Assignment #3

Competition using decision tree

Introduction to Big Data Analytics

TA: Dongmin Hyun and Junyoung Hwang

(dm.hyun@postech.ac.kr, jyhwang@postech.ac.kr)

Machine Learning Practice with Kaggle

- In this Homework, you will practice to implement, train, and tune a decision tree on two datasets using Kaggle.

- Kaggle is a framework for evaluating the performance of ML models (you can find a lot of tutorials for Kaggle).

- General procedures
  - Train your ML model using given training data from given links
  - Produce predictions using the trained model on test data (You can find the output format as '[data name]_sample_submission.csv' from the given links).
  - Submit the output for test data to Kaggle to evaluate the performance of your model.

Notes

- Your score is based on your ranking at Kaggle (private leader board).
  - The ranking of public and private leader board could be different.

- You have to submit your code and report (1~2 pages) to LMS.
  - The report includes summary of your process and your submission score.

- You can use ***any*** python library for this competition.

- You must use your student ID number (ex. 20181234) as your team name.

- *You must read all competition rules carefully and comply with them.*

Kaggle links for two Datasets

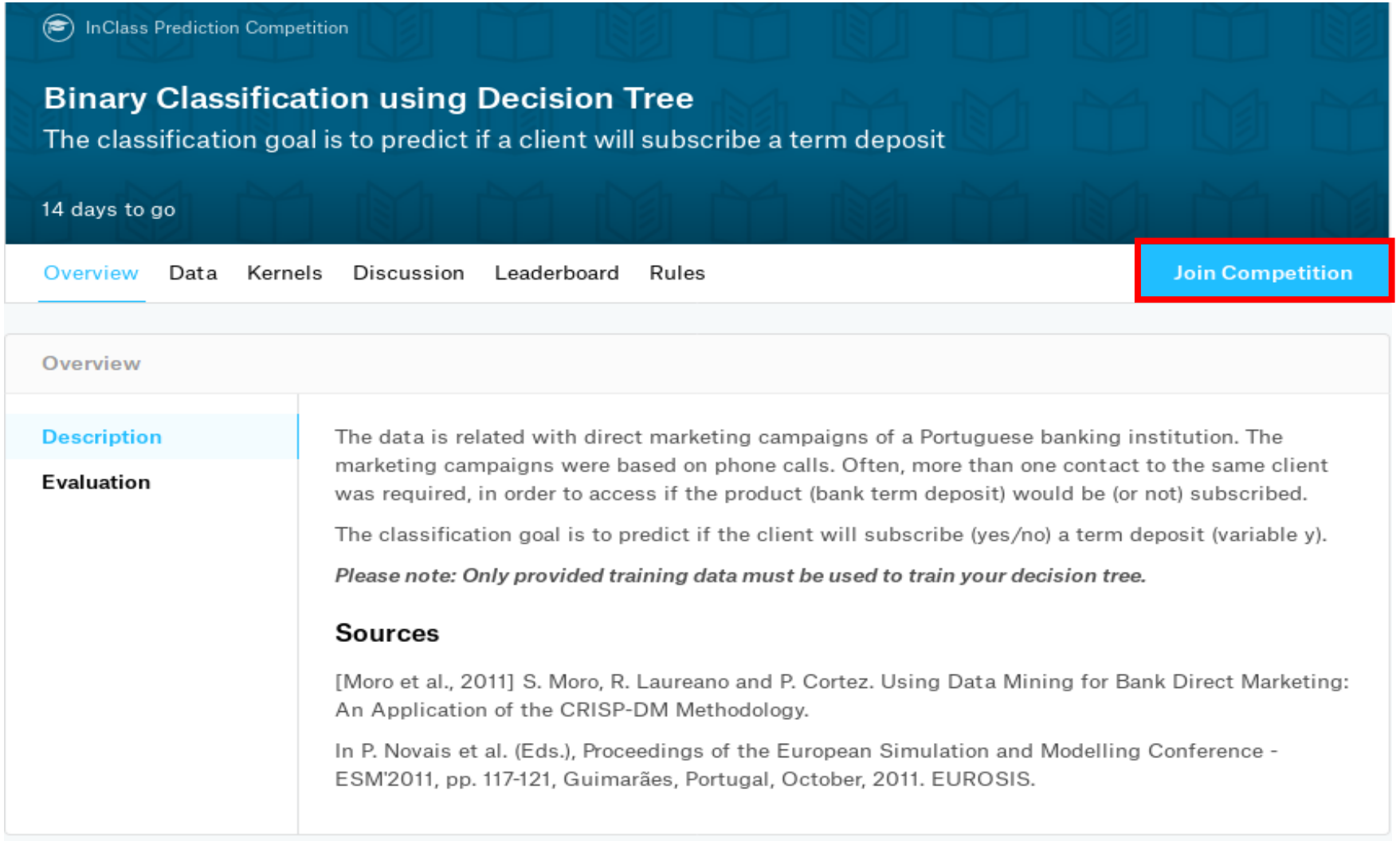
- San Francisco Crime Dataset (Multi-class classification)

https://www.kaggle.com/t/c73f61dbb1804cb7965fd7898e20efab

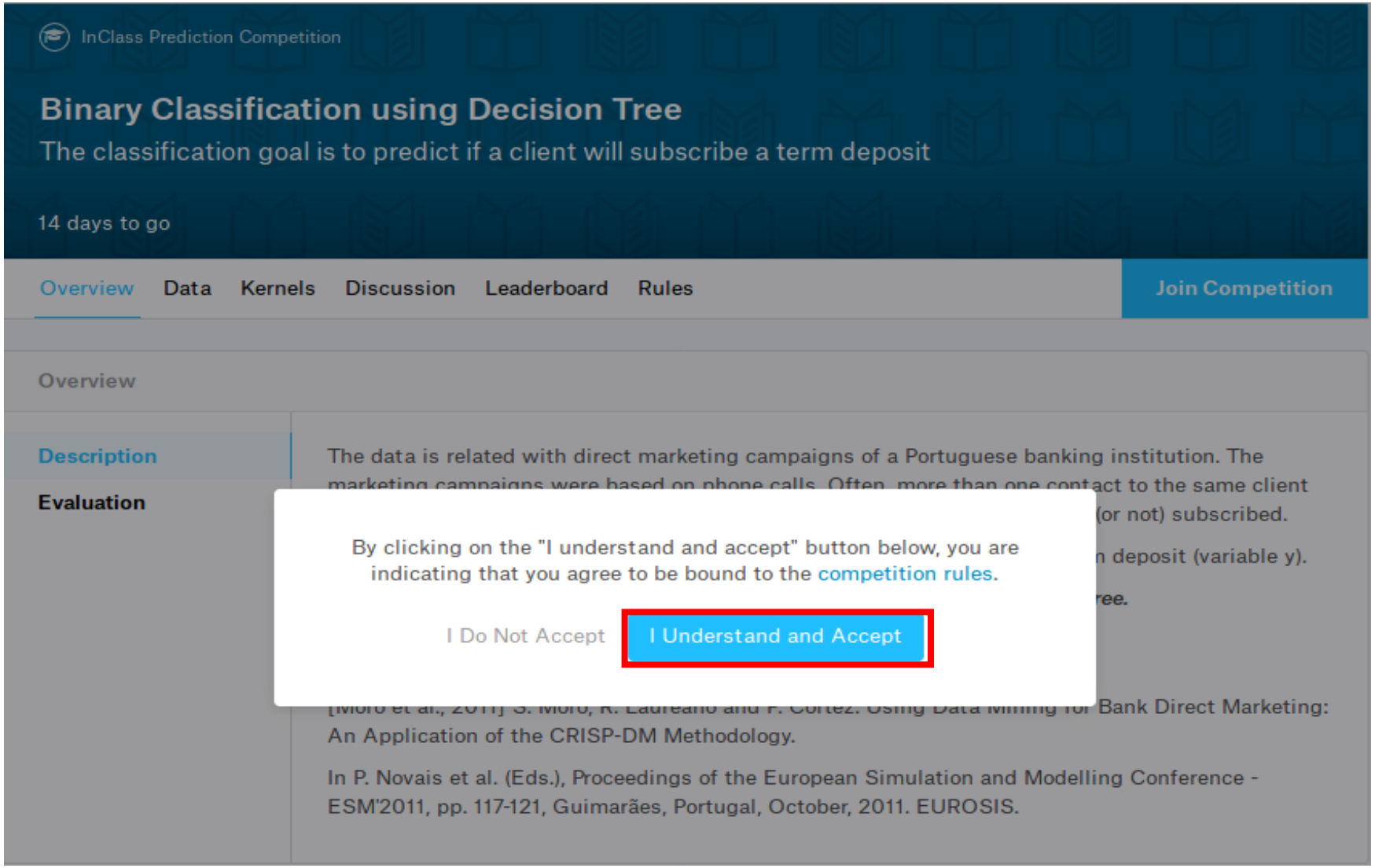
- Bank Dataset (Binary Classification)

https://www.kaggle.com/t/22b72934f5c342aaad360509efb995f4

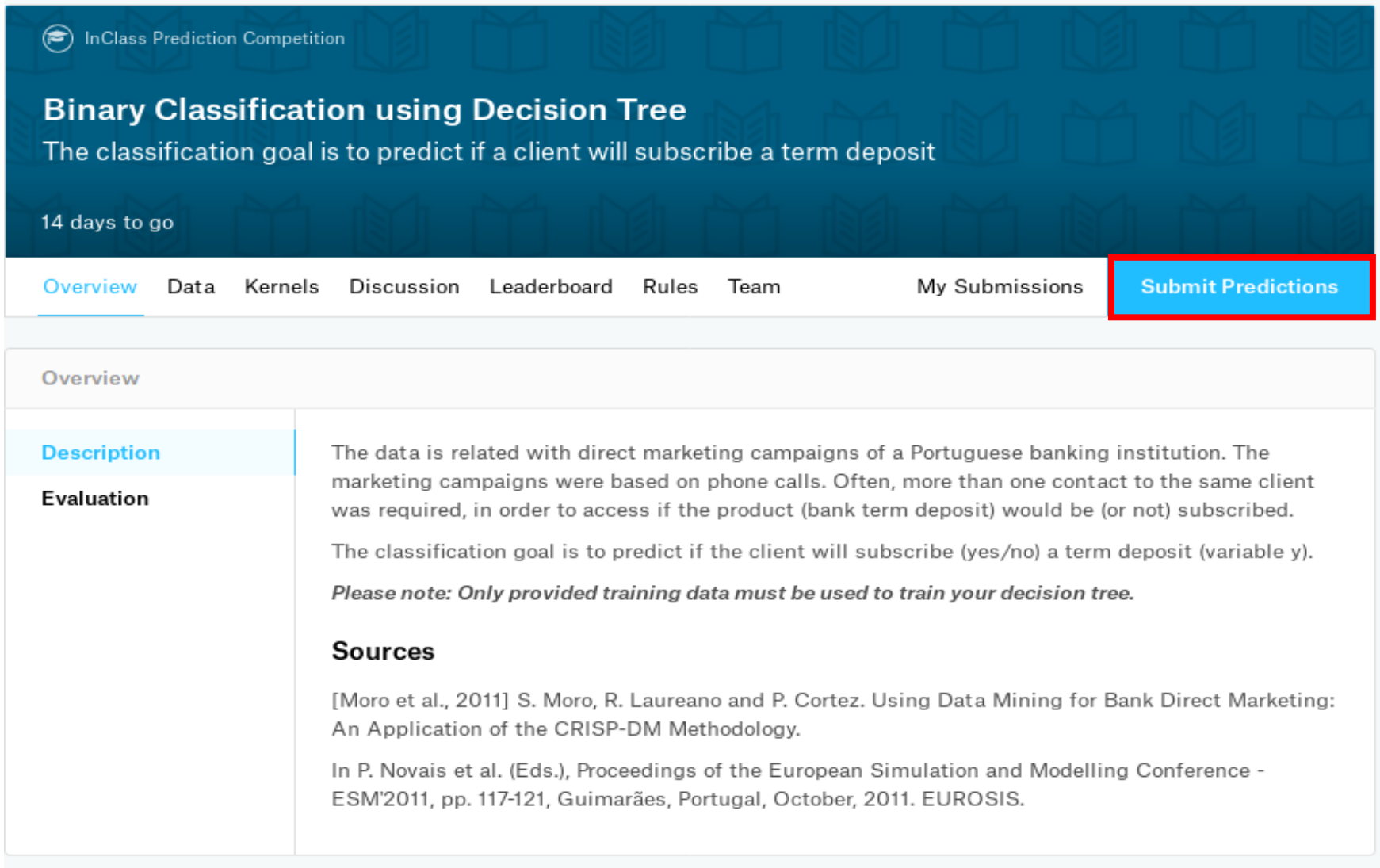
Kaggle Competition (Join using the given links)



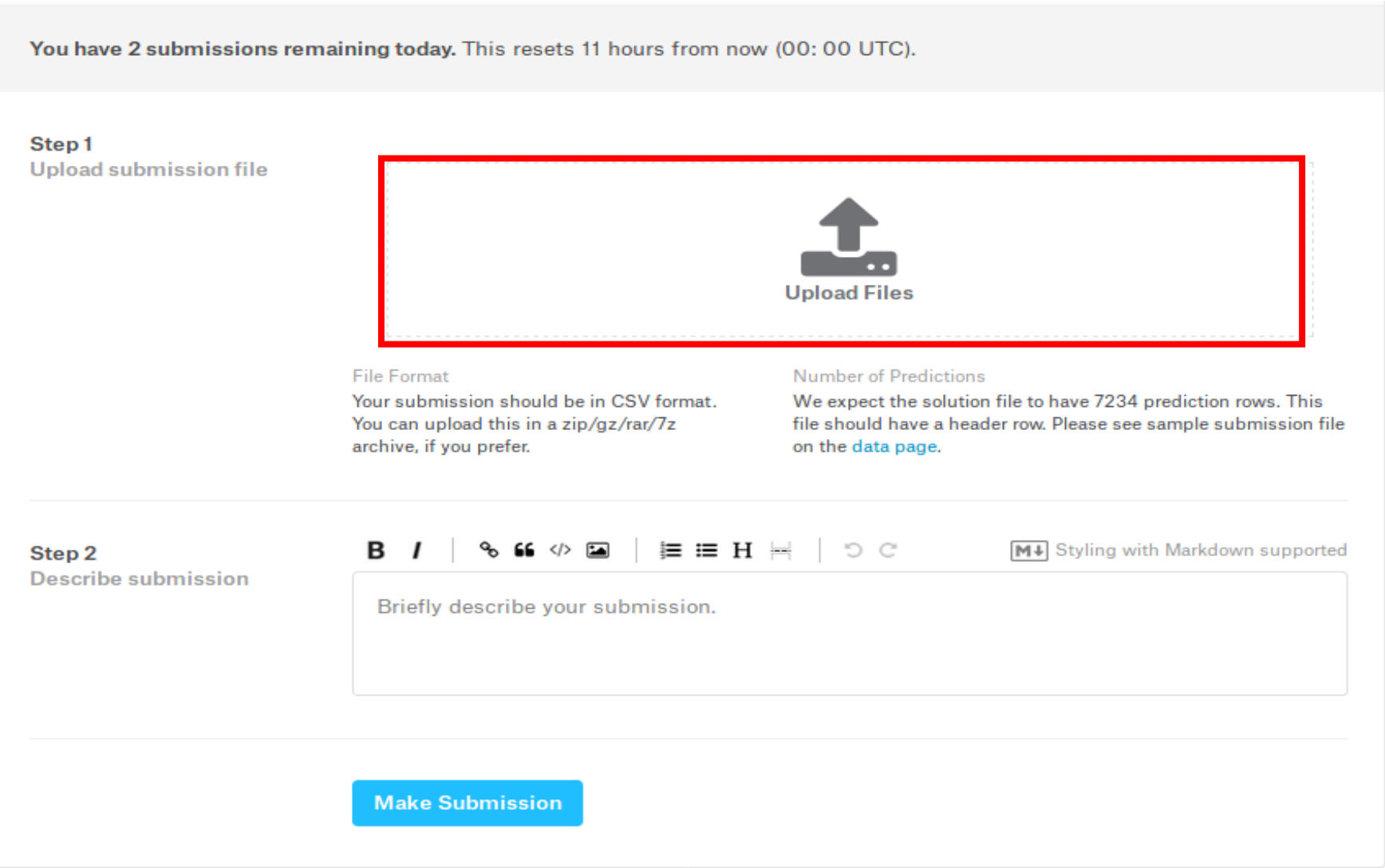
Kaggle Competition (Join Competition)



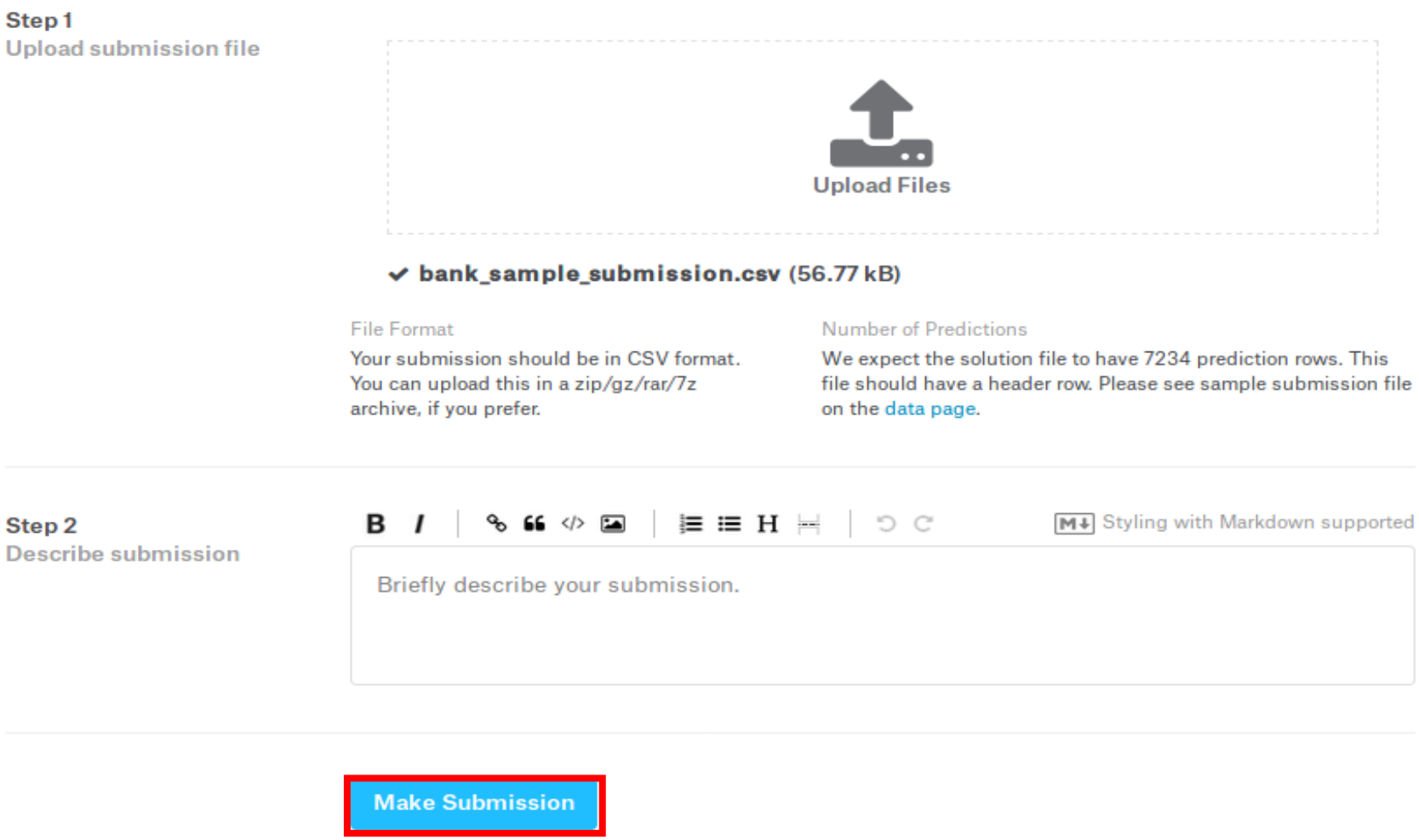
Kaggle Competition (Submit Predictions)



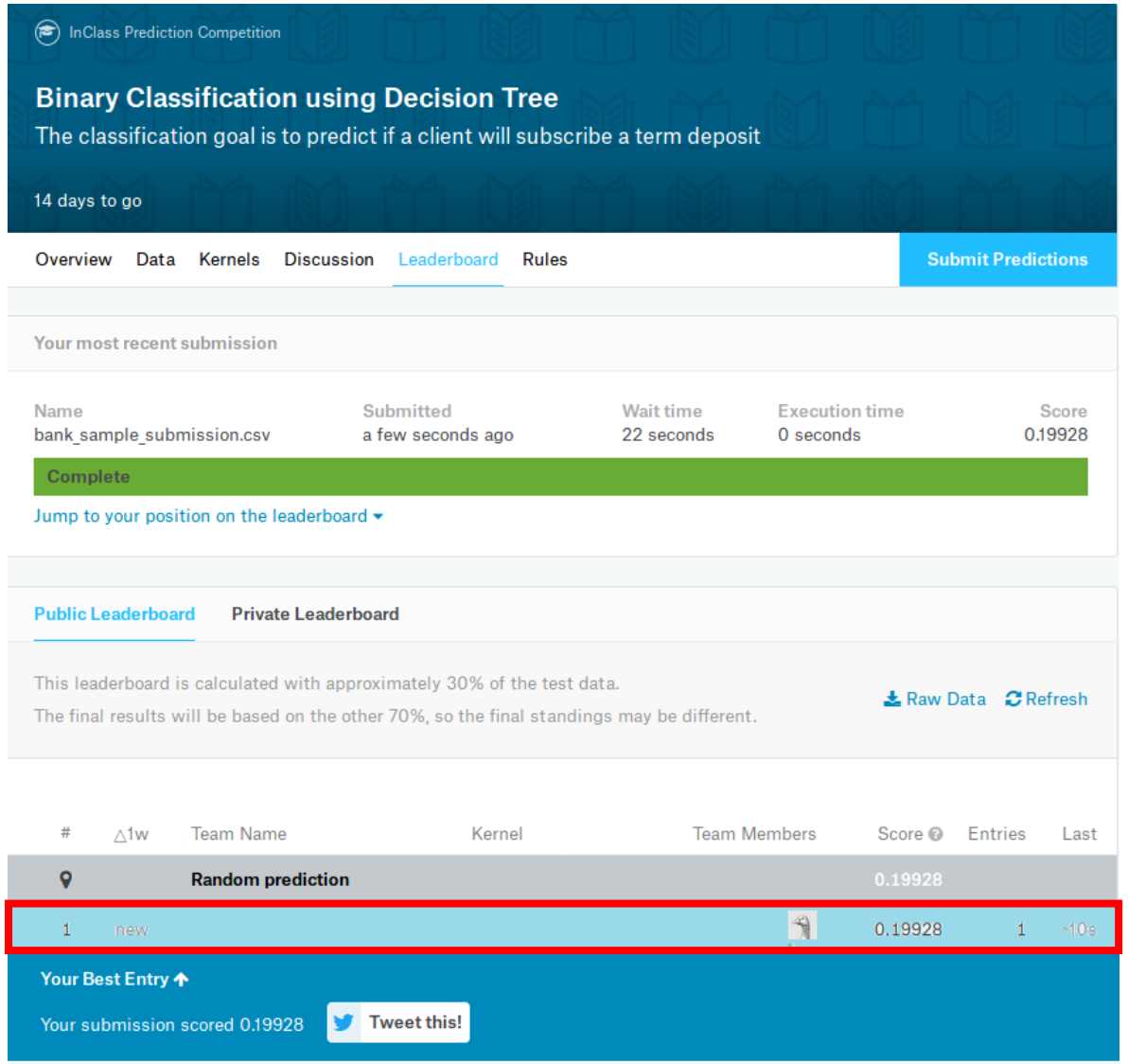
Kaggle Competition (Submit Predictions)



Kaggle Competition (Submit Predictions)



Kaggle Competition (Check Score)



Python Library for Competitions

- Anaconda(https://www.anaconda.com/download/)
  - Almost all packages are installed in anaconda including below.
- Data preprocessing
  - Pandas (http://pandas.pydata.org/)
  - matplotlib (https://matplotlib.org/)
- Machine Learning algorithm
  - Scikit-learn (http://scikit-learn.org/stable/)
  - Numpy (http://www.numpy.org/)
  - …