Interview 27 - PO

Interviewer 1

Okay. So, um, can you give us a bit of information about yourself, such as how much experience you have in general and specific to machine learning?

Interviewee

Yeah, so I, I started PhD in like advanced statistics related work. Um, so from where I kind of get the understanding of the machine learning, uh, where it's like application of the machine.

Interviewee

So after that, I work as a post doc. And also that was mostly like a statistical analysis, trend testing and stuff like that. But then I, um, came to the industry where exactly like, uh, the machine learning application bills, um, um, like, uh, the industry actually applied, uh, dna, DNA research. So I have some kind of a research background, but what we do.

Interviewee

In a research is completely different in a, uh, in a, in an industry prospect because industry always look for, um, like a, uh, like application testing. So what, like, you know, like in in, um, in a research we'll be doing train and testing where we already know that data sets as like the testing data sets, but in industry you have to put it in your production.

Interviewee

That is actually, you never saw the data is coming when. Any modeling a production. Right? So that's the, that's a difference. Is there, so I am working in industry almost three, four years by now. Before that, I work in a, um, like as a post-doctoral scientist, almost like, uh, another three, four years. And then, uh, my, uh, three years actually, yeah, three or four years actually.

Interviewee

Uh, and then, um, I have my PhD on that thing. So overall I have experience like water research. where, um, the researcher have a gap between, um, uh, when it goes to the industry, you know, and the reason behind of this conversion when you send the email, and I agree with that thing. So, um, I want to, like, normally I promote that thing too, like how, um, there could be this gap between what researcher doing in a machine learning, um, or AI prospect and what industry is doing.

Interviewee

So there's a gap. There's a, there's a good amount of gap understanding, and there's a, uh, good amount of gap is, um, like implementation, uh, from a, both, both of the side mainly, I will say, I I'll say like, uh, uh, in a research prospect because many thing we do in a research, uh, that is need to be a different level when we are implementing in there.

Interviewee

So, as I said, like the, the reason I. Agreed to these things. So I want to share something. If it is, that helps your thing and then who will be leader of your thesis or, or your research paper in future will get, um, some understanding out of that thing. And I actually promote, promote means I, I talk about this thing in many places because what I believe is data science.

Interviewee

There are very. , less amount of actual data scientists are there. Like everybody will say like, I, I am a data scientist. I'm doing that thing and that thing. But actual, when comes to the, uh, uh, comes to the work delivery, there are very few people with the knowledge is there. And as a, as a peer or as a, as a practitioner of the data science or machine learning, I feel like, uh, for my, my own duty to, uh, take these, um, these, uh, um, Uh, this proficient to any, uh, innovative level, you know, like the help and the other people in the profession.

Interviewee

So yeah, that's, uh, that's me.

Interviewer 1

Super interesting. Uh, we are happy to have you with us today.

Interviewee

Yeah.

Interviewer 1

Um, so to start off, I will ask you a really general question. What are the main quality issue you have encountered with your data model or system so far?

Interviewee

So, uh, the quality issue in says in a data model, and then the third one, what you are ask for,

Interviewer 1

uh, system. So the software system in general.

Interviewee

Uh, okay. So most of the time the quality issue data comes when the data is ingesting, okay? Uh, instead of, uh, the model part. So what does it mean? Means is. There could be like, suppose like somebody have a Google form, you know, like, uh, they're, they're taking, uh, a survey out of that form, right?

Interviewee

And that form is whatever the survey columns are there, or the row are there that is actually, um, going as a SQL server, like the SQL table over there, right? So now the thing is, if somebody in future add a new column, Or row or something, a new variable on there that is not before there. So that could be a quality issue if you build a, build a model over there.

Interviewee

So like extra, extra there. The second end thing is, uh, the NA data sets. N means if there's a no data, data is there, right? Sometime happen that, um, system is. , suppose like a iot device you are using that system is down for a couple of couple of days or something like that. So then you don't know, don't have that, the machine data basically.

Interviewee

Right? So, uh, one is that then sometimes it happen that uh uh, the anomalies, anomalies. Animal is also good in a use case. So suppose like if you're working in a financial industry, you need the animal to look into that. Like if there are like, uh, money laundering activity or illegal activities happening on that.

Interviewee
But then, uh, if you are working in a time series forecasting, then if you have anomal animal is, which is totally out of the blue things, which, which could be a problem. But when you are forecasting that model could get into a different, different job, you. So that thing is there. The third thing is, um, uh, the imbalance dataset, but although there's, there's a even, uh, there's a imbalance.

Interviewee
If there's imbalance data sets, there are processes out there that you can use it. So imbalance data sets, like I, I had to use cases for like, most of the people who use that, use cases when they have a marketing, uh, marketing use cases out there. So where they have like, try to predict like churn model.

Interviewee
Churn model means like how many, uh, customer would like to churn in next month or next two months or something like that. So in that cases, although it's email data, but they need that, I balance data because you think about that, the pupil chart. So maybe like if you have a 2 million customer base and next month probably 20 people or 30 people can charge, or maybe 200 not both.

Interviewee
So that's the case is imbalance. Like you have a total customer is of 2 million, but 200 people are churn, right? So how you predict that 200, right? Because next one, 200, next 1, 200, 200. If you do this, this one. So it'll be end of the year, it's 200 people are churn, right? So see there are like difficulties in data.

Interviewee
There are animals in the data there. Um, uh, like now in the data, but there are processes out there that, how we can that one. The second thing is in the model, so I, I kind of little bit go overhead. So what, what was the question? Like, uh, what is the data in the model or, I'm sorry, what is the question? No, no, I'm saying like, what was your actual question?

Interviewee

Was that, what is the. , uh, what are

Interviewer 1

the quality issues with the data? Yeah.

Interviewee

Okay. Okay. So, so the data quality and the model and the system. Yeah. So the next, uh, the, the, the question is, uh, the, the things comes is, uh, the quality issue in the model? Yeah. There are different types of models out there right now, like from a simple cation model to the, uh, the three waste model to the mural network model.

Interviewee

It's also depends upon like which type of the works are. So quality issue is again, uh, in a model prospect, I will say is like choosing, like the, selecting the model is the, the, the another issue can be created, right? So in that prospect, if you can select a very complicated model, right? And if you, you can select a very simple.

Interviewee

uh, auto integration or AMA model very, very simply like moving, moving average or linear integration model out of that thing. Right. So exactly what, like, I, I would like to be interactive. So exactly what you are looking for that prospect. Like what, what exactly you want the comment on that thing? What I am looking

Interviewer 1

for, uh, quality issues in the model?

Interviewer 1

Yeah. Yeah. Well, it really depends on how you define quality. And, um, for example, if explainability is something important to you, maybe a quality issue with the model will be that, uh, the model are not interpretable. Right. And you must use explainability, uh, techniques to, uh, right. Or if you want to scale, maybe your problem is that your models are too large to scale or, you know.

Interviewer 1
Yeah.

Interviewee
So in, in this 2.1 is explain. . So sometimes everybody thinks about like a regression model, right? Regression model, easy to explainable because it's a coefficient, it's a, um, like a slow, it's a correlation. It's easy, but when the three waste comes, it's very difficult to, everybody thinks about, it's a black box, but it's not a black box.

Interviewee
Right? Even like neural network, like, or propagation, backward propagation and, uh, the weight changing and all the stuff. So it's very hard to. Hard to describe to the business. Like when we're talking about the, the, the how, um, how, how we can explain to them. Uh, they always turns to want to, to become like very, uh, very simplistic, uh, in a manner like they wants to know.

Interviewee
So something like, right now I can give you one example, right? Right now I'm doing a exhibition where I'm predicting. Uh, and also I'm looking for the feature imports. Every time I say to them, there is a feature, feature, imports, uh, and then the next time they will come back, there's a correlation. They are thinking, what is a correlation?

Interviewee
But I say like, okay, it's a multi problem, but when you are thinking about the correlation, what I'm giving you is not a correlation. It's a feature importance. So the future importance is basically which feature have the more importance in that, that. then, uh, the less importance, uh, features are there, right?

Interviewee
And they always think about the coalition. So that's also the understanding. People have to be, uh, uh, people have to be understand, uh, in that prospect that that what is a use case. And also those who are like data scientists working as like me, have to explain to them every time that where things could be, um, different than other model.

Interviewee

This is not just a simple linear integrion model and, uh, you work on that thing, then the system prospect. If you think about the whole system, when you design the whole system, then uh, it's come. Uh, so are, are, are you familiar with the whole like, uh, uh, data science, like pipeline, like end-to-end? ?

Interviewer 1

I think so.

Interviewer 1

But if I, if I, there's something I don't understand, I will interrupt you and, and tell you. Okay,

Interviewee

no problem. So what happens is, first you have to understand, uh, normally a data scientist work or like when they call it a data science, is basically you get that data from iot device are anywhere like the data generator.

Interviewee

You have to build something out of that. The software, like if it is. Like, uh, uh, startup, they'll be building a software out of that. If it is a big company, they'll be building internal report out of that thing, right? So how to do this thing, the whole thing and make it, its automation, not, not like a manual intervention is the call is a data science pipeline end-to-end pipeline, which is the patient learning board.

Interviewee

So the data comes. Store in somewhere, which is kind of a database. It could be one frame and as cloud, whatever, the different cloud there. From there, the data engineer takes all the data and, uh, match multiple cable, put it, uh, put it another like similar type of a database or data database where the big data can be accessible.

Interviewee

The PI spot concept, how do concept. . The next is the data scientists. Take that data, look into the data, understand them. Uh, what is, what is understand, understandable meaning is coming out of that thing. Right? And the final parties, uh, no, there's the last two parties. One is DevOps. So DevOps is basically the automation part, which is run the whole, uh, hold the pipeline that this thing designed.

Interviewee

And the final part is, uh, the data analyst who will be building this dashboard right out of that thing. So this is the whole system, like the, the whole system. Now complexity goes in a whole system as like if you introduce multiple component on that, right? Like if you have a, uh, on-prem versus, uh, like a cloud practices.

Interviewee

Some people will be doing the whole thing in a one single type of cloud. Like I suppose everything. Either Databricks or Azure. Azure or AWS or gcp. But some people also, because they have, they're already in, build something in their, um, um, they are, how I'll say, uh, in system. So they want to combine that.

Interviewee

So in this process, this, when you are combining multiple things, there are uh, uh, there are complexity growths. Like if you just. keep a one single platform less complex than if you are b building in a multiple complex platform, a multiple, uh, different platform. It picks, uh, make it, make it more, more complex on that thing, right?

Interviewee

So in that prospect actually, uh, so the complexity will be if another, another complexity will be in the system, the whole, those who are the stakeholder, like those who are working on that thing. . So if, uh, they are in the same team, right. So they'll be pretty much knows what they're doing. So I'll, I'll, I'll, I'll put it as a resource problem too.

Interviewee

So I'll keep it up to here. And then if you have any questions and particularly then we can go in in that direction. . Yeah. Thank you.

Interviewer 1

It was really interesting. And, uh, if you, if you would like, I would like you to go more in detail with the system part. I think you mentioned a few interesting things, and, uh, also I have a question regarding the system part.

Interviewer 1

Um, so, so you said one, one problem is people are on different platforms, if I'm correct, and it makes it complicated. Okay. So is it, is it like people are, are on aws, Google Cloud, and a and Asia is. That's a problem.

Interviewee

Uh, exactly. Not like that. Like people don't use like the cloud together, but there are multiple service provider out there.

Interviewee

The service provider means what? I'm saying somebody's building in Azure. Right. Which is fine, but there are, um, like, uh, uh, like when you are scheduling, scheduling a pipeline, so there's a quality scheduling a pipeline, what does it mean is basically whatever you machine learning product you. It'll be automatically running in a particularly time of the period of the day.

Interviewee

It could be weekly, monthly, biweekly, any, any time of the day. So when you are scheduling as Azure, DevOps can do the scheduling. AWS can do the scheduling on the, some people like third party software. I dunno why. But it's also depends upon, uh, those are architecture out there. And also most of the time I feel like it's the old team, those who are working there because they are used to it.

Interviewee

Some other soft. So suppose like, uh, there's a software called Control name. So I, I, I know some company will be just putting a control name software, uh, to schedule that jobs, but where Azure, Azure DevOps can, sorry. Uh, yeah. Azure. Azure DevOps can do this. It's by itself, even like GCP can run, uh, run its scheduling by itself or data can do this by itself.

Interviewee

So these are the, these are the challenges out there. So if you ask from a research prospect, why. There are no single questions at there. They will say like, oh no, this thing, that thing happens. But frankly speaking, uh, it's if it is, if you ask me that, if it is not doable in one platform, yes, it's doable in one platform, how the companies are coming.

Interviewee

But it's also depends upon that. I already told you those who are using those who are managing things and those were like taking in decision. So I know companies one day they are in aws, uh, director. , uh, they will be moving towards Azure. Azure, like three months. They were, everything I was doing in AWS director change.

Interviewee

Some reason the director come in, they'll be moving towards Azure. I have no idea why that things happens, but things, things, things moves like that. So whatever you build in that part of the things or whatever you building, you have to drop it. You have to go to the and you have to feel like whatever in build in.

Interviewee

um, some, some software they already have or some system they already have in that company. Uh, they will take that one some part, but the some part will be coming from that. And that makes more complex in, uh, in, in general.

Interviewer 1

Okay. See, so basically technical depth is building because, um, people do not want to summarize.

Interviewer 1

Actually, I will go even simpler. . Sorry, I, I forgot what I was going to ask you

Interviewee

anyway. Yeah, of course. Yeah, yeah. , like, I got a brain

Interviewer 1

freeze. Yeah, yeah. Anyway, uh, but, uh, maybe this question will seem obvious to you, but why is it, why do you think it's complicated? Or why is it complex to have, uh, to use different technology on the same cloud?

Interviewer 1
Why is it better to be conclusive and only use, for example? Aw.

Interviewee
See why it's only aws. It's not aws. I'm just saying like either cloud or one, one sort of a technology. The reason is if, uh, the compatibility and uh, um, the data moving part. So suppose like the, the data moving part means I'm saying like, suppose you have storage, like blob storage or something like that, right.

Interviewee
Where the data is sitting there. Right. Or other SQL pool or, um, sql. It could be from Microsoft or somebody else. Now, if you are moving in a, uh, in a different, um, different software, then you have to either, you have to call that api and then when you are calling that api, how that is linked to this, the system, the core system.

Interviewee
So now, uh, the, um, it's called like a time lapse basically. So when you are calling, how much time it taking from these to transfer to that? . So if you think about that, um, you are, um, so suppose like it's a big airplane, uh, and instead of you're putting a, uh, uh, in middle, some, some seats, you put a bus seats inside there, right?

Interviewee
Uh, so is it, is it going to be what? Yes. You can sit and go in a bus seat, in an airplane seat, like you can replace with a bus seat and it can go, um, like military Hughes. So. In a adverse, uh, like a somewhere, um, in where it's not possible but ideal. No, not ideal. The reason is it's not much compatible when you are, uh, asking, uh, something output from there, uh, output from, uh, from, uh, from the system.

Interviewee

So now also it comes in a use case basis. So suppose like I give a, uh, one interview with the, uh, um, I, I give you one interview with. Uh, a company who gives you credit, credit score, right? Uh, so they are asking because I was coding there, and uh, they says like, what you are writing it's right. But the thing, the thing is what they want that output in a millisecond.

Interviewee

So you put it some output, like suppose your bank, you are putting some, uh, looking for like Soho, uh, grade score. So they want to keep output in a millisecond, less than a millisecond, right? Because you cannot just. in that moment. If you have a multiple different, um, cyber is going in background and if it is taking multiple time on, if there's a glitch, so then your end danger is suffering and that not going to happen somewhere.

Interviewee

But then, so that's why when they design, they design very cautiously like a financial I. Even like a biomedical institute and all this thing, when they're design, their system design, they're, they're very much, uh, very much, how I say, uh, very much cautiously. They don't go in a multiple different software where you have a multiple, like a multiple platform where you, you get a multiple, uh, like issues anytime, but suppose like a detail, uh, uh, warehouse logistic, those who are there.

Interviewee

If their dashboard is not like refreshing, uh, within order, if they have to wait for 10 seconds, it's fine for them. Nothing will happen if you refresh 10 seconds later that dashboard, right? So they actually, if you look at the industry trained, these, these guys, the retail, those are not like emergency service.

Interviewee

They actually do these lot of experiment in a different thing. Another thing is a cost. Why that is mean cost. . If you go with, suppose like aws, right? Everything you build in aws, so what AWS will say next, next day, you have to pay this much amount to use the whole thing, Mike, right? They will be, go, keep going up.

Interviewee

But if you have a multiple vendor on, on your, on your site, so you can go leverage some system over there, some system over there, and you can negotiate with the cost. And that guy knows that these guy, if I charge these one more, the account manager. , then he will be moving to some other cloud who will be giving it the whole thing.

Interviewee
So that's another thing. But again, those who have the bank financial institute, um, the fine tech, what we call it, and also like a hospitality, their neighbor can do that whatever they want, like whatever they want. They have to do it in a very smooth, efficient manner where no glitches. is, is it Answer your questions.

Interviewee
Yeah.

Interviewer 1
It, it's super, uh, comprehensive and it's perfect. Thank you. Uh, so just to make sure I understand everything or, or I, I'm saying everything, uh, so one of the issues, one, one problem you see is people problem or just something you observe, is that, uh, the, the, the, the pipelines are often deployed on different platform.

Interviewer 1
That's, that's the initial thing you mentioned. Uh, this is com. This makes it comp complicate complex to it's complex to understand everything. Uh, the reason why they don't, why it happens is because, uh, sometime people prefer using the, um, online service. They know, uh, and one advantage, one advantage of.

Interviewer 1
Having the pipeline on the cloud is that, um, you're able to negotiate with the account manager as well. So it's, it's a plus. It's a, it's a minus. So, so it's complex to have the service everywhere, but at the same time, if there's an issue, you're able to switch between cloud provider or, or, uh, more or less.

Interviewer 1
And the la another thing you mentioned is also the Latins. You said if we use a lot of services, um, generally it's gonna be less efficient than if everything is on one cloud. Am I.

Interviewee

Yes, yes. Latency will be more, um, if you are using multiple, multiple things. Because suppose like as I say, like control limb someday cannot work, right?

Interviewee

Some, some software update is happening. Control limb or Informatica where you schedule your, uh, pipeline, but your as Azure supporting fine or aw, right? So you have to just wait for that thing, right? And sometimes happen, like things doesn't work. Um, like in the morning. Um, comes in and until I left, like manually intervene and see like, okay, these production issues have production lead and add.

Interviewee

So they have to do the manual running. So normally what company do after, uh, after, uh, like everything is done, there is a, uh, there is a maintenance team. They keep it there. Let's think about as they're building a house type house, right? So when they need to build a house or like the whole architecture or end-to-end thing.

Interviewee

Ask for architecture. So same like software architecture and the actual architecture seems architecture. So they comes in, they build design and all this thing. Then these, uh, the, the handyman and all the, the, the, the general contractor comes and all us to the house. So same thing like all the engineers comes in, data scientist comes in and all this thing.

Interviewee

When everything done, everybody clean the house, nice build house. After everything running, then you need room, the cleaning. They will come every day, clean maintenance guys, they will come, the electrician, maintenance and all that stuff, like the security, same thing. Company keeps as a maintenance people with some, some training, uh, that they will be just maintaining the whole platform every day, morning or whatever.

Interviewee

They will be having a listing that, okay, I have, uh, so if it is support, like you have a hundred different machine learning models are there, and then normally they do not schedule in one day because it'll be put the pressure. Uh, pressure on, on the server. So I, they'll be putting in a, across the time period, like, uh, either morning, afternoon, evening, based on the, uh, based on the things they need from the business and across the, the week, like when it happened.

Interviewee
So the maintenance knows exactly which model runs in the what time, and then every day they have a checkbox that this thing ran or not, or something like that. There'll be one or two guys will be always. and it's runs in 24 hours, uh, in a model. Most of the time, if it is a C and big company, uh, we'll be there.

Interviewee
So I see.

Interviewer 1
Interesting. Thank you. Um, I have a lot of question and I will skip. I see you have a lot of experience and the, like, the, the system and I really enjoy that. So if you don't mind, I will skip all the data questions, questions and I will ask you. Uh, good. Perfect. Uh, so now I'm focus specifically on.

Interviewer 1
So, um, mm-hmm. , what are the challenges you have encountered during the deployment of a machine learning software system

Interviewee
Deployment Challenges, uh, one is qa. So what is a quality assessment? So what is done means normally when anything builds, there are two types of data. They keep it, one is deep, so development data and one is broad, so like production.

Interviewee

and sometimes com company keep it like testing data too. But testing is come like a, uh, sometimes another day. So the data means whatever you are building in a development, right? So they have to be, uh, tasted well. And the qa, well, like after UQ a it'll be, and the other people will say like, okay, it's pass and now we can a production.

Interviewee

But the problem is when it goes to the product, First of all, whatever the data, again, the testing will be happening in a product data to final after the deployment because you have to see exactly the same replica is happening or not. Because sometimes, again, at the beginning I say that the data can be having a different glitch.

Interviewee

Some data do not update properly because when, um, the data is coming, coming in, either like a batch wise or either like a, um, like continuous real time injection is happen. If there are some sometimes happens that things did not run, the pipeline did not run, so the data will be not updated in your fraud.

Interviewee

So basically the fraud and day will be uh, two. Um, two similar, similar type of data. Ideal condition will be there. Whenever you develop anything. Then you have to taste in your product. So one is that thing. So you have to make sure that your data is there. Most of the time it happened that product data are not updated as a deep data updated.

Interviewee

So one issue is that then sometimes it happens. Suppose like if it is, um, uh, if it is taking a little longer time to update. So this is a one, one problem with the, uh, open source software. The reason is that if in the between, luckily and luckily what happen is, is the Python 3.8 now 3.9, right? And that is existing in your Deb, right?

Interviewee

Uh uh, and then when you move to the prod, prod have a different version. So you ran everything and you've got it like a different version. So like version, you have to make sure that you have exactly the same version in the tape and prod else it'll not work, it'll broke. Broke means sometimes it's broke.

Interviewee

Sometimes it's get totally different results, even though in version, because it's version change means optimization, algorithm change, optimization, algorithm change means whatever you put it in a training, testing, the whole thing will be changing. . So, uh, that's one thing. The second thing, uh, so one is quality issue, one is ing issue, uh, in the broad.

Interviewee

Uh, and the the second thing is, uh, the testing the whole thing, system, the testing. So testing is another laborious thing. The testing means, I'm talking about the whole system, like each component you have to test in a fraud to be run, even though you have like multiple other. , same, same processing follow for another 10 or 15, uh, models already there.

Interviewee

But if there's a new models comes in, you have to, again, testing the whole thing, whole pipeline. So that, that's another thing. Um, yeah. So these are the things are there, and there's one couple of things that they are, like, what happens is like if somebody write a different style of codes, that that is never used and test.

Interviewee

production, that could be a problem normally, uh, as a software. Um, so machine learning, data science part is different, but when you're putting in a production, that spot is different. That more like software engineering. So you have to think about is a totally, totally, like if you are distribute, if you're building a, like a water distribution system.

Interviewee

So where wherever water pipe brokes, you just take that area out and like the pipe out and put it. So similar like a software production system too. If somewhere, if something is different change, you only change that piece. So if the new people come or something like that. I, I was leading a team one time, uh, nine people and whenever they put it in production something, every time we'll be getting something, um, like a broke.

Interviewee

The reason is either they will be changing a sim small thing here and there, or either they will be not commenting out. Couple of codes in the line. . So these are the small, small thing, like, I don't know how you'll be putting there, uh, in your paper or thesis. So these, these are the things at there. They, uh, uh, that's, uh, that's actually, um, could, uh, have a production problem.

Interviewee
And also like, uh, another issue is like you normally, most of the company doesn't give you to put in production in every day of the week. So there will be a, uh, guy who will be taking all the. And there will be a one or two days in the production. So if you miss that one means that day. So you're done for that week.

Interviewee
So you have to go for the other other timeline. So there's a timing too, uh, when, um, people put it in your production. .

Interviewer 1
Okay. Thank you. Uh, what is a problem? I didn't understand. What is a problem with people commenting lines or, uh, you mentioned data.

Interviewee
So commenting lines means, suppose like what happened is when they are testing in a small amount of data sets, suppose like 2000 or 3000 rows of the data sets.

Interviewee
Uh, and you said like, print these data frame, right? So it's, it's printing that data frame. Now put it in your product. You have a 2 million of data, right? And you said like, print that data. And then it started printing. And that end is like, you take it, output is, and it text or CSV pile. Right. And the, they, uh, they said the, the whole, the production line sets or DevOps pipeline sets will say ideal time.

Interviewee

Ideal time means like after a certain time, if it is running, they'll shut it down. Like after like half an hour or something like that sometimes happen. Like it's just printing, printing, printing, printing out. It's taking output it, shut it down because you already fix, uh, uh, your virtual missing timing because, um, sometime it happens, like if somebody put it in something in a loop and the virtual missing is running, so you are binding money.

Interviewee
So there are like a data governance or model governance prospect is there that you have to put a certain ideal time for each VMs. So, um, if you know something is wrong, then you fix it. But if it is. So people forget, people forget to stop it, or people like something is running for a long time and it's just baring money because, uh, cloud service is take money with the time and the space is there.

Interviewee
And also like the space, uh, when you are the sprinting, printing, printing, it takes the space, um, in your, our actual mission, right? So this is just one example out there, right? Uh, so there are multiple example could be happening, uh, examples are there in a different. Scenario when um, people don't comment it out or, uh, exactly.

Interviewee
Basically that's what I'm saying, like exactly. You have to keep the same. Shell like how you put a production in the previous one. If somebody, new guy coming in and do something differently, one or two lines here and there. Do not close it. Open it. Some like, uh, uh, comma colon is. , everything will be fail in a abroad, and you'll be like, multiple time, you'll be running between multiple departments.

Interviewee
Like why my fraud is failed.

Interviewer 1
I see. And is it, so you say you run between different department. Is it difficult sometime to see what, what is the mistake, uh, your team has done?

Interviewee

Uh, Uh, uh, yes. The reason is you very difficult to diagnostics di run a diagnostic as a model, uh, uh, as a, as a, as a model or as a developer to find out where is thetic, because then whole production code you have to take it out and then you'll see some, sometime your code is running very good in depth because they.

Interviewee

QA pass, that's how you go to your production. But when it's production, you don't have access to most of the things because the production access, normally people don't have, that, developer doesn't have the most of the, like access, you just have a certain system access, uh, that you can run these, these things.

Interviewee

So basically you have to then go to either a data engineer say like, Hey, can you please check, um, this thing that is running well from your side or. . Now, if they are busy in something else, they will put it in a hole. Then you have to go to the qa says like, Hey, did you check the data engineer? Whatever they did, um, it's in a product or not.

Interviewee

Then you have to go to, um, sometimes a data analyst who build that dashboard. The final dashboard, why these API is not calling from, uh, the, the SQL server where the, the final data is resting, right? So these are the, these are the, and also like DevOps sometimes happens that DevOps is not uploading, uh, properly.

Interviewee

Uh, the text value changes sometimes what happened, like you put it in a code, um, in a, in a GitHub, and you forgot to commit exactly what it need to be committed. Um, so the, the, there are so many issues are like that. The kitab issues are there. The pro like the QA issues are there. Uh, the data engineer, the table are there, so.

Interviewee

reading, writing, running, uh, everywhere could be any issue. And you, in a prod, you don't have all the accesses to just look into that. So you have to come and contact, uh, all the, all the administrator, all the, all the developer on that time. So if they are, uh, if, if they are out there, but I don't know, like, uh, if, if they're ready to help you, then that's fine on, on that time that if they are like, Uh, like a free or they're not working on something, then that's fine.

Interviewee

But there is another, another thing I just want to mention. I don't know like is it'll be helpful or not, but what happened in, in these most of the cases, this department normally, even though inside of in institution, if, um, all of them are from the same in institution, they have some internal, um, different policy to helping each other, right?

Interviewee

Because it's industry and everybody. want to, they want to keep their, uh, top of the game. Like they want to stay in their top of the game. And most of the times, most of the company doesn't have all the support. So they hire consultant. When the consultant comes, they are very competitive to each other. So the communicating with them, it's very difficult.

Interviewee

If you get stuck one places, it could take two days or three days to resolve just a small piece. Uh, problem, outcome there, you know?

Interviewer 1

I see. Interesting. Thank you. Um, yeah. Okay. I remember the follow up question I was going to ask you. Uh, so, so you mentioned earlier on that, um, Um, some, some, sometimes you build something in dev, in QA it works, but when you deploy, it doesn't work in prd.

Interviewer 1

And the reason is, uh, because you're part of the system depends on other part and you need to debug it with other team and other p teams are not necessarily, um, responsive. Okay. Uh, so, so my question is, is it an issue that is specific to machine learning software system or it's something that we can observe in, uh, general software system in general?

Interviewer 1

Just to make

Interviewee
sure it's, it's everywhere. Like any software system, anything you build system, they build, it's everywhere. Mm-hmm. , see, machine, machine learning. Machine learning is nothing. But we are developing something, eh, just add on to the, the software system. Right. It's sometimes it's, it's we, when you will build this software, then it'll be like full face software, but most of the time what the company keep it for their internal.

Interviewee
uh, reporting system, right? So it's, it's less complicated than their actual, like, I suppose like somebody build, um, like a whole website and looking at the booking data and stuff like that. That's a total different level, more complicated than this. But both of the system is follow the same pattern and have the same problems, um, on the production deployment.

Interviewee
Because even like I know my peers and. They stay even like, uh, weekends, evenings, uh, to just to put it in a production The other day, it's put it in a production. It's like the whole team stays like, almost like 50, 60 people just stays in a call and, uh, and put it in a production, uh, bank, uh, financial issue.

Interviewee
Mostly they reply either night or midnight or early morning, uh, weekends if they. Uh, uh, but other companies, uh, also doing it most of the times in the, in, in a night, if that is impacting directly to their customer. If it is a internal team reporting, then they can do it anytime.

Interviewer 1
I see. Perfect. Thank you. Um, I will ask you one question about maintenance and I think after that we will have, uh, spent our time.

Interviewer 1
So, um, have you encountered issue with data or data source during the maintenance of a machine or next software system

Interviewee

maintenance. So what, what, what, what, what, what, what? Oh, okay. So maintenance is like, when everything is deployed, you need to say like, uh, like you are just monitoring. So, uh, you want to know like if there is a data discrepancy or, uh, uh,

Interviewer 1

yeah, any issues with the data or data. anything. For example,

Interviewee

if you need to update the model, update the infrastructure, update the cloud system, update the operating system, yeah, yeah, yeah.

Interviewee

Anything, yeah. Or data model everything. Yeah. Uh, uh, yeah. Maintenance have a lot of, uh, lot of issues comes in. Um, so suppose like if you, uh, have a, uh, so one company I used to work, they had a, most of the model built out a random. and the random forest was not like, uh, so what happened is whenever you have, uh, one model or two model deployed, either like one year or two year, you have to change.

Interviewee

Like if that's, that's, I'm talking about max, like, uh, you have to change the model or either you have to, um, retrain the model or something like that because the data get change and. So that moment it's a huge problem. Like if you changing that, um, uh, that whole algorithm like that. So basically you have to bring whole thing down.

Interviewee

Yeah. You have to train testing and everything, and then you put it back. Some people do, uh, keep their model as a retraining model, like the retraining means like auto retrain model. So when it'll production after a certain time, it trigger an auto re. But also when they put it in a auto retrain, sometime if it is a neural network model, it could take weeks, sometimes like a couple of weeks kind of thing thinking, because what happened is like the fraud data comes in and it's a huge amount of data and it started retraining on, on, on, on the stop.

Interviewee

So yeah, so one is that that issue sometimes comes, that retraining model will have that. So that kind of point of the time, what they do. Uh, they reach in on a deep data and take all the parameter freeze in a file and then put it, uh, put it back in a, uh, in a production. So that's, uh, basically the model Prospect data.

Interviewee

Yeah. Data is like, if their data source is changing, of course there will be a, uh, maintenance problem. But normally what happen is, uh, when the whole, the end-to-end pipeline. They, that thing doesn't exactly take, uh, data from a iot device. So iot device data stores in a, uh, in a, in a server, SQL server or somewhere or database.

Interviewee

From that database, the things is put, uh, put it out. So if you change the data, So that time, the, the data engineering pipeline, which is basically the, the etl, which is called extraction transformation load, that could get impacted. So that's a different pipeline, but this machine learning pipeline normally stays, uh, the, the way it's designed normally stays intact from a data prospect until, , uh, those who are ingesting from a data engineering perspective, they doesn't know what tables, uh, they are ingesting or what the views are there.

Interviewee

Uh, they, they are updating. Sometimes hast happened by mistake. They rewrite the whole table. So suppose like you have only 10 column, I'm just taking over 10 column and your machine learning pipeline working on based on. and somebody, new data engineer comes and rewrite the data, have 11 column, just one single column, more, the whole thing crash.

Interviewee

So in that moment, probably they will not touch the machine learning pipeline. Mostly they will be going and touching, uh, touching the data, data pipeline. But before to go to that direction, you have to debug, uh, what is the issue. So first we'll be starting from. from the beginning or middle or whatever.

Interviewee

It's dependent upon experience and the team,

Interviewer 1

you know? Okay. And, uh, is it a challenging process to debug? Where is it? Where Oh yeah. .

Interviewee

Oh, yeah. It's a huge challenging process. It's the whole thing I'm describing. Every piece of the, uh, uh, area is very challenging. You need experience, uh, and, and, uh, you really need.

Interviewee

Uh, what you are doing, um, in that area from a manager perspective? Uh, from a developer perspective, it's very nightmare actually. Uh, if things doesn't work, because again, I'm talking about the bank, right? Suppose like bank or medical, like biomedical and all that stuff. So suppose like something goes down, what do you do?

Interviewee

Like you have to fix it. I'm saying like, okay, I, I work for a bottle industry, which is fine. Like something goes down, I say like, okay, take this Excel sheet and. But what happened with the financial data finance sector, um, aerospace, like, uh, those were working in, um, like the flight and all the stuff so that, that area, it's, it's, it's very critical that that cannot be, um, like, uh, like uh, you have to fix it.

Interviewee

And there I see.

Interviewer 1

Thank you. Uh, um, just quickly, uh, why is it complex? Maybe it's obvious. So I don't put words in your mouth. Uh, why is it complex to, to debug the pipeline?

Interviewee

Um, see if you find where is the problem, it's very easy to fix it. But the finding out the problem why it is, is the biggest problem. So suppose like if you have something, uh, caus causing, causing a bruise in your.

Interviewee

So if you find out what is the reason behind that, or is it a skin or inside the skin and all this thing, then, you know, like, I have to treat here. But until Alice, you find out what is, where is happening, where exactly happening, what is the reason that is the main thing because it's a huge line of work.

Interviewee

The channel of work and piece of the puzzles are there, uh, uh, any like machine learning? Any kind of, uh, pipeline, you pick it up any kind of a software development system. So you pick it up and there are so many, uh, people are involved, like one good team, like just a machine learning, uh, a machine learning software development team at least consists 10 to 12 people with experience.

Interviewee

At least like 10. I'm talking about 10 to 12 people with 10 to 12 years of experience, good experience. Then, uh, then, uh, they can build something very sustainable in a. , most of the company right now are struggling to build that, uh, build that pipeline or build that solution. I see. So let me, if we reward it, you are basically, you are saying that usually the symptom is obvious, but finding the root cause is very difficult.

Interviewee

Yes. Yes. Finding the, uh, root cause is very difficult and the location where it is. . Yeah. Localization obviously is a challenge. Yeah. Localization is a challenge. The responsible component, actually The responsible component or default ex. Exactly. Yeah, exactly. I'm talking about too much about the practical thing, but when you put it in the research, you'll have some what?

Interviewee

You have to put it in. Yeah. Have to ship it. The point is that you are doing your job. I'm doing my job. . Yeah, yeah, yeah, exactly. Yeah. Sometimes, you know, like, uh, what happens, like, I, I still write some paper, but not exactly in machine learning and stuff. I, in future, I would like to write in that stuff, but mostly my, my postdoc still is going on, which is, uh, it's a little different than, uh, what I do, uh, uh, regular.

Interviewee

Uh, but when I go in that zone, I have to find a word for that thing. And when I'm working in a, uh, when I'm working in the industry, corporate world, corporate world is different. And, uh, that's, I, I, I'm, I, I come, I, I can still say like, I'm still struggling, uh, to finding a word for corporate world, corporate world is totally different than, uh, than, uh, than the research.

Interviewee
You are right. Thank you. Go

ahead.

Interviewer 1
All right. Um, is there any other issue, quality issue we, we didn't cover about, uh, deployment monitoring that you think will be relevant for us to,

Interviewee
I think we covered mostly because, uh, a data model, uh, uh, the selection of the model, the whole. Uh, these are the 3, 3, 3, uh, three areas out there. Uh, one area is also, um, the quality issue comes around, um, the, the optimization technique, uh, people use for, particularly when they, uh, they're, uh, they are training the model.

Interviewee
Uh, basically the, the hyper parameter tuning and parameter tuning and all, all, all these type, most stuff there. So, uh, there could be a, a issue there, but most of the time I will say people look for the trade off, like trade off between, um, between the time they had, because most of the time what you are getting in a research, you will not get it in, uh, to.

Interviewee
That kind of experiment in, in industry so people find a trade up. So, uh, um, a trade up between the time and the delivery and, and your knowledge. So you have to find a trade up between that. The reason is what I'm saying, uh, you could find a very good optimization model like P C S L, like pro. global search, loosen, uh, or something other like, uh, uh, um, genetic algorithm based optimization algorithm for your model.

Interviewee

You can do research on that thing, right? How things is happening, changing the learning rate and all that stuff. But outside of the world, it doesn't care about what you are changing, little bit changing how the documention is happening and all that stuff. So you have to come up with something which is good in training, very easily training special.

Interviewee

The neural network and uh, um, neural network and method, and easily come up with something that actually they need. So you have to find the trade up between your consciousness. Also, another thing is as a practitioner, , you have to think about that you are delivering, um, that, um, I see that you are engineer, you have that, uh, iron ring, right?

Interviewee

You'll be going for the pH uh, in future. So, um, I, I dunno, do you have a pH or not? But anyway, so you, you'll go for a pH in future. So from your prospect, you, when you are delivering something, you have to keep, um, keep in that way that you are, you are not delivering something totally different from their prospect with the.

Interviewee

And in combination of is there have to be coming up a trade off, you know, uh, uh, between a based model versus what you are delivering. So that's another, you could put a quality issue. No, but it's kind of trade off issue. I will say . Well, it was, uh,

Interviewer 1

Maybe I have one small question for you, and, and you can Yeah, yeah.

Interviewer 1

Answer it in the time you want. Yeah. Uh, so in your opinion, what is the most pressing quality issue researchers should try to solve? Most? What, uh, sorry. What is the most pressing quality issue, uh, researchers should try to solve?

Interviewee

Uh, I'll say like the how you are. In putting data into a machine learning, um, machine learning, um, model.

Interviewee

And it's not quality issue, I will say, uh, because whatever the data comes, you have to build a framework, actually company already build a framework and um, uh, so you have to build a framework, how you treat all sort of data. Whatever comes out, either it's an all value. , um, uh, different, um, like uh, imbalance data sets or, um, something all sort of data.

Interviewee

So you have to build a framework of that thing like data framework where any sort of data comes in that will be robust for your machine learning data input. ,

Interviewer 1

sorry. It's, it's a, a framework that ingests every types of data and make sure it's robust for the model

Interviewee

that that's what you said. Yeah. Model. So, so basically what I'm saying is today, if I'm using a linear integration model to model using like exhibits or random forest ratio, I don't have to change the data, data structure or whatever, uh, the formality is.

Interviewee

The model is coming in, the data is coming in. I don't have to change anything out of that. It just comes in and I can. Use it in, in, in my model, you know, and also, uh, the same model. So suppose like I will take a piece card from a training piece card from a testing and piece card, from a credit, um, like a, uh, uh, forecasting.

Interviewee

All right? So any of these piece done have to touch in that framework differently. It's a similar way you'll be touching and it'll be going in input to your, um, your model. .

Interviewer 1

Okay. Thank you. That's super interesting. I, I, I'd like to thank you for, for the time you spent, we, we went a little bit overboard.

Interviewer 1
Yeah. But I think you, you had a lot of, uh, interesting things to say, so, uh, I think it was worth it, at least for us. For sure. It was really interesting. So, yeah,

Interviewee
it's, uh, very, uh, very interactive with, uh, um, with, uh, you guys by the way, like, how did you find me like. ,

Interviewer 1
it was on, uh, GitHub. We searched for ml, engineering Data Scientist, uh, profiles.

Interviewee
Oh, okay. Okay. No, I, I used to work for, um, one, um, um, one small company, uh, in, um, in City X. Um, uh, it's called Company X. Do you know that company? Uh, anyway, chance? Yeah. Yeah. Um, start, yeah, I know. Yeah. Yeah, I know. But we haven't signed you from. Oh, okay. Yeah, no, no. I thought, uh, maybe, uh, maybe, uh, I, I find you guys find , but anyway, like, uh, yeah, it's nice to talk to you anytime if you need, um, any, any kind of, uh, specific, uh, probably, I, I give totally like very, very broad view.

Interviewee
Sometimes I kind of froze and I go in a, these direction, that direction. But if you, if you need any very specific questions, you can chat. Thank you so much. Thank you for Yeah, no worries. Thank you. Thank you guys. Have a good one. All right. Have a good day. Thank you. Have a good day. Bye. Bye.

Created with the Delve Qualitative Analysis Tool (http://www.delvetool.com)