Road to:
"Classication of data from the ATLAS experiments"

Michael Janschek

January 14, 2016

Overview

kaggle

"Kaggle is the world's largest community of data scientists. They compete with each other to solve complex data science problems, and the top competitors are invited to work on the most interesting and sensitive business problems from some of the worlds biggest companies through Masters competitions." [2]

The Challenge

- 12th May 2014 - 15th September 2014
- 13,000$ prize money
- 1,943 participants in 1,785 teams

The Challenge - Goals

Promote data science in physics

"The Higgs boson machine learning challenge [...] has been set up to promote collaboration between high energy physicists and data scientists." [4]

The Challenge - Goals

Improve classification

"We expect that signicant improvements are possible by re-visiting some of the ad hoc choices in the standard procedure [...]." [4]

The Challenge - Goals

Strengthen the discovery

"The goal of the Higgs Boson Machine Learning Challenge is [...] to improve the discovery significance of the experiment." [1]

The Challenge - Data

Provided Data is simulated in a two-step procedure.
Technical properties of ATLAS are actually visible in some features.

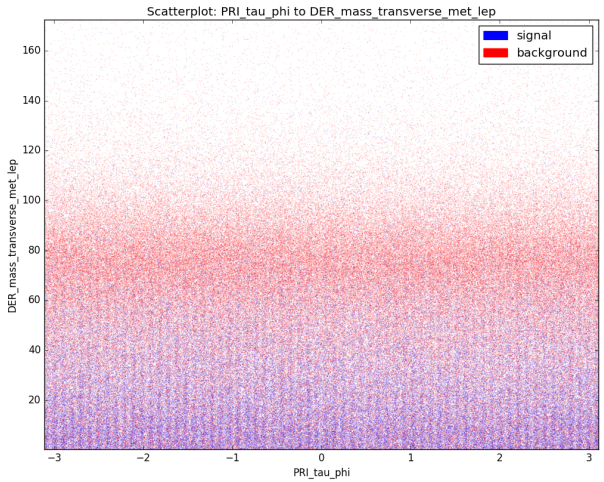


Figure: Scatterplot of PRI_tau_phi to a feature beneficial for demonstration

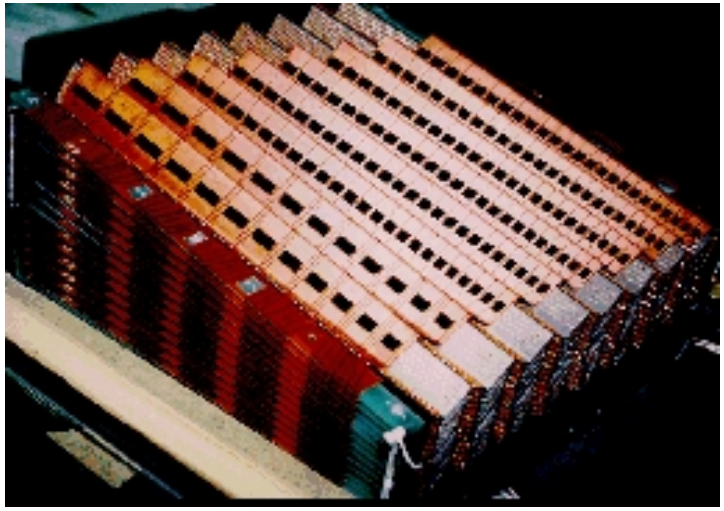


Figure: Inner sections of the calorimeter in ATLAS [3]

The Challenge - Data

- Five features with challenge-relevant information
  (only needed for submission and its evaluation)
- 30 features with simulated data
  (classification-relevant)
- simulated data has dimension $dim(data) = 800000 * 30$

The Challenge - Evaluation

Approximate Median Significance (AMS)

$$AMS = \sqrt{2(s + b + b_r)log[1 + (s/(b + b_r))] - s}$$

where:

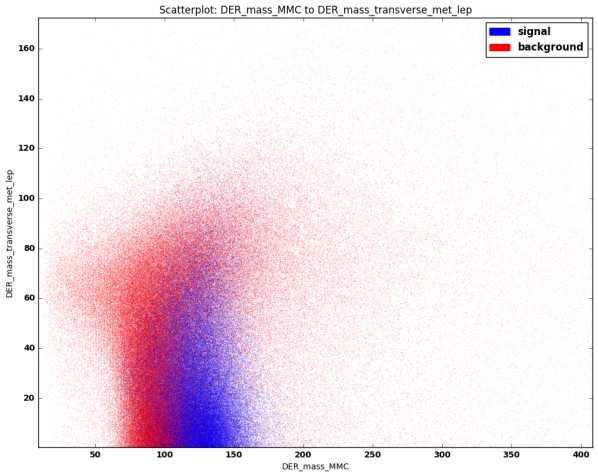
- $b_r = 10$ is a regulization term (set by the contest),
- $b = \sum_{i=1}^{n} w_i, y_i = 0$ is sum of weighted background (incorrectly classified as signal),
- $s = \sum_{i=1}^{n} w_i, y_i = 1$ is sum of weighted signals (correctly classified as signal),
- log is natural logarithm

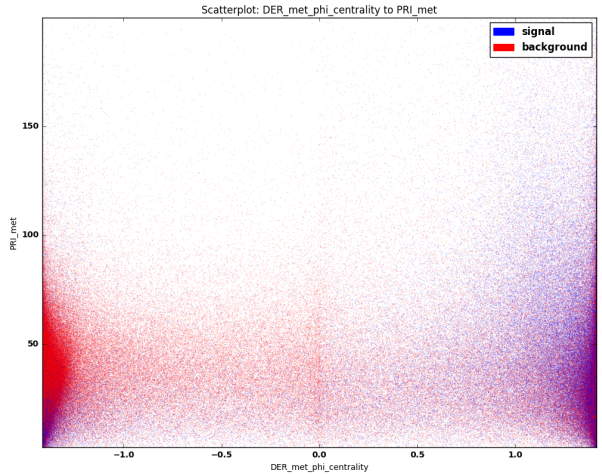
The Challenge - Task

1. learn connection between data and signal-/background-likelihood
2. classify the test-data (550,000 events)
3. submit in format
   $EventId, RankOrder, Class$
   $1, 2, b$
   $2, 541234, s$
   ...

Scatterplots

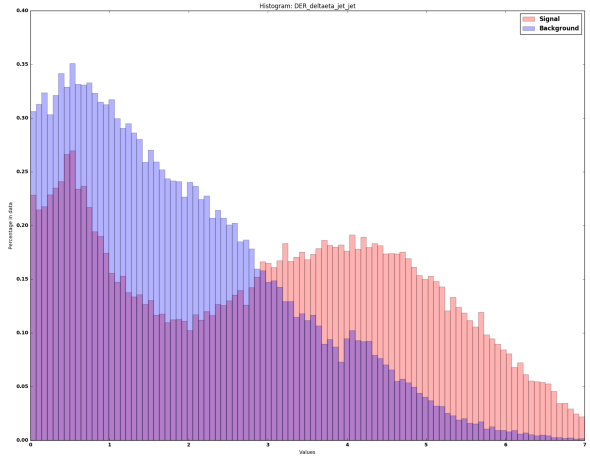
We are looking for features with clustering

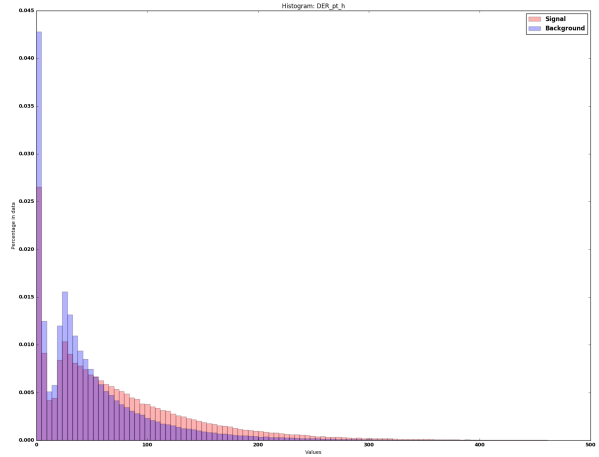




Histograms

We are looking for good-separable signal-distribution





PRI_jet_num and error-values

- 62% of events contain features with value -999.0
- No flaw of simulation, but values are *structurally absent*

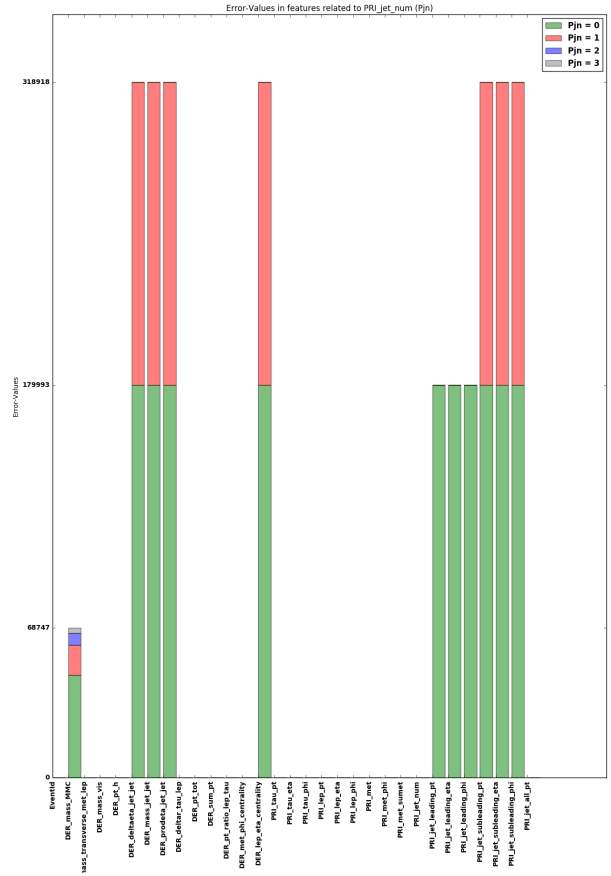


Figure: Histogram of errors related to PRI_jet_num

PRI_jet_num and error-values

"For instance, in events where there is no jet [...], there is no such a thing as a leading jet, thus the associated quantities [...] are structurally undefined, and the features derived from these as well." [5]

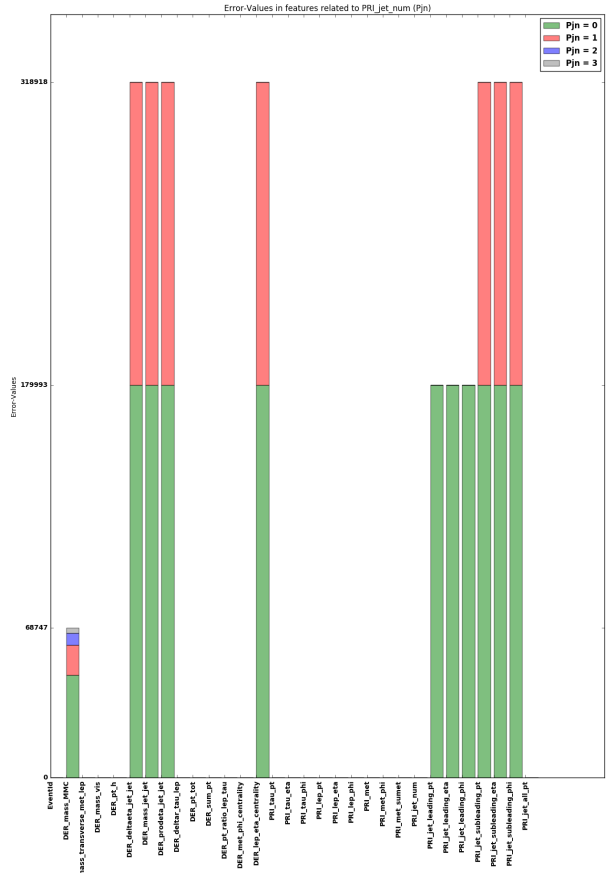


Figure: Histogram of errors related to PRI_jet_num

PRI_jet_num and error-values

"In fact, all missing features are related to PRI_jet_num, except DER_mass_MMC." [5]

We conclude, that seven features contain no information for 25% of events. It will help us to optimize our classifiers.

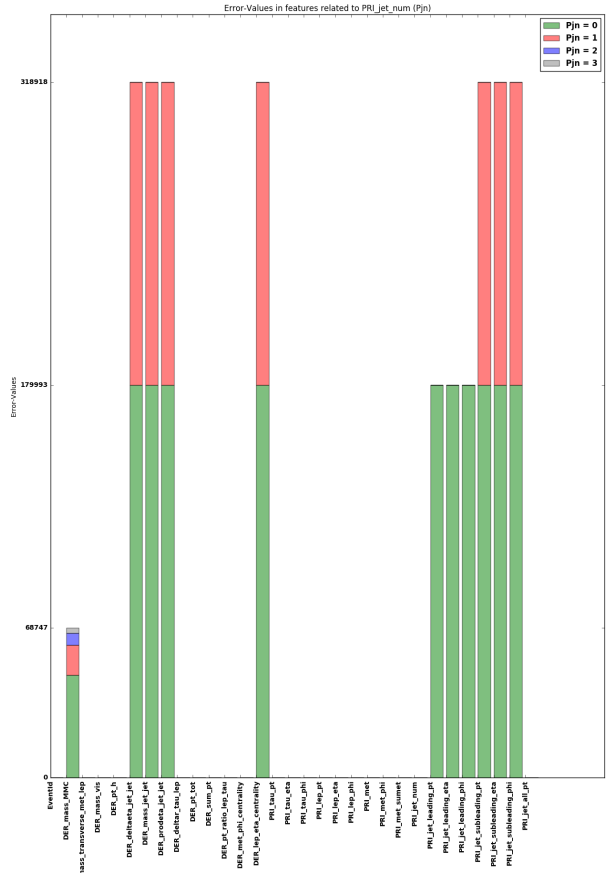
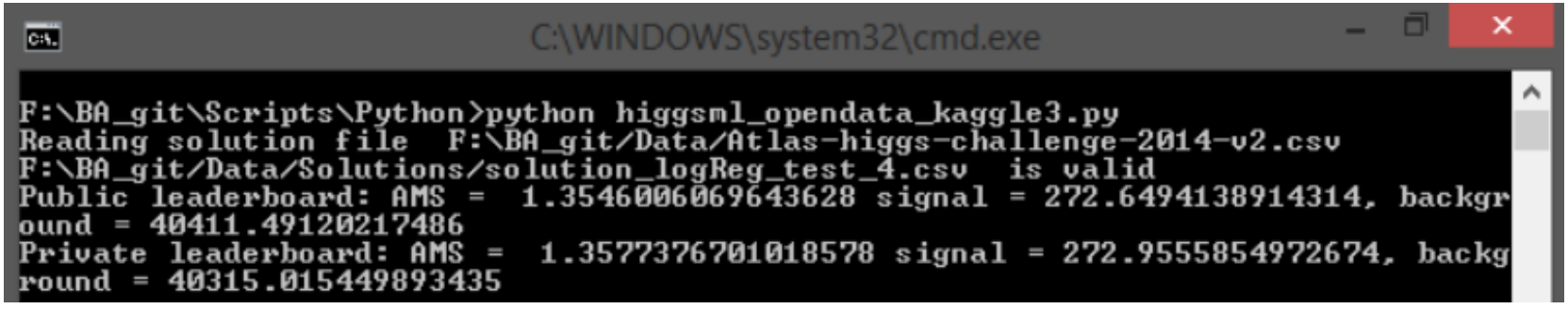


Figure: Histogram of errors related to PRI_jet_num

Logistic Regression

Logistic regression - Score

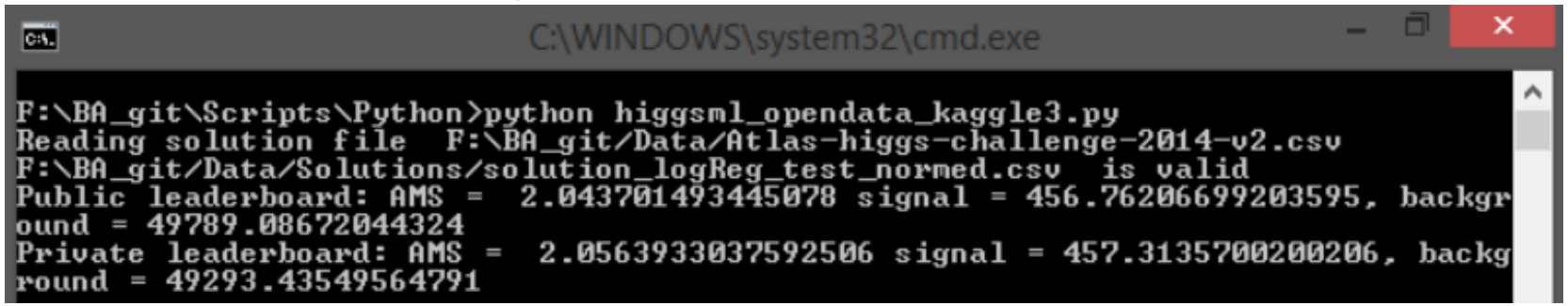
First, we try only few features.



Really low AMS, we would make rank 1625

Logistic regression - Score

We simply use all our data and normalize it

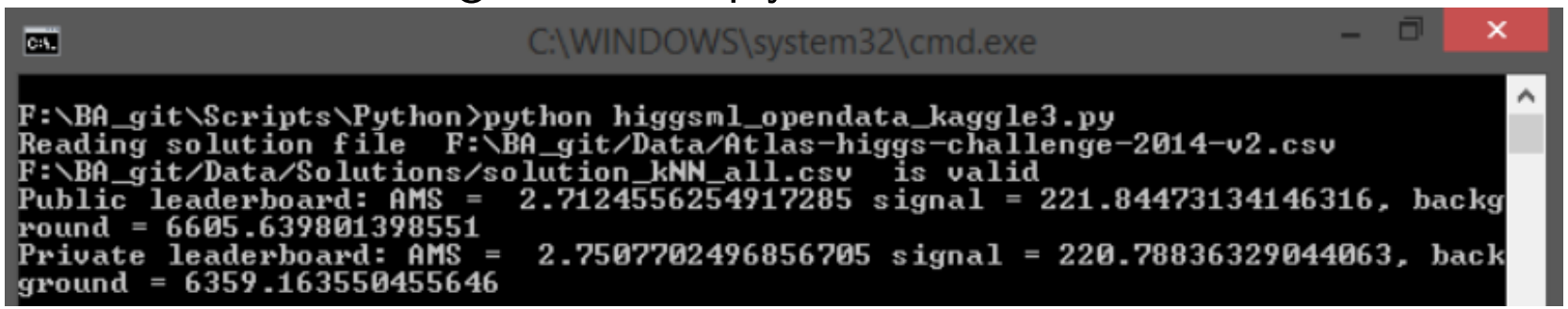


(matches with rank 1473)

k-Nearest-Neighbours

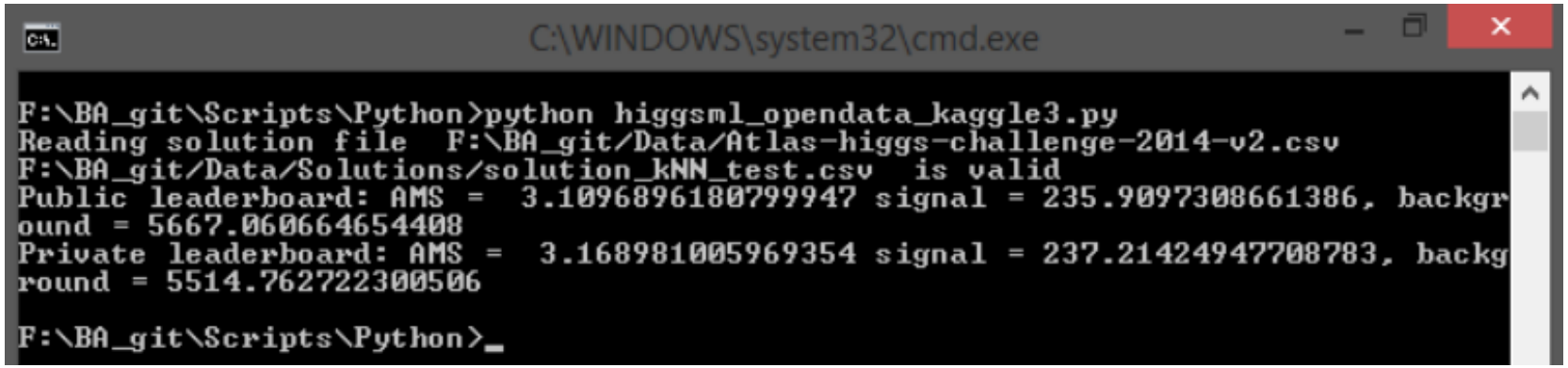
k-Nearest-Neighbours - Score

Again, we simply use all our data



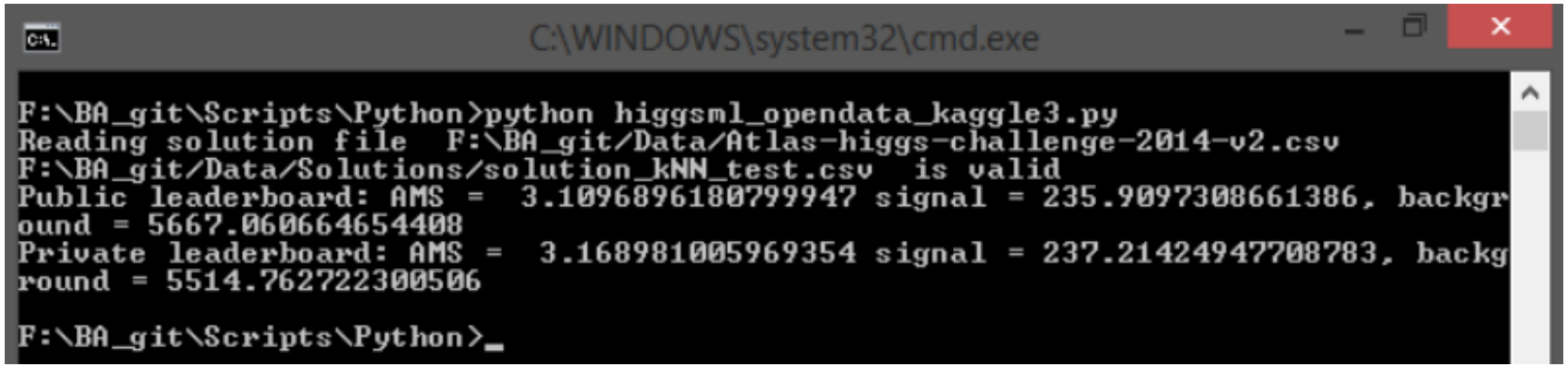
Not very exciting . . . Also kNN rapidly slows down with increasing feature number, we should remove "bad" features.

k-Nearest-Neighbours - Score



HOLY AMS, BATMAN!

k-Nearest-Neighbours - Score

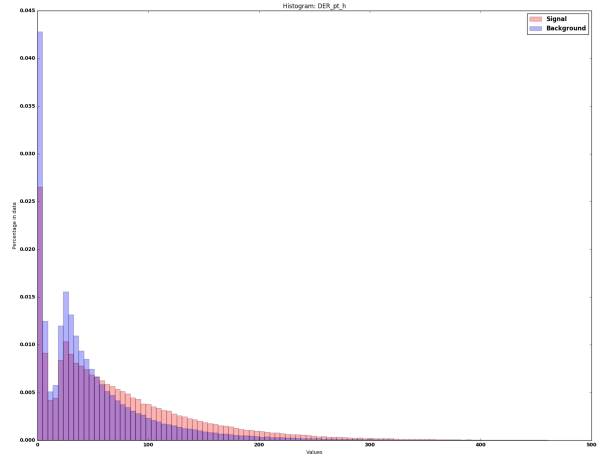


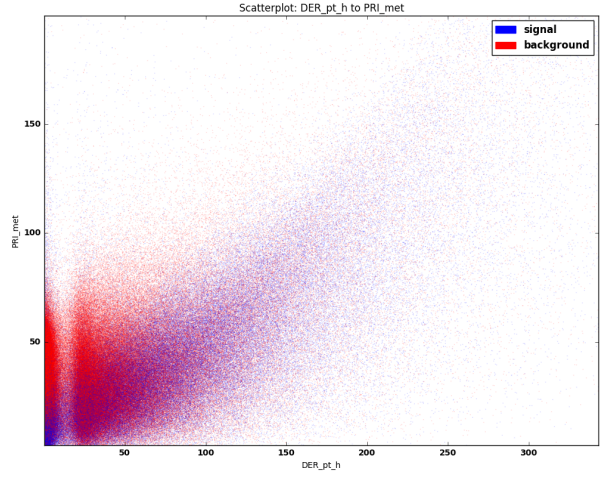
HOLY AMS, BATMAN!

(regarding the challenge, this would place us on rank 998)

2do

- WHY DER_pt_h ?!





2do

- WHY DER_pt_h ?!
- understand top submissions

2do

- WHY DER_pt_h ?!
- understand top submissions
- write the dang thing

2do

- WHY DER_pt_h ?!
- understand top submissions
- write the dang thing
- push the AMS

References I

[1] Higgs boson machine learning challenge.
https://www.kaggle.com/c/higgs-boson.
Accessed: 2016-01-03.

[2] Kaggle homepage.
https://www.kaggle.com/.
Accessed: 2016-01-14.

[3] Official page of the atlas-experiment.
http://www.atlas.ch/calorimeter.html.
Accessed: 2016-01-14.

[4] Claire Adam-Bourdarios, Glen Cowan, Cecile Germain, Isabelle Guyon, Balazs Kegl, and David Rousseau.
Learning to discover: the higgs boson machine learning challenge.
http://www.opendata.cern.ch/record/329, January 2015.
Version 2.3.

References II

[5] Cecile Germain.
Missing features: to impute or not?
https://www.kaggle.com/c/higgs-boson/forums/t/9552/missing-features-to-impute-or-not.
Accessed: 2016-01-14.