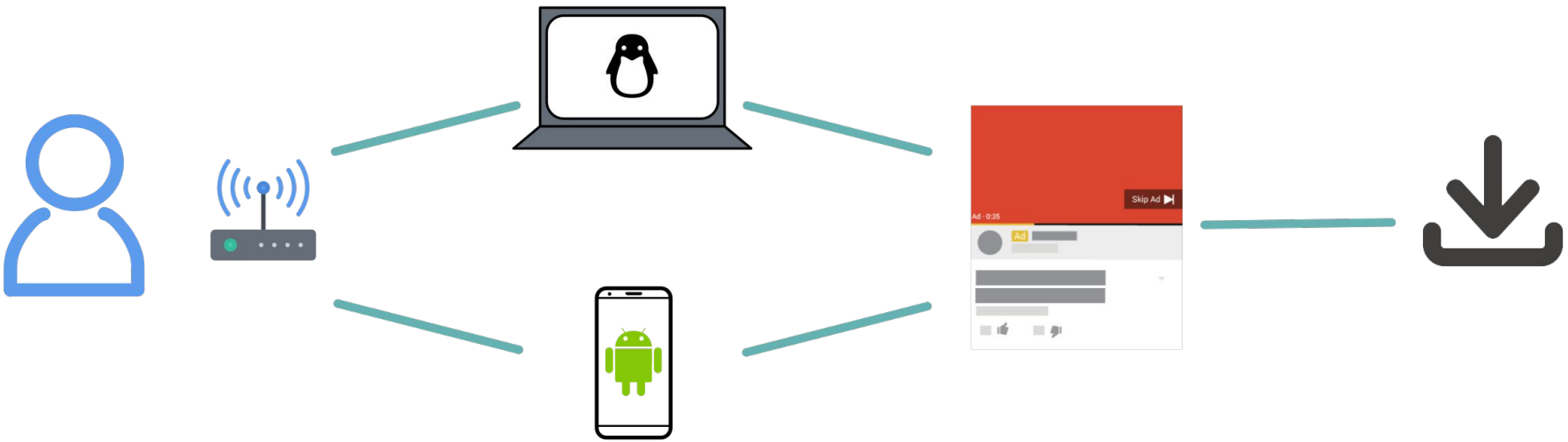
Talking Data Kaggle competition

Sowmya Vasan
Daisy Du
Erika Pelaez
Mentor: Nick Janetos

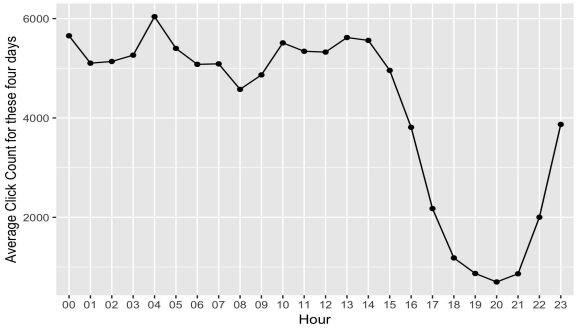
Click journey

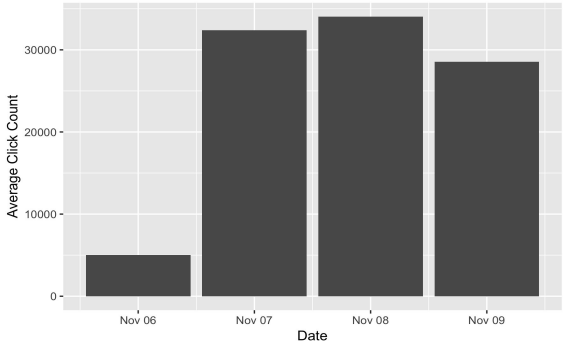


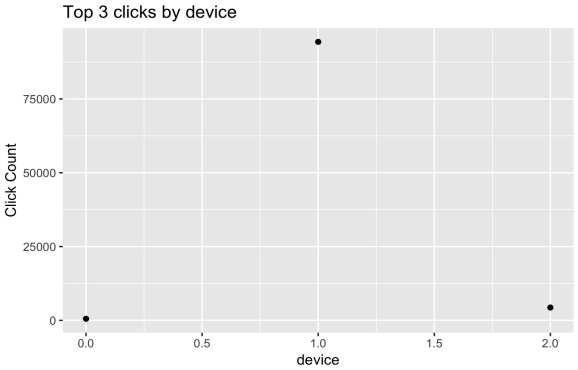
Explanatory Analysis

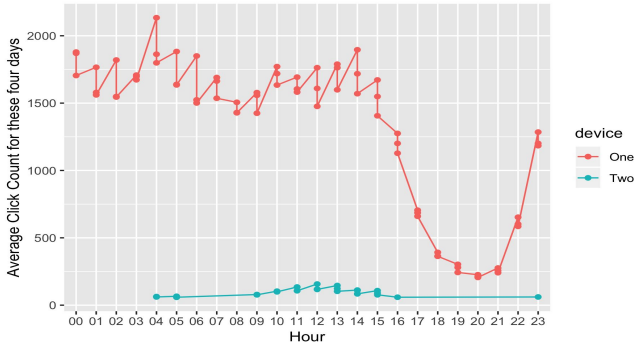
Feature	Unique Count
IP address	277396
Device	3475
OS	800
App type	706
Channel	202

Anatomy of the dataset









Feature Engineering

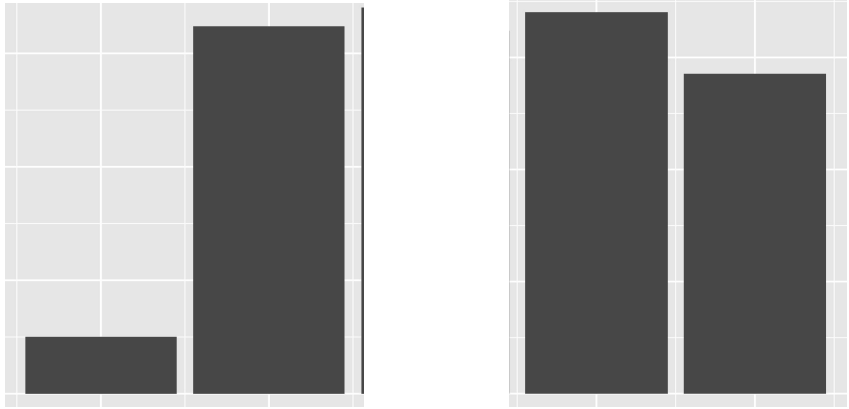
IP is an important feature but is highly dynamic. We decided to create various frequency based features on IP, to retain the importance of IP without actually having the feature, namely:

- Number of clicks every hour by an IP
- Number of clicks every hour by an IP-App combination
- Number of clicks every hour by an IP-OS combination
- Number of clicks every hour by an IP-Device combination
- Number of clicks every hour by an IP-Channel combination

Another feature we thought as interesting was to see if time from the previous click of the same click journey had any relation with app downloads hence we incorporated the following feature:

- Time from the previous click for every unique IP-App-OS-Device combination

Sampling



All Positives +

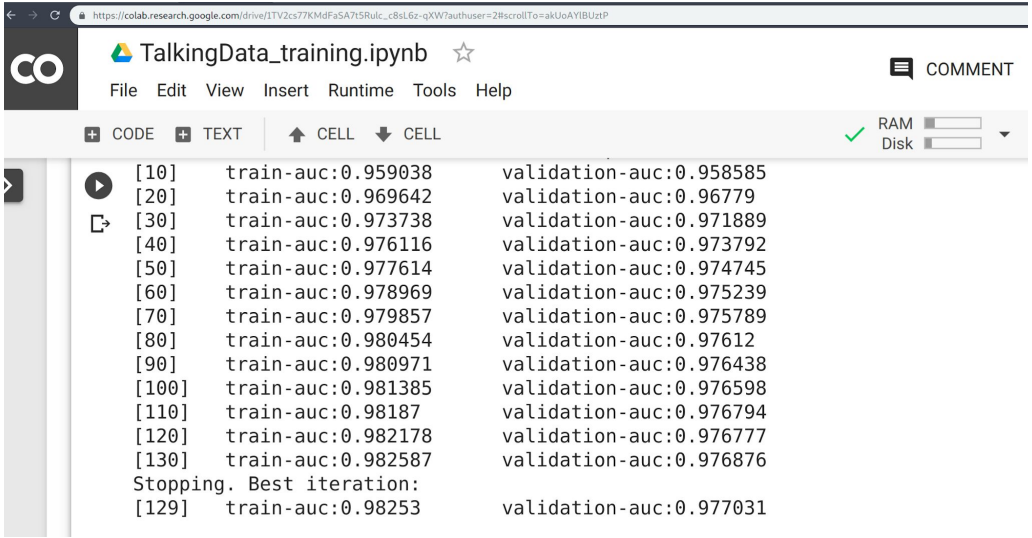
Training
20%

Validation
20%

Model selection and performance

Xgboost

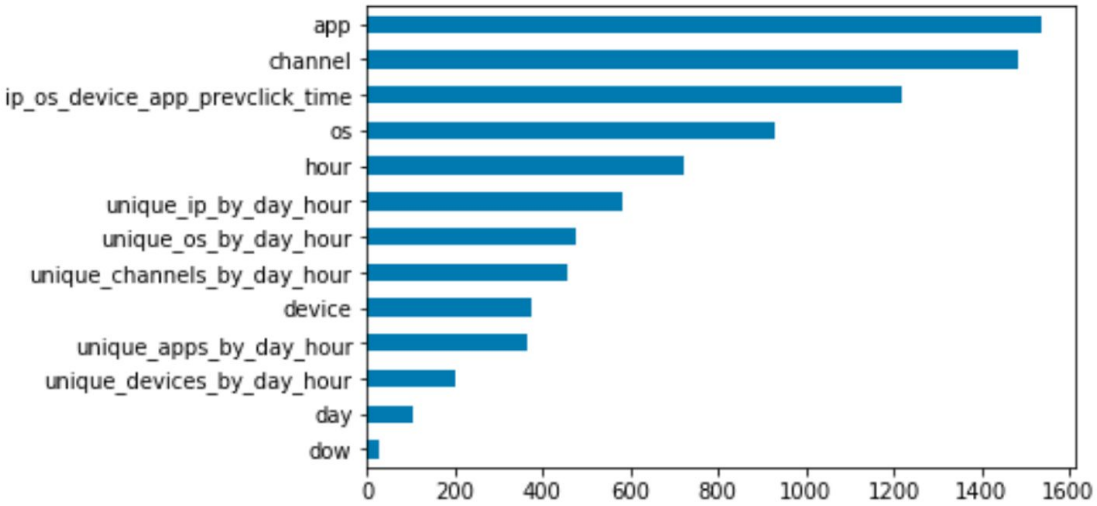
- Learning Rate:0.1
- Max Depth : 6
- 1000 estimators
- Early stopping rounds: 10
- Subsample : 80% (0.8)



Final results

Kaggle AUC Score:

0.82768



Following steps

- Hyper parameter tuning
- More effective sample selection
- Different evaluation metrics (F1 score, suggestions?)
- Rolling frequency features which is applicable in a production setting

Things Learnt

- Effective feature engineering when dealing with big data and limited memory
- Efficient way to handle big data with Dask
- Creative ways of handling sample selection with limited memory
- Modeling with large scale imbalanced dataset

Questions?

Thank you!!