General Assembly: Data Science Review

When writing Python code, why would you write a function? (30 sec)
- to avoid repeating yourself, and to create reusable code

What is the difference between a bar plot and a histogram? (30 sec)
- histogram shows the distribution of a numerical variable
- bar plot shows a numerical comparison across different categories

What is the difference between supervised and unsupervised learning? (60 sec)
- supervised learning has a response you are trying to predict, and goal is generalization
- unsupervised learning has no response, and goal is representation

How could you convert any regression problem into a classification problem? (60 sec)
- cut the range of possible response values into "bins" and treat those bins as ordered categories

In machine learning, what concepts are commonly represented by the following letters? (30 sec)
- n: number of observations
- p: number of features
- X: matrix of features
- y: vector of responses

How does KNN work for classification? (60 sec)
1. pick value for K
2. tally response of K nearest neighbors
3. use most common response as predicted class

What is the bias-variance trade-off, and why should we care about it? (120 sec)
- increasing model complexity increases variance but decreases bias, whereas decreasing model complexity decreases variance but increases bias
- total generalization error of a model is determined by both bias and variance, thus optimum model complexity requires balancing the two

What's wrong with training and testing on the same data? (60 sec)
- you can create an arbitrarily complex model that will perform well on the training data but won't generalize to out-of-sample data (known as "overfitting")

What are two procedures we used to estimate out-of-sample error? What are the strengths of each? (90 sec)
- train/test split is simple to code and fast to run

- cross-validation is more accurate for estimating out-of-sample error

What are some reasons that linear regression is popular? (60 sec)
- runs fast, easy to use, highly interpretable, well-understood

When using a classification model, what is the relationship between predicted probabilities and class predictions? (90 sec)
- predicted probabilities are the probabilities that each observation belongs to a given class
- they can be mapped to class predictions by selecting the class with the highest probability

Why is a confusion matrix useful for measuring the performance of a classifier? (60 sec)
- gives a much more nuanced picture of classification performance than classification accuracy
- allows you to calculate sensitivity, specificity, etc.

What is null accuracy, and why is it useful to know the null accuracy of your classifier? (60 sec)
- accuracy that could be achieved by always predicting the most frequent class
- gives you a baseline to compare your model against

What makes AUC better than accuracy as a single number measure of classifier performance? (90 sec)
- AUC is useful even when your classes are highly unbalanced
- accuracy requires setting a classification threshold, whereas AUC does not

What are some general strategies for handling missing values in your data? (90 sec)
- drop rows containing missing values, impute missing values, treat missing values as another category (for categorical features)

What are the different ways we encode categorical features for use with a model? (90 sec)
- 2 categories: encode as 0/1
- more than 2 unordered categories: create dummy variables and drop the baseline level
- more than 2 ordered categories: encode as a single numbered variable

How do you represent text documents as data for use with a machine learning model? (90 sec)
- create a document term matrix, in which each row represents a document and each column represents a word
- for each document, count the number of times that each word appears, or use a TF-IDF representation

Why is Naive Bayes popular for spam classification? (60 sec)
- text generates lots of features, and Naive Bayes handles lots of features well

- Naive Bayes is fast, and thus is appropriate for real-time applications

What are some advantages and disadvantages of decision trees, compared to other models? (90 sec)
- advantages: interpretable, can be displayed graphically or specified as a series of rules, non-parametric, automatically learns feature interactions
- disadvantages: high variance, low predictive accuracy

What is feature engineering, and what is the goal of feature engineering? (90 sec)
- creating features that don't natively exist in the dataset
- goal is to add new features that contain the "signal" from the data (with respect to the response value), rather than the "noise"

What are two conditions that must be met for ensembling of models to be useful? (30 sec)
- models should be independent and more accurate than the null model

How do Random Forests work? (120 sec)
- grow a lot of decision trees using bootstrapped training sets, and grow them deep
- when building each tree, each time a split is considered, only consider a random subset of predictors
- all trees make predictions, and those predictions are averaged

How does K-means clustering work? (90 sec)
1. choose k initial centroids
2. assign each point to the nearest centroid
3. recalculate centroids
4. repeat steps 2 and 3 until stopping criteria are met

How does regularization reduce overfitting? (60 sec)
- it constrains the size of coefficients, which tends to reduce variance more than increasing bias