

Midterm Assignment

Kaggle Prediction Competition

Itamar Caspi

May 10, 2019 (updated: 2020-05-06)

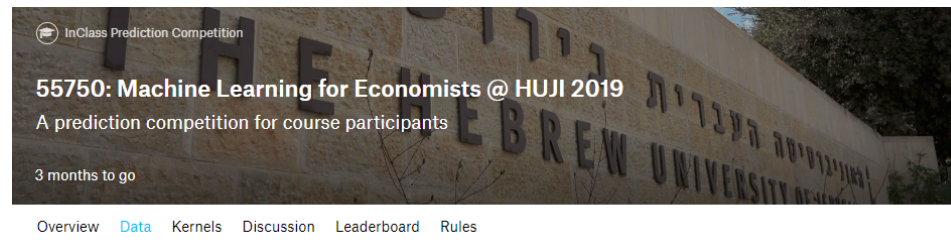
What is Kaggle?

- Kaggle is a huge data science community where machine learning practitioners around the world compete against each other.
- The datasets used in Kaggle are uploaded by public companies as well as private users.
- A "kaggler" wins if her algorithm is the most accurate on a particular data set.
- Kaggle competitions are one of the best places to practice your ML skills and learn about state-of-the-art ML method.



Introduce yourself to Kaggle

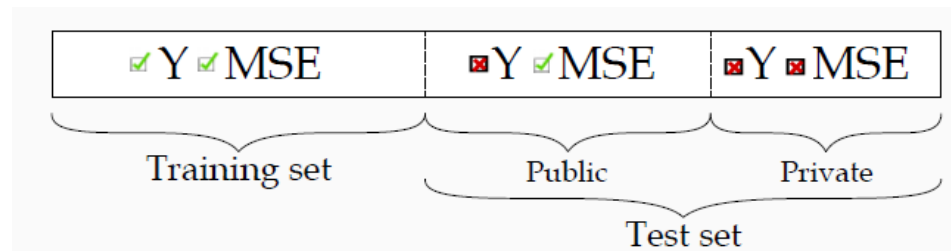
1. Visit www.kaggle.com and sign-up.
2. Go to the ml4econ course competition [webpage](#).
3. Review competition details: objectives, deadline, data, evaluation, submission rules, etc.



Kaggle competition data structure

- MSE for the public test set (30%) immediately available at submission.
- MSE for the private test set (70%) available only once the competition closes.
- The split between public and private test sets is arbitrary and unknown in advance to all competitors.

Your final ranking is based on how well you perform on the *private* test set.

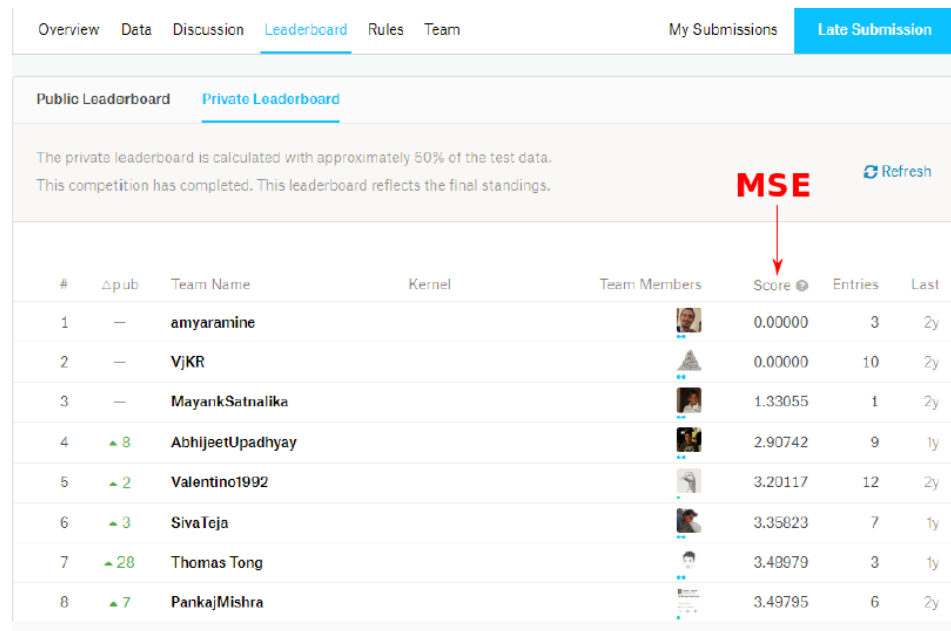


The basic Kaggle competition workflow

1. Acquire domain knowledge.
 2. Explore the data.
 3. Preprocessing (standardization, dummies, interactions, etc.).
 4. Choose a model class (asso, ridge, trees, etc.).
 5. Tune complexity (Cross validation).
 6. Submit your prediction.
 7. **Document your workflow (R Markdown)**
-

Tracking your performance

- Use the public lead-board to track your performance.
- Your ranking ("scores" column) is based on your MSE on the public test set.
- Once the competition is closed, the final ranking will be based on the MSE on the private test set.
- You can submit multiple predictions but be careful not to overfit the public test set!



The screenshot shows a competition interface with a top navigation bar containing 'Overview', 'Data', 'Discussion', 'Leaderboard' (selected), 'Rules', and 'Team'. On the right, there are links for 'My Submissions' and a blue 'Late Submission' button. Below the navigation bar, there are tabs for 'Public Leaderboard' and 'Private Leaderboard'. A message states: 'The private leaderboard is calculated with approximately 50% of the test data. This competition has completed. This leaderboard reflects the final standings.' A red arrow points to the 'Score' column header, which is labeled 'MSE' in red text. A 'Refresh' button is also visible. The table lists 8 teams with their rankings, public score changes, names, kernels, team members, scores, entries, and last update times.

#	Δ pub	Team Name	Kernel	Team Members	Score	Entries	Last
1	—	amyaramine			0.00000	3	2y
2	—	VJKR			0.00000	10	2y
3	—	MayankSatnalika			1.33055	1	2y
4	▲ 8	AbhijeetUpadhyay			2.90742	9	1y
5	▲ 2	Valentino1992			3.20117	12	2y
6	▲ 3	SivaTeja			3.35823	7	1y
7	▲ 28	Thomas Tong			3.48979	3	1y
8	▲ 7	PankajMishra			3.49795	6	2y

Getting started

Running the following code chunk will automatically download the data (train, test, and a sample submission file) you'll need for our Kaggle competition:

```
library(tidyverse)

train <- read.csv("https://raw.githubusercontent.com/ml4econ/lecture-notes-2020/master/a-1")
test <- read.csv("https://raw.githubusercontent.com/ml4econ/lecture-notes-2020/master/a-1")
sample_submission <- read.csv("https://raw.githubusercontent.com/ml4econ/lecture-notes-2020/master/sample_submission.csv")
```

NOTE: By default, a new project will be created on your desktop.

```
slides %>% end()
```

 [Source code](#)