Predicting Intubation Need for COVID Postive Inpatients

Kai Zhang

2020/7/19

Introduction

COVID+ patients suffer from acute respiratory distress and have acute difficulties in breathing. Such critically ill patients then require artificial respiratory support through an invasive mechanical ventilator, a process known as intubation. Since there was a surge in COVID+ cases in New York City during the period of March, the hospital authorities were concerned that there might be a shortage in availability of mechanical ventilators. The authorities wanted to predict the need for intubation for each patient in order to better manage their resources.

The simulated data on COVID+ patients who were hospitalized in the New York Presbyterian hopital (NYP), contains patient level information on baseline clinical measures that were determined at the time of admission to the hospital, including demographic variables, relevant clinical history and certain diagnosis tests. Besides, the vital signs of each patients (e.g. blood pressure, heart rate, SpO2 or blood oxygen levels, etc) were measured on regular basis during the course of hospitalization.

The goal is to build a predictive model predicting the binary event of intubation, based on baseline clinical measures and the longitudinal vital signs measures.

Feature Engineering

Vital signs are measurements of the body's most basic function. The five main vital signs routinely monitored by NYP include diastolic blood pressure, systolic blood pressure, heart rate, respiratory rate, and SpO2 which is a measure of pulse oxygen saturation.

New features with clinical significance were extracted from the longitudinal vital signs measures. The mean, standard deviation, maximum and minimum are computed for each inpatient and each vital signs, creating 20 features describing the average level, unstability and extremes of the vitals signs. In addition, the length of hospital stay (in days) was captured by time difference between the first and last timestamp recorded for each patients.

The final data contains 1345 COVID+ inpatients with 25 baseline variables, 21 vital signs variable, and binary outcome variable.

Exploratory Data Analysis

The distribution of continuous data grouped by intubation need is visualized in *Figure 1.* It is observed that age, bmi, diastolic blood pressure, sytolic blood pressue and heart rate are more preditive variables for intubation need. And also, all these variables follows normal distribution approximately, therefore no extra transformation is needed.

Figure 1 − Distribution of continuous variables by intubation

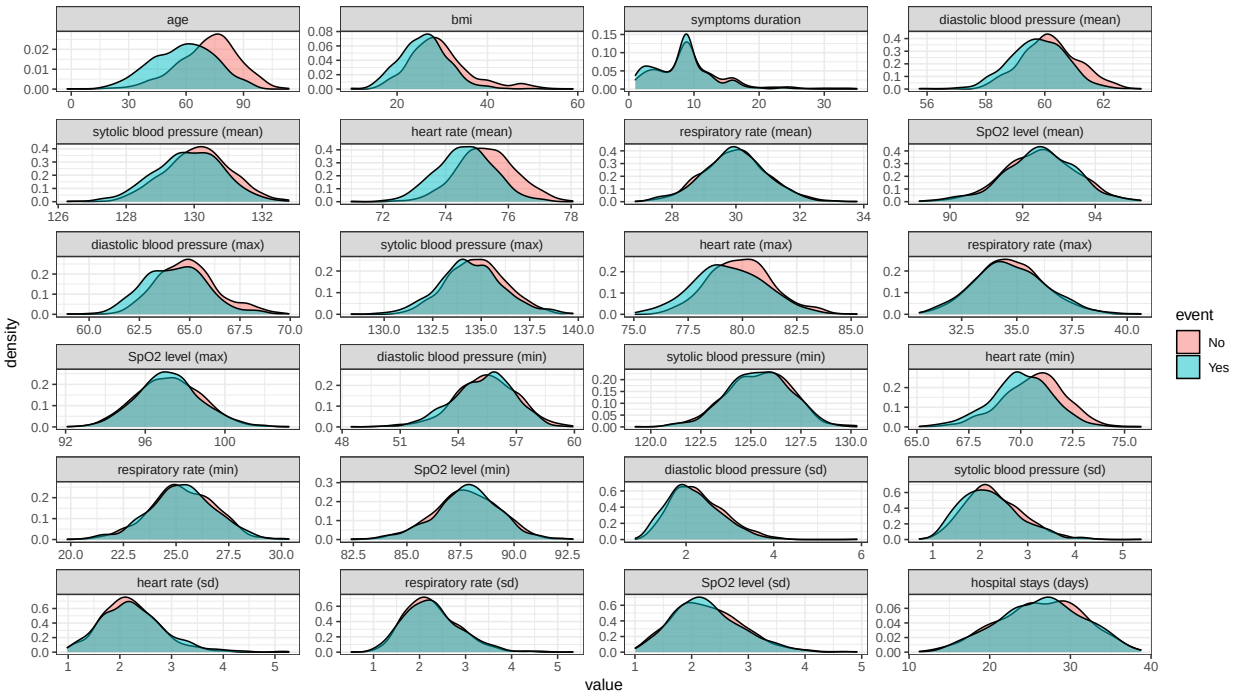


Figure 2 − Distribution of categorical variables by intubation needs

The distribution of categorical data is visualized in *Figure 2*.It is observed that diabetes mellitus, pulmonary disease, cancer, diarrhea symptoms and chest X-ray finding are probability more predictive variables in predicting intubation needs.

Data Preparation

Partitioning Train and Test Data

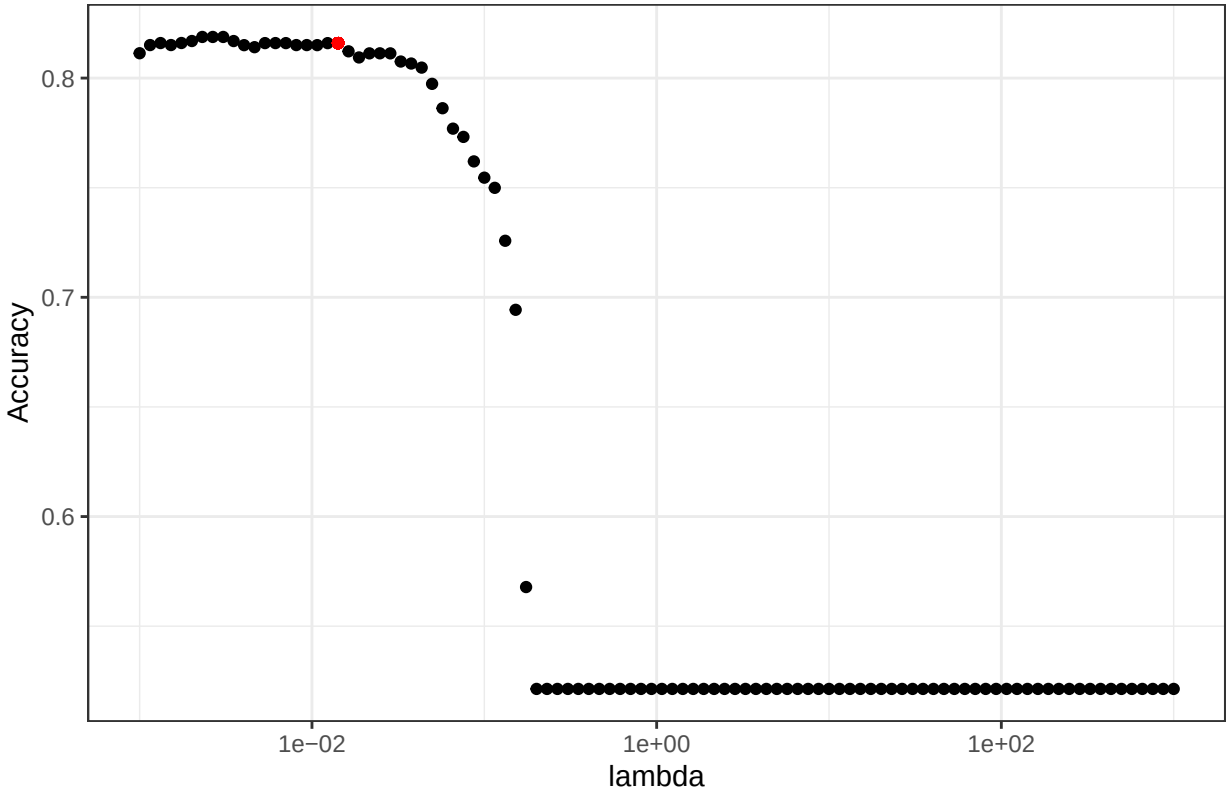
Stratified sampling is used to create train and test data. The resampling occurs within each outcome class (Yes/No), so 80% of sample within each class is split into train data.

Feature Selection

Lasso Regression is used to select predictive features. To tune the parameter λ, 10-fold cross validation is used to estimate the generalization errors of lasso regression with λ over 10^{-3} to 10^{3}.

Using one-standard-error rule, the optimal λ is 0.014175, and its cross-validation accuracy is 81.59%.

Figure 3 − Plot of lasso regression's estimated test accuracy versus lambda



The lasso regression selects 14 out of 46 features.

```
##  [1] "age"                                "bmi"
##  [3] "smoke"                              "diabetes.mellitus"
##  [5] "cancer"                             "fever.symptoms"
##  [7] "symptoms.duration"                  "Admission.to.ED"
##  [9] "diastolic.blood.pressure..mean." "sytolic.blood.pressure..mean."
## [11] "heart.rate..mean."                  "diastolic.blood.pressure..max."
## [13] "sytolic.blood.pressure..min."       "sytolic.blood.pressure..sd."
```

Method

Multiple machine learning techniques are used to predict the need for intubation for each patients. R *caret* package is used to perform a grid search to build a model for every combination of hyperparameters specified and 10-fold cross validation is used to evaluate each model.R *xgboost* package is used to build XGBoost model.

Elastic Net Regression

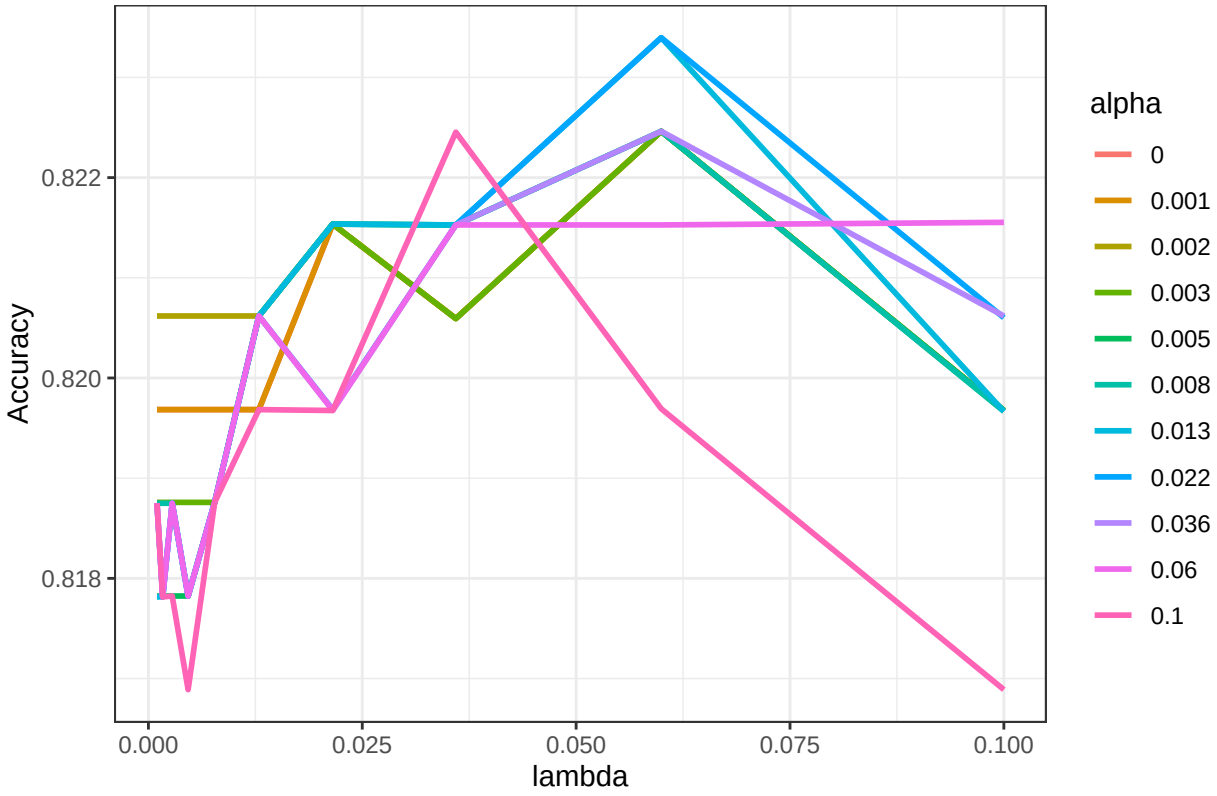
Elastic net is a regularized regression method that linearly combines L_1 and L_2 penalties of lasso and ridge regression. There are two hyperparameters need tuning:

- α: penalty factor determining relative importance between L_1 and L_2.
- λ: shrinkage coefficient.

A grid search is performed across the combinations of λ and α over 10^{-3} to 10^{-1}. The combination of λ and α with highest 10-fold cross validation accuracy is selected to be the optimal (See *Figure 4*).

As a result, the optimal elastic net regression is with alpha = 0.1292 and lambda = 0.05995, which has cross-validation accuracy = 82.23%.

Figure 4 − Grid search results for elastic net regression



Support Vector Machine

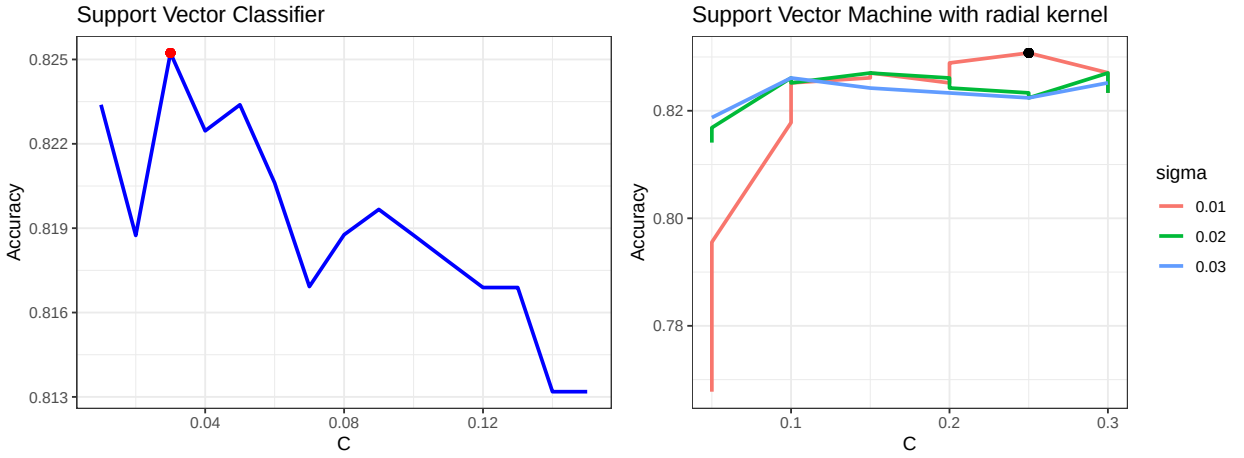
Super Vector Machine (SVM) finds a optimal hyperplane in a feature space enlarged by kernels, given one buget that allows some misclassified points.

Super Vector Machine with linear kernel is equivalent to Super Vecetor Classifier (SVC), which find a optimal hyperlanee in the original feature space. Its C parameter tells the algorithm how much the user care about misclassified points. In comparison, SVM with nonlinear kernel is equivalent to performing SVC in an enlarged feature space. It requires one additionl parameter σ, which determine the spread of the kernel as well as the decision boundary.

By a grid search on C parameter over 0.01 to 0.15, the optimal SVC is with $C = 0.03$ and has highest validation accuracy = 82.52% (see left of *Figure 5*).

Similarly, SVM with Radial kernel (C = 0.25, sigma = 0.010) have optimal cross-validated accuracy = 83.08% (see right of *Figure 5*).

Figure 5 − Grid Search Results of Support Vector Machine



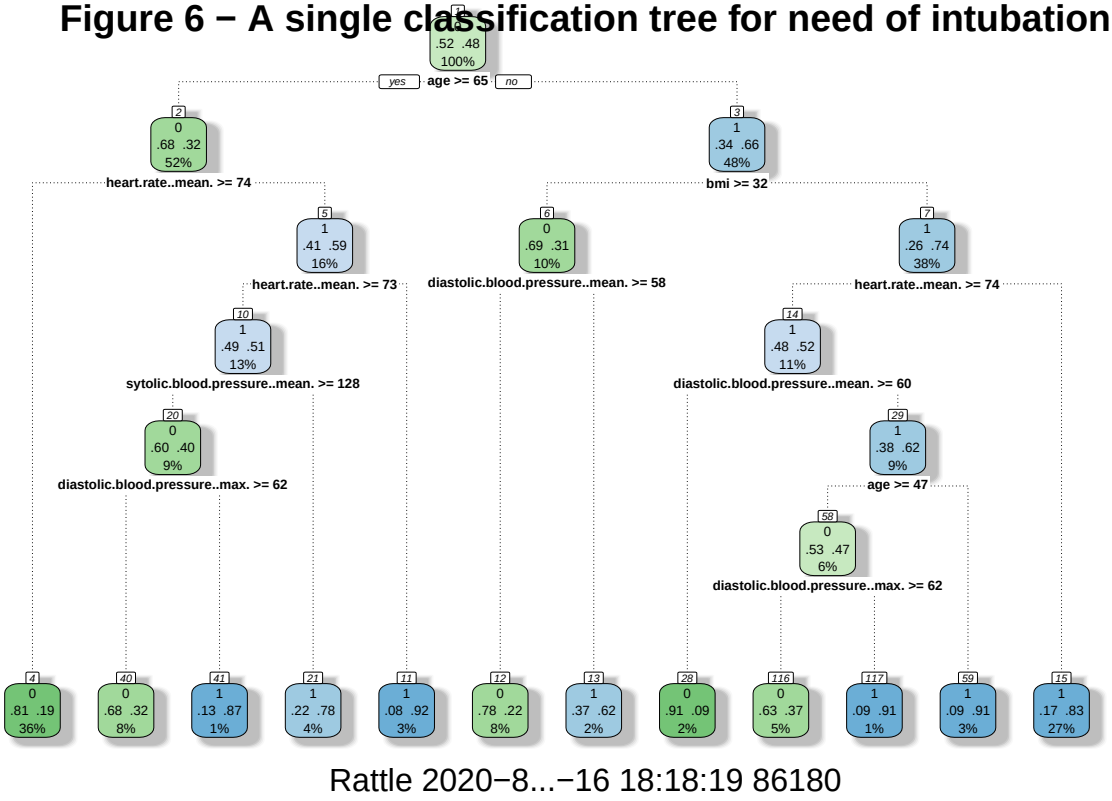
Decision Trees (CART)

A classification tree is an iterative process of splitting the data into partitions, and then splitting it up further on each of the braches. The strategy of finding optimal split is binary recursive partitioning. And also, pruning can prevent CART from overfitting.

CART have one hyperparameter complexity parameter, which requires a minimum improvement in the model needed at each node. This stops the tree continuing growing before overfitting. By 10-fold cross validation, a optimal CART is with complexity parameter = 0.01, which have validation accuracy 73.97%.

The final CART model is visualized in *Figure 6*.

Figure 6 − A single classification tree for need of intubation



Bagging and Random Forest

Bagging trees aggregates multiple CARTs so as to reduce the variance of the predictive model. In addition, Random Forest (RF) is one type of Bagging trees. At each split, RF randomly considers a subset of features, which decorrelate different CARTs and makes the model even more resistant.

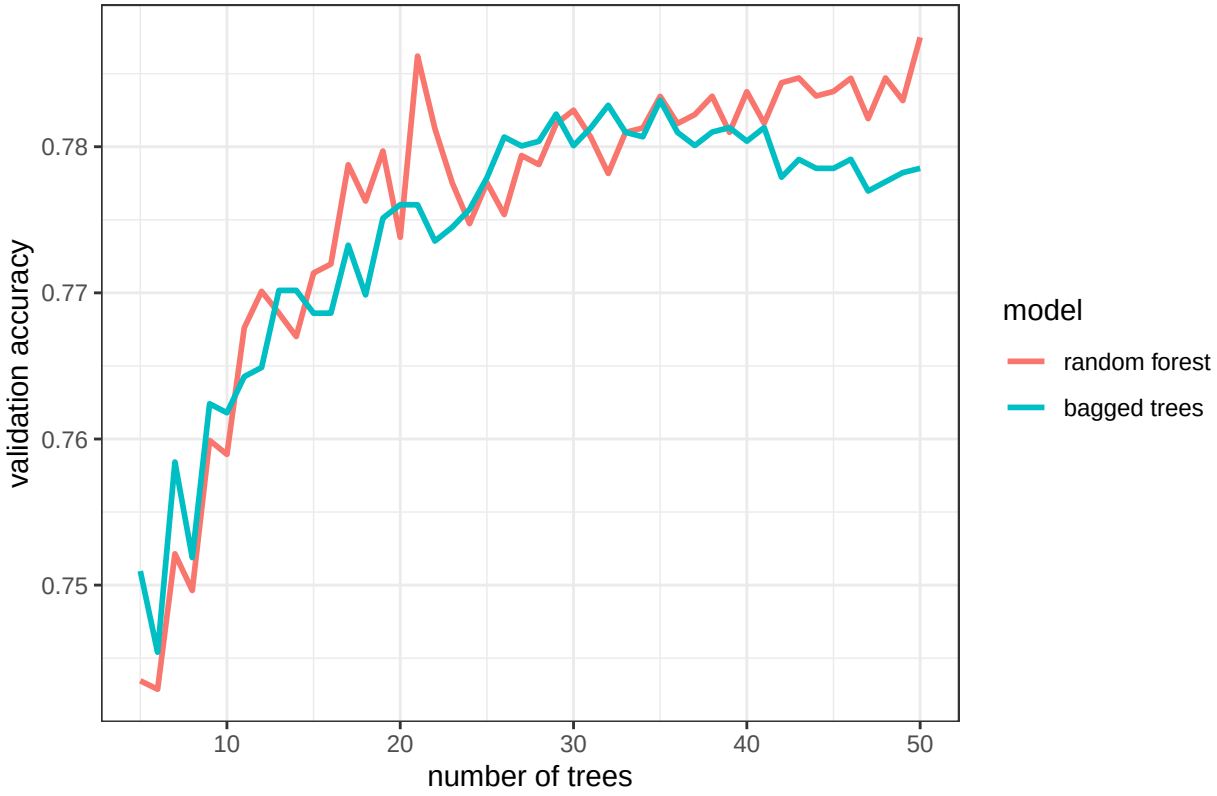
There are two main hyperparameters for bagging trees, which include:

- *mtry*: number of features used per split.
- *n.tree*: number of trees.

In general, Bagging with a large number of trees doesn't harm prediction accuracy, but it requires long computation time. In *Figure 6*, a optimal *mtry* should be a value after which the accuracy doesn't improve much any more.

By 10-fold optimal validation, the optimal random forest model is with $mtry = sqrt(p)$ and $n.tree = 21$, which is of accuracy 78.62%.

Figure 7 − Grid search results of bagging trees and random forest



XGBoost

XGBoost is another ensemble learning method - Boosting. Gradient boosting update the model by training on the residuals of the current model. Besides, XGBoost adds regularization term in its loss function, avoiding overfitting.

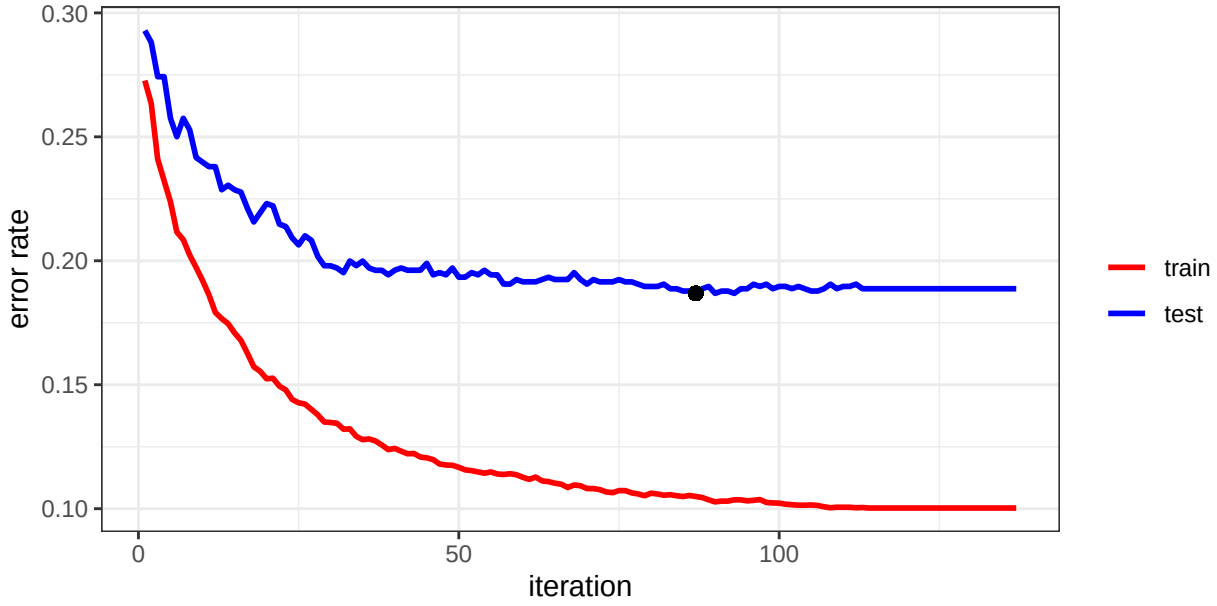
The main hyperparameters of XGBoost includes:

- Learning rate (η)
- max_depth: maximal depth of each tree.
- λ: L_2 regularization on leaf weights.
- α: L_1 regularization on leaf weight.
- colsample_bytree: percentage of features used per tree.

Grid search is performed over the combination of these hyperparameters. Moreover, early stopping strategy is used so that the model can stop learning if its validation accuracy is not improved in last 25 rounds.

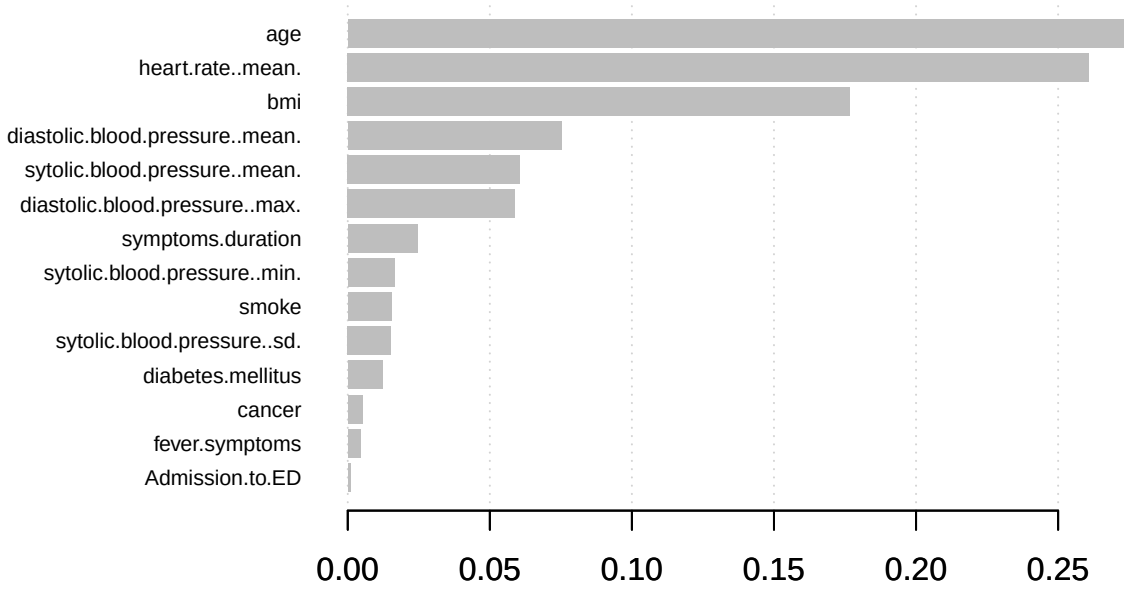
The grid search found XGBoost with learing rate (η) = 0.3, max_depth = 2, $\lambda = 1$, $\alpha = 5$, colsample_bytree = 1 has highest validation accuracy 81.31%. In addition, it uses 87 trees in total (see *Figure 8*).

Figure 8 − Learning Curve of XGBoost with hyperparameters tuned



Further, the *Figure 9* shows the feature importance. It measures the total information gain based on the gini index of the splits by each features.

Figure 9 − Feature importance in XGBoost



Neural Network

A Neural Network with two inner layers is trained using R *keras* package. Because the outcome variable is binary, the activation function of output layer is sigmoid function and the loss function is summation of binary cross entropy. In addition, the number of input layer and output layer should be 14 and 1.

There other main hyperparameters includes:

- units of two input layer
- activation function of innner layers
- optimizer: Adam, SGD and RMSprop, for exmaple.
- batch_size: the number of sample used at each updates.
- epoch number: number of complete passes through the training data.

In this model, *leaky_relu* function is used to be the activation function of two inner layers. Additionally, epoch size is set to be 10000 with early stoping monitor stopping the training process once the validation loss doesn't decrease within past 25 updates.

Then, a random search is used to select a optimal combination of the remaining hyperparameters on a normalized data. As a result, the optimal model have validation accuracy 77.78%, and its hyperparameters settings is the following:

- units number of the 1st layer: 50.
- units number of the 2nd layer: 25.
- optimizer: Adam.
- epoch size = 1371.

The learning curve of the neural network is shown in *Figure 10*.

Figure 10 − Learning Curve of Neural Network

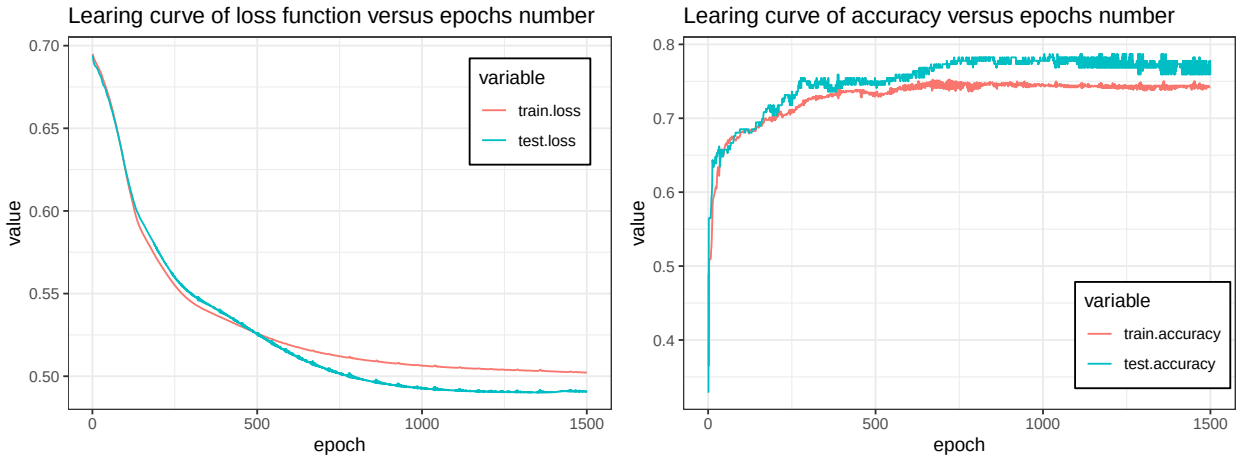


Table 1: Performance evaluation of the models

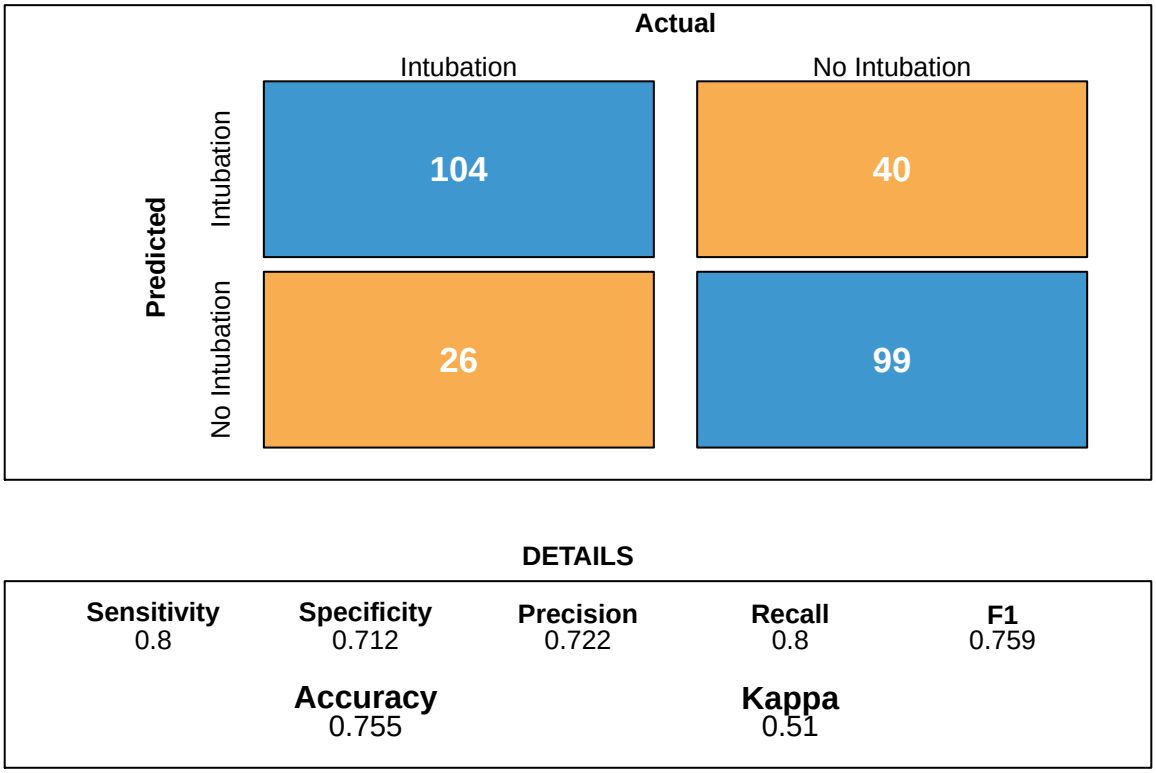
Model	Accuracy	Sensitivity	Specificity	Precision
Elastic net regression	0.792	0.803	0.780	0.791
SVM-linear	0.788	0.801	0.774	0.784
SVM-RBF	0.773	0.783	0.763	0.777
Classification tree	0.717	0.714	0.721	0.755
Random forest	0.762	0.786	0.739	0.741
XGBoost	0.755	0.792	0.722	0.712
Neural network	0.714	0.698	0.731	0.735

Final Model and Performance Evaluation

The result of training models is shown in *Table 1.* Accordingly, the Support Vector Machine with radial kernel have highest validation accuracy 83.08%. However, it is not easy to extract the feature importance from nonlinear SVM. By contrast, XGBoost have comparable validation accuracy 81.31% and provides estimates of feature importance (see *Figure 9*)

Using test data to evaluate the final XGBoost, the overall accuracy is 75.5% with sensitivity 80.00% and specificity 71.2% (seel *Figure 11*). The six most important factors includes age, heart rate average level, bmi, diastolic blood pressure average level, sytolic blood pressure average level, and daistolic blood pressure maximal value.

Figure 11 − CONFUSION MATRIX



Conclusion

In conclusion, XGBoost is selected as the final model predicting the intubation need of COVID+ inpatients. The XGBoost use an ensemble of decision trees where the new tree is added to the output of the existing sequence of trees in an efforts to improve the final output. Like other tree model, XGBoost have multiple ways to measure the contribution of each feature in predicting the label. The most commonly used one is to rank the features by the information gain they earn at each split.

The final model have estimated accuracy 75.5%. And also, it advices doctors to focus more on paitients' age, heart rate average level, bmi, diastolic and sytolic blood pressure average and extreme level.