# Random Forest; Regression Task; Example in Python; AMS 580

# Name:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_SBU ID:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Name:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_SBU ID:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Name:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_SBU ID:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Name:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_SBU ID:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Please include (1) Python file; (2) Output from Python with answers to all the questions asked; (3) Comparison of the results to those using R; (4) Recommended websites for Random Forest using Python.

#### Random Forest with the Boston Data – Regression Task

The **Boston** data set is a built-in data in R [**MASS package**]. **There are 14 variables:**

crim - per capita crime rate by town

zn - proportion of residential land zoned for lots over 25,000 sq.ft.

indus - proportion of non-retail business acres per town.

chas - Charles River dummy variable (1 if tract bounds river; 0 otherwise)

nox - nitric oxides concentration (parts per 10 million)

rm - average number of rooms per dwelling

age - proportion of owner-occupied units built prior to 1940

dis - weighted distances to five Boston employment centres

rad - index of accessibility to radial highways

tax - full-value property-tax rate per $10,000

ptrat10 - pupil-teacher ratio by town

black - the proportion of African Americans by town

lstat - % lower status of the population

**medv - Median value of owner-occupied homes in $1000's**

Your task is to split the data randomly into training (80%) and testing (20%), first build the best random forest to predict ‘**medv**’ using the training data, then use the out-of-bag (OOB) data to measure its performance, and then use that model to predic ‘**medv**’in the testing data. Please note that we use *randomforest* function in R to build the random forest model.

Please review the following website for related methods and procedures in R:

<http://www.sthda.com/english/articles/35-statistical-machine-learning-essentials/140-bagging-and-random-forest-essentials/>

1. How many observations do we have in the data?
2. Please use the random seed 123 to divide the cleaned data into 80% training and 20% testing.

1. Please first build the best random forest to predict ‘**medv**’ using the training data. Please use 10-fold cross-validation to obtain the best tuning parameter **mtry**, which is the best number of random variables to select (without replacement) for each tree node.
2. Next please use this random forest to predict ‘**medv**’ in the testing data. Please compute the **RMSE** for the testing data. Please add the predicted class label to the testing dataset.
3. Please use the **randomForest** function to refit the training data using the optimal **mtry** parameter and with 500 trees.
   1. Please print out the model summary, and write down the **RMSE ()** of the **out of bag (OOB)** samples.
   2. Show the variable importance values.
   3. Plot the variables importance measures using **MSE** and Node Impurity.
4. In a regression task using the random forest, suppose we have 12 variables (as predictors) in the original data set – then at each node split, what is the number of variables we should (as commonly recommended) to select, at random, to be considered for that node split?

