UNIT 6 ASSIGNMENT

Improving Performance with   
Ensemble Methods

## 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, exercises, and coding portion of the assignment. Now, let’s formulate your programming into well-thought 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 assignment page. Information about the grading rubric is available on any of the course assignment pages online. Do not hesitate to contact your facilitator if you have any questions about the assignment.*

Week 6 Written Portion

# Choosing Your Model

Answer the questions below about ensemble methods.

1. Explain ensemble modeling. What is the advantage of using this technique?

|  |
| --- |
| Ensemble modeling is when multiple models are combined to make more accurate predictions. Instead of relying on just one model, utilizing multiple allows overall performance to increase. Some advantages of using this technique are improved accuracy, increased robustness, reduced overfitting, better generalization, and a more diverse and interpretable model. |

1. Explain what bias and variance are, along with the bias-variance tradeoff.

|  |
| --- |
| Bias is the error introduced by approximating real-world problems with a model. Variance refers to the model’s sensitivity to fluctuations in the training data. It measures how much the predictions vary for different training sets. Bias variance trade-offs show the balance between the two—decreasing bias increases variance and decreasing variance increases bias. A model with high bias and low variance is simpler and less prone to overfitting, while a model with low bias and high variance is stronger and more flexible and therefore able to handle complex patterns (while risks overfitting). Finding the right balance allows for a model with ideal performance and generalization. |

1. Explain the differences among the ensemble methods bagging, boosting, and stacking.

|  |
| --- |
| Bagging is a technique to reduce variance by combining predictions from multiple models. To do so you train models independently on different subsets. Boosting also focuses on improving performance by training multiple models. However, it is done sequentially to correct the errors of the previous model. Weights are assigned to each training sample and misclassified samples receive higher weights in order for them to be fixed in the next iteration. Stacking, lastly, combines multiple models through a high-level model. To do so you train multiple base models and then use their predictions are inputs for the higher-level model—which makes the final predictions. |

1. Explain the random forest algorithm and how it relates to decision trees and bagging.

|  |
| --- |
| The random forest algorithm uses a number of decision trees trained through bagging with random feature selection in order to make accurate classifications. First, decision trees are trained. Instead of considering all features, the Random Forest algorithm randomly selects a subset of features in order to introduce diversity. Then, randomly selected data points are used for bagging. Each decision tree is trained independently and then once all of the trees are trained; the Random Forest algorithm combines their predictions in order to make the final predictions. For classification problems, this is done by a majority vote among the decision trees, while for regression problems the average of predicted values is taken. |

1. What’s the difference between gradient boosting decision trees and random forest?

|  |
| --- |
| Gradient boosting decision trees and Random Forest both utilize decision trees, however, they do so in different ways. Random Forest builds a number of decision trees in parallel and uses random feature selection and bagging. GBDT (Gradient Boosting Decision Trees) uses decision trees sequentially instead and corrects the errors of the previous iterations. GBDT tends to also use a smaller number of decision trees and focuses on refining the model’s weaknesses. Random Forest instead creates a large number of decision trees and creates a more complex model. |

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