Distill knowledge of additive tree models into GAMs

Arthur MAILLART and Christian ROBERT

16th june 2023

Agenda

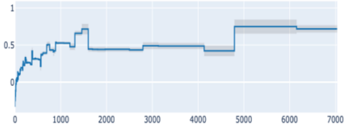
- Distilltrees
- Synthetic data
- Use case: Trade Credit Insurance
- Conclusion

Let us denote by Y the target variable and $x_1, \ldots, x_p$ the covariates, Generalized Additive Models Hastie and Tibshirani 1990 are defined:

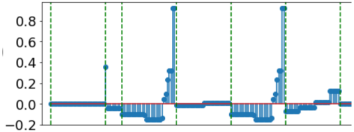
$$g(E(Y)) = \beta_0 + \sum_{k=1}^{p} f_k(x_k)$$

- GAM (smooth): In practice unknown smooth functions are replaced in the model by basis expansions $f_j(x_j) = \sum_{k=1}^{K_j} \beta_{jk} b_{jk}(x_j)$ and with the smoothness constraint $D(\beta) + \sum_j \lambda_j \int f_j''(x)^2 dx$ the problem becomes $\hat{\beta} = \arg\min_\beta \{D(\beta) + \sum_j \lambda_j \beta^T S_j \beta\}$.
- Binarsity: Weights w are defined by $\arg\min_w \{f(w) + g(w)\}$ where f is a goodness of fit function and $g(w) = s \sum_{j=1}^{d} \left(\sum_{k=2}^{d_j} |w_{j,k} - w_{j,k-1}| + \delta_C(w_{j,\bullet}) \right)$.
- EBM: Explainable Boosting Machine is fitted with Cyclic boosting. It begins by growing a shallow decision tree on the first feature in the dataset. It is the initial shape function. Then the algorithm iterate through predictors to fit a shallow tree (with a single variable at a time) on the residuals provided by the predictions of the current model (sum of shallow trees). Once a shallow tree has been learned for every feature, the boosting process cycles back to the first feature and continues in a round robin fashion for all epochs to jointly optimize all functions.

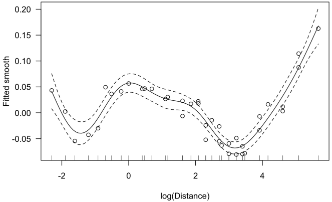
Examples



Explainable Boosting Machine explanation



Binarsity explanation

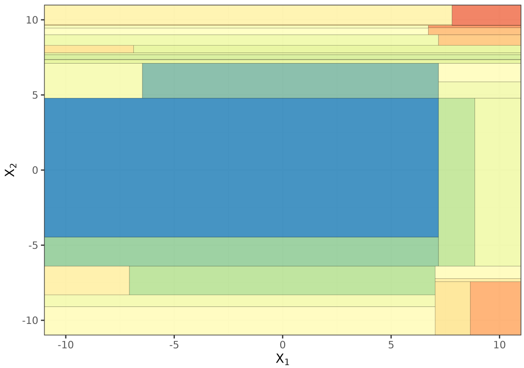


GAM (smooth) explanation

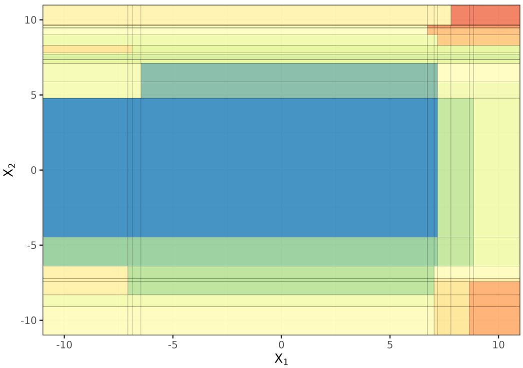
Detralytics

With Distilltrees, we want to encode the knowledge of additive tree ensemble models (e.g. GBM, XGBoost, Random Forest,...) into an additive model in piecewise constant functions.
To do this, we propose the following approach:

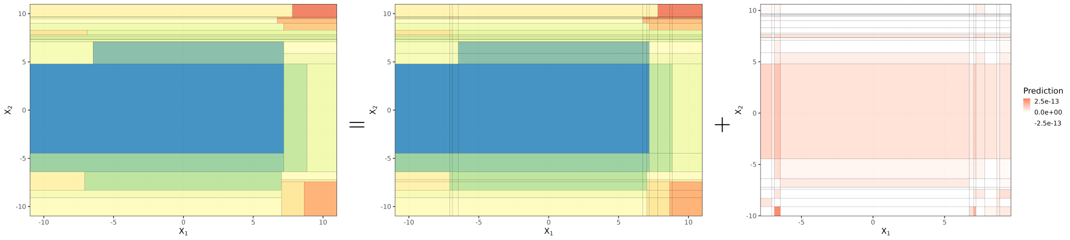
- Collect each tree in the set
- For each of these trees, identify the set of splits and create a new partition defined by crossing these splits thresholds.
- For each tree, get the training sample used to build the tree.
- Train a linear model on this new dataset.
- Combine linear models.



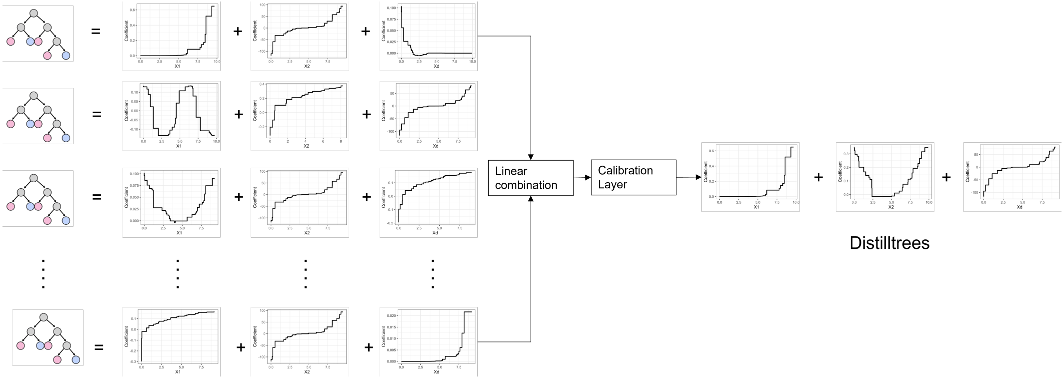
Tree-induced partition



Partition induced by the linear model



(a) Tree's response in covariates space (c) GLM's response in covariates space (e) Errors in covariates space


Linear combination
Calibration Layer
Distilltrees

True model

We simulated training and testing sets following this process:

- randomly drawing 50,000 values for $X_1, \dots, X_{10}$ with $X_i \sim \mathcal{U}([0,1])$ for $i \in \{1..10\}$
- building the target variable as:

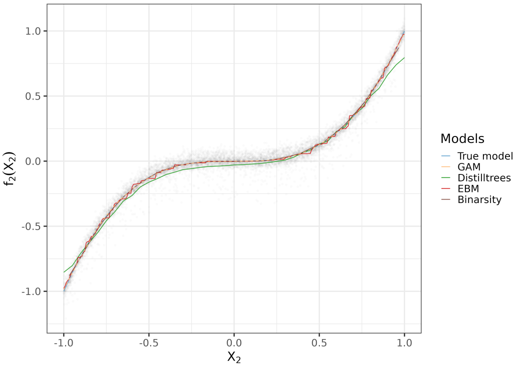
$$Y = 3X_1 + X_2^3 + \pi^{X_3} + \exp\left(-2X_4^2\right) + \frac{1}{1+|X_5|} + X_6 \log\left(X_6\right) + \sqrt{2|X_7|} + \max\left(0, X_7\right) + X_8^4 + 2\cos\left(\pi X_8\right).$$

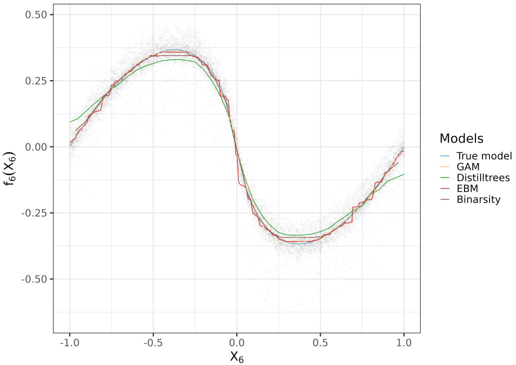
This means that X_9 and X_{10} are not linked to the target and are just noise variables.

Performance reported on test set

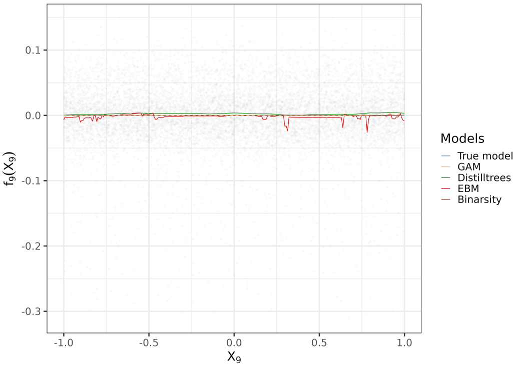
	R^2
GAM (smooth) Hastie and Tibshirani 1990	99.96%
EBM Nori et al. 2019	99.96%
Binarsity Alaya et al. 2019	99.76%
Distilltrees (our method)	99.69%

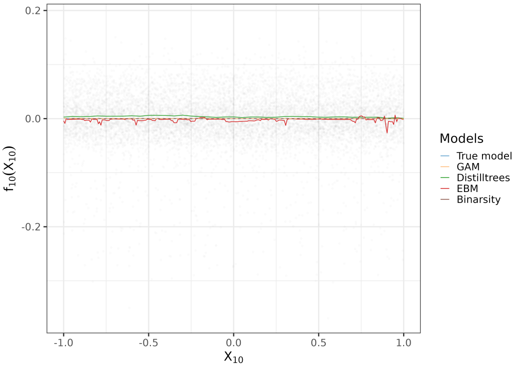
Below is a comparison of the explanations provided by the different models.





Noise variables are correctly handled by distilltrees.





We apply distilltrees to data from a credit insurer (Allianz Trade) to explain a binary model linked to an underwriting task. The aim is to build a model that can be explained to business experts and is competitive with XGBoost.

- Data containing 17 variables containing policyholder related informations,
- Motivations: getting an explainable model from XGBoost.

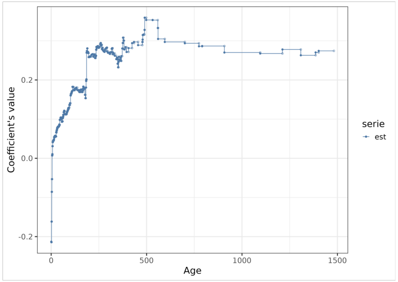
Performance comparison

Performance reported on test set

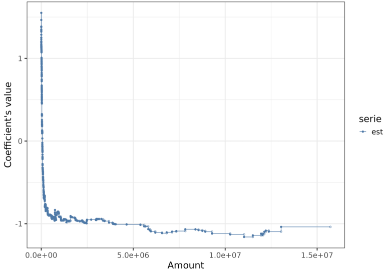
	train AUC	out of sample AUC	out of time AUC
XGBoost	90.28%	89.37%	83.59%
EBM	88.12%	87.69%	81.81%
Distilltrees	87.95%	87.42%	81.43%
Binarsity	87.22%	85.90%	79.63%
GAM	82.67%	81.83%	75.12%

Use case Trade Credit Insurance

Score = 1.17 +



+ ... +



Distilltrees model (17 variables)

Detralytics

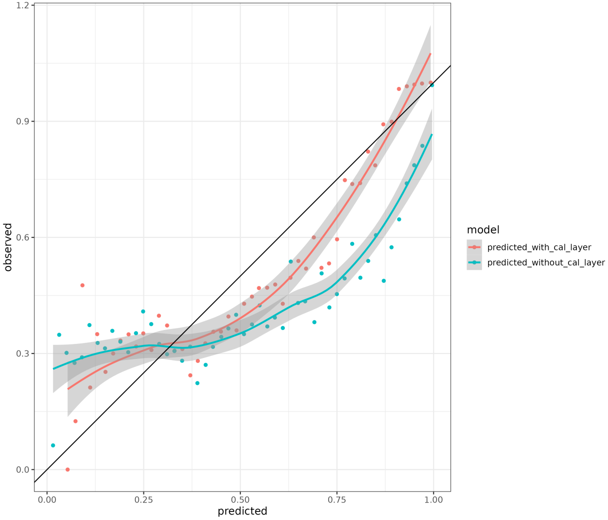
- If the performance is close to that of the initial model (XGBoost here), it is possible to create an explainable prediction model that replaces our initial model.
- The approach can be used for classification and regression.
- This method makes it possible to explain an additive model of tree predictions when the fidelity is sufficiently high.
- The approach can be applied to a wide range of models: random forest, gradient boosting and variants. In addition, although we have not specifically tested it, this technique can be applied to a linear combination of additive tree prediction models (stacking).

Bibliography I

Alaya, Mokhtar Z. et al. (2019). *Binarsity: a penalization for one-hot encoded features in linear supervised learning.* arXiv: 1703.08619 [stat.ML].

Hastie, Trevor J and Robert J Tibshirani (1990). *Generalized additive models.* Vol. 43. CRC press.

Nori, Harsha et al. (2019). "InterpretML: A Unified Framework for Machine Learning Interpretability". In: *ArXiv* abs/1909.09223.



Detralytics