Machine Learning Worksheet

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 : 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?

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 : Decision trees are highly prone to overfitting**.**

3. Which of the following is an ensemble technique?

A) SVM B) Logistic Regression

C) Random Forest D) Decision tree

Answer : 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?

A) Accuracy B) Sensitivity

C) Precision D) None of the above.

Answer : Accuracy

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 : Model B

6. Which of the following are the regularization technique in Linear Regression??

A) Ridge B) R-squared

C) MSE D) Lasso

Answer : Lasso & Ridge

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

A) Adaboost B) Decision Tree

C) Random Forest D) Xgboost

Answer : Decision Tree & Random Forest

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 : All of the above

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 : 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?

Answer : Adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases when the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected. Typically, the adjusted R-squared is positive, not negative. It is always lower than the R-squared.

Adding more independent variables or predictors to a regression model tends to increase the R-squared value, which tempts makers of the model to add even more variables. This is called overfitting and can return an unwarranted high R-squared value. Adjusted R-squared is used to determine how reliable the correlation is and how much it is determined by the addition of independent variables.

11. Differentiate between Ridge and Lasso Regression.

Answer : Lasso Regression : The word "LASSO" denotes Least Absolute Shrinkage and Selection Operator. Lasso regression follows the regularization technique to create prediction. It is given more priority over the other regression methods because it gives an accurate prediction. Lasso regression model uses shrinkage technique. In this technique, the data values are shrunk towards a central point similar to the concept of mean. The lasso regression algorithm suggests a simple, sparse models (i.e. models with fewer parameters), which is well-suited for models or data showing high levels of multicollinearity or when we would like to automate certain parts of model selection, like variable selection or parameter elimination using feature engineering.

Lasso Regression algorithm utilises L1 regularization technique It is taken into consideration when there are more number of features because it automatically performs feature selection.

Ridge Regression : Ridge Regression is another type of regression algorithm in data science and is usually considered when there is a high correlation between the independent variables or model parameters. As the value of correlation increases the least square estimates evaluates unbiased values. But if the collinearity in the dataset is very high, there can be some bias value. Therefore, we create a bias matrix in the equation of Ridge Regression algorithm. It is a useful regression method in which the model is less susceptible to overfitting and hence the model works well even if the dataset is very small.

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

Answer : The Variance Inflation Factor (VIF) measures the severity of multicollinearity in regression analysis. It is a statistical concept that indicates the increase in the variance of a regression coefficient as a result of collinearity.

VIF is another commonly used tool to detect whether multicollinearity exists in a regression model. It measures how much the variance (or standard error) of the estimated regression coefficient is inflated due to collinearity.

Generally, a VIF above 4 or tolerance below 0.25 indicates that multicollinearity might exist, and further investigation is required. When VIF is higher than 10 or tolerance is lower than 0.1, there is significant multicollinearity that needs to be corrected.

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

Answer : Scaling the target value is a good idea in regression modelling; scaling of the data makes it easy for a model to learn and understand the problem. Scaling of the data comes under the set of steps of data pre-processing when we are performing machine learning algorithms in the data set.

Scaling of the data comes under the set of steps of data pre-processing when we are performing machine learning algorithms in the data set. As we know most of the supervised and unsupervised learning methods make decisions according to the data sets applied to them and often the algorithms calculate the distance between the data points to make better inferences out of the data.

Similarly in the machine learning algorithms if the values of the features are closer to each other there are chances for the algorithm to get trained well and faster instead of the data set where the data points or features values have high differences with each other will take more time to understand the data and the accuracy will be lower.

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

Answer : There are three error metrics that are commonly used for evaluating and reporting the performance of a regression model , they are:

Mean Squared Error (MSE).

Root Mean Squared Error (RMSE).

Mean Absolute Error (MAE)

Mean Squared Error

Mean Squared Error, or MSE for short, is a popular error metric for regression problems.

It is also an important loss function for algorithms fit or optimized using the least squares framing of a regression problem. Here “least squares” refers to minimizing the mean squared error between predictions and expected values.

The MSE is calculated as the mean or average of the squared differences between predicted and expected target values in a dataset.

Root Mean Squared Error

The Root Mean Squared Error, or RMSE, is an extension of the mean squared error.

Importantly, the square root of the error is calculated, which means that the units of the RMSE are the same as the original units of the target value that is being predicted.

For example, if your target variable has the units “dollars,” then the RMSE error score will also have the unit “dollars” and not “squared dollars” like the MSE.

As such, it may be common to use MSE loss to train a regression predictive model, and to use RMSE to evaluate and report its performance.

Mean Absolute Error

Mean Absolute Error, or MAE, is a popular metric because, like RMSE, the units of the error score match the units of the target value that is being predicted.

Unlike the RMSE, the changes in MAE are linear and therefore intuitive.

That is, MSE and RMSE punish larger errors more than smaller errors, inflating or magnifying the mean error score. This is due to the square of the error value. The MAE does not give more or less weight to different types of errors and instead the scores increase linearly with increases in error.

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

Actual/Predicted	**True**	**False**
True	**1000**	**50**
False	**250**	**1200**

Answer : **Sensitivity :** TP / TP + FN = 1000/1000+250 = 1000/1250 = **80%**

Accuracy = TP + TN / TP + TN + FP + FN= 1000+50/1000+50+1200+250 = 1050/2500 = **42%**

Specificity = TN / TN + FP = 50/50+1200 = 50/1250 = **4%**

Precision = TP / TP + FP = 1000/1000+1200 = 1000/2200=**45.45%**

Recall = TP / (TP+FN) = 1000/1000+250 = 1000/1250 = **80%**