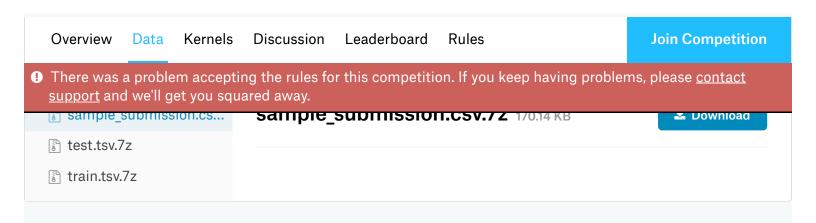


Mercari · 906 teams · 2 months to go (2 months to go until merger deadline)



Data Description

In this competition, you will predict the sale price of a listing based on information a user provides for this listing. This is a Kernels-only competition, the files in this Data section are downloadable just for your reference in Stage 1. Stage 2 files will only be available in Kernels and not available for download here.

Data fields

train.tsv, test.tsv

The files consist of a list of product listings. These files are tab-delimited.

train_id or test_id - the id of the listing

- name the title of the listing. Note that we have cleaned the data to remove text that look like prices (e.g. \$20) to avoid leakage. These removed prices are represented as [rm]
- item_condition_id the condition of the items provided by the seller
- category_name category of the listing
- brand_name
- price the price that the item was sold for. This is the target variable that you will predict. The unit is USD. This column doesn't exist in test.tsv since that is what you will predict.
- shipping 1 if shipping fee is paid by seller and 0 by buyer
- item_description the full description of the item. Note that we have cleaned the data to remove text that look like prices (e.g. \$20) to avoid leakage. These removed prices are represented as [rm]

Please note that in stage 1, all the test data will be calculated on the public leaderboard. In stage 2, we will swap the test.tsv file to the complete test dataset that includes the private leaderboard data.

sample_submission.csv

A sample submission file in the correct format.

- test_id matches the test_id column in test.tsv
- price

What will be available in the 2nd stage of the competition?

In the second stage of the competition, we will re-run your selected Kernels. The following files will be swapped with new data:

- test.tsv this will be swapped with the complete test dataset. You have ~700K rows in stage 1 test data, and will have ~3.5M rows in the stage 2 test data. The public leaderboard data remains the same. The file name will be the same (both test.tsv) to ensure that your code will run.
- sample_submission.csv similar to test.tsv, this will be changed from ~700K rows to ~3.5M rows. The file name will remain the same.

© 2017 Kaggle Inc

Our Team Terms Privacy Contact/Support





