MACHINE LEARNING ASSIGNMENT – 6

1. B

2. B

3. C

4. B

5. B

6. A, D

7. B, C

8. D

9. B

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

The adjusted R-squared can help us to identify whether adding more predictors to our model is actually improving the model's fit to the data or not. It's a good practice to prefer the models with higher adjusted R-squared, as it gives us the information about the model's performance considering the number of predictors used.

Adjusted R-squared = 1 - ((1 - R-squared) * (n - 1) / (n - p - 1))

Where n is the number of observations and p is the number of predictors in the model.

11. **Differentiate between Ridge and Lasso Regression.**

Ridge Regression (L2 regularization): It adds a penalty term to the cost function that is the sum of the square of the coefficients. This penalty term is called L2 regularization because it is the sum of the square of the coefficients (also known as the L2 norm). Ridge regression tends to distribute the penalty equally among all the coefficients, which can result in all the coefficients shrinking by a similar amount, but not necessarily becoming zero.

Lasso Regression (L1 regularization): It adds a penalty term to the cost function that is the sum of the absolute value of the coefficients. This penalty term is called L1 regularization because it is the sum of the absolute value of the coefficients (also known as the L1 norm). Lasso regression tends to shrink some of the coefficients towards zero, which can be useful for feature selection.

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

VIF stands for Variance Inflation Factor, it's a measure of the correlation between the independent variables in a multiple linear regression model. It is used to check the multicollinearity, which is a situation where two or more independent variables are highly correlated with each other. Multicollinearity can cause problems in the estimation of regression coefficients, by making them unstable and imprecise.

A VIF value of 1 indicates no correlation between the feature and other features in the model, while a VIF greater than 1 indicates that the feature is correlated with one or more other features in the model. The suitable value of a VIF for a feature to be included in a regression modeling is typically considered to

be less than 5, and a value of 10 or more is considered to indicate a high level of multicollinearity. However, it's important to consider the context of the problem, and to evaluate the VIF value in combination with other statistics such as correlation matrix, and in the light of the subject matter expertise.

It's important to note that, if a feature has a high VIF, it does not mean that the feature is not useful for the model, it's just that it is highly correlated with other features in the model, so one of them should be removed.

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

Scaling the data is an important step in preprocessing the data before training a machine learning model. There are a few reasons why we need to scale the data-

Many machine learning algorithms, such as linear regression and support vector machines, are sensitive to the scale of the input features. Without scaling, the model would be more sensitive to larger scale features, which would make it difficult to compare the importance of different features.

Scaling the data can help to improve the convergence rate of the optimization algorithms used to train the model. This is because many optimization algorithms use the gradient of the cost function to adjust the model parameters, and the scale of the input features can affect the size of the gradient and the learning rate.

Scaling the data can also help to improve the performance of some models like K-means clustering, Neural Networks, etc.

Scaling the data can help to reduce the impact of outliers, as they will be scaled down along with the rest of the data.

There are several ways to scale the data, such as normalization and standardization. Normalization scales the data to a range between 0 and 1, while standardization scales the data so that it has a mean of 0 and a standard deviation of 1. The choice of scaling method will depend on the specific machine learning algorithm and the characteristics of the data.

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

There are several metrics that can be used to check the goodness of fit of a linear regression model, some of the most common ones are-

R-squared (Coefficient of determination): It is a measure of the proportion of variance in the dependent variable that is explained by the independent variables in the model. It ranges from 0 to 1, with a value of 1 indicating a perfect fit and a value of 0 indicating that the model does not explain any of the variation in the dependent variable.

Adjusted R-squared: It is a modified version of R-squared that takes into account the number of predictors in the model, it's a better indicator than R-squared when comparing models with different number of predictors.

Root Mean Squared Error (RMSE): It is the square root of the mean of the squared differences between the predicted and actual values. It measures the average magnitude of the error in the predictions.

Mean Absolute Error (MAE): It is the average of the absolute differences between the predicted and actual values. It measures the average magnitude of the error in the predictions.

15. **Precision= 0.8, Recall/Sensitivity = 0.95, Specificity = 0.82 & Accuracy = 0.88**