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 are supervised learning algorithms that can improve predictive performance and model generalization. They are a class of techniques that can train multiple models and aggregates them by combining multiple independent models that are predicting the same outcome into a single prediction. The advantage of this technique is to achieve generalization by having low estimation bias and variance. |

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

|  |
| --- |
| Bias is defined as model rigidity that PREVENTS adaptation to nuances of the data. Variance is the model’s flexibility that causes the estimated model to be sensitive to data nuances. In Logistic regression we have higher bias but lower variance. In decision trees we have low bias but high variance. Every algorithm will have their own bias and variance tradeoff as different hyperparameters we choose for an algorithm will have their own bias and variance tradeoff. We explain the process of building multiple models with each of its own bias and variance tradeoff then averaging the predictions as the ensemble modeling. |

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

|  |
| --- |
| The differences among the ensemble methods: bagging, boosting, and stacking is that in stacking we take the weighted combination of the predictions of a total number of K different models. It is more of a general procedure that doesn’t have a specific supervised learning attached to it. It doesn’t have a specific algorithm implementation. Meanwhile in boosting there’s a specific algorithm implementation and we use the technique of bootstrapping and we have the random forest algorithm which utilizes the bagging technique. Random forest helps to find the optimal balance of low bias and variance for a decision tree. The random forest trees also trains on randomly sampled data and only utilizes a subset of features. In boosting we iteratively build models by focusing on the cumulative errors from prior iterations predictions. It can be done using a combination of any common supervised learning algorithms and this is considered to be the most math rigorous out of the three. Moreover, it utilizes the Gradient Boosted Decision Tree which consists of shallow (really simple) trees, trains on residuals, and uses all features. |

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

|  |
| --- |
| The bagging technique relies on the bootstrapping technique and the random forest algorithm utilizes the bagging technique to implement the algorithm. The random forest algorithm essentially is a set of decision trees and it combines them into an ensemble. We start with a full set of data and it’s split into training and testing set. We then take bootstrap samples by randomly sampling the training set into a k number of subset of features being randomly sampled. When making predictions on the test set, we make predictions from each of the trees and combine the results. The combined tree will tend to cancel out the error due to variance among one another, thus leading to good generalization. |

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

|  |
| --- |
| The differences between GBDT and random forest is that random forest consists of using deep trees, it trains on randomly sampled data and it uses a subset of features. Meanwhile GBDT consists of very simple (shallow) trees, it trains on residuals, and it utilizes all features. |

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