Machine Learning Worksheet

1. We cannot use R-squared to determine whether the coefficient estimates and predictions are biased, which is why we must assess the residual plots. R-squared does not indicate if a regression model provides an adequate fit to our data. A good model can have a low R^2 value. On the other hand, a biased model can have a high R^2 value! Hence, it is better to use Adjusted R Squared method where RSS is implied.
2. **TSS (total sum of squares)**is the squared sum of difference between the actual value and average value of target variable.
   TSS= = ∑(yi – ȳ)2 , here yi=Actual,ȳ average value of y.(also known as variance of target)
   RSS is the **Residual Sum of Squares** taken to get the variance of the target value around the best fit line.
   RSS= ∑(yi – ŷ)2.here yi=actual ,ŷ predicted value
   ESS, Explained Sum of Squares is the squares of deviation of predicted values from the mean value.
   ESS= ∑(ŷ - ȳ)2

   R^2 = 1 - RSS / SST = 1- RSS/ (ESS+RSS)

3. Regularization reduces the error by fitting the function appropriately on the dataset and reduces overfitting problem. Two main regularization techniques are L1 and L2 regularization techniques in machine learning.
4. **Gini index** or **Gini impurity** measures the degree or probability of a particular variable being wrongly classified when it is randomly chosen.
5. Unlike other regression models, decision tree doesn't use regularization to fight against overfitting. Instead, it employs tree pruning. Selecting the right hyperparameters (tree depth and leaf size) also requires experimentation, e.g. doing cross-validation with a hyperparameter matrix.
6. Ensemble Techniques combine the decisions from multiple models to improve the overall performance. They are of 2 types: Bagging and Boosting.
7. **Bagging**: Its full name is bootstrap aggregation. As its name, bagging applies bootstrap method on data set. First bagging uses bootstrap with replacement to generate a new data set, and fits a decision tree on it. Then bagging will repeat doing this to generate many decision trees. The final result will be the average of the decision trees.
   Boosting. Boosting mainly has two types:

   Adaboost: Adaboost changes the weight of samples to let algorithm pays more attention on wrong classified samples. Take a regression problem as an example. It starts with equal weight for all simples, such as 1/n. Then it fits a decision tree on the data, sum the weight of misclassified samples. Compute the contribution of this classifier by sum weight of misclassified

samples. Update the weight by contribution to all samples. Fit the residual of the first model by another decision tree, and update the weight as previous step.

Gradient boosting. The whole structure of Gradient boosting is same as Adaboosting, but the main difference is how it calculates the residual. The residual is defined as this: h(x)=y−Fm(x)h(x)=y−Fm(x)But gradient boost treats it as this 12(y−F(x))212(y−F(x))2, so gradient descent method can be used here to calculate the residual.

8. Out of bag error is the measure of prediction error of decision tree in random forest models. It is obtained by averaging the number of errors for out of bag samples in each estimator.
9. K in cross validation refers to the number of groups the data needs to be split into. Cross validation is a technique to evaluate machine learning models. The performance of each group is noted and finally the mean of all metric outcomes is taken as the final score of the model. This techniques helps to check and remove overfitting problem in the models.
10. Hyperparamter tuning is used to get the best set of parameters of any algorithm for the particular data being used. If no hyperparameter tuning is done then algorithm chooses default parameter for all type of data which of course might not be giving the best possible outcome of the model. In short the hyperparameter tuning brings out the best possible version of model by giving the right parameters for the data being used.
11. Large learning rate may introduce the issue of fast jumping of the derivates such that it did not reach the global minima. It is advised to choose the learning rate appropriately so that the gradient descent operation can be smooth and global minima point can be obtained easily. Large learning rate causes problem for the derivate to converge on the global minima.
12. Logistic regression is meant to be used in linear data problems where the classes are linearly separable by a line (plane or hyperplane). Non linear data will not give a linear decision boundary and hence errors will be more and more prediction will be incorrect.
    In logistic regression the output is taken as **sigmoid(best fit line)** OR **sigmoid(mx+c).**
    Since sigmoid function is used , there need to be a threshold value for x after which sig(mx+c) results to 1 else 0. This threshold value cannot be defined in case of non linear data. Even if it is defined we will get most wrong predictions.
13. Both Ada Boost and gradient boosting converts set of weak learners into a single strong learner. The difference is in the way they create weak learners during the iterative process.
    Ada Boost changes the sample distribution by modifying the weights attached to each instances. It increases the weights of the incorrect instances and decreases the weight of correctly predicted instances. This was the weak learners will focus more on the difficult instances . After getting trained the weak learners gets added to strong learner based on its performance. The higher it performs the more it contributes to strong learner.
    Gradient Boost do not change the sample distribution instead the weak learners get trained on only the remaining errors (incorrect predictions) made by the strong learner.
    At each iteration the remaining errors are computed and a weak learner is fitted to these remaining errors. Finally contribution of weak learner to strong learner is not decided by its performance (as in Ada grad) instead a gradient optimization process is used for this purpose.
14. Bias error occurs when model undergoes underfitting . Due to underfitting ,model try to make assumptions about the target function. Hence model gets biased.
    Low Bias: Less assumptions are made about the form of target function

High bias: More assumptions are made about the form of target function
Variance error is the amount that the estimated target function by the model will change with respect to new data.
Low variance: small changes to the estimated target function on occurrence of new data
High variance: large changes to the estimated target function on occurrence of new data

To get good prediction results a model should have low bias and low variance. Process of achieving this low bias and low variance by changing or choosing right parameters is known as **bias variance Trade off.**

15. **Linear-** Used for linear model where data is linearly separable.
    Polynomial- Decision boundary is of a polynomial function.
    RBF- RBF kernel is a function whose value depends on the distance from the origin or from some point.