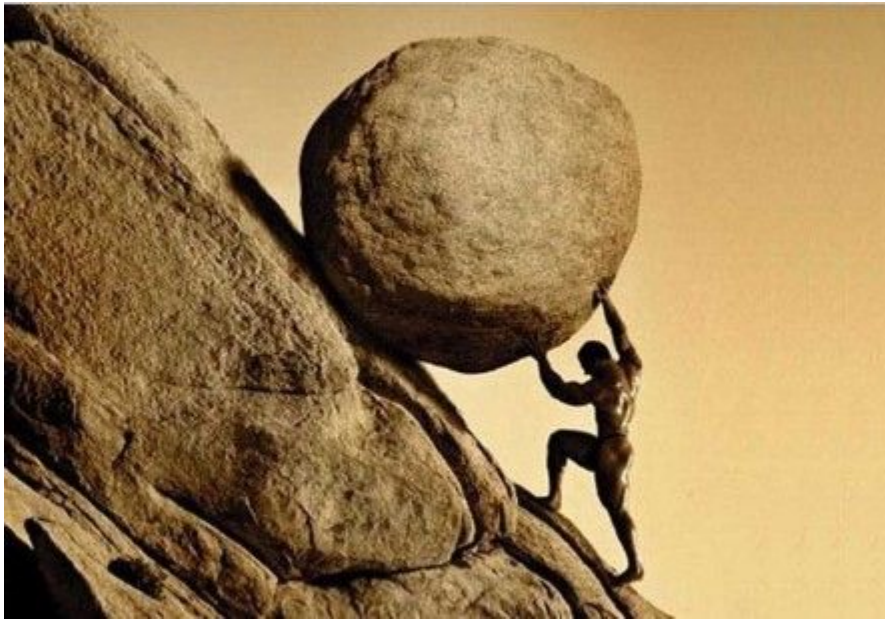
TalkingData Kaggle Challenge

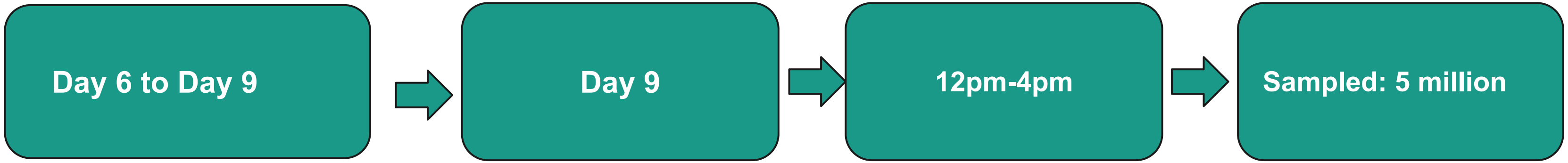
Analysis and lessons learned

Elyse Kadokura, Jenifer De Figueiredo, Keyuri Raodeo

Analysis

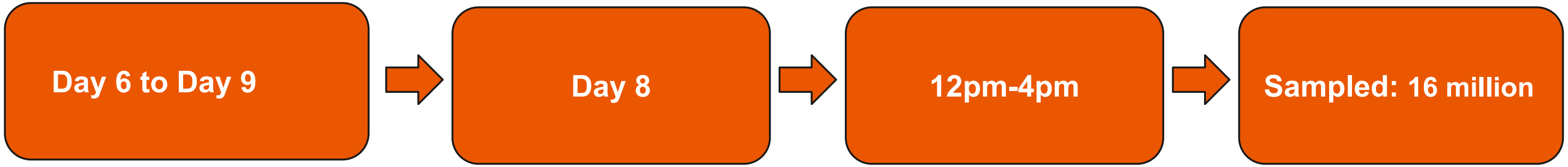


Creating test dataset



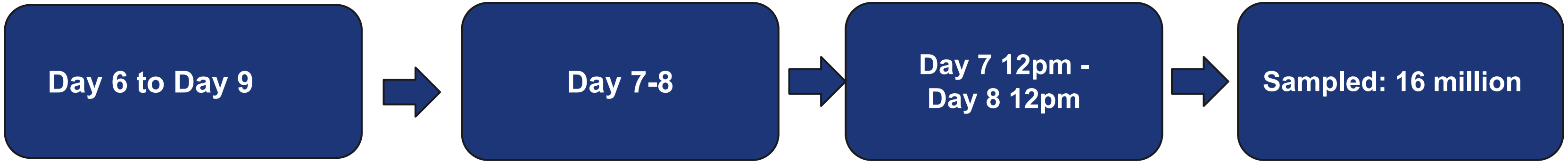
Day 6 to Day 9 → Day 9 → 12pm-4pm → Sampled: 5 million

Creating val dataset



Day 6 to Day 9 → Day 8 → 12pm-4pm → Sampled: 16 million

Creating train dataset



Day 6 to Day 9 → Day 7-8 → Day 7 12pm - Day 8 12pm → Sampled: 16 million

Features:

Grouped by date-hour, ip:

- Number of unique apps
- Number of unique devices
- Number of unique channels
- Number of unique OS
- Number of clicks

Attribution rate for OS, device, channel, and app

LightGBM

	Actual +	Actual -	Total
Predicted +	10,227	137,026	147,253
Predicted -	1,192	4,851,555	4,852,747
Total	11,419	4,988,581	5,000,000

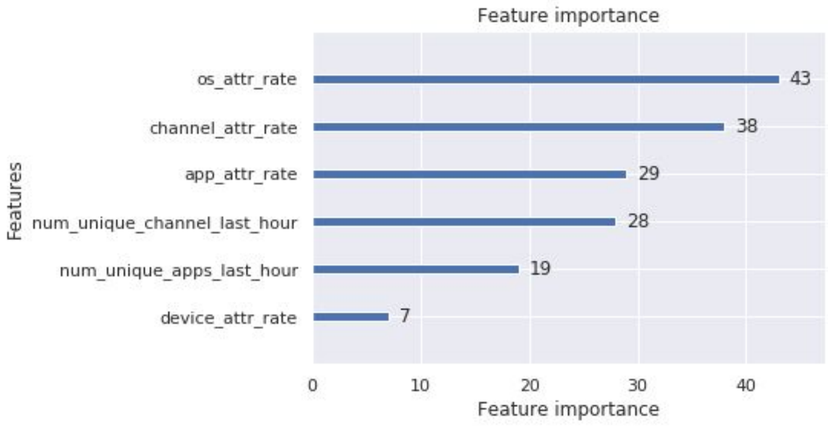
Recall (TP/TP+FN): 0.90

Precision (TP/TP+FP): 0.07

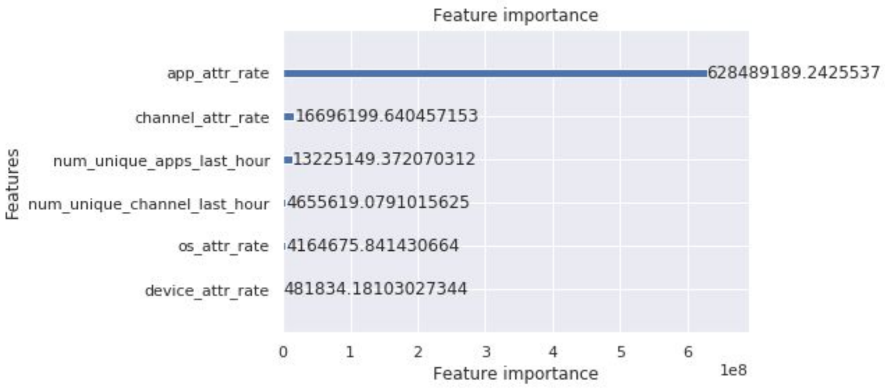
F1-score (2 * (P*R/P+R)): 0.13

Feature importance

Split:



Gain:

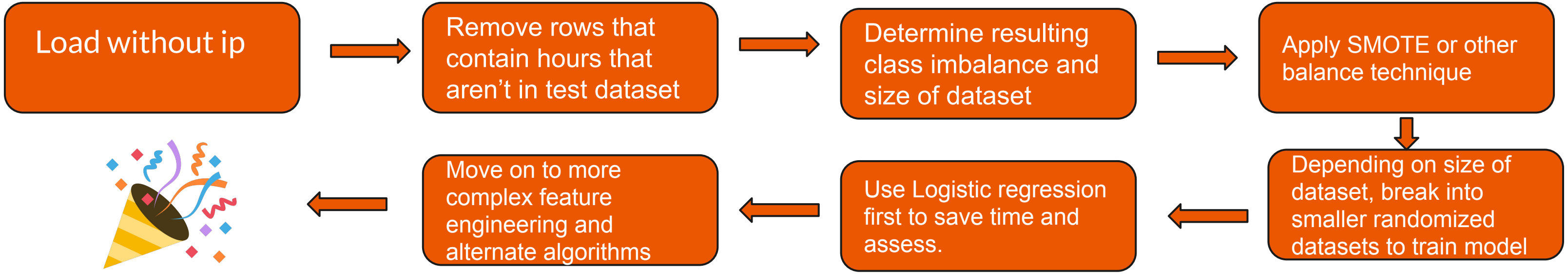


To do:

1. More feature engineering
2. Hyperparameter tuning
3. Explore other models
4. Explore cloud computing

Jeny’s Takeaways...

1. Always look at how big the training and test datasets are before deciding on a Kaggle project!
    a. Dask
    b. Hvplot
    c. Holo Views
2. Azure ML studio has space limitations, even if they don’t think they do, but attentive customer service...
3. Get a good grasp on test dataset and features before moving forward with model



Load without ip

Remove rows that contain hours that aren’t in test dataset

Determine resulting class imbalance and size of dataset

Apply SMOTE or other balance technique

Depending on size of dataset, break into smaller randomized datasets to train model

Use Logistic regression first to save time and assess.

Move on to more complex feature engineering and alternate algorithms

Jeny’s Takeaways cont’d…

4. The Confusion Matrix Enlightenment

Meteorological Contingency Table

	Observed Yes	Observed No
Forecast Yes	True Positive = “Hit” 	False Positive = “False Alarm”
Forecast No	True Negative = “Non-Severe-Weather Event” 	False Negative = “Miss” Let’s not go there…

Data Science Confusion Matrix

	Predicted Yes	Predicted No
Actual Yes	True Positive = “Hit” 	False Negative = “Miss” Let’s not go there…
Actual No	False Positive = “False Alarm” 	True Negative = “Non-Severe-Weather Event”