Advanced ML Final Project

By Keep Learning

The problem or challenge

Predict if a user is likely to make a purchase in the next 7 and 14 days?

First problems Big Data

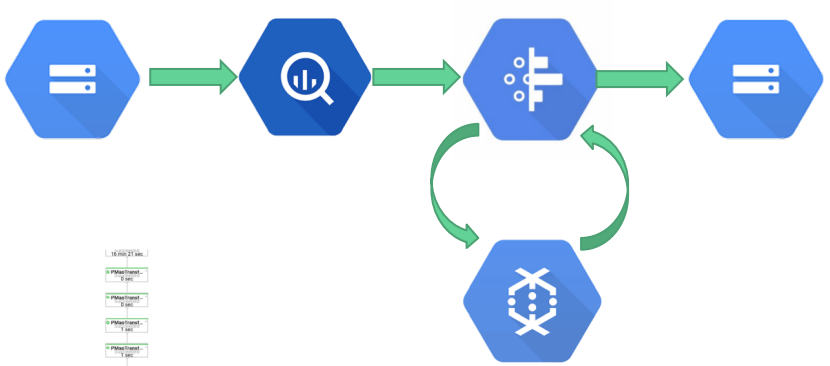
> 27 GB of data

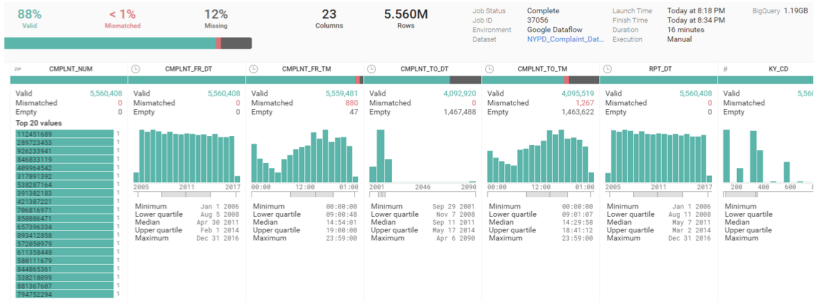
- Read Attributes

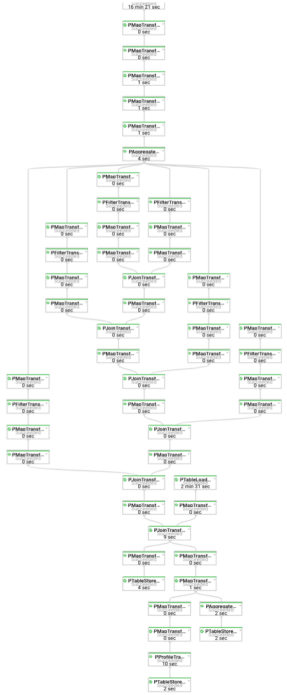
```
$ sed s/\\\\\"/\'/g attributes.csv > attribute_reformated.csv
```

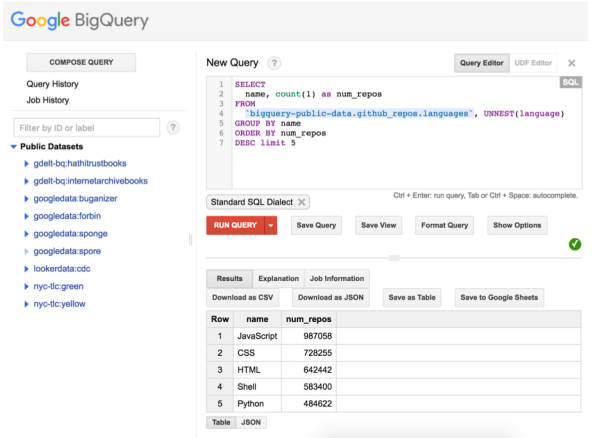
What we tried

- Working on the cloud
- Processing Pipeline
  - Fast & Cheap









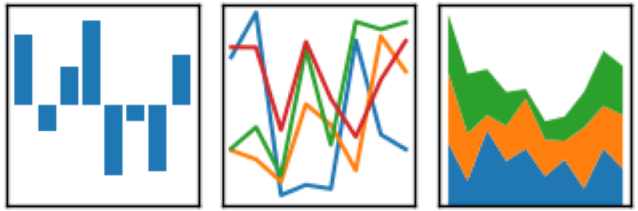
What we used

- Even faster and cheaper:
  - Local
  - Read specific columns

Local

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



Features

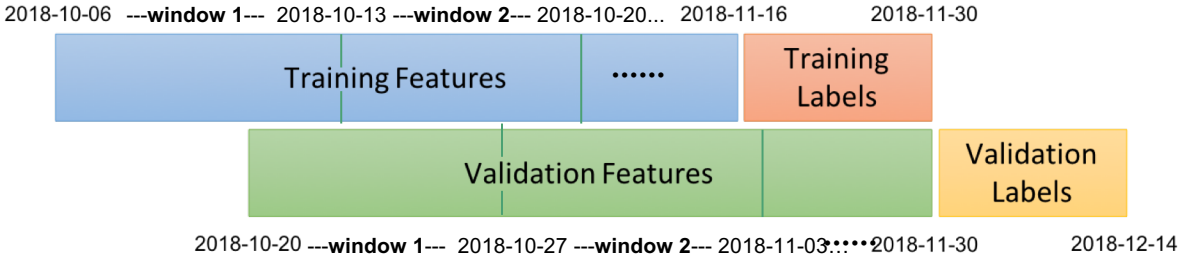
- EDA
- Feature Engineering

EDA

- Sessions
  - **Drop some columns that we thought trivial. For example, the column 'is_mau' and column 'is_wau', only False value**
- Events
  - **Only 5% of the record in Events dataset are purchase event**
- Attribute
  - One column has less than 100 unique values with integer data type (churn scores?)
  - One attribute_value has many 0, wide range of value, and format of xxx.xx(money-related?)
  - Column with float data type, range from 0 to 100 (LTV?)

Feature Engineering

- Training/ Validation Data Split



  - Time-related, money-related, purchase-related(event equals to '8'), and users' score-related columns
- Kaggle Submission Preparation
  - Some of the user ids from Sample_submission file don't exist in the datasets
    i. Add users to the training data
    ii. Impute all the columns with zero

Experimental Results

Each data scientist uses different methods of experimentation

What algorithms did you use in your experiment?

- Random Forest
- Xgboost
- LightGBM

Experimental results

	Validation Score	LB Score
Random Forest	0.947	**0.9433**
xgboost	0.953	**0.9502**
LightGBM	0.952	0.9489

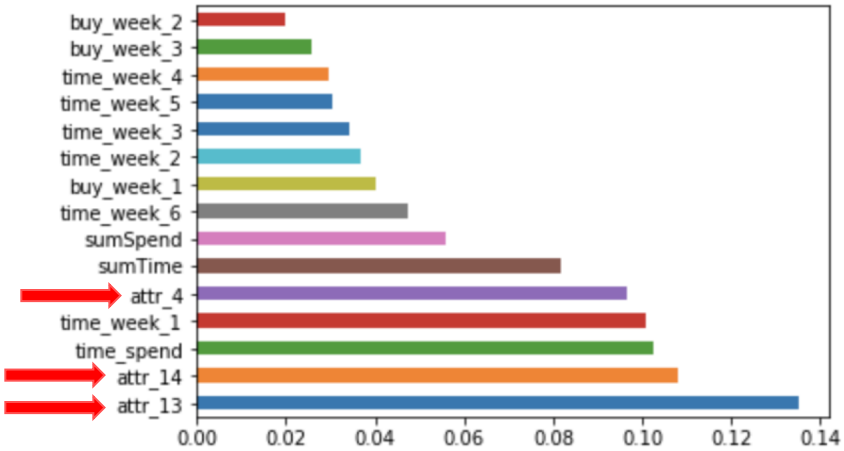
Adding attribute features

	Validation Score	LB Score
xgboost	0.9946	**0.9906**
LightGBM	0.9946	0.9902
Stacking	0.9742	**0.9704**

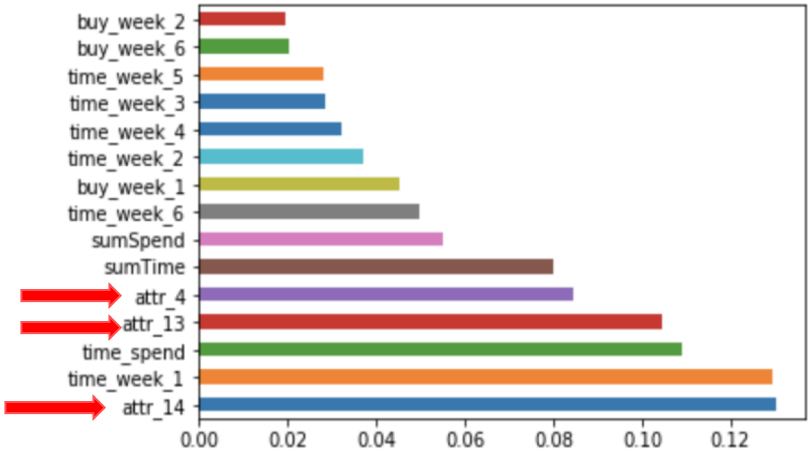
Feature Importance

- Why we improved score in the LB?

7 days xgboost model



14 days xgboost model



Running time for 7 days and 14 days model

	Validation Time for Hyper-parameters	Final Model Training Time	LB Score
Random Forest (No attr)	120 mins	5 mins	0.9433
xgboost	90 mins	2.8 mins	0.9906
LightGBM	4 mins	40 seconds	0.9902
Stacking	120 mins	40 mins	0.9742

On my MacPro 2017, 16GB RAM, 8 Cores

Lessons Learned

Have a right validation strategy and trust it!

Stacking model is not always better than single model!

Start with lightGBM for Kaggle Competition

Thank you.