MACHINE LEARNING

Ans1: High R-squared value for train-set and Low R-squared value for test-set.

Ans2: Decision trees are prone to outliers.

Ans3: Random Forest

Ans4: Sensitivity

Ans5: Model B

Ans6: Ridge, Lasso

Ans7: Decision Tree, Random Forest

Ans8: Pruning, Restricting the max depth of the tree

Ans9: 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

Ans10: 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 increase 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. Adjusted R-squared is positive, not negative. It is always lower than the R-squared. 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.

Ans11: Lasso regression is also called Penalized regression method. This method is usually used in machine learning for the selection of the subset of variables. It provides greater prediction accuracy as compared to other regression models. Lasso regularization helps to increase model interpretation. Ridge regression is a

model tuning method that is used to analyze any data that suffers from multicollinearity. This method performs L2 regularization. Ridge and Lasso regression are powerful techniques generally used for creating parsimonious models in presence of a large number of features. The main difference between Ridge and Lasso regression is that if ridge regression can shrink the coefficient close to 0 so that all predictor variables are retained. Where as Lasso can shrink the coefficient to exactly 0 so that Lasso can select and discard the predictor variables that have the right coefficient of 0.

Ans12: Variance Inflation Factor (VIF) measures the severity 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. Thus, the variance inflation factor (VIF) can estimate how much the variance of a regression coefficient is inflated due to multicollinearity. In general, a VIF above 10 indicates high correlation and is cause for concern. Some authors suggest a more conservative level of 2.5 or above. Sometimes a high VIF is no cause for concern at all. For example, you can get a high VIF by including products or powers from other variables in your regression model.

Ans13: We need to scale the data before feeding it to the train the model because Scaling of the data come under the set of steps of data preprocessing when we are performing machine learning algorithms in the dataset. As we know most of the supervised and unsupervised learning methods make decisions according to the datasets applied to them and often the algorithms calculate the distance between the datapoints to make better inferences out of the data. Scaling of the data makes it easy for a model to learn and understand the problem. 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.

Ans14: There are many other metrics but there are three error metrics that are commonly used for evaluating and reporting the performance to check the goodness of fit in linear regression they are:

Mean Squared Error (MSE): The most common metric is MSE. It is the average of the squared difference between the predicted and actual value. Since it is differentiable and has a convex shape, it is easier to optimize. MSE penalize large errors.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

Mean Absolute Error (MAE): This is simply the average of the absolute difference between the target value and the value predicted by the model. It is not preferred in cases where outliers are prominent. MAE does not penalize large errors.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|$$

Root Mean Squared Error (RMSE): This is the square root of the squared difference of the predicted and actual value. RMSE is just the root of the average pf squared residuals. Residuals are a measure of how distant the points are from the regression line. Thus, RMSE measures the scatter of these residuals. RMSE penalizes large errors.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

Ans15: So here we have:

True Positive (TP): 1000

False Positive (FP): 50

True Negative (TN): 1200

False Negative (FN): 250

We need to calculate sensitivity, specificity, precision, recall and accuracy.

Accuracy = TP + TN / TP + FP + TN + FN

1000 + 1200 / 1000 + 50 + 1200 + 250

Accuracy = 0.88

Sensitivity = TP / TP +FN

1000 / 1000 + 250

Sensitivity = 0.8

Specificity = TN / TN + FP

1200 / 1200 + 50

Specificity = 0.96

Precision = TP / TP + FP

1000 / 1000 + 50

Precision = 0.95

Recall = TP / TP + FN

1000 / 1000 + 250

Recall = 0.8