MISSING DATA IMPUTATION FOR SUPERVISED LEARNING

Jason Poulos and Rafael Valle

University of California, Berkeley

Supplementary Material

This file contains descriptive plots of feature correlation and missing data patterns in the benchmark datasets; plots of Bayesian hyperparameter optimization for training ANNs; and test set error plots for classifiers trained on MNAR-perturbed data.

S1 Descriptive plots

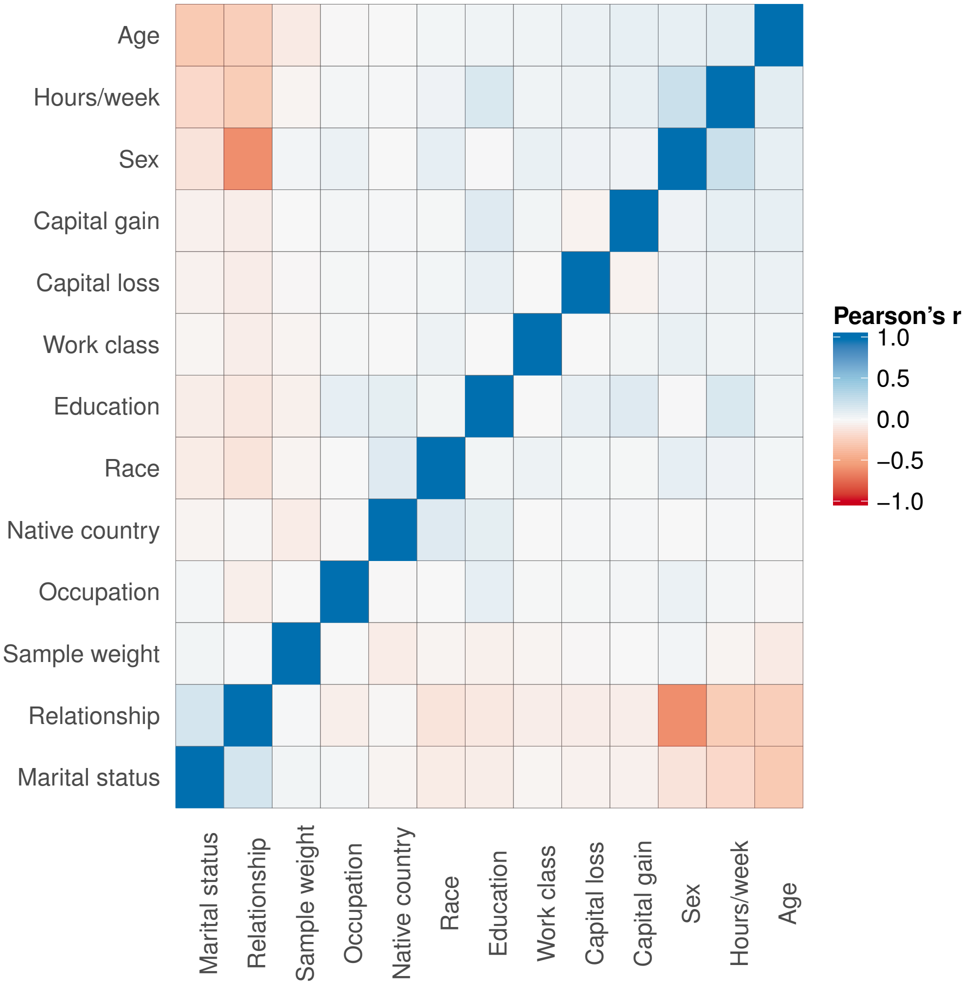

Figure 1: Correlation matrix for Adult training set. Pearsoepsn product-moment correlation coefficients (Pearson's r) are computed with listwise-deletion of missing values.

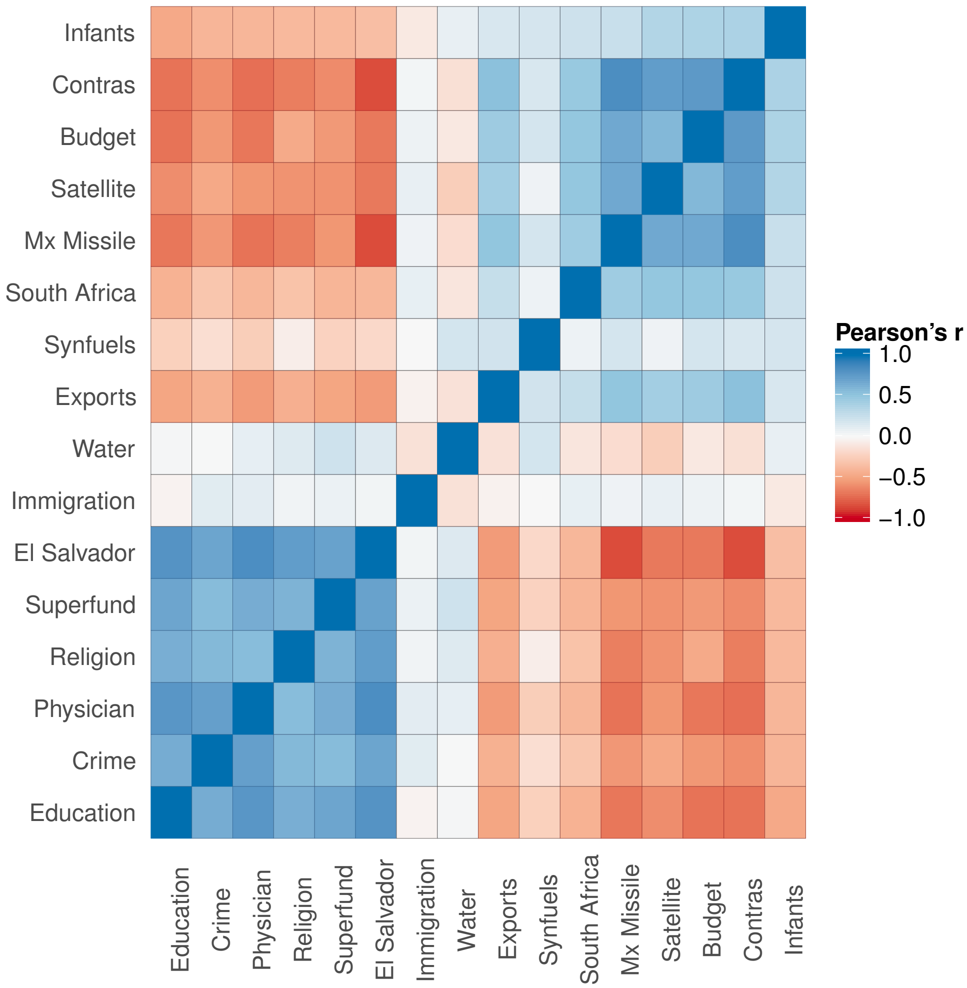

Figure 2: Correlation matrix for CVRs training set. See footnotes for Figure SM-1.

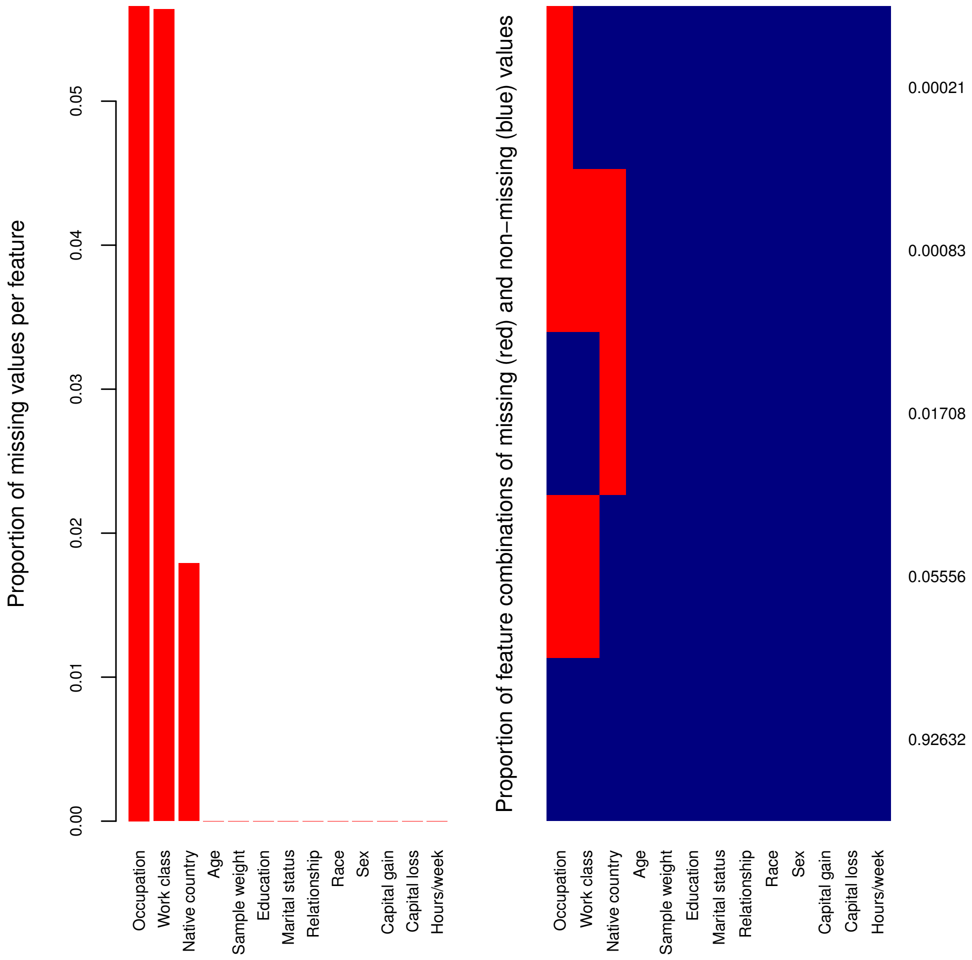

Figure 3: Histogram of proportion of missing values in each feature (Left) of Adult training set and aggregation plot of all existing combinations of missing and non-missing values in the samples (Right).

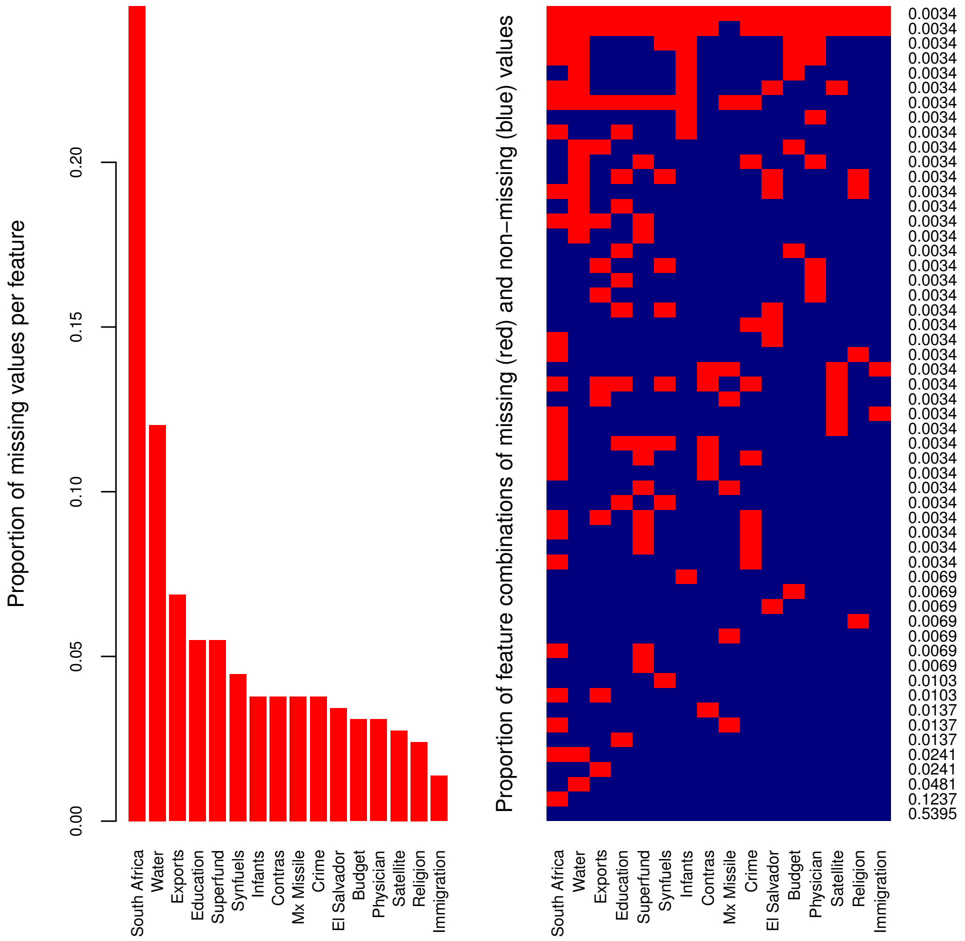

Figure 4: Histogram of proportion of missing values in each feature (Left) of CVRs training set and aggregation plot of all existing combinations of missing and non-missing values in the samples (Right).

S2 Bayesian hyperparameter optimization

The goal of Bayesian optimization is to choose a point in the hyperparameter space that appropriately balances information gain and exploitation. Figure SM-5 shows the exploration of hyperparameter space during Bayesian optimization for both Adult and CVRs datasets. Each circle represents a candidate ANNs classifier trained on a differently imputed and perturbed dataset. More circles appear in the plot for CVRs simply due to the fact that the training set is smaller. We see that most of the candidate models use dropout and have an initial learning rate close to the maximum of 0.01. The plurality of candidate models appear to either have momentum (1) or not (0).

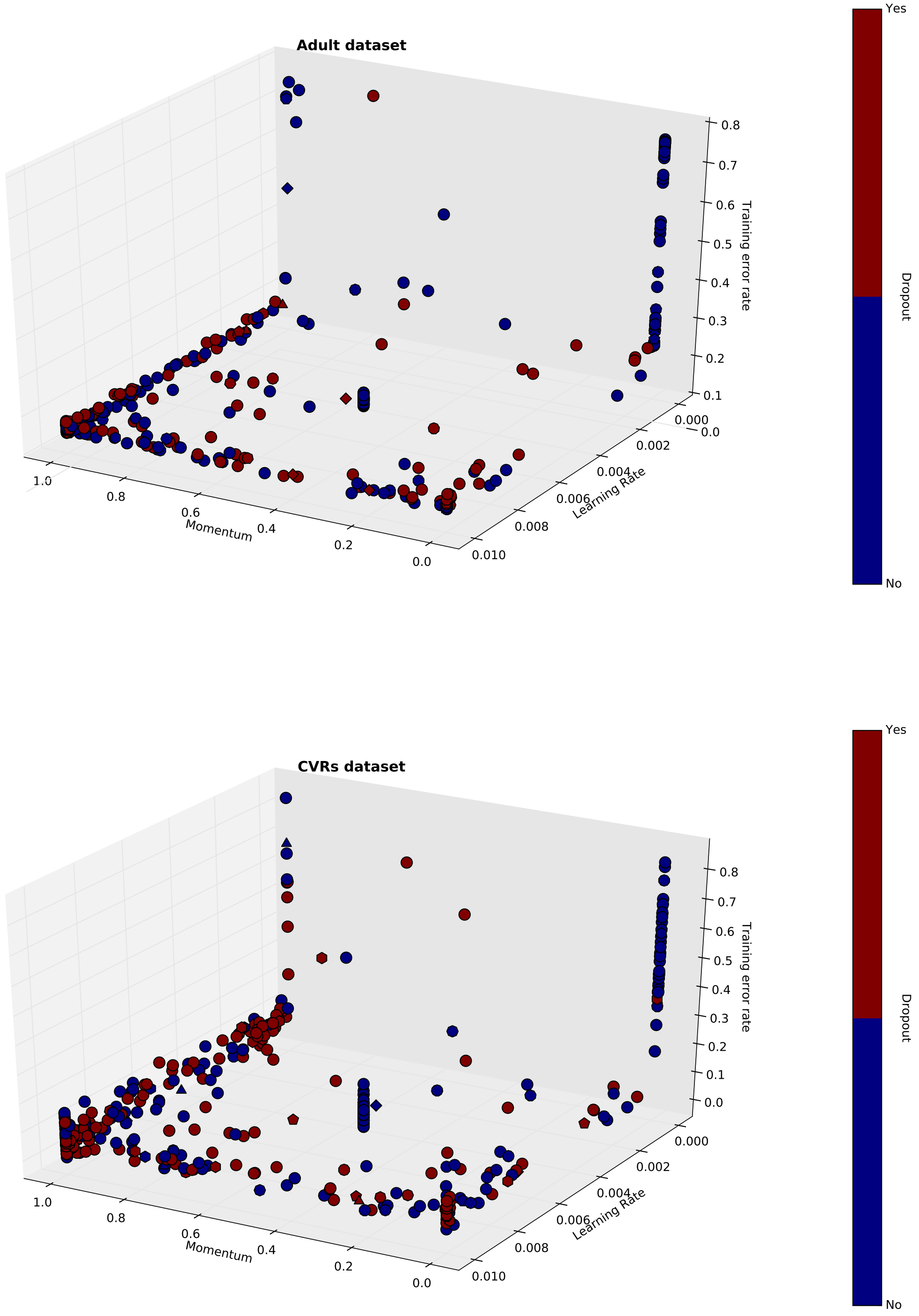

Figure 5: Exploration of hyperparameter space during Bayesian optimization. Each circle represents a candidate ANNs classifier trained on a differently imputed and perturbed dataset.

S3 Results with MNAR perturbation

We perturb the training data according to the MNAR mechanism

$$\Pr(M_{ij} = 1 | y_{ij}, \phi) = \begin{cases} \delta, & \text{if } y_{ij} \in A \\ 0, & \text{otherwise,} \end{cases}$$

where A is a vector containing at least one value from each categorical feature that we determine likely to be missing. We select categorical values in the Adult dataset that are theoretically correlated with low socioeconomic status, such as the values “Without pay” and “Never worked” for the feature *Work class.* The existing literature suggests item nonresponse in surveys is correlated with low income and low education (Rubin et al., 1995). We include in A only “nay” votes, under the assumption that refusing to take position on an issue or missing a vote is akin to voting against the issue.

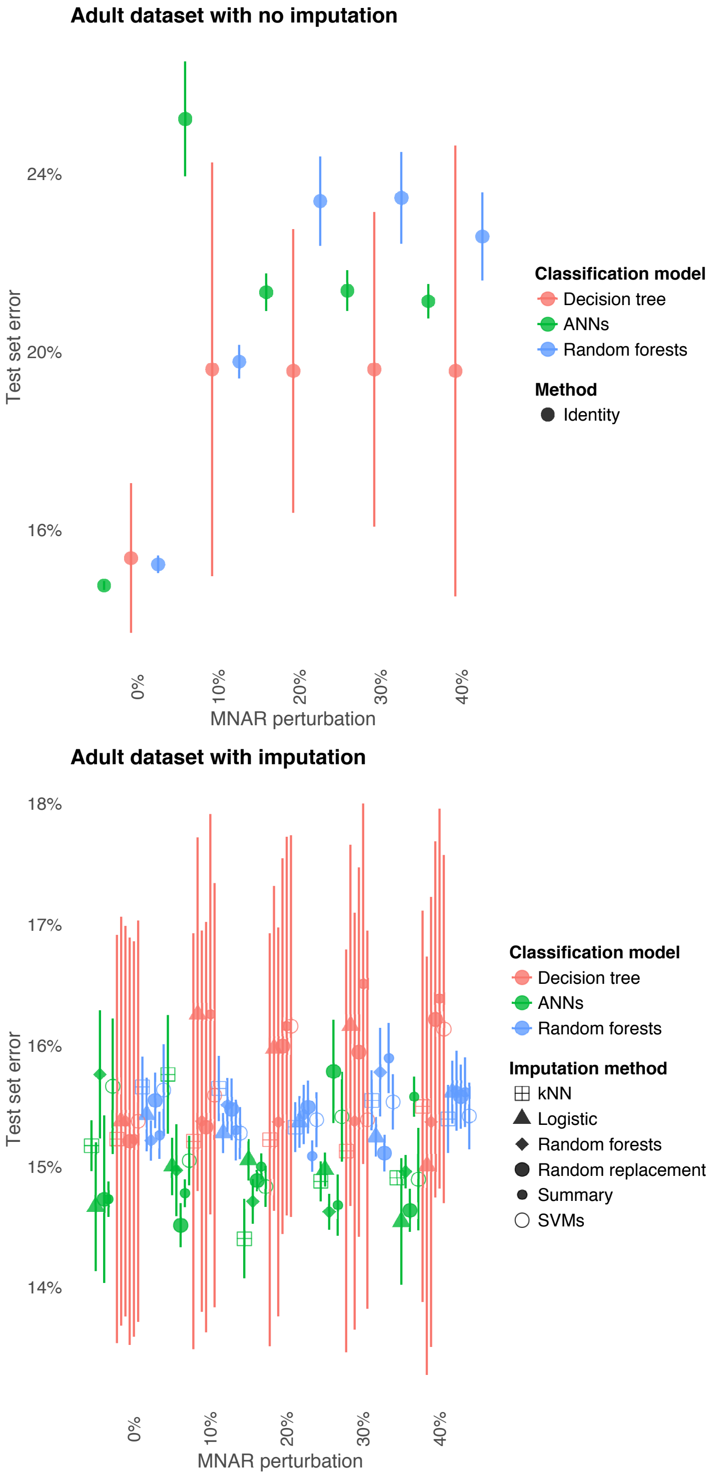

Figure 6: Error rates on the Adult test set with (bottom) and without (top) missing data imputation, for various levels of MNAR-perturbed categorical training features (x-axis). Error bars represent one standard deviation from the test error prediction.

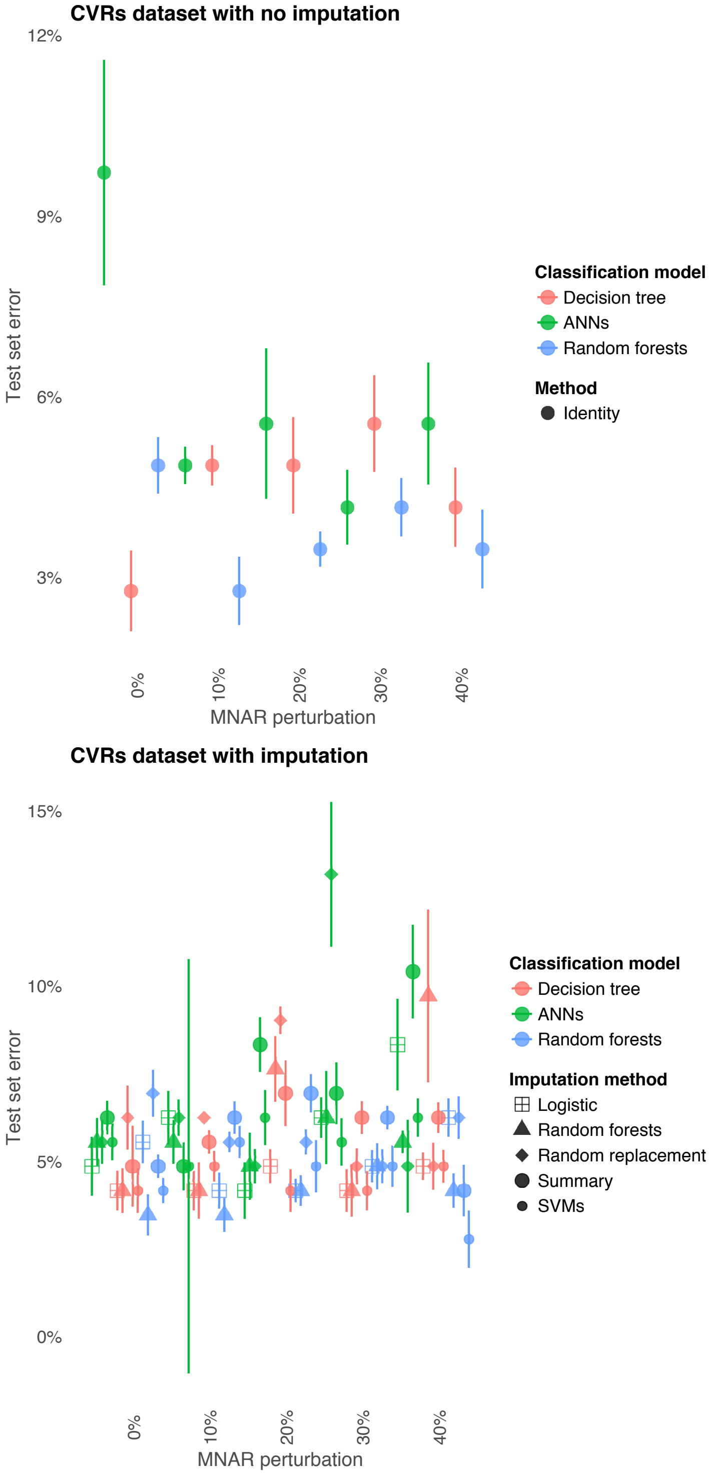

Figure 7: Error rates on the CVRs test set with (bottom) and without (top) missing data imputation. See footnotes for Figure SM-6.