**Homework #2 Instructions**

**Casey Bennett, PhD**

**DSC540, Winter 2019**

**DePaul University**

**Overview**

There are two python scripts, implementing Scikit using the same basic template you were shown in class. The first one uses a dataset related to Diabetes from Pima Indians for classification, the target variable being the presence/absence of diabetes in individuals. The second script uses a dataset of quality ratings for various red wines rated 1-10, for regression. We will explore ways of using Random Forests and Bagging to create models, as well as talk about the effects of feature normalization.

For classification, we will be creating an object we’ll name ‘clf’, and for regression we’ll name ‘rgr’. These are objects we can call methods on (such as fitting a model to some data), and access their internal variables (such as getting predicted class labels). Scikit API links for Random Forests describing methods and variables available can be found in the included links document.

*\*Follow the steps below, record answers to questions in a word document, and turn in both your completed code and the word doc.*

**Pima Diabetes**

Open up HW2\_Diabetes.py

1. First, let’s run a simple test/train split using a random forest. To do so, we need to do two things, first create a Random Forest Classifier object (clf), then “fit” some data using that object.
   1. Note on line 191, the test/training data is already split for you into separate arrays for both the features (data) and the target
   2. On line 196, we need to create a Random Forest Classifier. We can do this using RandomForestClassifier(). Using the API link in the accompanying document, call that function, and pass in the following parameters:
      1. Set number of trees to 100
      2. Set maximum depth = None
      3. Set minimum # of samples for split to occur = 3
      4. Set splitting criterion = 'entropy'
      5. Set random\_state variable to rand\_st
   3. Now we need to fit data. On line 197, call a fit() function on the clf object. It takes two arrays, the first for the training set feature data (data\_train) and the second for the training set target data (target\_train).
   4. Note that performance is already be calculated here for you on lines 199 and 201, using the test set (data\_test and target\_test)

*\*Question #1a: Run the code 5 times, record the accuracy and AUC score. What do you notice about the scores?*

*Record score 5 times*

1. *﻿﻿Random Forest Acc: 0.7695167286245354*

*Random Forest AUC: 0.8293552569542668*

1. *﻿Random Forest Acc: 0.7769516728624535*

*Random Forest AUC: 0.8474382968768277*

1. *﻿Random Forest Acc: 0.7397769516728625*

*Random Forest AUC: 0.8058109382366808*

1. *﻿Random Forest Acc: 0.7695167286245354*

*Random Forest AUC: 0.8117152184831743*

1. *﻿Random Forest Acc: 0.7843866171003717*

*Random Forest AUC: 0.8520904382973349*

*For Acc score, the range is around 0.73-0.78 which is a good score. The range is around +/- 0.05. And, AUC score is around 0.80-0.85 which is also a good score with +/- 0.05 range.*

*\*Question #1b: For the fit() method of a RandomForestClassifier, it lists three possible parameters on the API webpage, what are they? Define what you could pass in to each one?*

*X (Training input sample) ,*

*y(target value) class labels in classification or real numbers in regression.*

*sample\_weight - …*

1. Let’s repeat step 1 above, this time using cross-validation. To do so, we need to do two things, first create a Random Forest Classifier object (clf), then second pass that object and some data arrays into a “cross-validate” function.
   1. To turn on cross-validation, we need to first on line 32 change the cross\_val flag to equal 1 instead of 0
   2. On line 211, we need to create a Random Forest Classifier. We can do this using RandomForestClassifier(). Using the API link in the accompanying document, call that function, and pass in the following parameters:
      1. Set number of trees = 100
      2. Set maximum depth = None
      3. Set minimum # of samples for split to occur = 3
      4. Set splitting criterion = 'entropy'
      5. Set random\_state variable to rand\_st
   3. Now we need to run cross-validation. On line 212, call a cross\_validate() function and return the value to a scores object (you can do this by setting the scores = cross\_validate() ). For the cross\_validate function, pass in the following parameters:
      1. Set the estimator = clf
      2. Set the data to be fit = data\_np
      3. Set the target variable = target\_np
      4. Set scoring = scorers
      5. Set the # of cv folds = 5
   4. Note we are passing in the WHOLE dataset here, not just training data or training target values
   5. Note that performance is already be calculated here for you with the scorers being setup on line 207, and the results pulled out and printed below.

*\*Question #2: Run the code once, record the accuracy and AUC score.*

*﻿Random Forest Acc: 0.77 (+/- 0.08)*

*Random Forest AUC: 0.83 (+/- 0.07)*

*﻿CV Runtime: 0.9143397808074951*

1. Let’s explore how the number of trees affects performance of a Random Forest.
   1. On line 211, change the number of trees from 100 to 5
   2. Now set the number of trees to 10
   3. Now set the number of trees to 20
   4. Now set the number of trees to 50
   5. Now set the number of trees to 200
   6. Now set the number of trees to 500
   7. Now set the number of trees to 1000

*\*Question #3: Run the code once for each setting of the number of trees (5,10,20,50,100,200,500, 1000), record the accuracy and AUC scores. What do you notice about the scores? How do they change as the number of trees increases?*

5 trees﻿

﻿ ﻿Random Forest Acc: 0.74 (+/- 0.06)

Random Forest AUC: 0.77 (+/- 0.07)

CV Runtime: 0.056884050369262695

10 trees

﻿ Random Forest Acc: 0.74 (+/- 0.05)

Random Forest AUC: 0.80 (+/- 0.08)

CV Runtime: 0.1005849838256836

20 trees

﻿ ﻿Random Forest Acc: 0.76 (+/- 0.06)

Random Forest AUC: 0.81 (+/- 0.07)

CV Runtime: 0.18755507469177246

50 trees

﻿ ﻿Random Forest Acc: 0.77 (+/- 0.06)

Random Forest AUC: 0.83 (+/- 0.07)

CV Runtime: 0.4405341148376465

200 trees

﻿ ﻿Random Forest Acc: 0.77 (+/- 0.07)

Random Forest AUC: 0.83 (+/- 0.06)

CV Runtime: 1.765631914138794

500 trees

﻿ ﻿Random Forest Acc: 0.78 (+/- 0.07)

Random Forest AUC: 0.83 (+/- 0.06)

CV Runtime: 4.273550033569336

1000 trees

﻿ ﻿Random Forest Acc: 0.77 (+/- 0.07)

Random Forest AUC: 0.83 (+/- 0.07)

CV Runtime: 8.495339155197144

First of all, I noticed that both Acc and Auc are increase when the number of tree increase until the model meet the stopping point where It will not increase the score anymore even increasing more trees. Also, the more tree that is increase in the model the more time that need to run the model.

1. Now let’s try applying feature selection method we used for the wine dataset in Homework #1 to the diabetes dataset. We will turn on the Wrapper-Based Feature Selection, which essentially builds lots of models with different subsets of features, and picks the subset that performs the best. For simplicity here though, we will just build a single subset and select the top variables. We will use the same Random Forest model for this.
   1. First, on line 211, change the number of trees back to 100
   2. To turn on feature selection, we need to first on line 37 change the feat\_select flag to equal 1 instead of 0
   3. Note that there is an option to change the feature selection type on line 38, but the homework code is hard-coded to only use wrapper-based, so this doesn’t matter for now
   4. You will need to add a RandomForestClassifier() call to pass to the clf object on line 148, you can use something similar to the calls used elsewhere in the code (e.g. line 196). Don’t forget to set the parameters, particularly the random\_state and number of trees to 100.
   5. Note the SelectFromModel() function being called on line 149, this is where the actual feature selection occurs, with the clf object being passed in

*\*Question #4a: Run the code once, record the accuracy and AUC scores. What do you notice about the scores? How do they compare to the performance above in question #2?*

*﻿Random Forest Acc: 0.76 (+/- 0.05)*

*Random Forest AUC: 0.82 (+/- 0.06)*

*CV Runtime: 0.9646050930023193*

*I noticed that the accuracy is around 0.76 with +/- 0.05 range which is moderate to strong performance for the model. And, 0.82 +/- 0.06 on AUC score which is a good compliment to the accuracy. To compare this performance to question 2 performance, with 100 trees, it performed very close to the best performance in the question 2 with the similar range of score fluctuation.*

*\*Question #4b: What features were selected, and which were removed?*

Selected Feature : Blood Glucose, BMI, Family History, Age

Not Selected Feature : Time Pregnant , Blood Pressure, Skin Fold Thickness,

2-Hour Insulin

1. Random Forests and similar tree methods also produce a “feature importance” score that can also be used for feature selection. Let’s try manually setting up a feature selection section for that.
   1. Feature selection should already be turned on, with line 37 having feat\_select flag set to equal 1
   2. On line 38, change fs\_type from 2 to 4
   3. You will need to add a RandomForestClassifier() call to pass to the clf object on line 156 (under where it says “if fs\_type=4”), you can use something similar to the calls used elsewhere in the code. Don’t forget to set the parameters, particularly the random\_state and number of trees to 100.
   4. You will then need a line to fit the data. On line 157, call a fit() function on the clf object.
   5. On the line below that, add a line creating an empty array, and call it sel\_idx.
   6. Now we need to create a For loop, taking each value out of the feature\_importances\_ in the clf object in turn. Call this value ‘x’.
   7. Within the For loop, create an If-Else statement. Set “If” so that if x >= mean of the feature\_importances\_ array, then append the value of 1 to sel\_idx. For “Else”, append a 0.
   8. Hint, you can use numpy to calculate means and other statistics of arrays, e.g. np.mean (numpy is given the alias of ‘np’ in the code).
   9. Try to run the code.

*\* Question #5: Run the code once, record the accuracy and AUC scores. What features were selected, and which were removed? How do the selected features compare to what you saw in Question #4 above? Was the performance (accuracy, AUC) different than in Question #4?*

﻿ ﻿Random Forest Acc: 0.76 (+/- 0.05)

Random Forest AUC: 0.82 (+/- 0.06)

CV Runtime: 0.9324469566345215

Selected Feature : Blood Glucose, BMI, Family History, Age

Not Selected Feature : Time Pregnant , Blood Pressure, Skin Fold Thickness,

2-Hour Insulin

The selected features show the same the score from both Acc and AUC, but CV runtimes in question 5 (manual coding for selecting feature) is slightly faster.

**Wine Quality Dataset**

Open up HW2\_Wine.py … First, let’s repeat the steps we did above for Diabetes, with some tweaks. We will skip the Train/Test version, and jump right to the Cross-Val version.

1. Let’s repeat what we did for the Diabetes dataset in Question #2 above here for the Wine dataset. To do so, we need to do two things, first create a Random Forest Classifier object (clf), then second pass that object and some data arrays into a “cross-validate” function.
   1. On line 32, cross-validation is already turned on for you (set to 1)
   2. On line 211, we need to create a Random Forest Regressor. We can do this using RandomForestRegressor(). Using the API link in the accompanying document, call that function, and pass in the following parameters:
      1. Set number of trees = 100
      2. Set max features (for each split) = .33
      3. Set maximum depth = None
      4. Set minimum # of samples for split to occur = 3
      5. Set splitting criterion = 'mse'
      6. Set random\_state variable to rand\_st
   3. Now we need to run cross-validation. On line 212, call a cross\_validate() function and return the value to a scores object (you can do this by setting the scores = cross\_validate() ). For the cross\_validate function, pass in the following parameters:
      1. Set the estimator = rgr
      2. Set the data to be fit = data\_np
      3. Set the target variable = target\_np
      4. Set scoring = scorers
      5. Set the # of cv folds = 5
   4. Note we are passing in the WHOLE dataset here, not just training data or training target values
   5. Note that performance is already be calculated here for you with the scorers being setup on
   6. line 207, and the results pulled out and printed below.

*\*Question #6: Run the code once, record the RMSE and Expl Variance.*

﻿Random Forest RMSE:: 0.65 (+/- 0.02)

Random Forest Expl Var: 0.33 (+/- 0.11)

CV Runtime: 1.4887888431549072

1. Let’s explore how the number of trees affects performance of a Random Forest.
   1. On line 211, change the number of trees from 100 to 5

﻿Random Forest RMSE:: 0.70 (+/- 0.04)

Random Forest Expl Var: 0.20 (+/- 0.13)

CV Runtime: 0.07713484764099121

* 1. Now set the number of trees to 10

﻿Random Forest RMSE:: 0.67 (+/- 0.04)

Random Forest Expl Var: 0.28 (+/- 0.09)

CV Runtime: 0.1434922218322754

* 1. Now set the number of trees to 20

﻿Random Forest RMSE:: 0.66 (+/- 0.04)

Random Forest Expl Var: 0.30 (+/- 0.09)

CV Runtime: 0.2764778137207031

* 1. Now set the number of trees to 50

﻿

Random Forest RMSE:: 0.65 (+/- 0.03)

Random Forest Expl Var: 0.32 (+/- 0.11)

CV Runtime: 0.6833019256591797

* 1. Now set the number of trees to 200

﻿Random Forest RMSE:: 0.64 (+/- 0.02)

Random Forest Expl Var: 0.34 (+/- 0.10)

CV Runtime: 2.7352540493011475

* 1. Now set the number of trees to 500

﻿Random Forest RMSE:: 0.64 (+/- 0.02)

Random Forest Expl Var: 0.34 (+/- 0.11)

CV Runtime: 6.572230100631714

* 1. Now set the number of trees to 1000

﻿Random Forest RMSE:: 0.64 (+/- 0.02)

Random Forest Expl Var: 0.34 (+/- 0.10)

CV Runtime: 14.008708715438843

*\*Question #7a: Run the code once for each setting of the number of trees (5,10,20,50,100,200,500, 1000), record the RMSE and Expl Variance. What do you notice about the scores? How do they change as the number of trees increases? Is this the same as you for the Diabetes dataset in Question #3?*

*The RMSE stopped decreasing at 200 trees with 0.64 and Explained Variance stopped increasing at 0.34. I noticed that the RMSE was decrease and Expl Var was increase along with the increasing of tree number. We tried to set up more tree number at 500 tree and 1000 to see the difference but got the same score with more runtimes. The idea is the same that the best model performance come with the right number of tree.*

*\*Question #7b: What about run-times, how do those change as you change the number of trees? What do the changes in scores and run-times tell us about choosing the right number of trees?*

The run-times made a different even if the dataset was small. In this case, the run-times showed the obvious effect around 500 trees in the model at 6.5 second and 1000 trees at 14 second. It told us that it was better to assign the right number of tree to prevent the high running time which it will show real effect on a huge dataset and can have a problem with cost of operation.

1. Now let’s turn on Wrapper-Based Feature Selection, which essentially builds lots of models with different subsets of features, and picks the subset that performs the best. For simplicity here though, we will just build a single subset and select the top variables. We will use the same Random Forest model for this.
   1. First, on line 211, change the number of trees back to 100
   2. To turn on feature selection, we need to first on line 37 change the feat\_select flag to equal 1 instead of 0
   3. Note that there is an option to change the feature selection type on line 38, but the homework code is hard-coded to only use wrapper-based, so this doesn’t matter for now
   4. You will need to add a RandomForestRegressor() call to pass to the clf (must be rgr) object on line 148, you can use something similar to the calls used elsewhere in the code (e.g. line 196). Don’t forget to set the parameters, particularly the random\_state and number of trees to 100.
   5. Note the SelectFromModel() function being called on line 149, this is where the actual feature selection occurs, with the clf (rgr) object being passed in

*\*Question #8a: Run the code once, record the RMSE and Expl Variance. What do you notice about the scores? How do they compare to the performance above in question #6?*

*Random Forest RMSE:: 0.68 (+/- 0.03)*

*Random Forest Expl Var: 0.25 (+/- 0.11)*

*CV Runtime: 0.876276969909668*

*With 100 trees and selected features, RMSE was 0.68 and Expl Var was 0.25. Compared the score to question 6, the performance with the selected features was worse than performance from question ,but it has better run-times.*

*\*Question #8b: What features were selected, and which were removed?*

Selected Feature : ﻿'volatile acidity', 'sulphates', 'alcohol'

Not Selected Feature : ﻿'Class', 'fixed acidity', 'citric acid', 'residual sugar', 'chlorides', 'free sulfur dioxide', 'total sulfur dioxide', 'density', 'pH'

1. Let’s see if normalizing the features changes performance. To do this we are going to use the simple scaling function, but there are many different types of normalization available in Scikit (see the preprocessing link in the accompanying API document).
   1. First, turn off feature selection, on line 37 change the feat\_select flag back to 0
   2. To turn on feature normalization, on line 34 change the norm\_features flag to equal 1
   3. In the Preprocessing Section, on line 96, we need to call the scale() function, and pass into it the WHOLE dataset of features, which is data\_np. See the API link in the accompanying document.
   4. Note that we are simply returning the output of the scale function (i.e. normalized features) back to the same object.

*\*Question #9: Run the code once and record the RMSE and Expl Variance. What do you notice about the scores? How do they differ from question #6?*

*﻿Random Forest RMSE:: 0.65 (+/- 0.02)*

*Random Forest Expl Var: 0.33 (+/- 0.11)*

*CV Runtime: 1.375281810760498*

*The RMSE was 0.65 and Exp Var was 0.33 which were the same score as in question 6. And, they also had the same range of score fluctuation. The difference was the run-times which in question 9 was slightly better.*

1. Finally, let’s try to setup Bagging, using Decision Tree Regressors. Remember though, Bagging can use any sort of classifier or regressor as an ensemble, but generally simple fast methods (decision trees, K-nearest neighbors, naïve bayes, etc) are best because you are going to build a whole bunch of them. A neural network, for instance, would probably take too long on any decent sized dataset.
   1. First, turn off feature normalization, on line 34 change the norm\_features flag back to 0
   2. We’ll do this in the single test/train split section, so turn cross-validation off on line 32 by setting the cross\_val flag to 0
   3. We need to import the function into our script before we can call it. On line 11, add BaggingRegressor to the comma-separated list of packages imported from the sklearn.ensemble module.
   4. Let you do the rest yourself. You can see between lines 195-202 I’ve left space for you code You’ll need to add a DecisionTreeRegressor() on line 195. You can use your code from HW1 if you want, with the same parameters. Make sure you set random\_state= rand\_st.
   5. Create a BaggingRegressor() on the line below, pass it to the bag object, then fit the bag to data. Make sure you set parameters:
      1. Set max\_samples =0.6
      2. Set random\_state = rand\_st
   6. You will need to create some scoring on the lines down below replacing the commented lines. Refer back to Homework 1, but protip: we need to score/predict using the bag object, not the rgr object.

*\*Question #10a: Run the code 5 times and record the RMSE and Expl Variance. What do you notice about the scores? How stable are they? How do they differ from question #6?*

*1. ﻿Decision Tree RMSE: 0.6236700134801141*

*Decision Tree Expl Var: 0.3818222549561163*

*2. ﻿Decision Tree RMSE: 0.6339079473506273*

*Decision Tree Expl Var: 0.3844499727939842*

*3. ﻿Decision Tree RMSE: 0.588521149031318*

*Decision Tree Expl Var: 0.39524983979270567*

*4. ﻿Decision Tree RMSE: 0.6361996541966995*

*Decision Tree Expl Var: 0.4009306676241048*

*5. ﻿Decision Tree RMSE: 0.5891428398226407*

*Decision Tree Expl Var: 0.4440516478287342*

*In this method, I noticed that RMSE was lower and Expl Var was higher which were better performance than question 6. Both RMSE and Expl Var were quite stable which RMSE was around .58 - .63 and Expl Var was around .38 - .44. Moreover, 3 out of 5 runs from both scores had a very close range. The difference between this two were obviously the methods that were used. In question 6, it used randomforest which computed multiple trees , but this question used decision tree with bagging method. The performance was also different that question 10 performed better.*

*\*Question #10b: Based on the API webpage for the BaggingRegressor() in the accompanying API links document, what two parameters do we need to change to create a Random Subspaces model?*

*bootstrap : set to ‘False’*

*bootstrap\_features : set to ‘True’*

*\*Question #10c: Based on the API webpage for the BaggingRegressor(), notice that the function when called also calculates the out-of-bag error (oob\_score). Should we be using that metric then rather than a test/train split or cross-validation with bagging, or can we use them together (hint: do some googling)?*

*As I did some research on OOB, CV and test/ train split, I believe we cannot use it at the same time. However, We can try to compare performance by using each one of them which has the difference way to validate model. Cross Validation will validate model based on the number of fold such as 10 folds will have 90% of training set and 10 % test set , on the other hand OOB calculates the error of from the tree which it splits training set around 66% and test set for 33%. However, we can set cv to 3 to do get the split as close as OOB to compare performance. In the bigger dataset, OOB may be useful than CV in term of time and cost.*

**Summary Questions**

*\*Question #11: Compare the performance of Random Forests here to the Decision Tree models for both datasets in Homework #1. Did Random Forest perform better, worse, or the same? If your boss or customer asked why that might be, how would you explain?*

*For Pima diabetes dataset, for the train / test split , RF tended to have a better performance than DTs with Acc was around .73 to .78 and Auc was around 0.8-0.85, but with DT the performance is roughly 0.05 lower in both score. For the cross-validation, It appeared to show the same trend that RT had a better performance with a range of Acc 0.06 and Auc 0.14 more in average score.*

*In Wine Quality dataset, Cross -validation 5 folds , the performance from DT was significant better than RT with RMSE 0.01 and Expl Var 1.00 , and also DT ran faster. In addition with the feature selection applied to the model, both method had difference performance. It appeared that RT performed better with feature selection with RMSE at 0.68 and Expl Var at 0.25, on the other hand, DT with feature selection had high RMSE at 0.91 and significant difference in Expl Var at -0.35.*

*I would tell my boss or my customer to do DTs first since it is easier to explain and modeling it, but I would also recommed them to try RT because of it has an overall better in term of performance. However, RT may need a higher computational power because it runs more trees and it will use more cost to operate.*

*\*Question #12: We’ve now seen several different sampling and evaluation techniques. When it comes to evaluating model performance, what is the “gold standard” approach?*

*The general ‘gold standard’ would be cross-validation with the respectively 10 fold would be a great start. But, it is still depend on the size of dataset and the method that use in the model. For example, OOB may be better suit for RT to observe the amount of error than cross validation*