Company Name - Deloitte
Role - Data Scientist

Q. Difference between Correlation and Regression?

A. The main difference in correlation vs regression is that the measures of the degree of a relationship between two variables; let them be x and y. Here, correlation is for the measurement of degree, whereas regression is a parameter to determine how one variable affects another.

Q. Why do we square the residuals instead of using modulus?

A. It is because of the extra penalty for higher errors and squaring the residuals for mean deviation were observed to be more efficient than mean absolute deviation.

Q. Which evaluation metric should you prefer to use for a dataset having a lot of outliers in it?

A. Mean Absolute Error(MAE) is preferred when we have too many outliers present in the dataset because MAE is robust to outliers whereas MSE and RMSE are very susceptible to outliers and these start penalizing the outliers by squaring the error terms.

Q. Heteroscedasticity? How to detect it?

A. Heteroskedasticity refers to situations where the variance of the residuals is unequal over a range of measured values. When running a regression analysis, heteroskedasticity results in an unequal scatter of the residuals (also known as the error term). To check for heteroscedasticity, you need to assess the residuals by fitted value plots specifically. Typically, the telltale pattern for heteroscedasticity is that as the fitted values increases, the variance of the residuals also increases.

Q. Describe is P value?

A. A p-value is a measure of the probability that an observed difference could have occurred just by random chance. The lower the p-value, the greater the statistical significance of the observed difference. P-value can be used as an alternative to or in addition to pre-selected confidence levels for hypothesis testing.

Q. What is Root Cause Analysis?

A. Root cause analysis (RCA) is defined as a collective term that describes a wide range of approaches used to uncover causes of problems. Some RCA approaches are geared more toward identifying true root causes than others, some are more general problem-solving techniques, and others simply offer support for the core activity of root cause analysis.

Q. Describe about Regularization?

A. Regularization is the process which regularizes or shrinks the coefficients towards zero. In simple words, regularization discourages learning a more complex or flexible model, to prevent overfitting.

Q. What is DBSCAN Clustering?

A. Density-based spatial clustering of applications with noise (DBSCAN) is a well-known data clustering algorithm that is commonly used in data mining and machine learning. The main idea behind DBSCAN is that a point belongs to a cluster if it is close to many points from that cluster. It is able to find arbitrary shaped clusters and clusters with noise (i.e. outliers).

Q. What is R2? What are some other metrics that could be better than R2 and why?

A. R-squared (R^2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. R-squared does not measure goodness of fit. R-squared does not measure predictive error. R-squared does not allow you to compare models using transformed responses. R-squared does not measure how one variable explains another. Some better metrics that could be better than R2 are:

- Mean Squared Error (MSE).
- Root Mean Squared Error (RMSE).
- Mean Absolute Error (MAE)

Q. What is the curse of dimensionality?

A. The curse of dimensionality basically means that the error increases with the increase in the number of features. It refers to the fact that algorithms are harder to design in high dimensions and often have a running time exponential in the dimensions.

Q. What are advantages of plotting your data before performing analysis?

A. It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, find interesting relations among the variables.

Q. How would you explain a confidence interval to an engineer with no statistics background? What does 95% confidence mean?

A. A 95% confidence interval, for example, implies that were the estimation process repeated again and again, then 95% of the calculated intervals would be expected to contain the true parameter value.

Q. How do you deal with some of your predictors being missing?

A. Simple approaches include taking the average of the column and use that value, or if there is a heavy skew the median or mode might be better. A better approach, you can perform regression or nearest neighbor imputation on the column to predict the missing values. Then continue on with your analysis/model.

Q. What is confounding variables?

A. Confounding variable is a variable that is not included in an experiment, yet affects the relationship between the two variables in an experiment.

Q. Algorithms which solve overfitting problem?

A. PCA, Ridge regression, L1/L2 regularization etc.

Q. What are Plots to evaluate models?

A. Residual plots, validation curve, gain and lift chart, kolmogorov smirnov chart

Q. What is Bimodel, unimodel and skewed data?

A. Unimodal Distribution, as the name suggests, is a single peaked distribution which means one value occurs with the greatest frequency than the other values. Distributions often have a clear peak to their shape. If a distribution has two fairly equal high points, it is called a bimodal distribution. It is a distribution where two values occur with the greatest frequency. The graph resembles two humps on a camel's back. A data is called as skewed when curve appears distorted or skewed either to the left or to the right, in a statistical distribution. In a normal distribution, the graph appears symmetry but here, the data is unsymmetrical.

Q. How complex would SVM become if you have many features (eg 30)?

A. SVM does not overfit when using a lot of features, provided that you regularize correctly. With less features also, sometimes SVM model can have less accuracy. So, if you have many features SVM wouldn't be much complex model.

Q. Why scaling useful?

A. So if the data in any conditions has data points far from each other, scaling is a technique to make them closer to each other or in simpler words, we can say that the scaling is used for making data points generalized so that the distance between them will be lower.

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.

NOTE:
If you want to learn Data Science from scratch and become interview ready, check out our popular Data science course.

https://ai.cloudyml.com/Learn-Data-Science-from-Scratch