The Journal of Open Source Software

Stacked Ensemble Modeling with Tidy Data Principles

Simon P. Couch[1, 2] **and Max Kuhn**[2]

1 Johns Hopkins, Department of Biostatistics **2** RStudio PBC

DOI:

Software

- Review
- Repository
- Archive

Submitted:
Published:

Summary

Model stacking is an ensemble modeling technique that involves training a model to combine the outputs of many constituent statistical models. {stacks} is a free and open-source R software package for stacked ensemble modeling that aligns with tidy data principles. The package's functionality is closely aligned with the {tidymodels}, a collection of packages providing a unified interface to diverse statistical modeling techniques. Beyond simply providing a mathematically robust interface to build stacked ensemble models, {stacks} adheres to a consistent grammar to interface with two object classes that promote an intuitive understanding of the underlying implementation.

Statement of Need

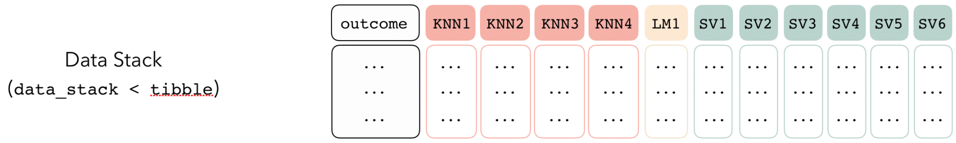
Ensemble learning (otherwise referred to as *model stacking*) is an increasingly popular approach in statistical and machine learning that involves training a models to generate predictions informed by many constituent models ("members"). Model stacking has been shown to increase predictive performance in a variety of settings and has thus greatly risen in popularity in recent years. In a spring 2020 community survey, the advanced statistical modeling technique was ranked as the first priority for future development among users of the {tidymodels}, a burgeoning software ecosystem for tidyverse-aligned predictive and inferential modeling used across many modern research and industrial applications (Kuhn & Wickham, 2022). {stacks} introduces model stacking to the {tidymodels}.

Packages implementing methods for predictive and inferential modeling in R are highly variable in their interfaces. The structure of inputted data, argument names, expected argument types, argument orders, output types, and spelling cases varies widely both within and among packages. This diversity in approaches obscures the intuition shared among common modeling procedures, makes details of usage difficult to remember, and prevents an expressive and idiomatic coding style. In contrast, {stacks} utilizes the consistent and unified interface of the {tidymodels} packages to implement a generalized and concise grammar for model stacking. The package supports ensembling using any member model type, cross-validation scheme, and error metric implemented in—or in alignment with—the {tidymodels}.

The principled and generalized approach of the package lends itself to diverse applications of predictive modeling. This capability, for one, was recognized with the receipt of the American Statistical Association's 2021 John M. Chambers Award, awarded "for the development and implementation of computational tools for the statistical profession by a graduate or undergraduate student" (ASA Section on Statistical Computing, 2021). The package's functionality has also been shared in venues such as R/Pharma 2020, rstudio::global(2021), and the 2021 Joint Statistical Meetings. To date, the package has been downloaded more than 10,000 times, evidencing its key contribution to a software ecosystem utilized in diverse research contexts.

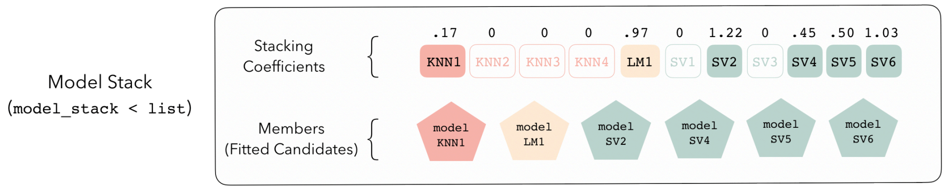
Underlying Principles

At a high level, stacked ensemble models in {stacks} are formed from *model definitions*, which specify a set of instructions to fit a model or set of models using {parsnip}, {recipes}, and {workflows}. Each model specified in a model definition is referred to as a *candidate member* in the package. To be used in the same model stack, all model definitions must share a *resampling object*, as defined in the {rsample} package, such as a cross-fold validation or set of bootstrap samples. After initializing a `data_stack` object with the `stacks()` function, model definitions can be iteratively added to the data stack with `add_candidates()`. This function collates the predictions on the assessment set, a subset of the training data used for preliminary model validation, from each candidate specified in the model definitions to the data stack.



The above diagram represents a data stack containing 11 candidate members, where 4 come from a shared K-nearest neighbors model definition, 1 arises from a linear model, and 6 come from a shared support vector machine model definition. Model definitions can result in multiple candidates when several possible hyperparameter values are being optimized over (e.g. the number of neighbors k or the precision parameter for the support vector machine's radial basis function).

After all candidates are collated to the data stack, the user can fit the meta-learner using the `blend_predictions()` function, which fits an elastic net model on the data stack, predicting the true outcome using the predictions from each of the candidate members. The coefficients of this model form the *stacking coefficients*, which are the weightings for each of the member models in ultimately predicting the outcome. Candidate members with non-zero stacking coefficients are *members*, and must be fitted on the entire training set with the `fit_members()` function. This function outputs a `model_stack` object, and is ready to predict on new data.



In addition to the aforementioned core verbs, the package supplies several helper functions to interface more effectively with model stacks. Notably, `collect_parameters()` juxtaposes model parameters with their stacking coefficients, and an `autoplot()` S3 method provides model diagnostic visualizations using {ggplot2}.

Comparison to Other Packages

A number of software packages on the Comprehensive R Archive Network share functionality with `stacks` (R Core Team, 2022). Notably, of course, the package integrates tightly with other packages in the {tidymodels} ecosystem, such as {tune}, {parsnip}, {rsample}, and {recipes}.

There are also other packages providing implementations of model ensembling in R. The {h2o} R package ports functionality from the H2O modeling ecosystem to R via a REST API, including an implementation of ensembling that supports meta-learners beyond the generalized linear model supported by {stacks} (LeDell et al., 2022). The {SuperLearner} package also provides an implementation of model ensembling, providing its own wrappers for member model types and also supplying a number of different options for meta-learners (Polley, LeDell, Kennedy, & van der Laan, 2022). Each of these packages differ in the number of model types, error metrics, and cross-validation schemes supported, as well as the modeling behaviors encouraged in their interfaces.

Acknowledgements

We acknowledge contributions from Julia Silge, Hannah Frick, and Asmae Toumi. Development of this package was supported by RStudio PBC, the Reed College Mathematics Department, and the American Statistical Association Sections on Statistical Computing & Statistical Graphics.

References

ASA Section on Statistical Computing. (2021). *John M. Chambers Statistical Software Award*. Retrieved from https://community.amstat.org/jointscsg-section/awards/john-m-chambers

Kuhn, M., & Wickham, H. (2022). tidymodels: A collection of packages for modeling and machine learning using tidyverse principles. *Boston, MA, USA*. Retrieved from https://tidymodels.org

LeDell, E., Gill, N., Aiello, S., Fu, A., Candel, A., Click, C., Kraljevic, T., et al. (2022). *h2o: R interface for the 'H2O' scalable machine learning platform*. Retrieved from https://CRAN.R-project.org/package=h2o

Polley, E., LeDell, E., Kennedy, C., & van der Laan, M. (2022). *SuperLearner: Super Learner Prediction*. Retrieved from https://CRAN.R-project.org/package=SuperLearner

R Core Team. (2022). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/