

# Tmall

理想生活上天猫



## **TMall Repeat Buyers Prediction**

Big Data Intelligence – Tsinghua University

Armando Fortes, David  
Pissarra, Gabriele Oliaro  
14 December 2021

# How promotions work

- Offer discounts on particular occasions
- Attract a large number of new customers
- Expect some of the new buyers to become regular customers
- Avoid one-time deal hunters as much as possible
- Maximize return on investment (ROI) by targeting people most likely to become loyal customers



# How promotions work

- Offer discounts on particular occasions
- Attract a large number of new customers
- Expect some of the new buyers to become regular customers
- Avoid one-time deal hunters as much as possible
- **Maximize return on investment (ROI) by targeting people most likely to become loyal customers**

**Big Data!**



# When to use Big Data

- Predicting repeated buyers requires a lot of data
- Difficult to do for individual stores
- Readily available for large e-commerce platforms, such as TMALL
- Models trained on data from large platforms can be used by individual stores, if made available



# The competition details

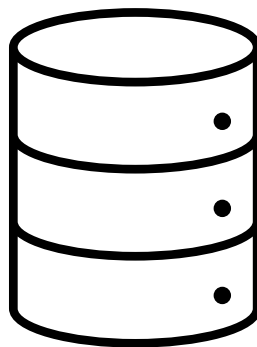
- Data from Tmall.com user behavior in the 6 months leading up to 11/11 promotion
- Given a training and testing dataset
- Need to predict labels for testing dataset and upload results in CSV format
- Use AUC (Area Under the ROC Curve) to evaluate predictions



# The competition dataset

The dataset contains two types of data:

- Customer demographic information, such as age and gender
- Customer-merchant interaction data:
  - Label indicating whether the customer is a repeated buyer (training dataset)
  - Activity log: one record (with timestamp, category, brand and item number, plus the action type) for each item that was clicked, added to cart, purchased or added to favorite



# The evaluation criteria

- We use the AUC (Area Under the ROC Curve) to benchmark our solutions
- The ROC curve is obtained by plotting the True Positive (TP) rate as a function of the False Positive (FP) rate, with one point for each classification threshold
- The AUC is the integral of the curve, evaluated from (0,0) to (1,1).
- AUC is a good measure because it is scale-invariant

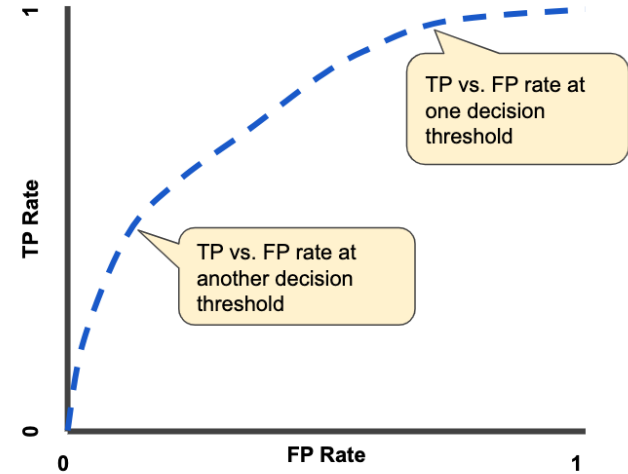


Image credits: developers.google.com

# The roadmap

**01**

**Clean data** 

The first step was to reorganize the dataset to optimize memory usage, and replace invalid (nan) values

**03**

**Training** 

We train several classifier models (CatBoostClassifier, LGBMClassifier, XGBClassifier) and optimize their hyperparameters

**02**

**Feature Engineering** 

Before training, we need to arrange the data in a way that makes training most effective. This involves creating features of interest by organizing and transforming the data with various techniques

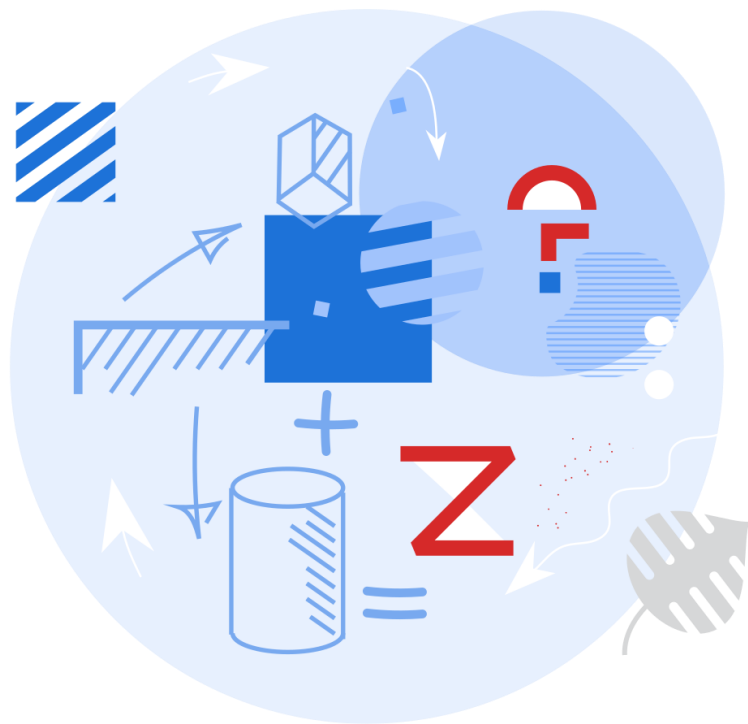
**04**

**Predicting** 

We predict using an ensemble of the trained classifier models, taking their best instances, and optimizing the weight of each model.



# Feature Engineering

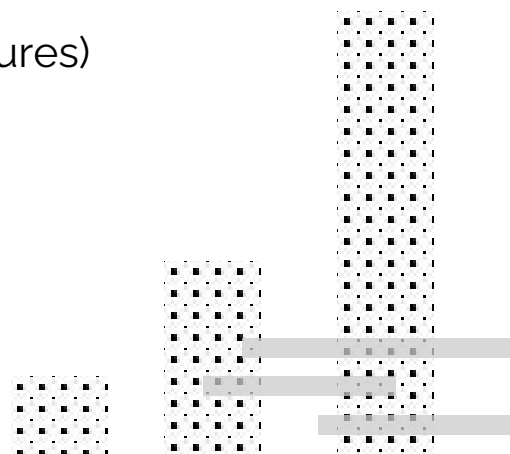




# Feature Engineering

- There's no strong correlation between users and merchants in the initial dataset

**Solution:** Create features!

1. Counting features (counting user purchase frequency, etc.)
  2. Statistical analysis features based on counting features
  3. Time period features (time span analysis, double 11 features)
  4. Principal component analysis features (PCA features)
- 

# Counting Features

Interactions regarding users, merchants or even both together.

User ID	Items	Categories	Brands	Days	Months
263947	36	26	20	22	6
338674	68	14	34	23	6
9058	42	18	19	16	5
237186	20	9	18	6	4
100871	56	22	35	17	6
...	...	...	...	...	...
49472	17	12	11	7	4
28974	129	43	61	21	4
359629	223	55	64	34	5
216711	1950	319	726	126	6
61119	153	46	72	34	5

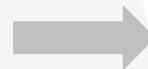


# Statistical Analysis Features

Over the calculated user-merchant counting features:

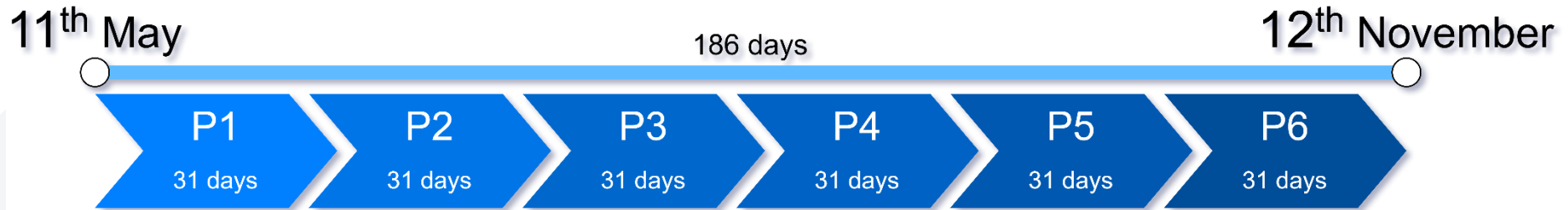
- Max
- Mean
- Median
- Standard deviation

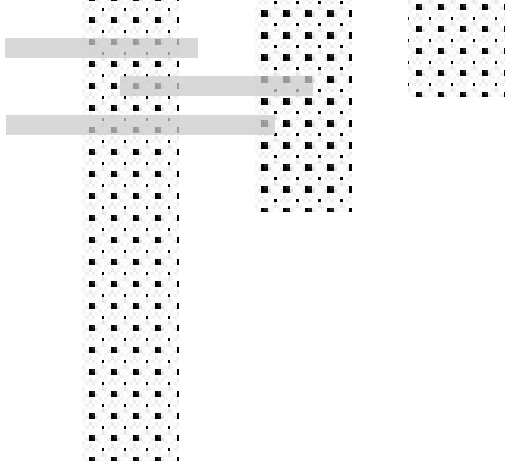
User ID	Merchant ID	Purchases
34176	3906	1
34176	121	1
34176	4356	6
34176	2217	1



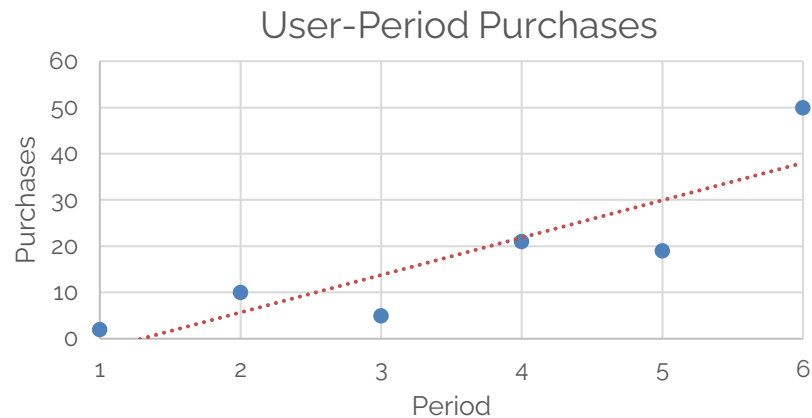
User ID	34176
max	6
mean	2.25
median	1
std	2.5

# Time Period Features



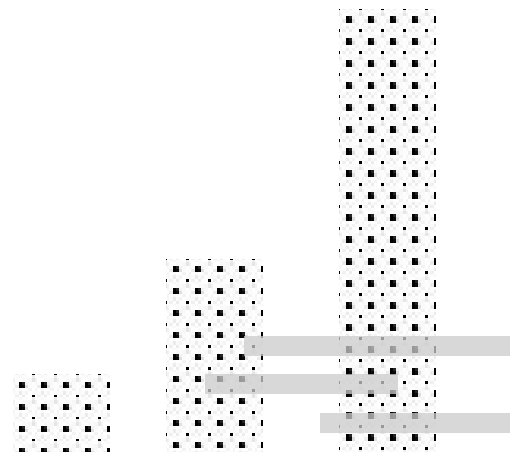
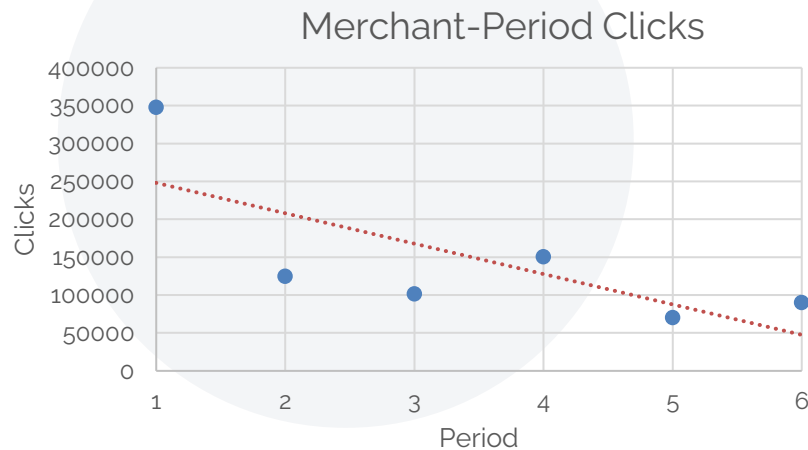


Period	Purchases
1	2
2	10
3	5
4	21
5	19
6	50



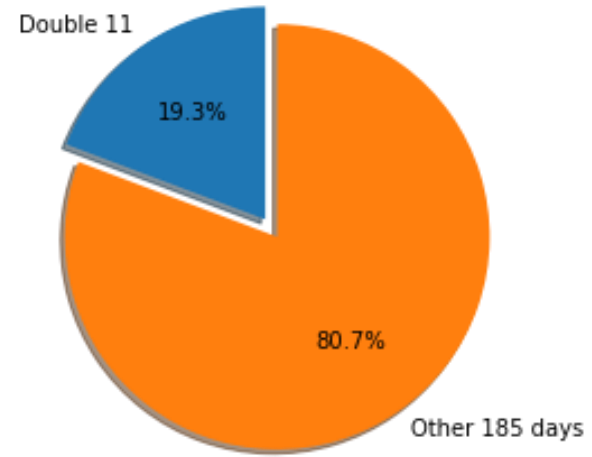
$$\frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

Period	Clicks
1	347901
2	125048
3	101852
4	150874
5	70247
6	90115

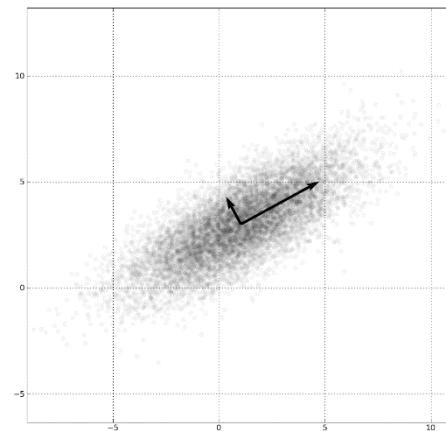
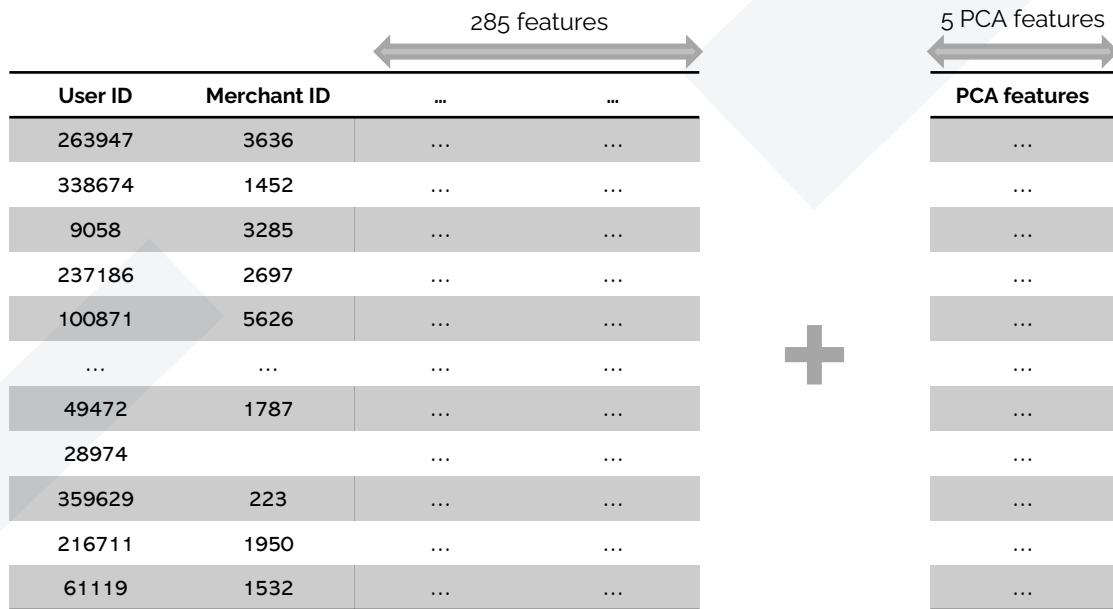


# Double 11 Features

- 11<sup>th</sup> November – shopping festival
- Huge volume of online transactions
- One-time deal hunters



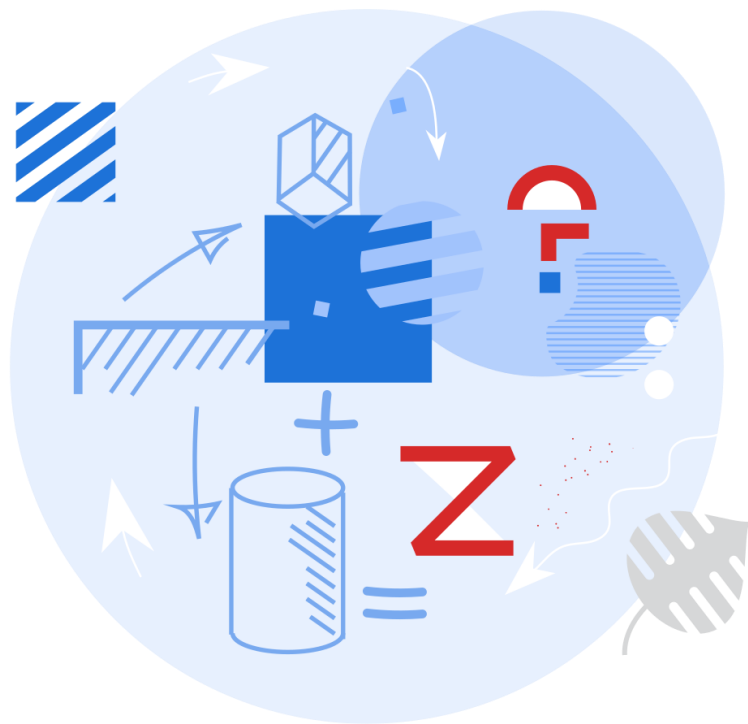
# PCA Features



Summarize all features in 5 dimensions using Principal Component Analysis and append it to the other features

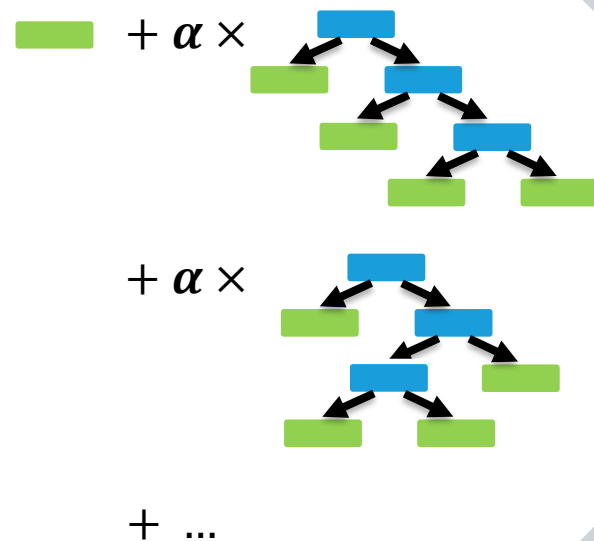


# Training Models



# Gradient Boosting

- Gradient Boosting is a stage-wise additive model which generates weak-learners during the learning process.
- Decision trees are gradually added one at a time, each of them being trained using the residual errors of their predecessors as labels.
- The contribution of a given decision tree to the ensemble is based on the gradient descent optimization process, where we attempt to minimize the overall error of the strong-learner.



# Gradient Boosting Implementations



Stands for e**X**treme **G**radient **B**oosting. Initially started as a research project by Tianqi Chen in March 2014, becoming famous by 2016.



Developed by Microsoft, being first released in January 2017. Specifically designed to achieve faster training speeds and higher efficiency.



Yandex  
CatBoost

Open-sourced by Yandex, one of Russia's leading tech companies in April 2017. Provides an innovative algorithm for processing **C**ategorical features.

# K-Fold Cross Validation

dmlc  
**XGBoost**

Avg Score: **0.6913**  
Best Score: **0.7067**



**LightGBM**

Avg Score: 0.6773  
Best Score: 0.6876



Yandex  
**CatBoost**

Avg Score: 0.6871  
Best Score: 0.7011

# Ensemble Model


$$P(u, m) = \sum_K w_k \cdot P_k(u, m)$$

# Ensemble Model

$$P(u, m) = \sum_K w_k \cdot P_k(u, m)$$


Model	ROC-AUC Score
XGBoost	0.6913
LightGBM	0.6773
CatBoost	0.6871
<b>Ensemble Model</b>	<b>0.6924</b>

# Fetching Best Features using XGBoost



Feature	Importance
items_user_merchant	5.358978
purchases_user_merchant_period_5	5.250384
purchases_user_merchant	5.186556
categories_user_merchant	4.047388
periods_user_merchant	4.018932
double11_periods_user_merchant_ratio	2.951180
categories_merchant	2.078619
favourites_merchant_user_max	2.042979
...	...

# Fetching Best Features using XGBoost



Feature	Importance
items_user_merchant	5.358978
purchases_user_merchant_period_5	5.250384
purchases_user_merchant	5.186556
categories_user_merchant	4.047388
periods_user_merchant	4.018932
double11_periods_user_merchant_ratio	2.951180
categories_merchant	2.078619
favourites_merchant_user_max	2.042979
...	...

**290  
Features**



**150  
Features**

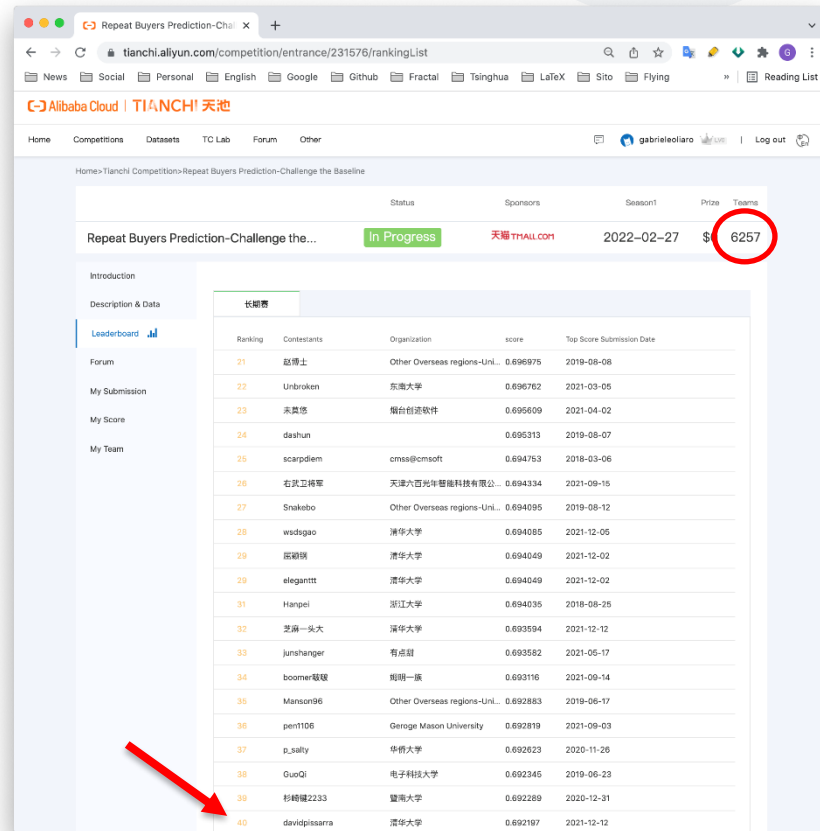


# Final Scores

Model	Every Feature (290)	Best Features (150)
XGBoost	0.6913	0.6916
LightGBM	0.6773	0.6757
CatBoost	0.6871	0.6882
<b>Ensemble Model</b>	0.6924	<b>0.6925</b>

# Conclusion

- Analyzed and cleaned large dataset
- Performed feature engineering to extract a variety of features of different complexity levels
- Implemented prediction model based on ensemble of classifiers
- Submitted our solution on tianchi.aliyun.com and made it to the Top 40 Leaderboard!



Repeat Buyers Prediction-Challenge the... In Progress 天福 THALL.COM 2022-02-27 \$6257

Ranking	Contestants	Organization	score	Top Score Submission Date
21	赵博士	Other Overseas regions-Unl...	0.696975	2019-06-08
22	Unbroken	东南大学	0.696762	2021-03-05
23	未菜葱	烟台创源软件	0.695609	2021-04-02
24	dashun		0.695313	2019-08-07
25	scarpdiem	cmss@cmsoft	0.694753	2018-03-06
26	右沃卫将军	天津六西格玛智能科技有限公司	0.694334	2021-09-16
27	Snakebo	Other Overseas regions-Unl...	0.694095	2019-08-12
28	wsdsgao	清华大学	0.694085	2021-12-05
29	雷福强	清华大学	0.694049	2021-12-02
29	eleganttt	清华大学	0.694049	2021-12-02
31	Hangei	浙江大學	0.694035	2018-08-25
32	芝罘—头大	清华大学	0.693594	2021-12-12
33	junahanger	有点甜	0.693582	2021-05-17
34	boomer望望	相闻一族	0.693116	2021-09-14
35	Manson96	Other Overseas regions-Unl...	0.692883	2019-06-17
36	pen1106	George Mason University	0.692819	2021-09-03
37	p_salty	华侨大学	0.692623	2020-11-26
38	GuoQi	电子科技大学	0.692345	2019-06-23
39	杉崎耀2233	暨南大学	0.692289	2020-12-31
40	davidpisarra	清华大学	0.692197	2021-12-12