ESE 542: Statistics for Data Science

Chapter 7 Project

Ensure that you follow the instructions to set the random seed values to ensure that your answers match the solutions. In addition to answering the **bolded** questions on Coursera, also attach your notebook, both as `.ipynb` and `.html`.

A hospital in Philadelphia is trying to better streamline its operations and plan for the future by reducing the number hospitalizations that could have been prevented. They have approached you, a data scientist, to help predict if patients will develop heart diseases so that healthcare providers have a chance to intervene much earlier. For this task, you have been provided with the Heart dataset, containing a binary outcome 'HD' that indicates the presence of heart disease in 303 patients. There are 13 predictors including 'Age', 'Sex', 'Chol' (a cholesterol measurement), and other heart and lung function measurements.

1. Load the Heart data and drop any rows with `NaN/null` values.
2. Binarize the 'HD' values such that `No=0` and `Yes=1`. One hot encode categorical features. Produce some numerical and graphical summaries of it. Do there appear to be any patterns?
3. What is the minimum age of the patients in the dataset? Assign this value to the variable `min_age`. **Input your answer onto Coursera.**
4. What is the maximum age of the patients in the dataset? Assign this value to the variable `max_age`. **Input your answer onto Coursera.**
5. Calculate the pairwise correlations between all the variables. Which predictor has the highest positive correlation with the response variable? Assign this to `highest_corr` and **input your answer onto Coursera.**
6. Since we are interested in building a predictive model, it is good practice to split the data into training and testing sets. Using `sklearn.model_selection.train_test_split`, divide the data into these sets using a 80/20 split with a `random_state=42`. These training and testing sets will be used for all the following parts.
7. Part A: Decision Tree

(a) Using all predictor variables, train a base classification tree to predict the response variable. Use `sklearn.tree.DecisionTreeClassifier` and set `random_state=42`.

(b) Plot the decision tree using the `Graphviz` package. Do you notice anything interesting?

(c) Evaluate your base model on the test set by calculating the precision, recall and accuracy. Assign these values to `dt_precision`, `dt_recall` and `dt_accuracy` respectively. **Input your answer onto Coursera.** *Hint*: A useful resource if you need a refresher of these terms is https://en.wikipedia.org/wiki/Precision_and_recall.

(d) Tune the hyperparameters of your model and evaluate this tuned model by calculating the precision, recall and accuracy on the test set. Assign these calculated values to `dt_tuned_precision`, `dt_tuned_recall` and `dt_tuned_accuracy` respectively. **Input your answer onto Coursera.** *Hint*: refer to the documentation to determine what can be tuned.

8. Part B: Bagging

(a) Using all predictor variables, train a base model to predict the response variable. Use `sklearn.ensemble.BaggingClassifier` and set `random_state=42`.

(b) Evaluate your base model on the test set by calculating the precision, recall and accuracy. Assign these values to `bag_precision`, `bag_recall` and `bag_accuracy` respectively. **Input your answer onto Coursera.**

(c) Do not modify `base_classifier`. Tune the hyperparameters of your model and evaluate this tuned model by calculating the precision, recall and accuracy on the test set. Assign these calculated values to `bag_tuned_precision`, `bag_tuned_recall` and `bag_tuned_accuracy` respectively. **Input your answer onto Coursera.** *Hint*: refer to the documentation to determine what can be tuned.

9. Part C: Random Forest

(a) Using all predictor variables, train a base random forest to predict the response variable. Use `sklearn.ensemble.RandomForestClassifier` and set `random_state=42`.

(b) Evaluate your base model on the test set by calculating the precision, recall and accuracy. Assign these values to `rf_precision`, `rf_recall` and `rf_accuracy` respectively. **Input your answer onto Coursera.**

(c) Tune the hyperparameters of your model and evaluate this tuned model by calculating the precision, recall and accuracy on the test set. Assign these calculated values to `rf_tuned_precision`, `rf_tuned_recall` and `rf_tuned_accuracy` respectively. **Input your answer onto Coursera.** *Hint*: refer to the documentation to determine what can be tuned.

10. Of the 3 base models that you trained, which model achieved the highest test accuracy? **Input your answer onto Coursera.**

11. Of the 3 models in which you tuned the parameters, which model achieved the highest test accuracy? **Input your answer onto Coursera.**

12. *Food for thought*: In this project, we asked you to pick the model with the highest test accuracy. For this task, do you think that accuracy is the best metric to use, or would precision, recall or even the F1-score be better?