Kaggle Challenge: How Many Shares?

HPC IA Introduction to Machine Learning

2019–2020

1 Challenge Presentation

Kaggle[1] is a platform dedicated to data mining. In particular, it hosts public data science challenges; the problems are posted by sponsors, and people from all over the world can compete for putting together the best solution.

Your project is centered on the How Many Shares (2019-2020)? challenge at Kaggle In Class. You can sign up for the challenge using the following invitation URL:
https://www.kaggle.com/t/f86f7284736444d48510ce125d189a36 .

The goal of the challenge is to predict how many times an online article is going to be-shared, using variables that describe the article (keywords, topic, length, sentiment analysis,etc.). A detailed description of all features is given on the competition webpage.

For this course, you will enter the competition either alone or in pairs. Your goal will be to try out algorithms seen throughout the course. You are also encouraged to try out other algorithms, combinations of those, and generally explore the challenge data.

The project will be graded on a 2-page report (figures and tables can be moved to an appendix) as well as the ranking of your final submissions to the leaderboard. Final submissions as well as the report are due on **on February 2, 2020, 23:59**.

You are strongly encouraged to use Jupyter + scikit-learn to complete the project, but you will not be graded on your code.

2 Instructions

1. Decide whether to work alone or in pairs.
2. Register on Kaggle and create a team.
3. Download the data.
4. Setup a cross-validation framework for the analysis of your data.
5. Use this framework for model selection. You are encouraged to explore:
   - Various transformations of the features, as suggested in the labs or as you see fit.
   - Various machine learning algorithms and their hyperparameters, particularly among those seen in class.

[1] https://www.kaggle.com/about

6. Submit as many models as you wish, in a limit of 5 a day, to the leaderboard, to see their performance on the *public validation set.*
7. Submit the predictions made by **2 optimized final models** to the leaderboard. Those are the models you believe should win the challenge. They will be judged according to their performanceon the *private validation set.*

3 Evaluation

Final report The report is to be deposited at `https://frama.link/hpc_ai-ml-2019_20` no later than **February 2, 2020 at 23:59**.

Please name your report file "Lastname1Initial_Lastname2Initial.pdf" (supposing you are a team of 2 people). If Jane Smith and Sarah Martin work together, their report should be named `MartinS_SmithJ.pdf`.

Your report should be **no more than 2 pages long.** Figures and tables can be moved to an appendix.

Your report should contain the following elements:

- **Your full names, your Kaggle user names, and your Kaggle team name**.
- A discussion of feature processing. Did you standardize the data, chose alternative representations for some features, discarded other features, and why?
- The cross-validated performance, on the training data, of the algorithms you explored. You are strongly encouraged to explore the space of parameters for each of these algorithms. Briefly explain how you did it. Discuss which algorithms/parameters work best.
- The performance, on the validation data (visible part of the leaderboard), of one model of each of the five families. Discuss whether the results match your expectations.
- A discussion of additional models you have tried, insights you have gained (e.g. "This method works well but is difficult to fit" or "This method is not very accurate but is really fast to train").
- A discussion of your choice of final model(s). You can submit up to 2 final models. What are these models, how did you construct them, why do you expect them to be your best proposals?

Include tables or figures as you see fit.

Leaderboard ranking The ranking of your team on the leaderboard on **February 2nd, 2020 at 23:59** will count towards part of your grade.