**Higgs Boson Machine Learning Challenge**

**Overview**

The goal of the Higgs Boson Machine Learning Challenge is to explore the potential of advanced machine learning methods to improve the discovery significance of the ATLAS experiment ([ATLAS](http://atlas.ch/) is a particle physics experimenttaking place at the Large Hadron Collider at CERN that searches for new particles and processes using head-on collisions of protons of extraordinarily high energy).  Using simulated data with features characterizing events detected by ATLAS, the task is to classify events into "tau tau decay of a Higgs boson" versus "background."

Training set of 250000 events, with an ID column, 30 feature columns, a weight column and a label column (b or s). All variables are floating point, except **PRI\_jet\_num**which is integer.

Variables prefixed with **PRI** (for PRImitives) are “raw” quantities about the bunch collision as measured by the detector. Variables prefixed with **DER** (for DERived) are quantities computed from the primitive features, which were selected by  the physicists of ATLAS. It can happen that for some entries some variables are meaningless or cannot be computed; in this case, their value is −999.0, which is outside the normal range of all variables.

*Solutions are evaluated by* the approximate median significance (AMS).

**Approach**

I split the training data set into 3 parts such as 60% of the data for training set, 20% for the cross validation and 20% for test set. I performed exploratory analysis over the training data set and found out the meaningless values (-999.0). I discarded the features which have more the 70% of the data points are -999.0 and replaced all other features meaningless values by their column mean. Then using the pre-processing function of R, I scaled the data.

I trained bagging (treebag), boosting (gbm) and random forest (rf) as a classifier then cross validated the models and tested the models on the cross validation set and test set respectively. I checked the accuracy of the models using ROC curve (pROC) and confusion matrix.

After that I predicted on the test data set and created submission file.

**Implementation**

I used R programming language specially R machine learning libraries such as caret, treebag, gbm, pROC packages etc.I trained bagging with decision tree (treebag), gradient boosted tree (gbm) and random forest (rf) algorithm. Treebag algorithm trained comparatively faster than gbm and rf. It took almost 1:40 hours to train (system: Intel Pentium 2010 2.8GHz, RAM 2GB). The accuracy of the model was a little bit higher than the other two models. The problem with treebag model was that it took a lot of memory space. It was a great challenge to fit the model in system’s RAM.

**Performance**

My approach was not bad, achieving a AMS score of 2.55743. I think I may get even better result, if I can select the features more wisely.

**Improvement**

I think**,** I need to find out the dependency of the features with another, gaining a bit more domain knowledge (particle physics) may help to select the features more effectively although the challenge mention it does not require any particle physics knowledge and tuning parameters may also help to get better accuracy.