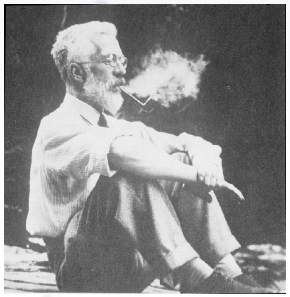
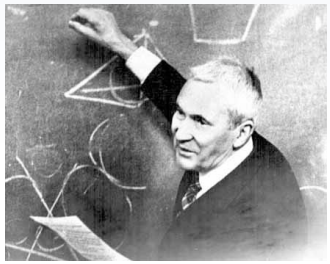
Applied ML

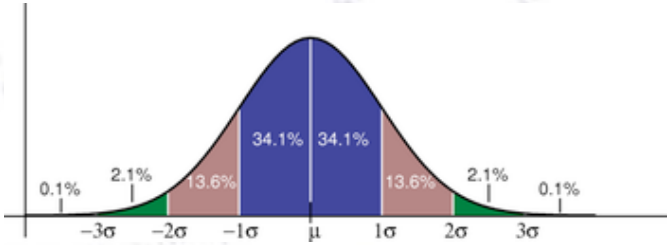
Results and Scores of Initial Project





Janni & Troels (NBI)



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

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 many ML algorithms from start to end, 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.

We also wanted you to feel the "insecurity" about not knowing if you had gotten everything out of the data. Especially with the **clustering**.

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.

Overall comments

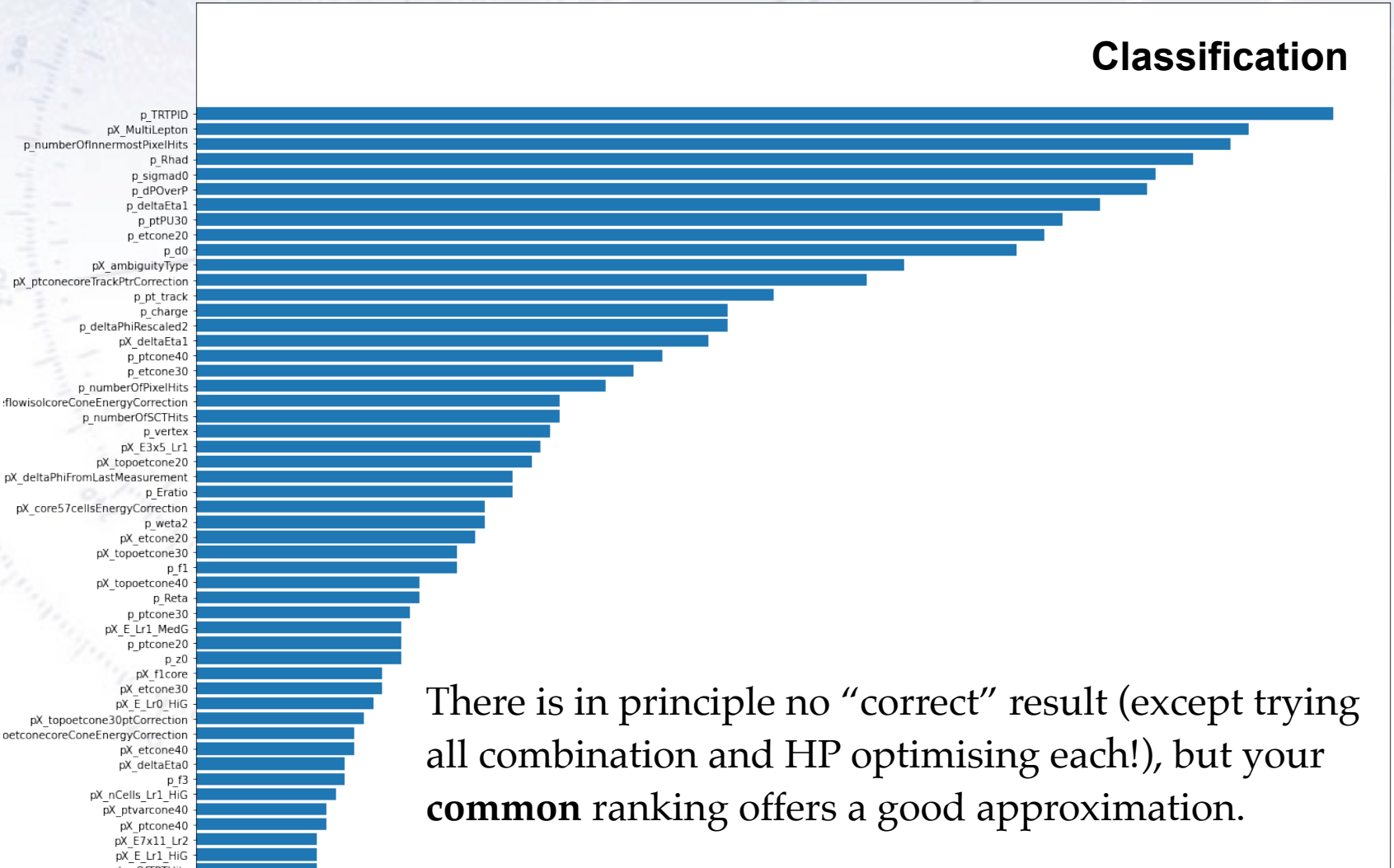
You generally did very well, and so let me start by gently stating, that you have nothing to fear - in fact, you did really great!

Grading it was perhaps comparable to 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.

Classification Results

Classification variable usage

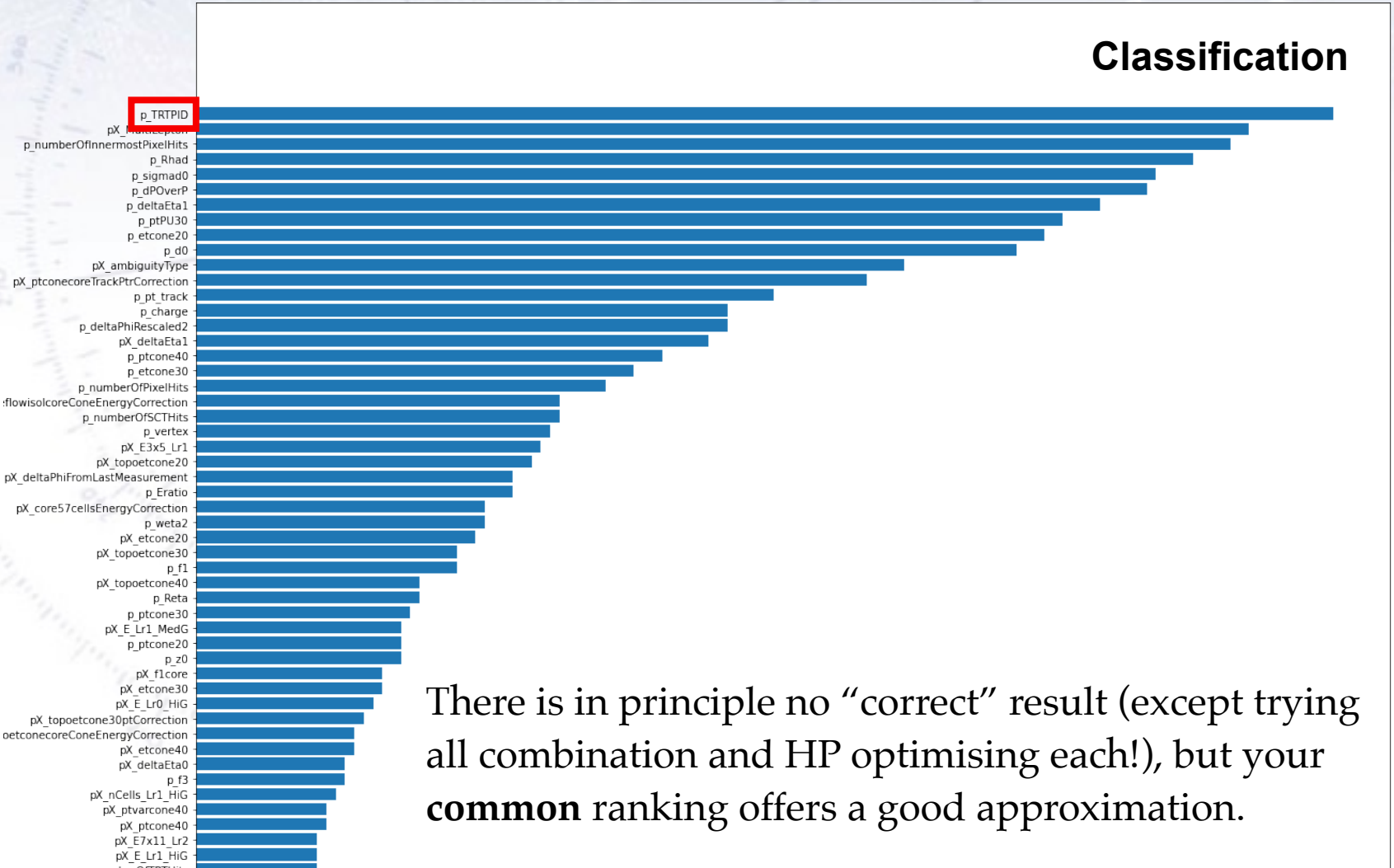
Many (most?) of you have made a good variable ranking. Below you find a variable usage **frequency** plot, showing how often a variable was used.



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

Classification variable usage

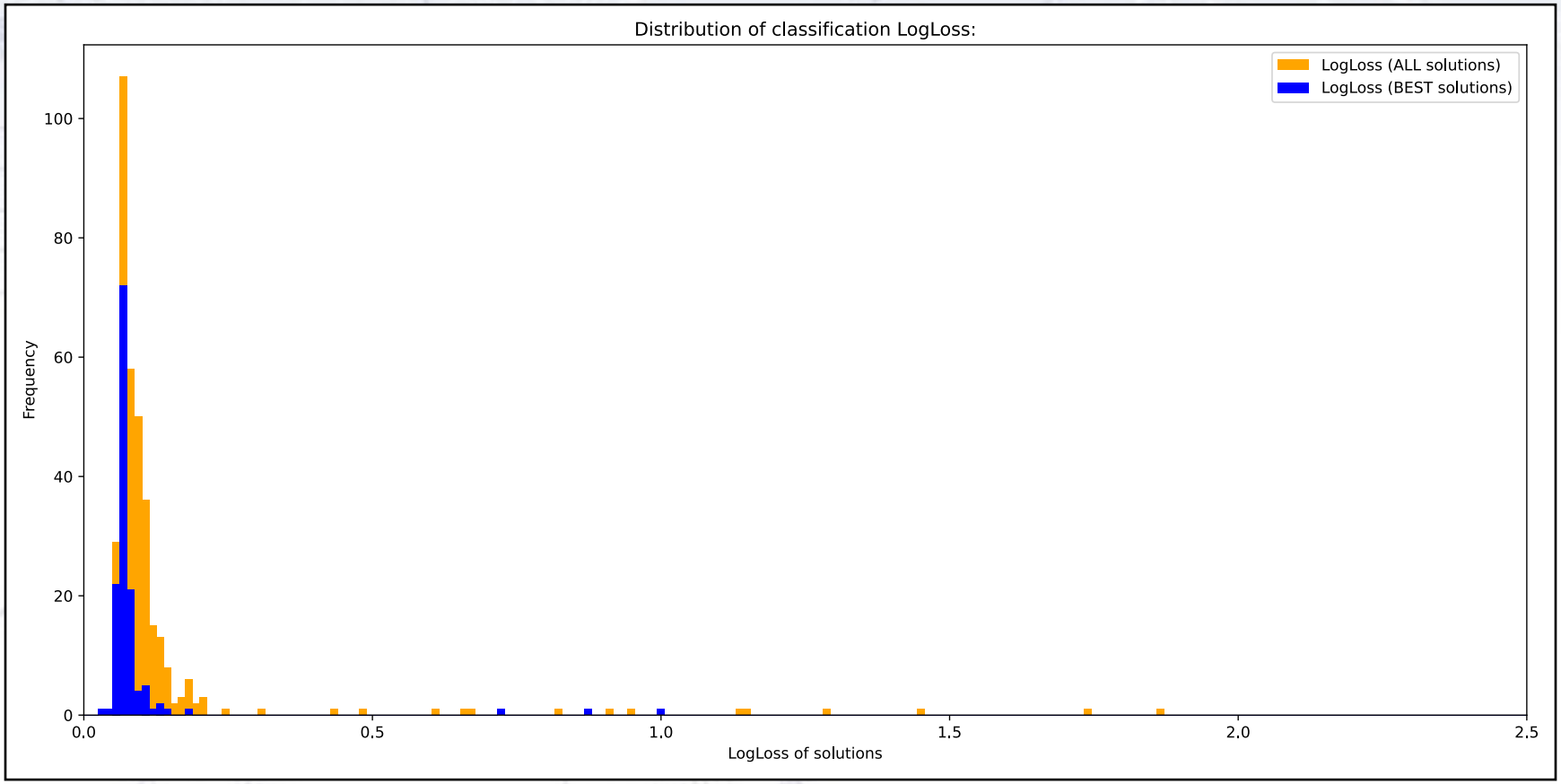
Many (most?) of you have made a good variable ranking. Below you find a variable usage **frequency** plot, showing how often a variable was used.



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

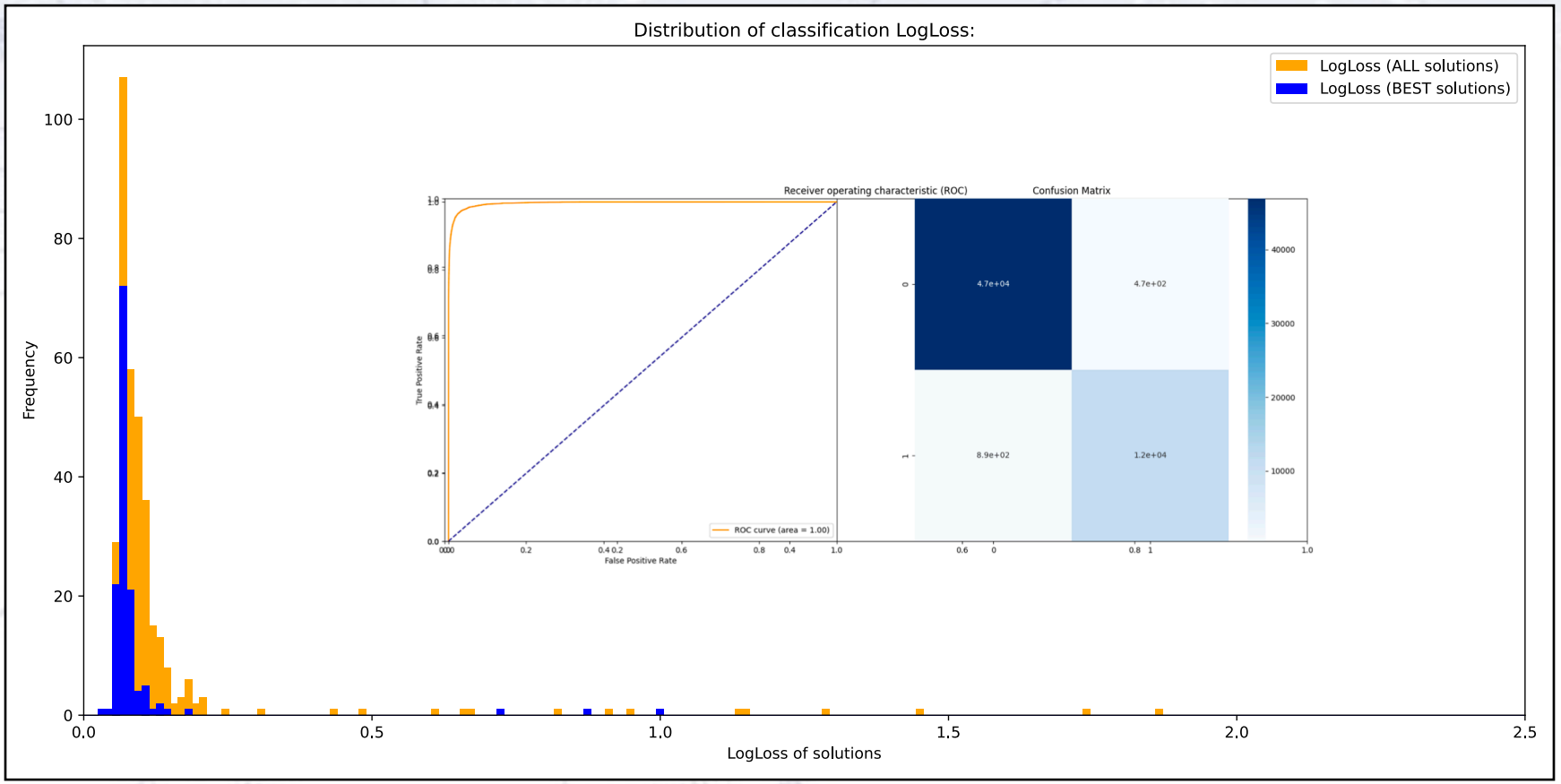
Classification score distribution

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



Classification score distribution

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



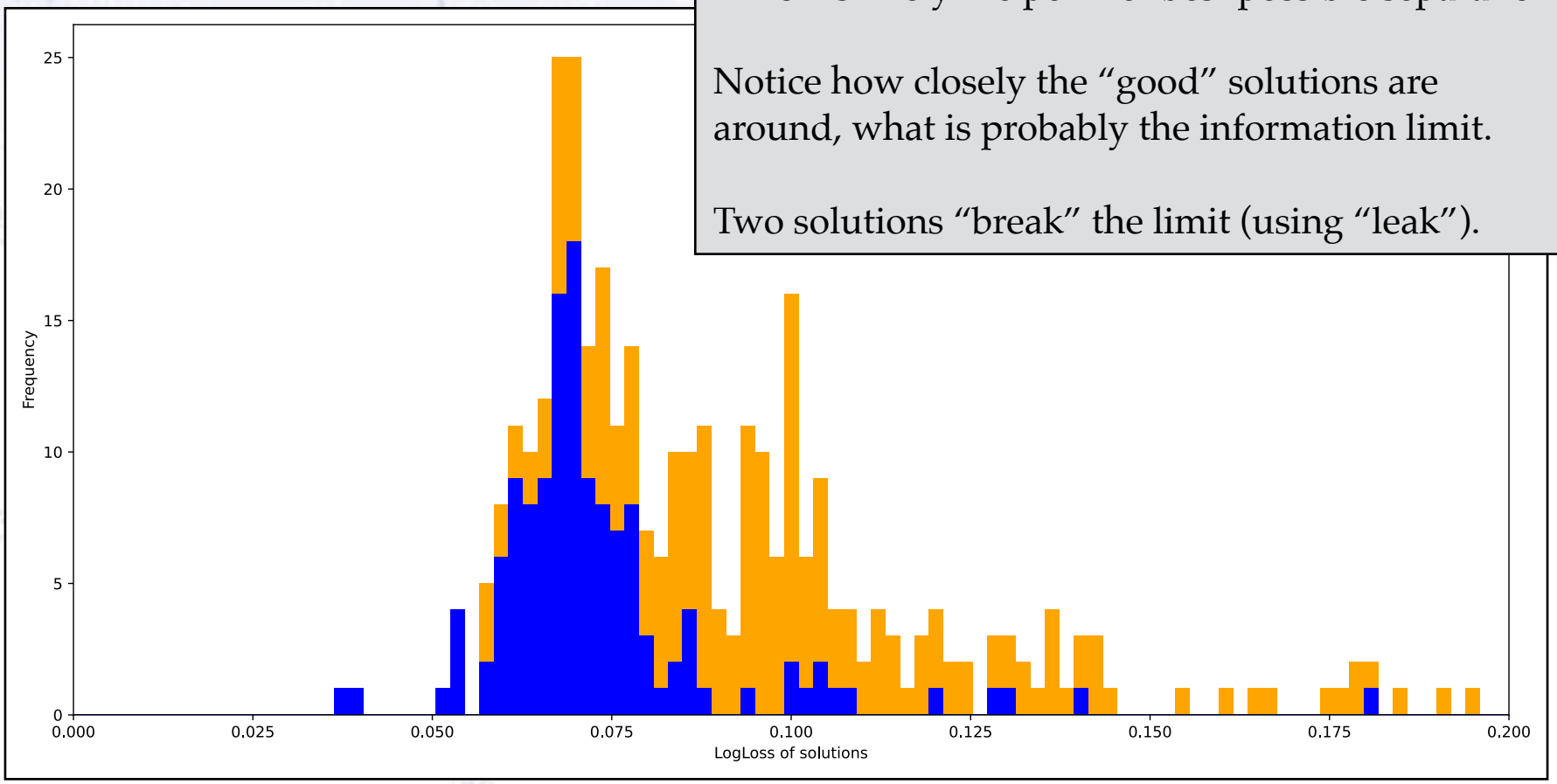
Classification score distribution

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.

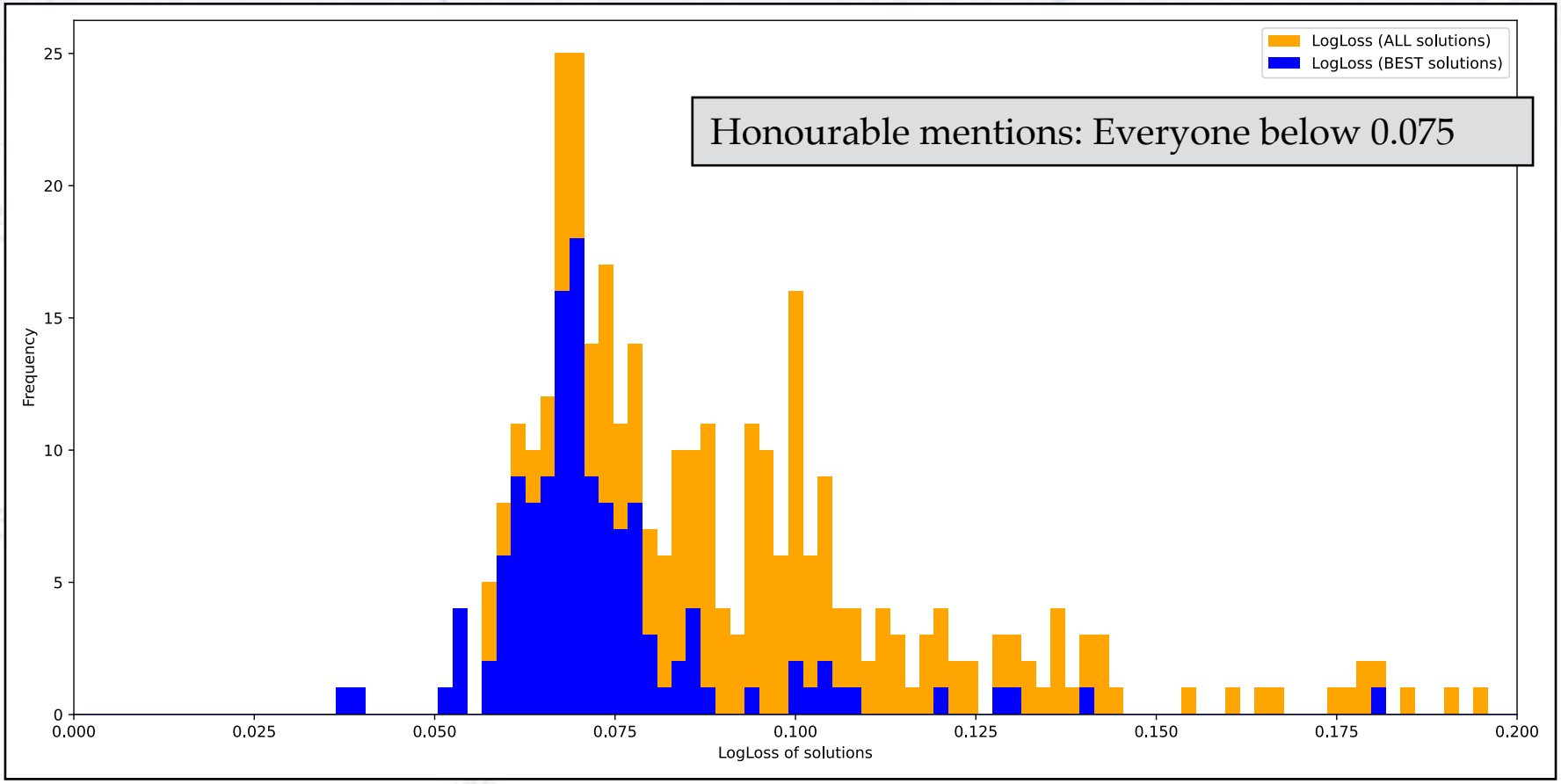
Notice how closely the "good" solutions are around, what is probably the information limit.

Two solutions "break" the limit (using "leak").



Classification score distribution

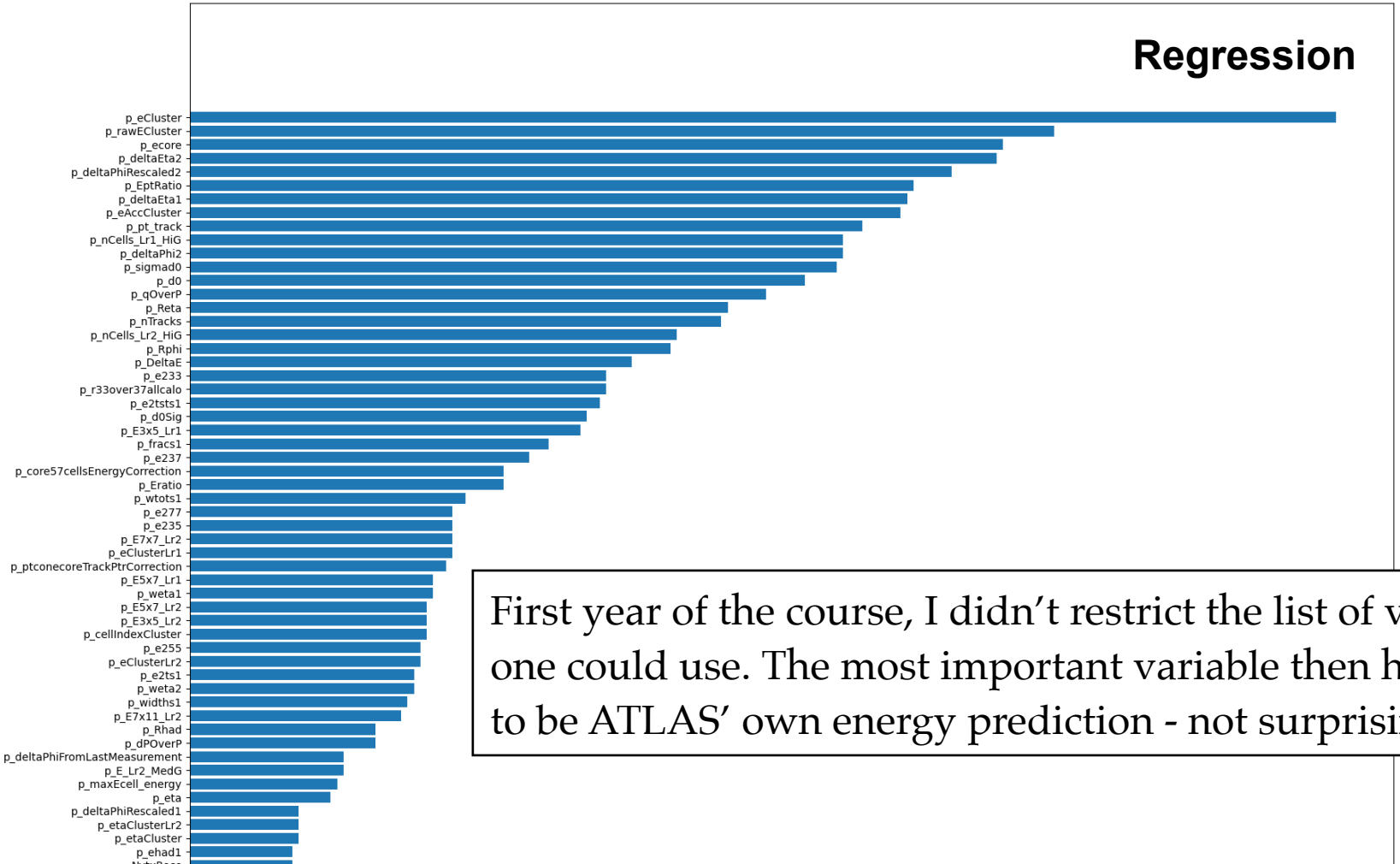
The distribution of the (Cross-Entropy) LogLoss values obtained was:



Regression Results

Regression variable usage

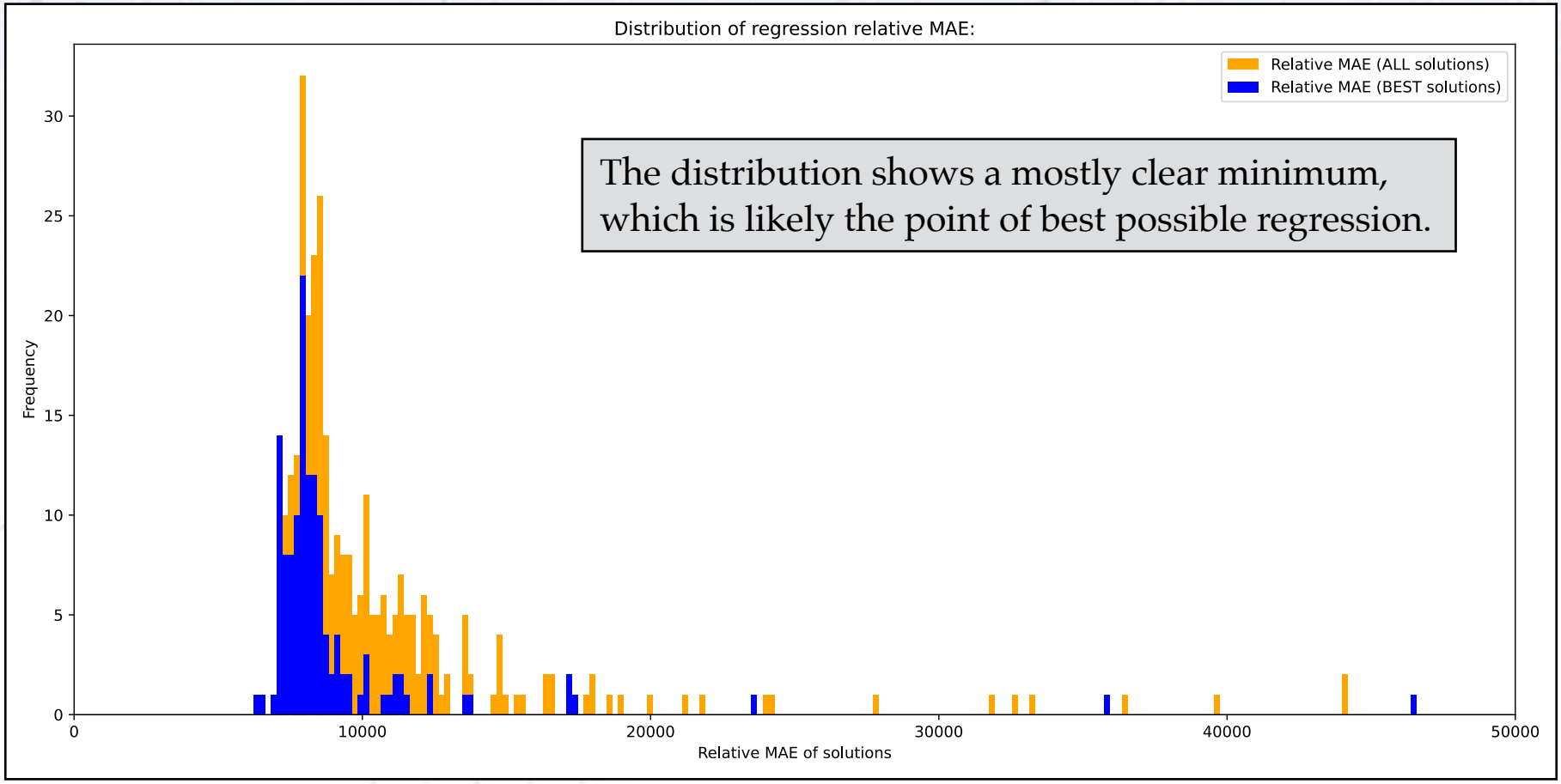
The variables have changed drastically from the classification case. There is NO overlap at all for the top 10-15 variables! Classification and Regression are in this case two very different tasks.



First year of the course, 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!

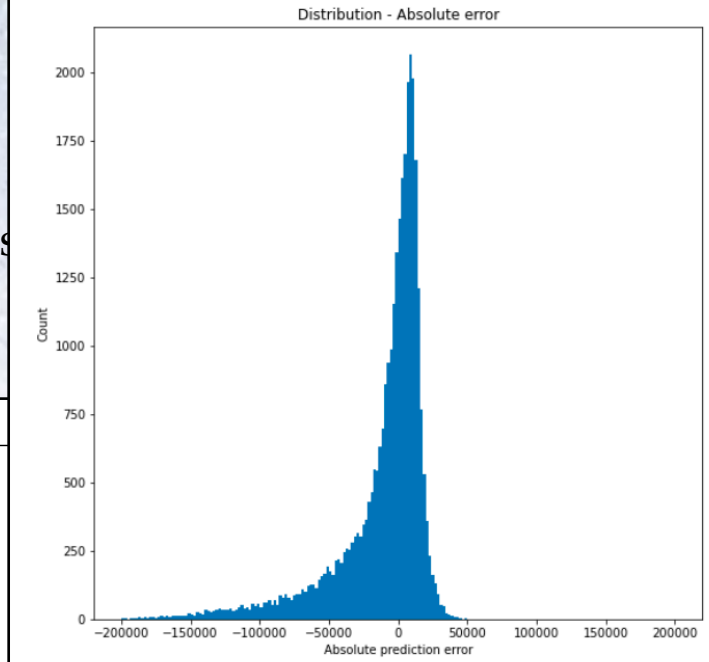
Regression score distribution

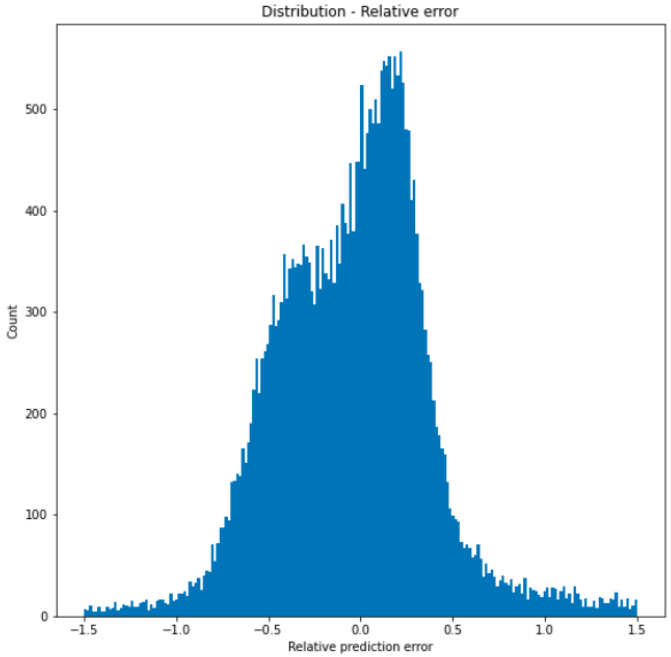
The distribution of the relative MAE (i.e. MAE((E-T)/T)) values obtained was:

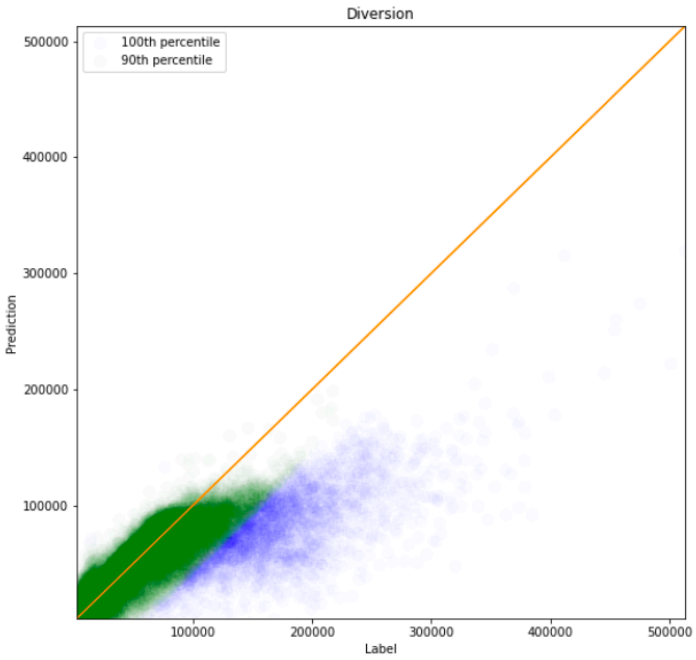


The dis

vas:



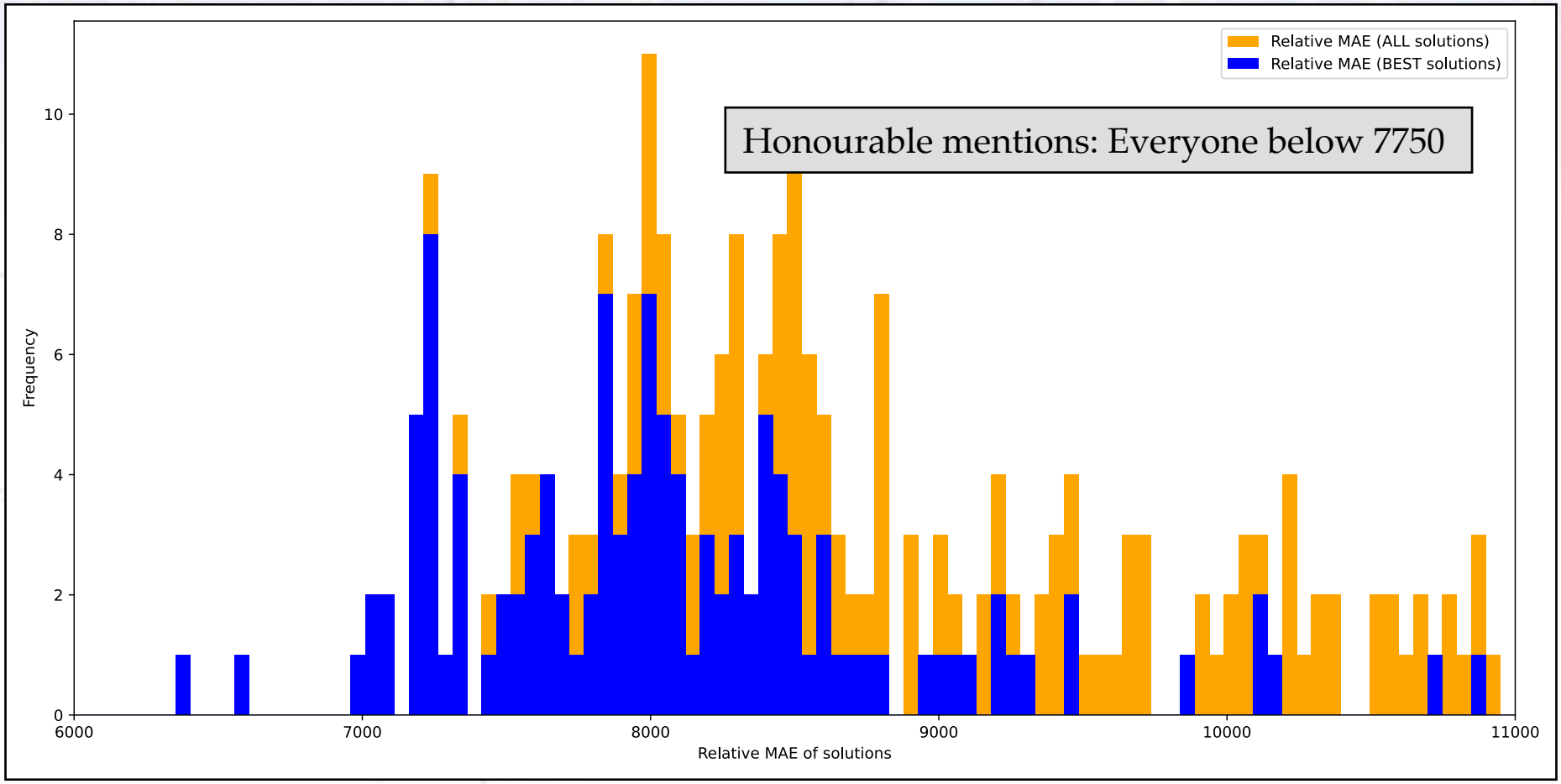




Looking at your own solution distribution might reveal issues (e.g. double bump).

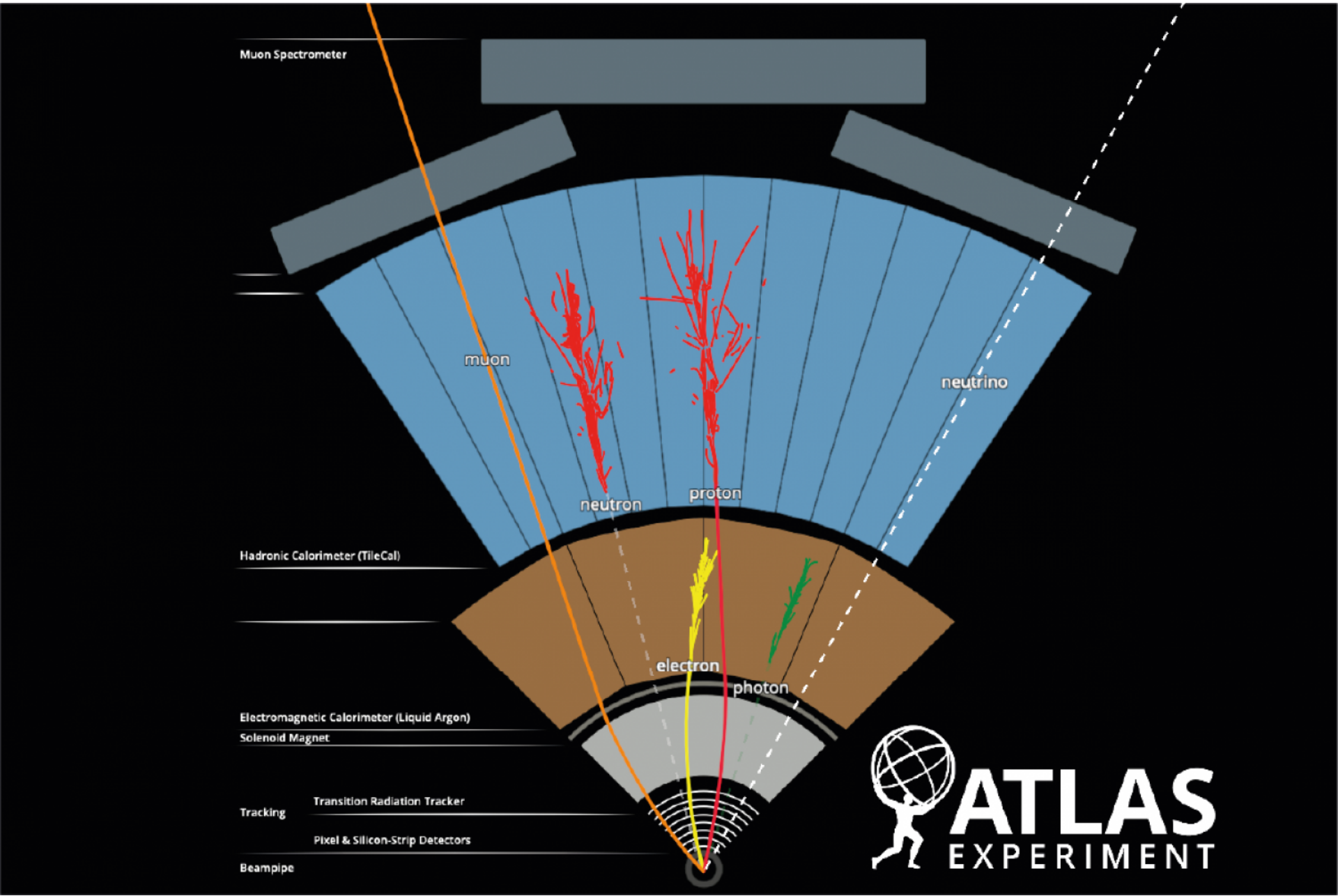
Regression score distribution

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

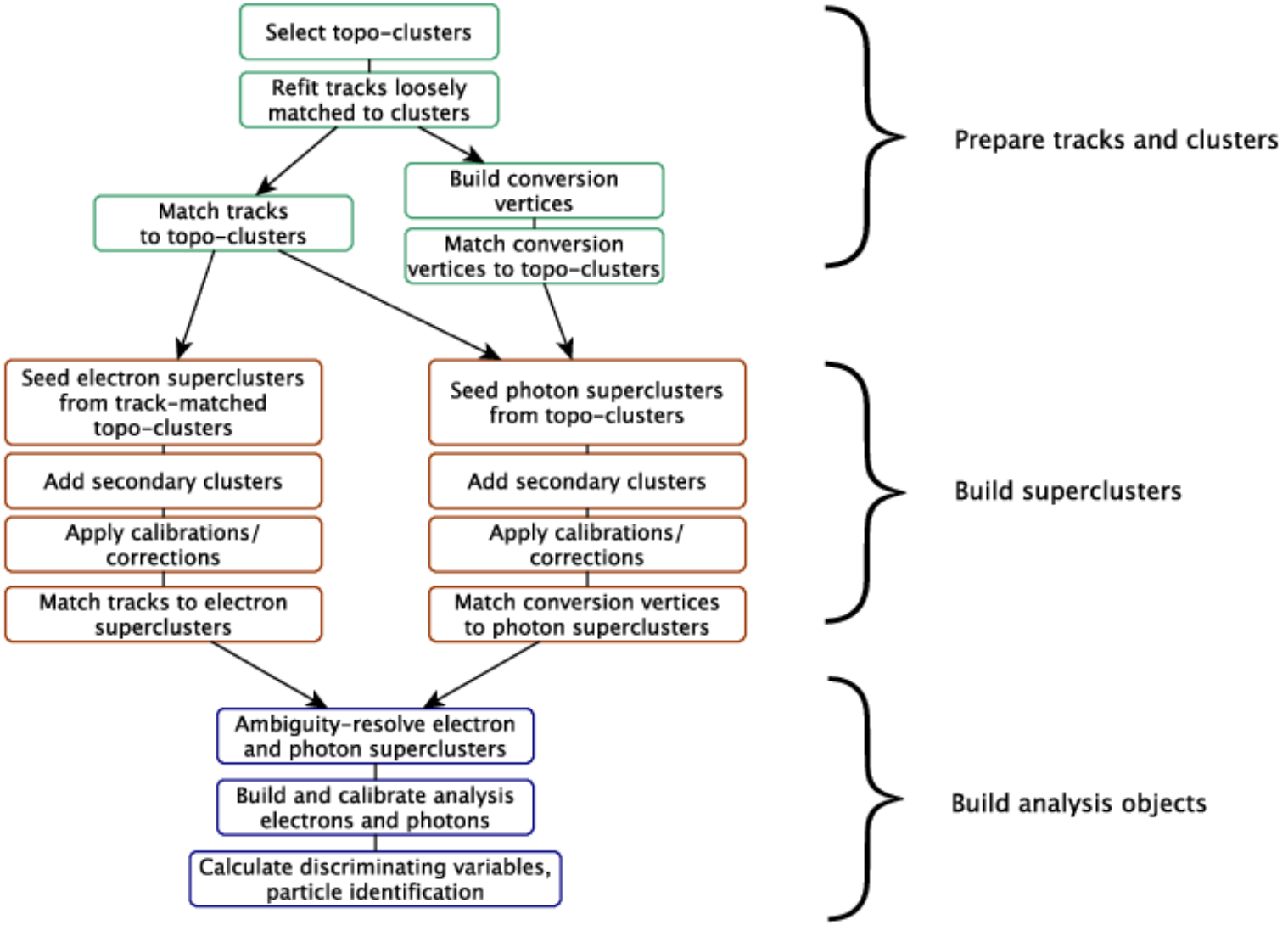


Notes about the data

Notes about the data



Notes about the data



New York Times - Front Page

Physicists Find Elusive Particle Seen as Key to Universe

POOL PHOTO BY DENIS BALIBOUSE

Scientists in Geneva on Wednesday applauded the discovery of a subatomic particle that looks like the Higgs boson.

WORLD'S BEST DESIGNED NEWSPAPER

POLITIKEN

Torsdag

5. juli 2012
Årgang 128. Nr. 275
Pris 25,00
Kundecenter
Politiken 70 15 01 01
1. udgave
www.politiken.dk

'Guds partikel' | Her ser du en historisk milepæl

Før den økonomiske krise steg fertiliteten, men nu falder den i 25 ud af 31 vestlige lande, viser nye tal. Danmark er hårdt ramt.

FERTILITET

PETER G. H. MADSEN

Når den økonomiske krise rammer, og fyresedlerne flyver i virksomhederne, skrues der op for præventionen.

Spørg bare i Letland. Da den finansielle krise skyllede ind over den vestlige verden i efteråret 2008, gik landets økonomi i koma. Og kort efter holdt lettiske kvinder nærmest op med at føde børn, og fertiliteten – som er antallet af børn en kvinde får i sit liv – styrtdykkede fra 1,44 i 2008 til 1,14 sidste år. Et enormt fald i demografiens verden, hvor ændringer ofte måles i mikroskopiske decimaler.

Letland er langt fra unik. I mange af de lande, herunder Danmark, hvor den økonomiske krise har sat sit spor i ledighedstal og nationalregnskaber, er der blevet længere mellem de glade forældre på landets fødestuer. Det fortæller seniorforsker Tomas Sobotka fra Vienna Institute of Demography.

»Det er en klar trend. Efter krisen ramte, er fertiliteten faldet i Europa, USA og en række andre af verdens rigeste lande. Jeg mener, forklaringen er, at den økonomiske krise, og den usikkerhed og frygt for arbejdsløshed, der følger med, får mange unge til at vente med at stifte familie«, siger han.

Tomas Sobotkas data taler også deres eget tydelige sprog. I 2008 havde 30 ud af flere børn end året før. Mest markant er faldet i blandt andet Spanien, Grækenland, Estland, Ungarn og Island. Lande, der har det til fælles, at de er blevet ramt hårdt af den økonomiske krise.

Seniorforsker Mogens Christoffersen fra SFI – Det Nationale Forskningscenter for Velfærd – er heller i tvivl om, at den økonomiske usikkerhed har fået mange unge til at overveje deres fremtidsplaner.

»Hvis du har svært ved at finde et job og et sted at bo, så er du også mere tilbageholdende med at stifte familie«, siger han.

Mogens Christoffersen fortæller, at samme udvikling kunne ses i 1930'erne, hvor datidens unge ventede med at få børn på grund af den økonomiske og politiske usikkerhed. Til gengæld kom der så et boom i fødselstallene efter Anden Verdenskrig.

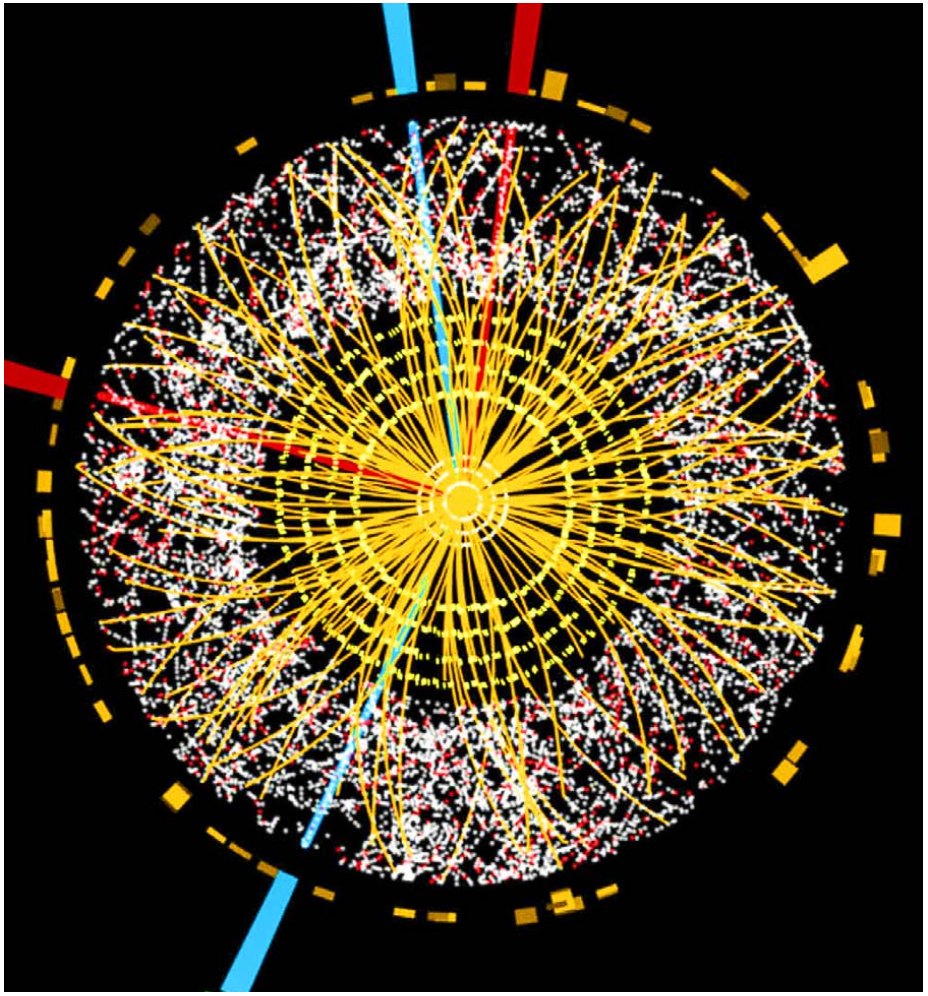
Om vi kommer til at se den udvikling gentaget, når og hvis økonomien vender i Europa, er svært at sige. Men der er en risiko for, at de kvinder, der i dag venter med at få børn, bliver fanget af det biologiske ur og aldrig kommer til at ligge i barselssengen, påpeger Mogens Christoffersen.

Også Danmark er ramt af fødselsrecessionen. Fra 2008 til 2011 faldt fertiliteten herhjemme fra 1,89 til 1,76.

Ifølge professor Jørgen Goul Andersen fra Aalborg Universitet handler den faldende fertilitet i Danmark dog mere om den politik, der føres på Christiansborg, end om den økonomiske udvikling. Resultatet er dog det samme. Nemlig at Danmark, såfremt udviklingen fortsætter i 2012, vil se et langtidsfald i arbejdsudbuddet på 10.000-12.000. Hvilket giver færre i den erhvervsaktive alder til at forsørge det stigende antal ældre.

»De danske tal for 2011 er skræmmende. Hvis udviklingen fortsætter, vil det få meget stor betydning for den fremtidige arbejdsstyrke«, siger han.

peter.g.madsen@pol.dk



I vil op for musikken, når den ikoniske Orange Scene åbner. Roskilde Festivalen har for længst slået sin status som en kulturbærende søjle fast. Hvert år introduceres nye generationer af unge for et bredt spektrum af den ypperste kvalitet i rytmisk musik netop nu og opdrages til at forstå, at det musikalske univers er langt dybere end tyggegummipop og pladeselskabernes nyeste teenagefænomen.

Det er heldigvis stadig musikken, der er i centrum. Men Roskilde Festival er kulturbærende i langt videre forstand end rent musikalsk. I en tid, hvor det politiske parnas er enige om, at kun skattelettelser og økonomiske incitamenter kan motivere mennesker, er en festival båret af frivillige kræfter en vigtig erindring om, at alt ikke passer ind i matematiske modeller.

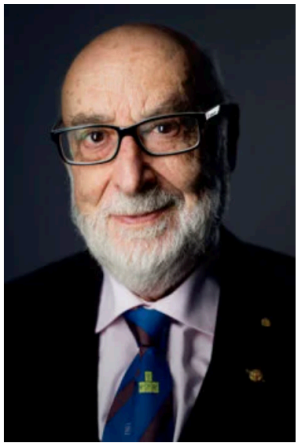
Det er også derfor, at Roskilde Festival virker grænseløst provokerende på avantgarden i tidens nyborgerlige bølge, der i foråret gik voldsomt til angreb mod Roskilde Kommunes ekspropriation af en nabogrund, som sikrer, at festivalen også i fremtiden har plads at boltre sig på. Magtmisbrug. Bestikkelse. Ingen anklage var for langt ude. Men kommunen blev pure frikendt af statsforvaltningen.

NÅR ANGREBET bliver så hårdt, er det fordi festivalen er en umulighed i det nyborgerlige verdensbillede. For de ved jo, at de bedste løsninger altid skabes af det frie marked. At al initiativ skabes i jagten på private profitter, og at effektivitet er umulig uden markedets usynlige hånd.

Tanken om, at tusindvis af frivillige arbejder for at stable festivalen på benene, passer ikke ind i cost-benefit-analyserne. Tanken om, at en nonprofitorganisation kan skabe et

What did we Classify and Regress?

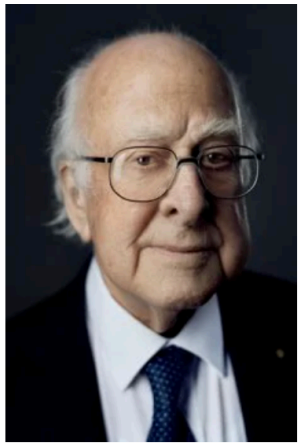
Nobel Prize in Physics 2013



© Nobel Media AB. Photo: A. Mahmoud

François Englert

Prize share: 1/2

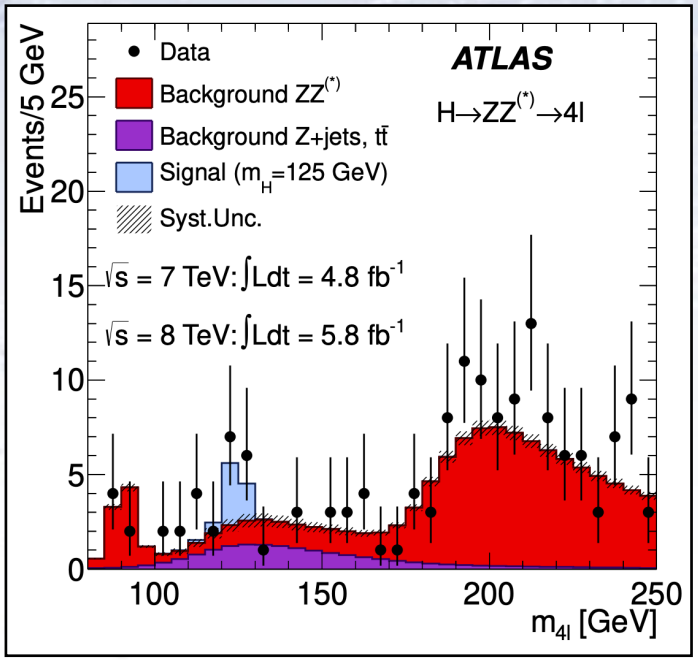


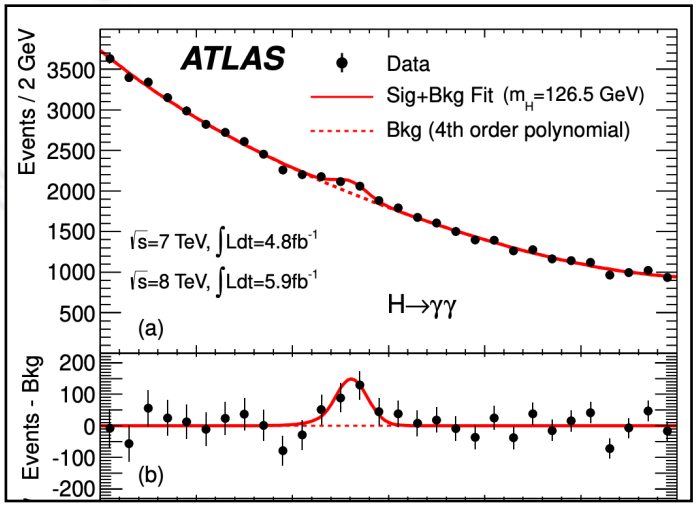
© Nobel Media AB. Photo: A. Mahmoud

Peter W. Higgs

Prize share: 1/2

The Nobel Prize in Physics 2013 was awarded jointly to François Englert and Peter W. Higgs "for the theoretical discovery of a mechanism that contributes to our understanding of the origin of mass of subatomic particles, and which recently was confirmed through the discovery of the predicted fundamental particle, by the ATLAS and CMS experiments at CERN's Large Hadron Collider"

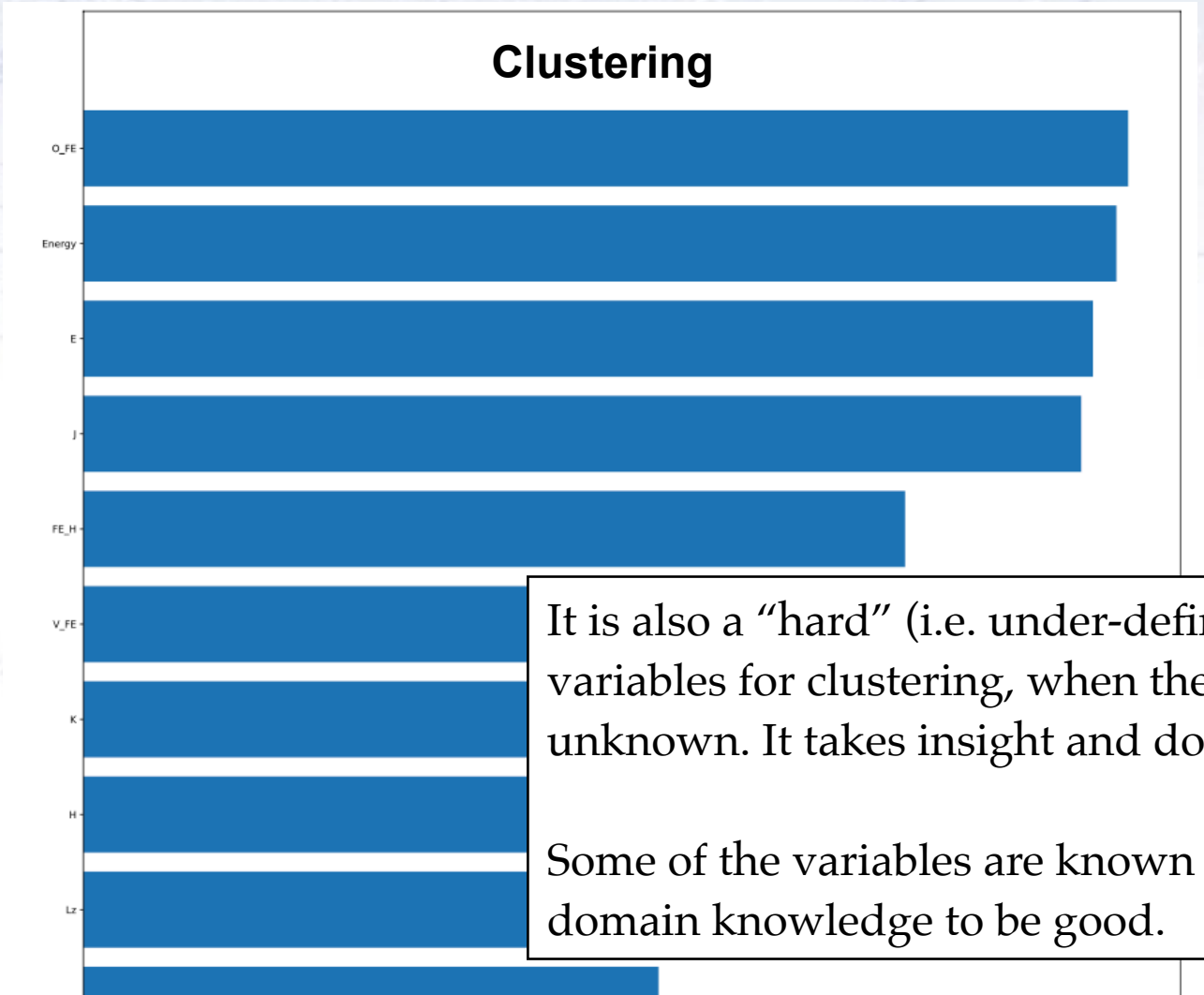




Clustering Results

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…

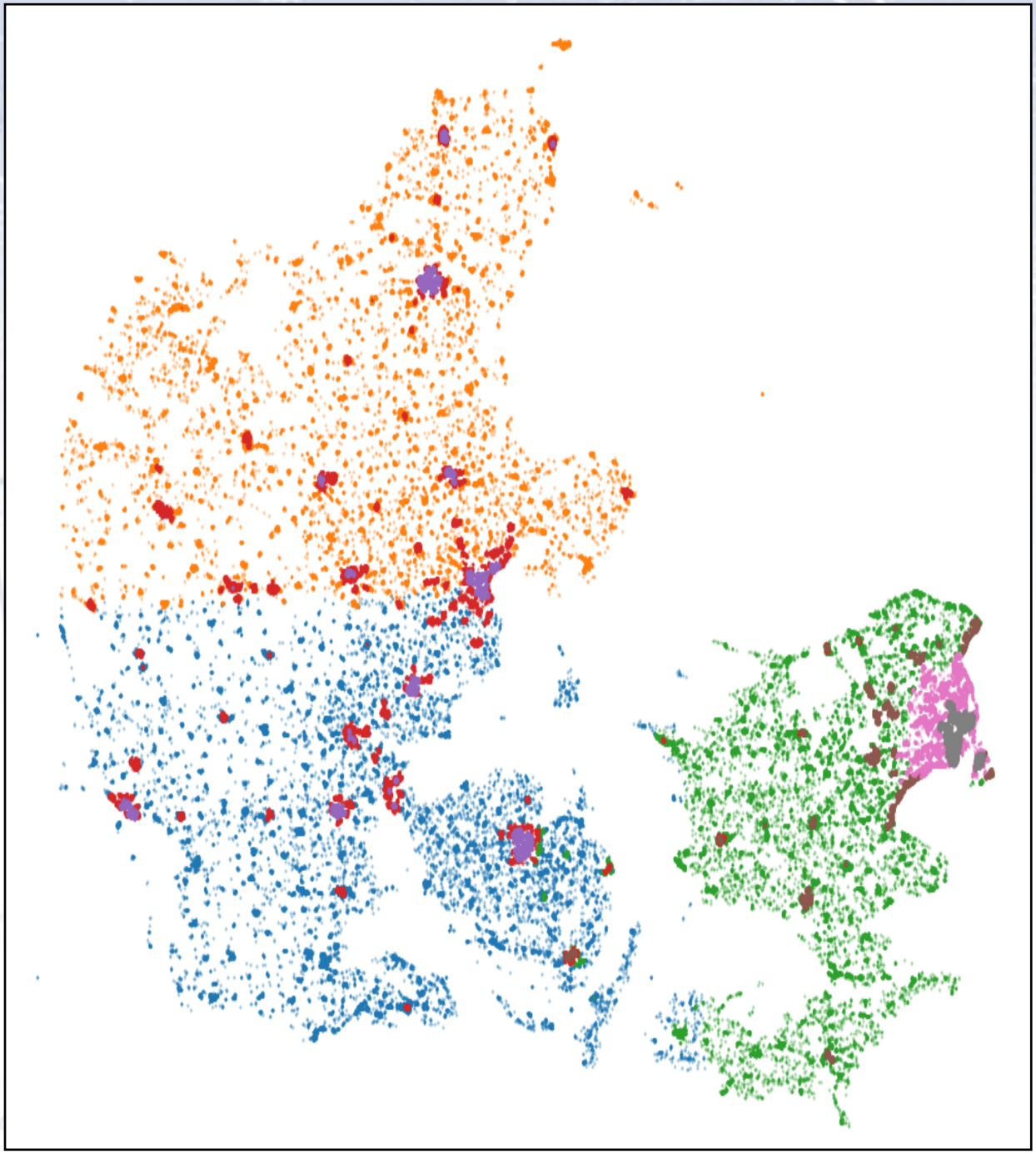
Some of the variables are known to those with domain knowledge to be good.

Clustering housing

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

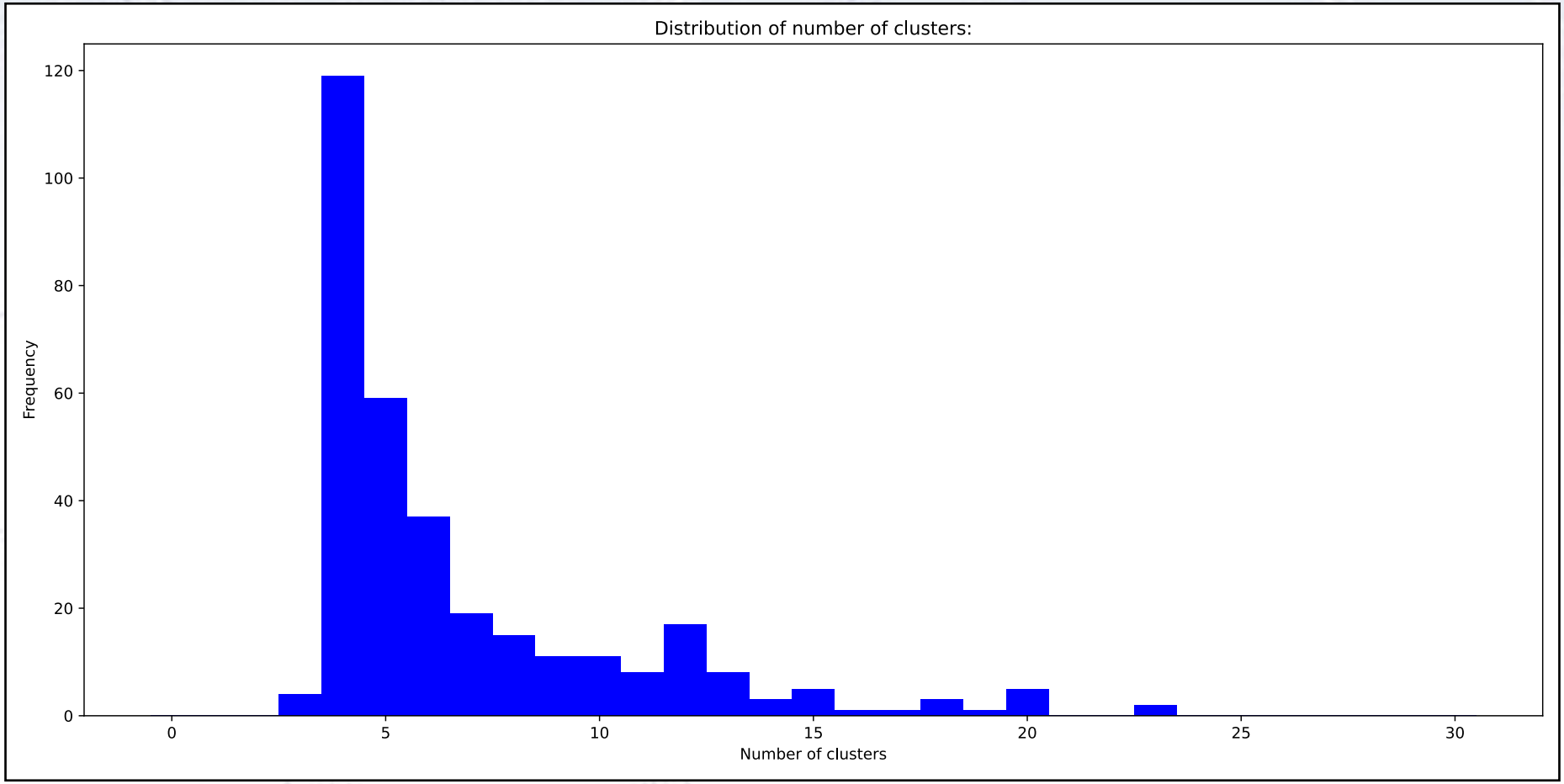
However, using just a few variables (x, y, density, price/m^2), one can cluster villas in Denmark very efficiently.

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



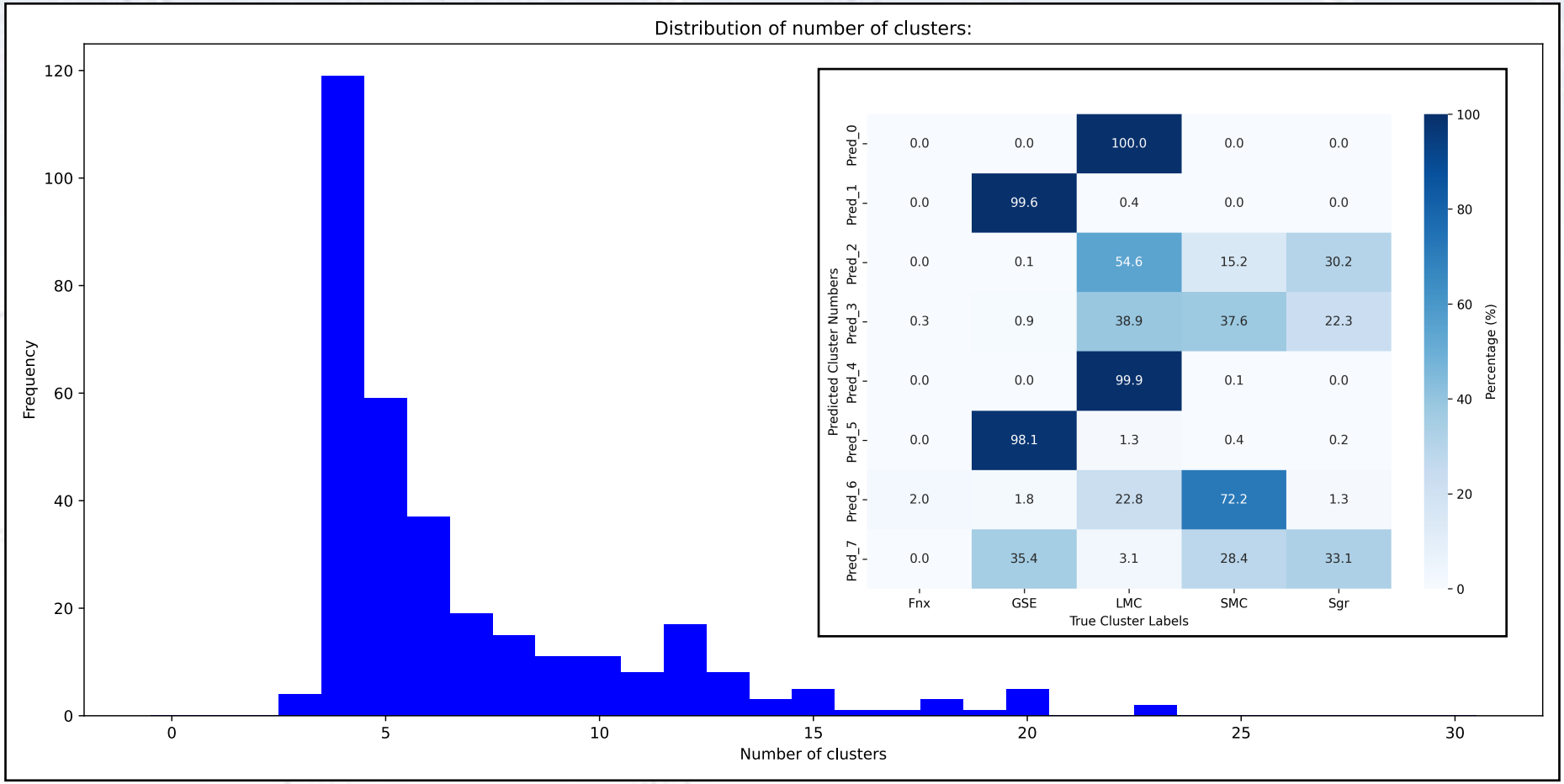
Clustering number distribution

The number of clusters was an interesting distribution (no-one went for 50!):



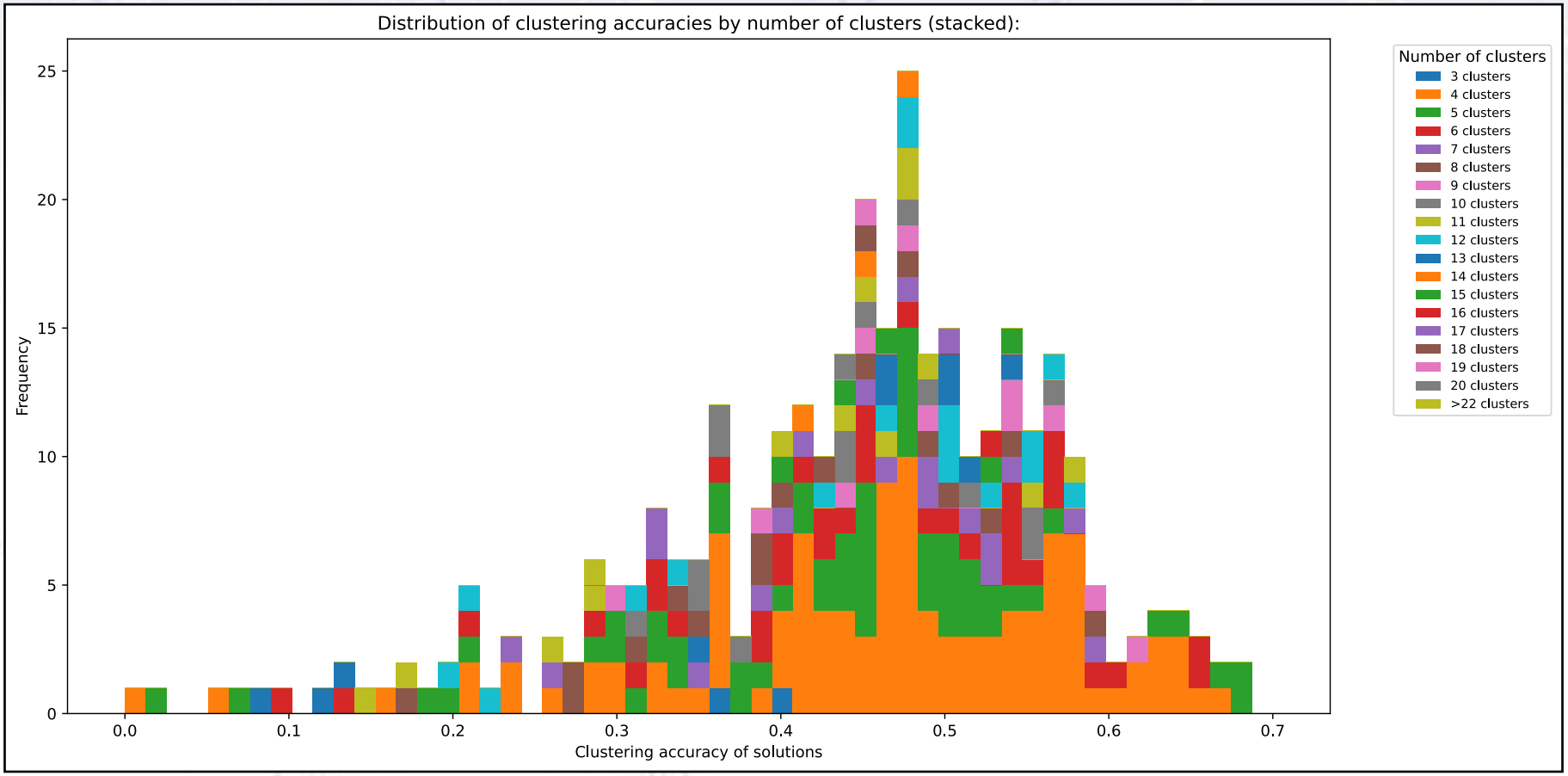
Clustering number distribution

The number of clusters was an interesting distribution (no-one went for 50!):



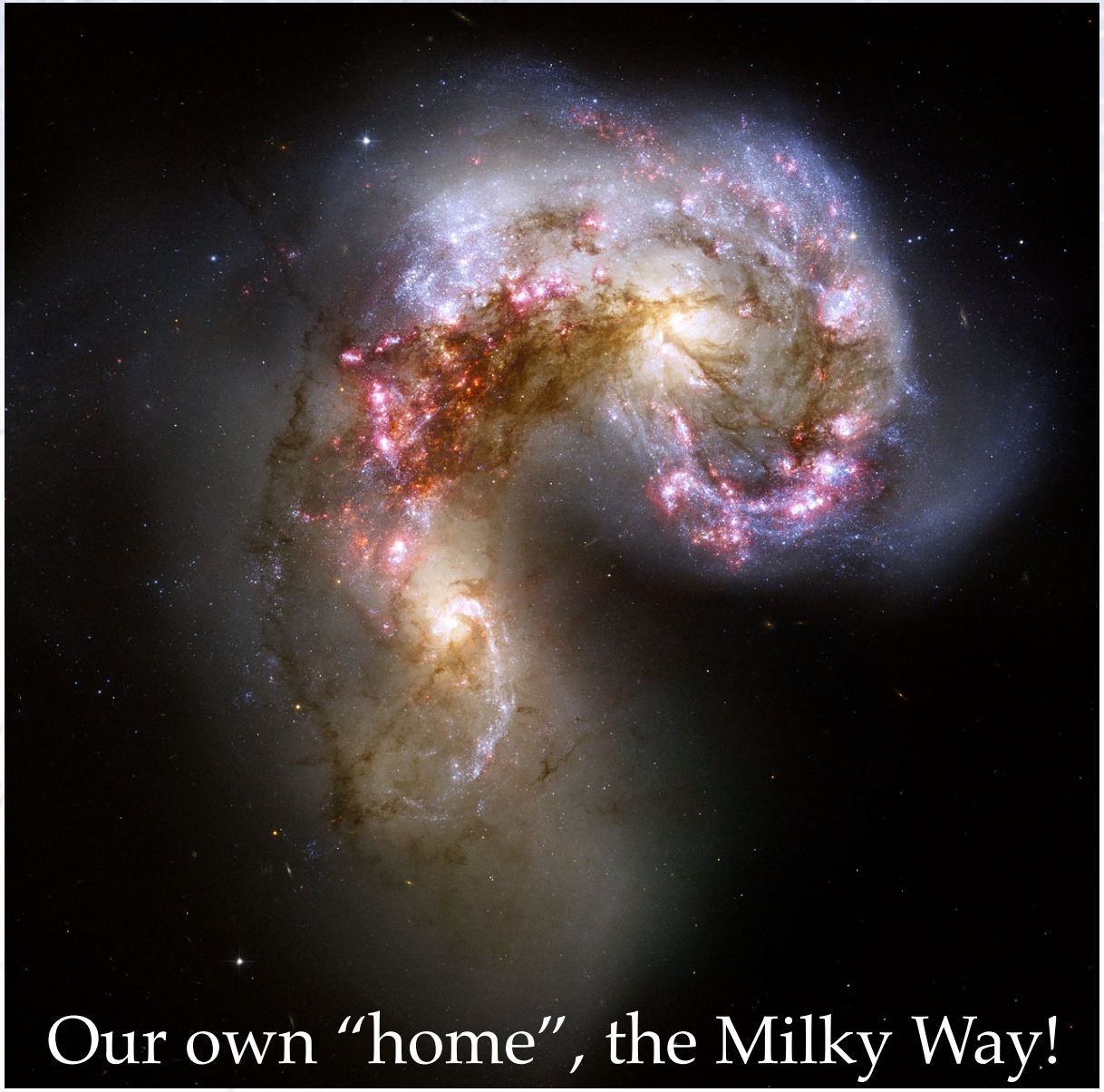
Clustering accuracy distribution

The accuracy of the clustering, given that there was five groups:



Notes about the data

What did we Cluster?



What did we Cluster?

The Milky Way has accreted many ultra-faint dwarf galaxies (UFDs), and stars from these galaxies can be found throughout our Galaxy today. Studying these stars provides insight into galaxy formation and early chemical enrichment, **but identifying them is difficult**. […]

Of the seven algorithms, **HDBSCAN most consistently balances UFD** recovery rates and cluster realness rates. We find that even in highly idealized cases, the vast majority of clusters found by clustering algorithms do not correspond to real accreted UFD remnants and we can generally **only recover 6% of UFDs remnants at best**.

arXiv:2206.07057

Our own "home", the Milky Way!

Scoring your solutions

How do we grade your projects?

Classification

- ✓ Complete **ResidualMLP**: 60000 predictions, 15 variables
  Variables: p_sigmad0, pX_MultiLepton, p_TRTPID...
- ✓ Complete **HistGBTree**: 60000 predictions, 15 variables
  Variables: p_sigmad0, pX_MultiLepton, p_TRTPID...
- ✓ Complete **DuplicateAware0999**: 60000 predictions, 15 variables
  Variables: p_sigmad0, pX_MultiLepton, p_TRTPID...
- ✓ Complete **RawFeatureBlend**: 60000 predictions, 15 variables
  Variables: p_sigmad0, pX_MultiLepton, p_TRTPID...

Regression

- ✓ Complete **TreeStackRelMAD**: 40000 predictions, 20 variables
  Variables: p_pt_track, pX_deltaPhiFromLastMeasurement, pX_deltaPhi2...
- ✓ Complete **FTTransformerMixup**: 40000 predictions, 20 variables
  Variables: pX_ecore, p_pt_track, pX_deltaPhi2...
- ✓ Complete **FTTransformerBlend**: 40000 predictions, 20 variables
  Variables: pX_ecore, p_pt_track, pX_deltaPhi2...

Clustering

- ✓ Complete **MetaConsensusK7**: 5950 predictions, 6 variables
  Variables: FE_H, SI_FE, AL_FE...
- ✓ Complete **GMMFullK8**: 5950 predictions, 6 variables
  Variables: FE_H, MG_FE, AL_FE...
- ✓ Complete **ConsensusK9**: 5950 predictions, 6 variables
  Variables: FE_H, SI_FE, AL_FE...

How do we grade your projects?

Ranking Summary

Best scoring classification: DuplicateAware0999 with a score of 7.0000

Your best classification solution ranked 1 out of 135 student best classification solutions.

Best scoring regression: TreeStackRelMAD with a score of 7.0000

Your best regression solution ranked 1 out of 135 student best regression solutions.

Best scoring clustering: ConsensusK9 with a score of 6.7700

Your best clustering solution ranked 27 out of 135 student best clustering solutions.

Your overall solution ranked 8 out of 135 students.

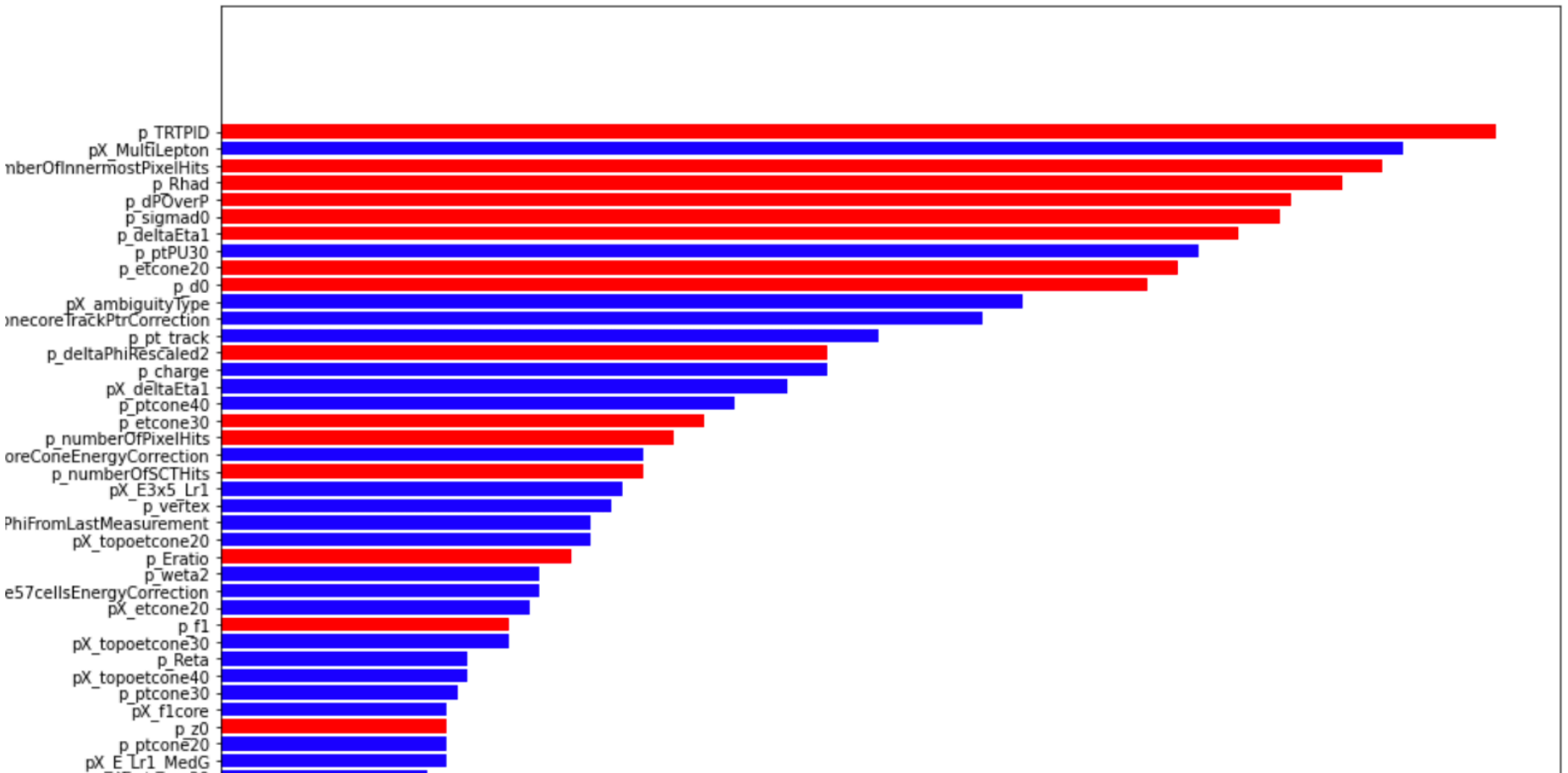
Scores Summary

You submitted a full solution, from which you get:	54.00	points
Your description was scored [0, 9]:	9.17	points
Your solution entailed 10 different algorithms, which gives you a score of [0, 6]:	6.00	points
Your best classification algorithm scored you [0,7]:	7.00	points
Your variable choice for classification scored you [0,4]:	4.00	points
Your best regression algorithm scored you [0,7]:	7.00	points
Your variable choice for regression scored you [0,4]:	4.00	points
Your best clustering algorithm scored you [0,7]:	6.77	points
Your variable choice for clustering scored you [0,4]:	3.15	points
Thus your total number of points was:	**101.09**	**points**

Your variable choice

Assuming, that the variable frequency reflected the actual ranking very well, your variable choice was scored as follows (factors were 4, 5, and 1):

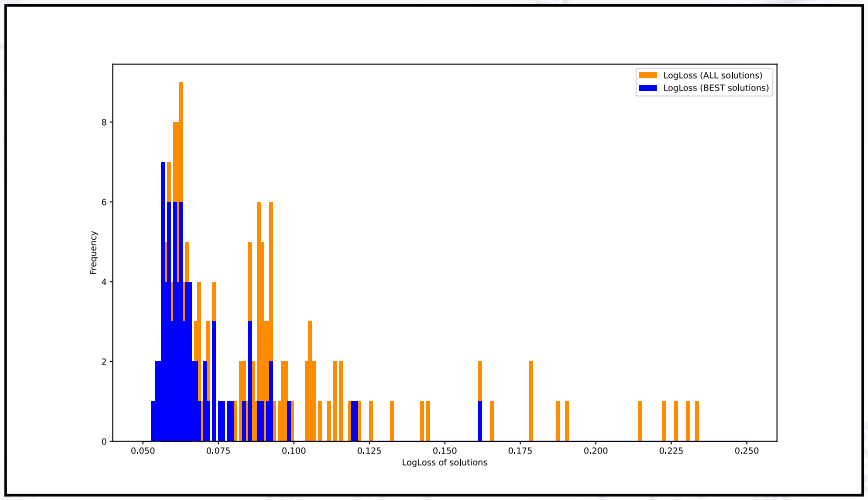
$$8 \times \left(\sum Freq(\text{Your variables}) / \sum Freq(\text{Top variables}) \right)$$

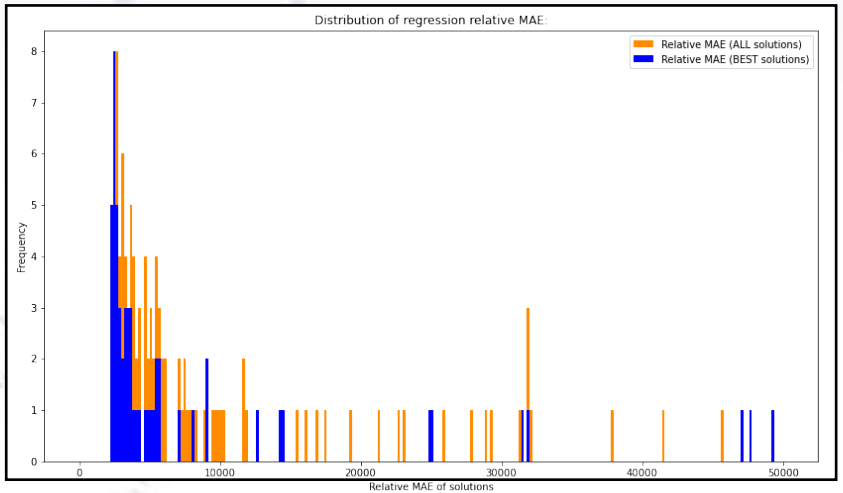


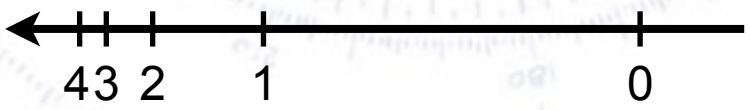
Performance scoring

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 [0,7+]:



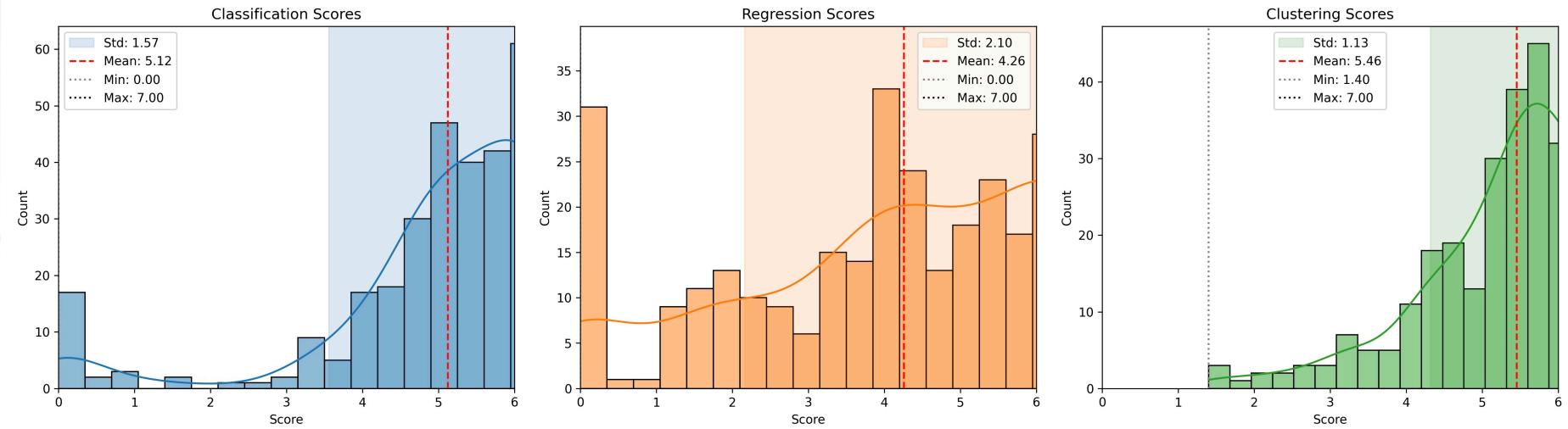




In all of this, you could of course not get negative points for an accepted solution!

The resulting score distributions

Score distributions for C**lassification, Regression, and Clustering** performance are shown below:

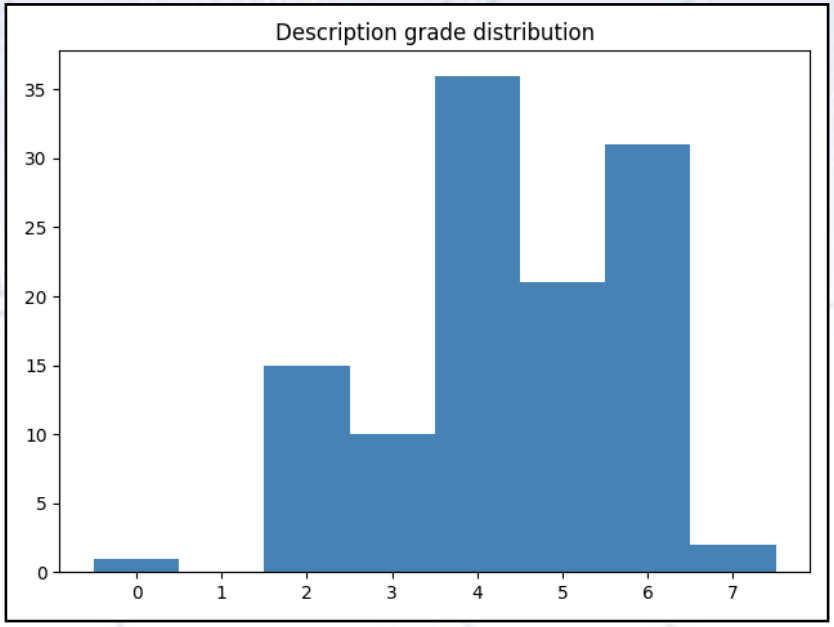


As you can see, most people got very nice scores.

In the regression task, training on non-electrons might be the (main) reason for reducing the performance (Note to self: Read the docs!).

The resulting score distributions

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



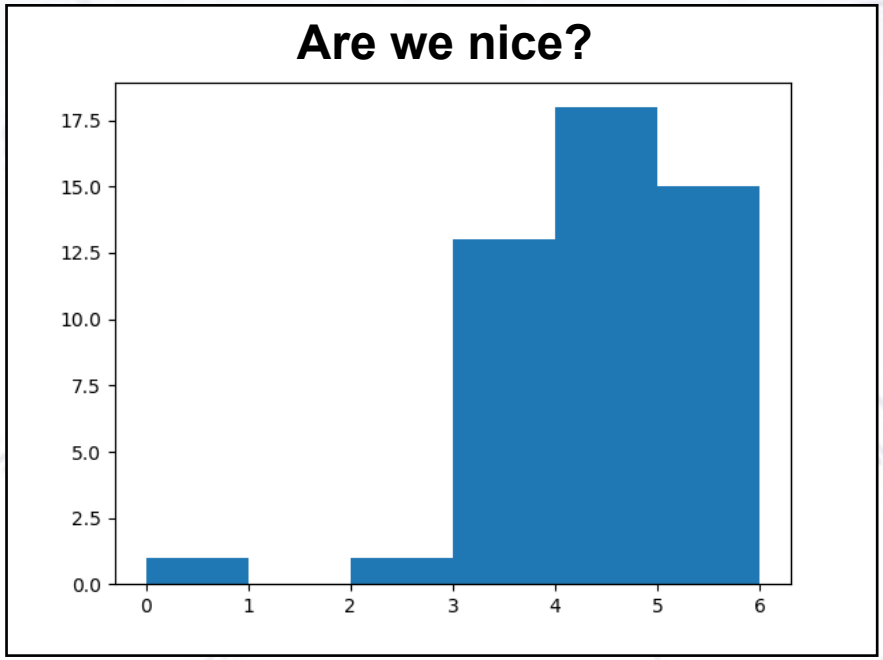
I read several of the "lower scoring" descriptions, but must say that I found them "reasonably acceptable", so in general the level was high (but don't do transformation of variables, when using a BDT!).

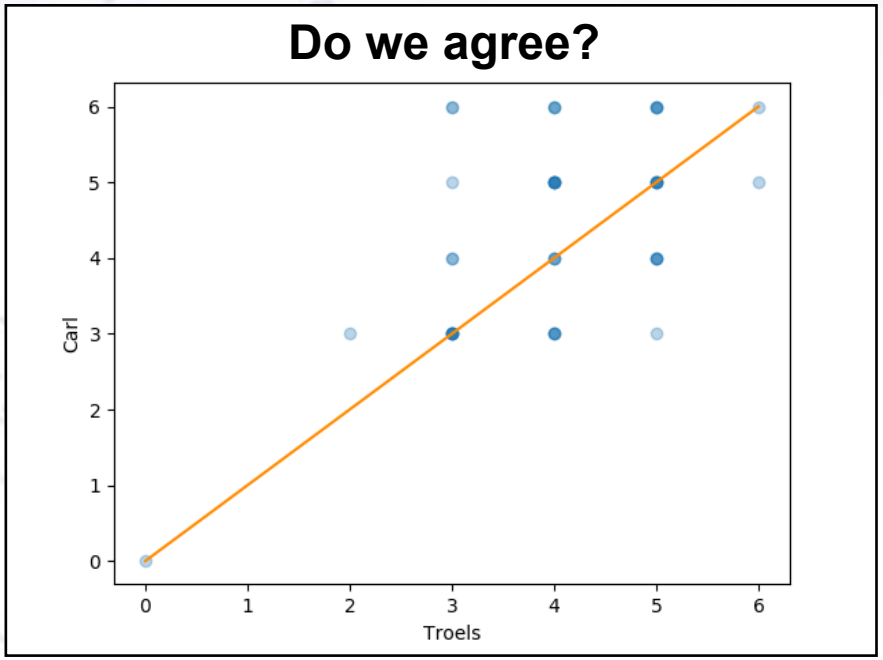
On algorithms, it was great to see that you both stuck with what you knew, but also explored new algorithms and got them working.

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.

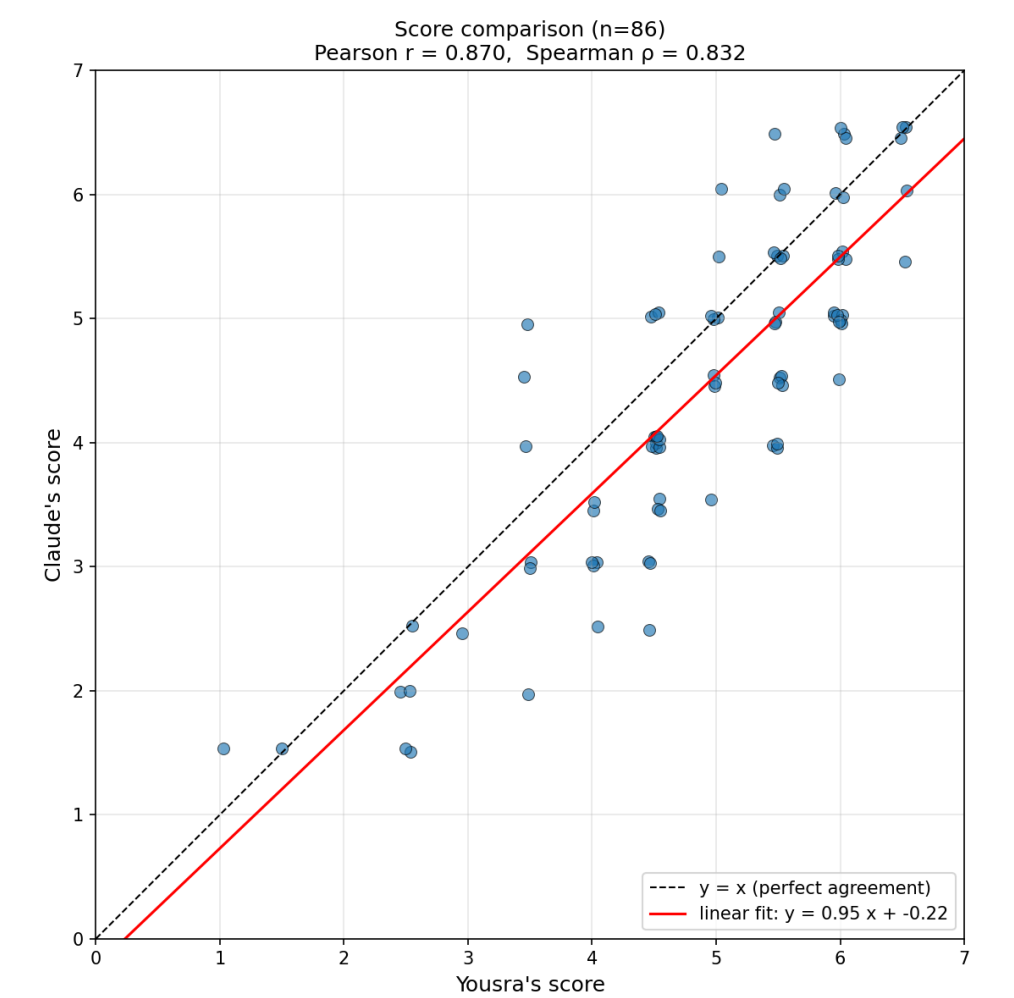
Numbers from 2021 (where Carl and I did it):





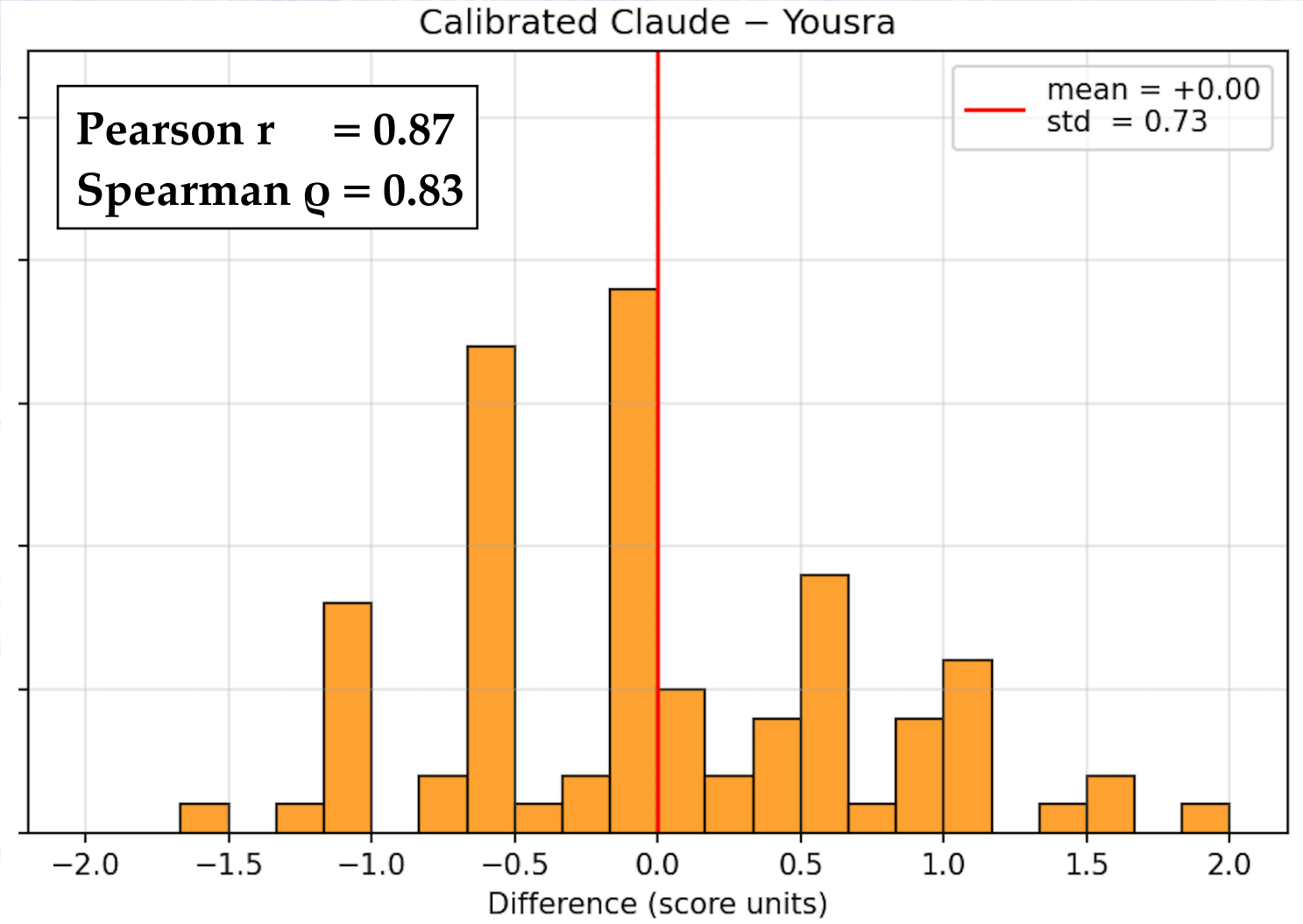
As you can see, we were generally satisfied. The descriptions were short and to the point, and give some insight into your line of thinking and working.

Does Claude agree with us?



Thanks to Yousra for scoring the majority of descriptions…

Does Claude agree with us?



Thanks to Yousra for scoring the majority of descriptions…

Reporting back to you

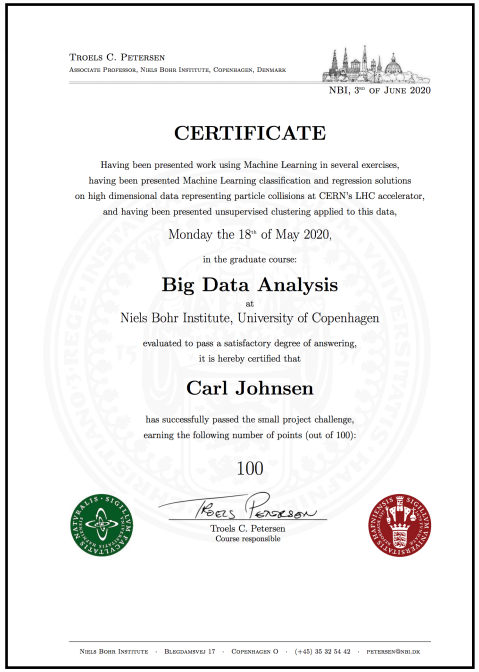
Feedback to you

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 during the exercises. 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…

Metric	Equation	Value
Accuracy	`sklearn.metrics.accuracy_score`	0.961083
AUC	`sklearn.metrics.auc`	0.987614
Cross entropy	`sklearn.metrics.log_loss`	0.108107

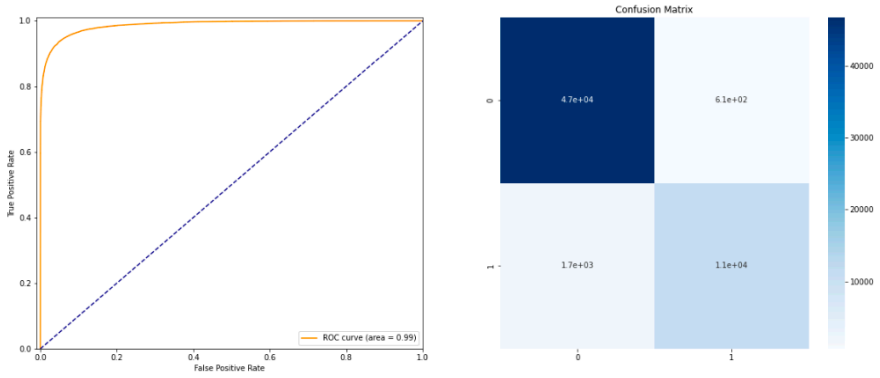
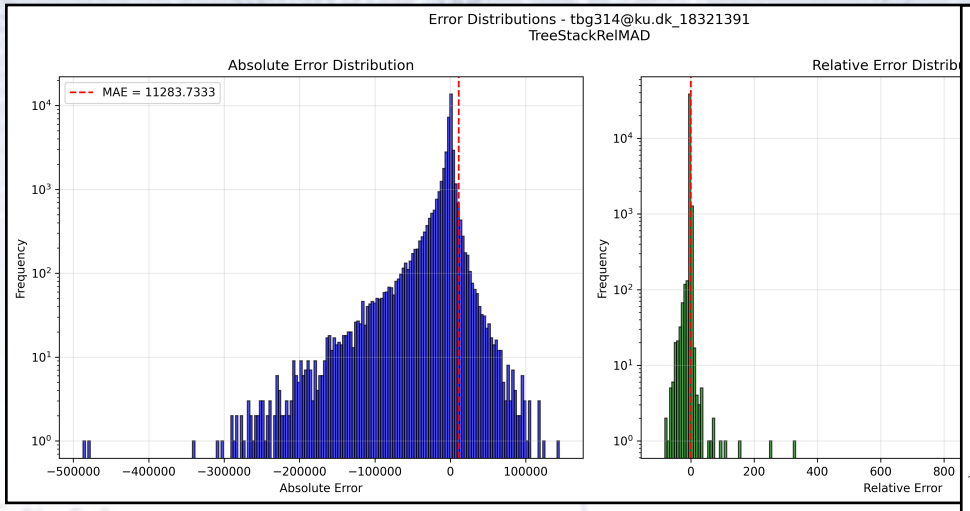
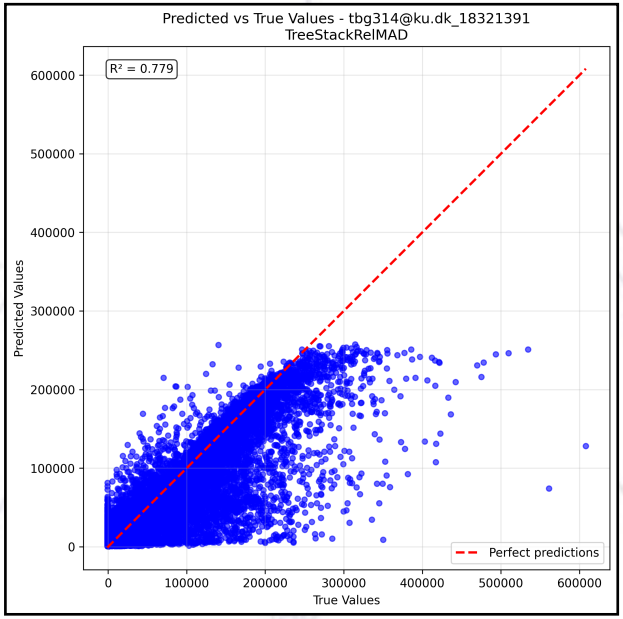
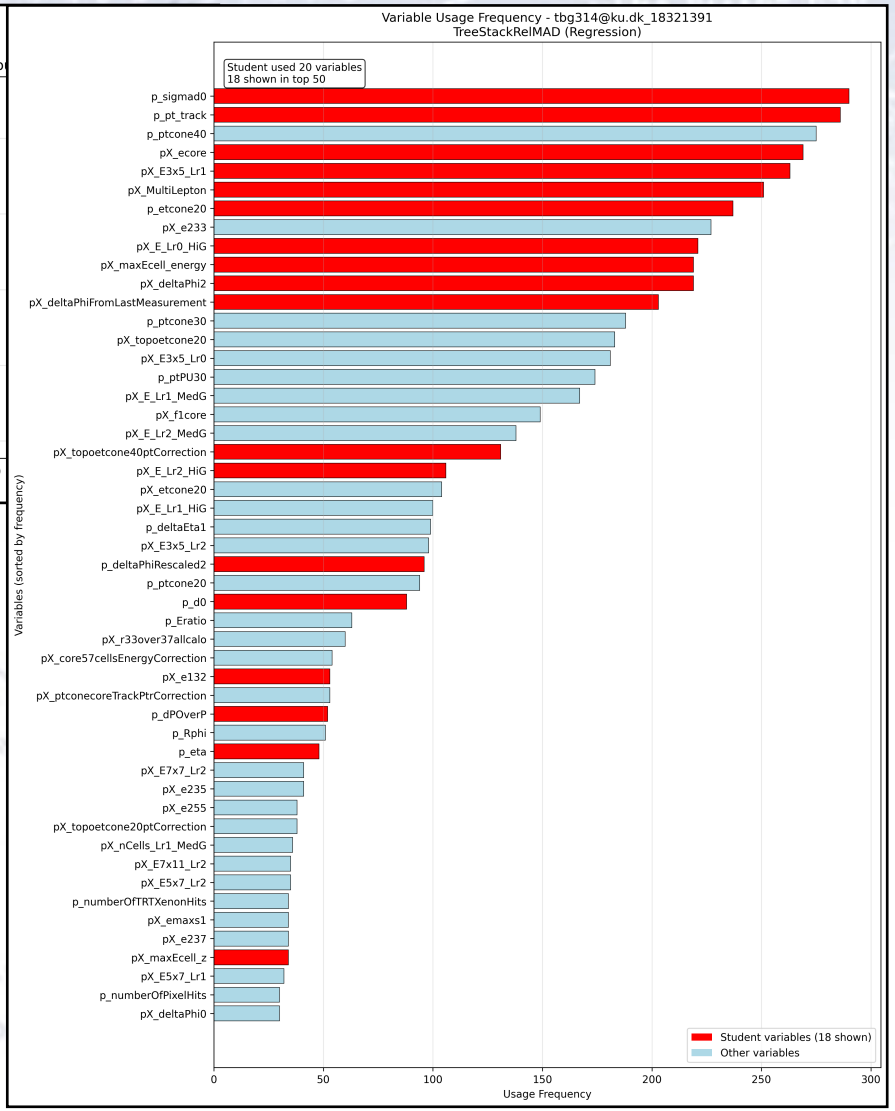


Figure 1: **Left:** ROC curve for the RandomForest implementation. The orange curve should be as close to the upper left corner as possible. **Right:** Confusion matrix for the RandomForest implementation. The diagonal squares ((0,0) and (1,1)) should have the higher values, compared to the squares in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

Regression report







Regression report

The solution gave the following metrics:

Metric	Equation	Value
MAE - Absolute	`sklearn.metrics.mean_absolute_error`	8120.6694
MAE - Relative	$\sum \lvert \frac{y_p - y_t}{y_t} \rvert$	8170.9401
RMS	$\sqrt{mean((y_p - y_t)^2)}$	13599.8964
RMS 98th percentile	$\sqrt{mean((y_p - y_t)^2)}$	10756.5212
RMS 90th percentile	$\sqrt{mean((y_p - y_t)^2)}$	8285.5263
RMS 70th percentile	$\sqrt{mean((y_p - y_t)^2)}$	6328.3174

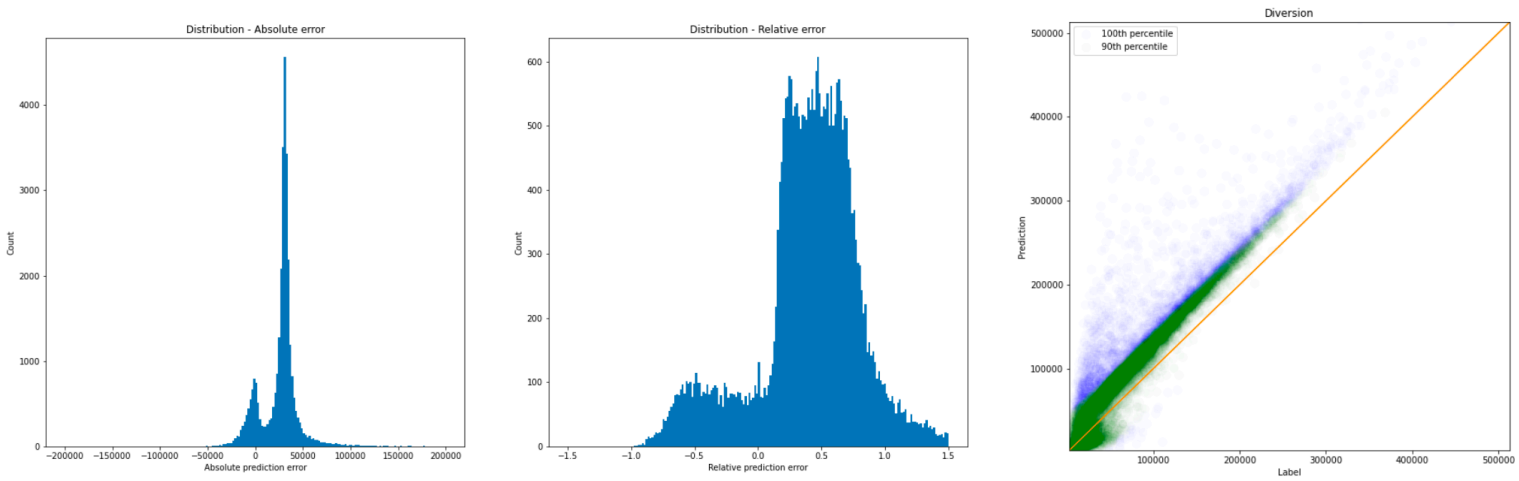
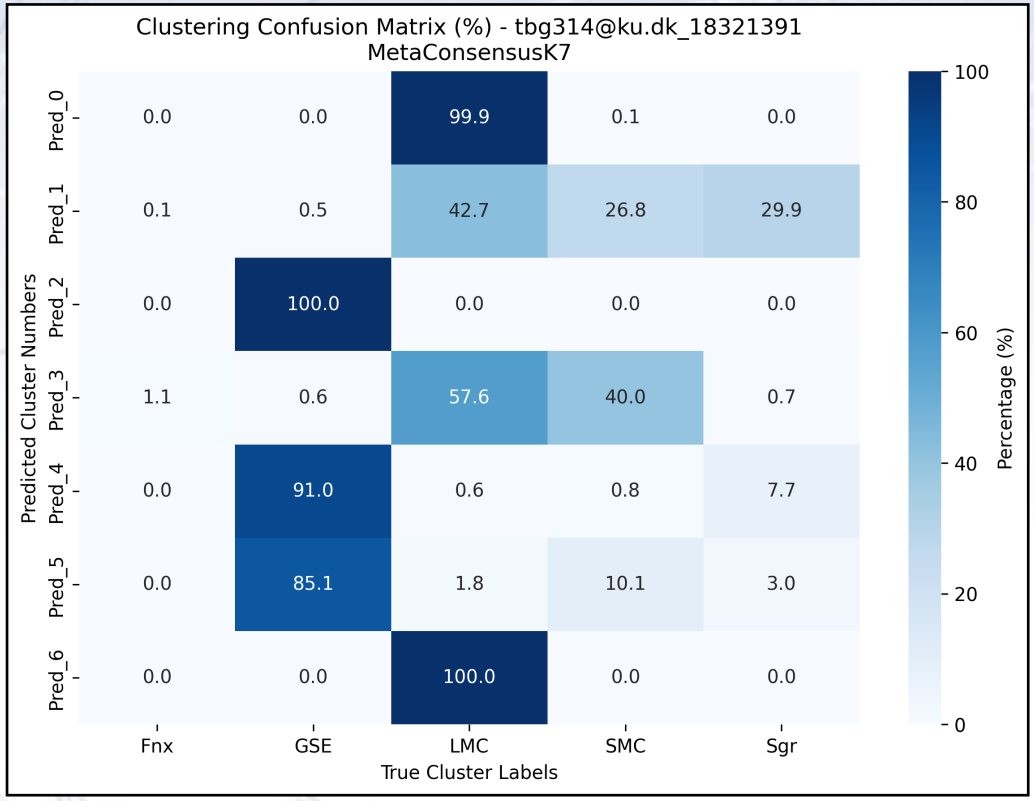


Figure 3: **Upper:** Distribution plots for the XGBoost implementation. The plots are for absolute error (*Left*) and relative error (*Right.*). Both plots should have a tall narrow curve, centered around 0. **Lower:** Diversion plot for the XGBoost implementation. The dots should be scattered close to the orange line - especially for the 90th percentile (green dots).

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."



Thank you, for all your hard work