FEATURE ENGINEERING

Our feature engineering of the training and test datasets consisted of:

* Feature Selection; and
* Accounting for Structurally Missing Values

FEATURE SELECTION

To deal with the risk of overfitting, we:

1. Subtracted the PRI\_tau\_phi column from the other 4 ‘phi’ columns.
2. Rotated the angle of the remaining 4 phi columns.
3. Deleted all 5 of the non-rotated phi columns, leaving us with 4 phi columns instead of 5.

This subtraction + rotation process makes the specific pattern of the phi variables more unique, so that, once the model is trained, it is better able to discriminate between signal and background in the test set.

For more on the logic of this process, see <http://www.jmlr.org/proceedings/papers/v42/diaz14.pdf>

STRUCTURALLY MISSING VALUES

The missing values in the dataset (-999) resulted from two sources:

1. Bad estimates of the mass of Higgs boson; and
2. Jets: particles that can appear 0, 1, 2, or 3 times in an event.

To deal with this structural missingness, we:

* Converted the -999s into NAs.
* Created a Boolean column to reflect missing(T)/present(F) values in the estimated mass of the Higgs boson (DER\_mass\_MMC).
* Created 8 binary columns reflecting each combination of presence(0)/absence(1) in the estimated mass of the Higgs boson and the number of jets (0,1,2,3).
  + J0 +M1, J0+M0, J1+M1, J1+M0…
* Imputed the column mean for all the NAs.

For more on the structural nature of the missing values in the Higgs boson dataset, see <http://www.jmlr.org/proceedings/papers/v42/cowa14.pdf>.