Company Wise Data Science Interview Questions

*Note: These questions are some of the commonly asked questions that our students have been asked during the interviews. The questions change and vary according to the interviewer.

COMPANY NAME: **INFOSYS**

ROLE: **DATA SCIENTIST**

- Curse of dimensionality? How would you handle it?
- How to find multicollinearity in the data set
- Explain the different ways to treat multicollinearity?
- How do you decide which feature to keep and which feature to eliminate after performing the multicollinearity test?
- Explain logistic regression
- We have a sigmoid function that gives the probability ability between 0-1 then what is the need for log loss in logistic regression?
- What is P-Value and its significance in statistical testing?
- How do you split the time series data and evaluation matrix for time series data?
- How did you deploy your model in production? How often do you retrain it?

COMPANY: **GOOGLE**

ROLE: **DATA SCIENTIST**

- Why do you use feature selection?
- What is the effect on the coefficients of logistics regression if two 3. Predictions are highly correlated?
- What are the confidence intervals of the coefficients?
- What is the difference between Gaussian Mixture Model and K-Means?
- How do you pick k in K-Means?
- How do you know when Gaussian Mixture Model is applicable?
- Assuming a clustering Model's labels are known, how do you evaluate the performance of the Model?

COMPANY: **XORIANT**

ROLE: **DATA SCIENTIST**

- List out data validation techniques
- Assumption of Linear Regression Model
- KNN imputation
- Why rotation of component in PCA?
- What is the role of groupby() function?
- Differentiate between Decision Tree and Random Forest
- .loc() VS .iloc()
- Handling outliers and missing values
- What technique is better to go with mean vs median vs mode?
- DataFrame VS Series
- Give me a solution for a problem where there would be a book name and you need to predict the accomplishment.

COMPANY: **WIPRO**

ROLE: **DATA SCIENTIST**

- Difference between WHERE and HAVING in SQL
- Basics if Logistic Regression
- How do you treat outliers?
- Explain confusion matrix
- Explain PCA { wanted me to explain the covariance matrix and Eigen vectors, values and mathematical expression and mathematical derivation for co-variance matrix}
- How do you cut a cake into 8 equal parts using only 3 straight cuts?
- Explain k-means clustering
- How is KNN different from k-means clustering?
- What would be your strategy to handle a situation indicating imbalanced datasets?
- Stock market predictions: If you are asked to predict whether or not a certain company will declare bankruptcy in the next 7 days. {Would you treat this as a classification or a Regression problem}

COMPANY: **TCS**

ROLE: **DATA SCIENTIST**

- Explain time series models, which have you used?
- SQL Questions- Group by top 2 salaries for employees-use Row num and Partition.
- Pandas and Numeric and Categorical Columns. For Numeric columns in DataFrame, find the mean of the entire column and add that mean value to each row of these numeric columns.
- What is Gradient Descent? What is the Learning rate and why do we need to reduce or increase it? Why Global minimum is reached and why it doesn't improve when increasing the LR after that point?
- What are Log-Loss and ROC-AUC?
- What is Multi-collinearity? How will you choose one feature if there are 2 highly correlated features? Give Examples with the techniques used
- VIF- Variance Inflation Factor explain.
- Do you know how to use Amazon SageMaker for MLOPS?
- Explain your projects end to end (15-20mins)

COMPANY: **UBER**

ROLE: **DATA SCIENTIST**

- Pick any product or app that you really like and describe how you would improve it.
- How would you find an anomaly in a distribution?
- How would you go about investigating if a certain trend in distribution is due to an anomaly?
- How would you estimate the impact Uber has on traffic and driving conditions?
- What metrics would you consider using to track if Ubers paid advertising strategy to acquire new customers actually works? How would you then approach figuring out an ideal customer acquiring cost?

COMPANY: **MINDTREE**

ROLE: **DATA SCIENTIST**

- What is a central tendency?
- Which central tendency is method is used if there exist any outliers?
- Central limit theorem
- Chi-Square test
- XYZ testing
- Difference between Z and T distribution(Linked to XYZ testing)
- Outlier treatment methods
- ANOVA test
- Cross-validation
- How will work in a machine learning project if there is a huge imbalance in the data
- Formulae of Sigmoid function
- Can we use the Sigmoid function in case of multiple classifications?
- What is Area under the curve?
- Which metric is used to split a node in the Decision tree?
- What is Ensembling learning?
- 3-4 situations based on these questions.

COMPANY: **GENPACT**

ROLE: **DATA SCIENTIST**

- Why do we select validation data other than test data?
- Difference between Linear- Logistic regression?
- Why do we take such complex cost functions for logistics?
- Difference between Random Forest and Decision tree?
- How do you decide when to stop spinning the tree?
- Measure of central tendency?
- What is the requirement of K-means algorithm?
- Which clustering techniques uses combining of clusters?
- Which is the oldest probability distribution?
- What all values does a random variable can take?
- Types of random variables?
- Normality of residuals

COMPANY: **FORD**

ROLE: **DATA SCIENTIST**

- How would you check if the model is suffering from multicollinearity?
- What is transfer learning? Steps you would take to perform transfer learning.
- Why is CNN architecture suitable for image classification? Why not RNN?
- What are the approaches for solving the class imbalance problems?
- When sampling what types of biases can be infected? How to control the biases?
- Explain concepts of epoch, batch iteration in machine learning?
- What type of performance matrix would you choose to evaluate the different classification models and why?
- What are some of the types of activation functions and specifically when to use them?
- What are the conditions that should be satisfied for a time series to be stationary?
- What is the difference between batch and Stochastic Gradient Descent?
- What is the difference between KNN and K-means clustering?

COMPANY: **QUANTIPHI**

ROLE: **MACHINE LEARNING ENGINEER**

- What happens when neural nets are too small? What happens when they are large enough?
- Why do we need a pooling layer in CNN? Common pooling methods?
- Are essemble models better than individual models? Why/why not?
- Use case- Considering you are working in a pen manufacturing company, how would you help the sales team with leads using data analysis?
- Assume you were given access to a website with google analytics data.
- In order to increase conversations, how do you perform A/B testing to identify the best page design?
- How is random forest different from Gradient boosting algorithm, given both are tree-based algorithms?
- Describe steps involved in creating a neural network?
- In brief, how would you perform the task of sentiment analysis?

COMPANY: **THE MATH COMPANY**

ROLE: **DATA SCIENTIST**

- Central limit theorem
- Hypothesis testing
- P value
- T-test
- Assumptions of linear regression
- Correlation and covariance
- How to identify and treat outliers and missing values
- Explain Box and Whiskers plot
- Explain any unsupervised learning algorithm
- Explain Random forest
- Business and technical questions related to your project
- Explain any scope of improvement in your project
- Questions based on case studies
- Write SQL query to find an employee with the highest salary in each department
- Write SQL query to find a unique email domain name and their respective count
- Solve questions using python (Usually 17 questions)

Rounds:

- Technical test (Python, SQL, Statistics) (Coding + MCQ) (90 min)
- Telephonic interview (10 min)
- Technical interview (45mins)
- Fitment interview (25mins)
- HR interview (30min)

COMPANY: **CAPITAL ONE**

ROLE: **DATA SCIENTIST**

- How would you build a model to predict credit card fraud?
- How do you handle missing or bad data?
- How do you derive new features from features that already exists?
- If you are attempting to predict a customer's gender, and you have only 100 data point, what problems could arise
- Suppose you were given two years of transaction history, what features would you use to predict credit risk?
- Design an AI program for Tic-tac-toe
- Explain over fitting and what steps you can take to prevent it?
- Why does SVM need to maximize the margin between support vectors?

COMPANY: **LATENT VIEW ANALYTICS**

ROLE: **DATA SCIENTIST**

- What is mean and median?
- Difference between normal and Gaussian distribution
- What is central limit theorem
- What is null hypothesis
- What is confidence interval?
- What is covariance and correlation and how will you interpret it?
- How will you find out the outliers in the dataset? And how is it always to remove outliers?
- Explain about Machine Learning
- Explain the algorithm of your choice
- Different methods of missing value imputation
- Explain your ML project
- How do you handle imbalance datasets?
- What is stratified sampling?
- Difference between standard scalar and normal scalar

COMPANY: **VERIZON**

ROLE: **DATA SCIENTIST**

- How many cars are there in Chennai? How do you structurally approach coming up with that number?
- Multiple linear regression?
- OLS vs MLE?
- R2 vs adjusted R2? During Model development which one do you consider?
- Lift chart and Drift charge?
- Sigmoid function in Logistic regression?
- ROC what is it? AUC and Differentiation?
- P-Value what is it and its significance? What does P in P-Value stand for? What is Hypothesis testing? Null Hypothesis vs Alternate Hypothesis?
- Bias Variance trade off?
- Over fitting vs Underfitting in Machine Learning?
- Estimation of Multiple Linear regression
- Forecasting vs Prediction difference? Regression vs time series?

COMPANY: **FRACTAL**

ROLE: **DATA SCIENTIST**

- Difference between array and list
- Map function
- Scenario
    - If coupon distributed randomly to customers of swiggy, how to check there buying behaviour?
    - Will you use segmenting the customers
    - Will you compare customers who got coupon and who didn't?
- Which is faster? List or Dict for look up
- How to merge two arrays?
- How much time SVM takes to complete if 1 iteration takes 10 sec for 1st class and there are 4 classes?
- Kernels in SVM. Their differences.

COMPANY: **TATA IQ**

ROLE: **DATA ANALYST**

- Why Data Science as a career?
- Stats:
    - What is P-Value?
    - What are Histograms?
    - What is confidence interval?
- Imagine you are a Sr data analyst at a new online cab booking startups
    - How will you do data collection and how will you leverage the data to give useful insights to the company?
    - Estimate: No of cabs bookings per day in Ranchi
- Imagine you are a product head manager at a NBFC which gives a secured loan what factors will you consider giving loan to?
- Inventory database on that have to be basic pandas/SQL query? Joins/ merge to get avg sales, its charts?
- You have a list of 3 numbers return the min diff, can use any python/SQL?
- What is Big data?

COMPANY: **TATA IQ**

ROLE: **JUNIOR DATA SCIENTIST**

- Explain the architecture of CNN
- If we put 3*3 filter over 6*6 image what will be the size of the output image?
- What will you do to reduce overfitting in deep learning models?
- Can you write a program for inverted star program in python?
- Write a program to create a data frame and remove elements from it
- Imagine you have 2 guns with 6 holes in each, and you load a single bullet in each gun, what is the probability that if I fire the guns simultaneously at least 1 gun will fire.
- There are 2 groups g1 and g2, g1 will ask g2 members to give them 1 member so that they both will equal in number, g1 will ask g1 members to give them 1 member so that they will be double of g1, how many members are there in the groups?

COMPANY: **TIGER ANALYTICS**

ROLE: **SENIOR DATA ANALYST**

- What is deep learning, and how does it contrast with other Machine Learning algorithms?
- When should you use classification over regression?
- Using Python how do you find rank, linear and tensor equations for an given array of elements? Explain your approach
- What exactly do you know about Bias-variance decomposition?
- What is the best recommendation technique you have learned and what type of recommendation technique helps to predict ratings?
- How can you access a good logistic model?
- How do you read the text from an image? Explain
- What are the options to convert speech to text? Explain and name few available tools to implement the same?