MACHINE LEARNING ASSIGNMENT -6

ANSWERS:

1. C) High R-squared value for train-set and Low R-squared value for test-set.
2. B) Decision trees are highly prone to overfitting.
3. C) Random Forest
4. B) Sensitivity
5. B) Model B
6. A) Ridge, D) Lasso
7. B) Decision Tree, C) Random Forest
8. A) Pruning, C) Restricting the max depth of the tree
9. A) We initialize the probabilities of the distribution as 1/n, where n is the number of data-points, B) A tree in the ensemble focuses more on the data points on which the previous tree was not performing well
10. The adjusted R-squared compensates for the addition of variables and only increases if the new predictor enhances the model above what would be obtained by probability. Conversely, it will decrease when a predictor improves the model less than what is predicted by chance.
11. Both lasso regression and ridge regression put a similar constraint on the coefficients by introducing a penalty factor. However, while lasso regression takes the magnitude of the coefficients, ridge regression takes the square.
12. Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables. A VIF value of 5 or below is not a cause for concern. As VIF increases, the less reliable your regression results are going to be.
13. Scaling gives equal weights/importance to each variable so that no single variable steers model performance in one direction just because they are bigger numbers.
14. Metrics used to check goodness of fit in linear regression model: R2 Squared, Adjusted R2 Squared, Mean Absolute Error, Mean Squared Error, Root Mean Squared Error
15. Calculations:
    - Sensitivity/Recall = TP/(TP+FN) = 1000/(1000+250) = 0.8
    - Specificity = TN/(TN+FP) = 1200/(1200+50) = 0.96
    - Precision = TP/(TP+FP) = 1000/(1000+50) = 0.95
    - Accuracy = (TP + TN)/(TP+FP+TN+FN) = (1000+1200)/(1000+1200+250+50)= 0.88