UNIT 5 ASSIGNMENT

Choosing Your Model

## Instructions

The questions below will prepare you for future interviews as they relate to concepts discussed throughout the week. You’ve practiced these concepts in the coding activities and the exercises   
as well as the coding portion of the assignment. Now let’s formulate your programming into well-reasoned responses.

Except as indicated, use this document to record all your assignment work and responses to any questions. At a minimum, you will need to turn in a digital copy of this document to your facilitator   
as part of your assignment completion. You may also have additional supporting documents that   
you will need to submit. Your facilitator will provide feedback to help you work through your findings.

**Note:** Though your work will only be seen by those grading the course and will not be used or   
shared outside the course, you should take care to obscure any information you feel might be of   
a sensitive or confidential nature.

*Begin your assignment by completing the questions below. Directions to submit your work can be found on the Unit 5 Assignment page online. Information about the grading rubric is available on any of the unit assignment pages online. Do not hesitate to contact your facilitator if you have any questions about the assignment.*

Week 5 Written Portion

# Choosing Your Model

Answer the questions below about selecting the correct models and approaches to solve your machine learning problems.

## Questions:

1. What is model selection and why is performing model selection important?

|  |
| --- |
| Model selection is like a type of search algorithm where we are searching for the supervised learning model that optimizes our out-of-sample loss. This is our final model, and we perform model selection because it is important that our model has good accuracy when generalizing new data and acceptable performance. |

1. What is out-of-sample validation and why is this key in helping us choose the best-performing model?

|  |
| --- |
| Out-of-sample validation is a way to assess the ability of a model to generalize by splitting up our data into subsets. Out-of-sample validation is the key in helping us choose the best performing model because in this moment, we can continue to tweak the model until the validation error improves, and it is difficult to tweak the ML algorithm once we use the test data set because the model is already overfitted. |

1. What is cross-validation and what is the benefit of performing cross-validation?

|  |
| --- |
| (K-fold) Cross-validation is a method of splitting the data into equally sized subsets where k is the number of subsets. The benefit of performing cross-validation is that we have many k chances to tweak our model (validate) so that the model can make more accurate predictions. |

1. What is the difference between feature engineering and feature selection? What are the benefits of feature selection?

|  |
| --- |
| Feature selection is the process of selecting the best features that contribute to a good, performing model. Feature engineering, on the other hand, is the process of transforming raw data into features and includes other tasks like handling outliers and missing data. The benefits of feature selection are that doing so can reduce overfitting, have an easier-to-understand model, and faster runtime, development, and lower maintenance costs. |

1. What are the differences among the classification evaluation metrics accuracy, precision, and recall?

|  |
| --- |
| The accuracy is the percentage of all examples that match the correct value (classified correctly) while precision only focuses on the percentage of true positives, out of the false and true positive predictions, that were classified correctly. Recall only focuses on the percentage of true positives that were correctly classified as positive. Additionally differences are that the accuracy metric is commonly used for multi-class problems while precision is favored for binary classification when false positives are worse than false negatives and recall is favored for binary classification, but only when false negatives are worse than false positives. |

*To submit this assignment, please refer to the instructions in the course*.