1. Select the option that is the most INACCURATE regarding the definition of Machine Learning:
   1. Machine Learning is automated and requires no programming
2. This is the type of Machine Learning that uses both data with labeled outcomes and data without labeled outcomes:
   1. Semi-Supervised Machine Learning
3. Predicting total revenue, number of customers, and percentage of returning customers are examples of:
   1. regression
4. Predicting payment default, whether a transaction is fraudulent, and whether a customer will be part of the top 5% spenders on a given year, are examples of:
   1. Classification
5. Which statement about evaluating a Machine Learning model is the most accurate?
   1. Model estimation involves choosing parameters that minimize the cost function.
6. (True/False) The unadjusted value from estimating a linear regression model will almost always increase if more features are added.
   1. True
7. (True/False) The Total Sum of Squares (TSS) can be used to select the best-fitting regression model.
   1. False
8. (True/False) The Sum of Squared Errors (SSE) can be used to select the best-fitting regression model.
   1. True

# End of Module Quiz

1. You can use supervised machine learning for all of the following examples, EXCEPT:
   1. Segment customers by their demographics.
2. The autocorrect on your phone is an example of:
   1. Supervised learning
3. This is the type of Machine Learning that uses both data with labeled outcomes and data without labeled outcomes:
   1. Semi-Supervised Machine Learning
4. This option describes a way of turning a regression problem into a classification problem:
   1. Create a new variable that flags 1 for above a certain value and 0 otherwise
5. This is the syntax you need to predict new data after you have trained a linear regression called *LR*:
   1. LR.predict(X\_test)
6. All of these options are useful error measures to compare regressions:
   1. ROC index
7. (True/False) It is less concerning to treat a Machine Learning model as a black box for prediction purposes, compared to interpretation purposes:
   1. True
8. Another common term for the testing split is:
   1. Validation split
9. Complete the following sentence: The training data is used to fit the model, while the test data is used to:
   1. measure error and performance of the model
10. Select the option that has the syntax to obtain the data splits you will need to train a model having a test split that is a third the size of your available data.
    1. X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.33)
11. Which statement about K-fold cross-validation below is TRUE?
    1. Each of the *k* subsamples in K-fold cross-validation is used as a test set.
12. (True/False) For a dataset with *M* observations and *N* features, Leave-one-out cross-validation is equivalent to k-fold cross-validation with *k* =*M-1* .
    1. True
13. If a low-complexity model is underfitting during estimation, which of the following is MOST LIKELY true (holding the model constant)?
    1. K-fold cross-validation will still lead to underfitting, for any *k*.
14. Which of the following statements about a high-complexity model in a linear regression setting is TRUE?
    1. A high variance of parameter estimates across cross-validation subsamples indicates likely overfitting.

Polynomial

1. What is the main goal of adding polynomial features to a linear regression?
   1. Capture the relation of the outcome with features of higher order.
2. What is the most common sklearn methods to add polynomial features to your data?
   1. polyFeat.fit and polyFeat.transform
3. The main purpose of splitting your data into a training and test sets is:
   1. To avoid overfitting
4. (True/False) For a dataset with *M* observations and *N* features, Stratified cross-validation is equivalent to k-fold cross-validation, where *k* =*N-1* .
   1. False
5. (True/False) A linear regression model is being tested by cross-validation. Relative to K-fold cross-validation, stratified cross-validation (with the same *k* ) will likely increase the variance of estimated parameters.
   1. False
6. In K-fold cross-validation, how will increasing *k* affect the variance (across subsamples) of estimated model parameters?
   1. Increasing *k* will usually increase the variance of estimated parameters.
7. Which of the following statements about model complexity is TRUE?
   1. Higher model complexity leads to a higher chance of overfitting.
8. Which of the following statements about model errors is TRUE?
   1. Underfitting is characterized by higher errors in both training and test samples.
9. Which of the following statements about regularization is TRUE?
   1. Regularization decreases the likelihood of overfitting relative to training data.
10. BOTH Ridge regression and Lasso regression
    1. add a term to the loss function proportional to a regularization parameter.
11. Compared with Lasso regression (assuming similar implementation), Ridge regression is:
    1. less likely to set feature coefficients to zero.
12. (True/False) The variance of a model is determined by the degree of irreducible error.
    1. False
13. (True/False) As more variables are added to a model, both its complexity and its variance generally increase.
    1. True
14. (True/False) Model adjustments that decrease bias also decrease variance, leading to a bias-variance tradeoff.
    1. False
15. Which of the following statements about scaling features prior to regularization is TRUE?
    1. The larger a feature’s scale, the more likely its estimated impact will be influenced by regularization.
16. Which of the following statements about model complexity is TRUE?
    1. Higher model complexity leads to a higher chance of overfitting.
17. (True/False) A model with high variance is characterized by sensitivity to small changes in input data.
    1. True
18. Which of the following statements about Elastic Net regression is TRUE?
    1. Elastic Net combines L1 and L2 regularization.