Assessing Feature Importance in Non-Linear Regression: An Ecology Perspective

Anthony Cannistra
Department of Biology
University of Washington
Seattle, WA 98105
tonycan@uw.edu

Abstract

The development of robust and accurate methods to predict the impact of anthropogenic global change on ecosystem function is a pressing challenge for quantitative ecologists. A delicate balance must be struck between models with high predictive performance and those with identifiable mechanism and foundation in theoretical ecology. Linear regression, a tried-and-true tool for ecologists to quantify observed ecological relationships, is an excellent experimental tool but a poor instrument for prediction. However, methods which can capture nonlinearities in ecological relationships, like kernel regression, support vector machines, and random forests, are opaque in their learned function and thus their predictions cannot be as easily verified by a skeptical ecologist. We use a tool known as a Shapley value to evaluate learned feature importance across several nonlinear methods and compare their top-ranked features to theoretical ecology results. We find that nonlinear methods in fact both have increased predictive performance and maintain afoundation in theoretical ecology.

1 Introduction

A perennial and pressing challenge in the field of ecology is the need to employ rigorously-derived scientific knowledge in service of the conservation of actual ecosystems and their properties. Increasingly it is being considered insufficient for ecologists to perform only basic science with outputs limited to those intended only for other ecologists—the burden of protecting biodiversity and ecosystem health in the face of anthropogenic global change is simply too much to allow for this narrow scope of expertise [Lubchenco, 1995, 2017]. Ecologists and other environmental scientists increasingly must bear the burden of using their expertise to inform policy and management decisions in service of life-sustaining systems.

This shift in ideology has a corresponding shift in experimental methodology. Though basic ecological exploration will always have a place in the field, ecologists must now strive to develop experimental methods which enable *prediction* as a core part of their outcome [Houlahan et al., 2017]. In this work we examine the performance of a particular methodological shift which utilizes existing methodology to enhance predictive performance of existing data.

Linear regression has traditionally been the tool of choice for ecologists in exploration of basic questions by providing statistically rigorous and easily interpretable estimates of effect size in analysis of a given ecological relationship. However, linear regression is insufficient for ecological predictions because many natural ecological relationships modeled by linear regression have significant nonlinearities as a result of unmodeled interactions and the modeled processes themselves. As a result, thresholds for acceptable goodness-of-fit measures in an ecological context are in the

CSE 546: Machine Learning. Autumn 2017 (Jaimeson). University of Washington, Seattle.

$R^2 = 0.14 - 0.30$ range on the low end, which is often enough to lend statistical significance to a result but almost always insufficient to allow for any kind of usable prediction.

The employment of regression approaches which can model these nonlinearities has been limited as a result of their relative lack of interpretability. The task of building predictive experimental methods in ecology requires a delicate balance between enhanced predictive performance and ensuring biological rigor. For example, a model which performs extraordinarily well in a purely predictive context (low error, high classification accuracy, etc.) but does not have any verifiable way to ensure that actual ecological theory is being leveraged in prediction will not be accepted by the community, let alone used to inform real-world policy and management practice.

In this work we examine whether a handful of performant but difficult-to-interpret nonlinear regression techniques can be trusted from a biological standpoint. We replicate results from a study [Angert et al., 2011] examining whether species' traits can explain observed shifts in those species' geographic ranges over the past century. We utilize two widely-studied datasets covering Swiss alpine plants [Holzinger et al., 2008] and Yosemite Valley small mammals [Moritz et al., 2008] in an effort to compare OLS regression coefficients to metrics of feature importance produced by alternative nonlinear regression methods with increased predictive performance. Using these coefficients and feature importance values we will evaluate whether models with increased predictive performance can be shown to still represent established biological theory in their predictions and thereby maintain their credibility and practical trustworthiness.

2 Methods and Data

Angert et al. [2011] used least-squares linear regression to evaluate whether species' traits can explain shifts of geographic range with climate change. The formulation of this question is motivated by the desire to create a predictive framework for understanding how species' traits might influence their future response to climate change. This work lays a solid foundation by identifying certain relevant traits across taxa, but the methods therein do not permit these kinds of actual predictions. We use this OLS framework as the baseline predictor to which we will compare additional nonlinear methods and their influential predictor variables.

We peform the below analyses on both a Swiss alpine plants' dataset [Holzinger et al., 2008] and a Yosemite small mammals dataset [Moritz et al., 2008]. The Swiss plants dataset contains $N = 139$ species in the original dataset. We remove samples which are missing data, one-hot encode categorical features for regression, and normalize/center the numeric features to have zero mean and unit norm, after which $N = 20, d = 43$. The Yosemite mammals dataset contains $N = 28$ species. After the same processing as above, we have $N = 20, d = 26$.

2.1 Non-Linear Regression Methods

Let $\hat{W}$ be the "canonical" coefficients learned by ordinary least squares regression. We train a $l2$-regularized linear (Ridge) regressor of the form

$$\hat{w} = \min_{w} ||Xw - y||_2^2 + \alpha||w||_2^2$$

and extract learned coefficients in $\hat{w}$ for comparison to $\hat{W}$. We also train the same regressor using an RBF kernel matrix $\mathbf{K}$:

$$\hat{\alpha} = \min_{\alpha} ||\mathbf{y} - \mathbf{K}\alpha||_2^2 + \lambda\alpha^T\mathbf{K}\alpha$$

$$\hat{w} = \sum_{i}^{N} \alpha_i x_i$$

As a result of the kernel trick we no longer have explicit features to extract coefficients for (hence the interpretability challenge referenced in Section 1). We address this challenge by calculating Shapley additive feature importance values, detailed in the next section.

We also train a support vector regressor with an RBF kernel, using grid search for kernel and SVM hyperparameters (ϵ) and a Random Forest of decision trees. For the SVM we extract Shapley additive feature importance values as above, and we use reported feature importance values (Gini scores) from the random forest training process as our feature importance scores.

2.2 Evaluation Metrics

To reiterate, a fundamental challenge in nonlinear regression methods in ecology is their lack of interpretability as a result of there being no easily-accessible feature importance values to explain the model's prediction basis. To address this, we utilize the Shapley additive feature value method, proposed in Lundberg and Lee [2017]. Shapley values are computed by treating the *explanation* of a given model's prediction as a model in and of itself—values are computed by training an additive method derived from cooperative game theory to learn each feature's contribution to a model's prediction. Explanations are generated from each prediction in the model training set to identify the most important learned features during training, and are averaged for each feature across all training examples. to generate a whole-model feature importance scale[1].

The compare all of the learning techniques, we use either regression coefficients (for OLS and Ridge regression), Shapley feature importance values (Kernel Ridge and SVM), or Gini feature importance values (Random Forest) to rank all features $\{a_i : i = 0 \rightarrow d\}$ such that each feature has an importance ranking for each of the several regression metrics.

In the next section we report results for all of the regression methods on both datasets. We also compute mean squared error for each of the regression techniques using a randomly-selected test set not included in the training data to facilitate performance comparison across methods.

3 Results

3.1 Swiss Alpine Plants

As expected, these various methods produced an improvement in predictive performance over ordinary least squares on this dataset, as seen in Figure 2).

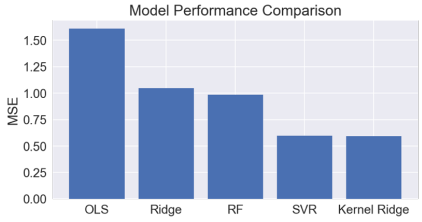

Figure 1: Mean Squared Error for various regression mechanisms on Plant Dataset.

In Figure 2 we present learned coefficients for the OLS regression, which represents our "canonical" feature rank ordering. Similar to results presented in Angert et al. [2011], a trait relevant to seed shed duration (in this case, `earliest_seed_shed_mo`) is the most influential coefficient as determined by its magnitude.

We show learned coefficients, Shapley Values, and Gini scores, respectively, for Ridge Regression, Kernel Regression, SVM, and Random Forest in Figure 3. Again, we notice a similar trend to the previous analysis using linear methods: these approaches all identify the temporal dimension of seed shedding as the most important predictive variable.

These results for the most part indicate preservation of theoretically-supported biology in these nonlinear methods with higher predictive accuracy. To illustrate this point we provide the rankings of all features in Figure 4. On average, seed shed duration-related trait `earliest_seed_shed_mo` is the highest-ranked feature, with `latest_seed_shed_mo` not far behind.

3.2 Mammals Data

Similarly to the analysis of the plants data, we observe a significant increase in predictive performance with the nonlinear prediction methods (Figure 5). The magnitude of difference between the OLS

[1]Shapley values are not intended to be aggregated—they represent a feature's contribution to a given single prediction, but their aggregation is a reasonable choice for comparing model performance [Lundberg and Lee, 2017]

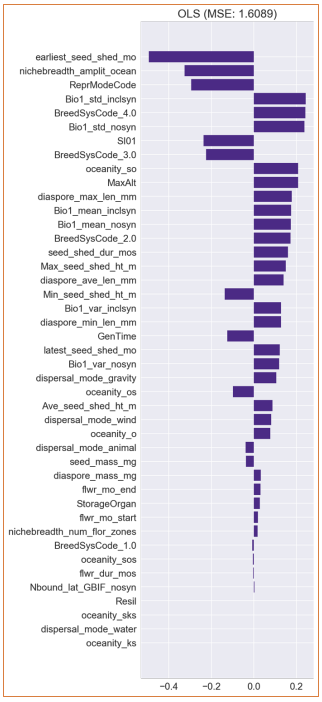

Figure 2: OLS Coefficients for Plant Data.

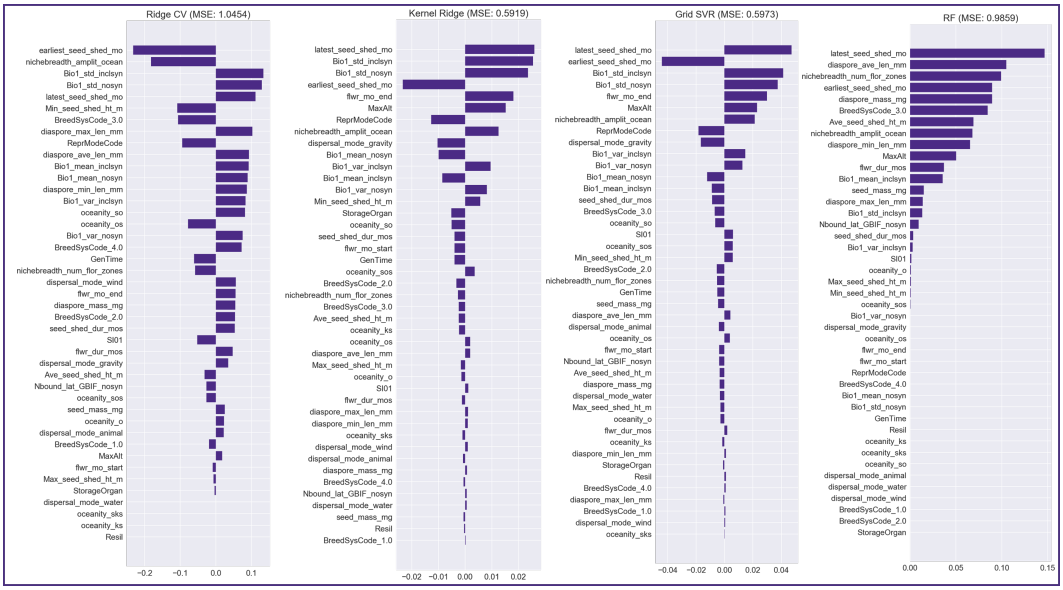

Figure 3: Feature importance values across nonlinear methods on plants dataset. Seed shed-related features (`earliest_seed_shed_mo` and `latest_seed_shed_mo`) have the highest rank across almost all methods.

.

error and the next lowest error is considerably greater in this case: about 37.08 versus 0.56 difference in MSE between OLS and the next lowest error for the mammals data and the plant data, respectively (Figures 5 and 1).

	OLS	Ridge	Kernel Ridge	RF	SVR	mean
earliest_seed_shed_mo	0.0	0.0	3.0	3.0	1.0	1.4
Bio1_std_inclsyn	3.0	2.0	1.0	14.0	2.0	4.4
nichebreadth_amplit_ocean	1.0	1.0	7.0	7.0	6.0	4.4
latest_seed_shed_mo	21.0	4.0	0.0	0.0	0.0	5.0
Bio1_std_nosyn	5.0	3.0	2.0	31.0	3.0	8.8
ReprModeCode	2.0	8.0	6.0	28.0	7.0	10.2
BreedSysCode_3.0	7.0	6.0	22.0	5.0	14.0	10.8
Bio1_mean_inclsyn	11.0	10.0	11.0	11.0	12.0	11.0
MaxAlt	9.0	35.0	5.0	9.0	5.0	12.6
Bio1_var_inclsyn	18.0	13.0	10.0	17.0	9.0	13.4
Bio1_mean_nosyn	12.0	11.0	9.0	30.0	11.0	14.6
Min_seed_shed_ht_m	17.0	5.0	13.0	21.0	18.0	14.8
diaspore_ave_len_mm	16.0	9.0	26.0	1.0	23.0	15.0
seed_shed_dur_mos	14.0	24.0	16.0	16.0	13.0	16.6
Bio1_var_nosyn	22.0	16.0	12.0	23.0	10.0	16.6
flwr_mo_end	31.0	21.0	4.0	26.0	4.0	17.2
oceanity_so	8.0	14.0	15.0	36.0	15.0	17.6
dispersal_mode_gravity	23.0	27.0	8.0	24.0	8.0	18.0
SI01	6.0	25.0	29.0	18.0	16.0	18.8
nichebreadth_num_flor_zones	34.0	19.0	21.0	2.0	20.0	19.2
diaspore_max_len_mm	10.0	7.0	31.0	13.0	39.0	20.0
diaspore_min_len_mm	19.0	12.0	32.0	8.0	35.0	21.2
GenTime	20.0	18.0	18.0	32.0	21.0	21.8
Ave_seed_shed_ht_m	25.0	28.0	23.0	6.0	28.0	22.0
oceanity_os	24.0	15.0	25.0	25.0	25.0	22.8
BreedSysCode_2.0	13.0	23.0	20.0	41.0	19.0	23.2
diaspore_mass_mg	30.0	22.0	36.0	4.0	29.0	24.2
oceanity_sos	36.0	30.0	19.0	22.0	17.0	24.8
BreedSysCode_4.0	4.0	17.0	37.0	29.0	38.0	25.0
Max_seed_shed_ht_m	15.0	37.0	27.0	20.0	31.0	26.0
seed_mass_mg	29.0	31.0	40.0	12.0	22.0	26.8
flwr_dur_mos	37.0	26.0	30.0	10.0	33.0	27.2
oceanity_o	27.0	32.0	28.0	19.0	32.0	27.6
flwr_mo_start	33.0	36.0	17.0	27.0	26.0	27.8
Nbound_lat_GBIF_nosyn	38.0	29.0	38.0	15.0	27.0	29.4
dispersal_mode_animal	28.0	33.0	35.0	37.0	24.0	31.4
dispersal_mode_wind	26.0	20.0	34.0	39.0	41.0	32.0
StorageOrgan	32.0	38.0	14.0	42.0	36.0	32.4
oceanity_ks	42.0	41.0	24.0	34.0	34.0	35.0
dispersal_mode_water	41.0	39.0	39.0	38.0	30.0	37.4
oceanity_sks	40.0	40.0	33.0	35.0	42.0	38.0
BreedSysCode_1.0	35.0	34.0	42.0	40.0	40.0	38.2
Resil	39.0	42.0	41.0	33.0	37.0	38.4

Figure 4: Feature Ranks by regressor, sorted by average rank, for Plants data.

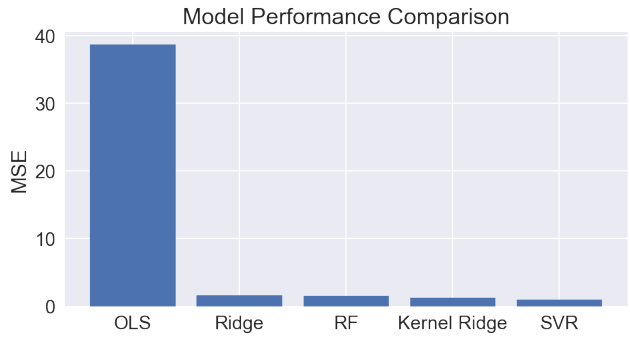

Figure 5: Mean Squared Error for various mechanisms on Mammals dataset.

In a departure from our expectation, our ordinary least squares regression coefficients do not correspond with results from the Angert et al. [2011] analysis. The features with the largest coefficients here are `Litter_size` and `Daily_rhythm_both` (Figure 6), which describe reproductive magnitude and daily temporal activity regime respectively. These features were not significant in Angert et al. [2011]—rather, we expected to see historical upper range limit (`Orig_high_limit`) as the feature with the highest performance. In this OLS regression, it was 5th.

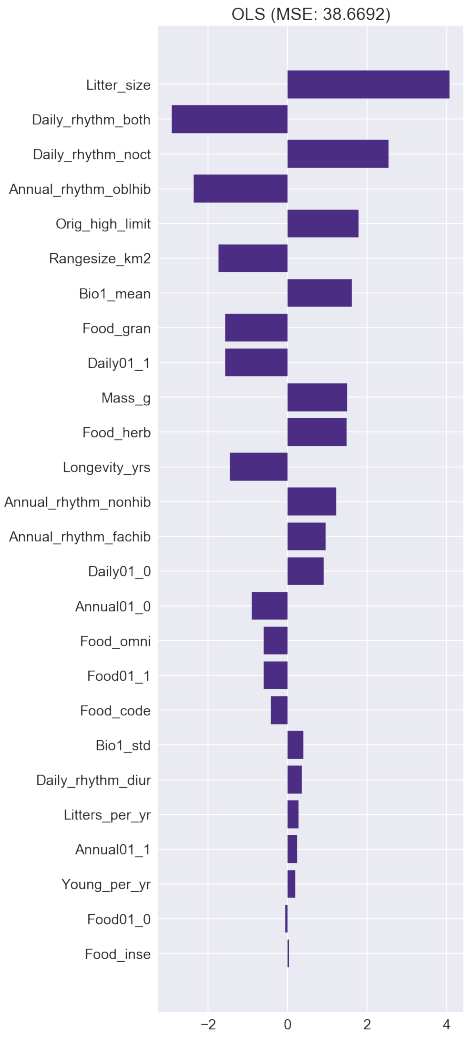

Figure 6: OLS Regression coefficients on Mammals data

We show learned coefficients, Shapley Values, and Gini scores, respectively, for Ridge Regression, Kernel Regression, SVM, and Random Forest in Figure 7. Here we observe that all other regression methods capture the expected result: that `orig_high_limit` is the most influential variable in these predictions. This is confirmed on average in Figure 8.

4 Discussion and Conclusion

These results lay the foundation for a case of "having our cake and eating it too"—we have shown that nonlinear regression methods on these two datasets not only achieve better predictive accuracy but also continue to maintain a theoretical foundation in biology in their predictions. This result has significant implications in the field of ecological forecasting as it supplements the existing literature's demonstration of the power of kernel-based and other nonlinear regression techniques with a preliminary suggestion that these models maintain a solid foundation in ecological theory.

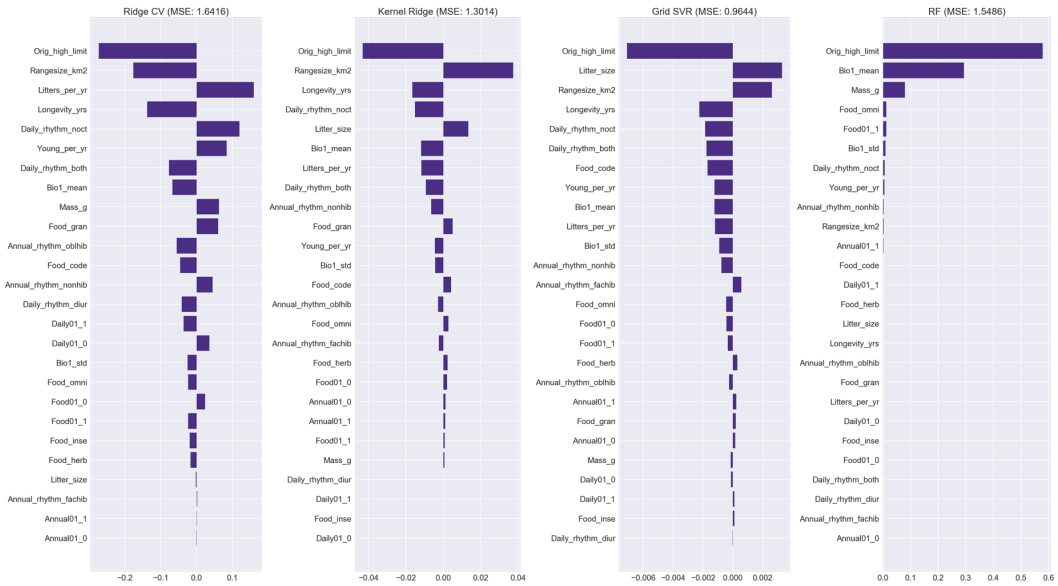

Figure 7: Feature importance values across nonlinear methods on mammals data.

	OLS	Ridge	Kernel Ridge	RF	SVR	mean
Orig_high_limit	4.0	0.0	0.0	0.0	0.0	0.8
Rangesize_km2	5.0	1.0	1.0	9.0	2.0	3.6
Daily_rhythm_noct	2.0	4.0	3.0	6.0	4.0	3.8
Bio1_mean	6.0	7.0	5.0	1.0	8.0	5.4
Longevity_yrs	11.0	3.0	2.0	15.0	3.0	6.8
Litter_size	0.0	22.0	4.0	14.0	1.0	8.2
Daily_rhythm_both	1.0	6.0	7.0	22.0	5.0	8.2
Annual_rhythm_nonhib	12.0	12.0	8.0	8.0	11.0	10.2
Young_per_yr	23.0	5.0	10.0	7.0	7.0	10.4
Litters_per_yr	21.0	2.0	6.0	18.0	9.0	11.2
Food_code	18.0	11.0	12.0	11.0	6.0	11.6
Annual_rhythm_oblhib	3.0	10.0	13.0	16.0	17.0	11.8
Mass_g	9.0	8.0	21.0	2.0	21.0	12.2
Bio1_std	19.0	16.0	11.0	5.0	10.0	12.2
Food_gran	7.0	9.0	9.0	17.0	19.0	12.2
Food_omni	16.0	17.0	14.0	3.0	13.0	12.6
Food01_1	17.0	19.0	20.0	4.0	15.0	15.0
Food_herb	10.0	21.0	16.0	13.0	16.0	15.2
Daily01_1	8.0	14.0	23.0	12.0	23.0	16.0
Annual_rhythm_fachib	13.0	23.0	15.0	24.0	12.0	17.4
Annual01_1	22.0	24.0	19.0	10.0	18.0	18.6
Food01_0	24.0	18.0	17.0	21.0	14.0	18.8
Daily01_0	14.0	15.0	25.0	19.0	22.0	19.0
Daily_rhythm_diur	20.0	13.0	22.0	23.0	25.0	20.6
Annual01_0	15.0	25.0	18.0	25.0	20.0	20.6
Food_inse	25.0	20.0	24.0	20.0	24.0	22.6

Figure 8: Feature Ranks by regressor, sorted by average rank, for mammals data.

There are several pieces of this analysis which could be bolstered in order to solidify this result. First and foremost the number of training and test samples in this analysis, in both datasets, is quite small. This induces significant bias when selecting training and test sets for validation. In addition, given

more data we would have been able to develop a bootstrap confidence interval on our coefficient values by sampling from the training set, but the very small nature of these data made this unrealistic.

There are also additional methods which could be used for the assessment of feature importance (and the stability of the observed results), such as the addition of random features to the data. The addition of random features has the potential to add robustness to this result by increasing the discrepancy between informative and noninformative features. In the earlier stages of this experiment we evaluated a sparse Lasso-regularized regressor, but almost all coefficients (except, usually, the Angert et al. [2011] result) were regularized to zero after cross-validation on the regularization strength. The convergence on such an extreme regularization is perhaps an artifact of either the chosen hyperparameter space or the small number of samples in the data. Either way, it is surprising that a quality result did not emerge from the Lasso, and this warrants further investigation.

Further, it remains to be discovered why in these examples we observe the *highest ranked* predictor to be in agreement with our expectation from Angert et al. [2011] but the subsequent predictors to be often quite disparate from the expectation. This observation is perhaps an artifact of a significant difference in methodology—for statistical significance reasons the original analysis used single-predictor linear regression to assess coefficient importance (i.e. the analysis only included 1-3 predictors in a single regressor to evaluate the coefficients, rather than using all predictors). We chose to use all predictors in our regressors as we were intending to emulate a predictive context wherein the *output* of the regressor was just as important as the learned coefficients. This difference (and the additive noise) may be a reason for this discrepancy.

Despite these shortcomings and open questions, this result remains quite promising in its suggestion that models with higher predictive performance as a result of their ability to capture nonlinearities in ecological relationships can be shown to still be using biological theory to underpin their predictions. We hope that these feature importance measures (especially the Shapley[2] value) are used more regularly in advanced ecological modeling techniques as a way to bolster their credibility and thereby increase their adoption and use.

References

A. L. Angert, L. G. Crozier, L. J. Rissler, S. E. Gilman, J. J. Tewksbury, and A. J. Chunco. Do species' traits predict recent shifts at expanding range edges? *Ecology Letters*, 14(7):677–689, Mar 2011. doi: 10.1111/j.1461-0248.2011.01620.x.

B. Holzinger, K. Hülber, M. Camenisch, and G. Grabherr. Changes in plant species richness over the last century in the eastern swiss alps: elevational gradient, bedrock effects and migration rates. *Plant Ecology*, 195(2):179–196, 2008. doi: 10.1007/s11258-007-9314-9.

J. E. Houlahan, S. T. McKinney, T. M. Anderson, and B. J. McGill. The priority of prediction in ecological understanding. *Oikos*, 126:1–7, 2017. doi: 10.1111/oik.03726.

J. Lubchenco. The relevance of ecology: The societal context and disciplinary implications of linkages across levels of ecological organization. *Linking Species & Ecosystems*, page 297–305, 1995. doi: 10.1007/978-1-4615-1773-3_28.

J. Lubchenco. Environmental science in a post-truth world. *Frontiers in Ecology and the Environment*, 15(1):3, 2017. doi: 10.1002/fee.1454.

S. Lundberg and S. Lee. A unified approach to interpreting model predictions. *NIPS 2017*, 2017. doi: arXiv:1705.07874.

C. Moritz, J. L. Patton, C. J. Conroy, J. L. Parra, G. C. White, and S. R. Beissinger. Impact of a century of climate change on small-mammal communities in yosemite national park, usa. *Science*, 322(5899):261–264, Oct 2008. doi: 10.1126/science.1163428.

[2] `http://github.com/slundberg/shap`