# Profiling Top Kagglers: Bestfitting, Currently #1 in the World

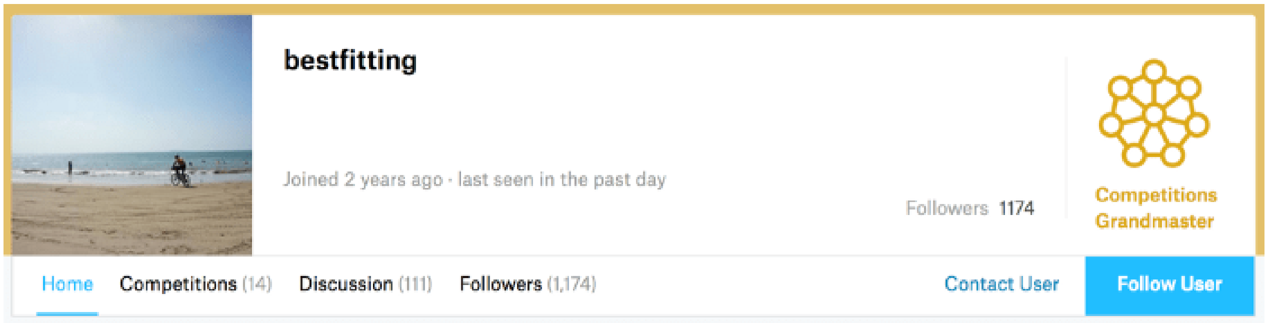
Kaggle Team Follow
May 7, 2018 · 8 min read

*We have a new #1 on our leaderboard — a competitor who surprisingly joined the platform just two years ago. Shubin Dai, better known as Bestfitting on Kaggle or Bingo by his friends, is a data scientist and engineering manager living in Changsha, China. He currently leads a company he founded that provides software solutions to banks. Outside of work, and off Kaggle, Dai's an avid mountain biker and enjoys spending time in nature. Here's Bestfitting:*

## Can you tell us a little bit about yourself and your background?

I majored in computer science and have more than 10 years of experience in software development. For work, I currently lead a team that provides data processing and analyzing solution for banks.

Since college, I've been interested in using math to building programs that solve problems. I continually read all kinds of computer science books and papers, and am very lucky to have followed the progress made on machine learning and deep learning within the past decade.



bestfitting

Joined 2 years ago · last seen in the past day

Followers 1174

Competitions Grandmaster

## How did you start with Kaggle competitions?

As mentioned before, I've been reading a lot of books and papers about machine learning and deep learning, but found it always hard to apply the algorithms I learned on small datasets that are readily available. So I found Kaggle a great platform with all sorts of interesting datasets, kernels, and great discussions. I couldn't wait to try something, and first entered the "Predicting Red Hat Business Value" competition.

## What is your first plan of action when working on a new competition?

Within the first week of a competition launch, I create a solution document which I follow and update as the competition continues on. To do so, I must first try to get an understanding of the data and the challenge at hand, then research similar Kaggle competitions and all related papers.

## What does your iteration cycle look like?

1. Read the overview and data description of the competition carefully

2. Find similar Kaggle competitions. As a relatively new comer, I have collected and done a basic analysis of all Kaggle competitions.

3. Read solutions of similar competitions.

4. Read papers to make sure I don't miss any progress in the field.

5. Analyze the data and build a stable CV.

6. Data pre-processing, feature engineering, model training.

7. Result analysis such as prediction distribution, error analysis, hard examples.

8. Elaborate models or design a new model based on the analysis.
9. Based on data analysis and result analysis, design models to add diversities or solve hard samples.
10. Return to a former step if necessary.

## What are your favorite machine learning algorithms?

I choose algorithms case by case, but I prefer to use simple algorithms such as ridge regression when ensemble, and I always like starting from resnet-50 or designing similar structure in deep learning competitions.

## What are your favorite machine learning libraries?

I like pytorch in computer vision competitions very much. I use tensorflow or keras in NLP or time-series competitions. I use seaborn and products in the scipy family when doing analysis. And, scikit-learn and XGB are always good tools.

## What is your approach to hyper-tuning parameters?

I try to tune parameters based on my understanding of the data and the theory behind an algorithm, I won't feel safe if I can't explain why the result is better or worse.

In a deep learning competition, I often search related papers and try to find what the authors did in a similar situation.

And, I will compare the result before and after making parameter changes, such as the prediction distribution, the examples affected, etc.

## What is your approach to solid cross-validation/final submission selection and LB fit?

A good CV is half of success. I won't go to the next step if I can't find a good way to evaluate my model.

To build a stable CV, you must have a good understanding of the data and the challenges faced. I'll also check and make sure the validation set has similar distribution to the training set and test set and I'll try to make sure my models improve both on my local CV and on the public LB.

In some time series competitions, I set aside data for a period of time as a validation set.

I often choose my final submissions in a conservative way, I always choose a weighted average ensemble of my safe models and select a relatively risky one (in my opinion, more parameters equate to more risks). But, I never chose a submission I can't explain, even with high public LB scores.

## In a few words, what wins competitions?

Good CV, learning from other competitions and reading related papers, discipline and mental toughness.

## What is your favorite Kaggle competition and why?

Nature protection and medical related competitions are my favorite ones. I feel I should, and perhaps can, do something to make our lives and planet better.

## What field in machine learning are you most excited about?

I am interested in all kinds of progress in deep learning. I want to use deep learning to solve problems besides computer vision or NLP, so I try to use them in competitions I enter and in my regular occupation.

## How important is domain expertise for you when solving data science problems?

To be frank, I don't think we can benefit from domain expertise too much, the reasons are as follows:

1. Kaggle prepared the competition data carefully, and it's fair to everyone;
2. It's very hard to win a competition just by using mature methods, especially in deep learning competitions, thus we need more creative solutions;
3. The data itself is more important, although we may need to read some materials related.

But, there are some exceptions. For example, in the Planet Amazon competition, I did get ideas from my personal rainforest experiences, but those experiences might not technically be called domain expertise.

## What do you consider your most creative trick/find/approach?

I think it is to prepare the solution document in the very beginning. I force myself to make a list that includes the challenges we faced, the solutions and papers I should read, possible risks, possible CV strategies, possible data augmentations, and the way to add model diversities. And, I keep updating the document. Fortunately, most of these documents turned out to be winning solutions I provided to the competition hosts.

## How are you currently using data science at work and does competing on Kaggle help with this?

We try to use machine learning in all kinds of problems in banking: to predict visitors of bank outlets, to predict cash we should prepare for ATMs, product recommendation, operation risk control, etc.

Competing on Kaggle also changed the way I work, when I want to find a solution to solve a problem, I will try to find similar Kaggle competitions as they are precious resources, and I also suggest to my colleagues to study similar, winning solutions so that we can glean ideas from them.

## What is your opinion on the trade-off between high model complexity and training/test runtime?

Here are my opinions:

1. Training/test runtime is important only when it's really a problem. When accuracy is most important, model complexity should not be too much of a concern. When the training data obtained resulted from months of hard work, we must make full use of them.

2. It's very hard to win a competition by only using ensemble of weak models now. If you want to be number 1, you often need very good single models. When I wanted to ensure first place in a competition solo, I often forced myself to design different models which could reach the top 10 on the LB, sometimes, even top 3. The organizers can select any one of them.

3. In my own experiences, I may design models in a competition to explore the upper limitation of this problem, and it's not too difficult to then choose a simple one to make it feasible in a real situation. I always try my best to provide a simple one to organizers and discuss with them in the winner's call. I found some organizers even use our solutions and ideas to solve other problems they face.

4. We can find that Kaggle has a lot of mechanisms to ensure the performance when the training/test runtime is important: kernel competitions, team size limitation,

adding more data that aren't calculated while scoring, etc. I am sure Kaggle will also improve the rules according to the goal of the challenge.

## How did you get better at Kaggle competitions?

Interesting competitions and great competitors on Kaggle make me better.

With so many great competitors here, winning a competition is very difficult, they pushed me to my limit. I tried to finish my competitions solo as many times as possible last year, and I must guess what all other competitors would do. To do this, I had to read a lot of materials and build versatile models. I read all the solutions from other competitors after a competition.

## Is there any recent or ongoing machine learning research that you are excited about?

I hope I can enter a deep reinforcement learning competition on Kaggle this year.

## You moved up the leaderboard to take the number 1 spot very quickly (in just 15 months). How did you do it?

First of all, №1 is a measurement of how much I learned on Kaggle and how lucky I was.

In my first several competitions, I tried to turn the theories I learned in recent years into skills, and learned a lot from others.

After I gained some understanding of Kaggle competitions, I began to think about how to compete in a systematic way, as I have many years of experience in software engineering.

About half a year later, I received my first prize and some confidence. I thought I might become a grandmaster in a year. In Planet Amazon competition, I tried to get a golden medal, so it came to me as a surprise when I found out I was in first place.

Then I felt I should keep using the strategies and methods I mentioned before and got more successes. After I won the Cdiscount competition, I climbed to the top of Users Rank board.

I think I benefited from the Kaggle platform, I learned so much from others and the rank system of Kaggle also play an important role in my progress. I also felt so lucky as I I never expected I could get 6 prizes in a row, my goals of many competitions were top 10 or top 1%. I don't think I could replicate the journey again.

However, I'm here not for a good rank. I always treat every competition as an opportunity to learn, so I try to select competitions from the field I am not so familiar with, which forced myself to read hundreds of papers last year.

### You've mentioned before that you enjoy reading top-scoring competition solutions from past competitions. Are there any you would highlight as being particularly insightful?

I respect all the winners and wonderful solution contributors, I know how much effort they put into it. I always read the solutions with an admirable attitude.

A few of the most memorable insights came from Data Science Bowl 2017: the pytorch, 3D segmentation of medical images, the solutions from the Web Traffic Time Series Forecasting which use sequence model from NLP to solve time series problem, and the beautiful solutions from Tom ( https://www.Kaggle.com/tvdwiele ), and Heng ( https://www.Kaggle.com/hengck23 ).

. . .

*Originally published at http://blog.kaggle.com on May 7, 2018.*

Machine Learning Kaggle Kaggle Competition Grandmaster

297

WRITTEN BY
Kaggle Team Follow

Official authors of Kaggle winner's interviews + more! Kaggle is the world's largest community of data scientists. Join us at kaggle.com.

Kaggle Blog Follow

Official Kaggle Blog ft. interviews from top data science competitors and more!

More From Medium

How can machines think ? Machine learning from scratch

Abidi Ghofrane

K-Nearest Neighbours

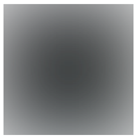
Himani Mogra

NLP Pipelines in a single line of code

Rahul Madan in Analytics Vidhya

Use C# and ML.NET Machine Learning To Predict Taxi Fares In New York

Mark Farragher in The Machine Learning Advantage



MRR vs MAP vs NDCG: Rank-Aware Evaluation Metrics And When To Use Them

Moussa Taifi, Ph.D. in The Startup

Computer Vision: Lane Finding Through Image Processing

Archit Rastogi in The Startup

Beginner’s Guide to Everything Image Recognition

Alexander Chow

These Frameworks Have Helped LinkedIn Build Machine Learning at Scale

Jesus Rodriguez in The Startup