Titanic Practice Exam Project Statement

General information for candidates

This assignment has two components. One is a statement of the business problem to be addressed. The other is a list of tasks to be done. Your report will consist of responses to ten specific tasks followed by an executive summary. The audience for the task responses is the examination grading team. Hence, technical language can be used. Each task will be graded individually, so be sure any work that addresses a given task is done within the writeup for that task. The final item in your report is an executive summary written for an audience not familiar with analytics concepts.

This document and the report template indicate the points assigned to each of the components. The total is 100 points. Each task will be graded on the quality of your thought process and conclusions. The executive summary will be graded on the quality of the presentation. Note that a component of the grading of the first ten tasks will also relate to the quality of the exposition, but these sections need not be written as formal reports.

At a minimum you must submit your completed report template and an Rmd file that supports your work. Graders expect that your Rmd code can be run from beginning to end. The code snippets provided should either be commented out or adapted for execution. Ensure that it is clear where in the code each of the tasks is addressed. Your thought process and conclusion for each task should be completely contained within your Word report. The Rmd code should be clear, contain commentary, and support your work.

You may submit other files as needed to support your work. In addition to Word (.docx) and RStudio (.Rmd) files, you may also submit Excel files (.xlsx or .csv). There is a limit of 10 files and no file can be larger than 25MB.

Business Problem

You are an actuary in charge of expanding your company's market for life insurance. Your Chief Actuary wants to launch a niche product targeting people who go on luxury cruises. To better understand the type of customers that are likely to purchase, you have been tasked with conducting market research based on historical records from people who traveled on the Titanic. This product will be called Life Jacket™ and will provide short term life insurance during cruises.

Your Chief Actuary wants to understand what factors contribute to a person surviving in the event of an oceanic catastrophe such as occurred with the Titanic.

Your assistant has already gone through and conducted preliminary work to

- Remove missing values (or impute them where needed)
- Relevel factor variables so that the base level has the most observations
- Split the data into training and test
- Binarize factor variables
- Set up code for building models

Specific Tasks

1. (5 points) Explore the survival rates by `sex`, `pclass`, and `embarked`. Apply a transform to `fare` if necessary.

2. (7 points) Create a new variable called "title" which captures passenger's title such as Mr., Mrs., Dr., and so forth.

3. (2 points) Create a new variable called `family_size` which counts the number of family members that a passenger has on board (including themselves).

4. (10 points) Use Kmeans to detect outliers.

Some of the data from the Titanic was of questionable quality. Some people had missing data filled in after the ship had sunk by contacting relatives, looking at social security records, tax info, and other public records. We don't trust this data and want to exclude it. We only know that this impacted between 5-20 passengers for the variables `age`, `fare`, `pclass`, and `family_size`.

One use of clustering algorithms is for detecting outliers. If a passenger has a very high or low value for certain variables, then they are dissimilar to the other passengers and they will be put into a cluster by themselves, or with others who have similar values.

Put each passenger into a cluster using kmeans. If a cluster has fewer than 5 passengers, then say everyone in that cluster is an outlier. There should be anywhere from 5 – 20 passengers that are classified as outliers. Choose values for the number of centers (k) and the number of starting values (`nstart`) and justify these choices based on the data set and the goal of detecting these outliers.

- Any cluster with fewer than 10 passengers in it should be counted as an "outlier" cluster
- Once you assign clusters, create a new column called `outlier_flag` which identifies these passengers. Then remove the cluster feature (aka the cluster number column).

5. (3 points) Select which variables should be used in modeling.

   Your assistant was not sure which variables to keep, and so they have included everything except `passengerid` because they realized that this has no relevance as to whether or not a person survived.

   Decide on which variables will be useful for predicting survival and delete the others.

6. (7 points) Fit a random forest for variable selection.