Learning Performance Maximizing Ensembles with Explainability Guarantees

Vincent Pisztora, Jia Li

Department of Statistics, Pennsylvania State University, USA
uxp5@psu.edu, jol2@psu.edu

Abstract

In this paper we propose a method for the optimal allocation of observations between an intrinsically explainable glass box model and a black box model. An optimal allocation being defined as one which, for any given explainability level (i.e. the proportion of observations for which the explainable model is the prediction function), maximizes the performance of the ensemble on the underlying task, and maximizes performance of the explainable model on the observations allocated to it, subject to the maximal ensemble performance condition. The proposed method is shown to produce such explainability optimal allocations on a benchmark suite of tabular datasets across a variety of explainable and black box model types. These learned allocations are found to consistently maintain ensemble performance at very high explainability levels (explaining 74% of observations on average), and in some cases even outperform both the component explainable and black box models while improving explainability.

Introduction

In most high stakes settings, such as medical diagnosis (Gulum, Trombley, and Kantardzic 2021) and criminal justice (Rudin 2019), model predictions have two viability requirements. Firstly, they must exceed a given global performance threshold, thus ensuring an adequate understanding of the underlying process. Secondly, model predictions must be explainable.

Explainability, however defined (Linardatos, Papastefanopoulos, and Kotsiantis 2020), is a desirable characteristic in any prediction function. Intrinsically interpretable "glass box" models ((Agarwal et al. 2021), (Lemhadri, Ruan, and Tibshirani 2021), (Rymarczyk et al. 2020)), which are explainable by construction, are particularly advantageous as they require no additional post-hoc processing ((Ribeiro, Singh, and Guestrin 2016), (Lundberg and Lee 2017)) to achieve explainability, and thus also avoid complications arising from post-hoc explanation learning ((Rudin 2019), (Garreau and Luxburg 2020)). Due to these advantages, glass box models are uniquely suited to settings where faithful explanations of predictions are required.

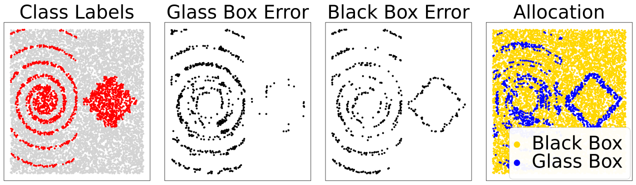


Figure 1: This figure shows a two-class classification task in which the areas of expertise (the diamond pattern for the glass box and the spiral pattern for the black box model) are complementary. The glass box achieves a 92.7% accuracy, the black box reaches 95.0% accuracy, and the allocated ensemble of the two exceeds both with a 95.8% accuracy. Thus, the resulting EEG allocation improves performance over both component models while also providing explainability (for 20% of observations in this case).

However, using an approach of "complete explainability", in which a glass box model is used as the prediction function across the entire feature space, may not be viable. It may be the case that, in a given setting, no glass box exists that can adequately model the relationship of interest in all regions of the feature space. Thus in some regions, the model's predictions will fail to exceed the performance threshold required by the use-case. If, as a consequence, the model exceeds the application's global error tolerance (e.g. a low accuracy in stroke prediction (Gage et al. 2001)), it may not be usable in practice.

An alternative, "partial explainability" approach requires instead that only a proportion of observations be provided intrinsically explainable predictions. We will refer to this proportion, which is the proportion of observations for which the explainable model is the prediction function, as the explainability level q. Such approaches, including our proposed method, Ensembles with Explainability Guarantees (EEG), can provide high performance while maximizing explainability, and work especially well in cases where the explainable model can be paired with an alternate model with complementary strengths. As demonstrated in Fig. 1, by identifying the areas of expertise of the glass box and black box models, the EEG approach can allocate predictions accordingly to improve both performance and explainability.

Generally, implementations of a partial explainability approach consist of an ensemble of models including at least one explainable model and alternate model (often a black box model), and an allocation scheme by which observations are distributed among the ensemble members for prediction. Individual methods are characterized by their heterogeneity in the following aspects.

Methods vary in the range of component models they can accommodate. Some are defined for only one set of glass box, black box, and allocator model types - for example LSP (Wang and Saligrama 2012) and OTSAM (Wang, Fujimaki, and Motohashi 2015), which use binary tree-type splitting to define regions, and linear models and sparse additive models respectively to predict within regions. Other methods are black box agnostic but still limited in glass box and allocator model type - for example HyRS (Wang 2019), HyPM (Wang and Lin 2021), CRL (Pan, Wang, and Hara 2020), and HybridCORELS (Ferry, Laberge, and Aïvodji 2023), which use rule-based models as both glass box and allocator. EEG is the only fully model-agnostic partial explainability method which can be implemented with any combination of glass box, black box, and allocator models.

Methods also vary in the approach used to learn each ensemble member model (i.e. glass box and black box). Most methods first learn the black box model globally (on the full dataset), and then learn the glass box model locally (on its allocated subset of the data), either simultaneously with the allocator (HyRS, HyPM, CRL, and HybridCORELS) or in an alternating EM-style (LSP, OTSAM, and AdaBudg (Nan and Saligrama 2017)). EEG on the other hand, learns both ensemble member models globally first before learning the allocations between them - similar to most general adaptive ensembling methods, e.g. (Gao et al. 2019), (Inoue 2019).

Finally, methods are characterized by their allocation criteria which commonly consist of an objective which combines one or more of the following - the explainability level, the underlying task performance of the ensemble, and the complexity of the glass box model. Most methods optimize a measure of post-allocation ensemble performance - LSP, HyRS, HyPM, and HybridCORELS minimize a 0/1 misclassification loss, AdaBudg uses a more flexible logistic loss, and CRL maximizes accuracy across a range of explainability levels. Several methods with rule-based glass box/allocator hybrid models (HyRS, HyPM, CRL, and HybridCORELS) also include a penalty on the complexity of these models. To control the explainability level, methods either include a reward term in the loss (HyRS, HyPM, and CRL), or directly restrict the model space to candidates which achieve the explainability level (HybridCORELS). In contrast, EEG optimizes an MSE loss between the predicted and actual "glass box allocation desirability" percentile of each observation.

More extensive reviews of the partial explainability approach and explainability methods in general are available in (Linardatos, Papastefanopoulos, and Kotsiantis 2020), (Nauta et al. 2022), and (Sahakyan, Aung, and Rahwan 2021).

As outlined above, our proposed method, Ensembles with Explainability Guarantees (EEG), differs from existing works in its approach to the partial explainability problem. The key novelties of this new approach, and their corresponding advantages are summarized below.

Independent and Global Component Models: The first key innovation of the EEG approach is the independent learning of each component model (i.e. the ensemble member models and allocator). As a result, EEG is agnostic to task, data, and component model type. Thus, the most powerful models can be used for each component as determined by the setting - in contrast with previous works which are more restricted.

Another important consequence of separate component model learning is that glass box predictions are independently explainable in the global context, and thus immune from "explainability collapse" - a scenario in which the allocator subsumes the glass box's prediction role, diminishing the value of the explainable prediction, in the extreme case reducing the glass box to an uninformative constant function. On the other hand, methods which either learn glass box models locally, or jointly with the allocator, are vulnerable to this type of degeneration.

Allocation Desirability Ranking: The second novel aspect of the EEG approach is the concept of allocation desirability. Given an ensemble of models, allocation desirability quantifies how beneficial it is for a given observation to be allocated to the default ensemble member model, say the glass box. Thus, it induces a preference for glass box allocation between all pairs of observations and consequently also defines a ranking of allocation preference across all observations that is optimal irrespective of the desired explainability level.

A key advantage of such a ranking is that it is independent of the training criteria of the ensemble member models, and thus can be adapted to score allocation desirability using metrics that best fit the setting. Indeed, the EEG desirability metric builds a ranking using a combination of relative sufficient performance and absolute performance measures which can natively accommodate any underlying problem type (e.g. regression, classification). This particular desirability metric also offers several additional benefits including allocation desirability percentile and sufficiency category estimates for each observation.

Q-Complete Allocation Optimality: The final key point of novelty of the EEG approach is the optimality of allocation, as defined in Proposition 1 and Proposition 2, which is encoded in the allocation desirability ranking for any explainability level. Thus, the learned allocator, which estimates this ranking, is an explicit function of q and provides the allocation solution to any explainability level after training only once. This capability is in contrast with previous works which provide, at most, several explainability level solutions with varying degrees of stability (Ferry, Laberge, and Aïvodji 2023).

These unique capabilities of the EEG method enable the following practical use cases:

- Given a minimum performance requirement on the underlying task, the method can be used to obtain the allocation with the highest explainability level that achieves or exceeds the performance threshold.

- Given a minimum explainability level requirement, the method can be used to obtain the allocation with the highest ensemble performance which meets or exceeds the required explainability level.
- Given a minimal level of post-allocation glass box-specific performance, the allocation that achieves the highest explainability level while meeting or exceeding this requirement can be found.
- Given a set of observations, sufficiency category estimates can be obtained for each, identifying which observations are likely to yield incorrect decisions and describing the likely failure mode for each such case to inform potential post-hoc remedies.

In the following sections, we first describe our method in detail and provide some theoretical assurances on the characteristics of the resulting allocator in the Methods section. Then, in the Experiments section, we describe the experimental settings and the estimation of the allocator, and demonstrate the method's favorable performance.

Methodology

Setting

First, we define the underlying task as the estimation of the function $f(x) = y$ where $x \in \mathcal{X}, y \in \mathcal{Y}$. We also define observations as $z = (x, y) \in \mathcal{X} \times \mathcal{Y} = \mathcal{Z}$, the training dataset $D^n = \{z_i : i \in \{1, ..., n\}, z_i \in \mathcal{Z}\}$, and loss function for the underlying task $l_U : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. Next, we define the ensemble component models - first, the intrinsically explainable glass box model as $g : \mathcal{X} \to \mathcal{Y}$ and the alternate, black box model as $b : \mathcal{X} \to \mathcal{Y}$, both of which are learned independently on the full training dataset D^n.

Next, we define the allocation task. We define the class of all allocator functions as $A = \{a : \mathcal{V} \to \{0, 1\}\}$ and the class of all "proper" allocator functions as $\hat{A} = \{a : \hat{\mathcal{V}} \to \{0, 1\}\}$, where $\mathcal{V}$ is a general space of inputs (typically $\mathcal{Z}$) and $\hat{\mathcal{V}} \subseteq \mathcal{V}$ containing only information available at allocation time. Next we define the class of q-explainable allocators as $A_q = \{a_q : a_q \in A, \frac{1}{n}\sum_{i=1}^{n} a_q(v_i) = q\}$ and the corresponding class of "proper" q-explainable allocators as $\hat{A}_q = \{a_q : a_q \in \hat{A}, \frac{1}{n}\sum_{i=1}^{n} a_q(v_i) = q\}$, for $q \in \mathcal{Q} = \{\frac{i}{n} : i \in \{1, ..., n\}\} \subseteq [0, 1]$, with q being the explainability level. Note, the set A is used to define the optimal allocator, whereas the set $\hat{A}$ is searched to obtain an estimator of this optimum.

We next define indicators of performance sufficiency. These functions $s : \mathcal{Z} \to \{0, 1\}$ should be thought of as context-dependent indicators of whether performance within a region of the feature space is sufficiently high to use the model in question reliably for explanation. Although the EEG approach holds for any such function s, sufficiency functions used in the Experiments section are defined as follows. For classification tasks, we define performance sufficiency as $s_f(z) = \mathbb{I}\{f(x) = y\}$, and for regression tasks as $s_f(z) = \mathbb{I}\{l_U(f(x), y) < \epsilon\}$, with $f : \mathcal{X} \to \mathcal{Y}$. In practice ϵ should be selected based on problem-specific context, however, lacking such context in the regression experiments conducted for this study, ϵ was selected to be the lower of the average validation losses of g and b, as a reasonable threshold for prediction correctness. These sufficiency indicators generate the following partition of the data: $Z_0 = \{z : z \in \mathcal{Z}, s_g(z) + s_b(z) = 0\}, Z_2 = \{z : z \in \mathcal{Z}, s_g(z) + s_b(z) = 2\}, Z_g = \{z : z \in \mathcal{Z}, s_g(z) = 1, s_b(z) = 0\}$, and $Z_b = \{z : z \in \mathcal{Z}, s_g(z) = 0, s_b(z) = 1\}$, with $n_0 = |Z_0|, n_2 = |Z_2|, n_g = |Z_g|, n_b = |Z_b|$, and $n_q = nq$.

Obs	$l_U(g)$	s_g	$l_U(b)$	s_b
z_1	0	1	2	1
z_2	3	1	4	0

Table 1: This table describes the loss values (l_U) and sufficiencies (s) of two observations (z_1, z_2), for both a glass box (g) and black box model (b). In the constrained allocation case, in which only one observation can be allocated to g, the optimal allocation changes depending on whether loss or sufficiency is used to determine allocation preference.

Next we motivate the use of the sufficiency perspective. Sufficiency functions are critical for defining coherent allocations when, as is often the case, the absolute performance measures used to learn ensemble component models do not match allocation preference (e.g. loss minimization vs accuracy maximization). Consider the constrained allocation decision in Table 1, in which only one observation can be allocated to g.

In this case, loss minimization dictates an allocation of z_1 to g and z_2 to b, which would allocate z_2 to an insufficient prediction. Sufficiency maximization would however yield a more satisfactory allocation of z_2 to g and z_1 to b. This example demonstrates the utility of sufficiency allocation - distinguishing between a case where the user is willing to sacrifice "a bit of performance" (as quantified by sufficiency) for explanation (z_1), and a case where even a small performance drop results in an explanation that is not sufficiently trustworthy to use (z_2).

In the next section, we define the objective of the allocation task and introduce our proposed approach for addressing it.

Optimal Allocation

In the allocation task, the objective is to construct an allocator a_q that will determine which model, either the explainable g or the black box b, is used for prediction on any given observation z, in a manner that is optimal relative to the following criteria. Firstly, for any given explainability level q, the allocator should distribute observations in a way that maximizes sufficient ensemble performance, defined as $\bar{t}(a_q) = \frac{1}{n}\sum_{i=1}^{n} s_g(z_i)a_q(v_i) + s_b(z_i)(1 - a_q(v_i))$. Secondly, and again for any q, the allocator should maximize sufficient explainable prediction $\bar{t}_g(a_q) = \frac{1}{n}\sum_{i=1}^{n} s_g(z_i)a_q(v_i)$, i.e. the performance of the model g on the subset of observations it has been allocated, subject to maintaining maximal $\bar{t}(a_q)$. Finally, the allocator should also be consistent in its allocations across the values of q, meaning that if an observation is allocated to g for a given q, it should remain allocated to the glass box for all higher explainability levels as well. Next, we define our allocator, and show that it meets the three criteria introduced above.

Our proposed allocation function is defined as $a'_q(z) = \mathbb{I}\{r(z) > 1 - q\}$, where rescaled ranking $r(z) = \frac{\text{rank}_{D^n}(\tilde{r}(z))}{n}$, and ranking $\tilde{r}(z) = 2s_g(z) - s_b(z) - \sigma(l_U(g(x), y) - l_U(b(x), y))$, with $\sigma(x) = \frac{1}{1+e^{-x}}$.

The intuition behind the allocator is as follows. First, all observations are sorted in sufficient performance maximizing order, i.e. allocation of observations in Z_g to g is prioritized over allocation of observations in Z_2 and Z_0, which in turn are prioritized over Z_b. Next, observations are sorted in explainable sufficient performance maximizing order, i.e. Z_2 is prioritized ahead of Z_0 for allocation to g. Then, within each sufficiency category, observations are ordered in absolute performance maximizing order, i.e. observations with large relative performance of g over b are prioritized for allocation to g. Next, this ranking is normalized, yielding the glass box allocation desirability percentile r. An important feature of this percentile is that it is constant with respect to q, thus the optimal observations to allocate to g, for any level of q, are simply the n_q most highly ranked, resulting in the allocator a'_q.

Note that in the described methodology, sufficiency based allocation can be viewed as a generalization of allocation via absolute performance, and can thus be reduced to the latter by selecting either $s_f(z) = 0$ or $s_f(z) = 1, \forall z, f$.

Next, we state the optimality properties of the proposed allocator a'_q. The proofs are available in the long form paper on arxiv.org in the Theoretical Results section of the Appendix.

Proposition 1. *(Maximal Sufficient Performance)* $\forall q \in \mathcal{Q}, a'_q \in A^*_q$ *where* $A^*_q = \{a^*_q \colon a^*_q = \arg\max_{a_q \in A_q} \bar{t}(a_q)\}$ *and* $\bar{t}(a_q) = \frac{1}{n}\sum_{i=1}^n t(a_q, z_i) = \frac{1}{n}\sum_{i=1}^n s_g(z_i)a_q(z_i) + s_b(z_i)(1 - a_q(z_i))$

Proposition 2. *(Maximal Sufficient Explainable Performance)* $\forall q \in \mathcal{Q}, a'_q \in A^*_{q|g}$ *where* $A^*_{q|g} = \{a^*_{q|g} \colon a^*_{q|g} = \arg\max_{a^*_q \in A^*_q} \bar{t}_g(a^*_q)\}$ *and* $\bar{t}_g(a^*_q) = \frac{1}{n}\sum_{i=1}^n t_g(a^*_q, z_i) = \frac{1}{n}\sum_{i=1}^n s_g(z_i)a^*_q(z_i)$

Proposition 3. *(Monotone Allocation)* $\forall q_i < q_j \in \mathcal{Q}$, $\{z : z \in \mathcal{Z}, a'_{q_i}(z) = 1\} \subseteq \{z : z \in \mathcal{Z}, a'_{q_j}(z) = 1\}$

Experiments

In this section we describe the data, model training procedures, performance evaluation metrics, and results of our experiments.

Datasets

Tabular data is used to evaluate the proposed methodology as it the setting for which the required intrinsically explainable glass box models are most readily available. Following the tabular data benchmarking framework proposed by (Grinsztajn, Oyallon, and Varoquaux 2022), we conduct experiments on a set of 31 datasets (13 classification, 18 regression). These datasets represent the full set of provided datasets with quantitative features less the four largest scale datasets (omitted due to computational limitations). These datasets are summarized in the Appendix of the long form paper available on arxiv.org.

Each dataset is split (70%, 9%, 21%) into training, validation, and test sets respectively, following (Grinsztajn, Oyallon, and Varoquaux 2022). All features and regression response variables are rescaled to the range [-1,1].

Models

Both glass box and black box models are learned on the full training dataset for each underlying task. For classification datasets, two types of glass box model are fitted, a logistic regression and a classification tree, as well as two types of black box model, a gradient boosting trees classifier and a neural network classifier. Analogously, for regression datasets, two types of glass box model are fitted, a linear regression and a regression tree, as well as two types of black box model, a gradient boosting trees regressor and a neural network regressor. In all cases, the architecture of the neural networks is the "Wide ResNet-28" model (Zagoruyko and Komodakis 2016) adapted to tabular data with the replacement of convolutional layers with fully connected layers.

An allocator is subsequently also learned on the full training dataset. Both gradient boosting trees regressors and neural networks are fitted as allocators for each allocation task. For allocator training, the features x are augmented with four additional constructed features, the predictions $g(x)$ and $b(x)$, and two distance measures $d(g(x), b(x))$ between them, the cross-entropy and MSE. In our experiments, inclusion of these features improved allocator learning - likely by removing the need for the allocator to attempt to learn these quantities on its own. Allocation performance is further improved by ensembling the feature-dependent learned allocator a'_q with a strong feature-independent allocator a''_q, where $a''_q(d(g(z), b(z))) = \mathbb{I}\{\frac{\text{rank}_{D^n} d(g(z),b(z))}{n} < q\}$. a''_q can be viewed as an "assume the black box is correct" allocation rule which is more likely to assign an observation to g if the distance between the predictions of g and b is low. Which of the two allocators is used for a given q is determined by their respective performances on the validation set.

Hyperparameter Tuning

Hyperparameter tuning for all models is done using 4-fold cross-validation, with the exception of the neural network tuning which is done using the validation set. A grid search is done to select the best hyperparameters for each model with search values available in the Appendix of the long form paper available on arxiv.org.

Each glass box and black box model is tuned on the full set of hyperparameters each time it is replicated. The gradient boosting trees allocator models are retuned on the full hyperparameter set each time as well. The neural network allocator is not retuned however, and instead uses the optimal settings found in the fitting of the black box on each dataset.

Metrics

We define the following metrics which are used to measure performance of our method. First we define the Percentage Performance Captured over Random (PPCR) for a given allocator as follows: $PPCR(a_q) = \frac{AUC(a_q) - AUC(r_q)}{AUC(o_q) - AUC(r_q)}$ where $AUC(f_q)$ is the area under the curve of function f_q over all values of q in its domain, o_q is the oracle allocator which has perfect information on the whole dataset, and r_q is the random allocator which selects a subset from the data being allocated uniformly at random. The PPCR metric is a percentage and represents the proportion of the oracle AUC, in excess of that also covered by random allocation, that the learned allocator is able to capture. Thus a value of zero indicates performance on par with r_q and a value of one represents perfect allocation.

Next, we define the Percent Q Equal or Over Max (PQEOM) as the percentage of q values for which the allocator is performing at least as well as the most accurate ensemble member model (i.e. g or b) and Percent Q Over Max (PQOM) as the percentage of q values for which the allocator is performing better than the most accurate ensemble member model (i.e. g or b).

Next, we define the Percent Contribution of Feature-dependent Allocator (PCFA) as the percentage of q values for which the feature dependent allocator a'_q is used for allocation decisions as opposed to the feature-independent allocator a''_q. A value close to one indicates that a'_q is used often, while a value close to zero indicates it is a''_q instead.

Next, we define the 95% Threshold Q Max (95TQM) as the highest value of q for which the ensemble performance meets or exceeds 95% of the performance of the better of g and b. Thus this is a measure of how much explainability can be utilized before the performance price becomes material.

Next, we define the maximum accuracy achieved by the allocator across all q (Max Acc), and the highest value of q for which this accuracy is maintained (Argmax q). The Max Acc can be benchmarked against the AUC, interpretable as the average accuracy across q. Each of these metrics is a percentage and higher values correspond with higher performance and higher explainability at this maximum performance level, respectively.

Finally, we define the accuracy with which the four sufficiency categories (Z_g, Z_b, Z_2, and Z_0) can be estimated as the sufficiency accuracy (s Acc). The higher this accuracy, the better able the allocator is to inform the user of which category a given observation is likely to be a member of.

Results

Evaluation of allocator performance using the metrics defined previously as well as visual inspection of the performance vs explainability trade-off curves (Fig. 2) revealed both the benefits and some of the limitations of learned allocation in the tabular data setting.

Firstly, performance was found to consistently and significantly outperform random allocation, as quantified by a cross-dataset PPCR of 37% (Table 2), indicating that the learned allocation captured close to 40% of the area under the curve available and in excess of random allocation. It was also found that on some datasets in particular, learned allocation performed close to oracle allocation (e.g. 89% and 71% on the IsoletR and BrazilianHousesR datasets).

Learned allocation was also found to perform at least at the level of the best ensemble member model across an average of 74% of the explainability range (PQEOM in Table 2). This indicates that for many datasets, there is a substantial explainability "free lunch" to be taken advantage of without performance loss. On a few datasets, performance of the allocated ensemble was found to outperform both g and b for a majority (93%) of the q range (PolR and FifaR PQOM). The 95TQM metric also supported these conclusions, with a cross-dataset average value of 94% indicating that allocation performance was within 5% of maximal individual model performance across approximately all values of q.

Assessing the PCFA metric suggests some limits to the upside of learned, feature dependent allocation - at least in the tested tabular data setting. A cross-dataset average value of $35\% \pm 34\%$ indicates that on average, the range for which the feature dependent allocator is used over the feature-independent one is indistinguishable from zero. This is consistent with a visual inspection of the representative performance-explainability curve e.g. Fig. 2 (b) where there is no improvement to be had in excess of a''_q. However, it is noted that the only possibility for "homerun" allocations is through the feature dependent a'_q as seen in Fig 2 (a) with the PolR dataset and also in Table 2 for datasets SulfurR, BikeSharingR, and FifaR. Thus the ensembled allocation scheme offers this upside without downside risk of low performance in either a'_q or a''_q.

Evaluation of the case in which a single allocation is needed is also positive. On a cross-dataset average, the 84% maximal accuracy achieved is quite high, and is also achieved at a high average explainability level (64%). Particularly strong individual results can be seen in the Pol and SulfurR datasets (Table 2). We also find that on a observation-level, the allocation is an accurate estimator of sufficiency category, with a cross-dataset average of 76% and with few datasets with accuracy under 60%.

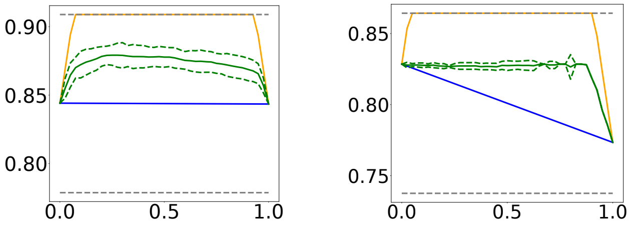
Ablation Studies

Allocator Feature Set Selection In addition to the features x used to learn the glass box and black box models, the allocation task also has access to their predictions $g(x)$ and $b(x)$, and any functions of the two - since the allocator is learned subsequent to the training of these models. To obtain the optimal feature set for allocation, standard tuning procedures (e.g. cross validation) can be employed to evaluate all feature sets of interest. However, as each candidate feature set requires the training of a corresponding allocator for evaluation, this approach can be prohibitively costly.

Thus, the following study was conducted to determine whether a consistently best feature set exists for the tabular data context used in the experiments. First, the universe of candidate features was selected to be the original features x used as inputs for the ensemble component models, the predictions of both of these models $g(x)$ and $b(x)$, and finally two measures of discrepancy between the predictions, the cross-entropy $d_{ce}(g(x), b(x))$ and the mean squared er-

Dataset	AUC	PPCR	PQEOM	PQOM	PCFA	95TQM	Max Acc	Argmax q	s Acc
Wine	79 ± 0	21 ± 0	71 ± 0	0 ± 0	7 ± 0	98 ± 0	80 ± 0	70 ± 0	78 ± 0
Phoneme	87 ± 0	12 ± 1	78 ± 4	7 ± 2	6 ± 1	100 ± 0	87 ± 0	50 ± 34	81 ± 2
KDDIPUMS	88 ± 0	17 ± 1	65 ± 3	34 ± 5	17 ± 7	100 ± 0	88 ± 0	66 ± 9	80 ± 1
EyeMovements	66 ± 0	33 ± 0	55 ± 1	7 ± 6	15 ± 2	70 ± 0	68 ± 0	31 ± 24	52 ± 0
Pol	98 ± 0	49 ± 0	98 ± 0	2 ± 0	0 ± 0	100 ± 0	99 ± 0	98 ± 0	96 ± 0
Bank	76 ± 0	-19 ± 0	4 ± 1	0 ± 0	0 ± 0	100 ± 0	79 ± 0	100 ± 0	71 ± 0
MagicTelescope	86 ± 0	39 ± 0	87 ± 2	12 ± 11	10 ± 0	100 ± 0	86 ± 0	47 ± 28	82 ± 1
House16H	89 ± 0	40 ± 0	84 ± 4	9 ± 6	6 ± 2	98 ± 0	89 ± 0	82 ± 8	86 ± 0
Credit	78 ± 0	5 ± 1	56 ± 4	14 ± 19	95 ± 0	100 ± 0	78 ± 0	76 ± 6	72 ± 0
California	90 ± 0	52 ± 0	88 ± 0	0 ± 0	7 ± 0	98 ± 0	91 ± 0	88 ± 0	90 ± 0
Electricity	92 ± 0	58 ± 0	88 ± 0	0 ± 0	7 ± 0	98 ± 0	93 ± 0	88 ± 0	92 ± 0
Jannis	79 ± 0	30 ± 0	53 ± 5	21 ± 5	14 ± 1	98 ± 0	79 ± 0	35 ± 6	76 ± 0
MiniBooNE	94 ± 0	54 ± 0	90 ± 0	0 ± 0	5 ± 0	100 ± 0	94 ± 0	90 ± 0	92 ± 0
WineR	73 ± 0	43 ± 0	85 ± 0	10 ± 0	20 ± 0	90 ± 0	74 ± 0	78 ± 0	67 ± 0
IsoletR	91 ± 0	89 ± 0	68 ± 0	0 ± 0	49 ± 10	73 ± 0	95 ± 0	68 ± 0	87 ± 1
CPUR	75 ± 0	40 ± 2	52 ± 18	0 ± 0	29 ± 11	80 ± 0	77 ± 0	70 ± 0	59 ± 1
SulfurR	98 ± 0	63 ± 2	73 ± 9	1 ± 2	79 ± 4	100 ± 0	98 ± 0	84 ± 22	96 ± 0
BrazilianHousesR	96 ± 0	71 ± 1	83 ± 0	64 ± 6	1 ± 2	93 ± 0	98 ± 0	8 ± 3	88 ± 0
AileronsR	75 ± 0	13 ± 0	70 ± 10	5 ± 7	4 ± 2	100 ± 0	76 ± 0	40 ± 35	64 ± 0
MiamiHousingR	76 ± 0	44 ± 0	76 ± 0	0 ± 0	67 ± 10	80 ± 0	78 ± 0	75 ± 0	65 ± 0
PolR	88 ± 0	44 ± 1	96 ± 1	93 ± 1	88 ± 0	100 ± 0	88 ± 0	81 ± 2	84 ± 0
ElevatorsR	75 ± 0	25 ± 0	64 ± 4	2 ± 2	17 ± 0	88 ± 0	75 ± 0	51 ± 29	63 ± 0
BikeSharingR	77 ± 0	26 ± 1	88 ± 4	22 ± 16	82 ± 1	100 ± 0	78 ± 0	21 ± 13	72 ± 0
FifaR	77 ± 0	33 ± 0	95 ± 0	93 ± 0	78 ± 0	100 ± 0	77 ± 0	50 ± 0	69 ± 0
CaliforniaR	78 ± 0	38 ± 0	73 ± 0	68 ± 0	68 ± 0	93 ± 0	79 ± 0	42 ± 0	63 ± 0
HousesR	78 ± 0	41 ± 0	73 ± 0	0 ± 0	48 ± 1	80 ± 0	79 ± 0	73 ± 0	62 ± 0
SuperconductR	83 ± 0	42 ± 0	60 ± 13	24 ± 11	95 ± 0	95 ± 0	83 ± 0	57 ± 1	76 ± 0
HouseSalesR	76 ± 0	50 ± 1	64 ± 12	0 ± 0	56 ± 6	85 ± 0	78 ± 0	78 ± 0	63 ± 0
House16HR	92 ± 0	55 ± 0	90 ± 0	0 ± 0	6 ± 0	98 ± 0	92 ± 0	90 ± 0	86 ± 0
DiamondsR	70 ± 0	19 ± 0	83 ± 0	34 ± 6	73 ± 0	100 ± 0	71 ± 0	20 ± 6	65 ± 0
MedicalChargesR	86 ± 0	24 ± 0	91 ± 1	85 ± 1	0 ± 0	100 ± 0	86 ± 0	67 ± 2	83 ± 0
Average	**83 ± 9**	**37 ± 21**	**74 ± 19**	**20 ± 29**	**35 ± 34**	**94 ± 9**	**84 ± 8**	**64 ± 24**	**76 ± 12**

Table 2: This table summarizes allocation performance across several metrics on each dataset and, in the bottom row, across the datasets (all metrics are reported as percentages). Averages and standard deviations are reported over 5 replicates. Metric definitions can be found in the Metrics section, and discussion of results can be found in the Results section.



(a) PolR Dataset

(b) SuperconductR Dataset

Figure 2: This figure shows two examples of the explainability (x-axis) vs sufficient performance (y-axis) trade-off, comparing the random (blue), oracle (orange), and learned (green) allocation curves. The PolR dataset is an example of complementary g and b models, resulting in an allocated ensemble that outperforms both component models across most of the q range. The SuperconductR dataset is an example of an explainability "free lunch" in which the b accuracy is maintained while increasing explainability using the allocator. Curves for all datasets are available in the Appendix of the long form paper available on arxiv.org.

ror $d_{mse}(g(x), b(x))$. The measures of disagreement were included as features as they translate to the "feature independent" strategy of allocation to model b for low values of $d(a(x), b(x))$ - in other words the optimal allocation strategy assuming a is always correct.

The candidate features were grouped into the sets listed in Table 3 and then used to train allocators on a subset of the benchmark datasets (Wine, WineR, Phoneme, SulfurR, Bank, BrazilianHousesR, FifaR, KDDIPUMS) with 6 replicates per model.

Next each feature set was evaluated as follows. First, within each dataset, each feature set's performance (defined as the AUC) was compared to the performance of the best alternative set of features. Then, the proportion of datasets for which the feature set being evaluated was not significantly worse (i.e. either significantly better or not significantly different) than the best alternative was recorded and reported in Table 3 for three significance levels (10%, 5%, 1%).

The results support the following three conclusions. First, no one feature set proved universally best across the tested datasets and thus a full search across feature sets would be advised in settings without resource constraint. Second,

Feature Set	$\alpha : 0.01$	$\alpha : 0.05$	$\alpha : 0.1$
x	18.75%	18.75%	12.50%
g, b	43.75%	43.75%	31.25%
d_{ce}	31.25%	18.75%	18.75%
d_{mse}	50.00%	43.75%	37.50%
x, d_{ce}	37.50%	25.00%	18.75%
x, d_{mse}	56.25%	37.50%	25.00%
g, b, d_{ce}	31.25%	25.00%	25.00%
g, b, d_{mse}	50.00%	50.00%	37.50%
x, g, b	56.25%	43.75%	37.50%
x, g, b, d_{ce}	50.00%	43.75%	37.50%
x, g, b, d_{mse}	56.25%	37.50%	37.50%
x, g, b, d_{ce}, d_{mse}	**75.00%**	**56.25%**	**43.75%**

Table 3: This table reports the percentage of datasets for which the allocator learned on the corresponding feature set is significantly better than or not significantly different from the best alternative feature set trained allocator. Results across three significance levels are reported and show that the "kitchen sink" x, g, b, d_{ce}, d_{mse} feature set is most consistently best (bolded) while the "unaugmented" original feature set of x is consistently the worst across all α.

although no universally best feature set was found, the "kitchen sink" set of all candidate features (x, g, b, d_{ce}, d_{mse}) was found to be best most consistently and was thus used to train all allocators reported in Table 2. Finally, allocators trained on just the original features x were found to be consistently worst among all alternatives thus supporting the augmentation of the original features in some form. This finding is consistent with the intuition that the predictions of the component models would be very useful to learning the optimal allocation and would be either very difficult or impossible to learn from r, the optimal allocation ranking response, alone during training.

Ensemble Component Model Selection The performance of any allocated ensemble is highly dependent not only on the individual performance of its component models (i.e. g and b) but on their level of synergy as well. In particular, it may be the case that the component model pair in (g_0, b_0, a_0) individually outperforms the respective component models in (g_1, b_1, a_1) but that the allocator a_0 trained with (g_0, b_0) underperforms a_1. In this case, the high relative advantages of (g_1, b_1) in different segments of the feature space overcome their global performance disadvantages as individual models compared to their counterparts in (g_0, b_0) to yield a stronger ensemble.

Thus, to determine how often high relative advantage overcomes superior individual performance in allocator training, the following study was conducted. For each dataset, an allocator was trained on each combination of available glass box (tree and regression) and black box (gradient boosting trees and neural network) models (i.e. four allocators per dataset). Then the allocator a_I, trained using the pair of component models (g_I, b_I) with the highest individual validation performance, was identified along with the allocator a_C, trained using the pair of component models (g_C, b_C) resulting in the highest ensemble validation performance. Finally the difference in test performance was measured between a_C and a_I ($AUC\Delta = AUC(a_C) - AUC(a_I)$).

The resulting $AUC\Delta$ values support the following two conclusions. Firstly, while a relatively high proportion (41.9%) of datasets yield different allocators depending on which of the two different component model selection processes (individual vs. combined performance) they utilize, the cross-dataset average difference in allocator performance is not significantly different from zero (0.01 ± 0.03). This result suggests that the glass box and black box model types used for the experiments did not exhibit high relative expertise in different parts of the feature space, indicating that it may be beneficial to use a more diverse set of component models in this setting. However, in rare cases (e.g. IsoletR, BrazilianHousesR) the combined performance selection method yields as much as 15% in additional performance. Thus, in resource constrained settings, or in cases in which many glass box and black box model types are under consideration, the individual performance selection method appears relatively low risk, although a full search across all component model combinations (the method used for Table 2) is recommended when feasible.

Acknowledgements

We would like to thank Jonathan Siegel for valuable discussion of the theoretical aspects of this work.

This research is supported by the National Science Foundation under grant number CCF-2205004. Computations for this research were performed on the Pennsylvania State University's Institute for Computational and Data Sciences' Roar supercomputer.

References

Agarwal, R.; Melnick, L.; Frosst, N.; Zhang, X.; Lengerich, B.; Caruana, R.; and Hinton, G. E. 2021. Neural additive models: Interpretable machine learning with neural nets. *Advances in neural information processing systems*, 34: 4699–4711.

Ferry, J.; Laberge, G.; and Aïvodji, U. 2023. Learning Hybrid Interpretable Models: Theory, Taxonomy, and Methods. *arXiv preprint arXiv:2303.04437*.

Gage, B. F.; Waterman, A. D.; Shannon, W.; Boechler, M.; Rich, M. W.; and Radford, M. J. 2001. Validation of clinical classification schemes for predicting stroke: results from the National Registry of Atrial Fibrillation. *Jama*, 285(22): 2864–2870.

Gao, X.; Shan, C.; Hu, C.; Niu, Z.; and Liu, Z. 2019. An adaptive ensemble machine learning model for intrusion detection. *Ieee Access*, 7: 82512–82521.

Garreau, D.; and Luxburg, U. 2020. Explaining the explainer: A first theoretical analysis of LIME. In *International conference on artificial intelligence and statistics*, 1287–1296. PMLR.

Grinsztajn, L.; Oyallon, E.; and Varoquaux, G. 2022. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in Neural Information Processing Systems*, 35: 507–520.

Gulum, M. A.; Trombley, C. M.; and Kantardzic, M. 2021. A review of explainable deep learning cancer detection models in medical imaging. *Applied Sciences*, 11(10): 4573.

Inoue, H. 2019. Adaptive ensemble prediction for deep neural networks based on confidence level. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 1284–1293. PMLR.

Lemhadri, I.; Ruan, F.; and Tibshirani, R. 2021. LassoNet: Neural Networks with Feature Sparsity. In Banerjee, A.; and Fukumizu, K., eds., *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, 10–18. PMLR.

Linardatos, P.; Papastefanopoulos, V.; and Kotsiantis, S. 2020. Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1): 18.

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Nan, F.; and Saligrama, V. 2017. Adaptive classification for prediction under a budget. *Advances in neural information processing systems*, 30.

Nauta, M.; Trienes, J.; Pathak, S.; Nguyen, E.; Peters, M.; Schmitt, Y.; Schlötterer, J.; van Keulen, M.; and Seifert, C. 2022. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*.

Pan, D.; Wang, T.; and Hara, S. 2020. Interpretable companions for black-box models. In *International conference on artificial intelligence and statistics*, 2444–2454. PMLR.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.

Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5): 206–215.

Rymarczyk, D.; Struski, Ł.; Tabor, J.; and Zieliński, B. 2020. Protopshare: Prototype sharing for interpretable image classification and similarity discovery. *arXiv preprint arXiv:2011.14340*.

Sahakyan, M.; Aung, Z.; and Rahwan, T. 2021. Explainable Artificial Intelligence for Tabular Data: A Survey. *IEEE Access*, 9: 135392–135422.

Wang, J.; Fujimaki, R.; and Motohashi, Y. 2015. Trading interpretability for accuracy: Oblique treed sparse additive models. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 1245–1254.

Wang, J.; and Saligrama, V. 2012. Local supervised learning through space partitioning. *Advances in Neural Information Processing Systems*, 25.

Wang, T. 2019. Gaining free or low-cost interpretability with interpretable partial substitute. In *International Conference on Machine Learning*, 6505–6514. PMLR.

Wang, T.; and Lin, Q. 2021. Hybrid predictive models: When an interpretable model collaborates with a black-box model. *The Journal of Machine Learning Research*, 22(1): 6085–6122.

Zagoruyko, S.; and Komodakis, N. 2016. Wide residual networks. *arXiv preprint arXiv:1605.07146*.