Data Science Interview

Tips and Tricks

Flavio Di Palo,
Applied Scientist @ Amazon US

About Me

- **Applied Scientist** at **Amazon** US

- **Previously:** Computer Science Graduate @ PoliMi.
- Previously: UIC Double Degree Program.

- In my spare time: Co-Founder @ TechTalents.

Add me on LinkedIn!

Disclaimer

Every content we are going to see in this slides is **NOT** related to my interview round or my employment with Amazon.com.

The following slides present different resources/materials that I have **discovered on my own** during my preparation process.

Please take it as friend's suggestion, **it is not the Holy Bible**.

I'm not speaking on behalf of Amazon.com.

My Mission

- While in the US I discovered **that there is a process to get a great tech job.**

- Wanted to share my learnings to help **PoliMi** students shine in the **international tech scene.**

- Started **Tech Talents** in 2021: a Community to share all you need to know about getting **a great tech job**!

Part I
Interviews Tips and Tricks

Looking for a job IS A JOB

- It took 6+ month for me to get a Job at Amazon
- The worst thing you can do is to start looking for a job when you actually **need** a job

Timeline

First year of Master's

- Choose **Machine Learning** related exams

- Choose exams with **projects**

Second year of Master's

- Choose a Thesis related to **Machine Learning**

- **Conferences** are a big opportunity

Start looking for a job 8-6 month BEFORE graduation.

Graduation First?
Interview First?

internships/startup experience is a plus.

Who?

Lorenzo Norcini

Software Development Engineer **@ Amazon Web Services, Seattle**

Matteo Biasielli

Data Scientist **@ King (Activision Blizzard), Sweden**

Mirko Mantovani

Software Engineer **@ Google, Silicon Valley**

Federico Sandrelli

Software Engineer **@ Bloomberg L.P., London**

Mattia Di Fatta

Applied Scientist **@ Amazon, Scotland**

Gabriele Galfre'

Software Engineer **@ Facebook, Silicon Valley**

Looking for a job: Resume

Polish you resume:

- One Page: Example/Guide
- Google's resume Tips
- Link to **GitHub** projects

Polish your **GitHub**:

- Insert a short README **for each project** you have done
- Example

Looking for a job: LinkedIn

Polish your LinkedIn:

- Professional Photo
- Insert summary of what you are **doing** at the moment and **what you are looking for**
- Attach resume on LinkedIn
- Describe **each project** you did and insert GitHub links
- Connect with **PoliMi Alumnis** working at your dream company
- Start following recruiters/managers at your dream company, they **will** post job opportunities

Job Search

- Learn about **different jobs** at **different companies**.

  Difference between Research Scientist, Data Scientist, Machine Learning Engineer?

- Understand which **job** do you like and **set jobs alert** on LinkedIn, Glassdoor for when new position opens

- Look for open position on Apple, Amazon, Google, Facebook and smaller companies you like!

Start applying early

You have followed all the steps, now what?

You should apply as much as you can!

*I Probably applied for **200+ jobs** over 6 months*

The truth is you **are getting no response** in 95% of the cases!

Keep polishing your Resume/Linkedin and **keep applying**

Adjust your Resume

- Recruiters look at your resume for only **6 seconds,** and they want to see what they need

- Adjust resume to put more emphasis on **your experiences** that reflect the Job Description

What if I get no response?

Referrals

- The best way to get an interview at a **Big Tech** is to know someone **working there** that suggests you for an interview

- It is not “your friend”, but just someone that sees value in your **academic/professional** path and thinks you may be a **good fit** for a job at that company

With a referral you have **higher chances of** getting an interview.

Referrals

- With a **Referral** your resume is put first on the stack of ones the **Hiring** **Manager** receives

- You do not have the guarantee to be **called at the interview**, but the guarantee that **your resume will be** actually **seen**.

Do not **WASTE REFERRALS!**

Questions?

Always ASK Questions!

- You may be given a problem to solve

  The problem may have few details and you **HAVE TO** ask for follow-up questions

- At the end of the interview you will be given time to ask questions to the interviewer, **take that chance!**

Is all this stuff important?

- I started with the boring stuff on purpose.

- It is **fundamental** to acquire the right **mindset** in order to be able to succeed in getting a great job.

- All the above mentioned steps are **as important as your technical preparation!** In this game the difficult part is getting the interview.

Data Jobs

Data Jobs

Relative to my personal opinion, I’m not speaking on behalf of Amazon.com.

Data Jobs

Job Title	Business Intelligence	Data Engineer	Data Scientist	Applied Scientist (ML Engineer)	Research Scientist
Degree (Generally)	Degree in Business, CS, Data Analytics	Degree in CS, Data Analytics	MS in CS, Math, Statistics	Ph.D or MS in CS, Math, Statistics	Ph.D Required
What do you do?	Defines KPI to understand business performance. Builds Dashboards to inform business Stakeholders.	Manages Data Pipelines to ensure quality and consistency of the data Builds ETL jobs to produce data consumed by other Roles.	Uses ML/Statistics to solve a Business problem. Writes reports and docs that influence business decisions.	Use ML to solve problems in production. Code for Alexa,RecSys etc. **Writes production code for ML Models (SDE bar)**	Research, publish paper to external ML/Science Conferences. Does research to advance the SOTA not necessarily applicable to production immediately.
Tools Used	SQL, Tableau.	SQL, ETL tools.	SQL, R, Python.	Python, AWS SageMaker.	Depends on the Research.

Data Jobs

Job Title	Business Intelligence	Data Engineer	Data Scientist	Applied Scientist (ML Engineer)	Research Scientist
Degree (Generally)	Degree in Business, CS, Data Analytics	Degree in CS, Data Analytics	MS in CS, Math, Statistics	Ph.D or MS in CS, Math, Statistics	Ph.D Required
What do you do?	Defines KPI to understand business performance. Builds Dashboards to inform business Stakeholders.	Manages Data Pipelines to ensure quality and consistency of the data Builds ETL jobs to produce data consumed by other Roles.	Uses ML/Statistics to solve a Business problem. Writes reports and docs that influence business decisions.	Use ML to solve problems in production. Code for Alexa,RecSys etc. **Writes production code for ML Models (SDE bar)**	Research, publish paper to external ML/Science Conferences. Does research to advance the SOTA not necessarily applicable to production immediately.
Tools Used	SQL, Tableau.	SQL, ETL tools.	SQL, R, Python.	Python, AWS SageMaker.	Depends on the Research.

Questions?

Join TechTalents Community!

Group for Tech interview preparation

LINK

Part II
Interview Questions

Interview Questions

- **Coding** Questions
- **Thesis/Project** related
- General **Machine Learning** Questions
- Data Science **Problems**

How to prepare?

Coding Questions

- Leetcode: start preparing on this as soon as you start looking for a job
- Cracking the coding interview Book (Chapters I, II and IV)

Walk me through your Thesis/Project

- Prepare SHORT AND CLEAR summary of each project on your resume

How to prepare?

Data Science Problems

- Mock interview website
- Read articles and blogpost about Machine Learning
- Stay updated on how companies are using ML
- Follow PMDS events

How to prepare?

General Machine Learning Questions

- Questions on ML interview book
- Questions on Glassdoor for your role
- High-level knowledge of anything on ML Cheatsheets
- Review your favourite ML book, course lectures, YouTube videos

Machine Learning Questions

What is the difference between Supervised Learning and Unsupervised Learning?

Supervised vs. Unsupervised

Supervised

- Output variable (y) is labeled in the training dataset
- Ex. Classification or Regression
- Algos: Decision Tree, SVM, etc.

Unsupervised

- Training dataset does not contain output variable (y)
- Based on underlying structure and distribution of the data
- Ex. Dimensionality Reduction, Clustering
- Algos: PCA, SVD, K-means, etc.

What is Stratified Sampling?
Have you ever used that?

Stratified Sampling

- **Sampling:**
  - Process of choosing a subset from a target population that will serve as its representative
  - Necessary if we cannot process the complete data in a reasonable time

- **Stratified Sampling:**
  - The entire population is divided into homogenous subgroups called strata
  - A sample is drawn from each stratum
  - Ex. Binary classification, ratio negative/positive is 9:1. Stratified sampling selects a subsample of the dataset that has the same negative/positive ratio
  - Used in? **Stratified CV**

What if your dataset has many more features (m) than samples (n)? How can it affect your model?

Curse of Dimensionality

- **Curse of Dimensionality:**
  - If I have a high number of features m and a low number of samples n, the feature space is very sparse
  - The model can easily overfit

- **How to solve?**
  - Dimensionality Reduction: PCA, SVD
  - Use L1 regularization to shrink the feature space

What is a Decision Tree?

Decision Tree

- **Decision Tree**
  - Tree-like structure to represent decisions and decision making
  - Each internal **node** is a feature
  - Each outgoing **edge** from a node represent the value that a feature can take
  - Each leaf node represent a **class label (classification)**

- **How is the tree generated?**
  - Examples in each node should have the same label, minimize the entropy in each leaf node
  - We decide on which feature to split first basing on **Information Gain** of that split

- Feature importance. Interpretable tree-like structure. **used in Ensemble**

What is an Ensemble?

Ensemble

- **Ensemble Learning**
  - Multiple individual Machine Learning models are strategically generated and combined to solve a particular task
- **Bagging**
  - Reduce variance, Ex. Random Forest
- **Boosting**
  - Reduce Bias, Ex. AdaBoost

How do you know when your model is overfitting?

Overfitting

- **Overfitting**
  - A model overfits the training data when it learns behaviour that arise from noise in the data, rather than the underlying distribution from which the data were drawn
  - Overfitting usually leads to loss of accuracy on **out-of-sample** data

- **How do you know?**
  - If the error measure chosen is low on the training dataset and high on the test and/or validation dataset
  - Monitoring training and validation loss during training you can identify when the model is starting to overfit the data

What methods can be used for Neural Networks regularization?

ANN regularization

- **Early Stopping**
    - We monitor validation loss and stop the training when training and validation loss start diverging too much

- **Dropout**
    - Models with an high number of parameters tend to overfit more
    - Dropout randomly “shuts down” a portion of the units in a layer during training

- **L1 or L2 penalty terms**
    - L1 and L2 regularization add a parameter λ in the loss function to penalize bigger weights

What is the difference between L1 and L2 regularization?

When Accuracy could not be a good performance measure?

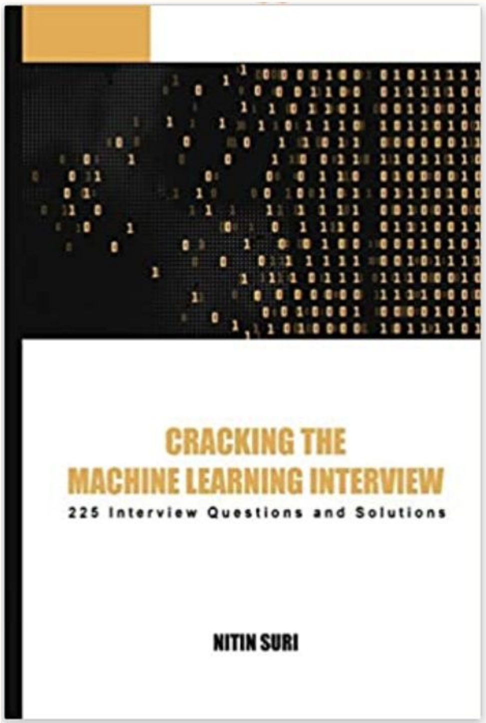
What is Area Under ROC Curve (AUC)?

Book

All the previous questions have been taken from

"Cracking the Machine Learning Interview"

Amazon Link



Data Science Problems

Musify

The notorious music streaming company “Musify.com” discovered that many of its users are **sharing premium accounts** between friends and family members

Musify is available for iOS, Android, Mac Os, Windows, Web App

Musify

- The company wants to implement some measures about it but before taking any action they want to understand **how common** this behaviour is

- Musify science team wants to design a Machine Learning system to spot:
    - The number of different users using the same account

Questions!

What can a user do on Musify?

- Musify is very similar to Spotify
- A user can listen to songs, artists, playlists
- A user can download specific songs, artists, playlists on specific device but **has to re-download** them on different devices
- **Two different users cannot use the same account in the same moment**

What data can we get?

Question: Do we have any legal/privacy restriction?

For this experiment we can collect:

- *Any possible data*
- *From any possible device or OS*

What data do you want?

Think about the problem:

- The number of different users using the same account

Granularity?

- For each song play?
- For each access to the app?

Data

For each session:

Data

For each session:

- User login ID
- Session Token ID
- Device Unique ID
- GPS Location
- Timestamp

Data

For each session:

- User login ID
- Session Token ID
- Device Unique ID
- GPS Location
- Timestamp

For each action:

Data

For each session:

- User login ID
- Session Token ID
- Device Unique ID
- GPS Location
- Timestamp

For each action:

- Song ID, Artist ID, Playlist ID for each play
- Song ID, Artist ID, Playlist ID for download

Data pre-processing?

Data pre-processing?

- **GPS Location.** Transform into:
  - State
  - City
  - Zip Code

Data pre-processing?

- **GPS Location.** Transform into:
  - State
  - City
  - Zip Code

- **Timestamp.** Transform into:
  - Day of the Week
  - Time of the day
    - Do we want to aggregate in Morning, afternoon, evening? Do we want specific time?

ACCESS Table.

ACCESS_TOKEN, HEX String.

DEVICE_ID, HEX String

GPS_CITY, ex. Milan

GPS_STATE, ex. Italy

GPS_ZIP_CODE, ex. 20142

ACCESS_DAY, ex. Monday

ACCESS_PART_DAY, ex. Morning

ACTION Table.

ACCESS_TOKEN, HEX String (to be joined with ACCESS table)

ACTION, ex. ‘Play’, ‘Download’

SONG_ID, HEX String

ARTIST_ID, HEX String

PLAYLIST_ID, HEX String

What Technique?

Problem:

Identify the number of different users using the same account

What technique should we use?

Supervised Technique? Unsupervised Technique?

What Technique?

Problem:

Identify the number of different users using the same account

What technique should we use?

Unsupervised Technique

What Technique?

Problem:

Identify the number of different users using the same account

What technique should we use?

Unsupervised Technique

Clustering!!

What Algorithm?

Always start simple!

Provide an answer that solves the problem, you can discuss about a more sophisticated solution later, **if you got time**

- K-means
- Hierarchical Clustering

What Algorithm?

It also depends on the features

We mainly have **categorical features**

What Algorithm?

It also depends on the features

We mainly have **categorical features**

- Do one-hot encoding, **very sparse space**

What Algorithm?

It also depends on the features

We mainly have **categorical features**

- Do one-hot encoding, **very sparse space**
- Use **Hierarchical Clustering** with Hamming distance

Improvements?

Homework

How can we understand who is using Musify at inference time?

We can compute the closest cluster to the new datapoint with Hamming Distance.

What if the behaviour changes? When should we re-train our algorithm?

Probably we can notice changes in a bi-weekly or monthly basis. Re-training involves cost for the company, we should asses how ***critical*** *having updated prediction is for business side.*

We can decide to train each 1, 3, 6 or 12 months depending on ***cost*** *and* ***business*** *priorities.*

Final Remarks

Knowledge is Power

Learning **what** to study for the interview is **more difficult** than the study itself

Interview process is always changing so **take time** to search the web for interview experiences and questions

Mindset

- You are **not a student** anymore! Think as a professional!

- Luck is always a variable, do not be discouraged if you are not able to **get** or **pass** an interview at the first shot, **just keep trying**

If you are able to graduate at **PoliMi** you have the ability to get your **dream job**!

Join TechTalents Community!

Group for Tech interview preparation

LINK

Thanks!

Questions?