DSI Instructor Challenge

Your Challenge, Should You Choose to Accept It

Welcome! Thanks for your interest in our Data Science Immersive program. In order to move forward, we'd like you to complete the following 2-part challenge.

Part 1: Modelling - Evaluates your ability to perform standard data science tasks in Python.

Part 2: Feedback - Evaluates your ability to review & debug student code while providing them with relevant feedback.

Submission

Please create a jupyter notebook and save your code in a github gist. In your notebook, create two markdown headings: #Part 1 & #Part 2.

Note: Please take *no more than 8 hours* to complete both of these challenges.

When you're finished, please submit a link to your Jupyter notebook using nbviewer. Afterward, one of our instructors or internal subject matter experts will review your submission and we'll follow up with you. Good luck!!

DSI Instructor Challenge

Part 1: Modelling

In this challenge, we are looking to assess your Python coding ability and data science skills.

Instructions

We'll be working on the following dataset, which includes measurements of breast cancer cells:

- DSI-Instructor-Challenge/part-1-dataset/breast-cancer.csv
- DSI-Instructor-Challenge/part-1-dataset/field_names.txt

The task is to predict for each cell, whether it is *malignant* or *benign*. Please follow the guidelines below.

Note: Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

Python Coding and Data Set

- Load in the data file and header file provided
    - The dataframe does not currently have a header, load in the header file and attach it to the dataframe
- Comment on any steps you might take to evaluate or transform the dataset.
- Compute the mean and median smoothness and compactness for benign and malignant tumors - do they differ? Explain how you would identify this.
- Write a function to generate bootstrap samples of the data.

Exploratory Analysis

- Identify 2-3 variables that are predictive of a malignant tumor.

    - Display the relationship visually and write 1-2 sentences explaining the relationship.

Modelling

- Build a model to predict the malignant tumors.
    - Use at least two classification techniques; compare and contrast the advantages and disadvantages of each.
    - Identify how you would control for overfitting in each classification technique.
    - Evaluate the performance of each model.
    - In each model, identify the most important predictive variables and explain how you identified them.

Explanation

- To Technical Audiences
    - Explain the limitations of your analysis and identify possible further steps you could take.
- To Non-Technical Audiences
    - Write a short summary of your analysis, explaining how your model works and how it performs.
    - Briefly explain the factors that contributed to malignant vs benign tumor identification.

DSI Instructor Challenge

Part 2: Feedback

During your role in our data science immersive, you'll need to be able to quickly scan and debug student code, while providing helpful feedback or tips. Therefore, this challenge is intended to help you demonstrate your instructional skills!

Instructions

In your notebook, move onto "Part 2" and log your responses accordingly. Please provide feedback for both of the sample student submissions provided.

- DSI-Instructor-task/part-2-student-samples/student-sample-1.py
- DSI-Instructor-task/part-2-student-samples/student-sample-2.py

Use the following guidelines to help structure your responses:

1. Code
    - Feel free to comment on style, library usage, or other improvements.
2. Methodology
    - Feel free to comment on the student's data setup, modeling methodology, and model evaluation.
3. Conceptual Understanding
    - Finally, feel free to add any suggestions or takeaways on how the student could continue to improve their understanding of these concepts.

Note: Assume the student is a relative beginner with novice programming skills. Their background may include college courses in data or statistics, but don't assume that they are comfortable with these concepts.

Note: For your reference, student samples refer to the following dataset:

- DSI-Instructor-task/part-2-dataset/part-2-data.train.csv

All set? When you're finished, please submit a link to your Jupyter notebook using nbviewer to render your gist.