1:34

Alexey

**This week, we'll talk about responsible and Explainable AI. We have a special guest today, Supreet. Supreet is an assistant VP in data strategy at Morgan Stanley. She's also the founder of DataBuzz. So she will probably tell us a bit about what that is. She's a writer, she's a speaker, and she loves communicating what she knows on different platforms. It's my pleasure to have you today. Hi.**

2:11

Supreet

Thank you for having me on the show.

# Supreet’s background

2:14

Alexey

**Before we go into our main topic of explainable and responsible AI, let's start with your background. Can you tell us about your career journey so far?**

2:25

Supreet

Sure. In 2017, I came to the states to pursue my Master’s as a student from India, the capital city. I started my Master's of business and science – it’s a combination of MBA and MS in data science – a very unique degree. Then I started my career in January 2019 as a data science consultant. I was a data science and technology consultant for three years before I decided to switch gears a little bit and be more involved in the strategy as I realized the importance of that in my early career years.

After these few years of chatting with so many ambitious people who want to pivot in this technology and data science, I realized that there was something that I could do out there to help everyone and hence, I started the community, DataBuzz, to mentor people who want to pivot in this way. I've had a very unique career journey, I would say, from a non-traditional, non-engineering background pivoting into this field. That’s what the goal of DataBuzz is.

3:34

Alexey

**Well, that's interesting. We will probably spend a bit more time talking about that. I'm quite curious to know more about this. I'm also interested in your title, which is Assistant VP in Data Strategy. I'm really curious, what does it mean? What do you do in your day-to-day job?**

3:54

Supreet

Yeah, of course. Morgan Stanley's definitely a big firm. We have different divisions – we have asset management, investment banking, wealth management and so forth. I'm in the analytics and data team for wealth management, and the data strategy and products team basically owns different AI-driven products.

We are the product owners who ensure that the data strategy is streamlined. We also introduce new data ideas into the stream, and obviously create different AI products, help launch different AI products in Morgan Stanley. We work very closely with data science to make everything happen.

4:34

Alexey

**So you're more on the analytical side of the data.**

3:39

Supreet

Definitely more on the analytical side, yeah.

# Responsible AI

4:43

Alexey

**Today’s topic is responsible AI. So what is responsible AI? How can AI be responsible?**

4:52

Supreet

Yeah. It's definitely a very interesting field and it's a very budding field, I would say. I've been in this regulated environment. I've been in finance for three plus years now, so I have definitely appreciated responsible AI and trustworthy AI. As the name suggests, it's all about developing algorithms and processes so that you can empower your employees and users or customers.

There have been numerous studies out there that say that if your consumers and customers trust you with the outcome, they'll invest in you. It's as simple as that. It's basically a collaborative way of working with other stakeholders, with keeping your end users informed and being able to integrate everyone's feedback, ultimately making this a very collaborative process.

5:51

Alexey

**What does that actually mean? Does it mean that if they trust us then we're responsible because we don't want to violate this trust? What does it mean to be responsible here? We're responsible for data, for the decision that our machine learning and the AI systems make, right? And we don't want to lose this trust that people put in us? Right?**

6:14

Supreet

Exactly. You basically have the right tools in place so that all your stakeholders and users feel confident about the decisions. So if anyone asks you, “How did you arrive at this outcome?” You have some tool – you have something to show that, “Okay, this is the step. This is what my algorithm followed and we were able to arrive at this decision.” So I would say that it’s tools and frameworks to empower you.

# Example of explainable AI

6:42

Alexey

**I'm just trying to think of an example. Can you maybe use some examples from your work, or maybe something that you can talk about where this is important? You said “We want to feel confident in this prediction, and we want to be able to explain how they happen, so that we have this trust.” Can you give us an example of that?**

7:06

Supreet

Of course. A few months back, an article came out where there was a husband and a wife – both of them are eligible for a credit card with the same limit, but the wife ended up receiving a lower limit than the husband. There was a big lawsuit against the company, because she felt that somewhere, the AI was biased towards her just because she was a female and maybe the algorithm ultimately assumed that she had a lower income and gave her a lower limit credit card.

So this is where, if you have the process in place, you are able to justify why the wife got that limit, and why the husband got a certain limit. Even before that, you are able to check – did your data actually have bias in it? And you can curb that so that you can produce a fair decision. There are two sides of the coin, I would say. The first is your data side, which has to be checked in terms of accuracy and fairness. Then there are your model’s predictions. Because if your data is biased, there’s no way that you're going to get unbiased predictions.

# Responsible AI vs explainable AI

8:20

Alexey

**What is the relationship between responsible AI and explainable AI? From what I hear now – if we want to feel confident in the predictions, we need to be able to explain them. Does this mean that responsible AI is explainable AI? Or what's the connection there?**

8:38

Supreet

Yeah, I think I will definitely deep dive into explainable AI and talk more about that. Explainable AI is more like a post mortem report, I would say. [chuckles] An incident has happened and now “Okay, what do you need to do? Okay, let's build a simple algorithm and justify that it’s good.” Responsible AI is more like, “Okay, we have these frameworks in place and your starting point or whenever you are building an algorithm or a project, you have all of these tools and frameworks in mind. For the incident to not happen anymore, you have these tools in place.” So it's more of a mindset. Explainable AI just gives you the authority, or I would say the tools, to build those frameworks. I can deep dive into some of the practical techniques that you can use to implement explainable AI in your day-to-day practice.

9:37

Alexey

**Okay. So what you're saying is that we should try to prevent these situations, like in the example of a husband and wife getting different credit proposals. Ideally, we should not let this even happen. We should catch all the bias we have in our data and try to mitigate it before we train our credit scoring system. When the credit scoring system is live, then we should be able to justify it. So if the wife, in this case, comes to the financial institution and asks, “Why did you give me a lower loan?” The bank would just say, “Okay. This, this, and this are the reasons.” But if we don't have that, then we have a problem. Right?**

10:29

Supreet

[laughs] Yeah.

# Explainable AI tools and frameworks (glass box approach)

10:30

Alexey

**You mentioned tools and you mentioned a framework. So how do we do this?**

10:36

Supreet

This is where I think it's the right point to explain some of the explainable AI techniques. So what is Explainable AI, essentially? It's a framework that can be integrated with your existing machine learning models, so that you can understand the output of your AI or ML algorithms. As I said, this is not only used to explain the results behind a machine learning algorithm, but you can use it to receive feedback so that you can retrain your model, you can use it to detect bias in your data. It's called the glass box approach. As they say, “Oh, AI/machine learning is a black box.” But here, you're giving transparency and it's called the glass box.

11:21

Alexey

**Glass box. I heard “white box” before, but I was always thinking and wondering, “Is white really transparent?” It’s just as opaque as black, right? [chuckles] Glass box makes much more sense.**

11:36

Supreet

[laughs] Yeah. I will start with the data level. First, when we talk about the data level, we talk about fairness and bias testing. There, you have a few data quality checks that I feel every data scientist does. They do some sort of exploratory analysis, dig into the data, see what they have. We can do a few checks there as well. One of them is skewness – you could check what your data looks like and it could be that you are missing out on one population or the other, then your data is highly skewed. The other could be missing data.

If you have too much missing data, it's important to analyze what you are missing. You might be able to talk to business stakeholders and get a sense of “Are we missing out on an entire population?” So these are some simple checks that can be done on the data side to ensure that your data is not biased and you're covering a wide array of populations. This is one of them. Then on the model side, you have different techniques. But I'm just going to pause to see if there are any questions or any comments on the data side?

# Checking for bias in data and handling personal data

12:48

Alexey

**Well, there is a question, “What tools do we use to check if there is any bias in the data?” I think this is related to what you were talking about. This is before we train a model, right? You talked about EDA (exploratory data analysis) , you talked about skewness analysis, and missing data. Are there any particular tools and techniques that we can use to check this?**

13:16

Supreet

This is more on the EDA side. You're not doing anything – you're just exploring the data, honestly. You're just analyzing “What's out there? What does my input look like?” Then when you actually dig deep into this, it's kind of more of like a conscious check plus a technical check. But here is where you will see, “What is happening? What is happening exactly?”

Then I would also say that there are these bias checks that you can do. Now, bias is a very subjective term. Bias can occur in a lot of shapes and forms. There’s a book called Trustworthy AI – I don't know how many of you have read it – where the author talks about different types of bias. She says that it can be like a gambler's fallacy, which is that the probability of a random event that occurs in the future is influenced by a past event. And that is what we assume when we are building AI models most of the time – that *my history* is the correct representation of *my future*. But that is not always true and that's where you need the human touch to check for such biases in your data and in your model.

14:39

Alexey

**So basically you need to, as a human – as an analyst or as a data scientist – you need to get your dataset from your database, CSV file, whatever, and just spend enough time trying to understand what's happening there. Right? Should we watch out for anything particular? Let's say, if we see columns like age and gender, should we already think “red flag”? Should we do something about these columns? Or what does it usually look like?**

15:14

Supreet

Yeah. I would actually say it depends, because age and gender and those things are very PII information – sensitive information. Most organizations wouldn't even let you touch [laughs] those attributes, especially if you're in a regulated organization. So I’m not sure if you would, first of all, have access to that kind of information to even make this judgment.

But there could be other factors – it could be the income of the person, or other factors about that person of whatever it is you are doing, that can lead you to make some assumptions, predictions, decision, which you would use to check back with your business stakeholders or some subject matter expert to see if they're even valid or needed.

16:05

Alexey

**Okay. So you think these financial institutions that we talked about – banks – don't have access to age and gender?**

16:14

Supreet

I think it depends on which organization we are talking about. Everyone works in a different way. It's kind of a recommendation – How are you handling that PII information? In the pharma world, we never had access to all of this data. [laughs] Or even if we had it, we wouldn't use it – we’d mask it out.

16:35

Alexey

**But in pharma, I guess it's important, right? Because the way drugs work on people depends on their age and depends on gender. There it is a justified use case. While when building a credit risk profile, maybe it’s less so.**

16:58

Supreet

Exactly. I don’t think it’s even a healthcare or a financial institution question – it's like a general industry thing. If in your use case, this type of data is not required, you wouldn't use sensitive information. Again, this is also responsible AI, because you're handling the data of a customer responsibly. Yeah, definitely – we do our due diligence in determining “Do we even need this data in the first place?”

# Understanding whether your company needs certain type of data

17:20

Alexey

**And how do we do this? Is it usually done by this… we talked about data checks – to discuss this, should we go into the model checks? To understand whether we even need this data?**

17:35

Supreet

Yeah, exactly. When you're determining whatever use case you want to solve, you will definitely have a detailed discussion with your product managers, business stakeholders, and that is where you will determine what data you need. And if it's sensitive information and if you can mask it out, that is where it will get out of your data pipeline for your AI models, so you can use other datasets.

18:02

Alexey

**Would it be wise to completely just throw away gender data, let’s say if it’s PII?**

18:07

Supreet

[chuckles] Yeah, it'll again depend on the use case. A very typical answer, but I don't think I can answer it any better. It, again, depends on your use case, what your SMEs say? What is your compliance? What is your Model Review Committee said before you do anything else?

# Data quality checks and automation

18:27

Alexey

**How much of this can be automated? At least when we talk about detecting biases in data *before* we come to modeling?**

18:37

Supreet

Yeah, I think for all the missing data, the data skewness, you can always develop DQ (data quality) tools that can do all of these simple tests for you and tell you – they can actually give you those alarms. Then you can go there and investigate – if there is an alarm and you feel that an investigation is needed. Most of this can be automated.

19:03

Alexey

**So what about this model part?**

19:08

Supreet

Yes. For models, also… I think a few years back, we didn't have as many open source technologies. But now we have a lot of those. Google, IBM, everyone has such good open source technologies that can easily be fetched into your Python pipeline or your PySpark – whichever tool you're using. There is a what-if tool by Google – it's almost an interactive tool – where you can understand the data. It gives you an interface of the dashboard, so you can basically fetch all your parameters, see what's the workings behind a machine learning model, especially if you're using TensorFlow, which is again, made by Google. So a what-if tool is a great tool to use, then we have other tools as well. We have Skater, which is an open source Python library. That was again designed so that you can see some of the module functioning. [cross-talk]

20:03

Alexey

**What is it called? I didn't hear the name.**

20:10

Supreet

Skater. S K A T E R.

20:11

Alexey

**Skater. Okay.**

20:14

Supreet

Yeah, that is another tool that helps you see the functioning behind the machine learning module. Now, the trick is, obviously, how you will integrate in your use case will, again, depend on what you're trying to achieve. You might not find that all of these tools can be integrated with your use case. There, it's more of an exploration, again. Then there's another one – AIX 360 by IBM. That is another amazing open source toolkit, which you can use to comprehend your predictions and your machine learning model.

20:55

Alexey

**So this one is… I'm just trying to understand. So “what-if” if I understood it correctly, you can see this as another exploratory data analysis tool, but smarter than the usual analytics, right? So it's smarter than just counting missing values. Then you can see… What kind of questions can you ask it? “What if I remove this column?” Or “What if I had this data?” What kind of ‘what-if’ questions can you ask?**

21:27

Supreet

Again, it's more like if you want to also see the importance of some of the features – if you want to see, “Okay, what if I remove this feature? How will my model get impacted?” If you just want to see what your data looks like, as I said, like the skewness and everything, it will easily allow you to do a plot and that will be very dynamic. You can do it over time, over months, or years, whatever use case you have in hand – so all of those things. And if you were using TensorFlow, then how did you arrive at that prediction? What if you remove the feature? How will it impact your model? How will it impact your accuracy? All of those things.

22:07

Alexey

**And Skater? Does it do something similar to that?**

22:12

Supreet

It’s similar, but it’s not just [audio cuts out] more broad. [chuckles] It's more on other models, if you want to use it. Basically, you can do it based on your complete data set. Explainable AI, whenever you are doing – you're using any framework. The idea of that is, if I were to apply this on a larger dataset, what will the impact look like? But whenever you build an explainable AI technique, you basically take a small dataset. With that, obviously, there's an assumption that this is a true representation of my bigger dataset. And then you build around that. Skater will facilitate some of those techniques for you.

22:56

Alexey

**The last one you mentioned is AIX 360 from IBM. My understanding was that it's used for interpreting the output – the predictions of the model. Or not only that?**

23:11

Supreet

Actually, all of these can be used for that. I mean, ultimately, you're using all of this. It's just that if you're not using TensorFlow, if you're using other open source packages, they'll help you facilitate that.

23:24

Alexey

**I've heard about tools like LIME and Shapley Values. Are these somehow related to that too?**

23:32

Supreet

The ones that I just spoke about are more open source Python packages, I would say. Now, if you actually want to build things from scratch – the way you will actually build any algorithm. That's where LIME comes into play. LIME is a combination of different things. You can do a linear regression model, or you can do a decision tree model. That is basically building a LIME model.

Actually, you might be doing a neural network model, but now you just want to be able to interpret that. So you, again, take a small data set and build a decision tree on that to come close to what the interpretations would look like, so that you're able to explain that to your stakeholders. Yes, LIME is very similar to that.

# Responsibility vs profitability

24:22

Alexey

**I'm quite interested in this. Let's say there are some features that we shouldn't use because they're PII (personally identifiable information) like gender and age. But in some cases, I kind of wish we could use them. For example, in an advertisement, maybe there is a product that would be more appealing to females than males.**

**If I decide to be responsible and trustworthy, then I should remove this feature, right? But I know that this feature is actually important if I want to be profitable. Then I have a dilemma. Do I want to be profitable? Do I want to make money? Or do I want to be responsible and remove bias from data? Do I want to get rich by exploiting these biases or not? So how can we still be profitable, but without exploiting all these biases?**

25:32

Supreet

I don't think all these decisions are made by one person. And they're definitely not made by data science practitioners. That is what responsible AI is. It's a very collaborative process. That's why they say it – because when you have 10 different people in a room, discussing such use cases and have 10 different perspectives, I would say that you will be able to meet halfway. That is where you will be able to decide the priorities of what you really want to do and how you can achieve it. That is exactly what we aim to do. It's beyond these techniques that I just mentioned. [chuckles]

26:12

Alexey

**But then I guess for this one, let's say if there are 10 people who have different perspectives, they should be different people. It's not like all data scientists are all white males. Otherwise they probably would think, “Okay, there is nothing wrong with using gender. Let's just use it.”**

26:32

Supreet

No, no, no, definitely. I meant different representations from different groups. Because data scientists think differently and project managers will think differently. Now, you add compliance people, they'll think differently. [laughs] That's what I meant and that is why like launching an AI product and doing all of this. Responsibility is definitely a journey.

26:56

Alexey

**The reason I’m mentioning it is because, in many courses, and also the courses we do at DataTalks.Club, there are datasets – there are some problems – where we do actually use age and gender. I think, actually, in the course we have right now there are two projects in which we use age and gender. You might think as a data scientist, “There is nothing wrong. What’s there to talk about here? Let's just use this.” Right? [chuckles]**

27:24

Supreet

Yeah, exactly. And we would never know. For us, it's all about features. “This is important. That's what our algorithm is saying.” We don't really think in that direction. [laughs] But yeah, it's important.

27:38

Alexey

**So what kind of people do we need to have in this room to be able to have these fruitful discussions? You said that we need, perhaps, data scientists, analysts, and people from compliance. Who else should we have?**

27:52

Supreet

You should also have subject matter experts who understand the data better. You should have your senior leadership, ultimately, as well. I think, obviously, it's not going to be easy [chuckles] to get an answer in one meeting, like “Oh, yeah. We’ve decided to come to a conclusion.” I think, at that time, you'll be thinking about some innovative workarounds. “Okay, let's use this one feature over this feature, and let's see how it performs. Let's drop *this* feature.” But at the end of the day, it's possible that the use case that you have in hand, like gender is one of the most important features. Then, again, it's a company's decision to make.

28:38

Alexey

**Then I guess – who is making this decision? Is it senior leadership, or everyone?**

28:46

Supreet

It’s collaborative.

28:48

Alexey

**But then if everyone is responsible, then nobody is, right? [chuckles**

28:52

Supreet

[laughs] I don't think so. I think they’re very productive and, obviously, everyone's responsible. But at the same time, you're in a business and you want to make money. [laughs]

29:02

Alexey

**But then, what happens at the end, I guess, is you come up with some sort of terms of service and then you say, “Okay. This is the data we collect. We collect gender.” And if you don't agree with this, you don't use the app? Or how does it usually work?**

29:15

Supreet

Yeah, I mean, it's also about the way we do those declarations and the way we sign all the terms and agreements, where somewhere the customer might give you the consent that “Okay, yeah. Just use my data.” And some customers might not give you their consent. So that’s also very competent.

29:37

Alexey

**But I'm still thinking about that example, where the wife and the husband – perhaps both of them gave consent to using their gender information. Then at the end, this is what happens. The company still got sued, even though they probably had paperwork in place. Right?**

29:55

Supreet

I mean, I'm not sure. I read it in the newspaper, so I don't know whether the company was able to justify that decision or not. We get the news from a third party, so you don't really know the facts. But that's what kind of made me think. Was the company able to justify? If the consumer is questioning you, are you able to justify what led to this difference in your decision? Can you say that it was beyond gender? Can you say it was not biased? And you can only say that if you've already done your due diligence before, in order to see that there's no bias in your data set.

30:35

Alexey

**And if you found out that there is actually bias in your dataset, and somebody's already suing you, it’s too late. Right?**

30:43

Supreet

[laughs] Exactly. It’s already too late.

# The human touch in AI

30:48

Alexey

**[laughs] I see. Per my understanding, we should analyze our data – we should have a human who actually goes through the dataset. Some things could be automated. But then, at the end, a human should actually go through this and do some analytics. After that, there are techniques that can automate this, like all these what-if tools and similar ones like Skater. That should be used by analysts and data scientists, I guess, to make sure that this actually doesn't happen. Also, at the end, we want to be profitable, but at the same time, we want to be trustworthy and responsible. Therefore, we should have this meeting with people from compliance, from data analytics, subject matter experts, from senior leadership, who get together and discuss the case. Am I missing anything?**

31:56

Supreet

No, no. We also highlighted the challenges on the way. This is one of the challenges, as you said – profitability versus being responsible. And then we have accuracy versus interpretability. Because for all of your complex models, you still might not be able to make an explainable AI model. All of those, I feel, are still the challenges and pitfalls of explainable AI and responsible AI?

# The trade-off between model complexity and explainability

32:29

Alexey

**Actually, we have a question from Shivam that is exactly about that. The question is, “How to manage the trade-off between model complexity and explainability? Complex models do not necessarily have good explainability, so how do we manage that?”**

32:45

Supreet

Exactly. Again, as I said, that's one of the challenges. In some use cases, it's a trade-off. As data scientists, we face other trade-offs – we have the bias versus variance trade-off. So this is one of those things where you need to, again, analyze what you're trying to achieve, who you're trying to target, and what your end goal is. What do you want to achieve? What do your business stakeholders want to achieve? And if we were to take a step back on accuracy, will that help us in the long term? Will our consumers trust us more? At the end of the day, you're also trying to build a brand – it might be a bank or it might be healthcare or whatever. You want people to indulge with you in some form or another. So yeah, it's one of those crossroads.

33:40

Alexey

**I guess, if you make a decision to prefer accuracy versus explainability, you should be prepared that eventually, somebody from senior management (like Mark Zuckerberg) will have to talk in front of Congress and explain things. [chuckles] If you don't want this to happen, then maybe you don’t… [laughs]**

34:03

Supreet

[laughs] I hope this doesn't happen, but yes. It's not that every time you will be choosing “Okay, let me make the most interpretable models.” Every time you encounter a use case, your decision making will be different. And that's why it's important to not make all of these decisions [inaudible].

34:26

Alexey

**I imagine that if you, let’s say, use linear or logistic regression for everything, then it will be not so biased and very explainable, but then the accuracy could suffer. But then if you go with deep learning, then it's the opposite side of the spectrum – you have little explainability, but a lot of accuracy.**

**So what I understand is that you just need to keep the long-term goal of the company in mind and ask yourself, “Are we ready to sacrifice our image with some extra profit? If people start talking bad about us, or we're an evil corporation – it's fine to store all the data about you and then sell it to others.” [chuckles] Not pointing fingers. [Supreet laughs]**

# Is completely automated AI out of the question?

35:28

Alexey

**There is a question from Raquel, “What does ‘you need a human touch’ mean?” I think this is related to our discussion, where the first step is always a human analyzing the data. Then the question goes on “Does this mean that responsible AI and checkpoints and alarms cannot be automated?” Maybe we can think of what can be automated and what cannot.**

35:57

Supreet

Yeah, as I said, all your data quality checks – you can be so creative. You can write pieces of codes and you can do all the checks in the world. But at the end of the day, you just cannot be like, “Oh, my DQ checks passed. I don't really need to look at it and see what's happening.” As your data keeps evolving, your tools and even all your production models need to evolve as well. Often, as a data scientist, you hear, “Oh, we have drift in our model.” And drift can also be in this form. It can be in the form of bias as well. You didn't have biased data before, and now you have biased data.

If you can build adoptable techniques, definitely do that. But at the end of the day, there has to be someone who can go in, analyze all of those – you obviously don’t have to go row by row [chuckles] but in a cumulative fashion, see what's happening with your data and your model performance. That's what I mean by “human touch”. At the end of the day, it's a human who's gonna write the code as well, to make that happen.

37:10

Alexey

**So a human automates it, right? We automate, but it's the human who is doing the automation and they need to put some thought in this process. Not just blindly take a tool from Google (or who knows what company) and just put it there and forget about it. Right?**

37:30

Supreet

Exactly.

# Detecting model drift and overfitting

37:31

Alexey

**To your point about drift, we should monitor for bias. There is a comment from Abhishek, “If the DataTalks.Club model sees more males attending the events, females will not get recommended the event and this can introduce a feedback loop, where the next time, the model will recommend the events only to males and ignore females.” This is something that happens gradually. At the beginning maybe it's equal, but then there is more and more and more and more – a feedback loop. And then at the end, it's only males attending the events. Are there tools that can help detect this kind of drift, or at least this kind of bias, that gets into your model and into your data because of these feedback loops?**

38:28

Supreet

Yeah. I feel like if your model is in production, you are already using some sort of tool to monitor your model’s performance. But you might just be measuring your performance in terms of accuracy. Now, you also need to build some sort of monitoring in terms of the population samples – if it's a male or a female, if you already have a unique identifier in your data set, how is your population drafting? Basically, all the DQ checks that you did during your input phase, you need to redo them even after every time you maybe launch a version of the product or every time you're monitoring your product.

It's basically about integrating those two pieces. I don't think there's just one tool that will be able to solve your problem. I feel like it will vary from problem to problem and then you will get creative when building. But basic statistics, I think, all of us can do. Mean, median, mode – basic statistics to find out, “Okay, what is the mean of my data? How is my population? What does the sample size look like?” All of those will also give you at least *a* view of what's happening. And then you can use that.

39:40

Alexey

**I imagine if we only monitor for accuracy, then in this example of this feedback loop, the model will maybe become 100% accurate.**

39:50

Supreet

Again, overfitting and that’s an alarm.

39:53

Alexey

**Yeah, yeah – exactly. And then from the model performance point of view, it's kind of doing well, right? Everyone who comes enjoys the event – maybe it has high precision, but the recall is bad. We are missing out on other people who might enjoy it, but simply because the model overfit, they don't get to see this in the recommendation kit. Then the way to monitor this is by checking these basic stats, like looking at some demographic information.**

40:28

Supreet

Demographic information, yes. And the way you actually do it for drift, also – there are these KS tools. I mean, there are a lot of tools in the market. Even AWS, if you're using it in your notebook instance, you have these tools where you can include all these basic statistics. So whatever is the identifier for you. That'll, again, vary for your data. So you do need to see how that is performing over time – the way you will monitor the accuracy, you can even monitor your data.

41:03

Alexey

**We probably covered it, but there is another question from Raquel, which is, “If we're using data to train our models and identify our checkpoints and alarm, how is the data protected from human bias?” I think we answered that – correct me if I'm wrong. We answered that by saying that we should monitor this demographic information. So we should take a population sample. Or maybe if we're not predicting for people, maybe there are some other indicators that we can monitor. It probably varies case by case, but we should have a metric that probably indicates some sort of diversity in the sample. Right?**

41:46

Supreet

Again, that human touch part comes in. When someone is analyzing those cumulative statistics and looking at that data, they will be able to see that if there's a pattern, the next time they're monitoring it, they'll easily be able to see it. The other thing that's also important is to check the source of your data. Sometimes the source that you are doing – I know that in some of the use cases, in some of the organizations, they collect all of this feedback from humans.

Basically, they do surveys and they collect that data. And they do it over a larger period of time and then use that data to build the model. Now, with survey data, what did your population look like? From that data, only you can start detecting the bias – the way you’re even collecting the data can be questionable. So check the source as well.

# How Supreet became interested in explainable AI

42:39

Alexey

**Just curious, how did you become interested in this topic? Like was it something you studied and then you became fascinated? Or is it something through practice that made you realize that this is a super important topic? How did it happen for you?**

42:51

Supreet

Yeah, it just happened because I was in two of the most regulated industries and everything was on high alarm. You have to take every step in a very thoughtful way because it impacts the lives of people and it can lead to lawsuits, easily [chuckles] I would say. So that is where it was kind of my observation and my reading about how important it is to be cautious and responsible and how companies are evolving. And that's why it became a topic of my interest.

43:28

Alexey

**So for areas like finance, insurance, health care, etc. this is especially important, right?**

43:36

Supreet

Exactly, exactly. Actually, anywhere where you were dealing with sensitive consumer information, which has been labeled as PII – it kind of becomes important. But now I would say that it's fanning out to all the industries because as people get more conscious of what they are getting, and people are very brand conscious for everything, even for the clothes we buy. We're like, “Okay, how are they treating animals? And what are they doing?” And so, people are becoming more brand conscious now.

# Trustworthy AI

44:07

Alexey

**Speaking again of age and gender, if you think about this, they are not really PII in the sense that, if I tell you “This person is a female, who is between 30 and 40 years of age,” you will not be able to find where this person lives, right? In this sense, it's not really personal identifiable information.**

44:40

Supreet

Yeah, it's again, subjective. I don't know if you've heard about this – there's a tool by Amazon, I think. They created this HR tool to screen their applicants. There, all of this information is easily available because on the resume, people will write about their date of birth and everything.

45:00

Alexey

**They shouldn't, probably. [chuckles]**

45:02

Supreet

Yeah. [laughs] They observed that for all the physically demanding jobs, the algorithm was just selecting males. It was ignoring females. Again, the bias – because data was so biased. Historical data was all males, so it assumed. In some of such use cases, you require all of those things, because age is an appropriate factor there, gender might be an appropriate factor there. So yeah.

45:30

Alexey

**Did you know how Amazon solved this problem? What did they do?**

45:35

Supreet

No. It was a tool that was developed by them, but then it was being used by other CEOs. It’s in the book that I was referring to – Trustworthy AI –where she kind of talks about this dilemma of CEOs trying to select candidates for this role. How they detected it is because she was like, “Why are you only giving me males to interview? Where are the females? For the past five months, you haven't got a single female candidate?" So again, the human touch – the conscious – someone was able to question the decisions and they were able to see how the model is biased. It was kind of like an ATS tracker that they used. But I'm not sure if it was… don't quote me on this if it was by Amazon or some other company. [chuckles]

46:23

Alexey

**Okay. I’ve seen that mentioned as well. Something like Fair Learn. You talked about this AIF 360. Abishek mentioned another tool – Fair Learn. Do you know what this tool is?**

46:40

Supreet

I probably haven't used it, but I'll check it out. What is it?

46:43

Alexey

**Fair learn? I guess it's like Scikit Learn, but instead of “Scikit” it's “Fair”.**

46:47

Supreet

Oh, okay. Nice.

# Reliability vs fairness

46:51

Alexey

**I assume it's something similar. I personally haven't used this tool. It’s probably something similar to Scikit Learn. At least the name suggests that it's probably something for fairness. Since we talked about fairness – I think we spent quite a lot of time talking about fairness –how is this related to reliable machine learning? Does a reliable model have to always be fair? And the other way around – if a model is fair, is it always reliable?**

47:24

Supreet

That's kind of a very thoughtful question. “Do all of my models need to be fair?” Comes down to “What is the use case I’m trying to solve? Am I trying to send someone personalized recommendations? Do I want to exclude a population?” Excluding a population also has some effects, right? Because you might be excluding some of the potential customers that were ready to engage with you, but just because your model was not fair enough, you just secluded a population and you secluded some potential customers.

So it's about that. The other thing is, again, it's also about the happiness of your users. Right? In the same family, there might be different people indulging with your product. One is receiving something the other is not. Obviously, everyone's not gonna put a lawsuit against you. But it's unhappy customers.

48:21

Alexey

**Some might. So you don't want to risk that, right? Abishek is writing a lot of comments and is probably also into this topic. He wrote “There is a cool paper titled ‘What would an Avenger do?’” Do you know this paper?**

48:40

Supreet

No, I haven't read that paper.

48:42

Alexey

**I'm really curious what it is. But maybe we can check it after the podcast and see what it is about. There are people saying that this is a nice paper – the title was amazing. I hope it's not a Rickroll. [laughs] Recently somebody sent me a Rickroll link, and I got Rickrolled. [chuckles]**

# Bias indicators

49:16

Alexey

**Okay. We have quite a few questions. So maybe we can cover them too. “Do you know if there are any synthetic indicators for bias in data similar to RMSE, which is an indicator of accuracy for models?” Is there a number that can just tell us, “Hey! Your model seems biased. Go check it.”?**

49:40

Supreet

I mean, I can tell you a number, but I feel like even if I was working on 10 use cases, that number will be different for all the 10 use cases. [chuckles] It will still depend, because you're obviously not using one dataset to build the model. You're probably using 15, especially if you're building a production AI model. So I don't think that there's any number like that. Then again, there are baselines that you can set, based on your company's data. If it goes beyond this baseline, it's an alarm, based on what you observe over time. But I don't think you can set up a static number. Again, you might be getting into the same situation.

# The future of explainable AI

50:17

Alexey

**So then you still need to have this human touch. A person needs to say, “Hey, why are you bringing me all the male candidates? Where are the females?” Right? So somebody still needs to raise this question and say, “Hey, what's going on?” So it's not possible to remove the human touch at all, right? The human always has to be in this process.**

50:42

Supreet

I mean, we don't know where the technology is heading, but…

50:44

Alexey

**But right now. [Supreet laughs approvingly] There is another point that in the age of AutoML – when you just get the data you throw it in a black box and then it gives back a good model and then you can just go and use it. When it becomes so simple to train a model, even for people who are maybe not data scientists, how can we still be cautious about this and not completely put a lot of trust in the model?**

51:21

Supreet

Again, I think even interpretability of all these explainable AI tools – that might not be a priority of everyone. It's a priority of *some* industries. So again, what's your company's goal? You might be just fine with developing one of the most wonderful and accurate AI algorithms and you just might be happy with that. You might not care about interpretability and that's where it comes. This versus that. It's an answer that you have to do. But I feel even in AutoML, you might give your input data, but you still have some power over the data. You can still explore that and see what's there before blindly just inserting it.

# About DataBuzz

52:08

Alexey

**Can you tell us about the DataBuzz community? What is it? How did you start?**

52:16

Supreet

DataBuzz . As I said, I am from a non-computer science background and I have had some unique challenges to get into this field – being from a non coding background, being able to justify that I can be a data scientist, and exploring other realms of data science. For people it might be, “Oh, there's a data scientist.” No – “There's a data strategist. There's an AI consultant. There's an MLOps engineer.” There are so many things you can do with your unique skill set, especially if you are building from a business and a business plus a science field, there is something unique that you have to offer.

DataBuzz – its goal is to mentor those people and engage those people. There's so much to learn. I am an avid reader, so whenever I see something cool, I try to post it on my page so that my community can benefit. Any cool resource that will help you upskill is posted on my LinkedIn page. So you can follow that and I also provide one-on-one consultations for free and I connect you with my large network of people that I have from the past five years. They can help you to pivot where you would like to pivot.

53:31

Alexey

**You've been doing this for five years. That's amazing.**

53:35

Supreet

No. I was doing it unofficially for a lot of years. But now, I started this community at the start of this year. Because I felt “Okay, now I can officially do this.” [chuckles]

# The diversity of data science roles

53:50

Alexey

**You said that it's not just data scientists – there are so many data roles. Quite often, I also get requests like, “Okay, I'm doing X. How can I become a data scientist?” And this X usually varies from… I don't know – baristas to software engineers. So first, what you need to explain to people is that it's not just data science. My question is – what is out there in addition to data science? You mentioned a few rolls, but what are the typical roles that people can look at?**

54:29

Supreet

I think first of all, the question to ask is “What is my skill set and what do I want to do with it?” Do you actually want to be building models? Some people are learning to research so these people are becoming machine learning engineers, or they're called research scientists or applied scientists, in the tech world. These are people who actually research the algorithms. Some people are interested in the aftermath of productionizing the model – the MLOps engineer. Those might be software developers who now want to pivot into the ML field – or DevOps engineers who want to pivot.

Now the other one is the people who want to relate the business to the AI. These are the AI consultants, who actually search all of the AI use cases and then pitch clients. Here, you typically have a consulting relationship with your customers. And then you have these data analysts – you have business analysts. Even those are a kind of data scientist, because now you're doing all the project management and you're doing the analytics and reporting for your data science results. And even before the data science results. Then you have data strategists [chuckles] who are strategizing on what data you need. All of these.

55:41

Alexey

**For example, a project manager or somebody who works as a project manager in some traditional industry, like manufacturing or anything. How can they select what works for them? It doesn't have to be a project manager, but usually, every person has unique skills. How do you understand this, if all you know is that “data scientist is like the sexiest job” and you might not know how about others?**

56:13

Supreet

Well, that's a tough question. Even when I was trying to decide my career, I think I have spoken to like 100+ people. I did coffee chats with people from different industries, different domains, and tried to explore “What do they actually do?” You cannot just read the job description and know, “This is where I fit.” I feel like all these networking events, conferences, coffee chats, and engaging in such communities will probably lead you where you want to go.

# Ethics in data science

56:44

Alexey

**From all these roles, does everyone need to know and care about responsible AI and explainable AI?**

56:53

Supreet

I don't think so. [laughs] I feel like all of these bring something unique to the table. And again, if your role does care about you, you do need to add that touch. But it will, again, depend on the organization. You might be in a startup who might say, “Eh, no. Just care about the profit. Just bring me the most profitable AI algorithm.” It depends. [cross-talk]

57:16

Alexey

**But then I, as a data scientist, probably have some moral duty to tell the CTO, “Hey, wait a minute. This is not how it's done. You should think about our reputation.”**

57:31

Supreet

It also depends on where you are in the stage of your career. Are you an influencer? Are you a decision maker in the decision making process? Or are you just the person who executes whatever is told to you? If someone's actually starting in the data science domain, it might be hard for them to convince people what needs to be done. Once you have established yourself, then it is your moral responsibility, if you’ve been in the industry for three or four years, to do everything responsibly.

57:59

Alexey

**At least everyone who gets into the data world should perhaps study ethics or something like this.**

58:07

Supreet

Yeah, but this is such a budding field that so many people don't… even though I don't build explainable AI models, but I feel like, as an ethical person, it's my responsibility to understand what I'm doing and how I'm doing it.

58:25

Alexey

**Actually, at work, if I want to have access to data – I, as a data scientist, have to have access to data – I need to go through special training. This training tells me what is good and what is not. I think more and more companies are actually implementing this. Of course, I can use my common sense. When there are quizzes and when I need to answer these quizzes, most of the time I just use my common sense. And more often than not, the answer is correct. But it's still a good idea to tell people, “You know what? This is how it's possible to do wrong with data.” All the data people should probably do this sort of thing.**

59:10

Supreet

Yeah, definitely. I think even when you join a firm, you go through some tricky things, like “What do you do in case of emergency?” And I feel like this is one of the things – when you have such data, how do you deal with this and become a responsible individual?

# Conclusion

59:27

Alexey

**I think that the time is up. So thanks for joining us. Maybe before we wrap up, is there anything you want to mention that maybe you forgot?**

59:40

Supreet

I just want to thank everyone for joining and for being so interactive. If there was any question that was not unheard, or if there's something that you would personally like to discuss – I'm very active on LinkedIn. Do connect with me there. You can even look into the DataBuzz page and all the views and everything that I spoke about today was my opinion and had nothing to do with Morgan Stanley. [chuckles] Disclaimer.

60:07

Alexey

**That was an important part to mention, right? [chuckles] Yeah, everyone for joining us too – for being active, for asking all these questions. There was a very lovely and lively discussion in the live chat. So thanks for doing that. And I wish everyone a great weekend.**

60:29

Supreet

Thank you.

60:30

Alexey

**Goodbye, everyone.**