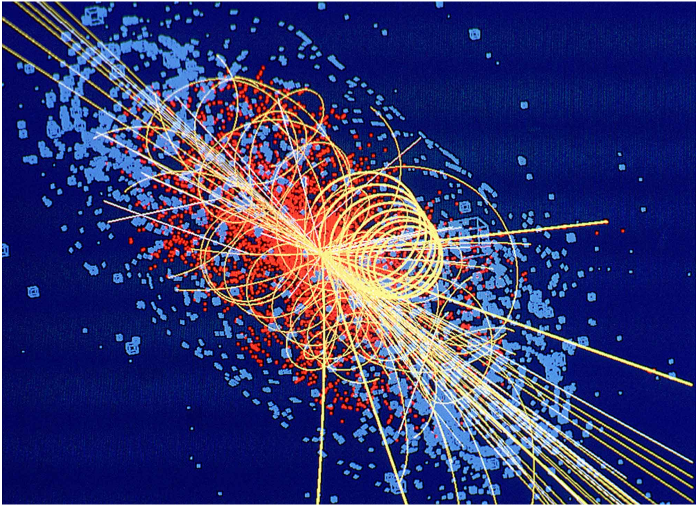
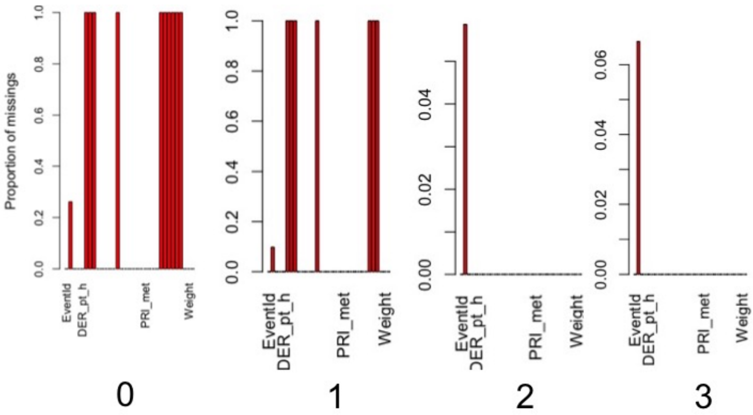
Higgs Boson dataset: From Description to Ensemble

Robert Castellano, Yannick Kimmel, Wanda Wang, Ho Fai Wong



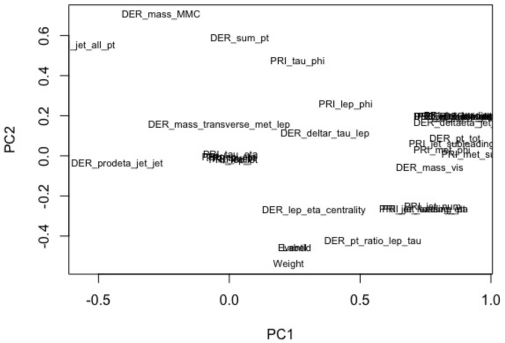
Exploratory data analysis

Sparse dataset



- Jet number can be treated as a factor for missingness.

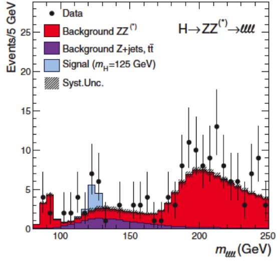
Principal component analysis

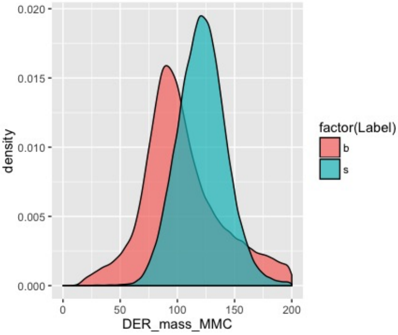


	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	h2	u2	com
DER_mass_MMC	0.25	-0.45	-0.02	0.55	0.04	-0.02	0.01	-0.24	0.00	0.63	0.3693	2.8
Label	0.23	-0.54	-0.07	0.11	0.24	0.01	-0.01	-0.36	-0.04	0.55	0.4476	2.8

- PCA shows that derived mass and label have a very strong relationship.

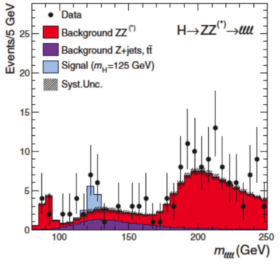
Mass as a predictor of Higgs Boson presence

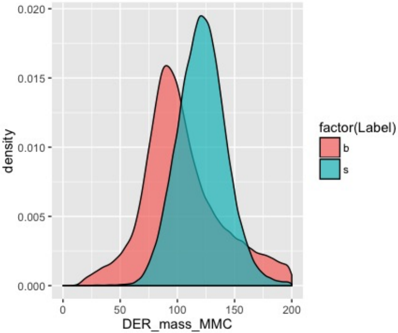




- Derived mass of Higgs Boson is different from other Bosons and subatomic particles.

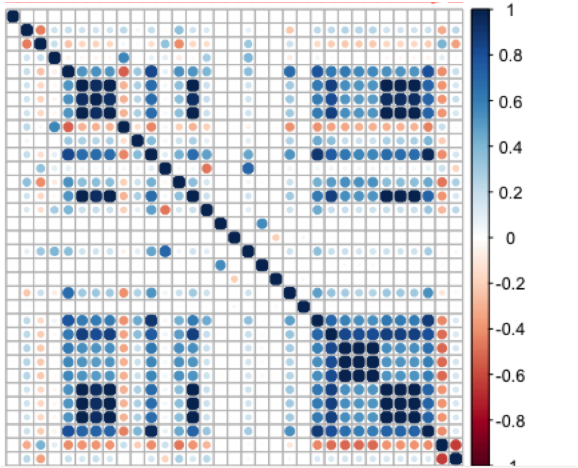
Mass as a predictor of Higgs Boson presence





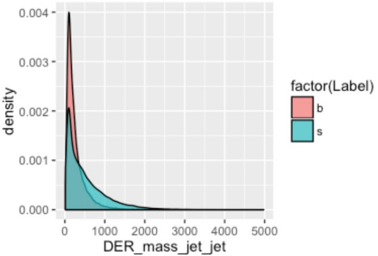
- Derived mass of Higgs Boson is different from other Bosons and subatomic particles.
- Simulated dataset increases signal, and must be offset using weights.

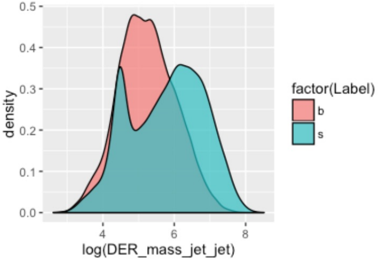
Correlation matrix



- There are several variables with strong covariance among the 33 variables.

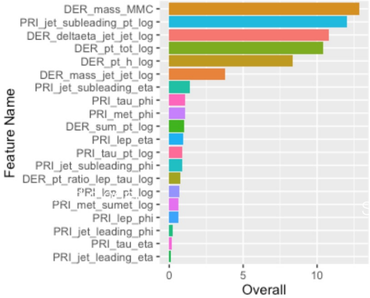
Initial Feature Engineering

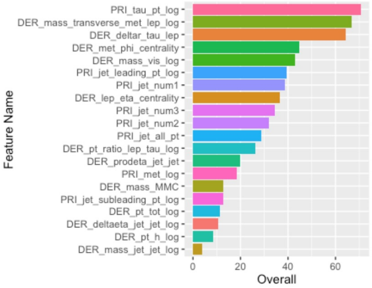




- 14 Features with long-tailed distributions were log transformed to reduce the positive skew towards smaller values, generating a more uniform distribution.. E.g. DER_mass_jet_jet: The invariant mass of the two jets.

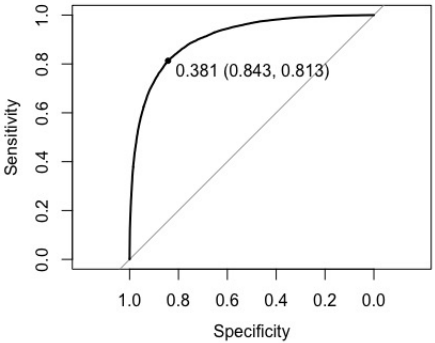
Logistic Regression - Variable Importance





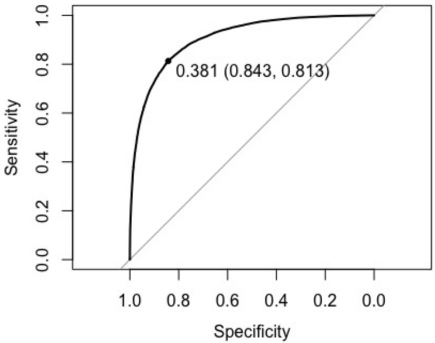
- Saturated Model vs. Stepwise BIC Model

Choice of AUC as model fit metric



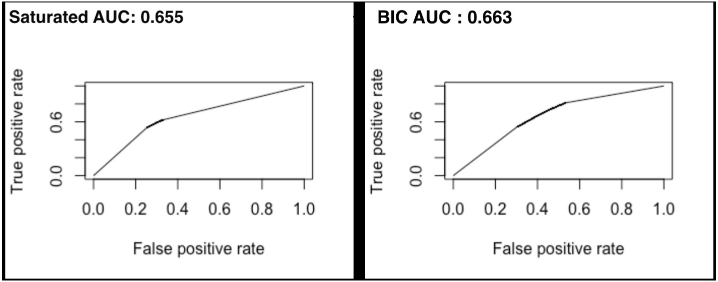
- Maximizes the true positive rate while also minimizes the false positive rate.

Choice of AUC as model fit metric



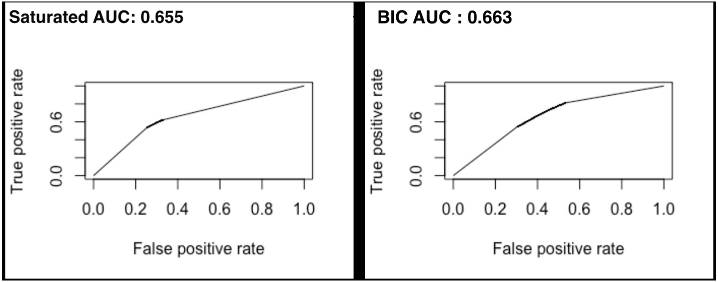
- Maximizes the true positive rate while also minimizes the false positive rate.
- Produces a smooth and continuous function unlike AMS.

Logistic Regression - Analysis



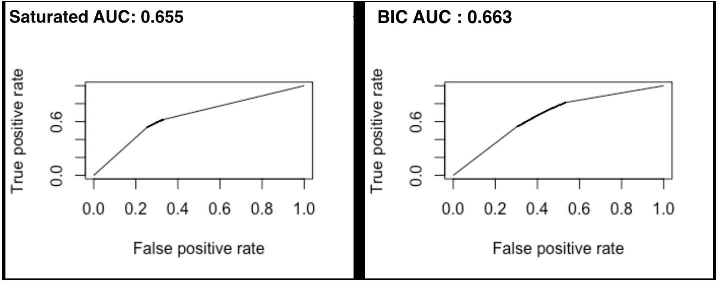
- Saturated Model: R.Squared: 0.20227; Stepwise BIC model: R.Squared: 0.20223.

Logistic Regression - Analysis



- Saturated Model: R.Squared: 0.20227; Stepwise BIC model: R.Squared: 0.20223.
- In comparing the deviance of the stepwise model to the deviance of the saturated model, the p-value for the overall test of deviance is > 0.65 (high)

Logistic Regression - Analysis



- Saturated Model: R.Squared: 0.20227; Stepwise BIC model: R.Squared: 0.20223.
- In comparing the deviance of the stepwise model to the deviance of the saturated model, the p-value for the overall test of deviance is > 0.65 (high)
- AUC plots are also not very different from one another.

Models

Our models

- Random forest

Our models

- Random forest
- Gbm

Our models

- Random forest
- Gbm
- Xgboost

Random forest model

Robert Castellano, Yannick Kimmel, Wanda Wang, Ho Fai Wong

- Tuning parameters
  - mtry: Number of splits per tree

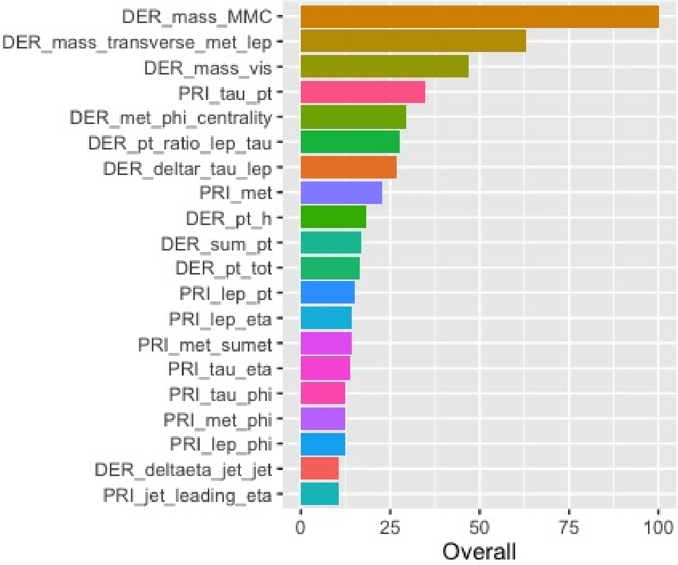
Random forest model

- Tuning parameters
  - mtry: Number of splits per tree
- Performed 5-fold CV to tune parameters.
  - 20% of training data for mtry gride of 1, 2, 3, 6, 9
  - 80% of training data for mtry gride of 4, 5, 6, 7, 8
  - mtry = 5

Random forest model

- Tuning parameters
  - mtry: Number of splits per tree
- Performed 5-fold CV to tune parameters.
  - 20% of training data for mtry gride of 1, 2, 3, 6, 9
  - 80% of training data for mtry gride of 4, 5, 6, 7, 8
  - mtry = 5
- AUC on training data = .9071
- Kaggle rank = 1311
- AMS = 2.57949

Random forest variable importance



Gbm model

- Gradient boosting model

Gbm model

- Gradient boosting model
- Tuning parameters
  - shrinkage: Learning rate
  - interaction_depth: Depth of variable interactions
  - n.trees: Number of trees
  - n.minobsinnode: Minimum number of observations in a terminal node

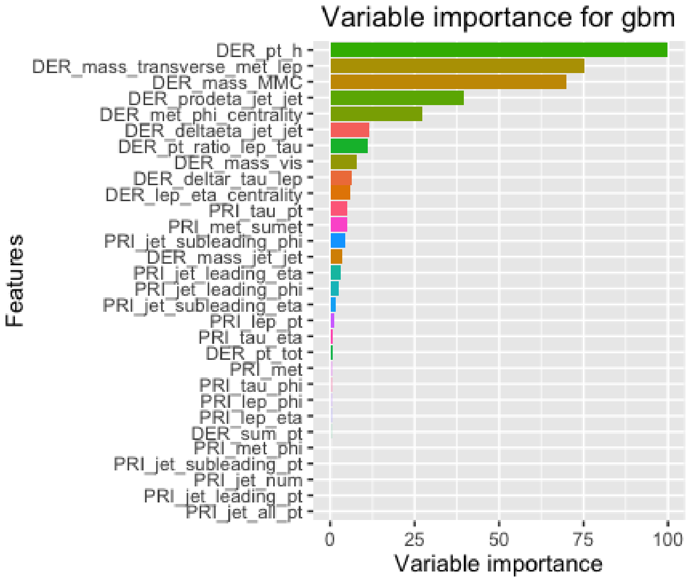
Gbm model

- Gradient boosting model
- Tuning parameters
  - shrinkage: Learning rate
  - interaction_depth: Depth of variable interactions
  - n.trees: Number of trees
  - n.minobsinnode: Minimum number of observations in a terminal node
- Performed 5-fold CV to tune parameters.
  - shrinkage = .1
  - interaction_depth = 3
  - n.trees = 150
  - n.minobsinnode = 10

Gbm model

- Gradient boosting model
- Tuning parameters
  - shrinkage: Learning rate
  - interaction_depth: Depth of variable interactions
  - n.trees: Number of trees
  - n.minobsinnode: Minimum number of observations in a terminal node
- Performed 5-fold CV to tune parameters.
  - shrinkage = .1
  - interaction_depth = 3
  - n.trees = 150
  - n.minobsinnode = 10
- AUC on training data = .855
- Kaggle rank = 1394
- AMS = 2.30069

Gbm variable importance



About xgboost

- Fast gradient boosting algorithm implementing in C++ by Tianqi Chen

About xgboost

- Fast gradient boosting algorithm implementing in C++ by Tianqi Chen
- Parallel computing

About xgboost

- Fast gradient boosting algorithm implementing in C++ by Tianqi Chen
- Parallel computing
- More tuning parameters

About xgboost

- Fast gradient boosting algorithm implementing in C++ by Tianqi Chen
- Parallel computing
- More tuning parameters
- Not completely greedy in tree creation

About xgboost

- Fast gradient boosting algorithm implementing in C++ by Tianqi Chen
- Parallel computing
- More tuning parameters
- Not completely greedy in tree creation
- Generally faster and performs better than gbm.

Xgboost model

- Parameters we tuned:
  - nrounds: Number of trees
  - max_depth
  - colsample_bytree: Percent of parameters used at each split. tree
  - eta: Learning rate

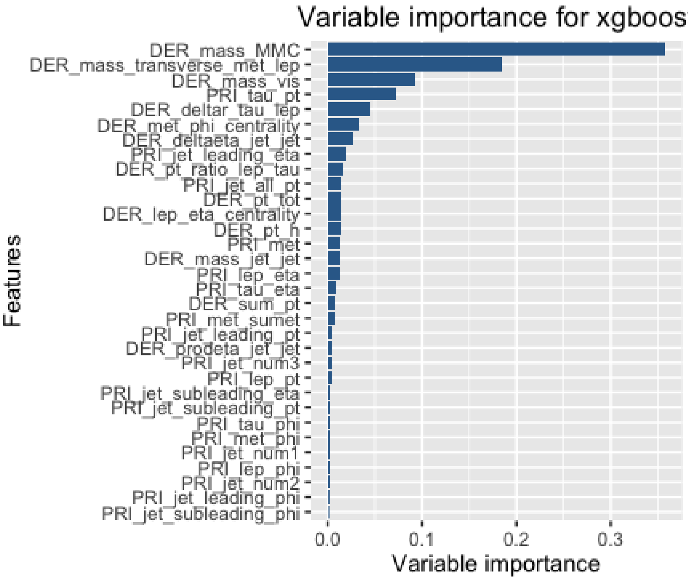
Xgboost model

- Parameters we tuned:
  - nrounds: Number of trees
  - max_depth
  - colsample_bytree: Percent of parameters used at each split. tree
  - eta: Learning rate
- Performed 5-fold CV to tune parameters.
  - nrounds = 200
  - max_depth = 5
  - colsample_bytree = .85
  - eta = .2

Xgboost model

- Parameters we tuned:
  - nrounds: Number of trees
  - max_depth
  - colsample_bytree: Percent of parameters used at each split. tree
  - eta: Learning rate
- Performed 5-fold CV to tune parameters.
  - nrounds = 200
  - max_depth = 5
  - colsample_bytree = .85
  - eta = .2
- AUC on training data = .9254
- Kaggle rank = 1340
- AMS = 2.49958

Xgboost variable importance



Ensemble

- Combined three models by majority vote

Ensemble

- Combined three models by majority vote
- Kaggle rank $= 1309$

Ensemble

- Combined three models by majority vote
- Kaggle rank = 1309
- AMS = 2.58510

Room for improvement

Feature engineering

- We did not include any additional variables
    - Basic physics. e.g. Cartesian coordinates of momentum

Feature engineering

- We did not include any additional variables
  - Basic physics. e.g. Cartesian coordinates of momentum
  - Advanced physics: e.g. CAKE variable

Feature engineering

- We did not include any additional variables
  - Basic physics. e.g. Cartesian coordinates of momentum
  - Advanced physics: e.g. CAKE variable
  - Better understand the physics of additional variables

Feature engineering

- We did not include any additional variables
  - Basic physics. e.g. Cartesian coordinates of momentum
  - Advanced physics: e.g. CAKE variable
  - Better understand the physics of additional variables
- Log transforms

Models

- More models

Models

- More models
- More sophisticated emsemble

Models

- More models
- More sophisticated emsemble
- Run different random seeds for the same model