Interview 19 - PO

Interviewee

Right pronunciation of your name. Interviewee. I'm sorry, what is the right pronunciation of your name?

Interviewee. Yes, yes. Yeah, . You pronounce it correctly? Yeah. Okay, so go ahead you. Yeah.

Interviewer 1

So like, uh, Interviewer 2 said what we want to do is to develop a catalog of quality suits in machine learning software system. I will break down that, that sentence.

Interviewer 1

So what is the quality issue, or more specifically, what is quality? Given, given two systems that accomplish the same thing. If you're able to tell that one is better than another, then what you're usually referring to is quality. So the one that is worse probably has some quality issues and a machine learning support system, it's a ease of our system with a machine learning component in it.

Interviewer 1

So for example, our recommend recommendation system. Anything else? Yeah. Um, we'll ask you about 20 question more, or. And, um, yes, if you have any question, feel free to, to interrupt, to ask your questions.

Interviewee

Yeah, sure. Perfect. Thank you.

Interviewer 1

Uh, so the first thing we would like is to have some more information on you, some background information.

Interviewer 1

So what is your current position? How many, how many years of experience you have in ML or in general?

Uh,

Interviewee

actually, Okay. Um, yeah, I have, um, maybe one year of experience in software engineering and, uh, like three years or a little bit more in, in the data science or in the data field. Um, I've worked in, uh, three companies, uh, as a data scientist and yeah, and that's relevant because.

Interviewee

Um, yeah, the, the domains are, are different. Um, yeah. And especially in the, in my first company, I, uh, I was lucky because it's an AI company. Um, or yeah, it's a company that serves other companies, uh, to, uh, optimize maybe the, uh, processes and using machine learning or using, uh, data science in general. So yeah, I was exposed to different, um, projects, um, in different domains.

Interviewee

Now I'm working as a data scientist in, um, in, um, in a supermarket chain and the e-commerce department. And yeah, um, I'm helping in, um, in the, in, in, yeah, in the advanced analytics in general for the, uh, operations and. For the upper management to, um, to just take decisions and yeah, exporting reports and, um, constructing data pipelines.

Interviewee

Um, yeah, I'm responsible for the, maybe the infrastructure because it's a relatively new team. Uh, the data team and also the e-commerce. Uh, maybe the, the, the, the, the supermarket chain itself is, is very old, but yeah. Um, when, um, when it comes to, to the e-commerce department, it's relatively new. Um, so yeah, we are, um, two years old now, maybe the e-commerce department and yeah, we are, we are halfway there maybe to.

Interviewee

Like, deploy all the solutions that you want to, to deploy or to be like satisfied, um, in terms of being data, data driven, uh, or take the, all the decisions using the data. Yeah. Okay. Thank you. Thanks

Interviewer 1

a lot. Um, so the first question will be, um, what are the main quality issues you have experienced in your data model or system?

Interviewer 1

So,

Interviewee

Okay. Um, yeah, I think that, yeah, if, if I am understanding correctly, maybe it's, it's a very general question, but, um, I think that the, the first problem that I face, uh, is the, maybe it's acquiring the data. Maybe it's as, uh, yeah.

Data-low-quantity

Interviewee

Yeah, acquiring the data itself. Uh, but yeah, maybe this is, uh, more of a, a business, um, problem than, than a technical one. Um, but yeah, when, when I, when I get the data, the first problem I, I I face is that, um, the diversity of the data, uh, formats, um, Yeah, sometimes I, I have, uh, connection string, or I have access to a database, and this is of course, uh, um, the best way to, to, to acquire the data, but sometimes it's on, uh, files, like Excel files.

Data-integration-issues

Interviewee

Sometimes I, sometimes it's APIs, so yeah, maybe this is the, the. The most pressing issue I am facing now because I'm dealing with multiple data sources and yeah, I have to, uh, like join all these data sources in a way, and I have to, um, make sure that um, every single step is done, uh, properly and there are.

Data-integration-issues

Interviewee

Uh, for, for example, the duplicates, the duplication, or that the, um, the joins between the data is, uh, is correctly done. Um, yeah. One of the most issues also is the data description because, uh, or the column descriptions. Um, because yeah, as far as I, um, yeah. Until, until now I have. Um, encountered the, a situation where, um, where the data description is, um, supplied along with the data itself.

Data-integration-issues   Data-metadata-issue

Interviewee

So yeah, that's, that's a lot of time consuming. It's all, it's a, it's a time consuming process to go back and forth and ask about the. Read the correct, uh, description of the, the columns, and, uh, sometimes, um, there is no clear answer. And sometimes you have to figure, uh, it yourself, uh, even if it's, uh, a database, uh, most of the times there, there, there is no clear, uh, schema for, for the database if the database is structured.

Data-metadata-issue

Interviewee

Um, And yeah, sometimes you have to dig deep into the database. Uh, it's, it's really hard problem, especially when the database is, is very, very big. And, um, yeah, sometimes even this database is built on, I don't know how to call it, but built on, uh, skeleton database maybe, and it has modifications. So there are tables which are, uh, active and tables which are inactive.

Data-metadata-issue

Interviewee

Um, so you don't know which one is, uh, yeah. Is the, is the most relevant to you, or which one is working and which one is, is not? So yeah. From, yeah, from this first step, uh, just acquiring the data or, um, uh, Yeah, reaching the point of reaching the point that you have, uh, structured data connected from multiple data sources.

Data-metadata-issue

Interviewee

It's, it's, it takes a lot of time and yeah, a lot of investigations and yeah, maybe interviews as well with the, uh, the people who are responsible for, for, for these data. Okay. I see.

Interviewer 1

So your main issues with data is first, the, the management of the data is often poor f from the people who owns the data.

Interviewer 1

And the second thing is that is difficult is that are the integration. So trying to merge all data service together. Yes. Uh, do you have some tools to help you? How do you address this problem? Is it like manually you, you contact people and you try to. Tables by your best guess, or you have some tools to, to guide you in this

Data-metadata-issue-fix

Interviewee

process?

Data-metadata-issue-fix

Interviewee

Uh, yeah, of course. I, after, um, I, uh, consult, uh, other people or are responsible for, for these data sets, um, or sometimes I, uh, I do it by, by guessing or sometimes. Yeah, I'm pretty sure that this is the correct way, but I have to of course, make sure a hundred percent, um, with, um, with the supplier of, of this data set.

Data-metadata-issue-fix

Interviewee

Um, but yeah, the tools, uh, you mean, um, like, uh, Python or are, or, uh,

Interviewer 1

I meant any library or framework you might use, but if you don't have it, it's uh, It's an exploratory

Interviewee

question, so, yeah. Yeah. I, I, I, I think for, for the data, uh, uh, for the database connections and for the data integration in general, I prefer Python, which is, it's, it's, uh, very mature in this, uh, area.

Interviewee

But for the data wrangling, the data processing, data cleaning, I prefer, uh, working with,

Interviewee

Okay, perfect. Thank you.

Interviewer 1

Um, so you mentioned that, uh, oftentimes you go fed data from, from external provider, external databases. Uh, did you ever fetch data in other ways, such as, uh, did you ever, uh, Use the services of someone who, who create data for you. So, so data collectors or did you ever use data that was generated by another system?

Interviewee

Um,

Interviewee

I'm not really sure. Um, okay. Yeah, that's fine. Thank you. Um, ,

Interviewer 1

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

Interviewee

to improve it? Yes. Um, yeah. For example, um, we had this, uh, data with, uh, phone numbers. It was a huge, uh, data set and yeah, we had to, um, yeah, make some, um, quality, um, No, uh, pipeline for the, the, the phone number.

Data-unnormalized-annotations   MLP-debuggability-issues

Interviewee

For each phone number to, to, to be. Um, um, yeah, it's a sanity check. Maybe a sanity check for, uh, phone numbers. Uh, for example, the phone number must, um, starts with, uh, zero, something like this. Uh, there must not be, um, Uh, con five consecutive ones or six consecutive ones in a phone number. Uh, otherwise we, we consider this phone number as, uh, a fake one.

Data-unnormalized-annotations   MLP-debuggability-issues

Interviewee

Uh, so yeah, something like this, uh, this is, yeah, in the top of my mind, this is the, um, yeah, and, and also, uh, yeah, I faced this, uh, issue a lot with date. Because sometimes, uh, the, the column is, is supposed to be a daytime, but it's not. It's only a date with, um, with empty time component or like 0, 0, 0, 0. And, uh, sometimes the date itself is in un it's, it's not, um, It's like, uh, an ambiguous fo format.

Data-unnormalized-annotations   MLP-debuggability-issues

Interviewee

Um, it's, it's not really clear whether it's, uh, the year, month, day or, um, or different formats. And yeah, I face this show when within the same, uh, Excel sheet there were different. Date formats, uh, within the same, uh, file. And it, uh, I didn't see it coming at all. So it was, yeah, this, that was one of the, the hardest issues.

Data-unnormalized-annotations   MLP-debuggability-issues

Interviewee

I, because yeah, it's, it's easy to, to solve a problem when you, uh, expected or, and yeah, you discover after so many steps. Uh, so many things gone wrong because, uh, one of the first steps were there were a problem was, uh, with, with the dates and maybe you will suffer a data loss because this, um, this data source is not, um, after processing, this is the source.

Data-unnormalized-annotations   MLP-debuggability-issues

Interviewee

Is, is, uh, It's not available anymore, so you have to open like a request again. Um, so yeah, maybe, uh, sometimes yeah, maybe sometimes you have to like, after each processing step, like save the, the data and Yeah. But it's, it's hard. It's hard to maintain. Yeah.

Data-unnormalized-annotations   MLP-debuggability-issues

Interviewer 1

Yeah. All right. Thank you. And, um, so, so you mentioned the issue with, uh, in a same column you had different date format.

Interviewer 1

Yes. How do you try, did, did you put any mechanism to detect these kind of issues earlier in the pipeline or now you're just more

Interviewee

careful? Yeah. Yes. Uh, I, I don't remember exactly, but I, I, I remember. For, um, yeah, I, I can detect the, the, the, for the, the formats, which, um, which for example, uh, year, months, day, and the formats which are day, months, uh, year, um, India, if I find this format, I will like convert it to, to the other one.

Interviewee

Um, so yes, that, that was. Step in the, the pre-processing maybe, um, um, the, the file, which is responsible for the pre pre-processing of the, of the data itself. So yes, it, it, it became a part of, of the pipeline. Okay. Perfect. Thank

Interviewer 1
you. Um, is there any other data quality issue we missed that you consider relevant?

Interviewee
Uh, maybe the missing data and, uh, especially when it's a date, uh, because yeah, you can imput if, if there is a machine learning, if you are going to feed this data into machine learning model or something, you can imput uh, missing data from features like, Demographic features, for example. But, uh, I don't think that, uh, including a date feature is, is, is relevant or is correct.

Interviewee
And, um, yeah, if, if this data will not be fed into a machine learning model and I want to subs the data according to the. There is a huge problem because there are missing values in the data itself. So these, um, observations will not, uh, be subsided from the data. So yeah, missing dates are, um, yeah, it's, it's the waste.

Data-missing-values

Interviewee
Okay, perfect. Thank you.

Interviewer 1
Um, how do you evaluate the quality of, of your models? And as a reminder, quality is not only ML performance, but you can also define it as scalability, uh, robustness, explainability,

Interviewee
uh, yeah. Um,

Interviewee
okay. You mean the machine learning models? Yeah. Your model.

Interviewee

Uh, yeah, it depends, uh, of course, because, um, really most of the times, uh, we don't care about the acceptability so much. Uh, maybe recently we, we began to like, um, um, put, put effort into, uh, , um, explain the, the model or explain the, the, uh, the predictions. Um, but yeah, the, the, the most important thing is just see the, the, the regular metrics.

Interviewee

The, yeah, the, the accuracy and, uh, of course the other metrics, the, the precision recall, the sensitivity, uh, specificity, the, the scores, the, the curves. Um, Yeah. And sometimes for, for the most important, um, maybe categories, sometimes we don't care about some cate of course, we care about all, all the categories.

Interviewee

Let, let me put you a little bit into context because we are, for example, uh, building a model for, um, for a demand for testing. It's a demand for testing model. Um, for predict prediction of the demand in, in the supermarket of each item. So we have a lot of items, um, adding all items to the model will of course introduce, um, noise.

Interviewee

And we have, we have to, um, use a deep learning model. Uh, yeah, for, for deep learning models, it's, yeah, they are bad and it's explainability. Um, but yeah, luckily there are, uh, other models which can explain the, the deep learning models. But yeah, when, when the problem itself is, is simple and does not require, um, Taking this approach of deep learning, I, I, I for sure prefer the most simple, um, machine learning models and the most incredible interpretable ones.

Interviewee

Uh, yeah. Maybe my favorite is the decision tree is because yeah, it, it's very powerful and it can be, um, It has high interpretability. Um, okay. Thank

you.

Interviewer 1

And why do you prefer explainable model or why at your company you're using? You said in the past you use not explainable model, but now

Interviewee
Yeah. Maybe a little bit more.

Interviewee
Yeah. We, we, yeah. In the past we did not care so much about explainability. Um, and, and that was not actually the, That was wrong. We should always, uh, care about explainability. But for, for example, uh, a credit scoring system, uh, it's not really a credit scoring, but yeah, it's something like that. But, um, so yeah, I think that the, it's, it's very important to explain the model of credit scoring system.

Interviewee
Um, But yeah, we, we did not care so much about the, we, we build a tree model. It was a very complex tree model. Um, and yeah, the most important thing for us was the, uh, the ability of this model, the prediction ability of this model. Uh, not, not the interpretability or the acception. Okay. Thank you.

Interviewer 1
Um, have you ever assessed the quality of your model, uh, with the user of your system?

Interviewee
I'm sorry, with the what?

Interviewer 1
Did you ever try to assess, did you ever try to assess the quality of your model's prediction with the user's of your system?

Interviewee
Uh, yes, but that was not actually my, uh, , uh, my, my job there was, uh, a very, um, yeah, um, powerful and, um, yeah, maybe a team, a very small team, which, um, measured the impact of the, the solution as a whole to to, to the other, to the client and.

Interviewee

Yeah, I did. I didn't, uh, have the the time to, to know how they do it exactly. But, but yeah, I think it's very important because the, um, because yeah, the, the prediction, uh, accuracy or the, the, these technical metrics, uh, are not the only. Metric to success or to, um, to, to a successful solution? Um, for, for the client.

Interviewee

Cause yeah, the client has other questions. The client, um, yeah, I don't know, but yeah. Measuring this, uh, impact is, Yeah, it was, maybe it was money with revenue. Was the solution able to, um, to get more revenue for the client? Or, or not? Yeah. I think this is one of the, the attributes that they, um, used to, to measure impact.

Interviewer 1

Okay. And what was the feedback from the user? What, what were the quality issues they told you they

Interviewee

saw? Uh,

Interviewee

I don't remember. Yeah, for sure. Uh, I don't have, yeah, a solid memory for that. Yeah, no worries. All right.

Interviewer 1

Uh, have you, have you encountered any other quality issue, uh, during the evaluation of your models?

Interviewee

Yes. Actually there, there, there was a, a recent one, um, I don't know if it's a quality issue or not, but yeah. This deep learning model. After the training, after training, the, the deep learning model using the, the validation set or, yeah, we had the training validation, um, uh, split. , uh, and it's a time series model.

Interviewee

So the validation or the test set is always the, for example, the last week or the last two weeks from, from, from the whole data set. So we are able to, uh, con, continuously measure the, the performance of, of the model, um, until we reach the point that we want. Predict the, uh, the new, uh, observation for, for example, the last week.

Interviewee

And, uh, in this case, we have to use the whole data set, uh, until yesterday, for example. So in this case, we don't have a validation, um, set. Uh, so. We did not know, know what to do, uh, actually, because it's a, it's a deep learning model and, uh, it, it continuously changed the, the, um, the hyper parameters and the, the number of s and the training, training, uh, error also is everything is continuously changing.

Interviewee

Uh, but yeah, the decision was to rely on the, um, the number of efforts. Yeah. Not the training, uh, error. The, yeah, we, we tried to like, rely on the training error to, uh, if we reached training error of, um, of a similar one based on the historical experiments, uh, we will stop. Training, and we will consider that this model will be, uh, good for, uh, prediction or good for, uh, inference.

Interviewee

But we found that this is not a good idea. And yeah, we relied on the number of airbox instead. Yeah. Okay.

Interviewer 1

But you, you do not use a validation, sir. If I understand correctly,

Interviewee

we, we don't. We use, of course, a validation set in Val invalidating the, the model performance of course. But we, uh, reach the point that we want to product new, um, we want to produce productions for the new week, for example.

Interviewee

So we have to use the whole data set. We can't, uh, split on training and validation. We have to use the whole data set. And yeah, we face this problem that, um, we cannot assess the performance of, of the model, uh, anymore. So, um, yeah, we, we, we decided that we will use the number of airs that we have seen before from, uh, from, um, Uh, validation experiments that we, that we did.

Interviewee

So yeah, for example, if, if we are using like, um, based on the best experiments, we have used like 40 apox and then 50 and then um, 45. So we will take the average number of AP that we used. Uh, during the, the past experiments and we will use it on, uh, the whole training data set. Yes. Okay, perfect. Thank you.

Interviewer 1

Um, what are the challenges you have encountered when you deploy a machine learning software system?

Interviewee

Uh, yeah, actually this is also, uh, is not part of my. Yeah, it's not part of my, my, it's not one of my responsibilities. Um, but yeah, I was part of, uh, the process at, at some point and when I was working, uh, at my first company, and, uh, I just had to containerize my model.

MD-issues-with-dependencies

Interviewee

It was on, in, in r it was, uh, a model which, which was, uh, built on. , which I, I, I think was not a very good idea to, to build a model on our, so this introduced, uh, a problem to, to the DevOps team, uh, because. They did not know how to deal with this. They, they are not familiar with, uh, models coming from models from, uh, our, so yeah, there was suggestions from their side to containerize.

MD-issues-with-dependencies

Interviewee

My, um, My model. And, and then they took care of the, the, the, the whole deployment process. So yeah, I had this, uh, experience containerizing the, the model. Um, and I think they made another container and the containers, uh, I don't know how they, they did it, how, how, how they orchestrated the, the containers to get, Okay, perfect.

MD-issues-with-dependencies

Interviewee

Thank you.

Interviewer 1

Uh, did you ever have a model that performed well locally but poorly once

Interviewee

deployed?

Interviewee

Uh, I think, yeah, it always happens. Um, especially when the, uh, I don't know, when the problem is, is hard. It's very hard. Um, When the, the, the, the quality, uh, metrics itself are not robust every time with every, um, with every experiment or with every, um, I don't know. Uh,

Interviewee

Yeah, maybe with, with, with every experiment or with every new distribution from the data, maybe this problem happens a lot because the, the, the distribution of the data keeps changing. Um, and yeah. I think I, I have not invested the, the enough time to, to investigate this, this problem, but yeah, it, it happens a lot, of course.

Interviewee

Okay. Thank

Interviewer 1

you. Have you encountered the issue with the data during the maintenance, um, learning software

Interviewee

system?

Interviewee

No. I, I don't remember. Okay. Perfect. Thank you.

Interviewer 1

Um, did you ever add issues with, with one of the following aspect, so fairness, robustness, explainability, you already mentioned it. Mm-hmm. , scalability, privacy, security. Any one of them that it rings. If it rings a bell, you can tell us about it.

Interviewee

Maybe robust. Yeah. Uh,

Interviewee

yeah, I, I, I think that that would be the, the most problem that, that we face. Um, yeah. Because yeah, as I told you, uh, things, um, keep changing. And this is very bad for a machine learning model. Um, for example, Yeah, if there is an item that's gone, um, inactive. So the, the machine that I model will predict that, um, the demand, uh, is X and in reality there is no demand at all for this model.

Interviewee

Cause it's, it's, um, it's inactive, it's not active anymore because of seasonality or something. And the model cannot. Um, the model did not, uh, capture this, um, seasonality, uh, component from the, the <INAUDIBLE> data. Um, so yeah, maybe I think the, the robustness of the model is, yeah. Okay. Thank you.

Interviewer 1

And in your opinion, what is the most pressing quality issue?

Interviewer 1

Researchers try to.

Interviewee

I think now, um, the explainability of the, the deep learning models, um, yeah, maybe this is the most topic

last-question-answers

Interviewer 1

right now. Thank you. All right. Do you have any other comment about the quality of machine learning software system?

Interviewer 1

All right. Well, so that's all for, uh, today. So yeah, thanks for your time and thank you to attend to this, uh, interview. I'm sure it'll be super useful for our study. So thank

Interviewee

you. Thank you so much. Thank you. Thank

Interviewer 1

you for your time. Have a good bye. You too. All right. Have a good day.

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