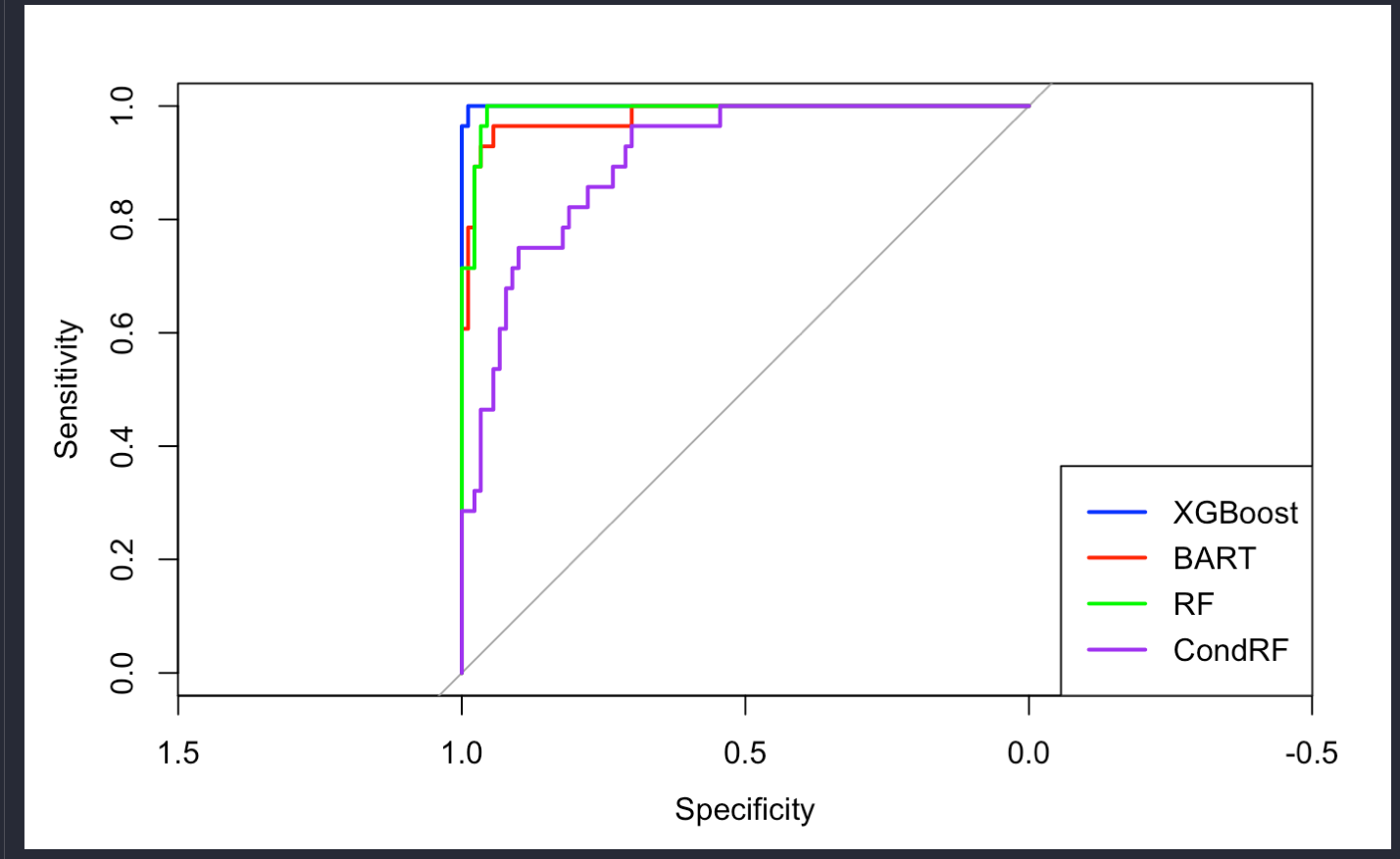
|  |  |  |
| --- | --- | --- |
| Model | Strategy for missing values | Best parameters tuned by 5-fold CV |
| XGBoost | To learn the splitting directions for samples with missing values based on the minimization of training loss [1] | max\_depth = 5  eta = 0.3  subsample = 1  colsample\_bytree = 0.8 |
| 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 |

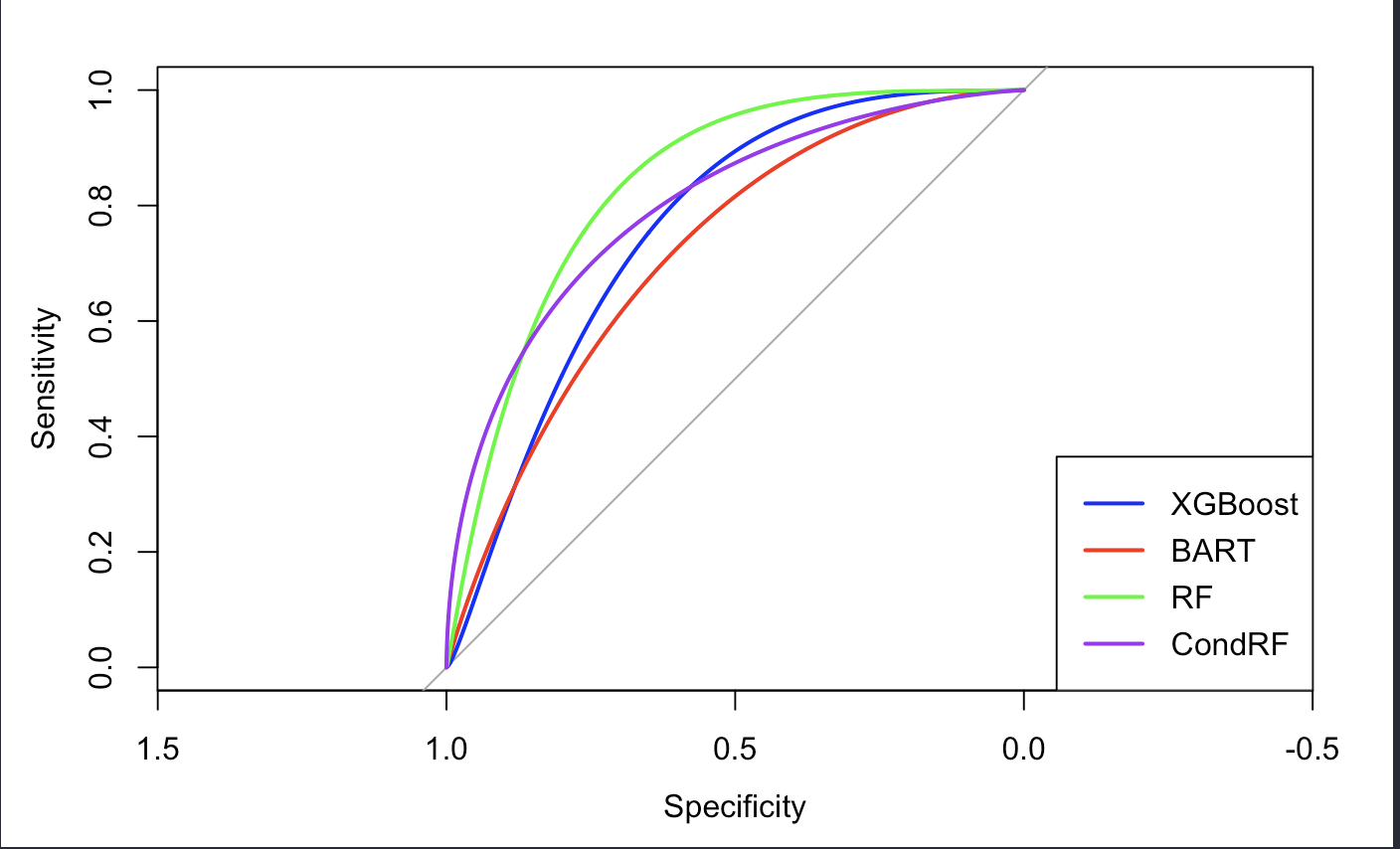
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | AUC | Sensitivity | Specificity | PPV | NPV |
| Training data (threshold) | | | | | |
| XGBoost  (0.312) | 1.000  (0.998,1.000) | 1.000  (0.877,1.000) | 0.989  (0.940,1.000) | 0.966  (0.822,0.999) | 1.000  (0.959,1.000) |
| BART  (0.312) | 0.983  (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) | 0.992  (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) | 0.914  (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 | 0.767  (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 | 0.746  (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 | 0.847  (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 | 0.806  (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) |

An optimal 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:



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.

To-do:

1. Not to tune too many variables

2. Add accuracy

3. Variable importance

4. Discard the variable with much missingness and try again

5. Discard the missing values and run the models

6. Multiple imputation: only report the model fitting based on one data but compare the results based on all imputed data; ask for the patient id from Jennifer

Monday 1-3pm

Wednesday 9:30 – 3