July 8, 2022

Response to Reviewer #3

Here is the detail of the work done in relation to the points raised by reviewer #3.

Reviewer's overview:

The package is the result of the implementation of new methodologies previously proposed by the authors and it is a very innovative proposal to address the problem proposed, they also carry out an extensive bibliographic review.

The article has a development according to the proposal of methodologies and functions presented that have been developed in the package.

Our reply: Thanks for your positive comments.

Response to Reviewer #4

We would like to thank the reviewer for his/her valuable comments. In our opinion, the paper has been improved as a result of these suggestions. We have done our best to accommodate all of the reviewer's comments, and as a result have carefully addressed all changes requested. We have duly acknowledged such help in the corresponding section.

Here is the detail of the work done in relation to the points raised by reviewer #4.

Reviewer's overview:

The introduction of the eat package by Esteve et al. solves a problem relevant to the readership of the R journal in a manner that is clearly articulated, methodologically sensible and well motivated. The package addresses a germane area of research, specifically, the estimation of production frontiers (i.e., the best producible outputs from given inputs, typically in engineering or economic contexts) and measurement of technical efficiency with respect to these frontiers, which are both topics of current interest in the operations research literature. The package makes a novel and helpful contribution, allowing the estimation of production frontiers using regularized tree based methods, namely multi-output pruned Decision Trees or Random Forests, which overcome issues of overfitting inherent in other non-parametric approaches. The approach implemented in the package is sensible and based upon the recently published method of Efficiency Analysis Trees (Esteve et al., 2020). The technology and methods in this package are up to date. One consideration for future enhancement could be to allow for the use of additional machine learning techniques which may outperform pruned Decision Trees or Random Forests. A reasonable state of the art alternative can be found in the stacked ensembles and neural networks in the H2O package, which also automates tuning of these algorithms to a decent extent. However, lack of these alternative techniques should not be an impediment to publication of this article, as Decision Trees and Random Forests likely are adequate for many use cases depending on the size of data, number of features, and relationships between features.

Our reply: Thanks for your positive comments.

Reviewer's comment on the article:

The article is clearly written and possesses consistent, intuitive mathematical notation. Sufficient context is given describing alternative nonparametric approaches to estimate production frontiers and efficiency, with clear explanation of how the method of Efficiency Analysis Trees differs and offers comparative value. A thorough list of alternative packages is also provided. Examples are reproducible and easy to follow. Only the following minor points could be clarified:

1. Which exact variable importance metric is used in rankingEAT? It seems to be based on the variable's impact on predictive performance but it would be good to clarify this.

Our reply: Thanks for your suggestion. Accordingly, we have substituted the paragraph:

These functions allow a selection of variables by calculating a score of importance through Efficiency Analysis Trees or Random Forest for Efficiency Analysis Trees, respectively. These importance scores represent how influential each variable is in the model. Regarding the available arguments of the functions, the user can specify the number of decimal units (digits) and include a barplot (from ggplot2) with the scores of importance (barplot). Additionally, the rankingEAT() function allows to display a horizontal line in the graph to facilitate the cut-off point between important and irrelevant variables (threshold).

with the following paragraph:

These functions allow a selection of variables by calculating a score of importance through Efficiency Analysis Trees or Random Forest for Efficiency Analysis Trees, respectively. These importance scores represent how influential each variable is in the model. Regarding the Efficiency Analysis Trees [RankingEAT()], the notion of surrogate splits by Breiman et al. (1984) was implemented. In this regard, the measure of importance of a variable x_j is defined as the sum over all nodes of the decrease in mean squared error produced by the best surrogate split on x_j at each node (see Definition 5.9 in Breiman et al., 1984). Since only the relative magnitudes of these measures are interesting for researchers, the actual measures of importance that we report are normalized. In this way, the most important variable has always a value of 100, and the others are in the range 0 to 100. As for the Random Forest for Efficiency Analysis Trees [RankingRFEAT()], Equation (9) was implemented for each input variable. Regarding the available arguments of the functions, the user can specify the number of decimal units (digits) and include a barplot (from ggplot2) with the scores of importance (barplot). Additionally, the rankingEAT() function allows to display a horizontal line in the graph to facilitate the cut-off point between important and irrelevant variables (threshold).

Reviewer's comment on the article:

2. Why does Free Disposal Hull overfit? It could help to provide a little more explanation to better delineate what EAT is doing differently.

Our reply: Thanks for your suggestion to improve the content of the manuscript. The Free Disposal Hull (FDH) technique is based on three microeconomic postulates (see the Background section). First, the technology determined by FDH satisfies free disposability in inputs and outputs. Second, it is assumed to be deterministic, that is, the production possibility set built by this technique always contains all the observations that belong to the data sample. Third, FDH meets the minimal extrapolation principle. This last postulate implies that FDH generates the smallest set that satisfies the first two postulates. Consequently, the derived efficient frontier is as close to the data as possible, generating overfitting problems. In contrast, the Efficiency Analysis Trees (EAT) technique meets the first two postulates but does not satisfy the minimal extrapolation principle. In this way, the EAT technique avoids overfitting problems. The difficulty for non-overfitted models lies in where to locate the production possibility set in such a way that it is close to the (unknown) technology associated with the underlying Data Generating Process. In our case, it is achieved through cross-validation and pruning.

This justification has been added to the Conclusion section of the updated version of the manuscript. In the current version, the first paragraph of that section is as follows:

The eat package allows the estimation of production frontiers in microeconomics and engineering through suitable adaptations of Regression Trees and Random Forest. In the first case, the package implements in R the so-called Efficiency Analysis Trees (EAT) by Esteve et al. (2020), which is a nonparametric technique that competes against the more standard Free Disposal Hull (FDH) technique. In this regard, the EAT technique overcomes the overfitting problem suffered by the FDH technique. FDH is based on three microeconomic postulates. First, the technology determined by FDH satisfies free disposability in inputs and outputs. Second, it is assumed to be deterministic, that is, the production possibility set built by this technique always contains all the observations that belong to the data sample. Third, FDH meets the minimal extrapolation principle. This last postulate implies that FDH generates the smallest set that satisfies the first two postulates. Consequently, the derived efficient frontier is as close to the data as possible, generating overfitting problems. In contrast, the Efficiency Analysis Trees (EAT) technique meets the first two postulates but does not satisfy the minimal extrapolation principle. This fact avoids possible overfitting problems. The difficulty for non-overfitted models lies in where to locate the production possibility set in such a way that it is close to the (unknown) technology associated with the underlying Data Generating Process. In the case of EAT, it is achieved through cross-validation and pruning. A subsequent convexification of the EAT estimation of the technology, known as CEAT by its acronym, yields an alternative estimate of the production possibility set in contrast to the traditional Data Envelopment Analysis (DEA) technique. In the second case, an ensemble of tree models is fitted and aggregated with the objective of achieving robustness in the estimation of the production frontier (Esteve et al., 2021).

Reviewer's comment on the article:

3. There is a LateX error in the table reference on page 11.

Our reply: Thanks. Done.

Reviewer's comment on the article:

4. The wording on page 7 ("will denote hereinafter the multi-dimensional predictor defined from $T^*(\aleph)$") is a little unclear. "Predictor" is often synonymous with feature or covariate, whereas here it seems to refer to predicted value unless I am mistaken.

Our reply: Thanks for the comment. We agree with you. We have substituted the word 'predictor' with 'estimator' in that sentence and other similar parts of the text. In the current version of the manuscript, we prefer to use the word 'predictor' only to denote features or covariates (i.e., inputs in our production context).

Reviewer's comment on the package:

The package's source code has been written to a high degree of technical quality. The essential standards of software development are met, including functionalized, clean code, consistent style, logically organized relationships between functions, built-in documentation, and helpful comments.

Our reply: Thanks for your positive comments.

Thanks for your help with the manuscript.

Regards,

Miriam Esteve
Center of Operations Research
Miguel Hernandez University
Elche, Spain
miriam.estevec@umh.es