MACHINE LEARNING (WORKSHEET SET- 6)

QUESTION-1 In which of the following you can say that the model is overfitting?

A) High R-squared value for train-set and High R-squared value for test-set.

B) Low R-squared value for train-set and High R-squared value for test-set.

C) High R-squared value for train-set and Low R-squared value for test-set.

D) None of the above

ANSWER-1- C) High R-squared value for train-set and Low R-squared value for test-set.

QUESTION-2 Which among the following is a disadvantage of decision trees?

A) Decision trees are prone to outliers.

B) Decision trees are highly prone to overfitting.

C) Decision trees are not easy to interpret

D) None of the above.

ANSWER 2- B) Decision trees are highly prone to overfitting.

QUESTION-3 Which of the following is an ensemble technique?

A) SVM

B) Logistic Regression

C) Random Forest

D) Decision tree

Answer-3 C) Random Forest

QUESTION-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?

A) Accuracy

B) Sensitivity

C) Precision

D) None of the above.

Answer-4 A) Accuracy

Question-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?

A) Model A

B) Model B

C) both are performing equal

D) Data Insufficient

ANSWER 5- B) Model B

QUESTION-6 Which of the following are the regularization technique in Linear Regression??
A) Ridge

B) R-squared

C) MSE

D) Lasso

ANSWER 6- A) Ridge and D) Lasso

QUESTION 7 Which of the following is not an example of boosting technique?

A) Adaboost

B) Decision Tree

C) Random Forest

D) Xgboost.

ANSWER 7- B) Decision Tree and C) Random Forest

QUESTION 8- Which of the techniques are used for regularization of Decision Trees?

A) Pruning

B) L2 regularization

C) Restricting the max depth of the tree

D) All of the above

ANSWER 8- A) Pruning

QUESTION 9- Which of the following statements is true regarding the Adaboost technique?
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

C) It is example of bagging technique

D) None of the above

ANSWER 9- D) None of the above

QUESTION 10 Explain how does the adjusted R-squared penalize the presence of unnecessary predictors in the model?

ANSWER 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.

QUESTION 11- Differentiate between Ridge and Lasso Regression.

ANSWER 11- Similar to the lasso regression, ridge regression puts 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. Ridge regression is also referred to as L2 Regularization.

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

ANSWER 12- A variance inflation factor (VIF) is a measure of the amount of multicollinearity in regression analysis. Multicollinearity exists when there is a correlation between multiple independent variables in a multiple regression model. This can adversely affect the regression results.

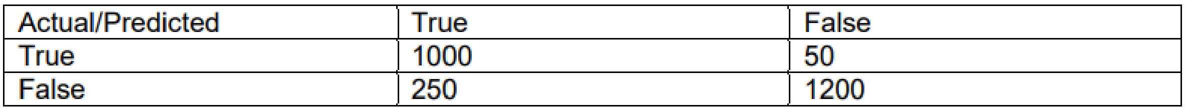
QUESTION 13. Why do we need to scale the data before feeding it to the train the model?

ANSWER-13- To ensure that the gradient descent moves smoothly towards the minima and that the steps for gradient descent are updated at the same rate for all the features, we scale the data before feeding it to the model.

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

ANSWER 14- The adjusted R-square statistic is generally the best indicator of the fit quality when you add additional coefficients to your model. The adjusted R-square statistic can take on any value less than or equal to 1, with a value closer to 1 indicating a better fit. A RMSE value closer to 0 indicates a better fit.

QUESTION-15 From the following confusion matrix calculate sensitivity, specificity, precision, recall and accuracy.



Actual/Predicted	True	False
True	1000	50
False	250	1200

ANSWER 15- ACCURACY-0.88

PRECISION-0.8

RECALL or SENSITIVITY-0.95

SPECIFICITY- 0.83