WOMEN IN DATA SCIENCE

#WiDS2019

Name: Rahul Mitra

Work Experience: 4.7 years of experience in Data Science and Software development.

Current Organization: Lead Data Sciences in Envestnet | Yodlee

Previous Organizations: IBM Software Lab, TCS

Academics: M.Tech in IT from IIT Kharagpur.

Achievement: Won 3 hackathons with my team during my time in IBM

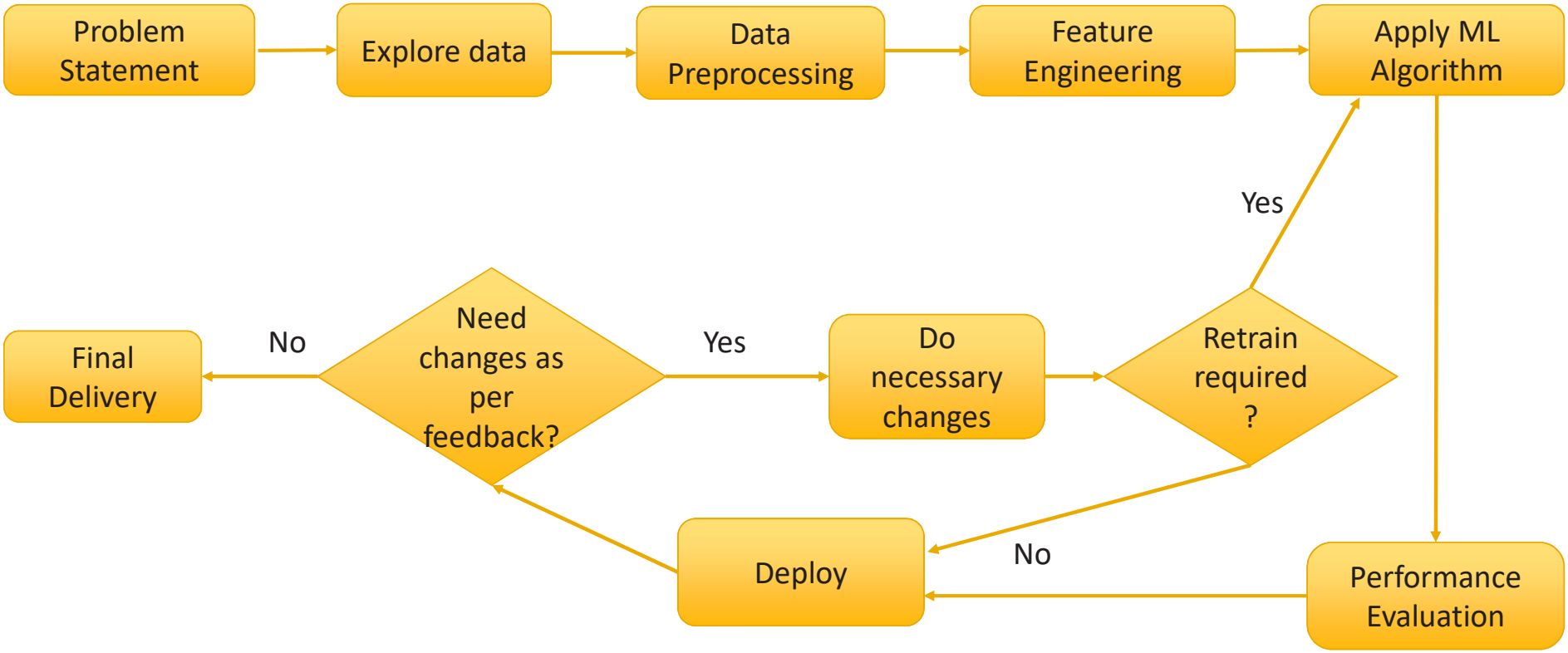
Skills: Machine Learning, Deep Learning, Java, Python

Hobby: Sports enthusiast.

Contents

- Basic steps
- Code walkthrough on CNN
- Introduction to Kaggle
- Workflow in Kaggle
- Top 5 points to succeed in Kaggle

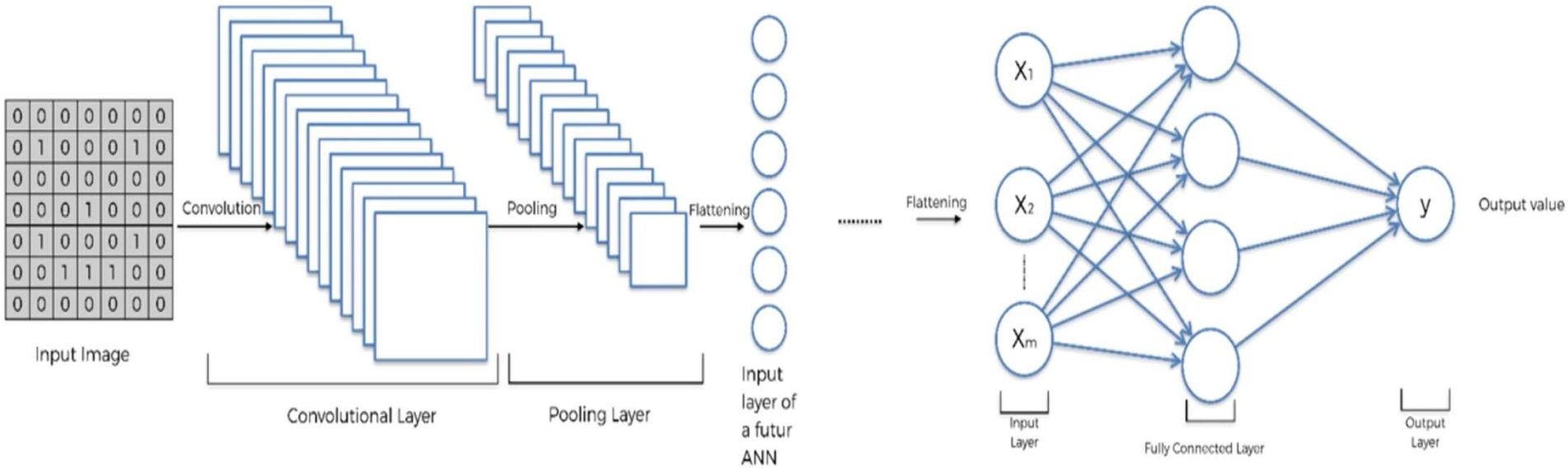
Steps to solve a problem with data science



Code walkthrough of building CNN on Fruits dataset

- Fruits dataset have images for 95 categories of fruits.
- In the demo we will see image classification between these using CNN.

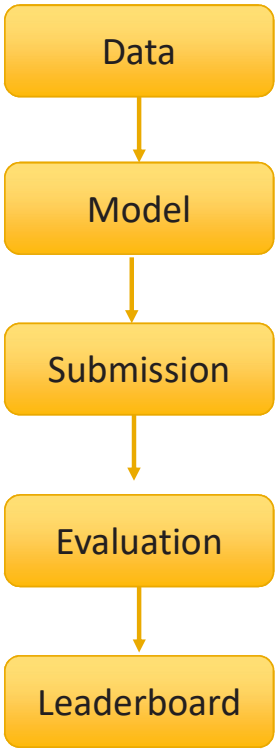
Code walkthrough of building CNN on Fruits dataset Contd..



https://github.com/rahulmitra-kgp/cnn_build_stepwise

Introduction to Kaggle

Workflow in Kaggle



Joining a competition

https://www.kaggle.com

Evaluation metric

LANL Earthquake Prediction

Can you predict upcoming laboratory earthquakes?

$50,000
Prize Money

Los Alamos National Laboratory · 696 teams · 4 months to go (4 months to go until merger deadline)

Overview | Data | Kernels | Discussion | Leaderboard | Rules | **Join Competition**

Overview

- **Description**
- **Evaluation**
- **Timeline**
- **Prizes**
- **Additional Information**

Submissions are evaluated using the mean absolute error between the predicted time remaining before the next lab earthquake and the act remaining time.

Submission File

For each `seg_id` in the test set folder, you must predict `time_to_failure`, which is the remaining time before the next lab earthquake. The file should contain a header and have the following format:

```
seg_id,time_to_failure
seg_00030f,0
seg_0012b5,0
seg_00184e,0
...
```

Competition timeline

LANL Earthquake Prediction

Can you predict upcoming laboratory earthquakes?

$50,000
Prize Money

Los Alamos National Laboratory · 696 teams · 4 months to go (4 months to go until merger deadline)

Overview | Data | Kernels | Discussion | Leaderboard | Rules | **Join Competition**

Overview

Description

Evaluation

Timeline

Prizes

Additional Information

- **May 27, 2019** - Entry deadline. You must accept the competition rules before this date in order to compete.
- **May 27, 2019** - Team Merger deadline. This is the last day participants may join or merge teams.
- **May 27, 2019** - External Data Disclosure deadline. All external data used in the competition must be disclosed in the forums by this date.
- **June 3, 2019** - Final submission deadline.

All deadlines are at 11:59 PM UTC on the corresponding day unless otherwise noted. The competition organizers reserve the right to update the contest timeline if they deem it necessary.

Dataset for the competition

Can you predict upcoming laboratory earthquakes?

Prize Money

Los Alamos National Laboratory · 696 teams · 4 months to go (4 months to go until merger deadline)

Overview | Data | Kernels | Discussion | Leaderboard | Rules | Join Competition

Data Description

The training data is a single, continuous segment of experimental data. The test data consists of a folder containing many small segments. The data *within* each test file is continuous, but the test files do not represent a continuous segment of the experiment; thus, the predictions cannot be assumed to follow the same regular pattern seen in the training file.

For each `seg_id` in the test folder, you should predict a *single* `time_to_failure` corresponding to the time between the *last row of the segment* and the next laboratory earthquake.

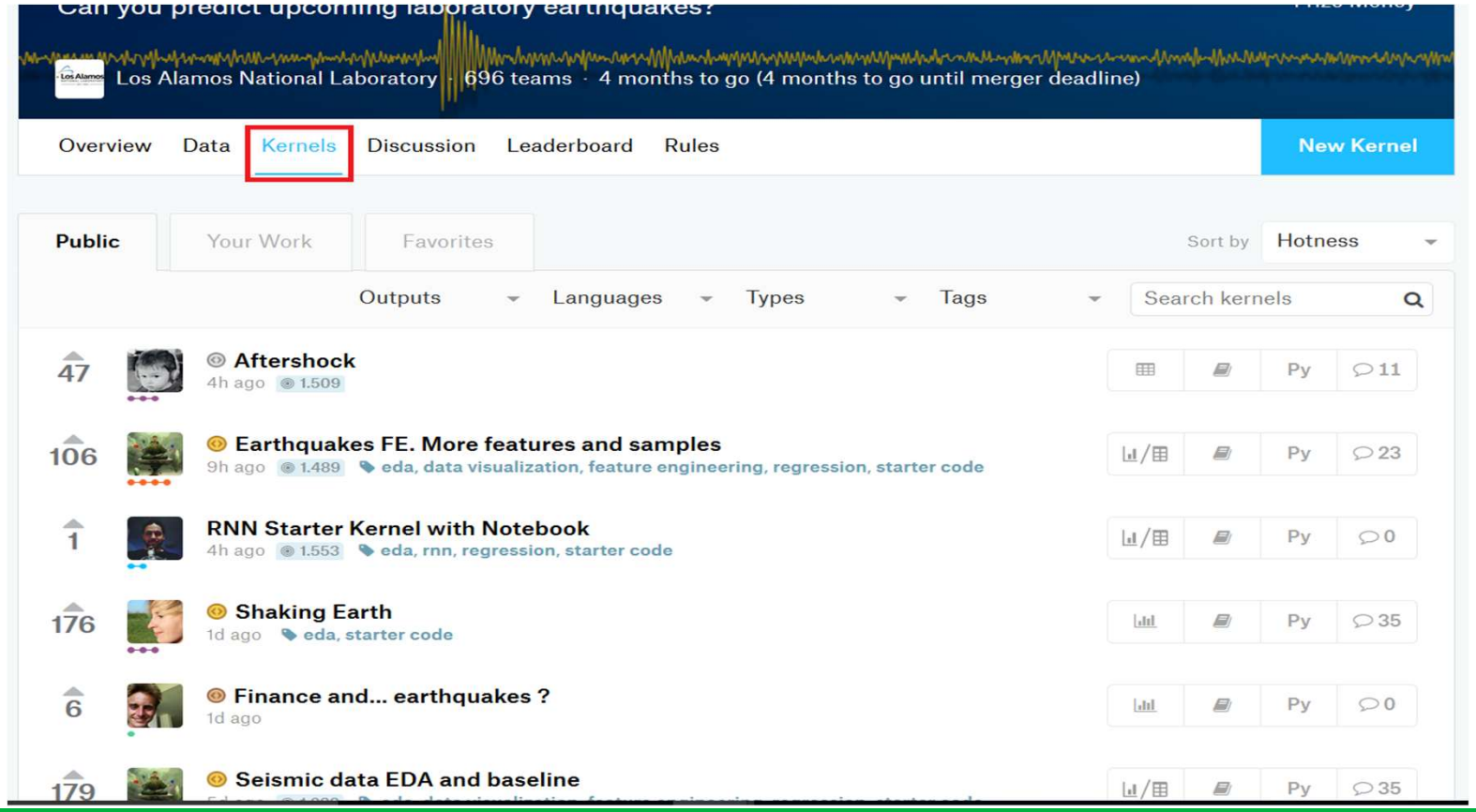
File descriptions

- **train.csv** - A single, continuous training segment of experimental data.
- **test** - A folder containing many small segments of test data.
- **sample_sumbission.csv** - A sample submission file in the correct format.

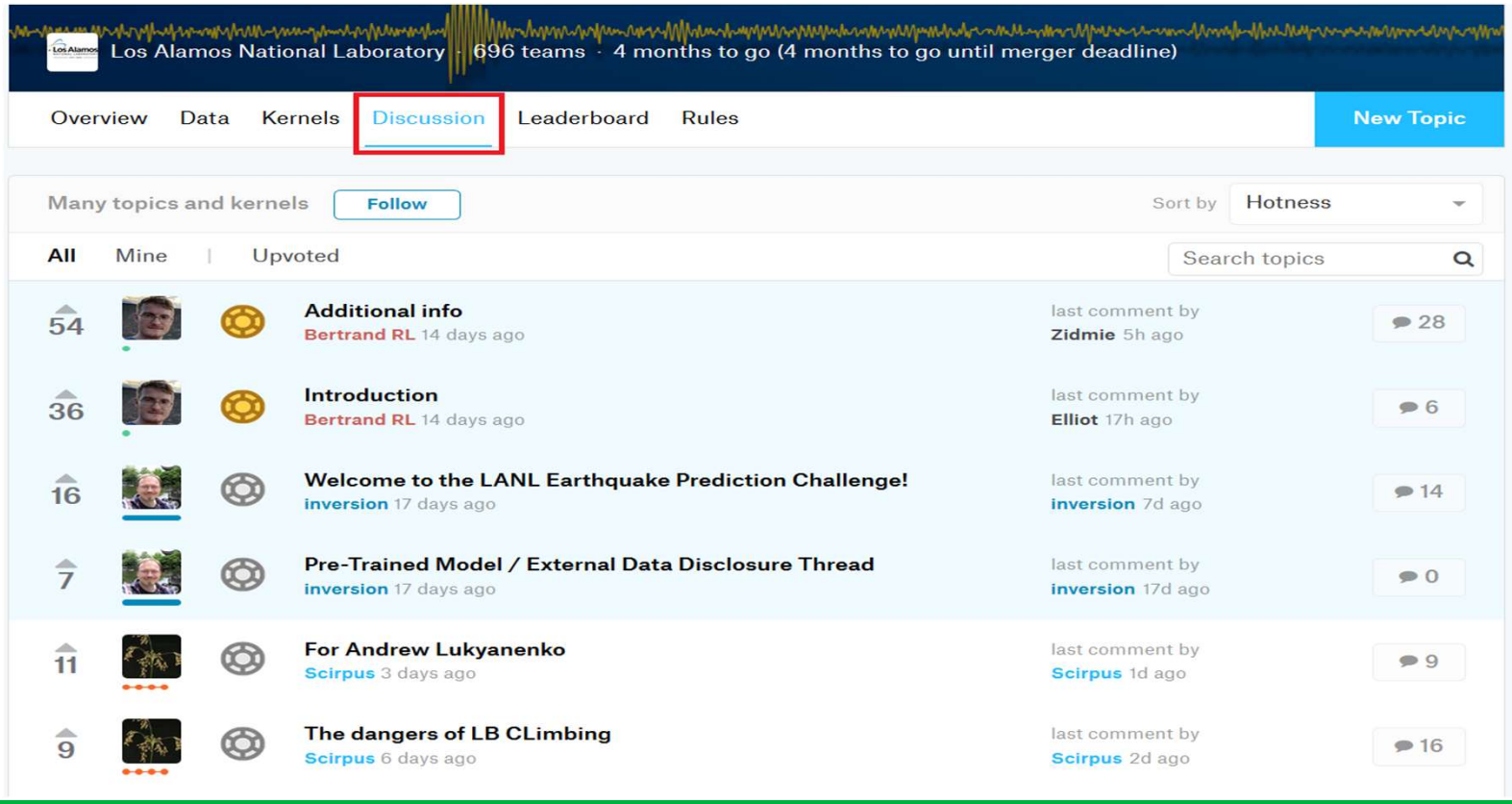
Data fields

- **acoustic_data** - the seismic signal [int16]
- **time_to_failure** - the time (in seconds) until the next laboratory earthquake [float64]
- **seg_id** - the test segment ids for which predictions should be made (one prediction per segment)

Kernel


Los Alamos National Laboratory · 696 teams · 4 months to go (4 months to go until merger deadline)
Overview
Data
Kernels
Discussion
Leaderboard
Rules
New Kernel
Public
Your Work
Favorites
Sort by
Hotness
Outputs
Languages
Types
Tags
Search kernels
47
Aftershock
4h ago 1.509
Py 11
106
Earthquakes FE. More features and samples
9h ago 1.489 eda, data visualization, feature engineering, regression, starter code
Py 23
1
RNN Starter Kernel with Notebook
4h ago 1.553 eda, rnn, regression, starter code
Py 0
176
Shaking Earth
1d ago eda, starter code
Py 35
6
Finance and... earthquakes ?
1d ago
Py 0
179
Seismic data EDA and baseline
Py 35

Discussion


Los Alamos National Laboratory · 696 teams · 4 months to go (4 months to go until merger deadline)
Overview
Data
Kernels
Discussion
Leaderboard
Rules
New Topic
Many topics and kernels
Follow
Sort by
Hotness
All
Mine
Upvoted
Search topics
54
Additional info
Bertrand RL 14 days ago
last comment by
Zidmie 5h ago
28
36
Introduction
Bertrand RL 14 days ago
last comment by
Elliot 17h ago
6
16
Welcome to the LANL Earthquake Prediction Challenge!
inversion 17 days ago
last comment by
inversion 7d ago
14
7
Pre-Trained Model / External Data Disclosure Thread
inversion 17 days ago
last comment by
inversion 17d ago
0
11
For Andrew Lukyanenko
Scirpus 3 days ago
last comment by
Scirpus 1d ago
9
9
The dangers of LB CLimbing
Scirpus 6 days ago
last comment by
Scirpus 2d ago
16

Leaderboard

Los Alamos National Laboratory · 696 teams · 4 months to go (4 months to go until merger deadline)

Overview | Data | Kernels | Discussion | Leaderboard | Rules | Join Competition

Public Leaderboard | Private Leaderboard

This leaderboard is calculated with approximately 13% of the test data.
The final results will be based on the other 87%, so the final standings may be different.

Raw Data | Refresh

In the money | Gold | Silver | Bronze

#	△1w	Team Name	Kernel	Team Members	Score	Entries	Last
1	new	Elliot			1.362	32	12h
2	new	Arik Ermshaus			1.394	22	1h
3	new	Zidmie			1.396	15	20h
4	new	Jun Koda			1.407	13	2d
5	new	ralphy			1.415	30	19h
6	new	dejtAR			1.432	13	4d
7	new	SinanKefeli			1.434	7	10h

Rules

Overview Data Kernels Discussion Leaderboard Rules Join Competition

Rules

One account per participant

You cannot sign up to Kaggle from multiple accounts and therefore you cannot submit from multiple accounts.

No private sharing outside teams

Privately sharing code or data outside of teams is not permitted. It's okay to share code if made available to all participants on the forums.

Team Mergers

Team mergers are allowed and can be performed by the team leader. In order to merge, the combined team must have a total submission count less than or equal to the maximum allowed as of the merge date. The maximum allowed is the number of submissions per day multiplied by the number of days the competition has been running.

Team Limits

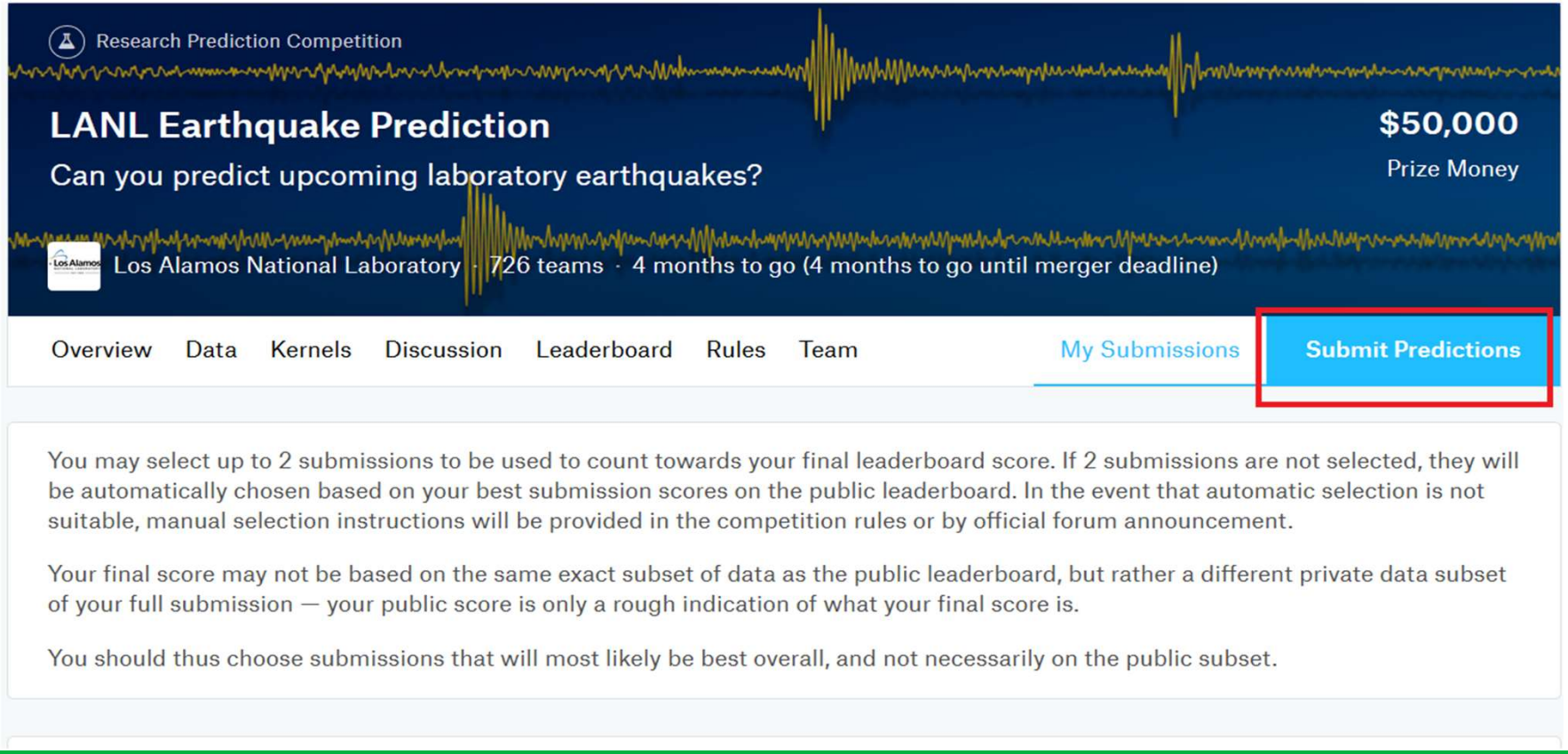
The maximum team size is 8.

Submission Limits

You may submit a maximum of 2 entries per day.

You may select up to 2 final submissions for judging.

Submission


Research Prediction Competition
LANL Earthquake Prediction
Can you predict upcoming laboratory earthquakes?
$50,000
Prize Money
Los Alamos National Laboratory · 726 teams · 4 months to go (4 months to go until merger deadline)
Overview
Data
Kernels
Discussion
Leaderboard
Rules
Team
My Submissions
Submit Predictions
You may select up to 2 submissions to be used to count towards your final leaderboard score. If 2 submissions are not selected, they will be automatically chosen based on your best submission scores on the public leaderboard. In the event that automatic selection is not suitable, manual selection instructions will be provided in the competition rules or by official forum announcement.
Your final score may not be based on the same exact subset of data as the public leaderboard, but rather a different private data subset of your full submission — your public score is only a rough indication of what your final score is.
You should thus choose submissions that will most likely be best overall, and not necessarily on the public subset.

Top 5 points to succeed in Kaggle

- Data preparation is very important and often takes considerable time.
- Special focus should be given on feature engineering.
- Domain knowledge in the specific field is crucial.
- Try tuning the hyper-parameters in the ML algorithm.
- Use a good validation set for evaluation.

Question?

Thank You