Agile Data Science Project Pattern

The following document outlines a methodology that I would like to follow for our Data Science project. Feel free to give feedback or suggestions for improvements. This is an attempt to translate or adapt successful agile practices to data science.

Assumptions: We have a decent understanding about the Figure of Merit (FOM). The FOM is a quantitative measure of success for a specific data science problem. For example for our Kaggle problem *Acquired Value Shopper Challenge*, <https://www.kaggle.com/c/acquire-valued-shoppers-challenge> , we have a well defined FOM given by the area under the ROC curve. The current leader board (May 12) shows a value of 0.62162 for first place and 0.43956 for last place.

Iteration 0 goals:

Download the data. Setup minimal infrastructure, a Git repo,a Mingle instance. Decode on methods of team communication. Begin thinking about the problem and how we might go about solving it. Have a look at the data to get a feel for the structure and data characteristics.

Iteration 1 goals:

Start two parallel tracts.

Tract 1:

The goal is to build three simple models. The first is called the Random Guess Model (RGM). Literally write an algorithm that makes a random guess at the solution. The next model is called the Cheat Model (CM). This model uses the training set to cheat and fetch the right answer. Such as model can only be run on validation data but has the characteristic that, it can’t be beaten by any other model. The third model is called the Minimally Viable Model (MVM). This is the fastest solution that we can implement that attempts to use the data is a real way and is expected to out-perform the RGM.

Tract 2:

Build a Validation Framework (VF) for the models being produced in Tract 1. Some coordination is required between the two groups concerning APIs. The goal of this track is to create a simple validation framework which can take some input file, read it, run each line through at least one model to make a prediction and then (if it is a training set) compare to the true value and calculate a FOM score for each model.

At the end of iteration 1, we have a validation framework, three simple models and have demonstrated that FOM scores for each model are ordered RGM, MVM and CM. The point of the RGM and CM are to test that the VF does what it is supposed to do and give us a feeling about the range of values for the FOM scores.

The running of the validation should be automated so that it can be verified again after any code change. The scores should be persisted each time it is run. Once these models are created and validated, they should not be changed so that the scores they produce do not change when run on the validation test data. The only exception should be if a bug in the VF is discovered and fixed. Otherwise refactoring of the VF might improve performance and produce more useful output but should not change these numbers.

Iteration 2 goals:

Develop and validate more advanced models expected to perform better than the MVM. Improve performance if necessary. Improve visualization of model results. Discuss with product owner about steps to move model scoring into production.

Iteration N goals:

Iterate on same Iteration 2 goals. Steadily increase FOM scores until product owner and team decide to stop. Launch into production.