kaggle

This section in the newsletter aims to introduce some of the platforms that will be used in the Algothon 2018. This week we'll be giving a balanced overview of Kaggle.

We encourage you participate in any Kaggle competitions to prepare for the 20th October.

What is Kaggle?

1. A platform used during competitions to carry out machine learning based executions.
2. In some ways, it resembles Codechef, Hackerrank but differentiates itself from these by being very specific to data science, machine learning etc.
3. In the case of the Algothon, you will be given a Kaggle data set to start building your own machine learning models and once the kernel is published it will be reviewed by the Hackathon panelist's. Other teams and kaggle users will also be given the chance to upvote on the models.

How does any data-scientist use Kaggle?

The platform can be used in 3 main ways:

1. Competitions - You can compete on many problems. Find the problems you find interesting and compete to provide better solution to the problem.
2. Datasets - The best place to discover and seamlessly analyze open data. You can find any dataset of your interest and apply analytical methods to fine tune your analytical capabilities.
3. Kernels - Allows, user to share and learn from different approaches.

Things to consider when beginning to use Kaggle

1. Kaggle's primary goal is to *find the algorithm that minimises a loss function (error between the expected and obtained result)* and so doesn't always follow how Machine Learning is used in the 'real world'. Unsurprisingly, this is not the only focus when considering the 'real world'. Given, the accuracy of model A is 91.0% and model B is 91.1% and they take 1 hour and 1 day respectively. We must *consider the trade-offs* of whether the slower output is worth the 0.1% increase in accuracy.

2. Although many people competing using Kaggle use ‘copy and paste’, Black Box machine learning algorithms, it is also possible to implement algorithms from scratch, to understand them better.

3. As solutions must be new and to be successful extended research, customised algorithms and trained advance models must be used. This continues to make the Kaggle less accessible to complete beginners.

4. Performance is always relative to others competing.
5. Most of the competitions on Kaggle involve very large datasets. *Our Algothon dataset is no exception at over 50GB.* This is often not the case in the real world. Algorithms that work well for large data (e.g. neural networks) are not suited for small datasets.

How can I use Kaggle to prepare for the Algothon?

Start experimenting with machine learning trends and techniques.

1. The article gives a **step by step guide** running users through the *‘installation of an environment’* to the *‘submitting predictions and retrieving scores on Kaggle’*. https://medium.com/@faizanahemad/participating-in-kaggle-data-science-competitions-part-1-step-by-step-guide-and-baseline-model-5b0c6973022a

2. This gives a high level and easy to understand intro to Machine Learning and the common statistical models used. It was taken from a Kaggle Q&A.

I. https://www.linkedin.com/pulse/machine-learning-whats-inside-box-randy-lao/?published=t

II. https://www.kaggle.com/questions-and-answers/41211

Read Kaggle blogs, interviews of grand masters and more general interviews with previous winners.

1. ‘No free Hunch’ Blog

This article from the blog gives a general analysis of how previous winners have been successful. For example, it talks about the *dangers of overfitting* to competition public datasets and the vast differences between public and private ranks, as a result. It also emphasises how pervious Kaggle winners have used specific techniques such as *the ensembling of models* and some of the preferred models from previous successful Kagglers such as the *‘Friedman's gradient boosting machine.’*
http://blog.kaggle.com/2014/08/01/learning-from-the-best/

2. Winners’ interview

This is article is an interview with winners of the ‘Two Sigma financial modelling competition’ ending in March 2017 to find signal in financial markets with limited hardware and computational time.
http://blog.kaggle.com/2017/05/25/two-sigma-financial-modeling-challenge-winners-interview-2nd-place-nima-shahbazi-chahhou-mohamed/

It is recommended that you start using Kaggle over the next few weeks before the Algothon to see which of the two competitions, Kaggle or Quandl/Quantopian you would like to focus your attention on.

The next Algothon primer email will give an overview of Quandl.