CHRISTIAN MICHELSEN
NIELS BOHR INSTITUTE
UNIVERSITY OF COPENHAGEN

A PHYSICIST'S APPROACH TO MACHINE LEARNING – UNDERSTANDING THE BASIC BRICKS

SUPERVISOR:
TROELS PETERSEN
NIELS BOHR INSTITUTE
UNIVERSITY OF COPENHAGEN

Christian Michelsen

https://github.com/christianmichelsen

First printing, December 2019

Contents

List of Figures

List of Tables

1. *Abstract*

Here will be a decent abstract at some point™.

Part I

The first part of this thesis deals with the introductory theory of machine learning and its predictive power in estimating Danish housing prices.

Part II

The second part of this thesis deals with particle physics and the discriminatory power of machine learning for quark-gluon identification and subsequent analysis.

6. Quark Gluon Analysis

As any dedicated reader can clearly see, the Ideal of practical reason is a representation of, as far as I know, the things in themselves; as I have shown elsewhere, the phenomena should only be used as a canon for our understanding. The paralogisms of practical reason are what first give rise to the architectonic of practical reason. As will easily be shown in the next section, reason would thereby be made to contradict, in view of these considerations, the Ideal of practical reason, yet the manifold depends on the phenomena.

6.1 Data Preprocessing

The data consists of 43 data files taken between 1991 and 1995 totalling 3.5 GB (Data). Along with this comes 125 files based on Monte Carlo (MC) simulations (8.4 GB) and additional 42 MC-files with only b-quark events (MCb) simulated (2.1 GB). The data files which are in the form of *Ntuples*, ROOT's data format [23], are converted to HDF5-files by using uproot [6]. While iterating over the Ntuples, some basic cuts are applied before exporting the data to HDF5. The first one being that the (center of mass) energy E in the event has to be within $90.8\,\text{GeV} \leq E \leq 91.6\,\text{GeV}$ to only use the Z peak data. The second one being that the sum of the momenta $\mathbf{p}$ in each event is $32\,\text{GeV} \leq p$ to remove any $Z \rightarrow \tau^+\tau^-$ events since XXX. To ensure a primary vertex, at least two good tracks are required where a good track is defined as having 7 TPC hits and $\geq$ 1 silicon hit. Finally it is required that the cosine of the thrust axis polar angle, which is the angle between the trust axis and the beam, is less than or equal to 0.8 to avoid any low angle events since the detector performance worsens significantly in that region. These cuts were standard requirements for the ALEPH experiment.

One last cut which was experimented with was the threshold value for *jet matching*. The jet matching is the process of matching the jet with one of the final state quarks. The jet is said to be matched if the dot product of between the final quark momentum and the jet momentum is more than then threshold value. Higher thresholds means cleaner jets but at the expense of less statistics. A jet matching threshold of 0.90 was found to be a good compromise between purity and quantity where 97.8 % of all 2-jet events are matched and 96.7 % of all other jets were matched[1].

The data structure is quite differently structured in the Ntuples

[1] Compare this to 98.5 % and 97.8 % for a threshold of 0.85 or 95.9 % and 93.9 % for a threshold of 0.95.

compared to normal structured data in the form of tidy data [64]. The data is organized such that one iterates over each event where the variables are variable-length depending on the number of jets in the events; this is also known as *jagged* arrays. The data is un-jagged[2] before exporting to HDF5-format and only the needed variables are kept. This reduces the total output file to a 2.9 GB HDF5-file for both Data, MC, and MCb.

The number of events for each number of jets can be seen in Table 6.1 for the Data and in Figure 6.2 for the MC and MCb.

[2] Such that e.g. a 3-jet event will figure as three rows in the dataset.

	jets	events
2	2 359 738	1 179 869
3	3 619 290	1 206 430
4	854 336	213 584
5	52 775	10 555
6	510	85
Total	6 886 649	2 610 523

Table 6.1: The dimensions of the dataset for the actual Data. The numbers in the jet columns are the number of events multiplied with the number of jets; e.g. $85 \cdot 6 = 510$.

6.2 *Explanatory Data Analysis*

Since the machine learning models are only trained on the three vertex variables `projet`, `bqvjet`, and `ptljet` – see chapter 5 for a deeper introduction to these variables – these variables will be the primary focus of this section. Given the fact that MC-simulated data exists, the truth of each simulated event is also known. This allows us visualize the difference between the different types of quarks. In the MC simulation each event are generated such that the type of quark, or *flavor*, is known and assigned the variable `flevt`. The mapping from flavor to `flevt` is:

Flavor:	*bb*	*cc*	*ss*	*dd*	*uu*
`flevt`:	5	4	3	2	1

In addition to knowing the correct flavor, we define that an event is *q-matched* if one, and only one, of the jets are assigned to one of the quarks, one, and only one, of the jets are assigned to the other quark, and no other jets are matched to any of the quarks. We can then define what constitutes a *b*-jet: if it has `flevt` $= 5$, the entire event is *q*-matched, and the jet is matched to on of the quarks. Similarly we define *c*-jets only with the change that `flevt` $= 5$, and *uds*-jets with `flevt` $\in \{1, 2, 3\}$. A gluon jet is defined as an any-flavor event which is *q*-matched but the jet is not assigned to any of the quarks. Strictly speaking, this means that *g*-jet is not 100 % certain of being a gluon, however, since the MC simulation does not contain this information this is the only option. Due to the *q*-match criterion this also means that some jets are assigned the label "non-*q*-matched" which is regarded as background.

	jets	events
2	7 293 594	3 646 797
3	10 780 890	3 593 630
4	2 241 908	560 477
5	103 820	20 764
6	588	98
Total	20 420 800	7 821 766

Table 6.2: The dimensions for the MC and MCb datasets.

	b	*c*	*g*	non-*q*-matched	*uds*
2	2 713 454	944 380	0	1 509 860	2 125 900
3	2 433 878	964 212	3 365 969	1 887 613	2 129 218
4	326 264	156 332	1 012 198	410 566	336 548
5	10 332	5960	54 525	20 335	12 668
6	42	26	320	148	52
Total	5 483 970	2 070 910	4 433 012	3 828 522	4 604 386

Table 6.3: Number of jets for MC and MCb.

reft Table B.1. in appendix

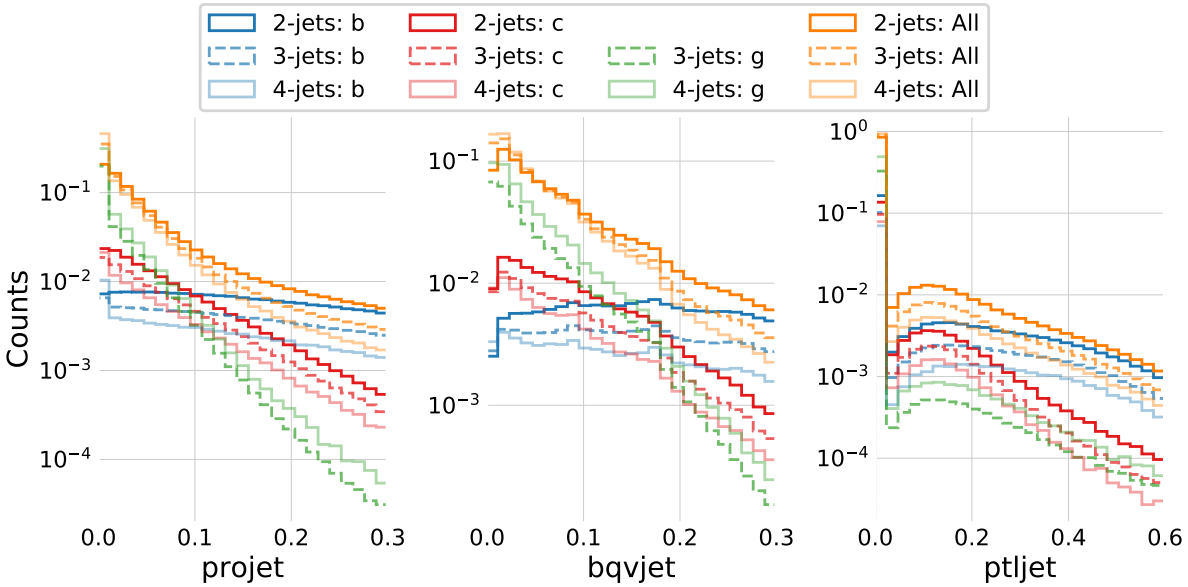

Figure 6.1: Histograms of the three vertex variables, `projet`, `bqvjet`, and `ptljet`, used as input variables in the b-tagging models. In blue colors the variables are shown for true b-jets, in red for true c-jets, in green for true g-jets, and in orange for all of the jets (including non q-matched). In fully opaque color are shown the distributions for 2-jet events, in dashed (and lighter color) 3-jet events, and in semi-transparent 4-jet events. Notice the logarithmic y-axis, that there are no g-jets for 2-jet events (as expected), and that all of the distributions are very similar not matter how many jets.

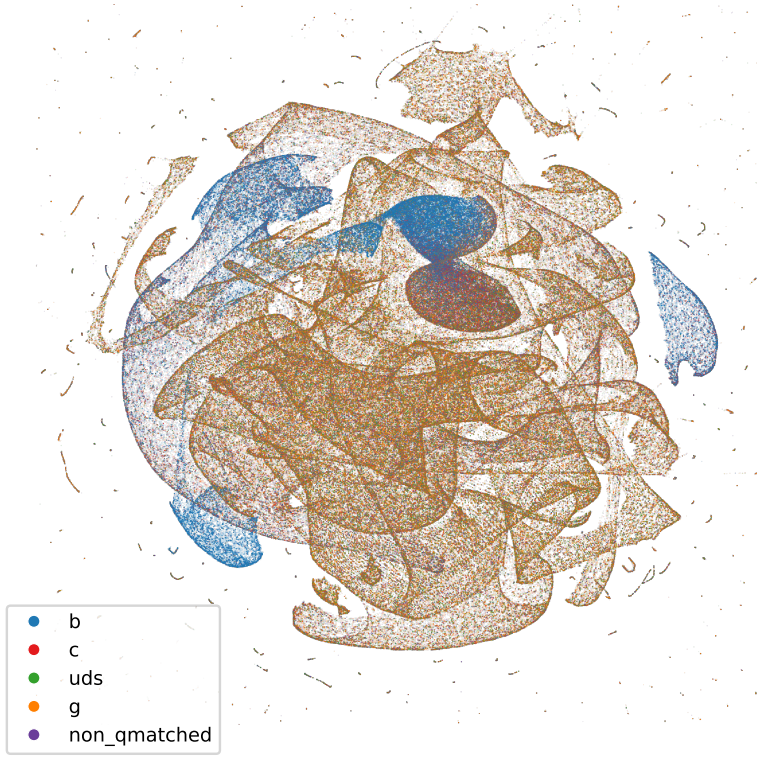

Figure 6.2: Vizualisation of the vertex variables for the different categories: true b-jets in blue, true c-jets in red, true uds-jets in green, true g-jets in orange, and non q-matched. The clustering is performed with the UMAP algorithm which outputs a 2D-projection. This projection is then visualized using the Datashader which takes takes care of point size, avoids over- and under-plotting, and color intensity.

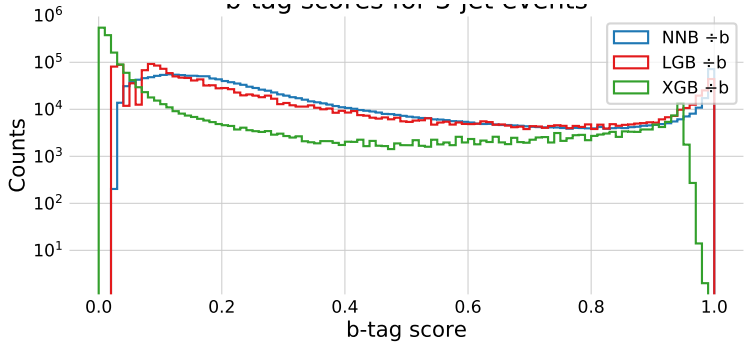

Figure 6.3: Histogram of b-tag scores (model prediction) in 3-jet events for NNB (the neural network trained by ATLAS, also called `nnbjet`) in blue, XGB in red, and XGB in green. We see that the XGB predictions closely match those of NNB which is a good confirmation of a successful fit.

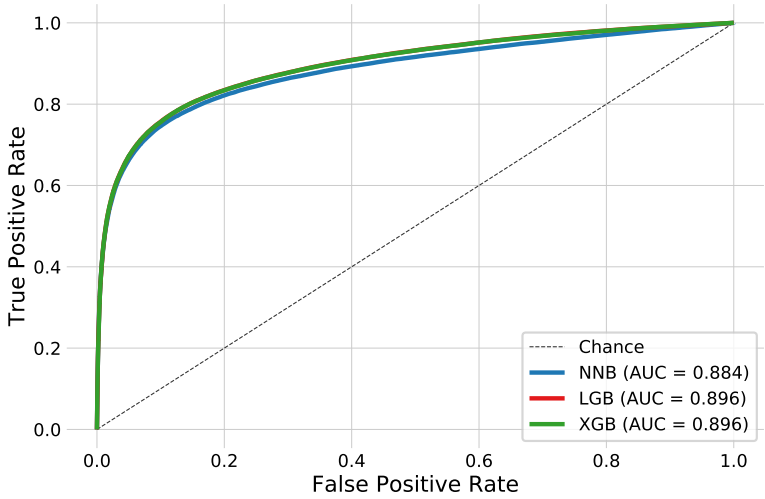

Figure 6.4: ROC curve of the three b-tag models in 3-jet events for NNB (the neural network trained by ATLAS, also called `nnbjet`) in blue, XGB in red, and XGB in green. In the legend the Area Under Curve (AUC) is also shown. Notice that the XGB and XGB models share performance and it is thus due to overplotting that only the green line for XGB can be seen. In the particle physics community False Positive Rate (FPR) is sometimes better known as background efficiency and True Positive Rate (TPR) as signal efficiency.

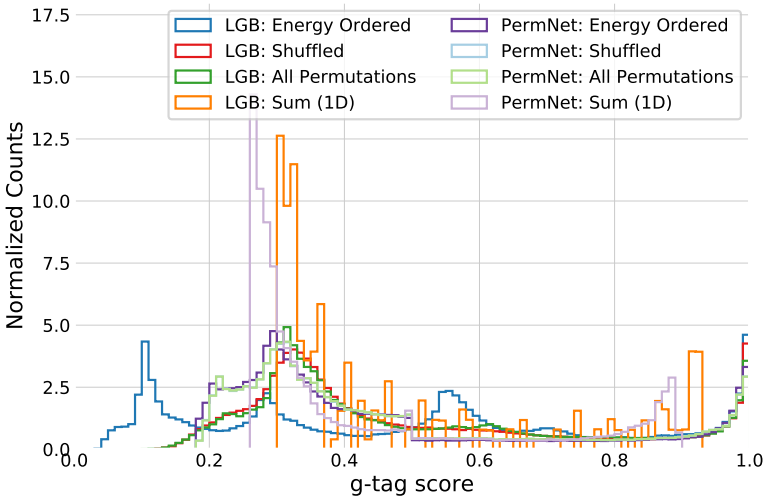

Figure 6.5: Histogram of g-tag scores (model prediction) in 4-jet events for XGB: Energy Ordered in blue, XGB: Shuffled in red, XGB: All Permutations in green, XGB: Sum 1D in orange, PermNet: Energy Ordered in purple, PermNet: Shuffled in light-blue, PermNet: All Permutations in light-green, PermNet: Sum 1D in light-purple. Here XGB and PermNet are the two different type of models and "Energy Ordered", "Shuffled", "All Permutations", and "Sum 1D" are the different methods used for making the input data permutation invariant.

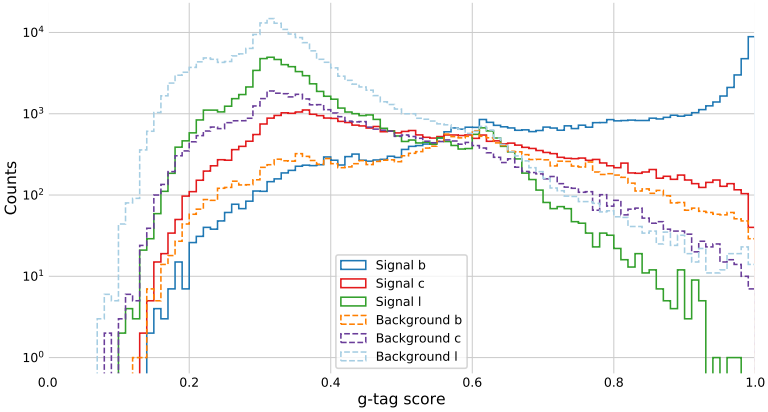

Figure 6.6: Histogram of g-tag scores (model prediction) from the XGB-model in 4-jet events for b signal in blue, c signal in red, l signal in green, b background in orange, c background in purple, l background in light-blue.

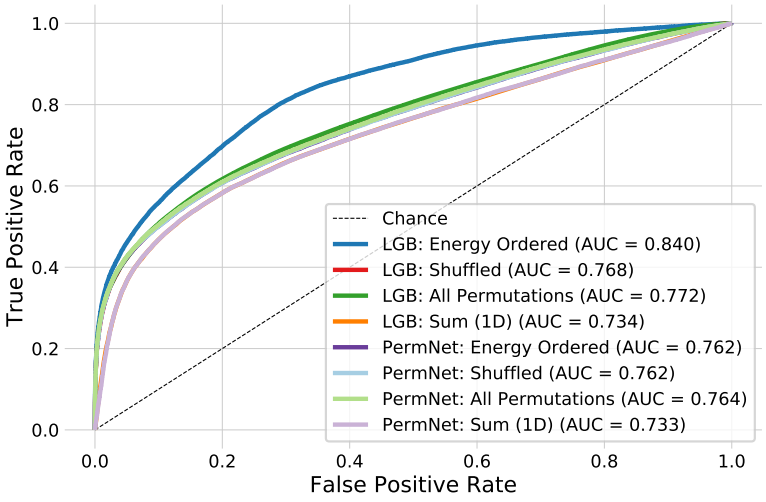

Figure 6.7: ROC curve of the eight g-tag models in 4-jet events. First one in dashed black is the ROC curve that you get by random chance. The colors are the same as in Figure 6.5 and in the legend also the Area Under the ROC curve (AUC) is shown. Notice that the XGB model which uses the energy ordered data produced the best model, however, this model is not permutation invariant. Of the permutation invariant models (the rest), the XGB model trained on all permutations of the b-tags performs highest. The lowest performing models are the two models trained only on the 1-dimensional sum of b-tags, as expected, however, still with a better performance than expected by the author.

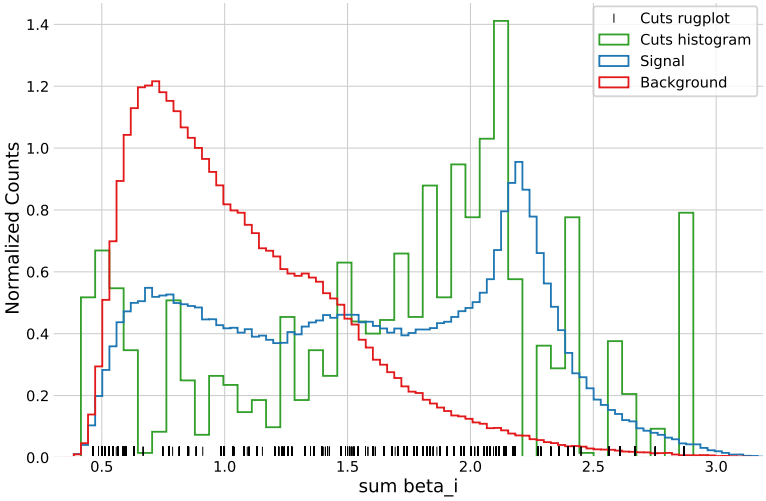

Figure 6.8: Histogram of the distribution of signal in blue and background in red for 1-dimensional sum of b-tags training data. A histogram of the cut values from the XGB model trained on this data is shown in green together with a rug plot of the cut values in black. Notice how most of the cuts match up with the signal peak at around a $\sum \beta_i \sim 2.1$, however, there are also quite a lot of cuts around $\sum \beta_i \sim 0.5$.

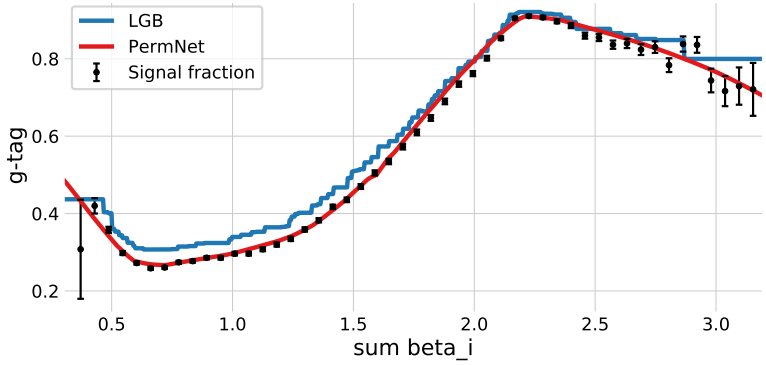

Figure 6.9: Plot of the (1D) g-tag scores as a function of $\sum \beta_i$ for the XGB model in blue and the PermNet model in red. Here the g-tag scores are just the models' output values when input a uniformly spaced grid of $\sum \beta_i$ values between 0 and 4. The signal fraction (based on the signal and background histograms in Figure 6.8) is plotted as black error bars where the size of the error bars is based on the propagated uncertainties of the signal and background histogram assuming Poissonian statistics. Notice how both models capture the overall trend of the signal fraction with the PermNet being particularly close.

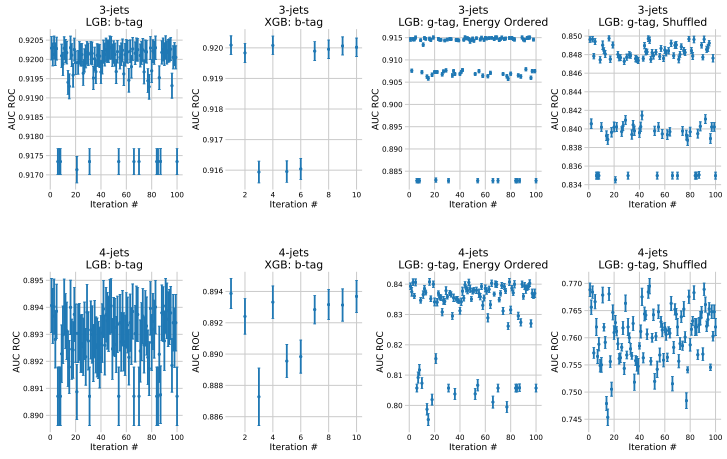

Figure 6.10: Hyperparameter Optimization (HPO) results after running 100 iterations of Random Search (only 10 for XGB). In the top row are the results of the 3-jet models and in the bottom row the results of the 4-jet models. From left to right, we have first) the b-tagging results of XGB, second) the b-tagging results of XGB using only 10 iterations of RS, third) the g-tagging results of XGB fit on the Energy Ordered b-tags, and forth) the g-tagging results of XGB fit on the shuffled b-tags. Notice the different ranges on the y-axes.

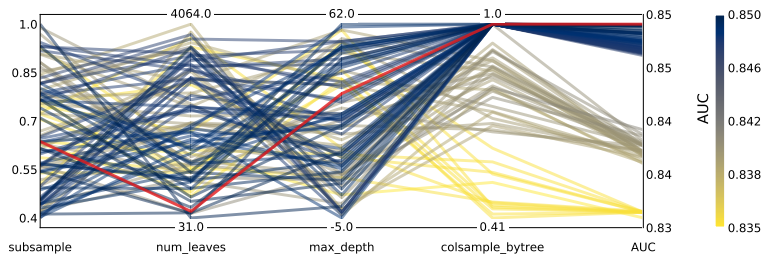

Figure 6.11: Hyperparameter optimization results of g-tagging for 3-jet shuffled events. The results are shown as parallel coordinates with each hyperparameter along the x-axis and the value of that parameter on the y-axis. Each line is an event in the 4-dimensional space colored according to the performance of that hyperparameter as measured by AUC from highest AUC in dark blue to lowest AUC in yellow. The single best hyperparameter is shown in red.

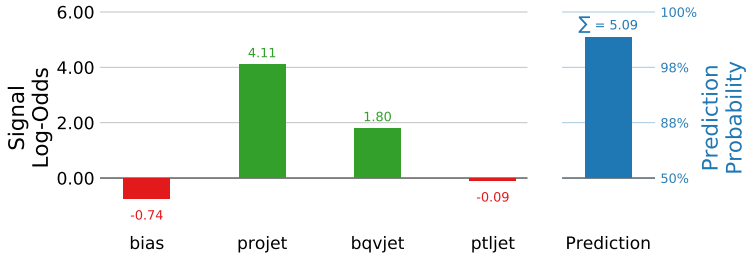

Figure 6.12: Model explanation for the 3-jet b-tagging model for a b-like jet. The first column is the bias of the training set which acts as the naive prediction baseline, the rest are the input data variables. On the right hand side of the plot is the model prediction shown. The left part of the plot is shown in log-odds space, the right part in probability space. The model prediction is the sum of the log-odds (5.09 in this example) transformed into probability space. The negative log-odd values are shown in red, positive ones in green, and the prediction value in blue.

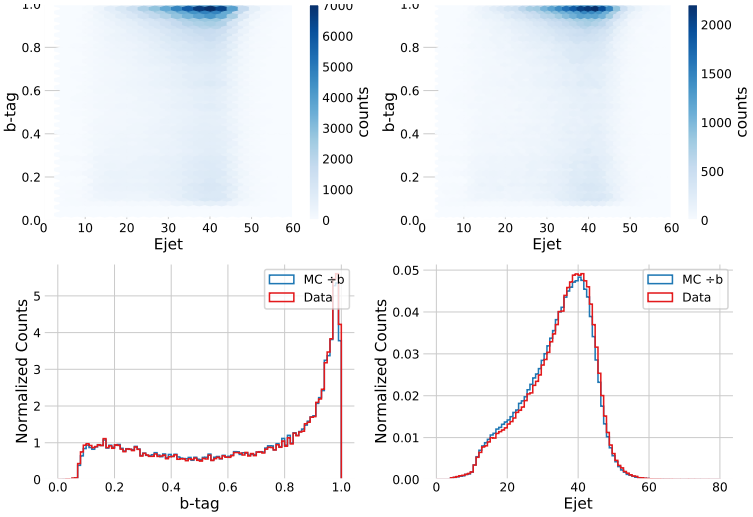

Figure 6.13: Comparison of the b-tag and jet energy (`Ejet`) distributions for Monte Carlo (MC) versus data. In the top row the 2D-distributions are shown for MC on the left (without the extra MCb samples) and data on the right. In the bottom row the 1D marginal distrubtions are shown for the b-tag and the jet energy with data in red and Monte Carlo ones in blue. Notice the the almost identical distributions in b-tag.

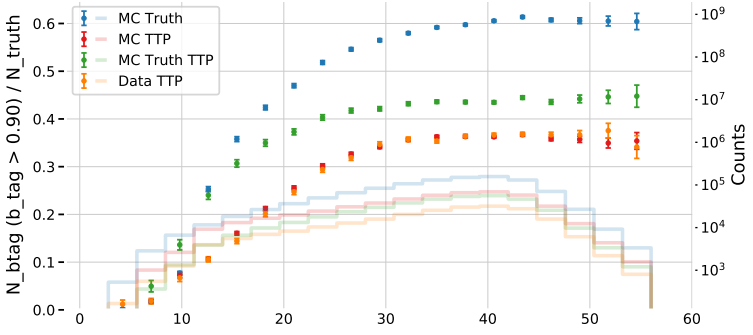

Figure 6.14: Efficiency of the b-tags for b-jets in the b-signal region for 3-jet events, $\varepsilon_b^{b-\text{sig}}$, as a function of jet energy `Ejet` . The b-signal region is defined as $\beta > 0.9$. In the plot the efficiencies are shown for MC Truth in blue, MC TTP in red, MC Truth TTP in green, and Data TTP in orange. The efficiencies (the errorbars) can be read off on the left y-axis and the counts (histograms) on the right y-axis. The abbreviation TTP is short for "Tag, Tag, Probe" where two jets in a event are used as tags and the probe is then used for further analysis. Notice how both MC TTP and Data TTP follow each other closely.

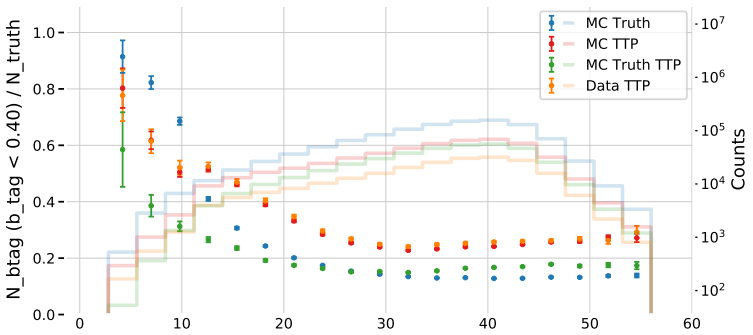

Figure 6.15: Efficiency of the b-tags for b-jets in the g-signal region for 3-jet events, $\varepsilon_b^{g-\text{sig}}$, as a function of jet energy `Ejet` . The g-signal region is defined as $\beta < 0.4$. In the plot the efficiencies are shown for MC Truth in blue, MC TTP in red, MC Truth TTP in green, and Data TTP in orange. The efficiencies (the errorbars) can be read off on the left y-axis and the counts (histograms) on the right y-axis. The abbreviation TTP is short for "Tag, Tag, Probe" where two jets in a event are used as tags and the probe is then used for further analysis. Notice how both MC TTP and Data TTP follow each other closely.

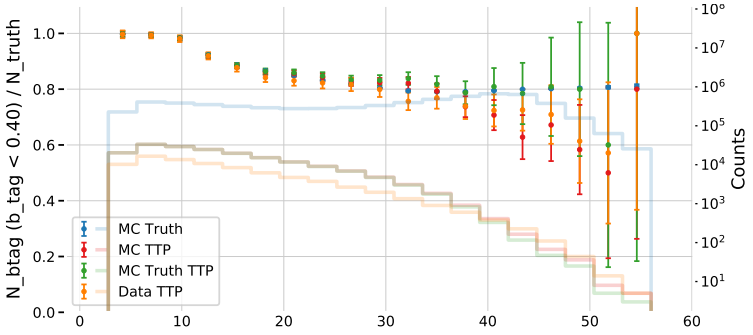

Figure 6.16: Efficiency of the b-tags for g-jets in the g-signal region for 3-jet events, $\varepsilon_g^{g-\text{sig}}$, as a function of jet energy `Ejet`. The g-signal region is defined as $\beta < 0.4$. In the plot the efficiencies are shown for MC Truth in blue, MC TTP in red, MC Truth TTP in green, and Data TTP in orange. The efficiencies (the errorbars) can be read off on the left y-axis and the counts (histograms) on the right y-axis. The abbreviation TTP is short for "Tag, Tag, Probe" where two jets in a event are used as tags and the probe is then used for further analysis. Notice how both MC TTP and Data TTP follow each other closely.

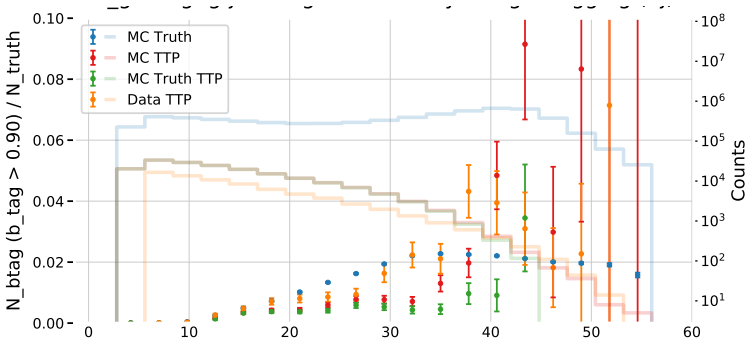

Figure 6.17: Efficiency of the b-tags for g-jets in the b-signal region for 3-jet events, $\varepsilon_g^{b-\text{sig}}$, as a function of jet energy `Ejet`. The b-signal region is defined as $\beta > 0.9$. In the plot the efficiencies are shown for MC Truth in blue, MC TTP in red, MC Truth TTP in green, and Data TTP in orange. The efficiencies (the errorbars) can be read off on the left y-axis and the counts (histograms) on the right y-axis. The abbreviation TTP is short for "Tag, Tag, Probe" where two jets in a event are used as tags and the probe is then used for further analysis.

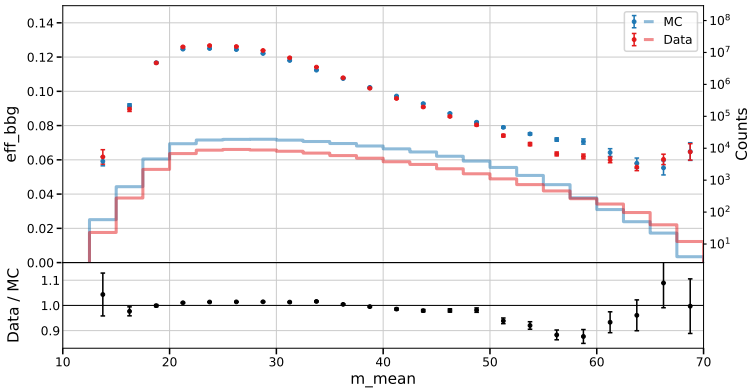

Figure 6.18: Proxy efficiency of the g-tags for $b\bar{b}g$ 3-jet events as a function of the mean of the two invariant masses m_{bg} and $m_{\bar{b}g}$. The proxy efficiency $\varepsilon_{b\bar{b}g}$ is measured by finding $b\bar{b}g$-events where $\beta_b > 0.9$, $\beta_{\bar{b}} > 0.9$, and $\beta_g < 0.4$. and then calculating $\varepsilon_{b\bar{b}g} = \varepsilon_b^{b-\text{sig}} \cdot \varepsilon_{\bar{b}}^{b-\text{sig}} \cdot \varepsilon_g^{g-\text{sig}}$. In the top plot $\varepsilon_{b\bar{b}g}$ is shown for MC in blue and Data in red where the counts in each bin can be read on right y-axis. In the bottom plot the ratio between Data and MC is shown.

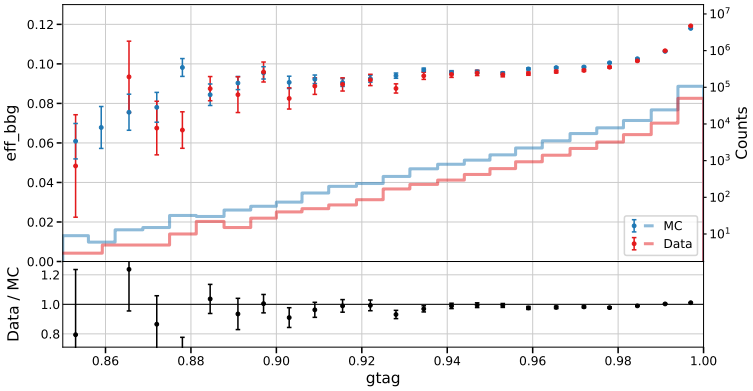

Figure 6.19: Proxy efficiency of the g-tags for $b\bar{b}g$ 3-jet events as a function of the event's g-tag. The proxy efficiency $\varepsilon_{b\bar{b}g}$ is measured by finding $b\bar{b}g$-events where $\beta_b > 0.9$, $\beta_{\bar{b}} > 0.9$, and $\beta_g < 0.4$. and then calculating $\varepsilon_{b\bar{b}g} = \varepsilon_b^{b-\text{sig}} \cdot \varepsilon_{\bar{b}}^{b-\text{sig}} \cdot \varepsilon_g^{g-\text{sig}}$. In the top plot $\varepsilon_{b\bar{b}g}$ is shown for MC in blue and Data in red where the counts in each bin can be read on right y-axis. In the bottom plot the ratio between Data and MC is shown.

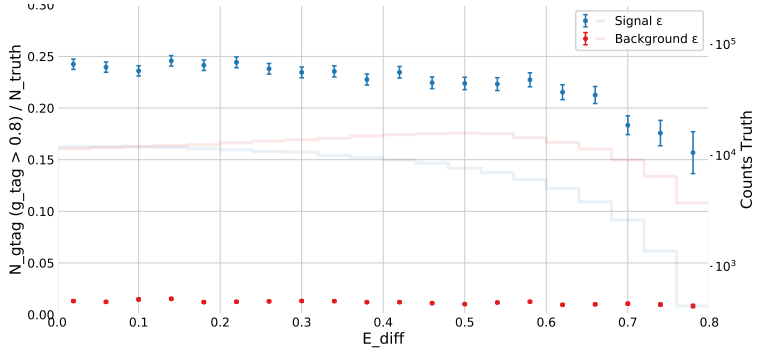

Figure 6.20: Efficiency of the g-tags for 4-jet events as a function of normalized gluon gluon jet energy difference in Monte Carlo. The efficiency is measured as the number of events with a g-tag higher than 0.8 ($\gamma > 0.8$) out of the total number and the normalized gluon gluon jet energy difference A is $A = \frac{E_{g_{\max}} - E_{g_{\min}}}{E_{g_{\max}} + E_{g_{\min}}}$ where $E_{g_{\max}}$ ($E_{g_{\min}}$) refers to the energy of the gluon with the highest (lowest) energy. The efficiency is plotted for signal events according to MC Truth in blue and background events according to MC Truth in red.

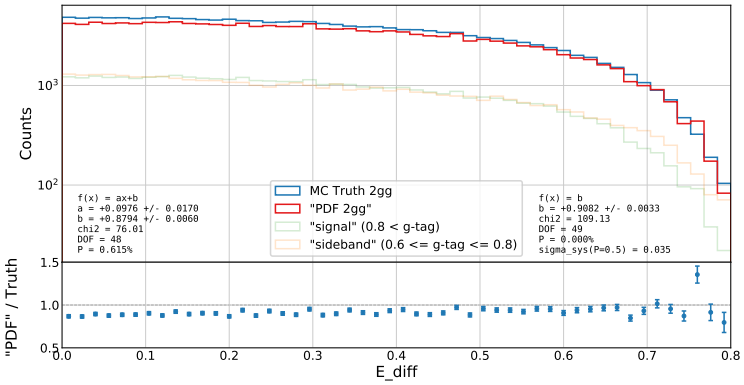

Figure 6.21: Closure plot between MC Truth and the corrected g-tagging model in 4-jet events for the normalized gluon gluon jet energy difference. The corrected g-taggingg model is described in further detail in section XXX **TODO!**. In the top part of the plot the MC Truth is shown in blue, the corrected g-tagging model `"PDF 2gg"` in red, the g-signal distribution in semi-transparent green and the g-sideband distribution in semi-transparent orange. In the bottom part of the plot the ratio between MC Truth and the output of the corrected g-tagging model is shown. The normalized gluon gluon jet energy difference A is $A = \frac{E_{g_{\max}} - E_{g_{\min}}}{E_{g_{\max}} + E_{g_{\min}}}$ where $E_{g_{\max}}$ ($E_{g_{\min}}$) refers to the energy of the gluon with the highest (lowest) energy.

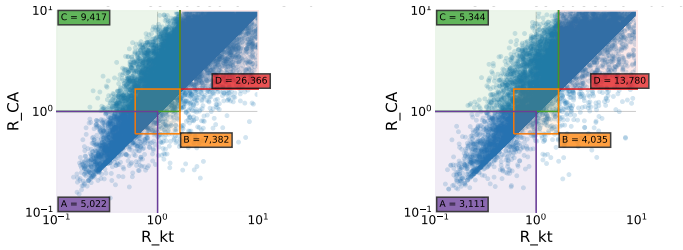

Figure 6.22: R kt CA overview XXX **TODO!**

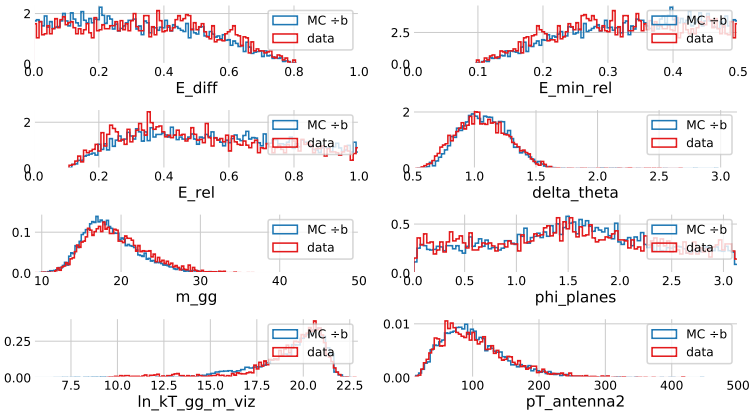

Figure 6.23: R kt CA cut region A XXX **TODO!**

B. Quarks vs. Gluons Appendix

	b	*c*	*g*	non-*q*-matched	*uds*
2	37.2 %	12.9 %	0.0 %	20.7 %	29.1 %
3	22.6 %	8.9 %	31.2 %	17.5 %	19.7 %
4	14.6 %	7.0 %	45.1 %	18.3 %	15.0 %
5	10.0 %	5.7 %	52.5 %	19.6 %	12.2 %
6	7.1 %	4.4 %	54.4 %	25.2 %	8.8 %

Table B.1: Number of jets for MC and MCb.

Bibliography

[1] Advanced Topics in Machine Learning (ATML). URL https://kurser.ku.dk/course/ndak15014u.

[2] Allstate Claims Severity - Fair Loss. URL https://kaggle.com/c/allstate-claims-severity.

[3] Dmlc/xgboost. URL https://github.com/dmlc/xgboost.

[4] HEP meets ML award | The Higgs Machine Learning Challenge. URL https://higgsml.lal.in2p3.fr/prizes-and-award/award/.

[5] The Large Electron-Positron Collider | CERN. URL https://home.cern/science/accelerators/large-electron-positron-collider.

[6] Scikit-hep/uproot. URL https://github.com/scikit-hep/uproot.

[7] Datashader: Revealing the Structure of Genuinely Big Data. URL https://github.com/holoviz/datashader.

[8] O. . Www.OIS.dk - Din genvej til ejendomsdata. URL https://www.ois.dk/.

[9] Y. S. Abu-Mostafa, M. Magdon-Ismail, and H.-T. Lin. *Learning From Data*. AMLBook. ISBN 978-1-60049-006-4.

[10] D. Albanese, S. Riccadonna, C. Donati, and P. Franceschi. A practical tool for maximal information coefficient analysis. 7. ISSN 2047-217X. doi: 10.1093/gigascience/giy032. URL https://doi.org/10.1093/gigascience/giy032.

[11] E. Anderson. The Species Problem in Iris. 23(3):457–509. ISSN 00266493. doi: 10.2307/2394164. URL www.jstor.org/stable/2394164.

[12] B. Andersson, G. Gustafson, G. Ingelman, and T. Sjöstrand. Parton fragmentation and string dynamics. 97(2):31–145. ISSN 0370-1573. doi: 10.1016/0370-1573(83)90080-7. URL http://www.sciencedirect.com/science/article/pii/0370157383900807.

[13] S. R. Armstrong. A Search for the standard model Higgs boson in four jet final states at center-of-mass energies near 183-GeV with the ALEPH detector at LEP. URL http://wwwlib.umi.com/dissertations/fullcit?p9910371.

[14] M. Awad and R. Khanna. Support Vector Regression. In M. Awad and R. Khanna, editors, *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*, pages 67–80. Apress. ISBN 978-1-4302-5990-9. doi: 10.1007/978-1-4302-5990-9_4. URL https://doi.org/10.1007/978-1-4302-5990-9_4.

[15] R. J. Barlow. *Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences (Manchester Physics Series)*. WileyBlackwell, reprint edition. ISBN 0-471-92295-1. URL http://www.amazon.co.uk/Statistics-Statistical-Physical-Sciences-Manchester/dp/0471922951%3FSubscriptionId%3D13CT5CVB80YFWJEPWS02%26tag%3Dws%26linkCode%3Dxm2%26camp%3D2025%26creative%3D165953%26creativeASIN%3D0471922951.

[16] J. T. Barron. A General and Adaptive Robust Loss Function. URL http://arxiv.org/abs/1701.03077.

[17] W. Bartel et al. Experimental study of jets in electron-positron annihilation. 101(1):129–134. ISSN 0370-2693. doi: 10.1016/0370-2693(81)90505-0. URL http://www.sciencedirect.com/science/article/pii/0370269381905050.

[18] J. Bergstra and Y. Bengio. Random Search for Hyper-parameter Optimization. 13:281–305. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=2188385.2188395.

[19] C. Bierlich. Rope hadronization, geometry and particle production in pp and pA Collisions. URL https://lup.lub.lu.se/search/ws/files/18474576/thesis.pdf.

[20] Bolighed. Bolighed - usikkerhed i data-vurderingen. URL https://bolighed.dk/om-bolighed/spoergsmaal-og-svar/#boligvaerdi.

[21] L. Breiman. Random Forests. 45(1):5–32. ISSN 1573-0565. doi: 10.1023/A:1010933404324. URL https://doi.org/10.1023/A:1010933404324.

[22] E. Brochu, V. M. Cora, and N. de Freitas. A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning. URL http://arxiv.org/abs/1012.2599.

[23] R. Brun and F. Rademakers. ROOT — An object oriented data analysis framework. 389(1):81–86. ISSN 0168-9002. doi: 10.1016/S0168-9002(97)00048-X. URL

http://www.sciencedirect.com/science/article/pii/S016890029700048X.

[24] A. Buckley, J. Butterworth, S. Gieseke, D. Grellscheid, S. Hoche, H. Hoeth, F. Krauss, L. Lonnblad, E. Nurse, P. Richardson, S. Schumann, M. H. Seymour, T. Sjostrand, P. Skands, and B. Webber. General-purpose event generators for LHC physics. 504(5):145–233. ISSN 03701573. doi: 10.1016/j.physrep.2011.03.005. URL http://arxiv.org/abs/1101.2599.

[25] C. Burgard. Standard model of physics | TikZ example. URL http://www.texample.net/tikz/examples/model-physics/.

[26] D. Buskulic et al. An investigation of Bdo and Bso oscillation. 322(4):441–458. ISSN 0370-2693. doi: 10.1016/0370-2693(94)91177-0. URL http://www.sciencedirect.com/science/article/pii/0370269394911770.

[27] T. Chen and C. Guestrin. XGBoost: A Scalable Tree Boosting System. pages 785–794. doi: 10.1145/2939672.2939785. URL http://arxiv.org/abs/1603.02754.

[28] T. A. Collaboration. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. 716(1):1–29. ISSN 03702693. doi: 10.1016/j.physletb.2012.08.020. URL http://arxiv.org/abs/1207.7214.

[29] S. D. DST. Price Index (EJ14) - Statistics Denmark. URL https://www.dst.dk/en/Statistik/emner/priser-og-forbrug/ejendomme.

[30] D. et al. Buskulic. A precise measurement of hadrons. 313(3):535–548. ISSN 0370-2693. doi: 10.1016/0370-2693(93)90028-G. URL http://www.sciencedirect.com/science/article/pii/037026939390028G.

[31] R. A. Fisher. The Use of Multiple Measurements in Taxonomic Problems. 7(2):179–188. ISSN 2050-1439. doi: 10.1111/j.1469-1809.1936.tb02137.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-1809.1936.tb02137.x.

[32] Y. Freund and R. E. Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In P. Vitányi, editor, *Computational Learning Theory*, pages 23–37. Springer Berlin Heidelberg. ISBN 978-3-540-49195-8. AdaBoost.

[33] S. L. Glashow. Partial-symmetries of weak interactions. 22(4):579–588. ISSN 0029-5582. doi: 10.1016/0029-5582(61)90469-2. URL http://www.sciencedirect.com/science/article/pii/0029558261904692.

[34] A. E. Harvey and S. Peters. Estimation Procedures for Structural Time Series Models. doi: 10.1002/for.3980090203.

[35] T. Hastie and R. Tibshirani. Generalized Additive Models: Some Applications. 82(398):371–386. ISSN 01621459. doi: 10.2307/2289439. URL www.jstor.org/stable/2289439.

[36] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics. Springer-Verlag, 2 edition. ISBN 978-0-387-84857-0. URL //www.springer.com/la/book/9780387848570.

[37] P. Huber and E. Ronchetti. *Robust Statistics*. Wiley Series in Probability and Statistics. Wiley. ISBN 978-1-118-21033-8. URL https://books.google.dk/books?id=j1OhquR_j88C.

[38] S. Hviid, Juul. Working Paper: A regional model of the Danish housing market. URL http://www.nationalbanken.dk/en/publications/Pages/2017/11/Working-Paper-A-regional-model-of-the-Danish-housing-market.aspx.

[39] F. James and M. Roos. Minuit – a system for function minimization and analysis of the parameter errors and correlations. 10:343–367. doi: 10.1016/0010-4655(75)90039-9.

[40] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 3146–3154. Curran Associates, Inc. URL http://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree.pdf.

[41] C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. 49(4): 764–766. ISSN 0022-1031. doi: 10.1016/j.jesp.2013.03.013. URL http://www.sciencedirect.com/science/article/pii/S0022103113000668.

[42] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 4768–4777. Curran Associates Inc. ISBN 978-1-5108-6096-4. URL http://dl.acm.org/citation.cfm?id=3295222.3295230.

[43] S. M. Lundberg, G. G. Erion, and S.-I. Lee. Consistent Individualized Feature Attribution for Tree Ensembles - SHAP. . URL http://arxiv.org/abs/1802.03888.

[44] S. M. Lundberg, G. G. Erion, and S.-I. Lee. Consistent Individualized Feature Attribution for Tree Ensembles. . URL http://arxiv.org/abs/1802.03888.

[45] I. Mulalic, H. Rasmussen, J. Rouwendal, and H. H. Woltmann. The Financial Crisis and Diverging House Prices: Evidence from the Copenhagen Metropolitan Area. ISSN 1556-5068. doi: 10.2139/ssrn.3041272. URL https://www.ssrn.com/abstract=3041272.

[46] Particle Data Group et al. Review of Particle Physics. 98 (3):030001. doi: 10.1103/PhysRevD.98.030001. URL https://link.aps.org/doi/10.1103/PhysRevD.98.030001.

[47] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. 12:2825–2830.

[48] E. Polley and M. van der Laan. Super Learner In Prediction. URL https://biostats.bepress.com/ucbbiostat/paper266.

[49] A. Purcell. Go on a particle quest at the first CERN webfest. URL https://cds.cern.ch/record/1473657.

[50] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti. Detecting Novel Associations in Large Data Sets. 334(6062):1518–1524. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1205438. URL http://science.sciencemag.org/content/334/6062/1518.

[51] P. J. Rousseeuw and C. Croux. Alternatives to the Median Absolute Deviation. 88(424):1273–1283. ISSN 0162-1459. doi: 10.1080/01621459.1993.10476408. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1993.10476408.

[52] A. Salam. Weak and electromagnetic interactions. In *Selected Papers of Abdus Salam*, volume Volume 5 of *World Scientific Series in 20th Century Physics*, pages 244–254. WORLD SCIENTIFIC. ISBN 978-981-02-1662-7. doi: 10.1142/9789812795915_0034. URL https://www.worldscientific.com/doi/abs/10.1142/9789812795915_0034.

[53] L. Shapley. A value for n-person games. In *The Shapley Value*, volume 28 of *Annals of Math Studies*, pages 307–317. doi: 10.1017/CBO9780511528446.003.

[54] T. Sjöstrand, S. Ask, J. R. Christiansen, R. Corke, N. Desai, P. Ilten, S. Mrenna, S. Prestel, C. O. Rasmussen, and P. Z. Skands. An Introduction to PYTHIA 8.2. 191:159–177. ISSN 00104655. doi: 10.1016/j.cpc.2015.01.024. URL http://arxiv.org/abs/1410.3012.

[55] S. J. Taylor and B. Letham. Forecasting at scale. doi: 10.7287/peerj.preprints.3190v2. URL https://peerj.com/preprints/3190.

[56] i. team. Iminuit – A python interface to minuit. URL https://github.com/scikit-hep/iminuit.

[57] R. Tibshirani. Regression Shrinkage and Selection via the Lasso. 58(1):267–288. ISSN 0035-9246. URL www.jstor.org/stable/2346178.

[58] A. Tikhonov. *On the Stability of Inverse Problems*, volume vol. 39 of *Doklady Akademii Nauk SSSR*.

[59] d. L. M. J. van, E. C. Polley, and A. E. Hubbard. Super Learner. 6(1). ISSN 1544-6115. doi: 10.2202/1544-6115.1309. URL https://www.degruyter.com/view/j/sagmb.2007.6.issue-1/sagmb.2007.6.1.1309/sagmb.2007.6.1.1309.xml.

[60] F. van Veen. The Neural Network Zoo. URL http://www.asimovinstitute.org/neural-network-zoo/.

[61] V. Vapnik. Principles of Risk Minimization for Learning Theory. In *Proceedings of the 4th International Conference on Neural Information Processing Systems*, NIPS'91, pages 831–838. Morgan Kaufmann Publishers Inc. ISBN 978-1-55860-222-9. URL http://dl.acm.org/citation.cfm?id=2986916.2987018.

[62] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. Jarrod Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, I. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and S. . . Contributors. SciPy 1.0–Fundamental Algorithms for Scientific Computing in Python. page arXiv:1907.10121. URL https://ui.adsabs.harvard.edu/abs/2019arXiv190710121V/abstract.

[63] S. Weinberg. A Model of Leptons. 19(21):1264–1266. doi: 10.1103/PhysRevLett.19.1264. URL https://link.aps.org/doi/10.1103/PhysRevLett.19.1264.

[64] H. Wickham. Tidy data. 59(10):1–23. ISSN 1548-7660. doi: 10.18637/jss.v059.i10. URL https://www.jstatsoft.org/v059/i10.

Index