Heart Disease

Random Forest Classification model

By Ahmad Qadri (arqchicago@gmail.com)

Heart Disease dataset

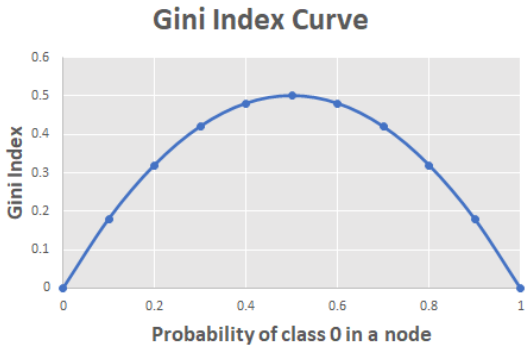
This dataset is collected by Hungarian Institute of Cardiology, University Hospital (Zurich, Switzerland), University Hospital (Zurich, Switzerland) and V.A. Medical Center (Long Beach and Cleveland Clinic Foundation). The dataset includes the target variable representing presence of some cardiovascular disease (0=no heart disease, 1=heart disease). The features measured in the dataset are Age, Sex, Type of Chest Pain, Resting Blood Pressure, Serum Cholesterol, Fasting Blood Sugar, Resting ECG, Max Heart Rate, Exercise Induced Angina, ST Depression induced by Exercise relative to rest, the Slope of peak exercise ST segment, Number of Vessels colored by Fluoroscopy and presence of Fixed or Reversible Defect in Stress Echocardiography.

Random Forest Classification

Random Forest is an ensemble method in which many decision trees are trained by bootstrapping the data and aggregating the results at the end. For classification problems, majority decision of each tree is taken to be the overall classification prediction. In each decision tree, data is continuously divided based on values of randomly selected features and purity of resulting nodes is computed to evaluate the effectiveness of the split. Various impurity measures, such as Gini Index, is used to measure purity of the nodes. Gini Index calculates probability of randomly picked data point that is classified incorrectly. The formula to calculate the Gini Index is

$$Gini\ Index = 1 - \sum_{i=1}^{n} p_i^2$$

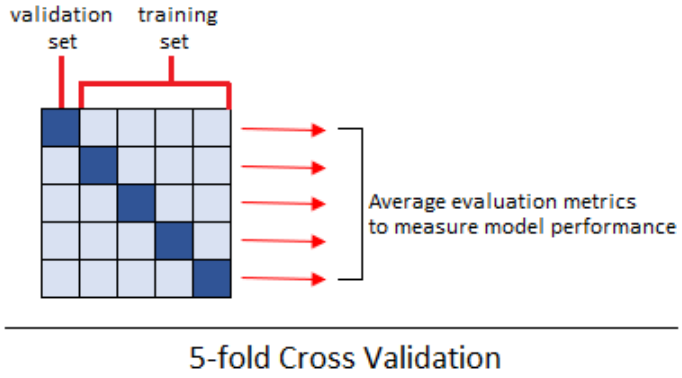
The following chart depicts values of Gini Index for nodes with different splits of a target classification variable.



For binary classification problems, if a node contains an even split of the target variable, then the Gini Index is 0.50. Hence the probability that a randomly picked data point is classified incorrect is random (p=0.50) which is expected due to the even split of two classes. If the node contains a higher percentage of one class, Gini Index decreases.

Cross Validation

This Random Forest classifier models presence of heart disease which is a binary variable. Standard classification metrics are used to optimize the model. This includes Area Under the Receiver Operating Curve, Recall, Precision and Accuracy.



These evaluation metrics are collected using 5-fold cross validation to avoid overfitting on the training set. This data is collected for each iteration in the hyperparameter tuning process.

Hyperparameter Optimization

Machine learning models use various parameter settings that have an impact on the cost function. Random Forest models involve a set of parameters that developers can optimize to evaluate the cost function. For example, one parameter setting is maximum depth of decision trees that are built. This setting allows developers to cap the depth of decision trees to avoid overfitting model on the training set. If this setting is not optimized, the tree is expanded until all leaves contain data points from the same class. This can lead to severe overfitting. The parameters tuned for modeling heart disease data optimized parameters including the number of decision trees, the number of features to consider for the best split, maximum depth of each tree, minimum number of data points required for

Heart Disease

Random Forest Classification model

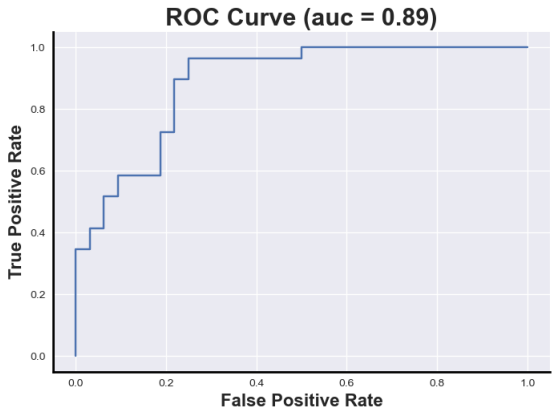
splitting a node, minimum number of data points required to be a leaf node and number of data points required to be a leaf node and whether a bootstrap should be used to build a tree instead of the full data set.

Model Results

Random Forest classifier model was run on the data set with 5-fold cross validation and hyperparameter tuning. Model evaluation metrics including Accuracy, Precision, Recall and ROC-AUC were collected for each iteration. The best parameter settings picked were based on maximizing the ROC-AUC during 5-fold cross validation. The metrics for the train and test sets based on these parameter settings are shown below.

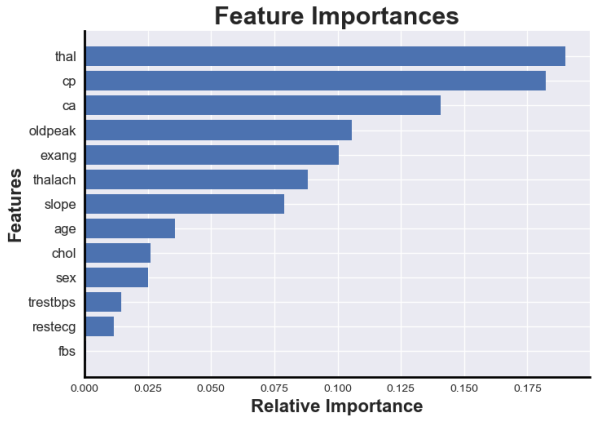
Metric	Training Set	Testing Set
ROC-AUC	0.886	0.883
Accuracy	0.892	0.885
Recall	0.947	0.938
Precision	0.868	0.857

An ROC curve was also obtained that plots the true positive rate (hit rate) against the false positive rate (false alarm rate) for different probability threshold values ranging from 0 to 1. A near perfect model would be one where the ROC curve is almost vertical (from x=0, y=0 to x=0, y=1) and then perfectly horizontal (from x=0, y=1 to x=1, y=1). Naturally, the curve for an optimal model is one that is closest to this behavior. The ROC curve for the model on heart disease data was obtained and is shown below.



Feature Importance

Once the best model is selected, the next step is to evaluate the significance of each feature in predicting the target variable using that model. The importance score can be calculated that is useful in understanding the model, patterns in the data and to reduce the number of features to simplify the model. Feature importance scores were calculated for the model trained on heart disease data set.



The 5 most important features adding value in the model included Presence of Fixed or Reversible Defect in Stress Echocardiography, Type of Chest Pain, Number of Vessels colored by Fluoroscopy, ST depression induced by exercise relative to rest, Exercised induced Angina. In general, all variables except Fasting Blood Sugar were important and should be included in the model.

Model Persistence

Once the best model is obtained and the evaluation metrics are satisfactory, the next step is to persist the model so that it can be used in the future without the need to retrain or reproduce it. This can be accomplished by object serialization in which a model is converted into byte stream and saved on the server. When the model is needed for predictions in the future, the byte stream is converted back into the model. Shell commands can be used to encrypt the model file and to allow only select users access to it.