3-MEAN BUSTERS
BEN TOWNSON, DAVID STEINMETZ, SHARAN DUGGAL

HIGGS-BOSON KAGGLE PROJECT

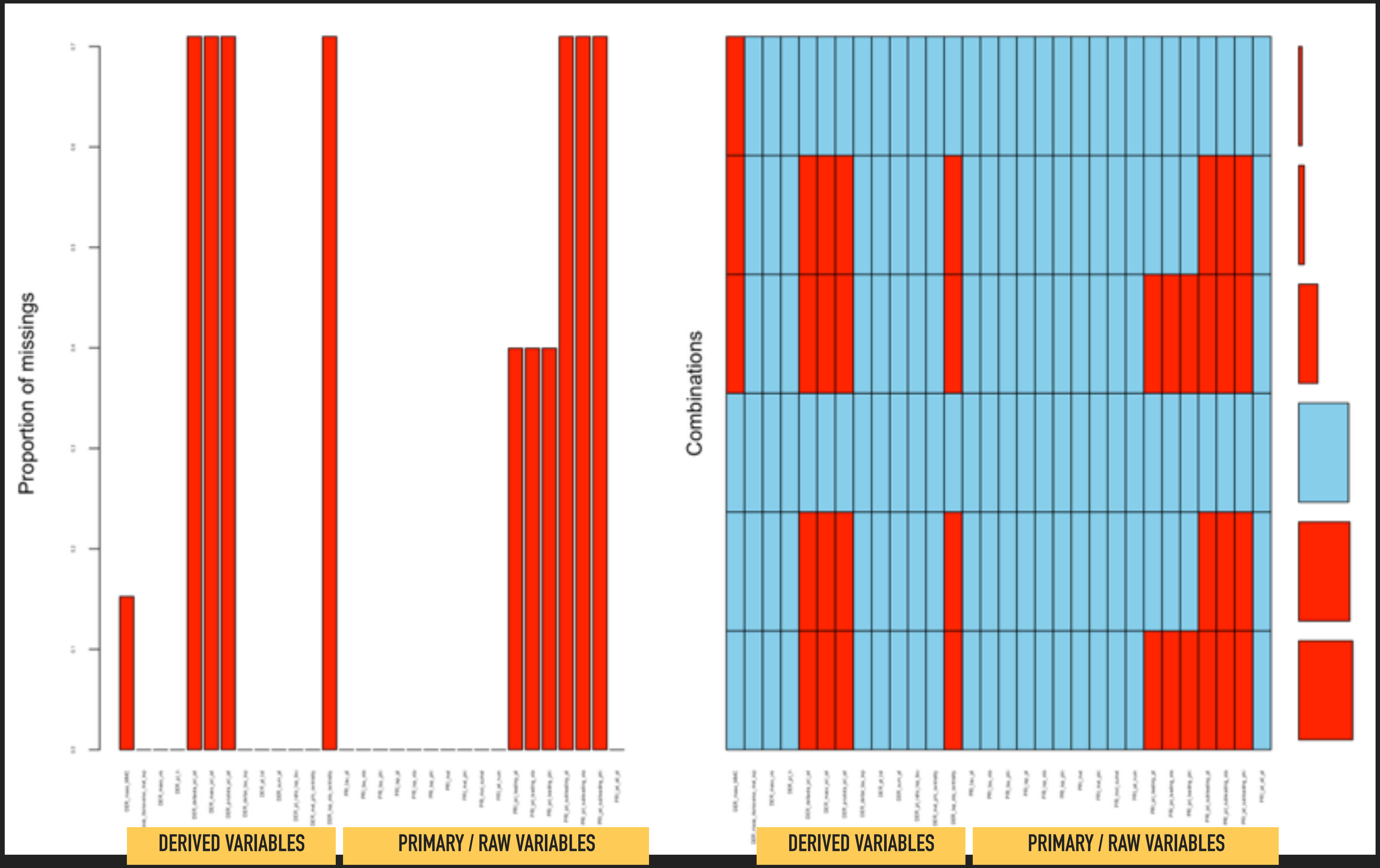
HIGGS-BOSON KAGGLE BACKGROUND & CHALLENGE

- Higgs-Boson particle had been theorized to exist 50 years ago, and was discovered in 2012.
- The particle can decays in various ways, and the discovery of multiple modes of decay increases confidence in the validity of the theory and the characteristics of the particle.
- The Kaggle challenge is to detect a tau tau decay of a Higgs boson "signal" versus other forms of particle decay classified as "background".
- The Kaggle data set included 800,000 simulated records which was split into a 250,000 record training set and a 550,000 record test set.

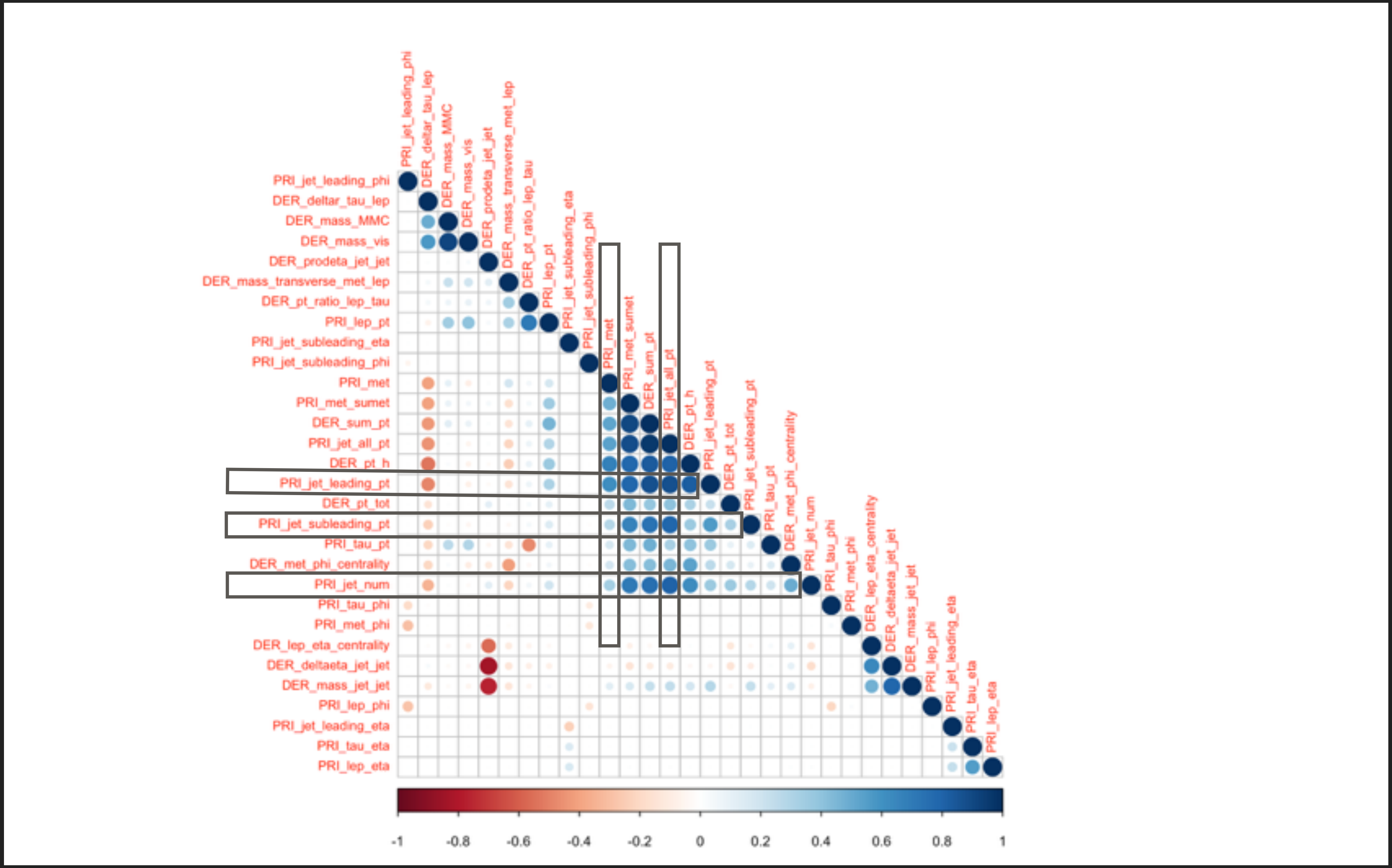
EXPLORATORY DATA ANALYSIS

A FAIR AMOUNT OF MISSING INFORMATION IN THE DATA CAN BE CLASSIFIED AS MISSING AT RANDOM, I.E. DEPENDENT ON OTHER VARIABLES IN THE DATA SET.

ANALYSIS OF MISSING DATA



CORRELATION MATRIX OF HIGGS-BOSON DATA SET



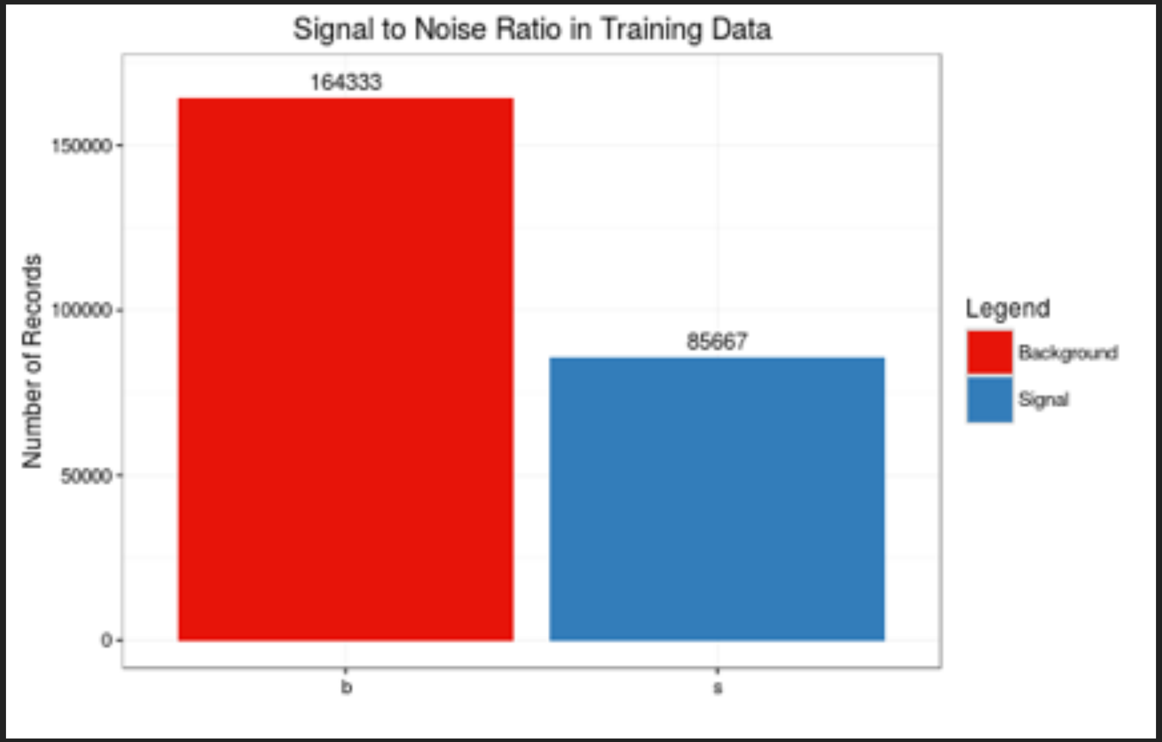
A CLOSER LOOK AT THE DATA SUGGESTS THAT A LOT OF THE MISSING DATA IS RELATED TO THE NUMBER OF JETS VARIABLE.

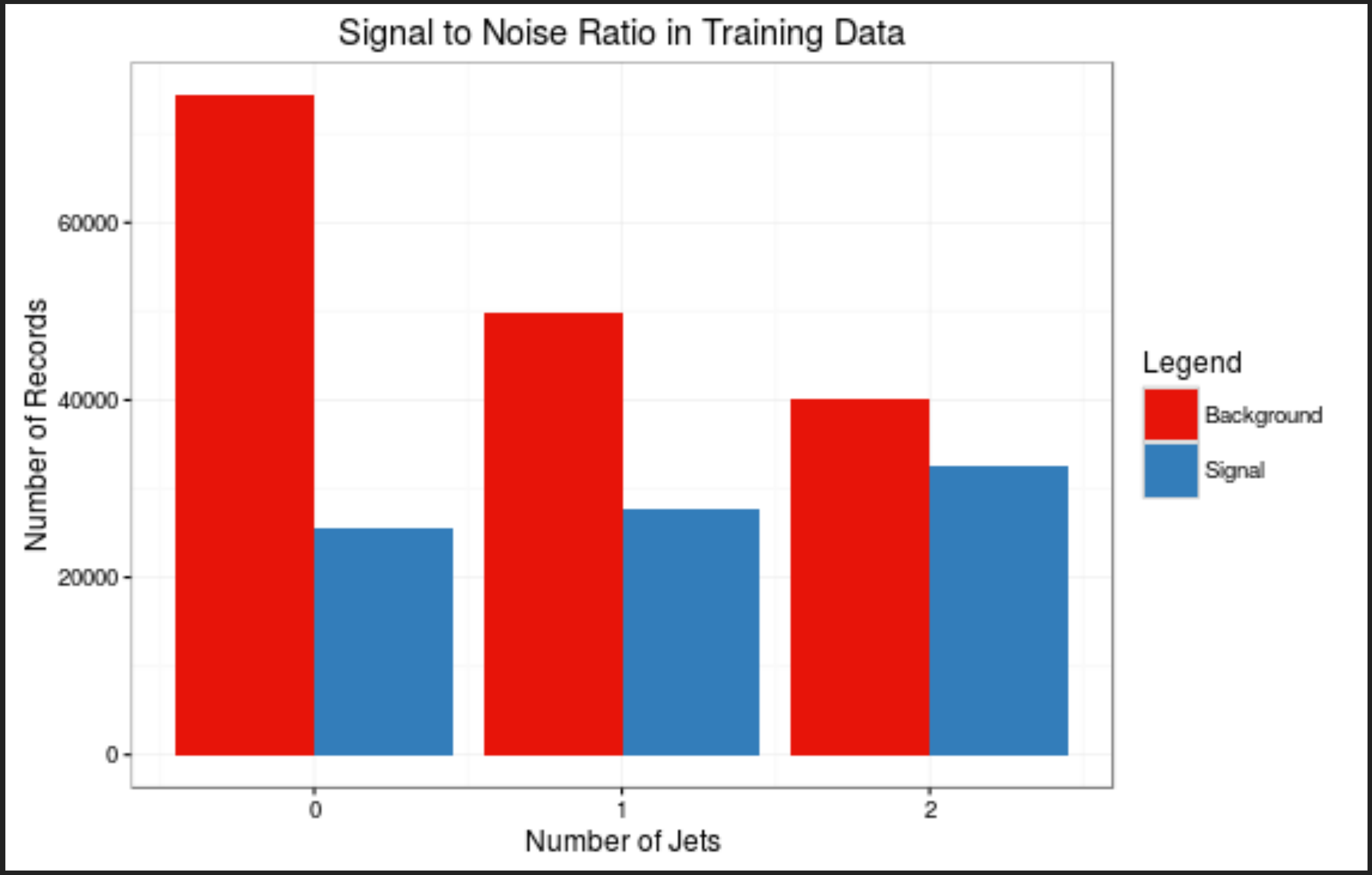
MISSING DATA BY NUMBER OF JETS

Variables	PRI_jet_num_0	PRI_jet_num_1	PRI_jet_num_2_3
DER_mass_MMC	26123	7562	4429
DER_mass_transverse_met_lep	0	0	0
DER_mass_vis	0	0	0
DER_pt_h	0	0	0
DER_deltaeta_jet_jet	99913	77544	0
DER_mass_jet_jet	99913	77544	0
DER_prodeta_jet_jet	99913	77544	0
DER_deltar_tau_lep	0	0	0
DER_pt_tot	0	0	0
DER_sum_pt	0	0	0
DER_pt_ratio_lep_tau	0	0	0
DER_met_phi_centrality	0	0	0
DER_lep_eta_centrality	99913	77544	0
PRI_tau_pt	0	0	0
PRI_tau_eta	0	0	0
PRI_tau_phi	0	0	0
PRI_lep_pt	0	0	0
PRI_lep_eta	0	0	0
PRI_lep_phi	0	0	0
PRI_met	0	0	0
PRI_met_phi	0	0	0
PRI_met_sumet	0	0	0
PRI_jet_num	0	0	0
PRI_jet_leading_pt	99913	0	0
PRI_jet_leading_eta	99913	0	0
PRI_jet_leading_phi	99913	0	0
PRI_jet_subleading_pt	99913	77544	0
PRI_jet_subleading_eta	99913	77544	0
PRI_jet_subleading_phi	99913	77544	0
PRI_jet_all_pt	0	0	0

THERE IS A 2:1 NOISE:SIGNAL RATIO IN THE OVERALL DATA SET, BUT THE RATIO IS SMALLER WHEN THE NUMBER OF JETS IS 0 OR LARGER WHEN THE NUMBER OF JETS IS > 1.

SIGNAL TO NOISE RATIO IN THE DATA





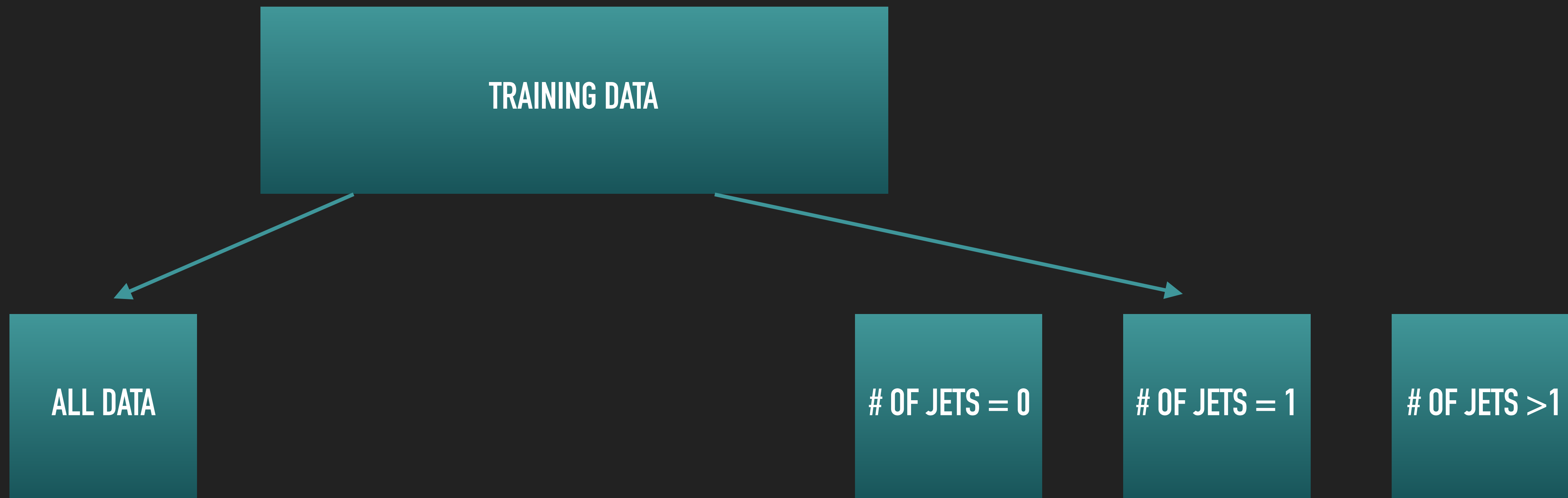
THE WEIGHTS ARE HALVED AS THE NUMBER OF JETS INCREASES. THIS MAY BE COMPENSATING FOR THE INCREASED NUMBER OF SIGNALS BY NUMBER OF JETS.

WEIGHT INFORMATION BY CLASS TYPE

# of Jets	Background Weights	Signal Weights
0 Jets	3.76	0.014
1 Jet	1.93	0.007
2+ Jets	0.88	0.004

ALL MODELS WERE RUN ON THE FULL TRAINING SET AS WELL AS ON A VERSION OF THE DATA THAT WAS SPLIT ON NUMBER OF JETS (AFTER REMOVING COLUMNS THAT ONLY REPRESENTED MISSING INFORMATION).

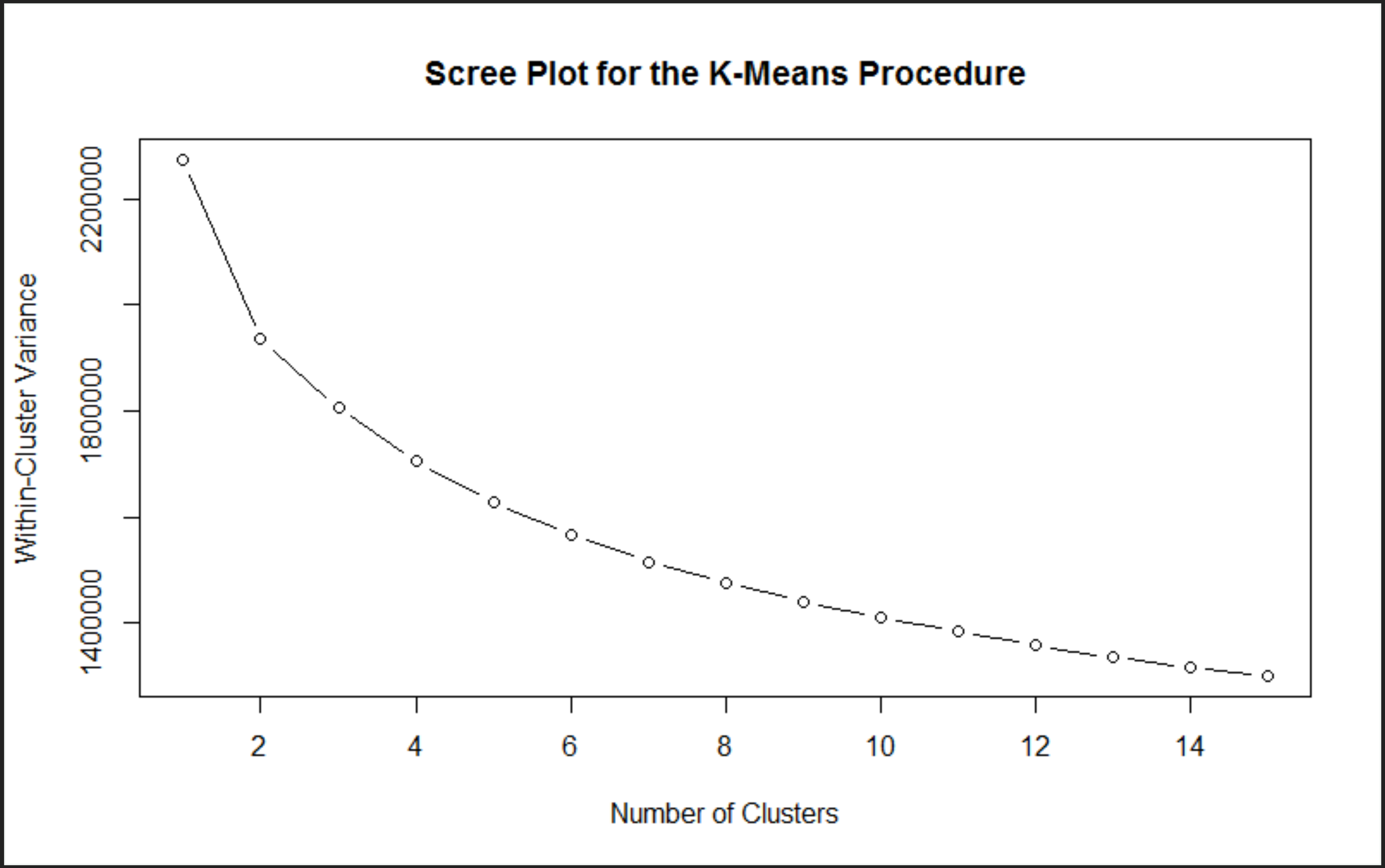
DATA SETS FOR MODELS



Removed Variables that were completely blank

DER_lep_eta_centrality	DER_lep_eta_centrality
DER_deltaeta_jet_jet	DER_deltaeta_jet_jet
DER_mass_jet_jet	DER_mass_jet_jet
DER_prodeta_jet_jet	DER_prodeta_jet_jet
PRI_jet_subleading_pt	PRI_jet_subleading_pt
PRI_jet_subleading_eta	PRI_jet_subleading_eta
PRI_jet_subleading_phi	PRI_jet_subleading_phi
PRI_jet_leading_pt	
PRI_jet_leading_eta	
PRI_jet_leading_phi	

K-MEANS CLUSTERING



MACHINE LEARNING MODELS

AFTER MEDIAN IMPUTING DATA FOR DER_MASS_MMC, THE RANDOM FOREST ON THE DATA SPLIT BY NUMBER OF JETS PRODUCED A LOW AMS SCORE

PARAMETER TUNING FOR RANDOM FORESTS ON SPLIT FILE

Records Tested: 5,000 w/ 1,000 trees	Jets = 0	Jets = 1	Jets = 2+
Predictors "mtry" (odd #s: 3-19)	7	7	7
Max # of nodes (3,5,8,10)	10	10	10
Threshold	0.35	0.5	0.5
OOB Error	14.95	18.33	15.04

TEST DATA AMS: 2.53

RANDOM FORESTS ON COMPLETE TRAINING SET

Records Tested: 5,000 w/ 1,000 trees	ALL TRAINING DATA with BAG-IMPUTED DATA
Predictors "mtry"	7
Max # of nodes	No Pruning
Threshold	0.5
OOB Error	16.01

TEST DATA AMS: 2.9

RELATIVE VARIABLE IMPORTANCE (TOP 20) - ALL TRAINING DATA

	Importance
DER_mass_MMC	100.00
DER_mass_transverse_met_lep	88.15
DER_deltar_tau_lep	58.94
DER_mass_vis	57.93
PRI_tau_pt	53.26
DER_met_phi_centrality	52.72
PRI_met	49.18
PRI_met_sumet	39.03
DER_pt_h	36.19
DER_pt_ratio_lep_tau	30.96
DER_pt_tot	30.61
PRI_lep_eta	30.19
PRI_lep_pt	29.10
PRI_jet_leading_eta	28.53
DER_lep_eta_centrality	25.01
PRI_tau_eta	23.98
DER_sum_pt	22.89
DER_mass_jet_jet	21.92
PRI_jet_all_pt	21.62
DER_deltaeta_jet_jet	21.20

LOGISTIC REGRESSION

- Fitted Logistic regression model to predict log odds of signal vs background noise
- First tried using all data, and chose only significant variables
- Produced max AMS on training data of 2.06
- Predicted 66,232 signals in test data

LOGISTIC REGRESSION: DATA SPLIT BY NUMBER OF JETS

- Fitted Logistic Regression model to predict log odds of three sub-sets of data (zero jets, one jet, two or more jets)
- Significant variables to the model were different for each subset (beyond just “absent” variables).
- Produced max AMS on training data of 1.07
- Predicted 70,470 signals in test data

VARIABLE SELECTION FOR LOGISTIC REGRESSION MODEL

Variable	Full Model	Two+ Jets	One Jet	Zero Jets
DER_mass_MMC	***	***	***	**
DER_mass_transverse_met_lep	***	***	***	***
DER_mass_vis	***	***	***	***
DER_pt_h	***	***		
DER_deltaeta_jet_jet	***	***	#N/A	#N/A
DER_mass_jet_jet	***	***	#N/A	#N/A
DER_prodeta_jet_jet	***	***	#N/A	#N/A
DER_deltar_tau_lep	***	***	***	***
DER_pt_tot		***	***	
DER_sum_pt				
DER_pt_ratio_lep_tau	***	***	***	***
DER_met_phi_centrality	***	***	***	
DER_lep_eta_centrality	***	***	#N/A	#N/A
PRI_tau_pt				
PRI_tau_eta				
PRI_tau_phi				
PRI_lep_pt				
PRI_lep_eta			*	
PRI_lep_phi				
PRI_met	***	***	***	
PRI_met_phi				*
PRI_met_sumet	***	***	***	***
PRI_jet_num	***	#N/A	#N/A	#N/A
PRI_jet_leading_pt		*		#N/A
PRI_jet_leading_eta				#N/A
PRI_jet_leading_phi				#N/A
PRI_jet_subleading_pt	**	***	#N/A	#N/A
PRI_jet_subleading_eta	***		#N/A	#N/A
PRI_jet_subleading_phi	***		#N/A	#N/A
PRI_jet_all_pt				

Signif. codes: '***' .1 '**' .1 '*' .5 '.' .1 ' ' 1

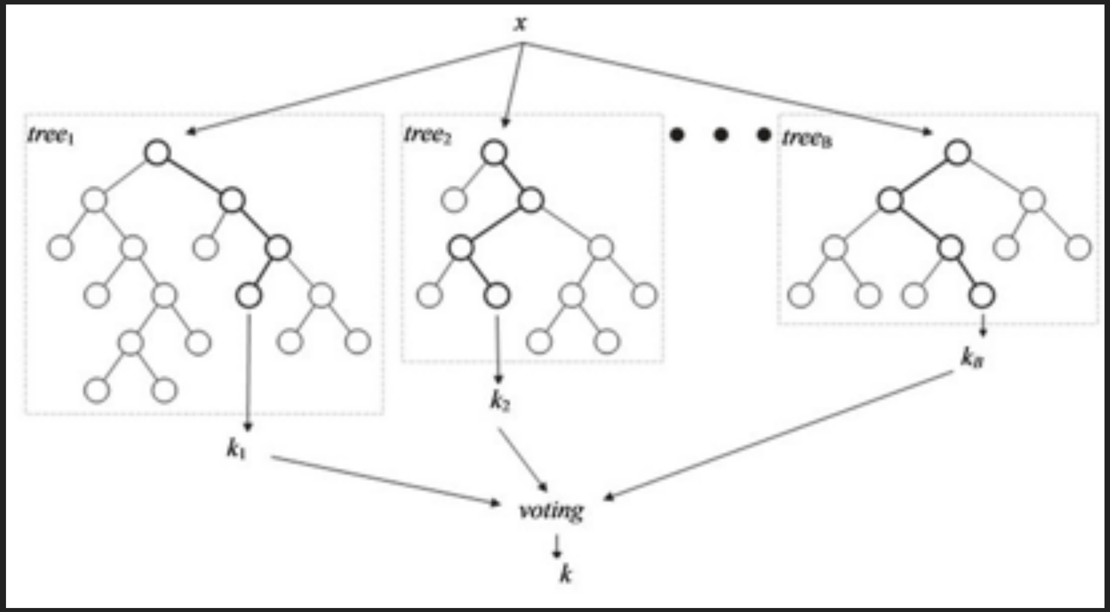
EXTREME GRADIENT BOOSTING (XG BOOST)

- Modified code published https://github.com/dmlc/xgboost/tree/master/demo/kaggle-higgs by Tainqi Chen, the author of the xgboost package
- Tuned parameters maximizing AMS score, focusing on eta (learning rate), max_depth (maximum depth of trees), and nrounds (the number of learning iterations).
- Ultimately chose eta = .05, max_depth=12, nrounds=120.
- Selection of Threshold unimportant, as probabilities will be fed to ensembling model
- Produced max AMS on training data of 8.42 (overfit?)
- Predicted 36,965 signals, but possible error in selecting optimal threshold

ENSEMBLING

MODELS IN THE ENSEMBLE

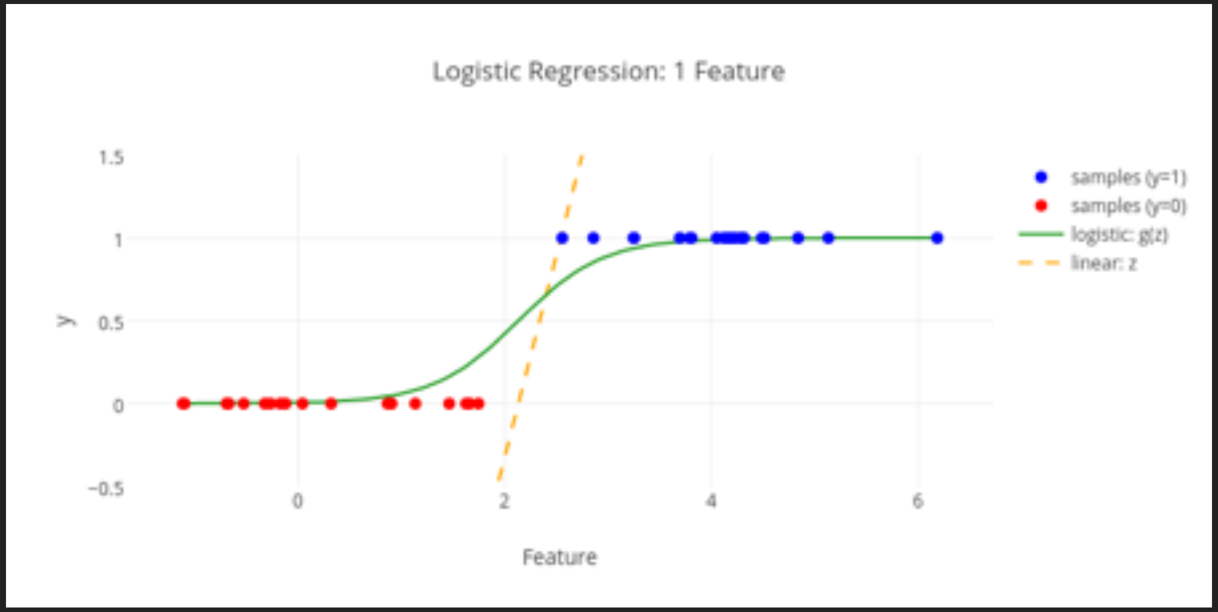
RANDOM FORESTS



XG BOOST

dmlc XGBoost

LOGISTIC REGRESSIONS



ENSEMBLING DETAILS

- Different kinds of ensembling:
  - averaging
  - voting
  - weighting
- Four variables to optimize
- α, β, γ: model weights
- η: signal threshold
- α logit + β rf + γ xgboost = ensemble probability
- ensemble probability > threshold η

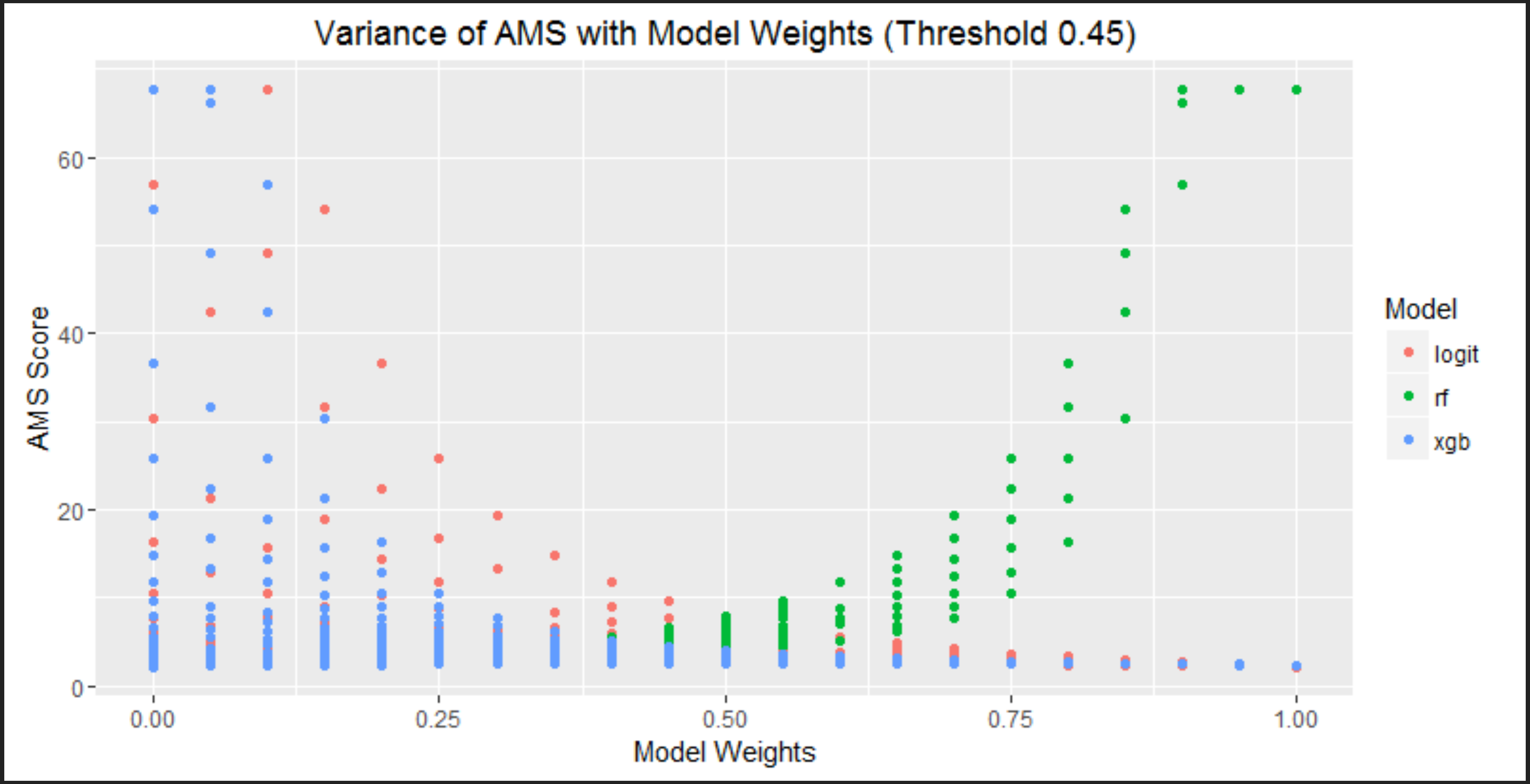
THE BEST RESULTS FROM A TWO-MODEL ENSEMBLE WERE DOMINATED BY THE XGBOOST MODEL

TWO-MODEL ENSEMBLE

Logit	XG Boost	Threshold
0.02	0.98	0.93

THREE-MODEL ENSEMBLE

Logit	RF	XG Boost	Threshold
0.1	0.9	0	0.45



CONCLUSIONS/NEXT STEPS

- The xgboost model performs best in terms of time and accuracy, but lacks interpretability
- The logistic regression model is easier to understand, but cannot match the random forest and boosted models in accuracy
- The random forest provides some interpretability but also cannot match the boosted model in accuracy
- The xgboost dominates a two-model weighted ensemble with the logistic model
- The random forest appears to dominate a three-model weighted ensemble, but more investigation needs to be done
- More models such as a neural network can be added to the ensemble to further reduce variance and improve accuracy