ML Project

Guilherme Medeiros Machado
Associate Professor at ECE
gmedeirosmachado@ece.fr

Announcements

- Final exam:
  - **Date:** December 17th
  - **Hour:** 15h15
  - **Length:** 2h
  - **Format:** Multiple Choice Questions
  - Small penalty for wrong, empty, or incomplete questions
  - Automatic correction

- Recap about the grade:
  - (Final exam grad * 0.5) + (Project grade * 0.3) + (Small MCQs * 0.2)

Opportunity

- Participate in a **final exam simulation**
  - **Different questions and modalities** (MCQ, written questions, oral questions...) than the real exam.
  - The content is all about machine learning.
  - You will receive a grade about your performance.
  - The simulation is **in person at Eiffel 1**
  - Two students at a time
    - Why: because we will collect some data during the exam for research purposes.

Practical Information

- **When:** From December 9th to 13th
- **Length:** 1h
- Mandatory inscription
  - Link sent by mail on Thursday (December 5th)
- Only **60 available time slots** (first-come, first-served)
  - Please honor your time-slot!
  - Or alert the organization team in advance, so other students could participate

Extra Motivation 😀

- Up to 2 points in your final exam:
  - 1 point for every participation
  - If your grade is higher than 16
    - +1 point
    - Total = 2 points in your DS grade
  - If your grade is lower than 16
    - No extra point
    - Total = 1 point (given by your participation)

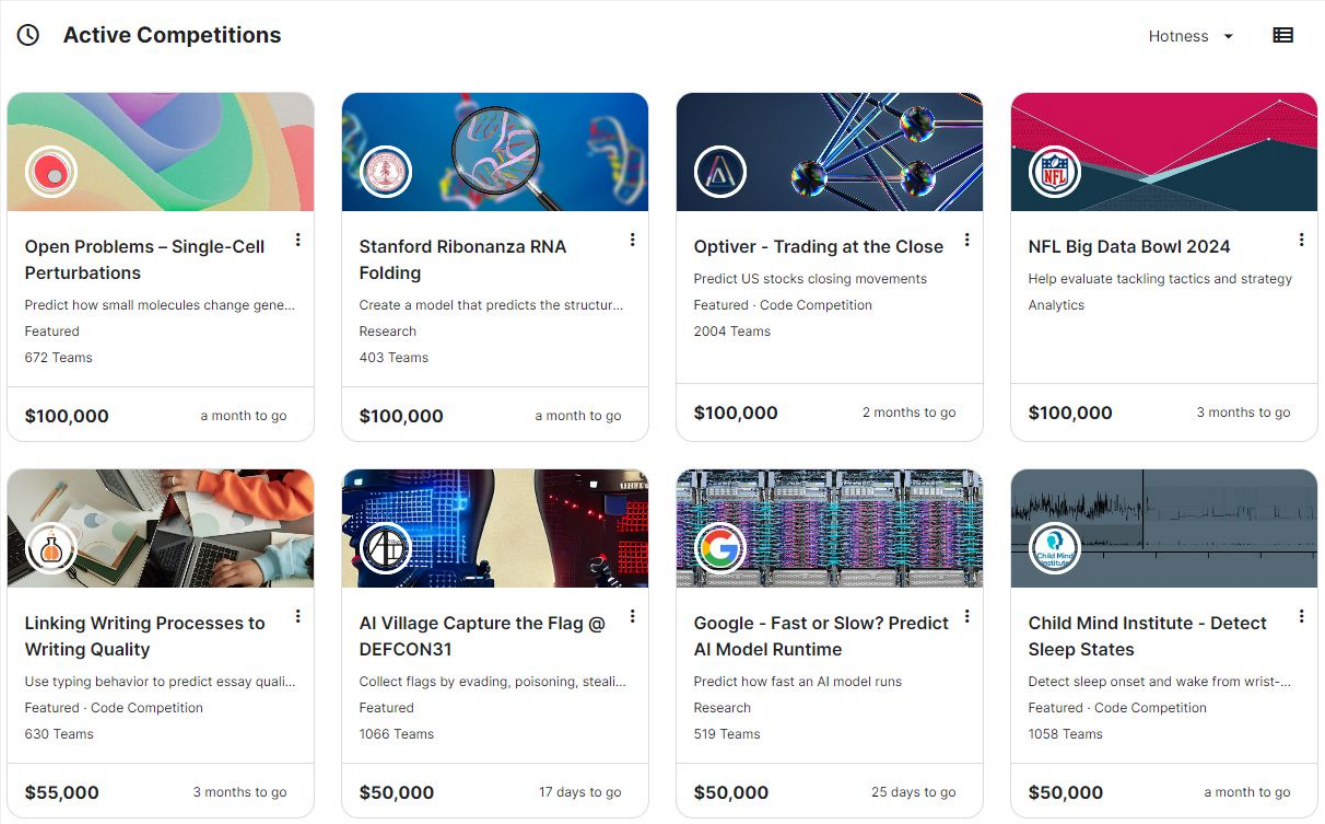
A kaggle Competition!

What is Kaggle?

- A subsidiary of Google.
- An online community of data scientists and machine learning engineers.

- In Kaggle you can:
  - Find useful datasets to build AI models.
  - Publish datasets.
  - Work with other data scientists and machine learning engineers.
  - Enter competitions to solve data science challenges.

A little extra Motivation



Coming back to reality

- We are going to join an easy competition.

- The goal is:
  - To assess your **capability to use all the tools** that you've learned during this course.

  - To test your **ability to propose new ways to solve a problem.**
    - You can create new features, normalize or not, play with hyperparameters, use differents versions of the algorithms...

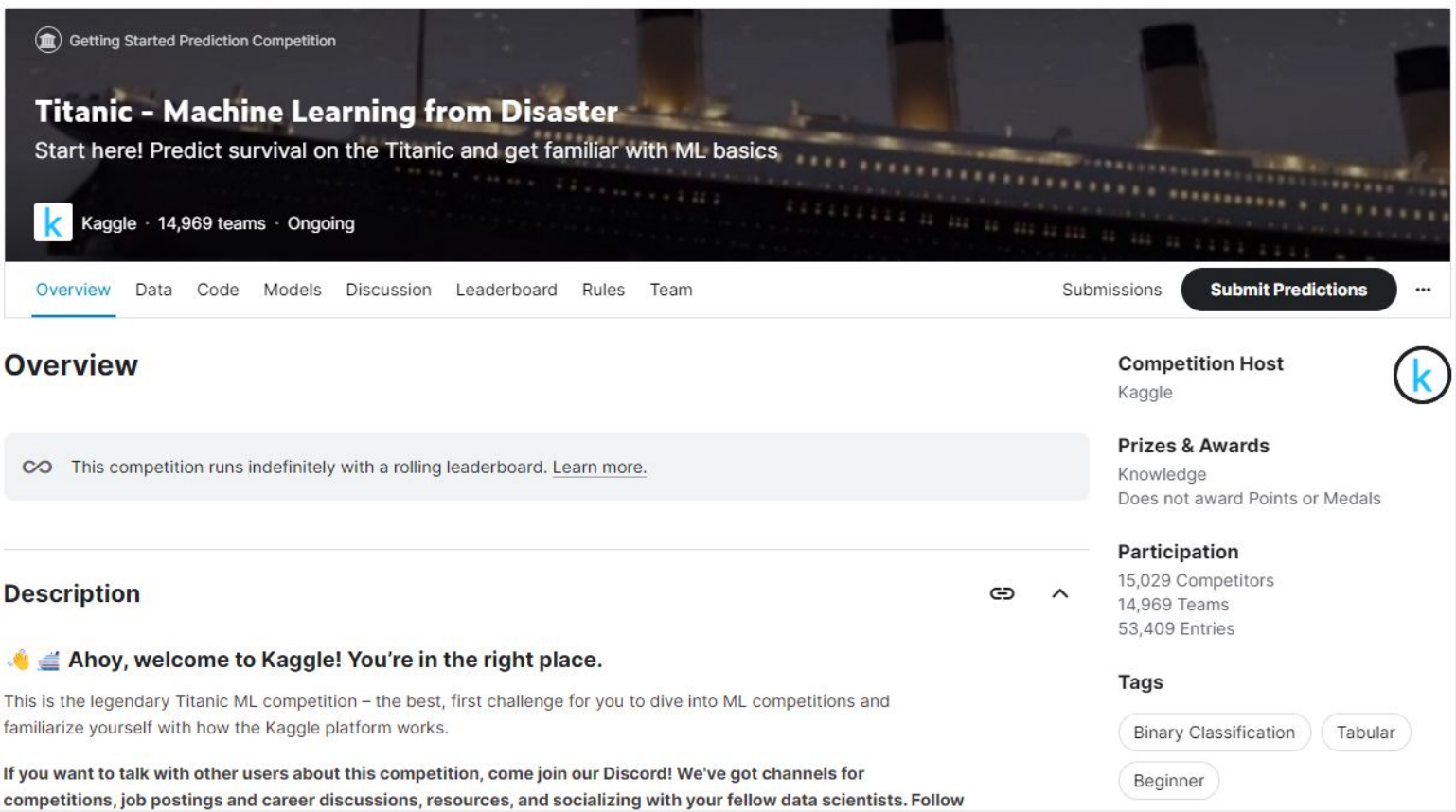
  - To assess your ability to deliver **good results in a short schedule**.

Our target competition

- Some Hints!
  - It is based on a movie and a real history;
  - The movie won 11 oscars;
  - It is the history of a tragedy;
  - The movie was created during the 90’s;
- Any guesses?

Our target competition

Our target competition



Titanic ML from disaster

- Step-by-step:
  1. Access https://www.kaggle.com/

  2. Register your account or Sign In

  3. Once you sign in access the competition page https://www.kaggle.com/competitions/titanic

  4. Click on “Code”, and then “New Notebook”
     - The first thing to do is to put your names as a comment inside the notebook

  5. Invite your team members by clicking in “Share”, and then searching their user names in the search bar. Give a name to your team!
     - Alternatively you can also use:
       - *Google Colab (https://colab.research.google.com/)*
       - or work in *Visual Studio using Live Share extension*
       - or work locally with github

  6. Finally you can write your code.

Titanic ML from disaster

- Step-by-step:
  7. You can access the data using pandas.

```
import pandas as pd

#reading the data
df_train = pd.read_csv("/kaggle/input/titanic/train.csv")
X_train = df_train.drop(['Survived'], axis=1)
X_test = pd.read_csv("/kaggle/input/titanic/test.csv")
y_train = df_train[["Survived"]]
```

  8. You should notice that **we do not have an "y_test"** variable, because your task is to create it.

  9. **Explore the dataframes of features X_train and X_test.** You will see that we have information about the passengers (name, ticket fare, gender, chamber's class....)

  10. If you want to know what is the meaning of the stored information you should look at the competition's webpage.

Titanic ML from disaster

- Step-by-step:

11. Proceed with the Feature Engineering, the scaling, the model implementation...

12. Generate a variable “y_pred” that contains your model prediction.

```
# predicting over training & testing datasets
y_train_pred = algo.predict(X_train)
y_test_pred = algo.predict(X_test)

algo.score(X_train, y_train)
```

13. Concatenate the predictions in “y_pred” with “PassengerId” of the variable “X_test”.

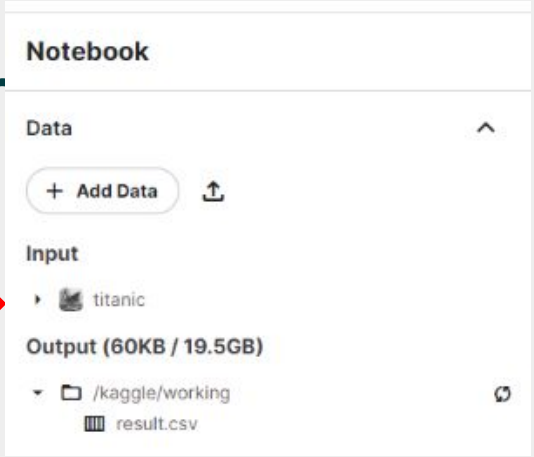
```
result = pd.concat([X_test["PassengerId"], pd.DataFrame(y_test_pred)], axis=1, ignore_index=True)

result.columns=["PassengerId","Survived"]

result
```

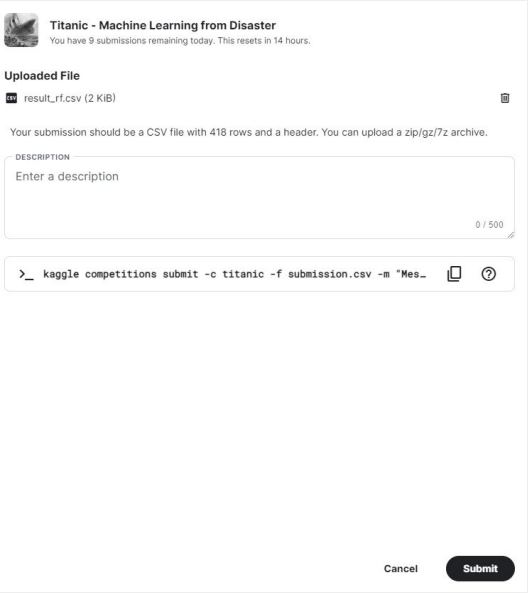
14. Save your result in a csv file.

```
result.to_csv("/kaggle/working/result.csv", index=False)
```

Titanic ML from disaster

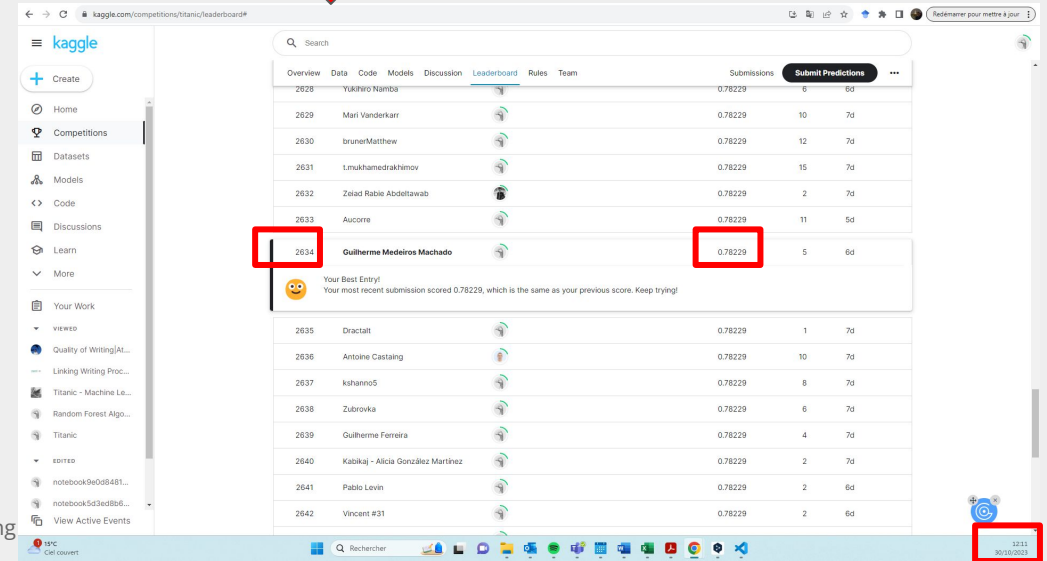
- Step-by-step:
  15. Download your csv file in the “Output” directory
  16. Return to the competition webpage https://www.kaggle.com/competitions/titanic/overview
  17. Click on “Submit Prediction” button, load your csv file and submit your results.
  18. Click on the “Leaderboard” tab, and take a screenshot of your results and of your position.





Titanic ML from disaster

- Step-by-step:
  19. Submit your **jupyter notebook (ipynb file)**, and the **screenshot** of your **score** and **ranking position** on Boostcamp **as soon as possible**.
      - For the **screenshot** submit a file where I can see your whole screen, with the time and the date like the example below.



  20. Submit your score on the Google Form, I will use the timestamp there to define the best group.

  21. You are competing to all the other groups.

What about my grade?

1. I will check your code:

    - 40% - Code Analysis
        - Here I will check for:
            - Correctness;
            - Plagiarism;
            - Code structure;
            - Your changings, your effort put on this code.

        - **If you get a zero in Code Analysis then your final grade is zero.**

What about my grade?

2. 40% - Accuracy score
   - Final accuracy = (Accuracy score + 0.13) * 20
   - Because we are establishing the higher accuracy you will get will be around 0.87

What about my grade?

2. 40% - Accuracy score
   - Final accuracy = (Accuracy score + 0.13) * 20

3. 20% - Leaderboard position bonus

🥇 If you are among the **top-3 groups** then you have 100% of bonus.

🥈 If you are at positions **4 to 6** then you have 60% of bonus.

🥉 If you are at positions **7 to 9** then you have 20% of bonus.

🙈 From the position **10 to ∞** you have 0% of bonus.

⌛ A tie will be settled by time.
If you arrived first in that position, you are ranked first.

What about my grade?

2. 40% - Accuracy score
   - Final Accuracy = (Accuracy score + 0.13) * 20

3. 20% - Leaderboard position bonus

🥇 If you are among the **top-3 groups** then you have 100% of bonus.

🥈 If you are at positions **4 to 6** then you have 60% of bonus.

🥉 If you are at positions **7 to 9** then you have 20% of bonus.

🙈 From the position **10 to ∞** you have 0% of bonus.

⌛ A tie will be settled by time.
If you arrived first in that position, you are ranked first.

4. Final Grade = (Code Analysis *0.4)+(Final Accuracy *0.4)+(Leaderboard*0.2)

Miscellaneous

- Each team has at most 2 questions to be asked during the Lab.

- You can submit your predictions at most 10 times to Kaggle.
  Be sure before submitting something.

- You lose your accuracy and ranking points (60% of the grade) if:
  - You send a screenshot that does not contain your rank position, your system date and time, and your accuracy score from Kaggle.

  - The date and time of your screenshot differs of more than 5 minutes of the date and time of your Boostcamp submission.

The Deadline

- You will have **24h** to submit your files on Boostcamp.
- You can submit your files many times. I will keep only your last submission.

On your mark, get set, go!

The timer starts now