Interview 30 - Ilan

Interviewer 1

We forgot to present ourself by the, by the way, my name is Pi. I'm a master's student at Polytechnic Moya. Ed, if you want, you can present yourself. Oh, hello. I'm, I'm

Interviewee

rash and I'm also doing my master at Polytechnic. Thank you for, for to Thank you.

Interviewer 1

Thanks. Yeah, sure. And to start off, maybe you, you can present yourself and tell how many years of experience you have machine learning.

Interviewer 1

and or anything

Interviewee

else? Uh, yeah. Yeah. Um, I have five years of experience with machine learning, uh, and um, I've worked on different projects on scoring system, on computer vision systems or NLP systems of classification and generation on some computer vision. Uh, so, um, cover a lot of, uh, fields of machine learning.

Interviewee

Um, uh, have experience with, uh, classical machine learning like, uh, logistical regression, linear regression, of course. Uh, so that's, there is a wide experience here. From my side.

Interviewer 1

Great. Super. Uh, so I'll start off with a general question. Really open-ended. Mm-hmm. , what are the main quality issues you have experienced regarding your data model or system

Interviewee

so far?

Interviewee

Mm, of course, uh, the main issue is the data, uh, because. Uh, I should understand what data should they use to get some results and, uh, I should understand how I can, uh, represent the data, uh, because, uh, their results at all depends on the data, uh, more than on some machine learning models. So for me, the data is the most important.

Interviewer 1

Yeah. Do, can you give us some specific example of issues you have encountered with your data?

Data-collection data-integration-difficult

Interviewee

Uh, the main issue is to find, uh, is to find relevant data. Uh, find and parse it. So, um, In the most cases, I should, uh, parse some data on my own, create some parcels. Uh, that is the first thing. And the second was thing when I found some data set, uh, I faced, uh, with some regulation issues means, uh, I can't just, uh, take it, take the data I should.

Data-collection data-integration-difficult

Interviewee

Connect with, um, with the peoples who made it and ask them. So, you know, the process. So that, that is the main issues

Data-collection data-integration-difficult

Interviewer 1

here. Yeah. Great. Thank you. Uh, so you talk about parsing data. What, what do you mean? You mean. having an un unstructured data and transforming it to a structured data.

Interviewee

Yeah. Yeah. I mean, um, the issue is to create parcels that, that we are gonna parse this data from database or maybe from the web.

Interviewee

Uh, so the issue, the main issue here is to create this sparer and, uh, make this sparer.

Interviewer 1

Okay, so if I under, if I understand correctly, it's to collect data from data source that is challenging. Yeah,

Interviewee

yeah, yeah,

Interviewer 1

yeah. Yeah. Okay. And can you give me a specific example of times you have encountered this issue?

Interviewee

Uh, times some, you mean, uh, how much time I, I spend it on On it? No,

Interviewer 1

just one specific, uh, time It happened to you, like

Interviewee

a Ah, yeah, yeah. Experience. Uh, so the most, um, difficult, uh, one was when I created market making bot. So I, I, I was faced with, uh, the goal to collect, um, different data from different crypto exchanges.

Data-metadata-issue

Interviewee

And, uh, there was a tough because, uh, the documentation of, uh, some exchanges. Does not really work. And um, the issue was to, to go on some resource to and find decisions, uh, how to solve this issue and how to collect the data from different exchange because, uh, because the guidance, uh, on the exchanges.

Data-metadata-issue

Interviewee

Does not work properly.

Data-metadata-issue

Interviewer 1

Sorry. What is an exchange? Is it, is it an exchange? Uh,

Interviewee

an exchange of, uh, I mean crypto exchange. The place where you can buy some cryptocurrency or change some cryptocurrency to another cryptocurrency.

Interviewer 1

Okay. See, and would you mind explaining what was the project you were working on work, working on at the.

Interviewee

Uh, market making. What? Uh, so the goal here was, um, to reduce, uh, working deposit, uh, I mean when market maker works. He spent his deposit, uh, to provide some liquidity on this exchange. Liquidity is, uh, for exchange, uh, is the thing. Uh, also as market maker, uh, also as market making, uh, when, uh, the people can came to exchange and buy or sell something with, uh, good prices and, um, If, if the exchange is bad, you will see big spread.

Interviewee

You mean you will buy, uh, on the very high and, and after it, uh, you'll try to sell it on very low prices. With the good exchange, with good liquidity, you will, uh, buy and sell, uh, at almost the same.

Interviewer 1

Okay. And when you were trying to get the exchange rate, were you, uh, I think you mentioned it before, but ju just to be sure, were you fetching the data from an API or were you scraping, uh, webpages?

Data-collection-API

Interviewee

Yeah, yeah. Uh, I'm here. I'm used, uh, the API to scrape the data and also here I'm used web sockets, uh, to get, uh, the data in real.

Data-collection-API

Interviewer 1

Okay, and the, and can you remind me what was the issue with the API and the website?

Data-metadata-issue

Interviewee

Yeah. Uh, you know, the data from web sockets, uh, is providing in some, uh, coded, uh, in some coded stream.

Data-metadata-issue

Interviewee

So I should use, uh, different coders for that. And, uh, If I see documentation, how to decode this information and, uh, how to get this information, uh, I'm not always, uh, see real, um, uh, real information. I mean, uh, documentation sometimes, uh, not. Maybe properly. Right? Or maybe something changed already. Uh, so, um, should go to different forums or, um, stack workshop or, um, other resources and trying to search what's going on here and why this documentation not working here.

Data-metadata-issue

Interviewee

So this is the issue.

Data-metadata-issue

Interviewer 1

Okay. I see. Thank you. And you also mentioned, uh, Some issue regarding how you use the data. So would you mind, uh, expanding in a bit more detail, what were the issues,

Interviewee

uh, about the using? Right.

Interviewer 1

Yeah,

Interviewee

yeah. Uh, you know, when I'm grab, uh, some data and, uh, I'm on, uh, some research stage. Uh, I need to understand how I can put the data, uh, into model in the proper way.

Interviewee

Uh, I mean, I can, uh, I, I can, uh, put the raw data into the model, uh, with the high, um, with, with, uh, a lot of dimensions, uh, just like, uh, um, Multi-dimensional, um, spots or samples, or I can, uh, prepare it and, uh, make it much more or less and, uh, put it into more smaller models. Uh, so on this stage, I should understand.

Interviewee

Um, uh, the, the main issue is here is to understand how I can prepare data, um, and, uh, put it into bottle and get a better result. Uh, Because sometimes raw data works, uh, better than some pre data. It's all depends. And, um, here, um, should use, uh, some explanation, uh, exploration analysis and some tests, uh, to understand what, how I can, um, make better.

Interviewee

That's okay. I sees how it. .

Interviewer 1

Mm-hmm. . Thank you. Uh, what was I wrong, or you, you mentioned some privacy issue with the data sets, like you, you can train on them. Maybe I misunderstood you earlier

Interviewee

on. Um, yeah. Uh, I mean, in the most cases, uh, I have a raw data. Raw data looks like, uh, let me check the board. Uh, some cast or,

Interviewee

or some snapshot of the event. And this event looks different, uh, for each sample. Uh, I mean, um, In this millisecond, I get, uh, the snapshot, it looks like three-dimensional. In the next millisecond, I get the next snapshot, and it's, and it's looks like four dimensionals. And, uh, the land of the first snapshot is, uh, for example, uh, 100.

Interviewee

And the lens of, uh, that second snapshot is, for example, 80. And, uh, here I should understand how I can, uh, uh, how can pre process, uh, all these different, uh, snapshots into one major view, into one major case. And then I should understand, uh, Is, it will be correct here to put into model or, um, maybe I should make

some preces here more.

Interviewee

Some pre, okay.

Interviewer 1

Perfect. Thank you. Mm-hmm. . Um, so we are moving on to data collection. Have you ever used data that was manually collected? So either someone who on the table Yeah. Go.

Data-collection-manual Data-low-quality Data-noisy-labels

Interviewee

Yeah, yeah. Manual Limitated, uh, you know that some services like, uh, Yandex, uh, and uh, amazing also has same service. I don't remember how it's, uh, how names, uh, so the, the people are sitting and, uh, label some data.

Data-collection-manual Data-low-quality Data-noisy-labels

Interviewee

Okay,

Data-collection-manual Data-low-quality Data-noisy-labels

Interviewer 1

perfect. And if she used that data, did you ever encounter quality issues with the.

Data-collection-manual Data-low-quality Data-noisy-labels

Interviewee

Uh, yeah. Yeah. Um, you know, in the most cases, um, we can, uh, the. The peoples who, who made this classification provide some metrics. Uh, I mean, uh, it's about 90% accuracy for labeling this data, or maybe 95 or maybe less, or maybe more.

Data-collection-manual Data-low-quality Data-noisy-labels

Interviewee

Uh, but not always. And uh, sometimes we. Uh, get, uh, we, we should make some our expertise by, uh, by, by my own. I mean, uh, we have, for example, 1000 samples, so let's get 50 samples here and, uh, trying to understand, uh, I, is it proper, uh, labeled here or maybe we get. Uh, bad, uh, bad labeled and, uh, yeah. This is the issue here, of course.

Data-collection-manual Data-low-quality Data-noisy-labels

Interviewee

Okay. Sometimes I can, I can use, uh, different, uh, models, uh, that's already portrayed and understand the quality of this data. All depends.

Data-collection-manual Data-low-quality Data-noisy-labels

Interviewer 1

Uh, use some models that are trained. What is a to,

Interviewee

to classify? Uh, I mean, um, if I have some data set, uh, for something, uh, and I want to learn another model on this data set, but, uh, I have, uh, another model that uh, already trained on this type of data and I can use this model to.

Interviewee

To check, uh, how well, uh, how well its dataset was labeled with. Okay, I see. So the machine learning model? Yeah. Okay.

Interviewer 1

So you use the machine learning model to make prediction, and if the prediction does not match the labels, I Okay. . Okay. I see it. I understand. Thank you. Um, have you ever used external data, uh, you already mentioned third party api, so scripting for the exchange rate, but have you ever use, uh, public data sets, for example?

Interviewee

Uh, public dataset? Yeah, of course. Something from cargo in the most cases for, for my, um, for my. In the most cases it's , uh, also open R here.

Interviewer 1

Yes. Yes. So I understand you are a PhD master of PhD student, something like that. Yeah. Yeah. Okay. I see. Okay. Thank you. Um, all right. And did you have any issue with the public data assets?

Interviewee

Um, no, um, use it only for some research. So there is not some issues here. Uh, just, just like, uh, every data science issues, like, uh, how to prepare data, how to feel nonverbal, non values and so on. Not nothing special. . Mm-hmm. .

Interviewer 1

Okay. Perfect. Thanks. Um, have you ever tried to measure the quality of your dataset and or tried to improve it?

Data-missing-values

Interviewer 1

The quality of the dataset?

Data-missing-values

Interviewee

Hmm. Yeah. You know, uh, when I'm making some service, uh, that, um, the pipeline that, uh, on the first stage par data on the second stage, uh, convert data to features, uh, yeah, I, I should, uh, make some metrics that, uh, will, will validate that data, uh, par that Parsi data is correct. Uh, the features that I prefers from, uh, raw data is correct.

Data-missing-values

Interviewee

Uh, so yeah. Um, I'm making some validation, um, processes here. Uh, and uh, sometimes, uh, I have, uh, data driven, uh, that I can see before I train the model. Uh, because, uh, I can see. New, new, uh, new type of categories here I can see, um, the different, uh, different, uh, data that I have never seen before, um, in numerical data.

Data-missing-values

Interviewee

So, yeah. Um, and, uh, I, I can see a lot of, uh, diff uh, different. And, and values maybe some empty strings or rows, some empty spaces, uh, on the data set. So, uh, of course when I'm making pipeline, I should control all these things. And, uh, yeah, this is the issue here when I'm creating the.

Data-missing-values

Interviewer 1

I see. Thank you. And which tool do you use to help you?

Interviewer 1

To help you? Yeah. Which tool do you use?

Data-missing-value-fix Data-missing-values MM-drift MM-drift-fix

Interviewee

Uh, we Which tools?

Data-missing-value-fix Data-missing-values MM-drift MM-drift-fix

Interviewer 1

Yes, to, um, Evaluate and improve the quality of a data set.

Data-missing-value-fix Data-missing-values MM-drift MM-drift-fix

Interviewee

Uh, you know, in the most cases, just some statistical, statistical methods, uh, that, uh, provide me information about the data drift. Uh, about the, uh, quantity of, uh, empty spaces and so on. So nothing special.

Data-missing-value-fix Data-missing-values MM-drift MM-drift-fix

Interviewee

Just, uh, something that they can, uh, try to use to find en.

Data-missing-value-fix Data-missing-values MM-drift MM-drift-fix

Interviewer 1

I see. Thank you. Uh, is there any other data quality issue we missed that you consider relevant?

Interviewee

Uh, you know, I can't remember just right now. So, uh, for now, so

Interviewer 1

perfect. Thanks. Um, how do you evaluate the quality of models? And as a reminder, quality is not only defined by the performance, ML performance, but you can also consider scalability, uh, robustness, explainability, and so many other aspects.

Interviewee

Uh, yeah. Yeah. Uh, so, you know, the main metrics is the business metrics, uh, when we create some models, uh, and, uh, that models, uh, are used in some services. And, uh, after the work of the service, we will collect some data how it, uh, service. Well or poor. So the main things here is the business metrics, uh, for market making, both, for example, it, uh, how much money was spent on, uh, on liquidity providing, uh, for, um, If, uh, we are trying to find the best, uh, uh, the best placement on some marketplace and, uh, use this information, uh, the business metric here is, uh, uh, how well we promote this, uh, posts and, uh, how much revenue we will get from.

Interviewee

So this is the main metrics here. And, uh, of course this metrics depends on, uh, the, oh, uh, these metrics depends of performance, of model, of speed, of performance, uh, of, uh, robustness, uh, of, uh, evaluation metrics like. Uh, a c for example, or some prior score if we will, uh, provide probabilities or some accuracy or, uh, g score if we'll talking about classification.

Interviewee

Um, so, uh, it, it's all just, um, Just the parts of, uh, mechanisms. So the main metrics, uh, is business metrics and for each tasks, uh, it's different. I see.

Interviewer 1

Thanks. Um, so you talk about user user metrics. I, I suppose you already test your machine learning software system with users of the system.

Interviewee

Uh, gimme, please tell me more.

Interviewer 1

Yeah, so since you mentioned user metrics, right? Mm-hmm. , I suppose you already tested your ML system with the user of the system, so mm-hmm. , right? And so my, my question was, um, do you have ever encountered quality issue? By testing your application with the users?

Interviewee

Um, yeah. Yeah. It's, uh, about the user experience, right?

Interviewer 1

Uh, yes. Or basically it's any quality issues you have encountered when you presented your, presented your ML application, ML software system to the

Interviewee

users? Mm-hmm. . Mm-hmm. . Yeah. Uh, Of course in the most cases is the speed of reaction, uh, and uh, the information of, uh, provided from a mail system for user. Uh, so how it's relevant.

Interviewee

Uh, and uh, here when we are talking about the users, uh, we. We are watching on some metrics, uh, how, how long time, uh, user was on this page and how, how much time he played with this feature provided by this ml. Think, uh, and, uh, of, of. The main thing is, uh, to make some action that will convert into money.

Interviewee

And, uh, we have some, uh, actions, uh, that move user to make, uh, the actions that will make money for service. And, uh, here, uh, we have. Some metrics that, um, uh, that is the sum of another metrics, you know, uh, the, it's, it's like staking. Uh, here is the first metrics. Here is the second here is assert, and here we have the final.

Interviewee

And, uh, in the most cases, that's how it works.

Interviewer 1

Okay, perfect. Thank you. Mm-hmm. , um, have you ever encountered issue during the deployment or the maintenance of a machine learning software system? Uh,

Interviewee

for me, no, because in the most cases, uh, it was on the DevOps or on envelopes. Uh, so I don't have here any big issues.

Interviewee

Um, no. It's, it was, uh, fine and fine and, uh, and, uh, When there was some services, uh, architecture, it was pretty good, like amazing or Azure and so on.

Interviewer 1

Okay, great. Perfect. Um, so you mentioned you had, you worked as, you had an experience, um, as an ops engineer.

Interviewee

Uh, not at all. Just, uh, for deployment on, uh, Uh, on some service list service like, uh, a Amazon aws, okay?

Interviewee

Models here, here is not, uh, here was not big service. Uh, for big, uh, account of users. Only for some small teams that can use this, uh, this tool for their own.

Interviewer 1

Okay. Uh, but was it a machine learning software system that you were deploying? So was there a machine learning component or not at all? Yeah,

Interviewee

yeah, yeah.

Interviewee

Uh, in the most cases, uh, all services that I developed was machine learning services based on different models and bicycle or, and other one.

Interviewer 1
I see. And would you like to talk a bit about them? What, what you deploy?

Interviewee
Um, yeah, of course. Some, uh, classification services, uh, means, uh, the user provide, uh, provide some table data and, uh, get the predictions on this data, some regression services the user provide.

Interviewee
Um, There's, uh, some documents and, uh, computer vision models read these documents and, uh, um, and, and eg it's, uh, and eeg, uh, Take, uh, some, uh, data and, uh, then count, uh, this and, uh, then use this data to make some progression predictions. So different ones.

Interviewer 1
Okay. Uh, would you mind giving me, um, more detail in one that you, you think, like what were some of the issues, uh, the quality issues with one of your product?

Interviewee
Uh, with deploying?

Interviewer 1
It can be anything, uh, deployment or monitoring or even before if there was some data issue or, uh, yeah. Ever.

Interviewee
Anything, uh, yep. Let's, uh, I will find something. Mm.

Interviewer 1
I mean, you can just tell me, yeah. What was your experience?

Interviewee

Yeah. Uh, you know, the biggest issue was, uh, when we are pushed, uh, some, uh, New things, uh, into repository.

Interviewee

It was, uh, Friday evening, uh, and uh, on the weekend it crashed and, um, it was terrible. So this is was the main issue? Uh, yeah.

Interviewer 1

Why, why did it crash? Was it a, um, a bug that was related to some machine learning component? unrelated.

Interviewee

Yeah. Yeah. It, it was, uh, machine learning service. Uh, it, uh, should, uh, predict some data based on the, another data.

Interviewee

And, um, there was a mistake, uh, me and my team, uh, was, um, uh, so it was, uh, Friday evening. So, uh, we are, uh, let me check the words. Uh, we are tired already and, uh, we'll lose some attention and make a tiny mistake in the, uh, future generation part. And, uh, after, after we push, commit into repository. Yeah. And uh, so we start working with a new source code.

Interviewee

Uh, it crashed on the weekend and it was very bad.

Interviewer 1

I see. And what was it? The problem, uh, with the future generation,

Interviewee

uh, just, uh, a sign, you know, we forget about the sign and. In the most cases is, uh, this does not, uh, make, uh, any big problem. But, um, it, uh, was a two person chance to crash, uh, all the process and it happens.

Interviewer 1

Okay. And why does crash, why does sign create, uh, result in crashing everything?

QM-robustness

Interviewee

Uh, because, uh, it was, uh, some equation, um, that generates some, some feature. And, uh, in the most cases, the data looks like a normal and, um, it, it was okay for, uh, plus sign. But, uh, if the data looks, uh, a bit abnormal, um, The feature change, uh, the features that, uh, are put into model change, uh, change significantly, and, uh, prediction predictions, uh, are, um, start making, uh, uh, very bad.

QM-robustness

Interviewer 1

So basically it was not robust.

QM-robustness

Interviewee

Not purpose, uh, was not robust. No, no, no. Yeah, it was just a mistake.

QM-robustness

Interviewer 1

Okay, perfect. Thank you. Um, so yeah, I will just list some quality aspect and you tell me if you ever had issue with one of them. Uh, so fairness, robustness, explainability, scalability, privacy of data, and security of your model.

Interviewee

Uh, yes. Scalability, I think is the issue, uh, because sometimes we have, uh, really big models and it's consume, uh, large of resources and we have, uh, Uh, budget, uh, fix it, budget for, um, our goals. And, uh, that is the real issue. And, uh, every time I should find some pass, how to make, uh, models smaller and, uh, protect the accuracy.

MLP-costly-platforms QM-scalability-issues QM-scalability-issues-fix

Interviewer 1

Okay. I see. Uh, so basically it, it costs a lot and you were trying to simplify your model to address that issue.

MLP-costly-platforms QM-scalability-issues QM-scalability-issues-fix

Interviewee

Yeah. Yeah. To, uh, to, to make, uh, the service, uh, bigger. I mean.

MLP-costly-platforms QM-scalability-issues QM-scalability-issues-fix

Interviewer 1

Okay. Oh, mm-hmm. , uh, bigger. Okay. I see. Mm-hmm. . Okay, perfect. Thank you. Uh, yeah. So I have two last questions for you, um, in your opinion.

Interviewer 1

Okay. What are the most pressing quality issues, uh, researchers should try to solve?

Data-dataset-evaluability

Interviewee

Uh, quality issue. You know, maybe, maybe we should, uh, find some method of, uh, how we can, uh, understand faster the quality of the data. Some, uh, tools that can, uh, that we can use to understand the quality of the data. I mean, just, uh, push one button, understand, uh, the. Uh, it so sounds like, uh, a miracle, but, uh, I think, uh, uh, kin can do something like this.

Data-dataset-evaluability

Interviewer 1

What do you mean understand the quality of your, of your data? Is it to grade the quality of your data or it's to pinpoint what are the error in a data set?

Interviewee

Um, All, all of this, I mean, uh, to push some button and, uh, understand your data, get the information of this data. Maybe we should, uh, create some, uh, machine learning models that will explain this data, uh, e explain what's going on in this data and what's the problems here with this data and so on.

Interviewer 1

Okay. I see. I underst. Thank you. Um, sure. And do you have any other comments about the quality of machine learning software system?

Interviewee

Um, no, I think no.

Interviewer 1

Perfect. Uh, so that's all for us. I think you provide some great insight. Uh, thank you. I'm sure it'll be

Interviewee

useful. Okay, Cole, and, uh, how much peoples already. Uh, talked with you.

Interviewee

We are at 30. Mm-hmm. . Okay. Nice. We need few more. Yeah. Uh, alright. Uh, will you send me information when you will, uh, start some movement with this? If, if we start what I, I mean, I, I mean the next stage. Uh, I, I, I, I mean, what will you, uh, announce, uh, some, uh, new stages and how, how deals go going?

Interviewer 1

Uh, do you mean if, if we are going to send you back the paper, the final

Interviewee

result?

Interviewee

Yeah. Final results or maybe some, um, middle stages, how it's going. Just curious.

Interviewer 1

Yeah, sure. I can. I can send it to you. Yeah, no problem. Okay, cool. So, all right, so we're glad, uh, thanks again to being, for being here. Thank you. And, uh, Okay, thanks. Have a good night. Oh, by the way, just before, uh, can you tell me in which region this is for demographic purpose, can

Interviewee

you tell me which, uh,

Interviewer 1

Okay, great. And, uh, where you worked, was it a startup or a large company or medium company? Uh, medium. Medium. Okay. Perfect. All right. Thanks. Have a good day. Have a good night. Okay,

Interviewee

you too. Bye-bye. Bye.
