Rules

* The questions are MCQ types. Each question carries equal marks. The time allotted is 120 minutes.
* There is no negative marking for any wrong answer.
* There is only one option correct for each question.
* Once an answer is submitted for a question, the participant is not allowed to visit the question again.

In this skilltest, we will take a case study approach; we will try to go through each and every aspect of Decision tree and try to understand it in more depth. Note that the first 8 questions are very basic, and targeted to evaluate that you know at least the basics. After the first 8 questions, questions will be randomly asked on topics pertaining tree based algorithms.

Topics covered:

Decision Tree

Random Forest

Gradient Boosting

Total number of questions: 45

**Q** 1) The data scientists at **"BigMart Inc"** have collected 2013 sales data for 1559 products across 10 stores in different cities. Also, certain attributes of each product based on these attributes and store have been defined. The aim is to build a predictive model and find out the sales of each product at a particular store during a defined period.

Which learning problem does this belong to?

supervised learning

unsupervised learning

reinforcement learning Bottom of Form

Q 2) Before building our model, we first look at our data and make predictions manually. Suppose we have only one feature as an independent variable (Outlet\_Location\_Type) along with a continuous dependent variable (Item\_Outlet\_Sales).

Outlet\_Location\_Type Item\_Outlet\_Sales

Tier 1 3735.14

Tier 3 443.42

Tier 1 2097.27

Tier 3 732.38

Tier 3 994.71

We see that we can possibly differentiate in Sales based on location (tier 1 or tier 3). We can write simple if-else statements to make predictions. Which of the following models could be used to generate predictions (may not be most accurate)?

if "Outlet\_Location" is Tier 1: then "Outlet\_Sales" is 2000, else "Outlet\_Sales" is 1000

if "Outlet\_Location" is Tier 1: then "Outlet\_Sales" is 1000, else "Outlet\_Sales" is 2000

if "Outlet\_Location" is Tier 3: then "Outlet\_Sales" is 500, else "Outlet\_Sales" is 5000

Any of the above

Q 13) Random forests (While solving a regression problem) have higher variance of predicted result in comparison to Boosted Trees (Assumption: both Random Forest and Boosted Tree are fully optimized). True or False?

TRUE

FALSE

Cannot be determined

Q 14) Assume everything else remains same, which of the following is the right statement about the predictions from decision tree in comparision with predictions from Random Forest and Boosted Decision Tree?

Lower Variance, Lower Bias

Lower Variance, Higher Bias

Higher Variance, Higher Bias

Lower Bias, Higher Variance

Q 15) Which of the following tree based algorithm uses some parallel (full or partial) implementation?

Random Forest

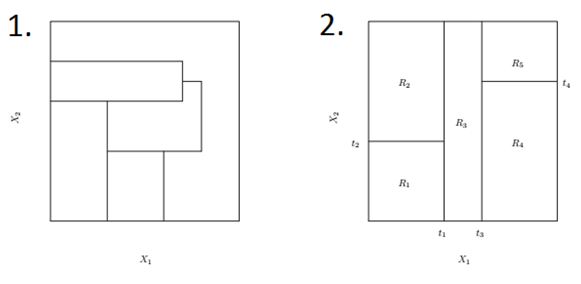
Gradient Boosted Trees

XGBOOST

Both A and C

A, B and C

Q 16) Which of the following could not be result of two-dimensional feature space from natural recursive binary split?

[](https://www.analyticsvidhya.com/wp-content/uploads/2016/11/Q16_Image.jpg)

Top of Form

1 only

2 only

1 and 2

None

Bottom of Form

Q 19) Let's say we have m number of estimators (trees) in a boosted tree. Now, how many intermediate trees will work on modified version (OR weighted) of data set?

1

m-1

m

Can't say

None of the above

Q 20) Boosted decision trees perform better than Logistic Regression on anomaly detection problems (Imbalanced Class problems). True or False?

True, because they give more weight for lesser weighted class in successive rounds

False, because boosted trees are based on Decision Tree, which will try to overfit the data

Q 21) Provided n < N and m < M. A Bagged Decision Tree with a dataset of N rows and M columns uses\_\_\_\_rows and \_\_\_\_ columns for training a individual intermediate tree.

N, M

N, m

n, M

n, m

Q 22) Given 1000 observations, Minimum observation required to split a node equals to 200 and minimum leaf size equals to 300 then what could be the maximum depth of a decision tree?

1

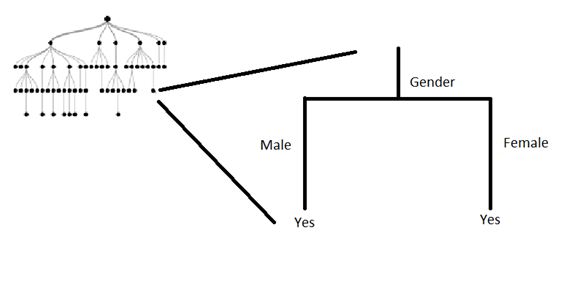
2

3

4

5

Q 23) Consider a classification tree for whether a person watches 'Game of Thrones' based on features like age, gender, qualification and salary. Is it possible to have following leaf node?

[](https://www.analyticsvidhya.com/wp-content/uploads/2016/11/Q25_Image.jpg)

Top of Form

Yes

No

Can't say

Q 25) In which of the following application(s), a tree based algorithm can be applied successfully?

Recognizing moving hand gestures in real time

Predicting next move in a chess game

Predicting sales values of a company based on their past sales

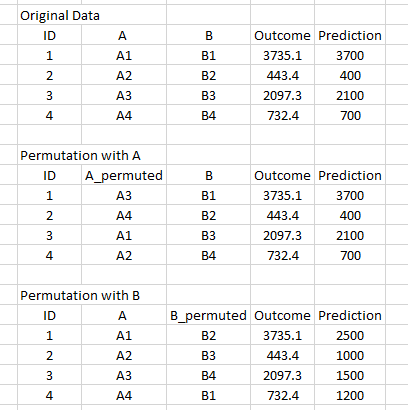
A and B

A, B, and C

Bottom of Form

Q 26) When using random forest for feature selection, suppose you permute values of two features - A and B. Permutation is such that you change the indices of individual values so that they do not remain associated with the same target as before.

For example:



You notice that permuting values does not affect the score of model built on A, whereas the score decreases on model trained on B.

Which of the following features would you select from the folowing solely based on the above finding?

A

B

Q 28) Which splitting algorithm is better with categorical variable having high cardinality?

Information Gain

Gain Ratio

Change in Variance

None of these

Q 29) There are "A" features in a dataset and a Random Forest model is built over it. It is given that there exists only one significant feature of the outcome - "Feature1" .

What would be the % of total splits that will not consider the "Feature1" as one of the features involved in that split (It is given that m is the number of maximum features for random forest)?

Note: Considering random forest select features space for every node split.

(A-m)/A

(m-A)/m

m/A

Cannot be determined

Q 32) While creating a Decision Tree, can we reuse a feature to split a node?

Yes

No

Q 33) Which of the following is a mandatory data pre-processing step(s) for XGBOOST?

Impute Missing Values

Remove Outliers

Convert data to numeric array / sparse matrix

Input variable must have normal distribution

Select the sample of records for each tree/ estimators

1 and 2

1, 2 and 3

3, 4 and 5

3

5

All of these

Q 34) Decision Trees are not affected by multicollinearity in features

TRUE

FALSE

Q 35) For parameter tuning in a boosting algorithm, which of the following search strategies may give best tuned model

Random Search.

Grid Search.

a or b

Can't say

Q 36) Imagine a two variable predictor space having 10 data points. A decision tree is built over it with 5 leaf nodes.

The number of distinct regions that will be formed in predictors space?

25

10

2

5

Q 37) In Random Forest, which of the following is randomly selected?

Number of decision trees

features to be taken into account when building a tree

samples to be given to train individual tree in a forest

B and C

A, B and C

Q 38) Which of the following are disadvantage of Decision Tree algorithm?

Decision tree is not easy to interpret

Decision tree is not a stable algorithm

Decision Tree will over fit if it perfectly memorizes it

Both B and C

Q 39) While tuning the parameters "Number of estimators" and "Shrinkage Parameter"/"Learning Rate" for boosting algorithm.

Which of the following relationship should be kept in mind?

Number of estimators is directly proportional to shrinkage paramter

Number of estimators is inversely proportional to shrinkage parameter

Both have polynomial relationship

Q 40) Let's say we have m number of estimators (trees) in a XGBOOST model.

Now, how many trees will work on bootstrapped data set?

1

m-1

m

Can't say

None of the above

Q 41) Which of the following statement is correct about XGBOOST parameters:

Learning rate can go upto 10

Sub Sampling / Row Sampling percentage should lie between 0 to 1

Number of trees / estimators can be 1

Max depth can not be greater than 10

1

1 and 3

1, 3 and 4

2 and 3

2

4

Q 42) What can be the maximum depth of decision tree (where k is the number of features and N is the number of samples)?

Our constraint is that we are considering a binary decision tree with no duplicate rows in sample (Splitting criterion is not fixed).

N

N - k - 1

N - 1

k - 1

Q 43) Boosting is a general approach that can be applied to many statistical learning methods for regression or classification. True or False?

TRUE

FALSE

Q 44) Predictions of individual trees of bagged decision trees have higher correlation in comparison to individual trees of random forest.

TRUE

FALSE

Q 45) Below is a list of parameters of Decision Tree. In which of the following cases higher is better?

Number of samples used for split

Depth of tree

Minimum samples for leaf

Can't Say

**Thank You for participating in Skilltest: Tree Based Algorithms**

**Your score is : 24.0**

**See** [**https://datahack.analyticsvidhya.com/contest/skilltest-tree-based-algorithms/lb**](https://datahack.analyticsvidhya.com/contest/skilltest-tree-based-algorithms/lb)

Bottom of Form