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

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

Submission

In your notebook, create two markdown headings: #Part 1 & #Part 2.

Please mention the time taken to complete both challenges.

**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:

● breast-cancer.csv

● 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.

**Part 2: Feedback**

**Instructions**

In your notebook, move onto "Part 2" and log your responses accordingly. Please provide

feedback for both of the sample submissions provided.

● sample-1.py

● 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 data setup, modeling methodology,

and model evaluation.

3. Conceptual Understanding

○ Finally, feel free to add any suggestions or takeaways on how to improve their understanding of these concepts.

Note: Assume the code is written by a novice. 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, sample code refers to the following dataset:

● part-2-data.train.csv