Data Scientist Challenge

Goal

As a data scientist you will need to have strong analytical skills as well as experience building software and strong communication skills. This technical challenge will cover all three of these components using a sample data set.

You will be provided with a data set about students. The data set includes information about the students' age, nationality, gender, as well as information about their studying habits. There is also a binary column expressing whether or not the students passed an important test.

Part 1: Analysis

- Describe the demographic details of people most likely to pass the test
- Describe the efficacy of the two interventions - the test prep course and the Dojo class
- Identify any other interesting trends from the data set and offer some analysis as to their importance or cause.

Part 2: Model Creation

Create a model that can predict whether or not a student will pass the test. Use the provided data set to train your model and test its accuracy. You have free choice of programming language, algorithm, and tools.

Part 3: Reporting

The goal in the final stage is to communicate your findings to less technical management staff. There are two requirements:

- Create visualizations to show the efficacy of your model. A non-data scientist should be able to infer at a glance how well it fits the data.
- Offer ideas for how we might help more people pass the test and create more accurate models based on your findings. Summarize with bullet points and consider adding more visuals.

What to return back to us

- A document of 2 to 4 pages in length describing your findings for Part 1. This can include graphs / visuals but should be mostly text. We want to see your writing ability.
- A software program for generating a model to predict whether or not a student will pass the test. This should include documentation as well as the source code. Ideally we should be able to run your program with a single command and verify your model's accuracy.
- A short technical explanation describing the software tools used as well as the statistical analysis techniques and algorithms you employed and the reason for their selection.
- 1 - 2 page document for Part 3 - visualizations plus some bullet points.