Kaggle (company) | Contests and Competitions | Competitive Programming | Data Science

List Question

What are some tips on becoming really good at data science competitions like Kaggle?

6 Answers

Giuliano Janson, Knows more coding than a statistician and more stats than an engineer.

Written 33w ago · Upvoted by Luis Argerich, Data Science Professor at UBA since 1997.

I've participated in Kaggle competitions for a couple of years, won two competitions, including a recent Kaggle Master competition and can share some of my tips.

Since you're asking about tips I'll spare you the fundamentals, like look at and understand your data, good modelling practices, good cross validation...

In terms of tips here's what you need to know before you can become really competitive (by really competitive I mean regularly doing top 10% and hitting top-10 once in a while):

-ensembling: know how to do it like a Kaggle pro. Use out of bag estimates to create predictions for the entire train set, and feed those into metamodels. Consider doing so with 2nd level meta-models as well. For example, given (X,y), predict y with a few models

Related Questions

Are Kaggle-like competition skills most valuable for data science jobs?
4,443 Views

What Kaggle competitions should a beginner start with?
24,553 Views

What are some prestigious/ popular data science competitions like Kaggle?
4,650 Views

How useful is a participation in Kaggle competitions for data science career?
3,951 Views

Is the top 25% in a Kaggle competition considered good for a beginner in the field of data science?
4,022 Views

What background do I need to do Kaggle competitions?
30,608 Views

How can I become data scientist + kaggle?
1,512 Views

Is a PhD reqt becoming increasingly irrelevant for Data Science?
12,308 Views

What are some good toy problems in data science?
228,630 Views

Would the following data science certificate programs be good preparation for Kaggle competitions, and a career switch into data science jobs?...

(Gradient Boosting, Random Forest, Neural Networks). Then use (X+predictions) to predict y using (i.e.) a Gradient Boosting. That is a metamodel. In a 2nd level metamodel, you predict y using a series o metamodels (GB, neural networks) and predict y for the 3rd time.

-semi-supervised learning: when you have high accuracy, you can predict the test set, then append train and test, train your algorithm on both and predict y again. I used this technique to win my first competition.

-get exposure to other tools and algorithms, like XGBoost, Vowpal Wabbit, libFM.

-team up with the right people at the right time. There is an art to that. It comes down to teaming up as early as you can, but not before the LB has settle down and you have truly figured out who is competitive and who just had a fast start. In many competitions you'll have hundreds of people that will be first to impress with some great progress. Most of those will not be able to sustain and eventually fade. But after a month or so of competition, you'll be able to see who is making real progress. The main problem though is, will you be in a position to ask one of these folks to work as a team or will you be too much behind already? So, as you see, it becomes an optimization problem. Get the best teammate you can given all the other moving pieces.

2.1k Views · View Upvotes

22,675 Views

Related Questions

More Answers Below

Are Kaggle-like competition skills most valuable for data science jobs?
4,443 Views

What Kaggle competitions should a beginner start with?
24,553 Views

What are some prestigious/ popular data science competitions like Kaggle?
4,650 Views

How useful is a participation in Kaggle competitions for data science career?
3,951 Views

Is the top 25% in a Kaggle competition considered good for a beginner in the field of data science?
4,022 Views

Roman Trusov, Regression is my profession

Written 33w ago

Apart the obvious excellence with your tools there are several things that you will often hear about in top kagglers' interviews but will not (likely) find in guides and tutorials.

There are good analysts that know their p-tests and their distributions. There are cool guys proficient in Theano who can construct neural networks after reading about them in a paper. There are matplotlib ninjas that can squeeze five-dimensional structure into a 2D plot in a representative way. Those skills are useful for real jobs, but that's what they are - skills. Instead if that I want to talk about art, because that's what those guys from the top do.

If you observe a leaderboard of a typical competition with a high prize you will notice that the majority falls far behind the top-20 or top-50. Those solutions are slightly tunes (if at all) baselines or quickly coded scripts that use off-the-shelve methods. What can you do to surpass them? They have the same libraries, they read the same forum and they have the same data. So, do what nobody else does.

Very often you can see that almost everyone at the competition uses the same core method. But if you find a non-obvious feature, you can get an advantage that is very hard to find - just because not many people look for it.

The winners excel at forming teams. The best teams don't consist of people who know the same technology - the participants complement each other both in terms of tech skills and in terms of general knowledge.

They know a lot about ensembling, stacking and model generalization. Those methods are not widely used in day-to-day tasks, in fact, they are endemic for kaggle competitions and some research cases. But they've proven to be very useful for getting that 0.0001% of score.

Hand-crafted features. Almost every applied state-of-the-art method that offers some significant improvement in a particular tasks uses hand-crafted features based on the insight about the inner structure of the problem. PCA is good. Autoencoders are good. But one cleverly engineered feature beats them every time.

After all, the success in the competition doesn't boil down to following some "tips" or guidelines on winning Kaggle competition, it's all about how deep you understand the problem and the data.

885 Views · View Upvotes · Answer requested by Sabrina Ali

Ani Rud, Blogger at Analyticscosm
Written 33w ago

For a beginner I would recommend all of the knowledge and getting started competitions which will give a good understanding of various techniques to be used and methods to be followed.

Specifically, I would recommend the following in order:

- Binary Classification: Titanic: Machine Learning from Disaster
- Multi-Class Classification: Forest Cover Type Prediction
- Regression with temporal component: Bike Sharing Demand
- Binary Classification with text data: Random Acts of Pizza

There's rich discussion on forums on these competitions and well as others. You may see how others have tried to approach the problems and get a fair idea on how to proceed. I would also suggest the below getting started tutorial and solutions from previous winners.

Start Solving Kaggle Problem With R: One Hour Tutorial

Learning Predictive Analytics: Kaggle Competition Solutions

388 Views · View Upvotes · Answer requested by Sabrina Ali

Yilun (Tom) Zhang, Aspirational data scientist (Actively using R and Python)
Updated 33w ago

I have done a few kaggle competitions (the ranking wasn't that good, around top 10-20%) and participated in 2 data science competitions offline. In my opinion, those the two types of competitions are completely different.

In Kaggle, the final goal is simple: to achieve a high prediction accuracy. Which requires a very deep understanding to machine learning algorithms. I think Giuliano Janson explained how you can improve your ranking in Kaggle competitions very well so I won't spend too many words explaining it (and he did much better than my in the competitions).

Basically I would recommend sticking on to one competitions (those in the tutorials would be good targets), and

- keep improving your result by trying something new (new model, new parameters, new/modified data columns)
- look at others' script and learn from them (learn the model, they way they handle parameters and variables), sometimes having a new variable will significantly improve your result

---

For offline data competitions, I was lucky to win both of the competitions I participated (thanks to my teammates).

The first one involves search engine marketing. We were given the data (multiple files) and the background and asked to predict the break-even bid for each advertisement the company has under each user search scenario. We won the competition by

- understanding the question very well
- had a very reasonable approach to the problem (very similar to the benchmark model provided by the host)
- had a decent behaving model
- explained everything very well in the presentations and Q&A

The second one is more like a computer science hackathon but with a data theme. Me and two of my teammates built a web-based recommender system. The recommendation was predicted by a machine learning algorithm in the backend using Python. The front end is

written using js (React). We won the competition by

- very creative idea
- solves or address a real world problem (or at least something looks useful in the industry)
- decent behaving model
- has an interface and interactive map in the front end
- great teamwork (model + backend API + frontend visualization)

To sum up, **in offline competitions, what matters the most is not the accuracy of the model but what can you do with what you designed; whereas in Kaggle, the accuracy is the only purpose.** This makes sense because most of the competitions on Kaggle are hosted by well known companies who are seeking for help from data scientists all around the world to improve their product (or service), so what they care the most are the accuracy, the result. In offline competitions (well, both of the competitions I participated were only open to students or recent grads), the idea is the most important, and then it comes the model since it is all about learning rather than winning.

Tahsin Mayeesha, newbie contestant
Written 33w ago

I can't remark on 'how to become really good at Kaggle', considering I'm not 'really good' at Kaggle.But I really admire people who are good at it like Triskelion | Kaggle. Here goes my remarks.

- Competitive machine learning is still at it's infancy right now and that's why people find Kaggle like competitions hard. There are only few resources for learning predictive modelling and it's not even taught in even most graduate schools. I believe after a while it'll be taught in high schools too like contest programming is available for high school students in the form of IOI or other school level contests.
- Kaggle is the type of place where it's super easy to get started and submit your first model, but it's really hard to get as good as other masters. There's no 'track' there

where a person can get step by step learning, each problem is from different domain and therefore feature selection is pretty hard, the really good competitors have been doing it for a while and they are just really good. I've seen top competitors just joining before the end of the competition and just take the top places. Not to mention it's valid to participate as teams in most contests so it's more or less hard for any single individual to beat them, specially when the difference between the first and second place can be 0.00001 difference in accuracy.

- But, Kaggle forum people are super helpful in general. I've not participated there for a while and I've decided to get better in theory first before participating further and I had huge knowledge gap first, but it's shrinking pretty fast now.The discussion forums are really active and people are sharing tutorials there which helps. Kaggle also recently launched https://www.kaggle.com/scripts which is obviously helpful for competitors.

- To get started in Kaggle, one basically needs proficiency in python, R or Julia and learn how to run basic techniques like linear/logistic regression, random forest, gradient boosting,neural networks etc to be sure that we can at least submit the output. R is easier to get started with and Coursera's John Hopkins specialization focuses on R. Python is harder to get started with IMO but there are tools like Scikit learn which makes it super easy. There are courses that focus on python too such as Harvard's CS109. I was using R first, but I've given up on that and I've decided to mostly use python for everything from now on and I've interest in Julia too because the style kind of seemed cool.

- But to do well in Kaggle , that seems more or less hard to me. Not hard in the sense "OMG IT"S IMPOOSSSIBLE FOR MORTALS!!!!!!!" (some idiots tend to do that) style hard, rather hard in the sense of it does require passion, persistence, breadth of knowledge and good data intuition. Above all, just giving 'time' is the best way to get better in practically anything.

- New contestants generally lack in areas like feature selection, feature preprocessing, feature engineering, model selection, tuning the model parameters, ensembling the models, dealing with missing data, and creating good submissions after the ensembling gets done. Needless to say, I'm new too, so I'm also lacking in all the areas. New contestants also lack knowledge of super cool tools like Vowpal Wabbit, XGboost and numerous other libraries that the expert contestants know about.

Everyone can use the tools like R with help of numerous packages and tools like Scikit Learn has great documentation too, so the feature + model selection can make or break the score. I'm not even going to comment on avoiding overfitting. That seemed like the hardest thing. Second hardest thing is to learn how to tune the parameters. This requires good understanding of the respective algorithm/technique, good understanding of the actual given data set and good intuition about what will happen if some action is taken.

- One demotivating factor of Kaggle is that they just want the predictions, they don't want production level code or they don't even make companies fight against each other to see who can achieve best results in a real market(that would be super cool IMO). Basically, since everyone knows that it's rare that their code will probably not be reused later, people focus more on beating each other instead of creating good scripts that can be reused by other people. Also, submitting over and over is more or less useless. The first five to six models are probably going to be the best reflection of the contestant's skill set, after that it's either good or bad luck.
- It's hard to get the sense of timing. How long should you try to work on one competitions? Should you switch or should you keep trying to just get a little bit more accuracy? Should you just study more books and algorithms or just participate in more contests? Should you try to team up with other people or should you just learn alone? Those depend on the particular person who's actually competing so there's not much point to think about what other people do and it's better to take an analytical approach.

Hope this helps! The most logical course of action is to just get started.

How can I become data scientist + kaggle?
1,512 Views

Is a PhD reqt becoming increasingly irrelevant for Data Science?
12,308 Views

What are some good toy problems in data science?
228,630 Views

Would the following data science certificate programs be good preparation for Kaggle competitions, and a career switch into data science jobs?...
22,675 Views

Can getting a middling score in kaggle competitions still help one get a data science position somewhere?
3,174 Views

Is there a data science competition website (Kaggle-like) in Japan?
803 Views

How do I use my Kaggle competition solution to get a data science job?
1,275 Views

Do highly ranked Kaggle users make for good data science hires?
6,041 Views

How similar are Kaggle competitions to what data scientists do?
41,820 Views

Will completing Data Science specialization on Coursera give good background to start competing on Kaggle?
4,133 Views

Is there a website grouping data science competitions (like Kaggle and others) by theme (computer vision, NLP, recommender systems)?
6,874 Views

How different is data in Kaggle competitions from real data?
7,321 Views

Should I mention taking part to a Kaggle competition in an interview for a position in data science?
848 Views

What are some good resources for preparing for Kaggle competitions?
6,160 Views

Top Stories