Lab 4: Telling a story about survival on the Titanic

Create a short reproducible report answering the questions below, The report should be concise, including only necessary figures, and should be submitted as both HTML and Rmd formats.

This lab is due at 5pm on Tuesday, March 30th. You should submit your assignment, in the form of both a knitted RMarkdown HTML file as well as the .Rmd file, by uploading them to Moodle. While you may collaborate with other students on this assignment, you must write up your own code and answers to the questions. Absolutely no cutting and pasting of any portion of the answers. This assignment, like the others, will be worth 20 points.

Getting started

In this exercise, we are going to look at a dataset describing characteristics (including survival) of passengers on the Titanic, which sunk in the North Atlantic Ocean in 1912.

Let's load the data into our current R session, and look at the variables available in the dataset:

```
library("Hmisc")
library(tidyverse)
getHdata(titanic)
head(titanic)
```

Exercise 1 **Understand your data** (1 pts) For this analysis, we are going to focus primarily on the impact of three predictor variables on survival: economic status (`pclass`), `sex`, and `age`. Examine your data carefully using some univariate plots and/or summaries of the variables to understand what the distributions look like. Choose no more than 2 tables and/or figures to summarize important features of the dataset.

Exercise 2 **Deal with missing data** (2 pts) There are a lot of missing data in the age variable. In real data analysis problems, missing data is a common and pesky problem. It can especially be difficult to deal with when the missingness is not "random", i.e. certain factors (whether they are variables you measure or not) can predict whether the data will be missing or not. For example, do you think that older or younger people might be more or less likely to be missing their age from this dataset? What other variables might determine whether age is observed or not? One way to start to look at this is to create a new factor variable that indicates whether 'age' is missing or not for each observation. Then we can create some simple tables to assess missingness across different groups. Try these types of tabulations out and determine whether you think missing age is predictable based on some of the other data in our dataset. Show one table or figure and summarize the results of your exploration of missing age data in a few sentences.

Exercise 3 **Remove missing data** (1 pt) For now, to make the rest of the lab easier to work through, we are going to ignore the observations that are missing age. This is rarely a great assumption to make in practice, especially if the missingness may be associated with other factors. So when we interpret our results, we will need to remember that our dataset may no longer be representative of the entire population of travelers on the Titanic. Run the following code to remove anyone missing age from our dataset.

```
titanic1 <- filter(titanic, !is.na(age))
```

Exercise 4 **Make some hypotheses** (2 pts) *Before making any multivariate plots*, think about (and if

This assignment is released under a Creative Commons Attribution-ShareAlike 3.0 Unported.

you are working with others, discuss with them) what relationships might exist between these variables of pclass, age, sex and survival. Make a short list, including directions of possible relationships and possible interactions. Write down what those ideas are here before exporing the data fully. It is good practice to "register" your ideas about what you are looking for, to prevent the analysis from turning into a "fishing expedition".

Exercise 5 **Look at your data** (2 pts) Create a few exploratory graphics and/or tables that illustrate the relationships between these variables and survival. (Hint: try adapting graphing code from the logistic regression lecture. Try using facets or colors to highlight important comparisons.)

Exercise 6 **Design and fit a model** (3 pts) Using the plots as your guide, write down a model that you'd like to fit to use to describe how this data predicts the outcome of survival. Then fit that model.

Exercise 7 **Examine your model performance** (2 pts) Now that you've fit a model, calculate, for each individual, an estimated probability of survival, using code similar to that below (Note: adding 'type='response'' ensures that if you fit a logistic regression model, the 'predict' function returns you predicted probabilities instead of predicted log-odds.):

```
titanic1$preds <- predict(fm1, type = "response")
```

Recall that we can define a probability threshold for the predicted outcome of survival, where if the predicted probability is greater than the specified threshold then we say the individal was predicted to have survived. Using the predicted probabilities from your model, compute your model's overall accuracy for a threshold of 0.5. We define accuracy as

$$accuracy = \sum_{i=1}^{N} \frac{1 - |\hat{y}_i - y_i|}{N}$$

where y_i is the observed outcome (either a 0 or a 1) and $\hat{y}_i$ is the predicted outcome (either a zero or a 1, computed based on the 0.5 threshold).

Exercise 8 **Where did your model not do well?** (3 pts)

Compute the accuracy of your model by each group of pclass and sex with the same 0.5 threshold. Print out a tabulation that has six rows, one for each combination of pclass (3 categories) and sex (2 categories) and the 0.5 threshold accuracy for each of those 6 groups. For which groups was your model less accurate? For the group of pclass and sex for which your model was least accurate, write a few sentences about why your model was not as accurate here. Before you write the sentences, look at the rows of your dataset corresponding to the not-as-accurate predictions and/or make some plots of your data to try to get a sense of what might have gone wrong here. (Note that you are not being evaluated here on how accurate your model is, so no need to fit and re-fit your model to make it better and better. Just going through the motions of the performance evaluation is good enough.)

Exercise 9 **ROC Curve** (2 pts) Use the `ROCR` package to plot an ROC curve for your model and calculate the area under the curve. Interpret the area under the curve in the context of the dataset.

Exercise 10 **Summary** (2 pts) Write 3-4 sentences that summarize the key conclusions of this analysis. Interpret the coefficients from your fitted model and describe your model's predictive performance.

Extra credit (5 pts) Fit a different kind of classification model (e.g. an XGBoost model or a Random Forest) to the same data that you did above, and compare it's results to your logistic regression. Was it more or less accurate? Credit earned regardless of it being a better forecaster or not, just getting the results is enough.