Interview 34 - Ilan

Interviewer 1

Okay. Think it's fine. Alright. Uh, so to start off, can you give us some information about you? So how many years of experience you have in machine learning? Okay. And what's your current

Interviewee

experience? Yeah. Uh, first, uh, my full name is, I'm from Region X. I graduated as computer science in 2020. And I have the skills in machine learning, uh, about three.

Interviewee

And data science special. I have, um, I have an internship with the Greenest company, uh, I think in 2022 in December, one month. And, uh, I ha I enjoy, uh, some courses with Google, Google Brain Achieve, and, and about learning and machine, the pro, the whole process of building. Okay. So I take the two path, uh, before building a model about data science and analyzing, uh, e d a atory data analysis and take the another path, which is the machine learning.

Interviewee

Okay.

Interviewer 1

Thank you. So great to

Interviewee

have you here. Thank you so much.

Interviewer 1

Um, so to start off, uh, we'll ask you a really general question. What are the main quality issues you have encountered with your data model or system so far?

Interviewee

Uh, so far, uh, actually when, when I enter any, uh, when, when I build or start to build any project in machine learning in general, The, the, the, the main goal, and the main issue I faced is, was the data.

Interviewee

As you, uh, as you know, gar, uh, we, we say in machine learning, in machine learning, community, garbage in, garbage out. So as the data, if it's garbage, the model will detect garbage. So the, uh, the best thing is to choose mo, uh, before choosing model is the data. Uh, when I analyze some data, I found, uh, uh, some sensors, uh, reading, uh, the, uh, reading, uh, I think, I think, uh, the weather.

Data-containing-outliers

Interviewee

The weather, I think reading the weather state, I don't remember. So I take, uh, the range be between, uh, 20 to 60 and, and the model prediction, it's give me negative, negative. When I see the model, uh, the model is perfect, but when I, uh, review, uh, the data, I, uh, I found the outliers. So, so, uh, when, uh, one of the most quality issues with data is the outliers and, uh, then the music value, that the two main effect on the data in general.

Data-containing-outliers

Interviewee

I see.

Interviewer 1

Thank you. Uh, moving on to data collection, have you ever used, um, data that was manually generated to train your models? So, um, I can give for example, someone who labels some image. Have you ever

Data-collection-web-scrape

Interviewee

used that? Oh, I, I use, uh, I do scraping data from some. And, uh, okay. Yeah. And, uh, and, uh, make model predicting, uh, uh, depend on questions and answer q and a.

Data-collection-web-scrape

Interviewer 1

Okay. Can you tell me some of the issues you had with

Data-collection-web-scrape Data-missing-values

Interviewee

the data? Mm, the data was, uh, text processing. So I have, uh, to remove, uh, first take the language and remove the stop. Library, uh, remove the snowboard and using limitation and steaming to return the, uh, to return the ver to the base, uh, like, like plane.

Data-collection-web-scrape Data-missing-values

Interviewee

It's play. Uh, and to make a search as smart, as smart as possible. Uh, and, and, uh, some, some data. And the same, the same issue I, I found with the same data and. Uh, uh, the more thing I face is missing data. Missing data. Okay. I see nightmares. Yeah. ,

Data-collection-web-scrape Data-missing-values

Interviewer 1

um, sorry, I forgot what I, I was about to ask you. Take that down.

Interviewer 1

Oh, yeah. That project where you, uh, you scrap, Did you, was it something that was deployed on the long term or, or it was a, like a one-time project,

Interviewee

Uhhuh in the long term and it's, yeah, it's updated. Uh, we can call it, uh, online bot or something online. Okay.

Interviewer 1

So I can, do you have any Yeah, go

Interviewee

for it. No.

Interviewer 1

Go ahead.

Interviewer 1

Okay. And did, did you have any issue with the, the wesc scraper? Like, did it work all, all the time or sometime it, it failed?

Interviewee

Uh, no. Uh, some, uh, it, uh, uh, harder, the hard thing is how to find the route and the idea of the, uh, scene question. Uh, but I, but I solve it programming by using some algorithm, simple algorithms.

Interviewer 1

Perfect. Thank you. Have you ever used public data sets or third party

Data-collection-system

Interviewee

api? Uh, uh. Like, like what?

Data-collection-system

Interviewer 1

Uh, for, for example, you want to do an image classification. Uh, you use, you pre-train your model on cifar and then you fine tune for your problem.

Data-collection-system

Interviewee

Oh, yeah. Like that? Yeah. I, uh, I think, uh, I work on a graduated project, uh, for, uh, for glasses for blind person.

Data-collection-system

Interviewee

Blind person. So instead of using sensors, regular sensors, I use camera. So how I train the camera, I use pre-trained model, uh, from depth. So I take the depth and make my own algorithm to detect a les. And so the blind person don't need to have a sensors around it, him. So on the. Camera to detect. Uh, and my project, uh, is, is succeed.

Data-collection-system

Interviewee

But, uh, the hardware is not very, because I don't have any experience with hardware, but the algorithm is, uh, doing well.

Interviewer 1

Okay. I see. And do you have any issue with the, the data you collected for, for that project?

Interviewee

Uh, no. Uh, no. Not, uh. Okay. Some, sometimes with the, with, uh, sorry. With environment, there are some, Because the weather, the lighting, so I have to fix this.

Interviewee

As you know, our environment are, are hard to make it, uh, detect anything or predicting anything, especially Forg classification.

Interviewer 1

Okay. Yes. Perfect. Thanks. Um, have you ever measured the quality of your data and or tried to?

Interviewee

Uh, uh, well, when I develop any machine learning or, uh, deep learning model, uh, after, after it, I made, uh, model evaluation by using, uh, the, the same metrics for, for accuracy and, uh, someone, uh, something.

Interviewee

And I see how the data affect and I think I work on a project. For detecting, uh, how can I say, uh, the crack in the street, the cracking, cracking line in the street, uh, how I say. Okay. Uh, so the street, the street, uh, and the street ground sometimes over time at, uh, at 12 o. Sorry, uh, depending on the weather.

Interviewee

So what I do is gathering data, depend on it to build, to build a drone project that detect, that detects, uh, some issues are found. Uh, uh, uh, Was delighting, delighting of the environment is very important because, you know, we have a real time in realtime video, sir.

Interviewer 1

Yeah. Okay. And, and so one of the issue you had with the data was that the lightning of the image wa was not good enough.

Interviewer 1

No.

Interviewee

Very good. Yeah. And, uh, the weather is, are changeable. Not a fixed, not a fixed, uh, light or fixed color. So you have, so you have to work on your, uh, uh, so you have little, little data and you must, uh, make your model complex, uh, to, to avoid overfitting under fitting. Sorry.

Data-low-quantity Data-low-quantity-fix

Interviewer 1

Okay. And, and so you, so I understand you mentioned two issue.

Data-low-quantity Data-low-quantity-fix

Interviewer 1

The first one is the LinkedIn and the second one is the size of the data set, which is really small. Yeah. And how do you, how do you address these two problems?

Data-low-quantity Data-low-quantity-fix

Interviewee

Yeah. Other, other bees? Uh, no, just this I, because I,

Data-low-quantity Data-low-quantity-fix

Interviewer 1

yeah. How, how did you try to solve the issue? Uh,

Data-low-quantity Data-low-quantity-fix

Interviewee

Uh, solve the issue by, uh, making some computer vision or, uh, image processing, sorry, uh, algorithm to fix the lighting.

Data-low-quantity Data-low-quantity-fix

Interviewee

Use some, uh, some aor, uh, some equations math equation to solve the lighting. Yeah. Oh, okay. Perfect. Thanks. Sorry. Excuse my language because, uh, no, no, no. Yeah, don't worry. Thank you.

Data-low-quantity Data-low-quantity-fix

Interviewer 1

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

Interviewee

Mm uh, no. Uh, uh, the, the best, the goal, the goal of, uh, the goal of building machine learning, the process or system anything is the focus on data before the model.

Interviewee

Uh, when I start working as machine learning, uh, project, simple project, I focus only on the model. And that's not correct because the data, as I mentioned before, it's like a, uh, we say garbage in, garbage out, clean in, clean out. It's depend on data. So, uh, the model sometimes may, uh, be. Depending on the, the data that we feed it, uh, I think I, I give, uh, give to the machine, uh, the learning model, make learning model to, uh, to generate outs.

Interviewee

So when I make it, uh, uh, it's go by us and, uh, uh, take, uh, part of un meaning. So the problem on the data, . Yeah.

Interviewer 1

Yes. Perfect. Thanks. Thank you. Um, moving on to model evaluation. How do you evaluate the quality of your models? And as a reminder, quality is not only defined by the ML performance, so accuracy upon score.

Interviewer 1

But there's also other aspects such as scalability, explainability, robustness, yeah.

Interviewee

You name it. Uh, when, when I work on any project and work before and still work, I using confusion metrics, uh, to, uh, uh, to compare, to compare the data between the, uh, validation set that I, and the training set. So, The first one I say is, uh, uh, simplicity data between a training, artistic and, uh, validation.

Interviewee

I give the validation with the training to imbalance the training to not go overfitting, so the model not, uh, save the, uh, not, uh, save the information as it I needed to generalize on, on different kind of data or information. I split the data and then when I, uh, evaluated with the, with the test, uh, I, I use confusion metrics, uh, as, uh, popular, uh, to see what, uh, what the difference is between testing and training data.

Interviewee

How the model, how the model, uh, act or effect by the date. Yeah.

Interviewer 1

Perfect. Thanks. Um, Have you, sorry. Have you ever assessed the quality of a model prediction with some benchmark model?

Interviewee

Uh, like what?

Interviewer 1

Um, for example, if you have, um, you are doing an image classification problem, you compare your model to, uh, some other model deployed somewhere that, uh, awesome result that you want

Interviewee

to achieve.

Interviewee

Ah-huh. Ah-huh. Yeah. Uh, by, by the accuracy. I see. How, how, uh, how the, uh, how my model accuracy, uh, doing with, with the other and try to enhance it depending on, uh, uh, sometimes I use the pre-trained model and, uh, find its unit and, uh, hyper, uh, play with hyper parameter to get better. As much as possible.

Interviewer 1

Yeah. Yes. Thank you. Thank you. Um, have you ever assessed the quality of your model with the user of the system?

Interviewee

Uh, no. Uh, I never tried with user yet. Yeah.

Interviewee

Sorry, I

Interviewer 1

think I'm muted, so,

Interviewee

yeah. Yeah, I can.

Interviewer 1

Have you encountered any other quality issue during the evaluation of your

Interviewee

models? Uh, well, uh, depending on the, uh, I think. Uh, when, uh, I try to choose, uh, when I have a problem and I need to solve it in machine learning, I use several kind of machine learning algorithm to find which one.

Interviewee

Uh, so for best practice, I, uh, I have listed of, uh, algorithm and for ion classification, supervised, unsupervised, and trying to solve and see what, what is the bestor I'd make for it, uh, hyper parameter. Then, then, uh, then make, uh, make random, random indexes between the data to avoid, uh, biases and, and

feature.

Interviewee

So I have the list of, uh, uh, list of algorithm and choose dependent on the accuracy. So we have experiment as we, uh, we can see the ML process is about experiment the date and found which algorithm fit on that. Yeah.

Interviewer 1

Yes. Perfect. Thank you. Um, moving on to model deploy. Um, what are the challenges you have encountered during the deployment of a machine learning software system?

Interviewee

Uh, for my personal view, I, I made a model for detecting handwriting, but not in English, in Arabic, so it was a challenge. Uh, what I, uh, I take, uh, I take a data, set, big data. I think, uh, 2000 images. Uh, 20,000, sorry. Sorry. And, uh, and, uh, and train the model. And when I save the model for model evaluation, I am, uh, I work as a web development, uh, web developer.

Interviewee

So I use sensor loader js. To, to take the model. I trained it in, uh, Python and visualize it on the internet. And, uh, I uploaded in GitHub and my portfolio. So anyone can use it now. Yeah. Okay,

Interviewer 1

great. Thank you. Um, did you ever have a model that performed well locally but poorly once deployed? Uh, sorry. Uh, did you ever have a, a model that it worked well locally, but when you deployed it, it didn't it, did it not work well?

Interviewee

Uh uh, no. It's depend on data you have. Uh, as I think if, uh, if I train it on locally, it'll doing well. But I, when uploaded the data sup, uh, sorry, change, uh, but I am not using that. I'm not, yeah, I'm not faced that. Yeah. Yeah.

Interviewer 1

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

Interviewer 1
Uh,

Interviewee
sorry.

Interviewer 1
How do you ensure that the quality of a mo, like a model does not decrease over.

Interviewee
Yeah, of course it's increased, uh, because of the, because the hardware, the hardware enhancement and some algorithm. We have a newly experimented, uh, uh, our, our machine learning accuracy order on model will be increased, no decrease.

Interviewer 1
Okay. Also, why, why does it.

Interviewee
Because of hardware, uh, cause of GPU enhancement.

Interviewer 1
Oh, yeah. Okay. Yeah. You mean what Every time

Interviewee
you train? Yeah, every time you train. Sorry.

Interviewer 1
Okay. Okay. I see. Yeah. But, but let's say you deploy a model, um, on your website. Yeah. How do you ensure that the, the model stay good throughout?

Interviewer 1

Once it deployed?

Interviewee

Yeah, yeah. Uh, I use, uh, I use, uh, sometimes between time and another, I, I take it and see what is the result. By using some sensor sensor board, it'll track your, uh, your, uh, model activity by, by the losses, by the, and you can see it's like a chart and you can analyze.

Interviewer 1

Yes. Yes. I see. So you, you, yeah.

Interviewer 1

I see you monitor some of the monitoring. Yeah. The metrics.

Interviewee

Yeah.

Interviewer 1

Some metrics. Mm-hmm. . Um, have you encountered issue with data during the maintenance of a machine learning software system?

Interviewee

Uh, uh, uh, no. Like what, sorry? Like what? Give me.

Interviewer 1

Um, well sometime you may have issue with the, um, I forgot the name, concept, thrift, right?

Interviewer 1

Like the data, it does not predict. The relation between the Xs and the, the features and the label, the

Interviewee

Uhhuh, , Uhhuh, . Good. Uh, uh, that's, uh, before, uh, when, when I see like this problem I make, uh, uh, feature, uh, selection by using, uh, some, some algorithm. Uh, I think the best, uh, one is choosing the decision three when we have a classification.

Interviewee

So it's take the best value for the, for, uh, and, uh, So it's take, uh, if we, uh, when I use, uh, feature engineering, we, we have two problems when the missing value, and we have the how, how the columns are relative to each other. Uh, for the missing value, we, we using, uh, decision three and, uh, some linear, uh, some linear algorithm to fit in the well.

Interviewee

And, uh, and for. Sorry for fe uh, for, uh, feature, uh, feature, uh, for column, for column relationship, we use either correlation or, uh, some, uh, or some algorithm, uh, uh, with machine learning to find which one is better to choose the, uh, to choose, uh, to choose, sorry, to choose the column that, that, uh, affect only each other, not, uh, have, uh, related.

Interviewer 1

I see. Yeah. Yeah. Perfect. Thank you. Um,

Interviewer 1

as you encountered issue with the model during the maintenance of a machine learning software system, so it's a bit similar, but, uh, for example, you can, you could have model, model stainless or a model which has performance, which are unreliable between each time you train the model.

Interviewee

Oh, yeah. Uh,

Interviewee

let me remember. It's a lot. Uh, yes. Take

Interviewer 1

your time. Don't worry.

Interviewee

And, uh, and the, and the ma uh, sorry. You, you ask me in the maintenance of machine learning algorithm. Yeah, yeah. Uh, well, it's, uh, depend. Sometimes when I do it like machine learning algorithm, I and I, uh, upload it and the maintenance, I see how it's affect, how, how the data are going, uh, how the model, how the model.

Interviewee

And, uh, when I see some issues or, or something, I change the metrics. I change the metrics of the . Okay, I see. Perfect.

Interviewer 1

Yeah. All right. Uh, thank you. So I will list out a number of quality aspect for machine learning model, and you tell me if you have ever experie. Experience one of them, uh, in issue with one of them, right?

Interviewer 1

Uh, so fairness, robustness, explainability, scalability, privacy and data. Secur, uh, security.

Interviewee

Yeah. Uh, uh, scalability when, uh, yeah. And, uh, uh, yeah. Scalable.

Interviewer 1

Okay. Yeah. Uh, would you mind giving me an example where you had an issue with scalability?

Interviewee

Uh, when, uh, when we try to make a model for, uh, uh, data, uh, simple, uh, sort some sample of, uh, sample of dataset.

Interviewee

When we take it, uh, some sample, we, uh, as you know, the data are, uh, are coming from population and we take it as a sample. When we train this model on this sample, sometimes, uh, have, uh, issues because not cover all, all the attributes of the, the, the meaning problem. So we, we have to scale it more and take some data and, uh, and, and to perform very.

Interviewee
Okay, perfect.

Interviewer 1
Thank you. Uh, I have three question, uh, since we have a little bit of time, I'll ask you one question about data collection. Yeah. Um, have you ever used data that was generated by another system? So for example, let's say you are in, uh, um, imagine you are at, uh, trying to implement the model for a market like a supermarket where you can bike groceries.

Interviewer 1
uh, someone recorded transaction of the groceries, Uhhuh , if you use that data, it, it'll be data generated by a system, right? The, the, the system records the transaction. Uh, so my question is, have you ever used data that was generated by a

Interviewee
system Ah-huh by a system? I use, uh, a data generated by medical system.

Interviewee
I think for a patient for a, I think cancer thy thyroid. Yeah, I take the data from, uh, uh, it's, it's came from, uh, uh, as you say, uh, as I say, sorry, uh, medical, uh, system and take the data, make some algorithm to, to predict, uh, how, how it's, how it's behaved. Okay.

Interviewer 1
And, uh, so my usual question, what were some of the data qualities should you add with, with the data?

Interviewee

Uh, Collecting the data is more important thing, but, uh, the, the more, more important is how do you analyze it and give it to model? How can you clean it? Uh, we, as we can see, say, sorry, how we clean it and, uh, push it to the, and feed it to the model. This, the, the quality issues I faced a lot between collecting and, and.

Interviewer 1
Perfect. And, and what, so, so you mean cleaning the data? What did you do to clean the data?

Interviewee
Uh, uh, using, uh, ity data analysis for, uh, uh, and feature engineering for, uh, removing some missing value, uh, uh, sorry, not removing. Remove the perfect way we use, uh, machine learning to, to, to fill the. Value removing the, uh, duplicates value.

Interviewee
Some, some record have duplicates, sorry. And take, uh, the, the un uh, take out the unrelated, uh, column. Some column are not like, uh, name or some, or some, uh, or some column generated by system like Id, the ID of the patient.

Interviewer 1
I see. Perfect. All right. And I have, so I have two final questions for you. Um, in your opinion, what is the most pressing quality issue researchers should try to solve?

Interviewee
In my opinion, I think the, the algorithm of the model itself, because we, uh, as I mentioned before, we don't have. Any, um, certain information about this algorithm is for this problem? No. We, we have to experiment with the problem and, uh, by, by making model evaluation, we choose the model that we, uh, think it's better depending on the occurs.

Interviewer 1
Yeah. I see. So auto ML will be something, so putting more effort in auto ML will be interesting for you. Yeah. Like auto ML is automatically generating.

Interviewee

Yeah. Yeah. I, I hear. Yeah. Okay, good.

Interviewer 1

Perfect. Um, and do you have any other comment about the quality of machine learning software system? Uh, Okay. Perfect.

Interviewer 1

All right, so that's all the question we had for you today. Uh, so thanks a lot Interviewee to be there. It's really appreciated to, to spend a bit of your time with us. Thank you. Um, so we wish you, I'm sure it'll be useful and, uh, we wish you the best of luck with your future and thank you. All right. Have a good day.

Interviewer 1

Thank you for your time. Have a good time. Thank

Interviewee

you for your time. Have a good day. Bye. Thank you so much. All right, bye-bye. Bye. Hi. Hi.
