Filling Missing Values for Incomplete Data Sets

The Client Challenge

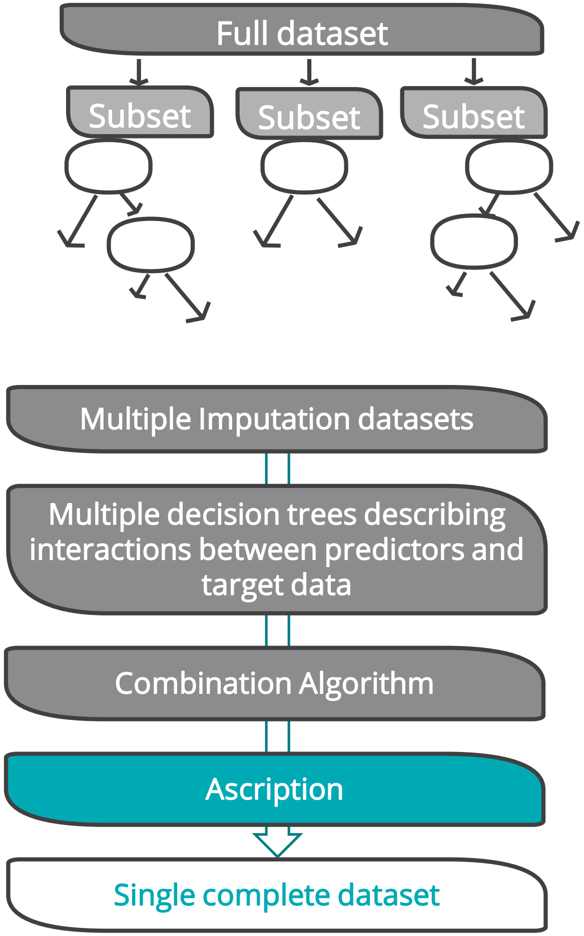
Our client, a global digital data collection enterprise, had a critical data set that was missing inputs. To ensure that the Client could report on important KPIs using robust sample sizes, we applied a mathematical technique we call 'ascription'. It enables the Client to develop full information for the entire sample employing data provided by only part of it.

The GemSeek Approach

GemSeek tackled this particular task utilizing the Conditional Random Forest Algorithm to predict values for all missing entries.

The Conditional Random Forest Algorithm revolves around training many relatively inaccurate decision trees and then using the information each one provides to estimate a very accurate final estimate. This form of the algorithm is well suited for problems with multiple descriptive variables, but few observations. There is no necessity to specify in advance a fixed list of predictor variables and interaction to infer the variable of interest. Output is in the form of a single complete dataset. Additionally a report card can be produced including importance measure per predictor variable.

The Deliverables



Full dataset

Subset

Subset

Subset

Multiple Imputation datasets

Multiple decision trees describing interactions between predictors and target data

Combination Algorithm

Ascription

Single complete dataset