DATA ANALYTICS 2020

Comparing Variable Importance in Prediction of Silence Behaviours between Random Forest and Conditional Inference Forest Models

Stephen Barrett

Technological University Dublin

Geraldine Gray

Technological University Dublin

Colm McGuinness

Technological University Dublin

Michael Knoll

University of Leipzig

Presenter: Dr. Geraldine Gray

- A Senior Lecturer in Informatics at Technological University Dublin, specialising in Data Science.
- Co-ordinator for TU Dublin's MSc in Computing in Applied Data Science and Analytics.
- Active in research related to psychometric data analysis for a number of years.
- Email: geraldine.gray@tudublin.ie

Note: The corresponding author for this paper is Dr. Stephen Barrett, email: S.Barrett@live.ie

CONSTRUCTS FROM QUESTIONAIRE

Acquiescent

Quiescent

Pro-Social

Opportunistic

Diffident

Disengaged

Authenticity

Psychological Safety Climate

Health & Mental Health

Organizational Citizenship Behavior

Humane Orientation

Gender Egalitarianism

Collectivism

Uncertainty Avoidance

Future Orientation

Performance Orientation

Power Distance

MODELS APPLIED

Random Forest

- Builds many trees on bootstrapped samples with replacement. Each tree is allowed to grow fully without being pruned back, producing many over-fit trees.
- A second step allows the model to split on a random selection of attributes at each node for each tree, decorrelating the trees
- The two randomization steps result in different trees over-fitting different sections of the dataset. Classification is based on a majority vote amongst all trees.
- Using the Gini Index as a splitting metric biases attribute selection towards attributes with more variation in the predictor space, more missing values or variables with many factor levels.

Conditional Inference Forest

- Use Conditional Inference Trees (CIT) as a base for the Random Forest.
- In CIT Attribute selection verses splitting criteria are separated removing bias in selection of attributes. A global statistical test is performed for where the null hypothesis states there is no difference between X_n and Y. If global null hypothesis rejected. A Bonferroni corrected Association test is undertaken for each attribute to see if there is a relationship with Y.
- The attribute to split on is selected based on the lowest p-value. The splitting point is then determined in the usual manner.
- In conditional Inference Forest, this method is applied across many trees where each tree is exposed to a subset of attributes and bootstrapped samples.

INTERPRETATION METHODS

Variable Importance (RF)

- Permutate column to break relationship between independent variables and dependant variable.
- The difference in accuracy is recorded per tree and aggregated across all the trees in the forest
- Bias towards variables with many unique values/missing values.
- Correlations inflate predictor performance.

Variable Importance (CIF)

- Permutate within the segmented data conditioned on the split point.
- Permutate columns together which have a correlation greater than 0.2 (default within cforest)
- Permutating on the conditional variable importance
- The difference of out of bag (OOB) error between permutated and non permutated trees and aggregated up to the Forest

Partial Dependency Plots

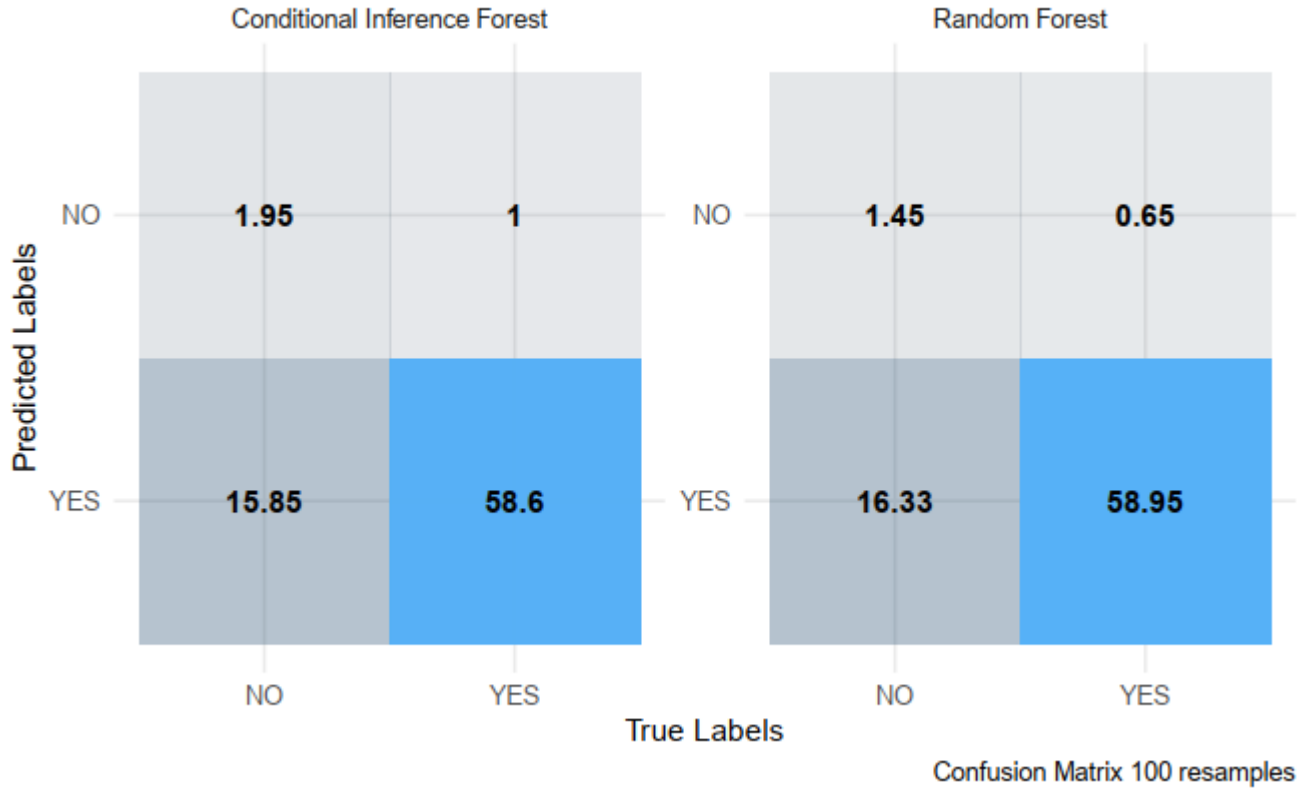
- Create a value range for the attribute you are interested in
- For every value in that range duplicate the dataset
- Insert the value into every row, run the model and record the output for each record. Take the mean of that result to produce a Partial Dependency Score.
- Problems with correlation which result in extrapolation

MODEL TUNING AND RESULTS

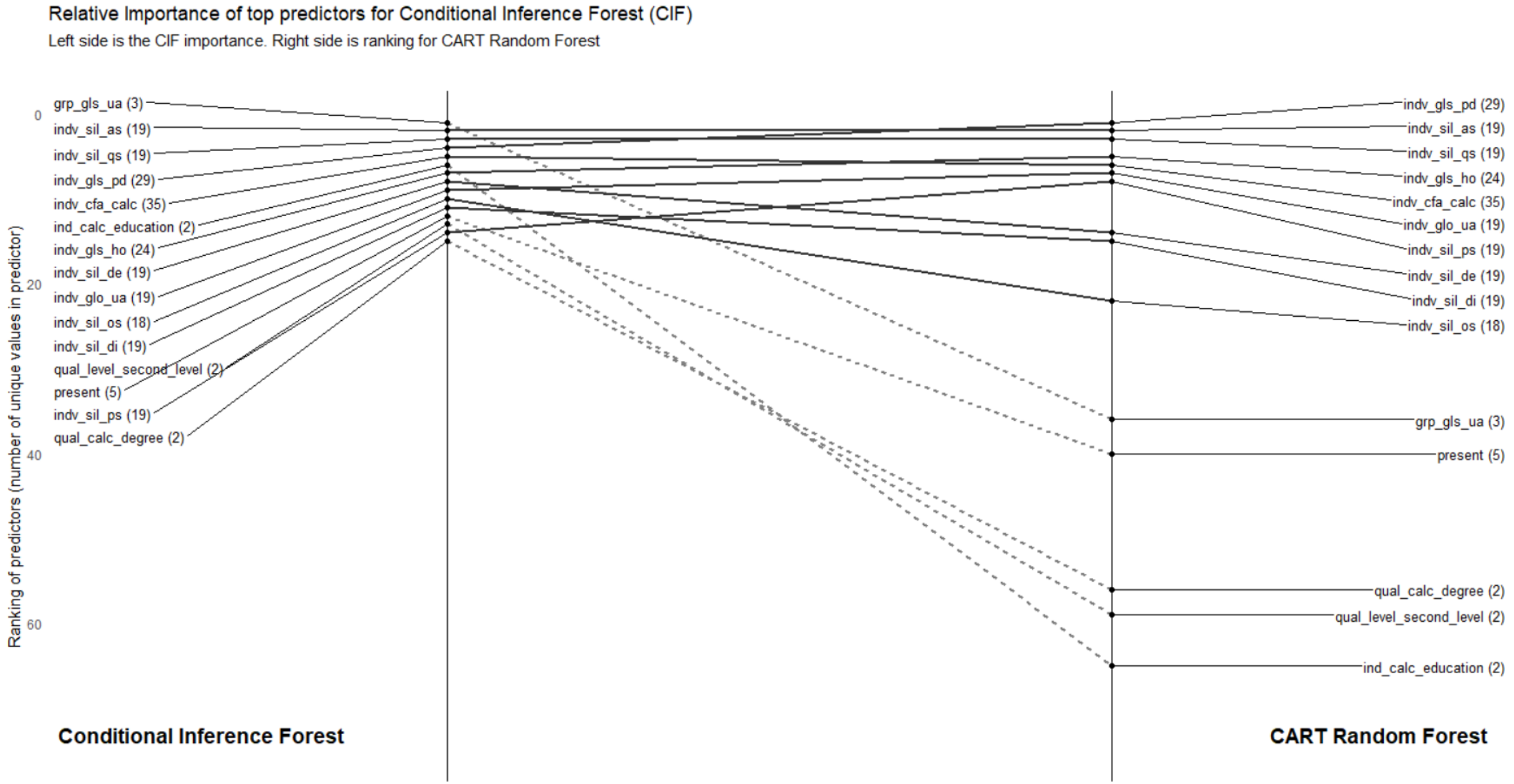
- **Random Forest** varied *mtry* from 2 to 80 with minimum node size from 2 to 20. Optimal values were 17 for *mtry* and 2 for node size. Model ***AUC was 0.65***

- Conditional Inference Forest was tuned across the same *mtry* range where the optimal *mtry* value was found to be 57. All other tuning parameters were left at their defaults. Model ***AUC was 0.65***

- Models had difficulty in identifying if someone would not engage in Silence. This could be attributed to the imbalanced nature of the dataset.

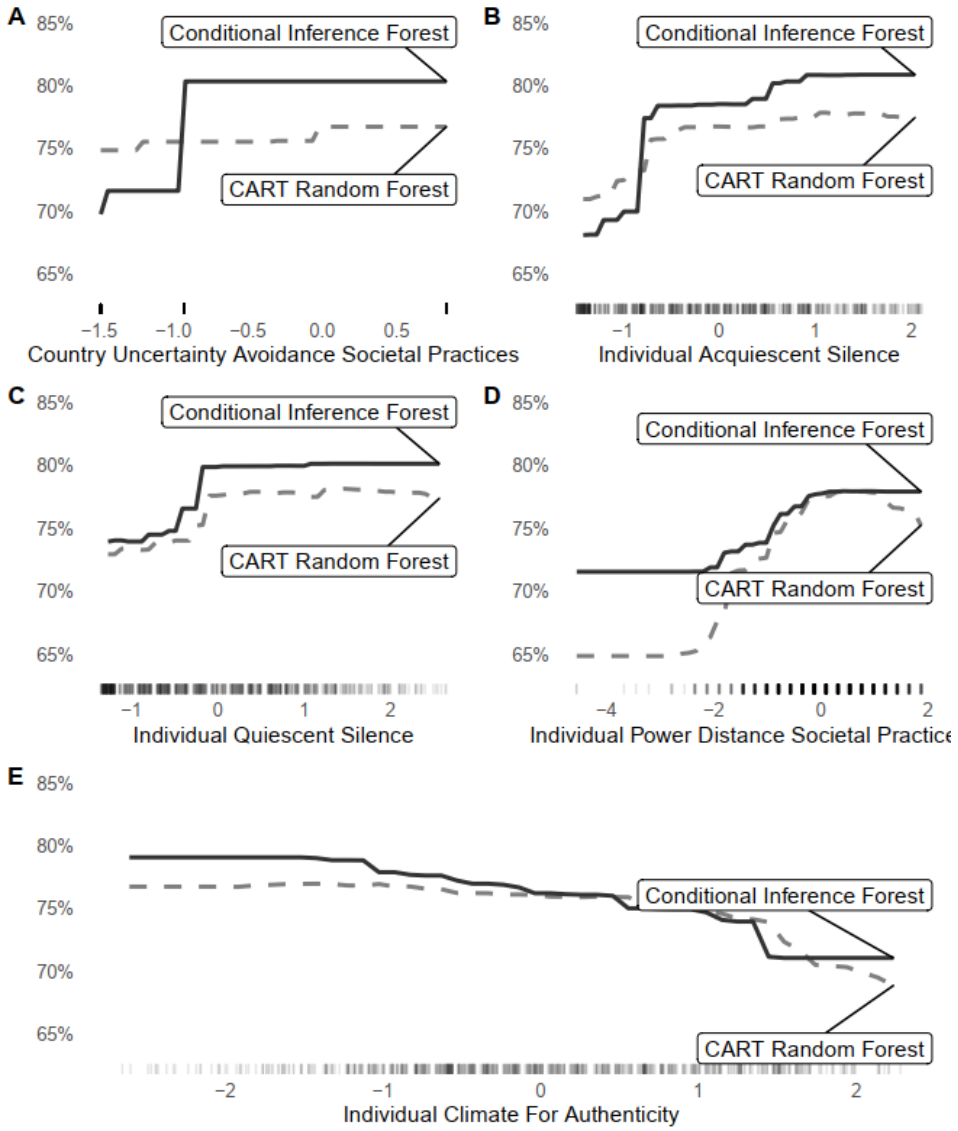


Top 15 ranked most important variables



*Dotted lines are predictors with less than 6 unique variables

INSIGHTS



- The more a society tries to avoid uncertainty the higher the probability of someone engaging in Silence **(A).**
- The more a person thinks their feedback will result no discernible change the higher their chances of engaging in Silence **(B).**
- The more a person fears that their feedback might result in negative consequences, the higher the probability they will engage in Silence **(C).**
- The higher the perceived power distance by an individual the less likely they are to voice their concerns **(D)**
- If an individual is part of a working group where they are free to be authentic to themselves and have the ability to express this, their probability of engaging in Silence is reduced **(E).**

CONCLUSIONS

Specific

- Random Forests in conjunction with PDPs can be used with variable importance measures to highlight non linear relationships between predictors and target variables.
- A CART based random forest showed a bias for predictors with more values. A CIT based forest did not have the same bias.

General

- Where the predictor space has varying number of distinct values per predictor, and model interpretation is the goal of the analysis, Conditional Inference Forest is better than Random Forest for exploring variable importance.
- This finding is particularly pertinent for researchers who wish to use tree based modelling for survey data where the questions pertaining to the constructs have a different number of available options.

Study limitations

- Conclusions are based on models with week AUC scores. Further work is needed to determine their generalisability.