Case Study 3

BSDS Spring 2021

Instructions

- You must work in your assigned group on Friday, 5/7 in a breakout room during the live lecture (unless you have made prior arrangements).
- On Wednesday, 5/12 each group must present one or two of the open-ended problems in the assignment and submit their Jupyter Notebook.
- Each group will submit a Jupyter Notebook file (.ipynb) and an associated .html file via Canvas. Each group member will submit the same two files, except for the "equal work pledge".
- The submitted notebook must start with a Markdown cell header indicating the Case Study number, dataset, group members, and any external sources. Everyone will additionally include an "equal work pledge" which states that they understand all code/answers in the assignment and that all group members contributing equally. If this is not the case please indicate.
- Each question must be annotated appropriately with Markdown cells. The Notebook file should written in a way that a third party with no knowledge of the questions can read it.
- All group members will receive the same grade unless the "equal work pledge" is violated.

The dataset for this assignment in available on Canvas and comes from the UCI Machine Learning Repository. We will see if it is possible to predict Wine Quality from different chemical properties of white wine. You may use `lm` and `glm` to perform the regressions below, but this is not necessary.

1. Import the dataset consisting of different chemical properties of each wine along with a Wine Quality rating then do the following:

- Set a random seed for the assignment (not 2021!)
- Make sure all the columns of your tibble are of type Double.

2. Write a function that takes in a tibble and outputs a list of three tibbles representing a train, validation, and test split of 60/20/20. (Bonus point if you write a function that also takes in three percentages for different splits!)

3. Pick 5 of the variables that are not Wine Quality.

(a) For each variable create a linear model to predict quality from this variable.

(b) Use your validation set and the Root Mean Squared Error metric to hypothesize which of these 5 variables is the most important.

(c) Does the test set validate your hypothesis?

4. Now create a linear model to predict quality with all five of the previous variable as inputs and report the RMSE.

5. Do the following using the same five variables as above.

(a) Create a logistic model to predict whether a wine has quality strictly above 5 or not. What is the accuracy of your model on the validation set?

(b) What is the accuracy of your linear model on the validation set from Question 4?

(c) Which model is better? Use your test set to validate your hypothesis.

6. Read the article here.

(a) Create a histogram of wine quality scores and also provide the count of wines with quality above 5 and those 5 or below. Is the binary classification task from Question 5 *imbalanced*? Why might using accuracy as a metric be an issue here?

(b) Create a confusion matrix (evaluated on the validation set) for both of your models from Question 5.

7. Use what you've done in the previous questions to for the following.

(a) Pick up to three variables that you think give a good linear model, meaning it generalizes well to unseen data. Why do you think this?

(b) Use those three variables to fit a linear model to the **entire** dataset. This is your final regression model! List the variables and the fitted model parameters.

(c) Choose the model type that you think will give the best accuracy on unseen data. Fit that model to the **entire** dataset. This is your final binary classification model! List the model (and variables) and the fitted model parameters.