Company Name - ZS
Role - Data Scientist

Q. Meaning when p values are high or low?

A. High p-values indicate that your evidence is not strong enough to suggest an effect exists in the population. An effect might exist but it's possible that the effect size is too small, the sample size is too small, or there is too much variability for the hypothesis test to detect it. The smaller the p-value, the stronger the evidence that you should reject the null hypothesis.

Q. Difference between expected and mean value?

A. While mean is the simple average of all the values, expected value of expectation is the average value of a random variable which is probability-weighted.

Q. How time series problems different from regression problem?

A. Regression is Interpolation. Time-series refers to an ordered series of data. Time-series models usually forecast what comes next in the series - much like our childhood puzzles where we extrapolate and fill patterns.

Q. what is RoC curve?

A. An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters: True Positive Rate. False Positive Rate.

Q. Random forest or multiple decision trees. Which is better?

A. Random Forest is suitable for situations when we have a large dataset, and interpretability is not a major concern. Decision trees are much easier to interpret and understand. Since a random forest combines multiple decision trees, it becomes more difficult to interpret.

Q. Example when false positive is important than false negative?

A. A false positive is where you receive a positive result for a test, when you should have received a negative result. Some examples of false positives: A pregnancy test is positive, when in fact you aren't pregnant. A cancer screening test comes back positive, but you don't have the disease. Innocent party is found guilty in such cases.

Q. Gridsearch vs Random search?
A. Random search differs from grid search in that we no longer provide an explicit set of possible values for each hyperparameter; rather, we provide a statistical distribution for each hyperparameter from which values are sampled. Essentially, we define a sampling distribution for each hyperparameter to carry out a randomized search.

Q. Hyperparameters in SVM?

A. kernel: It maps the observations into some feature space. Ideally the observations are more easily (linearly) separable after this transformation. There are multiple standard kernels for these transformations, e.g. the linear kernel, the polynomial kernel and the radial kernel.

C: It is a hypermeter in SVM to control error. The C parameter tells the SVM optimization how much you want to avoid misclassifying each training example. For large values of C, the optimization will choose a smaller-margin hyperplane if that hyperplane does a better job of getting all the training points classified correctly.
gamma: Gamma is used when we use the Gaussian RBF kernel. if you use linear or polynomial kernel then you do not need gamma only you need C hypermeter. Somewhere it is also used as sigma. Gamma decides that how much curvature we want in a decision boundary.

Q. Multicollinearity? Deal?

A. Multicollinearity occurs when two or more independent variables(also known as predictor) are highly correlated with one another in a regression model. This means that an independent variable can be predicted from another independent variable in a regression model.

To remove multicollinearities, we can do two things.

1. We can create new features

2. remove them from our data.

Q. Why use Decision Trees?

A. First, a decision tree is a visual representation of a decision situation (and hence aids communication). Second, the branches of a tree explicitly show all those factors within the analysis that are considered relevant to the decision (and implicitly those that are not).

Q. What is Time Series (ARIMA)?

A. ARIMA, short for 'AutoRegressive Integrated Moving Average', is a forecasting algorithm based on the idea that the information in the past values of the time series can alone be used to predict the future values.

Q. Detect outliers, what are quartiles?

A. The most effective way to find all of your outliers is by using the interquartile range (IQR). The IQR contains the middle bulk of your data, so outliers can be easily found once you know the IQR. Quartiles divide the entire set into four equal parts. So, there are three quartiles, first, second and third represented by Q_1 , Q_2 and Q_3 , respectively. Q_2 is nothing but the median.

Q. What is Conditional Probability?

A. Conditional probability formula gives the measure of the probability of an event given that another event has occurred.

Q. PCA Advantages and Disadvantages?

A. Advantages: Removes correlated features, Improves algorithm features, Reduces overfitting, Improves Visualization.

Disadvantages: Independent variables become less interpretable, Data Standardization is must, Information loss.

Q. What is Pruning?

A. 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.
