**Machine-Learning worksheet 2**

1. C) High R-squared value for train-set and Low R-squared value for test-set
2. B) Decision trees are highly prone to overfitting.
3. C) Random Forest
4. A) Accuracy
5. B) Model B
6. A) Ridge

D) Lasso

7. B) Decision Tree

C) Random Forest

8. A) pruning

C) Restricting the max depth of the tree

9. 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.

10. The adjusted R-squared looks at whether additional input variables are contributing to the model.

Compared to a model with additional input variables, a lower adjusted R-squared indicates that the additional input variables are not adding value to the model.

Compared to a model with additional input variables, a higher adjusted R-squared indicates that the additional input variables are adding value to the model.

11. Ridge and Lasso regression uses two different penalty functions. Ridge uses l2 where as lasso uses l1. In ridge regression, the penalty is the sum of the squares of the coefficients and for the Lasso, it’s the sum of the absolute values of the coefficients.

**Ridge regression shrinks the coefficients and it helps to reduce the model complexity and multi-collinearity.**

**Lasso regression not only helps in reducing over-fitting but it can help us in feature selection.**

12. A variance inflation factor(VIF) detects multicolinearity in regression analysis. The VIF estimates how much the variance of a regression coefficient is inflated due to multicollinearity in the model.

The suitable value of a VIF for a feature to be included in a regression modelling is less than 10.

13. Feature scaling in machine learning is one of the most critical steps during the pre-processing of data before creating a machine learning model. Scaling can make a difference between a weak machine learning model and a better one.

Machine learning algorithm just sees number — if there is a vast difference in the range say few ranging in thousands and few ranging in the tens, and it makes the underlying assumption that higher ranging numbers have superiority of some sort. So these more significant number starts playing a more decisive role while training the model. Feature scaling is needed to bring every feature in the same footing without any upfront importance.

14. The following are some of the metrics that can be used to check the goodness of fit in regression models:

* R squared
* Adjusted R squared
* MAE
* RMSE etc.