Ensemble ML Algorithms - Boosting Interview Questions

We believe that you have learned both theoritical and practical knowledge on Naive Bayes classification algorithm through your assignment.

So let's test your knowledge here. This will help you to be prepared for interviews too!

Best with Quest

1. What are the advantages and disadvantages of XGBoost?

```
Advantages:

XGB consists of a number of hyper-parameters that can be tuned – a primary advantage over gradient boosting machines.
XGBoost has an in-built capability to handle missing values.
It provides various intuitive features, such as parallelisation, distributed computing, cache optimisation, and more.
Disadvantages:

Like any other boosting method, XGB is sensitive to outliers.
Unlike LightGBM, in XGB, one has to manually create dummy variable/ label encoding for categorical features before feeding them into the models.
```

2. What is Gradient Boosting Machine (GBM)?

```
A Gradient Boosting Machine or GBM combines the predictions from multiple decision trees to generate the final predictions. Keep in mind that all the weak learners in a gradient boosting machine are decision trees.

But if we are using the same algorithm, then how is using a hundred decision trees better than using a single decision tree? How do different decision trees capture different signals/information from the data?

Here is the trick – the nodes in every decision tree take a different subset of features for selecting the best split. This means that the individual trees aren’t all the same and hence they are able to capture different signals from the data.

Additionally, each new tree takes into account the errors or mistakes made by the previous trees. So, every successive decision tree is built on the errors of the previous trees. This is how the trees in a gradient boosting machine algorithm are built sequentially.
```

3. How does AdaBoost algorithm works?

AdaBoost (Adaptive Boosting) : It works on similar method as discussed above. It fits a sequence of weak learners on different weighted training data. It starts by predicting original data set and gives equal weight to each observation. If prediction is incorrect using the first learner, then it gives higher weight to observation which have been predicted incorrectly. Being an iterative process, it continues to add learner(s) until a limit is reached in the number of models or accuracy.

Mostly, we use decision stamps with AdaBoost. But, we can use any machine learning algorithms as base learner if it accepts weight on training data set. We can use AdaBoost algorithms for both classification and regression problem.

4. What is LightGBM?

It is a gradient boosting framework that makes use of tree based learning algorithms that is considered to be a very powerful algorithm when it comes to computation. It is considered to be a fast processing algorithm.

While other algorithms trees grow horizontally, LightGBM algorithm grows vertically meaning it grows leaf-wise and other algorithms grow level-wise. LightGBM chooses the leaf with large loss to grow. It can lower down more loss than a level wise algorithm when growing the same leaf.

LightGBM is not for a small volume of datasets. It can easily overfit small data due to its sensitivity. It can be used for data having more than 10,000+ rows. There is no fixed threshold that helps in deciding the usage of LightGBM. It can be used for large volumes of data especially when one needs to achieve a high accuracy.

5. What is Catboost?

CatBoost algorithm is another member of the gradient boosting technique on decision trees. One of the many unique features that the CatBoost algorithm offers is the integration to work with diverse data types to solve a wide range of data problems faced by numerous businesses.

Not just that, but CatBoost also offers accuracy just like the other algorithm in the tree family.

The term CatBoost is an acronym that stands for "Category" and "Boosting." Does this mean the 'Category' in CatBoost means it only works for categorical features?

The answer is, "No."
According to the CatBoost documentation, CatBoost supports numerical, categorical, and text features but has a good handling technique for categorical data.

The CatBoost algorithm has quite a number of parameters to tune the features in the processing stage.

"Boosting" in CatBoost refers to the gradient boosting machine learning. Gradient boosting is a machine learning technique for regression and classification problems. Which produces a prediction model in an ensemble of weak prediction models, typically decision trees.

6. Explain the differences between Random Forest and Gradient Boosting machines.

Random forests are a significant number of decision trees pooled using averages or majority rules at the end. Gradient boosting machines also combine decision trees but at the beginning of the process unlike Random forests. Random forest creates each tree independent of the others while gradient boosting develops one tree at a time. Gradient boosting yields better outcomes than random forests if parameters are carefully tuned but it's not a good option if the data set contains a lot of outliers/anomalies/noise as it can result in overfitting of the model.Random forests perform well for multiclass object detection. Gradient Boosting performs well when there is data which is not balanced such as in real time risk assessment.

7. Why boosting is a more stable algorithm as compared to other ensemble algorithms? Model accuracy or Model performance? Which one will you prefer and why?

Boosting focuses on errors found in previous iterations until they become obsolete. Whereas in bagging there is no corrective loop. This is why boosting is a more stable algorithm compared to other ensemble algorithms.

This is a trick question, one should first get a clear idea, what is Model Performance? If Performance means speed, then it depends upon the nature of the application, any application related to the real-time scenario will need high speed as an important feature. Example: The best of Search Results will lose its virtue if the Query results do not appear fast.

If Performance is hinted at Why Accuracy is not the most important virtue - For any imbalanced data set, more than Accuracy, it will be an F1 score than will explain the business case and in case data is imbalanced, then Precision and Recall will be more important than rest.

8. How would you handle an imbalanced dataset?

Sampling Techniques can help with an imbalanced dataset. There are two ways to perform sampling, Under Sample or Over Sampling.

In Under Sampling, we reduce the size of the majority class to match minority class thus help by improving performance w.r.t storage and run-time execution, but it potentially discards useful information.

For Over Sampling, we upsample the Minority class and thus solve the problem of information loss, however, we get into the trouble of having Overfitting.

There are other techniques as well -
Cluster-Based Over Sampling - In this case, the K-means clustering algorithm is independently applied to minority and majority class instances. This is to identify clusters in the dataset. Subsequently, each cluster is oversampled such that all clusters of the same class have an equal number of instances and all classes have the same size

Synthetic Minority Over-sampling Technique (SMOTE) - A subset of data is taken from the minority class as an example and then new synthetic similar instances are created which are then added to the original dataset. This technique is good for Numerical data points.

9. Mention why feature engineering is important in model building and list out some of the techniques used for feature engineering.

Algorithms necessitate features with some specific characteristics to work appropriately. The data is initially in a raw form. You need to extract features from this data before supplying it to the algorithm. This process is called feature engineering. When you have relevant features, the complexity of the algorithms reduces. Then, even if a non-ideal algorithm is used, results come out to be accurate.

Feature engineering primarily has two goals:

Prepare the suitable input data set to be compatible with the machine learning algorithm constraints.
Enhance the performance of machine learning models.
Some of the techniques used for feature engineering include Imputation, Binning, Outliers Handling, Log transform, grouping operations, One-Hot encoding, Feature split, Scaling, Extracting date.

10. Differentiate between Boosting and Bagging?

Bagging and Boosting are variants of Ensemble Techniques.

Bootstrap Aggregation or bagging is a method that is used to reduce the variance for algorithms having very high variance. Decision trees are a particular family of classifiers which are susceptible to having high bias.

Decision trees have a lot of sensitiveness to the type of data they are trained on. Hence generalization of results is often much more complex to achieve in them despite very high fine-tuning. The results vary greatly if the training data is changed in decision trees.

Hence bagging is utilised where multiple decision trees are made which are trained on samples of the original data and the final result is the average of all these individual models.

Boosting is the process of using an n-weak classifier system for prediction such that every weak classifier compensates for the weaknesses of its classifiers. By weak classifier, we imply a classifier which performs poorly on a given data set.

It’s evident that boosting is not an algorithm rather it’s a process. Weak classifiers used are generally logistic regression, shallow decision trees etc.

There are many algorithms which make use of boosting processes but two of them are mainly used: Adaboost and Gradient Boosting and XGBoost.

11. What are hyperparameters and how are they different from parameters? What is a voting model?

A parameter is a variable that is internal to the model and whose value is estimated from the training data. They are often saved as part of the learned model. Examples include weights, biases etc.

A hyperparameter is a variable that is external to the model whose value cannot be estimated from the data. They are often used to estimate model parameters. The choice of parameters is sensitive to implementation. Examples include learning rate, hidden layers etc.

A voting model is an ensemble model which combines several classifiers but to produce the final result, in case of a classification-based model, takes into account, the classification of a certain data point of all the models and picks the most vouched/voted/generated option from all the given classes in the target column.

12. What do you mean by the ROC curve? What do you mean by AUC curve?

Receiver operating characteristics (ROC curve): ROC curve illustrates the diagnostic ability of a binary classifier. It is calculated/created by plotting True Positive against False Positive at various threshold settings. The performance metric of ROC curve is AUC (area under curve). Higher the area under the curve, better the prediction power of the model.

AUC (area under curve). Higher the area under the curve, better the prediction power of the model.

Now Rest with this Quest :)