Interview 31 - Amin

Interviewer 1

All right. Uh, to start off, would you give us some background information about yourself and like how much experience you have in master learning or anything else?

Interviewee

Mm-hmm. , so I'm a master's student at University X in, uh, Country X. I am getting a master's in Applied Mathematics. I have been working also as a data scientist for a company called Company X, uh, for the previous four years.

Interviewee

Uh, in that time I've experimented with a variety of different machine learning techniques. Uh, I've been doing r and d for them, and I work with unsupervised clustering a lot. My thesis here at University X is in categorical based data and unsupervised clustering with that.

Interviewer 1

Okay, super far. Thank you.

Interviewer 1

Um, So I, we'll start off with a general question. What are the main quality issues you have encountered with either data model or system so far?

Interviewee

The data that I use is often, um, Often corrupted or some dirty where, because I deal with a lot of personally entered, entered data, so people misspell things a lot. Um, and in order to deal with that is, is fairly time or computationally intensive. Other than that, the, when I use other. Software systems that have machine learning implemented.

Interviewee

Like if I were to use a tool to help me, um, online with, with a project, it's often fairly simple. Um, I've, I've worked with a company, uh, that is primarily a data warehouse and they claim lots of machine learning models when they're just random forced and. Petition. I, I do seek a little more nuance in, in, in the models, but, um, so the, the, the most frustrating issues that I've run into is, is uncleaned data where we have to clean it ourselves, which can be computationally expensive and simple, um, models for plug and play that.

Interviewee

We, we can't use in production because it's, doesn't give it a lot of nuance.

Interviewer 1

Okay. See, so you cannot use them in production because they're not good enough basically. Correct.

Interviewee

Yeah. We too, too simple. We find they, they capture different features that need to be outlined in. So I, instead of, um, seeing the relationship between two people, it will see the relationship between.

Interviewee

To features that they have in common where we want to identify the relationship between the people themselves.

Interviewer 1

Okay. See, and what, what prevents the company from adopting a more, uh, adopting more complex models?

Interviewee

It's much more difficult to, to have more complex models, especially because they are more. Um, specific.

Interviewee

The more complex you get, oftentimes the more specific you have to get, and they try to offer a generalized plug and play option for us. Oh, okay. I see.

Interviewer 1

Okay, Henderson, thanks. Um, so I'll ask you a couple of questions about data collection, and I will be, I will try to look into, uh, well data issue with some data collection process basically.

Interviewer 1

Uh, so you mentioned earlier that, uh, Use data that was manually entered by, uh, some people. Right. The, so can you tell us some of the issues, uh, you encountered up to

Interviewee

now? Yeah, right. The most common is misspellings. Um, when, so I, I work with data that's somewhat similar to LinkedIn. So you can imagine if someone in, that's the kind of data that we have.

Interviewee

So oftentimes misspell. Um, synonyms is another thing that, that we have to deal with where they, they say the same exact 10 people will say the same exact thing, but 10 different ways. So we have to find a way to, to match those and, and then missing data, sometimes data. Quite be inferred. Um, like if they were to start a job at a certain point, we, we can't quite infer that.

Interviewee

We can usually infer when they end a job, if they start a job right afterwards. Um, but sometimes some data can't be inferred and, and it's missing.

Interviewer 1

Okay. See. And how do you address these problem? Or more specifically, which tool you use to Yeah. Address these.

Interviewee

That's a good question, Laura. It's an ongoing process. Um, in our, in our data cleaning, uh, we, we typically have to go through and. Um, we've created a couple simple regressions to map the timing of jobs. Uh, I have a in algorithm that clusters titles and skills, uh, that takes into account the spelling takes into account what other similar ones are already mentioned.

Interviewee

And so those ways we are able to. In further the missing or, or

Interviewer 1

incorrect data. Okay. I see. Uh, so, so you have maybe some technique, but they, they are aircraft, so you do not use any, you do not use any, uh, tool, for example,

Interviewee

correct. For, for that specific data cleaning, we have yet to, to use any, any general tools.

Interviewer 1

Okay. And, and so you, you, you said for that specific data cleaning, but do you use tool for other data cleaning activities?

Interviewee

I'm not sure. Perhaps elsewhere in the company they do. Uh, I, I, I don't know on my end though. Okay, perfect. Thank you.

Interviewer 1

Um, so have you ever used external data, so that includes public dataset, uh, third party API, or Yeah. Anything like that?

Interviewee

Nothing large scale. I, I've done simple projects with, uh, data sets from, uh, I've done a couple weather, weather projects, so data sets from some of the US based weather collection services.

Interviewer 1

Okay. See, thanks. Um, and have you ever used data that was generated by another, . Uh, so any system that generates a, for example, a transaction of, let's say you are given a table of transaction that were, that were recorded by some system, um, yeah, this is an example. Have you ever used Yeah.

Interviewee

Um, that weather data is, is recorded and, um, otherwise imputed for, to, to make it clean and consistent in the different time zones and the different regions.

Interviewer 1
Okay, perfect. Thank you. Um, have you ever measured the quality of your data and or tried to improve it?

Interviewee
Yes. Oftentimes when beginning a project, you, you look for outliers, uh, if, if, uh, the temperature of a, a specific. Region goes beyond a thousand degrees or something like that, that's obviously, um, an outlier.

Interviewee
Uh, we, we look for that and we regularize it with the, the seasonal trends that, that it should have. Okay, I see.

Interviewer 1
Thanks. Um, what are some of the issues you repetitively encounter when you are preparing data for machine learning?

Interviewee
Really the, the most frequent issue is, is missing data. That, that one is the, the one that comes to mind for most frequent that I have to deal with. Okay.

Interviewer 1
Perfect. Thanks. And is there any other data quality issue we missed that you consider relevant?

Interviewee
No. Okay. Thank you.

Interviewer 1
All right. Um, so how do you evaluate the quality of models?

Interviewer 1

And as a reminder, quality is not only defined by the ML performance, like accuracy, F1 score, but you also have other respect, such as robustness, uh, scalability, explainability,

Interviewee
uh, yeah, anything. Uh, yeah. So we use, we use all those, um, oftentimes in the unsupervised area that, that I work in, we. We do an eyeball check.

Interviewee
We, we look and make sure that the output is, is relevant and is consistent. Like we are grouping firefighters with ambulance drivers and not with lawyers. Um, so the eyeball test is, is one that I've used in the past. Uh, otherwise we, uh, we also look at the, the consistency of the model. If it's, um, if there's any randomness.

Interviewee
Within the model, we, we look to make sure that, um, answers generated are consistent and the variance between the answers and the input should be, um, continuous.

Interviewer 1
Okay. Could you gimme an example of, uh, what you just mentioned?

Interviewee
Uhhuh, . Uh huh in, uh, in one of the models I created, I, I was predicting whether an individual was going to leave a.

Interviewee
And so we take into account, uh, several different variables, including the, the typical time that they spend in a job, uh, the typical time anybody spends in that kind of a job. And so we would hope that as we test out to see, um, How the model does with small perturbations, we, we'll take the same individual and manipulate a couple different variables, like how long they typically spend or the, the frequency that they switch jobs.

Interviewee

And we look to see if that has a corresponding change, uh, relative to the scale that it. Okay. I see. In the output.

Interviewer 1

I see. Thank you. So basically, this is an issue you have often with models is that they're not, uh, like they, they're really, uh, any small fixation in the input might create large fluctuation in the output.

Interviewer 1

Yes.

Interviewee

Okay. We want to avoid the chaotic, the divergence there. Yeah. Yeah. I see.

Interviewer 1

Um, have you ever used benchmark model to evolve with the quality of your models?

Interviewee

No, I have not. Okay. Thank you.

Interviewer 1

Um, have you ever assessed a quality of a model prediction with the user of the system?

Interviewee

I personally have not. The, the models that, that I created for, for example, the, the, whether someone is going to leave their job or not, that, um, in our presentations for, um, To clients, they use that, and that gives a good understanding of what the company will provide to the client. Um, okay. So they, they give a, a, a realtime assessment there, but I have not been personally witness to that.

Interviewer 1

Okay. And are you aware of some issue that, uh, might, uh, risen up after the presenting the solution to the.

Interviewee

Um, aside from possible incorrectness of the model, no. Okay.

Interviewer 1

Thanks. Have you encountered any other quality issue during the evaluat evaluation of your model?

Interviewee

Not that I can think of now, no. Okay. Thanks.

Interviewer 1

Um, how and where are your models?

Interviewee

Um, are you, you saying like infrastructure are, we host them on Amazon servers? They're, they're, yeah. Everything I've done is, is on an AC two cluster, I believe.

Interviewer 1

Okay. Um, what are some of the challenge you have in culture during the, uh, during the deployment of a machine learning software system?

Interviewee

Uh, permissions, I guess.

Interviewee

I as just a, an employee, I don't have the permission sometimes necessary to make the changes that we need on our, on our cluster to, to like the file organization, things like that. Um, sometimes I, we use a docker to, to, to launch the. Um, the whole system and oftentimes I'll have trouble getting a specific piece of software installed on the Docker to make sure that everything runs correctly.

Interviewee

Um, but that, that's it. Okay.

Interviewer 1

So basically your permission on the cloud, this is, uh, the main issue when deploying. Okay. Yeah. Uh, did you ever have a model that performed well locally but poorly once deployed?

Interviewee

Yes, the, uh, the, the first iteration of our model to do the unsupervised clustering, um, we, I didn't have access to a large enough computer, so we ran out of memory and unsupervised clustering is, is, can be fairly memory intensive and not super scalable. So I trended on a smaller set and, uh, when we deployed it, It shows that I didn't train it on a very large data set.

Interviewer 1

Okay. See, what was the algorithm you were using?

Interviewee

It was a modularity based clustering called Lova. And so it, uh, we defined a metric to say whether two people are similar enough and then we. We assigned that metric to everybody in the database. We ran the clustering algorithm and then showed who was closest and who was most similar and, uh, and created a, a sort of way to project that onto the larger data.

Interviewer 1

Okay, I see. Oh, that's interesting. It's, um, it's an efficiency issue basically. You cannot run the algorithm on a large data asset because it's not scalable. Right. Yeah. All right. Um, so, and my last question, have you ever encountered, have you encountered any other quality issue during the deployment of machine learning software system?

Interviewee

Um, no. Amazon's been pretty stable and as long as we have our docker running locally, it typically runs on their clusters, so we're good there. Super. ,

Interviewer 1

how do you ensure that the quality of, of a machine learning software system does not decrease over time?

Interviewee

Um, regular updates and bug fixes. Uh, I think what we sometimes do is, um, We sometimes do a refresh of, of the system to make sure that the packages that we import, things like that, are all up to date. Um, we run a test locally and then redeploy.

Interviewer 1

Okay. And, and do you have some time issues when you update the packages?

Interviewer 1

Yes,

Interviewee

oftentimes, especially when if we were to update Python, uh, that would take a lot of the work that we've already done. And, uh, we'd have to go back and, and rerun, um, or retype a lot of the, the code.

Interviewer 1

Okay. And also previously you mentioned that sometime you have bug fix. Bug fixes. Bug fixes, yeah.

Interviewee

Um,

Interviewer 1

what, can you give me some example?

Interviewee

Graphics, when we need to render a specific kind of graph or generate a PDF for a client? Um, sometimes if they. If we have yet to, um, catch a error in the data or a, a type of error in the data, then that will stop the rendering process and they'll be missing part of the PDF that we generate. And so little bug fixes like that.

Interviewee

Um, getting the, the map so that it's a, uh, properly sized things like.

Interviewer 1

Okay, so, so it's not directly related to the machine learning component, it's a more of a software issues, if I understand correctly? Yes. Yeah. Okay, perfect. Um, have you encountered issue with data during the maintenance of a machine?

Interviewer 1
Machine learning software system?

Interviewee
What do you mean? Like the, the data that, that is fed into the machine learning.

Interviewer 1
The live, yeah, the live data, basically.

Interviewee
No, I don't know. Um, I don't, our company's fairly small, so we don't work with a lot, uh, a high flow of live data. Um, a lot of the time it's, it's one request at a time and so, uh, there's not a lot to, uh, to really maintain there.

Interviewer 1
Okay. I see. Thanks. Did you wanted to say something, Amin? No. Okay.

Interviewee
All right. Um,

Interviewer 1
and I'm not sure, um, do, do you have models in production?

Interviewer 1
Yes. Yes. Okay. Um, and do you have, do you, did you encounter any issue with the model when it was, uh, deployed? No.

Interviewee

Okay. Thanks.

Interviewer 1

Um, . Is there any other issue we miss about model maintenance, uh, machine learning, software system maintenance that you miss, that we miss and you consider relevant?

Interviewee

No, I don't think so.

Interviewee

Okay, perfect. Thank you.

Interviewer 1

All right. Uh, so I will basically, I will list a number of quality aspects and if you ever had experience, like if you ever experienced quality issue with one of these, uh, you please, you can. . Uh, so did you ever add issues with one of the following quality aspects, furnace, robustness, explainability, scalability, privacy of data and security of your model?

Interviewee

Yeah. Privacy is one of the things that, uh, that we have to deal with. Um, since we deal with, uh, individual data, we, we have access. Data like LinkedIn. Um, we also have access to things like credit card information. So we, um, anytime that I handle any of the data, I I have, I usually have to get it straight from the server.

Interviewee

And, um, I typically anonymize a lot of it so that, um, we don't deal with any leaks and. Uh, so that, that's one thing that we, that we've had to work with. And as a co as we've grown, there have been quite a few other measures of security. So that two factor authentication to, to log into our services, things like that.

Interviewee

Um, so that's one of the things that we've had to run into.

Interviewer 1

Okay. See, uh, I have two questions for you. So, how do you anonymize your data? And the second one is, um, what are. What are leaks? You, you mentioned leaks earlier. What are you referring to?

Interviewer 1
Uh, so for

Interviewee
your first question, uh, when I get to the data, a I take out any sort of identifier. So we have an identifier from the original database, an identifier for our clients. And, and there's one more Id, uh, for each individual. . And so I, I, I take that all out. Um, I do a couple of random shuffles of the data and that, that's usually sufficient.

Interviewee
Uh, I don't, I don't handle any of the real sensitive information, uh, in, in terms of the, the data leaks. Um, as companies grow, we, you know, we, we give graphics of. The data that we have, um, the results of our machine learning models, and we just have to be careful that nobody has access to the underlying data that generates those and nobody has access.

Interviewee
That shouldn't have access to be able to pull the data from our databases. So those, those are the things that we have to lock down. Okay, I see.

Interviewer 1
Perfect. Thank you. Um. , and if I understand correctly, your your answer to how you anonymize your data. You are manually re uh, removing columns or feature that have sensitive information.

Interviewer 1
Mm-hmm. . Okay, thanks. All right, so I have two last questions for you. Um, in your opinion, what are the most pressing quality issues researchers should try to solve?

Interviewee
Hmm, that's a good question.

Interviewee

In, in the field of machine learning, I, I, I, I think. We do a good job in stating our confidence and in predictions sometimes in other fields, uh, in, in the field of nutrition, for example, when people use machine learning models to, to say or to derive the outcomes of a specific study, and they say, because of these outcomes, we know.

Interviewee

This compound has an effect. And, uh, it's very difficult to have, um, confidence and things like that unless you know every single parameter. But I, in, in everything that I've read in, in my field, we are good about labeling the level of confidence that we have with regards to the claims that we make. So I, I, I don't know of any direct things that, uh, need to be addressed.

Interviewee

Okay. Perfect.

Interviewer 1

And do you have any other comment about the quality of ML system?

Interviewee

No. No. Perfect. I, uh, I'm excited for when random forests are not the most commonly used models. Yeah.

Interviewer 1

That's for sure. And, and what, what do you think is missing for this to change?

Interviewee

I think there has to be a lot of work in the studies of, of PDEs in order to understand the underlying systems that are going around and be able to, to solve those a little better.

Interviewee

I, I, I'm a big fan of, of trying to understand the underlying. And random forests do a great job at giving you a good answer, but their explainability is, is very poor and PDE is, there's, I don't know, also hard to explain if you, if you don't have a good enough background, but it's a lot more confident, you can be confident in the relationships.

Interviewee
I'm sorry, what

Interviewer 1
is the acronym? P d E for?

Interviewee
Oh, partial differential equation. . Perfect.

Interviewer 1
All right. Uh, so yeah, uh, that's it for us. I, I have maybe some demographic question. So I understand you're from, uh, you're from the Country X? Yes. Um, and what size is, is your company more or less, is it like a small or medium or large company?

Interviewer 1
Small.

Interviewee
It's small, yeah, small. We, we have probably around 75 or so employees. Okay.

Interviewer 1
And what is the, the, the field in which you are working? I, I don't mean data science, but I mean, uh, data science applied to, so, ,

Interviewee
we, um, we work with, uh, we're, we're like indeed, you know, indeed or LinkedIn jobs. We're, we're kind of in that, in that genre.

Interviewee

That's true.

Interviewer 1

You told me earlier. Okay. Perfect. All right. So thanks a lot for being there. I, I, uh, I mean, yeah, I, I, I think it was really great and, uh, I mean, you spent, uh, 30 minutes with us, so, uh, we really appreciate your.

Interviewee

Yes. Thank you very much. I hope you guys thank you so much.

Interviewer 1

You too. Thank you so much.

Interviewer 1

Have a good Yes.

Bye.
