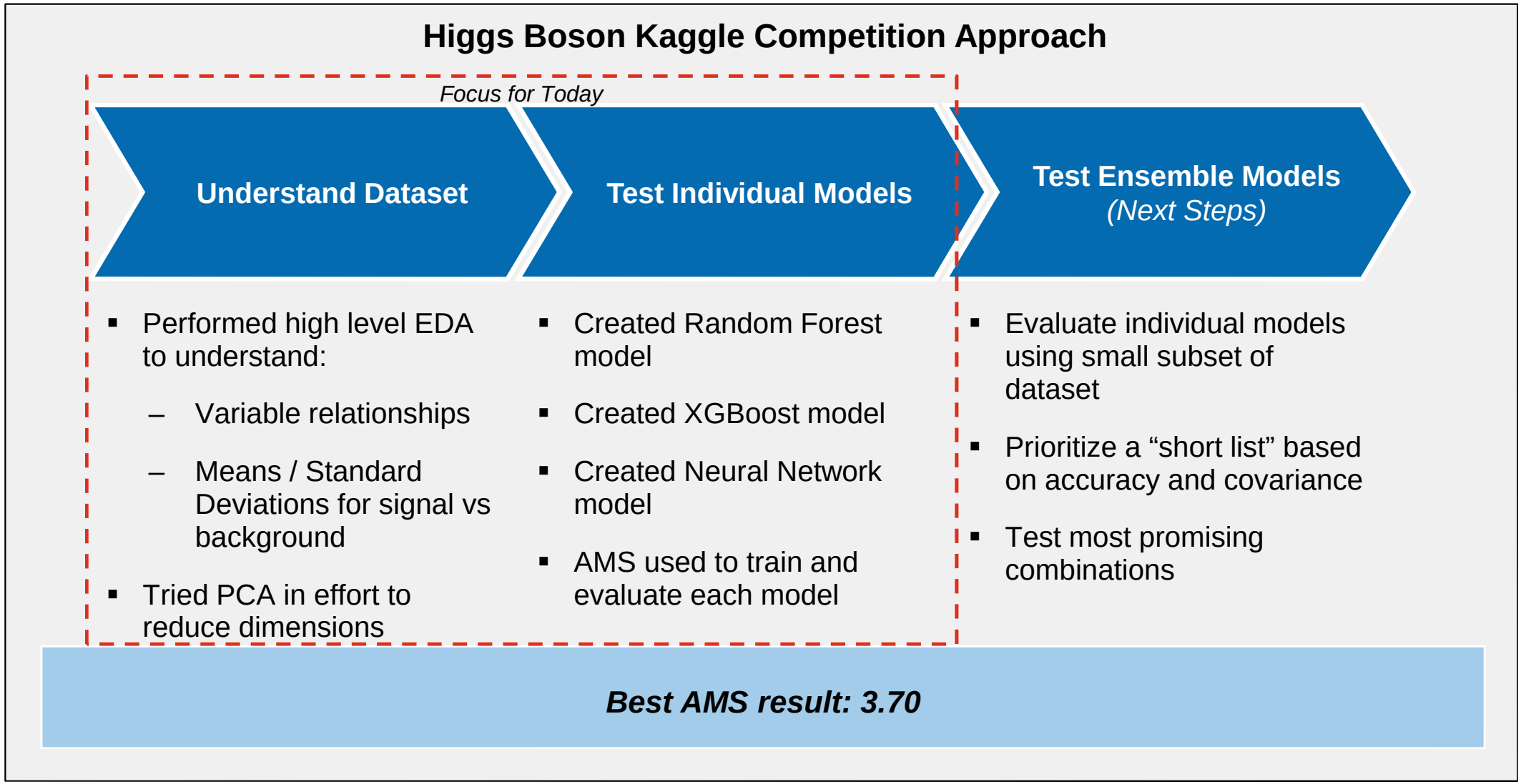
Kaggle Competition Project:
Team 7 Sigma

August 28, 2016

Our team completed 2 steps of a 3 step approach to the Higgs Boson Kaggle competition



Agenda

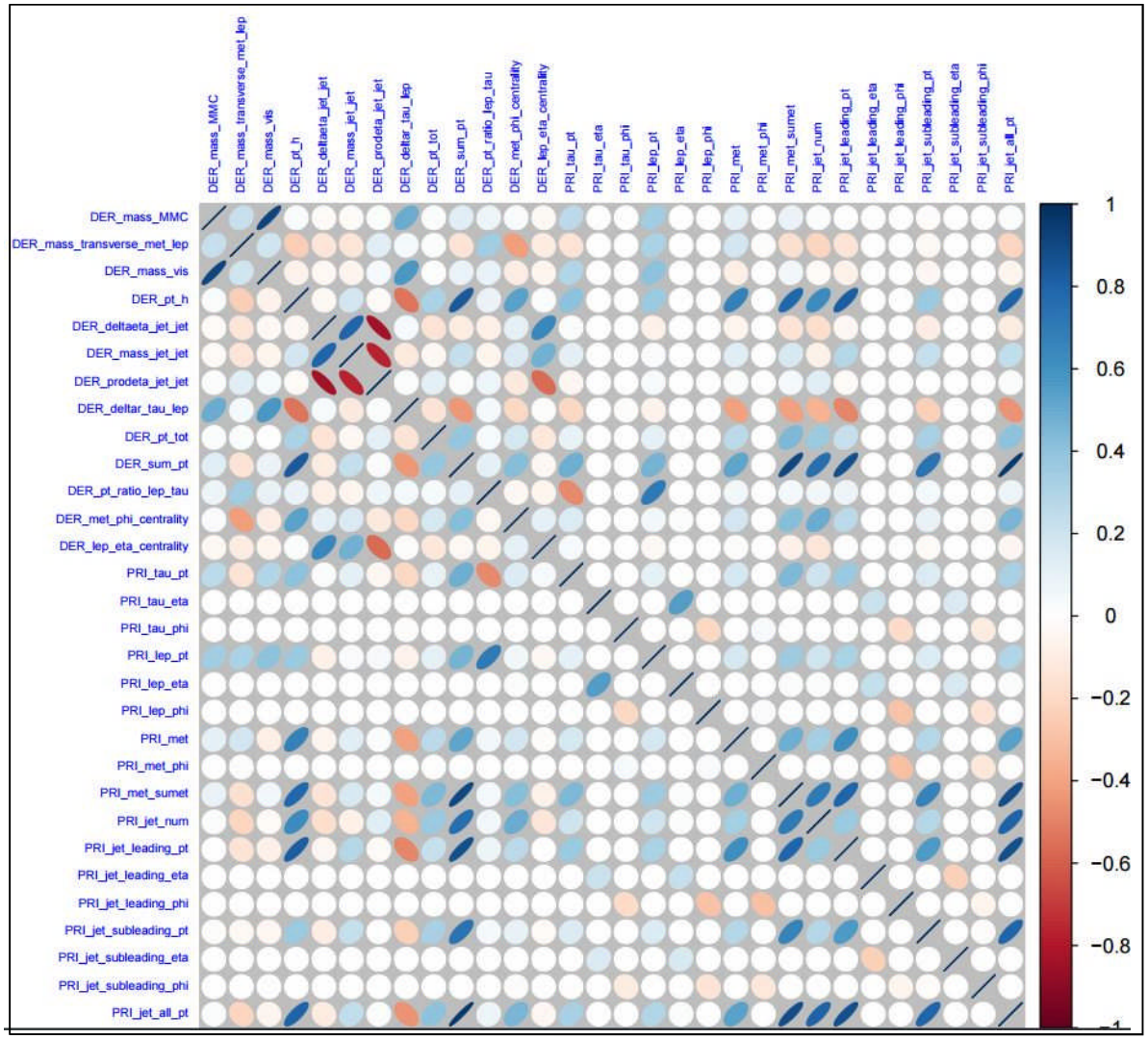
- Understand Dataset
- Test Individual Models
- Takeaways and Next Steps

Agenda

- **Understand Dataset**
- Test Individual Models
- Takeaways and Next Steps

Examining the relationships between variables reveals several strong correlations

Variable Correlation Matrix



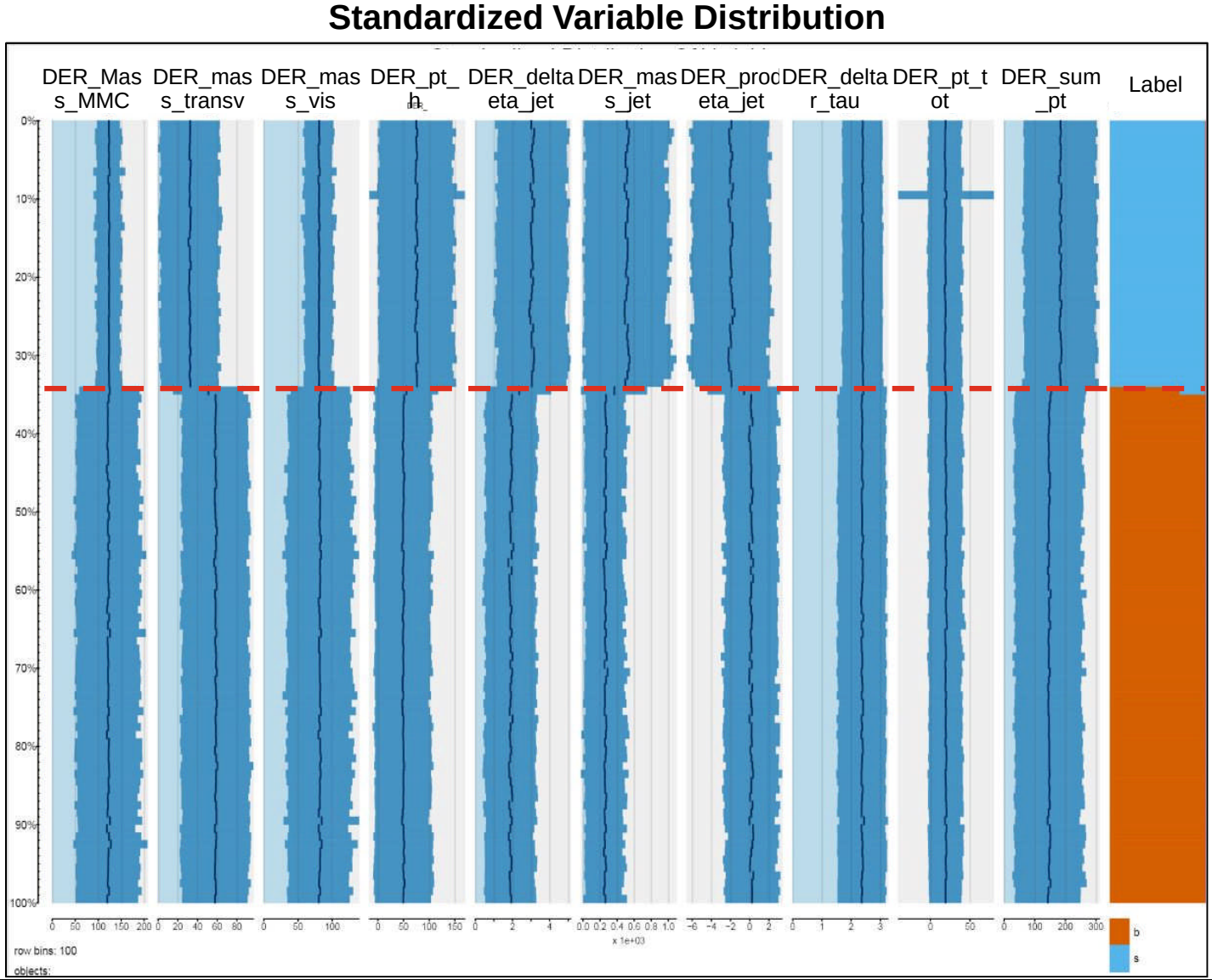
Strong Positive Correlations

- DER_deltaeta_jet and DER_mass_jet
- DER_mass_MMC and DER_mass_vis
- Most PRI and DER jet fields

Strong Negative Correlations

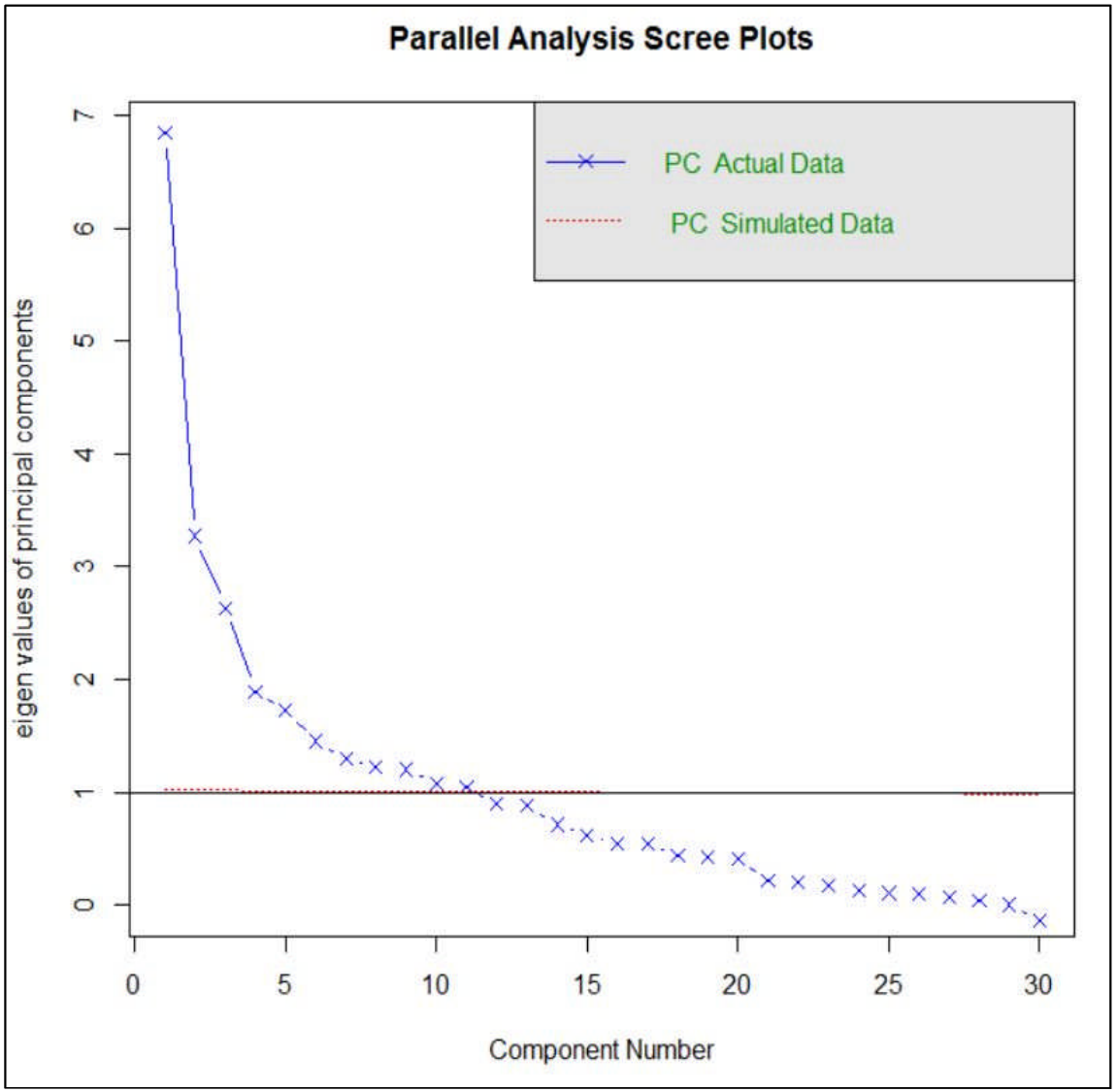
- DER_deltaeta_jet and DER_prodeta_jet
- DER_mass_jet and DER_prodeta_jet
- DER_lep_eta_centrality and DER_prodeta_jet

Mean and Standard Deviations of many variables tend to vary for signal vs background events



- ***Top part of distribution is signals***
- ***Most variables have noticeable difference in either mean or standard deviation comparing signal to background***
- ***Again, mass variables tend to vary greatly for signal vs background***

PCA results indicate it will be difficult to reduce dimensions as first principal component explains just 23% of variance

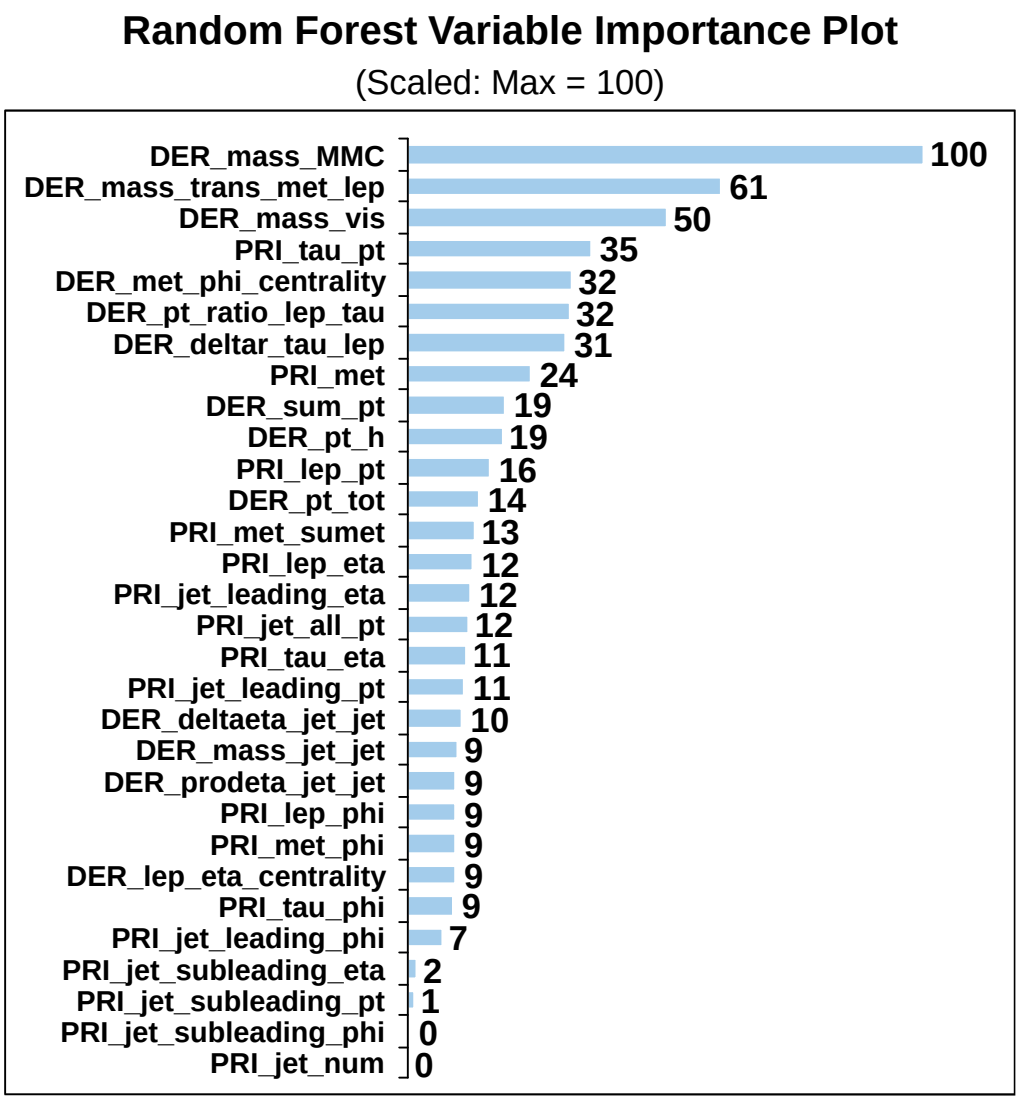


- ***Reducing to 10 principal components explains just 75% of variance***
- ***First principal component explains only 23% of variance***
- ***Will be difficult to eliminate variables → implies meaningful information contained***

Agenda

- Understand Dataset
- **Test Individual Models**
- Takeaways and Next Steps

First individual model attempted was a Random Forest which was not highly predictive but highlighted importance of mass variables

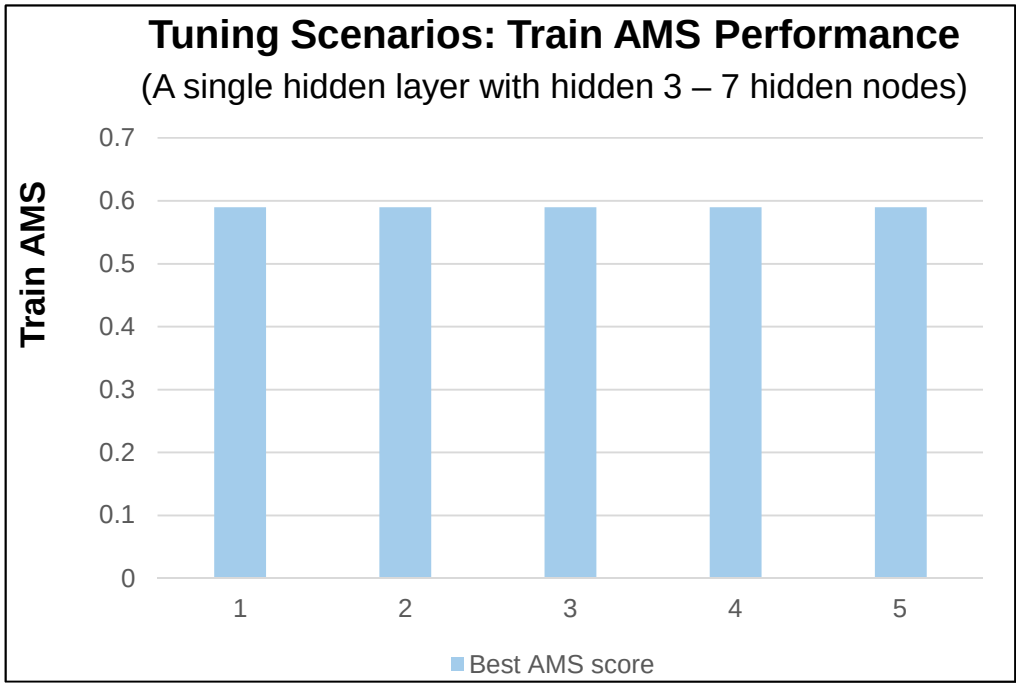


Discussion

- Trained model using:
  - 250,000 samples
  - Cross-validation (2 fold, repeated once)
  - 3 different “mtry”s (2, 16, 30)
- Algorithm selected in an attempt to:
  - Predict response on test set
  - Understand variable importance
- Relatively low AMS implies not a great model for prediction
- Seems that mass related variables are most important

Test AMS = 2.10

Neural Networks generally yielded poor results; cross validation tuning with large hidden node values was unsuccessful given time



Hidden Nodes	3	4	5	6	7
Best Threshold	1	1	1	1	1
Best AMS score	0.59	0.59	0.59	0.59	0.59

Discussion: Tuning Scenarios

- Interesting to note that for 3-7 hidden nodes the AMS is the same
- A single hidden layer with 20 hidden nodes yielded **0.59** AMS as well
- It appears that although the function to cross validate different hidden nodes and threshold work, all models converge on threshold of 0.6 and predict all background

Next Steps:

- Manually tune the best training model (baseline) by adjusting the number of hidden nodes on single hidden layer
- Unable to complete 20-50 hidden node cross validation on training data because of time

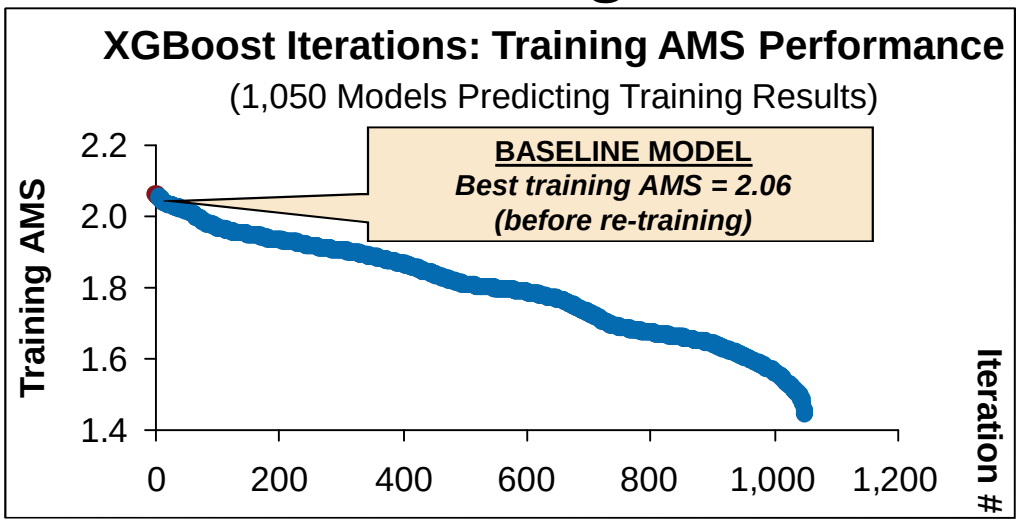
Detail: with 2 nodes, best Neural Network AMS produced ~0.6

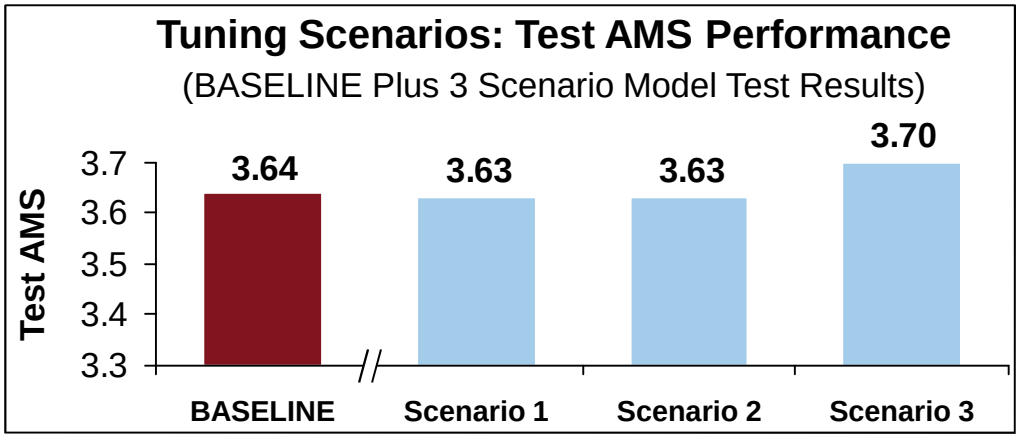
Addendum: Neural Network Tuning Output

(Max Nodes = 2)

```
                          AMS
2.0000000 0.7750000 0.5245716
[1] "Predicting layer with 2 nodes & threshold:  0.8"
                          AMS
2.0000000 0.8000000 0.5287238
[1] "Predicting layer with 2 nodes & threshold:  0.825"
                          AMS
2.0000000 0.8250000 0.5414986
[1] "Predicting layer with 2 nodes & threshold:  0.85"
                          AMS
2.0000000 0.8500000 0.5507575
[1] "Predicting layer with 2 nodes & threshold:  0.875"
                          AMS
2.0000000 0.8750000 0.5543306
[1] "Predicting layer with 2 nodes & threshold:  0.9"
                          AMS
2.0000000 0.9000000 0.5649238
[1] "Predicting layer with 2 nodes & threshold:  0.925"
                         AMS
2.000000 0.925000 0.569943
[1] "Predicting layer with 2 nodes & threshold:  0.95"
                          AMS
2.0000000 0.9500000 0.5795095
[1] "Predicting layer with 2 nodes & threshold:  0.975"
                          AMS
2.0000000 0.9750000 0.5881239
[1] "Predicting layer with 2 nodes & threshold:  1"
                          AMS
2.0000000 1.0000000 0.5949289
[1] "Choosing best model with ams score: 0.594928948248174"
```

XGBoost produced a much better model – particularly after tuning to avoid overfitting – and resulted in a test AMS of 3.64 to 3.70





ETA	0.10	0.10	0.12	0.10
Depth	9	9	10	10
Rounds	85	75	85	75

Discussion: 1,050 Iterations

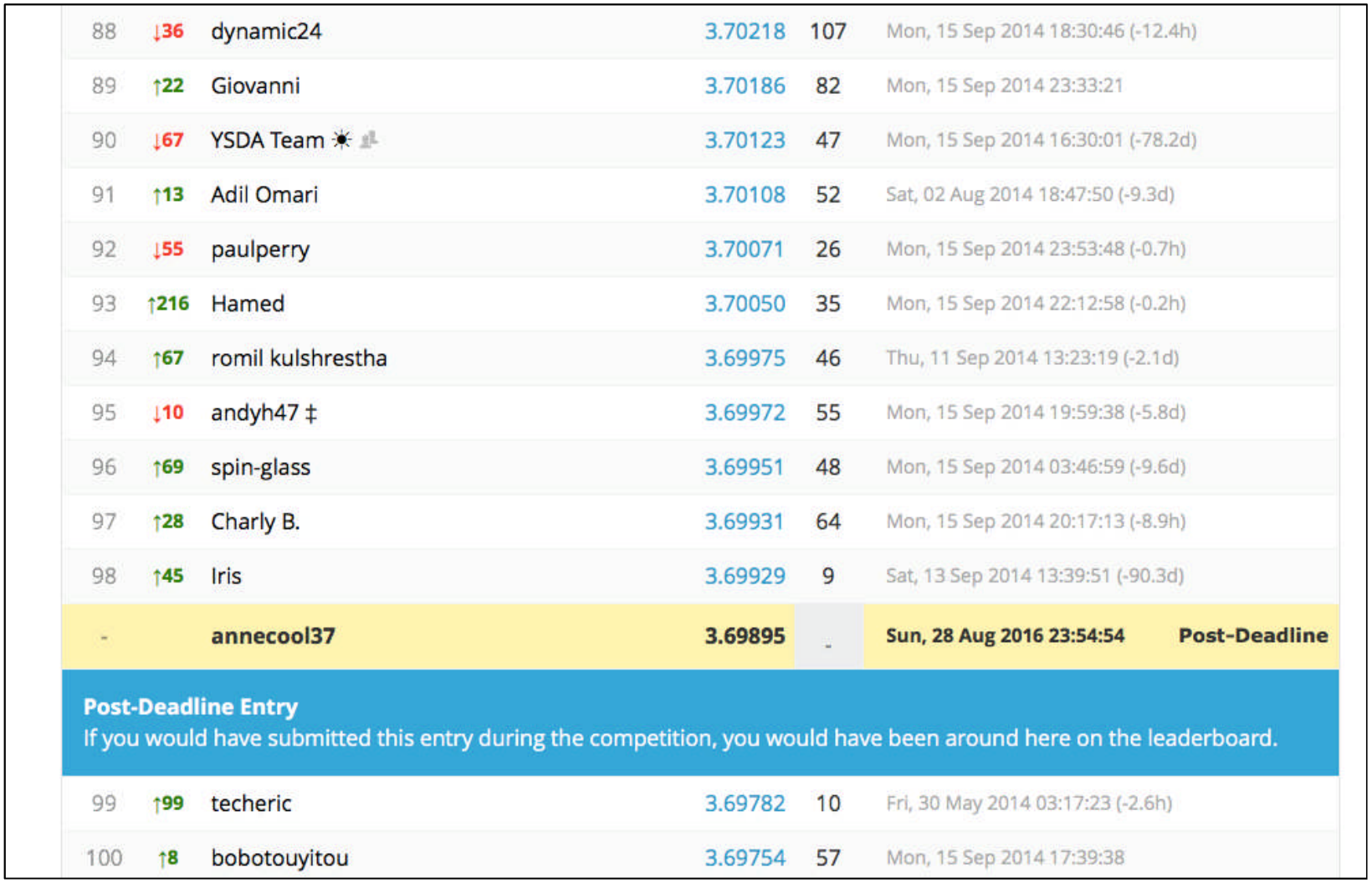
- Ran 1,050 iterations of XGBoost model, varying parameters
- Best training AMS was **2.06**, but when re-trained against all observations, improved to **4.89** (training AMS)
- However, AMS drops to **3.64** against test dataset (used as baseline for re-tuning)

Discussion: Tuning Scenarios

- Next, manually tuned the best training model (baseline) by varying parameters
- Unable to systematically submit large number of versions to Kaggle, so manually selected 3 parameter combinations
- Best one produced test AMS = **3.7**; reduced overfitting by decreasing # of rounds

Best XGBoost predictions yielded a top 100 Kaggle score

Kaggle Leaderboard Submission Ranking



88	↓36	dynamic24	3.70218	107	Mon, 15 Sep 2014 18:30:46 (-12.4h)	
89	↑22	Giovanni	3.70186	82	Mon, 15 Sep 2014 23:33:21	
90	↓67	YSDA Team ☀	3.70123	47	Mon, 15 Sep 2014 16:30:01 (-78.2d)	
91	↑13	Adil Omari	3.70108	52	Sat, 02 Aug 2014 18:47:50 (-9.3d)	
92	↓55	paulperry	3.70071	26	Mon, 15 Sep 2014 23:53:48 (-0.7h)	
93	↑216	Hamed	3.70050	35	Mon, 15 Sep 2014 22:12:58 (-0.2h)	
94	↑67	romil kulshrestha	3.69975	46	Thu, 11 Sep 2014 13:23:19 (-2.1d)	
95	↓10	andyh47 ‡	3.69972	55	Mon, 15 Sep 2014 19:59:38 (-5.8d)	
96	↑69	spin-glass	3.69951	48	Mon, 15 Sep 2014 03:46:59 (-9.6d)	
97	↑28	Charly B.	3.69931	64	Mon, 15 Sep 2014 20:17:13 (-8.9h)	
98	↑45	Iris	3.69929	9	Sat, 13 Sep 2014 13:39:51 (-90.3d)	
-		**annecool37**	**3.69895**	-	**Sun, 28 Aug 2016 23:54:54**	**Post-Deadline**

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

99	↑99	techeric	3.69782	10	Fri, 30 May 2014 03:17:23 (-2.6h)	
100	↑8	bobotouyitou	3.69754	57	Mon, 15 Sep 2014 17:39:38	

Agenda

- Understand Dataset
- Test Individual Models
- **Takeaways and Next Steps**

Takeaways and Next Steps

Takeaways

- EDA indicates that mass related variables are most important
- EDA shows mean and standard deviation indicate that of many variables tend to vary for signal vs background events
- Tuned xgbBoost model yield highest AMS score of our tested models

Next Steps

- Continue to tune neural network model
- Attempt ensemble model given best neural net and xgbBoost models
- Attempt other ensemble model combinations