Consumer Finance Monitor (Season 4, Episode 26): A Deep Dive Into the Federal Agencies' Request for Information on Artificial Intelligence, With Special Guest Nicholas Schmidt, Chief Executive Officer, SolasAI

Speakers: Chris Willis

Chris Willis:

Welcome to the Consumer Finance Monitor podcast, where we explore important new developments in the world of consumer financial services and what they mean for your business, your customers and the industry. I'm your host Chris Willis, the co-practice leader of Ballard Spahr's Consumer Financial Services Group. And today, our podcast is going to feature a recent conversation between me and Nicholas Schmidt, a leading expert on artificial intelligence on the combined federal agencies' request for information about artificial intelligence. We'll be talking about what kind of information the agencies are asking for, what kind of issues they may be interested in, and what they might do with the information once they get it. So let's turn to that conversation and give it a listen.

Chris Willis:

So today, we're joined by a very special guest to talk about the federal agencies' request for information on artificial intelligence. And that's the Nick Schmidt. Nick is here with me and Nick wears two hats, I first got to know him as the AI practice leader at BLDS, LLC. BLDS is a firm that provides statistical fair lending analysis for creditors and others, and has been doing so for many years. And Nick is the AI practice leader there. He's a nationally recognized expert in artificial intelligence and machine learning. And so I've worked with him over the years in that capacity.

Chris Willis:

But Nick also launched a new venture called SolasAI, of which he is the chief executive officer. And Solas is a company that provides software to creditors to allow them to test and optimize machine learning models to reduce disparate impact to test for and reduce disparate impact during the model development process. It's an exciting new venture that Nick just launched, I think, within the past couple of months and offers a significant avenue for creditors to fold in fair lending compliance considerations into their machine learning model development process. So Nick, thanks a lot for being with us today.

Nicholas Schmidt:

Thank you. I'm looking forward to our conversation.

Chris Willis:

Okay. So that's what we're going to do today, Nick and I are going to have a conversation about the request for information from the federal regulators. And as we all know, a month or so ago, the federal regulators all jointly, the federal banking regulators, plus the CFPB put out a request for information seeking knowledge and responses about how the industry is using artificial intelligence and machine learning. And in fact, just yesterday, the agencies jointly extended the response deadline for that request for information. I think it's now July 21st, or something like that. So there's still time to get comments on.

Chris Willis:

But when that request for information came out, it got me thinking, "What will the federal regulators do in this relatively uncharted area?" Sort of acknowledge that machine learning is there, but they haven't really done that much in terms of making pronouncements about its use in the consumer credit industry. So I think the right place to start is, how much do the federal regulators know today about machine learning model development and testing? Is it a totally black box that they don't understand? Is the market 10 years ahead of them? Or do they have a firm level of understanding of what's going on with machine learning, what its advantages and potential pitfalls are? In other words, what is the state of their knowledge? And Nick, you interact with them a lot in the course of your work. So what's your perspective on the level of expertise with this? Both the CFPB and maybe the federal banking regulators with respect to this topic?

Nicholas Schmidt:

I think that the experience I've had working with the regulators and talking to them at conferences and things like that, is that they have a really good body of knowledge. Now, many of them are non technical, they're lawyers, so they may not have the kind of knowledge that a data scientist would have, but that's really not required in most situations. I think the days when someone could go into the CFPB and or any other regulator and say, "Oh, we're doing this, it's just too complicated for you to understand," are very much over and in a way, the Trump administration by taking away the ability to do much of any enforcement allowed the CFPB in particular to just learn. And what I saw was that there were many vendors, many vendors, many other groups going in and giving presentations to the CFPB on machine learning, AI explainability, and those presentations have been absorbed.

Nicholas Schmidt:

Now, when I look at the RFI, what I see is that they're asking exactly the right questions. And to me that indicates that they've got a pretty good body of knowledge. I don't think they're much ahead of many people in the industry. And in fact, I think the fact that they do see things from every angle means that they probably have a pretty good handle on lots of the different ways that things are handled and the pitfalls and benefits.

Chris Willis:

So in your view, and I agree with this but I think it's important for the audience to hear this, this RFI isn't an ignorant set of regulators grasping in the darkness to learn the basic facts about machine learning, model development and testing. This is an informed regulator asking targeted questions that will inform whatever the regulator decides to do next. Do you agree with that?

Nicholas Schmidt:

Absolutely. I mean, among some of my nerdy friends, we talk about the absurdity of some AI models and the way that people are putting them together. And one of the things that we all are horrified by is how over fit models are put into production. And it's particularly a problem with AI and ML, and that is, in fact, one of the things that the RFI talks about is what are you doing about over fitting? That's a technical question that goes into a little bit of depth, it shows that they know what's going on.

Chris Willis:

So just to give the audience some context for over fitting, my understanding of over fitting is basically where you train a model on a particular data set and you test it on that same data set. So its predictive power is very tied to the idiosyncrasies of that one data set. And the problem there is then it doesn't generalize out into the real world very well, because it's been trained on, and then tested on a two narrow data set. Is that a good working definition of overfit?

Nicholas Schmidt:

Absolutely. The way I like to describe it, is an overfit model is a conspiracy theorist. It finds connections where there really are not. And so what happens is it finds all these connections in the data that it's been trained on. But then when it goes out into the real world, those connections are no longer there. They were just an artifact of that original data. And so the model ends up failing in ways that are surprising to the modeler.

Chris Willis:

Right. And you don't get the predictive performance you expect. And then you may get probably anticipated effects. It could even affect the fair lending performance of the model, because you could get disparate impact in ways that weren't demonstrated by your development data set.

Nicholas Schmidt:

Yeah, absolutely. There is, it is an all encompassing risk. It is a business risk and it's a regulatory risk, it's a reputational risk. If you are putting overfit models into production, you're not engaged in certain practices.

Chris Willis:

Yeah. And it's really interesting how the overfit question fits into the existing older, for example, OCC model risk management guidance. Where that guidance is very clearly devoted to making an accurate model that will work well into production. And overfit phenomenon is one that is particular to machine learning and it shows that the OCC, for example, which is very concerned with model, predictive performance, and safety and soundness would be very naturally concerned with overfit, I think.

Nicholas Schmidt:

Yeah. I wouldn't say that overfitting is a concern just for machine learning, it's more of a concern and it can be harder to diagnose. The model governance procedures that are in place and most lending institutions do a very good job handling that risk in traditional models. And that can be handled in machine learning, but it's still a developing science to some degree.

Chris Willis:

Yeah. We can talk all day honestly, probably about overfit. But let's go on to some the other stuff that we were going to talk about on the webinar. And when you and I talked to prepare for the webinar, we isolated four potential priority areas for the federal regulators. I'm just going to introduce them now and then we'll talk about them one at a time.

Chris Willis:

So you and I think that four areas where the regulators may be interested and may take some action are the following, on the screen now, explainability and adverse action notices. In other words, can you tell why a model made a decision and what the principal decision were? Second, the consumer group criticism of machine learning models containing hidden or inherent biases and what that means and how the regulators might deal with it. Third, how would you test a machine learning model for disparate impact, both traditional underwriting and fraud type models as well as newer types of interactive models that we'll talk about when we get to that. And then finally, what I'm going to call the big enchilada of this discussion, which is machine learning and less discriminatory alternatives. One of the elements of disparate impact is if there's a less discriminatory alternative that equally serves the business need involved, then you have to adopt that, else you're at risk of a disparate impact.

Chris Willis:

And there's something revolutionary about machine learning models and less discriminatory alternatives that I think the regulators are going to need to address. And so we're going to save that for the last part of our discussion. But I'm going to call that the big enchilada. So let's take those Nick one at a time. Let's talk about explainability and adverse action. This is one of the few areas where we've seen one of the regulators actually say something publicly.

Chris Willis:

So the problem here is that there's guidance in regulation B and the commentary about how to develop and report adverse action reasons. The principal reasons for the decline. And the examples and then commentary of reg B have been there for decades, and are really built into the assumption of a logistic regression underwriting model. And they don't really work that well in a machine learning model. And the CFPB has acknowledged this, there was a July 2020 blog post by the agency discussing this problem and then the CFPB held a tech sprint in October of 2020, to allow people to showcase potential solutions to this problem. So let's give the audience your take on the explainability and adverse action issue. What it is and what you think the regulators may do with it.

Nicholas Schmidt:

One of the things I think is really important that was pointed out in the blog post is the principles that underlie the requirement, the adverse action notice and in it they said that the adverse action notices serve anti discrimination, education and accuracy. And I absolutely agree that there are all these issues about how do you apply the existing framework to machine learning AI? What does it mean, but I think if we come back to these three purposes, we can ask, whether the methodology we're using actually does serve those purposes. And the thing that really needs to get thrown out, is the example form with I think it's 18, 20 possible reasons for declining credit, those just are rarely helpful. Particularly in models that have and use alternative data.

Nicholas Schmidt:

I don't see machine learning and AI as being such a big problem in developing adverse action notices, particularly for the most commonly used types of machine learning models, which are typically gradient boosted trees. The techniques that have been developed to explain those are really, very good. And are not any worse than what is used in logistic regression. The question that is pertinent is, if you have a model 200, 300 variables, how do you combine them in a way to give that up to four reason codes. And that's a difficult question, because you may have things that are in one, that reasonably go in one reason code and another and there's a real opportunity to get it wrong and a real opportunity for obfuscation. And so what is required is a lot of care at that point of a model building things.

Chris Willis:

So do you think that... First of all, have you seen any movement in the industry away from gradient boosted trees as predominant underwriting model? Do we think we're going to continue to see that technique used, so that the robust explainability principles that accompany those models are good for us for the time being? Or do we have to develop something new?

Nicholas Schmidt:

There are a lot of techniques for neural nets. And to the extent that people move away from gradient receding trees to neural nets, there are opportunities for explanations. Because my clients tend to use trees, my area of focus has really bang on that. As I have seen move people move into the deep neural networks, we're going to get into that more. I think that the explainability is catching on. And so the issue of aggregation is probably way bigger of an issue than... And that's the human part of aggregating and deciding which four codes to give people or which, four reasons, is probably a bigger issue and a bigger problem than the explainability techniques and their relative inaccuracy. So I see it as sort of it's an issue. But it's not something that cannot be overcome today.

Chris Willis:

Yeah. That's my impression, too, is that I feel like this explainability adverse action concern arose in the earliest days of machine learning models, when explainability principles were not very well implemented or available. And a lot of machine learning models to me were black boxed, because nobody knew why they were making the decision providing the score that they did. But model development technology seems to have very quickly adapted to that, and gives us explainable, highly explainable models now. So this feels like yesterday's problem to me. Do you agree with that?

Nicholas Schmidt:

I don't know that I do so far as yesterday's problem. But yes, I think for the most part, the question of how you interpret or how you create explanations, technically, is good enough. A lot of the confusion, I believe, came out of academic work in fairness and AI and it's excellent work, but it really muddied the water, in terms of how does industry use explanations. There was one article I read that started with the Socratic concept of an explanation and then went downhill from there. And I'm trying to figure out how my clients should create an adverse action notice, I just threw the paper away. Because it's such a good start, but it's not actionable. And when we think about what is actionable, what can be actionable for a lender, I think we're in pretty good shape to use these techniques.

Chris Willis:

Yeah, that's my impression as well. So that's... To me we put the easiest issue first. And surely the regulators are going to address this, again, the CFPB has publicly manifested interest in it, twice last year. So we expect them to do something about it. But this isn't that hard of a nut to crack. So let's move on to, progressively a little bit more difficult nut to crack. This idea of hidden bias in machine learning models. And so here, there's a couple of viewpoints. And I hope you can help the audience understand them. On the one hand, you have people in the industry who say, "Look, I've got a data set of empirical data showing various attributes of borrowers and their empirical repayment performance. If I go train a machine learning model on that, what is the problem? All I'm doing is using empirical data. I'm not putting any bias in it. I'm just asking which of the input variables are most predictive of the eventual repayment performance in my data set? How can you criticize that? What's wrong with that?"

Chris Willis:

On the other hand, you have consumer advocates, saying quite loudly, "Machine learning models can do incredible harm to society because they will have hidden biases in them and will perpetuate or exacerbate inequalities among protected groups within society, like African Americans or Hispanics or other protected groups." What is actually the problem? What does it mean to have hidden bias in a model? Where does it come from? And how would you respond to someone who says, "I don't get the problem, all I'm using is empirical data?"

Nicholas Schmidt:

I have to have these arguments very frequently. And that the idea that the data are the data and that's what it is, is really not reasonable. Because at the end of the day, a model is, I'm going to oversimplify, but it's looking for average and facts. And if you put data in, even if it's good data, it's going to average the effects across people. And so you can have something where it is adversely affecting a group, African-Americans, more so than let's say Whites, even though it's just truly unfair. And an example I've given and I apologize Chris, because you've heard me say this one, is I know a model that that found that the best predictor of charge offs for a credit card was how many times you've shopped at a convenience store in the first 30 days you had a car. And I laughed about it. I thought, oh, that makes sense. We've got... What do you buy at a convenience store? Cigarettes, beer, candy, lottery tickets. Those are pretty well correlated with a high risk behavior.

Nicholas Schmidt:

But then what I thought of is, what about food deserts? What about cities just generally, where you shop at convenience stores more often? Well, that's also correlated with race. So what you end up having is people, let's say, more frequently African Americans who live in cities more often, who are a low risk, but they shop at a convenience store because it means convenient. The fact is, is in this model, the White people in the suburbs who go to convenience stores engage in high risk behavior are contaminating the effect of people on this data. So yes, that variable maybe predictive, but it's causing harm to a subgroup because things are unequally distributed. And so I don't think that that argument of the data is predictive is nearly sufficient.

Chris Willis:

So having set forth the problem, and I understand the problem is that the data set may create the kinds of associations, on average that you just mentioned, that both fail to predict for part of the population and which unfairly penalize part of the population. Let's talk about what we do about it as a society and maybe as the regulators as well. There are some extreme solutions out there that have been proposed by consumer advocates to this problem. One thing that comes to mind is Cathy O'Neil's Weapons of Math Destruction. So tell us what those extreme proposals are, before we get to what might be more realistic.

Nicholas Schmidt:

The main idea I understand is getting rid of credit scores, getting rid of algorithms et al and that leads to two possibilities. One is that the bank lets everybody who walks in the door get a loan. And that's probably not going to result in very good outcomes for the bank and maybe even for the borrowers, depending on what happens when they don't repay.

Chris Willis:

Right.

Nicholas Schmidt:

The next one that I often hear is, well, there needs to be more personalization. And I don't argue about that. But part of the benefit of an algorithm is it takes away the biases associated with personalization. So effective non discriminatory personalization, that absolutely would make a model better. But if that's not done carefully, then you're just making the situation worse.

Chris Willis:

Right. It seems like a throwback to the regulatory concerns historically, with judgmental underwriting, which is, of course, highly personal. But it's done in a way that the regulators consistently attacked as leading to disparate impact over a series of decades.

Nicholas Schmidt:

I think that's right. Now there is counter method which is that are a number of smaller lenders that have focused on more judgmental lending in certain communities where there may not be enough data on immigrant communities, things like that, where models might not be that effective. That's great. And that's probably very likely to be increasing credit opportunities. But that's a very highly specialized area. It's difficult for me to imagine that it's really feasible for national lenders to implement those kinds of policies and modeling with some fallback and appealability and adverse action notices, and maybe second look models is probably the best alternative.

Chris Willis:

Yeah. I think that's a good segue into talking about realistic solutions for a national scale, and doesn't want to do judgmental underwriting tailored to every community and need something that can scale, but can also be fair and meet the requirements of passing fair lending standards. So give the audience some ideas of, if I'm a creditor and I'm interested in trying to prevent, detect, undo, counteract hidden biases in my machine learning model development, what would I do?

Nicholas Schmidt:

There are a lot of opportunities. The first thing you can do, and possibly the most important one is to get a diverse group of people in a room and talk about the variables. And given what I do, my tendency is towards a quantitative solution. But really that first pass is what's important. And what most of our clients do is they get fair lending lawyers and outside counsel, inside counsel compliance stakeholders, and hopefully, it's a diverse group for people to look at data and say, wait a sec, I'm coming from my community, I know that this is going to have an effect. And it might not be something that you or I catch, but it would be something that other people catch. So that's just getting through the door. And that's a very important thing.

Nicholas Schmidt:

But then once you get into data that are being used in a model or a potential model, then you really do have to start doing testing. You want to test on the model outcome first, and see whether or not there's bias or disparate impact or whatever appropriate measure might be. And then if there is, figure out where it's coming from. And of course, ultimately do something about it. But looking for what might be there is the first step.

Chris Willis:

Yeah. And so, I mean, what I hear you saying is, obviously having a good eye on variable selection in the first place is important. So somebody should catch your convenience store variable, probably before it even gets into development, because that would seem reasonably likely to have a disparate impact and would not seem very intuitively related to credit repayment performance, as opposed to something that more directly measures the consumer's experience repaying credit, as opposed to shopping habits, for example. And then I think what I hear you saying is, do the fair lending testing on the model, through the development cycle, to make sure that you can see any disparate outcomes of the model, identify where they come from, and make appropriate adjustments as you go. And that, in my mind seems like it's an integral part of the model development process, now that we know how machine learning models work. Would you agree with that?

Nicholas Schmidt:

Yeah. I think it's integral, it's essential and it's expected by regulators. And all of your peers, all of your largest peers are doing it, you probably should be as well is what I would tell a lender. And so the idea that you can skate under the surface without doing this testing, I don't think is viable. And the risk of not doing it, there's not only a regulatory risk, but there's also a reputational risk now. I think probably everybody in the audience can think of at least one or two of these algorithms going on, and then becoming very public. And so if you don't have the ability to say, "We looked at this, we know why there was a disparity. Here's why it is, it's reasonable. Taking it out would have caused us harm and maybe it would have caused consumers harm overall. Do you see why no?"

Chris Willis:

What is the role, do you think of the data set, the training data set in managing the hidden bias issue? Is it true that a larger, more diverse data set helps address or counteract the possibility of hidden bias implication?

Nicholas Schmidt:

It can. It's possible. So having a more diverse data set will absolutely be able to cache idiosyncratic things for smaller groups. And one of the best examples is in race, let's say that African American borrowers represent 15% of the population. If there are things about their credit underwriting process that are idiosyncratic to them to some degree, if you don't have a sufficient number of African American borrowers in a data set, the model cannot pick that up. So having a bigger data set is great in that way. It can be harmful then, in terms of or it can be good, it is not determined whether or not you're including a lot of variables. If you're just including 10 variables that are reasonable, they're causally related to credit, they're understandable, all of that, that's very good. If you add the 11th variable and it's, do you subscribe to Avenue magazine? Adding data has not done anything good there and it's probably only going to increase bias and disparate impact.

Chris Willis:

Right. And that, of course, that's the kind of variable that you would mentally exclude at the beginning before you even start training the model, I think.

Nicholas Schmidt:

Yeah. And it's a good example of a hard to find variable in a technical setting because... I haven't done a study of this. But I'm willing to bet that the number of people who subscribe to Avenue magazine is relatively low, which means that the variable probably has relatively little impact. And there as much as it is probably the vast, vast majority of people who subscribe to it are African American, there are a lot of African Americans who do not subscribe to it. And so if-

Chris Willis:

Right.

Nicholas Schmidt:

... you just have a technical methodology, it's just going to wash out with a lot of the other variants. And that differs from that review that you do by hand, where Chris Willis says, "Wait, that's not right. That's going to be a problem." And so I think that's where this handholding between people involvement and technical involvement is necessary.

Chris Willis:

Yeah, it makes sense. I think that's a great segue. We've talked about fair lending, testing in models as part of the development process. Let's talk about that. That's another area that we think the regulators will be interested in. And before we get started about this, we have a question from the audience, which makes a natural point to address here, if a creditor doesn't collect race data on borrowers of a particular product, which of course it's illegal to do for a product except mortgage, how do we determine or assign race and ethnicity for the purpose of doing fair lending testing? I mean, I think the answer to that is you just use BISG, you use the name and the address to infer probability of race or ethnicity. And even though that's an imperfect method, it's what the regulators use. So that's what we should use for testing. Would you agree with that, Nick?

Nicholas Schmidt:

Yeah, absolutely. The regulators are not completely wagged to BISG, as I understand it, but there's no reason not to use it. It does have widespread acceptance, it's what virtually everyone does use, it's pretty easy to do. And it does a very good job explaining group differences. There's so much controversy about it, and I think that's very much overblown when we're talking about disparate impact analysis for a model. Because you're averaging things out, errors to a large degree get averaged out. And you do end up with a very sound and pretty solid measure of the race, of facts.

Chris Willis:

Okay. Thanks for that. So let's talk about how to do fair lending testing of a model that was developed using machine learning technology. Let's take it in pieces. First, let's take more traditional things you'd use a model for as a creditor, so making an underwriting decision or a pricing decision or spotting a high risk of fraud or perhaps having a model that will prioritize collection efforts, models that are basically designed to influence credit related decisions, where there's a specific target variable of like probability of fraud, probability to charge off, probability of repayment, things like that. If I want to test a machine learning build model, of one of those types for disparate impact, how do I do it? Is it any different from testing an old logistic regression model?

Nicholas Schmidt:

It really is not. The outcomes are the same, right? And with first case disparate impact testing, all you're interested in is looking at the outcomes and their effect on race. And you're not conditioning on any other factors. And this is one of the things that modelers really hate, is that disparate impact doesn't take into consideration whether or not you're going to repay the loan or the accuracy of the model or anything like that, all it does is say, is the average score or the average outcome for Blacks less favorable than that of Whites? Or women and men. Things like that. And if that is the case, it's statistically significant. It's beyond a threshold that I normally typically set to say, "Okay, this is something for review," then there is disparate impact. And it's done in the same way that we do judgmental processes, it's the same way that you do logistic processes, really very, very minimal difference there.

Chris Willis:

And even beyond the raw disparate impact of, is the score distribution different across protected classes? I assume that you could also add control credit worthiness controls, let's say it's an underwriting model, that are designed to root out differences where there is an unexplained difference on the basis of race, even holding credit worthiness constant via controls. If you just do that, you can do the same thing with the output of a machine learning model that you could with a logit, right?

Nicholas Schmidt:

Yeah, you absolutely can. But I've always felt that that is missing the point, and particularly with the logit model, where understanding why the model is giving a prediction is such an amazing thing. But even with machine learning with explainability techniques, let's say you have a machine learning model with a bunch of alternative credit variables and you find those disparate impact and so you say, "Okay, I want to see if credit worthiness is driving it." So I take the FICO and I control for FICO. Well, FICO is not in the model. But we know exactly what is in the model. We can determine with 100% precision why an outcome is occurring. And so the idea that you control for other variables outside the model is, I think, not a ideal way to go. Because at the end of the day, all of those factors that you might control for are not going to control for the entire prediction.

Nicholas Schmidt:

And what that tells you is that if there is any disparity left over, then it must be your model causing discrimination. And that's not really true or not necessarily true. It may be that those other factors are entirely valid. And so you've just gone down a road, where it looks like you have potentially intentional, or well, not potentially intentional, but you have disparate treatment, because you haven't been able to explain away the differences. And you can't do anything further. And so I'd say just start from the model, stick with the model, understand why the model's giving a different prediction, and then determine whether or not it's reasonable.

Chris Willis:

Reasonable in terms of being justified by the predictive-

Nicholas Schmidt:

Yeah.

Chris Willis:

... power it involves in product?

Nicholas Schmidt:

The predictive power of the model, the correlation with race or whatever class it may be, and also the reputational power and ability of it. They are things very predictive, they're highly predictive, they may be really good valid factors, but it's just not worth the risk. And-

Chris Willis:

Got it.

Nicholas Schmidt:

Yeah.

Chris Willis:

So let's talk about other types of models. So we've talked about the more standard traditional models, underwriting fraud collection, pricing, things like that. What about AI models that would underlie more interactive systems that a creditor might use? Like speech recognition or emotion recognition? Or the AI model behind a chatbot. There you don't have this clear target variable, like we were talking about a minute ago, you have a more interactive and more fluid and dynamic situation. How would you test one of those for potential differential treatment disparate impact?

Nicholas Schmidt:

There are two things here, and those are really important. And they have the potential to change the fair lending landscape. Because right now, our clients and we focus really on this disparate impact notion of differential distributions in outcomes. And it's all predicated on being able to define a favorable outcome. If there's a model and it's saying get product A or product B, and they're not rentable, one gives you a bunch of miles, one gives you a bunch of points for a credit card, who knows which one is favorable overall? There's really not any disparate impact testing that can be done. Those models have just generally been put in lower risk tiers and not tested. In these new models like the language processing, we may be in a similar situation where there's no offer being given. It is routing you to the right place and say, "Hi, I want to change my address." And so it sends you to that place.

Nicholas Schmidt:

The issue is that these models have been shown to be hugely differentially predictive. And this is actually what we call bias. Bias with a differential prediction is saying that the model is just not as good for one group as it is for another. And I just found a paper that came out, I think it was last year in the National Academy of Sciences journal that found that speech recognition systems for Google, Microsoft, Apple, were twice as likely to make errors for African Americans relative to Whites. And what was interesting was, it actually wasn't in words. They believe it was the rhythm, accents on the syllables and some of the word orders that tend to be used in African American vernacular English. Basically, the model couldn't figure it out. And there's no offer there, but if you make those kinds of mistakes basically, if your model makes that kind of things to a particular community, it's going to be problematic.

Chris Willis:

Sure.

Nicholas Schmidt:

And while these models may only be used for servicing now, eventually, they will be used for credit decisions or steering or things like that. And that's when it really will become a problem.

Chris Willis:

Sure.

Nicholas Schmidt:

In the short-term I think bias testing is going to be important.

Chris Willis:

Yeah. Bias testing in the sense of testing the model on diverse user groups to make sure the experience is consistent across it. And again, this also seems like something where a diverse data set for training is essential to prevent this kind of thing from happening.

Nicholas Schmidt:

That's absolutely right. This is the place where diversity in data is essential. This is a problem in the facial recognition software. There's a famous paper by a woman at MIT, on the front page of the New York Times found that the error rate for identifying gender from women of color was 35 times that of White men. And that has profound effects. And the problem was, was that there were not enough women of color in the development data.

Chris Willis:

Yeah. Makes sense. So we have a bunch of questions from the audience on BISG and other sorts of issues that I don't think we're going to have time to address. So you and I will address them by email. So don't worry, we will answer your questions by email after the program. But let's move into what I announced as the big enchilada at the beginning of the program, which is less discriminatory alternatives, there can be a disparate impact, then the defendant has to show there's a business justification. But then if the regulator shows that there's a less discriminatory alternative that serves the business justification equally well, then it's still a violation of the law.

Chris Willis:

And as I perceive the problem here is the traditional way of searching for less discriminatory alternatives that grew up in the days of logistic regression models was to take each individual by itself, drop it out of the model and then look at the impact on both the model's predictive efficiency, and on the level of disparate impact in the model. And where we use drop one theory to identify variables that contribute little or none to the predictive power that contributes significantly to disparate impact, we know to drop those variables out of the model. But that's a lot for machine learning models, because in machine learning models the variables interact, they don't sit there one by one and operate independently of one another. That's really the defining feature of a machine learning model. So what are the techniques for doing machine learning model development with less discrimination. How do you do it?

Nicholas Schmidt:

So to give you an illustration of the problem that you're talking about, there is early on in my days with machine learning with lenders and looking at these less discriminatory alternatives, we looked at a model and there was a variable that had 60% of the total importance in the model, which is very high. It seems like you drop that variable, the model falls apart. I suggested trying dropping that variable and I thought the model would fall apart, but just I wanted to see it. They did, the model stayed exactly the same in terms of its quality, the variable number two that might have been 15% of the importance was now 60%. And so the idea that even dropping a very important variable is going to have some big effect on disparate impact or equality, just not true, necessarily. Absolutely, sometimes it is.

Nicholas Schmidt:

So what you have to do is use these other techniques. And there's been tremendous amount of very good work done on developing techniques that either change the data, change the model architecture, to train the machine to learn unbiased results or to change the predictions, all to try and minimize the disparate impact. And each of those has potentially big problems in terms of compliance, I believe. Because they are more explicitly using race in the modeling process, the modeling development process, than is traditionally used the drop one approach.

Nicholas Schmidt:

And given our history, BLDS's history in developing some of those methodologies, that's really where we started, and even machine learning. And so I developed, with my team, this technique that instead of doing a drop one, it was drop a lot, add a lot, try different combinations. And what we found is that, even though that may not be surgically precise, it does a really good job of finding less desperate alternatives. And it turns out, if you don't mind me, potentially anticipating your next question, it turns out the machine learning actually is better at finding less discriminatory alternatives than other methods, or sorry, than traditional logistics or ordinary least squares and-

Chris Willis:

-that process is not judgmental.

Nicholas Schmidt:

Yeah. Well, that may be possible. But that's not where I was going. What it is, is that machine learning is so malleable. And it can take so many variables. And it has so many little knobs and dials that can change, there's really an opportunity to go in and extract variables that are similarly predictive, but causing different levels of disparate impact and off the effect or lower the effect so that you end up with virtually identically a predictive model, but is finding or is causing much less concern.

Chris Willis:

Right. And so it seems like if that's the situation, then the creditor is presented with an easy problem and the regulator is probably presented with an easy problem of if I have a choice between three models, all of similar or identical predictive efficiency, but one has less disparate impact than the other two, I'm going to choose the one with less disparate impact, and I'm legally required, probably, to choose the one with less disparate impact. That seems like the easy question. But here's the hard question Nick, let's say that I go through a model development process and I have a choice between models where I have the most predictive model that has a certain amount of disparate impact. And I can give up some of my predictive performance to reduce disparate impact. Am I legally required to do that? Is the regulator going to say that I'm legally required to do that? I don't understand the Equal Credit Opportunity Act to require that currently. But what do we think the answer to that question is? Is there a requirement to give up prediction to some disparate impact or will there be?

Nicholas Schmidt:

I think that there is some expectation among regulators that you look into that, and if you are able to find a model that has lower disparate impact and a marginal decrease in the quality, you should have a very good reason why you're not adopting it. And we have a lot of suggestions we make about how to define different thresholds and trade offs that should use to figure out how much of a quality drop you're willing to take. And I don't think it's a one size fits all. But, again, it's one of those issues that even if it's not explicitly required by the regulators, there is some expectation that you would do it, and everybody else is doing. Do you want to be the aggressive lender in terms of your compliance strategy? Maybe, maybe not. If you are not, or if you are and really are putting yourself out there, in an environment that right now is not friendly.

Chris Willis:

Got it. So let's conclude. We've just got about five minutes left. The federal regulators have issued this RFI, they've extended the comment period on it again, until sometime towards the end of July. Let's speculate together for a minute, about what will the output of this RFI be? So they're going to gather all this information from the public, from the industry, from consumer advocates, and everybody else, what are they going to then do with it? Let's start with rulemaking. Do you think there's any possibility we're going to see a full on notice and comment rulemaking from the CFPB or anybody else about how to properly use machine learning or artificial intelligence in consumer credit?

Nicholas Schmidt:

I don't see that happening. I think that there will be comments, and I think there will be a lot of advisory. Suddenly I'm forgetting the... I think there'll be a lot of commentary on it. But I don't expect any specific regulations to come out of it in terms of fair lending. I think that the process of developing compliance is probably going to be pretty organic and not led by rulemaking.

Chris Willis:

Yeah I-

Nicholas Schmidt:

I am curious if you... Okay. Yeah.

Chris Willis:

I was just about-

Nicholas Schmidt:

I think the loaning thing is just too hard.

Chris Willis:

Yeah. It's too complicated, there's too much in the future that's uncertain. There's still a lot of technological development going on. And there's no reason for the regulators to tie themselves down with a regulation. So I think, I agree with you, there's virtually zero chance of rulemaking here. I do think you're right, that there'll be some informal guidance statements made, along the lines of that guidance statement that we got from the regulators jointly about a year ago, on the use of alternative data in underwriting models. Not terribly specific, not terribly prescriptive, just general discussion of the issue. That's what I'm thinking is going to happen in this area. What do you think?

Nicholas Schmidt:

I think so. And I think actually looking to model governance, generally, in the process that has come out of SR 11-7, is probably going to be something similar that comes out for fair lending in AI. SR 11-7 is not very specific, it's principles. And it has worked really well. And you talk to modelers and risk managers at banks and they all hate the documentation they have to put up with, but there's generally recognition that it's a pretty good piece of regulation, because it just specifies the principles. And my expectation is more of that will come out, because as you said, and you're absolutely right, they don't want to tie themselves down to a specific technology. And it's far too fluid to saying this is right or this is wrong.

Chris Willis:

Yeah. And so I share your view in that regard. And then I think the agencies will just use their supervisory and enforcement tools to both learn about and understand and take action with respect to any practices that they don't think are appropriate. Last question for you, Nick, do you think that the federal regulators are going to evolve into a position of hostility towards machine learning models? Will there be a war declared on machine learning models saying, this is just bad and we don't want you to do it? Or do you think it'll be a more moderate response from the regulators? What's your take on that? That's going to be our parting question, by the way.

Nicholas Schmidt:

I think it is going to be more moderate. Because among the regulators I talk to, and I don't talk to everyone, but even the true deep, deep liberals, I think, recognize that there is value to this technology. And also, that it's not going away, no matter how much anybody may want it to. So I don't think it's disappearing.

Chris Willis:

Yeah, I agree with that too. And in fact, I think that the regulators believe that there's a lot of social good that can be done with machine learning models by making credit more available to traditionally underserved groups that if used properly, machine learning models, not only can be more accurate, more predictive, etc, but can make credit decisions more inclusive by their capacity to digest and deal with more data than your traditional 10 variable, which is the progression credit bureau based model. So for that reason alone, I think they're not going to squash machine learning models. I think they are going to insist that they be properly developed and tested as we've been talking about throughout the course of this webinar.

Nicholas Schmidt:

Yeah. I agree.

Chris Willis:

I really want to thank Nick for having that conversation with me and for the opportunity to share it with you all. And of course, thanks to all our listeners for tuning in to today's show. Be sure to visit our website, ballardspahr.com, where you can subscribe to the show in Apple Podcasts, Google, Spotify, or your favorite podcast platform. And don't forget to check out our blog, Consumer Finance Monitor, for daily insights about the financial services industry. If you have any questions or suggestions for our podcast, please email us at podcast@ballardspahr.com. And stay tuned each Thursday for a great new episode. Thank you all for listening.