

Higgs Boson
dataset: From
Description to
Ensemble

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

Exploratory
data analysis

Models

Room for
improvement

Higgs Boson dataset: From Description to Ensemble

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

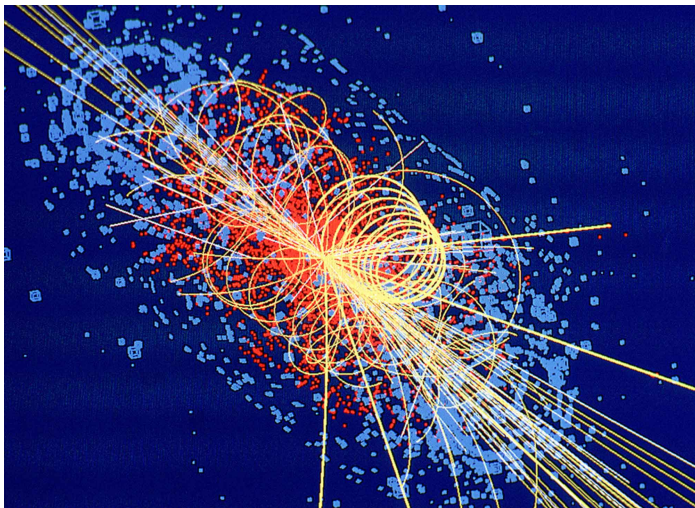
Higgs Boson
dataset: From
Description to
Ensemble

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

Exploratory
data analysis

Models

Room for
improvement



Higgs Boson
dataset: From
Description to
Ensemble

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

Exploratory
data analysis

Models

Room for
improvement

Exploratory data analysis

Sparse dataset

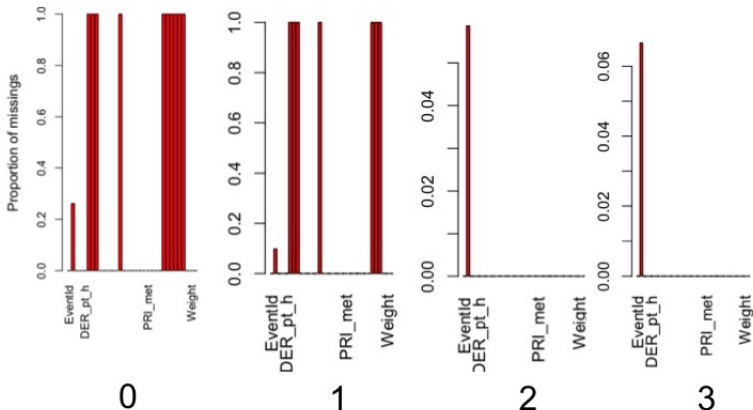
Higgs Boson
dataset: From
Description to
Ensemble

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

Exploratory
data analysis

Models

Room for
improvement



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

Mass as a predictor of Higgs Boson presence

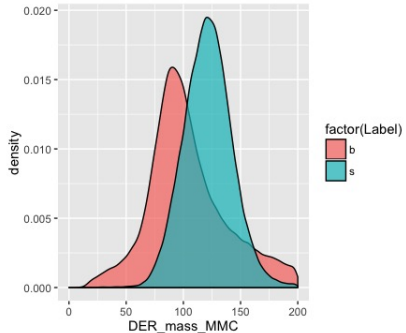
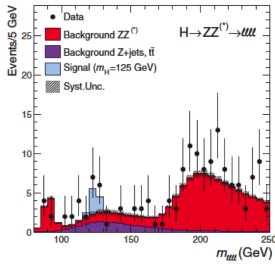
Higgs Boson
dataset: From
Description to
Ensemble

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

Exploratory
data analysis

Models

Room for
improvement



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

Mass as a predictor of Higgs Boson presence

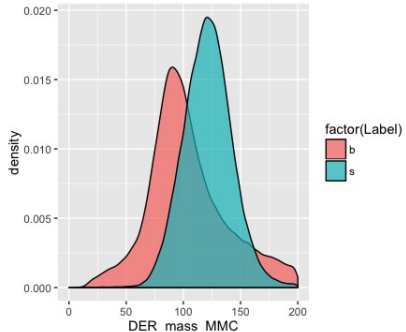
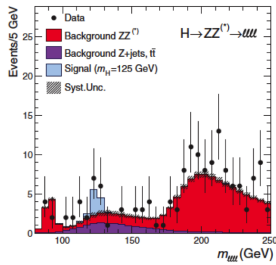
Higgs Boson
dataset: From
Description to
Ensemble

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

Exploratory
data analysis

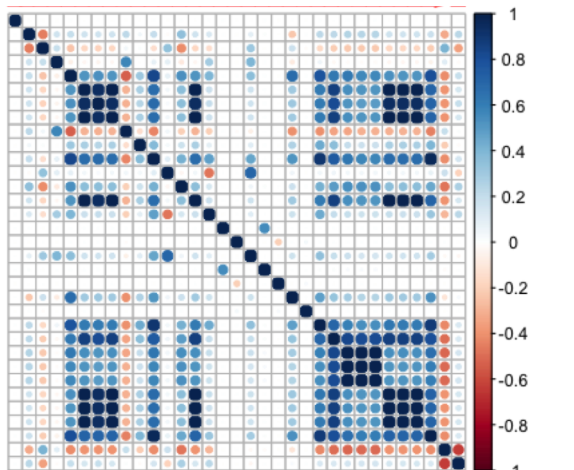
Models

Room for
improvement



- 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.

Higgs Boson
dataset: From
Description to
Ensemble

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

Exploratory
data analysis

Models

Room for
improvement

Choice of AUC as model fit metric

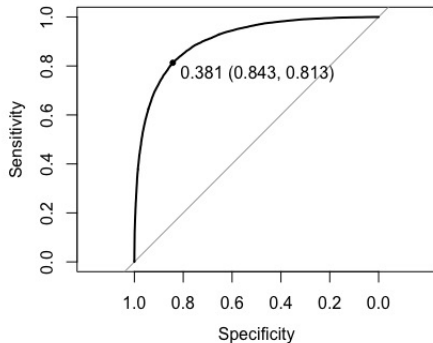
Higgs Boson
dataset: From
Description to
Ensemble

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

Exploratory
data analysis

Models

Room for
improvement



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

Choice of AUC as model fit metric

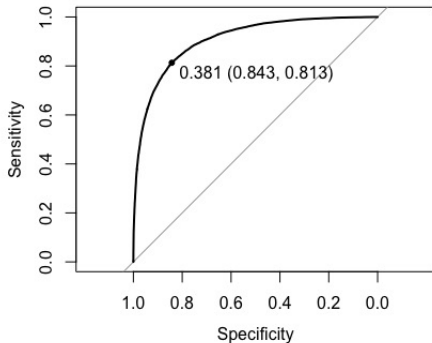
Higgs Boson
dataset: From
Description to
Ensemble

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

Exploratory
data analysis

Models

Room for
improvement



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

Higgs Boson
dataset: From
Description to
Ensemble

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

Exploratory
data analysis

Models

Room for
improvement

Models

Our models

Higgs Boson
dataset: From
Description to
Ensemble

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

Exploratory
data analysis

Models

Room for
improvement

■ Random forest

Our models

Higgs Boson
dataset: From
Description to
Ensemble

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

Exploratory
data analysis

Models

Room for
improvement

- Random forest
- Gbm

Our models

Higgs Boson
dataset: From
Description to
Ensemble

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

Exploratory
data analysis

Models

Room for
improvement

- Random forest
- Gbm
- Xgboost

Random forest model

Higgs Boson
dataset: From
Description to
Ensemble

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

Exploratory
data analysis

Models

Room for
improvement

- Tuning parameters
 - mtry: Number of splits per tree

Random forest model

Higgs Boson
dataset: From
Description to
Ensemble

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

Exploratory
data analysis

Models

Room for
improvement

- Tuning parameters
 - mtry: Number of splits per tree
- Performed 5-fold CV to tune parameters.
 - 20
 - 80
 - $mtry = 5$

Random forest model

Higgs Boson
dataset: From
Description to
Ensemble

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

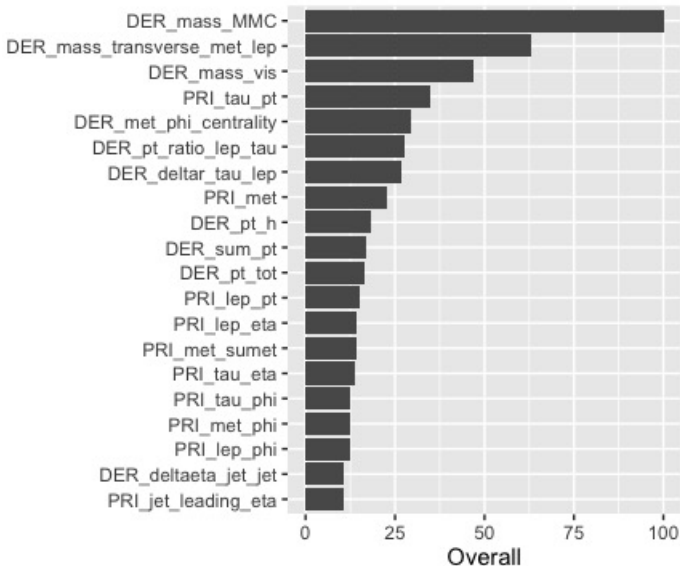
Exploratory
data analysis

Models

Room for
improvement

- Tuning parameters
 - mtry: Number of splits per tree
- Performed 5-fold CV to tune parameters.
 - 20
 - 80
 - $mtry = 5$
- AUC on training data = .9071
- Kaggle rank = 1311
- AMS = 2.57949

Random forest variable importance



Higgs Boson
dataset: From
Description to
Ensemble

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

Exploratory
data analysis

Models

Room for
improvement

Gbm model

■ Gradient boosting model

Higgs Boson
dataset: From
Description to
Ensemble

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

Exploratory
data analysis

Models

Room for
improvement

Gbm model

Higgs Boson
dataset: From
Description to
Ensemble

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

Exploratory
data analysis

Models

Room for
improvement

- 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

Higgs Boson
dataset: From
Description to
Ensemble

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

Exploratory
data analysis

Models

Room for
improvement

- 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

Higgs Boson
dataset: From
Description to
Ensemble

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

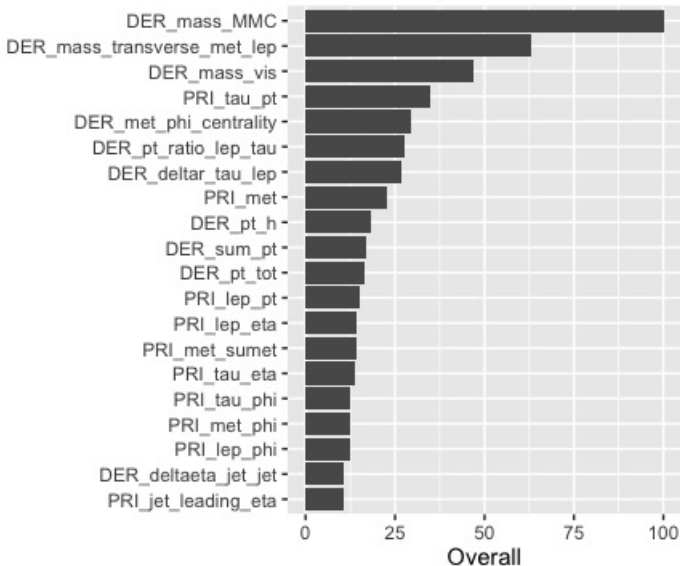
Exploratory
data analysis

Models

Room for
improvement

- 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 = .9071
- Kaggle rank = 1311
- AMS = 2.57949

Random forest variable importance



Higgs Boson
dataset: From
Description to
Ensemble

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

Exploratory
data analysis

Models

Room for
improvement

About xgboost

Higgs Boson
dataset: From
Description to
Ensemble

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

Exploratory
data analysis

Models

Room for
improvement

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

About xgboost

Higgs Boson
dataset: From
Description to
Ensemble

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

Exploratory
data analysis

Models

Room for
improvement

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

About xgboost

Higgs Boson
dataset: From
Description to
Ensemble

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

Exploratory
data analysis

Models

Room for
improvement

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

About xgboost

Higgs Boson
dataset: From
Description to
Ensemble

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

Exploratory
data analysis

Models

Room for
improvement

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

About xgboost

Higgs Boson
dataset: From
Description to
Ensemble

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

Exploratory
data analysis

Models

Room for
improvement

- 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

Higgs Boson
dataset: From
Description to
Ensemble

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

Exploratory
data analysis

Models

Room for
improvement

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

Xgboost model

Higgs Boson
dataset: From
Description to
Ensemble

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

Exploratory
data analysis

Models

Room for
improvement

- 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

Higgs Boson
dataset: From
Description to
Ensemble

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

Exploratory
data analysis

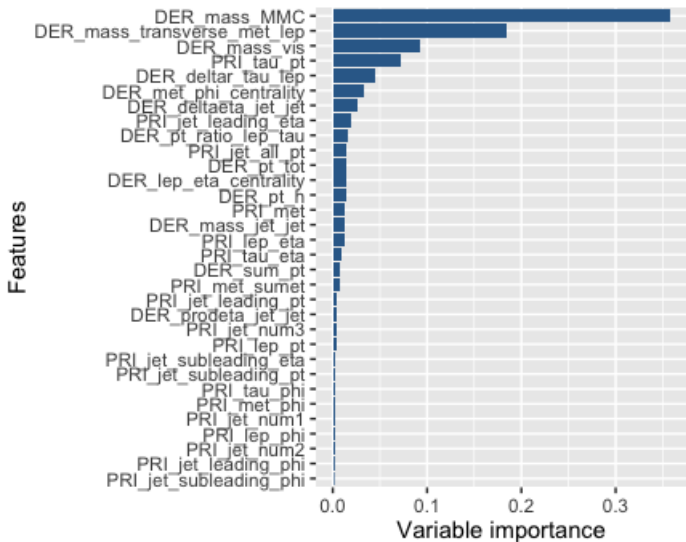
Models

Room for
improvement

- 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

Variable importance for xgboost



Higgs Boson
dataset: From
Description to
Ensemble

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

Exploratory
data analysis

Models

Room for
improvement

Ensemble

Higgs Boson
dataset: From
Description to
Ensemble

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

Exploratory
data analysis

Models

Room for
improvement

- Combined three models by majority vote

Ensemble

Higgs Boson
dataset: From
Description to
Ensemble

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

Exploratory
data analysis

Models

Room for
improvement

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

Ensemble

Higgs Boson
dataset: From
Description to
Ensemble

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

Exploratory
data analysis

Models

Room for
improvement

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

Higgs Boson
dataset: From
Description to
Ensemble

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

Exploratory
data analysis

Models

Room for
improvement

Room for improvement

Feature engineering

Higgs Boson
dataset: From
Description to
Ensemble

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

Exploratory
data analysis

Models

Room for
improvement

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

Feature engineering

Higgs Boson
dataset: From
Description to
Ensemble

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

Exploratory
data analysis

Models

Room for
improvement

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

Feature engineering

Higgs Boson
dataset: From
Description to
Ensemble

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

Exploratory
data analysis

Models

Room for
improvement

- 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 models

Feature engineering

Higgs Boson
dataset: From
Description to
Ensemble

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

Exploratory
data analysis

Models

Room for
improvement

- 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 models
- Log transforms

Models

Higgs Boson
dataset: From
Description to
Ensemble

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

Exploratory
data analysis

Models

Room for
improvement

■ More models

Models

Higgs Boson
dataset: From
Description to
Ensemble

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

Exploratory
data analysis

Models

Room for
improvement

- More models
- More sophisticated ensemble

Models

Higgs Boson
dataset: From
Description to
Ensemble

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

Exploratory
data analysis

Models

Room for
improvement

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