

# Competitions Setup

**Create a new competition or competition metric**

## Overview

Anybody can launch a machine learning competition using Kaggle's Community Competitions platform, including educators, researchers, companies, meetup groups, hackathon hosts, or inquisitive individuals! In this guide, you will learn how to set up your own competition, step-by-step.

Before diving in, it's helpful to understand how a Kaggle competition works.

## How Kaggle competitions work

### Overview

Every competition has two things, a) a clearly defined problem that participants need to solve using a machine learning model and b) a dataset that's used both for training and evaluating the effectiveness of these models.

For example, in the Store Sales – Time Series Forecasting competition, participants must accurately predict how many of each grocery item will sell using a dataset of past product and sales information from a grocery retailer.

Once the competition starts participants can submit their predictions, Kaggle will score them for accuracy, and the team will be placed on a ranked leaderboard. The team at the top of the leaderboard at the deadline wins!

## **Datasets, Submissions & Leaderboards**

Every competition's dataset is split into two smaller datasets.

One of these smaller datasets will be given to participants to train their models, typically named `train.csv`.

The other dataset will be mostly hidden from participants and used by Kaggle for testing and scoring, named `test.csv` and `solution.csv` (`test.csv` is the same as `solution.csv` except that `test.csv` contains the feature values and `solution.csv` contains the ground truth variable(s) – participants will never, ever see `solution.csv`).

When a participant feels ready to make a submission to the competition, they will use `test.csv` to generate a prediction and upload a CSV file. Kaggle will automatically score the submission for accuracy using the hidden `solution.csv` file.

Most competitions have a maximum number of submissions that a participant can make each day and a final deadline at which point the leaderboard will be frozen.

It's conceivable that a participant could use the mechanics of a Kaggle competition to overfit a solution - which would be great for winning a competition, but not valuable for a real-world application.

To help prevent this, Kaggle has two leaderboards – the public and private leaderboard. The competition host splits the `solution.csv` dataset into two parts, using one part for the public leaderboard and another part for the private leaderboard. Participants generally will now know which samples are public vs private. The private leaderboard is kept a secret until after the competition deadline and is used as the official leaderboard for determining the final ranking.

## Create your competition

To create a new competition, click on the “Create new competition” button at the top of the Kaggle Community landing page.

Then, enter a descriptive title, subtitle and URL for your competition. Be as descriptive and to the point as possible. In our example above, the title “Store Sales - Time Series Forecasting” quickly outlines the type of data, the industry of the dataset, and the type of problem to be solved.

If you want to create a competition with more privacy, you can limit your competition's visibility and restrict who can join on this page.

**Visibility:** Competitions with their visibility set to public are viewable on Kaggle and appear in Kaggle search results. Competitions with visibility set to private are hidden and only accessible via invitation URLs from the host.

**Who Can Join:** Competitions access can be set to three levels: anyone, only people with a link and restricted email list. If you select anyone, all Kagglers can join your competition. Selecting only people with a link, will restrict access to those users you provide a special URL. Finally, restricted email list is the most private competition. Only Kagglers with accounts that match the emails or email domains you specify will be able to join. Note: if you select restricted email list, notebooks will be turned off. This provides a way to ensure that any private data that you have in a competition is not accidentally leaked through shared notebooks. You can choose to re-enable notebooks if you choose.

Review and accept our terms of service, then click “Create Competition”.

Your competition listing is now in draft mode. You can take your time to prepare the details before making the competition public.

# Prepare the dataset

## Overview

You will typically need to prepare and split your chosen dataset into four CSV files with different purposes and formatting requirements:

- `train.csv` will be given to participants to train their models. It includes the inputs and the ground truth. For example, in the grocery store competition, `train.csv` contains columns of product data and the solution columns – whether or not the product sold. Typically this is roughly 70% of the original dataset.
- `test.csv` is given to participants and includes the features of the test set so they can create a submission file with their predictions.
- `solution.csv` is always hidden from participants and used by Kaggle's platform to score submissions. The rows should correspond with those of `test.csv` and typically comprises roughly 30% of the original dataset.
- `sample_submission.csv` is a placeholder CSV file with the correct formatting, which helps participants understand the expected submission format for the competition.

It's up to you to determine how exactly you'd like to split your dataset into train and test files but it's typically best practice to ensure both train and test have the same type of data represented. Also, most people go with a 70/30 or 75/25 train/test split but it's problem and dataset dependent.

Note: this guide provides instructions for tabular data. Other problem types like image data are possible using similar steps.

## Implement a unique ID column

Before splitting the dataset, make sure that your dataset has an `Id` column with unique values. The `Id` column is how the scoring system knows which rows of a submission correspond to which rows of the solution. Make sure that the `Id` column is the very first column of your solution file.

## Prepare the `train.csv` file

Take a large chunk of your dataset, typically 70% and split it into its own dataset named `train.csv`. Be sure not to remove the ground truth column(s) because participants need that information to train their models. Save and set aside for upload later. For example:

```
train.csv
input_feature1,input_feature2,target_feature
100,52.12,1
192,203.2,1
64,-59.1,0
```

## Prepare the `test.csv` and `solution.csv` files

Take the rest of your dataset and duplicate it to create two identical files.

Then take one file and remove the ground truth column(s) and save it as `test.csv`.

Next, take the other copy and delete all columns except the unique id column and the ground truth column(s). Save it as `solution.csv`.

Your solution file needs to specify which rows will be used for the public leaderboard and which will be used for the private leaderboard. You'll need to add a `Usage` column to your solution file where each row contains one of three values: `Public`, `Private` or `Ignored`. This step is not strictly necessary for competitions that use legacy metrics.

Examples:

`test.csv`

```
id,input_feature1,input_feature2
```

```
0,93,34.82
```

```
1,104,74.3
```

```
2,89,-12.0solution.csv
```

```
id,target_feature,Usage
```

```
0,1,Public
```

```
1,0,Private
```

```
2,1,Ignored
```

## **Prepare the `sample_submission.csv` file**

Duplicate the `solution.csv` file, delete the `Usage` column, and replace all ground truth values with placeholders that have valid values. Save this as `submission.csv`. This file will be given to users as an example of how to format submissions for evaluation. For example:

```
sample_submission.csv
```

```
id,target_feature
```

```
0,0
```

```
1,0
```

```
2,0
```


## Set up scoring

Navigate to the Host tab > Evaluation Metric page in the right side navigation to set up scoring.

### Designate your scoring metric

Choose the scoring metric you'd like to use for your competition in the drop down menu, or see below for how to write your own metric in Python.

There are many ways to determine “how accurate” a submission may be. In the grocery store competition example, you may want to reward underestimates more than overestimates, or reward predictions exponentially more the closer they get to the ground truth. If you are unfamiliar with the types of common evaluation metrics used in machine learning, we'd encourage you to take a look at the details of common evaluation metrics to find the right fit.

Kaggle provides two types of metrics: Python (tagged with the  icon ) and Legacy (no icon). There are a few key differences. The source code for Legacy metrics is not publicly available and they typically have limited documentation. The setup process is also slightly different: Legacy



metrics require manually mapping every column. However, Legacy metrics do offer speed advantages in some circumstances.

When a metric is selected, your competition will be tied to the latest version of that metric. If a newer version is later published, you must manually update your competition to use it.

## **Upload the `solution.csv` file**

Click on the upload icon to upload your `solution.csv` file.

If you've chosen a Python metric, check that your solution file's format matches that expected by the metric's documentation, or just continue to testing a submission to see if it matches.

If you've chosen a Legacy metric, then after uploading the `solution.csv` file the column headers will auto populate the Solution Mapping table below. Mapping allows our metric code to understand which columns to use for calculations. Choose the correct “Expected Column” values. Note, some evaluation metrics let you score multiple columns simultaneously.

## **Upload the `sample_submission.csv` file and map the verification**

Click on the upload icon to upload your `sample_submission.csv` file.

If you've chosen a Legacy metric, then after uploading you'll again need to complete the same process of column mapping for the submission format.

## Upload data for participants

Click on the Data tab and “Upload first version” button on the bottom of your screen to upload all data that participants can access – `test.csv`, `train.csv` files and `sample_submission.csv` file. Note: you will have additional data files if creating an image/video/etc. competition. Kaggle will process your data and create a versioned dataset, which will also be made accessible via Kaggle notebooks.

## Creating a New Metric

You can implement a new metric in a Python notebook at [this link](#) or from the Host > Evaluation Metric tab on a competition. Metric notebooks can be published and shared, but currently only Kaggle staff can add metrics to the public metric listing. If you think your metric is a good candidate for general use, please make the notebook public and post in the competition hosting forum.

Before your metric executes, Kaggle automatically reads the solution and submission file into Pandas dataframes, aligns the solution and submission rows based on a provided id column, and calls a `score()` function. Your metric code needs to define this `score()` function and it must return a single float. Almost all solution files are split into a `Public` and `Private` set by way of a `Usage` column in the file. The `score()` function is called separately for each of these respective sets.

Your `score()` function must satisfy the following constraints:

- Accept the arguments `solution: pd.DataFrame`, `submission: pd.DataFrame`, `row_id_column_name: str`, in that order. You can add any other keyword arguments that you need after those three. Any additional keyword arguments are configured on a per-competition basis on the Evaluation Metric page.
- All arguments and the return value of `score` must have type annotations.
- Default argument values are encouraged but not required.
- `score()` must return a finite float.
- `score()` must have a docstring. The docstring will be shown to competition hosts on the evaluation tab after they have selected a metric. We encourage you to include at least the same sections covered in our example metric's docstring: a general description of the metric, explanations of each of the `score` arguments, references for the metric math, and examples of valid use.
- In order to prevent data leaks from the solution file, errors must specify who will see the details. Only errors raised as `ParticipantVisibleError` will be visible to all participants.
- Error messages will be truncated to 280 characters.
- The scoring runtime is limited to 30 minutes total for the `Public` and `Private` splits combined.
- Metric notebooks do not have internet access and can not use accelerators, so your `score()` function must not rely on these notebook features.

Once your code is ready, you will also need to define some metadata in the `Metric` section of the notebook sidebar. You must save this metadata separately from the rest of the notebook.

- Name: your metric will use the metric notebook's name. Save the metadata to update the name.
- Description: a short (less than 255 characters) description of the metric.
- Category: the main use of the metric, such as clustering or regression.
- Leaderboard sort order: toggle this to indicate if a higher score is better or worse.
- Pass complete submission: Advanced use only. You almost certainly only want to use this if your submission can have a different number of rows than the solution file. When enabled, your metric will receive the entire submission file for both the public and private scoring rounds. Your metric will need to manage matching the solution and submission rows using the `row_id_column_name`.

You will need to use the dedicated Save button in the Metric section of the notebook sidebar for this metadata, in addition to the Save & Validate button used to save the notebook's source code. When you save your metric, your notebook will first be committed like any other notebook, followed by a series of metric-specific validation checks. This validation step will also re-run any unit test functions and doctests that are discoverable with Pytest. We strongly encourage you to include test cases, but they are not mandatory.

If the validation step fails, your notebook code will still save, but no new metric version will be created. We recommend reviewing this example metric or metric template before you begin coding.

# Test your competition

## Sandbox Testing

Once you set up the solution and submission files you can test submissions in the submission sandbox. You will need at least one sample submission that successfully generates a score in order to launch your competition.

Verify that the scoring is working as intended (e.g. a random submission should have a random score, a perfect submission should have a perfect score, etc.). You may have to experiment to understand what is and is not allowed in submission formats, but the system should provide clear error messages in the event something is wrong with a file.

## Benchmarking a Solution (Optional)

To create a benchmark score for your participants to meet or exceed, check the box next to the submission you'd like to use as a benchmark. You'll then see that score listed as a benchmark on the leaderboard.

## Finalize your settings and descriptions

Most of the heavy lifting is now complete for the competition and it's now time to craft all the final details and settings.

First navigate to the Host tab and complete your configuration in the Basic Details, Images and Evaluation Metric pages.

Then click through the Overview, Data, and Rules tabs and make sure all text descriptions are polished and ready for participants.

You can also go to the Launch Checklist page which shows your remaining steps.

## **Score Decimals to Display**

The "Score Decimals to Display" setting on the Basic Details page controls how many decimal places are shown in the user interface. We always use full-precision scores for calculations and ranking comparisons, but it can be useful to truncate the displayed scores to make them look cleaner or to prevent leaderboard probing. For example, if participants can see full-precision scores, they could make small changes to their submission and examine the score difference to infer the ground truth of the public test set, or reverse engineer the split between public and private leaderboards.

## **Launch and invite participants**

Go to Host > Launch Checklist and confirm that all the boxes are checked green. Once they are, you're good to go! Buttons allowing you to launch the competition now or schedule launch in the future will appear – choose according to your needs.

You'll know your competition is live when it says "Competition is active."

You can invite participants to your competition by sharing the URL at the bottom of the Launch Checklist or Basic Details. This link respects the access settings you specified when creating the competition. If you selected anyone can join, this link will be the competition URL. If you selected only people with a link, anyone with this URL can participate in the competition, so make sure you share the link with the right audience. If you'd like a select group to participate, send the URL via email. If you'd like broad participation, use social media or encourage participants to invite their friends. If you selected restricted email access, the link will only work if the Kaggle's email address appears on the list of restricted emails you specified.

## **FAQs**

### **Creating Your Competition**

#### **Where can I get a dataset for my competition?**

We recommend that you source your own, since it's typically best to use data to which the participants do not have access (to minimize the temptations to cheat).

But, if you don't don't mind it being fully accessible by participants (e.g. for a purely educational competition), consider browsing Kaggle's Datasets platform. It hosts thousands of public datasets and has rich search and

filter tools to help you find something that fits your needs. Each dataset should include a data use license, which will indicate if you can use it for your competition.

### **I'm receiving [an error]. How can I resolve it?**

Start by reading through this setup guide. If you still can't resolve the issue, try asking other Community Competition hosts in the Kaggle forums.

### **I want to run the same competition again. Do I need to start from scratch?**

For now, you are not able to clone a past competition. You'll need to start setup from the beginning.

### **Who can see my competition?**

It depends on the privacy setting that you chose. Kaggle has 2 privacy settings – public and limited. Public means that your competition will be listed and discoverable on kaggle.com. Limited means that only people with the provided URL can view and join the competition.

### **Where can I find the invitation link?**

If you selected Public, you can share your competition from your browser tab – anyone can see the competition. If your competition is set to Limited privacy, visit your competition > Host > Privacy > URL for Sharing (if you've selected Limited).



## **How do I contact support?**

Unfortunately, we aren't able to provide hands-on support for setting up or troubleshooting your competition. But, if you are experiencing an issue that you believe is affecting the entire platform, please contact us. We also encourage connecting with other community competition host on Kaggle's forum.

## **Can I offer a prize for a Community Competition?**

Unfortunately, a cash prize cannot be offered without additional paperwork with Kaggle. If you'd like to run a competition with a cash prize, please reach out to our Kaggle Competitions Team, who can walk you through the necessary steps.

## **During Your Competition**

### **Can I invalidate or delete a participant's submissions?**

Yes, go to your competition and navigate to: Host > All Submissions. There you can hide specific submissions.

### **Can I upload a new solution file and rescore the competition?**

You can upload a new solution file, but you cannot rescore a competition on your own. Please upload a new solution file and contact support. An administrator can rescore your competition. Competitors' new submissions will be scored against the new solution file.

**I would like to download my participants' email addresses so I can email them for a new competition. How do I do this?**

Due to privacy regulations, you cannot currently download the email addresses of participants.

**I want to give participants more time to compete, how do I change my competition deadline?**

If the competition has already ended, you should set up a new competition, as participants will have seen the private leaderboard. If the competition is still active, you can change the deadline by going to: Your competition > Host > Settings > Deadline