4

Leveraging Discussion Forums

Discussion forums are the primary means of information exchange on Kaggle. Whether it’s discussing an ongoing competition, engaging in a conversation about a dataset, or a notebook presenting a novel approach, Kagglers talk about things all the time.

In this chapter we present the Discussion forums: how they are organized, the code of conduct and the manners in which the wealth of information therein can be used. We cover the following topics:

* How forums work
* Discussion points for an example competition
* Netiquette

# How forums work

You can enter the discussion forum in several ways. The most direct way is by clicking on Discussions in the left-hand side panel:

Obraz zawierający tekst

Opis wygenerowany automatycznie

Figure 4.1: Entering the Discussion page from the main menu

The top section contains Forums, aggregations of general topics. Perusing those is useful whether you are participating in your first competition, have a suggestion to make, or just have a general question because you feel lost. Below the forums, you can find the combined view of discussions across Kaggle: mostly conversations related to competitions (which form the bulk of activity on Kaggle), but also notebooks or notable datasets. By default, they are sorted by "hotness"; in other words, those with highest participation and most activity are shown closer to the top.

The section below Forums is the one where you can find content more relevant to the dynamic nature of the field: a collection of discussions from different subsets of Kaggle, with the ability to filter on specific criteria:

Obraz zawierający tekst

Opis wygenerowany automatycznie

Figure 4.2: Discussions from across Kaggle

Depending on your interest, you can start personalizing the content by using the filters: Based on your preferences, you can filter by:

* “recency” – which allows to e.g. control the range of information you are catching up on
* “my activity” – if you need an overview of your comments / publications / views across all forums; useful if you are involved in multiple discussion simultaneously
* “admin” – provides a quick overview of announcement from Kaggle admins
* “types” – discussions can take place in the general forums, in specific competitions or around datasets
* “tags” – while not present everywhere, several discussions are tagged and this functionality allows a user to make use of that fact

Obraz zawierający tekst

Opis wygenerowany automatycznie

Figure 4.3: Available filters for discussions

The next figure shows a sample output of filtering on discussions on the tag “beginner”:

Obraz zawierający tekst

Opis wygenerowany automatycznie

Figure 4.4: Filtering discussions to those tagged “Beginner”

As an alternative, you can also focus on a specific topic; since topics like computer vision attract a lot of interest, it is probably useful to sort the topics. You can sort by Hotness, Recent Comments, Recently Posted, Most Votes, and Most Comments:

Obraz zawierający tekst

Opis wygenerowany automatycznie

Figure 4.5: Computer vision topics subset of the general discussion forum

People come to Kaggle for diverse reasons but, despite the growth in popularity of notebooks, competitions remain the primary attraction. Each Kaggle competition has its own dedicated discussion forum, which you can enter by going into the competition page and selecting Discussion.

Obraz zawierający tekst

Opis wygenerowany automatycznie

Figure 4.6: Discussion forum for a competition

It was not always the case, but these days virtually all the competitions have a FAQ topic pinned at the top of their dedicated discussion forum; starting there is a good idea for two main reasons:

* It saves you time; the most popular queries are probably addressed there.

You avoid asking redundant or duplicate questions in the remainder of the forum, making everyone’s experience better.

Like notebooks, discussion forums have an option for you to bookmark particularly relevant topics for later reference:

Obraz zawierający tekst

Opis wygenerowany automatycznie

Figure 4.7: Bookmarking a topic in a discussion forum

An overview of all your bookmarked topics can be found under your profile page:

Obraz zawierający tekst

Opis wygenerowany automatycznie

Figure 4.8: Bookmarking a topic in a discussion forum

It is a completely normal thing to feel lost in a competition at some point: you came in, tried a few ideas, got some traction on the leaderboard, and then you hit the Kaggle version of a runner's wall. This is the moment when discussion forums are the place to go.

As an example, we look at the Optiver Realized Volatility Prediction competition <https://www.kaggle.com/c/optiver-realized-volatility-prediction>, characterized by the organizers like this:

*“In the first three months of this competition, you’ll build models that predict short-term volatility for hundreds of stocks across different sectors. You will have hundreds of millions of rows of highly granular financial data at your fingertips, with which you'll design your model forecasting volatility over 10-minute periods. Your models will be evaluated against real market data collected in the three-month evaluation period after training”.*

There is quite a lot to unpack here, so we will walk over the main components of this challenge and show how they can be approached via the discussion forums. First, participation in this competition requires some level of financial knowledge – not quite experienced trader level maybe, but understanding the different manners of calculating volatility is not trivial for a layman (which most of Kagglers are in this specific matter). Luckily for the participants, the organizers were very active during the competition and provided guidance to resources intended to help newcomers to the field:

<https://www.kaggle.com/c/optiver-realized-volatility-prediction/discussion/273923>

If the entry knowledge still proves insufficient to get started, do not hesitate to figure things out in public and ask for help, like here:

<https://www.kaggle.com/c/optiver-realized-volatility-prediction/discussion/263039>

or here:

<https://www.kaggle.com/c/optiver-realized-volatility-prediction/discussion/250612>

Over time, people started developing increasingly sophisticated models to handle the problem. There is a balance to strike here: on the one hand, you might want to give something back if you have learned from veterans sharing their findings before – but on the other, you do not want to give away your (potential) advantage by publishing all the great code as a notebook. A reasonable compromise is discussing e.g. your feature ideas in a post in the forum competition – along the lines of this one:

<https://www.kaggle.com/c/optiver-realized-volatility-prediction/discussion/273915>

In recent years, more competitions are moving away from the fixed test dataset format and introduce some sort of variation: sometimes they enforce the usage of the Kaggle API (such competitions require submission from a notebook), others introduce a special timetable split into training phase and evaluation against live data. This was the case with Optiver:

*Starting after the final submission deadline there will be periodic updates to the leaderboard to reflect market data updates that will be run against selected notebooks. Updates will take place roughly every two weeks, with an adjustment to avoid the winter holidays.*

While straightforward to formulate, this setup generated a few challenges for re-training and updating the models. Should you encounter such situation, feel free to inquire – as did participants in this competition:

<https://www.kaggle.com/c/optiver-realized-volatility-prediction/discussion/249752>

Validation scheme for your trained model is always an important topic in a Kaggle competition – usually coupled with the perennial “CV vs LB” discussion. The Optiver competition was no exception to that rule:

<https://www.kaggle.com/c/optiver-realized-volatility-prediction/discussion/250650>

Unless the thread is already present – and it’s always a good idea to check, so that redundancy can be minimized – you might want to consider a related type of thread: single model performance. Sooner or later, everybody starts using ensembles of models, but they are not very efficient without good single model components. The collaborative quest for knowledge did not stop there: if you think you have found a better way of approaching the problem, it is probably a good idea to share it – either you will have done something useful for others, or you will find out why you were wrong (saving you time and effort); either way, a win – as shown for instance in this discussion

<https://www.kaggle.com/c/optiver-realized-volatility-prediction/discussion/260694>

Apart from the obvious personal benefit (you get a sort of peek into how the competitors are doing), such threads allow for information exchange in the community, facilitating the collaborative element and being helpful for beginners. An example of such a discussion can be found under <https://www.kaggle.com/c/optiver-realized-volatility-prediction/discussion/250695> .

If you have gone through the topics such as the ones listed above, there is a possibility you still find yourself wondering: am I missing anything important? Kaggle is the kind of place where it is perfectly fine to ask:

<https://www.kaggle.com/c/optiver-realized-volatility-prediction/discussion/262203>

We mentioned validation above, which always links – at least for a Kaggler – to the topic of information leakage and overfitting. Leaks are discussed extensively in Chapter 6, dedicated to designing validation schemes. Here, we touch briefly on how they are approached via discussions. Kaggle being a community of inquisitive people, if there is suspicion of leakage, somebody is likely to raise the topic.

For example, names of the files or ids of records may contain timestamps, which means they can be reverse engineered to effectively peak into the future and produce an unrealistically low error metric value. Such a situation took place in the “Two Sigma Connect competition” <https://www.kaggle.com/c/two-sigma-connect-rental-listing-inquiries/> - you can read up on the details in Kazanova’s post:

<https://www.kaggle.com/c/two-sigma-connect-rental-listing-inquiries/discussion/31870#176513>

Another example is the Airbus Ship Detection Challenge <https://www.kaggle.com/c/airbus-ship-detection> in which the participants were tasked with locating ships in satellite images. It turned out a significant proportion of test images were (random) crops of the images in the training images and matching the two was relatively straightforward: <https://www.kaggle.com/c/airbus-ship-detection/discussion/64355#377037>

A rather infamous series of competitions were the ones sponsored by Santander: of the three instances when the company organized a Kaggle contest, two involved data leakage:

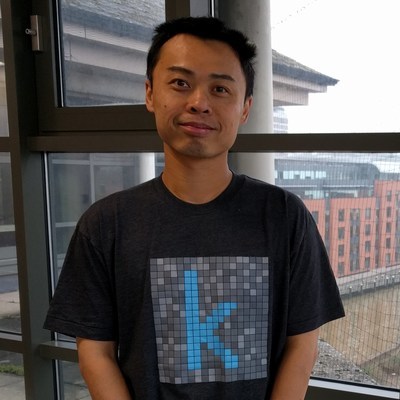
<https://www.kaggle.com/c/santander-value-prediction-challenge/discussion/61172>

What happens next varies per competition: there have been instances when Kaggle decided to reset the comp with new / cleaned up data, but also when they allowed it to continue (because they perceived the impact as minimal). An example of handling such a situation was the Red Hat competition:

<https://www.kaggle.com/c/predicting-red-hat-business-value/discussion/23788>

Although leaks in data can disturb a competition severely, the good news is that over the last 2-3 years leakage has all but disappeared from Kaggle – so with any luck, this section will be read once but not become a staple of your experience on the platform.

The topic of experience on the platform is an excellent segue into a Grandmaster interview:

Yifan Xie (https://www.kaggle.com/yifanxie) is a Discussions and Competitions Master, as well as the co-founder of Arion.ai. Here's what he had to say about competing in competitions and working with other Kagglers.

What's your favourite kind of competition and why? In terms of techniques and solving approaches, what is your speciality on Kaggle?

I don’t really have a favorite type; I like tackling problems of all kinds. In terms of techniques, I have built up a solid pipeline of machine learning modules that allow me to quickly apply typical techniques and algorithms on most data problems. I would say this is a kind of competitive advantage for me: a focus on standardizing, both in terms of work routine and technical artefacts over time. This allows for quicker iteration and in turn helps improve efficiency when conducting data experiments, which is a core component of Kaggle.

How do you approach a Kaggle competition? How different is this approach to what you do in your day-to-day work?

Over time, I have developed a specific way of managing and gathering information for most my major data endeavors. This is applicable to work, Kaggle competitions, and other side projects. Typically, I capture useful information such as bookmarks, data dictionaries, to-do lists, useful commands, and experiment results in a standardized format dedicated to each competition, and when competing in a team, I will share this info with my teammates.

Tell us about a particularly challenging competition you entered, and what insights you used to tackle the task.

For me, it has always been useful to understand the wider context of the competition; for instance, what are the social/engineering/financial processes that underpin and bring about the data we are working on? For competitions in which one can meaningfully observe individual data points, such as the *Deepfake Detection Challenge*, I would build a specific dashboard (usually using Streamlit) that allows me to check individual data points (in this case, it was pair of true and fake videos), as well as building simple stat gathering into the dashboard to allow me a better feel of the data.

Has Kaggle helped you in your career? If so, how?

I would say Kaggle is the platform that contributed the most to my current career path as a co-owner of a data science consultancy firm. It allowed me to build over several years the skillset and methodology to tackle data problems in different domains. I have both customers and colleagues who I got to know from forming teams on Kaggle competitions, and it has always served me very well as a source of knowledge, even though I am less active on it these days.

In your experience, what do inexperienced Kagglers often overlook? What do you know now that you wish you'd known when you first started?

For newcomers on Kaggle, the one error I can see is overlooking critical non-technical matters: rules on teaming, data usage, sharing of private information, usage of multiple accounts for innocuous reasons, etc. These are the types of error that could completely invalidate one’s often multi-month competition efforts.

The one thing I wish I knew at the beginning would be not to worry about the day-to-day position on the public leaderboard – it brings unnecessary pressure on oneself, and causes overfitting.

Are there any particular tools or libraries that you would recommend using for data analysis or machine learning?

The usual: scikit-learn, XGB/LGB, PyTorch, etc. The one tool I would recommend that people learn to master beyond basic usage would be NumPy, especially for more advanced ways to sort and subset information; stuff that a lazy approach via pandas makes easy, but for which a more elaborate equivalent version in NumPy would bring much better efficiency.

What's the most important thing someone should keep in mind or do when they're entering a competition?

There are four reasons to do any data science-related stuff, in my book: for profit, for knowledge, for fun, and for good. Kaggle for me is always a great source of knowledge and very often a great memory to draw upon, so my recommendation would always be to remind oneself that ranking is temporary, but knowledge/memory are permanent :)

Do you use other competition platforms? How do they compare to Kaggle?

I am a very active participant on Numerai. For me, based on my four reasons to do data science, it is more for profit, as they provide a payout via their cryptocurrency. It is more of a solitary effort, as there is not really an advantage to teaming; they don’t encourage or forbid it, but it is just that more human resources don’t always equate to better profit on a trading competition platform like Numerai.

Numerai for me is a more sustainable activity than Kaggle during busy periods of my working calendar, because the training data is usually unchanged at each round, and I can productionize to a high degree to automate the prediction and submission once the initial models are built.

The continuity feature of Numerai also makes it better suited for people who want to build dedicated machine learning pipelines for tabular datasets.

# Netiquette

Anybody who has been online for longer than 15 minutes knows this: during a discussion, no matter how innocent the topic, there is always a possibility that people will become emotional, and a conversation will leave the civilized parts of the spectrum. Kaggle is no exception to the rule, so the community has guidelines for appropriate conduct:

<https://www.kaggle.com/community-guidelines>

Those apply not just to discussions, but also to notebooks and other forms of communication. The main points you should keep in mind when interacting on Kaggle are:

* Don’t slip into what Scott Adams calls “mind reading illusion”: Kaggle is an extremely diverse community of people from all over the world (for many of them, English is not a first language), so maintaining nuance is a massive challenge. Don’t make assumptions and try to clarify whenever possible.
* Do not make things personal; Godwin’s law is there for a reason. In particular, references to protected immutable characteristics are an absolute no-go area.
* Your mileage might vary, but the fact remains: this is not the Internet Wild West of the 1990s, when telling somebody online to RTFM was completely normal; putdowns tend to alienate people

Do not attempt to manipulate the progression system (which is a basis for awarding Kaggle medals): this aspect covers an entire spectrum of platform abuse, from explicitly asking for upvotes, to collusion, to outright cheating.

In short, do towards others as you would have them do to you – and things should work out fine.

# Summary

In this chapter we have talked about discussion forums, the primary manner of communication on the Kaggle platform. We demonstrate the forum mechanics, show an example of how discussions can be leveraged in more advanced competitions and briefly summarize the netiquette.

This concludes the first, introductory part of this book. The next chapter marks the start of a more in-depth exploration of how to maximize what you get out of Kaggle and looks at getting to grips with the huge variety of different tasks and metrics you must wrestle with in competitions.