Competitions

Find challenges for every interest level

Types of Competitions

Kaggle Competitions are designed to provide challenges for competitors at all different stages of their machine learning careers. As a result, they are very diverse, with a range of broad types.

Featured

Featured competitions are the types of competitions that Kaggle is probably best known for. These are full-scale machine learning challenges which pose difficult, generally commercially-purposed prediction problems. For example, past featured competitions have included:

- Allstate Claim Prediction Challenge - Use customers' shopping history to predict which insurance policy they purchase
- Jigsaw Toxic Comment Classification Challenge - Predict the existence and type of toxic comments on Wikipedia
- Zillow Prize - Build a machine learning algorithm that can challenge Zestimates, the Zillow real estate price estimation algorithm

Featured competitions attract some of the most formidable experts, and offer prize pools going as high as a million dollars. However, they remain accessible to anyone and

everyone. Whether you're an expert in the field or a complete novice, featured competitions are a valuable opportunity to learn skills and techniques from the very best in the field.

Research

Research competitions are another common type of competition on Kaggle. Research competitions feature problems which are more experimental than featured competition problems. For example, some past research competitions have included:

- Google Landmark Retrieval Challenge - Given an image, can you find all the same landmarks in a dataset?
- Right Whale Recognition - Identify endangered right whales in aerial photographs
- Large Scale Hierarchical Text Classification - Classify Wikipedia documents into one of ~300,000 categories

Research competitions do not usually offer prizes or points due to their experimental nature. But they offer an opportunity to work on problems which may not have a clean or easy solution and which are integral to a specific domain or area in a slightly less competitive environment.

Getting Started

Getting Started competitions are the easiest, most approachable competitions on Kaggle. These are semi-permanent competitions that are meant to be used by new users just getting their foot in the door in the field of machine learning. They offer no prizes or points. Because of their long-running nature, Getting Started competitions are perhaps the most heavily tutorialized problems in machine learning - just what a newcomer needs to get started!

- Digit Recognizer
- Titanic: Machine Learning from Disaster - Predict survival on the Titanic
- Housing Prices: Advanced Regression Techniques

Getting Started competitions have two-month rolling leaderboards. Once a submission is more than two months old, it is automatically invalidated and no longer counts towards the leaderboard. Similarly, your team will drop from the leaderboard if all its submissions are older than two months. This gives new Kagglers the opportunity to see how their scores stack up against a cohort of competitors, rather than many tens of thousands of users. If your team is removed from a Getting Started competition due to the rolling expiry and wishes to rejoin, creating a new submission will cause it to show again on the leaderboard.

Additionally, the Kaggle Learn platform has several tracks for beginners interested in free hands-on data science learning from pandas to deep learning. Lessons within a track are separated into easily digestible chunks and contain Notebook exercises for you to practise building models and new techniques. You'll learn all the skills you need to dive into Kaggle Competitions.

Playground

Playground competitions are a "for fun" type of Kaggle competition that is one step above Getting Started in difficulty. These are competitions which often provide relatively simple machine learning tasks, and are similarly targeted at newcomers or Kagglers interested in practicing a new type of problem in a lower-stakes setting. Prizes range from kudos to small cash prizes. Some examples of Playground competitions are:

- Dogs versus Cats - Create an algorithm to distinguish dogs from cats
- Leaf Classification - Can you see the random forest for the leaves?
- New York City Taxi Trip Duration - Share code and data to improve ride time predictions

Competition Formats

In addition to the different categories of competitions (e.g., “featured”), there are also a handful of different formats competitions are run in.

Simple Competitions

Simple (or “classic”) competitions are those which follow the standard Kaggle format. In a simple competition, users can access the complete datasets at the beginning of the competition, after accepting the competition’s rules. As a competitor you will download the data, build models on it locally or in Notebooks, generate a prediction file, then upload your predictions as a submission on Kaggle. By far most competitions on Kaggle follow this format.

One example of a simple competition is the Porto Seguro Safe Driver Prediction Competition.

Two-stage Competitions

In two-stage competitions the challenge is split into two parts: Stage 1 and Stage 2, with the second stage building on the results teams achieved in Stage 1. Stage 2 involves a new test dataset that is released at the start of the stage. Eligibility for Stage 2 typically requires making a submission in Stage 1. In two-stage competitions, it’s especially important to read and understand the competition’s specific rules and timeline.

One example of such a competition is the Nature Conservancy Fisheries Monitoring Competition.

Code Competitions

Some competitions are code competitions. In these competitions all submissions are made from inside of a Kaggle Notebook, and it is not possible to upload submissions to the Competition directly.

These competitions have two attractive features. The competition is more balanced, as all users have the same hardware allowances. And the winning models tend to be far simpler than the winning models in other competitions, as they must be made to run within the compute constraints imposed by the platform.

Code competitions are configured with their own unique constraints on the Notebooks you can submit. These may be restricted by characteristics like: CPU or GPU runtime, ability to use external data, and access to the internet. To learn the constraints you must adhere to, review the Requirements for that specific competition.

An example of a code competition is Quora Insincere Questions Classification.

Code Competition FAQ

I'm getting errors when submitting. What should I do?

1. Please see our page on code competition debugging for tips on understanding and preventing submission errors.

1. First you'll need to write a Notebook which reads the Competition's dataset and makes predictions on the test set. Specifically, have your Notebook write your predictions to a "submission file", which is typically a submission.csv file, though some competitions have special formats. See the competition's Evaluation page, or look for sample_submission.csv (or similar) in the Data page for more information on the expected name and format of your submission file.

2. Save a full version of your Notebook by clicking "Save Version" and selecting "Save & Run All". This saves your code, runs it, and creates a version of the code and output. Once your save finishes, navigate to the Viewer page for your new Notebook Version.

3. In the Notebook Viewer, navigate to the Output section, find and select the submission file you created, and click the "Submit" button.

Can I upload external data?

Some competitions allow external data and some do not. If a competition allows external data, you can attach it to your Notebook by adding it as a data source. If a competition does not allow external data, attaching it to your Notebook will deactivate the "Submit" button on the associated saved version.

What are the compute limits of Notebooks?

The compute limits of the Notebooks workers are subject to change. You can view the site-wide memory, CPU, runtime limits, and other limits from the editor.

Code competitions come in many shapes and sizes, and will often impose limits specific to a competition. You should view the competition description to understand if these limits are activated and what they are. Example variations include:

- Specific runtime limits
- Specific limits that apply to Notebooks using GPUs
- Internet access allowed or disallowed
- External data allowed or disallowed
- Custom package installs allowed or disallowed
- Submission file naming expectations

How do I team up in a code competition?

All the competitions setup is the same as normal competitions, except that submissions are only made through Notebooks. To team up, go to the "Team" tab and invite others.

How will winners be determined?

In some code competitions, winners will be determined by re-running selected submissions' associated Notebooks on a private test set.

In such competitions, you will create your models in Notebooks and make submissions based on the test set provided on the Data page. You will make submissions from your Notebook using the above steps and select submissions for final judging from the "My Submissions" page, in the same manner as a regular competition.

Following the competition deadline, your code will be rerun by Kaggle on a private test set that is not provided to you. Your model's score against this private test set will determine your ranking on the private leaderboard and final standing in the competition.

Joining a Competition

Kaggle runs a variety of different kinds of competitions, each featuring problems from different domains and having different difficulties. Before you start, navigate to the Competitions listing. It lists all of the currently active competitions.

Public competitions are viewable on Kaggle and appear in Kaggle search results. Depending on the privacy and access set by the host, some competitions may be unavailble for you to see or join. If a host set a competition's visibility to private, you would only see the competition's details if they shared a unique URL with you.

If you click on a specific Competition in the listing, you will go to the Competition's homepage.

The first element worth calling out is the Rules tab. This contains the rules that govern your participation in the sponsor's competition. You must accept the competition's rules before downloading the data or making any submissions. It's extremely important to read the rules before you start. This is doubly true if you are a new user. Users who do not abide by the rules may have their submissions invalidated at the end of the competition or banned from the platform. So please make sure to read and understand the rules before choosing to participate.

If anything is unclear or you have a question about participating, the competition's forums are the perfect place to ask.

The information provided in the Overview tabs will vary from Competition to Competition. Five elements which are almost always included and should be reviewed are the "Description," "Data", "Evaluation," "Timeline," & "Prizes" sections.

The **description** gives an introduction into the competition's objective and the sponsor's goal in hosting it.

The **data** tab is where you can download and learn more about the data used in the competition. You'll use a training set to train models and a test set for which you'll need to make your predictions. In most cases, the data or a subset of it is also accessible in Notebooks.

The **evaluation** section describes how to format your submission file and how your submissions will be evaluated. Each competition employs a metric that serves as the objective measure for how competitors are ranked on the leaderboard.

The **timeline** has detailed information on the competition timeline. Most Kaggle Competitions include, at a minimum, two deadlines: a rules acceptance deadline (after which point no new teams can join or merge in the competition), and a submission deadline (after which no new submissions will be accepted). It is very, very important to keep these deadlines in mind.

The **prizes** section provides a breakdown of what prizes will be awarded to the winners, if prizes are relevant. This may come in the form of monetary, swag, or other perks. In addition to prizes, competitions may also award ranking points towards the Kaggle progression system. This is shown on the Overview page.

Ready to join? If the competition allows anyone to join, you should be able to click "Join" and accept the competition's rules. If the competition has restricted access, the host will share a private link with you that allows you to join.

Once you have chosen a competition, read and accepted the rules, and made yourself aware of the competition deadlines, you are ready to submit!

Forming a Team

Everyone that competes in a Competition does so as a team. A team is a group of one or more users who collaborate on the competition. Joining a team of other users around the same level as you in machine learning is a great way to learn new things, combine your different approaches, and generally improve your overall score.

It's important to keep in mind that team size does not affect the limit on how many submissions you may make to a competition per day: whether you are a team of one or a team of five, you will have the same daily submission limit.

When you accept the rules and join a Competition, you automatically do so as part of a new team consisting solely of yourself. You can then adjust your team settings in various ways by visiting the "Team" tab on the Competition page:

You can perform a number of different team-related actions on this tab.

Types of Team Memberships

There are two team membership statuses. One person serves as the Team Leader. They are the primary point of contact when we need to communicate with a team, and also have some additional team modification privileges (to be discussed shortly). Every other person in the team is a Member.

If you are the Team Leader you will see a box next to every other team member's name on the Team page that says "Make Leader". You may click on this at any time to designate someone else on your team the Team Leader.

Changing your Team Name

The team name is distinct from the names of its members, even if the team only consists of a single person (yourself). You can always change your team name to something custom, and other users will see that custom name when they visit the competition leaderboard. Most teams customize their names!

Anyone in the team can modify the team name by visiting the Team tab.

Merging Teams

You may invite another team to your team or, reciprocally, accept a merge request from another team. If you propose a merger, the merger can be accepted or rejected by the Team Leader of the other team. If you are proposed a merger, the Team Leader may choose to accept or reject it.

There are some limits on when you can merge teams:

- Most competitions have a team merger deadline: a point in time by which all teams must be finalized. No mergers may occur after this date
- Some competitions specify a maximum team size; you will not be able to merge teams whose cumulative number of members exceeds this cap
- You will not be able to merge teams whose combined daily submission count exceeds the total submission limit to that date (daily limit x number of days).

All of this can be managed through the Team tab.

Disbanding a Team

Choose your teammates wisely as only teams that have not made any submissions can be disbanded. This can be done through the Team tab

Making a Submission

You will need to submit your model predictions in order to receive a score and a leaderboard position in a Competition. How you go about doing so depends on the format of the competition.

Either way, remember that your team is limited to a certain number of submissions per day. This number is five, on average, but varies from competition to competition.

Leaderboard

One of the most important aspects of Kaggle Competitions is the Leaderboard. The Competition leaderboard has two parts.

The public leaderboard provides publicly visible submission scores based on a representative sample of the test data. This leaderboard is visible throughout the competition.

The private leaderboard, by contrast, tracks model performance using the remainder of the test data. The private leaderboard thus has final say on whose models are best, and hence,

who the winners and losers of the Competition will be. Which subset of data is calculated on the private leaderboard or a submission's performance on the private leaderboard is not released to users until the competition has been closed.

Many users watch the public leaderboard closely, as breakthroughs in the competition are announced by score gains in the leaderboard. These jumps in turn motivate other teams working on the competition in search of those advancements. But it's important to keep the public leaderboard in perspective. It's very easy to overfit a model, creating something that performs very well on the public leaderboard, but very badly on the private one. This is called overfitting.

In the event of an exact score tie, the tiebreaker is the team which submitted earlier. Kaggle always uses full precision when determining rankings, not just the truncated precision shown on the Leaderboard.

Submitting Predictions

Submitting by Uploading a File

For most competitions, submitting predictions means uploading a set of predictions (known as a "submission file") to Kaggle.

Any competition which supports this submission style will have "Submit Predictions" and "My Submissions" buttons in the Competition homepage header.

To submit a new prediction use the Submit Prediction button. This will open a modal that will allow you to upload your submission file. We will attempt to score this file, then add it to My Submissions once it is done being processed.

Note that to count, your submission must first pass processing. If your submission fails during the processing step, it will not be counted and not receive a score; nor will it count against your daily submission limit. If you encounter problems with your submission file, your best course of action is to ask for advice on the Competition's discussion forum.

If you click on the My Submissions tab you will see a list of every submission you have ever made to this competition. You may also use this tab to select which submission file(s) to submit for scoring before the Competition closes. Your final score and placement at the end of the competition will be whichever selected submission performed best on the private leaderboard. If you do not select submission(s) to be scored before the competition closes, the platform will automatically select those which performed the highest on the public leaderboard, unless otherwise communicated in the competition.

Submitting by Uploading from a Notebook

In addition to our usual Competitions, Kaggle may also allow competition submissions from Kaggle Notebooks. Notebooks are an interactive in-browser code editing environment; to learn more about them, see the documentation sections on Notebooks.

To build a model, start by initializing a new Notebook with the Competition Dataset as a data source. This is easily done by going to the "Notebooks" tab within a competition's page and then clicking "New Notebook." That competition's dataset will automatically be used as the data source. New Notebooks will default as private but can be toggled to public or shared with individual users (for example, others on your team).

Build your model and test its performance using the interactive editor. Once you are happy with your model, use it to generate a submission file within the Notebook, and write that submission file to disk in the default working directory (/kaggle/working). Then click "Save Version" and select "Save & Run All" to build a new Notebook version using your code.

Once the new Notebook Version is done (it must run top-to-bottom within the Notebooks platform constraints), navigate to the Notebook Viewer page to see the execution results, then find and select your submission file in the Output section, and you should see a “Submit” button to submit it to the Competition.

Leakage

What is Leakage?

Data Leakage is the presence of unexpected additional information in the training data, allowing a model or machine learning algorithm to make unrealistically good predictions.

Leakage is a pervasive challenge in applied machine learning, causing models to over-represent their generalization error and often rendering them useless in the real world. It can be caused by human or mechanical error, and can be intentional or unintentional in both cases.

Some types of data leakage include:

- Leaking test data into the training data
- Leaking the correct prediction or ground truth into the test data
- Leaking of information from the future into the past
- Retaining proxies for removed variables a model is restricted from knowing
- Reversing of intentional obfuscation, randomization or anonymization
- Inclusion of data not present in the model’s operational environment
- Distorting information from samples outside of scope of the model’s intended use

- Any of the above present in third party data joined to the training set

Examples

One concrete example we've seen occurred in a dataset used to predict whether a patient had prostate cancer. Hidden among hundreds of variables in the training data was a variable named PROSSURG. It turned out this represented whether the patient had received prostate surgery, an incredibly predictive but out-of-scope value.

The resulting model was highly predictive of whether the patient had prostate cancer but was useless for making predictions on new patients.

This is an extreme example - many more instances of leakage occur in subtle and hard-to-detect ways. An early Kaggle competition, Link Prediction for Social Networks, makes a good case study in this.

There was a sampling error in the script that created that dataset for the competition: a > sign instead of a >= sign meant that, when a candidate edge pair had a certain property, the edge pair was guaranteed to be true. A team exploited this leakage to take second in the competition.

Furthermore, the winning team won not by using the best machine-learned model, but by scraping the underlying true social network and then defeated anonymization of the nodes with a very clever methodology.

Outside of Kaggle, we've heard war stories of models with leakage running in production systems for years before the bugs in the data creation or model training scripts were detected.

Leakage in Competitions

Leakage is especially challenging in machine learning competitions. In normal situations, leaked information is typically only used accidentally. But in competitions, participants often find and intentionally exploit leakage where it is present.

Participants may also leverage external data sources to provide more information on the ground truth. In fact, “the concept of identifying and harnessing leakage has been openly addressed as one of three key aspects for winning data mining competitions” (source paper).

Identifying leakage beforehand and correcting for it is an important part of improving the definition of a machine learning problem. Many forms of leakage are subtle and are best detected by trying to extract features and train state-of-the-art models on the problem. This means that there are no guarantees that competitions will launch free of leakage, especially for Research competitions (which have minimal checks on the underlying data prior to launch).

When leakage is found in a competition, there are many ways that we can address it. These may include:

- Let the competition continue as is (especially if the leakage only has a small impact)
- Remove the leakage from the set and relaunch the competition
- Generate a new test set that does not have the leakage present

Updating the competitions isn’t possible in all cases. It would be better for the competition, the participants, and the hosts if leakage became public knowledge when it was discovered. This would help remove leakage as a competitive advantage and give the host more flexibility in addressing the issue.

Resources for Getting Started

Getting Started

- The Getting Started Competitions are specifically targeted at new users getting their feet wet with Kaggle and/or machine learning:

- Binary classification: Titanic: Machine Learning from Disaster
- Regression: House Prices: Advanced Regression Techniques
    - The Kaggle Learn platform has several tracks for beginners interested in free hands-on data science learning from pandas to deep learning. Lessons within a track are separated into easily digestible chunks and contain Notebook exercises for you to practise building models and new techniques hands-on. It is a great way to start deep diving into data science and quickly get familiar with the field!
    - What Kaggle has learned from almost 2MM machine learning models on Youtube. This data.bythebay.io talk by Kaggle founder Anthony Goldbloom lays out what Kaggle competitions are all about.
    - How to (almost) win at Kaggle on Youtube. In this talk competitor Kiri Nichols summarizes the appeal of Competitions as a data science learner.

Discussion

- General Discussion: There are six general site Discussion Forums:
    - Kaggle Forum: Events and topics specific to the Kaggle community
    - Getting Started: The first stop for questions and discussion for new Kagglers
    - Product Feedback: Tell us what you love, hate, or wish for
    - Questions & Answers: Technical advice from other data scientists

    - Datasets: Requests for and discussion of open data
    - Learn: Questions, answers, and requests related to Kaggle Learn courses
- Competition Discussion Forums: No matter the competition you are participating in, you can count on plenty of active community members making posts to the forums. If you get stuck on a particular aspect of the problem, Discussions are a great place to ask questions.
- Competition Notebooks: Similar to Discussions, Notebooks shared within a competition are an excellent source of Exploratory Data Analyses (EDAs) & basic starter models which can be forked and built upon for applied learning.
- The Kaggle Noobs Slack channel: This Slack channel is a popular watering hole for general banter among Kaggle ML practitioners from Novice to Grandmaster.

Techniques

- Public, reproducible code examples in Notebooks are a great way to learn and put to practice new techniques. Search for techniques in Notebooks by tag using the search syntax `tag:classification`. Fork Notebooks to make a copy of the code to modify and experiment with.
- The No Free Hunch blog. No Free Hunch is a great way of keeping up with goings-on on Kaggle. Many past Competitions winners have been interviewed about and presented their winning models on No Free Hunch. Here are some examples of past winner's interviews:
    - NOAA Right Whale Identification
    - Instacart Market Basket Analysis, Winner's Interview: 2nd place, Kazuki Onodera
    - Two Sigma Financial Modeling Code Competition
- Various tutorials have been published on No Free Hunch:
    - An Intuitive Introduction to Generative Adversarial Networks
    - Introduction To Neural Networks

    - A Kaggle Master Explains Gradient Boosting
    - A Kaggler's Guide to Model Stacking in Practice
- Marios Michailidis: How to become a Kaggle #1: An introduction to model stacking: In this Data Science Festival talk top Kaggler Marios Michailidis (Kasanova) explains model stacking, a key feature of winning competition models, in great detail.
- Kaggle Grandmaster Panel: A panel Q&A from H2O World 2017 featuring some top Kagglers.
- How to Win A Kaggle Competition - Learn From Top Kagglers: This Coursera course, put together by high-ranking Kagglers, going into great detail on the tools and techniques used by winning Competitions models.

Cheating

Cheating is not taken lightly on Kaggle. We monitor our compliance account (the formal channel for reporting cheaters, or appealing a removal for cheating) during competitions. We also spend a considerable amount of time at the close of each competition to review suspicious activity and remove people who have violated the rules from the leaderboard. When we believe we have sufficient evidence, we take action through removal or possibly even an account ban.

We also monitor and investigate moderation reports (plagiarism, voting rings, etc.) throughout the week, and take action as appropriate, which includes removing medals as well as full-out blocking accounts.

If you believe you have evidence that suggests a team violated competition rules, please report it to the Competitions compliance account for a thorough investigation.