

# Midterm Assignment

Kaggle Prediction Competition

Itamar Caspi

June 30, 2024 (updated: 2024-06-30)

# Kaggle: A Global Hub for Data Science Competitions

- **Kaggle** is a *vibrant data science community* where machine learning practitioners worldwide compete.
- Public companies and private users alike upload the datasets used in Kaggle competitions.
- A "kaggler" clinches victory by developing the most *accurate* algorithm for a specific dataset.
- Kaggle competitions serve as *platforms* for practicing ML skills and keeping abreast of state-of-the-art ML methods.



# Getting Started with Kaggle

1. **Step One:** Visit [www.kaggle.com](https://www.kaggle.com) and sign-up.
2. **Step Two:** Navigate to the ml4econ course competition (link on Moodle)
3. **Step Three:** Thoroughly review the competition details including objectives, deadlines, data, evaluation criteria, submission rules, and so on.

## Machine Learning for Economists @ HUJI 2024

Prediction competition for the ML4Econ course participants



Overview Data Code Models Discussion Leaderboard Rules Team Submissions Settings

### Overview

#### Start

22 seconds ago

#### Close

a month to go

### Description

In this competition, course participants will use a sample from the 1981 PSID (Panel Study of Income Dynamics) dataset to train and test machine learning models learned in the course. Specifically, participants are required to apply tools introduced in the first part of the course to predict log wages based on a set of individual characteristics (such as experience, education, and occupation).

### Competition Host

Itamar Caspi

### Prizes & Awards

Kudos

Does not award Points or Medals

### Participation

0 Entrants

0 Participants

0 Teams

0 Submissions

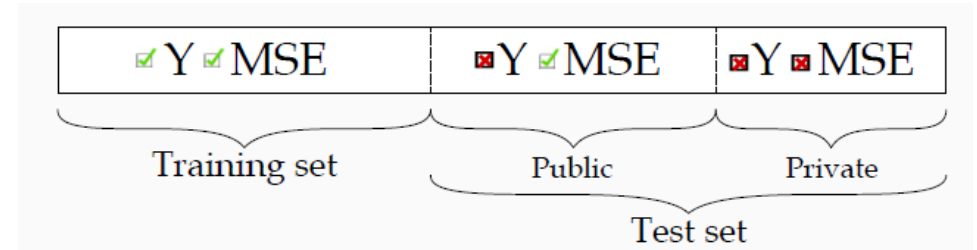
### Tags

Mean Squared Error

# Understanding the Kaggle Competition Data Structure

- *Immediate Feedback*: The Mean Squared Error (MSE) for the public test set (30%) is available immediately upon submission.
- *Delayed Feedback*: The MSE for the private test set (70%) is disclosed only after the competition closes.
- *Unpredictable Split*: The division between the public and private test sets is arbitrary and undisclosed to competitors in advance.

Remember, your *final ranking* hinges on your performance on the *private* test set.

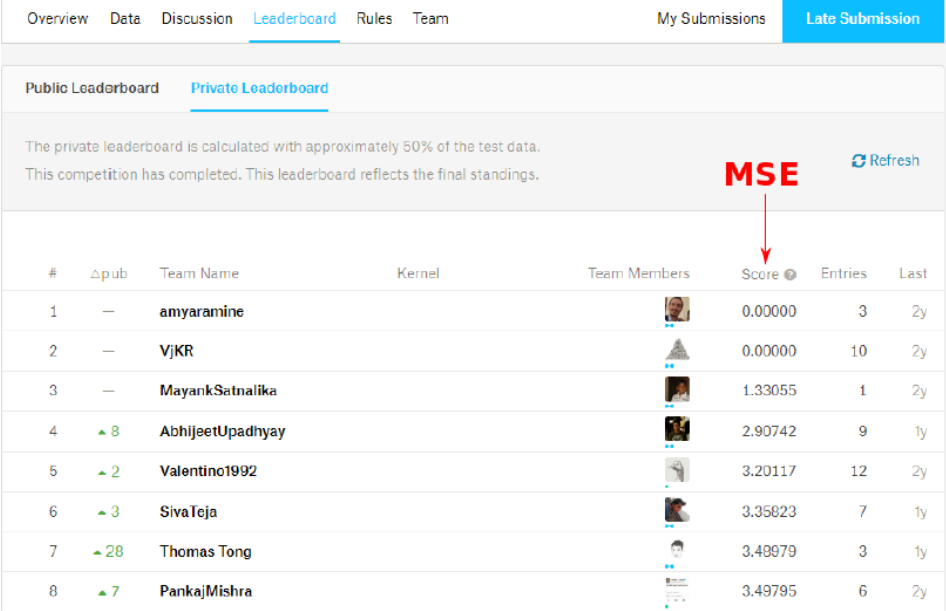


# Mastering the Basic Kaggle Competition Workflow









1. **Acquire Domain Knowledge:** Build understanding around the problem domain.
2. **Explore the Data:** Perform an initial data analysis.
3. **Preprocessing:** Employ techniques such as standardization, creating dummy variables, determining interactions, etc.
4. **Choose a Model Class:** Decide on a model class, like Lasso, Ridge, Trees, and so forth.
5. **Tune Complexity:** Use cross-validation for optimizing model parameters.
6. **Submit Your Prediction:** Forward your model's prediction for evaluation.
7. **Document Your Workflow:** Keep a well-structured record of your process using *R Markdown*.

# Monitoring Your Performance

- Leverage the public leaderboard to *track your performance*.
- Your interim ranking (reflected in the "scores" column) is determined by your MSE on the public test set.
- After the competition closes, the final ranking will hinge on the MSE on the private test set.
- While you may submit multiple predictions, exercise caution to avoid overfitting the public test set!



The screenshot shows a competition interface with tabs for Overview, Data, Discussion, Leaderboard, Rules, and Team. On the right, there are links for My Submissions and a blue button for Late Submission. Below the tabs, there are links for Public Leaderboard and Private Leaderboard. A note 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 below lists the top 8 teams with their rankings, public score changes, names, kernels, team members, scores, entries, and last update times.

#	Δpub	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

# Kickstarting Your Kaggle Journey

Executing this code chunk will automatically download the essential data for our Kaggle competition. This includes train data, test data, and a sample submission file.

```
train <- read.csv("https://raw.githubusercontent.com/ml4econ/lecture-notes-2024/master/a-  
test <- read.csv("https://raw.githubusercontent.com/ml4econ/lecture-notes-2023/master/a-l  
sample_submission <- read.csv("https://raw.githubusercontent.com/ml4econ/lecture-notes-20
```

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

 [Source code](#)