Model evaluation TLDR explainer

### Ahmed etal 2019 Forecasting GRACE

* Root mean square error (RMSE)
* scaled RMSE (R\*)
* correlation coefficient (r) [“cor”, aka COR in metrics csv], Correlation: correlation of full model prediction (whole thing, all data used to generate model) predicted back to the full dataset: sensitive to the differences between modeled and observed data including the extreme values (i.e., outliers) [116].
* Nash–Sutcliffe efficiency (NSE) coefficient: best objective function for evaluating the overall fit between the predictive and observed values [115]
* seasonal adjusted NSE (NSE\*)

Binomial only or does Gaus/Pois also? RMSE & r/COR do non bin. Looks like NSE KGE KGE’ do continuous.

### Servat & Dezetter 2009 selection

* Crec
* CrecBi
* Fortin
* Nash / Nash-Sutcliffe Efficiency NSE. Only the Nash function can be linked directly to a statistical measure, viz. the percentage of residual variance compared to the total variance observed
* SExpER

Which won? NSE.

Improved upon by the Kling Gupta: <https://en.wikipedia.org/wiki/Kling%E2%80%93Gupta_efficiency>

Which is slightly improved upon by KGE’ <https://www.sciencedirect.com/science/article/pii/S0022169412000431?via%3Dihub>

But are KGE’, KGE, and NSE, all RESTRICTED to hydrological data, or are they simply FROM that world?

Elith et al 2008:

* CV correlation: correlation of training set (4/5ths) model predictions, against testing set (1/5ths) actual data, aver-aged across the number of reshuffles (default k=5, 5 shuffles. IIRC. Might be k=10)

Binomial evaluation metrics

* TSS: True Skill Statistic/Total Skill Score, Elizabeth Becker: “TSS is a threshold independence measure meaning you have to set a threshold for what constitutes a presence vs. absence. she used the sensitivity-sum maximization approach described by Liu, C., Berry, P.M., Dawson, T.P., & Pearson, R.G. (2005). She also provided her R code (in [assumedly Bonnie’s] analysis notes folder).” Becker et al. 2020 [ask Bonnie. Also do I know Elizabeth Becker?]
* Sensitivity
* Specificity
* Accuracy
* Precision: X% of the predictions are right
* Recall: Y% of actually existing things are captured
* OverallAccuracy
* BalancedAccuracy
* F1score : P & R equally rated, score importance depends on project
* F2score: weighted average of P & R
* Threshold: Threshold which produced best combo of TPR & TNR
* Prevalence: Prevalence
* ODP: Overall diagnostic power
* CCR: Correct classification rate
* TPR: True positive rate
* TNR: True negative rate
* FPR: False positive rate
* FNR: False negative rate
* PPP: Positive predictive power
* NPP: Negative predictive power
* MCR: Misclassification rate
* OR: Odds-ratio
* Kappa: Cohen's kappa
* Dev: deviance from 2 vecs, obs & pred vals. What does DEVIANCE actually MEAN though? Deviance explained, variance accounted for… what is it actually SAYING about the model performance? Is seems to exist outside this list, above it all somehow. Why does “how much did the model explain?” not equal “how well did the model perform?”?.
  + And if it really does live outside/above this list, is Deviance Explained the caveat which contextualises the result score? If the (possibly cleverly combined delta-to-one) Chosen Test Metric score is, 0.8, say, but deviance explained is only 0.5 / 50%, is the Deviance-Adjusted Chosen Test Metric score therefore 0.4?
    - And if so, what does THAT mean in context? 0.4 sounds bad. 0.8 was supposedly great. Can 0.4 be used in production?
  + Is Deviance Explained THE caveat, *or A caveat*? Are there other uber/caveat/context metrics?
  + Presumably Deviance Explained’s contextual performance isn’t linear but asymptotic. Systems tend to resemble themselves, the world is quite predictably ordered such that sampling captures most of the whole quickly, i.e. 80/20 i.e. asymptotic, AND THEREFORE a deviance explained score of 0.8 will be less than twice as good as a DE score of 0.4.
    - Presuming DE really is a metric which describes how much of the entire system has been captured. But it can’t be, not in a true sense. It only has a single vector of values to work from, it doesn’t know what it doesn’t know, therefore it can only tell you how the model prediction compares to the original data. So how can it be anything other than another performance metric? Look up DE, deviant from what, it has a deviance score before the model is run, which has to be a statistical description of a property the input data.

Bonnie notes

* Harrell’s C-index (concordance index) explained here: <https://statisticaloddsandends.wordpress.com/2019/10/26/what-is-harrells-c-index/> used by Sergi Perez Jorge, binomial, widely used in med: “C estimates the probability that, of two randomly chosen patients, the patient with the higher prognostic score will outlive the patient with the lower prognostic score. Values of c near 0.5 indicate that the prognostic score is no better than a coin-flip in determining which patient will live longer. (Harrell et al).
  + 0.5 = random, 1 = perfect, binary choice, sounds like AUC/TSS/frends.
  + But for binary data, presence-absence, do we actually WANT this metric which judges whether the comparison of continues value prediction scores is reliable. Does this essentially rate the performance of continuous data?
  + IS THIS WHAT I WANT?
  + See also RMSE, NSE though, there are others.

Questions to myself

* What about hurdle/delta models, i.e.: even if/when I deduce/decide which evaluation metric is best/I prefer for Gaus/nonbin, how/can I synthesise the two results? If bin is good & gaus is bad, what is the overall score/rating/evaluation?
* Presumably this changes if you change either of the component parts, i.e. TSS+NSE != AUC+RMSE?
* If so, does the whole industry have to choose and stick to it in order for results to be meaningful across studies?
* Does the quality of the predictive performance of the model hinge on the author’s choice of metrics and how well they understand and describe those results?
* Actually, these are two different things for two different audiences:
  + (The correctly chosen and applied) stats: the objective numerical truth/result.
  + The choice, presentation, and *description* of specific stats: contextualisation of the predictive performance as it will be applied to the real world; crucial for production (management (layer of company, or of a species, e.g.)).
* So therefore: pick the “best”/most apt, explain why you chose them (statistical skill e.g. unbiased, balanced; and contextual utility), then present the scores, explain what your score values mean (0.6 – 0.7 = good, lane et al 2009), then explain what that value means in the context of predicting to the real world, *including all known caveats*. If your model performed amazingly but you know it’s super data limited, you can’t naively/optimistically tell trusting readers that it will succeed in production/the real world.
* This is NOT a death sentence. This does NOT mean that we throw our papers up into the air and abandon all hope of ever being able to trust statistics. It JUST means that we have to, again,
  + Find/choose the one/few *best* tests for bin / nonbin (possibly per distribution)
  + Test, present results, explain contextual meaning including data, model performance, and performance evaluation metric caveats/limitations.

ToDo

* List bin & continuous metrics. This exercise.
* Can/should everything be cross validated?
  + Elith et al boiled AUC & some others into gbm.step, possible as sub functions; could this be extended to other metrics?
  + Should I stop thinking from a gbm.auto-centric POV and think tidymodels-style, and if so, what provision is there for CV within that framework?
  + Does tidymodels have some tidyMLeval function with a list of families? If so, that potentially populates the Item 1 list for me?
* Species-conservation-contextualised description of bin & continuous metrics, in isolation where appropriate (AUC vs TSS, e.g., where 2+ metrics have the same contextual meaning), and in concert where appropriate (F1, F2, TPR, TNR etc etc).
* Therefore: which metrics I’ve chosen, why, and how to interpret their results in the context of one’s data and the production environment/world where it will be applied.
* If this process remains an open question i.e. I don’t completely solve it on this flight, move this to a github issue OR discussion/forum post, are there such things? Do I have access to this feature on github basic for gbm.auto? Email BCC people and invite them to weigh in if they’re so minded
  + Bonnie
  + Chuck
  + (invite) Other OTN crew?
  + Colin Minto
  + Steph
  + NFF
  + Lauran
  + Trevor Hastie????
  + Liberty / Kristine / Courtney / Cat / Lauren / other helpees
  + John Froeschke
  + Dovi
  + Robin
  + Jane Elith???
  + Dan Crear
  + San Diego IATTC crew?
  + StackExchange, CrossValidated
* Ask Bonnie cc Dan Crear for the email chain with Elizabeth Becker NOAA.

Synthesis funnel concept:

First person does a deep dive on a subject, adds to the field, develops, etc.

Their ‘student’/mentee/etc synthesises their teachings from a learner’s perspective, writes/improves documentation.

The next person improves, adds smaller elements, etc.