Introduction

In the field of machine learning, the relationship between accuracy and interpretability is often seen as a trade-off. An ensemble of trees, like for example, a random forest, often provides highly accurate results but the final predictive model is too complex for human understanding (James et al., 2013). A high-accuracy, low-interpretability model like this will from now on be referred to as a *black box* model.

As the usage of black box models has become more commonplace it is increasingly important to understand how these models make their predictions (Bodó, & Janssen, 2021; Varshney & Alemzadeh, 2017). As pointed out by Rudin (2019): “ the lack of transparency and accountability of predictive models can have (and has already had) severe consequences”. In the USA, for example, people have been incorrectly denied parole, or incorrectly released on bail (Wexler, 2017). Recently, the Dutch government came under scrutiny in an event called “Dutch childcare benefits scandal” (Dutch: “Toeslagenaffaire”), where, between 2013 and 2019, an estimated 26,000 parents of Bulgarian descent were wrongly accused of making fraudulent benefit claims, driving many families into severe (financial) hardships (van Bruxvoort & van Keulen, 2021; Huisman, 2020). The European Digital Rights (EDRi) organization, a network of more than 40 European civil organizations, advocate for more regulation and transparency on artificial intelligence decision-making systems in the European Union, especially in the area of law enforcement and criminal justice.

While highly accurate black box models are very useful, the call for models with higher interpretability is valid and cannot be ignored. In practice, post-hoc explanation techniques that attempt to interpret black box models are often used. Local Interpretable Model-agnostic Explanations (LIME; Ribeiro, et al., 2016) and Shapley additive explanations (SHAP; Lundberg & Lee, 2017) are popular examples. Kaur et al. (2020) however, point out that while these tools are useful in uncovering issues with datasets or models, they are often over-trust and misused for interpretation. Rudin (2019) argues that instead of trying to explain black models, the way forward is to create *glass box* models that are not only accurate, but also inherently interpretable. She disregards the trade-off between accuracy and interpretability as a myth. She shows that in numerous domains, highly interpretable glass box models exist that have accuracy close to, equal, or higher than black box models. Furthermore, she notes that:

Generally, in the practice of data science, the small difference in performance between machine learning algorithms can be overwhelmed by the ability to interpret results and process the data better at the next iteration. In those cases, the accuracy/interpretability tradeoff is reversed – more interpretability leads to better overall accuracy, not worse.

That said, high accuracy is of course still needed. Perfectly interpreting a model that is not accurate would not be a fruitful enterprise. In the current thesis, I will attempt to improve the accuracy of the Generalized Linear Mixed-Model (GLMM) tree model. As explained below, this glass box model is easily interpretable, as it results in a single decision tree. It can also produce more accurate results as compared to decision trees created using the CART mechanism (Fokkema et al., 2018; Hajjem et al., 2017; Sela & Simonoff, 2012). GLMM trees however, fall short on the accuracy of black box models. Using the *born-again tree* algorithm, combined with a Bayesian tree ensemble method, I will attempt to improve the accuracy of the GLMM tree to be comparable to that of a black box model, while maintaining the high interpretability.

**Born-Again Trees**

In 1996, Breiman and Shang introduced the Born-Again (BA) tree algorithm. The BA approach allows the user to create a single tree model, which has an accuracy that is close to the black box model, but is much easier to interpret and can be used by humans to make predictions using only an image of the tree and the predictor variables occurring in the tree.

A BA algorithm follows these four steps:

1. A black box model is fitted on the original predictor variables (**X**).
2. Based on **X**,a large number of observations are artificially generated (**X­gen**).
3. The black box model is used on **X­gen** to create predictions for the generated data (**Ygen**).
4. A single tree is fitted on **Ygen** to create a BA tree.

Researchers studying BA trees can thus focus on three aspects:

* The black box model (steps 1 and 3)
* The data generation method (step 2)
* The single BA tree (step 4)

Below, I will describe how these aspects are studied in the current thesis.

***The Black box model: Baysian Aditive Regression Trees***

Previous studies focusing on black box models in the context of BA trees, found accurate results with boosted tree ensembles (Breiman & Shang, 1996; …), support vector machines (BRON BRON), and neural networks (Craven & Shavlik, 1995; …). In the current study, I will focus only on tree ensembles. In previous research, Classification And Regression Trees (CART) based ensembles were used. In the current study however, I will use Bayesian Additive Regression Trees (BART) based ensembles. Not only have BART based ensembles been shown to outperform CART based ensembles in terms of predictive accuracy (Chipman, George & McCulloch, 2010), they provide the added benefits of calculating a posterior probability distribution (PPD), and allowing mixed-effects structures to be incorporated. The PPD will be used in the data generation method, and the mixed-effects structure will be used for accounting for multilevel and longitudinal data structures.

***The Data Generation Method***

In order for the BA tree to have a high accuracy, it needs a large sample size to build the tree (Breiman & Shang 1996). Breiman and Shang therefore suggest manufacturing data to increase the sample size. In their study, they apply *smearing* on the original set of predictor variables **X**, to create a larger dataset **X­gen.** The smearing algorithm repeats the following steps:

1. A row *i* is randomly selected from **X**(*i, j*), with [*i* = 1, … , *N*].
2. A random number *r* is selected for every predictor *j*, with [*j* = 1, … , *P*].
3. If *r* > *palt*, **X­gen**(*i, j*) = **X**(*i, j*). Else, a random value of *j* in [*j* = 1, … , *P*] is selected for **Xgen**(*i, j*) = **X**(*i, jrandom*).
4. Using the black box model defined earlier, **Xgen**(*i, j*) is used to create **Ygen**(*i, j*)

*N* is the number of participants, *P* is the amount of predictors, and *palt* is a threshold number that is set beforehand. Note that for *palt* = 0, we are performing permutation and when *palt* = 1, we are sampling random participants as is. *Palt* thus represents the degree of permutation. *Palt* can be optimized for more accurate results. Breiman and Shang (1996) use *palt* = 0.25 and *palt* = 0.50 in their study. In the current thesis, I will compare the effectivity of *palt* = 0, *palt* = 0.25, and *palt* = 0.50, with the use of a new data generation method: the PPD method.

There has not been significant improvement to smearing as the data generation method, but the use of permutation may lead to unrealistic data patterns and has been criticized (Hooker et al., 2021). In the current study, I will sample the **Y­gen** from the PPD acquired when using BART ensembles.Data generate in this manner should more closely resemble **X** and thus lead to a more accurate BA tree. As stated before, this is possible when using a BART ensemble as a black box.

***The BA Tree***

As stated earlier, the BART based ensembles allow for incorporation of mixed-effects structures. To take advantage of that, I will use a Generalized Linear Mixed-Model (GLMM) tree model as the single BA tree. GLMM trees account for correlated structures in decision-tree analyses, and have been shown to yield more accurate, as well as less complex trees (e.g., Fokkema et al., 2018; Hajjem et al., 2017; Sela & Simonoff, 2012).

***Research Questions***

The main goal of the current paper is to study the data generation method in the creation of a BART based BA GLMM tree (working title). I will compare using the PPD to generate data, with the traditional smearing. When smearing, different values of palt will be used. The main research question is thus:

**RQ1:** How does a BART based BA GLMM tree perform in terms of predictive accuracy and interpretability when using the PPD to generate data, compared to using smearing or using a single GLMM tree?

When the optimal data generation method is found, I will compare the resulting BART based BA GLMM tree to a regular GLMM tree, and to a multilevel BART ensemble. This leads us to the second research question:

**RQ2:** How does a BART based BA GLMM tree perform in terms of predictive accuracy and interpretability compared to a GLMM tree, and to a multilevel BART ensemble?

I will answer these questions by using real datasets that have a multilevel structure.

**Methods**

***Datasets***

*Safety*

*ECLSK*

*ACT*

***Working Hypotheses***

**References**

Bodó, B., & Janssen, H. (2021). Here Be Dragons–Maintaining Trust in the Technologized Public Sector. *Amsterdam Law School Research Paper*, (2021-23).

Breiman, L., & Shang, N. (1996). Born again trees. *University of California, Berkeley, Berkeley, CA, Technical Report*, *1*(2), 4.

van Bruxvoort, X., & van Keulen, M. (2021). Framework for Assessing Ethical Aspects of Algorithms and Their Encompassing Socio-Technical System. *Applied Sciences*, *11*(23), 11187.

Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, *4*(1), 266-298.

Craven, M., & Shavlik, J. (1995). Extracting tree-structured representations of trained networks. *Advances in neural information processing systems*, *8*.

Fair Trials and European Digital Rights (EDRi), (2022, March 1). *Civil society calls on the EU to ban predictive AI systems in policing and criminal justice in the AI Act*. Retrieved from <https://edri.org/our-work/civil-society-calls-on-the-eu-to-ban-predictive-ai-systems-in-policing-and-criminal-justice-in-the-ai-act/>

Fokkema, M. (2017). Fitting prediction rule ensembles with R package pre. *arXiv preprint arXiv:1707.07149*.

Fokkema, M., Edbrooke-Childs, J., & Wolpert, M. (2021). Generalized linear mixed-model (GLMM) trees: A flexible decision-tree method for multilevel and longitudinal data. *Psychotherapy Research*, *31*(3), 329-341.

Fokkema, M., Smits, N., Zeileis, A., Hothorn, T., & Kelderman, H. (2018). Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees. *Behavior research methods*, *50*(5), 2016-2034.

Hooker, G., Mentch, L., & Zhou, S. (2021). Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance. *Statistics and Computing*, *31*(6), 1-16.

Huisman, P. (2020). Hoe de toeslagenaffaire kon gebeuren. *Management Kinderopvang*, *26*(2), 36-37.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). New York: springer.

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., & Wortman Vaughan, J. (2020, April). Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1-14).

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., & Wortman Vaughan, J. (2020, April). Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1-14).

**Klievink, A. J. (2021). *Hollen én stilstaan: hoe data en digitalisering de overheid veranderen* (Doctoral dissertation, Leiden University).**

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " 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* (pp. 1135-1144).

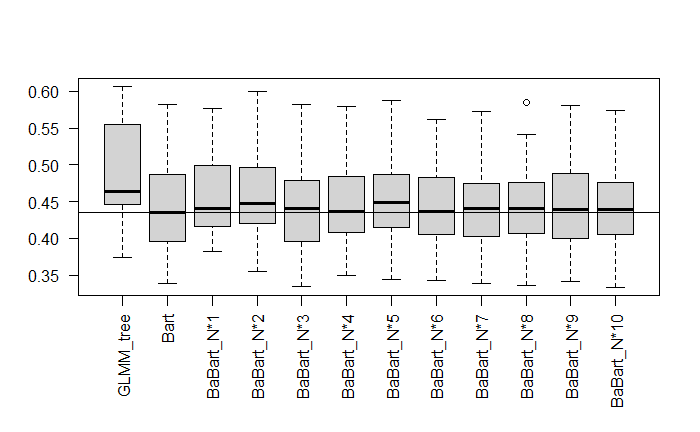
Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206-215.

Semenova, L., Rudin, C., & Parr, R. (2019). A study in Rashomon curves and volumes: A new perspective on generalization and model simplicity in machine learning. *arXiv preprint arXiv:1908.01755*.

Varshney, K. R., & Alemzadeh, H. (2017). On the safety of machine learning: Cyber-physical systems, decision sciences, and data products. *Big data*, *5*(3), 246-255.

Wexler, R. (2017). When a computer program keeps you in jail. *New York Times*.

Figuur 1: MSE\_weighted\_par\_1to10



Figuur 2: MSE NOT weighted

