1. What is difference between parametric and non parametric algorithms?

- A model parameter is a configuration variable that is internal to the model and whose value can be estimated from the given data.

They are required by the model when making predictions.

Their values define the skill of the model on your problem.

They are estimated or learned from historical training data.

They are often not set manually by the practitioner.

They are often saved as part of the learned model.

The examples of model parameters include:

The weights in an artificial neural network.

The support vectors in a support vector machine.

The coefficients in linear regression or logistic regression.

Nonparametric machine learning algorithms are those which do not make specific assumptions

about the type of the mapping function. They are prepared to choose any functional form from the training data, by not making assumptions.

Some more examples of popular nonparametric machine learning algorithms are:

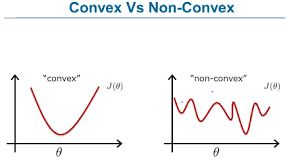
k-Nearest Neighbors

Decision Trees like CART and C4.5

Support Vector Machines

My observation is that pramametric makes assumptions and nonparametric does not.

2. Difference between convex and non convex cost functions.



A convex function: given any two points on the curve there will be no intersection with any other points, for non convex function there will be at least one intersection. In terms of cost function with a convex type you are always guaranteed to have a global minimum, whilst for a non convex only local minima.

3. How do you decide when to go for a deep learning project.

1. **Large amount of data:** Deep learning models require a large amount of data to learn from, and they usually perform better as the data increases. If you have access to a large dataset, deep learning might be a good option.
2. **Non-linear relationships:** Deep learning models excel in capturing non-linear relationships between the inputs and outputs. If you suspect that your problem involves complex non-linear relationships, deep learning might be a good choice.
3. **Complex tasks:** Deep learning models can handle complex tasks, such as image or speech recognition, natural language processing, and predictive modeling. If your project involves a complex task, deep learning might be the way to go.
4. **Sufficient computational resources:** Deep learning models require significant computational resources, including powerful hardware and time-consuming training processes. If you have access to sufficient computational resources, deep learning might be a viable option.
5. **Benchmarking with other models:** Finally, it's always a good idea to benchmark deep learning models against other models, such as traditional machine learning models, to see if they offer a significant performance improvement.

4. why is naïve bayes naïve.

 it makes the assumption that features of a measurement are independent of each other.

5. example where median is better than mean.

The mean is used for normal distributions. The median is generally used for skewed distributions. The mean is not a robust tool since it is largely influenced by outliers. **The median is better suited for skewed distributions to derive at central tendency since it is much more robust and sensible**.

1. Unreasonable effictiveness of data

The more data you have about things, the more things you can analyze. But, **the growth of data also has a deeper and more subtle impact**, which has become known as the “Unreasonable Effectiveness of Data”. Google made an experiment to make the machine learning model learn Arabic and Chinese.

1. Why KNN is lazy learning technique.

KNN is called a lazy learner because when we supply training data to this algorithm, the algorithm does not train itself at all.

Yes, that’s true!

KNN does not learn any discriminative function from the training data. But it memorizes the entire training dataset instead.

**There is no training time in KNN.**

But, this skipping of training time comes with a cost.

**Each time a new data point comes in and we want to make a prediction, the KNN algorithm will search for the nearest neighbors in the entire training set**.

1. Semi supervised machine learning.

**Examples of Semi-Supervised Learning**

* [**Text classification**](https://www.geeksforgeeks.org/sentiment-classification-using-bert/): In text classification, the goal is to classify a given text into one or more predefined categories. Semi-supervised learning can be used to train a text classification model using a small amount of labeled data and a large amount of unlabeled text data.
* [**Image classification**](https://www.geeksforgeeks.org/python-image-classification-using-keras/): In image classification, the goal is to classify a given image into one or more predefined categories. Semi-supervised learning can be used to train an image classification model using a small amount of labeled data and a large amount of unlabeled image data.
* [**Anomaly** **detection**](https://www.geeksforgeeks.org/machine-learning-for-anomaly-detection/): In anomaly detection, the goal is to detect patterns or observations that are unusual or different from the norm

1. OOB error.

What is the difference between OOB error and accuracy?

In a Random Forest, I know that the Out Of Bag Error is described as the fraction of number incorrect classifications over number of out of bag samples. Accuracy is defined as the number of correct classifications divided by the number of samples.

1. In what scenario decision tree is used than random forest.

Interpretable models: Decision trees can be easily visualized and understood, making them a good choice when interpretability is important.

Small datasets: Decision trees can perform well on small datasets, where the increased complexity of a random forest may lead to overfitting.

**Low computational resources**: Training a random forest can be computationally expensive, especially for large datasets. In situations where computational resources are limited, a decision tree may be a more practical choice.

**Highly imbalanced datasets**: Random forests can struggle with highly imbalanced datasets, where one class is much more prevalent than the others.

1. Why logistic regression called regression?

Logistic Regression is one of the basic and popular algorithms to solve a classification problem. It is named 'Logistic Regression' because its underlying technique is quite the same as Linear Regression. The term “Logistic” is taken from the Logit function that is used in this method of classification

**Yes, logistic regression is a regression algorithm** and it does predict a continuous outcome: the probability of an event. That we use it as a binary classifier is due to the interpretation of the outcome. Logistic regression is a type of generalize linear regression model.

1. What is online machine learning?

**Online learning means that you are doing it as the data comes in.** **Offline means that you have a static dataset**. So, for online learning, you (typically) have more data, but you have time constraints. Another wrinkle that can affect online learning is that your concepts might change through time.

1. What is no free lunch theorem?

The theorem states that all optimization algorithms perform equally well when their performance is averaged across all possible problems.

1. How will a laptop of 2gb ram process data of 10 gb?
2. **Sampling** = but information might get lost
3. **Using coud computation** = aws and gcp but it is costly
4. **Out of core ML** = 3.1 streaming the data 3.2 extracting the features 3.3 training the model

**You have to divide the data into chunks and perform feature extraction on them after that while training the data you have to use “incremental learning” model that will divide the data into batches while training. Example: SGD model.**

1. Difference between structured and unstructured data.

 structured data include **customer relationship management (CRM), invoicing systems, product databases, and contact lists**. Unstructured data includes various content such as documents, videos, audio files, posts on social media, and emails.

1. Difference b/w bagging and boosting.

**Bagging**: It is a homogeneous weak learners’ model that learns from each other independently in parallel and combines them for determining the model average.

**Boosting**: It is also a homogeneous weak learners’ model but works differently from Bagging. In this model, learners learn sequentially and adaptively to improve model predictions of a learning algorithm.

1. Assumptions of linear regression.

# 1,  **Independence of observations** = the dependent observations are time series, independent are experimental.

# 2**,  No Hidden or Missing Variables**

The second assumption of the linear regression model is that you have used **all relevant explanatory variables** in your model.

3, **Linear relationship** between dependent and independent variable.

4, Normality of the residuals = you can check by using Q-Q plot

5,  **No or little Multicollinearity** The fifth assumption of linear regression is that there is no or little multicollinearity. Multicollinearity is the phenomenon when a number of the explanatory variables are strongly correlated.

# 6**. Homoscedasticity**

The sixth assumption of linear regression is homoscedasticity. Homoscedasticity in a model means that the error is constant along the values of the dependent variable.

# 7.  All independent variables are uncorrelated with the error term

# 8. Observations of the error term are uncorrelated with each other

1. Assumptions of Logistic regression.

### **The Response Variable is Binary**

### **The Observations are Independent**

### **There is No Multicollinearity Among Explanatory Variables**

### **There are No Extreme Outliers**

### **There is a Linear Relationship Between Explanatory Variables and the Logit of the Response Variable**

1. How to measure accuracy of clustering algorithm.

#### **Silhouette Coefficient**

Silhouette Coefficient

#### **Dunn’s Index**

Dunn’s Index (DI) is another metric for evaluating a clustering algorithm. Dunn’s Index is equal to the minimum inter-cluster distance divided by the maximum cluster size. Note that large inter-cluster distances (better separation) and smaller cluster sizes (more compact clusters) lead to a higher DI value.

1. How to deal with imbalanced dataset?

# 1. Use the right evaluation metrics

Applying inappropriate evaluation metrics for model generated using imbalanced data can be dangerous. Imagine our training data is the one illustrated in graph above. If accuracy is used to measure the goodness of a model, a model which classifies all testing samples into “0” will have an excellent accuracy (99.8%), but obviously, this model won’t provide any valuable information for us.

In this case, other alternative evaluation metrics can be applied such as:

* Precision/Specificity: how many selected instances are relevant.
* Recall/Sensitivity: how many relevant instances are selected.
* F1 score: harmonic mean of precision and recall.
* MCC: correlation coefficient between the observed and predicted binary classifications.
* AUC: relation between true-positive rate and false positive rate.

# 2. Resample the training set

Apart from using different evaluation criteria, one can also work on getting different dataset. Two approaches to make a balanced dataset out of an imbalanced one are under-sampling and over-sampling.

## 2.1. Under-sampling

## 2.2. Over-sampling

**3. Use K-fold Cross-Validation in the Right Way**

**4. Ensemble Different Resampled Datasets**

 e.g. the 1.000 cases of the rare class and randomly sample 10.000 cases of the abundant class. Then you just split the 10.000 cases in 10 chunks and train 10 different models.

1. **Resample with Different Ratios:** if 10 models are trained, it might make sense to have a model that has a ratio of 1:1 (rare:abundant) and another one with 1:3, or even 2:1. Depending on the model used this can influence the weight that one class gets.
2. How do you measure the accuracy of a recommendation engine

### Metrics used in Explicit Recommender Systems

For such a system, the metrics used could be pretty simular to that used in a standard regression problem since the target is really a score that you could be predicting, and the actual score is available to measure how good the prediction is.

1. Mean Absolute Error: Mean over all data points, absolute value of difference between actual rating and predicted rating.
2. Root Mean Square Error: Square root of Mean over all data points, square of difference between the actual rating and predicted rating.
3. R2 score

### Metrics used in Implicit Recommender systems

Since the implicit feedback available is usually binary – in the form of “yes the item is relevant” or “no the item is not relevant”, metrics are simimar to that for classification algorithms.

* Precision: Of all recommended items, how many are relevant ?
* Recall : Of all relevant items, how many made it into the recommendations?
* Precision@K: Of the top K recommended items, how many are relevant?
* Recall@K: Of all the relevant items, how many made it into the top K recommendations.
* Mean Average Precision (MAP): If there are K items recommended, precision@k averaged from k=1…K averaged over all data points.
* Mean Average Recall (MAR): If there are K items recommended, recall@k averaged from k=1…K averaged over all data points.

1. What are some ways to make your model more robust to outliers
2. You can use a model that's resistant to outliers. Tree-based models are generally not affected by outliers, while regression-based models are. If you are performing a statistical test, try a non-parametric test instead of a parametric one.
3. You can use a robust error metric: Switching from mean squared error to mean absolute difference reduces the influence of outliers
4. Winsorize your data- Artificially cap your data at some threshold.
5. Transform your data - If your data has a very pronounced right tail, try a log transformation.
6. How can you measure the performance of a dimensionality reduction algorithm on your dataset?
7. Reconstruction error: This metric measures how well the original data can be reconstructed from the reduced-dimensional data. It can be calculated as the difference between the original data and the reconstructed data.
8. Variance retained: This metric measures the amount of variance in the original data that is retained in the reduced-dimensional data. It can be calculated as the ratio of the variance in the reduced-dimensional data to the variance in the original data.
9. Computation time: This metric measures the time required to perform the dimensionality reduction algorithm on the dataset.
10. Clustering accuracy: This metric measures how well the reduced-dimensional data can be clustered into meaningful groups. It can be calculated by comparing the clustering results of the original data to the clustering results of the reduced-dimensional data.
11. Visualization: Visualization is a qualitative metric that can be used to assess how well the reduced-dimensional data can be visualized in a lower-dimensional space, such as 2D or 3D.
12. What is Multicollinearity? How to detect it? List some techniques to overcome.

Multicollinearity (or collinearity) **occurs when one independent variable in a regression model is linearly correlated with another independent variable**. An example of this is if we used “Age” and “Number of Rings” in a regression model for predicting the weight of a tree.

1. Simply drop some of the correlated predictors.

2. We can try to standardize the predictors by subtracting their mean from each of the observations.

3. Do some linear transformation e.g., add/subtract 2 predictors to create a new bespoke predictor.

4. As an extension of the previous 2 points, another very popular technique is to perform Principal components analysis (PCA).

25. List some ways using which you can reduce overfitting in a model.

* Train with more data. With the increase in the training data, the crucial features to be extracted become prominent. ...
* Data augmentation. ...
* Addition of noise to the input data. ...
* Feature selection. ...
* Cross-validation. ...
* Simplify data. ...
* Regularization. ...
* Ensembling.

1. Types of bias

**The formal definition of bias is an inclination or prejudice for or against one person or group**

* **Prejudicial Bias:** Fundamentally, biases make their way into an application because those of us designing them carry these biases knowingly or unknowingly. application. Hence, this is perhaps the most complex and important source to correct.
* **Sampling Bias:** Another common source of bias is how we collect data to train our model. Intentionally or unintentionally,
* **Algorithm Bias:** The next step is choosing an algorithm that we’ll use to create the model to train.
* **Confirmation Bias:** Once we start to train our model and evaluate its predictions, we may tend to retain information that affirms our preconceived notions.

1. How do you approach a categorical feature with high cardinality

**converted to numerical values**.

1. Explain Pruning in Decision Trees and how it is done.

Pruning is a data compression technique in machine learning and search algorithms that reduces the size of decision trees by removing sections of the tree that are non-critical and redundant to classify instances.

How do you prune a tree in decision tree?

We can prune our decision tree by using information gain in both post-pruning and pre-pruning. In pre-pruning, we check whether information gain at a particular node is greater than minimum gain. In post-pruning, we prune the subtrees with the least information gain until we reach a desired number of leaves.

1. What is ROC-AUC curve? List some of it’s benefits.

AUC - ROC curve is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes.

 It tells how much the model is capable of distinguishing between classes.

1. What are kernels in SVM? Can you list some popular SVM kernels.

In Support Vector Machines (SVM), a kernel is a function that takes in two inputs, computes the dot product between them, and maps them to a higher-dimensional space. The transformed data in this higher-dimensional space can make it easier to separate the data into different classes using a linear classifier.

SVMs use kernels to transform the input data into a higher-dimensional feature space. The transformed data is then used to find the maximum margin hyperplane that separates the data into different classes. The kernel function allows the SVM to work in a high-dimensional feature space without explicitly computing the coordinates of the data in that space.

The most commonly used kernels in SVM are:

1. Linear kernel: This kernel performs a simple dot product between the input features.
2. Polynomial kernel: This kernel maps the input features to a higher-dimensional space using a polynomial function.
3. Radial basis function (RBF) kernel: This kernel maps the input features to an infinite-dimensional space using a Gaussian function.
4. Sigmoid kernel: This kernel maps the input features to a higher-dimensional space using a sigmoid function.

Why does L2 regularization give sparse coefficients?

L1 regularization penalizes the sum of absolute values of the weights, whereas L2 regularization penalizes the sum of squares of the weights. The L1 regularization solution is sparse. The L2 regularization solution is **non-sparse**.

List some ways using which you can improve a model’s performance

* Add More Data. Having more data is always a good idea. ...
* Treat Missing and Outlier Values. ...
* Feature Engineering. ...
* Feature Selection. ...
* Multiple Algorithms. ...
* Algorithm Tuning. ...
* Ensemble Methods. ...
* Cross Validation.

Can PCA be used to reduce the dimensionality of a highly nonlinear dataset?

PCA can be used to significantly reduce the dimensionality of most datasets, even if they are highly nonlinear because it can at least get rid of useless dimensions. However, if there are no useless dimensions, reducing dimensionality with PCA will lose too much information.

What’s the difference between probability and likelihood?

**If we flip the coin one time, the probability that it will land on heads is 0.5**. Now suppose we flip the coin 100 times and it only lands on heads 17 times. We would say that the likelihood that the coin is fair is quite low.

Once a dataset’s dimensionality has been reduced, is it possible to reverse the operation? If so, how? If not, why?

**No, dimensionality reduction is not reversible in general**. Because information is lost

Why do we always need the intercept term in a regression model??

We always need the intercept term in a regression model because it ensures that the regression line passes through the mean of the dependent variable. If we did not include an intercept term, the regression line would be forced to pass through the origin (0,0), and this may not be appropriate for many datasets. For example, if we are modeling the relationship between a person's height and weight, it is unlikely that someone with zero height would have zero weight.

1. When Your Dataset Is Suffering From High Variance, How Would You Handle It?
2. Collect more data: Adding more data to the dataset can help in reducing the variance as it provides a more representative sample of the population.
3. Feature selection: High variance may indicate that some features in the dataset are not relevant to the problem at hand. By removing these features, we can simplify the model and reduce variance.
4. Regularization: Regularization is a technique that adds a penalty term to the loss function, which discourages large weights in the model.
5. Cross-validation: Cross-validation is a technique used to evaluate the performance of the model on unseen data.
6. Ensemble methods: Ensemble methods combine multiple models to improve the overall performance.
7. Which Among These Is More Important Model Accuracy Or Model Performance?

Model accuracy is important because it can help gauge a model's ability to process, understand, and predict. If model accuracy is off, it's a pretty big warning sign that there might be some serious issues with your model.

What is active learning and where is it useful?

Active learning is a type of machine learning technique where a model learns from a small set of labeled data and then selects the most informative samples to label from a larger pool of unlabeled data.

Active learning is particularly useful in situations where obtaining labeled data is expensive or time-consuming

Some specific applications of active learning include:

1. Object detection: Active learning can be used to improve the accuracy of object detection models by selecting the most informative images to annotate.
2. Text classification: Active learning can be used to improve the accuracy of text classification models by selecting the most informative documents to label.

Why is Ridge Regression called Ridge?

Ridge regression adds a ridge parameter (k), of the identity matrix to the cross product matrix, forming a new matrix (X`X + kI). It's called ridge regression **because the diagonal of ones in the correlation matrix can be described as a ridge**.

1. State the differences between causality and correlation?

A correlation between variables, however, does not automatically mean that the change in one variable is the cause of the change in the values of the other variable. **Causation indicates that one event is the result of the occurrence of the other event**; i.e. there is a causal relationship between the two events.

Is it possible to speed up training of a bagging ensemble by distributing it across multiple servers?

Yes it is possible to speed up the training time of bagging by distributing them on multiple servers because **all of them are independent of each other**.

* If a Decision Tree is underfitting the training set, is it a good idea to try scaling the input features?

Decision Trees don't care whether or not the training data is scaled or centered; scaling the input features will just be a waste of time.

1. What is cross validation and it's types?

**a machine learning technique where the training data is split into two parts: A training set and a test set**.

1. How do we interpret weights in linear models?

In linear models, the weights represent the coefficients of the independent variables (also called features or predictors) in the linear equation. These weights determine the magnitude and direction of the effect of each independent variable on the dependent variable (also called the response or target variable).

For example, consider a simple linear regression model that predicts a person's salary based on their years of experience. In this case, the weight associated with the "years of experience" variable represents the amount by which the salary changes for each additional year of experience. If the weight is positive, it indicates that the salary tends to increase with more experience, while a negative weight indicates the opposite.

Which Gradient Descent algorithm (among those we discussed) will reach the vicinity of the optimal solution the fastest? Which will actually converge?

The convergence speed and the ability to reach the optimal solution depend on the problem's characteristics, such as the cost function's shape, the learning rate, and the initialization. However, among the three gradient descent algorithms that we typically discuss (batch gradient descent, stochastic gradient descent, and mini-batch gradient descent), stochastic gradient descent (SGD) tends to reach the vicinity of the optimal solution the fastest, while batch gradient descent is more likely to converge.

Why is it important to scale the inputs when using SVMs?

SVMs try to find the maximum margin hyperplane that separates the classes. The margin depends on the distances between the support vectors and the hyperplane. If the input features have different scales, some features will dominate the distance calculation, while others will be ignored,

What is OvR and OvO for multiclass classification and which machine learning algorithm supports this.

One-vs-Rest (OvR) strategy: Also known as One-vs-All, this strategy trains a binary classifier for each class, where the positive class is the target class, and the negative classes are all the other classes combined.

One-vs-One (OvO) strategy: This strategy trains a binary classifier for every pair of classes, where the positive class is the target class, and the negative class is the other class in the pair.

How will you do feature selection using Lasso Regression?

After fitting the Lasso model, we can identify the important features by examining the coefficients of the model. Features with non-zero coefficients are considered important, and features with zero coefficients are considered unimportant and can be removed from the model.

What is the difference between loss function and cost function?

**the loss function is to capture the difference between the actual and predicted values for a single record whereas cost functions aggregate the difference for the entire training dataset**.

What are the common ways to handle missing data in a dataset?

DELETION, IMPUTATION, PREDICTION,

What is the difference between standard scaler and minmax scaler? What you will do if there is a categorical variable?

* StandardScaler scales the data to have zero mean and unit variance, which means it transforms the data to have a mean of 0 and a standard deviation of 1. It is recommended when the data is normally distributed or when the algorithm used for modeling is not sensitive to the scale of the features.
* MinMaxScaler scales the data to a fixed range, typically [0, 1]. It is recommended when the data is not normally distributed or when the algorithm used for modeling requires the features to be on a similar scale.

When dealing with categorical variables, we need to encode them into a numerical form before preprocessing. One common method is one-hot encoding, which creates a new binary variable for each category in the original variable. After encoding, we can preprocess the numerical data using either StandardScaler or MinMaxScaler. We can then combine the encoded categorical variables and the preprocessed numerical variables to create the final dataset for modeling.

1. What types of model tend to overfit?
2. High Complexity Models: Models with high complexity, such as deep neural networks and decision trees with many nodes.
3. Small Datasets: Models trained on small datasets tend to overfit more than models trained on large datasets.
4. Models with Many Features: Models with many features, such as linear models with many coefficients or SVMs with many support vectors,
5. Ensemble Models: Ensemble models, such as bagging, boosting, and stacking, can overfit if the individual models in the ensemble are overfitting.
6. Can you tell the complete life cycle of a data science project?
7. Problem definition: The first step in any data science project is to clearly define the problem you are trying to solve. This involves understanding the business objectives, defining the scope of the project, and identifying the key stakeholders.
8. Data collection: Once you have defined the problem, the next step is to collect the data you will need to solve it. This may involve obtaining data from internal databases, external sources, or collecting new data through surveys or experiments.
9. Data cleaning and preparation: After collecting the data, it's important to clean and prepare it for analysis. This involves tasks such as removing duplicates, handling missing values, and transforming the data into a suitable format.
10. Exploratory data analysis: With the data cleaned and prepared, the next step is to explore the data and gain insights that will inform the modeling process. This may involve visualizing the data, calculating summary statistics, and identifying patterns or anomalies.
11. Feature engineering: Feature engineering involves selecting and transforming the variables in the dataset to create new features that may be more predictive of the target variable. This may involve tasks such as scaling, normalization, or encoding categorical variables.
12. Model selection and training: With the data prepared and the features engineered, the next step is to select a suitable model and train it on the data. This may involve trying different algorithms, tuning hyperparameters, and evaluating the performance of the model on the training data.
13. Model evaluation:
14. Model deployment:
15. Model monitoring and maintenance:

What are the different evaluation metrices for a regression model?

1. Mean Absolute Error (MAE): MAE is the average absolute difference between the predicted and actual values. It measures the average magnitude of errors in the predictions without considering their direction.
2. Mean Squared Error (MSE): MSE is the average squared difference between the predicted and actual values. It measures the average of the squared differences between the predicted and actual values, and is more sensitive to large errors than MAE.
3. Root Mean Squared Error (RMSE): RMSE is the square root of the MSE, which gives the estimate of the standard deviation of the prediction errors.
4. R-squared (R2): R-squared is a statistical measure that represents the proportion of the variance in the dependent variable that is explained by the independent variables. It ranges from 0 to 1, with higher values indicating a better fit of the model to the data.
5. Adjusted R-squared: Adjusted R-squared is a modified version of R-squared that adjusts for the number of independent variables in the model. It penalizes the inclusion of unnecessary variables in the model and is a better measure of model fit when comparing models with different numbers of variables.
6. Explained Variance Score: Explained Variance Score is a metric that measures the proportion of the variance in the target variable that is explained by the model.
7. Mean Absolute Percentage Error (MAPE): MAPE is a percentage-based metric that measures the average absolute percentage difference between the predicted and actual values. It is often used in forecasting models.

When to use rsquare and mse?

If the dataset contains outliers or extreme values that might disproportionately affect the model's performance, you may prefer R-squared, which is less sensitive to outliers. MSE, on the other hand, is sensitive to outliers because it squares the differences between predicted and observed values.

What are the different evaluation metrices for a classification model?

List some of the drawbacks of a Linear model

1. Limited Complexity: Linear models assume that the relationship between the input variables and the output variable is linear. This assumption can limit the ability of the model to capture complex relationships and patterns in the data.
2. Sensitivity to Outliers: Linear models are sensitive to outliers, which can have a disproportionate impact on the model's coefficients and predictions. Outliers can distort the linearity of the relationship between the input variables and the output variable, leading to poor model performance.
3. Assumption of Normality: Linear models assume that the error terms of the model are normally distributed. This assumption may not hold in some cases, which can affect the accuracy and reliability of the model.
4. Multicollinearity: Linear models can suffer from multicollinearity, which occurs when two or more input variables are highly correlated. This can lead to unstable and unreliable model coefficients and make it difficult to interpret the model.
5. Lack of Robustness: Linear models can be sensitive to changes in the data, and small changes in the input variables can lead to large changes in the model coefficients and predictions. This lack of robustness can make linear models less suitable for certain types of data and situations.

     What do you mean by Bias variance tradeoff?

The bias-variance tradeoff arises because decreasing the bias of a model usually increases its variance

     Explain Kernel trick in SVM

**allows us to project data from a training set which isn't linearly separable into a higher dimensional space where it becomes linearly separable**.

 What is the main difference between Machine Learning and Data Mining?

Data mining is used on an existing dataset (like a data warehouse) to find patterns. Machine learning, on the other hand, is trained on a 'training' data set, which teaches the computer how to make sense of data, and then to make predictions about new data sets.

Why sometimes it is needed to scale or normalise features?

Why do we need to scale features?

**Without scaling features, the algorithm may be biased toward the feature with values higher in magnitude**. Hence we scale features that bring every feature in the same range, and the model uses every feature wisely.

1. What is the difference between Type 1 and Type 2 error?

Type I error (false positive): the test result says you have coronavirus, but you actually don't. Type II error (false negative): the test result says you don't have coronavirus, but you actually do.

1. What is the difference between a Generative model vs a Discriminative model?

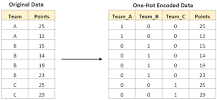
A discriminative model can estimate the probability that an instance belongs to a class. A generative model can estimate the probability of the instance, and also the probability of a class label.

### **Why binary\_crossentropy and categorical\_crossentropy give different performances for the same problem?**

binary\_crossentropy: Used as a loss function for binary classification model. The binary\_crossentropy function computes the cross-entropy loss between true labels and predicted labels. categorical\_crossentropy: Used as a loss function for multi-class classification model where there are two or more output labels

[Why does one hot encoding improve machine learning performance?](https://stackoverflow.com/questions/17469835/why-does-one-hot-encoding-improve-machine-learning-performance)

Why one-hot encoding is better than label encoding?



In most scenarios, one hot encoding is the preferred way to convert a categorical variable into a numeric variable because **label encoding makes it seem that there is a ranking between values**

Differentiate between wide and tall data formats?

 A dataset can be written in two different formats: wide and long. A wide format contains values that do not repeat in the first column. A long format contains values that do repeat in the first column. Notice that in the wide dataset, each value in the first column is unique.

What is the difference between Covariance and Correlation

Covariance indicates the direction of the linear relationship between variables. Correlation measures both the strength and direction of the linear relationship between two variables.

1. How will you find the correlation between a categorical variable and a continuous variable ?

To find the correlation between a categorical variable and a continuous variable, you can use methods such as ANOVA (Analysis of Variance) or chi-square tests. The choice of method depends on the number of categories in the categorical variable and the distribution of the continuous variable.

Is more data always better?

More data is not always better in machine learning. In general, having more data can help to improve the performance of machine learning models, as it allows the models to learn from a larger and more diverse set of examples. However, there are cases where having more data may not lead to better performance, or may even lead to worse performance

#### **How can you determine which features are the most im- portant in your model?**

1. Feature Importance from a Trained Model: Some machine learning models provide feature importance scores directly. For example, decision trees and random forests can provide feature importance scores based on how often a feature is used in the model to split the data. Linear models like linear regression and logistic regression can provide feature importance scores based on the magnitude and sign of the feature coefficients.
2. Permutation Importance: In this method, the feature importance is determined by shuffling the values of one feature at a time, re-fitting the model, and measuring the change in the model's performance. The idea is that if a feature is important, then shuffling its values should significantly reduce the model's performance.
3. Recursive Feature Elimination: This method involves repeatedly training a model on subsets of features and removing the least important feature in each iteration until the desired number of features is reached. The importance of a feature is determined by how much the model's performance drops when that feature is removed.
4. Correlation Analysis: This method involves analyzing the correlation between each feature and the target variable. If a feature has a high correlation with the target variable, it is likely to be important.
5. Domain Expertise:

Which hyper-parameter tuning strategies (in general) do you know

   
Grid Search: This involves creating a grid of possible hyperparameter values and trying all possible combinations of these values to find the combination that results in the best performance.

 Random Search: This involves randomly sampling hyperparameter values from predefined distributions and trying these combinations to find the best performing combination.

 Bayesian Optimization: This is a sequential model-based optimization technique that uses Bayesian inference to select hyperparameters that are likely to perform well based on past evaluations. It involves building a probabilistic model of the objective function and using this to guide the search for optimal hyperparameters.

 Gradient-based Optimization: This involves optimizing hyperparameters using gradient descent. This approach can be computationally expensive as it requires computing the gradients of the objective function with respect to the hyperparameters.

 Evolutionary Algorithms: This involves using a population-based approach to search for optimal hyperparameters. Hyperparameters are treated as individuals in a population and are evaluated based on their fitness, with the fittest individuals being selected for the next generation. This process continues until convergence is reached.

 Ensemble-based Optimization: This involves combining the predictions of multiple models, each with different hyperparameters,

# Describe the differences between and use cases for box plots and histograms

Histograms are preferred to determine the underlying probability distribution of a data. Box plots on the other hand are more useful when comparing between several data sets.

#### **How would you differentiate between Multilabel and MultiClass classification?**

In multi-class classification, each input will have only one output class, but in multi-label classification, each input can have multi-output classes.

Normalization is a technique used to rescale a dataset so that all values are between 0 and 1. It involves subtracting the minimum value in the dataset from each data point and then dividing by the range of the dataset. Normalization is useful when the range of the data is not important, but the relative relationships between the values are.

Standardization, on the other hand, is a technique used to transform a dataset so that it has a mean of 0 and a standard deviation of 1. It involves subtracting the mean from each data point and then dividing by the standard deviation. Standardization is useful when the scale of the data is important, and you want to compare different features that have different units.

There are several popular distribution curves that are commonly used in data science and statistics. Here are some of the most commonly used distributions along with scenarios where you might use them in an algorithm:

1. Normal Distribution: This is a bell-shaped curve that is symmetric around the mean. It is commonly used in algorithms that involve continuous variables that have a natural clustering around a central value, such as the heights of people, weights of objects, and test scores.
2. Binomial Distribution: This is a discrete distribution that models the number of successes in a fixed number of trials. It is commonly used in algorithms that involve binary outcomes, such as the probability of a coin flip or the success rate of a drug trial.
3. Poisson Distribution: This is a discrete distribution that models the number of occurrences of an event in a fixed interval of time or space. It is commonly used in algorithms that involve count data, such as the number of customer complaints, accidents, or arrivals at a store.
4. Exponential Distribution: This is a continuous distribution that models the time between events in a Poisson process. It is commonly used in algorithms that involve waiting times, such as the time between customer arrivals or the time between failures of a machine.
5. Gamma Distribution: This is a continuous distribution that generalizes the exponential distribution and models the sum of exponential random variables. It is commonly used in algorithms that involve waiting times, such as the time between events that follow a Poisson process, or the duration of a project that is composed of multiple tasks.
6. Content-Based Filtering: This type of recommendation system recommends items to users based on their preferences for certain attributes or features of the items. For example, if a user likes action movies, a content-based filtering system would recommend other action movies.
7. Collaborative Filtering: This type of recommendation system recommends items to users based on the preferences of other users with similar tastes. Collaborative filtering can be further divided into two sub-types
8. Collaborative Filtering: This type of recommendation system recommends items to users based on the preferences of other users with similar tastes. Collaborative filtering can be further divided into two sub-types:

a. User-based Collaborative Filtering: In this approach, the system recommends items to a user based on the preferences of other users who are similar to them.

b. Item-based Collaborative Filtering: In this approach, the system recommends items to a user based on the preferences of other users who have liked similar items.

Which metrics can be used to measure correlation of categorical data?

1. Chi-squared test: This is a statistical test that measures the independence between two categorical variables. It calculates the difference between the observed and expected frequencies of the categories and produces a test statistic that can be used to determine whether the two variables are related.

Types of sampling for categorical data

How to handle multicollinearity in linear regression

Multicollinearity occurs when there is a high correlation between two or more predictor variables in a linear regression model. This can lead to unstable and inaccurate coefficient estimates, which can affect the interpretation and prediction power of the model. Here are some ways to handle multicollinearity in linear regression:

1. Remove one of the correlated variables: If two or more variables are highly correlated, it may be possible to remove one of them from the model without affecting the model's overall performance.
2. Use principal component analysis (PCA): PCA is a technique that reduces the dimensionality of the data by transforming the correlated variables into a new set of uncorrelated variables called principal components. The principal components can then be used as predictors in the regression model.
3. Ridge regression: Ridge regression is a technique that adds a penalty term to the regression coefficients to shrink them towards zero, reducing the impact of the correlated variables.
4. Lasso regression: Lasso regression is similar to ridge regression, but it adds a penalty term that can shrink the coefficients to exactly zero, effectively removing some of the correlated variables from the model.
5. Collect more data: Collecting more data can help reduce the effect of multicollinearity by increasing the sample size and providing more information about the relationships between the variables.
6. Use a different regression model: If none of the above approaches are effective, it may be necessary to use a different regression model, such as a nonlinear regression model or a generalized linear model.

**Linear regression has two methods:**Top of Form

* 1. OLS = minimizing the difference between actual and predicted
  2. Gradient decent =global minima

**How PCA works?**

1. Standardize the range of continuous initial variables
2. Compute the covariance matrix to identify correlations
3. Compute the eigenvectors and eigenvalues of the covariance matrix to identify the principal components
4. Create a feature vector to decide which principal components to keep
5. Recast the data along the principal components axes

**ARIMA and SARIMAX**

ARIMA (AutoRegressive Integrated Moving Average) and SARIMAX (Seasonal AutoRegressive Integrated Moving Average with eXogenous variables) are statistical models used for time series forecasting.

ARIMA is a time series model that captures the linear relationship between an observation and a certain number of lagged observations, as well as the differences between observations to make the time series data stationary. It consists of three components:

1. Autoregressive (AR): This component models the relationship between an observation and a certain number of lagged observations (i.e., its own past values). It is denoted by AR(p), where 'p' represents the order of autoregression.
2. Moving Average (MA): This component models the dependency between an observation and a residual error from a moving average model applied to lagged observations. It is denoted by MA(q), where 'q' represents the order of moving average.
3. Integrated (I): This component models the differencing needed to make the time series data stationary. It is denoted by I(d), where 'd' represents the order of differencing.

SARIMAX is an extension of ARIMA that incorporates seasonality and exogenous variables. Seasonality refers to patterns that repeat at fixed intervals, such as daily, weekly, or yearly patterns. Exogenous variables are external variables that may influence the time series, but are not influenced by the time series itself. SARIMAX includes additional components compared to ARIMA:

1. Seasonal Autoregressive (SAR): This component models the relationship between an observation and a certain number of lagged observations at the seasonal frequency. It is denoted by SAR(P), where 'P' represents the order of seasonal autoregression.
2. Seasonal Moving Average (SMA): This component models the dependency between an observation and a residual error from a seasonal moving average model. It is denoted by SMA(Q), where 'Q' represents the order of seasonal moving average.
3. Exogenous (X): This component models the relationship between an observation and exogenous variables. It is denoted by X, and exogenous variables are included as additional inputs in the model.

ARIMA and SARIMAX models are widely used for time series forecasting in various domains, such as finance, economics, and demand forecasting, to make predictions based on historical data and potentially external factors.

**Rolling mean and dickey fuller method:**

1. Rolling Mean: The rolling mean, also known as moving average, is a technique used to smooth out time series data by calculating the average of a fixed-size window of consecutive data points. This is done to identify trends or patterns in the data by reducing the noise or randomness present in the raw data. Rolling mean can be calculated for different window sizes, such as 5, 10, or 20 data points, depending on the data frequency and the level of smoothness desired. Rolling mean is often used to visualize the trend component of a time series data, which can be useful in identifying whether the data has a clear upward or downward trend.
2. Dickey-Fuller Method: The Dickey-Fuller method, specifically the Dickey-Fuller test, is a statistical test used to determine if a time series data is stationary or not. If the test statistic is smaller than the critical value, it indicates that the data is stationary, implying that it does not have a trend or seasonal component.

In ARIMA modeling, these techniques can be used as part of the model-building process to identify the appropriate values for the ARIMA parameters (p, d, q), where:

* p: The order of the AutoRegressive (AR) component, which models the relationship between an observation and a fixed number of lagged observations.
* d: The order of differencing needed to make the time series stationary, which can be determined using the Dickey-Fuller method or other techniques.
* q: The order of the Moving Average (MA) component, which models the dependency between an observation and a residual error from a moving average model applied to lagged observations.

**If our data is non stationary we do the shifting**

### 51. How will you combine different pandas dataframes?

The dataframes can be combines using the below approaches:

* **append() method**: This is used to stack the dataframes horizontally. Syntax:

df1.append(df2)

* **concat() method:**This is used to stack dataframes vertically. This is best used when the dataframes have the same columns and similar fields. Syntax:

pd.concat([df1, df2])

* **join() method:**This is used for extracting data from various dataframes having one or more common columns.

**Metrics for time series evaluation**

There are several metrics that can be used to evaluate time series forecasting models, depending on the specific requirements and characteristics of the data. Here are some commonly used metrics:

1. Mean Absolute Error (MAE): This is the average of the absolute differences between the predicted values and the actual values. It gives an idea of the magnitude of the forecasting errors without considering their direction. Lower MAE values indicate better performance.
2. Root Mean Squared Error (RMSE): This is the square root of the average of the squared differences between the predicted values and the actual values. RMSE penalizes larger errors more heavily compared to MAE, as it involves taking the square root. Lower RMSE values indicate better performance.
3. Mean Absolute Percentage Error (MAPE): This is the average of the percentage differences between the predicted values and the actual values, expressed as a percentage. MAPE is useful when you want to evaluate the forecasting accuracy in relative terms, as it takes into account the scale of the data. Lower MAPE values indicate better performance.
4. R-squared (R2): This is a measure of how well the predicted values explain the variance in the actual values. It ranges from 0 to 1, with higher values indicating better performance. R2 can be used to assess the proportion of the variance in the data that is explained by the forecasting model.
5. Forecast Bias: This is the average difference between the predicted values and the actual values. A forecast bias close to zero indicates that the model is not systematically over- or under-estimating the actual values. However, a significant forecast bias can indicate a systematic error in the model.
6. Forecast Horizon Accuracy: This is a measure of how accurate the model's forecasts are at different time horizons. It can help assess the stability and consistency of the model's performance over time.
7. Forecast Interval Coverage: This is the proportion of actual values that fall within the forecast intervals generated by the model. Forecast intervals provide a measure of uncertainty around the point forecasts. Higher coverage indicates better calibration of the forecast intervals.