Kaggle: The good, the bad and the code

Brief Presentation

28th August 2018

What is Kaggle?

- Kaggle is a data science competition hosting website.
- There are 3 major categories of competitions:
  - Learning
  - Research
  - Industry funded (Featured)
- Kaggle cannot be used to learn to code. If you know how to code, you can learn how to code ML algorithms on Kaggle.

My Experience on Kaggle

- PCA + Random Forests on MNIST dataset: https://www.kaggle.com/zbpvarun/pca-random-forests
- XGBoost + LightGBM on Zillow housing prices
- Neural Networks on Cdiscount Image classification.
- Humpback Whale Image Identification using bounding boxes and neural networks.

The advantages of Kaggle

- Immense computational power on servers.
- State of the art ML and data analysis packages pre-installed without compatibility issues.
- Incentives for collaboration has fostered an open and welcoming community that encourages helping each other to produce better algorithms / results.
- Is becoming a repository of vast, clean and easy to work with datasets.

The disadvantages of Kaggle

- Immense computational power and seamless preinstalled packages.
- Overemphasizes the machine learning part of data science which is a minority part of the job.
- As welcoming as the community is, competition culture has resulted in a lot of crowding.