GRADIENT ASCENT

A comprehensive intern guide
on Data Science and Analytics

Consulting & Analytics Club
IIT Guwahati

Resources

Topic	Link
Linear Algebra	Link
Stats/Probability	Link
DeepLearning.AI Specialization Course Notes	Link
Linear Regression	Link
Logistic Regression	Link
Multi-Class vs Multi-Label Classification	Link
Multi-Label Classification	Link
Multi-Class Classification with Imbalanced Dataset	Link
Naive Bayes	Link
Bias Variance Trade-off	Link
Support Vector Machine	Link
Support Vector Machine Code	Link
Ensemble Methods: Bagging, Boosting and Bootstrapping	Link
Feature Engineering for ML Models	Link
Principal Component Analysis	Link
T-distributed Stochastic Neighbor Embedding(t-SNE)	Link
K-means Clustering	Link
K-means Clustering Code	Link
K-Nearest Neighbour(KNN)	Link
KNN Code	Link
Feature Engineering in Images	Link
All about Natural Language Processing(Watch according to your needs)	Link
Feature Scaling	Link
Gaussian Distribution	Link
Mini-Batch Gradient Descent	Link
Gradient Descent with Momentum	Link
Grid Search	Link
Batch Normalization	Link
Recurrent Neural Network (RNN)	Link

Resources

Topic	Link
Long-Short Term Memory (LSTM)	Link
Different Types of Losses and significance	Link1 Link2 Link3 Link4
Evaluation Metrics and their significance in particular cases	Link1 Link2 Link3 Link4
Regularization and Optimization Techniques (Blogs related to Andrew Ng (DL Course 2)	Link1 Link2 Link3
Ensemble Models	Link1 Link2 Link3
Data Handling (Train, Dev and Test)	Link1 Link2 Link3
Machine Learning Case Studies	Link
Interview Prep Playlist (Krish Naik) - Checkout his other relevant playlists too	Link

Interview Experiences:

1. **Publicis Sapient**
    - Daksh Kaushik | 2023
    - Lokesh Nahar | 2023
2. **Decimal Point Analytics**
    - Amish Agarwal | 2023
3. **Providence**
    - Akshay Chintala | 2023
4. **Nimbleedge**
    - Umang Jain | 2023
5. **Adobe**
    - Pragyan Banerjee | 2023
    - Varun Yerram | 2022
6. **Amazon**
    - Debarshi Chanda | 2022
    - Shreya Sajal | 2022
7. **Microsoft**
    - Arsh Kandroo | 2022
    - Roshan Shaji | 2022
    - Rishon D' Souza | 2022
8. **Fractal Analytics**
    - Ishant Khurana | 2022
    - Aryan Meshram | 2022
9. **Envestnet Yodlee**
    - Pranjal Verma | 2022

Publicis Sapient

Daksh Kaushik (Role: Data Scientist | LinkedIn)

INTERVIEW PROCESS:

Round 1

Duration ~ 60 mins

- A DSA question based on shortest path DP. Although DSA questions don't carry much weight in the data science interview still you should try to crack these questions.
- Moving on they asked me about my project, I told them one of my projects which was license plate detection where me and my peers used the YOLOV5 algorithm. So they asked me about the same in deep. I want to say that whatever project you do you must have a deep understanding of the underlying algorithms used in that project.
- There were questions about situational ML which I had to answer on the spot. These types of questions are easy if we know the model-building pipeline. (examples from Kaggle)
- The interviewer asked me to estimate the number of customers coming to the mall in a year. He gave me some parameters and asked me which algorithm I would use for the same, how will you check outliers, and check the efficiency of the model.

Then we moved on to some standard HR questions.

Publicis Sapient

Lokesh Nahar(Role: Data Scientist | LinkedIn)

INTERVIEW PROCESS:

Round 1

Duration ~ 60 mins

- I was asked about OOP's, DBMS particularly SQL, and the Basic difference between SDE and data science.
- They also asked some technical questions on ML like Bias Variance, supervised unsupervised learning, Use of NLP, and what all was covered in the NLP course I took(like word embeddings, Word2Vec)
- Then we moved towards project discussion. They asked me about my favorite project(about face recognition)and then questioned me about the same.
- Then they asked some basic SQL questions along with some jee types probability questions. I also told them about my JEE Adv rank and some research work I was doing. I also told them about my case study under C&A.
- Then we discussed in what ways I can go deeper into the fields. For e.g., We can use Recommender systems and tools like NoSQL, Azure, etc.
- Be aware as well as humble in the HR round. , they asked me about 5 company values. It was mainly based on a resume. Sidenote puzzles can also be useful

Decimal Point

Amish Agarwal (Role: ML Intern | LinkedIn)

INTERVIEW PROCESS:

Round 1: Aptitude Test and Coding Test

- He started with a basic introduction and then moved on to my Resume.
- He asked about one of the projects and asked me to explain it to him. Ensure you know everything about your project and can respond with any cross-questions based on implementation details. Whatever I was explaining, he cross-questioned me on some things related to the project. This went on for about 20 mins.
- After that, he asked me about ML basics. Questions include the difference between bias and variance, what we mean by high bias/ variance, how it can be tackled, and so on. He expected a fundamental understanding of the concepts, cross-questioning me whenever I used a big word and asking me to break it down in simpler terms.

(Try to use examples to explain your point, I used the whiteboard to draw and explain high bias/variance; similarly, you can take the help of such tools to explain yourself.)

- Finally, he asked one question that he was asking everyone in the end. It was based on PCA. On the whiteboard, he made 2D axes and plotted some data points. Then he drew 3 lines naming them PC1, PC2, and PC3. He asked which ones were the principal components of the data. The one capturing the max variance(you can look at the diag and tell) was the first, and we know that PCs are orthogonal to each other, and for n-dimensional space, there is max. N PCs, so the 2nd PC was perpendicular to the 1st one, and there is no 3rd PC.
- In the end, he asked me if I had any questions, so I asked about the kind of projects that are allotted to interns.

Decimal Point

Round 2 (HR Round)

I got a call from HR after all the other interviews were completed. It was basically an HR round in which they asked typical questions like what are your strengths, gave you a situation, and asked what you will do. It was just formality, and it meant that you were selected.

Providence

Akshay Chintala (Role: ML Intern | LinkedIn)

INTERVIEW PROCESS:

Round 1: Aptitude Test and Coding Test

- They asked me what I'm good at to which I said probability. They asked some basic questions in probability and statistics.
- They asked linear algebra although not that much.

Round 2: Interview

- They asked me about the project I made.
- Mainly stuff like what is the purpose of the project and the problems I faced during it.
- I made a project on Braille to speech converter which was a CNN model and I got questions related to that project only.
- They asked me first what the project does and why I made it. They asked what problems I faced when I was working on it and what was my contribution in that project.

Nimbleedge

Umang Jain(Role: ML Intern | LinkedIn)

INTERVIEW PROCESS

Round 1:

- In the first half, he asked me various ML-related questions, like Regularization, Bias Variance, Kmeans, and KNN, etc.
- In the second half, he asked me to explain him any one of my projects. So I had a project which was a combination of both NLP and CV domains, so I explained that project to him , and he asked me various questions that he had about the project.
- This round wasn't that difficult (Also they were mainly searching for people with NLP backgrounds)

Round 2:

Duration ~ 90 mins

The second round was with the founder. In that round, they were mainly checking how strong the basic knowledge is, and they asked me several situation-based questions.

- Suppose you have two decision trees D1 and D2, and a dataset T1 and T2 which were a subset of a common large dataset. D1 and D2 are trained on T1 and T2 respectively. Now you have to make a third decision tree D3, such that it is robust to both datasets T1 and T2, but T1 and T2 are not accessible to us, and only have access to D1 and D2. This question was asked to check the thinking in case of an unseen problem. There was no right or wrong answer in this case.
- Another question he asked me was, the difference between SGD and GD, which one is better, and why it is better, on an intuitive understanding

Nimbleedge

- Next was, to consider SVM. Suppose we train an SVM multiple times, with different random initializations at the start, still, why do we get the same weights once the training is complete every time? The answer was, that the loss function of SVM is Convex, having a global minimum, and SVM always reaches the global minimum perfectly, giving the same weights at the end.
- Then he asked me to write pseudo code of a two-layered neural network for binary class classification, from scratch, both forward and backward propagation.

Adobe

Pragyan Banerjee(Role: MDSR Intern | LinkedIn)

INTERVIEW PROCESS

Round 1:

Duration ~ 45 min

- I already had a research internship before that, so for the first 30 mins , the discussion mainly revolved around that only , we discussed what I did in that internship and what more could have been done. They asked me what was the outcome of the project? Did anyone continued the work after my internship was over, or if there was any chance of a publication.
- Then they asked simple questions on CNN (I told them that I was primarily interested in CV)
- They also asked me a question in which there were 2 matrices each representing part of an image ... What could have been the best ways to concatenate them to make a single image
- There were lots of variations to this question... And we sort of had a discussion on this topic. Discussed some ways to do them and their pros and cons.
- Simple stats-related questions like suppose there is a flow of data (data that is coming real-time), how to calculate its mean and variance on the go (real-time).

I'd suggest focusing on the basics and how you'd apply them to solve simple real-world problems because apart from projects they usually ask basic questions only.

Adobe

Varun Yerram (Role: MDSR Intern | LinkedIn)

INTERVIEW PROCESS

Round 1 (ML Round)

Aptitude, Probability and Statistics

• Explanation of one of the projects, with cross questions on it.
• From a stream of input numbers, find the following at any point of time:
    • Mean
    • Variance
• A Sample number from the stream with probability 1/n (n = number of numbers read)
• Given an unbiased coin, design an experiment to select one of the three numbers uniformly:
    • Follow up - what type of distribution will it form.
    • Follow up - Calculate the probability mathematically and prove that it converges to 1/3.
• Given a sequence of heads, calculate and find the number of times you need to flip to get HH.
• Given a matrix where rows and columns are sorted independently, design an algorithm to search a number in an optimal way.(O(2n) i.e. O(n) possible)
• Given an unknown matrix A(nxn). You are allowed to choose a column vector X (nx1) and get the output result. What is the minimum number of times, you need to choose vectors to know the values of the complete matrix (Ans - n)

ML/AI Questions Based on Project

• Explain cross connections in UNET and what is the difference between them and residual connections in ResNet.

• Are the parameters shared in a Unet encoder and Unet Decoder, what is the effect of sharing them?
• Explain Backward pass of Unet, more specifically how do gradients propagate through decoder to encoder.

Other ML/AI Questions

• Given an RGB Image, and 64 kernels of 5x5 size. You apply the convolution operation with stride 2 and padding 0. What is the number of weights that are updated in every pass of the model? How will the answer change if you add bias to convolutions?
• Imagine you have a GPU and you are training convolutional networks, the GPU can fit a maximum batch size of 64 images at a time. You need to reproduce a paper in which they have used a batch size of 128 images. How would you do it? (Ans - Explain gradient accumulation in detail)
• What does loss.backward() do in pytorch? how will its functioning change in the above question. (Depending on your choice of framework question may vary)
• At last I was asked to explain an NLP project I had done, as he had an NLP background.

Amazon

Shortlisting:

30 July - 1 August: Amazon ML Challenge (Problem Statement)
4 August: Amazon ML Challenge Finale (Top 10 teams in the Hackathon)

1. Debarshi Chanda (Role: Applied Scientist Intern | LinkedIn)

INTERVIEW PROCESS:

Round 1: SDE Round (Coding/DSA)

TL;DR: Trees + DP Duration ~ 45 mins

Q. Write a function to print the Spiral Order Traversal of a Binary tree.
A. Video Link

- Already knew the solution, said the solution using 2 stacks.
- Interviewer asked if the same can be implemented using a single queue.
- Said about doubly ended queue but hadn't used it before so told the interviewer honestly that I hadn't used deques before. Went on to code the 2 stack solution.

Q. An array is given, Find the length of the subarray having maximum sum.
A. Told the Brute Force approach: *O(n^2)*
Modified Kadane's algorithm to maintain length of the array as well.

Round 2: ML Round

Interviewer will check your knowledge of ML Depth & Breadth

ML Depth: Project + ML Design ~ 30 minutes
ML Breadth: ML Fundamentals ~ 40 minutes

Amazon

Interviewer was constantly refining my thoughts and answers.
Expect a lot of follow-up questions.

Discussion on Project (Amazon ML Challenge)

- Explain Problem Statement of Amazon ML Challenge
- Explain the Approach, Follow-up: Mathematical formulation of Arcface, told him didn't know and gave the intuitive explanation.
- Asked me about the performance of various models we tried.
- Follow-up: XLNet performed better than BERT. Why? Said about Masked Language Modelling vs Permuted Language Modelling.

ML Design (Interviewer said that there was no right or wrong answer)

- Suppose you want to use all the columns (Title, Description, Bullet Points).What can you do?
    1. 1st Design: Use 3 BERT models for the 3 columns respectively, concatenate embeddings and use a fully connected layer.
    2. 2nd Design: Use 1 BERT and use [SEP] token as a separator between different columns.
- Pros and Cons of Both the design.
- Which model will you choose for your leaderboard? 1st
- Is there any case when the 2nd model will outperform the 1st model? When the columns have a similar distribution.
- Which model should we use if we only have 3000 rows instead of 30 lakh?

Amazon

ML Fundamentals

- Metrics: Suggest a metric other than accuracy. F1-score. Follow-up: Define F1- score. Define Precision or recall. How would you calculate F1-score for 10000 classes?
- Suggest something other than F1-score. AUC score. Follow-up: Define AUC. Statistical significance of AUC.
- Regularization in NN: Dropout, How to deal with different behavior of dropout in train and test time. Scaling.
- Batch-Normalization: What is Batch-Normalization, Why is it useful? Same scale and minimize covariate shift.
- Difference between SGD, Batch Gradient Descent and Mini-batch Gradient Descent.
- Why does mini-batch GD and SGD work despite only using a very small number of samples? Come from the same distribution. Follow-up: How can we check the quality of this estimate? Mean and standard deviation can be a good way to check quality. Is this estimate biased or unbiased? How to get an unbiased estimate?
- Linear Models: Difference between L1 and L2 regularization. Why L1 induces sparsity and L2 doesn't? Why use L2 at all?
- Can linear models overfit to the training data, it is just a line(Outliers)
- Unsupervised: Tell any Unsupervised learning algorithm to decompose 10000 dimensions in 10. PCA. Explain PCA in 2 lines. How do you get the eigenvectors? You have the 10 eigenvectors, now how do you convert the 10000 dimensions to these 10 dimensions.

Amazon

2. Shreya Sajal (Role: Applied Scientist Intern | LinkedIn)

INTERVIEW PROCESS

Round 1: SDE Round (Coding/DSA)

Duration ~ 45 mins

2 DSA questions, interviewer shared the link to the live coding platform. Fortunately I was able to solve both.

- Intro
- Longest Common Substring (brute force, then explained DP approach, Time-space complexity, code and dry run)
- Density of Binary Tree in one traversal (started thinking using DFS but BFS approached clicked at the right moment, wrote the code)

Round 2: ML Round

Duration ~ 1 h 10 mins

The interviewer started with his intro, he was a Senior Applied Scientist at Amazon and he mainly worked in the NLP domain.

- Moved onto my intro, ML exposure overview, Brief overview of 2 CV projects + Amazon ML Challenge approach and my contribution.
- Went on with discussion on my Amazon ML Challenge approach further as we had used 2 approaches (End-to-End Classification and Transformer model head with KNN, and and ensemble for the final predictions):
  - What was KNN over embeddings exactly doing (Inference time Euclidean distance calculation and voting).
  - How did you train 3M points? (Stratified K-Fold)

Amazon

- If you didn't have any memory limitations and and could train all the points, then I claim End-to-End Classification should work better than KNN + embeddings. Justify my claim. (He assumed that the extracted embeddings were not finetuned and we were directly using the pretrained weights for embedding extraction, so I just justified it based on his assumption only as it was easier that way).
- If the max_length was some fixed value for your classification model that couldn't be changed and text length > max_length what would you do? (Trim- info loss, so best I could come up with text length // max_length rows for one row with labels same for training, max vote for segments during inference, he seemed satisfied).
- What else could you have done if there was no such time limit for the hackathon? (Stack embeddings + KNN and few more points from our future prospects sections mainly).
- What if you were not allowed to use BERT or any transformer based models? (I explained him a solution using Bi-LSTM many-to-one, the main focus was on the basic architecture not much depth into the u, f, o gates, and also he focused on the output dimensions and loss function for training).
- What if it was multilabel and not multiclass? (He mainly wanted to know the output layer dimensions, activations and loss function to be used)
- How are word embeddings trained? As I had mentioned about word embeddings in my Bi-LSTM approach, he wanted me to explain the training of any word embedding of my choice. (I explained him word2vec both skipgram and CBOW variants, how is the input output taken and last layer activations mainly).

Amazon

- Follow-Ups: If the vocab is very large then the training will be very expensive because the softmax denominator was to be calculated over the entire vector? How is that dealt? (I mentioned hierarchical softmax is used as a solution to this but I didn't know how it works exactly so couldn't tell that)
- You have a sentence and you have to do POS tagging for every word in that (3 POS were there noun, verb and pronoun. I started with basic RNN approach and explained, again the main focus was on output activations, dimensions and loss function at every time step)
- Design a search engine for Amazon, you have product descriptions and a text query given. (I explained a sentence embedding + semantic similarity + ranking approach, he was satisfied). Follow-Ups: What if you had a binary column as well for every product? (I first said that we can multiply each row's binary with it's cosine similarity we had calculated, so the disliked one's get filtered, he said it was okay as a baseline but there will be some info loss in this, I then said we can change the weight from 0 to 1 and from 1 to 2, then none of the rows will be completely lost and the not similar-liked would not get an edge over the similar-disliked ones mostly, he sounded satisfied)
- Explain BERT architecture (I explained mainly how it is trained and the tasks it is trained on- NSP and MLM and the input it takes, input word embeddings and some basic transformer encoder stack attention overview, he was satisfied).
- Follow-Ups: What is masked language model? What if the masking was not done? (I couldn't answer) What is language modelling?

Microsoft

Shortlisting (Coding Test)

- Section 1 (Python): Write a custom distance function (in python) based on a given formula to be used for the KNN algorithm of sklearn. This was an easy question solved by almost everyone. No knowledge of sklearn was necessary.

- Section 2 (Sklearn): This was a very lengthy question. We were given test and training datasets on which we had to train a Decision tree classifier (all parameters given) and get predictions. We also had to do this while using Cross-validation and identify the best model (from K folds) and store the best parameters of those models. There were many such complicated tasks. While we were allowed to use the documentation and try the code on our local machine, being familiar with the documentation of sklearn is very important.

- Section 3 (MCQs): 17 MCQs on ML, confusion matrix etc. No questions from DL, Probability, Statistics. Easy.

1. Arsh Kandroo (Role: Data Scientist | LinkedIn)

INTERVIEW PROCESS

Round 1

- Picked up a project from my resume, and asked me to explain the approach. Inquired about the model that I used and how it works, which was XGBoost.
- Basic questions on Gradient Descent (slope of derivative etc)
- Asked me to share the screen and write the formula for Gradient Descent optimization with momentum. Further, he asked questions about the intuition of the same and the significance of every term in the formula. Moving ahead, he asked me to recursively open up the formula and asked some questions based on that.

Microsoft

- Basic questions about Mean, Variance and their formulas.
- Eigenvalues and Eigenvectors, positive definite matrix and Hessian matrix.
- Lastly, he asked me to share my screen again and code the function for cosine similarity between two vectors.
- Interview ended with me asking him about his projects and stuff

This guy's camera was off and didn't seem jovial. So, I tried to talk to him in the end about his projects and work to ease things off a bit.

Round 2

Duration ~ 50 mins

- Discussed 2 projects from my resume thoroughly, from motivation to the models used. Make sure you know the functioning and intuition behind each and every model or the library that you have used in the project. Since one of my projects was based on NLP, which resulted in him asking me questions about BERT and LSTMs. Also, there were some questions based on shortcomings of the project; make sure you have already tapped them and you should know how to overcome them.
- Gave me a situation in which I've to classify queries on a search engine as technical or not. Told him my approach and the models that I could use for better results.
- Asked me to share my screen and gave me a small DataFrame. He wanted me to implement an easy method in pandas on the DataFrame. Though my answer wasn't what he was expecting but he was satisfied.
- Again, the interview ended with me asking him about his work in the MS and what team he has been working under

This guy was really cool. Smiled throughout the interview. He had done his MTech in Data Science from IIT Madras. Also, we talked about his journey in MS.

Microsoft

Round 3

Duration ~ 50 mins

- Started with him asking me about my journey in ML and explaining all my projects in chronological order.
- He picked up a project from my resume and asked me to tell its shortcomings and how'd I handle them.
- Gave me an ML case study: "Whenever we search diseases related to symptoms on a search engine, it may show us extreme diseases on top, hiding the real disease. We want to develop an app when given symptoms as input shows us all the probable diseases along with their probabilities."
- I explained my approach from data collection to deployment. He asked me the challenges I'll face and how will I handle them.
- Towards the end, we talked about life at MS and the projects that he had worked on. Also, I asked him for feedback of my interviews, which turned out to be pretty good.

This guy was really cool as well with 17 years of experience at MS.

General Tips

- Whenever explaining your approach, make sure you start from the beginning like data collection and move ahead step by step towards data cleaning, preprocessing, model development and metrics.
- Make sure to tap the shortcomings of your project beforehand and the answers to tackle them.
- Communicate clearly and try to discuss thoughts with your interviewer. He/ She will definitely guide you towards the correct answer.

Microsoft

2. Roshan Shaji (Role: Data Scientist | LinkedIn)

INTERVIEW PROCESS

Round 1

Duration ~ 40 mins

- First asked me about one of my projects. Described what it was, when it was done(as part of Summer Analytics ka final hackathon). Described what was special about the data(duplicates, class imbalance). Told him I used SMOTE to solve the imbalance. Described how it works in brief. He asked a counter question, which was actually one of the drawbacks of the SMOTE method.
- Asked if I know Python. Said yes, but not OOPS. Asked me to code a basic function on Notepad (find cosine similarity between two vectors).
- Asked me about momentum in gradient descent. I did not remember the formula(and I told him that). Opened Paint and drew the diagram of what momentum does. He realized I knew how it works, so he told me the formula and asked me to explain why momentum works(i.e. Why those mathematical formulas speed up gradient descent). After a few hints, I was able to solve it.
- Asked a question on statistics, was not able to answer it satisfactorily so he moved on. Asked me to write the formula for expectation, variance of a Continuous Random Variable.
- Asked me if I have any questions for him, I asked him about the work he does.

Round 2

- Chillest interview ever. Interviewer was a friendly, young guy. Asked me to describe one of my projects and what I had done(data cleaning). Explained the scraping and cleaning part in detail.
-

Microsoft

- What is supervised, unsupervised, semi-supervised. Give a real life example of supervised(eg: learning what each color is called)
- Explain bias and variance. Explained it, and drew the graph of bias-variance tradeoff.
- What are activation functions? Why use them? List the common ones.
- Gave me a diagram showing some data in 2 dimensions, asked if I would use logistic regression or decision trees for classification. The key was that the data points were not linearly separable.
- Suppose you want to detect diabetes, cancer, heart disease etc. using neural networks.
- A person could have multiple diseases. Do you build one model for each disease or one model to detect all the diseases? Why?
- Most interesting question: Given the weight, length, region and color of a mushroom, predict if it is poisonous. For such questions, make sure to ask questions(do we already have the data? If yes, do we have a lot of data? Do we have time and resource constraints? etc.). Told him my approach, how I'd preprocess(nulls, duplicates, outliers), what model I'd use(and which ones you would not), and why. Make your thought process known to the interviewer. Communicate.
- Then he asked if I had questions. I asked him what work he does, how interns are allotted teams, and what an intern's day looks like.

Round 3

- This guy was the boss of my second interviewer. Again, very chill. Asked me to give an intro. Told him how I started doing ML(because I like math). He then asked me why I still stuck to ML after one year. While answering this, I made sure to talk about Microsoft for brownie points :D

Microsoft

- This interview was discussion-based, with him giving me a problem to solve. I had to describe the entire pipeline, from obtaining data, to model selection and tuning, to deployment. He asked questions in between. Asked me how I'd assess my product. Had a brief discussion on that. Discussed additional features I could create. He seemed satisfied, and said we were done(25 mins into the interview). Then he asked me if I had questions for him. Asked him about the work he does, his opinion on things like AutoML, Github CoPilot etc, and finally asked for feedback. In the end, it was clear to me that I was through.

This guy was really cool as well with 17 years of experience at MS.

General Tips:

- Do not fool around if you do not know something. Tell them you do not know it, they'll help you.
- Communicate clearly. Explain all your ideas.
- Be yourself, do not act fake. The interviewers want you to do well, they are not your enemies.

Microsoft

3. Rishon D' Souza (Role: Data Scientist | LinkedIn)

INTERVIEW PROCESS

Round 1

There were 2 parts to the interview: coding and machine learning.

Coding section: Very standard parentheses matching question based on the stack DS was asked (Link). I was asked to code it.

Machine Learning section: In this section you control the interview. He asks you whether you are familiar with a topic and then questions you based on the terms you use while explaining the topic. They asked me to explain Linear & Logistic regression. Since this was a very basic question I was expected to go into the advanced mathematical ideas of it. Some topics discussed were-

- Type of problems LR solves
- Probabilistic Interpretation (Maximum Likelihood Estimate)
- Polynomial LR
- Assumptions made when using LR
- Loss function used and why
- Closed form analytic solution
- Gradient descent (all types/optimizers)
- Convex optimization problems
- Bias Variance
- Metrics used in classification

Round 2

This interview while technical didn't go into direct theory questions in ML. These were application based questions designed to test innovative thinking and good practices as a practical data scientist.

Microsoft

Question was open-ended. You were expected to ask for further clarifications. Ask as many questions as you want and modify your approach based on how the interviewer responds. Every step of the problem you solve the interviewer throws in another piece of the problem.

Question:

You are a wildlife photographer and you have captured 20 million photos with 5 million photos from each continent. A lot of the photos are random trees, grass etc. There could be multiple photos of the same species. Devise a system (taking the scale of the problem in mind) to identify exotic species (species not seen before in that continent). You also have a wildlife expert at your disposal for consultation. However, being a human he can only look at a limited number of photos and charges about 500rs per photograph given for consultation.

My approach:

This was a complicated question and required several clarifications. It has to be solved respecting the scale (compute capabilities given size of dataset) and integrity (accuracy and non-redundancy) of the problem.

I used a pre-trained CNN (removing the last few layers) to extract features from the images. Then used the unrolled feature vectors on a clustering algorithm to form object clumps of the same type. This way similar species and unwanted images(like grass) are grouped. Taking a representative from each group we can manually weed out the bad clusters. Now we can use the expert. Interviewer then added a constraint that you can only ask the expert 10 questions. Here we further reduce the clusters to only the “exotic” species of that continent by using Anomaly Detection (using Gaussian similarity).

Microsoft

Round 3

This was the final interview taken by the head of the ML division. This was a mix of an HR and Technical interview. We went deep into my projects (about 30 min). Since my projects had a visual component I offered to run them in front of him. He also asked how I would approach building a chat bot for IITG freshers. I had to mention what sort of data I would need, where I would collect the data from and what features I would add. He didn't expect theoretical knowledge of NLP, just wanted to know my thinking process and creativity. Overall this interview relied on soft-skills and thinking out loud.

General Tips:

- Don’t try to specifically prepare for a test as formats and topics asked could wildly differ. Prepare in general for ML by doing courses and projects. The skills you learn on the way (sklearn, git, dbms) will be tested.
- Don’t use terms you are not familiar with. Most interviews are cross questioning based rather than a fixed set of questions. ML is a large field; be honest if you've not worked on a specific area.
- Ask questions and think out loud.
- At the end of the interview they will ask if you have any questions for them. Ask as many questions as you can. This shows interest and sincerity.

Fractal Analytics

1. Ishant Khurana (Role: Imagineer / Trainee Data Scientist | LinkedIn)

INTERVIEW PROCESS

Round 1 (Aptitude Test and Coding Test)

In the first step of this round, we have to solve some basic Aptitude questions:-

• The test contains 70 Aptitude questions, separated by 4 sections (in 75 min):- Data Analysis, Reasoning Ability, Quantitative Ability, and Verbal Ability.

(According to me, this is the first and most important part of the interview process as you have to divide your time wisely and should invest your time in the right questions, else you can waste a lot of it.)

In the next step, we have to solve three programming questions:-

• There are 3 coding problems, level of these problems is medium level, there's one more thing that you have to solve these questions in some specific language, (you can't use C++ or C) (in my case the language we are supposed to do code was python)

(This is an elimination round, so please be sure you do all the three questions else your chances are less to get selected for the further interview process.)

Round 2 (Technical Interview)

This interview round lasted for around 30 minutes. Most of the questions were related to my projects. They are expecting that you should have detailed information about your project, from it's codebase to it's working. Then the interviewer asked me some simple

probability and stats problems, kinda medium level. (In this round they just check your confidence and your knowledge in the things you showed in your resume. Be prepared with a nice Introduction for yourself, to set a good tone for the interview at the start.)

Round 3 (HR Interview)

This is the last round of the interview progress. In this round too first there's a bit of question on my projects. After that it's all normal HR round questions. (In both rounds (2 & 3) don't try to act too smart if you are wrong somewhere just accept it and always show that you are a learner.)

Fractal Analytics

2. Aryan Meshram (Role: Imagineer / Trainee Data Scientist | LinkedIn)

INTERVIEW PROCESS

Round 1 (Aptitude Test and Coding Test)

It was an online test consisting of two rounds. First round was an aptitude test which had questions of logical reasoning, mathematical thinking, English as well as statistics.

• Second round was a coding round which had moderate to tough questions. I was able to solve 2 out of 3 questions.

• My performance in the coding round was enough to make me believe that I would not get selected (once again) for sure.

Round 2 (Technical Interview)

Technical round was the one I was most scared of. It was a 15 minute round (much to my relief that I would not be asked to code a question). I nervously joined the call at 4 pm.

The interviewer was a man who didn't turn on his video. He started off with my introduction. I introduced myself and told him the past experience I had in DL. Much to my amusement he asked me about the second project I had in my CV. I explained to him whatever I could. He seemed satisfied and based on my answers he asked me some basics of neural networks, loss function, gradient descent etc.

By the time I finished my answer, time had already ended for my interview and seeing his hurriedness, I chose not to ask him any questions at the end (which I regretted a lot at that time).

Round 3 (HR Interview)

• This was a very interesting round. The interviewer was a lady who asked me to introduce myself. After she heard my prepared answer, she started a casual conversation about me, how my friends thought of me, my hobbies, what makes me different, what expectations I had from the company, where do I see myself in the next five years etc.

• This time I made sure to ask a question or two before leaving. Well... I made a blunder ! I asked the wrong question. The interviewer was also perplexed when I asked her about the time when the company got listed on the Fortune 500 list. Actually, the company is NOT in the Fortune 500 list !!! She ended the call ending all the expectations I had before the interview

Envestnet Yodlee

Pranjal Verma (Role: Project Trainee(Data Science) | LinkedIn)

INTERVIEW PROCESS

Round 1

• Sort numbers in an array
• Select all those cells in the data frame under given conditions and change the value of only those
• Reverse the string using a one-liner
• How to deal with Missing data?
• Techniques one can use to find outliers, and how do we treat them?
• Some technical questions about seasonality, trends in a time series data, and how to find them?
• What is True positive, false negative, sensitivity, recall?
• Define one case when a recall is preferred over precision and explain regularization.
• Explain the working of Random forest
• Explain the working of linear regression and how the loss is calculated?
• Name some of the charts and explain the boxplot
• Rest all were resume-based

Consulting & Analytics Club
IIT Guwahati

caciitg.com/ga