**Interview Questions**

Company: Axtria

Role : Data Scientist

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: Altimetrik

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. Two Logistic Regression Models - Which one will you choose - One is trained on 70% and other on 80% data. Accuracy is almost same.

6. What is Log-Loss and ROC-AUC?

7. What is Multi-collinearity? How will you choose one features if there are 2 highly correlated features? Give Examples with the techniques used.

8. VIF – Variance Inflation Factor – Explain.

9. Do you know to use Amazon SageMaker for MLOPS?

10. Explain your Projects end to end (15-20mins).

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?

Role: 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)

.

.

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?

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?

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: 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

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: Latentview Analytics

Role : Data Scientist

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

Company : Unknown

Role : 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.

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

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

Company :Verizon

Role: 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: 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: 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: 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