Data Scientist Technical Skills Assessment
CSP Informatics Center, VA Boston Healthcare System
August, 2019

Background

Lung cancer is the leading cause of cancer-related deaths worldwide. In 2018, the United States is estimated to have had approximately 234,030 new diagnoses of lung cancer, with an estimated 154,050 deaths. As lung cancer is a highly heterogeneous disease with variable survival outcomes, there has been wide interest in developing prognostic models that integrate varying sources of patient information (e.g. demographic, clinical, genomic, imaging) to predict patient survival following diagnosis. An eventual goal for these models is to deploy them as part of decision support systems to facilitate improved decision making in routine clinical practice.

Task

Using the simulated clinical and genomic data provided, build and evaluate a predictive model of one-year survival after diagnosis with NSCLC. This is the core task.

To do so, feel free to use whatever language and/or packages you feel most comfortable with. Our group primarily uses R and Python for such analyses.

Please provide a report detailing your approach and results, as well as whatever code you used. The code should be provided in a way that makes it easy to understand its relationship to the report's results. For example, using knitr or a Jupyter notebook makes this easy.

Please address the following questions in your report:

1. This data, like our real data, may be messy, incomplete, and/or sparsely documented. Please walk us through how you cleaned up this dataset. Do you see any interesting trends?
2. Tell us how you decided which features to use for your model. Are there any new features it might be productive to engineer from the current features?
3. Which algorithm did you use for the prediction and why?
4. How did you assess the predictive model's quality? Summarize your findings.
5. Next steps? What might you do with more time or access to additional data or expertise?

Data Description

The file clinical.csv contains clinical data on each patient. Its columns are as follows:

1. ID: A unique identifier for the patient.

2. Outcome: Whether the patient is alive or dead at the followup time.
3. Survival.months: The followup time in months.
4. Age: The patient's age (in years) at diagnosis.
5. Primary Site: Location of primary tumor.
6. Histology: The tumor histology.
7. Stage: Stage at diagnosis.
8. Grade: Tumor grade (1-4 or missing).
9. Num.Primaries: Number of primary tumors.
10. Tumor.Size: Size of the tumor at diagnosis.
11. T: Tumor Stage.
12. N: Number of metastasis to lymph nodes.
13. M: Number of distant metastases.
14. Radiation: Whether radiation took place (5) or not (0).
15. Num.Mutations: The total number mutations found in the tumor.
16. Num.Mutated.Genes: The total number of genes with mutation.

The file genomics.csv contains information as to which genes were found to have mutation in each patient's tumor sequencing data. Only genes with mutation are listed.

1. ID: A unique identifier for the patient.
2. Gene: The name of the gene.