Instructions

OBJECTIVE

You are trying to determine the 7-year survival of prostate cancer patients. A patient survived if they are still alive 7 years after diagnosis. This means that a patient is counted as dead whether or not the death was due to their cancer. You have been given details about the patients and their cancers to help you with your prediction.

DIRECTIONS

Dataset

You have been given a folder labeled ‘participant_files’. In it there are two data sets. The set labeled ‘training_data’ has details of patients, the state of their cancer at time of diagnosis, and some information about the progression of their disease. You will use this data to train any models or create any rules you consider relevant. The second data set, labeled ‘(name)_score’, is the set you will score and submit to finish. There is also a file called ‘Data Dictionary.xlsx’ which contains information about the data in the dataset.

Submission

You have until the next class to submit the scored set. A full submission contains the following:

- The scored data:
    - Replace (name) with your name
    - Do NOT change the starting column names. If you do we cannot score your data
    - Final data set must be in the form ‘.csv’.
    - Populate the column ‘survival_7_years’ with your prediction. This must be a 0 or a 1, not a probability.
- A PPT description of how you explored the data, your prediction process and your final model. This can be in whatever format you like. It’s helpful for us to understand the ‘why’ behind what you did.

Scoring

We will measure the performance of your model/rules by accuracy. In other words, we will take predictions from your scored data and compute the following:

$$\frac{number\ correct\ predictions}{total\ number\ of\ patients\ in\ scoring\ data}$$

This being said, accuracy is not the only aspect of your model we will examine. We care about your approach and knowledge more and weigh your write up and description just as heavily as your final accuracy score.

Tips

- Be sure to examine the scoring data before beginning and reflect on what data might be important to consider when solving the problem.
- Feel free to supplement the data dictionary with outside research on prostate cancer.
- The symptom codes are predictive! We suggest you think about how to use the symptom field in your prediction.
- If you don't know how to build a statistical model, that's fine! Try to think about a set of rules that can help you figure out who might survive and who might die. A good, data-driven, rule-based prediction can show your skill with data as well.
- The PPT is your chance to walk us through, not just the exact steps you took to solve the assessment, but the journey you took to get there. For example, feel free to include details on why you selected a particular method or anything you tried that didn't work.