Mining Frik

Peter Us

June 8, 2016

Outline

- Preprocessing
- Model
- Results
- Other methods

Preprocessing 1/3

- ExternalCreativeId, NetworkType, ExternalSuplierId
  - Removed, to many missing values.
- TopMostReachable*, HostWindow*
  - Numerical features, missing values replaced with means
- Timestamp
  - Parsed into: dayOfMonth, dayOfWeek, hour, minute, second, millisecond
- JSON features
  - Parsed into: size and length

Preprocessing 2/3

- GEO LAT, GEO LNG
  - Added features: lat+lng, lat*lng, lat-lng
- # of missing values
  - Added feature that counts the number of missing values for each sample
- Categorical attributes
  - One-hot-encoding
  - Only encoded categories that have at least 100.000 samples

Preprocessing 3/3

- 835 features, more than 2.000.000 samples
  - Most of the values are 0 (one-hot-encoding)
- Use sparse matrix representation
  - We only store the non-0 values from the dataset
  - Makes whole dataset require only ~5GB of memory

Model 1/2

XGBoost

- eXtreme Gradient Boosting
  - supports sparse matrices
  - C++ implementation, bindings for other languages
  - provides us with feature importances
  - state of the art and winning model in recent competitions

Also...

"I think it's because most companies noticed that for other applications they can bruteforce their way with XGBoost and achieve unbeatable performance for free. Whether this sentence is a sarcasm or not it's up to you."

— user @ Kaggle forums

Model 2/2

Learning

- ~10 parameters
  - using naive CV would require $\sim O(n^{10})$ iterations)
- Learning method:
  1. Start with high learning rate, and find the optimal number of estimators
  2. Use CV to tune "tree" specific parameters
  3. For above selected parameters use CV to tune regularization parameters (L1, L2)
  4. Select lower learning rate, find the optimal number of estimators (for above tuned parameters), and train the final model.

Results

- Score: 2.9760241295
- Feature analysis:

feature name	score
TIMESTAMP	0.211131257606
GEOIP	0.19498215531
TOPMOSTREACHABLEWINDOWHEIGHT	0.0704352463388
PLATFORMVERSION	0.0396138436504
ACCOUNTID	0.0337613324035
UABROWSERVERSION	0.0329135387182
TOPMOSTREACHABLEWINDOWWIDTH	0.0325169900589
FILESJSON	0.0302060685619
HOSTWINDOWHEIGHT	0.0270336792879
CDNNAME	0.0213315830496

Other

- Tried learning/predicting each day separately with an unique model
  - 14 models, worse results.
- Tried using neural networks (Keras, Theano)
  - Comparable results, much longer computation time on desktop PC.
  - No information about feature importances
- Other models (random forest, linear SVM)
  - results obtained were not comparable with XGBoost results

Thank you for your attention

Code: ptrus

Contact: ptrusr@gmail.com