Variable Importance Confidence Intervals within Random Forest

DR. HEATHER COOK, UNIVERSITY OF SOUTHERN INDIANA (PRESENTER)

DR. DANIEL KEENAN, UNIVERSITY OF VIRGINIA

DR. DOUGLAS LAKE, UNIVERSITY OF VIRGINIA

Background: Random Forest Steps

- Select the number of decision trees to build
- For each tree:
  - Select a random sample with replacement
  - Build a decision tree and for each split:
    - Randomly select k predictors
    - Select the best predictor among those k selected to split the data
  - Observations out-of-bag (OOB) used to calculate variable importance (VIMP) per predictor
- Collectively, these trees create the forest
- Per variable, the VIMP is aggregated over all the trees

Background: Issues

Bootstrapping cannot be directly implemented to calculate VIMP confidence intervals

- Random forest already uses bootstrapping
  - Cannot guarantee that the OOB samples will be OOB and not also used to grow the tree
  - Currently available VIMP confidence interval methods are complex

Goals

1. Focus on standard R packages for random forest
   - *randomForest*
   - *randomForestSRC*

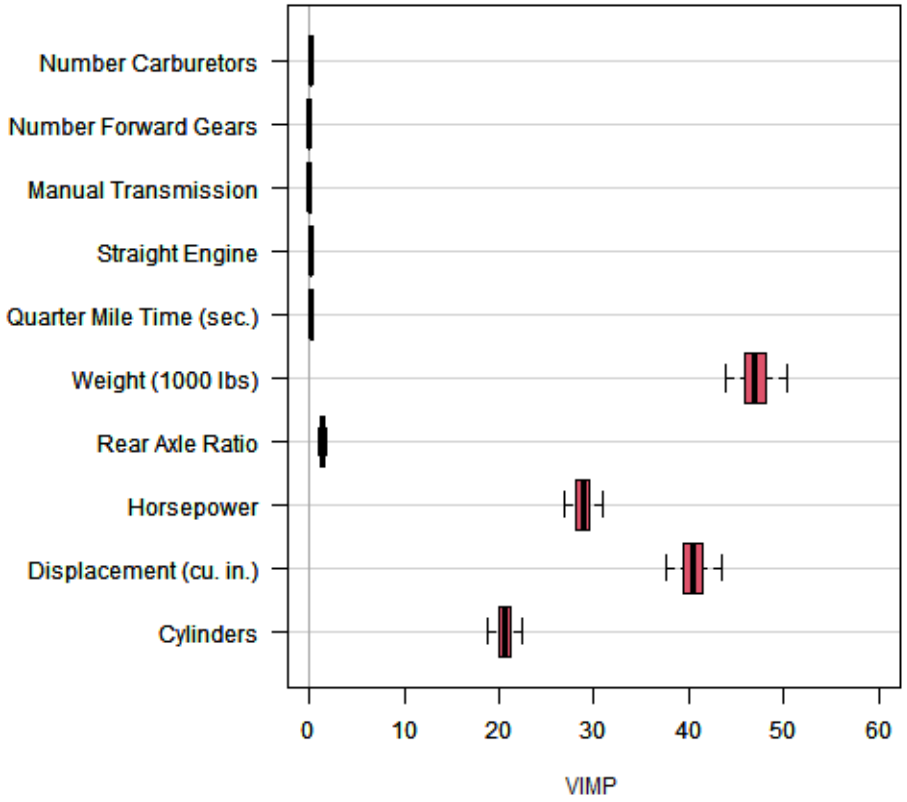
2. Explain our new method of calculating VIMP confidence intervals within a random forest model

3. Compare our new method to existing methods of VIMP confidence intervals
   - Existing methods (Ishwaran & Lu, 2018, “Standard Errors and Confidence Intervals for Variable Importance in Random Forest Regression, Classification, and Survival.”)

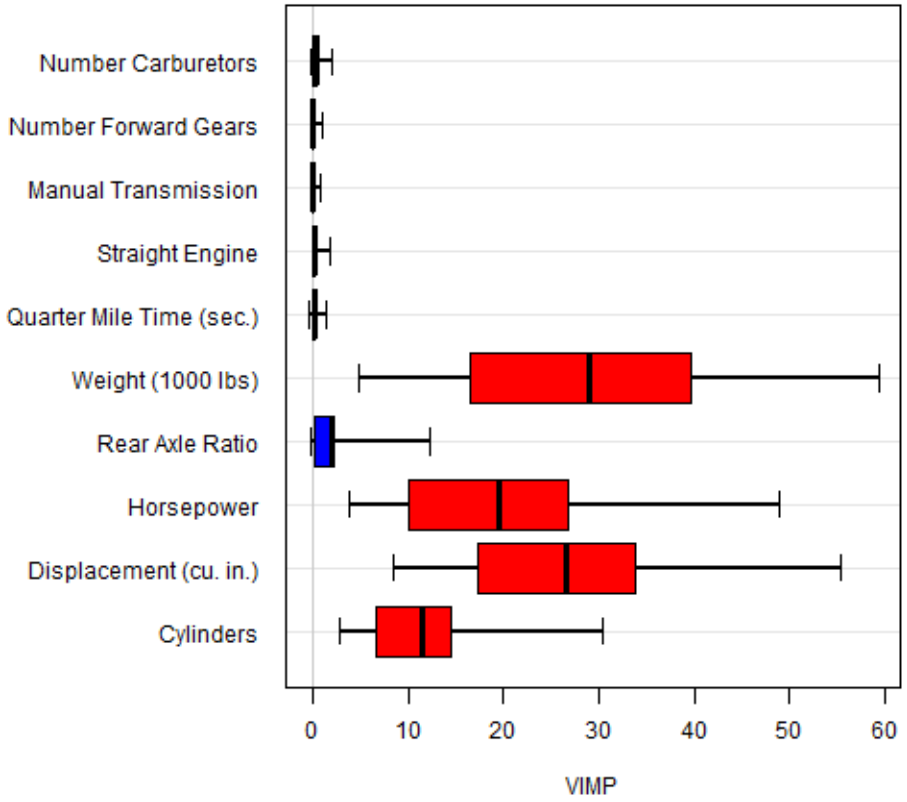
Our Method

1. Train a random forest for parameters
2. Create the random forest model with the selected parameter values
3. Extract the VIMP per each tree
4. Implement bootstrapping with the per tree VIMP values for each variable
   i. Take a random sample with replacement of the per tree VIMP values
      - Calculate the mean VIMP from these values
   ii. Repeat the previous step several times, say 1000 times
   iii. Take the 2.5^{th} and 97.5^{th} percentiles of these 1000 means to create the 95% confidence interval for a variable's importance

OUR METHOD RESULTS

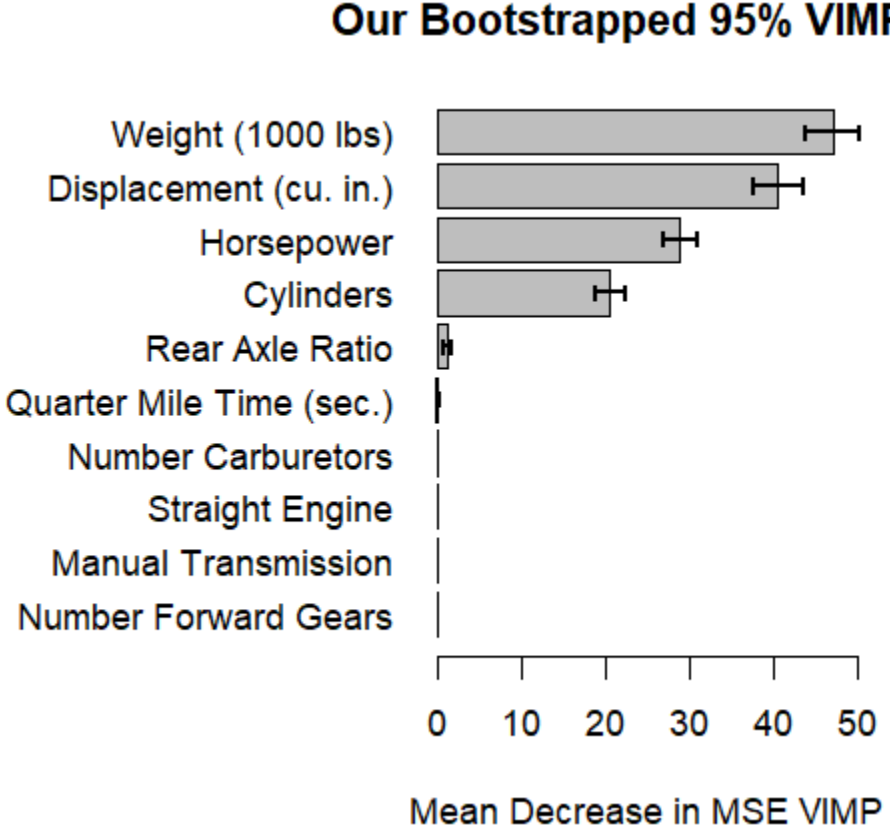


ONE EXISTING METHOD RESULTS



Conclusion

- Addition to interpretations of predictors and their order of importance:
  1-2. Weight or displacement (overlap)
  3. Horsepower
  4. Number of cylinders
- Our method vs current methods is:
  - Faster than current methods
  - Easier to compute
  - Easier to plot and manipulate results in R



Future Work & Author Contacts

- Explore behavior via simulations & further compare to current methods
- Release R code to the public

Dr. Heather Cook
Assistant Professor of Statistics
University of Southern Indiana
Department of Mathematical Sciences

8600 University Boulevard
Evansville, IN 47712
USA

hlcook1@usi.edu

Dr. Daniel Keenan
Professor Emeritus
University of Virginia
Department of Statistics

1827 University Avenue
Charlottesville, VA 22904
USA

dmk7b@virginia.edu

Dr. Douglas Lake
Professor
University of Virginia
Department of Medicine, Cardiovascular Medicine

1827 University Avenue
Charlottesville, VA 22904
USA

del2k@virginia.edu

References

- Archer, Kellie J.; Kimes, Ryan V. “Empirical characterization of random forest variable importance measures.” 2008
- Grömping, Ulrike. “Variable Importance Assessment in Regression: Linear Regression versus Random Forest.” 2009
- Ishwaran, Hemant; Lu, Min. “Standard Errors and Confidence Intervals for Variable Importance in Random Forest Regression, Classification, and Survival.” 2018
- Janitza, Silke; Strobl, Carolin; Boulestix, Anne-Laure. “An AUC-based permutation variable importance measure for random forests.” 2013
- Sandri, Marco; Zuccolotto, Paola; “A Bias Correction Algorithm for the Gini Variable Importance Measure in Classification Trees.” 2008
- Strobl, Carolin; Boulesteix, Anne-Laure; Zeileis, Achim; Hothorn, Torsten. “Bias in random forest variable importance measures: Illustrations, sources and a solution.” 2007