MACHINE LEARNING -5

1) BOTH R-SQUARED AND RESIDUAL SUM OF SQUARES ARE USEFUL MEASURES IN EVALUATING THE GOODNESS OF FIT OF A MODEL,BUT THEY BOTH SERVE DIFFERENT PURPOSES. R-SQUARED IS A USEFUL MEASURE TO ASSESS THE OVERALL FIT OF THE MODEL AND TO COMPARE DIFFERENT MODELS WHILE RESIDUAL SUM OF SQUARES(RSS) IS USEFUL TO IDENTIFY THE DEGREE OF ERROR IN THE MODEL'S PREDICTION. THE CHOICE BETWEEN THEM DEPENDS ON THE CONTEXT AND WHAT YOU WANT TO EVALUATE.

2) TSS(TOTAL SUM OF SQUARES)- IT IS THE SUM OF SQUARED DIFFERENCES BETWEEN THE OBSERVED DEPENDENT VARIABLE AND THE OVERALL MEAN.
   ESS(EXPLAINED SUM OF SQUARES)- IT IS THE SUM OF DIFFERENCES BETWEEN THE PREDICTED VALUE AND THE MEAN OF DEPENDENT VARIABLE.IT DESCRIBES HOW WELL OUR LINE FITS THE DATA
   RSS(RESIDUAL SUM OF SQUARES)- IT MEASURES THE TOTGAL SUM OF SQUARED DIFFERENCES BETWEEN THE ACTUAL VALUES OF THE DEPENDENT VARIABLE AND THE PREDICTED VALUES.
   TSS=ESS+RSS

3) WHILE TRAINING A MACHINE LEARNING MODEL, THE MODEL CAN EASILY BE OVERFITTED OR UNFITTED.TO AVOID THIS, WE USE REGULARIZATION IN MACHINE LEARNING TO PROPERLY FIT A MODEL INTO OUR TEST SET. REGULARIZATION TECHNIQUE HELP REDUCE THE CHANCES OF OVERFITTING AND HELP US GET AN OPTIMAL MODEL.

4) GINI IMPURITY IS A MEASUREMENT USED TO BUILD DECISION TREES TO DETERMINE HOW THE FEATURES OF A DATASET SHOULD SPLIT NODES TO FORM THE TREE.GINI IMPURITY RANGES BETWEEN 0-0.5. THE FEATURE WITH THE LOWEST IMPURITY WOULD DETERMINE THE BEST FEATURE FOR SPLITTING THE CURRENT NODE.

5) YES UNREGULARIZED DECISION TREES ARE PRONE TO OVERFITTINGSPECIALLY WHEN A TREE IS DEEP. THIS IS DUE TO THE AMOUNT OF SPECIFICITY WE LOOK AT LEADING TO SMALLER SAMPLE OF EVENTS THAT MEET THE PREVIOUS ASSUMPTIONS. THIS SMALL SAMPLE COULD LEAD TO AN UNSOUND CONCLUSIONS.

6) ENSEMBLE TECHNIQUE IS A MACHINE LEARNING TECHNIQUE THAT ENHANCES ACCURACY AND RESILIENCE IN FORECASTING BY MERGING PREDICTIONS FROM MULTIPLE MODELS. IT AIMS TO MITIGATE ERRORS OR BIASES THAT MAY EXIST IN INDIVIDUAL MODELS BY LEVERAGING THE COLLECTIVE INTELLIGENCE OF THE ENSEMBLE.

7) BOTH BAGGING AND BOOSTING ARE DIFFERENT ENSEMBLE TECHNIQUE THE MAJOR DIFFERENCES ARE

BAGGING-1) IT COMBINES MULTIPLE MODELS TRAINED ON DIFFERENT SUBSETS OF DATA.

2) TO REDUCE VARIANCE BY AVERAGING OUT INDIVIDUAL MODEL ERROR.

3) EACH MOLDEL SERVES EQUAL WEIGHT IN THE FINAL DECISION.

4) IT IMPROVES ACCURACY BY REDUCING VARIANCE.

BOOSTING-1) TRAIN MODELS SEQUENTIALLY FOCUSING ON THE ERROR MADE BY THE PREVIOUS MODEL.

2) REDUCES BOTH BIAS AND VARIANCE BY CORRECTING MISCLASSIFICATION OF PREVIOUS MODEL.

3)MODELS AREWEIGHTED ON THE BASIS OF ACCURACY, BETTER ACCURACY MODEL WILL HAVE A HIGHER WEIGHT.

4ACHIEVES HIGHER ACCURACY BY REDUCING BOTH BIAS AND VARIANCE.

8) THE OUT OF BAG ERROR IS AN AVERAGE ERROR FOR EACH CALCULATION USING PREDICTIONS FROM THE TREES THAT DONOT CONTAIN IN THEIR RESPECTIVE BOOT STRAP SAMPLE. THIS ALLOWS RANDOM FOREST CLASSIFIER TO BE FIT AND VALIDATED WHILE BEING TRAINED.

9) CROSS VALIDATION IS AN EVALUATION METHOD USED IN MACHINE LEARNING TO FIND OUT HOW WELL YOUR MACHINE LEARNING MODEL CAN PREDICT THE OUTCOME OF UNSEEN DATA. IT IS A METHOD THAT IS EASY TO COMPREHEND, WORKS WELL FOR A LIMITED DATA SIMPLE AND ALSO OFFERS AN EVALUATION THAT IS LESS BIASED. THE DATA SAMPLE IS SPLIT INTO “K” NO OF SMALLER SAMPLES, HENCE THE NAME K-FOLD CROSS VALIDATION.

10) HYPERPARAMETERS DIRECTLY CONTROL MODEL STRUCTURE,FUNCTION AND PERFORMANCE.HYPERPARAMETER TUNING ALLOWS DATA SCIENTIST TO TWEAK MODEL PERFORMANCE FOR OPTIMAL RESULTS. THIS PROCESS IS AN ESSENTIAL PART OF MACHINE LEARNING, AND CHOOSING APPROPRIATE HYPERPARAMETER VALUE WHICH IS CRUCIAL FOR SUCCESS.

11) THE CHOICE OF LEARNING RATE CAN SIGNIFICANTLY IMPACT THE PERFORMANCE OF GRADIENT DESCENT. IF THE LEARNING RATE IS TOO HIGH, THE ALGORITHM MAY OVERSHOOT THE MINIMUM.

12) NO, WE CANNOT USE LOGISTIC REGRESSION FOR THE CLASSIFICATION OF NON-LINEAR DATA AS LOGISTIC REGRESSION RUNS ON AASSUMPTION THAT THERE IS A LINEAR RELATIONSHIP BETWEEN INPUT AND OUTPUT. THEREFORE, IT FAILS TO CAPTURE THE NON-LINEARITY OF THE DATA.

13) ADABOOST IS COMPUTED WITH A SPECIFIC LOSS FUNCTION AND BECOMES MORE RIGID WHEN COMES TO FEW ITERATIONS.
BUT IN GRADIENT BOOSTING IT ASSIST IN FINDING THE PROPER SOLUTION TO THE ADDITIONAL ITERATION MODELLING PROBLEM AS IT IS BUILT WITH SOME GENERIC FEATURES

14) IN MACHINE LEARNING,AS YOU TRY TO MINIMIZE ONE COMPONENT OF THE ERROE(the bias), THE OTHER COMPONENT(variance) TENDS TO INCREASE , AND VICEVERSA.FINDING THE RIGHT BALANCE OF BIAS AND VARIANCE IS KEY TO CREATING AN EFFECTIVE AND ACCURATE MODEL.THIS IS CALLED THE BIAS-VARIANCE TRADEOFF.
15) GIVE SHORT DESCRIPTION OF

LINEAR -LINEAR KERNEL IS USED WHEN THE DATA IS LINEARLY SEPARABLE THAT IS IT CAN BE SEPARATED USING A SINGLE LINE. IT IS MOSTLY USED WHEN THERE ARE LARGE NUMBER OF FEATURES IN A PARTICULAR DATASET.

RBF- RADIAL BASIS FUNCTION KERNEL IS A POWERFUL KERNEL USED IN SVM. UNLIKE LINEAR OR POLYNOMIAL KERNELS,RBF IS MORE COMPLEX AND EFFICIENT AT THE SAME TIME THAT IT CAN COMBINE MULTIPLEPOLYNOMIAL KERNEL MULTIPLE TIMES OF DIFFERENT DEGREES TO PROJECT THE NON-LINEARLY SEPARABLE DATA INTO HIGHER DIMENSIONAL SPACE SO THAT IT CAN BE SEPARABLED USING A HYPERPLANE.
POLYNOMIAL-IT REPRESENTS THE SIMILARITY OF VECTORS IN THE TRAINING SET OF DATA IN A FEATURE SPACE OVER POLYNOMIALS OF THE ORIGINAL VARIABLES USED IN THE KERNEL.