|  |  |  |
| --- | --- | --- |
| Model | Strategy for missing values | Best parameters tuned by 5-fold CV |
| XGBoost | To learn the splitting directions for samples with missing values based on the minimization of training loss [1] | max\_depth = 5  eta = 0.3  subsample = 1  colsample\_bytree = 0.8 |
| BART | To incorporate missingness by augmenting the nodes’ splitting rules to (a) use missingness itself as a variable to be considered in a splitting rule and (b) also handle sorting the missing data to the left or right [2] | k = 1  num\_trees = 150  prob\_rule\_class = 0.2 |
| Random Forest | To conduct split in a node, the algorithm imputes missing values by drawing randomly from the set of nonmissing in-bag data within the current node [3] | mtry = 2  nodesize = 5  splitrule = “gini”  sampprop = 0.8 |
| Conditional random forest | Splits are determined by only observed variables; then at one node, to pass down observations missing that variable, the algorithm considers surrogate splits. As an example, assume that the split (age < 40, age >= 40) has been chosen. The surrogate variables are found by re-applying the partitioning algorithm (without recursion) to predict the two categories “age <40” vs “age >=40” using the other independent variables. [4] | mtry = 5  maxsurrogate = 0 |

**1. Method 1 to deal with missing data:**

**1.1 More parameter tuning**

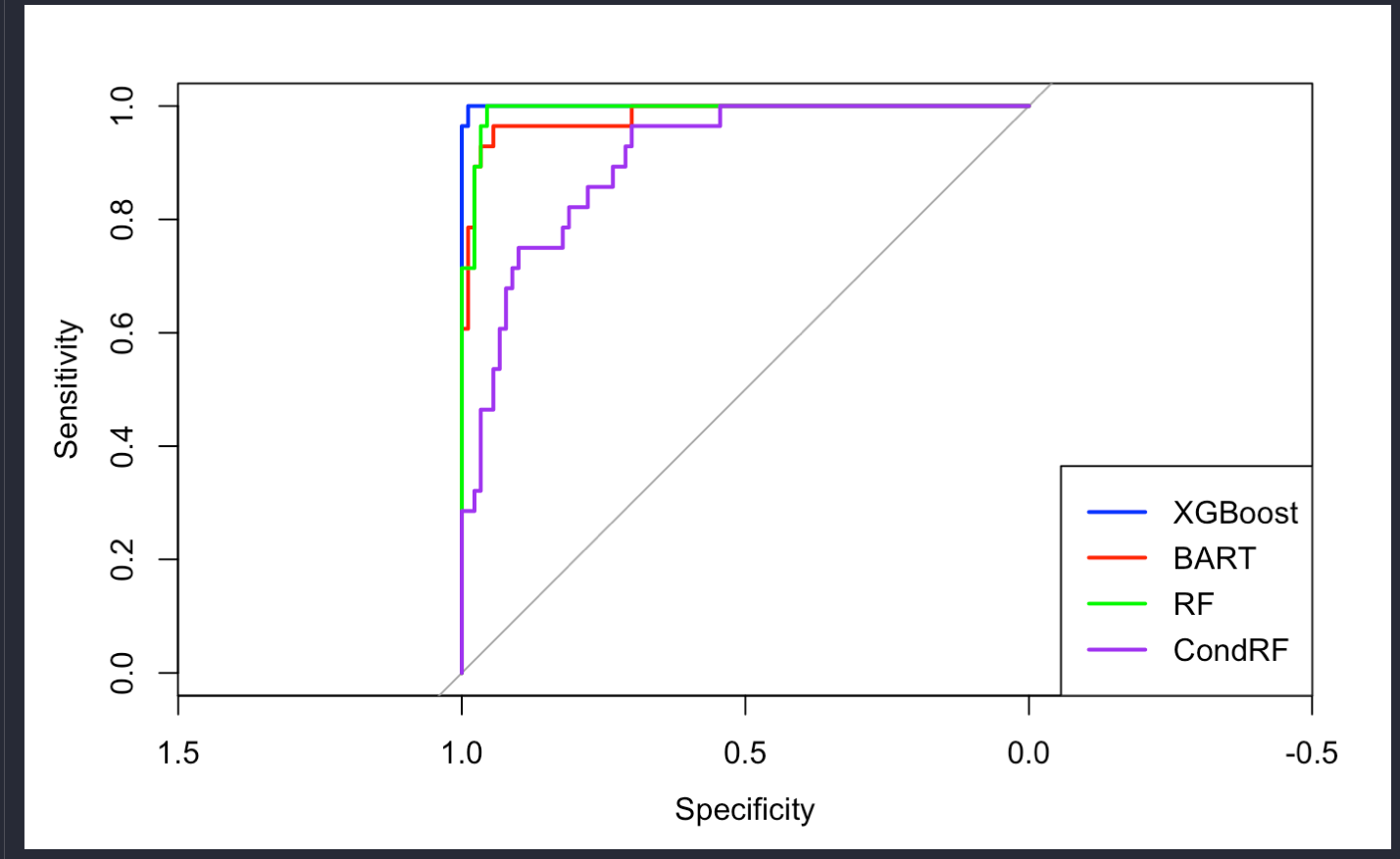
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | Accuracy | AUC | Sensitivity | Specificity | PPV | NPV |
| Training data (threshold) | | | | | | |
| XGBoost  (0.312) | 0.992 | 1.000  (0.998,1.000) | 1.000  (0.877,1.000) | 0.989  (0.940,1.000) | 0.966  (0.822,0.999) | 1.000  (0.959,1.000) |
| BART  (0.312) | 0.924 | 0.983  (0.955,0.999) | 0.964  (0.817,0.999) | 0.944  (0.875,0.982) | 0.844  (0.672,0.947) | 0.988  (0.937,1.000) |
| Random Forest  (0.303) | 0.966 | 0.992  (0.979,1.000) | 1.000  (0.877,1.000) | 0.956  (0.890,0.988) | 0.875  (0.710,0.965) | 1.000  (0.958,1.000) |
| Conditional Random Forest  (0.230) | 0.780 | 0.914  (0.856,0.957) | 0.964  (0.817,0.999) | 0.722  (0.618,0.811) | 0.519  (0.376,0.660) | 0.985  (0.918,1.000) |
| Test data | | | | | | |
| XGBoost | 0.714 | 0.767  (0.603,0.895) | 0.455  (0.167,0.766) | 0.789  (0.627,0.904) | 0.385  (0.139,0.684) | 0.833  (0.672,0.936) |
| BART | 0.776 | 0.746  (0.567,0.900) | 0.545  (0.234,0.833) | 0.842  (0.687,0.940) | 0.500  (0.211,0.789) | 0.865  (0.712, 0.955) |
| Random Forest | 0.755 | 0.847  (0.711,0.950) | 0.727  (0.390,0.940) | 0.763  (0.598,0.886) | 0.471  (0.230,0.722) | 0.906  (0.750,0.980) |
| Conditional Random Forest | 0.714 | 0.806  (0.639,0.947) | 0.818  (0.482,0.977) | 0.684  (0.513,0.825) | 0.429  (0.218,0.660) | 0.929  (0.765,0.991) |

Model performance:

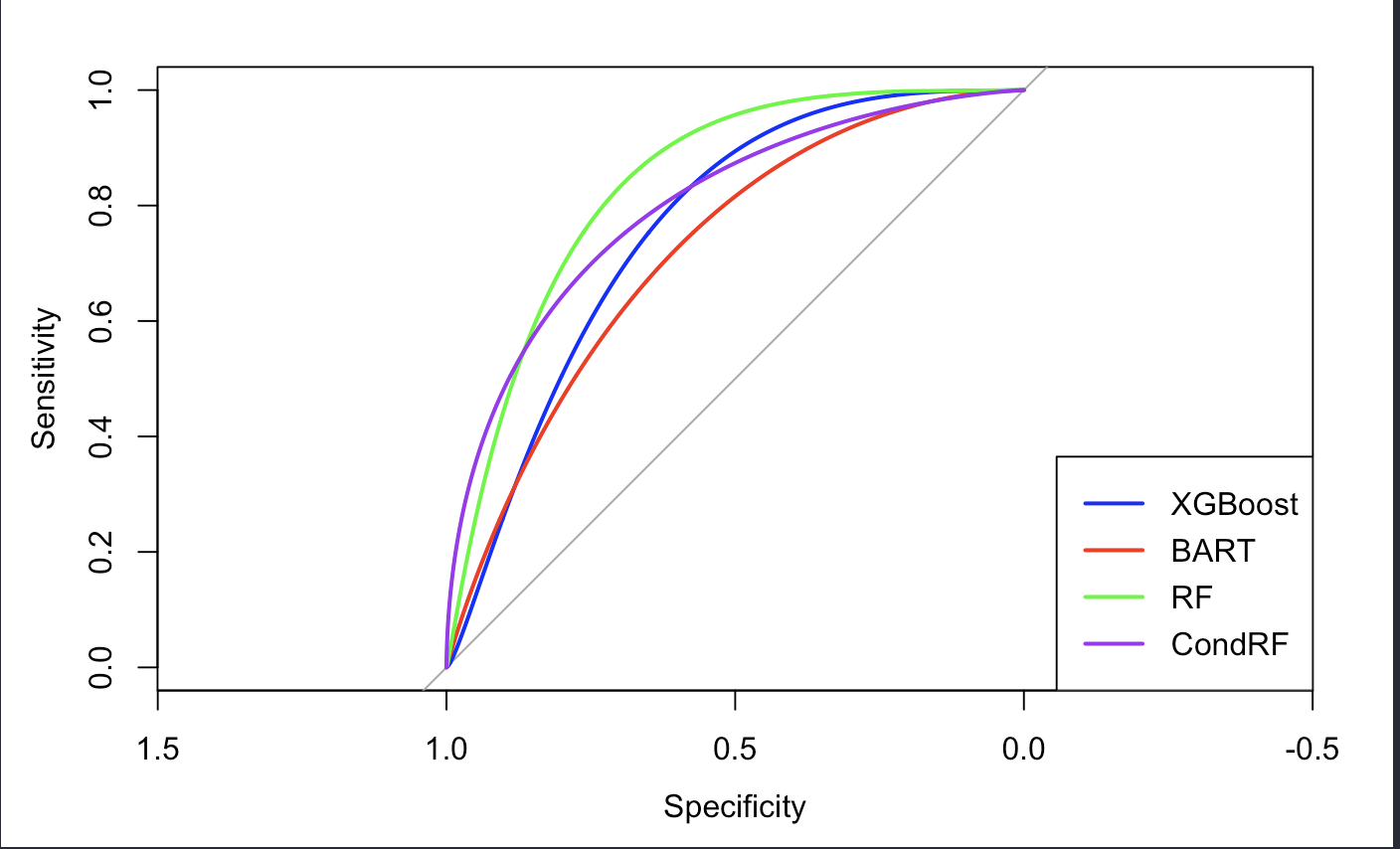
To calculate accuracy, sensitivity, specificity, PPV, and NPV, we need to determine a threshold for classifying a subject as a case. Specifically, we need to find a threshold such that a subject is classified as a case if  .

Here, a reasonable threshold is determined based on the ROC curve estimated with the training data, such that the summation of sensitivity and specificity is the largest at the threshold [5]. This is because Youden Index = sensitivity + specificity – 1.

ROC curve based on training data:



Smoothed ROC curve based on test data:



The boxplot for predicted probability of Y=1:

On training data:

A diagram of a graph

Description automatically generated with medium confidence

On testing data:

A diagram of a graph

Description automatically generated

**1.2 Less parameter tuning:**

|  |  |
| --- | --- |
| Model | Best parameters tuned by 5-fold CV |
| XGBoost | max\_depth = 3 |
| BART | num\_trees = 20 |
| Random Forest | mtry = 2  nodesize = 5 |
| Conditional random forest | mtry = 5  maxsurrogate = 0 |

Model performace:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | Accuracy | AUC | Sensitivity | Specificity | PPV | NPV |
| Training data (threshold) | | | | | | |
| XGBoost  (0.333) | 0.983 | 0.999  (0.996,1.000) | 1.000  (0.877,1.000) | 0.978  (0.922,0.997) | 0.933  (0.779,0.992) | 1.000  (0.959,1.000) |
| BART  (0.281) | 0.898 | 0.917  (0.846,0.972) | 0.750  (0.551,0.893) | 0.944  (0.875,0.982) | 0.808  (0.606,0.934) | 0.924  (0.849,0.969) |
| Random Forest  (0.309) | 0.932 | 0.980  (0.956,0.995) | 0.964 (0.817,0.999) | 0.922 (0.846,0.968) | 0.794  (0.621,0.913) | 0.988 (0.935,1.000) |
| Conditional Random Forest  (0.230) | 0.780 | 0.914 (0.857,0.958) | 0.964  (0.817,0.999) | 0.722  (0.618,0.811) | 0.519  (0.376,0.660) | 0.985  (0.918,1.000) |
| Test data | | | | | | |
| XGBoost | 0.776 | 0.788  (0.644,0.907) | 0.455  (0.167,0.766) | 0.868  (0.719,0.956) | 0.500  (0.187,0.813) | 0.846  (0.695,0.941) |
| BART | 0.816 | 0.766  (0.601,0.921) | 0.455  (0.167,0.766) | 0.921  (0.786,0.983) | 0.625  (0.245,0.915) | 0.854  (0.708,0.944) |
| Random Forest | 0.776 | 0.833  (0.687,0.952) | 0.727  (0.390,0.940) | 0.789 (0.627,0.904) | 0.500  (0.247,0.753) | 0.909 (0.757,0.981) |
| Conditional Random Forest | 0.714 | 0.806  (0.629,0.945) | 0.818  (0.482,0.977) | 0.684  (0.513,0.825) | 0.429  (0.218,0.660) | 0.929  (0.765,0.991) |

ROC curve based on training data:

**A graph with different colored lines

Description automatically generated**

Smoothed ROC curve based on testing data:

**A diagram of a function

Description automatically generated with medium confidence**

The boxplot for predicted probability of Y=1:

On training data:

A diagram of a graph

Description automatically generated

On testing data:

**A diagram of a graph

Description automatically generated with medium confidence**

**2. Method 2 to deal with missing data: MissForest**

|  |  |
| --- | --- |
| Model | Best parameters tuned by 5-fold CV |
| XGBoost | max\_depth = 7 |
| BART | num\_trees = 20 |
| Random Forest | mtry = 5  nodesize = 3 |
| Conditional random forest | mtry = 4 |

We use MissForest [6] to impute missing data in training and testing data.

Model performance:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Model | Accuracy | AUC | Sensitivity | Specificity | PPV | NPV |
| Training data (threshold) | | | | | | |
| XGBoost  (0.429) | 1.000 | 1.000  (1.000,1.000) | 1.000  (0.877,1.000) | 1.000  (0.960,1.000) | 1.000  (0.877,0.999) | 1.000  (0.960,1.000) |
| BART  (0.267) | 0.831 | 0.893  (0.822,0.954) | 0.857  (0.673,0.960) | 0.822  (0.727,0.895) | 0.600  (0.433,0.751) | 0.949  (0.874,0.986) |
| Random Forest  (0.416) | 1.000 | 1.000  (1.000,1.000) | 1.000  (0.877,1.000) | 1.000  (0.960,1.000) | 1.000  (0.877,1.000) | 1.000  (0.960,1.000) |
| Conditional Random Forest  (0.252) | 0.822 | 0.914 (0.859,0.958) | 0.929  (0.765,0.991) | 0.789  (0.690,0.868) | 0.578  (0.422,0.723) | 0.973  (0.905,0.997) |
| Test data | | | | | | |
| XGBoost | 0.837 | 0.798  (0.646,0.919) | 0.455  (0.167,0.766) | 0.947  (0.823,0.994) | 0.714  (0.290,0.963) | 0.857  (0.715,0.946) |
| BART | 0.796 | 0.835  (0.682,0.945) | 0.727  (0.390,0.940) | 0.816  (0.657,0.923) | 0.533  (0.266, 0.787) | 0.912  (0.763,0.981) |
| Random Forest | 0.796 | 0.792  (0.629,0.912) | 0.455  (0.167,0.766) | 0.895  (0.752,0.971) | 0.556  (0.212,0.863) | 0.850  (0.702,0.943) |
| Conditional Random Forest | 0.755 | 0.811  (0.663,0.940) | 0.636  (0.308,0.891) | 0.789  (0.627,0.904) | 0.467  (0.213,0.734) | 0.882  (0.725,0.967) |

ROC curve based on training data:

A graph with different colored lines

Description automatically generated

Smoothed ROC curve based on testing data:

A diagram with colored lines

Description automatically generated

The boxplot for predicted probability of Y=1:

On training data:

A diagram of a graph

Description automatically generated

On testing data:

A diagram of a graph

Description automatically generated with medium confidence

**3. Summary**

|  |  |  |  |
| --- | --- | --- | --- |
| Model | AUC  No imputation + more tuning | AUC  No imputation + less tuning | AUC  Imputation + less tuning |
| Training | | | |
| XGBoost | 1.000  (0.998,1.000) | 0.999  (0.996,1.000) | 1.000  (1.000,1.000) |
| BART | 0.983  (0.955,0.999) | 0.917  (0.846,0.972) | 0.893  (0.822,0.954) |
| Random Forest | 0.992  (0.979,1.000) | 0.980  (0.956,0.995) | 1.000  (1.000,1.000) |
| Conditional Random Forest | 0.914  (0.856,0.957) | 0.914  (0.857,0.958) | 0.914  (0.859,0.958) |
| Testing | | | |
| XGBoost | 0.767  (0.603,0.895) | 0.788  (0.644,0.907) | 0.798  (0.646,0.919) |
| BART | 0.746  (0.567,0.900) | 0.766  (0.601,0.921) | 0.835  (0.682,0.945) |
| Random Forest | 0.847  (0.711,0.950) | 0.833  (0.687,0.952) | 0.792  (0.629,0.912) |
| Conditional Random Forest | 0.806  (0.639,0.947) | 0.806  (0.629,0.945) | 0.811  (0.663,0.940) |

Feature importance:

A graph of a bar chart

Description automatically generated with medium confidence

Reference

[1] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

[2] Kapelner, A., & Bleich, J. (2015). Prediction with missing data via Bayesian additive regression trees. *Canadian Journal of Statistics*, *43*(2), 224-239.

[3] Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests.

[4] Breiman, L. (2017). *Classification and regression trees*. Routledge.

[5] Schisterman, E. F., Perkins, N. J., Liu, A., & Bondell, H. (2005). Optimal cut-point and its corresponding Youden Index to discriminate individuals using pooled blood samples. *Epidemiology*, *16*(1), 73-81.

[6] Stekhoven, D. J., & Bühlmann, P. (2012). MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics, 28(1), 112-118.

Add DART

Add importance p-value

Analyze CRS

Multiple imputation – need patient id.