MISSING DATA IMPUTATION FOR SUPERVISED CLASSIFICATION

Jason Poulos and Rafael Valle

University of California, Berkeley

Supplementary Material

This file contains descriptive plots of missing data patterns in the benchmark datasets, a description and plots of Bayesian hyperparameter optimization for training neural networks on the benchmark datasets, and test error plots for classifiers trained on MCAR-perturbed categorical training features.

S1 Patterns of missing data

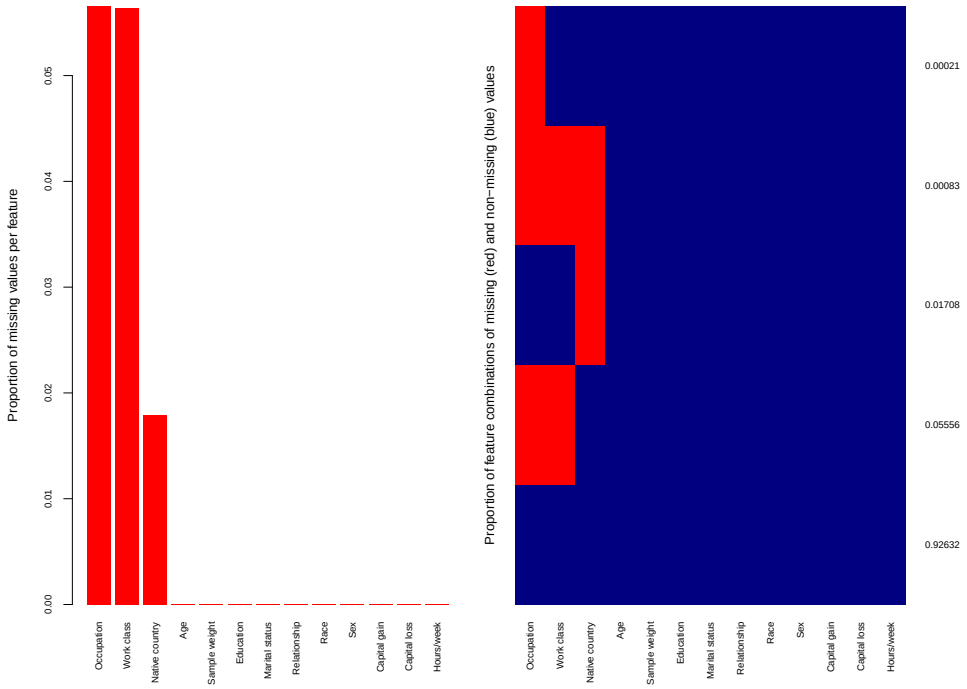


Figure 1: Histogram of proportion of missing values in each feature (Left) of Adult training set and aggregation plot of all existing combinations of missing and non-missing values in the samples (Right).

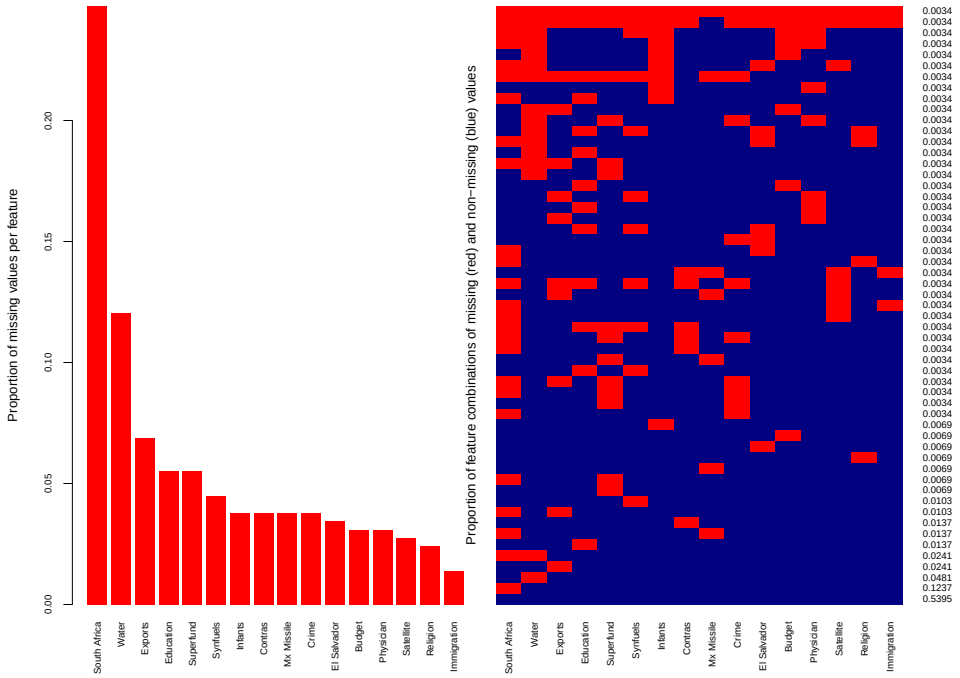
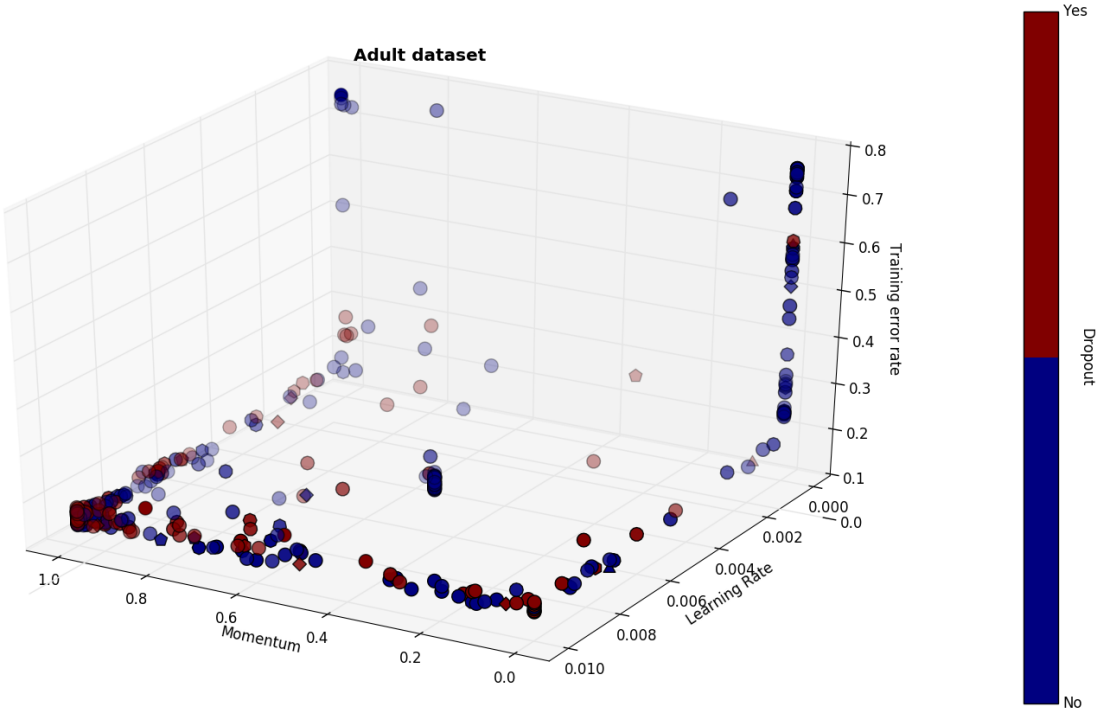
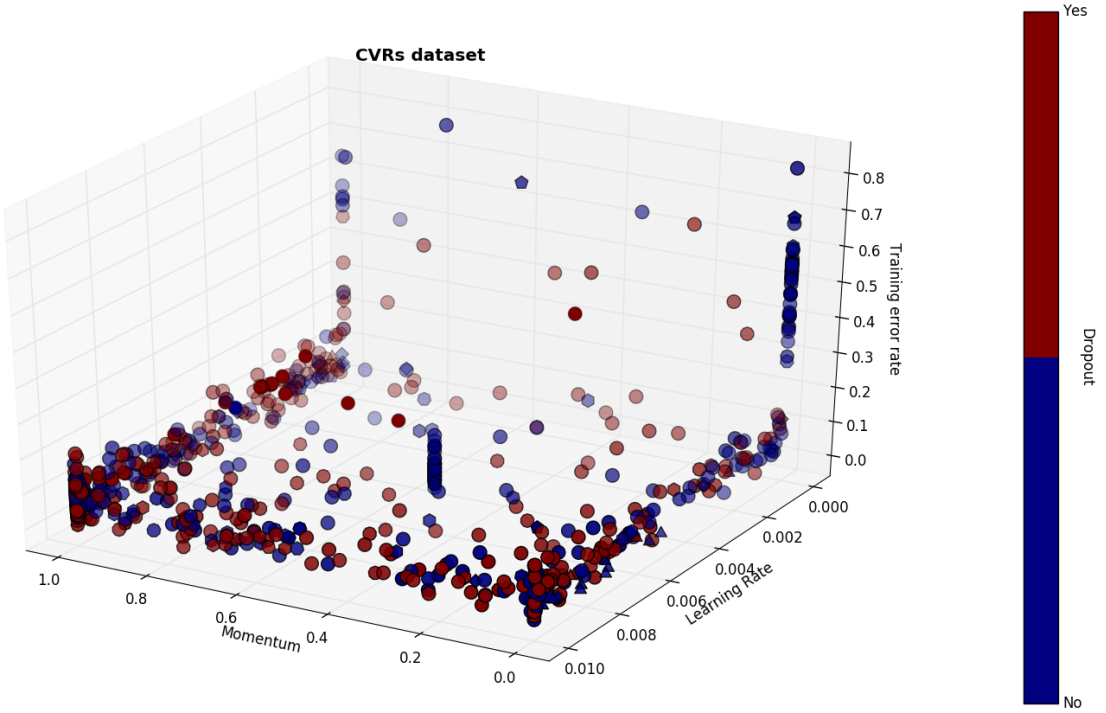


Figure 2: Histogram of proportion of missing values in each feature (Left) of CVRs training set and aggregation plot of all existing combinations of missing and non-missing values in the samples (Right).

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 3 shows the exploration of hyperparameter space during Bayesian optimization for both Adult and CVRs datasets. Each circle represents a candidate neural network 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).


Adult dataset
Momentum
Learning Rate
Training error rate
Dropout
Yes
No


CVRs dataset
Momentum
Learning Rate
Training error rate
Dropout
Yes
No

S3 Results with MCAR perturbation

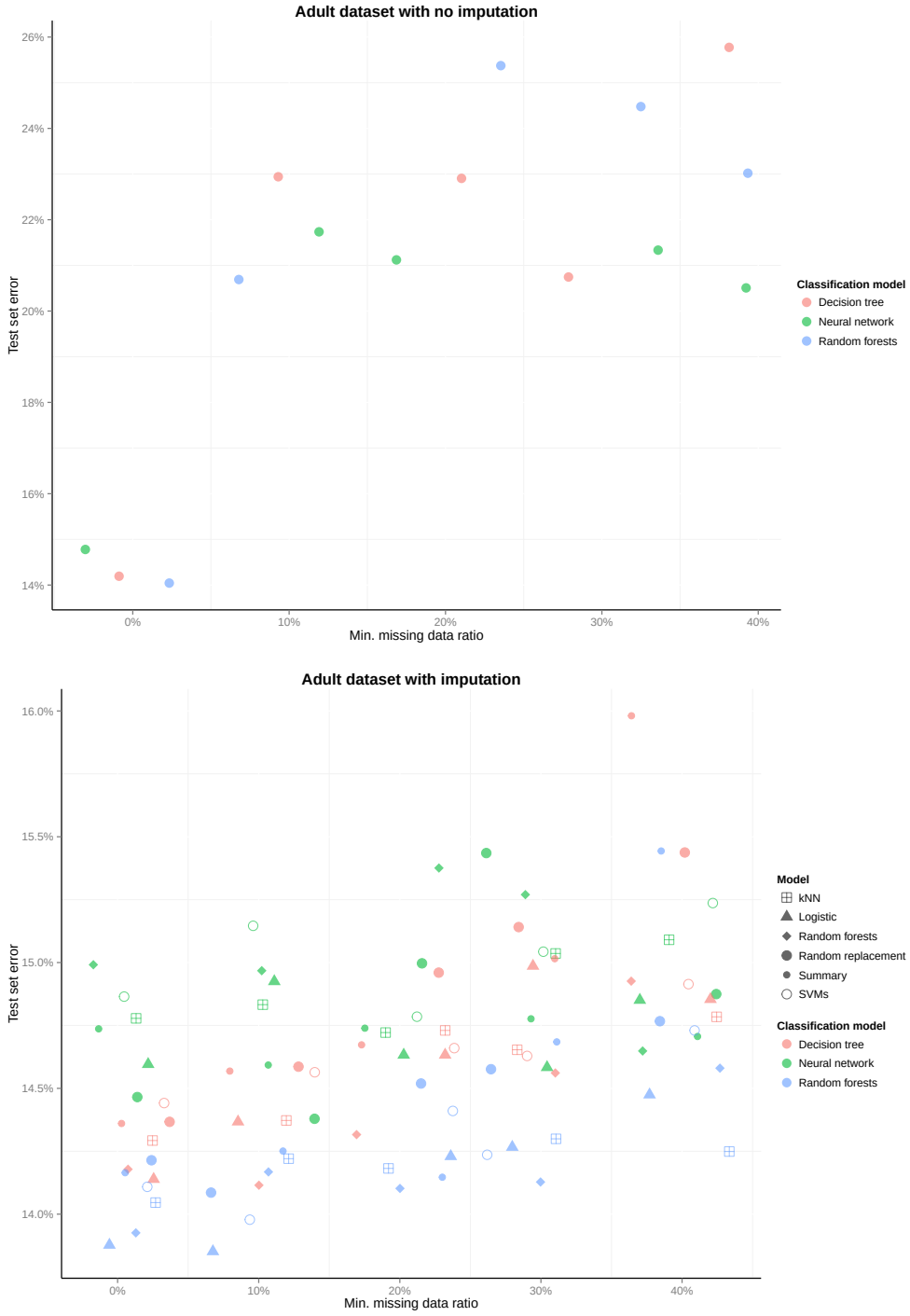


Figure 4: Error rates on the Adult test set with (bottom) and without (top) missing data imputation, for various levels of MCAR-perturbed categorical training features (x-axis). One-hot encoding is used to represent missing data in the absence of imputation. The decision tree and random forests classifiers are trained with maximum depths of 8 and 16, respectively.

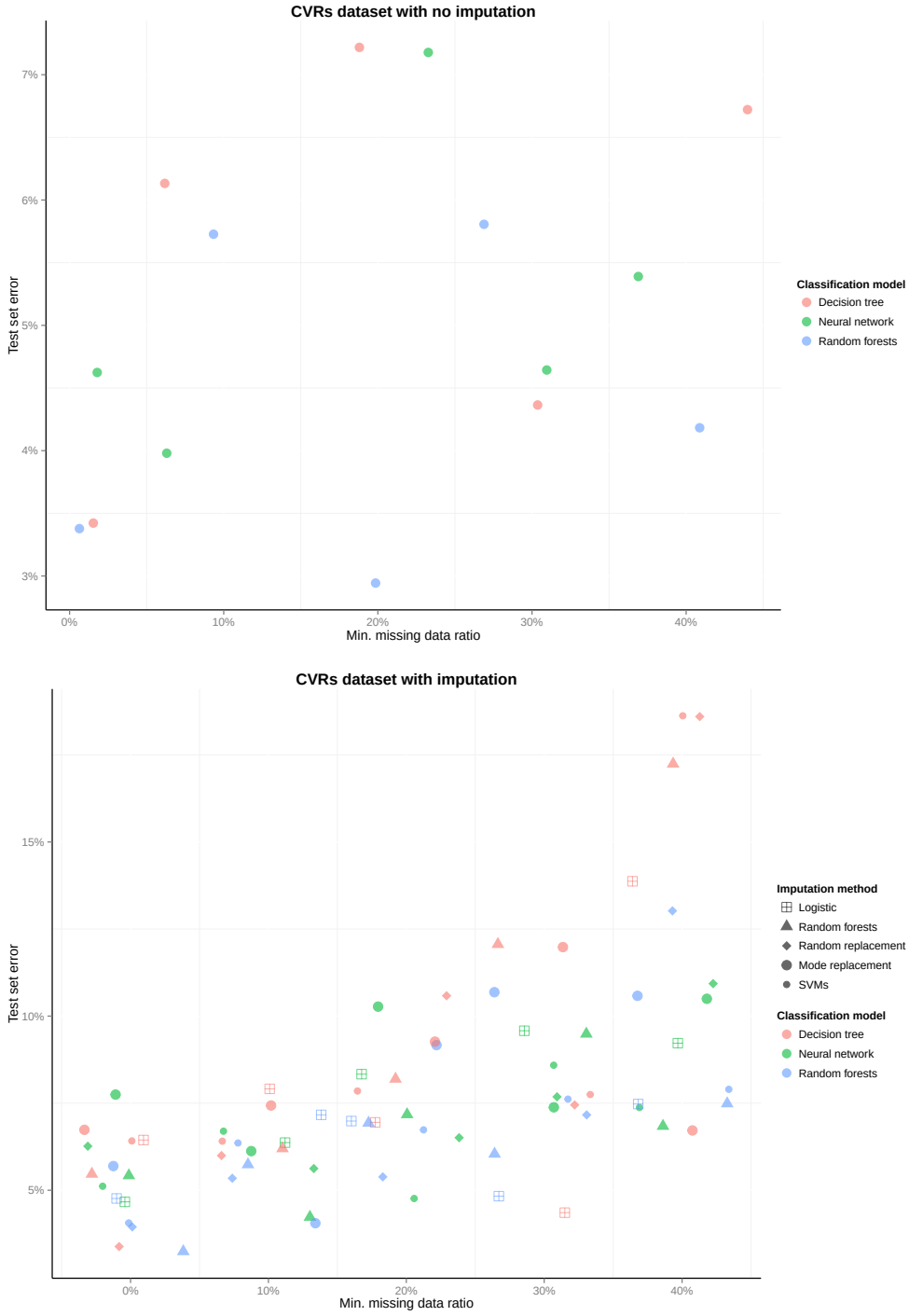


Figure 5: Error rates on the CVRs test set with (bottom) and without (top) missing data imputation.

See footnotes for Figure 4.