Statistical Learning Assignment 2 - Semester 1, 2022

- **INSTRUCTIONS**:

  1. The assignment must be typed (not handwritten). You may use either Microsoft Word (or similar) or R markdown in RStudio for the assignment. Note that the final project will require the use of R markdown. **When answering this question, it should be no longer than 10 A4 pages [single sided] with a font size no smaller than 11 point.**
  2. The assignment due date is listed on the Wattle (Turn-it-in) site. Upload the assignment through Wattle using Turn-it-in. You should submit your assignment in **two different parts.** **Do not submit a zipped file.** **If you are using R markdown:**

     (a) A pdf file [you should convert an HTML file to pdf] of your assignment (this should include important R code to highlight what you have done).

     (b) A '.Rmd' file [an R markdown file].

     If you are using Microsoft Word (or similar):

     (a) A Word file of your assignment (this should include important R code to highlight what you have done).

     (b) A '.R' file of your R code.
  3. In answering the questions, write your answers clearly and succinctly. Use appropriate graphs and tables when you think they help to describe your point or thinking process. Do not just "print" a set of results. Every result should be discussed and have a reason for being presented. **No points will be awarded unless you clearly discuss what you are doing.**
  4. No late assignments will be accepted.
  5. **The assignment you turn in must be your own work. This includes all computer code, writing, and mathematics. Please see the university resources on Academic Integrity** `https://www.anu.edu.au/students/academic-skills/academic-integrity` **for more details.** For this assignment, you do not have to cite material from either of the two textbooks or class slides. Any other material should be appropriately cited.
  6. **Have fun with the exploration!**

---

1. (100 points) We will explore some of the techniques you have learned thus far by examining data from the Kaggle learning competition Spaceship Titanic. As detailed on the Kaggle site:

> Welcome to the year 2912, where your data science skills are needed to solve a cosmic mystery. We've received a transmission from four lightyears away and things aren't looking good.
>
> The Spaceship Titanic was an interstellar passenger liner launched a month ago. With almost 13,000 passengers on board, the vessel set out on its maiden voyage transporting emigrants from our solar system to three newly habitable exoplanets orbiting nearby stars.
>
> While rounding Alpha Centauri en route to its first destination—the torrid 55 Cancri E—the unwary Spaceship Titanic collided with a spacetime anomaly hidden within a dust cloud. Sadly, it met a similar fate as its namesake from 1000 years before. Though the ship stayed intact, almost half of the passengers were transported to an alternate dimension!

We are interested in modeling and predicting which passengers are transported to an alternative dimension. The data are on Wattle, as well as on Kaggle. The variables are:

- **PassengerId** - A unique Id for each passenger. Each Id contains gggg which indicates a group the passenger is travelling with and pp which is their number within the group. People in a group are often family members, but not always.
- **HomePlanet** - The planet the passenger departed from, typically their planet of permanent residence.
- **CryoSleep** - Indicates whether the passenger elected to be put into suspended animation for the duration of the voyage. Passengers in cryosleep are confined to their cabins.
- **Cabin** - The cabin number where the passenger is staying. Takes the form deck/num/side, where side can be either P for Port or S for Starboard. **For this assignment you only need to consider the side of the ship and not the other two pieces of information. See below for code to create the variable.**
- **Destination** - The planet the passenger will be debarking to.
- **Age** - The age of the passenger.
- **VIP** - Whether the passenger has paid for special VIP service during the voyage.
- **RoomService**, **FoodCourt**, **ShoppingMall**, **Spa**, **VRDeck** - Amount the passenger has billed at each of the Spaceship Titanic's many luxury amenities.
- **Name** - The first and last names of the passenger.
- (Y) **Transported** - Whether the passenger was transported to another dimension. This is the response (target variable) you are trying to model and predict.

The data sets from Kaggle (also on Wattle) include:

- ***train.csv*** - training data set
- ***test.csv*** - test data set. Note: this data set does not contain Y, as you are using the predictors to try to predict Y and submit those predictions to Kaggle.
- ***sample_submission.csv*** - this is an example submission file. When preparing your prediction submissions, you should follow the format with your own *True* or *False* values.

- Reading in the training data:

  ```
  train <- read.csv("train.csv", na.strings="", header=TRUE)
  ```

  Note that there are missing data. For this assignment you do not need to consider multiple imputation.

- Here is an example of how to disentangle the information about the side of the ship from the *Cabin* variable. For this assignment you only need to consider the information about the side of the ship.

  ```
  Side <- rep(NA, nrow(train))
  Port <- grep("P", train$Cabin)
  Starb <-   grep("S", train$Cabin)
  Side[Port] <- "P"
  Side[Starb] <- "S"
  Side <- as.factor(Side)
  train <- data.frame(Side, train)
  ```

(a) Create an account on Kaggle. What is your Kaggle public name?

(b) From the **training data set**, remove the first **1,693** cases to create a **validation data set**.

(c) (15 points) Using the training data, conduct an exploratory data analysis. In doing your analysis make sure to identify any unusual points and discuss why they are unusual. For this assignment do not remove any unusual points, only comment on them (if they exist). You may also consider any transformations of the covariates. For the rest of the assignment, if you believe the transformations are appropriate (provide justification - this can simply be a discussion), use those transformations.

(d) Consider a logistic regression model to examine the relationship between whether a passenger was *transported* ($Y = True$) or not ($Y = False$) and their covariate information ($\boldsymbol{x}$).

i. (8 points) Use k-fold cross-validation to determine your "best" model based on the lowest miss-classification rate. Let $k = 10$. While you may use the **glm()** function in R, write your own code for the cross-validation. Provide the estimated cross-validation miss-classification rate $CV_{(k)}$ and standard error. Present this as $[CV_{(k)}, CV_{(k)} - SE(CV_{(k)}), CV_{(k)} + SE(CV_{(k)})]$. Use a forward selection search process for your model search. You do not need to consider any interactions.

ii. (6 points) Using the **validation data set**, provide the confusion matrix based on the "best" model from 1(d)i. Compute the overall miss-classification rate, as well as the false-positive and false negative rates. Does changing the "threshold" help with these?

iii. (4 points) From your "best" model in 1(d)i, provide provide 95% confidence intervals for the regression coefficients.

iv. (6 points) From your "best" model in 1(d)i, without using the **boot()** function, write your own R function which takes the **training data set** and the number of bootstrap samples as inputs, and returns bootstrap standard errors and 95% confidence intervals for the regression coefficients from your "best" model. Compare these results to the estimated asymptotic standard errors and confidence intervals produced from using **glm()**. Make sure to clearly outline your algorithm.

v. (6 points) From your "best" model in 1(d)i, using the bootstrap approach (using any R functions you believe will help), provide 95% confidence intervals for the predicted probability of being transported for the first 5 individuals in the **test data set** and first 5 individuals in the **validation data set**.

vi. (4 points) From your "best" model in 1(d)i, using the **test data set** submit your predictions to Kaggle. What was the miss-classification rate? What was your ranking?

(e) (12 points) Repeat 1(d)i, 1(d)ii, 1(d)v, and 1(d)vi using linear discriminant analysis instead of logistic regression. You may use any R functions that you believe will help. This means you can write your own functions or use those already in R. Discuss your modelling assumptions. If the assumptions are violated, then what is or is not a concern?

(f) (12 points) Repeat 1(d)i, 1(d)ii, 1(d)v, and 1(d)vi using quadratic discriminant analysis instead of logistic regression. You may use any R functions that you believe will help. This means you can write your own functions or use those already in R. Discuss your modelling assumptions. If the assumptions are violated, then what is or is not a concern?

(g) (12 points) Repeat 1(d)i, 1(d)ii, 1(d)v, and 1(d)vi using nearest-neighbour analysis instead of logistic regression. You may use any R functions that you believe will help. This means you can write your own functions or use those already in R. Discuss your modelling assumptions. If the assumptions are violated, then what is or is not a concern?

(h) (15 points) Compare and contrast the four different models. Based on the four different models, provide a statistically relevant discussion of the scientific question of which passengers are being transported. The captain needs to try to figure out a solution!