A Comparison of K Nearest Neighbors (KNN) and Random Forest on predicting ten years risk of coronary heart disease

Glossary

accuracy	Percentage of correct predictions made by the model
attributes	A quality describing an observation
classification	Predicting a categorical output
class	One of a set of enumerated target values for a label
Confusion Matrix	Table that describes the performance of a classification model by grouping predictions into 4 categories
continuous features	Variables with a range of possible values defined by a number scale
data set	A collection of examples
decision trees	A model represented as a sequence of branching statements
ensemble	A merger of the predictions of multiple models
epoch	A full training pass over the entire dataset such that each example has been seen once
false positives	An example in which the model mistakenly predicted the positive class
features	An input variable used in making predictions
grid search	A tuning technique that attempts to compute the optimum values of hyperparameters
hyperparameter	Hyperparameters are higher-level properties of a model such as how fast it can learn (learning rate) or complexity of a model
majority class	the more common label in a class-imbalanced dataset
minority class	The less common label in a class-imbalanced dataset
objective	A metric that your algorithm is trying to optimize

objective function	The mathematical formula or metric that a model aims to optimize
overfitting set	Creating a model that matches the training data so closely that the model fails to make correct predictions on new data
oversampling	Reusing the examples of a minority class in a class-imbalanced dataset in order to create a more balanced training set
precision	In the context of binary classification (Yes/No), precision measures the model's performance at classifying positive observations (i.e. "Yes")
predictor	Synonym for feature
rank	The ordinal position of a class in a machine learning problem that categorizes classes from highest to lowest
Recall	Also called sensitivity. In the context of binary classification (Yes/No), recall measures how "sensitive" the classifier is at detecting positive instances
ROC curve	**Receiver Operating Characteristic curve.** A plot of the true positive rate against the false positive rate at all classification thresholds
supervised model	Training a model using a labeled dataset
synthetic data	A data not present among the input data, but created from one or more of them
testing set	A set of observations used at the end of model training and validation to assess the predictive power of your model
Training Set	A set of observations used to generate machine learning models
true positives	An example in which the model *correctly* predicted the positive class

Reference:

https://ml-cheatsheet.readthedocs.io/en/latest/glossary.html

https://developers.google.com/machine-learning/glossary#objective_function

Intermediate results including any negative results

Training both models with imbalanced data resulted in models predicting minority class incorrectly as seen on below confusion charts in figure 1 and figure 2.

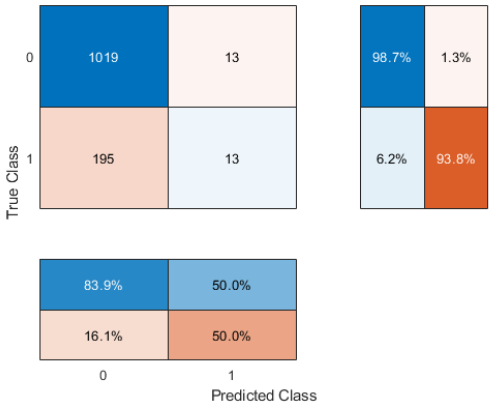

Figure 1. Random Forest Confusion chart

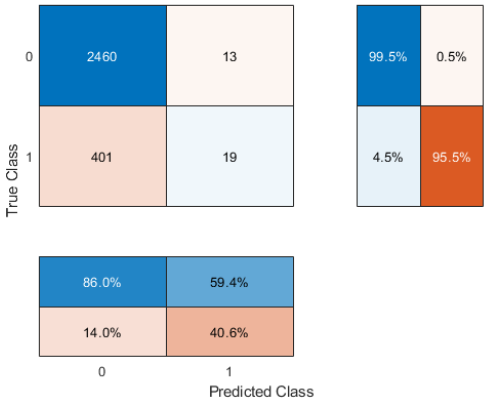

Figure 2. KNN Confusion chart

Implementation details including a brief description of main implementation choices

Python was used for initial analysis and visualization of histograms of attributes and target and exploring correlation between variables in a heat map. Also, SMOTE (Synthetic Minority Oversampling Technique) was implemented to dataset to get balanced classes for target.

The building process of machine learning functions was implemented in Matlab. For KNN classification fitcknn() built in function was used to fit the classification model to training data with chosen hyperparameters. For implementing Random Forest TreeBagger() built-in function were chosen. Both built-in functions were suitable for Classification problems.