Exam PA December 2018 Project Statement

From: Peter Stone, Account Manager
To: Exam PA Candidate
Re: Mine Safety Ratings

I hope you can help me out on short notice. Beryl had just started a project for the mine workers union, but a family emergency has come up and she needs to fly across the Pacific today. Moreover, a union leader called me just now and said he would stop by later today to talk about our progress—I thought we had another week!

The union would like to give functioning mines in the United States a simple five-star safety rating to help their members when choosing where to work and negotiating hazard pay. They think they have a good understanding of what factors drive mine safety, but they want us to offer an independent, data-driven analysis so that they can validate and refine their opinions. They pointed us to national mine data in the attached csv file. They asked our analytics firm for the following:

1. Two models, using different approaches, that will predict the rate of injuries per 2000 employee hours for a given mine; and
2. A report that identifies the key factors resulting in higher or lower injury rates.

As I said, the union leader will be here this afternoon and I need a draft of the report in five hours, not a minute more. It should contain the following items, as seen in the report template you will use:

1. **Executive Summary**: Based on your two models, indicate the factors that influence mine safety. This should be written for the union leadership, an intelligent but non-technical audience. All other sections can include technical writing.
2. **Data Exploration and Feature Selection**: Present key elements of the data, including tables and graphs that help the reader understand the important variables in the dataset. Describe how the data was cleaned and prepared, including feature selection, transformations, interactions, and other approaches you considered.
3. **Model Selection and Validation**: Describe the model fitting and validation process used. State the two models you selected and why they are preferable to other choices.
4. **Findings and Recommendations**: Interpret the results of the selected models and discuss additional steps that might improve the analysis.

Also provide the Rmd file that backs up your work and any other relevant files you create so Beryl can pick this up upon her return. (Remember that our system will only accept four types of file: Word (.docx), Excel (.xlsx and .csv), and RStudio (.Rmd). Also, you cannot provide more than ten files, and no file can be larger than 25MB in size.)

I know this is your first analysis project for us and I'm putting you under a lot of pressure…just stay calm and work through the steps methodically while keeping track of time. Spending half the time on analysis and half on writing is generally good guidance. Beryl sent me the note below to pass along, along with an Rmd file she had started. She also provided two R cheat sheets…I hope these are helpful. When I talked to her after she boarded and told her the timing, she said that the two models we should be using are decision trees and GLMs, as stated in her memo.

I'll look for your report in five hours.

Thanks,

Peter

From: Beryl Gold, Actuary
To: Peter Stone
Re: Family Emergency

Peter, I'm sorry, but a family emergency has come up and I need to fly home immediately. I'm sending this e-mail from the airport and won't be in touch for several days. Everything else can wait but I hope you can find someone to work on the mine safety analysis. Here's what the analyst needs to know:

- The data is from the U.S. Mine Safety and Health Administration (MSHA) from 2013 to 2016. Each row represents the safety experience of one mine for one year. There are 20 variables—I have reworked them so most do not need extra interpretation, but a data dictionary is attached just in case. A couple notes:
    - MINE_STATUS: it seems like coal miners use different terms than the others
    - PCT_HRS_###: the proportion, between 0 and 1, of employee hours in each type of mining work, with all these fields adding up to 1…not sure what all these mean…the union can help after our first draft
- I started work in an Rmd file, which I am passing along. It is a good idea to read through the entire file before modifying it with further analysis.
    - There is more data checking and cleaning to do—some records and/or fields should probably be eliminated.
    - For both the decision tree and GLM, a Poisson distribution is used to predict the count variable NUM_INJURIES, with EMP_HRS_TOTAL/2000 (for "per 2000 employee hours") being used as an offset, as is common when predicting a rate with variable exposure. I know this is not the most familiar setup, but the code for doing this is given and explained in the Rmd file, and similar functions also accept offsets.
    - For the decision tree, stick with a decision tree and do NOT try to run a random forest or do any sort of boosting—a random forest froze my computer earlier today!
    - For the GLM, given time constraints, do NOT employ elastic net (or ridge or lasso). It will be easier to explain selection of variables with simpler methods.
    - Validation should be on NUM_INJURIES, but then the requested injury rate can be obtained by dividing out the offset EMP_HRS_TOTAL/2000.

- I included the loglikelihood function for the Poisson distribution in case it is helpful for validation.
- I can never remember how to do partitions with the caret package, so some code for doing this is included. Other methods for validating your models are fully acceptable—how you use a method is more important than what method you use.
- Random number seeds should be used so that I can reproduce the results later—this applies to any random split of data for validation, including those embedded in analytics functions.

They're calling my boarding group—gotta go…

Good luck!!!

Beryl

Data Dictionary for Mine Data

The data is from the U.S. National Institute for Occupational Safety and Health Mining Program.

Variable	Description
YEAR	Calendar year of experience
US_STATE	US state where mine is located
COMMODITY	Class of commodity mined
PRIMARY	Primary commodity mined
SEAM_HEIGHT	Coal seam height in inches (coal mines only)
TYPE_OF_MINE	Type of mine
MINE_STATUS	Status of operation of mine
AVG_EMP_TOTAL	Average number of employees
EMP_HRS_TOTAL	Total number of employee hours
PCT_HRS_UNDERGROUND	Proportion of employee hours in underground operations
PCT_HRS_SURFACE	Proportion of employee hours at surface operations of underground mine
PCT_HRS_STRIP	Proportion of employee hours at strip mine
PCT_HRS_AUGER	Proportion of employee hours in auger mining
PCT_HRS_CULM_BANK	Proportion of employee hours in culm bank operations
PCT_HRS_DREDGE	Proportion of employee hours in dredge operations
PCT_HRS_OTHER_SURFACE	Proportion of employee hours in other surface mining operations
PCT_HRS_SHOP_YARD	Proportion of employee hours in independent shops and yards
PCT_HRS_MILL_PREP	Proportion of employee hours in mills or prep plants
PCT_HRS_OFFICE	Proportion of employee hours in offices
NUM_INJURIES	Total number of accidents reported