Company wise Data Science Interview Questions

Company: Google
Role: Data Scientist

1. Why do you use feature selection?
2. What is the effect on the coefficients of logistic regression if two 3. predictors are highly correlated?
4. What are the confidence intervals of the coefficients?
5. What's the difference between Gaussian Mixture Model and K-Means?
6. How do you pick k for K-Means?
7. How do you know when Gaussian Mixture Model is applicable?
8. Assuming a clustering model's labels are known, how do you evaluate the performance of the model?

Xoriant Interview Questions

1. List out Data Validation techniques
2. Assumption of Linear Regression Model
3. KNN imputation
4. Why rotation of component in PCA?
5. what is the role of groupby() function?
6. Decision Tree vs Random Forest
7. loc vs iloc
8. Handle outliers and filling missing values
9. What techique is better to go with mean vs median vs mode?
10. DataFrame vs Series
11. Give me a solution for a problem where there would be a book name and you need to predict the accomplishment.

Company wise Data Science Interview Questions

Follow me for more resources

@arjun-panwar
https://www.linkedin.com/in/arjun-panwar

Company: Uber
Role: Data Scientist

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

Company: TCS
Role: Data Scientist

1. Explain about Time series models you have used?
2. SQL Questions - Group by Top 2 Salaries for Employees - use Row num and Partition
3. Pandas find Numeric and Categorical Columns. For Numeric columns in Data frame, find the mean of the entire column and add that mean value to each row of those numeric columns.
4. What is Gradient Descent? What is Learning Rate and Why we need to reduce or increase? Why Global minimum is reached and Why it doesn't improve when increasing the LR after that point?
5. What is Log-Loss and ROC-AUC?
6. What is Multi-collinearity? How will you choose one features if there are 2 highly correlated features? Give Examples with the techniques used.
7. VIF – Variance Inflation Factor – Explain.
8. Do you know to use Amazon SageMaker for MLOPS?
9. Explain your Projects end to end (15-20mins).

Company: Capital One
Role: Data Scientist

1. How would you build a model to predict credit card fraud?
2. How do you handle missing or bad data?
3. How would you derive new features from features that already exist?
4. If you're attempting to predict a customer's gender, and you only have 100 data points, what problems could arise?
5. Suppose you were given two years of transaction history. What features would you use to predict credit risk?
6. Design an AI program for Tic-tac-toe
7. Explain overfitting and what steps you can take to prevent it.
8. Why does SVM need to maximize the margin between support vectors?

Company: Latentview Analytics
Role: Data Scientist
Experience: 2 years

1. What is mean and median
2. Difference between normal and gaussian distribution
3. What is central limit theorem
4. What is null hypothesis
5. What is confidence interval
6. What is covariance and correlation and how will u interpret it.
7. How will you find out the outliers in the dataset and is it always to remove outliers
8. Explain about Machine Learning
9. Explain the algorithm of your choice
10. Different methods of missing values imputation
11. Explain me your ml project
12. How did you handle imbalance dataset
13. What is stratified samplings
14. Difference between standard scalar and normal scalar

Company: Verizon
Role: Data Scientist

1. How many cars are there in Chennai? How do u structurally approach coming up with that number?
2. Multiple Linear Regression?
3. OLS vs MLE?
4. R2 vs Adjusted R2? During Model Development which one do we consider?
5. Lift chart, drift chart
6. Sigmoid Function in Logistic regression
7. ROC what is it? AUC and Differentiation?
8. Linear Regression from Multiple Linear Regression
9. 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?
10. Bias Variance Trade off?
11. Over fitting vs Underfitting in Machine learning?
12. Estimation of Multiple Linear Regression
13. Forecasting vs Prediction difference? Regression vs Time Series?
14. p,d,q values in ARIMA models

Company: Fractal
Role: Data Scientist

1.Difference between array and list
2.Map function
3. Scenario,
if coupon distributed randomly to customers of swiggy, how to check there buying behavior.
Use segmenting customers
Compare customers who got coupon and who did not
4. Which is faster dictionary or list for look up
5. How to merge two arrays
6. How much time svm takes to complete if 1 iteration takes 10sec for 1st class. And there are 4 classes.
7. Kernals in svm, there difference

Company name: Infosys
Role: Data scientist

1) curse of dimensionality? How would you handle it?
2) How to find the multi collinearity in the data set
3)Explain the difference ways to treat multi collinearity!
4) How you decide which feature to keep and which feature to eliminate after performing multi collinearity test?
5)Explain logistic regression
6)we have sigmoid function which gives us the probabilty between 0-1 then what is the need of logloss in logistic regression?
7) P value and its significance in statistical testing?
8) How do you split the time series data and evaluation metrics for time series data
9) How did you deploy your model in production? How often do you retrain it?

Company: Wipro
Role: Data Scientist

1. Difference between WHERE and HAVING in SQL
2. Basics of Logistics Regression
3. How do you treat outliers ?
4. Explain confusion matrix ?
5. Explain PCA (Wanted me to explain the co-variance matrix and eigen vectors and values and the mathematical expression and mathematical derivation for co-variance matrix)
6. How do you cut a cake into 8 equal parts using only 3 straight cuts ?
7. Explain kmeans clustering
8. How is KNN different from k-means clustering?
9. What would be your strategy to handle a situation indicating an imbalanced dataset?
10. Stock market prediction: You would like to predict whether or not a certain company will declare bankruptcy within the next 7 days (by training on data of similar companies that had previously been at risk of bankruptcy). Would you treat this as a classification or a regression problem?

Company: Accenture
Role: Data Scientist

1. What is difference between K-NN and K-Means clustering?
2. How to handle missing data? What imputation techniques can be used?
3. Explain topic modelling in NLP and various methods in performing topic modeling.
4. Explain how you would find and tackle an outlier in the dataset.
5. Follow up: What about inlier?
6. Explain back propagation in few words and its variants?
7. Is interpretability important for machine learning model? If so, ways to achieve interpretability for a machine learning models?
8. Is interpretability important for machine learning model? If so, ways to achieve interpretability for a machine learning models?
9. How would you design a data science pipeline?
10. Explain bias - variance trade off. How does this affect the model?
11. What does a statistical test do?
12. How to determine if a coin is biased? Hint: Hypothesis testing

Company: Tiger Analytics
Role: Senior Analyst

1. What is deep learning, and how does it contrast with other machine learning algorithms?
2.When should you use classification over regression?
3.Using Python how do you find Rank, linear and tensor equations for an given array of elements? Explain your approach.
4.What exactly do you know about Bias-Variance decomposition? 5.What is the best recommendation technique you have learnt and what type of recommendation technique helps to predict ratings? 6.How can you assess a good logistic model?
7.How to you read the text from an image? Explain?
8.What are all the options to convert speech to text? Explain and name few available tools to implement the same?

Company Name : Tata IQ
Role: Data Analyst

1. Why data science as a career?
2. Stats:
3. What is p value?
4. What is histograms?
5. What is confidence interval?
6. You are a Sr data analyst at a new Online Cab booking Startups
7. How you will do data collection and how you will leverage the data to give useful insights to the Company?
8. Guestimate: No Of cabs booking per day in Ranchi
9. You are product head manager(not remember exactly) at a NBFC which gives a Secured loans what factors will you consider giving loan to ?
10. Inventory Database based on that have to do basic pandas/sql query? Joins / merge to get avg sales, its chart?
11. You have a list of 3 numbers return the min diff. Can use any python/sql
12. What is Big Data?

Role: Junior Data Scientist

1) Explain the architecture of CNN
2)If we put a 3×3 filter over 6×6 image what will be the size of the output image
3) What will you do to reduce overfitting In deep learning models
3) Can you write a program for inverted star program in python
4)Write a program to create a dataframe and remove elements from it
5) I have 2 guns with 6 holes in each, and I load a single bullet In each gun, what is the probability that if I fire the guns simultaneously atleast 1 gun will fire (atleast means one or more than one)
5) There are 2 groups g1 and g2, g1 will ask g2 members to give them 1 member so thay they both will be equal in number, g2 will ask g1 members to give them 1 member so thay they will be double of g1, how many members are there in the groups (I'm not sure of this question as I tried to solve but didnt get correct answer)

Company: Mindtree
Role: Data Scientist

1. What is central tendency
2. Which central tendency method is used If there exists any outliers
3. Central limit theorem
4. Chi-Square test
5. A/B testing
6. Difference between Z and t distribution (Linked to A/B testing)
7. Outlier treatment method
8. ANOVA test
9. Cross validation
10. How will you work in a machine learning project if there is a huge imbalance in the data
11. Formula of sigmoid function
12. Can we use sigmoid function in case of multiple classification
13. What is Area under the curve
14. Which metric is used to split a node in Decision Tree
15. What is ensemble learning
16. 3 situation based questions

Company: Genpact
Role: Data Scientist

1. Why do we select validation data other than test data?
2. Difference between linear logistic regression?
3. Why do we take such a complex cost function for logistic?
4. Difference between random forest and decision tree?
5. How would you decide when to stop splitting the tree?
6. Measures of central tendency
7. What is the requirement of k means algorithm
8. Which clustering technique uses combining of clusters
9. Which is the oldest probability distribution
10. What all values does a random variable can take
11. Types of random variables
12. Normality of residuals

Company wise Data Science Interview Questions

Follow me for more resources

@arjun-panwar
https://www.linkedin.com/in/arjun-panwar

Company: Ford
Role: Data Scientist

1. How would you check if the model is suffering from multi Collinearity?
2. What is transfer learning? Steps you would take to perform transfer learning.
3. Why is CNN architecture suitable for image classification? Not an RNN?
4. What are the approaches for solving class imbalance problem?
5. When sampling what types of biases can be inflected? How to control the biases?
6. Explain concepts of epoch, batch, iteration in machine learning.
7. What type of performance metrics would you choose to evaluate the different classification models and why?
8. What are some of the types of activation functions and specifically when to use them?
9. What are the conditions that should be satisfied for a time series to be stationary?
10. What is the difference between Batch and Stochastic Gradient Descent?
11. What is difference between K-NN and K-Means clustering?

Company: Quantiphi
Role: Machine Learning Engineer

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

Company: TheMathCompany
Role: Analyst (Data Science)

1. Central limit theorem
2. Hypotheses testing
3. P value
4. T-test
5. Assumptions of linear regression.
6. Correlation and covariance.
7. How to identify & treat outliers and missing values.
8. Explain Box and whisker plot.
9. Explain any unsupervised learning algorithm.
10. Explain Random forest.
12. Business and technical questions related to your project.
13. Explain any scope of improvement in your project.
14. Questions based on case studies.
16. Write SQL query to find employee with highest salary in each department.
17. Write SQL query to find unique email domain name & their respective count
18. Solve question (17) using Python.

Rounds:
1. Technical Test (Python, SQL, Statistics) (Coding+MCQ) (90 min).
2. Telephonic interview (10 min).
3. Technical interview (45 min).
4. Fitment interview (25 min).
5. HR interview (30 min).

Company: Cognizant
Role: Data Scientist

1. SQL question on inner join and cross join
2. SQL question on group-by
3. Case study question on customer optimization of records for different marketing promotional offers
4. Tuple and list
5. Linear regression
6. Logistic regression steps and process
7. Tell me about your passion for data science? Or What brought you to this field?
8. What is the most common problems you face whilst working on data science projects?
9. Describe the steps to take to forecast quarterly sales trends. What specific models are most appropriate in this case?
10. What is the difference between gradient and slope, differentiation and integration?
11. When to use deep learning instead of machine learning. Advantages, Disadvantages of using deep learning?
12. What are vanishing and exploding gradients in neural networks?

Company: Husqvarna Group
Role: Data Scientist

1. Telecom Customer Churn Prediction. Explain the project end to end?
2. Data Pre-Processing Steps used.
3. Sales forecasting how is it done using Statistical vs DL models - Efficiency.
4. Logistic Regression - How much percent of Customer has churned and how much have not churned?
5. What are the Evaluation Metric parameters for testing Logistic Regression?
6. What packages in Python can be used for ML? Why do we prefer one over another?
7. Numpy vs Pandas basic difference.
8. Feature on which this Imputation was done, and which method did we use there?
9. Tuple vs Dictionary. Where do we use them?
10. What is NER - Named Entity Recognition?

Company wise Data Science Interview Questions

Follow me for more resources

@arjun-panwar
https://www.linkedin.com/in/arjun-panwar

Company: Deloitte
Role: Data Scientist

1. Conditional Probability
2. Can Linear Regression be used for Classification? If Yes, why if No why?
3. Hypothesis Testing. Null and Alternate hypothesis
4. Derivation of Formula for Linear and logistic Regression
5. Why use Decision Trees?
6. PCA Advantages and Disadvantages?
7. What is Naive Bayes Theorem? Multinomial, Bernoulli, Gaussian Naive Bayes.
8. Central Limit Theorem?
9. Scenario based question on when to use which ML model?
10. Over Sampling and Under Sampling
11. Over Fitting and Under Fitting
12. Core Concepts behind Each ML model mentioned in my Resume.
13. Genie Index Vs Entropy
14. how to deal with imbalance data in classification modelling?

Company: Wipro
Role: Data Scientist

1. What is a Python Package, and Have you created your own Python Package?
2. Explain about Time series models you have used?
3. SQL Questions - Group by Top 2 Salaries for Employees - use Row num and Partition
4. Pandas find Numeric and Categorical Columns. For Numeric columns in Data frame, find the mean of the entire column and add that mean value to each row of those numeric columns.
5. What is Gradient Descent? What is Learning Rate and Why we need to reduce or increase? Why Global minimum is reached and Why it doesn't improve when increasing the LR after that point?
6. Two Logistic Regression Models - Which one will you choose - One is trained on 70% and other on 80% data. Accuracy is almost same.
8. What is Log-Loss and ROC-AUC?
9. Do you know to use Amazon SageMaker for MLOPS?
10. Explain your Projects end to end (15-20mins).

Company: Infosys
Role: Data Scientist

1) Measures of central tendency
2) What is the requirement of k means algorithm
3) Which clustering technique uses combining of clusters
4) Which is the oldest probability distribution
5) What all values does a random variable can take
6) Types of random variables
7) Normality of residuals
8) Probability questions
9) Sensitivity and specificity etc.
10) Explain bias - variance trade off. How does this affect the model?
11) What is multi collinearity? How to identify and remove it.

Company: Tiger Analytics
Role: Data Scientist

1. What are the projects done by you.
2. Suppose there is a client who wants to know if giving discounts is beneficial or not. How would you approach this problem?
3. The same client want to know how much discount he should give in the next month for maximum profits.
4. Can you have a modeling approach to say in last year what mistakes client did in giving discounts. Meaning if they should have have a different discount and increased sales.
5. What feature engineering techniques you used in past projects.
6. What models you used and selected the final model.

Company: Genpact
Role: Data Scientist

1. What makes you feel that you would be suitable for this role, since you come from a different background?
2. What is an imbalanced data set??
3. What are the factors you will consider in order to predict the population of a city in the future?
4. Basic statistics questions?
5. What are the approaches for treating the missing values?
6. Evaluation metrics for Classification?
7. Bagging vs Boosting with examples
8. Handling of imbalanced datasets
9. What are your career aspirations?
10. What's the graph of y = |x|-2
11. esstimate on no. Of petrol cars in Delhi
12. Case study on opening a retail store
13. Order of execution of SQL

Company: Ericsson
Role: Data Scientist

Round No: 1st Round

1. How to reverse a linked list
2. Give a logistic regression model in production, how would you find out the coefficients of different input features.
3. What is the p- value in OLS regression
4. What's the reason for high bias or variance
5. Which models are generally high biased or high variance
6. Write code to find the 8 highest value in the DataFrame
7. What's difference between array and list
8. Whats the difference between Gradient boosting and Xgboost
9. Is XOR data linearly separable
10. How do we classify XOR data using logistic regression
11. Some questions from my previous projects
12. Given a sand timer of 4 and 7 mins how would you calculate 10 mins duration.
13. What's the angle between hour and minute hand in clock as 3:15

Company: FISERVE
Role: Data Scientist

1. How would you check if the model is suffering from multi Collinearity?
2. What is transfer learning? Steps you would take to perform transfer learning.
3. Why is CNN architecture suitable for image classification? Not an RNN?
4. What are the approaches for solving class imbalance problem?
5. When sampling what types of biases can be inflected? How to control the biases?
6. Explain concepts of epoch, batch, iteration in machine learning.
7. What type of performance metrics would you choose to evaluate the different classification models and why?
8. What are some of the types of activation functions and specifically when to use them?
9. What is the difference between Batch and Stochastic Gradient Descent?
10. What is difference between K-NN and K-Means clustering?
11. How to handle missing data? What imputation techniques can be used?

Company: Landmark group
Role: Data Scientist

1. Use Case - Consider you are working for pen manufacturing company. How would you help sales team with leads using Data analysis?
2. Interviewers ask about scenarios or use-case based questions to know interviewee thought process and problem-solving skills.
3. Assume you were given access to a website google analytics data.
4. In order to increase conversions, how do you perform A/B testing to identify best page design.
5. How is random forest different from Gradient boosting algorithm, given both are tree-based algorithm?
6. Describe steps involved in creating a neural network?
7. LSTM solves the vanishing gradient problem, that RNN primarily have. How?
8. In brief, how would you perform the task of sentiment analysis?

Company: Axtria

1.RNN, NN and CNN difference.
2. Supervised, unsupervised and reinforcement learning with there algo example.
3. Difference between ai, ml and dl
4. How u do dimentionality reduction.
5. What is Multicollinearity
6. Parameters of random forest
7 . Parameters of deep learning algos
8. Different feature selection methods
9. Confusion matrix

Company: Latentview Analytics
Role: Data Scientist
Experience: 2 years

1. What is mean and median
2. Difference between normal and gaussian distribution
3. What is central limit theorem
4. What is null hypothesis
5. What is confidence interval
6. What is covariance and correlation and how will u interpret it.
7. How will you find out the outliers in the dataset and is it always to remove outliers
8. Explain about Machine Learning
9. Explain the algorithm of your choice
10. Different methods of missing values imputation
11. Explain me your ml project
12. How did you handle imbalance dataset
13. What is stratified samplings
14. Difference between standard scalar and normal scalar
15. Different type of visualization in Dl project
16. What architecture have you used
17. Why have u not used RNN in your nlp project
18. Why we don't prefer CNN in nlp based project

19 What is exploding gradient and vanishing gradient and how to rectify it
20. Difference between LSTM and GRU
21. What is precision and recall
22 What is auc metic
23. What if your precision and recall are same
25.What is Bias Variance Trade Off?

Company: Bridgei2i
Role: Senior Analytics Consultant

1) What is the difference between Cluster and Systematic Sampling?
2) Differentiate between a multi-label classification problem and a multi-class classification problem.
3) How can you iterate over a list and also retrieve element indices at the same time?
4) What is Regularization and what kind of problems does regularization solve?
5) If the training loss of your model is high and almost equal to the validation loss, what does it mean? What should you do?
6) Explain evaluation protocols for testing your models? Compare hold-out vs k-fold cross validation vs iterated k-fold cross-validation methods of testing.
7) Can you cite some examples where a false positive is important than a false negative?
8) What is the advantage of performing dimensionality reduction before fitting an SVM?
9) How will you find the correlation between a categorical variable and a continuous variable ?
10) How will you calculate the accuracy of a model using a confusion matrix?
11) You are given a dataset with 1500 observations and 15 features. How many observations you will select in each decision tree in a random forest?
12) Given that you let the models run long enough, will all gradient descent algorithms lead to the same model when working with Logistic or Linear regression problems?
13) What do you understand by statistical power of sensitivity and how do you calculate it?
14) What is pruning, entropy and information gain in decision tree algorithm?
15) What are the types of biases that can occur during sampling?

Company: Prodapt Solutions
Role: Data Scientist

1. Telecom Customer Churn Prediction. Explain the project end to end?
2. Data Pre-Processing Steps used.
3. Sales forecasting how is it done using Statistical vs DL models - Efficiency.
4. Logistic Regression - How much percent of Customer has churned and how much have not churned?
5. What are the Evaluation Metric parameters for testing Logistic Regression?
6. What packages in Python can be used for ML? Why do we prefer one over another?
7. Numpy vs Pandas basic difference.
8. Feature on which this Imputation was done, and which method did we use there?
9. Tuple vs Dictionary. Where do we use them?
10. What is NER - Named Entity Recognition?

Company: Landmark group
Role: Data Scientist

1. SQL question on inner join and cross join
2. SQL question on group-by
3. Case study question on customer optimization of records for different marketing promotional offers
4. Tuple and list
5. Linear regression
6. Logistic regression steps and process
7. Tell me about your passion for data science? Or What brought you to this field?
8. What is the most common problems you face whilst working on data science projects?
9. Describe the steps to take to forecast quarterly sales trends. What specific models are most appropriate in this case?
10. What is the difference between gradient and slope, differentiation and integration?
11. When to use deep learning instead of machine learning. Advantages, Disadvantages of using deep learning?
12. What are vanishing and exploding gradients in neural networks?
13. What happens when neural nets are too small? What happens when they are large enough?
14. Why do we need pooling layer in CNN? Common pooling methods?
15. Are ensemble models better than individual models? Why/why - not?

Company: Mindtree
Role: Data Scientist

1. What is central tendency
2. Which central tendency method is used If there exists any outliers
3. Central limit theorem
4. Chi-Square test
5. A/B testing
6. Difference between Z and t distribution (Linked to A/B testing)
7. Outlier treatment method
8. ANOVA test
9. Cross validation
10. How will you work in a machine learning project if there is a huge imbalance in the data
11. Formula of sigmoid function
12. Can we use sigmoid function in case of multiple classification (I said no)
13. Then which function is used
14. What is Area under the curve
15. Which metric is used to split a node in Decision Tree
16. What is ensemble learning
17. 3 situation based questions

Company: CodeBase Solutions
Role: Data Scientist

1. What are the ML techniques you've used in projects?
2. Very first question was PCA? Why use PCA?
3. Types of Clustering techniques (Not algorithms)? Which Clustering techniques will you use in which Scenario - example with a Program?
4. OCR - What type of OCR did you use in your project - Graphical or Non - Graphical?
5. OCR - What is a Noise? What types of noise will you face when performing OCR? Handwritten can give more than 70% accuracy when I wrote in 2012 but you're saying 40%.
6. Logistic Regression vs Linear Regression with a real-life example - explain?
7. Is Decision tree Binary or multiple why use them?
8. Do you know Map Reduce and ETL concepts?
9. What is a Dictionary or Corpus in NLP and how do you build it?

10. How do you basically build a Dictionary, Semantic Engine, Processing Engine in a NLP project, where does all the Synonyms (Thesaurus words go).
11. What are the Types of Forecasting? What are the ML and DL models for forecasting (He said Fast-forwarding models as example) other than Statistical (ARIMA) models you've used in your projects?
12. What is a Neural Network? Types of Neural Networks you know?
13. Write a Decision Tree model with a Python Program.
14. How do you build an AZURE ML model? What are all the Azure products you've used? I said Azure ML Studio.
15. Cibil score is an example for Fuzzy model and not a Classification model.
16. What is an outlier give a real life example? how do you find them and eliminate them? I gave an example of calculating Average salary of an IT employee.

Company: Deloitte
Role: Data Scientist

1. G values, P values, T values
2. Conditional Probability
3. Central Values of Tendency
4. Can Linear Regression be used for Classification? If Yes, why if No why?
5. Hypothesis Testing. Null and Alternate hypothesis
6. Derivation of Formula for Linear and logistic Regression
7. Where to start a Decision Tree. Why use Decision Trees?
8. PCA Advantages and Disadvantages?
9. Why Bayes theorem? DB Bayes and Naïve Bayes Theorem?
10. Central Limit Theorem?
11. R packages in and out? For us it's Python Packages in and out.
12. Scenario based question on when to use which ML model?
13. Over Sampling and Under Sampling
14. Over Fitting and Under Fitting
15. Core Concepts behind Each ML model.
16. Genie Index Vs Entropy
17. how to deal with imbalance data in classification modelling? SMOTHE techniques

Verizon Data Science Interview Questions

1. How many cars are there in Chennai? How do u structurally approach coming up with that number?
2. Multiple Linear Regression?
3. OLS vs MLE?
4. R2 vs Adjusted R2? During Model Development which one do we consider?
5. Lift chart, drift chart
6. Sigmoid Function in Logistic regression
7. ROC what is it? AUC and Differentiation?
8. Linear Regression from Multiple Linear Regression
9. 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?
10. Bias Variance Trade off?
11. Over fitting vs Underfitting in Machine learning?
12. Estimation of Multiple Linear Regression
13. Forecasting vs Prediction difference? Regression vs Time Series?
14. p,d,q values in ARIMA models
    1. What will happen if d=0
    2. What is the meaning of p,d,q values?
15. Is your data for Forecasting Uni or multi-dimensional?
16. How to find the nose to start with in a Decision tree.
17. TYPES of Decision trees - CART vs C4.5 vs ID3
18. Genie index vs entropy
19. Linear vs Logistic Regression
20. Decision Trees vs Random Forests
21. Questions on liner regression, how it works and all
22. Asked to write some SQL queries
23. Asked about past work experience
24. Some questions on inferential statistics (hypothesis testing, sampling techniques)
25. Some questions on table (how to filter, how to add calculated fields etc)
26. Why do u use Licensed Platform when other Open source packages are available?
27. What certification Have u done?
28. What is a Confidence Interval?
29. What are Outliers? How to Detect Outliers?
30. How to Handle Outliers?

Company: L&T Financial Services
Role: Data Scientist

1. Explain your Projects
2. Assumptions in Multiple linear regression
3. Decision tree algorithm
4. Gini index
5. Entropy
6. Formulas of gini and entropy
7. Random forest algorithm
8. XGBoost Algorithm
9. Central Limit theorem
10. R2
11. Adj R2
12. VIF
13. Different Methods to measure Accuracy
14. Explain Bagging and Boosting
15. Difference Between Bagging and Boosting
16. Various Ensemble techniques
17. P value and it's significance
18. F1 Score
19. Type 1 and Type II error
20. Logical questions for Type 1 and Type II error
21. Logical questions for Null and alternate Hypothesis

Role: Data Scientist

1. Decorators in Python- Live Example
2. Generators in Python- Live Example
3. SQL Questions
3.1 Group by Top 2 Salaries for Employees
3.2 use Row num and Partition
4. Pandas find Numeric and Categorical Columns.
4.1 For Numeric columns, find the mean of the entire column and add that value to each row of the column.
5. What is Gradient Descent?
5.1 What is Learning Rate and Why is it reduce sometimes

6. Two Logistic Regression Models - Which one will you choose - One is trained on 70% and other on 80% data. Accuracy is almost same.
7. What is LogLoss ?
8. Explain your Projects end to end.(15-20mins)

Role: Data Science Intern

1. Tell me about your journey as a Data Science aspirant
2. What was the one challenging project or task that you did in this domain and why was it challenging?
3. What model did you use for that? I replied Random Forests
4. What is Random Forest and how is it used?
5. How are Random Forest different from Decision Trees and what problems do they solve that decision trees can't?
6. Multi class Classification and which metric is preferred for it
7. Given a banking scenario to predict Loan Defaulters, which metric will you use?
8. How will you handle the class imbalance in this case?

Company: Latentview Analytics

1. What is mean and median
2. Difference between normal and gaussian distribution
3. What is central limit theorem
4. What is null hypothesis
5. What is confidence interval
6. What is covariance and correlation and how will u interpret it.
7. How will you find out the outliers in the dataset and is it always to remove outliers
8. Explain about Machine Learning
9. Explain the algorithm of your choice
10. Different methods of missing values imputation

11. Explain me your ml project
12. How did you handle imbalance dataset
13. What is stratified samplings
14. Difference between standard scalar and normal scalar
15. Different type of visualization in Dl project
16. What architecture have you used
17. Why have u not used RNN in your nlp project
18. Why we don't prefer CNN in nlp based project
19 What is exploding gradient and vanishing gradient and how to rectify it
20. Difference between LSTM and GRU
21. What is precision and recall
22 What is auc metic
23. What if your precision and recall are same

Data Science Interview Questions

1. Naive bayes assumptions
2. What are the approaches for solving class imbalance problem?
3. When sampling what types of biases can be inflected? How to control the biases?
4. GRU is faster compared to LSTM. Why?
5. What is difference between K-NN and K-Means clustering ?
6. How to determine if a coin is biased ? Hint: Hypothesis testing
7. How will u present the statistical inference of a particular numerical column?
8. How would you design a data science pipeline ?
9. Explain back propagation in few words and it’s variants?
10. Explain topic modeling in NLP and various methods in performing topic modeling.

Company: Myntra
Role: Data Analyst

Introduce yourself.
One complex sql query- 2 table are there, Table1(cust_id,Name)
Table2(cust_id,Transaction_amt)
Write a query to return the name of customers with 8th highest lifetime purchase.
Achieve the same using python.
ML questions:
What's the problem in having multi collinearity in data set.
If there is business requirement to keep two corelated features in model, what would you do.
How would you deal with feature of 4 categories and 20% null values
Some questions based on my project.

Company: Nira Finance
Role: Data Scientist

Asked to explain my project.
Have you not done any classification problem as your Resume only mentions regression tasks.

Explain the working of Gradient boosting.
Difference between boosting and bagging.
What would you do when output is imbalanced.
What is more preferred over sampling or under sampling.
In what case under sampling a non harmful approach.
How would you measure the performance of models built on imbalanced dataset.
Whats the meaning of precision and recall.
Tell me about a task you did and are very proud of.
Do you have any questions for me
Post interview I was given a ML assignment to solve and submit within next 2 weeks.

Company: Myntra
Role: Data Analyst
Round type: Use case Round

Problem Statement:

Given 2 teams of Myntra namely:
1. Finance Team: They focus to take decisions which are Money driven.
2. Customer Experience Team : They focus to improve the Customer Experience with Myntra
Whenever Customer places a refund request Myntra can process it in 2 different ways:
1. Directly accept the Return request.
2. Put the request on hold and verify the product for damages or manhandling by customer. Only if the products are found to be in proper state, accept the return.
Now, there is a conflict of opinion between these two teams.
Finance Team likes the 2nd option as it minimize the chances of loss.
But Customer Experience teams likes the 1st option as their main aim is to improve Customer Experience.
Now suppose you are part of the Customer Experience team. How would you convince the Finance team to follow the 1st step.
What kind of Data you would be looking for solving this task.
Is there any need for model building for this use case.

Company: Ericsson
Role: Data Scientist

Round No: 1st Round
How to reverse a linked list
Give a logistic regression model in production, how would you find out the coefficients of different input features.
What is the p- value in OLS regression
What's the reason for high bias or variance
Which models are generally high biased or high variance
Write code to find the 8 highest value in the DataFrame
What's difference between array and list
Whats the difference between Gradient boosting and Xgboost
Is XOR data linearly separable
How do we classify XOR data using logistic regression

Some questions from my previous projects
Given a sand timer of 4 and 7 mins how would you calculate 10 mins duration.
What's the angle between hour and minute hand in clock as 3:15

Company: Legato Health Technologies
Role: MLOps Engineer
Experience: 2-3 years

Round 2:
Complete ML technical stack used in project?
Different activation function?
How do you handle imbalance data ?
Difference between sigmoid and softmax ?
Explain about optimizers ?
Precision-Recall Trade off ?
How do you handle False Positives ?
Explain LSTM architecture by taking example of 2 sentences and how it will be processed?
Decision Tree Parameters?
Bagging and boosting ?
Explain bagging internals
Write a program by taking an url and give a rough code approach how you will pass payload and make a post request?
Different modules used in python ?
Another coding problem of checking balanced parentheses?

Role: Junior Data Scientist

1) Explain the architecture of CNN
2)If we put a 3×3 filter over 6×6 image what will be the size of the output image
3) What will you do to reduce overfitting In deep learning models
3) Can you write a program for inverted star program in python
4)Write a program to create a dataframe and remove elements from it
5) I have 2 guns with 6 holes in each, and I load a single bullet In each gun, what is the probability that if I fire the guns simultaneously atleast 1 gun will fire (atleast means one or more than one)

6) There are 2 groups g1 and g2, g1 will ask g2 members to give them 1 member so thay they both will be equal in number, g2 will ask g1 members to give them 1 member so thay they will be double of g1, how many members are there in the groups (I'm not sure of this question as I tried to solve but didnt get correct answer)

Data Science Interview Questions:

1. How do check the Normality of a dataset?
2. Difference Between Sigmoid and Softmax functions?
3. Can logistic regression use for more than 2 classes?
4. What are Loss Function and Cost Functions? Explain the key Difference Between them?
5. What is F1 score? How would you use it?
6. In a neural network, what if all the weights are initialized with the same value?
7. Why should we use Batch Normalization?
8. In a CNN, if the input size 5 X 5 and the filter size is 7 X 7, then what would be the size of the output?
9. What do you mean by exploding and vanishing gradients?
10. What are the applications of transfer learning in Deep Learning?
11. Why does a Convolutional Neural Network (CNN) work better with image data?

Data Science Interview Questions:

1. What is the Central Limit Theorem and why is it important?
2. What is the difference between type I vs type II error?
3. Tell me the difference between an inner join, left join/right join, and union.
4. Explain the 80/20 rule, and tell me about its importance in model validation.
5. What is one way that you would handle an imbalanced data set that's being used for prediction (i.e., vastly more negative classes than positive classes)?
6. Is it better to spend five days developing a 90-percent accurate solution or 10 days for 100-percent accuracy?
7. Most common characteristics used in descriptive statistics?
8. What do you mean by degree of freedom?
9. Why is the t-value same for 90% two tail and 95% one tail test?
10. What does it mean if a model is heteroscedastic? what about homoscedastic?

11. You roll a biased coin (p(head)=0.8) five times. What's the probability of getting three or more heads?
12. What does interpolation and extrapolation mean? Which is generally more accurate?

Data Science Interview Questions:

1. What the aim of conducting A/B Testing?
2. Explain p-value.
3. Explain how a ROC curve works?
4. What is pruning in Decision Tree?
5. How will you define the number of clusters in a clustering algorithm?
6. When to use Precision and when to use Recall?
7. What are the assumptions required for linear regression? What if some of these assumptions are violated?
8. How are covariance and correlation different from one another?
9. How can we relate standard deviation and variance?
10. Explain the phrase "Curse of Dimensionality".
11. What does the term Variance Inflation Factor mean?
12. What is the significance of Gamma and Regularization in SVM?

Data Science Interview questions:

How will you calculate the Sensitivity of machine learning models?
What do you mean by cluster sampling and systematic sampling?
Explain Eigenvectors and Eigenvalues.
Explain Gradient Descent.
How does Backpropagation work? Also, it states its various variants.
What do you know about Autoencoders?
What is Dropout in Neural Networks?
What is the difference between Batch and Stochastic Gradient Descent?
What are the different kinds of Ensemble learning?
What is entropy, information gain and gini index in decision tree classifier and regression?

Role: Data Scientist

1. What is central tendency
2. Which central tendency method is used If there exists any outliers
3. Central limit theorem
4. Chi-Square test
5. A/B testing
6. Difference between Z and t distribution (Linked to A/B testing)
7. Outlier treatment method
8. ANOVA test
9. Cross validation
10. How will you work in a machine learning project if there is a huge imbalance in the data
11. Formula of sigmoid function
12. Can we use sigmoid function in case of multiple classification (I said no)
13. Then which function is used
14. What is Area under the curve
15. Which metric is used to split a node in Decision Tree
16. What is ensemble learning

Role: Data Scientist

1. What is central tendency
2. Which central tendency method is used If there exists any outliers
3. Central limit theorem
4. Chi-Square test
5. A/B testing
6. Difference between Z and t distribution (Linked to A/B testing)
7. Outlier treatment method
8. ANOVA test
9. Cross validation
10. How will you work in a machine learning project if there is a huge imbalance in the data
11. Formula of sigmoid function
12. Can we use sigmoid function in case of multiple classification (I said no)
13. Then which function is used
14. What is Area under the curve
15. Which metric is used to split a node in Decision Tree

16. What is ensemble learning
17. 3 situation based questions

Company: Legato Health Technologies
Role: MLOps Engineer

Round 2:
Complete ML technical stack used in project?
Different activation function?
How do you handle imbalance data ?
Difference between sigmoid and softmax ?
Explain about optimisers ?
Precision-Recall Trade off ?
How do you handle False Positives ?
Explain LSTM architecture by taking example of 2 sentences and how it will be processed?
Decision Tree Parameters?
Bagging and boosting ?
Explain bagging internals
Write a program by taking an url and give a rough code approach how you will pass payload and make a post request?
Different modules used in python ?
Another coding problem of checking balanced parentheses?

Company: Cerence
Role: NLU Developer

Question1 :
Write a function that take two strings as inputs and return true if they are anagrams of each other and false otherwise
e.g.
(hello, hlleo) --> true
(hello, helo) --> false

Question 2 :

Write a function that take an array of strings "A" and an integer "n",
that return the list of all strings of length "n" from the array "A" that can be constructed as the concatenation of two strings from the same array "A"
e.g.
A = [dog, tail, sky, or, hotdog, tailor, hot] and n=6
output should be "hotdog" and "tailor"

Question 3 :
Given an array "arr" of numbers and a starting number "x",
Find "x" such that the running sums of "x" and the elements of the array "arr" are never lower than 1.
e.g.
arr = [-2, 3, 1, -5].
The running sums will be x-2, x-2+3, x-2+3+1 and x-2+3+1-5.
So, the output should be 4.

Company: GEOTAB

Python :
1. Is python a language that follows pass by value, or pass by reference or pass by object reference
2. What are lambda functions and how to use them
3. Difference between mutable and immutable objects with example.
4. What are Python decorators? Why do we use them

SQL :
1. What is the difference between Inner join and left inner join ?
2. What are window functions ?
3. What is the use of groupby ?

SQL Round
3 tables given as below:
TRIPS
trip_id
vehicle_id
start_time
stop_time

VEHICLE_MAKE

vehicle_id
make_id

MAKES
make_id
make_name

There is a table which contains vehicle trips. Trips are not necessarily in order.

There is a table which contains vehicle makes. Makes are not necessarily known.

PROBLEM: Write SQL code that provides the number of trips that started on September 1st, 2020 for each vehicle with a KNOWN make.
Order the results by the trip count.

op
vehicle_id | trip_count
4 | 2
1 | 1
2 | 1

Role: MLOps Engineer

1st round:-
Introduction
Current NLP architecture used in my project
How will you identify Data Drift? Once identified how would you automate the handling of Data Drift
Data Pipeline used
Fasttext word embedding vs word2vec
When should we use Tf-IDF and when predictive based word embedding will be advantageous over Tf-IDF
Metrics used to validate our model
In MongoDB write a query to find employee names from a collection
In Python write a program to separate 0s and 1s from an array- (0,1,0,1,1,0,1,0)

Company: Latentview.

Initial they had asked for the explaining the project which I had done. I explained the Customer prediction case . Then I was asked with python questions by sharing my screen.

1. How do you handle the correlated variables without removing them
2. Explain the SMOTE, ADAYSN technique
3. What is stratified sampling technique
4. Explain the working of random forest and xgboost
5. How do you optimise the Recall of your output
6. What are chisquare and ANOVA test
7. In python they asked for LOC,ILOC, how do you remove duplicate,How to unique values in column,
8. In SQL they asked for the query for having matches between different teams

Company: Myntra
Role: Data Analyst

Introduce yourself.

One complex sql query- 2 table are there, Table1(cust_id,Name)
Table2(cust_id,Transaction_amt)
Write a query to return the name of customers with 8th highest lifetime purchase.
Achieve the same using python.
ML questions:
What's the problem in having multi collinearity in data set.
If there is business requirement to keep two corelated features in model, what would you do.
How would you deal with feature of 4 categories and 20% null values.

Company: Enquero Global
Role: Data Scientist

Previous job role and responsibilities
Problem statement of your project and How do you overcome challenges
How do you handle feature which had many categories.
When to use precision and recall.
What are outliers & how do you handle them
Joins, self joins, said me to write sql queries on self joins
How good your with python
Logic for reverse string.
Data collection- how do you collect data and data preprocessing.
Focused on EDA part
Have you deveployed any project in cloud if you which cloud you had used and how do you do that.
How do you interact with domain expertise and business analytics people.
How do you replace Missing values for continuous variables.
What is your largest data set you have handled till now and tell me size of dataset.
Overall Mostly focused on SQL, DATA COLLECTION, EDA, feature engineering and selection.

A continuous variable is having missing values, so how will you decide that the missing values should be imputed by mean or median?
What is PCA and what each component means? Also, what is the maximum value for number of components?
What is test of independence? How do you calculate Chi-square value?
When precision is preferred over recall or vice-versa?
Advantages and disadvantages of Random forest over Decision Tree?
What is the c hyperparameter in SVM algorithm and how it affects bias variance tradeoff?
What are the assumptions of linear regression?
Difference between Stemming and Lemmatization?
Difference between Correlation and Regression?
What is p-value and confidence interval?
What is multicollinearity and how do you deal with multicollinearity? What is VIF?
What is the difference between apply, applymap and map function in python?

Deloitte Interview :

Role: Data Scientist
Candidate Name: Wanted to remain anonymous

ROUND 1 :

Introduction

- Started with Classification particularly Imbalance , oversampling.
Which class should i oversample etc.
Telecom Churn Case Study Questions like Evaluation metric for imbalance data
what threshold to choose to diving the classes (0.5 in case of balanced else sensitivity / Specifivity etc.
What if i don't use SMOTE() for handling imbalance how should i select the threshold now (messed up by me, roc , auc etc) Ans = Presion - Recall Curve

- NLP Questions
Sentiment analysis, preprocessing like (TFID, BOW), Embeddings, stemming, Lemmatization
libraries in know : nltk, spacy

- Regression Preprocessing
answered outlier, missing value immputation, Distribution, dummies, multicolinearity etc
You have two highy co-related columns which one will you drop? : "Based on Business Problem i will see accordingly.",

- Naive Bayes Explanation , Drawback of Naive Bayes(couln't answere drawback of Naive Bayes, 'Assume all are independent', him)
- Hand Gesture Recognition Techniques (End to End)
- Resource Timesheet Forecasting . (What is it?? what you do on this?, " Explained with a story based on what i do in TCS".

- Do you know any Boosting Algorithms : YES
where have you used?? in Telecom Churn and Healthcare Analytics by AV

- Gradient Descent (How it works)

- KNN related. How do we choose value of K ??

- Stastical Computing:
Type 1 and Type 2 error
Alternate name of Type 1 error (couldn't answer alternate name of Type 1 error, 'False +ive, him)
What is p-Value (Explaiend with the example of Linear Regression from statsmodel)

- Do you have exposure of TimeSeires analysis : NO (didn't ask anything and seems fine with him)

My checklist before going for an SQL round of interview:

1. WHERE , AND, OR, NOT, IN
2. ORDER BY, ASC, DESC
3. IS NULL
4. LIMIT
5. MIN, MAX, COUNT, AVG, SUM
6. LIKE, WILDCARDS
7. IN BETWEEN
8. INNER JOIN
9. LEFT JOIN
10. Subqueries(most important)
11. UNION
12. GROUP BY
13. HAVING
14. LEFT, RIGHT, MID, CONCAT
15. PARTITION BY, OVER
16. LEAD,LAG
17. RANK, DENSE_RANK, PERCENT_RANK
18. ROW_NUMBER, CUME_DIST
19. FIRST_VALUE, LAST_VALUE
20. AS

Company wise Data Science Interview Questions

Follow me for more resources

@arjun-panwar
https://www.linkedin.com/in/arjun-panwar

My Statistics Checklist before going for a Data Science Interview:

1. Inferential and descriptive Statistics
2. Sample
3. Population
4. Random variables
5. Probability Distribution Function
6. Probability Mass Function
7. Cumulative Distribution Function
8. Expectation and Variance
9. Binomial Distribution
10. Bernoulli Distribution
11. Normal Distribution
12. Z-score
13. Central Limit Theorem
14. Hypothesis Testing
15. Confidence Interval
16. Chi Square Test
17. Anova Test
18. F-Stats

Some Data Science Companies(not ranked) for Job Hunting:

1. Genpact
2. Tredence Analytics
3. Fractal Analytics
4. Tiger Analytics
5. Bridgei2i
6. Ugam
7. Latent View
8. Brillio
9. Abzooba
10. AbsolutData
11. Gramemer

12. BluePi
13. Knowledge Foundry
14. Wipro
15. TCS
16. Accenture
17. Purplle
18. AbsoluteData
19. Hansa CEquity
20. Lymbyc
21. IBM
22. PwC
23. EY
24. KPMG
25. Sibia
26. ZS
27. ZF
28. TechVantage
29. L&T Infotech
30. Cognizant
31. Amazon
32. Microsoft
33. Walmart
34. Philips
35. Ford
36. JP Morgan
37. Deloitte
38. Shell
39. Mu Sigma
40. Postman
41. Altrix
42. HP
43. HCL
44. Dell
45. Paypal
46. Fidelity Investments
47. Rakuten
48. Infosys
49. Flipkart
50. Myntra