COMP4432 - Assignment 2

(due by midnight MST the day prior to Live Session 4)

Part 1: Data Exploration.

Load the scikit-learn diabetes bunch object into a variable. Print out the description of the dataset. Load the diabetes features into a pandas dataframe with the proper column names. Add the target variable to this same dataframe. Run a command to look at the data types of your dataframe to see if there is any missing data. Perform descriptive statistics on the numeric columns of your dataframe. Plot histograms of your data to get a feel for each column's distribution. Split your dataframe into a training and test set with 20% of your data being in the test set. Define a correlation matrix. Look at values highly correlated with the target. Plot the correlation matrix with a Seaborn heatmap. Use a Seaborn pairplot to look at the scatter plots of the three values with the highest target correlation. Prepare a feature set by dropping the target from your training dataframe. Copy your training target into a new dataframe.

Part 2: Model Training.

Train a linear regression model using your training set. Print the RMSE of your regression model on your training set. Implement a cross_val_score on a decision tree regressor on your training set. Print out root mean and standard deviation of the cross-validation scores. Do the same for a RandomForestRegressor. Record which model performs better.

Part 3: Model Tuning.

Print out the parameters of your random forest model. Do a grid search cross-validation with the following values: n_estimators: 3,10,30 and max_features: 2,4,6,8, as well as the following experiment: bootstrap: False, n_estimators: 3,10 and max_features: 2,3,4. Print out the best parameters and the best performing model based on this grid search. Using the cv_results dictionary, print out the rmse of each feature combination for comparison. Also print out the feature importances of the best performing grid search model. Describe how it compares with the correlation matrix we implemented earlier.

Part 4: Model Evaluation.

Document the best-performing model between the single feature model you trained in Assignment 1, and the models you trained in part 2 and 3 of this assignment. Evaluate the best performing model against your test set. Save your model for future use.