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.

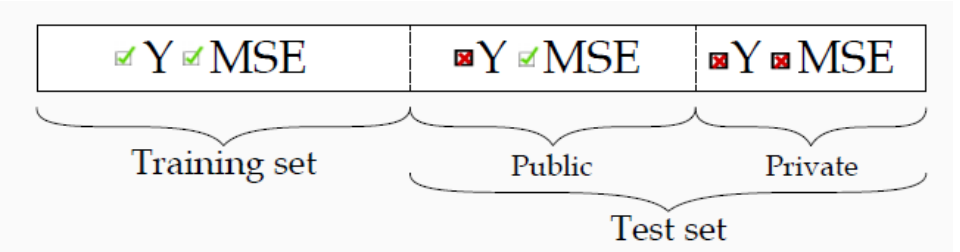
Introduce yourself to Kaggle

1. Visit www.kaggle.com and sing-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.

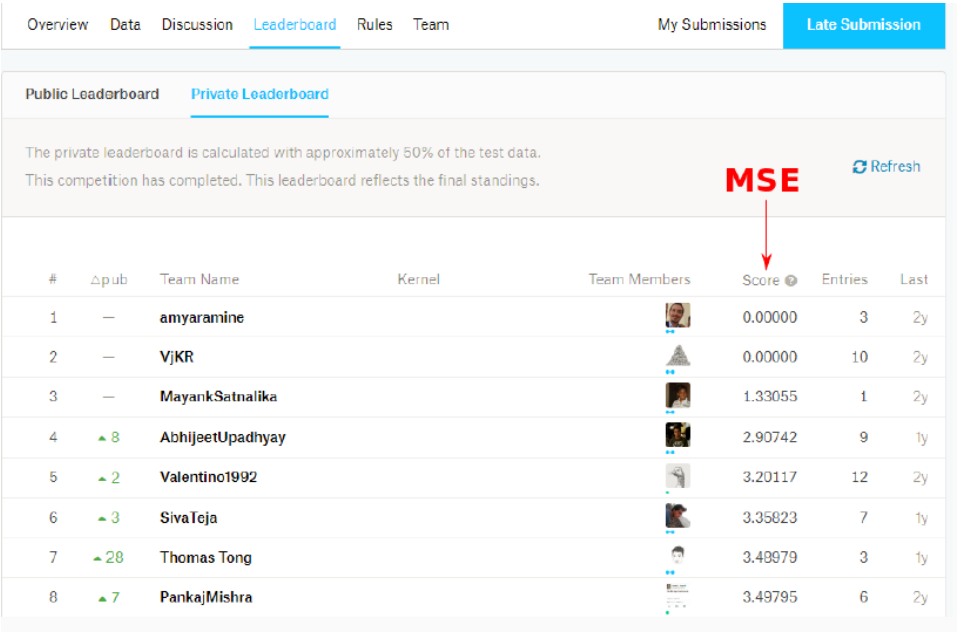


The basic Kaggle competition workflow

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.
- Your can submit multiple predictions but be careful not to overfit the public test set!



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

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

sample_submission <- read.csv("https://raw.githubusercontent.com/ml4econ/lecture-notes-2
```

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

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

Source code