Worksheet\_Set\_4

Machine Learning

. In which of the following you can say that the model is overfitting?

Ans:- Low R-squared value for train-set and High R-squared value for test-set.

Which among the following is a disadvantage of decision trees?

Ans:- Decision trees are not easy to interpret

. Which of the following is an ensemble technique?

Ans:- Random Forest

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:- Accuracy

. 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:- Model B

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

Ans:-Ridge,Lasso

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

Ans:- Random Forest.

. Which of the techniques are used for regularization of Decision Trees?

Ans:- Pruning, Restricting the max depth of the tree

. Which of the following statements is true regarding the Adaboost technique?

Ans:-None of the above

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

Ans:-Adjusted R-squared is a measure of the goodness of fit of a multiple regression model that accounts for the number of predictors in the model. It penalizes the presence of unnecessary predictors in the model by reducing the R-squared value. When the predictor is added to the model it will increase the R-squared value. If unnecessary predictors are added then it will decrease the adjusted R-squared.

the adjusted R-squared is a useful tool for evaluating the goodness of fit of a multiple regression model.

Differentiate between Ridge and Lasso Regression?

Ans:- Ridge - Ridge regression works by attempting at increasing the bias to improve variance (generalization capability) This works by changing the slope of the line.The model performance might be little poor on the training set but it will perform consistently well on both the training and testing datasets.

Lasso:-Lasso Regression is a type of linear regression that adds a penalty term to the cost function to reduce the magnitude of the coefficients of the predictors. The penalty term is the L1 norm of the coefficients, which is the sum of the absolute values of the coefficients. Lasso Regression is a useful tool for feature selection, as it can automatically select a subset of the predictors that are most relevant to the response variable. This is particularly useful in situations where there are many predictors, as it can help to reduce overfitting and improve the interpretability of the model.

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

Ans:- Variance Inflation Factor. Variance Inflation Factor, is a measure of the multicollinearity between independent variables in a multiple regression model. It measures how much the variance of an estimated regression coefficient is increased due to the presence of other correlated predictors in the model. It is important to note that a high VIF value for a predictor does not necessarily indicate that it is unimportant or should be removed from the model, as it may still have a significant impact on the response variable even if it is highly correlated with other predictors. In such cases, it may be more appropriate to keep the predictor in the model and to interpret the results carefully, taking into account the high levels of multicollinearity.

Overall, VIF is a useful tool for identifying and diagnosing multicollinearity in multiple regression models, and can help to guide the selection of predictors for the model. A VIF value of 10 or less is generally considered to be a suitable threshold for including a feature in a regression model.

Top of Form

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

Ans:- Improve model performance. Scaling the data to have a mean of zero and a standard deviation of one, or scaling it to have a minimum value of zero and a maximum value of one, can help improve the performance of these algorithms by reducing the impact of the scale of the input data on the model.

Preventing One feature dominating the other, If the features in the data have different scales, then one feature could dominate the others and make it difficult for the machine learning algorithm to accurately learn the relationship between the features and the target variable.

Speed up convergence:- learning algorithms, such as gradient descent, use an optimization process to find the best parameters for the model.

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

Ans:- R-squared- R-squared measures the proportion of the variation in the response variable that is explained by the model. The value of R-squared ranges from 0 to 1, where a value of 1 indicates that the model perfectly fits the data, and a value of 0 indicates that the model does not explain any of the variation in the response variable.

Mean Squared Error:- MSE measures the average squared difference between the predicted values and the actual values of the response variable. The smaller the MSE, the better the model fits the data.

Mean Absolute Error:- MAE measures the average absolute difference between the predicted values and the actual values of the response variable. The smaller the MAE, the better the model fits the data.

Root Mean Squared Error:- RMSE is the square root of the MSE and measures the average magnitude of the error between the predicted values and the actual values of the response variable.

Adjusted R-squared:-Adjusted R-squared takes into account the number of predictors in the model and adjusts R-squared to reflect the quality of the model in terms of its ability to explain the variation in the response variable. A higher adjusted R-squared value indicates a better fit of the model to the data.

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

Ans:- Accuracy-TP+TN/ALL SUM= 0.88

Sensitivity:- TP/TP+FN= 0.95

Specificity (TNR)- TN/TN+FP = 0.82

Precision – TP/TP+FP = 0.8

Recall - 0.95