Homework #5

Kamryn Parker

1. Matt wants to know if exercise has an impact on resting heart rate – he thinks it might. So he collects data on resting heart rate (a continuous variable) from three groups of volunteer college students. One group regularly runs, one group regularly lifts weights, and one group doesn’t really exercise at all.
   1. **What is Matt’s hypothesis?**

Matt’s hypothesis is that exercise does have an impact on a person’s resting heart rate

* 1. **What is the null hypothesis?**

The null hypothesis is that exercise does not have an effect on a person’s resting heart rate.

* 1. **What statistical test can Matt use to evaluate his hypothesis?**

I think the best statistical test would be an ANOVA test.

* 1. **In 3-5 sentences, elaborate on why the test you picked is the right one:**

The ANOVA test is the best statistical test for more than two data sets. It will also tell the difference in means between groups. ANOVA can also reduce the variability between data sets.

* 1. **Are there any disadvantages/challenges with the test you picked?**

The disadvantages to ANOVA are that if the means of the data sets differ it can’t exactly tell us which means differ so, it can be hard to quantify results sometimes.

1. Laura is working for a company that produces bulletproof vests. Based on a large collection of data, including continuous and categorical variables, Laura is trying to predict whether a vest will “fail completely”, “succeed with significant damage” or “succeed with minimal damage” in any given situation.
   1. **Is Laura attempting a classification or regression task?**

Laura is attempting a classification task because she has three different categories she is attempting to predict.

* 1. **What kind of algorithm might Laura use to accomplish this task?**

The algorithm I would probably use in this situation is a decision tree classifier or a random forest model.

* 1. **In 3-5 sentences, elaborate on why the model you picked is a good one:**

I would use either of these models because they can handle both categorical and continuous variables with ease. A decision tree is also really good because it can take on more than just binary classification so, it can predict many different categories of your data. If you experiment with both decision trees and random forest classifiers, it can be a great way to do feature engineering by experimenting off each model to see what fits best for the data.

* 1. **Are there any disadvantages to the model you picked?**

Decision trees can become extensively large if you do not prune it or have restricting hyperparameter. Also, any small change in the data can cause a massive change in the structure of the model which leads to instability.

* 1. **Does your model require any pre-processing/cleaning of the data?**

Decision trees are very intuitive so there isn’t any need to scale data beforehand. Also missing data does not affect the tree either so there isn’t necessarily a need to drop data (only a preference).

1. Michelle is working at a manufacturing plant that wants to predict **how long** a type of production machine will run before breaking down, in order to better design maintenance schedules. Michelle has a large amount of mostly continuous data from machine sensors to help her build this model.
   1. **Is Michelle attempting a classification or regression task?**

I believe Michelle is attempting to use a classification model.

* 1. **What kind of algorithm might Michelle use to accomplish this task?**

I believe the best algorithm Michelle could use is Time Series Linear Model

* 1. **In 3-5 sentences, elaborate on why the model you picked is a good one:**

I believe this would be the best model because it is great for time series analysis. The data presented is trying to predict the lifespan of different machinery so, using time series analysis of already broken-down machines would be a great way to predict future breakdowns. The tslm() function in R also has a great breakdown of what features are significant or not in the data.

* 1. **Are there any disadvantages to the model you picked?**

Time series analysis and be pretty difficult to understand and pick up the patterns for. Time series analysis can also require a lot of set up before analyzing which can be very time consuming and not always realistic.

* 1. **Does your model require any pre-processing/cleaning of the data?**

Time series analysis requires a lot of important components. First, it should only include continuous data so, you may need to cut down on important pieces of data. It is also extremely important to make sure your time series is accurate so being able to format it first is key.