ASSIGNMENT 6

MACHINE LEARNING

1. In which of the following you can say that the model is overfitting?
Ans: Option C; High R-squared value for train-set and Low R-squared value for test-set.

2. Which among the following is a disadvantage of decision trees?
Ans: Option B; Decision trees are highly prone to overfitting.

3. Which of the following is an ensemble technique?
Ans: Option C; Random Forest

4. Suppose you are building a classification model for detection of a fatal disease where detection of the disease is most important. In this case which of the following metrics you would focus on?
Ans: Option C, Precision

5. The value of AUC (Area under Curve) value for ROC curve of model A is 0.70 and of model B is 0.85. Which of these two models is doing better job in classification?
Ans: Opition B; model B

6. Which of the following are the regularization technique in Linear Regression??
Ans: Option A and D; Ridge and Lasso

7. Which of the following is not an example of boosting technique?
Ans: Option B and C; Decision Tree and Random Forest

8. Which of the techniques are used for regularization of Decision Trees?
Ans: Option A and C; Pruning and Restricting the max depth of the tree

9. Which of the following statements is true regarding the Adaboost technique?
Ans: Option A and B; --We initialize the probabilities of the distribution as 1/n, where n is the number of data-points
--A tree in the ensemble focuses more on the data points on which the previous tree was not performing well

10. Explain how does the adjusted R-squared penalize the presence of unnecessary predictors in the model?
Ans: 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. Differentiate between Ridge and Lasso Regression.
Ans: Lasso regression penalises the model based on the sum of magnitude of the coefficients. Lasso acts like feature selection I.e. it only account features that are necessary for model evaluation.
Ridge regression penalizes the model based on the sum of squares of magnitude of the coefficients. Ridge does not act as feature selection tool.

12. What is VIF? What is the suitable value of a VIF for a feature to be included in a regression modelling?

Ans: Variace Inflation factor (VIF) is the measure of multicollinearity seen within the independent features. The VIF value within 5 is considerd non multicollinear while VIF values above 5 shows that those features are multicorrelated and these features need to be dropped before model building.

13. Why do we need to scale the data before feeding it to the train the model?

Ans: Scaling the data using standard scaler method ensures that all the features data have same units which is helpful for model building.

14. What are the different metrics which are used to check the goodness of fit in linear regression?

Ans: Five metrics gives us some hints about goodness of fit of our model.They are mean absolute error, root mean squared error, relative absolute error, relative squared error and coefficient of determination(R squared).

From the following confusion matrix calculate sensitivity, specificity, precision, recall and accuracy.

Actual/Predicted	True	False
True	1000	50
False	250	1200

Ans: Accuracy = TP+TN/TP+TN+FP+FN = 1000+1200/1000+1200+250+50 =2200/2500 = 0.88
Recall = TP/TP+FN = 1000/1000+50 = 0.95
Precision = TP/TP+FP = 1000/1000+250 = 0.8
Sensitivity = TP/TP+FN = 1000/1000+50 = 0.95
Specificity = TN/FP+TN = 1200/50+1200 = 1200/1250 = 0.96