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 selecting an optimal machine learning model for a problem. It involves evaluating and comparing different models based on their performance metrics with the training set and test set and selecting the one that exhibits the best generalization with complexity and accuracy.  You want your model to be the best for performance and succeed for previously unseen data. You also want it to be efficient because the algorithms can have different complexities, memory requirements, and training times. Additionally, you want to avoid overfitting and increase the model’s interpretability. |
| --- |

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 an evaluation metric on examples that were not part of model training. It is a way to properly evaluate how well a model generalizes to new, unseen data. It sections the data into training, validation, and test sets. Training set: used to actually fit the model to the data. Validation set: used to evaluate model candidates for model selection; continue tweaking this set until the model is satisfactorily accurate. Test set: used for estimating the generalization performance of the best selected model.  This helps with bias because the model has never seen the data. You can also measure different models equally because they’re all performing on the same validation data. It helps prevent overfitting and helps hyperparameter tuning. Also, it provides insight on how the model will evaluate and run in real-world deployment. |
| --- |

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

| Cross-validation assesses the performance of a model by splitting the available data into multiple subsets or "folds". It then iterates through and trains the model on different combinations of these folds. The most common form of cross-validation is k-fold cross-validation, where the data is divided into k equally sized folds.  **K-Fold Cross-Validation**: separate remaining data into K Equal Size partitions (folds), iterate through each fold to create a training and validation split, each fold computes a loss, average of validation becomes main input for comparisons.  It provides a more reliable estimate of the model's performance by averaging the performance across multiple iterations. It also reuses the data allowing for the model to be assessed on a diverse range of data. And, it provides insights to model stability because, if it trains differently to each fold it might be very sensitive to the specific data. |
| --- |

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

| Feature Engineering is creating new features or changing existing features to enhance the representation of the data. It allows meaningful, informative, and relevant features to better train the model. You can create new features, handle missing data, encode the categorical variables, etc.  Feature selection is selecting a subset of the available features to be used as inputs to the machine learning model. It identifies the most important features to reduce dimensionality/complexity and improve model interpretability. You can select the most statistically significant ones with univariate selection. Or, simply see how the model computes its coefficients (Model-based) among other methods.  It helps with model performance by choosing the most important features, it can reduce overfitting, and help the generalization of the model. It is also less complex which allows for increased interpretability, therefore making it easier/quicker to train. |
| --- |

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

| Accuracy measures the overall success of the predictions made by a classification model. It calculates the proportion of correctly classified instances (both true positives and true negatives) out of the total number of instances in the dataset. It allows for an overall evaluation but may not be suitable when the classes are imbalanced.  Precision focuses on the accuracy of the positive predictions made by a classification model. It calculates the proportion of correctly predicted positive instances (true positives) out of the total number of instances predicted as positive (true positives + false positives). It shows the model's ability to avoid false positives which is helpful for medical diagnoses or fraud detection.  Recall measures the proportion of correctly predicted positive instances (true positives) out of the total number of actual positive instances (true positives + false negatives). It quantifies the model's ability to identify positive instances correctly. Higher value means lower rate of false negatives. |
| --- |

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