291 WANT ANSWERS

Latest activity: Mon

QUESTION TOPICS

Kaggle (company)

Machine Learning

Edit Topics

SHARE QUESTION

Twitter

Facebook

QUESTION STATS

Views	26,978
Want Answers	291
Edits	

What do top Kaggle competitors focus on?

To do well in a competition, clearly many aspects are important.
But, what do you think helped you (or a top competitor you know) do better than others?

At a high level, I'm curious if you approach a competition with one-size-fits-all approach or try to develop a particular insight.

More specifically, I'm interested in your thoughts on utility of the following:
- bringing more data to the party
- feature engineering
- data prep/normalization
- superior knowledge of subtleties of ML
- know-how of tackling predictive modeling problems
- proprietary tools
- picking a problem you had particular insight about
- picking less competitive competition(s)
- forming a good team
- seek specialist's advice
- persistence
- luck

Edit

Want Answers 291 Comment Share 6 Downvote

4 ANSWERS ASK TO ANSWER

Cari Kaufman
Edit Biography • Make Anonymous

Write your answer, or answer later

Chris Hefele
118 upvotes by Kornél Csernai, Leo Polovets, Olivier Grisel, (more)

I've been fortunate enough to have landed in the top few spots of a few Kaggle competitions, and I've been surprised by what factors that do not matter much, as well as those that do. Here are my observations & views:

Persistence & Enthusiasm: First, I agree with Vik's comments that persistence plays a big role. These contests are both addictive and frustrating. Many times I've coded up what I thought was a great idea, only to be dumbfounded about why it ended up *hurting* my score. Converging on a good model takes time, so to keep going, it really helps to be enthusiastic about the problem, or about learning something new, or about working with your teammates.

Know your data (aka Exploratory Data Analysis): Each dataset is unique, so I think it's really, really important to roll around in the data & get to know its quirks & inconsistencies inside out. You should do this *before* you even think of throwing the data into some ML algorithm or doing pre-processing. Write code that generates thousands of graphs that you can scan, so you get a good "feel" for what the variables in the raw dataset are doing. Sometimes quirks can provide insights, while other times they provide opportunities. Large

Set Up Your Account

~~Visit your feed~~

Follow 10 more topics

Find your friends on Quora

Upvote 5 good answers

Ask your first question

Add info about what you know

Answer a question

RELATED QUESTIONS

What tools do top Kaggle participants use?

How has the Heritage Health Prize and Kaggle successfully minimised public concerns around patient privacy risk?

What tools was Kaggle developed with?

How do TopCoder and Kaggle compare?

Only 3 Indians are found in the top 100 of Kaggle rankings. What are Indians missing?

Is it worth hosting a Kaggle money competition?

Do people who do really well at Topcoder also do well on Kaggle?

What are some alternatives to Kaggle?

Why are the number of public Kaggle competitions with cash prizes decreasing?

How do I start doing Kaggle competitions?

More Related Questions

outliers have burned me more than once, so one lesson I've learned is to always have a sensible strategy for finding and dealing with them, rather than ignoring them.

Feature Engineering: I think this is one of the most important skills, if not the most important one. The good news, though, is that it's not exclusively a highly-technical skill. A good dose of creativity and common sense can take you a long way when dreaming up new features. Also, if the contest domain is completely new to me, I usually just Google around for some papers in the field, and scan them to get a general idea of what features/factors are already known to be very important. But I don't dwell on doing *too* much research (i.e. no more than 5%-10% of the competition duration). Most of the time, I have fun generating as many crazy ideas for features as possible, and then let a features-selection or feature-prioritization algorithm prune away the ones that don't work well.

Avoiding Overfitting: When every 0.000001 matters, a crucial technical skill is avoiding overfitting your model to the data. When the leaders of a contest are separated by only a very small amount, even a little overfitting can cause you to lose many places in the final standings.

Picking a problem you had particular insight about: I've been surprised that this hasn't been particularly important. Sometimes, I've done poorly on problems that I thought I had insights on, while other times I've done well on areas completely new to me. So for me, I think there's a very loose correlation (at best) between insights/experience and success in a contest. Many Kaggle winners come from fields outside of the contest's domain, so previous insights don't seem to be mandatory. But knowing a problem area already may mean you're interested in it, which means you'll likely work harder on the contest & be more successful.

On using external data: It's soooo tempting to use external data, but generally, I haven't found using them useful. One really interesting blog post about this topic came up during the Netflix Prize, where members of the winning team commented that external data helped their weaker models at first, but as they progressed, it didn't help them at all. http://pragmatictheory.blogspot....

Superior knowledge of the subtleties of ML: Now-a-days, with everyone using libraries of ML routines, you can use techniques you don't fully understand & do fairly well. This is both good and bad. Knowing the subtleties & details of each algorithm is helpful for tuning, but one could conceivably stumble your way towards the best algorithm or parameters by just trial and error. I'd say most most competitors probably have access to the same set of 'standard' machine-learning algorithms that are provided in ML libraries, so this really doesn't provide a huge competitive advantage.

Proprietary tools: I don't think this is an advantage at all, given all the open-source tools and libraries that are out there. I've never used a proprietary tool for these contests (well, actually that's not true, I sometimes make charts in Excel...)

Updated 19 Jul, 2012. 10,585 views.

Upvote 118 Downvote Comment 1 Share 4

Vik Paruchuri, Machine Learning Person
169 upvotes by Zachary Reiss-Davis, Changqi Cai, Leo Polovets, (more)

Thanks for asking me to answer this question (I guess at least one person thinks I am a top Kaggle competitor!). Anyone please feel free to correct anything inaccurate or off base here.

This is a tough question to answer, because much like any competitive endeavor, any given Kaggle competition requires a unique blend of skills and several different factors. In some competitions, luck plays a large part. In others, an element that you had not considered at all will play a large part.

For example, I was first and/or second for most of the time that the Personality Prediction Competition [1] ran, but I ended up 18th, due to overfitting in the feature selection stage, something that I has never encountered before with the method I used. A good post on some of the seemingly semi-random shifts that happen at the end of a competition can be found on the Kaggle blog [2].

Persistence, Persistence, and more Persistence

You have outlined some key factors to success. Not all of them are applicable to all competitions, but finding the one that does apply is key. In this, persistence is very important. It is easy to become discouraged when you don't get into the top 5 right away, but it is definitely worth it to keep trying. In one competition, I think that I literally tried every single published method on a topic.

In my first ever Kaggle competition, the Photo Quality Prediction [3] competition, I ended up in 50th place, and had no idea what the top competitors had done differently from me.

I managed to learn from this experience, however, and did much better in the my second competition, the Algorithmic Trading Challenge [4].

What changed the result from the Photo Quality competition to the Algorithmic Trading competition was learning and persistence. I did not really spend much time on the former competition, and it showed in the results.

Expect to make many bad submissions that do not score well. You should absolutely be reading as much relevant literature (and blog posts, etc), as you can while the competition is running. As long as you learn something new that you can apply to the competition later, or you learn something from your failed submission (maybe that a particular algorithm or approach is ill-suited to the data), you are on the right track.

This persistence needs to come from within, though. In order to make yourself willing to do this, you have to ask yourself why you are engaging in a particular competition. Do you want to learn? Do you want to gain opportunities by placing highly? Do you just want to prove yourself? The monetary reward in most Kaggle competitions is not enough to motivate a significant time investment, so unless you clearly know what you want and how to motivate yourself, it can be tough to keep trying. Does rank matter to you? If not, you have the luxury of learning about interesting things that may or may not impact score, but you don't if you are trying for first place.

The Rest of the Factors

Now that I have addressed what I think is in the single most important factor (persistence), I will address the rest of your question:

1. The most important data-related factor (to me) is how you prepare the data,

and what features you engineer. Algorithm selection is important, but much less so. I haven't really seen the use of any proprietary tools among top competitors, although a couple of first place finishers have used open-source tools that they coded/maintain.

2. I have had poor results with external data, typically. Unless you notice someone on the leaderboard who has a huge amount of separation from the rest of the pack (or a group that has separation), it is unlikely that anyone has found "killer" external data. That said, you should try to use all the data you are given, and there are often innovative ways to use what you are given to generate larger training sets. An example is the Benchmark Bond Competition [5], where the competition hosts released two datasets because the first one could be reverse-engineered easily. Using both more than doubled the available training data (this did not help score, and I did not use it in the final model, but it it an illustration of the point).

3. Initial domain-specific knowledge can be helpful (some bond pricing formulas, etc, helped me in the Benchmark Bond competition), but it is not critical, and what you need can generally be picked up by learning while you are competing. For example, I learned NLP methods while I competed in the Hewlett Foundation ASAP Competition. That said, you definitely need to quickly learn the relevant domain-specific elements that you don't know, or you will not really be able to compete in most competitions.

4. Picking a less competitive competition can definitely be useful at first. The research competitions tend to have less competitors than the ones with large prizes. Later on, I find it useful to compete in more competitive competitions because it forces you to learn more and step outside your comfort zone.

5. Forming a good team is critical. I have been lucky enough to work with great people on two different competitions (ASAP and Bond), and I learned a lot from them. People tend to be split into those that almost always work alone and those that almost always team up, but it is useful to try to do both. You can learn a lot from working in a team, but working on your own can make you learn things that you might otherwise rely on a teammate for.

6. Luck plays a part as well. In some competitions, .001% separates 3rd and 4th place, for example. At that point, its hard to say whose approach is "better", but only one is generally recognized as a winner. A fact of Kaggle, I suppose.

7. The great thing about machine learning is that you can apply similar techniques to almost any problem. I don't think that you need to pick problems that you have a particular insight about or particular knowledge about, because frankly, it's more interesting to do something new and learn about it as you go along. Even if you have a great insight on day one, others will likely think of it, but they may do so on day 20 or day 60.

8. Don't be afraid to get a low rank. Sometimes you see an interesting competition, but think that you won't be able to spend much time on it, and may not get a decent rank. Don't worry about this. Nobody is going to judge you!

9. Every winning Kaggle entry is the combination of dozens of small insights. There is rarely one large aha moment that wins you everything. If you do all of the above, make sure you keep learning, and keep working to iterate your solution, you will do well.

Learning is Fun?

I think that the two main elements that I stressed here are persistence and learning. I think that these two concepts encapsulate my Kaggle experience nicely, and even if you don't win a competition, as long as you learned something, you spent your time wisely.

References

1. http://www.kaggle.com/c/twitter-...
2. http://blog.kaggle.com/2012/07/0...
3. http://www.kaggle.com/c/PhotoQua...
4. http://www.kaggle.com/c/Algorith...
5. http://www.kaggle.com/c/benchmar...

Updated 22 Jul, 2013. 15,072 views. Asked to answer by Leo Polovets.

Upvote 169 Downvote Comments 3+ Share 2

Dan Becker, On currently leading team in Heritage... (more)
11 upvotes by Scott Hendrickson, Raghavan Muthuregunathan, Olivier Grisel, (more)

There are a lot of different techniques, and many of them have their own tuning parameters. In my opinion, choosing the best parameters or the best off-the-shelf technique is overrated.

Instead, the keys are:
1) Feature Engineering: Extracting the data into a format that creates good predictive variable/features. Depending on the situation, it's usually better to create too many features rather than too few... most algorithms will sort out what's relevant and what isn't. Data prep and feature engineering is the biggest determinant of final rank.

2) Invest time early in the competition in building a workflow/infrastructure for working with the data.
For example, I worked hard streamlining my workflow on Amazon's EC2, so I don't need to futz with their web interface when I need to use a lot of memory. Time spent setting version control and automating repetitive tasks will pay off quickly.

3) Blending/ensembling can lead to huge improvements. Some of the blending algorithms have been discussed on the Kaggle forums. Be a little careful not to overfit when ensembling... but any ensembling you do is likely to improve your score over your best individual model.

Updated 4 Feb. 2,549 views.

Upvote 11 Downvote Comment Share

Jacob Jensen, Machine Learning, Data Visualization,... (more)
2 upvotes by Deepak Pant and Andy Chung.

From my observations of Kaggle's blog, domain-specific knowledge is lightly used in general. Normalization and feature processing along with a lot of testing to tune a classification method (random forests more often than not) yields victory.

Updated 24 Nov, 2012. 2,087 views.

Upvote 2 Downvote Comment Share

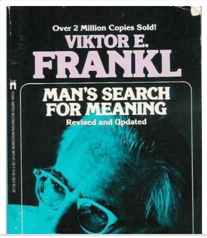
Top Stories from Your Feed

Answer written • 2012

What books are best to help people who are suffering?

Tomas Flodr, Creative Professional, Buccaneer-Scholar

39 upvotes by Nan Waldman, Dane Watts, Ryan Taylo Helsing, (more)



Viktor Frankl's Man's Search for Meaning is short (only 168 pages) but hugely influential book on the subject of human suffering. First part is deeply moving

Read In Feed

Answer written • 28 Dec

Is 32 too old to start going to the gym for a 6 pack?

Ron Blouch

4.8k upvotes by Sorna Kumar S, Anand Ramakrishnan, Samuel Liu, (more)

Read In Feed

Answer written • Sat

I want to start a quail farm integrator company that sells quail meat in Nigeria. What do I need to get started? How much money do I need?

Ade Taiwo, Individualist

1 upvote by Hamza EL Ghazi.

To get started, Do the necessary research and run a pilot farm to get real feedback on various factors from operations to sales. Or pay someone to figure it out for you. ...

Read In Feed