xgboost

An R package for Fast and Accurate Gradient Boosting

Tong He

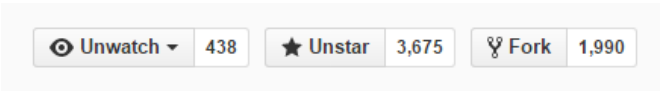
Simon Fraser University

useR! 2016

The Project *XGBoost*

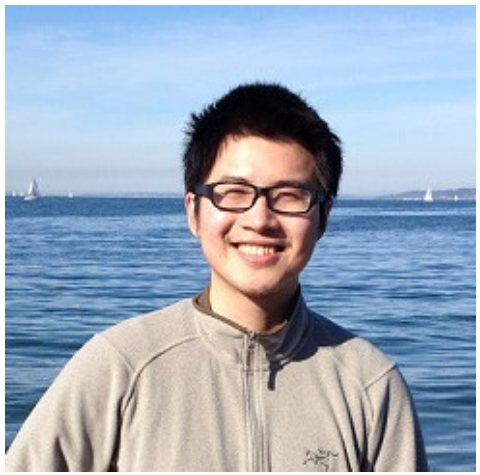
eXtreme Gradient Boosting for supervised learning.

- on Github: dmlc/xgboost



- Written in C++
- interfaces in *R*, python, Julia, Java
- Widely applied in competitions and industrial activities

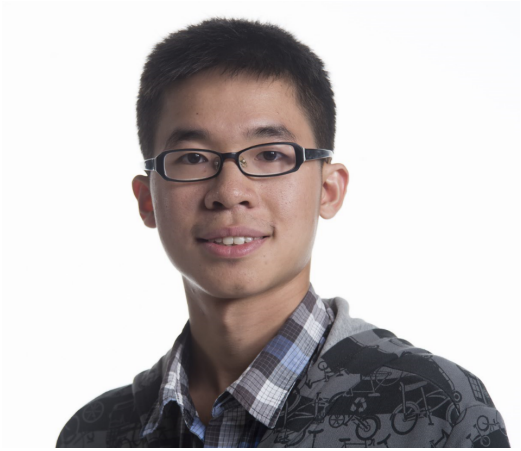
- Project Creator: *Tianqi Chen* from University of Washington



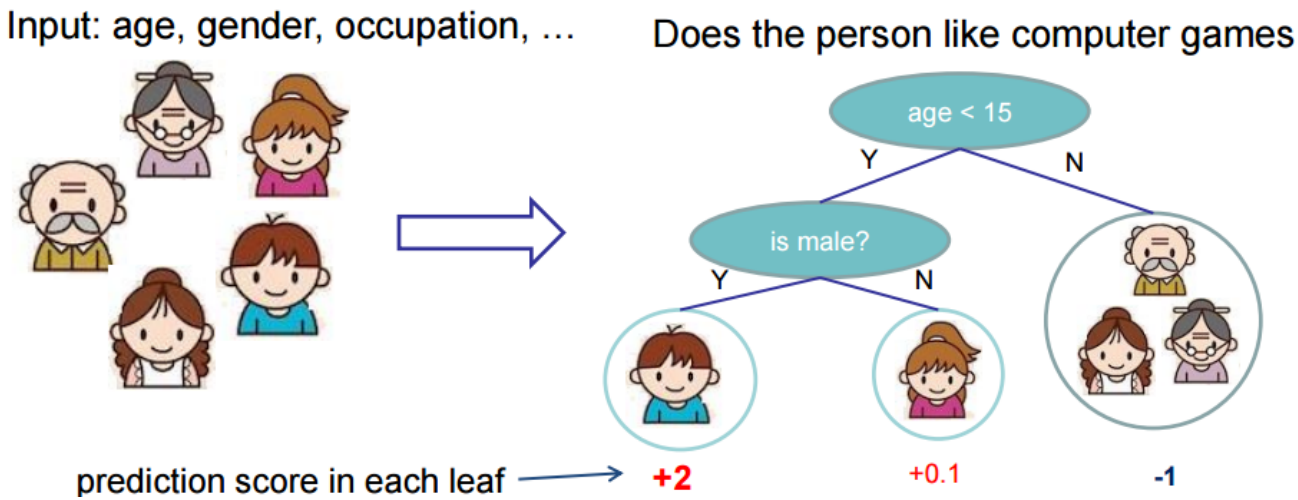
The R package *xgboost*

- John M. Chambers Statistical Software Award in 2016
- On CRAN, over 5k downloads in the last month

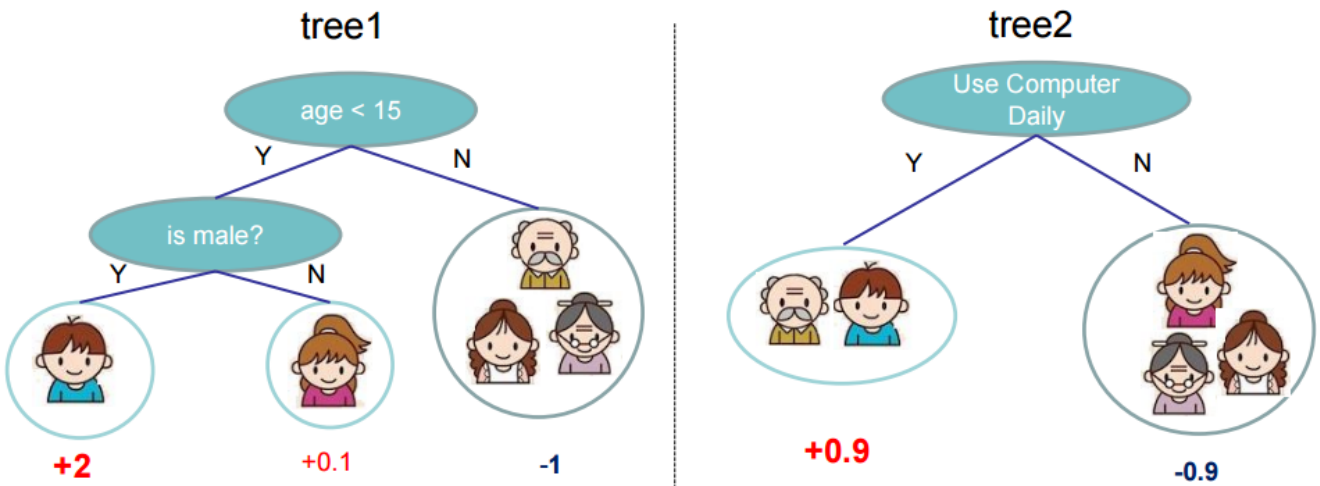
- R-package maintainer: *Tong He* from Simon Fraser University



Tree Model


Input: age, gender, occupation, …
Does the person like computer games
age < 15
Y
N
is male?
Y
N
prediction score in each leaf
+2
+0.1
-1

Tree Ensemble



f( ) = 2 + 0.9= 2.9 f( )= -1 + 0.9= -0.1

Prediction of is sum of scores predicted by each of the tree

Tree Ensemble

- Parallel: Random Forests
- Iterative: Boosting

Gradient Boosting

- Calculate the gradient of the current ensemble
- Add a new tree: take a step along the gradient

eXtreme Gradient Boosting

- L_1 and L_2 regularization
- Using both first and second order gradient
- Prune on a full binary tree

Accuracy

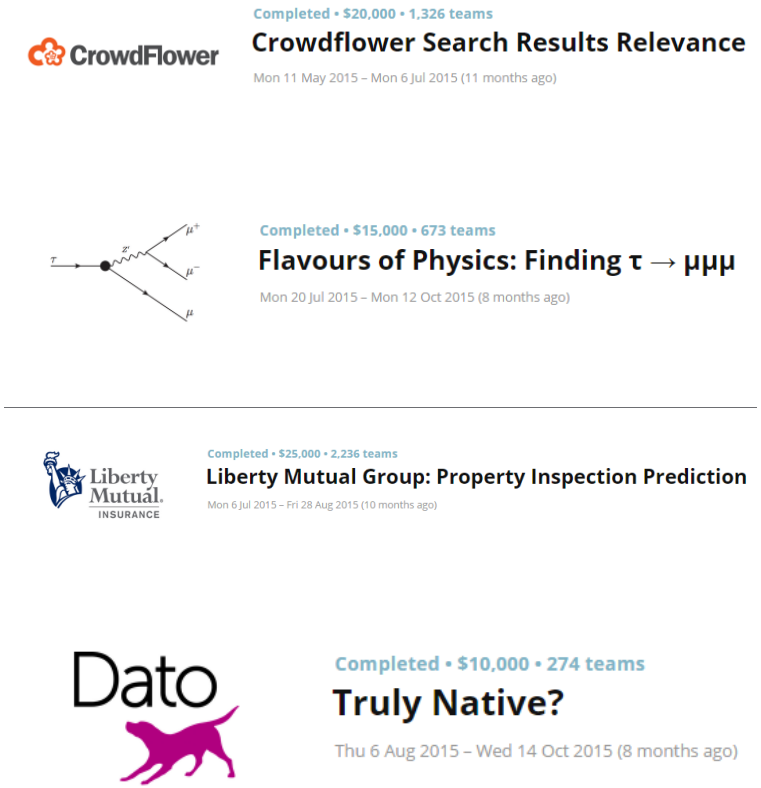
Tested on the Higgs Boson Machine Learning Competition:

	Original Data	Physical Features
R-gbm	3.38356	3.38422
python-sklearn	3.55236	3.56985
XGBoost	3.64655	3.65860
XGBoost(tuned)	3.71142	3.72370

Table: 5-fold crossvalidation result, in Approximate Median Significance

Winning Solutions

There are a number of machine learning competition winners using xgboost as a part of their solution

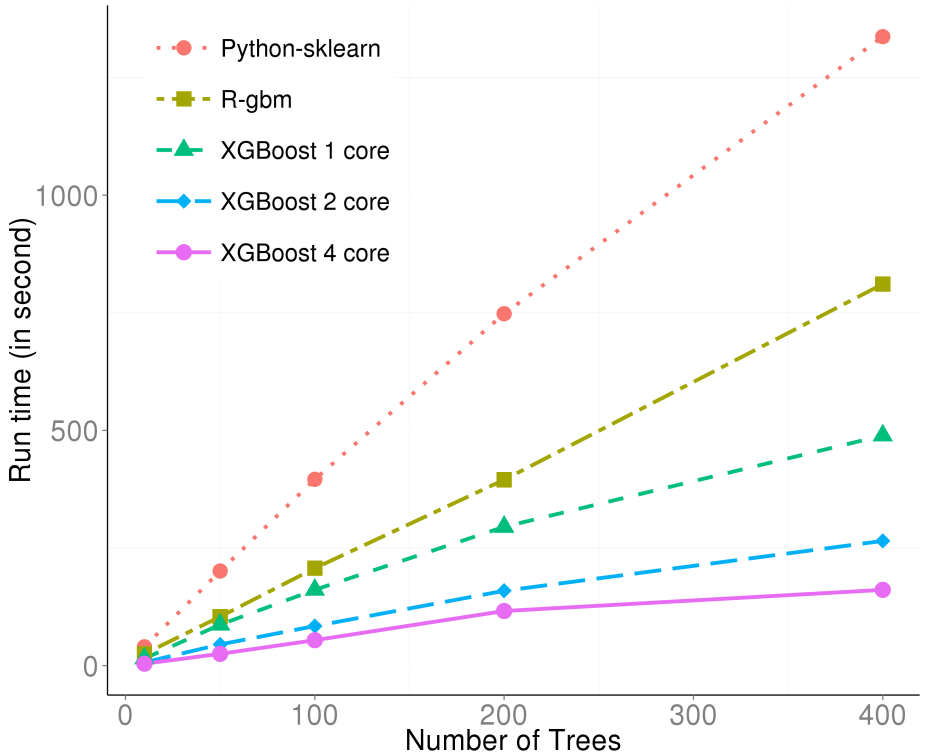


and much more

Time Efficiency

XGBoost is also known to be very fast

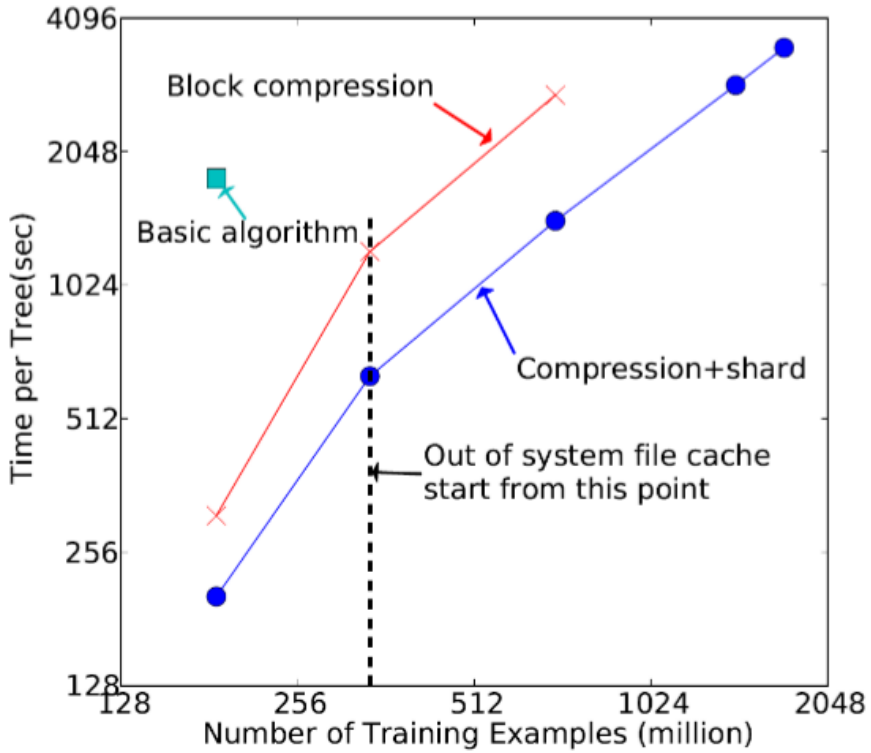
- Multi-threading by OpenMP
- Internal data structure in C++
- Pre-sort the features



External Memory Training

External memory training for larger-than-memory data

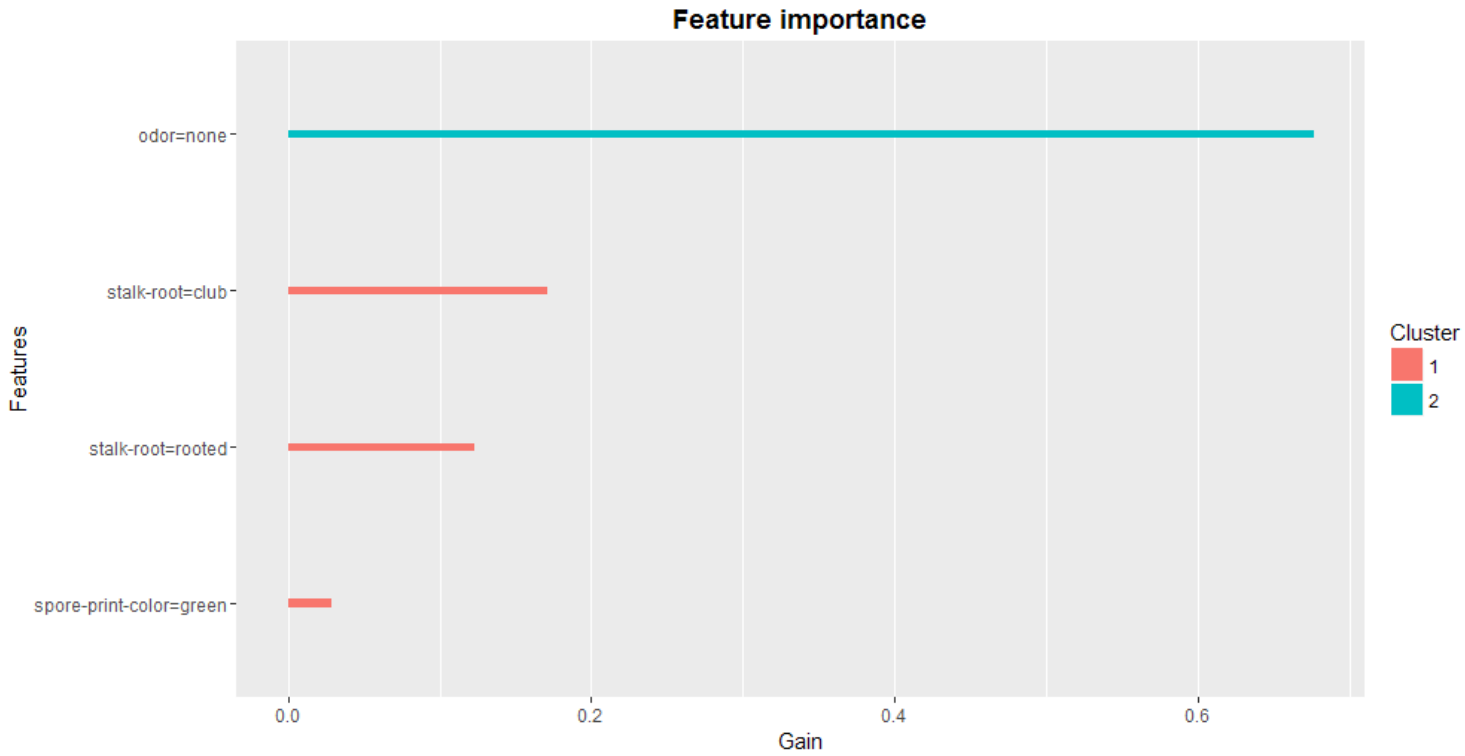
- Block compression
- Shard data onto 2 SSD disks



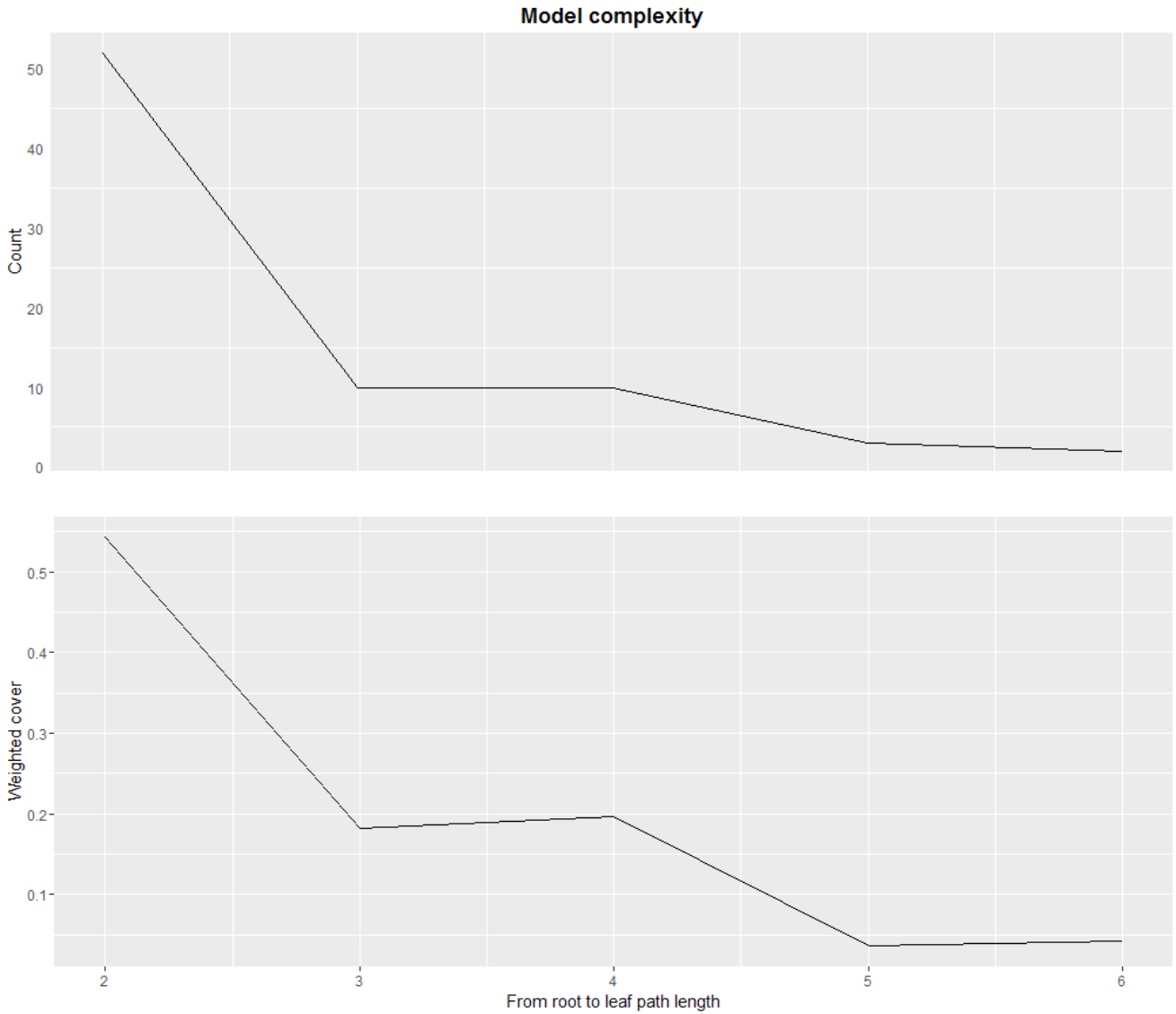
Visualization

- Feature importance
- Number of leaf per level
  - inspired by Aysen Tatarinov
- Merge multiple tree plots into one
  - inspired by Sean Welleck

Visualization



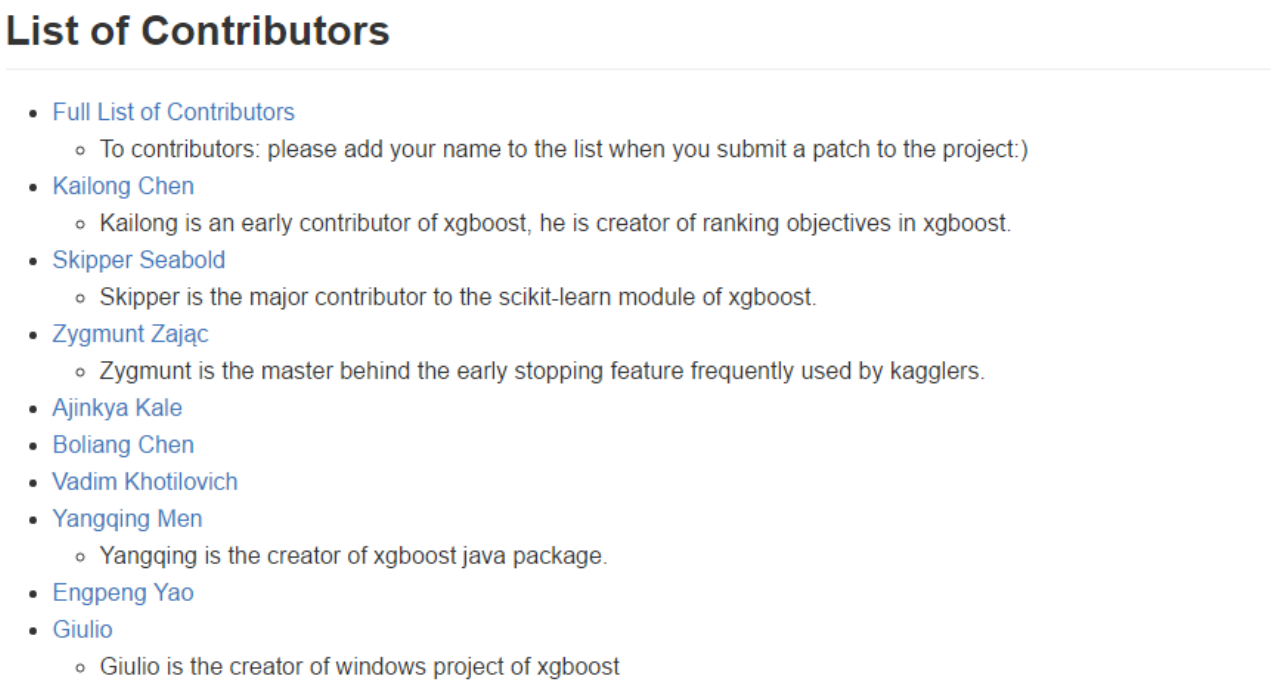
Visualization



Welcome to Contribute

- Michael, Vadim: Contributors of the R packages
- A long list of contributors all over the world!

- Pull Request
  - Contribute your code
- Issue
  - Contribute your ideas
  - Ask us a questions
  - Report a bug

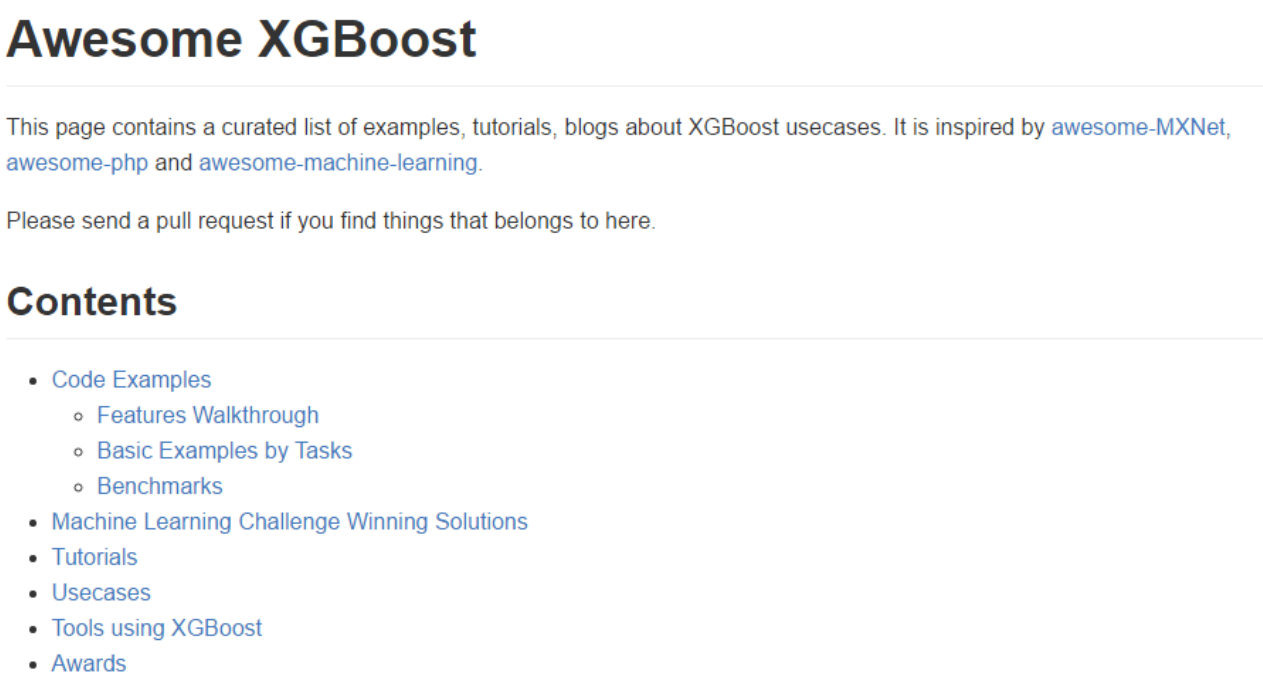


List of Contributors

- Full List of Contributors
  - To contributors: please add your name to the list when you submit a patch to the project:)
- Kailong Chen
  - Kailong is an early contributor of xgboost, he is creator of ranking objectives in xgboost.
- Skipper Seabold
  - Skipper is the major contributor to the scikit-learn module of xgboost.
- Zygmunt Zając
  - Zygmunt is the master behind the early stopping feature frequently used by kagglers.
- Ajinkya Kale
- Boliang Chen
- Vadim Khotilovich
- Yangqing Men
  - Yangqing is the creator of xgboost java package.
- Engpeng Yao
- Giulio
  - Giulio is the creator of windows project of xgboost

Resources

Demo of XGBoost:



Awesome XGBoost

This page contains a curated list of examples, tutorials, blogs about XGBoost usecases. It is inspired by awesome-MXNet, awesome-php and awesome-machine-learning.

Please send a pull request if you find things that belongs to here.

Contents

- Code Examples
  - Features Walkthrough
  - Basic Examples by Tasks
  - Benchmarks
- Machine Learning Challenge Winning Solutions
- Tutorials
- Usecases
- Tools using XGBoost
- Awards

The slides for this talk: github.com/hetong007/useR2016

References

- Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." arXiv preprint arXiv:1603.02754 (2016).
- Chen, Tianqi, and Tong He. "Higgs boson discovery with boosted trees." Cowan et al., editor, JMLR: Workshop and Conference Proceedings. No. 42. 2015.