Interview 8 - PO

Interviewee

Great. Thanks. Yeah, I will. I actually, I started very perfect.

Interviewer 1

All right. Uh, so Interviewee if you can give us some background Infor, uh, information about you, how much experience you have in a med or in general?

Interviewee

Okay. Uh, I have as of now, I guess, uh, 2017, so that would mean, uh, five, five to six years experience as a software engineer.

Interviewee

Technically not bachelor's in software engineering. Uh, so I've been working, uh, in, uh, business intelligence until 2022. I did a bit of machine learning for, uh, proof of concepts, mostly and integration into, uh, big data stacks. So I've seen a bit of everything, uh, in that, uh, business intelligence job. And I've been more focusing on, uh, machine learning since joining, uh, Company 1, uh, in, uh, 2022.

Interviewee

So mostly, uh, my, my work around, uh, machine learning has been focused on the pipelining and auto automation of training and, uh, and evaluation of the models in order to apply them in the real life situations. Great. Thanks.

Interviewer 1

Uh, so I'm gonna ask you some question we have around 25 questions. Uh, you're not expected to answer all of them if you don't know you just so you don't know when we will pass the next one.

Interviewer 1

So, alright. Uh, so what are the main quality issues that you have encountered with your data model or system so far in

Interviewer 1

your

Interviewee

experience? Any data model or systems? Okay. Let's start with the first one data. Uh, data quality is always an issue. Uh, anyone who's done, uh, ETLs knows what, what a paint. It can be sometimes, uh, ranging from anything from, uh, uh, I'd say environment specific, uh, uh, issues like, uh, local time used on servers.

Interviewee

Uh, the using UTC versus, uh, Versus a local time or encoding issues, uh, stuff like that. There's also more often than not missing data in, in the data sets. So even though the data still looks usable, sometimes it can affect your model because, uh, you didn't notice that, uh, either some entries were generated through, um, test situations from the, the developers and or the assignment and, or the DBA end, but haven't been clean from the, the data set.

Interviewee

Uh, so usually you need to clean those in DTL as far as how it's represents in itself in the model. One thing that's specific to machine learning or AI in this case, I think is that you don't necessarily have. A great usual, uh, a great baseline every time to check the results of the, of the model. So even though, you know, approximately what the precision you're aiming for should be, you don't necessarily see the, the impact of, uh, bad data.

Interviewee

If you just look at the model output. So there's a lot of data validation to be done from the, the raw data source to the creation of your, your model input. That's mostly based on, I'd say the developer's experience with several of those issues and, uh, just, uh, basic, uh, verification. That's done. I wouldn't say necessarily manually, cuz you can add steps to your pipeline to just do some data validation, but there's never an end to how much validation you can do.

Interviewee

So there's always trying to size up. What's the right amount of work to dedicate to that, to that part. So for the data, that's mostly the, the issue I'm seeing for the model. There's always, uh, the model drift. So no matter what type of model you're doing, you always have to, uh, account for that and build your pipeline to.

Interviewee

Whenever you, you retrain, recheck what the, the performance of the model is. Make sure it's not going down. And if it's going down, try to find out the reason why it could be data. Cuz if, if you're putting it in a prediction environment, usually you retrain on new data. That's been acquired since the last training and you don't want to do a new release of your, of your whole pipeline.

Interviewee

Every time you retrain the model, it's usually automated. So you need to set up alerts. It's not really, uh, it's not really an issue as in it's impossible to do, but it has to be taken into account and integrated into your, your forecasting of the, of the amount of work required. Because if you skip that step, the client's gonna come back in and bite you.

Interviewee

My and what was the third part of I'm sorry, uh, system systems. Oh, um, yeah, most machine learning situations I've encountered had integrated an insanely absurd amount of systems together because you're gathering data from usually like minimum five, six sources, and they're usually completely different.

Interviewee

So you're impacted by not only the, the, I would say the, the tenacity of your, your current systems that you're using for your Mo for your, uh, machine learning solution, but also those from the source and the destination, cuz, well, if you're predicting something, you probably wanna push it from your pipeline to.

Interviewee

Let's say a, a production system, uh, where the, for example, a dashboard that's used by the end user or a reporting system that's used by managers to monitor the, the efficacy of their, their

Interviewer 1

company. Yeah, I see. I see. Thanks. Uh, well last question. So you, you bring, bring up a lot of good points. Uh, in my, in my next question, I will be able to go more in detail into each thing you said.

Interviewer 1
All right. Um, so do you use any of the following data, that data collection techniques. So do you have any, um, data collectors, people that manually fetch and create data for you to train

Interviewee
your systems? Usually not. Cuz we try to automate that part. Uh, it's something I saw quite a lot, uh, while working at my last employer.

Interviewee
I'm not sure if, uh, Well, I, I could tell you, but I'm not sure if something I want written so I'm not sure if you yeah, yeah. Anonymize that part. But, uh, while I was working for the Company 1, that was probably one of like the biggest issue with how we collected data is everything was siloed from departments.

Interviewee
So oftentimes we had data collector, which gave us dumps for specific time, uh, that, uh, time, uh, of, uh, the request. But that makes your system completely useless on the long term because okay, you can train your solution on that specific point of data set, but it's not gonna learn over time. It's not gonna adapt changes.

Interviewee
So you're gonna have to go back to that person, ask, request it for the new data again. And if. It wastes like 50% of your time when you work that way. But when you are able to automate it, like I've seen mostly, we do a here at Company 2, then you don't have that middle man. And it's not an issue for, for the long term.

Interviewee
Maybe there are issues as in maybe they update their data source and they rename or move some of data, the data you use, but it's much easier to change in an automated system than it is to actually train someone, to do their job a different way. Yeah. Yeah. That's for sure.

Interviewer 1

Sure. Thank you. Um, did you, I think you, do you, do, did you ever use external data?

Interviewer 1

So for example, public data set yep. At third party API

Interviewee

or web scrap data. Uh, yeah, I did a bit, not too much, but I'm familiar with the format. Okay. Is there a specific question with the,

Interviewer 1

yeah, there's a follow up question. So if you did, uh, what were the quality issues you have encountered

Interviewee

specifically with XML data?

Interviewee

Hmm. Uh, external data, not external data. Oh, external. I, I, I heard XML. Oh, external data. Oh boy. That there's a, , there's a bunch of them. Uh, you don't always have the reliability of the source. Uh, if you're doing internally, usually you can do your own, uh, quality control on the data. You can do your own lineage, everything else, uh, everything around that.

Interviewee

But with external data, there's a. There's a disconnect between what the provider gives you and what you use internally. So you end up kind of doing your own, uh, data quality on it, which doesn't necessarily reflect what they have on their side. So you have to evaluate, uh, maybe you you'll give more confidence to some data that they themselves know is not good.

Interviewee

And you're also vulnerable to changes from their, from their end. That's especially true with, uh, external sources that you're not in communication with. So for that, I'm thinking, uh, let's say, uh, you use open data from the city 1 as an example, or you use, uh, material data or, uh, precipitation. Uh, news feeds, anything like that.

Interviewee
There's always the possibility that it changes and the more you automate, the more it risks stopping your execution for a later date. So it's, it adds a layer of complexity to yeah, sure. To that position.

Interviewer 1
Yeah. And how do you verify the quality of the data and you ensure that it stays yeah.

Interviewee
Of good quality.

Interviewee
It depends on what you have. Uh, from my experience, it's not something I it's something I know I should do, and I want to do more, but there's often not enough time to do it, but, uh, mostly I'd say you have like, uh, you can do some batch testing to check the let's say for, uh, if you have, uh, numerical values, the maximum minimum you're expected warn you, if you go over or under a certain threshold, uh, a certain mean, for example, uh, or if it's strings, you can just check if it's empty, if, if it's, uh, expected to be empty or if it's no, that kind of, of things.

Interviewee
But as far as the true quality of the data, the best use case I've seen were completely internal we're. We tracked every system, the data went through and add, uh, what was software's name? Uh, I think it wasn't our big data system. I think it was Atlas or something like that. Uh, if you can, if you let me check, uh, one second, I can tell you,

Interviewee
yeah, it was Apache Atlas. So we could put tags on data sets that, uh, went that, uh, followed through the data in every step of the transformation. So if we found some data to be lacking and merged it with another data set, it impacted the quality of that, that other data set. There are a few tools for that, but it's not something I've seen.

Interviewee

Put into application on a large scale or at least not as much as I want to . Yeah.

Interviewer 1

Um, thank you. And have you ever used data generated by another system which may be ML based or not?

Interviewee

Uh, yes. In both cases actually um, in non ML based pretty much everything we do. Like there's more often, often than not the data's generated by another, another system, often the transactional one.

Interviewee

Um, it's very rare that we get data directly from users or something like that. Uh, even, even if we take web as an example, it's still processed through the website. So I'd still consider that a system instead of a. Direct data entry and for, uh, machine learning system, I had one specific case in, um, in one of our clients where we had two different models that generated the two different predictions.

Interviewee

And we had to take the results from one of the model to set as a default value. When the second model did not predict a correct, uh, something that was within the acceptable range, uh, for, for the user. Uh, I, in that case, I don't think it was directly in the model input though. So there was no impact on the, the predictions of the second model.

Interviewee

Okay. It was just on the pipeline itself. Yeah. Okay.

Interviewer 1

And you mentioned, uh, sometime you used transaction generated by systems. Uh, do you ever encounter issues with this data?

Interviewee

Uh, yes. . Yes. Especially when there are version changes on those transactional systems. Cuz you change the version. Usually you change the data structures too, so it affects your model.

Interviewee

But also, like I said earlier, there's the, the possibility of systems that still have, uh, manual testing that have been done that still have records in them. But probably the most difficult case I've had of that was, uh, when I worked on, um, on IOT devices. So anything Evan based is much more prone to error because.

Interviewee

Whenever there's a latency issue or there's an issue with the micro controllers or anything like that, you get ARN data and that can really mess up, uh, not only your model itself, but also the, the full pipeline. Sometimes your data just will crash in your transformations before even getting to your model.

Interviewee

So it's, uh, okay. It's a common.

Interviewer 1

Okay. I see. And how, how can you prevent these? How do you prevent these kind of issues from happening? Do, do you have ever, do you ever put in place mechanism to fix

Interviewee

these issues? Yeah. Yeah. Usually it's an ingestion for the IOT example. Uh, I, I gave, I designed a notification and, uh, retention system for a tool that's called, uh, called NiFi.

Interviewee

I don't know if you're, uh, familiar with it. It's a tool for, uh, realtime data ingestion. So, um, I think it's Apache too. Well, it allows you to basically, um, connect to a source that in my case was a message cube and ingest the data in real time all the time, do transformations on them and put them on whatever storage you want.

Interviewee

So I designed, um, a system to. Check the data while it was transiting at certain key stages. And if it didn't meet specific parameters, the data flow still continued. But the data that was, uh, let's say corrupted, but that was not of the quality that we, uh, assessed was good enough. It was stored elsewhere for review by one of the developer.

Interviewee
And the developer was notified by email and instant messaging. So we could go back on that, that specific data that did not meet the quality of requirements, check what the issue was and possibly raise it with the provider of the data because of more often than not. It was just, like I said, in IOT, just let's say, uh, a specific, um, a specific controller lost connection to the internet.

Interviewee
well, the server that collected the information from those controllers kept sending again and again, the same value that it had at the last message it received, for example, or it just sent an art beat that shouldn't be there. Uh it's sometimes it even sent like the same, it sent a different message, but the, the controller was not connected to the time servers anymore.

Interviewee
So the internal clock of the, the device was stuck on the same time and date. So it had it added new entries, but it was all at the same date time. So it's the kind of thing we could catch. in very few cases, we were about, we were able to just correct the data and add it back to our data lake. But in most cases, it was there to notify the owner of the, the system that, oh, you have a problem right now.

Interviewee
Go fix it. And then it helps, uh, in the long term.

Interviewer 1
Okay. Understand. So, so to, to fix this kind of issue you have described, uh, you have a data validation system in which you, you wrote some Rob, if data is outside of expected boundaries, put it in, put it aside. And when it is aside, while someone will check, what is the problem?

Interviewee

Yeah. Perfect. I see. Great super in a perfect world. I would've liked to automate that part too, but at some point you, you gotta end somewhere, so yeah. Good

Interviewer 1

data. Yeah. Difficult. All right. Um, okay. Have you ever measured the quality of your data and, or tried to improve it?

Interviewee

Uh, yeah, all the time. Um, the best example of that would be, uh, while there was the, the tool I mentioned, uh, on Hadoop that allowed us to, uh, tag the data.

Interviewee

And according to the checks were made set of data quality and returned back on it. But there was also, there was also a project where we had an issue that from ex we used a pretty much only data from external data sources that we didn't have contact or control over. So all web, uh, web provided and. We had kind of an issue that, uh, some of them sometimes went online or had missing missing days of data, things like that.

Interviewee

So, and we had about, I think it was over 50 different sources on a, on that project. So we had a lot of issues tracking, which data came from who and which data was, let's say, uh, production grade, or, uh, just experimental. So I made in the, in the database we had, that was a more primitive, uh, primitive, less automated, uh, system.

Interviewee

But I made, uh, uh, some, uh, stored procedure on our database that automatically forced the use, the. The user, the data scientist that added the da, the data to specify exactly from where you took it. What's the, the level of quality he or she estimated for that data and, uh, a general, general, uh, idea of what was contained inside of it.

Interviewee

So, and when it could be used. So we used that those store procedure added everything to, uh, I think it was three or four different tables afterwards. And whenever we generated new data sets from those sources, we added that were collected and added to the database internally. The new table generated from the, the, the joins was also took the, took the information from those table and generated a new report for that one, depending on, let's say the worst data available or things like that.

Interviewer 1
Okay. I see. Great, great. Super. So, in, in some way, you first, you first data scientist to input metadata for each of your yeah, exactly.

Interviewee
That you're putting nice. Exactly. Both meta metadata and lineage also. Yeah. Okay. Sorry. I have a question. And was that enough? Were you performed? Um, user adoption was an issue oh, let's put like this.

Interviewee
Uh, it was enough for, I think a year, let's say, let's say 16 months, something like that. But, but afterwards there was a point where it became too much of a burden for the data scientists mm-hmm and they pushed to have it removed. And I can understand why, because we reached, I think about, uh, 120 hundred and 30 tables, and sometimes we were mostly in the proof of concepts at that, at that moment.

Interviewee
So especially considering it wasn't really automated well, like I take responsibility for that, but. We didn't have the resources or the time at that moment, it was pure built, uh, by me, uh, me alone. So that made it, so the users, uh, while the data scientists decided that, okay, we're not going to use that anymore from now.

Interviewee
I thank you. No problem.

Interviewer 1
Thank you. Uh, and is there any other data quality issue we missed that you cons consider

Interviewee

relevant?

Interviewee

Uh, let me think about that.

Interviewee

I'm not sure if it would count in data quality itself, but it's adjacent the there's often an issue with the quantity of. , uh, especially since I've worked, uh, a big chunk of time on big data. A lot of clients want the results from years and years of data, but start collecting it the day they start the project.

Interviewee

So that's usually a big issue too, where, okay. We have data that is good technically, but we don't have enough for the model to give a good result yet. And trying to skirt around that with clients and explain to them that, okay, yeah, we can generate something, but it's not necessarily gonna be as good as it would've been with a year, two years, three years of, of data is something that's, uh, that I've found quite hard to explain to people are not.

Interviewee

as knowledgeable in AR or machine learning. I

Interviewer 1

see. I see. And do you have any, um, I mean, in a way to not prevent this kind of issue, but to attune weight, this kind of issue to the, the lack of data

Interviewee

quantity, uh, there are a few ways, depending on what project you're working on, uh, you can generate dummy data from a range that you, you know, throughout through the business rules.

Interviewee

Like if you're in con in contact with, uh, clients that are mostly, uh, it's a mostly hands on specific in their own field, let's say, uh, if you're talking, uh, to someone who works in construction or a grocery or, or anyone in the supply chain, they kind of have. A gut feeling of what they should expect. And they have business rules that allow them to gauge about what they expect.

Interviewee

Uh, so we can generate data from that to alleviate, but that makes our model maybe good enough for the first year until we have the actual data itself. So that's one, the, the, the other solution is to get that depends on, on the use case, but get the, the client to generate data himself. And I'm thinking of, of another client, uh, right now that we're talking to that basically works with classified information or information that it's not, uh, that they cannot record.

Interviewee

So one thing they will do is they'll go on the Terra and. Play out scenarios that would generate about the same data as they would have in a real life situation. But it's not that specific data. Okay. That's interesting. And last one, I just had a flash. Yeah. Go for, uh, I did a bit of, uh, machine learning on, uh, pictures and videos, so deep learning on those.

Interviewee

And one thing we did quite a bit to generate data was, um, do simulations through, uh, game engines. So we had basic textures for different key, uh, elements we wanted to have in the, in the picture and just had it running in installation in a game engine and take, uh, videos or screenshot nonstop for. Two three months and use that as extra footage to, uh, to gauge the, the quality of, uh, of our system.

Interviewer 1

Okay. That, that's really interesting. So you're saying that you reproduce reality in a video video game. Yeah. And you put your AI model in it, or, or you extracted some data visual data

Interviewee

for that, that specific, uh, case I'm thinking about? Uh, we had to determine features in a video. Uh, I ended up using the, using something called Yolo.

Interviewee

I don't know if you, you know, about it, it's real time object detection, but before we started the, on that path, cuz we had the video camera fees that we wanted to analyze in real. what we did was we had, uh, and we didn't have any annotation for the videos we had at the time. So what we did, we had, uh, a, what is it?

Interviewee

An English, uh, an intern, uh, an intern was generating data on Newt with just, let's say a street houses, trees, uh, sidewalk, and everything. And we tried to apply the detection on that generated feed, and it wasn't as perfect as it would've been with real data with, but it was still actually pretty decent.

Interviewee

That was before we used the, we used the out of the box solution, like you.

Interviewer 1

Okay. And how, and how do you label? So I understand you put your model in the virtual world, but out yeah. So do you train your model in the virtual world or do you test it in the virtual work?

Interviewee

Uh, both at first, but then we ended up carrying the, the same model over and retraining.

Interviewee

And, uh, we started, at some point we trained it, trained it with like 80%, uh, model generated and 20% real life, and then tested it with the same proportions on data. And we tried to dial down the quantity of generated, uh, input and, uh, the quantity of, uh, generated test data.

Interviewer 1

Okay. Great. And how do you label the data in the virtual work

Interviewee

for, for that specific instance?

Interviewee

It was, uh, we had, uh, I don't remember the, the, the website, but there was, uh, a service for data labeling that we hired. It was, I think something like, uh, thousand dollars for 500 pictures or something like that. Okay. I see. So we sent key frames from the, the feed, but that was literally pixel by pixel, always labeling the, the full, uh, the fuller objects.

Interviewee

There's also the possibility of trying to label it through another machine learning, uh, algorithm. But, uh, when we tried that we didn't have the time to. Actually fine tune it. So it would give us a result we wanted. So we preferred to resort to manual, uh, labeling. Okay. I see.

Interviewer 1

And did you have any issue with the data collectors?

Interviewer 1

The one that labeled the data?

Interviewee

Uh, yes and no. Yes. We had issues, but not enough to actually actually impact greatly the results. Like we skimmed through what they gave us and sometimes corrected like two or three things and went back and, but it was not even like 10% of the pictures we had to edit, so it wasn't that big of video.

Interviewer 1

Okay. Thank, thank you a lot. That's very interesting. Uh, so we're done with data. I will go through model evaluation and everything related to model. Um, If, I guess you have less experience in model evaluation than in data. So if you don't know, you just say, I dunno. All right. Uh, so how do you evaluate the quality of

Interviewee

models from what I've seen?

Interviewee

And it's not my expertise since I'm not data scientist, but more on the engineering side. Uh, usually the way we do it is our data scientist that's defined on the project, identifies the, the metrics that are the key metrics for that specific model, because every model is different, uh, metrics that are more efficient on it, I guess.

Interviewee
And then we have it generated automatically during the testing phase at the end of the, the pipeline we compared that, uh, those metrics. Let's say the 10 last executions, the 10 last trainings of the model. See which one is see if it's better or worse than it was before. And we have thresholds where, okay.

Interviewee
Maybe we still use it. It's okay. Maybe it, it dipped a bit because the data and the input changed, it's fine. It needs some time to readjust or maybe there's something clearly wrong. And then we go back and try to find that out. What was the issue, uh, with the, the metrics?

Interviewer 1
Yeah. I see. Thanks. Uh, so you're, you're really looking at accuracy and precision.

Interviewee
Yeah, it depends. Ed precisions sometimes. Uh, also, uh, sometimes the mean error rate, uh, was used another project, but to be honest, I don't really, uh, Look at which metric we use. It's just, okay. You give me the, the mathematical formula. I apply it and I compare the results afterwards. Yeah. All right. Me on that follow up.

Interviewee
Sorry. There are people that are smarter and smarter than me that have this job.

Interviewer 1
everyone else is expertise. Yeah. Uh, so I see. Time is sleeping. So I will go, I will jump to, you mentioned earlier that you had maybe some experience in maintenance of ML model and maybe deployment. So I will go to that. And, uh, I think after that we'll be finished.

Interviewer 1

All right. Perfect. Uh, so how and where do you do, does your model are deployed?

Interviewee

Sorry, where my models are deployed. Right now, uh, since working at Company 2, it's pretty much all on the cloud. So both the training and the inference is usually done on the cloud, but it always depends on the client. Uh, like the other example I, I gave earlier that had, uh, data that was, uh, sensitive.

Interviewee

Well, they're gonna have to deploy their model on micro controllers. So that's gonna be a big challenge because we're gonna have to make sure it works without access, internet access, gonna have to make sure it builds on that specific controller. Cuz we don't know yet what the operating system is. We don't know what's the requirements in term of Ram hard drive space, uh, computing power and everything.

Interviewee

So that's one we're looking to do soon. . And when I was at my old, uh, job at Company 1, uh, we mostly did deployments internally either through, uh, microservices, uh, on ES or literally on bare metal, uh, instance. And I also did technically deploy some on big, big data clusters.

Interviewer 1

Okay. And have you started or not the, the project on

Interviewee

microcontrollers?

Interviewee

No, it's not started yet.

Interviewer 1

Okay. I would've loved to ask

Interviewee

you a question about it. It's it's not even signed yet, so okay. No, I, I'm not giving names of clients specifically for that reason, but yeah. Yeah. It's, it's the kind of project I I've seen float around and I know a bit a thing or two about. Okay. So,

Interviewer 1
um, what are the challenge you encounter while deploying model or model me learning software system?

Interviewee
Yeah, uh, specifically, uh, for machine learning or in general, uh,

Interviewer 1
machine learning software system. So, uh, software system with a machinery component,

Interviewee
cause there most issues personally I've encountered with those while deploying, let's say from a technical standpoint are issues that I've seen with the deployment of any software.

Interviewee
Like it's be it net network issues, data issues, uh, just cost issues sometimes. Well, actually that's one, that's probably even more relevant to machine learning. Uh, cloud makes it a lot easier. Cost is usually a bit, a big factor for those because clients are used to the transactional cost of systems. But when you go towards machine learning, just the training time sometimes costs a lot.

Interviewee
Like, uh, there was a project, uh, which was, I'm not sure if I would label it big data, but it was close to, uh, to te well, 20 terabytes of data in entry that we had to, uh, scale horizontally for the, both the training and the inference because was the model input was just too big. And we were racking up probably around nine to $10,000 of computing time every month.

Interviewee
So. You gotta justify it to the client. That's that's kind of an issue. Yeah.

Interviewer 1

Okay. That, that's a really, that's really interesting. So, so you're saying, um, there was too much data coming at the same time, something like that?

Interviewee

No, it's just that we had, yeah. Oh, well, yeah. For the model training. Yeah. Sorry. Yeah, there was too much data in the model input

Interviewee

to train on a single node. So we had to train it in parallel on, uh, we ended up going for 20 nodes on AWS, which cost about, I think was 40 cent a node an hour. So it racked up the cost more or $4 an note an hour, sir. So cost a lot more than the, the initial system the client had, which was one node, a single node of the same size, but we scaled from.

Interviewee

Uh, let's say, uh, five stores to 87 stores. Okay. I see. And

Interviewer 1

did you just go with the, so, so it was casting a lot, did you? Yeah. Sorry. My question is not clear. So it, it cast a lot. Were you trying to video consider going to a simpler model, so cost less where you, you kept the, the solution that

Interviewee

was more expensive?

Interviewee

Uh, the problem, I don't think the problem was the, the model type we used, cuz we used the light GBM, which isn't really a heavy one. It's just the quantity of data. We're talking about a classic service industry chain and 90 stores, a lot of stores. Like if you wanna scale your solution. Sometimes you'll have to, to deal with the fact that you need to increase the, the costs.

Interviewer 1

I see. I see.

Interviewee

Thank you. Scaling is also an issue that can happen in real environments where you just, if you wanna apply your solution to your whole business and your business is big, your solution's gonna be big.

Interviewee

That's interesting.

Interviewer 1

I wanna ask you follow up, follow up question,

Interviewee

but I gotta think I, I can extend a bit if, uh, if need be. Yeah.

Interviewer 1

Uh, I mean we have maybe, yeah, maybe for two minutes, I'll never will ask you a few question and then

Interviewee

we will be done. Perfect.

Interviewee

Um,

Interviewer 1

yeah. I, I thought you were going to continue

Interviewee

speaking. Oh, continue on the scaling. Oh, sorry. Yeah, well, yeah, there's always the limit. Of both the hardware and also the, the software side of things. So either hardware as in, okay. Do you have enough Ram CPU on your machine depending on what model you're training?

Interviewee

Sometimes it's more CPU intensive, sometimes it's more Ram intensive. Uh, but there is also on the software side of things where some libraries for some model types are good until you reach a certain quantity of data where literally it becomes exponentially slower to train it. So you have to either fine tune your data, set in the input to make sure you have less data, but that would also technically reduce your accuracy.

Interviewee

So you don't necessarily want to do that or switch to another library that. Does more or less the same thing uses the same algorithm or, but differently. So for the example I was giving the that's a real, a good representation, a good use case for that, where we use the live GBM model on both the initial solution and the new version, but the new version was killed in parallel with the PI spark.

Interviewee

So it allowed us to technically deploy it on 100, 200 machines. If we wanted, without having the exponential increase, we switched to a model that allowed us to have linear, uh, linear cost in resources. And that's still horizontally instead of vertically.

Interviewer 1

Okay. Uh, so basically if I'm interested correctly, you add one library and you add PI spark.

Interviewer 1

And by putting it at PI spark, you were able to reduce the cost because it paralyzed better

Interviewee

than it's not reduced the cost. It it's not reduced the cost. It's literally that would've been impossible without it, because we were getting to around 256 gigabytes of Ram with the single node and literally the biggest node available.

Interviewee

So you can't really go higher unless you, you pay for, let's say you rent a super computer, but okay. The cost's gonna be bigger, but even then the super computer's not necessarily gonna be available every day at X hour. So you need to find a way to use cheaper hardware, more available hardware to, to run your, your solution.

Interviewee
Okay. I see. Thanks. Thanks.

Interviewer 1
Um, and okay, we'll move on to model maintenance. How do you ensure that the quality of machine learning software system does not decre decrease over time?

Interviewee
Well, well, the, the metrics that, uh, we mentioned earlier are, are the first part. Uh, usually there is a certain point where we kind of assume it's always gonna be the case cuz data's ever changing.

Interviewee
So the models and the parameters we used are not gonna be necessarily the, the right one for two years from now or three years now, from now in a, in a production environment. So we keep a VI visualization on the, the, the results of those metrics. And if we see. An actual trend and not an outlier with this specific training, then it's time to go reevaluate the, the hyper parameters used for the training itself.

Interviewee
Okay, perfect.

Interviewer 1
And have you had any other issue regarding the maintenance of your model

Interviewee
or system? Um, I'm not sure it fits in maintenance, but somewhere between maintenance and deployment. One of the issue I faced a lot was user acceptance because users are very quick to blame the algorithm for things that are outside its control or for their own misinterpretation of the data.

Interviewee

So especially with very technical people in their field, they have their own point of view of. What the results should be, and they expect the algorithm to give them what they expect. So if you get a result that is better than what they expect, they tend to think that it's wrong instead of being happy that they get a better a result.

Interviewee

So the maintenance part of that is that every time you deploy a new version or you retrain a new version of the model, there's always this small feedback of, oh, I'm not getting what I used to be getting. Why is that? What changed? And then they start not trusting the result and not using it for their day to day activities.

Interviewee

So it's more of a psychological one, but still, still there. Yeah, I have a question. How do you address it? Sorry. How do you address such situations? One explaining to the user? Yeah, there's always the, there's always the outtake of explaining when they found out. Usually you, you need to take some time with them.

Interviewee

Usually two, three sessions of like 30 minutes an hour, just to more, more often than not. It's just to reassure them that it's normal and try to convince them that the result is good. Uh, we usually, we usually, uh, generate some graphs that show them the progression in time. So what they they're used to seeing versus what is now, and we try to point out what the, the source of the slight change was, even though it's usually found by the model itself.

Interviewee

We can still try and explain it as best we can as to why what changed. So either the input data, we, we have more, or, uh, it's more precise than it used to be, or maybe just, uh, the models getting better at this, this specific situation instead of all the editors. So it happens a lot with the edge cases where there's a lot of variance, but, uh, usually the, the way I feel, uh, works the best is to not have them notice it.

Interviewee

I know it sounds weird in a way, but, uh, I've tried to not alert them of any changes on the back end. Because it's be, it's kind of a self fulfilling prophecy at some point that if you tell them, oh, we have a big new, uh, uh, model that's coming, they're gonna nitpick it. So if you don't tell them, you wait like two weeks and you tell them, oh, it's been rolled out for two weeks.

Interviewee
Usually, uh, they don't contest anymore. because they should have noticed at that point, probably. So it is mostly a psychological reaction. Yeah. Yeah. Mostly it's very, and it's very present with experts in their field. And I totally understand, like, uh, usually the interactions we have with them is very thorough and very, uh, very often during the exploration part of the project where we need to understand their.

Interviewee
Business logic, their, their comprehension of their specific field. But when we get to production, we try to tone it down a bit and be more black boxy around our, our solution, just to make sure they don't bias themselves out of it.

Interviewee
nice. Thank you.

Interviewer 1
You should write the guide, the best practices.

Interviewee
that's more project management than the than engineering though. Yeah.

Interviewer 1
Uh, you mentioned one thing I, I, I think was interesting. You say, you said, um, I will go quickly. Uh, yep. So when there's a change in the model, you try to explain to the client why it happened or, or change the data, something like this.

Interviewer 1

Um, do you not

Interviewee

change, not change the data, but just explain what changed? Uh, we had a few issues, uh, uh, with the last, uh, last time, uh, we put a model in production that were, that was specifically that case. And we ended up like going through every single data set. We added input. See if there was any change that could, uh, oh, losing everything.

Interviewee

Do you still hear me? Yeah, we hear you. Oh, okay. It's just, I don't see any cameras anymore. Uh, Okay, I'm back. Uh, so yeah, we went, we combed through the data that was used in input just to make sure everything was right. We, we isolated the specific situations in which they were uncomfortable with the result.

Interviewee

So we could trace back where that came from. Of course, the model part itself. We, we can't just go inside the model and see what, what the, what transformation was made to, to get that prediction. But we could monitor the changes on, on the input side and we could give some counter arguments on other, uh, other results in the output side.

Interviewee

So for example, if we, uh, if we, since we scaled from like five stores to 80, 87 stores, Well, the model, there's one model for all of them, the model itself, it's kind of expected that it's gonna be a little bit less precise for those five stores, because it's not all the data that's specific to them, but it's gonna be more precise or it's gonna be almost as precise for all the other 82 stores that were added.

Interviewee

So it's the, that kind of, a bit of, uh, reverse engineering that we do to justify it to client. It's more, more about narrative most often than about actually finding out if there isn't really a technical issue.

Interviewer 1

So you check the data and you try to reason or, oh

Interviewee

yeah, I can hear you. Can you hear me

Interviewer 1

okay. Yeah. So what, what I was saying, I was summarizing, summarizing what you said. So you check out the data and, uh, you tried to reason why, uh, it changed like on your best intuition physically. Okay. Thank you. All right.

Interviewer 1

Uh, so that's our last question. I will ask you the next one. Um, so in your opinion, what is the most pressing quality issue? Researchers should

Interviewee

try to solve researchers specifically as in, okay. As in developing a new, uh, let's say solution or a new way of thinking that would help in business practices, I'd say, and most important quality issue that should be addressed.

Interviewee

Yeah. Most pressing quality problem. I I'd say probably. Um, data quality and input. Give us a way to actually get a good sense of the quality of the data without necessarily knowing what the data is because yeah, we can evaluate the quality of any data, but when I have 25 terabytes of data to analyze, I can do it manually everywhere, especially when you have multiple sources and everything.

Interviewee

So yeah, having a way to highlight problematic data sets, and I'm not talking about, uh, specifically about missing data or something like that. That's pretty easy to rationalize, but more like seemingly good data that is bad. if it makes sense that for us, I think would be the best it's the, the most lacking part.

Interviewee

Right now. Cause yes, there are tools, but most that I've tried, always require so much time investment that they become less worth it than just discarding data that we find is suspicious. So that's I see. Probably I guess. Yeah,

Interviewer 1
it's a really, really

Interviewee
good input also, especially. Yeah, go ahead. Less, less footnote.

Interviewee
Uh, especially since we're hitting a point where data collection has been great for the last 15 years, I'd say maybe 10, but data usage itself is still at kind of its infancy from, from my, my perspective. Like we have a lot of data from million million of sources that are just unused right now. There's.

Interviewee
Maybe what 2% of the data collected that is actually used for something. So being able to just filter out everything that's bad from the get go without too much time investment would allow us to multiply our inputs for machine learning and then be able to have like way more complex models without having to waste 50% of our time to just clean up the, the data beforehand.

Interviewer 1
Yes, it's a, a great, uh, closing, uh, idea. I agree with you. All right. Uh, so thanks a lot.

Interviewee
for, uh, my

Interviewer 1
pleasure spent your time tonight, especially after work. Uh, so it was great talking to you and I'm sure it'll be really impressing for

Interviewee

our study. All right. Well, thanks a lot to you guys, too. And, uh, good luck with your, with your.

Interviewee

Project slash thesis, not sure which is, but, uh, yeah, I, I hope to see your solution as soon as possible to try and try it out, uh, in the, in the production con context, I don't expect us to add to an, the release of any YouTube. Yeah. But you know, I'm still young and I'll still be able to use it. My point that even if you have some update, we not an

Interviewee

you so much for your time. Thank you too. Have a great day. Thank you. Thank you.

Interviewer 1

Bye bye.

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