

Interview 4 - Amin

Interviewer 1

okay. I will start. Um, so what we want is to develop a catalog of quality issues for machine learning software system - uh, what is a quality issue? Well, any issue you have with, uh, some software system - machine learning software system, and that affected his quality. So for example, you could have a machine learning software system that is not explainable.

Interviewer 1

Uh, it could consume a lot of. Resources. So memory, uh, you know, uh, these kind of concerns and, and, and just a brief parenthesis, a machine learning software system is just a software system that has a machine learning component in it. So for example, the recommender system of Netflix for like suggestion of movies, well, this is a, a machine learning software system.

Interviewer 1

So. As long as it has a machine learning component. Yeah. That's it. Okay. And why, why we are doing this is to improve future work on the, well, no, to help future work on improving the quality of ML systems.

Interviewer 1

So we already asked for your permission to record the interview. So that's perfect. And, um, I will. If you're, could you tell us, I will ask to use some background information. Uh, can you tell us what is your current position, how much experience you have in ML or even in general? Um, and yeah,

Interviewee

so I've been working as a data scientist that Company 1 for almost two years.

Interviewee

And I started like, before that I was working for six years in market research, dealing with the data analysis, modeling and stuff, but like more simple stuff. Modeling like linear regressions and stuff. And then I decided I want to just deepen my knowledge and move into data science and machine learning.

Interviewee

So I went to School 1 in 2018 and like, then my journey started. So, um, um, that's about it. If you need any more details, let me know. So, but like I'm very familiar with, uh, data overall.

Interviewee

Exploration and stuff. And, uh, for the modeling part, it's probably the most difficult part in the data scientist life, make it work and to, to, to make, to produce a good

Interviewer 1

model. Yeah. Yes. So, so you said you went to, um, isn't the master in data science and business analytics at School 1 you did?.

Interviewee

Yes.

Interviewer 1

All right.

Interviewer 1

Um, okay. And, uh, I already know the answer, but for what purpose you use artificial intelligence at, Copmany 1.

Interviewee

To predict the future. Let's say so most of the time we want to predict something and that's something normally would be in the future. Um, yeah, most of the time we are dealing with some times-series problems.

Interviewee

Um, up to this moment that I had at Company 1.

Interviewer 1

Okay. A lot of time series problem. That's what you said. Well, an

Interviewee

anyway, you know, even if it's not a pure time, serious problem, you're just trying to predict a success or failure of a student on a platform. It's still taking into account. The past progress of the student.

Interviewee

There is some history, so it's like semi time series, time series is always present in the background. Even if you want to predict, like now in our current pro project, again, uh, if, uh, something will succeed or not. We are dealing with the data that was formed over the years. And there is some shifts and the distribution in the data and in some categories that we want to predict, there is a lot of times seriously.

Interviewee

Anyway.

Interviewer 1

Yes. Yes. Thank you. Um, and generally, uh, well, Yeah, you, you already answered this one. I was, I'm going to ask you which type of data you usually uh, have, but, uh, you you're usually with time series and, uh, and tabular data from, from what I understand. Uh,

Interviewee

yeah. In majority of projects. Yes. Um, sometimes it's just the binary target that we want to predict.

Interviewee

Sometimes it's a numeric time series once we had, we dealt with the predictions for the in language processing,

Interviewer 1

so. Okay. Interesting. And how is the data? Well, that's a bit large. Um, I will, sorry, I'll move on to the next question. um, which model algorithms and framework do you us usually use? Uh, so for example, do you use psyche learn?

Interviewer 1

Do you use, uh, TensorFlow, Pytorch?

Interviewee

I would

Interviewee

say, um, like for the, maybe for the baseline models, you could produce something with psyche learn, which is appropriate. It might be logistic progression, some really basic algorithm. Um, we really like, uh, L G B. Um, models for, uh, the, Pytorch. We did use it in the NLP, um, project, but for materials, I personally did not use it that most.

Interviewee

So, uh, like the powers of L GBM are good enough. They, they are actually very well. Very good. So you, they need to go into, into the deep learning stuff and.

Interviewer 1

And sorry to ask you, um, L G B M. Is it in, does it has its own library?

Interviewee

Um, yeah, it's like a standalone library. It, it has like wrappers around it, so it works as a scikit-learn model.

Interviewee

So you can use the same commands, fit, transform, and so on. So, um, you, you put the, you, you put their `x_train`, `y_train`. Test sets and everything. So it works as a psychic learn model. It's just like, it's a standalone package and it has a lot of, uh, configurations that you have, uh, um, that you can manipulate those hyper parameters that you cover.

Interviewee

And so, and gives a lot of flexibility and it's like a tree based model. So it can give you a lot, a lot of, um, as I'm saying flexibility, in terms of, uh, finding. Um, very particular, let's say the zones of the distribution and predicting accurately there, so,

Interviewer 1

huh. Great, great. Um, where are your models? Uh, usually usually deployed.

Interviewer 1

Do you deploy them on Azure, AWS or GCP? Something like this?

Interviewee

Let's say maybe so far, we did not have so many projects where we did deploy the model. Let's say in my experience, it's like five projects in which three we deployed. So it was once on GCP and, um, Azure we are using Azure of the time, but my colleagues can be used AWS as well. So it's, it's just my particular sample of projects, which is rarely needed.

Interviewee

Okay, good. Nice. Thank you. And that, that also depends on the infrastructure that the clients have. So you might need to deploy in the clients infrastructure and that's, that's the, that's what determines the solution, where to.

Interviewer 1

Okay. Perfect. Thank you. And maybe I misheard you, but did you say that not a lot of, uh, the project you work on were deployed on a platform?

Interviewee

Uh, some projects were like, you know, proof of concept. We just need to produce a model and prove that it can work. So we don't deploy is just a very short and sweet project. And. That's it. Um, one project was not about even predicting at all. Um, it was more like business analysis project and we did not have enough data, so we just stopped it for that reason.

Interviewee

Um, the rest of the projects here, it was like one of GCP. Second on Azure. Now we are going to deploy on. And, uh, that's, that's it for my projects in particular. So Azure data bricks.

Interviewer 1

Okay. Okay, perfect. I, I it's a bit random, but I, I was thinking about, uh, I should have said in the intro, but I think I went to Company 1 and I saw the, not the painting, but, uh, You did something with a model and they produce a painting.

Interviewee

Oh yeah. That was really, uh, a fun stuff. It was just, uh, an attempt to utilize the client's data. And it was like, let's see everything in one plug, but I just plugged in the wrong data. So let's say I had two data sets, which had this similar variables, but one was like aggregated and another one was not.

Interviewee

And so I plugged in non aggregated and bump. And we started joking. Can we sell it as NFT and so on? So after that, I just,

Interviewee

I didn't know what that he was going to do.

Interviewee

Yeah.

===== Connection got interrupted =====

Interviewer 1

All right. So I'll just repeat the question for the transcript. Uh, so what are the main quality issues, uh, you have encountered with your data model or system so far?

Interviewee

So with regard to data, let's say we're collecting external data. Um, and if it's positioned like as the final one and. After some time they're having a model electrons in production bump, they changed the historical data

Interviewee

For the weather, it was just like so unexpected and change it a lot. Like for the temperature, they changed by four degrees, let's say for some region regions. And it was like, okay, so this data source just unreliable. How can you change the temperature data for a specific region for four degrees? Just like that, you know, for the past data for the past year of data.

Interviewee

So that was, um, um, very annoying, let's say. And that was unexpected because when we were working on the model, we didn't see those big shifts. We saw that the data would, was updated for maybe the past two weeks or so, but then the end of here revision came and everything shift. and all the predictions that we are getting with the previous data labor were correct.

Interviewee

I'm sorry to say that because the revised data was good, but what they produced in the past month was just bad and the predictions were very off versus what we expected to see. For example, then, um, quality the issues in data. If we continue data, let's say the data that we can work with. Can be biased to where it's, uh, third elements.

Interviewee

I don't know why it can be product that is very widespread versus other products that are low represented. Um, it can be, uh, vendor that is overrepresented. So let's say we have a huge data, but the bulk of it would belong just to one vendor or just a couple of products. And multiple products would be very heavily underrepresented.

Interviewee

And in such a case, Cannot expect that we can produce good predictions for, uh, products that are, um, appearing several times in this sample. Let's say so. And if the need of the client is to have good predictions for maybe some critical products that are appearing only twice a year, then. Then there is no question how then there is no answer how we can figure good predictions with this type of data or with this like, um, problem set up.

Interviewee

So on that's something we are dealing with right now, for example, then, um, a big, big question is they, um, not gift, like I mentioned that, uh, I've been participating in five projects. Three of them were having some problem with the target. Let's say, either target is not available and we are trying to construct it from the data available.

Interviewee

Let's say client's data, but it's not representative of the general reality, but we assume it's a good proxy. And then we start changing the definition. We take this subset of the data. We take this and that. And then like, you know, half the half of the project, we just spent redefining the target and fine tuning.

Interviewee

Because different people at different moment who were the users, they were saying, ah, that's not what we see. Your approximation is not good for this particular week. So I don't believe your model. So , you know, your target is not good. But then on average, we saw that the target is good as describing, but sometimes it's, uh, like it's missing the particularities of a given week.

Interviewee

Let's say the approach that we were using. Um, another example is, uh, like a client wants to get predictions on the texts, but they don't have the, uh, target code at all. We start the project, no target coded. And then we start coding the target. And then what I see in all those codings that the simple documents are coded, but more complex are not getting coded or it's like much more difficult and it takes much more effort to code them.

Interviewee

So let's say you produce a model. It prove that it works, but on just this sub-sample of data and we don't know how it'll perform. If you start coding more complex documents. Uh, so yeah, in the models, uh, uh, oftentimes is, uh, the issue that we are trying to be more precise, and that's why we are looking into a more complex model and maybe putting more factors inside.

Interviewee

and then bam, you cannot explain, well, not stepping inside the model. You can use sharp and this stuff, but in the end, when you look at, uh, those, um, sharp factors, when you like aggregate them and see the whole picture, sometimes you're looking at the variable. That should be, uh, lowering your forecast, but instead it's just making it go up for the value that you expect.

Interviewee

So it's just work and counter intuitively. So in those complex models, a model may try to squeeze out something from the variable that maybe it shouldn't be doing. So overall, the prediction let's say is very good for the model, but when it comes to the explanation just goes like, um, Nowhere because it produces more questions than answers if they are trying to explain it with those, um, sharp values, for example, and stuff.

Interviewee

Uh, so of course in this case, some kind of a linear model would be, uh, much more explanatory, but the quality is not comparable. It's just so much lower than for the more complex model. So there is this kind of trade off that we are dealing with. And I know that some of my colleagues were implementing those, uh, sharp explanations at the very high level.

Interviewee

Let's say they might have many variables and they would aggregate some of those and, uh, show up the results just in very, uh, like bulk view. Like, let's say contribution of this group of variables, whether bum, like, I don't know, distribution, BU something else bum. So in very aggregated and it was working for them, but not in our case.

Interviewee

So, and the, I, we were having multiple predictions. We were having predictions for multiple weeks ahead and each week would have their own explanations. And sometimes the weather would contribute positively to the predictions sometimes negatively. And it's very, really hard to explain because when you analyze those sharp values, you need to take into account.

Interviewee

Not only. Um, well, you see the contribution, but you don't know, uh, what kind of, what was the value of the variable that produced that contribution? So you need also to know the value of the variable to understand the, let's say the direction of an effect that it produces. Uh, yeah. So it's like, it's like really complex stuff to explain your model.

Interviewee

Well, what I find, but with the simpler models, Uh, it's much easier. Let's say for a logistic progression, it was very nice. You just see your, um, coefficients or whatever, and the client was satisfied with that already. So

Interviewer 1

thanks a lot for all this information. Um, I will start in, I will last question, follow up question for, uh, every point you've mentioned.

Interviewer 1

Yep. In the order you've mentioned them. Um, so far the weather problem with the four degree difference. Was it that your historical data somehow changed from, from, from, sorry. It changed with a four degree difference?

Interviewee

Uh, yeah, it was a very interesting situation. So we started project somewhere in March and we were like, by, by end of the year, we deployed the project.

Interviewee

It was running in production for so always good. But then when we come and our predictions are just so like, They are so missing the reality. And we were like, what's happening. And we started digging down the, the chain and figured out that just the entry data for the weather have changed a lot. And, um, yeah, it was like, we produced the comparisons with sent the comparisons to the vendors and they're saying, oh, we just revised the whole year of data.

Interviewee

And we were like, oh, nice. Before you were revising just the last maybe week or two, but now you revised the whole year. And sometimes it's just a huge difference. Uh, and so, yeah, and in the end we just stopped working with the vendor and found another one. Um, that issue should be solved for now, but we need that kind of, uh, data stability, at least the data on which we develop the model.

Interviewer 1

Yeah. Do you know why they changed so much or their historical data?

Interviewee

Um, They make some kind of revisions every year, that's it. And in some cases they were saying, ah, this data was just originally wrong. We don't know what happened. So. Part of the explanation was it's like, ah, we just did the lame job here.

Interviewee

So, uh, so we needed to, uh, repair it and, uh, it was, and that's like part that they corrected, let's say starting autumn, they corrected the data a lot. And those autumn data were used to make predictions for, let's say January of the next year. And the January comes, we see the predictions are very off. And when we plug in the new data, Predictions are good.

Interviewee

But with the old data we are saying like what what's happening guys? And they, there was a long group and, uh, in the end they just ended with the new provider. So at some point I stopped participating in the project. So I don't remember the whole picture. So all the details.

Interviewer 1

It's great. Thank. Uh, I think I don't, I don't remember what which project it was, but you mentioned you didn't have enough data or you had missing data, uh, and you, do you know which one I'm talking about?

Interviewer 1

Because I only noted missing data.

Interviewee

Uh, so, um, maybe you, it was about the target was missing when we started the project. Right. So we were expected to have a small sample of label data, but the project came and there was. The person that was supposed to do this on the client side was doing something else.

Interviewee

Uh, and then we didn't have any label data. So we started with just, you know, um, in some parts we could create our own labels. It was a text data, and we could just go and use some, uh, regular expressions to extract some keyword. What we did. And that was, and that was actually, um, one of the best approaches that we had on the project that were a simple rule based algorithm that produced very good predictions in there.

Interviewee

And the model was like trying to reach that level of precision. Uh, but I understand that it might be, uh, a connection between, uh, those keywords that we noted and the person that was labeling the documents at that point, when she was written to the same keyword. So it might be like, you know, favoring those rule based algorithm.

Interviewee

So, yeah, that was, uh, another problem. And then, uh, I was saying that, uh, when we started labeling those documents, it was, uh, that the simpler documents were labeled first more complex and with different formats. And with, let's say from different organizations were labeled very little. And, um, once let's say we have more samples, the same model, of course won't work, uh, on the data.

Interviewee

So there is always this question of, you know, you are trying to build the model, you need the data for it, but if you have garbage data, you will have the garbage output garbage in, garbage out. The quality of data that we're getting. And it's maybe not the, the build data essays, but like very biased data.

Interviewee

In this case, it was biased to where simpler documents of the same organization that had a very specific structure and so on. Um, other documents were not having a similar structure. Um, so I'm just saying that the same algorithm, I don't expect it to work properly. It needs to be explored more and fine tune a lot.

Interviewee

So it's just like, not that you plug in new data that you are getting new predictions. Uh,

Interviewer 1

Yeah. All right. Thank you. Uh, you also mentioned something about explainable model, but we have question on explainable later on. So we'll, we will come back to the, to that, uh, later on. Yeah. All right. So we'll jump into the first phase of the workflow.

Interviewer 1

Uh, so data collection, you already mentioned some problem yet. When collecting data. So for example, uh, some of your text data was not labeled. Um, and I will go with this. Uh, well, I will ask you question, well, let me rephrase this. My first question is how do you collect data to train your models? And, uh, I will go on every, uh, data collection method that I know of.

Interviewer 1

And if you ever use it, you tell me maybe if you had a problem with it and if you don't, we, we just are going to move to the next one. All right. So the first question is how do you collect data to train your model? And, and did you ever use data collectors? So you pay someone to collect data for you.

Interviewee

In that case for the weather data that we adjust it.

Interviewee

Right. And also, um, we use the data from some providers, like experts in data that we don't have the in free access, so you have to pay for it. So yeah, we did use this data, this kind of data. Mm-hmm mm-hmm but, um, well maybe it's more specialized. It's not just like generic, but

Interviewee

that's it.

Interviewer 1

Okay. Are, are you referring to the API?

Interviewee

The API?

Interviewee

What, what do you, yeah, yeah, it, it can be API. It can be, it depends on the provider. The best case for us is having the, an API that you can just query easily and that's it. But, um, we had to deal with. Um, just Excel files that are loaded on the platform and we have to create an algorithm that would go, you can download that file, save it, transform.

Interviewee

And then it goes into our data. Um, what else? We had to scrap the data for the text, for example, it's just, um, go in one by one for the necessary links and scraping everything there, or just the parts that we needed. Um, what else. API files. Yeah, it can be just a, um, a page. Where's the, um, table. Well, you just need to put some checks in.

Interviewee

Okay. I want to data for this period for this product, or I dunno, this countries. And then, um, it produces you press like button create and it produces some table, uh, on the screen and then you press another button download. So it's kind of this process that we needed to emulate. So, like, I know data engineers use use some survey and stuff, so to, to collect

Interviewer 1

Okay, thank you. Um, do you, do you use some time public data sets,

Interviewee

data sets?

Interviewee

I guess? Um, we did like, at least we explored a lot public data sets. Um, let's say United mentions. Uh, what I, I don't remember exactly, but the experts in first data between countries, we explore that a lot. It has an API sometimes it's it breaks down. Um, and, uh, the most important issue with that was that we know that the data exists for some years, some countries, because we saw it on their websites, but it's not represented on the, on that site.

Interviewee

So it's just like very, a lot, a lot of missing data for our countries of. So we know it exists for 2020, but on their website, it's just 2018. The last year that is available. The issue of, um, um, updating the data is very often the case with the public data. And, um, most of the time it would be coming out later than you need it.

Interviewee

If you need, let's say weekly predictions, but it comes once a month with the delay of half, uh, one month and a half. there is no use of the data for you. So yeah, we started a lot. We used a little, uh, what we did use, we did use, uh, United S D a data for agriculture, some agricultural stuff. So it's really ABL in weekly or even daily.

Interviewee

And, uh, it's not very easy to collect again. We had to go click here, click there, download this file. And so. but at least those data was, uh, off like reasonably good quality. They did have some revisions, but not as drastic as I mentioned for the weather stuff. And mostly for the, uh, later periods of data, newer data, um, and, uh, um, whether they done well, we paid for it.

Interviewee

I don't know if it's public or not, so. And what else. Maybe some transportation data, again, available from the us statistical websites. It's also can go through through an API. And so it was kind of good, but at the point when, when it was discussed, including the project, I was already out of the project.

Interviewee

So I don't remember, I don't know all the.

Interviewer 1

No worries. No worries. Um, uh, do you, do you remember if, if, if there was, uh, quality issues with the data, for example, missing data, dirty data, uh, these kind of problem.

Interviewee

That really depends on the data that you get. And let's say, um, the main source of data for us would be the client's data, right?

Interviewee

So, um, that would be normally some database and we would use some SQL clearing on the database. Um, and like, of course you can expect all kinds of data issues. And I. You know, as a person that works with data for many years, I know that there will be a problem with the data. You, you cannot bet on it having like all good and right.

Interviewee

So you need to get somehow like the future proof, but, well, that's not possible, uh, in the first place. So you need to know that the issues will come. You don't know when they will come, but like, this is something that you need to plan for, like maintenance of your project and so on. Uh, so, um, for the, um, bad quality data, I totally for the weather right then, uh, for, um, let's say open sources, uh, first they would publish the, uh, preliminary data and then they would update that would normally affect the past several weeks.

Interviewee

The problem with your models. When you train them, you will need the most recent data. Right? And the first data that is issued is not like 100% good. And you don't know, it'll be a revision of 5% or it'll be revision of 100% and you are making predictions with that latest data. Right. And then in three weeks it gets revised and you are like, ah, I see my predictions are off.

Interviewee

I see with the latest data, they would be good, but you. People already got the predictions they need to make use of those predictions or, or they cannot use of those predictions because they start trusting them, for example.

Interviewer 1

Okay. Okay. I see. And, and, sorry, I, I have one clarification question. Uh, so you mentioned there was revision.

Interviewer 1

Uh, I didn't catch up on it at first, but this means that the data that you consume in your model is a prediction of another model. Is it. Am

Interviewer 1

I understand, right?

Interviewee

No, it's not like, uh, a prediction of, uh, another model. It's rather, so let's say we are going to collect some data that is issued by us government, but they issue, uh, today they, they are issuing like, um, today for the yesterday situation.

Interviewee

It's very fast that they issue the data, but like two weeks comes and they revised this recent data. And sometimes these revisions are just like we little ones, like couple of percents, but sometimes it can be just, oh, we just made a mistake. At some point we saw some exports of fruit from Country 1 in December, which was like nonsense.

Interviewee

And, uh, we saw that in the revisions, they removed that, that data. But it was really like the bad data. And if you relied on it, that would be a problem for us. And the problem is like, so they issued preliminary data. You use it, you use the most recent data for your predictions. You can go maybe and save.

Interviewee

Um, instead of this week of data, you can use the past one, but you need the recent data, but it's still not the final one. And, uh, depending on the vision that they are going to, uh, produce, it can be, uh, a big shift. In the numbers. And if this variable is crucial for you, then good luck with your forecast, you know

Interviewee

Yes. So did, did, did, was it clear for you.

Interviewer 1

Yes. I understand they, they have some revision, uh, because they miscalculated something or

Interviewee

yes, it's not necessarily miscalculation, but it's just the pro process that statistical offices would have, you know, so first they issue the, the, all the available data this moment, but then, um, they kind of, um, go through probably some better internal process of validating that data.

Interviewee

And, uh, they can, that can introduce changes sometimes important.

Interviewer 1

Mm-hmm okay. I see. And, uh, I will ask really quickly. I see. Time is, uh, going fast, so I will ask, ask very quickly. Uh, so for, but for the weather API, did you get, get, were you getting prediction the future for the temperature or it was always historical data.

Interviewee

Um, We were using historical data in our case. Okay. And it was getting revised, but yeah, there is also this issue, you know, then maybe you want to build a model with the weather, but, uh, for the history you can use already the weather that materialized, but for the future, you might need the weather forecast, which, which are not of the same quality as the data that, uh, would be already recorded as actual, right.

Interviewee

So. That, uh, that is like, so we're evaluating the model on the past data, all the evaluations that we do are on the past data, but then, um, when you start using forecasts, you should expect errors to grow. And that's why we're always making this disclaimer. So, you know, like if they're using stable data in the past, but unstable for the future, then that's part of, uh, why our model may start fluctuate more than we want it.

Interviewee

Yes.

Interviewer 1

Yes. That, that, that, that must be a challenge. All right. Uh, thank you. And, uh, so the last question for the data collection section, uh, have you encountered any other data quality issues caused by data collection?

Interviewee

Um,

Interviewee

Data collection. It might be, you know, uh, as, as ugly as having decimals in your numbers, uh, like coms or dots and, uh, it maybe messed up from source to source and you need to be extremely careful.

Interviewee

And some source are like in letting America, for example, they were even having. You know, in English like system, you would, could come up for, uh, thousands to separate solvents and dots for the decimals. It was completely opposite. So like this kind of is, they are annoying. Um, but yeah, yeah, you can spot them, uh, missing data.

Interviewee

Of course we have missing data. If, if there is a lot of such missing data, we cannot use it. Um, But like every data set would have some missing data. I would say I, I did not see an ideal data set where every variable would be filled 100%. Um, it's a question if you need it, maybe they don't need it. We just drop it.

Interviewee

Um, but sometimes yeah, sometimes it's important. Well, then in that case, we, we can just add an indicator variable saying, okay, this main variable was missing, then plug it into model and see if it, if it helps us explain any better or predict any better. So, um, What other data issues, or it may be a very complex data, for example, that you, from the business rules, when you are discussing with the client, they are telling you one thing, but you open the data and you see, oh gosh, I have like four entities and then I try to join them and just like, Explodes my data because like, wow, that's, it is just a complex data that can give you a lot of pain for, let's say 5% of cases, it would be irregular and you need to develop special algorithms to treat it.

Interviewee

If you want to treat it and make separate decisions for it, how to, um, how to run it in production and so on, because it does not fit into your regular case of just like simpler kind of data.

Interviewer 1

Nice. All right. Thank you. Uh, so we will move on to data preparation. Mm-hmm the first question is, uh, which tools framework do use when preparing your data and by preparation, we mean data cleaning and data transformation.

Interviewer 1

So both of the, both of that,

Interviewee

um, Not something very specific is just, you know, we explore each data set first and then we pick up the variables that we need and we decide what transformations we need. So we just use I D Panas and basic algorithms to. To produce data urges, uh, cleanups and stuff. So there is nothing really sophisticated.

Interviewee

If let's say for data exploration, you could use like pandas profiling, for example, to see just like the general picture on your data, but that would not be even enough. You know, we would need to dig into the data more. And then based on the digging, we decide what transformations we need, um, and code them most of the time.

Interviewee

It's not very difficult to. Right. Some code to clean up, but, uh, the, the most important part is to understand where you need those cleanups to find the problems in your data to be able to repair them.

Interviewer 1

Mm-hmm mm-hmm so, so I understand, uh, data cleaning is most often manually done instead of automatically

Interviewee

with the tool.

Interviewee

Uh, yes. Um,

Interviewee

Yes, we, we don't have subscription for one tool I was supposed to try, but, but life stood in my way and I didn't try it. It's dataware or something. I don't remember exactly the name, but it can help you, um, clean and repair the data. Um, and maybe that would be something that I would look into more, but when I, when I saw this data that I have right now, it won't help me deal with it.

Interviewee

It's just like the data is complex. We need some, a lot of, uh, little transformations here, there, and there. And before we can create our final data set, where on some data, we need to aggregate it first, we're merging with the rest of the data. And so it's like, it's not something that is easily do. Okay.

Interviewee

With such a tool.

Interviewer 1

Thank you. I will move on to model evaluation section mm-hmm um, so how do you evaluate the quality of your models? Very large question. I

Interviewee

know. Yeah. So like normally you would build your data on this, on train test and you will evaluate on your test set. Uh, so, um, with time series models, like, uh, there is.

Interviewee

Um, specific thing that we do, let's say we want to update our model in production. So we would emulate this one, evaluating the model quality. Let's say we are. We know that we will predict once a week for three weeks ahead. So for our past data, we would repeat the same process. We would train them all on like be up to, I don't know, 2020, and then predict for one week ahead or for three weeks ahead, then move one week ahead.

Interviewee

Retrain the models. So like emulate the whole process. So, um, we can see how on the past data, not this specific model worked, but the whole process of updating the model. Uh, I know that sometimes my colleagues would just have this train and test set. That was done. Like for time series, you split it sequentially.

Interviewee

The most recent data would be your test data. Um, but like that's, uh, what we found useful, you know, um, not having just one model that predicts you, like, well, in this case, we are kind of training multiple, multiple models behind the scenes. Each week, we are training a new model in our evaluation. Um, this is something like more complex to design, even your experiments with.

Interviewee

And if you try to, um, to optimize your hyper that's, that's taking a lot of time, let's say. Um, but, uh, it, it gives you the kind of, in our opinion gives you the best estimate of, um, what you could expect in the best. When we train on the past data, as I told we are having already the stable data for the future, it's not stable.

Interviewee

It may change. And so this affects the predictions.

Interviewer 1

Hmm. Interesting. Um, the use, um, existing qualifying model is to evaluate your model. So, uh, a baseline out in the wild that you can't refer to, like not, not your own baseline, but some other models, someone did. And you use this to compare, uh, your performance.

Interviewer 1

It's really

Interviewee

specific. Well, you know, I would say maybe in an field that would. Will be easier and more relevant when you're using some, uh, um, pre-trained model and you're just ING it on your data, then you might see, or your quality maybe compares to some, uh, standards that were declared for that model.

Interviewee

But, but again, um, requires you having the same goals as we, for that particular model that was trained before you, uh, it's it's rare that you would. The exactly the same goal in our projects. Normally it's very, it's very like we are trying to customize our projects to the client's needs and it might be to get creative or very creative with, uh, implementations of, uh, algorithms and models that we.

Interviewer 1

I see. I see. Thanks. Um, the next question is, did you use any user acceptance measurements for train model? Sorry. So in other words, do you sometime go to the client and you say, here's my here's my model. Um, how do you like it? Something like

Interviewee

this? So, uh, this is sometimes we are constantly talking about the user experience, user acceptance, and there are high risks when it comes to the final user that they won't be able to use the model or will be unwilling to use it.

Interviewee

So we actually, in one big project, we saw it that. We were quite, maybe we should have started this earlier. We were always saying this conclusion, we should be engaging with the final user as early as possible, explaining all the details all in the way. Because when the day came, the first, the three produced the first predictions that were already good enough and could be used, uh, the clients, uh, people who would need to use the model were not ready.

Interviewee

They have some, um, processes that they followed maybe for decades. And then you come with your new model and say, oh, now your, our model will predict to the future. And you may adjust your, maybe I dunno, by inhibits based on this model. And they were like, we don't understand what it's about. Then they started doubting the target.

Interviewee

Then there was a lot of issues. Like, you know, that I saw that people are just conservative. They don't want to take risks. It was the, um, uh, project owner who wanted to take the risks, but not the people on the place who, who would need to use that model. So there was a big discrepancy between, between the, um, what we were targeting and.

Interviewee

People were, uh, capable of doing that, that particular moment in time. So it was a very long process of talking with those guys, explaining them. It like took us really weeks and months to reach to the point where the acceptance was better. Um, yeah. And, um, now starting a new project with discuss already.

Interviewee

We, we start discussions about the user interface, how the predictions will, I dunno, pop up on the screen or where would they be produced in an email, the alert that would come or like some pop up and whatever. And so this is where necessary to take into account right away and, uh, to prepare people along the way who are supposed to, to use those model predictions.

Interviewer 1

Okay. See, that's interesting. So you're not only responsible to develop the model. You're also responsible to make sure it is. used by the, the

Interviewee

end user somehow. Well, it's not like we are responsible normally. Like we don't have a user experience, um, like depart or, or person on the team, but this is something that we heavily consider and discussed with the client.

Interviewee

Um, in one project we have an external consultant that was having some discussions with the, um, people on the client side, trying to understand their moods and how they would use the model and how they overall, how their own. Works because they work like using some algorithms, uh, uh, and then you come and say, oh, now I give you this pre pre like predictions.

Interviewee

And they start to compare your predictions to their own predictions, what they heard from their clients and the other counterparts. And so, um, in the end, the, there is a lot of evaluation on their side and they may just mistrust it and, uh, either for reason or not. So. Um, or this week, like our constructed target was not mentioned their estimates and they discard the whole work.

Interviewee

Like they are tending to like saying, ah, your model produces some bit predictions, but when we dig into the data and when we see how we selected the data, it was like a regular process. It was good as just a right on a regular. Where the trends were shifting too fast, for example. And so, um, and that led to a lot of like back and forth between us and them explaining them what's happening and that, like, you know, it, it was actually a part of the budget that was redistributed to, um, the user experience and talks with the guys on, uh, that would make decisions using those model results.

Interviewee

Uh, and we did not accomplish maybe some other parts of modeling or some other products because this part of the budget went. Um, yeah. So now we try to devote more time to that. It's really important that that just like determines the success of your model in the.

Interviewer 1

I see. I see. Thank you. Uh, we'll jump to model deployment questions.

Interviewer 1

Uh, and the only question I will ask is, um, did you ever add a problem that performed well locally, but bad once deploy? So

Interviewee

yes, that around with the, with the weather each year? Uh, it was, um, well, our deployment process was a bit. Lengthy in time. So let's say over four months that we were deploying, the predictions were good and then comes together issue. And we just like finished all the deployment. And it, the model should be running on its own.

Interviewee

Some dashboard should be built on their own and everything. And then BU they see the prediction like what's happening and we spend the week digging into the issue. Why is it so weird? Even our predictions look weird on the plot and what's going on. Yeah. That was the case. And, uh, that was resolved finally with some amendments to the data providers and stuff, but it's like, A very big risk in our project.

Interviewee

But something like this happens and then like your month of work, uh, they will doubt if like, if the model is good at all and you need to find, to look for explanations and stuff. Mm-hmm

Interviewer 1

I see. Uh, did, did you have any, uh, did you take any measure to prevent this kind of things from happening? For example, you could maybe use, uh, data validation tool that checks if the, the, the there's a distribution.

Interviewer 1

Well, a shift in the distribution of the data.

Interviewee

Um, well, you know, we had a lot of discussions about this shift and the distribution of data and some data you would expect some shifts like, well, even your yearly better, it would be just fluctuating. One year. You may have, uh, hotter summers one year it'll be colder, but is it a shift in the distribution?

Interviewee

And so if you get values that are within your distribution, But then they get revised. Like, I don't know, twice then, um, it's not in the formal shift in distribution. It's just some problem with the data, you know? And, um, it's not even possible to, to. To track this down such kind of issues. Uh, I realized that sometimes it'll be possible if let's say they reported, uh, I dunno, um, humidity in one unit and then they just in one type of units and then they just decided that yeah, we, we will use another unit and then a few zeros.

Interviewee

That of course total grade would, uh, would, um, trace that down. But in, in the case when it's within you're bound of, uh, stuff, then yes. And also for some, uh, variables that are just changing over time in one direction. So your distribution is always shifting to where it's, let's say higher levels of something.

Interviewee

So, uh, the tool will tell you that this is something abnormal. You know, it's just a real, a reality and you don't need to react to that. So, uh, I, um, I would love to have such a tool, but, um, I don't see that it's like 100% automatizable

Interviewer 1

yes, I understand. Yes. Um, I will move on to, I see we have like five minutes left.

Interviewer 1

Um, if you wanna leave at 10. No worries. I think we are, we will finish at like 10, five.

Interviewee

Oh, it's okay. We can finish. Okay. We can take five or 10 more minutes of no problems. Okay. Thanks.

Thanks. If you think that my input is valuable, then yes, absolutely.

Interviewer 1

Um,

Interviewer 1

sorry. Yes. Um, do you have some models that have gone still after some time? So you deploy them and at some point you realize they're, they're not good anymore because, uh, there's a, like there's a shift in distribution or data is, uh, not good anymore.

Interviewee

Like, you know, I have very little sample of models. Um, I know that one project that we had, um, at some prediction at some time, prediction went very unreliable because again, of issues with the data, and I'm not sure it was fixed because maybe there was no much interest for that model on the client's team.

Interviewee

And it was a rather short process pro problem to project. And, uh, there was not a lot of few as for IAnd, so people were not using it. And when the problem arrived, there's just. They were trying to fix it, but then realized that, uh, they were not using it for example. So, um, it can happen. Of course. Um, on the other hand, I was surprised like in my first project that I had years ago at Company 1, that it was, uh, running very well, that people very making some really good views with it.

Interviewee

And it was even like, uh, competing for some awards. it was very fun to learn recently. Because I did not hear about the project for a while. And, uh, that's where I had a lot of dots, you know, like to put a simpler model in production or more complex, but then to have more power, to explain, uh, with a simpler model and stuff.

Interviewee

And in the end simpler model one, and the project is still running and the algorithm, you know, it's not about the model itself. More about how you approach this modeling stuff. As I told you, we tends to, for the time series, we tends to emulate our processing production by retraining our model retreat on the past data.

Interviewee

And, uh, so that's what we did in this case as well. And, uh, yeah. And I'm pretty happy with the results so far.

Interviewer 1

Yeah, that's great. That's great. Thank you. Uh, did you add any other. Any other issue regarding the main tenants of your ML

Interviewee

model

Interviewee

maintenance? You know, I participate less in the maintenance. It's more on the, um, data engineers side and they may have some issue like, you know, This source that we need to query just right now for the model. It's not available. It's under maintenance right now. So you need to run your, your pipeline later, or, well, maybe our guys can program that it reruns three times or three times it doesn't run them.

Interviewee

And you need to double check maybe manually or, uh, I dunno. Well, there are some solutions for that, but well, most of the problems would come from something UN. Suppose your, your file you are using, they just changed all. I don't know the names of columns and stock. Then you need to go in your project and change those column names somewhere in the parameters or whatever.

Interviewee

So it, it requires a manual input on your site, or let's say they completely redesign their website and all your scraping algorithms are down. you need to write the new one.

Interviewer 1

I see, I see that. That's a good point. We that's a good point. We'll take cons. Anyway, I was going into the, into detail, but thank you for getting input on maintenance of, uh, data pipeline.

Interviewer 1

So I will move on to the last section we have, uh, it's about quality measure of ML models. Um, so. Earlier on you talk about explainability. Uh, could you like, uh, just summarize what you said in general,

Interviewee

please? Well, I find it, uh, very difficult to have, um, good explainability for the complex models. That, um, by complexity, I mean the number of variables that will we use, sometimes it may be really elevated, uh, and, uh, the type of model that we use.

Interviewee

So suppose we are using sharp values and sharp values would, uh, show you how much your, uh, what is the contribution. Of this particular variable to this particular prediction versus the, let's say some average baseline that it produces behind the scenes. So suppose on average, you should be getting a value of 10, but you, your prediction is 15.

Interviewee

So in that difference how much it adds to your average. And, um, then in this case, Well, you can aggregate those explanations and maybe it works for you, but maybe it's not working for you. If let's say you see that your predictions that you're trying to explain well, and you take into account the values of the variables that explain them and you see, okay, my contribution is positive of this variable, but it should be negative.

Interviewee

What's happening. And then it produces more questions than answers. And this is really a difficult task to have good explanations for everything, especially for a complex model. In this regard, of course, simpler models are better. So we should be targeting as let's say, as, uh, low complex, smoother as possible.

Interviewee

But you know, that's not always possible from the point of view of. Precision that you are getting in predictions. So it's kind of a trade we're always wanting. The client always wants to see the higher maybe metric in the first right place, the higher accuracy. And then, um, the explanations are in this case more good to have.

Interviewer 1

Okay, sorry. Yeah. Okay. I see. And the reason why you needed explainable model, what it was, if I remember correctly, it was to explain some prediction to the client,

Interviewee

right? Uh, yeah. Normally that's why, well, it's, it's always good to, well, if you can look inside your model, uh, that's always good to have that insight.

Interviewee

Um, But the same saying like with the complex model, it's hard, even for yourself and you are beginning to the data trying to understand. Okay. So my prediction was made out of this data. These are my variables. This is like the value of variable X. And it says that it's like contributing five, uh, points to, I don't know, to this prediction and you're looking at it and you are like, like, wait a minute.

Interviewee

You know, is it really well what you would expect? Or it's just you coming back to the model itself, this like GBM model, it's just a three based model that is, that can split your data. As you know, as it wants to explain your data better. And sometimes it maybe like capturing maybe some random patterns in some parts of.

Interviewee

So like what I like to see with this explanatory part is that on general, on average, you see, okay. Let's say better is important. I know it's important and it's reflected in high sharp values, for example. Um, and sometimes you see it's positive, sometimes it's negative, but when you start digging into a concrete prediction, then it starts being a bit.

Interviewee

But on general. Yes. On average it works. Uh, you see, okay. Yeah, this is important. And I understand why this is not important. And I also know why, because I saw there is very little correlation let's say on stuff. Um, so, uh, yeah, at high level it might work at low level explaining the concrete prediction. It might go not.

Interviewer 1

Okay, super interesting. Um, I will move on to the next question. Have you ever faced problem related to the scalability of the model?

Interviewee

Well, not me in particular that that's yeah, I did not face, but like my other cliques might have faced.

Interviewer 1

That's perfect. That's a good

Interviewee

answer as well. Well, but we need to, to make note that, uh, I didn't have, so maybe big data in the first place, um, that I wanted to run my model on. Um, so most of the time it was, um, rather we needed data. And so it was not any issue, scalability. It was never any issue of other, like other quality issue.

Interviewer 2

Okay. Perfect.

Interviewer 1

Thank you. Do you have robustness issues when building ML model. So, um, this is maybe a more difficult quality to explain, but, um, basically if your model is used to some - I will give you a concrete example. For example, if you have a model that detects wolves in picture, it only saw picture of wolves with a background of snow. Well, if you put a wolf into, uh, I dunno, uh, a desert, well, it will not detect the wolf because maybe not detect the wolf because, um, it has a different background.

Interviewee

say, yeah, it's always a chance that your model will learn not from what you expect.

Interviewee

Like in this case with the. Um, and it was like Wolf versus asking my example that I know, like, it was just looking at the snow. It was not looking at the features of the wall and shape of the eyes and the areas and stuff. It was just taking. Okay. It's snow bomb. it's evolve. And then, yeah. So there is always this risk and the, maybe you, you can tackle it, uh, with some training down your.

Interviewee

And, um, just thinking logically what you want to put in your model, but sometimes you, well, sometimes it's just such a drastic, um, prediction quality when you plug more variables and, but then in the end you are getting those weird explanations. Like snow explains you. The role, like in our case was maybe like, um, some variable explaining, not in the way that we.

Interviewee

Like the model, it uses the shortcuts and there is no way to prevent those shortcuts. Maybe just streaming down your variables.

Interviewer 1

Okay. Perfect. Thank you. And that's the last question, uh, to conclude in your opinion, uh, sorry. I will rephrase that question. Uh, so, so if you, um, if there was one issue, you, I will rephrase that question again.

Interviewer 1

What is the biggest pain point in your daily work day that you would like to go away?

Interviewee

My biggest pain point is that, um, lack of, or inadequate targets or targets, they change the change in time because clients understanding changes and they want to adjust the target or whatever. So like, yeah, I see it as a big point and big pain point.

Interviewee

It's hard often, it's hard to obtain your targets. You need to derive it manually for some kind of tasks. In another case, it can be just like, you need to construct your target from the available data. And then depending on some business rules, uh, or like, you know, it may change. So let's say you, you, you have trained your model.

Interviewee

You have optimized your per. And then they come and say, okay, we just need to exclude like how the data from what we are modeling because of this business rule. And then you are like, okay, good. It's, it's good. If your model is still good on this subset of data, if you retrain it on it, but it might not be the case.

Interviewee

And you might need to, uh, restart your process of finding in the model. So yeah, I see the target, the good target as the main prerequisite of the good project. Part, hopefully like good data is the prerequisite of a good project, but good target is probably one of the important parts of, uh, the predictions of the models and everything.

Interviewee

Okay.

Interviewer 1

I see. Thank you. Uh, so that's all for the question. Um, yeah, I would like to thank you for, we took 10 minute. More than what you, what we said, who should we would take? Sorry. so, anyway, thanks a lot for your time. Um, great answers and I'm sure it's gonna be really useful, um, for our study. So thanks. And I will maybe see each other at the, at the sorry, movie eyes sometime I go there to work.

Interviewer 2

All

Interviewee

right. Okay. Well, thank you very much for this. Thank you. It was interesting to talk to you guys. Thank you.

Interviewer 2

Thank you for your participation. Sorry. I have connection problems. So

Interviewee

thank you again. Uh, weed the meeting. So, I mean, you can have a look at it later.

Interviewer 2

Thank

Interviewer 1

you. Thank you.

Interviewer 2

All right. Bye. You too. Bye-bye.

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