## Visually Exploring Random Forests with ggRandomForests

## John Ehrlinger

Department of Quantitative Health Sciences Lerner Research Institute Cleveland Clinic john.ehrlinger@gmail.com

**Keywords:** Machine Learning, Random Forests, survival analysis, **ggplot2**, **randomForestSRC** 

Random Forests [1] (RF) are a fully non-parametric statistical method requiring no distributional assumptions on covariate relation to the response. RF are robust, optimizing predictive accuracy by fitting an ensemble of trees to stabilize model estimates. RF utilizes all variables in predicting the specified outcome, effectively weighting the most important covariates by assessing their impact on separating dissimilar groups of observations. Random Forests for survival [3, 5] (RF-S) are an extension of RF techniques to survival settings, allowing efficient non-parametric analysis of time to event data. The **randomForestSRC** [4] package is a unified treatment of Breiman's random forests for survival, regression and classification problems.

Predictive accuracy make RF an attractive alternative to parametric models, though complexity and interpretability of the forest hinder wider application of the method. We introduce the **ggRandomForests** package, an implementation of **ggplot2** [7] graphics for exploring **randomForestSRC** objects. Using both classification (RF-C) and survival (RF-S) examples from our research at the Cleveland Clinic, we will demonstrate the **randomForestSRC** package. We use Variable Importance measure (VIMP) [1] as well as Minimal Depth [6], a property derived from the construction of each tree within the forest, to assess the impact of variables on forest prediction. We will also demonstrate the use of variable dependence plots [2] to aid interpretation RF results in different response settings.

## References

- [1] Breiman, L. (2001). Random Forests. *Machine Learning* 45(1), 5–32.
- [2] Friedman, J. H. (2000). Greedy function approximation: A gradient boosting machine. *Annals of Statistics* 29, 1189–1232.
- [3] Ishwaran, H. and U. B. Kogalur (2007). Random survival forests for R. R News 7, 25-31.
- [4] Ishwaran, H. and U. B. Kogalur (2013). Random Forests for Survival, Regression and Classification (RF-SRC), R package version 1.4.
- [5] Ishwaran, H., U. B. Kogalur, E. H. Blackstone, and M. S. Lauer (2008). Random survival forests. *The Annals of Applied Statistics* 2(3), 841–860.
- [6] Ishwaran, H., U. B. Kogalur, E. Z. Gorodeski, A. J. Minn, and M. S. Lauer (2010). High-dimensional variable selection for survival data. *J. Amer. Statist. Assoc.* 105, 205–217.
- [7] Wickham, H. (2009). ggplot2: elegant graphics for data analysis. Springer New York.