Model Algorithm

GLM

* Strengths
  + Can be used with path analysis
* Weaknesses
* Presence vs Abundance
* Path Analysis

GAM

* Strengths
* Weaknesses
* Presence vs Abundance
* Path Analysis

Machine learning

* Strengths
* Weaknesses
* Presence vs Abundance
* Path Analysis

Ensemble model

* Strengths
* Weaknesses
* Presence vs Abundance
* Path Analysis

-What are the costs of false positives and false negatives?

-Accuracy works best if false positives and false negatives have similar cost. If the cost of false positives and false negatives are very different, it’s better to look at both Precision and Recall (<https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/>)

*Alldredge and Griswold 2006 - Resource selection studies for categorical resource variables*

-Design 1: individuals not marked, population level habitat availability

-Design 2: individuals marked, population level habitat availability

-Design 3: individuals marked, individual level habitat availability

-analyses applicable to design 1: bonferroni adjusted confidence intervals, logistic regression, log-linear modelling (Poisson regression with all categorical variables?), discrete choice (between selected location and a set of randomly selected points).

-compositional analysis is NOT applicable. Only makes sense for individuals, not for populations.

-unit sum constraint – all add up to 1

-recommended Manley et al 2002 paper on resource selection functions

*Rammer and Seidl 2019 - Harnessing Deep Learning in Ecology*

-comparing GLM to neural network, they had similar accuracy (correct predictions / total observations), but glm had notably worse precision (higher false positive rate) and recall (how many true positives were correctly identified)

-ML is a “black box” – parameters can’t be interpreted intuitively, less good for inference. But less accurate representation of reality

-ML makes fewer assumptions. More accurate description of relationship between drivers and responses. Hierarchical model structure reflects ecological structure

*Breiman 2001 - Statistical Modelling: The two cultures*

-Data modelling (try to figure out the stochastic model that generated the data) vs algorithmic modelling (try to find an algorithm that predicts the data)

-Argues that weakness of data modelling is that we assume we can figure out a good parametric model. But if the model is wrong, we draw the wrong conclusions. Goodness of fit tests and residual tests lack power when there are more than 4 or 5 dimensions. Often predictive accuracy is not tested. Cross-validation recommended by several statisticians.

-Multiplicity of models – often there are several equally plausible explanations/models. This is a problem for both data models and algorithms, but maybe more of a problem for data models where it’s common to reduce the number of variables. Makes the model unstable. Can end up with different models depending on how you carry this out / slight differences in data. Bagging is one solution?

-for prediction, interpretability and prediction often conflict. “a model doesn’t have to be simple to provide reliable information”

-“Usually, the initial evaluation of which variablesare important is based on examining the absolutevalues of the coefficients of the variables in the logis-tic regression divided b ytheir standard deviations.Figure 1 is a plot of these values”

-RF doesn’t have the same problem with multicollinearity masking variable importance.

-ML allows you to discover clusters more easily

---related blog (<https://towardsdatascience.com/thoughts-on-the-two-cultures-of-statistical-modeling-72d75a9e06c2>)

---overfitting is an issue, but the solution isn’t to simplify, but to use more robust validation

---for validation, an accuracy based approach is less subjective than a weight based approach (??)

Other ML reading online:

-what is the difference between machine learning and deep learning?

-ML requires extensive training data – how much?

-difficult to interpret – well, how do you interpret them?

-data can’t be biased

-more exploratory than confirmatory

-https://towardsdatascience.com/the-limitations-of-machine-learning-a00e0c3040c6

-should I consider using GAMs?

Response variable: zero-inflated counts

Predictor variables:

temp – normal

Landcover – left-skewed (zero-inflated?), proportions

~~-import the data I created yesterday with the missing temperate data~~

-re-learn how to check logistic regressions

~~-compare plot location improvement from initial version~~

-~500 plots to 2500 plots

-stop and think out what I’m going to do and be more organized

-can I find any examples online?

-double check plot inclusion and exclusion criteria

-spatial autocorrelation?

-how will I check scaling?

-look at the actual PRISM habitat data to see if it’s useful. Does it align with NLCC in some way?

-do a pca on all the habitat proportions?

<https://damariszurell.github.io/SDM-Intro/>

Also, the model objective will affect modelling decisions. We can distinguish three main objectives for SDMs: (a) inference and explanation, (b) mapping and interpolation, and (c) forecast and transfer.

-I want to do all three of these things in different parts of my project

-scaling predictors

-revisit Johnson levels of habitat selection