William Randall
none

Homework 1

Q1

A. False
    a. There can be a pattern in data but it might be because of an external factor and therefore the correlation doesn't equal causation.
B. False
    a. They can use C++ for better performance or any other language!
C. False
    a. Data Scientists usually use already created models and they spend more time cleaning and getting data.
D. True
    a. The data from the past can include biases from the time it was collected. If this was not the case then we would be able to predict the election way better.
E. False
    a. Some data scientists must collect data.
F. False
    a. Linear regression is a supervised model because it doesn't generate new data as a GAN would. We also feed in labeled data to a linear regression model.
G. True
    a. Linear regression can confidently make predictions inside the range of data on which it was trained because this ensures that no outliers will throw off the linear regression model which might be outside of the range it was trained on.

Q2

This isn't a good idea because this is essentially using a model on an untrained range of data. Where in this case the model was trained on a dataset of people from the UK and USA, and this model might work very poorly unless trained on people who are from Brazil. People in the UK and USA share many more similar traits than when either country is compared to Brazil.

Q3

The issue with this would be that people who talk on Twitter and in the comment sections of News articles usually are driven to complain online. So, only looking at these two places will most likely result in a much more negative average sentiment. A better way would be to poll students randomly from many different buildings and at many different times of the day. This suggested way of using Twitter and other comments is not a random sample.

Q4

A. Student's Year, Student's GPA, Student's Major, Student's Current Completed Unit Count, Professor Teaching the Course, Professor Rating, Percentage of A's given in past years, Pre Recs taken, If the course is offered that quarter
B. Labels
    a. Student's Year
        i. Numerical 1-5
    b. Student's GPA
        i. Numerical 1-5 in 0.5 increments
    c. Student's Major
        i. One hot encoding
    d. Student's Current Completed Unit Count
        i. Numerical
    e. Professor Teaching the Course
        i. One Hot Encoding
    f. Professor Rating
        i. Numerical Range 0 - 100
    g. Percentage of A's given in past years
        i. Numerical Range 0 - 1 in 0.05 increments
    h. Pre Recs taken
        i. Boolean T or F
    i. If the course is offered that quarter
        i. Boolean T or F
C. I would poll students at random through email and through posting a reward for completing my poll around campus, and I would also send out an email to the students for more involvement. I would also walk around campus, Westwood, and the dorms to randomly ask people to answer the questions. And get information from the UCLA registrar.

Q5

A

A. Remove the row that contains a null value.
    a. Advantage
        i. Computationally easy.
    b. Disadvantage
        i. Lose every row with a null value which could be a huge amount of data.
    c. Advisable data set
        i. Any large data set that has very few nulls. Columns that are not good predictors.
    d. Inadvisable data set
        i. A small data set with a lot of null values or any data set with a lot of nulls.
B. Replace null with 0
    a. Advantage
        i. Computationally easy
    b. Disadvantage
        i. Can skew data downward
    c. Advisable data set
        i. Any data set with a feature that has very low numbers and a null. Or you can do this for a feature that has a negative to positive spread so a 0 would do less damage to that feature. Columns that are not good predictors.
    d. Inadvisable data set
        i. Small data set with nulls as well as large numbers. Using this method would skew the mean down towards 0.
C. Replace null with average
    a. Advantage
        i. It will not skew your mean.
    b. Disadvantage
        i. It will pull the standard deviation towards your mean, which will make it seem like your data is more tightly centered around the mean when it might not be. It is also heavily affected by outliers.
    c. Advisable data set
        i. Large data sets with few nulls. Columns that are not good predictors.
    d. Inadvisable data set
        i. Small data sets with a lot of nulls.
D. Replace null with the median
    a. Advantage
        i. It will not skew your data set if there are outliers.
    b. Disadvantage

        i. It is still introducing data that might be very wrong. It will continue to pull your data towards your median. It might also skew your mean (this might now be an issue).
    c. Advisable data set
        i. Large Data sets. Columns that are not good predictors.
    d. Inadvisable data set
        i. Small Data set. Data sets with large amounts of nulls.

B

Replace null with median would work the best with outliers because it will not be affected by outliers as much as the other options like replacing with the mean.

Q6

Info:
3 red
2 yellow
1 blue
2 drawn at random without replacement

X = number of red marbles drawn
Y = number of blue marbles drawn

A. P(X=1) = 3/6 * 3/5 + 3/6 * 3/5 = 0.6
B. P(X=0,Y=1)
   = P(draw a blue, draw a non red) + P(draw a non blue, draw a red)
   = 1/6 * 2/5 + 2/6 * 1/5
   = 2/15
    a. Draw 0 red & draw 1 blue
C. P(X = 1 | Y = 1)
   = P(given a blue was drawn, prob of drawing a red)
   = 3/5

Q7

A. I would not. Walking pace is a numerical value that would be better represented in a float of MPH.
B. I would one-hot encode class name because there are 4 options.
C. I would not one-hot encode city in the US because that would add too many dimensions to the model.
D. No, I would use a single boolean, not a one-hot encoding for having diabetes and not having diabetes.
E. I would one hot encode this if there were less than 100 brands represented in the data set. If there were more then it would get to be adding too much dimensionality.
F. I would not one hot encode this specific interaction term because it would be too many dimensions assuming that the number of hours of sleep is not discrete and is a continuous float of a value.
    a. But I would separate them and one hot encode test difficulty.