Company Name - Hotstar
Role - Data Analyst

Q. What is the statistical power of sensitivity?

A. The statistical power of an A/B test refers to the test's sensitivity to certain magnitudes of effect sizes. More precisely, it is the probability of observing a statistically significant result at level alpha (α) if a true effect of a certain magnitude (MEI) is in fact present.

Q. What is the difference between covariance and correlation?

A. "Covariance" indicates the direction of the linear relationship between variables. "Correlation" on the other hand measures both the strength and direction of the linear relationship between two variables.

Q. What is negative indexing? Why is it needed? Can you give an example for the same in python?

A. This means that the index value of -1 gives the last element, and -2 gives the second last element of an array. The negative indexing starts from where the array ends.
example: for list L = [0,2,35,3]; L[-1] will print 3 in Python.

Q. What is the condition for using a t-test or a z-test?

A. z-test is used for it when sample size is large, generally n >30. Whereas t-test is used for hypothesis testing when sample size is small, usually n < 30 where n is used to quantify the sample size.

Q. What is the main difference between overfitting and underfitting?

A. Overfitting is a modeling error which occurs when a function is too closely fit to a limited set of data points. Underfitting refers to a model that can neither model the training data nor generalize to new data.

Q. What is the KNN imputation method?

A. The idea in kNN methods is to identify 'k' samples in the dataset that are similar or close in the space. Then we use these 'k' samples to estimate the value of the missing data points. Each sample's missing values are imputed using the mean value of the 'k'-neighbors found in the dataset.

Q. GMM vs K means?

A. Gaussian mixture models (GMMs) are often used for data clustering. You can use GMMs to perform either hard clustering or soft clustering on query data. To perform hard clustering, the GMM assigns query data points to the multivariate normal components that maximize the component posterior probability, given the data. K-means clustering aims to partition data into k clusters in a way that data points in the same cluster are similar and data points in the different clusters are farther apart. Similarity of two points is determined by the distance between them.
Officially, k-means is one application of Vector-Quantification (VQ), and GMM is of Expectation-Maximize (EM) algorithm. But in my opinion, both k-means and GMM can be seen as a version of with different possibility density distribution: k-means uses uniform distribution while GMM uses gaussian.

Q. Regression imputation?

A. Regression imputation fits a statistical model on a variable with missing values. Predictions of this regression model are used to substitute the missing values in this variable.

Q. Z score for outliers' treatment?

A. Z score test is one of the most commonly used methods to detect outliers. It measures the number of standard deviations away the observation is from the mean value. A z score of 1.5 indicated that the observation is 1.5 standard deviations above the mean and -1.5 means that the observation is 1.5 standard deviations below or less than the mean.

Q. Forward and Backward selection? It's working?

A. Forward Selection chooses a subset of the predictor variables for the final model. Unlike forward stepwise selection, it begins with the full least squares model containing all p predictors, and then iteratively removes the least useful predictor, one-at-a-time. forward selection starts with a null model (with no predictors) and proceeds to add variables one at a time, and so unlike backward selection, it DOES NOT have to consider the full model (which includes all the predictors).

Q. Statistical power of sensitivity?

A. The statistical power of an A/B test refers to the test's sensitivity to certain magnitudes of effect sizes. More precisely, it is the probability of observing a statistically significant result at level alpha (α) if a true effect of a certain magnitude (MEI) is in fact present.

Q. How do you handle imbalance data?

A. Follow these techniques:

1. Use the right evaluation metrics.
2. Use K-fold Cross-Validation in the right way.
3. Ensemble different resampled datasets.
4. Resample with different ratios.
5. Cluster the abundant class.
6. Design your own models. Sigmoid

Q. Difference between sigmoid and softmax ?

A. The sigmoid function is used for the two-class logistic regression, whereas the softmax function is used for the multiclass logistic regression (a.k.a. MaxEnt, multinomial logistic regression, softmax Regression, Maximum Entropy Classifier).

Q. Precision and Recall? How they are related to ROC curve?

A. The recall is the ratio of the relevant results returned by the search engine to the total number of the relevant results that could have been returned. The precision is the proportion of relevant results in the list of all returned search results. When dealing with highly skewed datasets, Precision-Recall (PR) curves give a more informative picture of an algorithm's performance. We show that a deep connection exists between ROC space and PR space, such that a curve dominates in ROC space if and only if it dominates in PR space.
