Higgs Boson Machine Learning Challenge

Group Project

Team 2WD

Ruonan Ding

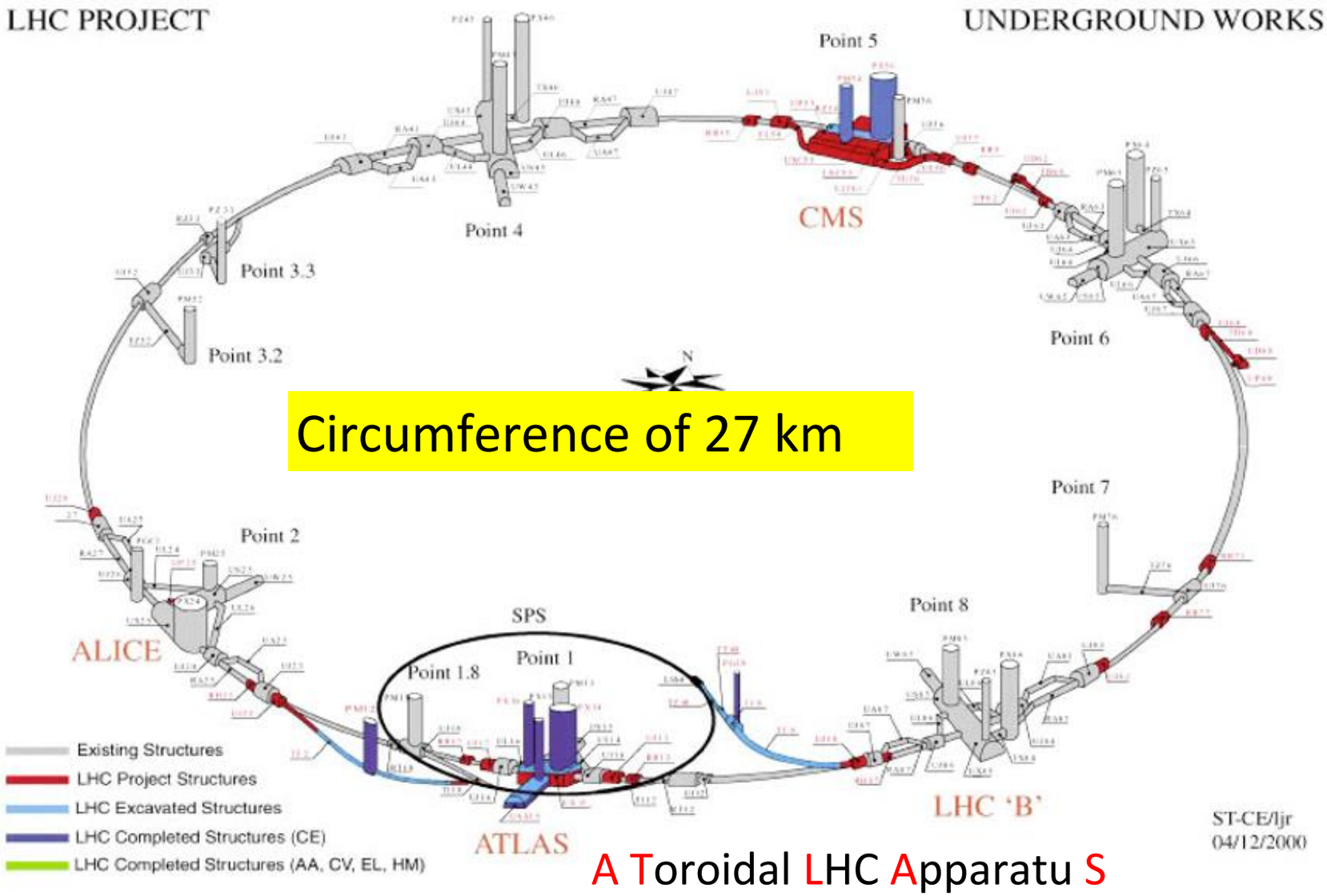
Joseph Wang

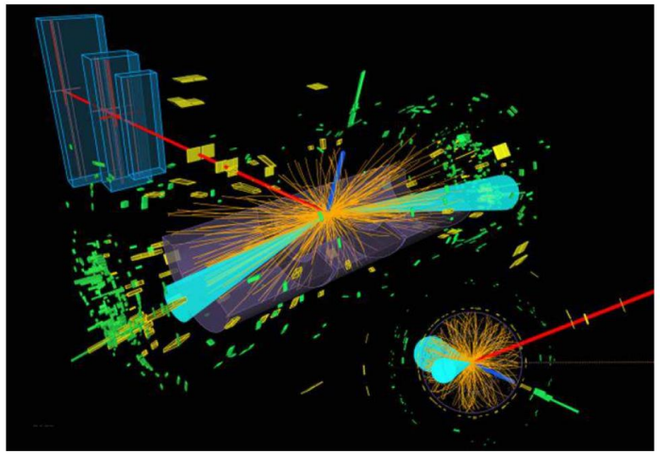
Frank Wang

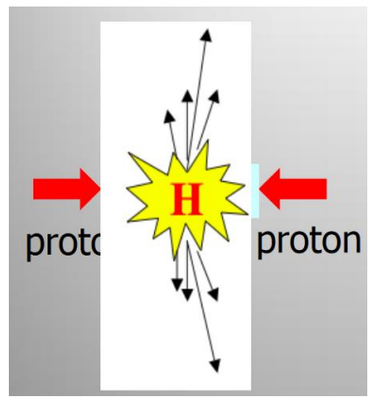
AGENDA

- Data Exploration and Preprocessing
  - data and feature overview
  - handling missing value
- Model
  - Logistic Regression with PCA
  - Random Forest
  - Gradient Boosting
  - Neural Network
  - XGboost
- Conclusion

LHC Underground Layout

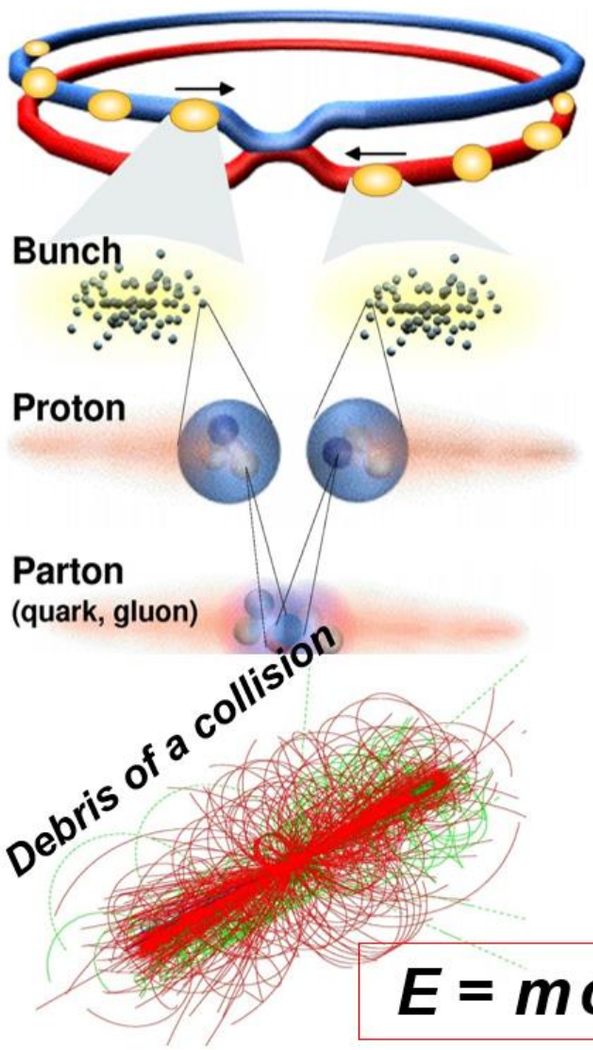






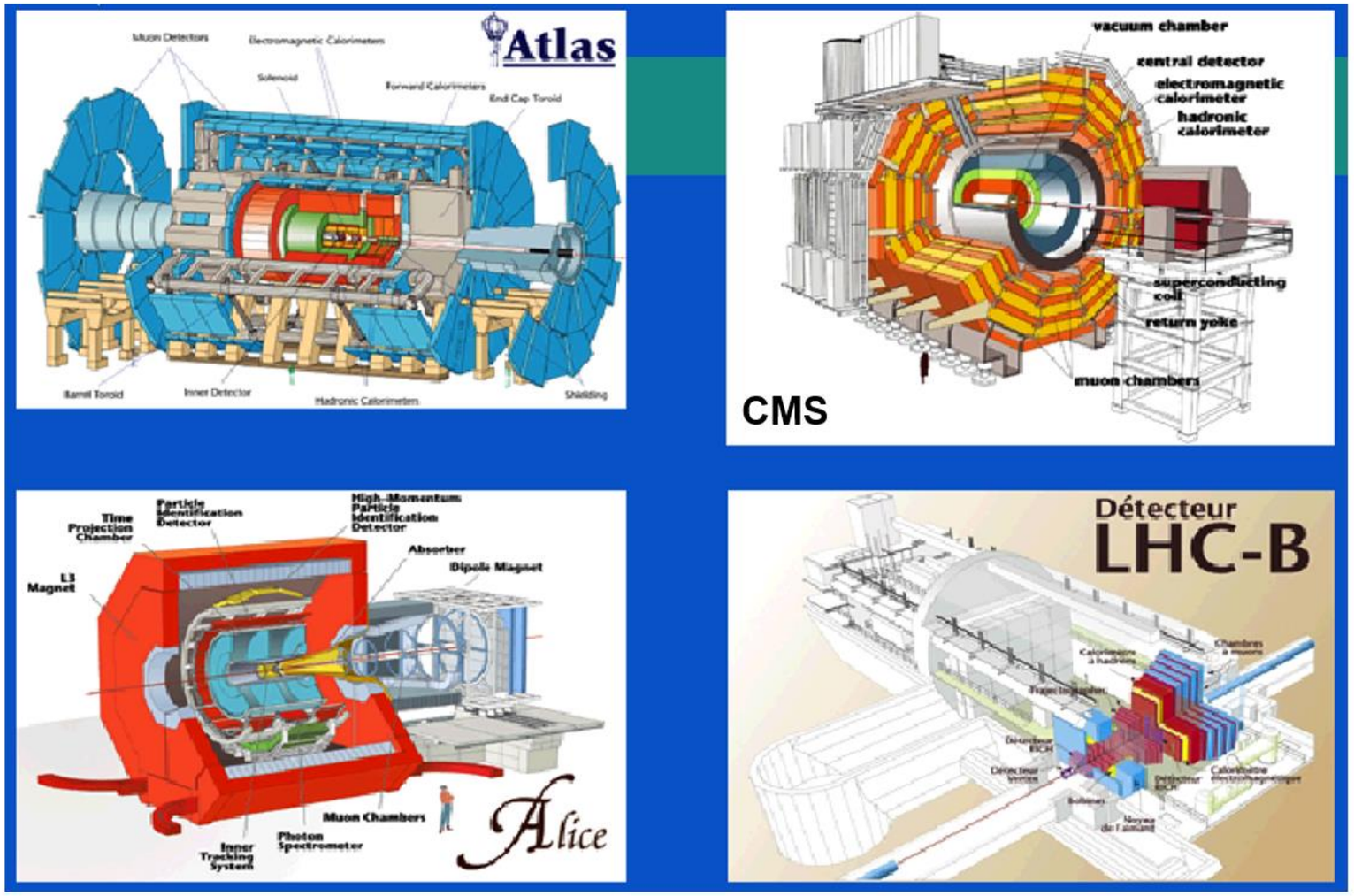
A Toroidal LHC Apparatu S

LHC



$E = mc^2$

LHC Detectors



Max proton energy: E = 7 TeV

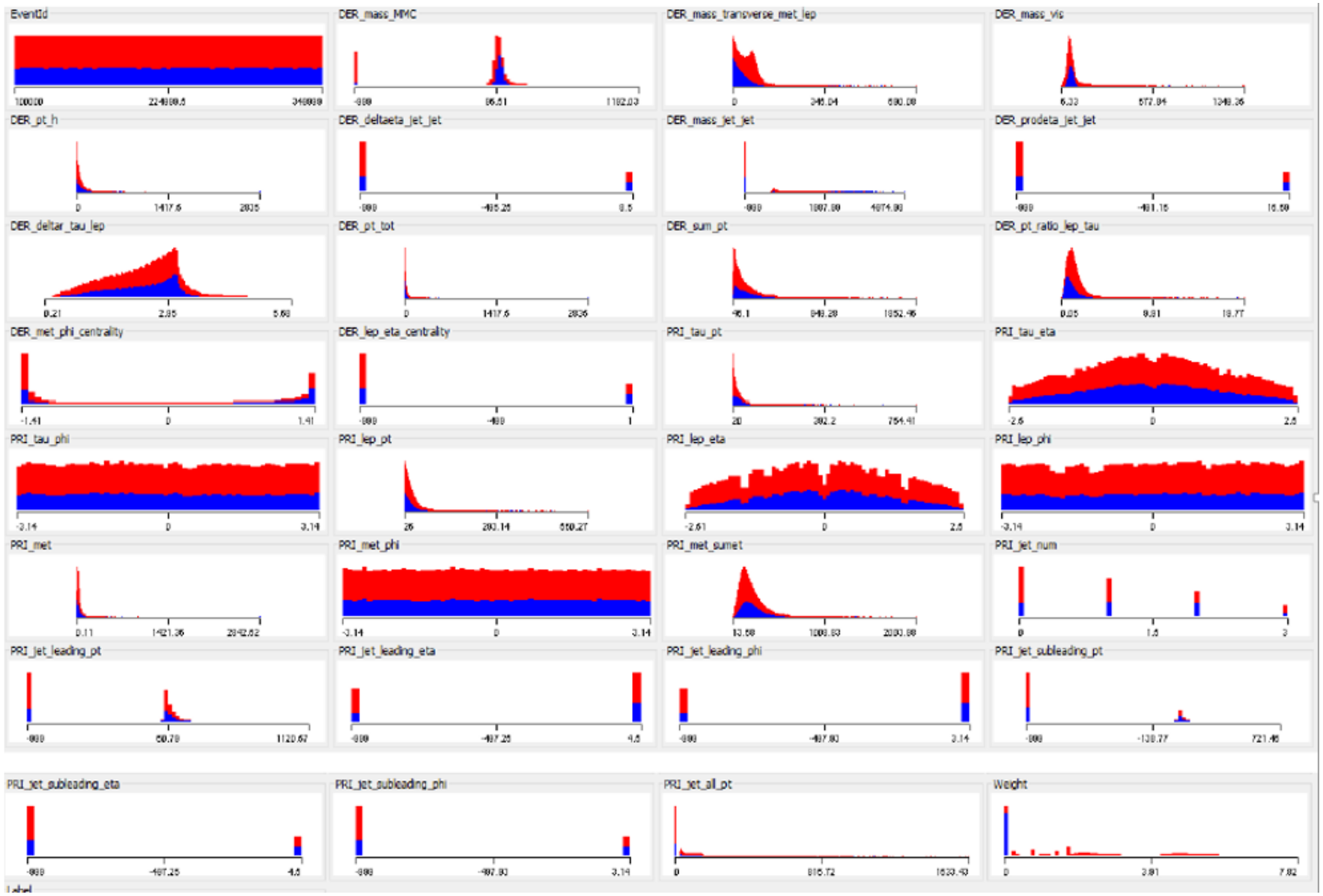
v = 0.99999991 c

Bunch diameter = 16 µm (hair 50 µm)

# of p / bunch = 10^{11}

Features

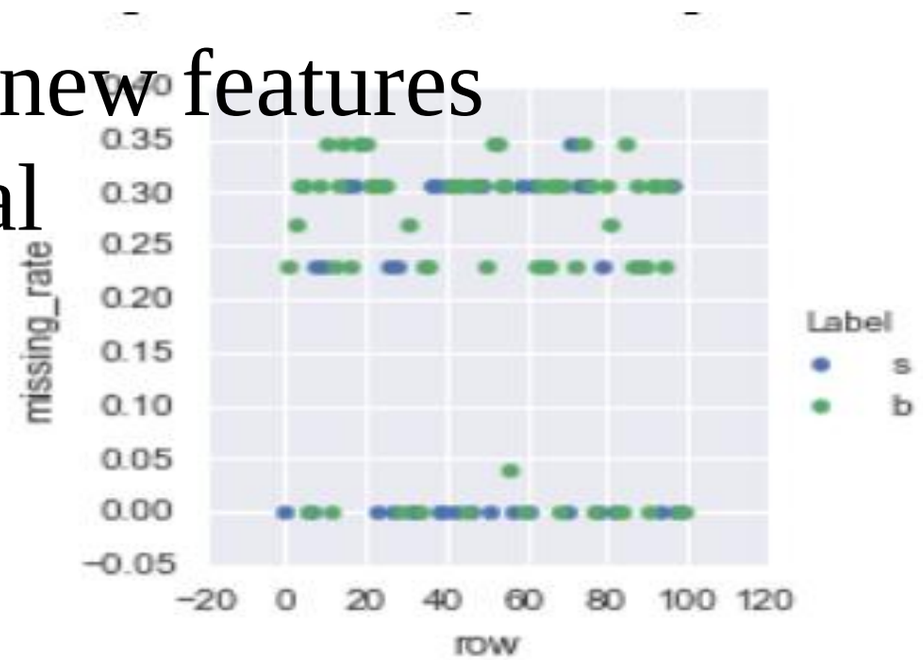
- **Missing**
- **Weak Signal and Strong background**
- **Energy sensitive**



Red: background, Blue: signal

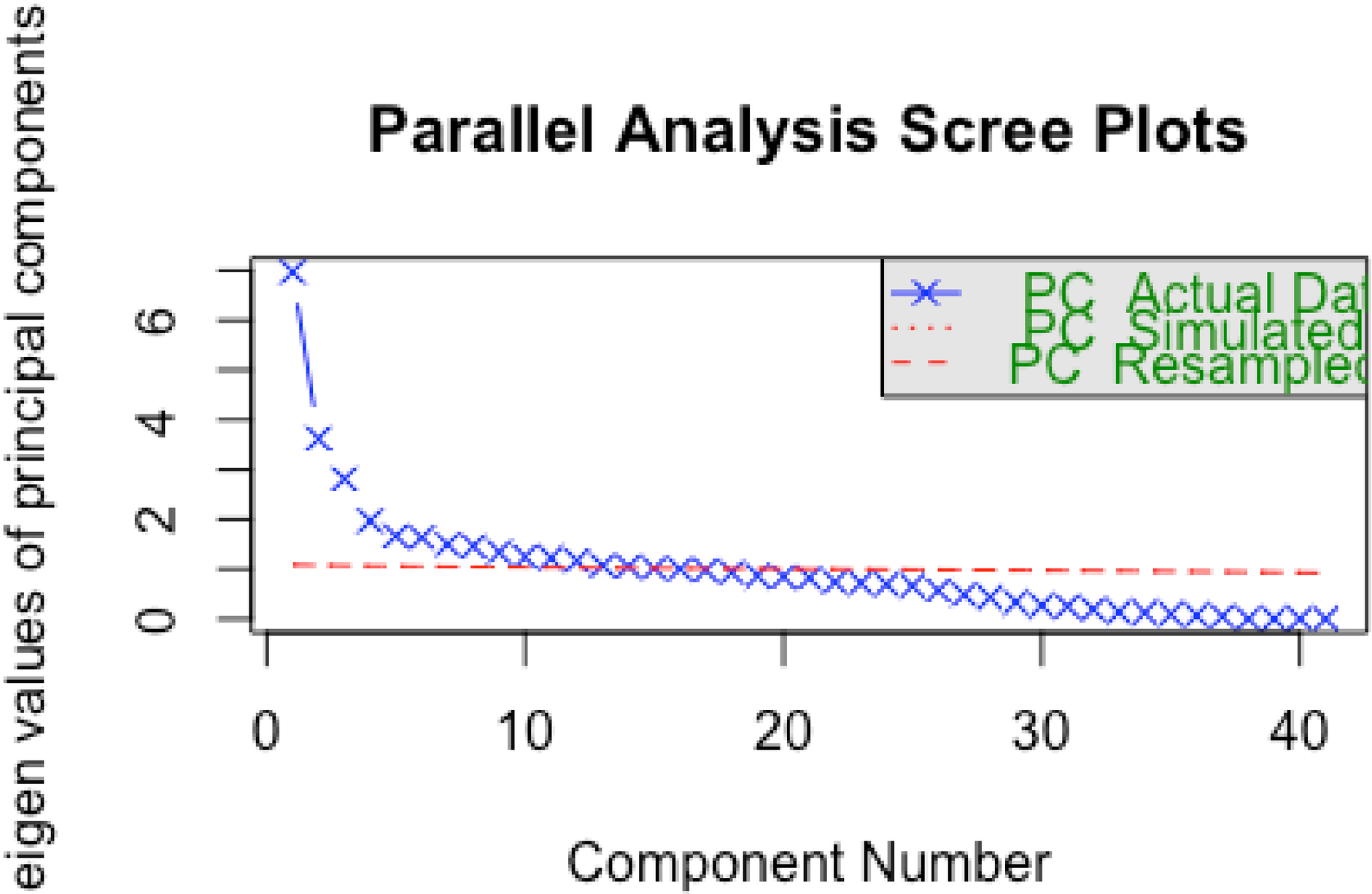
Data Features

- Large missing data
  - 7 columns missing up to 70%
  - 3 columns missing up to 40%
  - Missing data includes both signal and background
  - Each sample can has missing data up to 35%
- Adding 16 new features
- Weak signal



miss data in each col

DER_mass_MMC	0.152456
DER_mass_transverse_met_lep	0.000000
DER_mass_vis	0.000000
DER_pt_h	0.000000
DER_deltaeta_jet_jet	0.709828
DER_mass_jet_jet	0.709828
DER_prodeta_jet_jet	0.709828
DER_deltar_tau_lep	0.000000
DER_pt_tot	0.000000
DER_sum_pt	0.000000
DER_pt_ratio_lep_tau	0.000000
DER_met_phi_centrality	0.000000
DER_lep_eta_centrality	0.709828
PRI_tau_pt	0.000000
PRI_tau_eta	0.000000
PRI_tau_phi	0.000000
PRI_lep_pt	0.000000
PRI_lep_eta	0.000000
PRI_lep_phi	0.000000
PRI_met	0.000000
PRI_met_phi	0.000000
PRI_met_sumet	0.000000
PRI_jet_num	0.000000
PRI_jet_leading_pt	0.399652
PRI_jet_leading_eta	0.399652
PRI_jet_leading_phi	0.399652
PRI_jet_subleading_pt	0.709828
PRI_jet_subleading_eta	0.709828
PRI_jet_subleading_phi	0.709828
PRI_jet_all_pt	0.000000

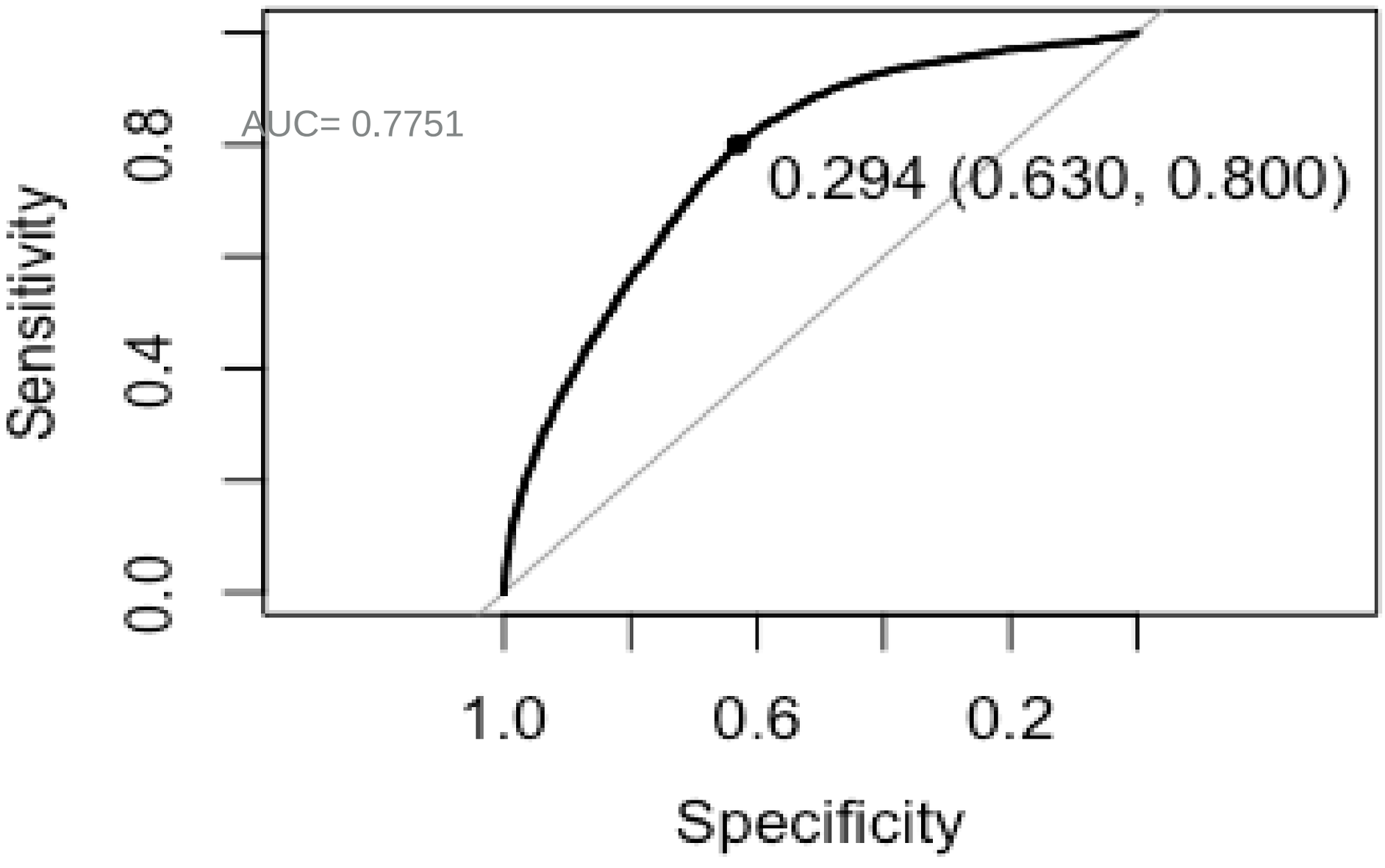


With the 42 variables we currently have, we were able to reduce the dimension down to the 15 synthetic variables. Next step is to reconstruct the given variables in terms of newly created Principle Components.

LOGISTIC REGRESSION USING DIMENSION REDUCTION

Next step is to fit a Logistic Regression after using the newly constructed Principle Components.

Using 5 folds cross-validation and repeated 5 times on the sample data. Note that the metric to optimize is accuracy in this case. This could potentially be the reason of a lower AMS in the overalls



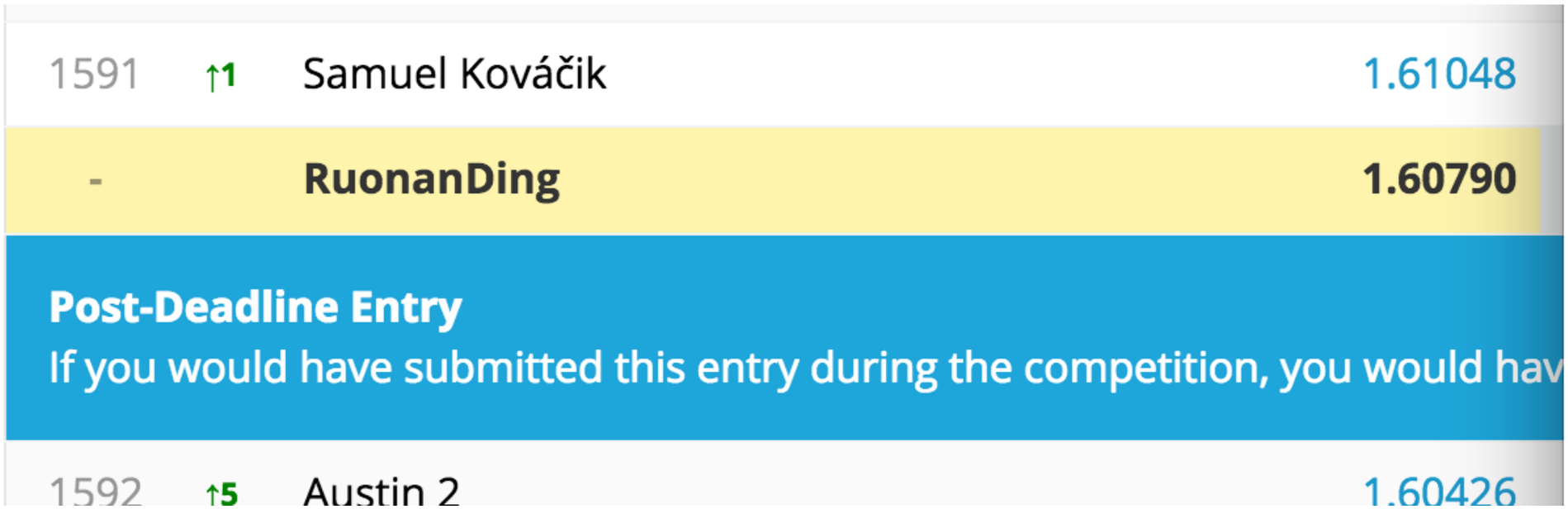
LOGISTIC REGRESSION USING DIMENSION REDUCTION

Result:

Using 80% of the df.Train: The accuracy is 68.4%. AMS score is 1.545

20% of the df.Test: The accuracy is 68.7%. AMS score is 0.781.

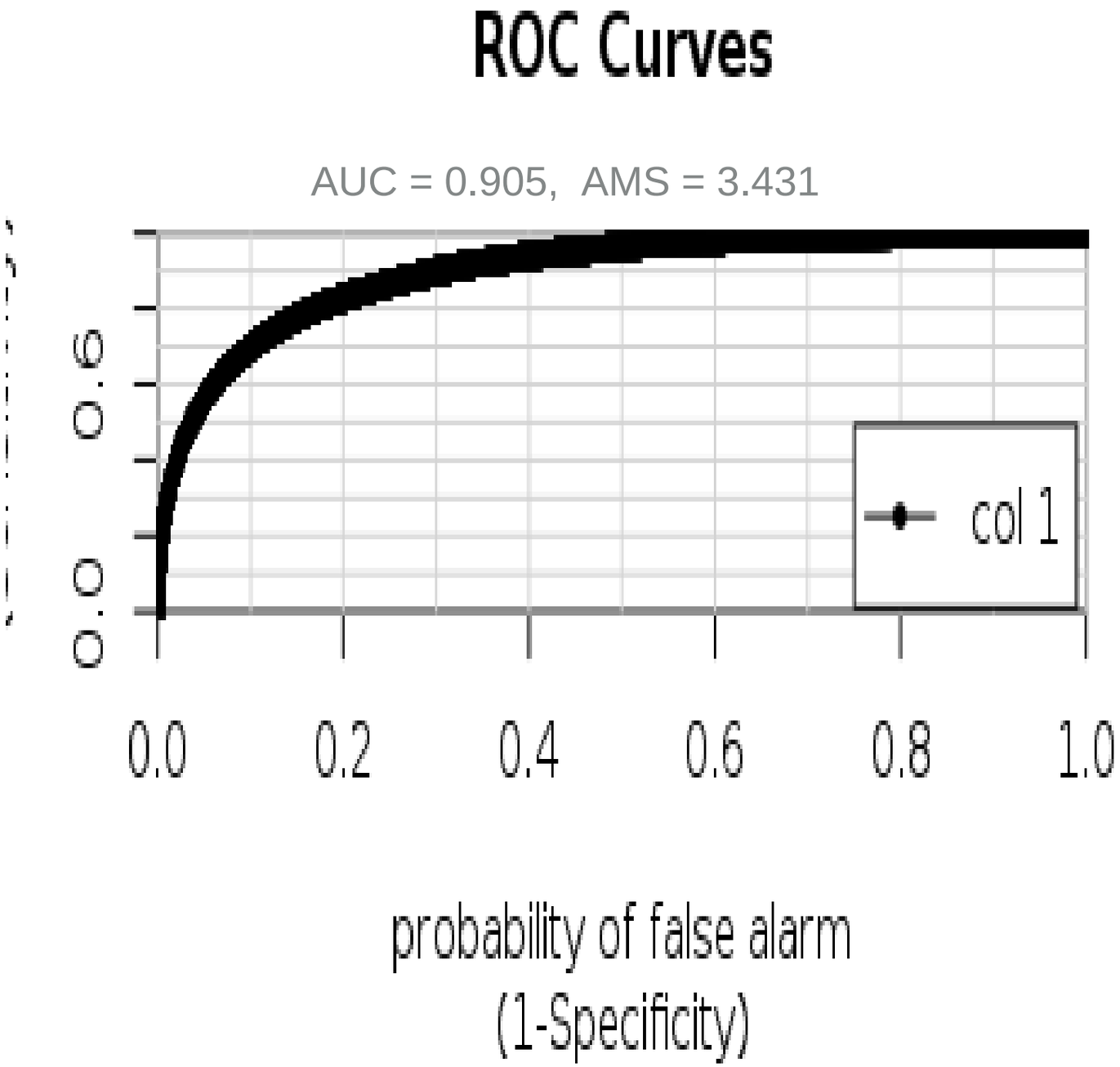
The submission result:



Learning Experience:

- For PCA, data imputation and scaling is very important. Need to validate my imputation and scaling
- In terms of model tuning, I should have used AMS. Accuracy is not the same as AMS.
- Cross validation could have done in a bigger scale for this model since the running time was ok.

GRADIENT BOOSTING MODEL



Left is the result on the training data set.

ntree= 500 ;

Learning Rate = 0.1;

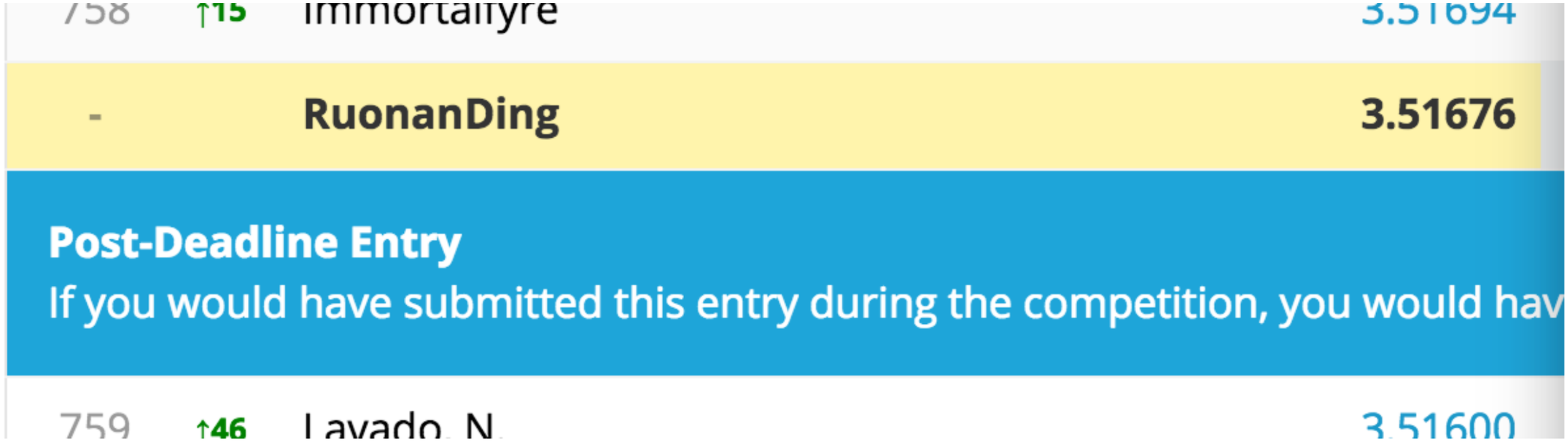
Interaction.depth = 10;

n.minobsinnode = 10;

cv.folds = 2

Performed a 2 fold cross-validation twice on the entire training data.

GRADIENT BOOSTING MODEL



Lessons learned:

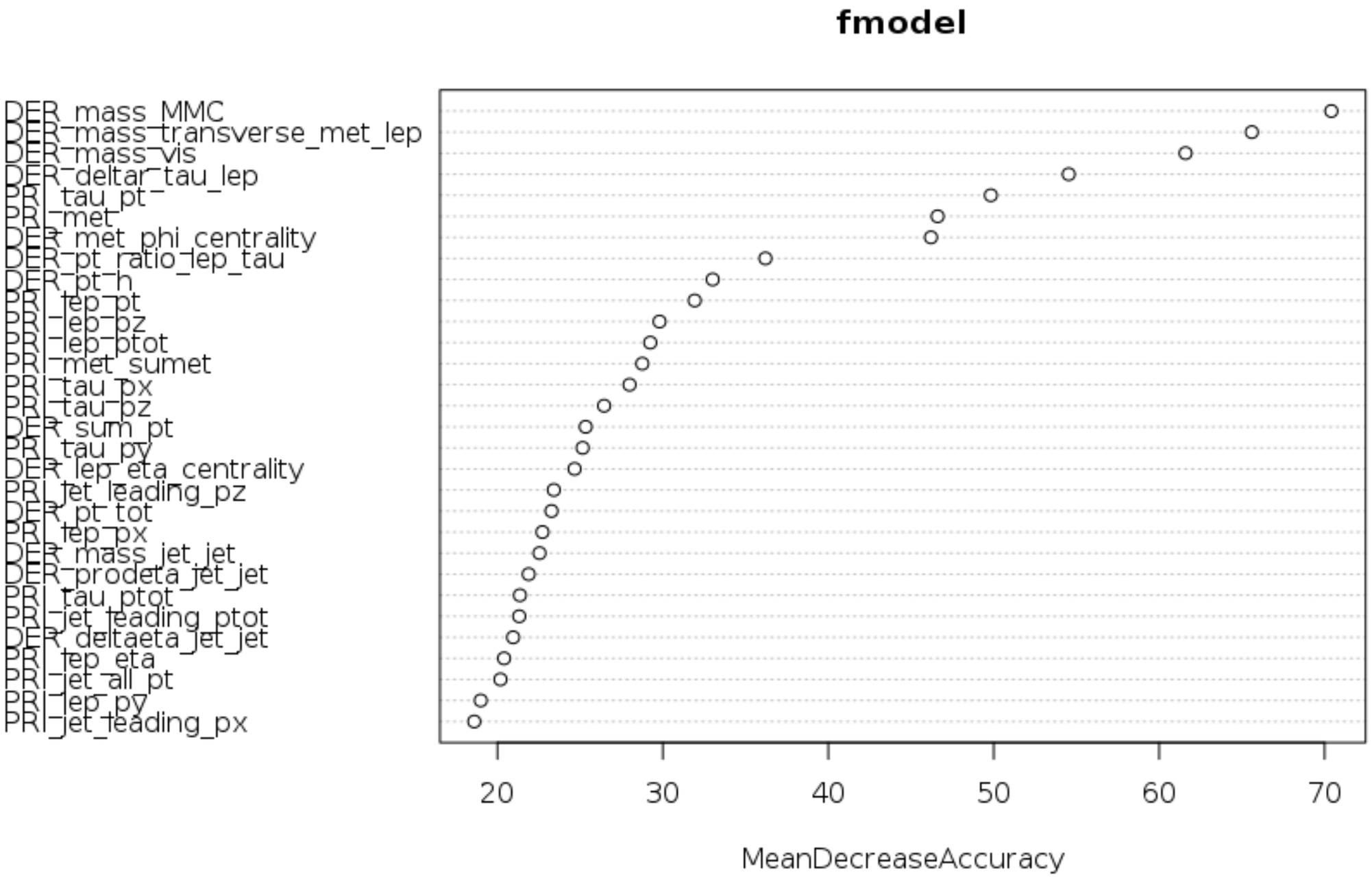
After doing some research, I make a cutoff on the probability prediction and call the upper 14% of events as signal. You can optimize this threshold to maximize the AMS. I used a testing grid. After running several tests around, the best cutoff is the top 14% (0.86).

referece: https://dbaumgartel.wordpress.com/2014/06/15/the-kaggle-higgs-challenge-beat-the-benchmarks-with-s

RANDOM FOREST

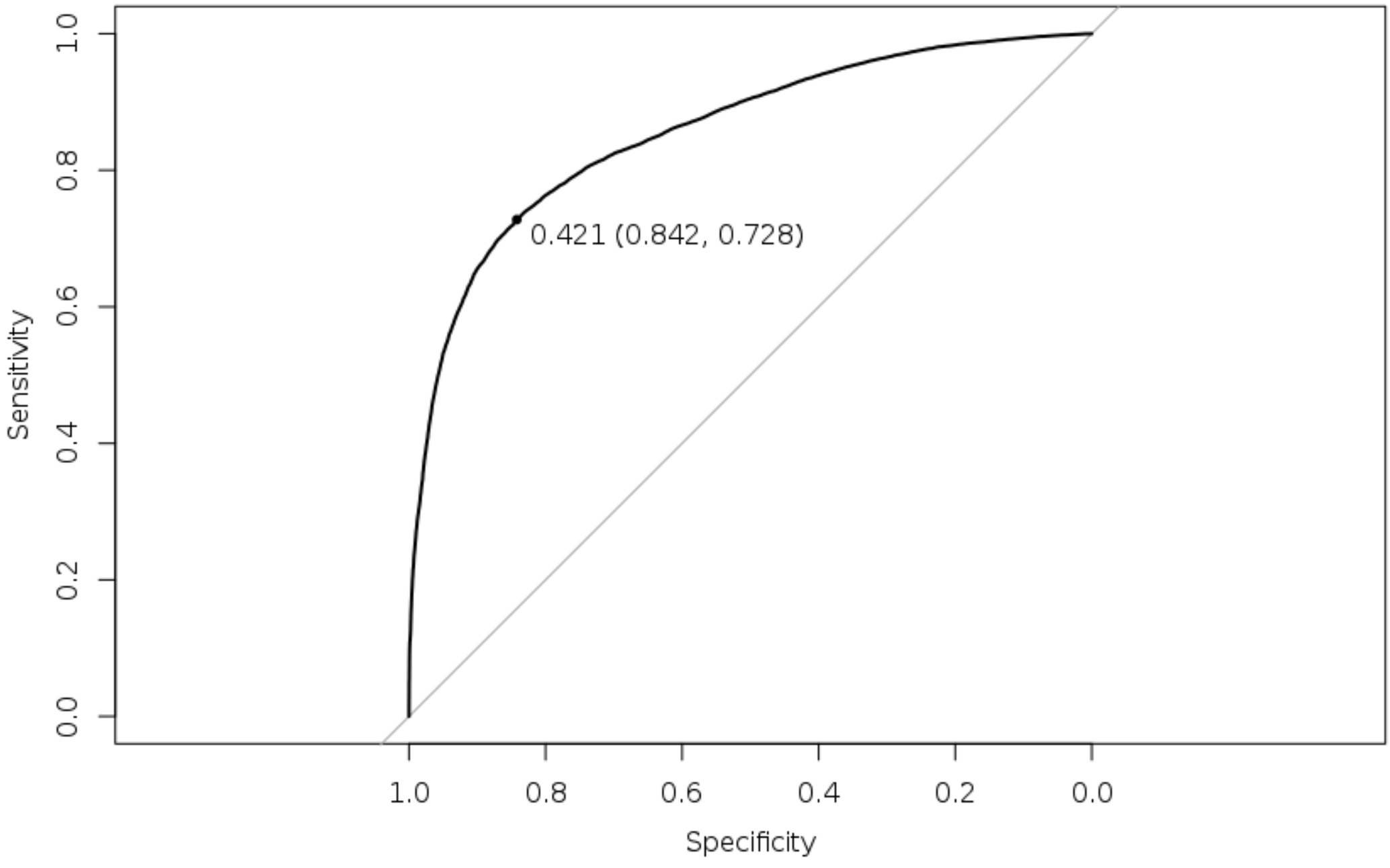
Number of trees: 500

No. of variables tried at each split: 7

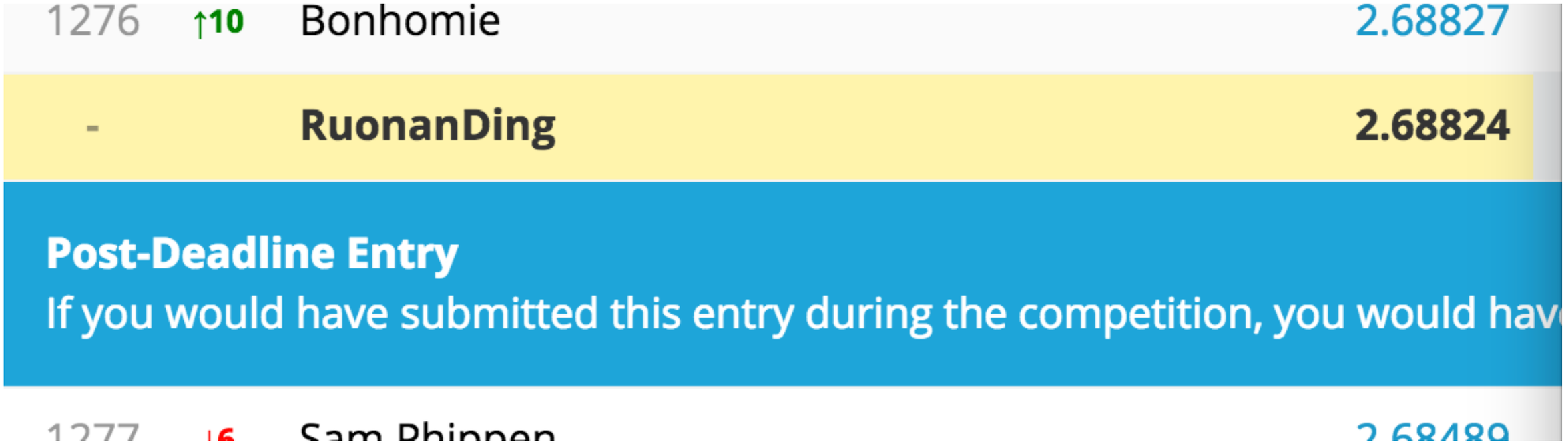


RANDOM FOREST

Test dataset : Area under the curve: 0.8563. Accuracy = 0.8026. (vs. 0.8476 in training)
AMS = 1.243051



RANDOM FOREST



Lessons Learned:

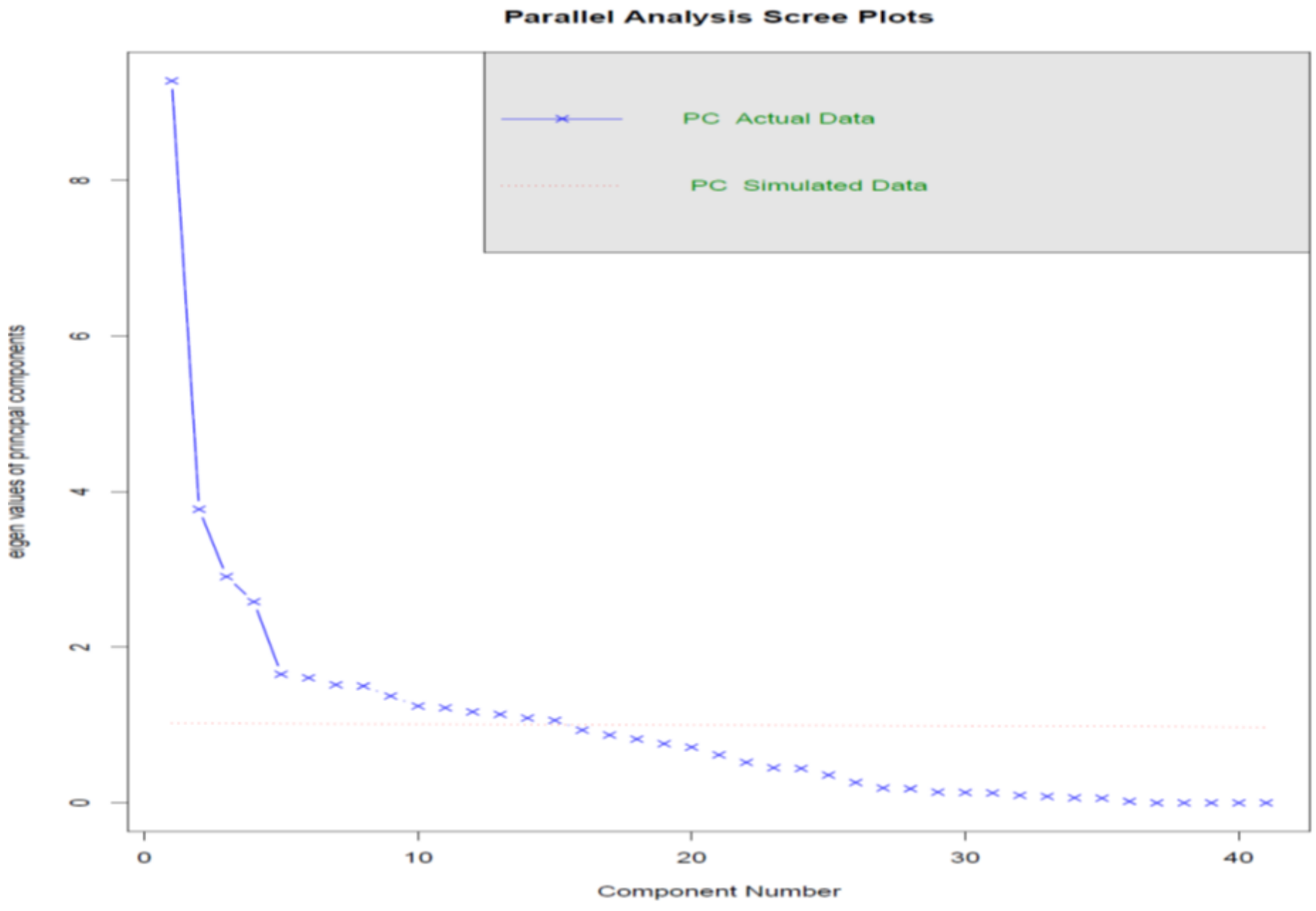
mtry matters. So does the number of tree.

After cutdown some the variables, the RF might be performing better in term of accuracy.

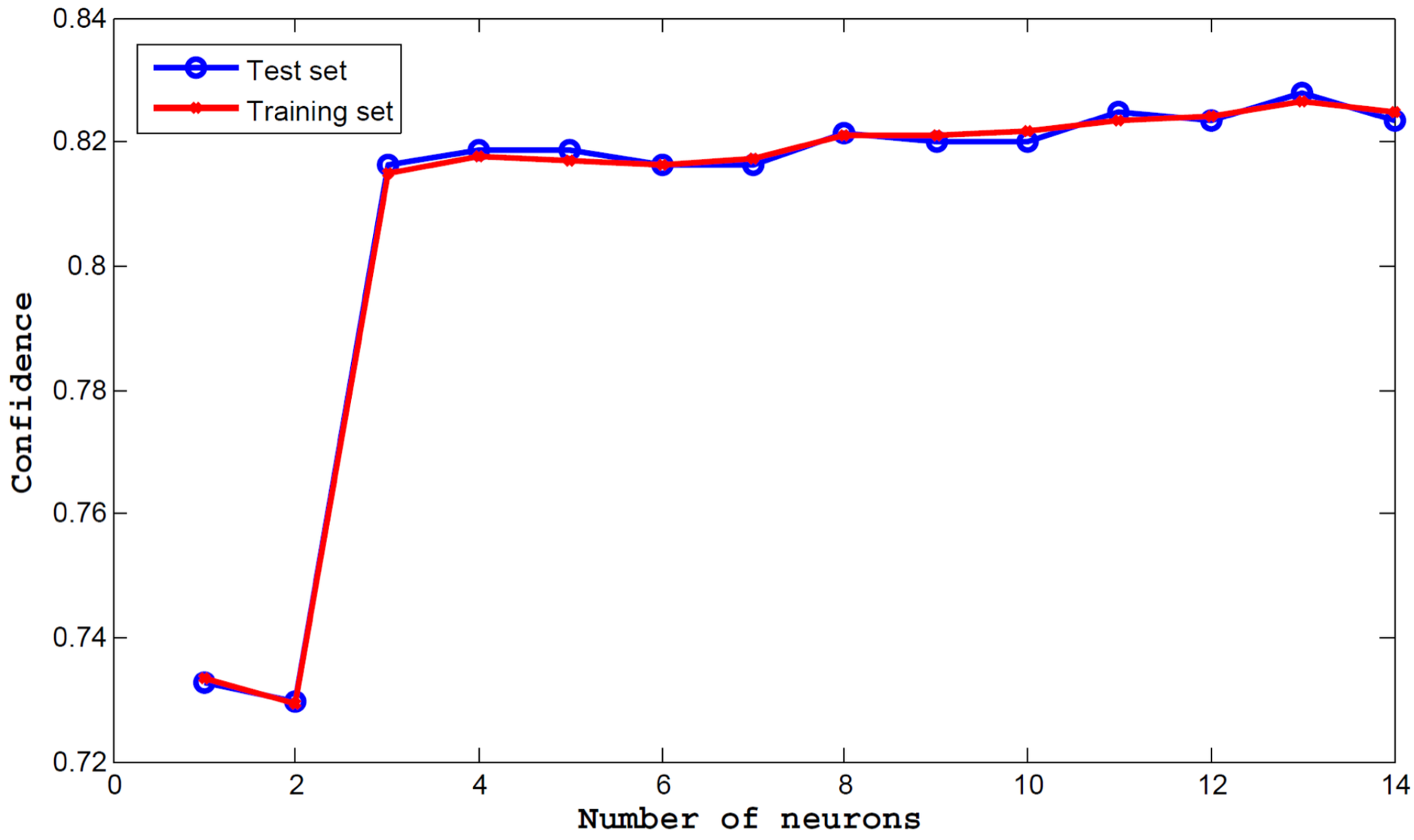
The tricky case is to find the threshold. I improved in the next try.

PCA ANALYSIS

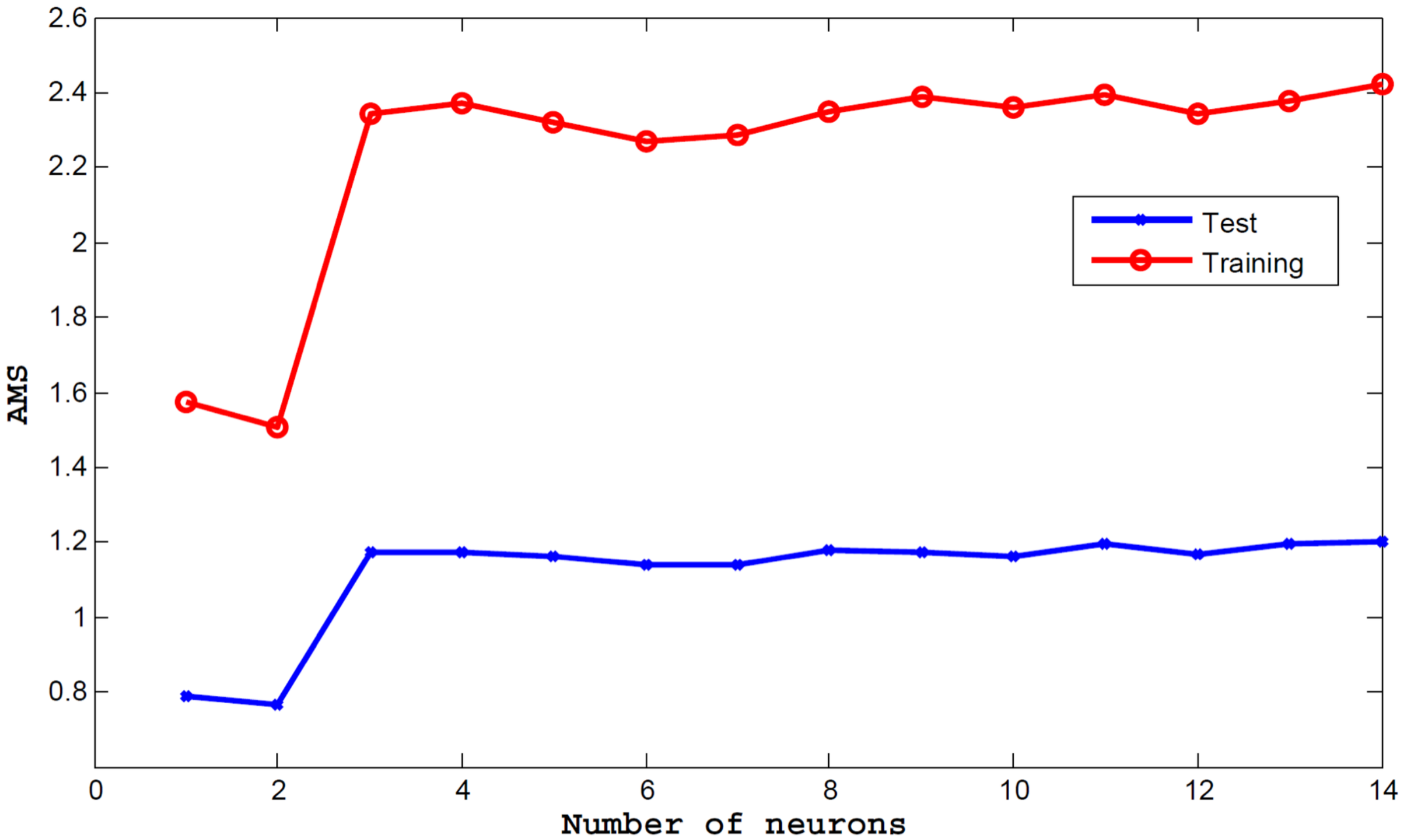
20 components are taken !!



PREDICTION MEASURES BY NEURAL NETWORK



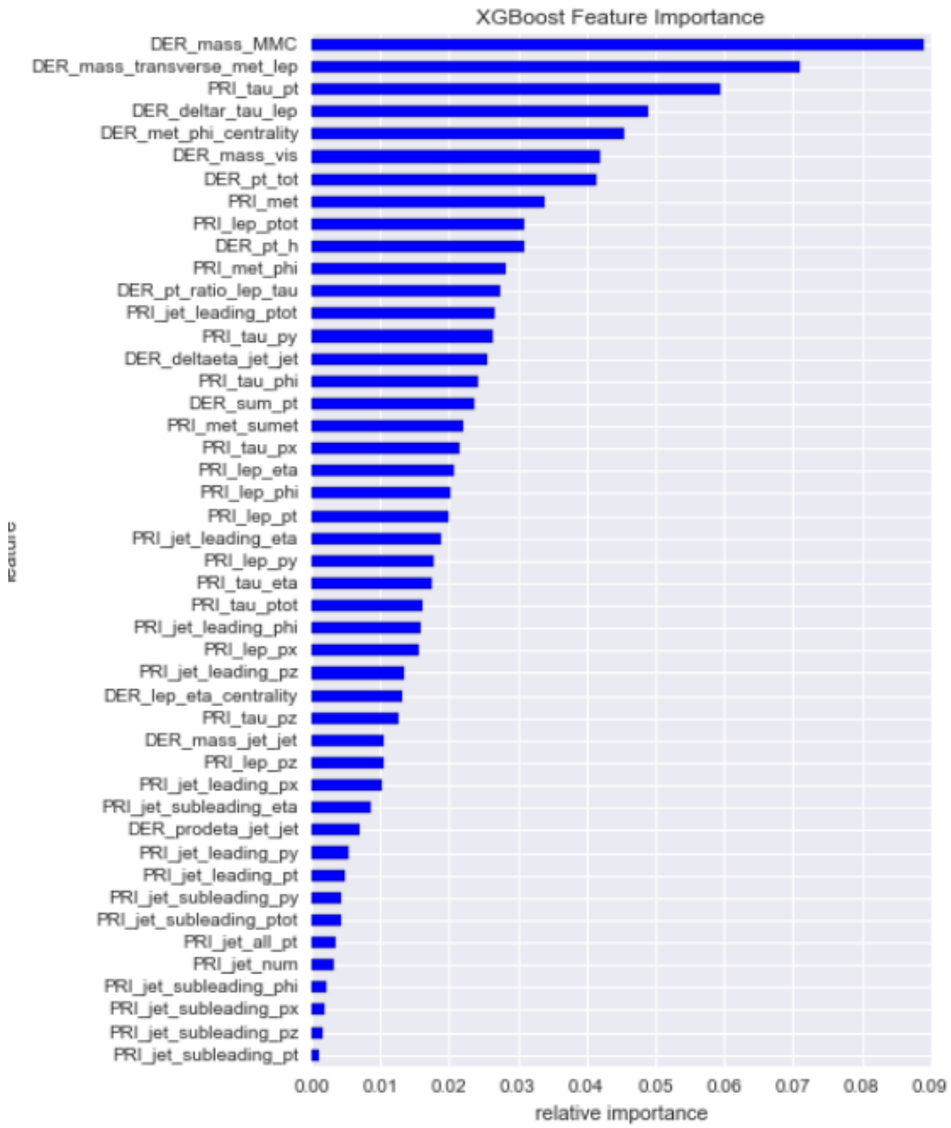
NEURAL NETWORK



CONCLUSION AND FURTHER IMPROVEMENT:

- The one layer neural network is not good enough for the data we have
- The problem is due to strong background observations than Higgs boson particles
- We can identify these weak learners and put it back to our training data to compensate the imbalance.

XGBoost



Parameters tuning:

- eta, max_depth, subsample, gamma, colsample_bytree..
- CPU time; step by step tuning
- Best AMS=3.6X

Lesson learned:

- Accuracy!=AMS
- Easy to overfitting, CV
- Variation of the score

584	↑76	AlKhwarizmi ‡	3.59685	30	Tue, 02 Sep 2014 01:12:28 (-18d)	
-		**wang frank**	**3.59660**	-	**Sun, 12 Jun 2016 18:20:00**	**Post-Deadline**

Post-Deadline Entry
If you would have submitted this entry during the competition, you would have been around here on the leaderboard.

XGBoost

Note

- Page 2(LHC layout), and 5 (data plot) can be dropped, if we have more than enough…..

CONCLUSION

- Got to practice various models.
- Learned from every submission.
- We want to explore more on every model individually.
- Didn't finish SVM tuning.
- Teamwork