True / False (total 5 pts)

1. **FALSE** Measuring accuracy on the Training Set is the best estimate for how the model will do in deployment. (need a holdout/Test set to estimate performance in the wild)
2. **TRUE** A good decision tree split is one that significantly reduces entropy AND increases information gain
3. **TRUE** Adding more complexity will typically improve performance as measured on the training set
4. **(T/F)** If you have missing data in your data set, you should ~~typically~~ always remove the rows that contain the missing data

**NOTE: FOR SECTION 1, I WILL ACCEPT TRUE OR FALSE, THERE WAS CONFUSION ON HOW THE WORD ‘TYPICALLY’ SHOULD BE INTERPRETED. FOR SECTION 2, I HAD THEM CHANGE “TYPICALLY” TO “ALWAYS”, IN WHICH CASE THE ONLY CORRECT ANSWER IS FALSE.**

1. **TRUE** You can build an unsupervised model without any specific target variable in the training data (this is the definition of unsupervised)

Identify the following tasks as classification (C), regression (R), or unsupervised (U): (3)

1. **Unsupervised – clustering is an unsupervised problem** I run a business with a customer call care center and I get a few hundred calls a day. I have data on these calls and I want to bucket them into different types to understand and track why my customers call.
2. **Regression – there is a clear numeric target** I have data on my customers’ spend going back many years. When I get a new customer I want to be able to predict their lifetime value, that is, how much money they will spend with us during their lifetime.
3. **Classification – binary target (pregnant or not)** The model used in the Target case study

**Multiple Choice – Choose one Answer: (8 pts)**

1. The more complex a model is:
2. the more interpretable it is
3. the better it will perform on the holdout set
4. **the more likely it is to overfit**
5. the easier it is to train
6. Why might we split our data into training and test sets?
7. to increase complexity
8. to measure complexity
9. **to get the best possible measure of the model's performance**
10. to improve the model's performance
11. How can we determine the most important feature in a decision tree model?
12. the one that results in the highest overall entropy
13. **the one that accounts for the most overall information gain**
14. the one that appears in most splits
15. the one that is in the most leaf nodes
16. Which of the following is NOT an example of complexity in a data science model
17. number of nodes in a tree
18. degree of a polynomial
19. **number of observations**
20. depth of a tree
21. Which of the following is NOT a feature of cross-validation?
22. It provides an estimate of generalization performance
23. **It is faster to compute than using a single holdout set**
24. Each data point is used for both training and testing
25. It provides an average across multiple estimates of performance
26. What makes a regression problem different from a classification problem?:
27. **it has a numeric target**
28. it has numeric features
29. it is unsupervised
30. it is a prediction task
31. Which of the following is an example of overfitting?

A) **fitting a regression with more than 100 features (will accept A or C)**

B) putting 5 or more symbols on a single scatterplot

C) **fitting a decision tree until every leaf node is pure (will accept A or C)**

D) using decision trees on a data set with less than 100 observations

1. Lets say you have collected IQ data on NYU students.  The two features you have for each student are the school they are in (Stern, Courant, etc) and their IQ (a numeric value ranging from 50-150).    What would be the most effective way to visually show the differences in IQ distribution across the different schools?
2. **side-by-side boxplots**
3. scatterplot
4. stacked bar plot
5. correlation plot

**Short Answer Questions: (8 pts)**

1. Explain a scenario when a data science model has an accuracy of 99% but that is not considered good performance.  (2 points)

**Best answer: a scenario where the base rate (total number of positives) is very low, (like ad click prediction) because you can get 99% accuracy by just predicting everyone will not click.**

**Other answers can be possible here – such as**

* **a model with bias (such as the amazon hiring case discussed in class)**
* **a model where even 1% error results in some kind of loss to the business**
* **if the 99% is calculated on the training set it would not be necessarily a good model.**
* **(these would be worthy of full credit)**

**Note – be gentle here – only take off 1 point if they say anything reasonable …**

Reach out on email thread if not sure how to grade.

1. Here is a piece of python code we might use to fit a predictive model, using a training and a test set.  This code splits your original data set and creates four subsets. (3 points)

from sklearn.model\_selection import train\_test\_split

(X\_train, X\_test, Y\_train, Y\_test) = train\_test\_split(features, labels, test\_size = .3)

1. First you need to fit the model.  Which two subsets will you use in the call to ".fit"?

**X\_train, Y\_train**

1. Then we need to use the model to make predictions.  Which subset will you use in the ".predict" call?

**X\_test (if they say X\_test AND Y\_test, this is wrong)**

1. What two values will you use in order to calculate the error (or accuracy) of the model?

There was some confusion on this question : the correct answer is 1) Y\_test and 2) the predicted vector from part B (they may call this y\_pred). Those are the two values you need to calculate error. I *will* also accept something like “number of correct predictions, and number of total observations” – technically that is correct.

Some answer “RMSE”, or some other measure of error. Those are wrong – they ARE measures of error, but you need the prediction vector (or the number of correct cases) in order to calculate it.

1. The plot below represents a key concept in data science.  Provide a sentence about each of the following three aspects of this plot: ( 3 points):

A diagram of a training and test results

Description automatically generated

1. the relationship between complexity and the training set

Must say something about the training set decreases with increasing complexity

1. the relationship between complexity and test set

Something abt how test set error decreases with increasing complexity to a point but will then increase due to overfitting.

Many here say something like “test set error increases with more complexity” – this is only partially right, so I have been giving ½ point for this. Im looking for “with more complexity, test set improves to a point and then gets worse due to overfitting”

1. the meaning of the vertical dashed line

that is the “Sweet spot” or optimal spot where the complexity balances performance and overfitting.