Deliverable 3

Marwan Jabbour

Due: November 13, 2020

What have you changed since the previous deliverable?

Instead of defining a validation set, I decided to split my data into an 80/20 training/test set and use RandomizedSearchCV on my training set. I decided to use this instead of GridSearchCV because GridSearchCV takes too much time to run, and the accuracy is quite high with Randomized Search. N_estimators and max_features are 2 of the most important hyperparameters so I made sure to include them here. I then defined the random grid and used it with my random forest classifier. I chose 90 iterations.

```
[12] bootstrap = [True, False]
     number_of_trees = [int(x) for x in np.linspace(start = 200, stop = 2000, num = 10)]
     max_features = ['auto', 'sqrt']
     split_min_samples = [2, 5, 10]
     leaf_min_samples = [1, 2, 4]
```

```
[13] random_grid = {'bootstrap': bootstrap,
                    'n_estimators': number_of_trees,
                    'max_features': max_features,
                    'min_samples_split': split_min_samples,
                    'min_samples_leaf': leaf_min_samples,
                    }
```

```
[15] rf = RandomForestClassifier()
     rf_random = RandomizedSearchCV(estimator = rf, param_distributions = random_grid, n_iter = 90, cv = 3,
     rf_random.fit(X_train, Y_train)
```

If so, how have your changes improved the results? Provide graphs.

```
                F1 score:
*********************************

0.8864507364507364
                Accuracy score:
*********************************

0.8975609756097561

                Confusion Matrix:
*********************************

[[0 0 0 ... 0 0 0]
 [0 3 0 ... 0 0 0]
 [0 0 1 ... 0 0 0]
 ...
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 1 0]
 [0 0 0 ... 0 0 1]]


          Classification report:
***********************************

       precision    recall  f1-score   support

126.0       0.00      0.00      0.00         1
149.0       1.00      1.00      1.00         3
157.0       1.00      1.00      1.00         1
164.0       1.00      1.00      1.00         1
167.0       1.00      1.00      1.00         1
168.0       0.00      0.00      0.00         0
169.0       1.00      1.00      1.00         1
174.0       1.00      1.00      1.00         1
175.0       1.00      1.00      1.00         4
176.0       1.00      1.00      1.00         2
177.0       1.00      1.00      1.00         4
178.0       1.00      1.00      1.00         1
183.0       1.00      1.00      1.00         1
184.0       1.00      1.00      1.00         1
188.0       1.00      1.00      1.00         2
192.0       1.00      1.00      1.00         2


              Mean Squared Error:
***********************************

167.94634146341463
              Rand Score:
***********************************

0.7983882225116387
```

I also noticed that the mean squared error is a bit high, but after doing some research I discovered it might be a rescaling issue. I will ask my TPM about this.

Final demonstration proposal: *Draft*

Welcome to the Heart Disease Predictor

How it works

Try it now!

How it Works

1) Algorithm behind it
2) Mention Dataset used
3) Word of caution: not official …

Heart Disease Predictor

Age

Sex

m ◯ f ◯

Chest Pain Type (1-4)

1 ◯ 2 ◯ 3 ◯ 4 ◯

...

Results

Warning: **Not official test. Consult physician**

You are not at risk of heart disease.

You are at risk of heart disease.