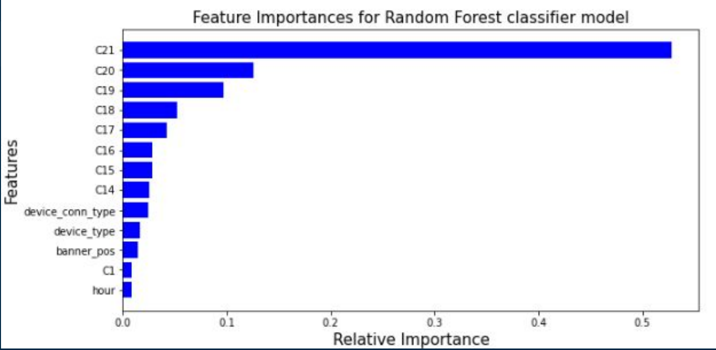
2

	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

GRADIENT BOOSTING

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

Best Model: Random Forest Model with balanced data and best hyperparameters using GridSearchCV.

Accuracy score: 0.67

ROC-AUC score: 0.72

Stakeholders: Marketing, advertising, design directors

Unfortunate that feature C21 is hidden by Avazu, but nonetheless, when trying to improve click through rate for advertisements, companies should focus on feature C21

Do you have any questions?

amandaahnn@gmail.com
925 989 4599

THANKS