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 defined as a process to optimize the out-of-sample loss. It aims to search this base of model design options to find the perfect balance between overfitting and underfitting. So, it chooses 1 final model from among many candidates that results in the optimal model for a given data set and machine learning problem. Performing model selection is important because the goal of machine learning is to have our model generalize to new, previously unseen data. Model selection allows us to choose the model that best generalizes. |

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 defined as when we evaluate our models so we use data that wasn’t used in model training. It’s a technique used for evaluating a model’s ability to generalize to new data. This is key in helping us choose the best-performing model because it can properly evaluate how our model generalizes to new, previously unseen data. It splits our data into 3 data sets: training, validation, and test. Each with their own role, training data set for model fitting, validation for performing model selection, and test for the final evaluation. |

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

|  |
| --- |
| Cross validation is defined as a method used to measure/estimate the performance and accuracy of a model. The benefit of performing cross validation is that it’s used to prevent overfitting of a model, which is something we should avoid in machine learning. |

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

|  |
| --- |
| The difference between feature engineering and feature selection is that in feature selection we choose specific number of features that are considered relevant data in the training of our model. The benefits of feature selection is that this will help improve our model’s performance. In contrast, feature engineering uses any amount of data to build more interpretable models. |

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

|  |
| --- |
| The differences among the classification evaluation metrics are, in accuracy we utilize kappa to adjust our model accuracy by the formula. Moreover, to measure the model accuracy we do: true positive + true negative divided by the total number of predictions. In precision we observe how well the model can predict a specific outcome. It helps us better understand the model’s precision tradeoff for various values and determine optimal threshold to use, which we can use the tool called AUC-ROC. Its formula is the number of correct predictions divided by the total number of predictions. It’s more of a measure of quality. In recall it observes the number of times time model correctly predicted a specific outcome. And this focuses more on quantity. Recall’s formula is the number of correct predictions divided by the total number of actual purchases. |

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