Kaggle Higgs-Boson competition

My assigned role as part of the Quantum Jitters team was to find a model outside of the standard selections. I tried a couple of inference engines, boosted, bagged, SVM, standard regression, etc.

In the end I came back to the XGBoost routine for a couple of reasons; first, it ran relatively quickly; and second there seemed to be a fair number of tuning parameters available to play with.

Playing with the model and actually getting it to a good fit are two completely different concepts. In general, I took 80% of the training set to train the model, and the remaining 20% as a test case. In all cases I used the AMS as the deciding metric, and measured accuracy as the percentage of the 20% test data correctly ID’d as my own performance indicator.

While I won’t go into every iteration, some of the tuning parameters I used were:

* Number of cv folds, between 2 and 5
* Number of rounds, between 50 and 400
* Eta (shrinkage) between .001 and .6
* Gamma between .01 and .6
* Max tree depth between 6 and 10

Generally, I received AMS scores on the training (80%) set of between 1.2 all the way up to 1.8+. These resulted in % accuracy rates of 80%+ when applied to the test (20%) dataset. My first instinct is this was overfitting, and that turned out to be the case. The AMS scores on the Kaggle full test data site were much lower, with the highest being about .5.