Module 2 Quiz

1. What are the three primary considerations when selecting an algorithm to use in modeling (select three)?

   - Performance
   - Interpretability
   - Validation & testing approach
   - Computational efficiency

2. What does the term "variance" in modelling refer to?
   1. The sensitivity of the model to small fluctuations in the training data
   2. The error introduced by modelling a real life problem using an over-simplified model
   3. The range of the output values of the model
   4. The total error of the model predictions

3. "Underfitting" refers to the situation in modelling when:
   1. The model is overly complex and unable to generalize well to make predictions on new data
   2. The model consistently generates predictions which are lower than the target values
   3. The model is too simple to fully capture the patterns in the data
   4. The model was not trained on enough data

4. Why do we split our data into a training set and test set, and then hold back the test set while training our model?
   1. We can then use the test set to compare versions of our model and select the best version as our final model
   2. We use the test set to calculate the performance of models with different sets of features, to help us with feature selection
   3. We select our model using the training set and then add our test set in, re-train our model, and calculate the model's performance across our full dataset
   4. We use the test set to evaluate performance of our final model as an unbiased indicator of its ability to generate quality predictions on new data

5. When we split our data to create a test set, how much of our data do we generally use for training and how much for testing?

   1. Typically 70-90% for training and 10-30% for testing
   2. Typically 50% for training and 50% for testing
   3. Typically 70-90% for testing and 10-30% for training
   4. Typically 95% for training and 5% for testing

6. Why would we use cross-validation instead of a fixed validation set (select all that apply)?

- It is computationally less expensive
- It maximizes the data available for training the model, which is particularly important for smaller datasets
- It enables us to skip the use of a test set for evaluating model performance
- It provides a better evaluation of how well the model can generalize to new data because the validation performance is not biased by the choice of datapoints to use in a fixed validation set

7. If we use standard K-folds cross-validation with k=5, how many times do we use each observation in our data as part of a validation fold?

1. Once
2. Five times
3. Four times
4. It varies by observation

8. What does the term "data leakage" mean?

1. We use some of our training data in the test set to evaluate model performance
2. We lose valuable data by removing rows with missing values
3. We reduce the data available for training our model by carving a portion out to use as a test set
4. We use our test set data at some point during the model building process and it influences the development of our model

9. Why is data leakage a dangerous thing in the modelling process?

1. It can invalidate the estimated performance of our model based on the test set, and cause our performance estimate to be overly optimistic relative to the model's ability to generate predictions for new data
2. It can significantly reduce the amount of data available for model training
3. It can artificially reduce our model's performance on the test set, causing us to believe our model is not as good as it actually is in generating predictions for new data
4. It commonly leads to underfitting on the training and test data, resulting in models that are too simple

10. If we are using cross-validation to compare multiple models and select the best one, how do we compare the models?

1. For each model, we run cross-validation and calculate the error on the validation fold of each iteration. We then take the minimum error on a validation fold as representative of the cross-validation error of the model. We compare each model's minimum error and choose the best one
2. For each model, we run cross-validation and calculate the error on the training folds for each iteration. We then take the average error across all training folds as representative of the model's performance, compare the average error of each model and select the model with the minimum average error
3. For each model, we run cross-validation and calculate the error on the validation fold of each iteration. We then take the average error across iterations as representative

of the model's performance, compare the average error of each model and select the model with the minimum average error

4. We run cross-validation and during each iteration of the cross-validation we try a different model, calculate the validation fold error, and then choose the model with the lowest error on the validation fold.