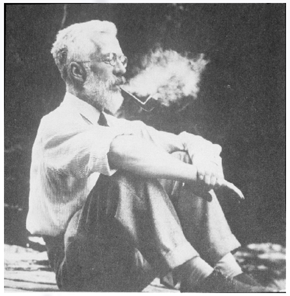
Applied ML

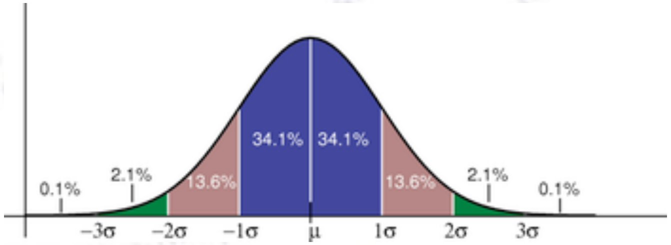
Results and Scores of Small Project








Troels C. Petersen & Carl Johnsen (NBI)



"Statistics is merely a quantisation of common sense - Machine Learning is a sharpening of it!"

Overall comments

The name “Small Project” is misleading, and should have been “Initial project”, because it is by no means small. But you did very well, and so let me start by gently stating, that you have little/nothing to fear - in fact, you did really great!

Grading it was perhaps harder than the project itself, but we have done our best to be as open as possible about the scoring. And to give you a maximum of feedback, we have produced a report for each of you.

The motivation

We wanted you to try the very **real challenge** of optimising models, without knowing their performance on the data it is applied to.

We also wanted you to **individually** run ML algorithms, so that you have the machinery in place after the course.

We insisted that you tried **both tree- and NN-based algorithms**, to get a feel for their differences and similarities.

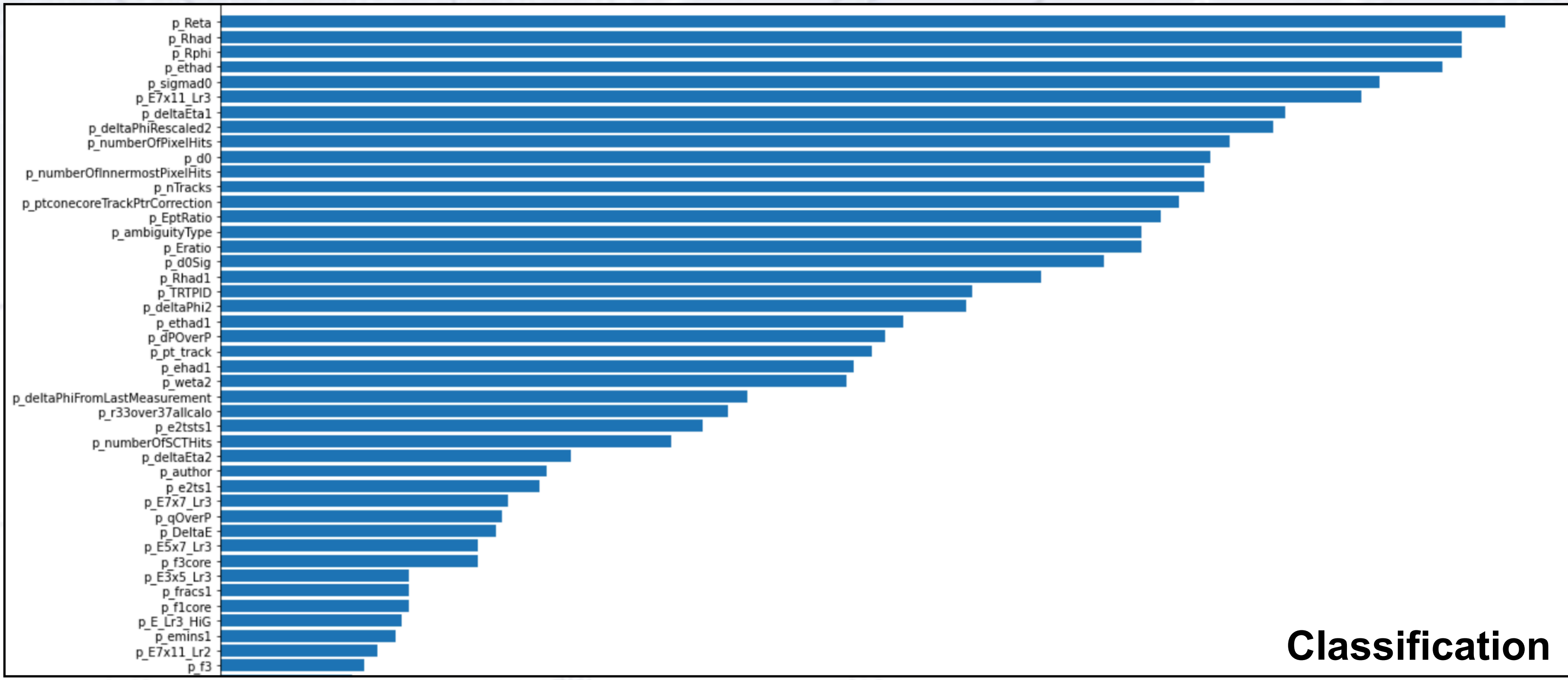
The description file was meant to trigger you to **think about your models**, and what you tried. Also, considerations of size and performance are in place.

Finally, we wanted to **ensure** that you yourself tried all the work and things to consider, to put together ML models and apply them.

Classification Results

Classification variable usage

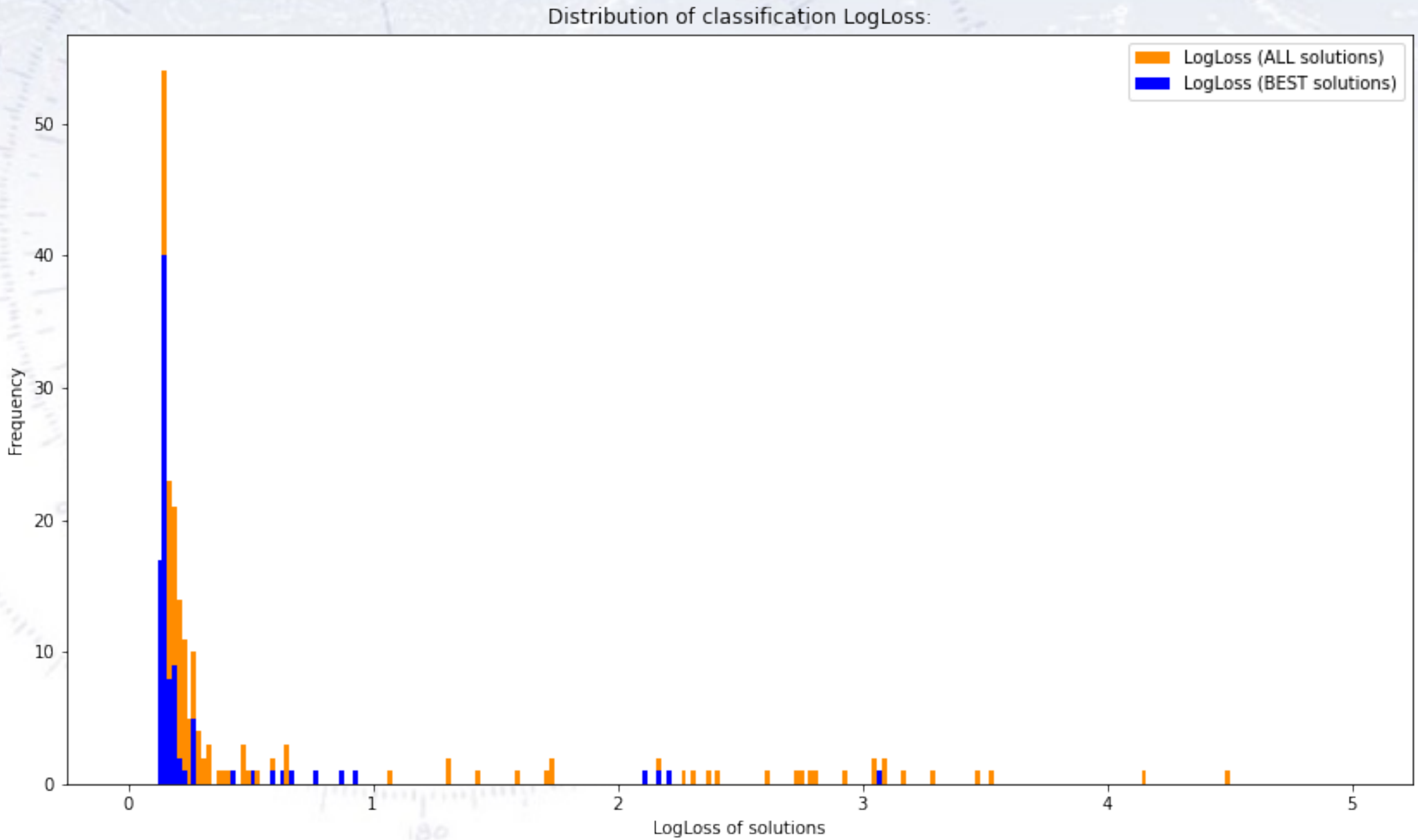
Many (most?) of you have made a good variable ranking. Below you find a variable usage **frequency** plot, showing how often a variable was used.



There is in principle no "correct" result (except trying all combination and HP optimising each!), but your **common** ranking offers a good approximation.

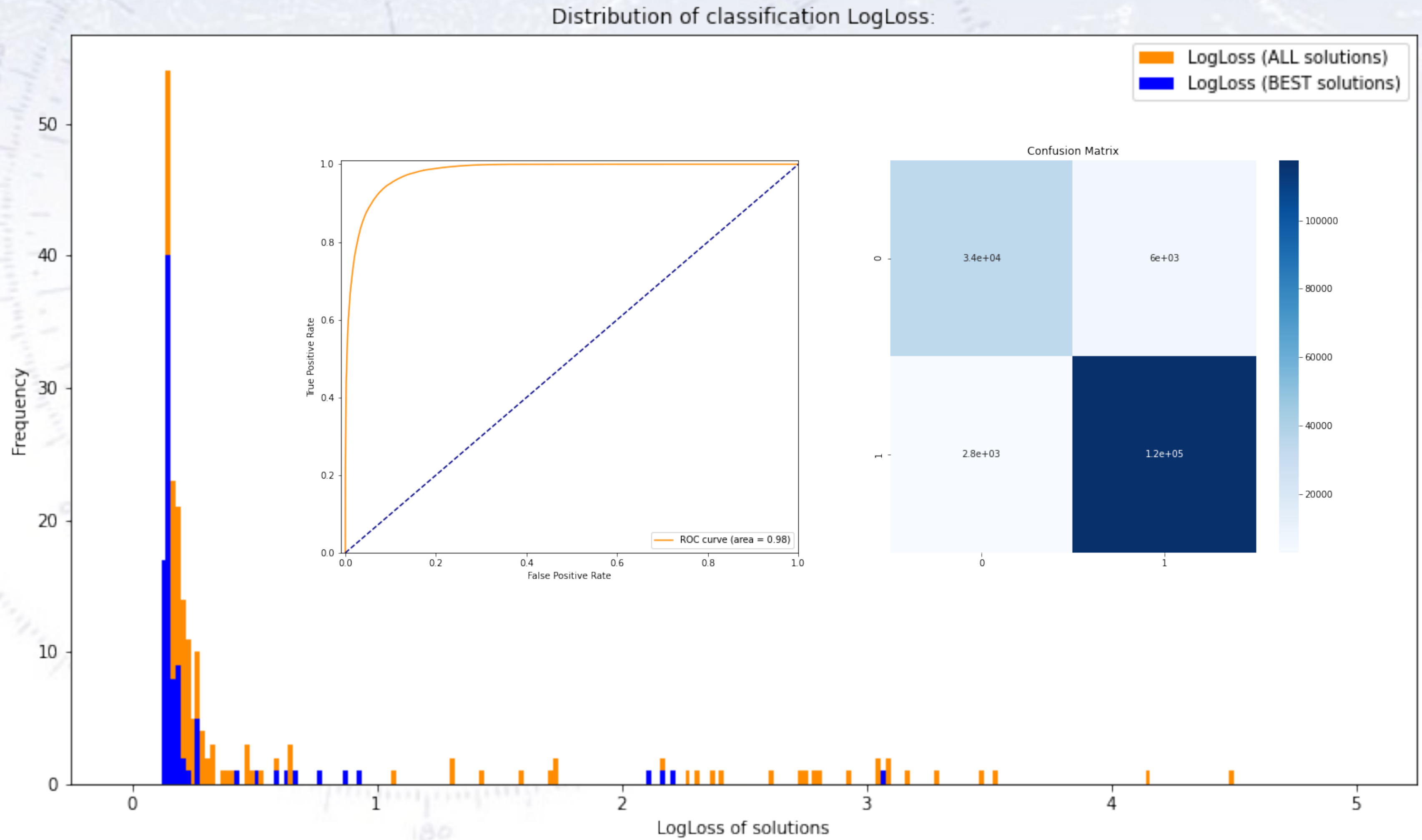
Classification score distribution

The distribution of the (Cross-Entropy) LogLoss values obtained was:



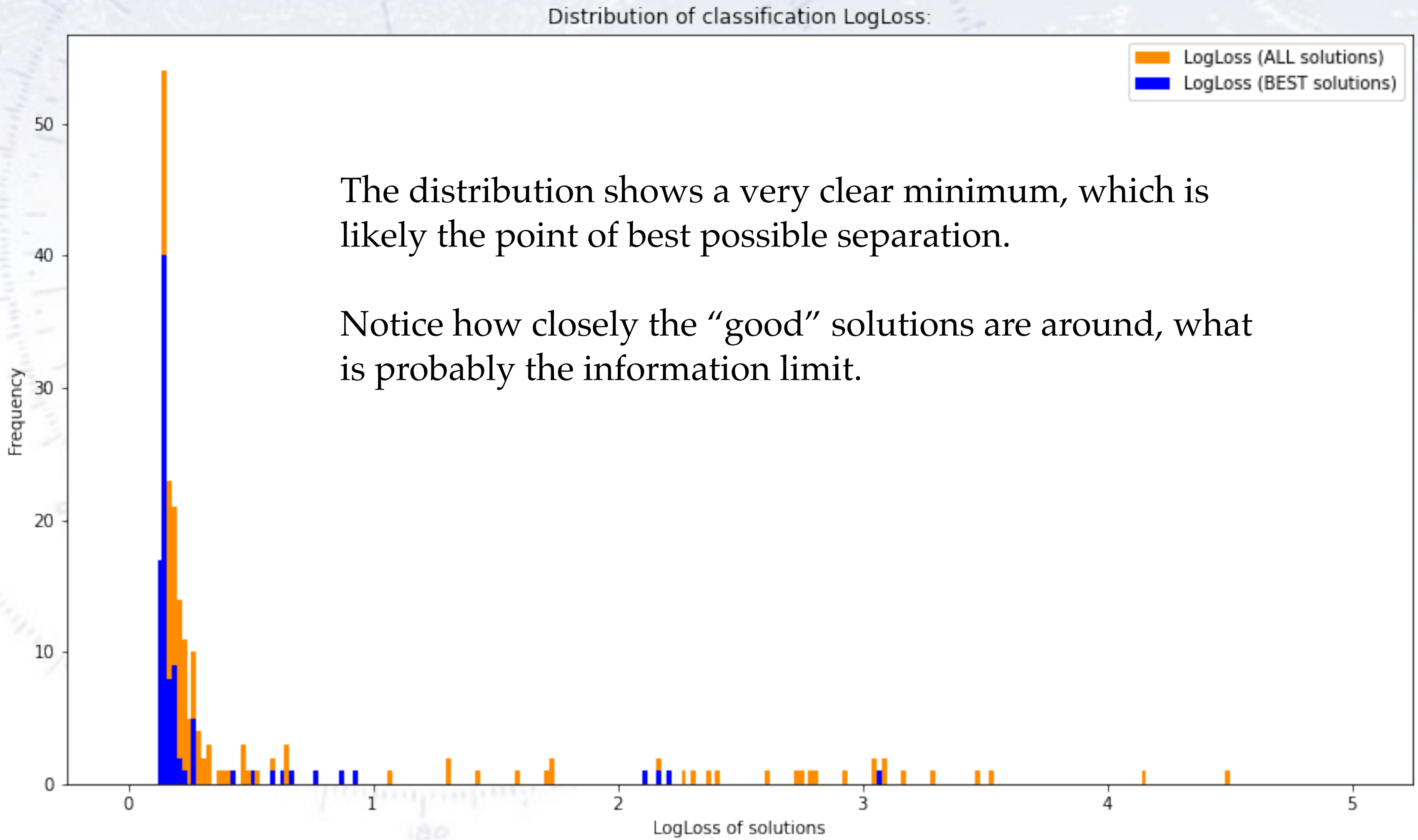
Classification score distribution

The distribution of the (Cross-Entropy) LogLoss values obtained was:



Classification score distribution

The distribution of the (Cross-Entropy) LogLoss values obtained was:

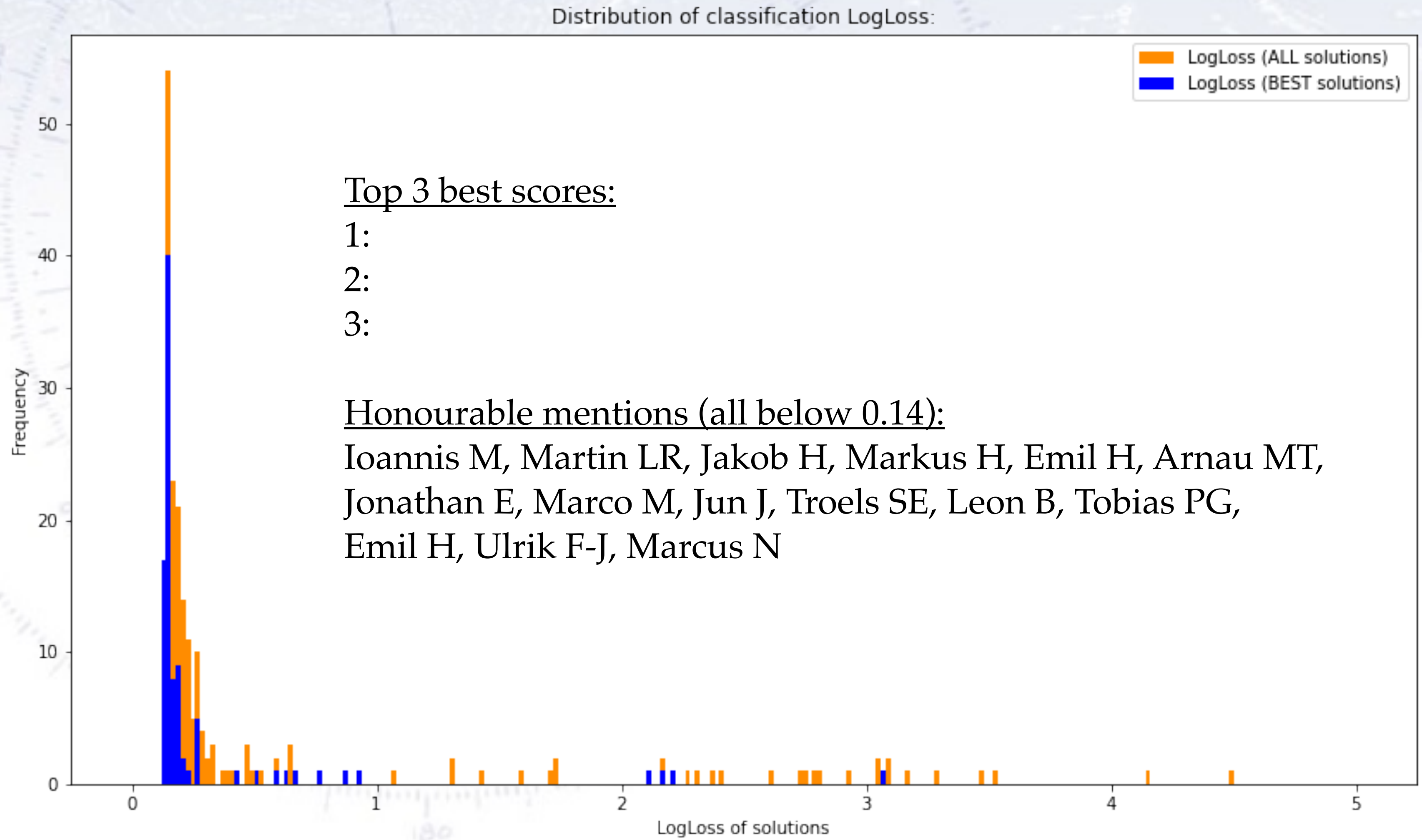


The distribution shows a very clear minimum, which is likely the point of best possible separation.

Notice how closely the “good” solutions are around, what is probably the information limit.

Classification score distribution

The distribution of the (Cross-Entropy) LogLoss values obtained was:



Top 3 best scores:

1:

2:

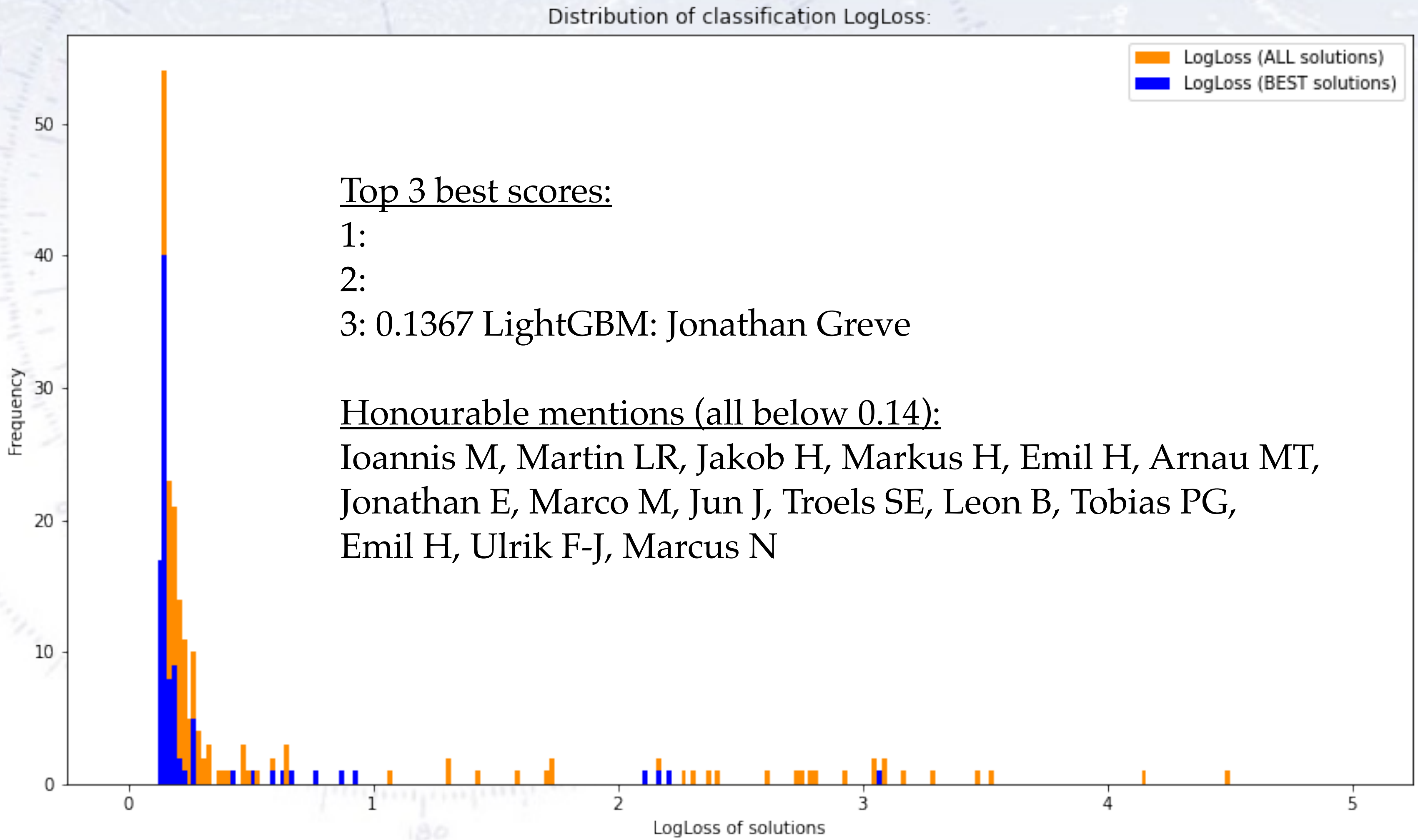
3:

Honourable mentions (all below 0.14):

Ioannis M, Martin LR, Jakob H, Markus H, Emil H, Arnau MT, Jonathan E, Marco M, Jun J, Troels SE, Leon B, Tobias PG, Emil H, Ulrik F-J, Marcus N

Classification score distribution

The distribution of the (Cross-Entropy) LogLoss values obtained was:

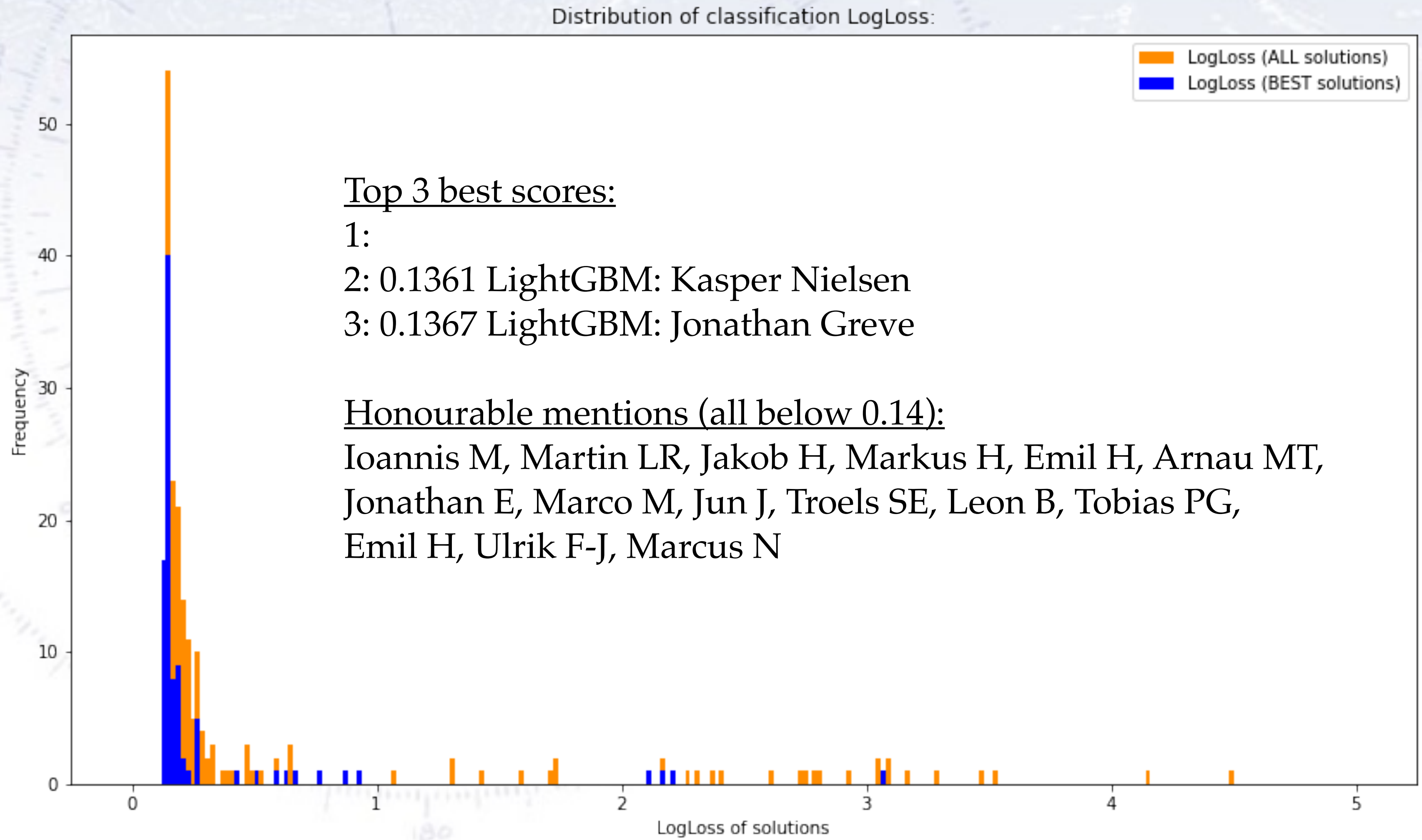


<u>Top 3 best scores:</u>
1:
2:
3: 0.1367 LightGBM: Jonathan Greve

<u>Honourable mentions (all below 0.14):</u>
Ioannis M, Martin LR, Jakob H, Markus H, Emil H, Arnau MT, Jonathan E, Marco M, Jun J, Troels SE, Leon B, Tobias PG, Emil H, Ulrik F-J, Marcus N

Classification score distribution

The distribution of the (Cross-Entropy) LogLoss values obtained was:

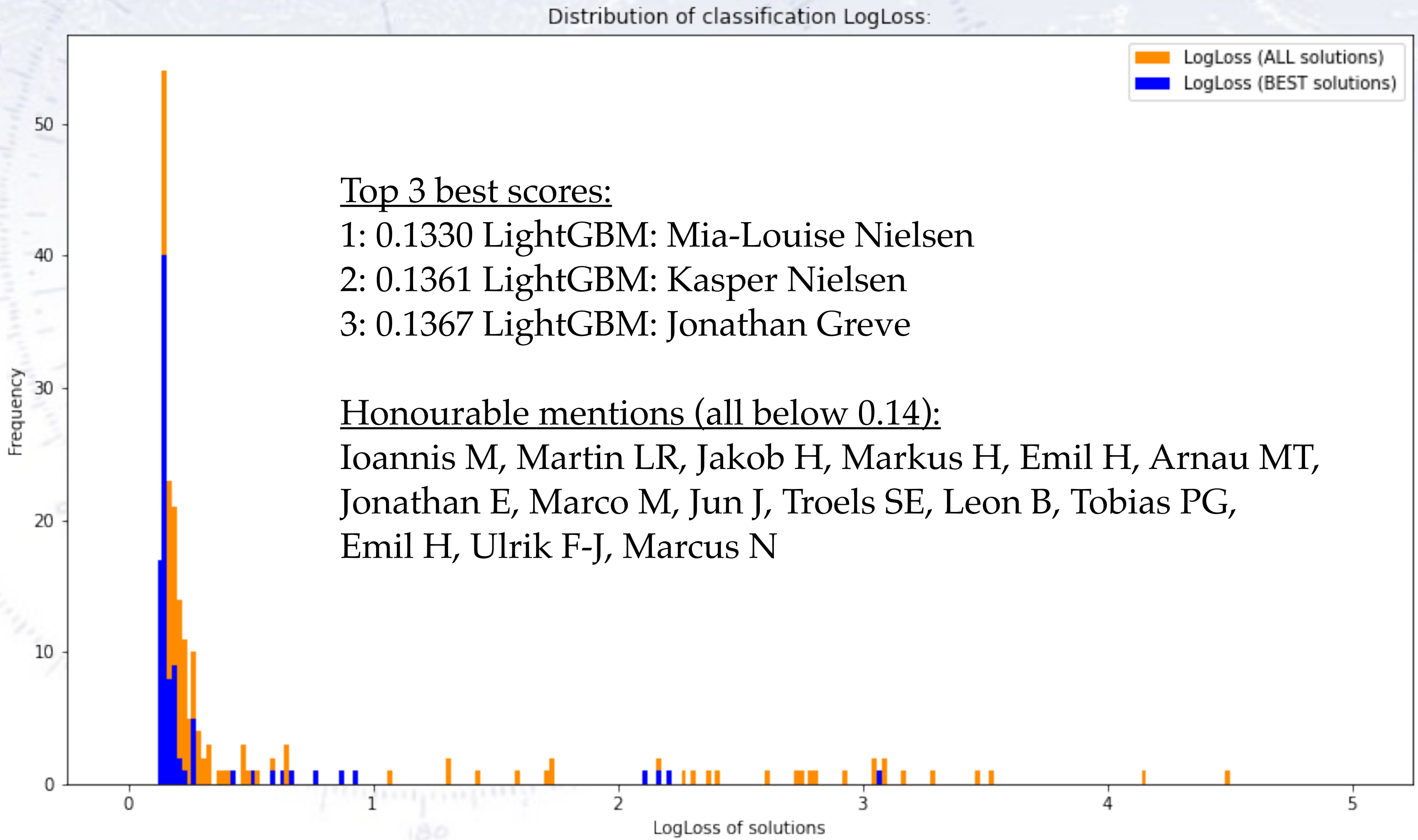


<u>Top 3 best scores:</u>
1:
2: 0.1361 LightGBM: Kasper Nielsen
3: 0.1367 LightGBM: Jonathan Greve

<u>Honourable mentions (all below 0.14):</u>
Ioannis M, Martin LR, Jakob H, Markus H, Emil H, Arnau MT, Jonathan E, Marco M, Jun J, Troels SE, Leon B, Tobias PG, Emil H, Ulrik F-J, Marcus N

Classification score distribution

The distribution of the (Cross-Entropy) LogLoss values obtained was:



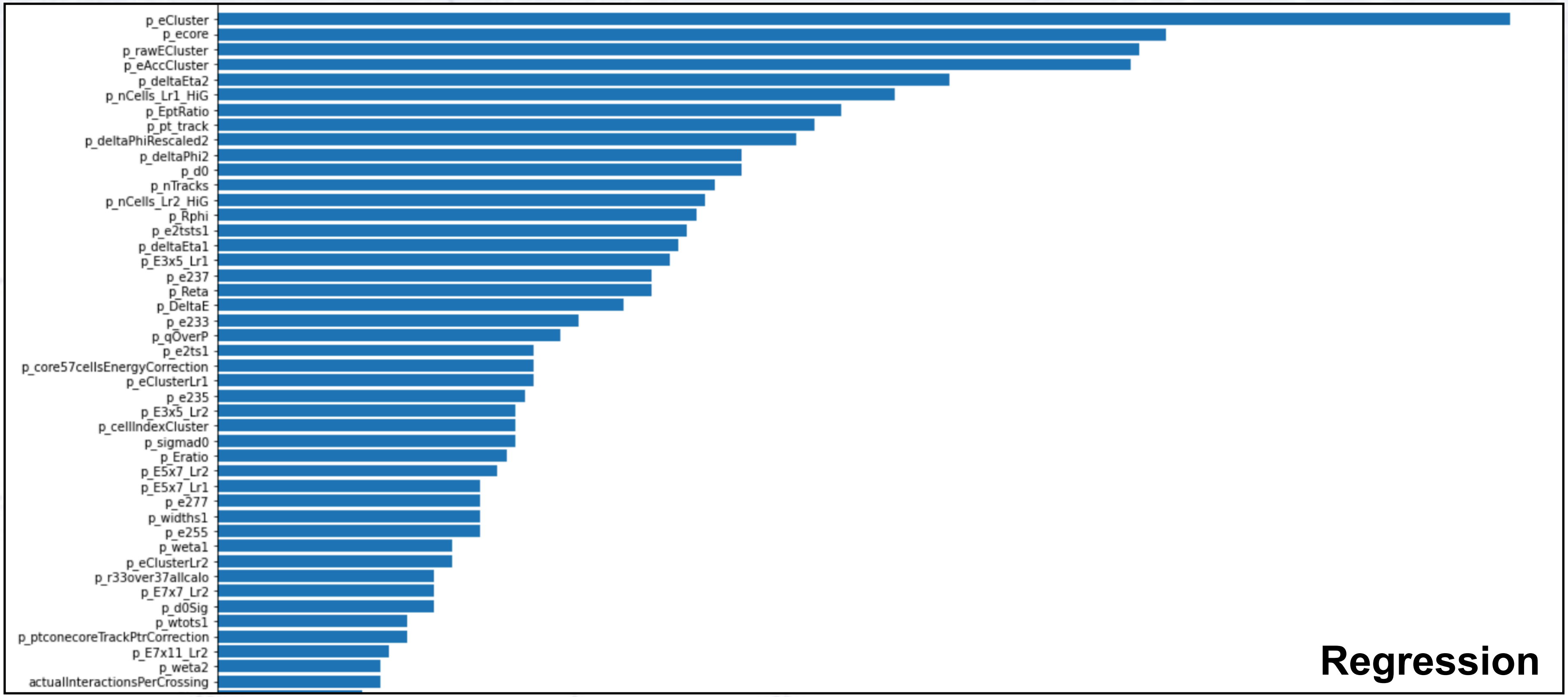
<u>Top 3 best scores:</u>
1: 0.1330 LightGBM: Mia-Louise Nielsen
2: 0.1361 LightGBM: Kasper Nielsen
3: 0.1367 LightGBM: Jonathan Greve

<u>Honourable mentions (all below 0.14):</u>
Ioannis M, Martin LR, Jakob H, Markus H, Emil H, Arnau MT, Jonathan E, Marco M, Jun J, Troels SE, Leon B, Tobias PG, Emil H, Ulrik F-J, Marcus N

Regression Results

Regression variable usage

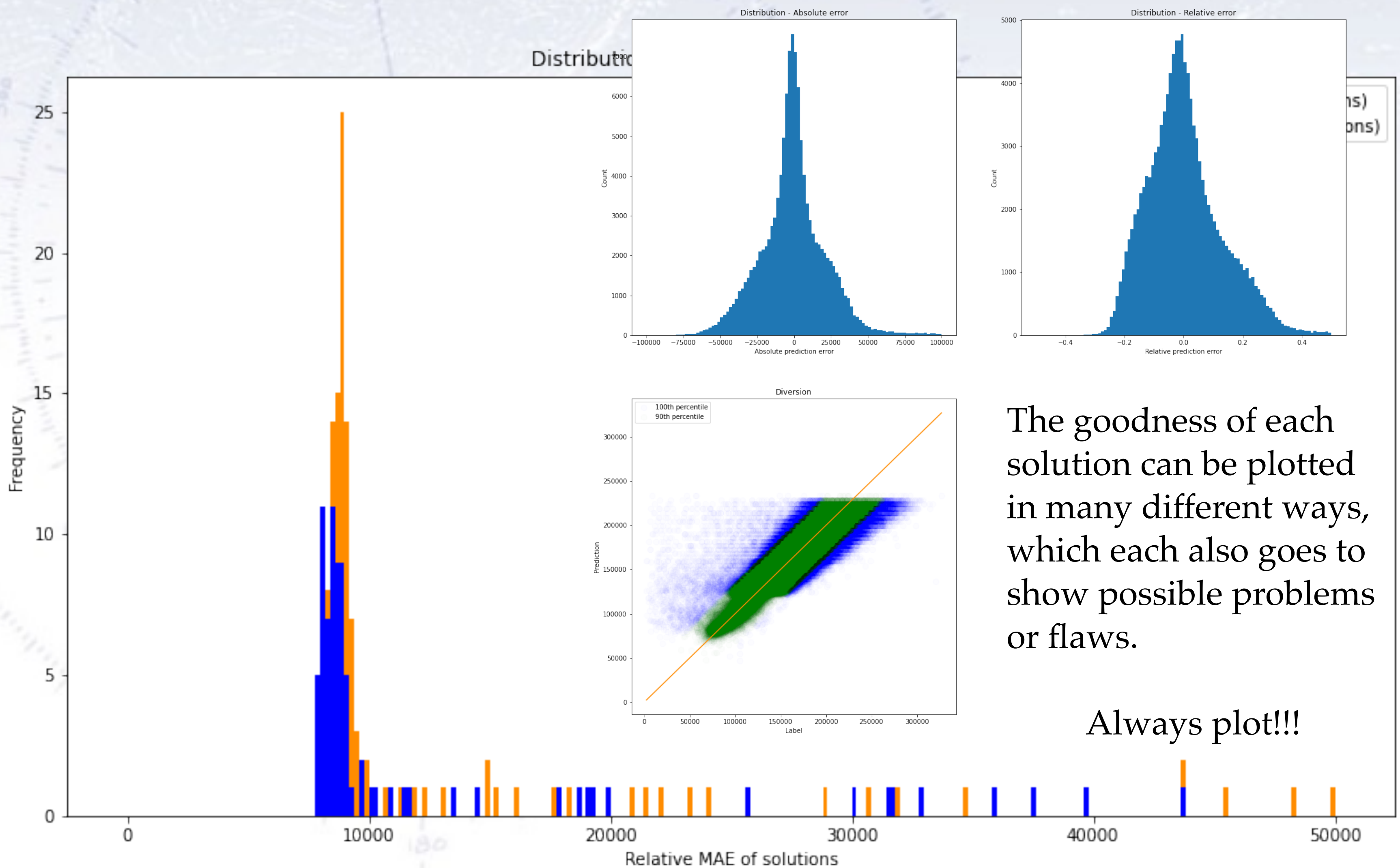
The variables have changed drastically from the classification case. There is NO overlap at all for the top 10-15 variables! Classification and Regression are two very different tasks.



Last year, I didn't restrict the list of variables one could use. The most important variable then happens to be ATLAS' own energy prediction - not surprisingly!

Regression score distribution

The distribution of the relative MAE (i.e. MAE((E-T)/T)) values obtained was:

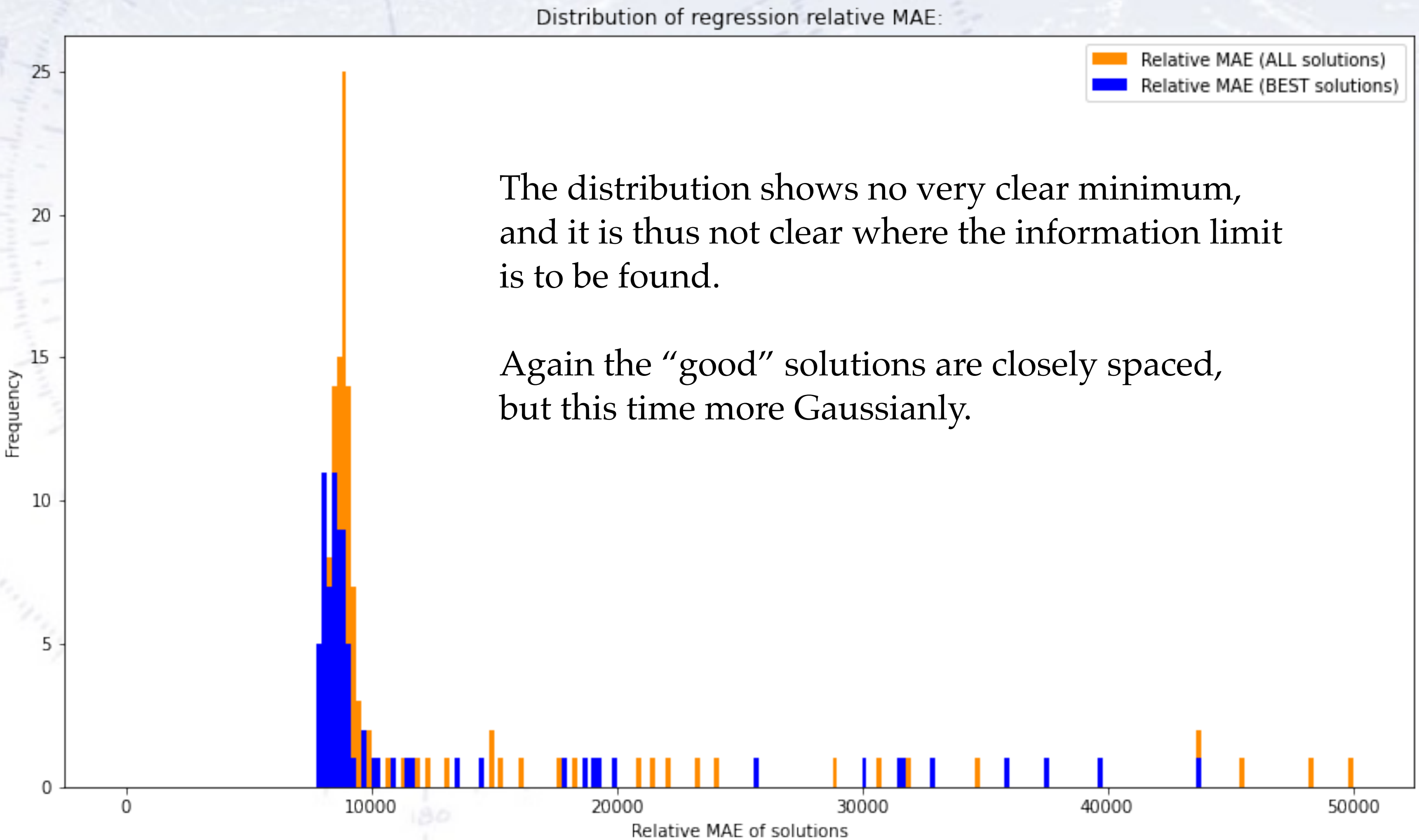


The goodness of each solution can be plotted in many different ways, which each also goes to show possible problems or flaws.

Always plot!!!

Regression score distribution

The distribution of the relative MAE (i.e. MAE((E-T)/T)) values obtained was:

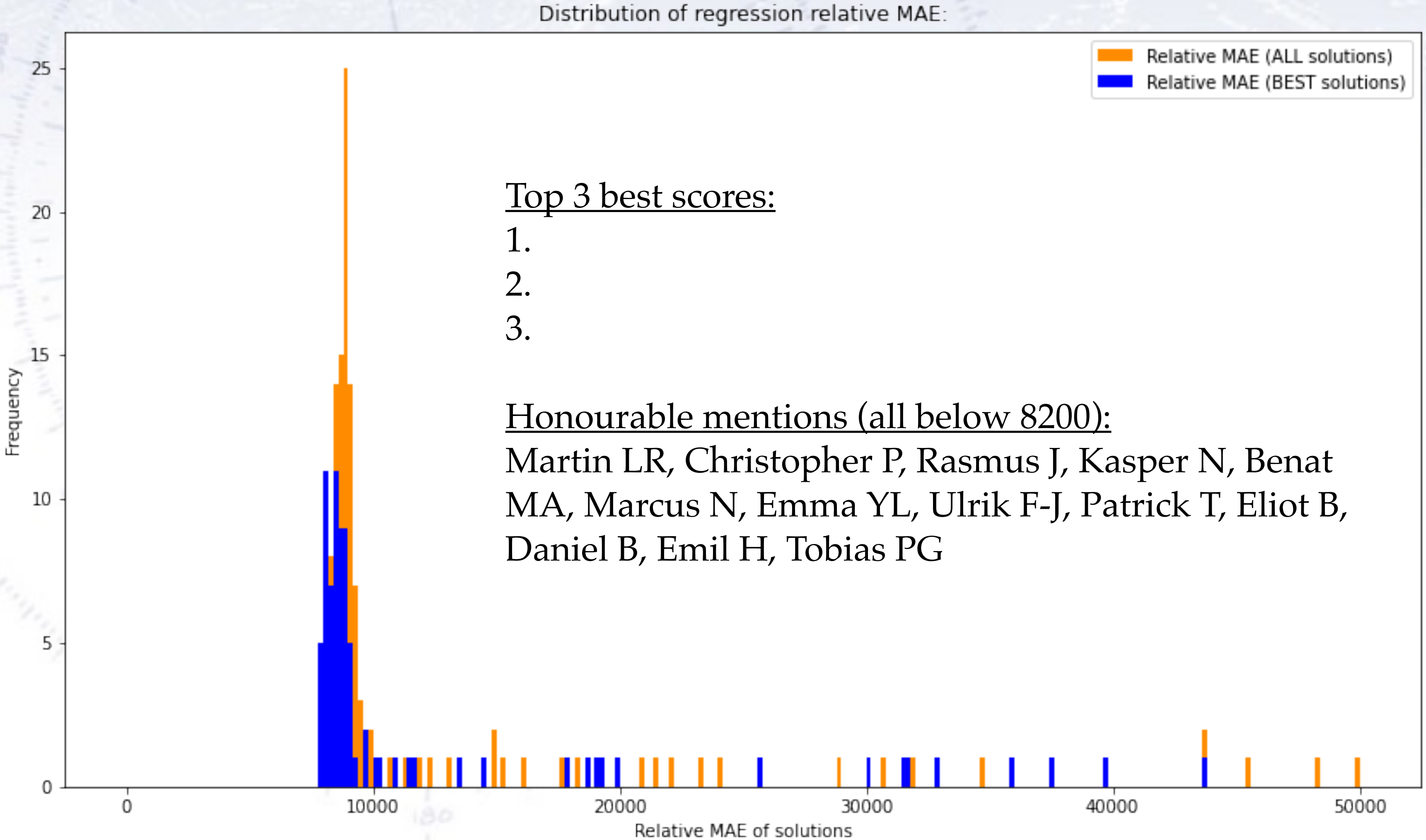


The distribution shows no very clear minimum, and it is thus not clear where the information limit is to be found.

Again the "good" solutions are closely spaced, but this time more Gaussianly.

Regression score distribution

The distribution of the relative MAE (i.e. MAE((E-T)/T)) values obtained was:



<u>Top 3 best scores:</u>

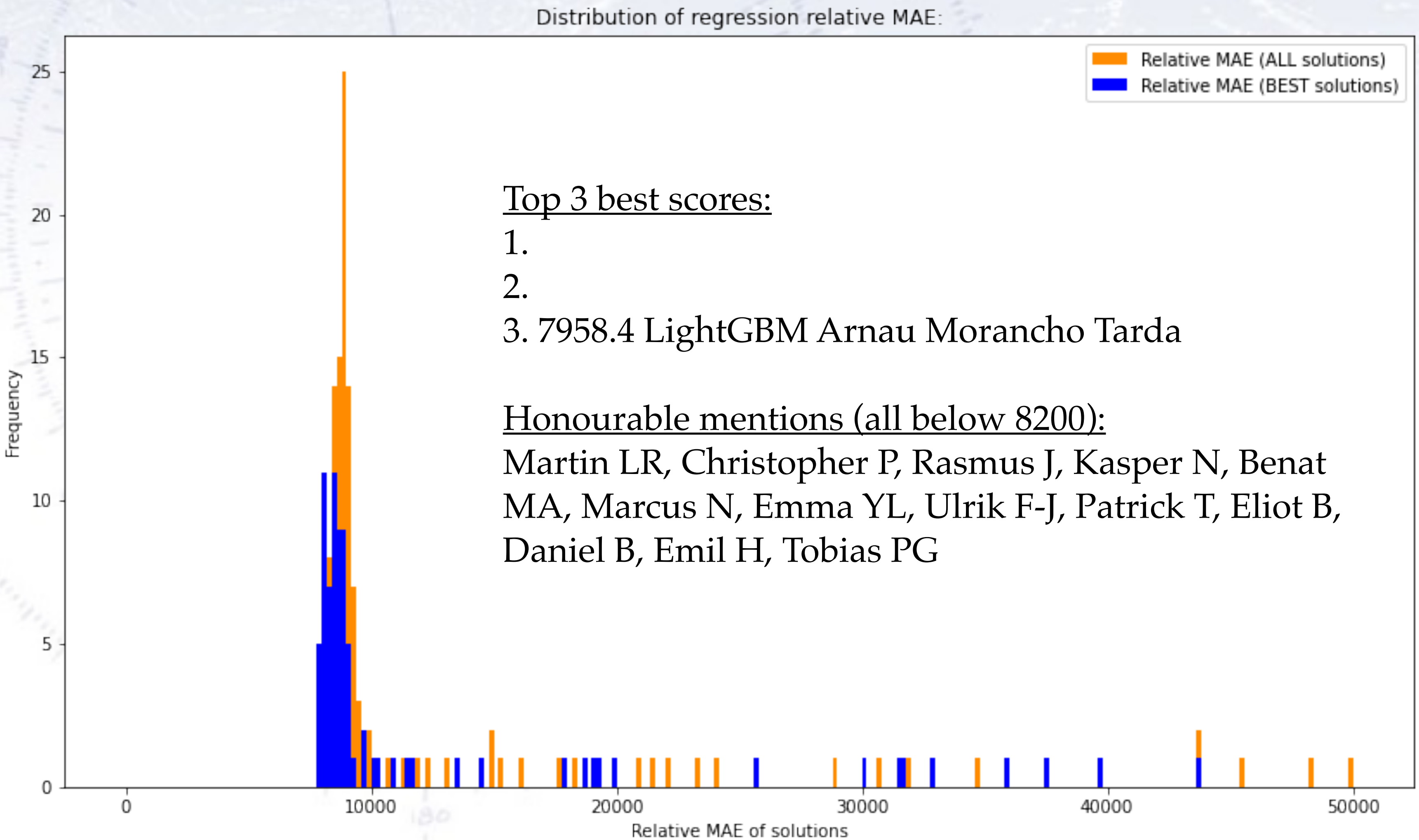
1.
2.
3.

<u>Honourable mentions (all below 8200):</u>

Martin LR, Christopher P, Rasmus J, Kasper N, Benat MA, Marcus N, Emma YL, Ulrik F-J, Patrick T, Eliot B, Daniel B, Emil H, Tobias PG

Regression score distribution

The distribution of the relative MAE (i.e. MAE((E-T)/T)) values obtained was:



Top 3 best scores:

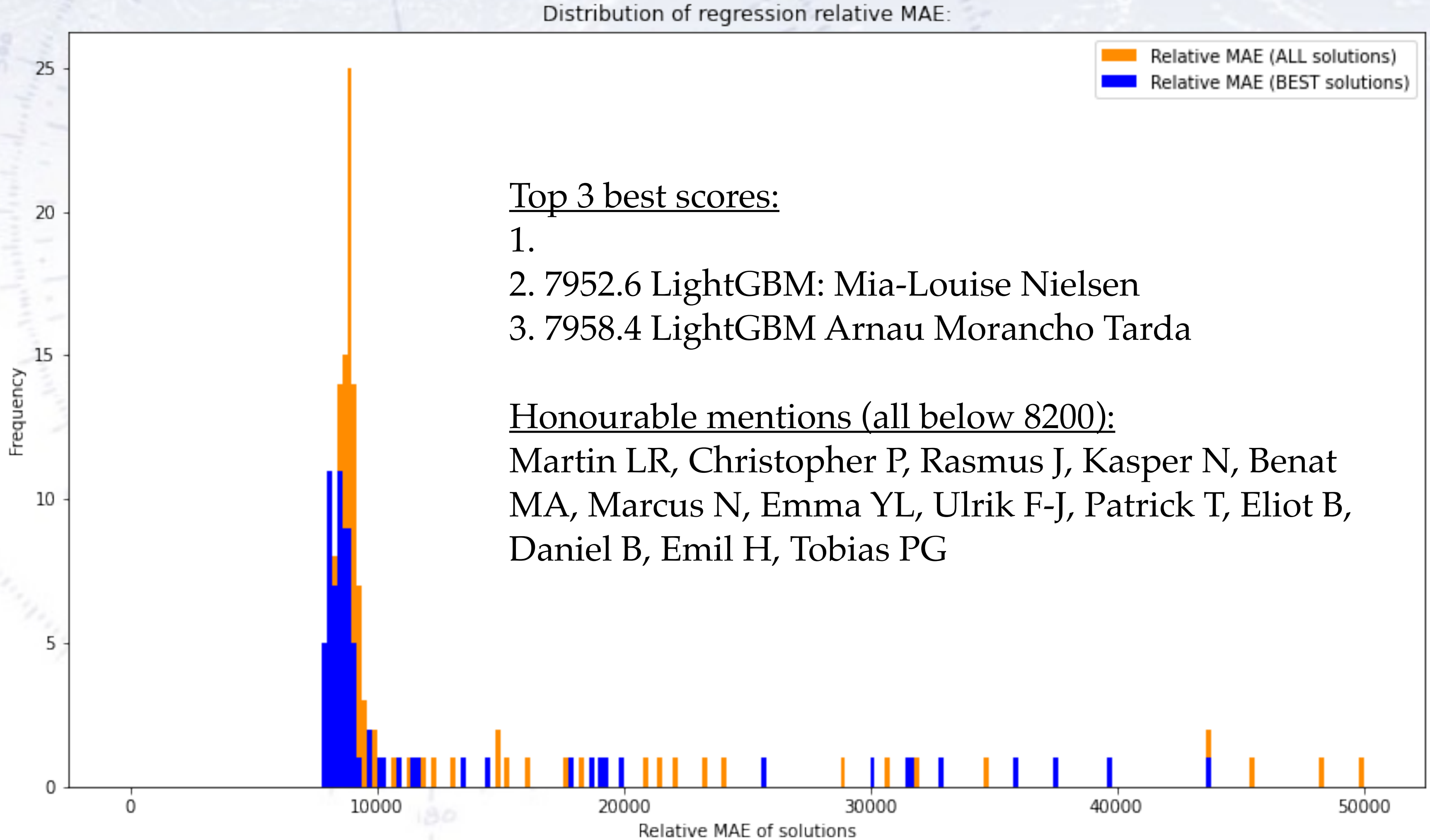
1.
2.
3. 7958.4 LightGBM Arnau Morancho Tarda

Honourable mentions (all below 8200):

Martin LR, Christopher P, Rasmus J, Kasper N, Benat MA, Marcus N, Emma YL, Ulrik F-J, Patrick T, Eliot B, Daniel B, Emil H, Tobias PG

Regression score distribution

The distribution of the relative MAE (i.e. MAE((E-T)/T)) values obtained was:



<u>Top 3 best scores:</u>

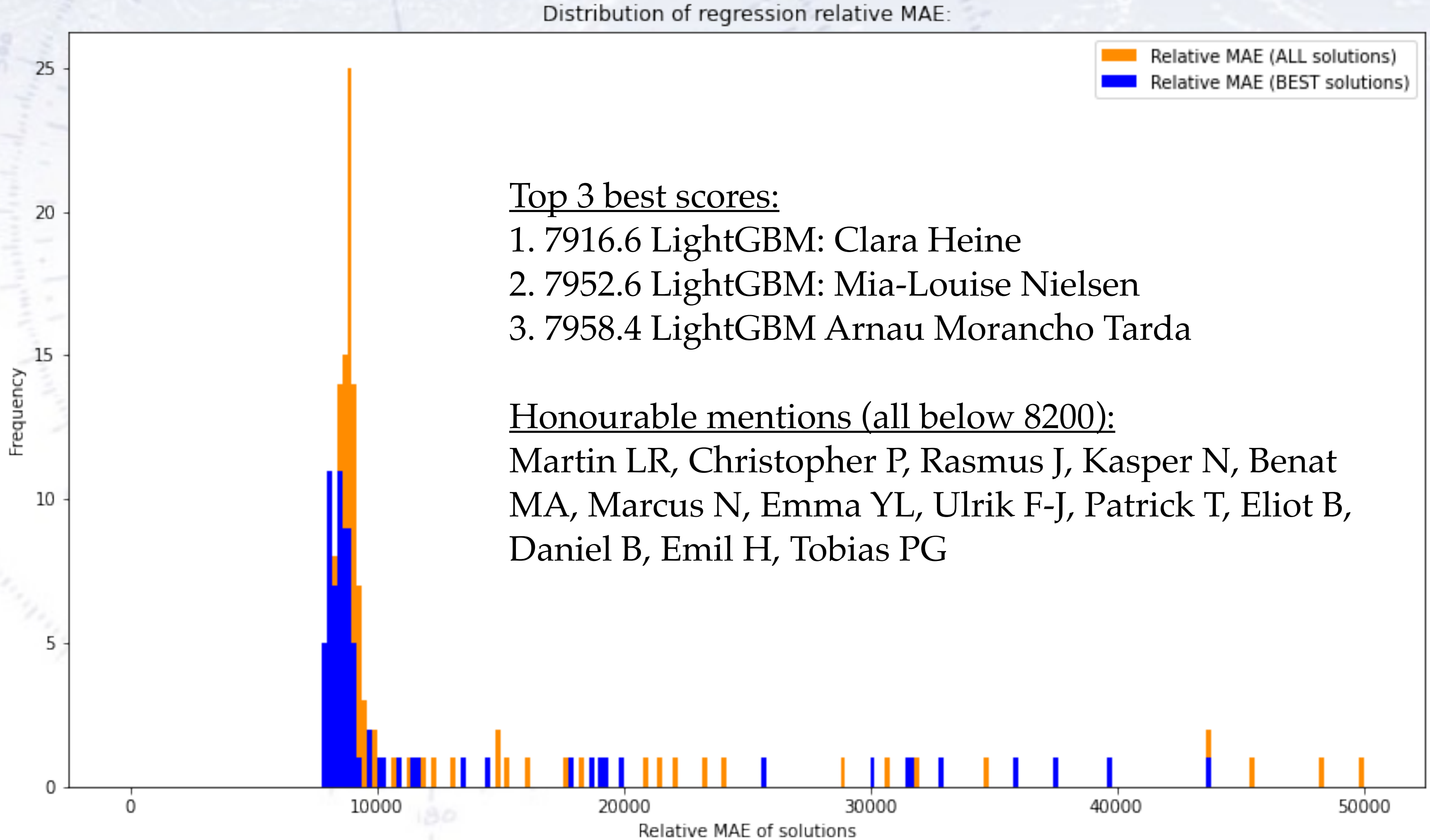
1.
2. 7952.6 LightGBM: Mia-Louise Nielsen
3. 7958.4 LightGBM Arnau Morancho Tarda

<u>Honourable mentions (all below 8200):</u>

Martin LR, Christopher P, Rasmus J, Kasper N, Benat MA, Marcus N, Emma YL, Ulrik F-J, Patrick T, Eliot B, Daniel B, Emil H, Tobias PG

Regression score distribution

The distribution of the relative MAE (i.e. MAE((E-T)/T)) values obtained was:



Top 3 best scores:

1. 7916.6 LightGBM: Clara Heine
2. 7952.6 LightGBM: Mia-Louise Nielsen
3. 7958.4 LightGBM Arnau Morancho Tarda

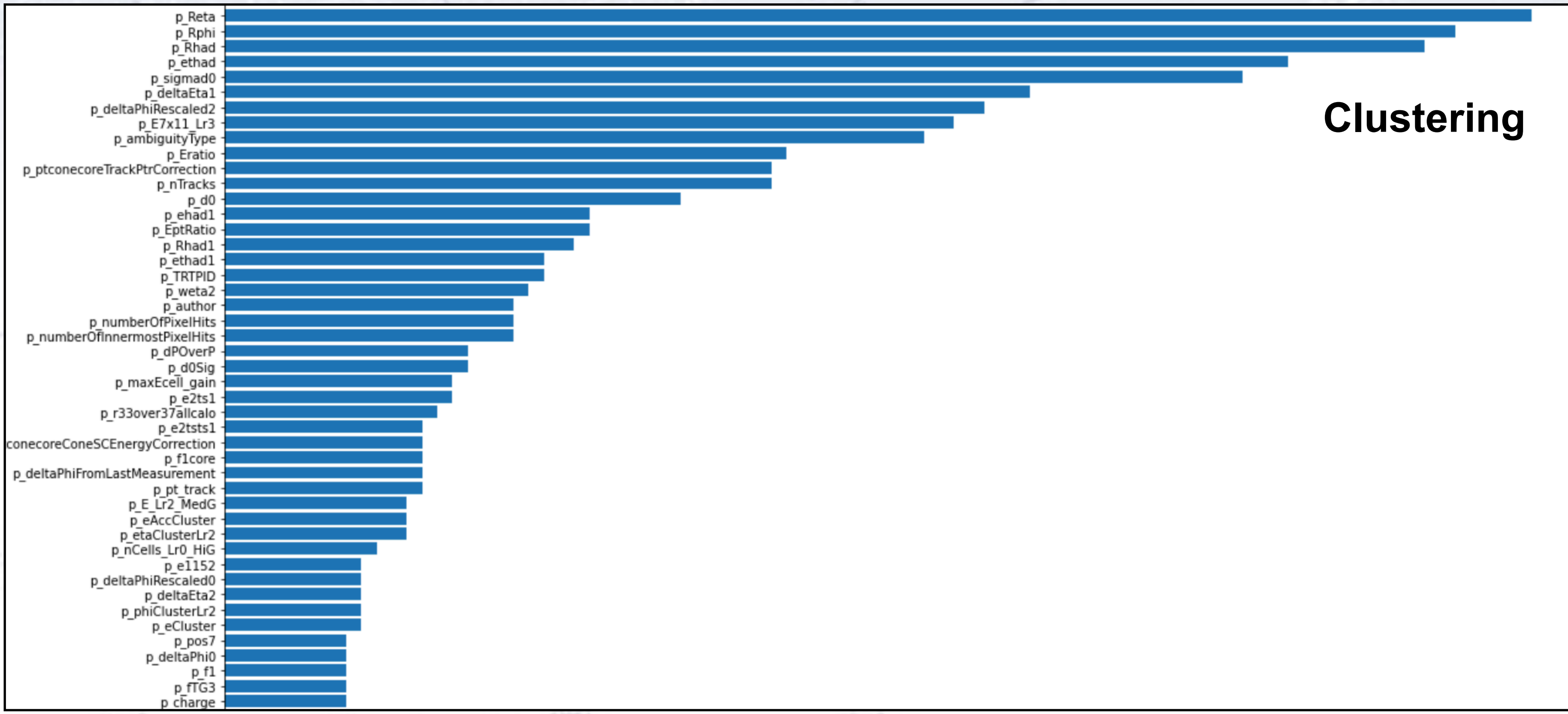
Honourable mentions (all below 8200):

Martin LR, Christopher P, Rasmus J, Kasper N, Benat MA, Marcus N, Emma YL, Ulrik F-J, Patrick T, Eliot B, Daniel B, Emil H, Tobias PG

Clustering Results

Clustering variable usage

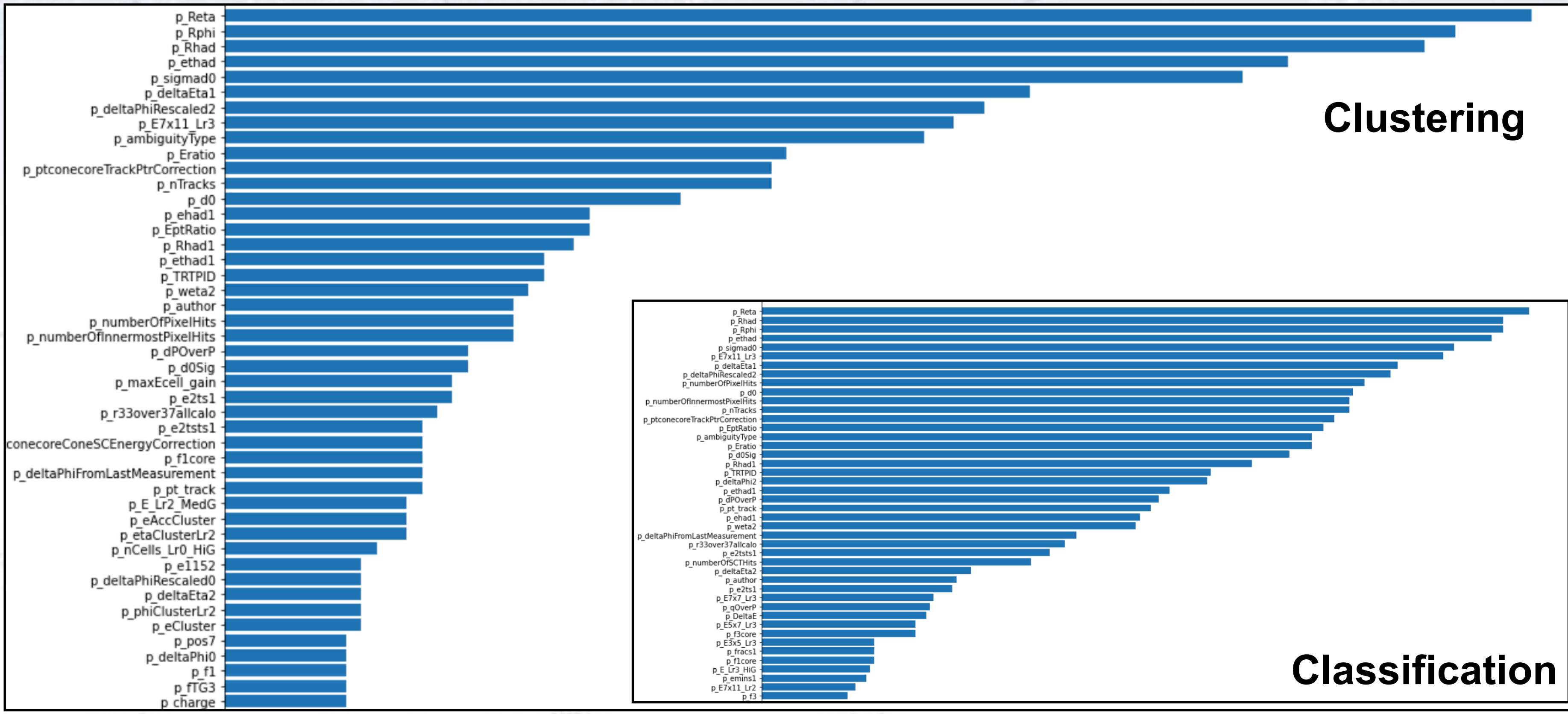
I would have thought, that the clustering variable usage would be near-identical to that of the (supervised) classification task. However, it is not entirely…



It is also a "hard" (i.e. under defined) task of choosing variables for clustering, when the task/target is unknown. It takes insight and domain knowledge…

Clustering variable usage

I would have thought, that the clustering variable usage would be near-identical to that of the (supervised) classification task. However, it is not entirely…



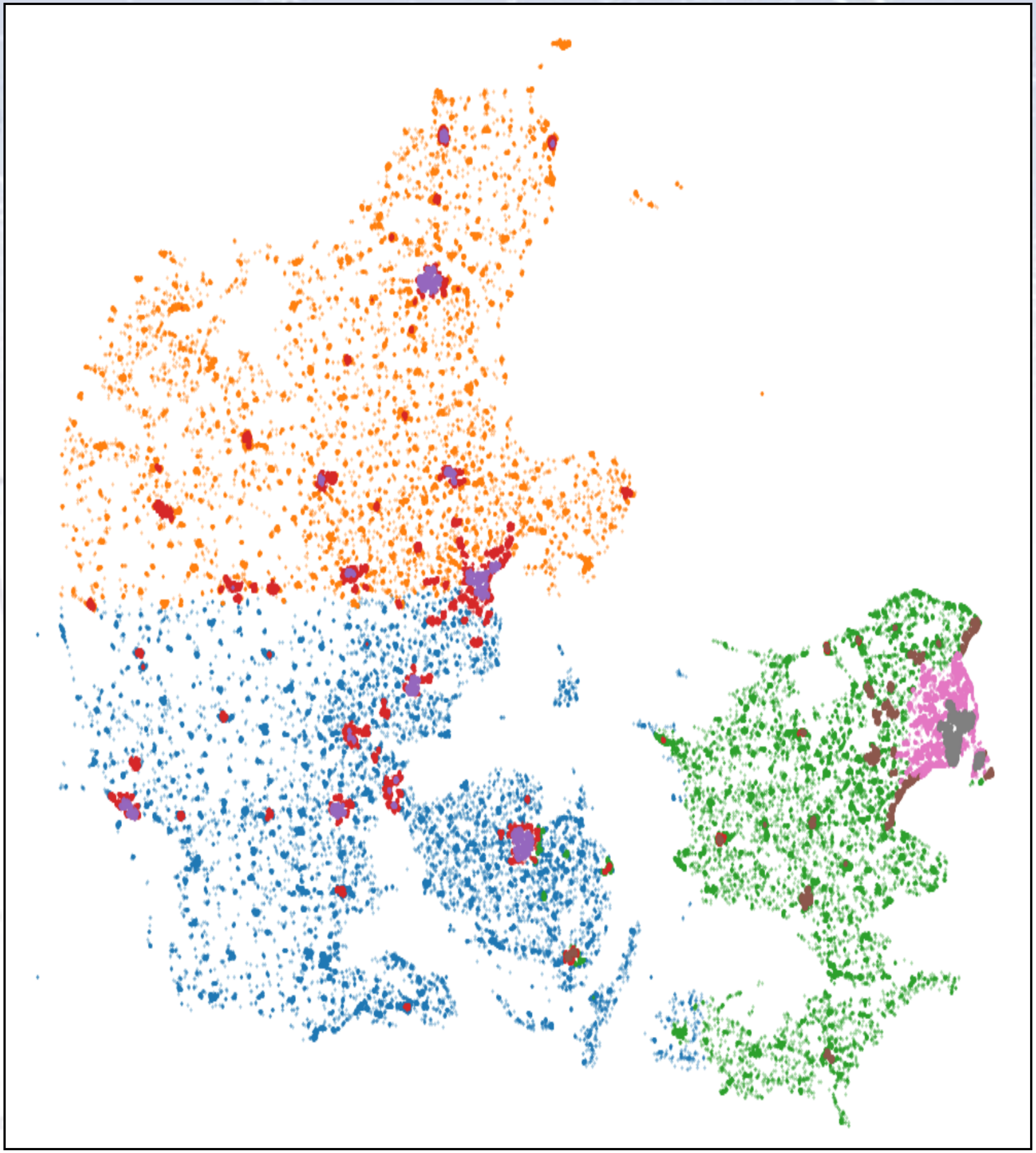
It is also a "hard" (i.e. under defined) task of choosing variables for clustering, when the task/target is unknown. It takes insight and domain knowledge…

Clustering housing

While postal codes are good, they are not very useful in clustering Denmark.

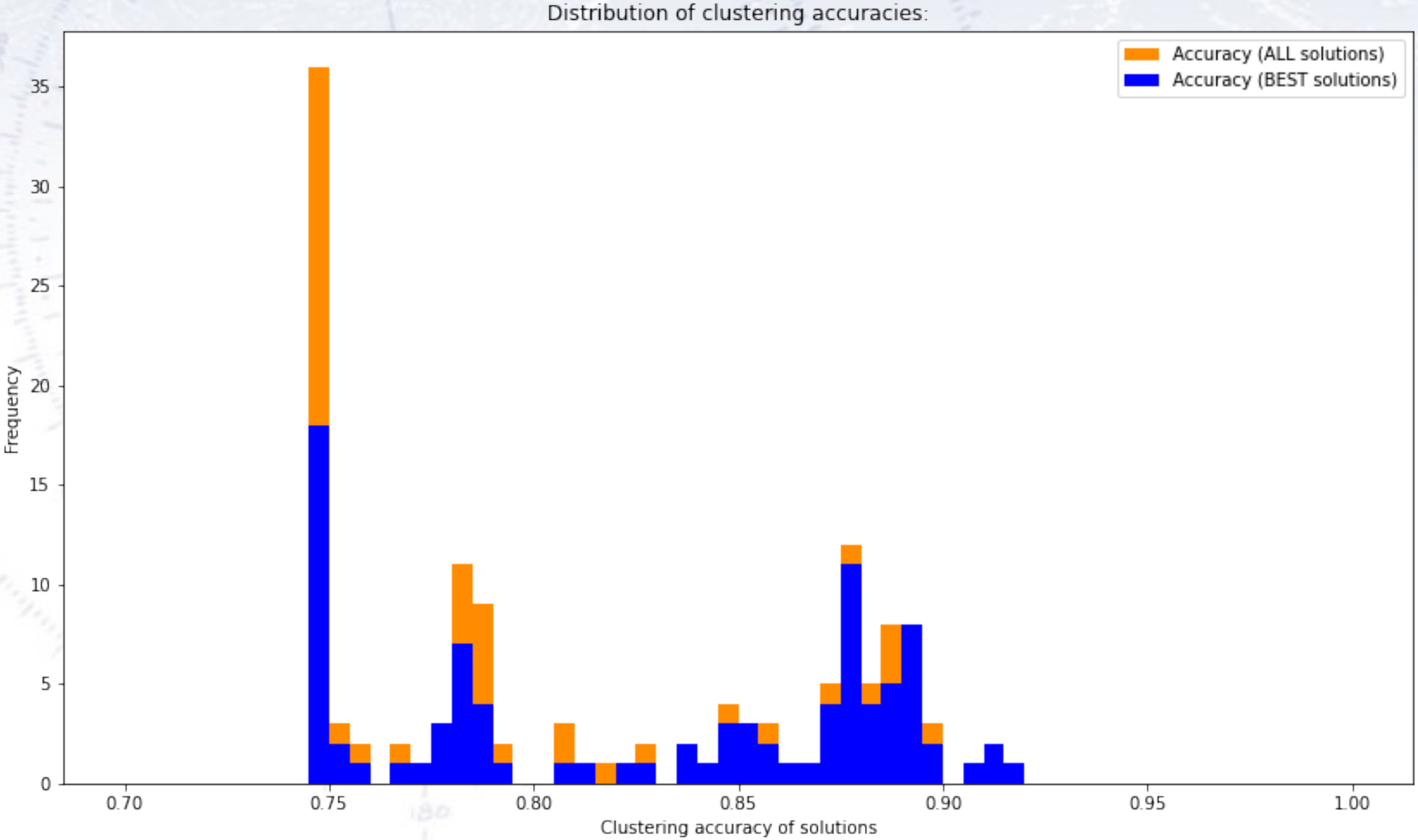
However, using just a few variables (x, y, density, price/m^2), one can cluster villas in Denmark very efficiently.

In this way, one can follow trends for a type of house much better.



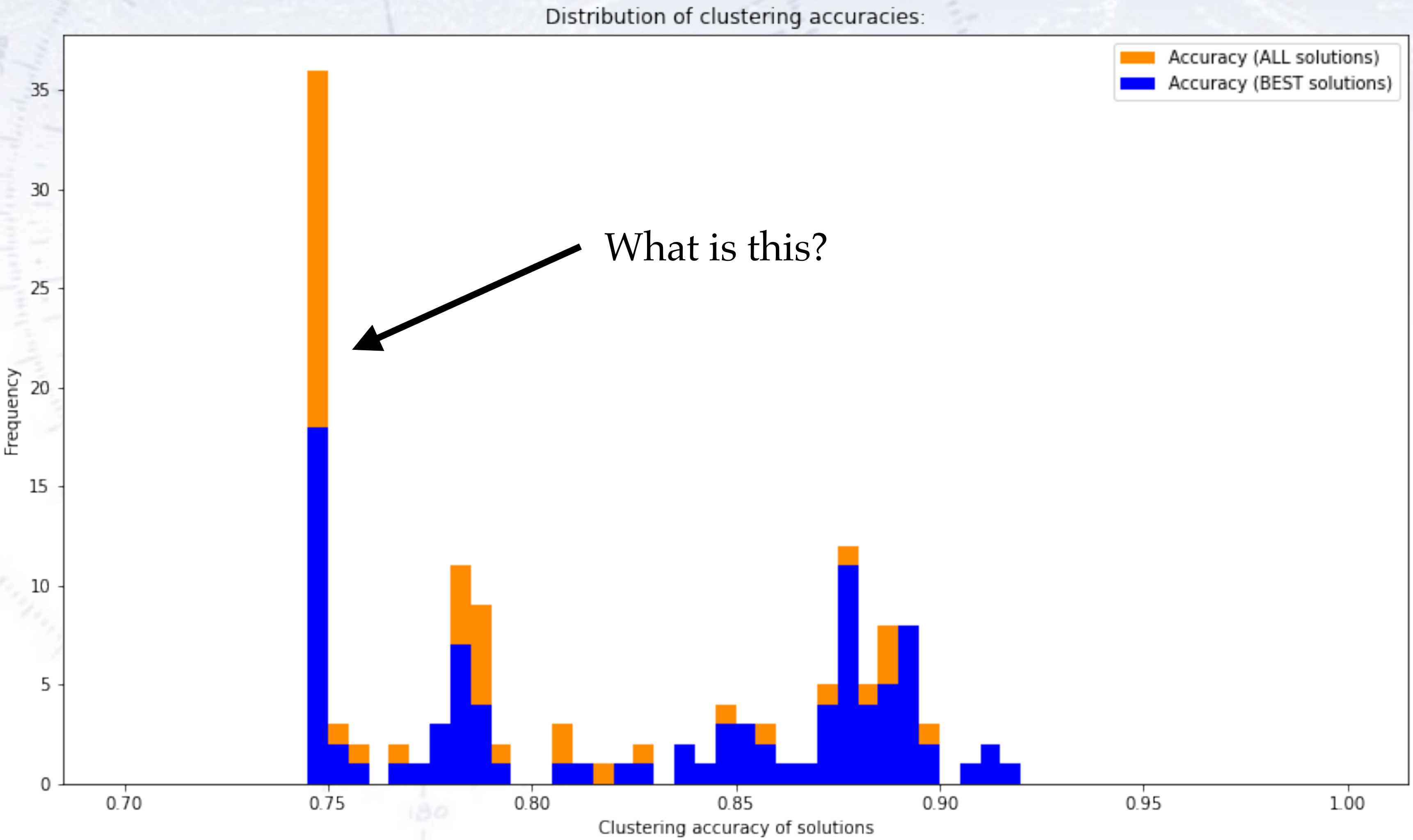
Clustering accuracy distribution

The accuracy of the clustering (when assigned either electron or not) was:



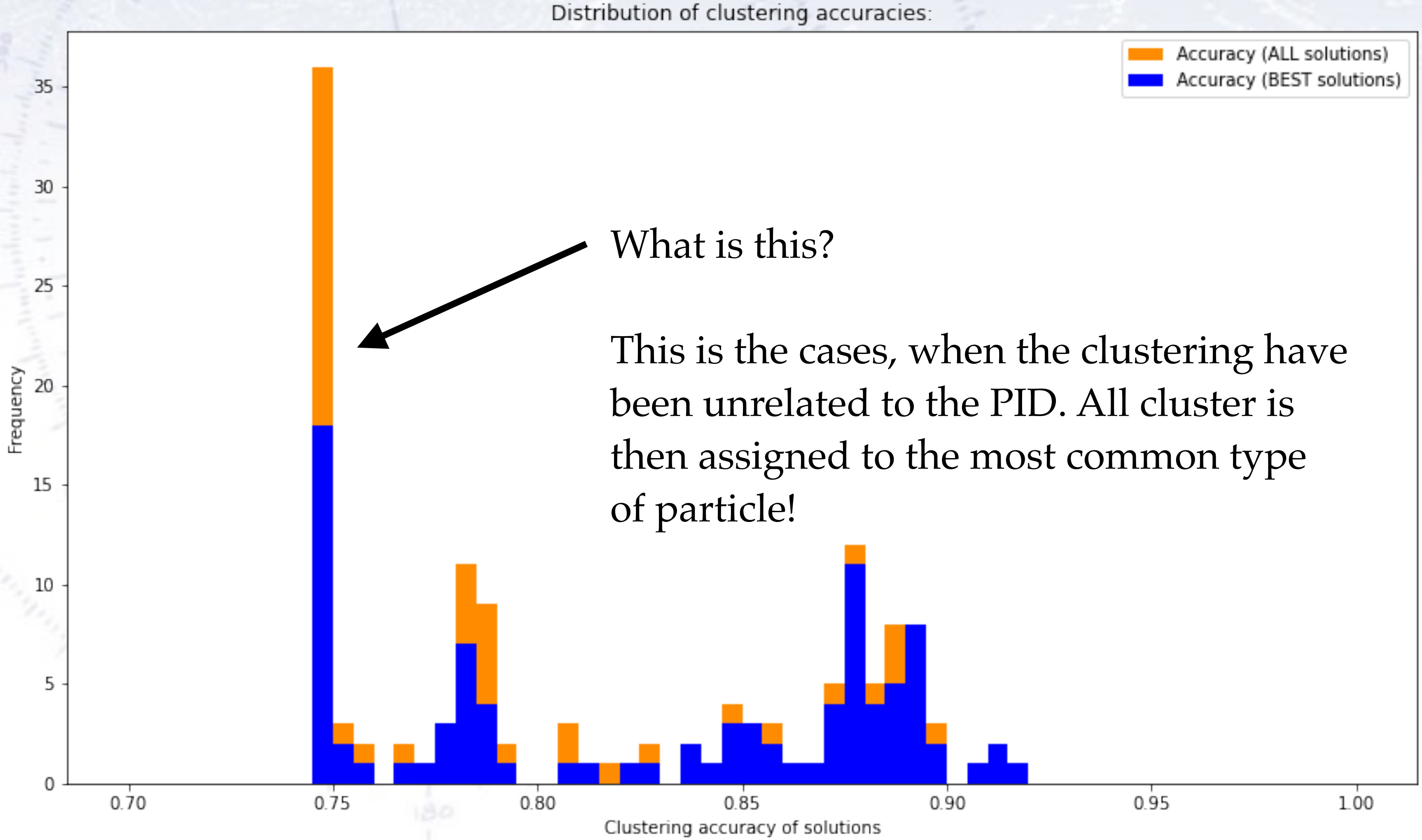
Clustering accuracy distribution

The accuracy of the clustering (when assigned either electron or not) was:



Clustering accuracy distribution

The accuracy of the clustering (when assigned either electron or not) was:

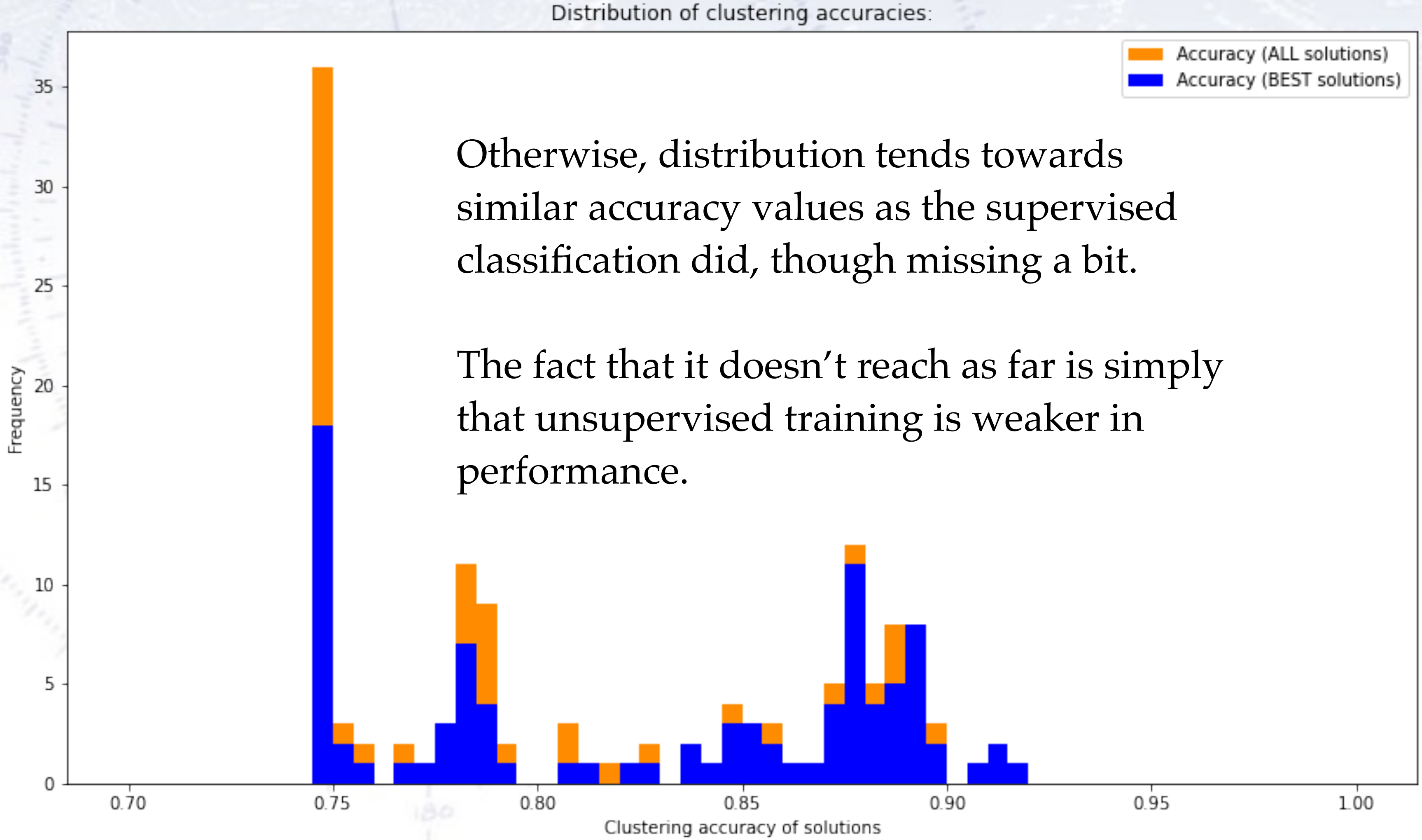


What is this?

This is the cases, when the clustering have been unrelated to the PID. All cluster is then assigned to the most common type of particle!

Clustering accuracy distribution

The accuracy of the clustering (when assigned either electron or not) was:



Otherwise, distribution tends towards similar accuracy values as the supervised classification did, though missing a bit.

The fact that it doesn't reach as far is simply that unsupervised training is weaker in performance.

Scoring your solutions

How do we grade your projects?

Final Score:

You submitted a full solution, from which you get:

Your choice of methods based on your description was scored as follows [0, 6]:
Your solution entailed N different algorithms, which gives you a score of [0, 6]:

Your best performance for classification gave: $\max(0, (-\log(\text{CrossEntropy} - 0.12)) \times 1.4)$:
Your variable choice for classification was scored 4 x (VarFreq(you) / VarFreq(top)):
Your classification had 0 penalties, totalling to:

Your best performance for regression gave: $\max(0, -\log(\text{MAD}((E-T)/T)/7800-1) \times 1.8)$:
Your variable choice for regression was scored 5 × (VarFreq(you)/VarFreq(top)):
Your regression had 0 penalties, totalling to:

Your best performance for clustering gave: $\max(0, (\text{Accuracy} - 0.75) \times 20)$:
Your variable choice for clustering was scored (VarFreq(you) / VarFreq(top)):
Your clustering had 0 penalties, totalling to:

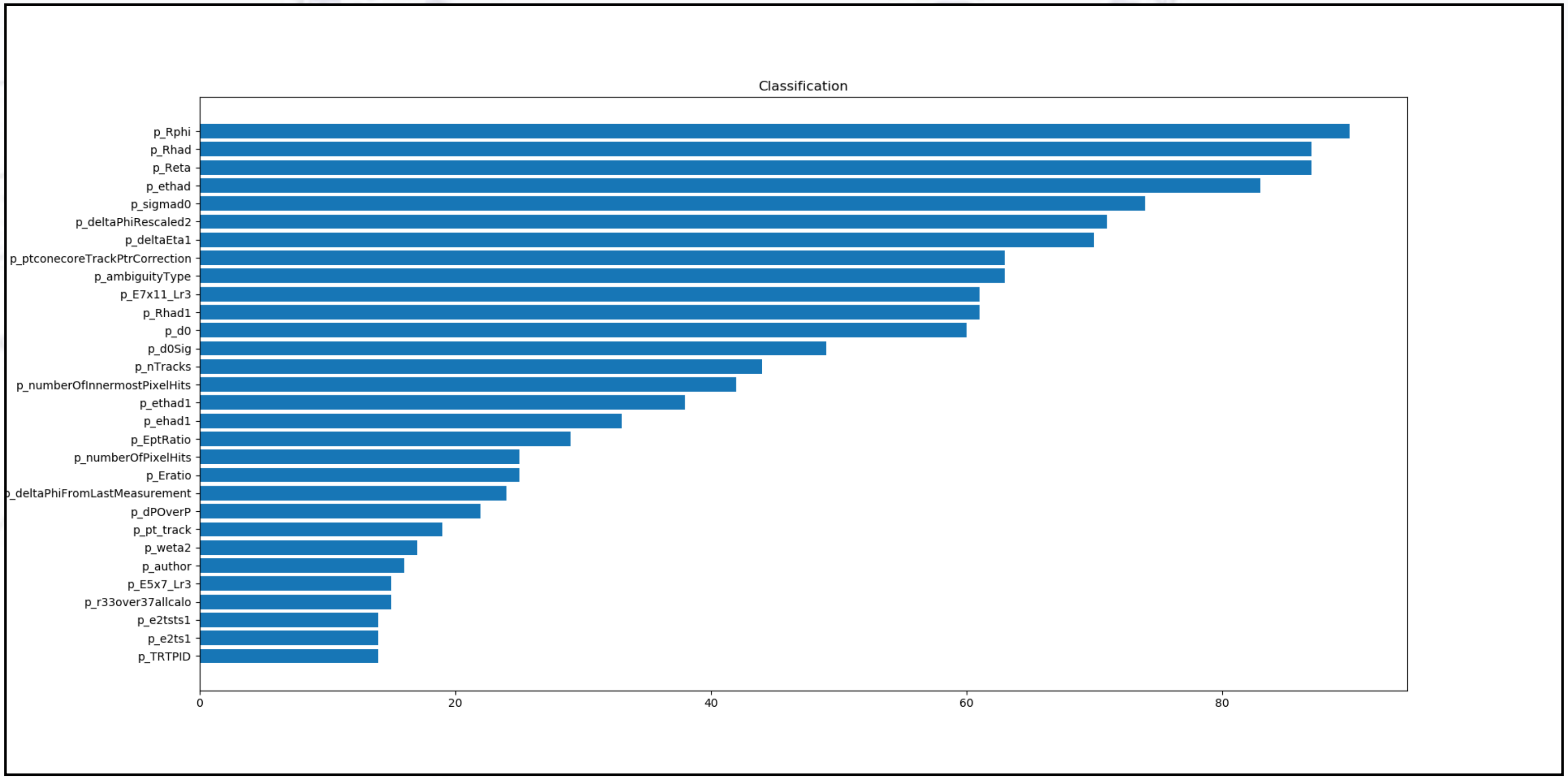
Thus your total number of points was:

Your variable choice

Assuming, that the variable frequency reflected the actual ranking very well, your variable choice was scored as follows:

$$8 \times \left(\sum Freq(\text{Your variables}) / \sum Freq(\text{Top variables}) \right)$$

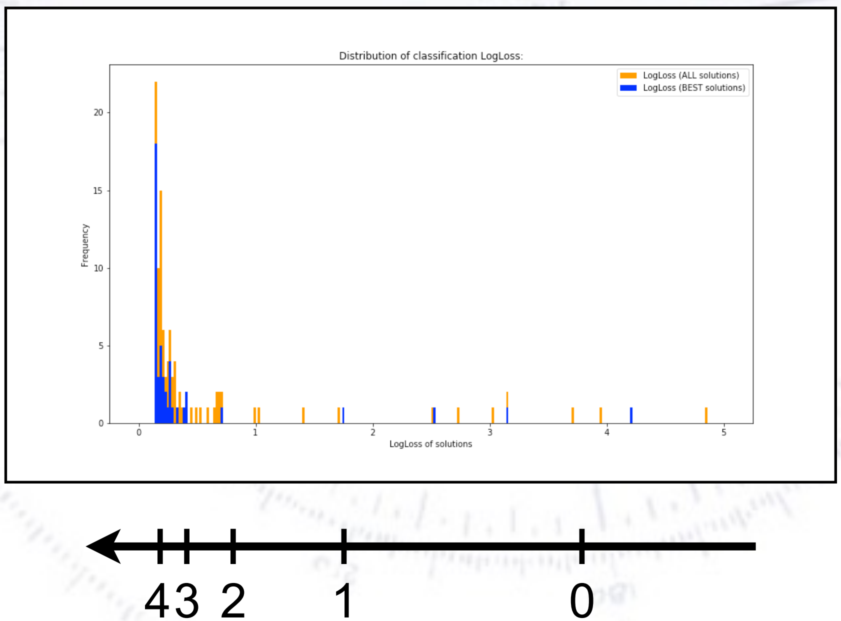
…so if you picked the top variables, you would get full points.

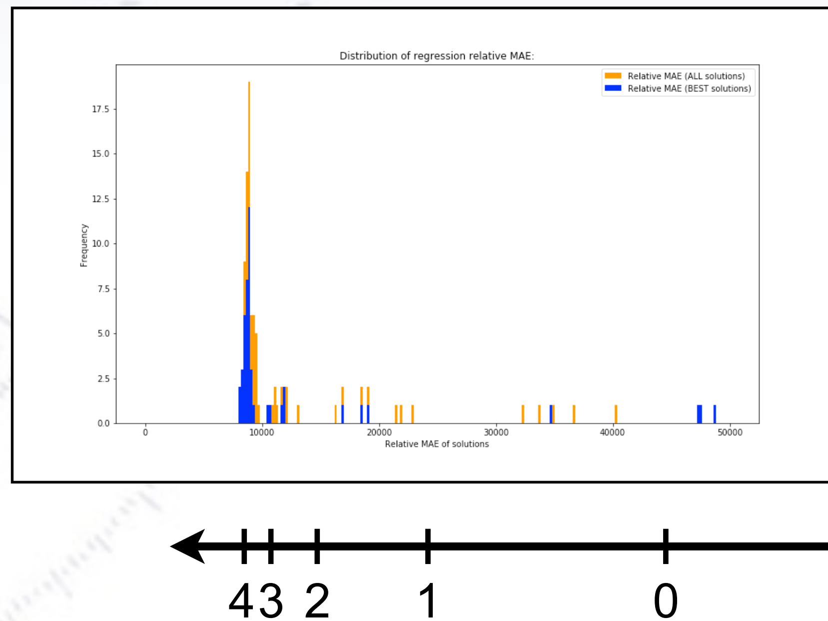


Performance scoring

As mentioned, performance isn't everything, and we certainly didn't want it to be for the small project. Getting close to the information limit is just great.

This was reflected by using a logarithmic scoring, which turned your best key performance parameter into a score in the (open) range [0,5+]:

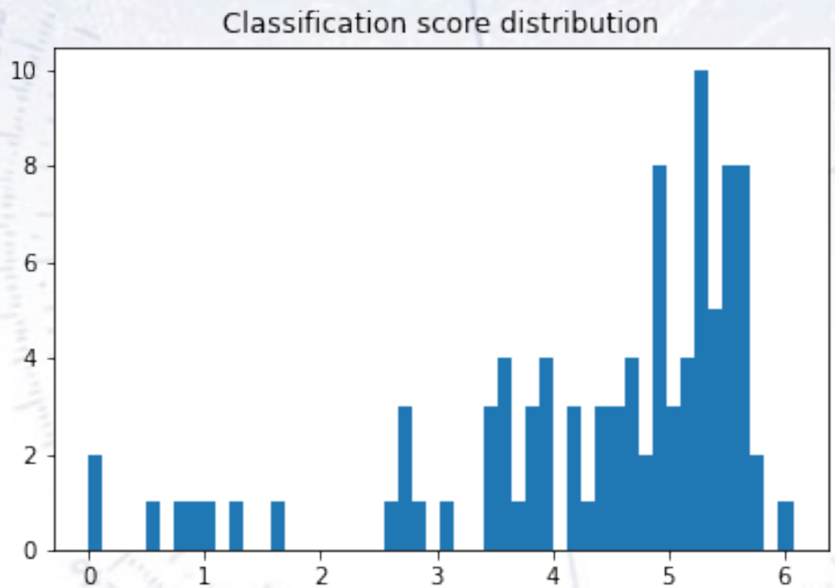


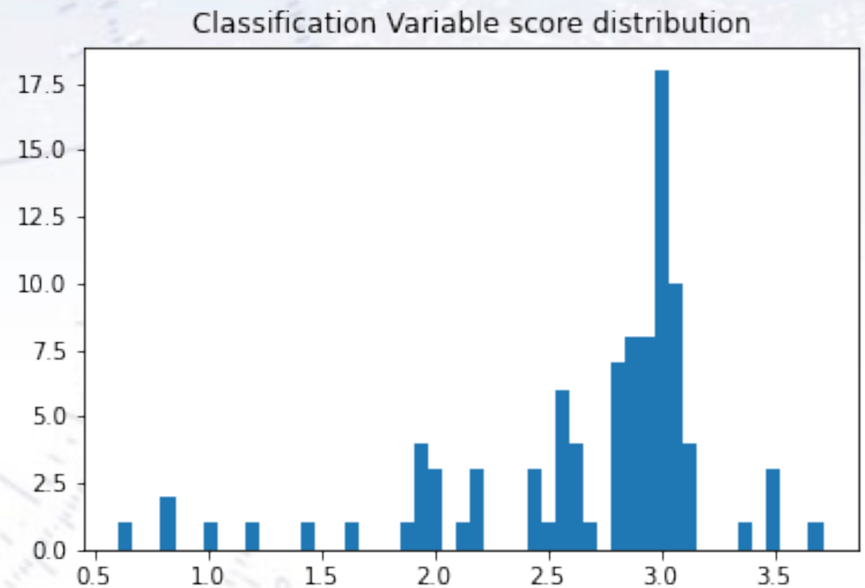


In all of this, you could of course not get negative points for an accepted solution!

The resulting score distributions

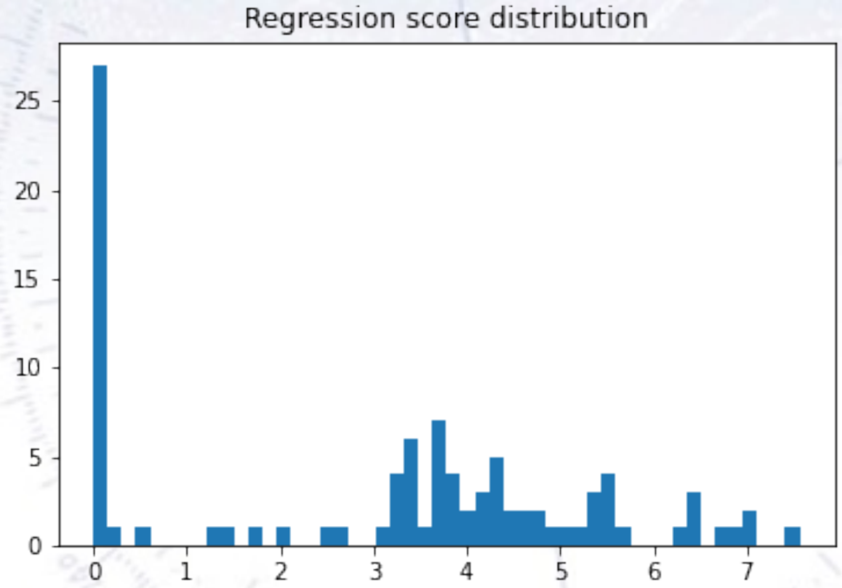
Score distributions for **classification** performance and variable choice:

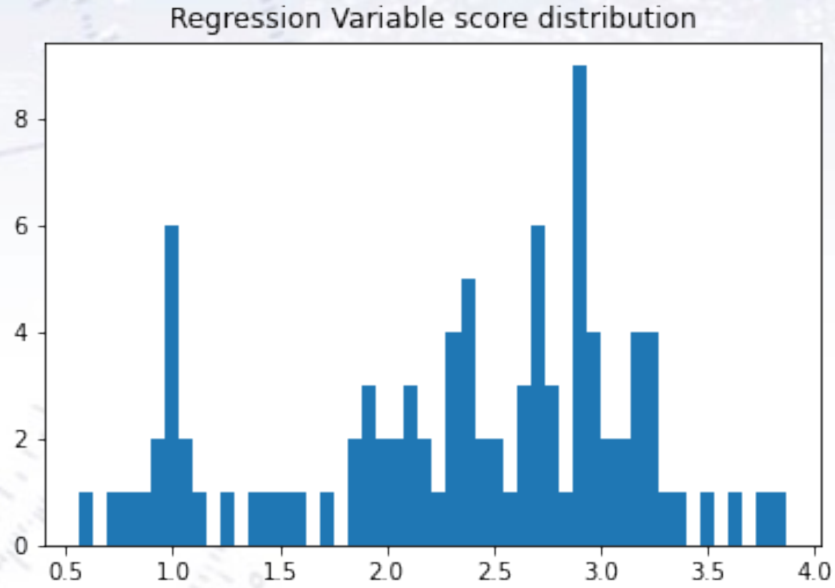




The resulting score distributions

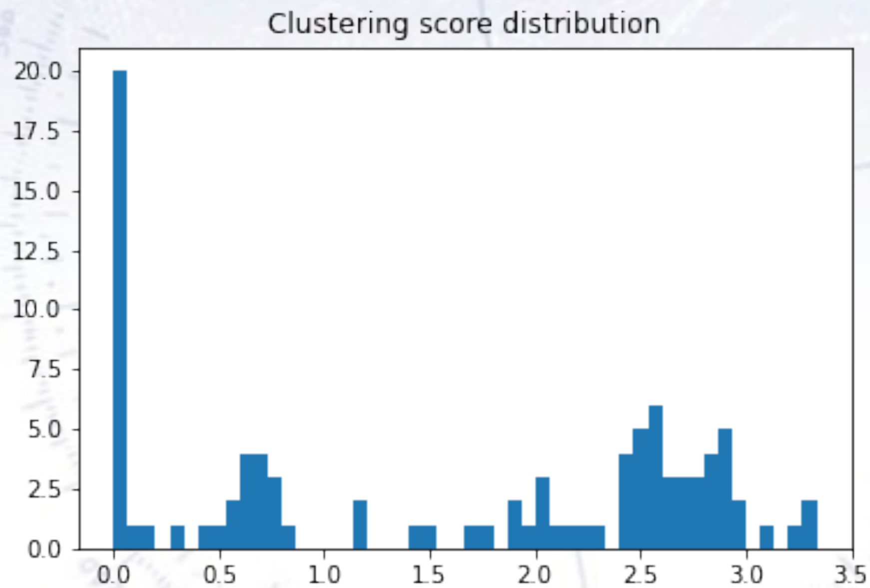
Score distributions for **regression** performance and variable choice:

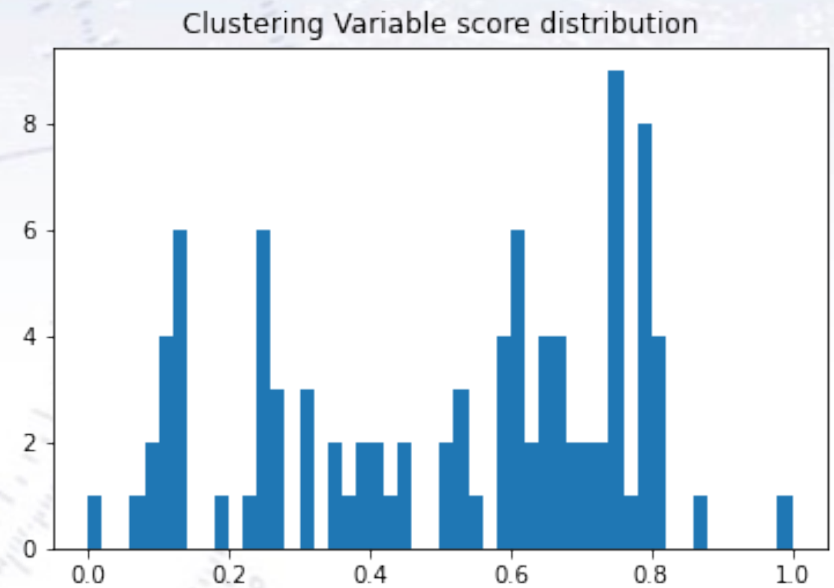




The resulting score distributions

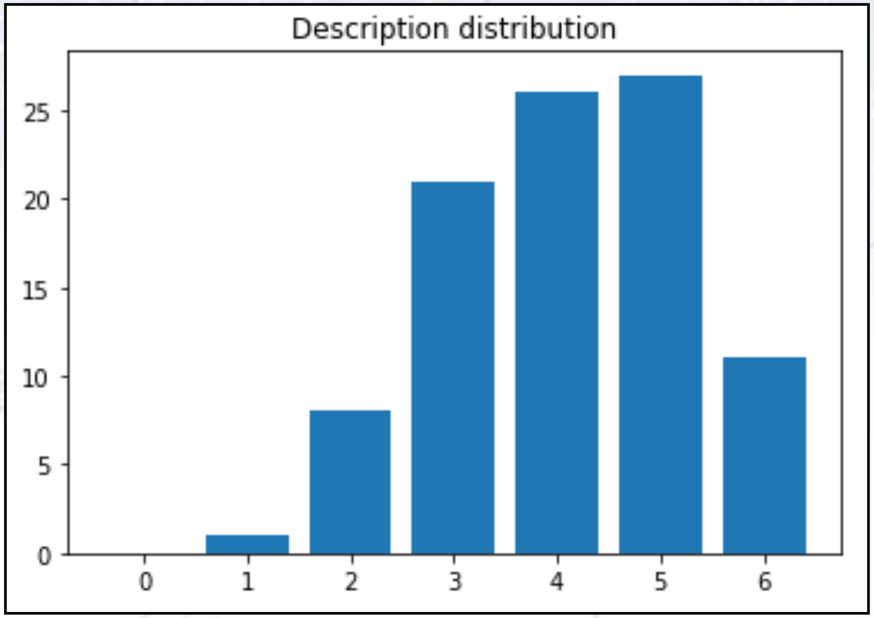
Score distributions for **clustering** performance and variable choice:

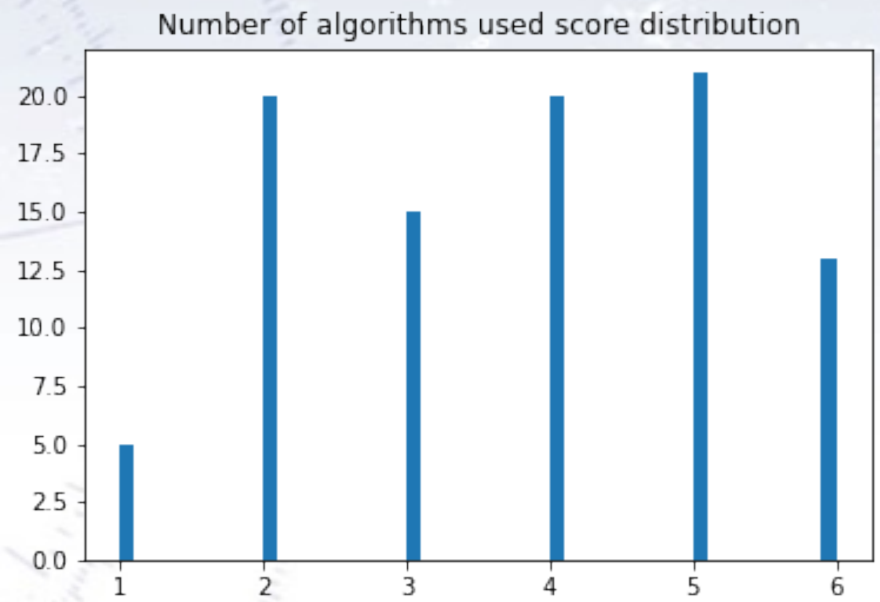




The resulting score distributions

The scores for descriptions and number of different algorithms (that work!) are:





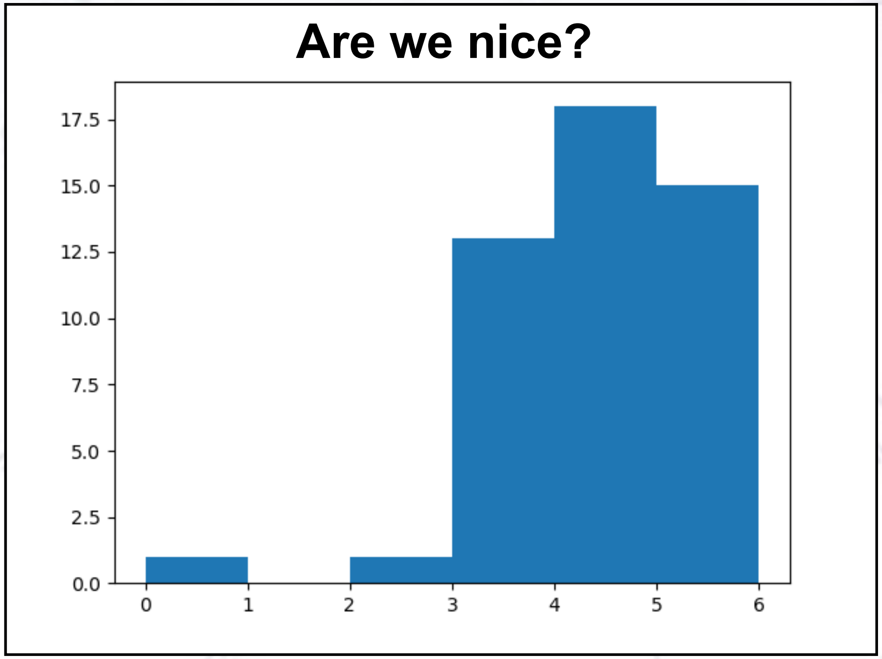
I read one of the "lower scoring" descriptions, but must say that I found it reasonably good, so in general the level was high.

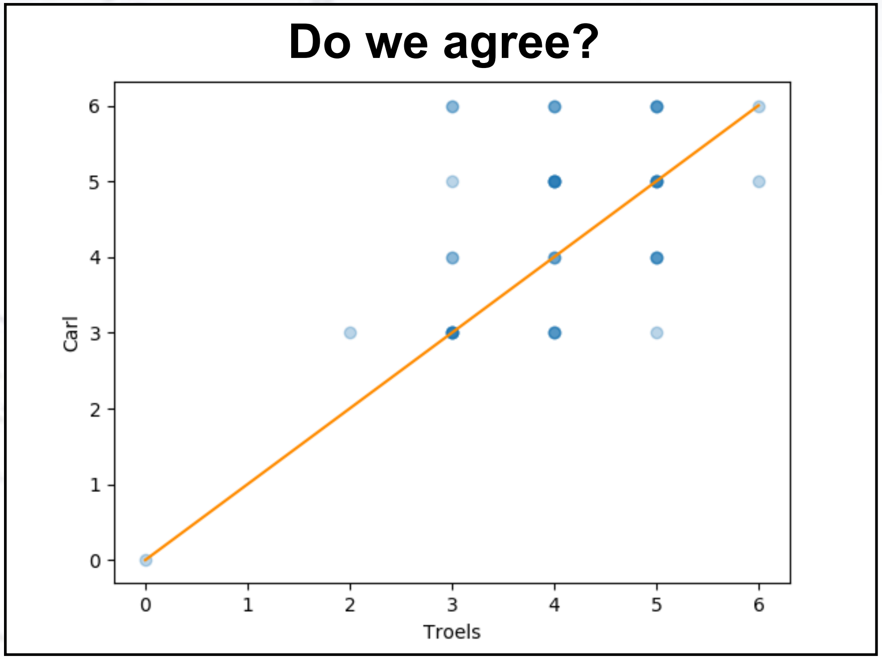
On algorithms, it was great to see that you both stuck with what you knew, but also explored new algorithms and got them working.

Your description reports

We read through your descriptions, and did a manual scoring (the only) based on choice of algorithms, hyperparameter optimisation, and data division (e.g. cross validation). Each yielded a score of 0-2, giving a total score of 0-6 points.

Numbers from last year (where Carl and I did it):





As you can see, we were generally satisfied. The descriptions were short and to the point, and give some insight into your line of thinking and working.

Reporting back to you

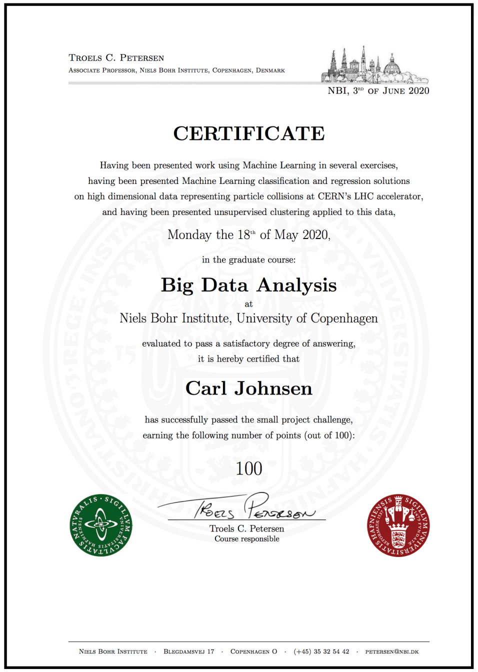
Feedback to you

We have created a small report back to you, which consists of:

- A certificate - for you to be proud of handing in…
- A summary - for you to know how you did…
- A solution scoring with key numbers and illustrations - for you to understand how your model performed.

These are (hopefully) being mailed to you by all of us right now. Please sit down after class and look through them.

Also, don't hesitate to discuss them with your peers. Perhaps you have already done this (great), **but this feedback and reflection is the process through which you learn the most**… please use it.



Troels C. Petersen
Associate Professor, Niels Bohr Institute, Copenhagen, Denmark

NBI, 3rd of June 2020

CERTIFICATE

Having been presented work using Machine Learning in several exercises,
having been presented Machine Learning classification and regression solutions
on high dimensional data representing particle collisions at CERN's LHC accelerator,
and having been presented unsupervised clustering applied to this data,

Monday the 18th of May 2020,

in the graduate course:

Big Data Analysis

at

Niels Bohr Institute, University of Copenhagen

evaluated to pass a satisfactory degree of answering,
it is hereby certified that

Carl Johnsen

has successfully passed the small project challenge,
earning the following number of points (out of 100):

100

Troels C. Petersen
Course responsible

Niels Bohr Institute · Blegdamsvej 17 · Copenhagen O · (+45) 35 32 54 42 · petersen@nbi.dk

Classification report

By now you should know what all the different plots and number are…

The solution gave the following metrics:

Metric	Equation	Value
Accuracy	`sklearn.metrics.accuracy_score`	0.940735
AUC	`sklearn.metrics.auc`	0.976952
Cross entropy	`sklearn.metrics.log_loss`	0.153488

The solution produced the following plots:

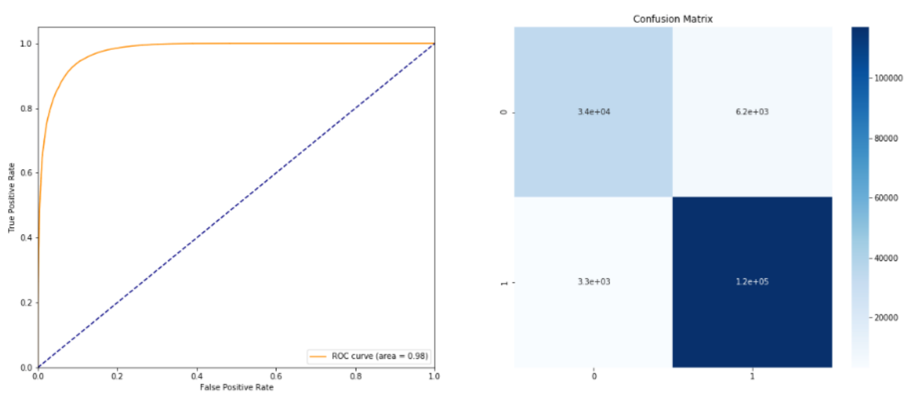


Figure 1: **Left:** ROC curve for the tensorflow2 implementation. The orange curve should be as close to the upper left corner as possible. **Right:** Confusion matrix for the tensorflow2 implementation. The diagonal squares ((0,0) and (1,1)) should have the higher values.

Regression report

The solution gave the following metrics:

Metric	Equation	Value
MAE - Absolute	`sklearn.metrics.mean_absolute_error`	6953.2194
MAE - Relative	$\sum \lvert \frac{y_p - y_t}{y_t} \rvert$	9060.6884
RMS	$\sqrt{mean((y_p - y_t)^2)}$	14261.8800
RMS 98th percentile	$\sqrt{mean((y_p - y_t)^2)}$	9238.8301
RMS 90th percentile	$\sqrt{mean((y_p - y_t)^2)}$	6074.5612
RMS 70th percentile	$\sqrt{mean((y_p - y_t)^2)}$	4586.3129

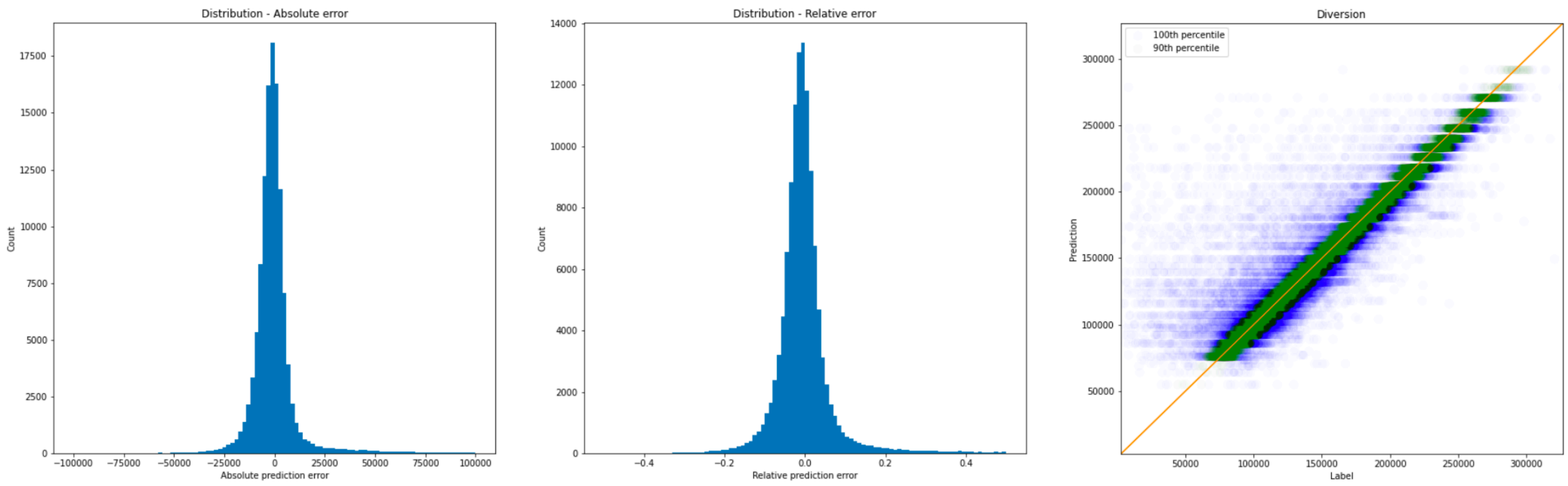


Figure 2: **Upper:** Distribution plots for the xgboost1 implementation. The plots are for absolute error (*Left*) and relative error (*Right*.). Both plots should have a tall narrow curve, centered around 0. **Lower:** Diversion plot for the xgboost1 implementation. The dots should be scattered close to the line - especially for the 90th percentile.

Clustering report

The clustering report is necessarily not very detailed, as unsupervised learning carries a great deal of uncertainty on what you're doing.

However, remember the remark by Alexander Nielsen about t-SNE & UMAP, but applied more generally:
"I always start by throwing a clustering algorithm at data, just to see what structures turn up, if any.
Even the latter result tells me something valuable for the further analysis."

clustering - KMeans

The solution produced the following metrics:

Metric	Equation	Value
Accuracy	`sklearn.metrics.accuracy_score`	0.7492

To compute the accuracy, the following mapping was used, based on the clusters resemblance to electron classification:

Cluster	0	1	2	3
is electron	1	1	1	1

The solution provided the following plot:

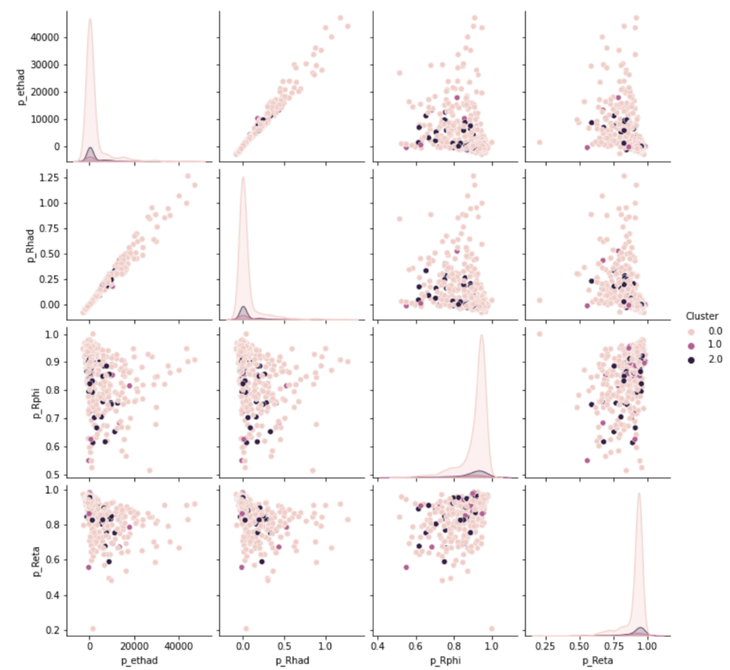


Figure 6: Pairplot for the KMeans implementation. The variables chosen are the top 4 most used variables for clustering. There should be a clear distinction of the clusters.

Thank you, for all your hard work

Thank you, for all your hard work