Interview 17- Rached

Interviewer 1

Describe the goal of the goals of the interview. Uh, so what we want is to develop a catalog of quality issues in machine learnings software system. So let's, uh, break down that sentence. What is a quality issue, or more specifically, what is quality, uh, given two system that achieved the same thing? Uh, if you can say that one system is better than a than another, usually what you're referring to is.

Interviewer 1

And so the system that is worse may have quality issues. A machine learning software system is just a software system that has, that has a machine learning component in it. So we want to develop a, a catalog of quality issues in machine learning software system. Uh, we'll ask about 20 questions. Mm-hmm. , uh, so yeah, and if you have any questions during the interview, feel free to ask them.

Interviewer 1

Uh, if you want to clarify at some points or anything. Sure. No

Interviewee

problem.

Interviewer 1

All right. Uh, so to begin, can you give us some background information about you? Like how many experience, how much experience do you have in ml? What is your current role,

Interviewee

et cetera? Okay. Um, I joined my current company. Um, about six years ago, like more than six years ago.

Interviewee

And before that, before I joined my company, I was doing physics. So it's a different field. So I, uh, before, um, right before I joined my company, I went to a training, uh, training program to become a data scientist. So this is my current company is my first data scientist. and I stay in this company, um, as a data scientist, um, all the way until now.

Interviewee

So I didn't change my position and our, uh, so my, my learning was quite, uh, general. So I, well, in a, in a training itself, it's quite general. It didn't. Um, focus too specifically. Um, any topic. So it's quite general and our company, after I joined the company, the, uh, focus, uh, is developing the core product, uh, which is a recommendation, uh, system.

Interviewee

It's a size recommendation system, so it now becomes a specific, more specific topic, and I focus on that as, uh, in e-commerce. In fashion. So basically, uh, we have to know some, something about fashion, uh, some about closing, and also something about selling things online. So that becomes a narrower topic and we focus on this, um, until now.

Interviewee

So that's my, um, my background in data science really regener. Okay, great.

Interviewer 1

That's really interest. Uh, so to begin with, uh, I will ask you the first question. What are the main quality issues you have encountered with your data model or system so far?

Interviewee

Um, the, the data has quite a bit of issues because it's not very stable and it's not very reliable.

Interviewee

Because we are a B2B company. Um, so we get the data from, um, not directly. So part of it is directly from the end users, the shoppers, online shoppers. We have a widget in place so that how we can shop quite well. So the user enters anything, it goes through, um, the internet and go to our phone space. Uh, Our, uh, cloud, uh, system, and there's not much issue.

Interviewee

So it's direct. Um, and we control the questions our company. So, uh, let's say it's multi optional question. Then whatever the user clicks on it must be one of the three, uh, one of the few multi uh, options. Then, uh, there are are a bunch of, um, other. Kind of kinds of data from the, the business we're dealing with, like the o online shopping companies.

Interviewee
They send us the purchase data of the, their customers, and then se send us the garments data, like the measurement of the clothes and so on. So, such part of the, this part of data is not so reliable because, It depends on the, the unit in the business. If they have a lot of good engineers, good data people, they can send really, uh, more, more reliable data.

Interviewee
And sometimes, uh, they don't even, um, well, it's probably more so for smaller companies. They just, uh, say, okay, the purchase record is, uh, we have a list of who are. Purchasing WA clothes and uh, maybe the format is not correct or there are some missing, uh, boxes in the sheets or something. So they still send us and we, um, well, it depends on.

Interviewee
If our people can just simply figure it out. We have some automation just to convert different, uh, maybe we, we see several mistakes several times and we, we know what, what this kind of mistake is and we just convert automatically and sometimes we have to make out, uh, just talk to the clients to send more accurate data.

Interviewee
and for, for the garment measurement is similar. Uh, you, you, you, they measure the garments, for example, and they say is, uh, let's say the, the, the sleeve lens is, is what, how much the, the, uh, color, uh, diameter is, how much and so on. And sometimes you see obvious mistakes. So, so that kind of thing. When it enters, uh, it enters our system.

Interviewee

It's not, uh, we are not totally sure the, we can detect, detect such error. So sometimes it is a number. If a number is 10 times, it's quite obvious a mistake, but the machine doesn't know because there is no hard limits. Like, what's the size of your lens? It can be, uh, maybe they, they, they took inch, uh, by mistake for centimeters or something, but they, they didn't notice and just put the number in that, that kind of thing.

Interviewee
So, um, we try to have the, a better automation in detecting such error, but it's not always successful and yeah, so, so this kind of data you can kind of imagine it relies on who you talk. It's not in our control, just what you think. That's

Interviewer 1
really interesting. Thank you for all this information. Uh, so, so what are the tools you use to try to detect issues in the data?

Interviewer 1
I think you mentioned that you have some process to kind of verify.

Interviewee
Um, after, uh, we have a bunch of people in our company who are more, uh, garment experts. So they more or less set some rules. Um, so for example, the, um, the, the lens of the, the back of a jacket, if it's more than some number, then you kind of labeled it.

Interviewee
Uh, so, so they are se some of such rules. We so far don't. Um, well, as far as I know, because I'm not in, in the team, but as far as I know, there is no machine learning in detecting such, such error. So our team, we are producing, uh, we are making a recommendation system, but the input part of the data, we just ask some other team to prepare the data for us.

Interviewee
And the other team, as far as I know, they don. Uh, machine learning experts in the team, they just set some rules. They are more like government experts. Okay, that

Interviewer 1

makes sense. Thank you. Thanks.

Interviewee

Um,

Interviewer 1

so I'll ask you, so do you use any of the following data collection techniques, whether it is data collectors, so people that create data for you manually?

Interviewer 1

Uh, external data. Well, you just mentioned, uh, one, so an external data source can also be, for example, a public dataset, a third party api, I, or web script data. So data collector, external data and data generated by another system. Uh, this one you already mentioned it

too,

Interviewee

so Yeah. Okay. Yes. And, and you are also asking public available data.

Interviewee

Yes, yes. Uh, we, we have one simple, um, uh, a sub recommendation system for, for children who are growing in their, uh, height end weights. And we, uh, just by our own data collection, we cannot have a good enough recommendation system. To predict like the, the growth, uh, because when children buy clothes, you, you know, they, they kind of expect to wear, uh, for some period of time.

Interviewee

It's not only right now, but like in several months they see, want to wear it. So we, we have to have a little bit of this, um, prediction of how much they will grow in the following months. Uh, so this part of, we, we cannot. Collect just by asking our clients. They don't have neither. So we, uh, have this, uh, I think it is.

Interviewee

Um, c d C data is, uh, American, um, public. Public health data. So they have the, the, uh, trend of, of the children growth, uh, statistically collected in some databases is their government data. So, so we can just go online and, and just find the data sets and for this combined with our own collection from our clients, then we make this part of recommendation for children.

Interviewer 1

Okay, great. And do you have, is there any issue in the data, uh, that you're given that you reflect from the C D C uh,

Interviewee

webpage? Um, not much problem. I will say just, uh, um, the formats, um, we, we carry it from a table in the, I think, table in a pdf. And as, uh, there is a bit of conversion, but it is pure engineering, so it's not much of an issue.

Interviewee

Okay. Yeah. So, so not, not much. Okay, thanks.

Interviewer 1

Um, have you ever measured the quality of your data and or tried to

Interviewee

improve it? Um, yes. The, there's a quality of, of data. By, um, so, so, so this part is, is in, uh, in our company is mainly handled by another team, but we have some, um, some role of consulting them. Um, so the, the, uh, the control is, well, the, the detection of the quality of data is.

Interviewee

Um, they will see the fluctuation of some, uh, KPIs. Uh, for example from a shop, the amount of daily purchase or daily use of the widgets, for example. Uh, it can fluctuate in. In time. So every day maybe there are a different number of how many users use the widget. If, if it changes too much, then um, then, um, well, so, so you observe for a certain period of time and you kind of assume this shop, the, the.

Interviewee
Uh, the shop has certain normal fluctuation, and if it then change too much away from such fluctuation, then you question whether there is a. Uh, problem on their shop, sending us data, uh, something like this. So there are several KPIs like as IED already mentioned, uh, the amount of data they send to us, uh, because it's continuous in in time.

Interviewee
Um, some shops send daily data. Some are like weekly, but it's always continuous in time. So we expect the amount of data to be roughly, roughly in a range. And other things. I think, um, we, so far don't, don't detect much, uh, many on different segments of, uh, more or less the amount. So this amount, like we have a lot of different, um, kinds of, Collections, uh, purchase is one and use of widget and, uh, the return of, of their products and so on.

Interviewee
So that there are different things and we can segment them to different categories like, uh, male and female products, and, uh, different like diff, just subcategory. And, uh, basically the amounts is the, the, the check. And the inside amount, the, the actual number. Um, what, like the, the feature of each data point?

Interviewee
Right now we don't have a alarm, uh, because it's, it's really hard to differentiate whether it is the real situation or it is a mistake by, uh, in between the client and. So they're somewhere in Theran. Uh, the, the transfer of the data. There's some something wrong or it's actually they are faithfully recording the real situation.

Interviewee
The change is just, uh, reflecting the real world. So, so it's hard to differentiate, but this drop of data amounts, this is harder to believe. That's natural. So we, we use this more.

Interviewer 1
Okay, I understand. So the main strategy you use to see if there is issues in the data is changing the behavioral behavior of the model, like changing in KPIs or, or things like that, basically.

Interviewer 1

So indirectly,

Interviewee

yes. It, this, this part of check is before the model, before the data enters the model. So the, the data sent by the. We have a check just to check the, like the daily amount of something or DA daily. Mm-hmm. K of something. Uh, then is before, before it gets into the model. So check is not the result of the model, but the data right before it enters the model.

Interviewee

So basically you are checking the correctness of the data before entering into the model? Yes. Okay. Thank you. Thank you.

Interviewer 1

Uh, , what are some of the issues you repetitively encounter when you try to prepare data for ml?

Interviewee

Um, the issues, well, um,

Interviewee

I, I will say

Interviewer 1

the main issues. Yeah. The main

Interviewee

issues. Yeah. Yeah. Um, just to, to trust. Um, just to trust the data is the, the whole pipeline of data. There could be some problem that, but we don't know and we are still struggling in coming up with a better idea to detect any issues. Like it's connected to your previous question, what is already in place?

Interviewee

What's the alert system to say this data? Has no mistake and we can use it to do a machine learning. Uh, we are still, uh, sometimes see science, um, indicating maybe there is a problem, but we don't have evidence. Um, like. Mm. For example, we, we received the data of people wearing different size of clothes and maybe starting, uh, some, starting some point in time that we observe this.

Interviewee
Uh, most people wear much smaller size, so it could be we are missing the larger. So there's data missing problem or, uh, there's some mistake, wrong labeling of something. Uh, so, so someone in in our client's ends label the data, the, the science wrongly. It could be the case, but there is so, so this kind of things, we just don't have any evidence saying which way it is.

Interviewee
and, um, but there, there's always a suspicion if you see some, something changes too fast because it's, it just feels unnatural. If, if you, um, if the time range is so short, uh, you, you see sudden data change. Then it leads to us wondering, uh, whether is, is, uh, there is any mistake in the pipeline and we just don't know.

Interviewee
And it's hard to ask. Um, what the only thing you can do is just to recheck, which I, if you can trace back the data pipeline from the entry point of our machine learning tracing back. Up to the point you can, you can reach, um, like from, from the, from the, uh, time the client sends the data to our team, our team processes the data.

Interviewee
So, so that's the point. When our company controls, we, we, we, uh, we get the data in front, the client and start At this point we have the pipeline, it, it, the pipeline. Is, uh, held by our own people. So this part we can control, we can, we can check each step to see this big change in time happens in which step, if we find them, then we can say, uh, we can look deeper.

Interviewee

Just tell the responsible person, look deeper, why there is such a big. In this step of the data processing, but, uh, often it is beyond our reach. Uh, like this change in time. This weird behavior is before, uh, is is already in the first, let's say first file the client sends us. So it's not after our company's processing.

Interviewee
It's already already there than we, it's harder to ask the client. To recheck because they, maybe they are not as technical or they just believe that's their real, um, record of data. So their part is, um, just beyond our reach. Yes. I see.

Interviewer 1
And is it difficult sometimes to, so you mentioned earlier on that in the data processing pipeline, sometimes you have.

Interviewer 1
Inspect each time step or each step to see if there's an issue? Yes. Is it sometimes something that is different, difficult to do, to like debug the data

Interviewee
pipeline? Uh, it is, is, um, difficult in the way that, um, many steps are not static steps. It's like data streaming. So, so streaming is, uh, at each point, each step.

Interviewee
You don't see the whole data set at that point. So it's harder to do statistics. Uh, for example, the average, uh, let's say we collect the, collect our end users height in centimeters, coming to our, our data, just, just one example. So one day all the centimeter values are in average, lower by 10 centimeters.

Interviewee
How can it be our customers? Suddenly, all of them becomes shorter by 10 centimeters. Is it possible then in the, the pipeline we can see, uh, we can try, try to see in which step in the pipeline this 10 centimeter drop happens. But if it is a streaming, Then at each step you don't have a complete data file having all the people that day.

Interviewee

Um, so there is no, no concrete file to look at because the streaming is like constantly flowing in time. So, so that part is, uh, yes, I will say there are some difficulties like this. Okay. I see.

Interviewer 1

But, uh, could you use historical data to perform the statistical checks,

Interviewee

uh, to use historical data? Um, yes. We, we can use historical data.

Interviewee

Uh, that historical data has to be saved somewhere first. So if it is streaming, then of course we, we save. Safe data in the table, maybe before, before some streaming and after some streaming, not, not in the middle steps. So if you want to inspect the middle steps, you have to take another, uh, another effort to say, okay, at this step we want to have a stop point to send all the data from the streaming additionally to another table.

Interviewee

And we stand there. Uh, after, um, after this flow is done, then we take a look at the table. So, so that's, I I'm not sure this is what you mean by historical. Yes,

Interviewer 1

exactly. It is really interesting. This is exact exactly what I wanted to

Interviewee

understand. Thank you. Um,

Interviewer 1

is there any other data quality issue we missed that you

Interviewee

consider relevant?

Interviewee

Um,

Interviewee

Uh, I think they are, uh, what I mentioned is just the, the main concern. Uh, they are maybe really small things, but I, yeah, I cannot pinpoint it. That's smaller things. That's perfect. Thank you.

Interviewer 1

How do you, how do you evaluate the quality of your models? As a reminder, quality is not only defined by ML performance, but also by other aspects such as explainability, robustness, scalability, fairness.

Interviewer 1

There is many others.

Interviewee

Mm-hmm. , um, uh, of course there are, um, tests right after the training, so that's very typical. Everybody does that. Um, and, and that deck can come up with, uh, come, uh, down to one number. Just say, let's say error rate of the training. So that's definitely, uh, something everyone knows. Uh, other quality kind of, uh, why you describe model is good.

Interviewee

Um, there is the training, uh, cost issue as well. Uh, like we are using, um, this tool TensorFlow as a mentor, uh, to trend our, our model. And, and finally we, we de deployed the model also in TensorFlow ecosystem. So it is TF serving as the final, uh, service of the model. Then, uh, so, so with the whole training pipeline, including.

Interviewee

The, the pre-processing and the training itself, uh, there's a cost we want to reduce because our, uh, Clients or, or we can say business partners. Some of them are much bigger than others. And for the really big sh uh, shops, they have a lot of data. So if we do training, regular training, let's say just weekly training each week, we collect a lot of data in, in training.

Interviewee

And, uh, there is a, there is a balance between taking. The extreme side is taking all the data that you can get and trend a model. Uh, their hope is with all the data, you, you have a really deep, really more precise model. And on the other extreme is you take only the necessary data. So the cost is much lower.

Interviewee

The resources you, you spend, um, training time is to, training time is also much slower. So there is a balance and we want. Uh, we want to get to the point that we think the data amount is enough, enough in a way that adding more data doesn't affect the behavior of the model too much. So our recommendation system is you giving me gi, give me a product.

Interviewee

That you want to buy and you enter some of your, uh, dimensions, like your height and weight and age and several things, and we give you back a size. So say, uh, which size you should buy. So the size is a concrete number. And if, if, uh, after collecting some amount of data, the size in average doesn't change much.

Interviewee

Like if you should buy l. , um, or you should buy m um, if I collect 10 times of da of the data, generally the model still tells the same person, you should buy M or you should buy L. Uh, if it doesn't change anymore, then we, uh, well probably, uh, we don't want, don't want to. Too much data, just adding the cost and the training will, will go on for, for, for a long time.

Interviewee

So, so this, uh, judgment, I, I think, um, that's, that's one quality. So the balance between the cost, uh, and training time and the performance of the model or the behavior of the model. Okay, I see that that's one, one. Okay.

Interviewer 1

And how are you able to add that balance between the size of your data set and the power of your model?

Interviewer 1

How, how, how do you choose how much data is enough and which

Interviewee
model is enough? Well, uh, um, I, I think right now it's not ideal. It's just someone, uh, not in our team. Maybe more like, uh, finance team or someone just come and say, Uh, your, your cost is, uh mm, disproportionate proportionately high or something like that.

Interviewee
Or, uh, we are, uh, seeing that your, your cost is growing. Then can you take a look? If there is a, is a possibility that you can control this cost within certain range or something. It's more like external pressure. So for us, there is a. It's not something we, in, in my own opinion, that's something important.

Interviewee
Control, controlling cost is something, uh, you should put a part of the quality of the model, but in reality in our team, uh, usually it's by external pressure because we are so busy with other things, like in the model, the, the, the performance, the behavior, the model, and so on, and. So, so that's the, if you ask me this question directly, how we do this currently, our team is still more passive, uh, passively taking, uh, someone complain about cost.

Interviewee
Then we, we worry about cost. Then we see if we reduce by cost by half. Then those people will be happy. Then we can answer this question saying, okay, we cut the cost by by half. And then the, the, if the behavior is acceptable, uh, we can definitely check the behavior. Just run, run the amount of data, uh, right now and have the model save somewhere, model version eight, and cut the data to half.

Interviewee
And trend model, save the model, call it model version B. And do a comparison just by sending a whole bunch of random people, a random people like in, in, in our record, not not real people, just in our record. We have a lot of collection and we just send those collection to the models and see the outcome, the, the outcome of the, of the models.

Interviewee

We make a list like percentage difference, A good percentage difference is 1% difference, and you save the cost by, by. And those people are happy, they don't, won't bother us anymore. So, so that's, that's kind of our approach. I, I think it's, it's really hard to say what's the best approach, but we are like this, we are being pushed this way.

Interviewee

So that's, that's just a reality.

Interviewer 1

Yeah. That's a reasonable strategy. There is. Thank you. Um, , have you, have you ever used existing benchmark models to evaluate the quality of

Interviewee

your model? Um, uh, actually in my time here, no. I think before me, the company had some like this for our, uh, machine learning model is not so, So straightforward.

Interviewee

It's not, not typical supervised learning. Uh, there is no, we, we are re we are making recommendation and there is no correct answer to the recommendation. So, mm.

Interviewee

What, what we, uh, did before was, Uh, we, we have some comparison of simpler models, um, but we haven't used, um, uh, why you say the, the, the ready made model. Uh, we have a really simple model by just giving, um, uh, like in most popular answer. Uh, and we, we use the, the new model compared with. So-called most popular answer model.

Interviewee

And that there's of course, some, uh, improvement in, in percentage, um, something like this. But we, we didn't, well, in, at least in my time here, we didn't use what you said, the, the ready made. Um, model. Okay. Thank

Interviewer 1

you. Uh, have you ever encountered any other quality issue, uh, during the evaluation of our machine learning software system?

Interviewee

Um, well, the main issue is the validation of the model is really, uh, really hard to define. Um, Uh, as I say, we are making a recommendation system and we, we see the, um, so the evidence of what the best recommendation is is really hard to, to say. Uh, we have the AV test system. Um, So we deploy one model in for, in certain funnel.

Interviewee

So some people are using the, the newer model, some people are using the older model and the way we compare them, this is kind of validation process. It's, it's online validation, it's just AP test, which one will lead to lower return. So when people are happy with their purchase, they return less. So that's kind of a belief, just a belief there.

Interviewee

There is no strong definition of what's the, the best recommendation, right? The only thing we can do is to monitor how much people return. So that's kind of a proxy of the best recommendation, but no one has a, in their dictionary say the best means. Yeah, so, so now we, we use this proxy and this proxy is only, uh, is only checked on, on, uh, online, which means after we already deployed the model.

Interviewee

So before we deploy ma, deploy the model, if we want to, just to save, save, uh, just check. Uh, ju just to double check, this is a model ready. For deployment, then we have less tools because you cannot do AB test. AB test is always already facing the U end user. So internally we have even less tools to validate the model.

Interviewee

The way we validate is to, of course, we, we can compare with the previous version of model to see how much. Change it is. That's one way at least. It's not too, uh, radical. So if it is radical, then you have the high risk. Maybe it's radically better, maybe it's worse. But that's always a concern if it's changed too much from the previous version of a model.

Interviewee
So now, so after that, we have some, uh, sanity check. Uh, like at least it's making something reasonable. Like you give. Our recommendation is spreading for different body sizes or high end weight combinations. At least you spread the, the, the recommendation to different sizes and then the, the, uh, Simple, reasonable behavior is enforced.

Interviewee
Like if you are, if you're same heights, but your waste rolls, you gain weight. It means, it means gradually you will get bigger clothes. So that's some really, really simple sanity check. And it is, uh, it is of course automatic. There are too many, uh, products to check. So we have some automatic check and this, and also the comparison with the previous model.

Interviewee
And, uh, like the, uh, distribution, like a r it's, it is a general, uh, distribution across possible, possible recommendations. So this kind of, kind of, uh, really simple checks, it's really hard to say, uh, the, the percentage of correct recommendation because recommendation has no correctness. Yeah. And the o, the only thing we can possibly reach closer to correctness is only after we deploy online and monitor the AP test.

Interviewee
If no one after, if before we deploy the model before the new model, a lot of people return to the shop and after we deploy a new model, almost no one returned to the. Then, then we can call it a perfect model. But that's only after you, you do it online. So once it's online, it is already in the risk regime.

Interviewee
Right. So that's hard to check before you deploy. Yeah, that's a

Interviewer 1

really good point. Thank you. Um, what are some of the challenges you have encountered during deploying a machine learning software

Interviewee

system? Um, the difficulty in, in the serving. I think it's traffic, it's more engineering part. So you, you we, we build a model and model will be sent to different servers or, or instances in the cloud.

Interviewee

So the cloud will have several, uh, um, server like machines serving the traffic. And we use, uh, Google Cloud and there are some automation. I think that's, That's very engineering. I, I'm, I'm not, there are teams just handling that, uh, just from, from the information I get from them. There's a lot of automation.

Interviewee

They can scale up, they can still scale down based on the traffic they receive. So they, if they need more servers, they just open new server and, uh, replicate our model to a new server. So to handle that traffic and traffic, of course, when it comes in, it was split to different servers to be served. Uh, and there are still some, uh, Scaling issues because we, our models are, uh, built for each different, each shop.

Interviewee

So there are some really big shops. There are some really small shops, and we deploy for them and they are, uh, scaling systems and if, well, so, so it is, uh, how you, how you, um, deploy the bigger shop. To a enough scalable system, but then separates the smaller shops to a less scalable system is again, a little bit coupled with the cost, uh, because the, of course you can, if you say you don't have any budget limit.

Interviewee

You can just deploy each shop to the best scalable system, then it's, it's solved, but it's not So our deployment, uh, well, I, I, I think the, the biggest concern, or the thing that's, that takes the most time from our engineers is the traffic. To handle the traffic and to, uh, deploy the models according to the size of shop.

Interviewee

And you kind of have to predict beforehand, uh, because you, what if you already have to assign sys assign models to different kinds of systems? The time of assignment is before you really receive the traffic, right? So, so that you have to have a judgment. Before you, you do this, uh, and judge based on a little bit of guess.

Interviewee
I think guess well it's just you see different shops, how much their purchase is and so on, and you kind of, uh, guess the, the traffic, how much the traffic will be.

Interviewer 2
Thank you for your answers. I have a follow up question. Are you relying on your cloud service provider for these. For example, using some autoscale techniques or, um, autoscale, uh, um, facility from your provider, or you are trying to do it on your side? Uh, we, well, uh, we have another team mm-hmm. focusing on this, and we use the cloud system.

Interviewee
So, so the, the deployment itself is, Completely in Google Cloud service. Mm-hmm. But e, even in Google Cloud service, there are some experts knowing how to best handle it. So in our company, there is a small team who will handle this, uh, deployment, like choosing, choosing, uh, what service to use in a clouds and choose which type of machine and so on.

Interviewee
I see. Thank you. So we don't have an in-house, uh, server or something like this. We just use the cloud service. No, that was my

Interviewer 1
question. Thank you. Okay. And just to clarify the issue, uh, the issue is you have small clients, you have big clients, you do have autoscaling systems. But what you want is to prevent that these autos scaling systems scale too much and it costs a lot.

Interviewer 1
Am I

Interviewee

understanding Correct. Uh, we, we have, uh, auto scaling systems, but we, because each shop, uh, for some bigger, some small, each shop has its own model. So when you deploy a new model for a specific shop, you already want to know which level of scaling you want for this model. So some model will be on the less scalable system, some will be.

Interviewee

More scalable system. Okay, I

Interviewer 1

see. So you, you prevent scaling too much for shop that cannot pay too much for machine learning?

Interviewee

Uh, yes.

Interviewer 1

Okay. And do you sometimes change a model you use? If it is a small shop, maybe you will use a simpler model that consumes less resources such as a tree or

Interviewee

right now the type of model.

Interviewee

Is the theme. We, we have, uh, different generations, generations of models, but the whole team is working on, uh, let's say the next generation of model which has better features or the interior calculation is better. Usually we, once we have this development and we are ready, we change all the shops to.

Interviewee

Newer generations of newer generation of model. Uh, so this is the, the development is for all the shops, uh, equally. Yeah. Okay. Is it, is it your question? Yeah, yeah, it's fine. And, um,

Interviewer 1

yeah, that's perfect. That's perfect.

Interviewee

Thank you. Um,

Interviewer 1

have you add any other issue regarding the maintenance of your model or,

Interviewee

There are a lot of cleaning up of the model.

Interviewee

Um, right. As I, uh, told you just a few minutes ago, our, the deployment and so on, the intensive flow in TF serving, this is the newer assistant, which is well packaged, like each model for each shop. Is, uh, packed into a, a file, a model file, and it's served independently or individually by TF serving, but before our, a lot of our secondary logic or, or even right now, our secondary logic, like some, uh, small shifts for some kind of garment or something like that, this kind of, uh, logic, it also affects the recommendation.

Interviewee

But it exists in the backend system. It's not packaged in the model, uh, which means the backend system is calling, is calling the model, but after calling the model and get the result, the backend will still shifts the recommendation a little bit. Uh, this part is, has a maintenance issue because it's not, this is business logic.

Interviewee

Should be ideally inside a model, right? Because it still affects your final results. So if you don't pack it into the model, it's harder to, to maintain. It's just a one big repository. And sometimes the repository, uh, different people are working on the standard repository and everyone wants to merge it to, to the, uh, master branch.

Interviewee

Um, so there's a, like a. Not, not the best situation. Uh, we are trying to move all the business logic or recommendation anything about the algorithm into the model so the backend doesn't have to do anything. So they are just a repository, which, which is always well maintained. It's just like passing, passing the front end of the widget.

Interviewee
To the model, uh, to the correct model. Like just checking the shop, which shop it is a call the model and get, get, result out. Yeah. So that this is the current effort. Um, yeah. So, so that's one maintenance issue, but I'm not sure is, uh, why you defined the maintenance of the model itself because I, I can consider this maintenance, maintenance of the algorithm.

Interviewee
It's not formally defined in the model, but we want to put it into the model.

Interviewer 1
Yes. Well, it, it's, it's a great, thank you. It's really interesting. Um,

Interviewee
so yeah. Yeah.

Interviewer 1
Perfect. Uh, w I was talking about the model because sometime some people mentioned, uh, data drift or things like this. It's more the data, but it's some somewhat related to the model, so.

Interviewer 1
Mm-hmm. Yeah. Uh, but thanks. Been interesting. Uh, we'll move on to the last section of the interview. Um, so did you ever add any issue regarding any of the following? Quality aspect? So, fairness, robustness, explainability, scalability we already touched upon. And privacy.

Interviewee

Um, for privacy. Um, We had a lot of, recently, we had a lot of discussions and we are just trying to understand whether the privacy more, more risk, uh, more um, strict privacy rules will affect our data.

Interviewee

And currently we are not super concerned, but as always, we are in close contact with the people handling the privacy and our training. Um, depends a bit on recognizing the, the same user who is putting the input to who is buying what product. So this link has to be there. And, uh, I think this, if you don't handle it well, it could be a privacy issue If someone can recognize the person.

Interviewee

The shopper, but we try, uh, according to my understanding, it probably won't affect us, uh, because our, uh, team, the privacy team can handle it in a way that the, the ID of the person, uh, whatever Id it is from a shop or from, from whoever, is not, um, exposing anything. Um, It's not exposing any information except, except we can link this person to who is doing the purchase.

Interviewee

So that's, that's already that Our purpose is just to make a link, you know, uh, we, we do machine learning model and the. You can, you can think it as a big, uh, data table. And each row is an event and each event has a link of the whole user journey, whole user journey up until is, is buying the product and it, it could, may return the product and so on.

Interviewee

So as long as the link exists, we don't have a problem. But if, if someone without understanding, well, she'll say this. There is a privacy issue, so we cannot identify the user does. There is no link. Uh, but our, I think our team, um, they are doing a good job. So the link is preserved, but the person is not identified.

Interviewee

That's what my understanding. So, so far there is no big issue. And for, uh, fairness, you're saying the, the, whether the model. Fear for different users. This is one thing I, I think we are still trying to find a better way. Uh, because our, um, our recommendation, if you don't, don't do anything, just let 'em all run without, uh, good pre-processing, then it will go toward the most popular option.

Interviewee

And in our. In some cases the most popular option for each, uh, input is fine. Uh, for, for our, uh, training because we train different sizes, uh, different recommendation results together. If you, uh, train them together and try train all different people with different sizes together, then the most popular, uh, Popular data points in the, in the whole dataset will be the people with the average high and average weight.

Interviewee

So it will look bad if you let someone who is much taller or much heavier to try to widget it. If that becomes obvious, then those people will not get the best recommendation. But the really average high and weight will get the best recommendation. It will look quite bad because they. In this dataset consider minority.

Interviewee

So it's becomes a almost like discrimination, although it's not our intention, it's just, uh, imbalance in the model. So we are, we have a little bit of balance by sampling different, uh, different clothes sizes. So if you sample, uh, from like from Xs, s m and L XL and so on, you have some kind of balance. It will make the situation a bit better, but we are still thinking, uh, maybe there are better ways.

Interviewee

Okay. Perfect.

Interviewer 1

Yeah. Thanks a lot, uh, for that complete answer. And, uh, the last question is, in your opinion, what is the most pressing quality issue researchers should try

Interviewee

to solve? , uh, in my experience, I think, um, the, the, the biggest quality issue is at the end of building a model. Um, there is no good enough validation process.

Interviewee

Uh, there is no good enough indication to say this model. Well perform as expected in the real world. Um, so there, this validation check is before your deployed model. So of course, before your deploy model, you cannot really know how it will perform, but at least you can guess. And our guessing it's not good enough.

Interviewee
It's not always not good enough. So there are cases in our, in our experience, we, we, the validation is very successful. The model looks fine, looks justly, normal, is and is not so. It's a bit different from the previous version, but our initial guess is, is, uh, of course it's much better. It is our gen, newer generation with better data quality and so on.

Interviewee
So the difference in w in our eyes was it's much better. But after deployment, the performance actually worse. So our validation system didn't catch it. Yeah, so, so the validation, offline validation, and the real online. Proof in ab test. The the distance is, is quite far. I, I wonder what other industries, they have similar issues.

Interviewee
In our case, it's fashion, like predicting what people wear in size. It's, it's of course important, but it's not life threatening. Right? It's just a size of clothes. People can see a word, but in other industries, I, I don't know what other. If someone really wants to know before deployment, the model should be good enough.

Interviewee
Uh, it's not dangerous, it's not making ridiculous recommendation. Maybe it's even like some, someone is dependent on it. Um, then they should have a really careful check. Yeah. But my, my connection with other industry so far is limited. Yeah.

Interviewer 1
Good. It's really interesting. Thank you for your input. And do you have any other comment about the quality of machine learning software system?

Interviewee

Um, before coming into this industry, my focus was always the, the learning itself, the machine learning model. Uh, After getting the data, you just trust the data and you say you check the machine learning model based on what you are, given, what data you're given. So you make a test set and you test the machine learning itself against the test set, and then you come up with the kpi.

Interviewee
So I think this is in, in school, the teaching was, uh, refined. That's, that's the way it should be considering just the machine learning. But now in the real world, there are, there are so many other. Um, other things concerning your final product is good or not. It's not only the central part. So, like I said, the.

Interviewee
Uh, the, the data coming in, there can be a lot of problems. Then, uh, just dis just, uh, distorting your, your whole testing as well. And then your, your final, uh, outcome is hard to, uh, interpret or is, is harder to, to check against the reality, uh, until you, you really deployed it and so on. So, there are many other components behind.

Interviewee
Model, uh, before the model and after model. Uh, they are also important parts in the whole product, but it's not strictly defined inside the machine learning. So I think that the school teaching and, uh, of course it's very valid points, but the, it's not enough to make a whole product, um, good enough. All right.

Interviewee
Well,

Interviewer 1
thank you for your, uh, for your information. Is this, your time has been very valuable to us and it'll be really useful for our study.

Interviewee

Uh, I also really enjoyed the, the, your questions. I think it is, uh, really, uh, inspiring. That's really good questions. Thank you. Thank you so much. Thank you for your time, your, your assets were wonderful and we'll use those information.

Interviewee

Thank you so much. Yeah, thank you for the interview. All right. Have a good day. I wish you, uh, a successful research outcome. Thank you so much. Hopefully. Thank you. Thank you. Yeah, we need it. Bye. Bye.
