1. In the sense of machine learning, what is a model? What is the best way to train a model?

Answer= To train a machine learning model, one needs a huge amount of pre-processed data. Here pre-processed data means data in structured form with reduced null values, etc. If we do not provide pre-processed data, then there are huge chances that our model may perform terribly.A machine learning model is a file that has been trained to recognize certain types of patterns. You train a model over a set of data, providing it an algorithm that it can use to reason over and learn from those data

3 steps to training a machine learning model

Step 1: Begin with existing data. Machine learning requires us to have existing data—not the data our application will use when we run it, but data to learn from. ...

Step 2: Analyze data to identify patterns. ...

Step 3: Make predictions.

2. In the sense of machine learning, explain the "No Free Lunch" theorem.

Answer: All of this theory is great, but what does “no free lunch” mean for you as a data scientist, a machine learning engineer, or someone who just wants to get started with machine learning?

Does it mean that all algorithms are equal? No, of course not. In practice, all algorithms are not created equal. This is because the entire set of machine learning problems is a theoretical concept in the NFL theorem and it is much larger than the set of practical machine learning problems that we will actually attempt to solve. Some algorithms may generally perform better than others on certain types of problems, but every algorithm has disadvantages and advantages due to the prior assumptions that come with that algorithm.In computational complexity and optimization the no free lunch theorem is a result that states that for certain types of mathematical problems, the computational cost of ...oblems in the class, is the same for any solution method. The name alludes to the saying "there ain't no ...method offers a "short cut". This is under the assumption that the search space is a probability density function. It does not apply to the case where the search space has underlying structure

Creating a copy...

3.Describe the K-fold cross-validation mechanism in detail.

Answer=Split the dataset into the number of k folds. Start off with using your k-1 fold as the test dataset and the remaining folds as the training dataset. Train the model on the training dataset and validate it on the test dataset. Save the validation score.Cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample. The procedure has a single parameter called k that refers to the number of groups that a given data sample is to be split into. As such, the procedure is often called k-fold cross-validation.To achieve this K-Fold Cross Validation, we have to split the data set into three sets, Training, Testing, and Validation, with the challenge of the volume of the data. Here Test and Train data set will support building model and hyperparameter assessments.

4. Describe the bootstrap sampling method. What is the aim of it?

Answer: The bootstrap method is a statistical technique for estimating quantities about a population by averaging estimates from multiple small data samples. Importantly, samples are constructed by drawing observations from a large data sample one at a time and returning them to the data sample after they have been chosen.It allows entrepreneurs to retain full ownership of their business. When investors support a business, they do so in exchange for a percentage of ownership. Bootstrapping enables startup owners to retain their share of the equity. It forces business owners to create a model that really works.

```
5. What is the significance of calculating the Kappa value for a classification model? Demonstrate
how to measure the Kappa value of a classification model using a sample collection of results.
```

5. What is the significance of calculating the Kappa value for a classification model?

Demonstratehow to measure the Kappa value of a classification model using a sample collection of results bold text.

Answer= The kappa score is an interesting metric. Its origins are in the field of psychology: it is used for measuring the agreement between two human evaluators or raters (e.g., psychologists) when rating subjects (patients). It was later “appropriated” by the machine-learning community to measure classification performance.It basically tells you how much better your classifier is performing over the performance of a classifier that simply guesses at random according to the frequency of each class. Cohen's kappa is always less than or equal to 1. Values of 0 or less, indicate that the classifier is useless.

The kappa statistic is frequently used to test interrater reliability. The importance of rater reliability lies in the fact that it represents the extent to which the data collected in the study are correct representations of the variables measured.

Cohen's kappa is a metric often used to assess the agreement between two raters. It can also be used to assess the performance of a classification model.

6. Describe the model ensemble method. In machine learning, what part does it play?

Answer= Ensemble methods are techniques that create multiple models and then combine them to produce improved results. Ensemble methods usually produces more accurate solutions than a single model would. This has been the case in a number of machine learning competitions, where the winning solutions used ensemble methods. Ensemble methods is a machine learning technique that combines several base models in order to produce one optimal predictive model . To better understand this definition lets take a step back into ultimate goal of machine learning and model building.

Ensemble modeling is a process where multiple diverse models are created to predict an outcome, either by using many different modeling algorithms or using different training data sets. The ensemble model then aggregates the prediction of each base model and results in once final prediction for the unseen data.

7. What is a descriptive model's main purpose? Give examples of real-world problems that descriptive models were used to solve.

Answer:A descriptive model describes a system or other entity and its relationship to its environment. It is generally used to help specify and/or understand what the system is, what it does, and how it does it. A geometric model or spatial model is a descriptive model that represents geometric and/or spatial relationships. Descriptive modeling is a mathematical process that describes real-world events and the relationships between factors responsible for them. The process is used by consumer-driven organizations to help them target their marketing and advertising efforts.

In descriptive modeling, customer groups are clustered according to demographics, purchasing behavior, expressed interests and other descriptive factors. Statistics can identify where the customer groups share similarities and where they differ. The most active customers get special attention because they offer the greatest ROI (return on investment).

The main aspects of descriptive modeling include:

Customer segmentation: Partitions a customer base into groups with various impacts on marketing and service. Value-based segmentation: Identifies and quantifies the value of a customer to the organization. Behavior-based segmentation: Analyzes customer product usage and purchasing patterns.

Needs-based segmentation: Identifies ways to capitalize on motives that drive customer behavior. Descriptive modeling can help an organization to understand its customers, but predictive modeling is necessary to facilitate the desired outcomes. Both descriptive and predictive modeling constitute key elements of data mining and Web mining.

8. Describe how to evaluate a linear regression model.

Answer=Linear Regression Analysis consists of more than just fitting a linear line through a cloud of data points. It consists of 3 stages – (1) analyzing the correlation and directionality of the data, (2) estimating the model, i.e., fitting the line, and (3) evaluating the validity and usefulness of the model. Mathematically, the RMSE is the square root of the mean squared error (MSE), which is the average squared difference between the observed actual outome values and the values predicted by the model. So, MSE = mean((observeds - predicteds)^2) and RMSE = sqrt(MSE). The lower the RMSE, the better the model There are 3 main metrics for model evaluation in regression:

R Square/Adjusted R Square.

Mean Square Error(MSE)/Root Mean Square Error(RMSE)

Mean Absolute Error(MAE)

There are two methods of evaluating models in data science, Hold-Out and Cross-Validation. To avoid overfitting, both methods use a test set (not seen by the model) to evaluate model performance

9. Distinguish :

1. Descriptive vs. predictive models

2. Underfitting vs. overfitting the model

3. Bootstrapping vs. cross-validation

Answer (a)--Descriptive mining is generally used to support correlation, cross-tabulation, frequency, etc. While descriptive analytics are used by companies to understand what has happened.The term 'Predictive' defines to predict something, so predictive data mining is the analysis done to predict the future event or multiple data or trends. It defines the features of the data in a target data set.Two of the most widely used predictive modeling techniques are regression and neural networks.

(b)--->A model that is underfit will have high training and high testing error while an overfit model will have extremely low training error but a high testing error.Overfitting is a modeling error which occurs when a function is too closely fit to a limited set of data points.Here we will discuss possible options to prevent overfitting, which helps improve the model performance. Train with more data. ... Data augmentation. ... Addition of ...tion. ... Cross-validation. ... Simplify data. ... Regularization. ... Ensembling.

Creating a copy...

Underfitting refers to a model that can neither model the training data nor generalize to new data.Increase model complexity. Increase the number of features, performing feature engineering. Remove noise from the data. Increase the number of epochs or increase the duration of training to get better results

(c)--->In summary, Bootstrapping is a statistical procedure that resamples a single dataset to create many simulated samples. This process allows you to calculate standard errors, construct confidence intervals, and perform hypothesis testing for numerous types of sample statistics.In essence, bootstrapping is random sampling with replacement from the available training data. Bagging (= bootstrap aggregation) is performing it many times and training an estimator for each bootstrapped dataset. It is available in modAL for both the base ActiveLearner model and the Committee model as well. Cross validation splits the available dataset to create multiple datasets, and Bootstrapping method uses the original dataset to create multiple datasets after resampling with replacement. Bootstrapping it is not as strong as Cross validation when it is used for model validation.Cross-validation is a model validation technique. So cross-validation and bagging solve different problems. You cannot replace cross-validation with bagging as it would not have any sense.

10. Make quick notes on:

1. LOOCV.

2. F-measurement

3. The width of the silhouette

Answer:1. LOOCV(Leave One Out Cross-Validation) is a type of cross-validation approach in which each observation is considered as the validation set and the rest (N-1) observations are considered as the training set. In LOOCV, fitting of the model is done and predicting using one

observation validation set.eave-one-out cross-validation, or LOOCV, is a configuration of k-fold cross-validation where k is set to the number of examples in the dataset. LOOCV is an extreme version of k-fold cross-validation that has the maximum computational cost

2.F-measurement...The F-measure is calculated as the harmonic mean of precision and recall, giving each the same weighting. It allows a model to be evaluated taking both the precision and recall into account using a single score, which is helpful when describing the performance of the model and in comparing models

An F-score is the harmonic mean of a system's precision and recall values. It can be calculated by the following formula: 2 x [(Precision x Recall) / (Precision + Recall)].

Precision and Recall are two measure that can be interpreted as percentages. Their arithmetic mean would be a percentage also. F1 score is actually the harmonic mean of the two; analogously it's still a percentage.