TalkingData全球算法大赛盘点

路瑶 数据科学部

What is the…

- Power of Data Science
- Mystery of Data Scientist
- Magic of Data Scientist
- The Arena of Data Scientist

The Arena of Data Scientist

Kaggle: leading platform for crowdsourcing data challenges.

- A community to build the best solution on problems posed by industry, government and academia.
- Over 1,200 data science challenges.
- More than 600,000 registered users over 194 countries from around the world, from a wide variety of educational backgrounds and are often experts in their fields.
- Platform of KDD CUP

Famous competition and scientist of Kaggle

Kaggle Drug Discovery Competition, 2012

Geoffrey Hinton

Deep Learning

TalkingData host the competition in Kaggle for

- Provide open data and open platform for
- Smartest scientist and smartest methodology
- To solve most challenging topics

TalkingData Competition was announced in Data Science Summit in San Francisco on July 13th

Acknowledge Turi

- A famous Machine Learning company who created GraphLab.
- Acquired by Apple Inc. in August

TalkingData Mobile User Demographics

Goal of the competition:

- To know your users' profile by
- Learning from their behavior.

Given

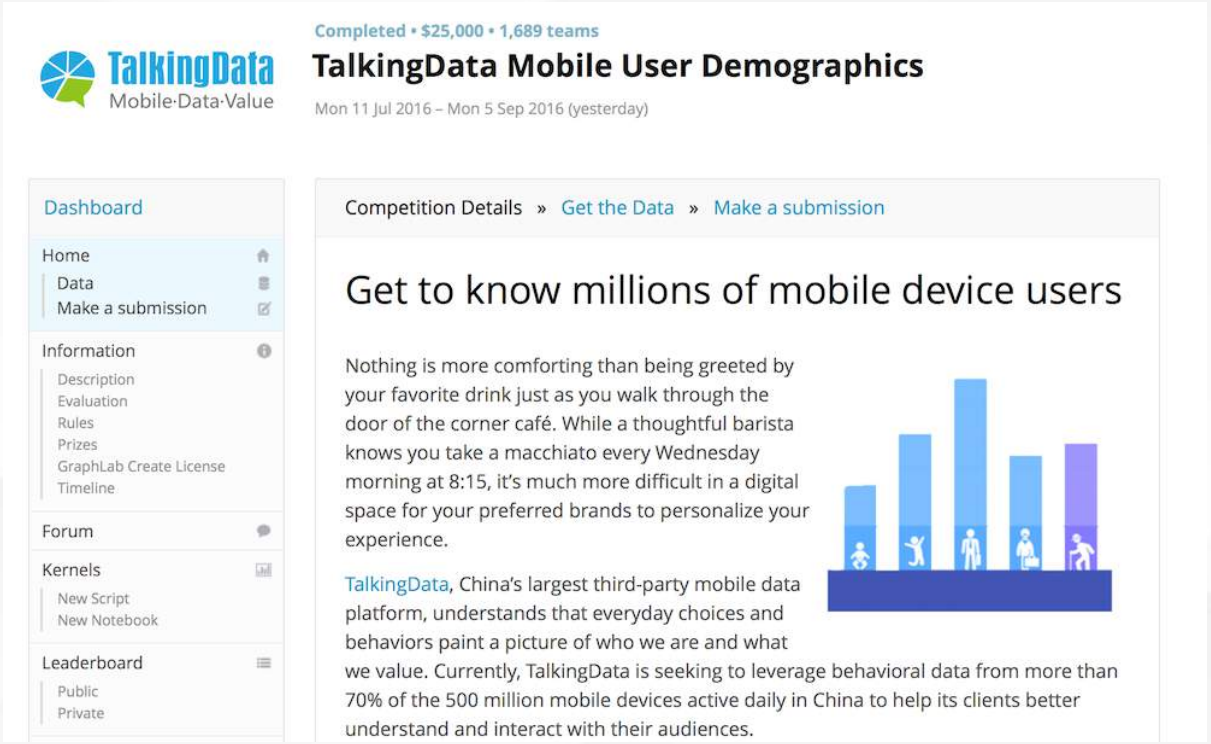
- Application usage and trace with time stamps
- Mobile brand and device mode

To optimize

- The estimation of their age and group

Evaluated by

$$logloss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log(p_{ij}),$$



Participation of the competition

1689 teams, 1961 players submitted, 24,629 entries, 2729 kernels

- **Largest** competition hosted by a Chinese company.
- Among the **highest** number of Kernels of all.
- Great proportion of **top** kagglers participated.
- **70+** countries and regions represented by the participant pool.

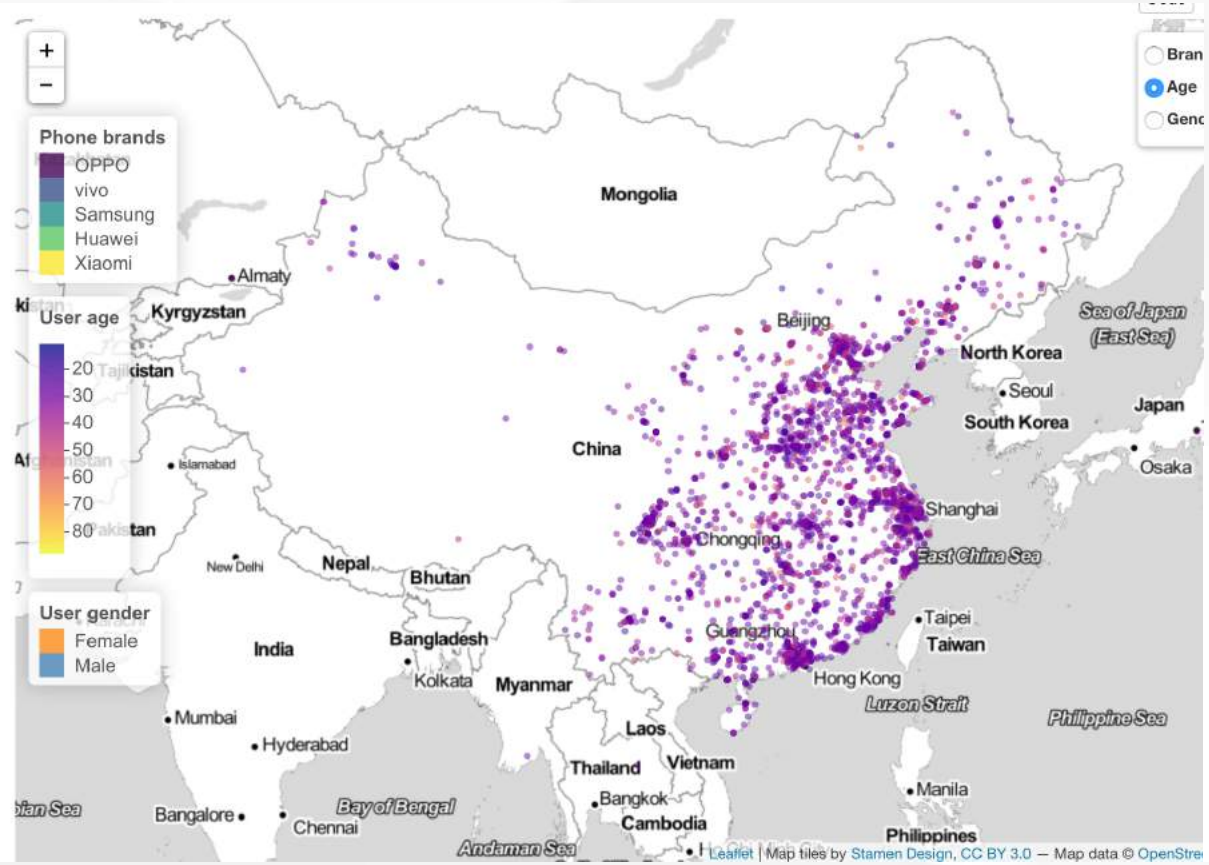
What is the…

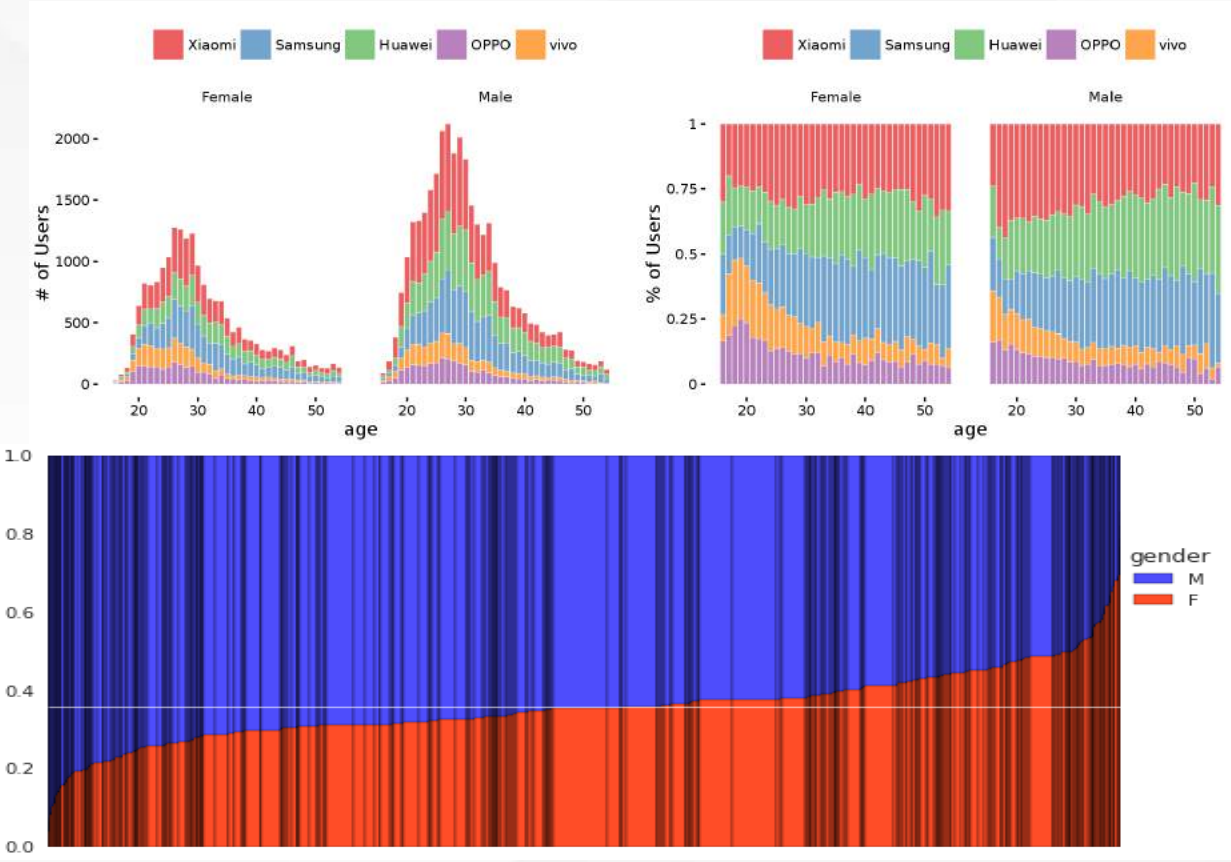
- Power of Data Science
- Mystery of Data Scientist
- Magic of Data Scientist
- The Arena of Data Scientist

Magic of Data Scientist

- Interpret the raw data
- Execute feature engineering
- Fine tune individual and ensemble models
- Avoid overfitting
- Fetch the best tools

Interpret your data

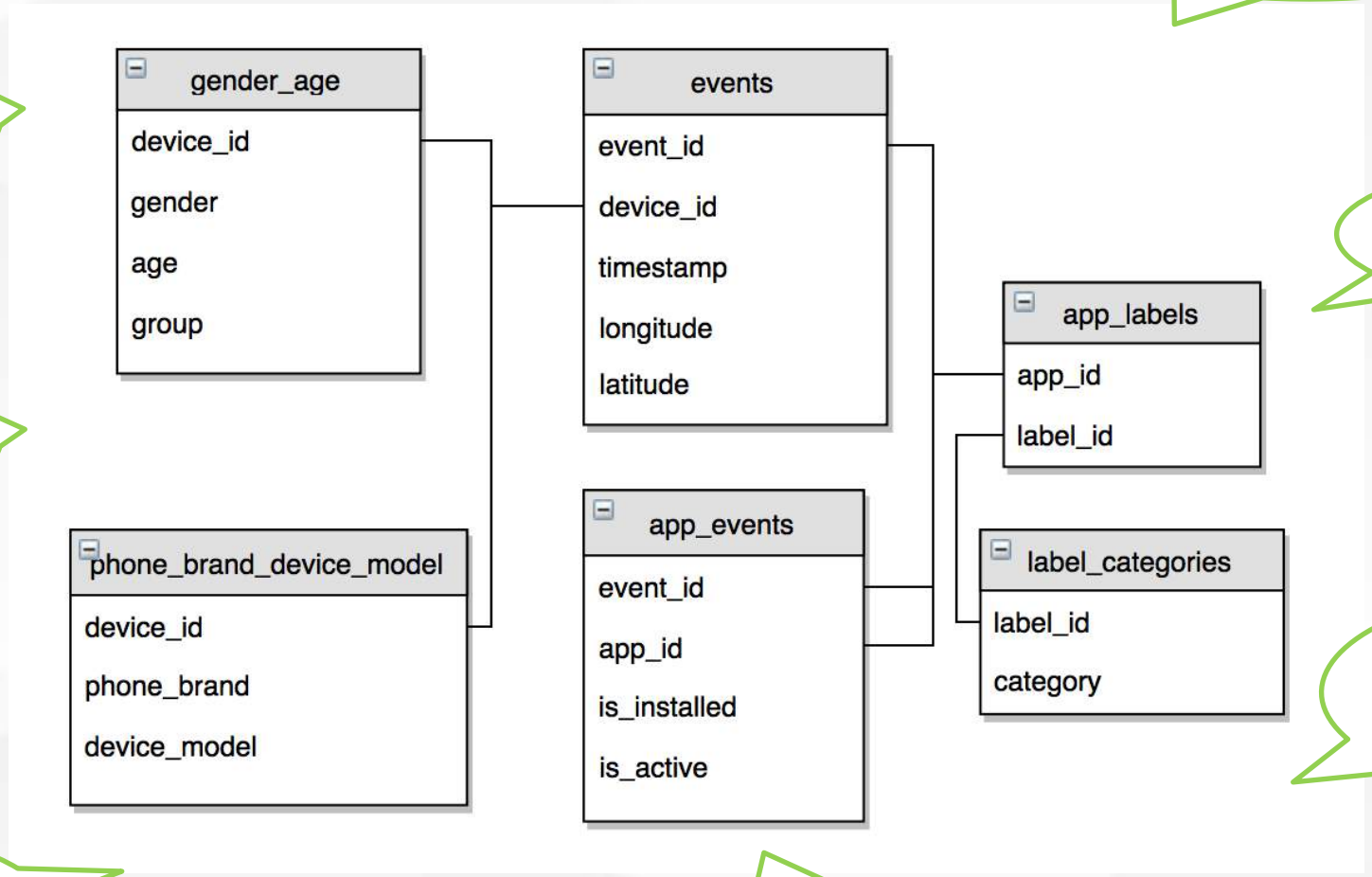




Feature engineering

Active time slots

Age/Gender as feature



Active area

App Label Patterns

District specific features

Binary/Weighted Feature

Install but not active

Fine tune individual and ensemble models

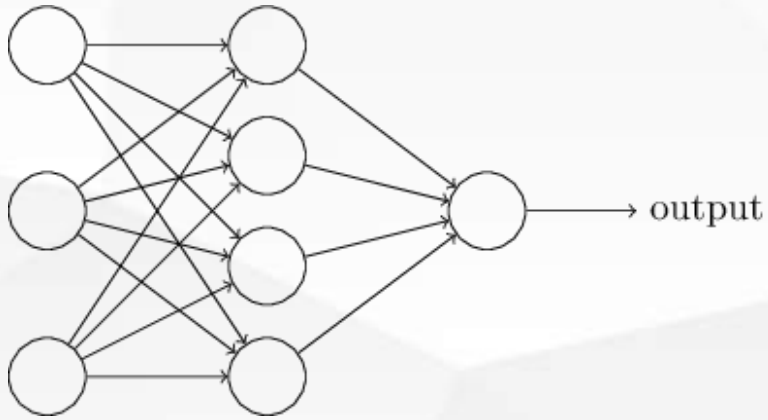
Random Forest

GBDT

SVM

Logistic Regression

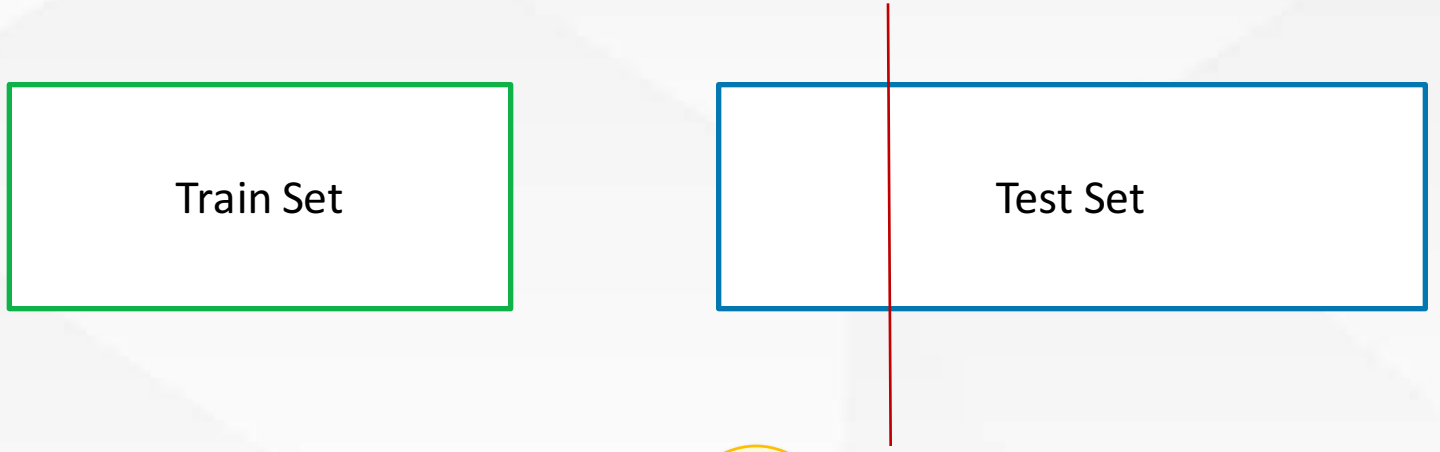
Neural Network



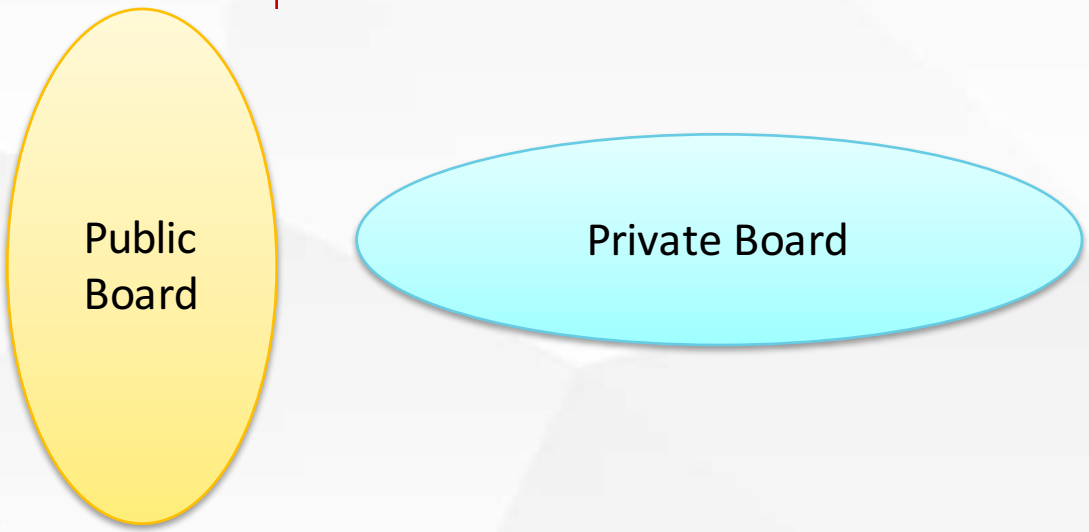
Adaboost

Stack

Avoid Overfitting



Always do
Cross Validation



Fetch the best tools

Keras

- highly modular neural networks library, written in Python
- capable of running on top of either Tensorflow or Theano.
- allows for easy and fast prototyping
- runs seamlessly on CPU and GPU.

XGboost

- Scalable, Portable and Distributed Gradient Boosting Library,
- Runs on single machine, Hadoop, Spark, Flink and DataFlow
- Higher precision, winner in Higgs Boson signal competition in Kaggle, 2014

What is the…

- Power of Data Science
- Mystery of Data Scientist
- Magic of Data Scientist
- The Arena of Data Scientist

How are our competitors

- Open and helpful
- Smart and creative, willing to solve problems in reality
- Elegant and rational
- Hard working

What we offered

- **Industrial data and valuable business problems**
- **Honor to Data Scientists’ professional skills and spirit**

Mind and Heart of the smartest scientist

Over the World

T11
2016

THANKS

Let’s rock data together!

hr@tendcloud.com

TalkingData
Mobile·Data·Value