Interview

Dan: So the first questions that I have are just about like your existing. What your existing approach is for model explainability. So I guess first question would be like. Do you use PDP and ICE plots for model explainability or do you tend to use other techniques?

Participant: Yeah, I don't use ICE plots, but I. I think that it's super helpful. I use PDP plots, but like not. It's kind of like the poor man's average marginal effect where I plot like maybe maybe I'll find two groups. Most of the work that I do is like. Heavy XXXXXXXXXX ML. So modeling XXXXXX XXXXXXXXX XXXXXXX XXXXXX XXXXXXXX XXXXXXXXXXXX. And so maybe like I will sort of like. Look at the marginal effect of different features and separate that by location. Right. So you could imagine that maybe like the clustering that you're doing. You could do that automatically or something. So I do that, but I'm really, I'm looking at like the average predicted value across. The actual features. So I'm not like doing a Monte Carlo simulation and like replacing. The actual value and holding everything constant. So it's like a little bit less work. And then I use, I would say like the main framework that I use is like SHAP values. To just to understand feature importance and then the the the shape of feature importance across the feature value distribution.

Participant: Do you use both one-way and two-way plots? Typically, yes, I would use two-way plots, but I I definitely don't have exposure to this. Like this extensive of an examination.

Dan: So just to before we move on. So. In the plots you're using what you call like these like marginal effect plots, is that basically like. Like do you like bin the feature and then like what's the average model prediction for this bin and then look at. Look at that.

Participant: Exactly.

Dan: Okay. Got it.

Participant: Yeah, but I'm not like taking so I'm not like taking the instance. Holding everything but the feature constant and iterating over every possible feature value.

Dan: Yep. Yep.

Participant: So it's sort of like not really it's not really the same thing, but it's like generally the same idea.

Dan: Yeah. Yeah. Got it. And so is there a reason why you use that as opposed to PDPs or is there a reason why you don't like look at ICE plots? Or like for why, like you prefer a SHAP over over that?

Participant: I would say that the reason is like convenience familiarity history of using the tools. But I mean going forward, I think I totally want to like incorporate using both of these. In in my sort of like model interrogation, I mean like probably like half of my work is is sort of model interrogation and I end up sort of doing code reviews for and model interrogation for other people's models that are more domain experts and less, you know, ML practitioners that maybe they would just like train a model, look at bulk error stats and then immediately deploy it for some. You know, bespoke solution client.

Dan: Got it.

Participant: And so they don't necessarily know to look to see if the models are fitting on like three observations, which happens all the time.

Dan: Right. So for these marginal effect plots. Do you do that for both one-way and two-way?

Participant: I would say like more commonly one-way. But I again, it's all sort of like hypothesis based. And so if I suspect that there is. A heavy interaction or there's confounding that's happening, then I would like maybe do it two-way. But it would have to be I wouldn't like my go to procedure wouldn't be iterate all over all. Features and look at every two-way table.

Dan: So that feeds into the next question of how do you determine which plots you create and look at?

Participant: I would, I mean, typically would start with like. The highest ranking features according to their feature importance that comes straight out of the methods. If it's like a tree based method or the SHAP plots and I would. I mean, the first table that I would look at for each of the most important features is how the predicted value varies with the feature value and then. Whether or not that's super in line with the observations in the training and validation set.

Dan: Okay. So in terms of like seeing if it's in line with the training data, what type of visualization would you use for that?

Participant: Yeah, I would look at just like two different scatter plots. So like comparing the predicted versus the actual

across the feature. So it'd be like a two side by side scatter plots really or overlapped something like that.

Dan: So one so one scatter plot would have like the feature value and one axis, the predict the value on the other and then compare that to the ground truth.

Participant: Yeah, exactly.

Dan: Okay.

Participant: Or or I would even, you know like, before that I would look at just like the univariate kernel density estimate for the actual and the predicted.

Dan: Can you describe a bit more about that?

Participant: Yeah. So just like to understand if the model is like predicting the right distribution shape. I would look at either like a fine grained histogram of the predicted and the actual overlaid or a kernel density estimate of those values.

Dan: Got it. And then for. For analyzing. How predicted value varies with feature value. Which, which particular SHAP plots are using for that?

Participant: I'm trying to remember like what the names of the SHAP plots are. But it's it's typically it looks like

Dan: I can pull up the the GitHub page as well if that makes easier.

Participant: Yeah, that would be perfect. I don't know exactly like what the what they call the different plots.

Dan: Right.

Participant: Oh, yeah. Okay. So like this.

Dan: So I think this is an instance level plot and then I guess this would be...

Participant: This. Bam. The bee swarm. I feel like this is normally what I look at right to understand. Okay, in this range of feature values, there's a. Negative or positive impact on predicted value. And the main takeaway, like how this would help me in some in deciding like, is this model good or not. Is in addition, of course, to like bulk error stats and looking at actual predictions is like, is does this makes sense. Is this mechanistically driven. Am I capturing the intuitive mechanism. For whatever is being modeled.

Dan: Got it. Okay. And what would you (unintelligible)? Are there any pain points in your current model interpretability approach?

Participant: I wouldn't say it's like. There's not like a one. There's not like a task that's always like, oh, this is like a bummer that I have to do this, but more like. You know, I don't know what I don't know. And so the huge upside to this tool is that it's like pretty exhaustive. And so the issue with the hypothesis driven approach is that. If you don't have the hypothesis, you don't think to look at it.

Dan: Mmm hmm.

Participant: Yeah. So I really like that part. I mean, automatically, just like to have every single. Plot right there. That's huge.

Dan: Okay. So now we'll shift into talking more about PDPilot. So in general, how would you say it supported or do not support your model analysis. Were there any questions that you weren't able to answer or any tasks that you weren't able to perform?

Participant: I mean, I think the number one takeaway is that like immediately my eye was drawn to the model having that weird. Effect being over fit to those three observations. I think I didn't come to like that. That wasn't my like complete conclusion. But, you know, given more time and if I understood the dataset a little bit better, like. I think that the. If I had, I think that would be like the goal. Right. By looking at these plots, like one of the really beneficial outcomes would be to understand that the model is over fit. But, you know, the plots did help me get halfway there. Any tasks that I was unable to perform, none that I can think of.

Dan: Okay, so I guess at the end of your exploration where where would you say you were at in terms of like. What that effect was with the above ground living area?

Participant: Well, I think I definitely my. I knew that something was probably wrong with the model or that there was like some other feature that was causing that that was super correlated with the below ground, the basement size or

whatever. So like. My thought was okay, this variable is confounding with something else that isn't popping up in this analysis. So maybe that's like. This house is super old or like something that's correlated with basement size. I don't know, maybe like a rotting or something.

43 **Dan:** Okay.

44 **Participant:** And so. Yeah.

45 **Dan:** Is there anything that you think that the tool could have done differently to like help you in that analysis?

46 **Participant:** I mean, I think really it's more the. No, I don't think so. I think the tool helped me see. That issue immediately and then like having a comprehensive diagnostic just like takes longer than 30 minutes. So like, yeah, I don't think so. I think the tool. And I think getting familiar with it. Would allow me to use it a little bit better, but I mean, in terms of like the usability is if I can even navigate around after just watching you do it for like 15 minutes. Like, I think that that's pretty, pretty usable. There are a lot of GUIs out there that are like, I don't even want to try because they just like take to learn to learn.

47 **Dan:** Okay. So were the visualizations useful and were any of them unclear?

48 **Participant:** The visualizations were good. I think there was one. What was I, I feel like I kept trying to do something that I couldn't do the. I think it would be nice to be able to see the multiple clust, like the automated clusters on the ICE plots. I think that's the only thing that would be pretty nice.

49 **Dan:** Right. So with that, would you want like. So what would you envision like the workflow for that being? Would it be like. You like look at the clusters and then like. I don't know. There's like some button that's like highlight these clusters in or highlight this cluster in.

50 **Participant:** Right.

51 **Dan:** Like these plots and then you'd have like those selected or would it be.

52 **Participant:** I think it would be like, okay, so if you go. So if you sort of toggle the plot to clusters.

53 **Dan:** Yep.

54 **Participant:** Here. So I would want to be able to take these clusters and then superimpose like separate the ICE. The ICE instances. By by the cluster. And so maybe what that would be like is, yeah, like have a. Have a. Even if you just like toggle back to centered or standard, like have them automatically be highlighted and then like. Click anywhere on the plot to. To remove that or something.

55 **Dan:** So would that be. Would that be highlighting like this cluster of instances across all of the other plots?

56 **Participant:** Ohh, yeah.

57 **Dan:** Or would it be be within each plot you would like color the lines by what cluster they're in.

58 **Participant:** Probably the latter. To be in line with the way that the clu- for for this page. For the way that would probably be the most intuitive scenario because that's what the cluster view is doing. So it makes sense to like extend that, but I like what. I mean the the. The selection kind of does that and that's really nice. But it would be it would also be cool to just like have it be right off the bat accessible.

59 **Dan:** So what would you see as like the. I guess what would you what would your use case be for like coloring the lines by their. Their cluster. So if if you had these ICE plots and each line was colored by its cluster. What would you then do with that?

60 **Participant:** So I think that the main the main use case and the thing that I would kept trying to go after is. So if let's just say that the above ground living area is colored by cluster, then what I would. This would be the workflow. So I I visually can see. The different how this how the instances are separated and then I would take the take the cursor thing and. Highlight one of the clusters approximately and then the rest of the page would perform the segmentation across. All of the features and so. It would be like that step right where I like see the cluster in one feature and then I select those instances and then it. It performs them for the rest.

61 **Dan:** Okay, got it.

62 **Participant:** So maybe it's not as important as I'm thinking, but it's just like a small.

63 **Dan:** Right.

64 **Participant:** A small thing.

65 **Dan:** Right. Where like right now the plots are clustered. But then. If you want to like. See this cluster of instances across all the other plots. There's not an easy way to do that.

66 **Participant:** Yeah.

67 **Dan:** Okay, I see. All right, so what impact do the different rankings have on your model analysis and which rankings did you find to be the most or least useful?

68 **Participant:** Well, importance for sure. Very useful. That's probably like most in line with like the. Kind of level of detail of the work that I do, but then the histogram comparison was probably the second most important. The clustering. I think I don't have enough of like an intuition built up to like. Use that very quickly, but. Yeah, probably the histogram is like the nicest one.

69 **Dan:** Okay

70 **Participant:** to show those distribution similarities.

71 **Dan:** Okay, so, got it, makes sense so for cluster difference you felt like you didn't have intuition built up to make good use of that. And would the same apply to the to the highlighted line similarity one?

72 **Participant:** The line similarity. Probably because as I said before, like I don't, I don't use. I don't look at the ice plots like very regularly. But yeah.

73 **Dan:** Right. Okay. And are there any additional ways you think could be helpful to rank the plots.

74 **Participant:** I don't think so.

75 **Dan:** Okay. How did analyzing subsets or clusters of instances impact your analysis and did you find the clustering and the highlighting to be useful?

76 **Participant:** Yeah, I mean, both of them that I mean, that's like probably the, the, the highlighting I think is like the most useful part because it allows you to detect whether or not, it basically gives you a very transparent view into how sensitive the model is, and if that is something that is really important for the domain that you're working in, which for me it is, then I can just see and follow the rabbit hole down to is the model sensitive to this feature, is it just a couple instances that represent the the pole. And the highlighting really is the key there because that highlighting view just shows you.

77 **Dan:** So in terms like understanding like how sense of the model is can you just explain a bit more about like the role that highlighting plays into that?

78 **Participant:** Yeah. Well, in the. I mean, really just like looking at the most important feature. I could immediately see that there is like a couple of the lines that had a more severe down trend in value associated with growing area in the house. So immediately like that change in relationship doesn't make sense. Okay, now I can see that that change in relationship is exaggerated in a couple observations. And so now I can highlight those observations and see. I think the. The part that's really helpful is that when you highlight when you go to the detailed view and highlight those observations, I can see them on the scatter plot. And that that view makes it very clear that they're like outliers. And then the ICE plot shows that they're probably. Causing an overfit to occur.

79 **Dan:** Right. So I guess with this one, it was like highlighting...

80 **Participant:** Those. Yep.

81 **Dan:** ...these.

82 **Participant:** Yep. And then go to the detail plot

83 **Dan:** And then the detail plot.

84 **Participant:** And so I guess it doesn't. Yeah, those three are blue are green, right?

85 **Dan:** Mm hmm.

86 **Participant:** Yeah.

87 **Dan:** Right. Right. Yeah. So those are green, but then we also do have like more like. I guess we do generally see that

they're like the higher value homes with the exception of these that we're highlighting.

88 **Participant:** Yeah.

89 **Dan:** Okay. Yeah. So I can see that. And then I guess you also at one point, then like. Looked at like. So I think at one point, you like, yeah, yeah, I think it's like at some point you noticed that like these were like the like, had the higher overall quality as well and.

90 **Participant:** Yeah

91 **Dan:** Dove down that line. Okay. Got it. Okay. So the last next were the filtering capabilities useful and are there any additional ways that you would like to be able to filter the plot?

92 **Participant:** Yeah, I think that I use the filter pretty early to look at all of the two-ways for that main feature. So that was super helpful just to like clean up the screen. Other filters. I mean, I can't really think of a way. But the increasing and decreasing is is definitely nice. I'm trying to think of like other other ways that would be nice too.

93 **Dan:** So at one point, you highlighted lines and then used the increasing and decreasing. Was your thought that that was going to like filter them based on the high-, like the shape of the highlighted?

94 **Participant:** I think it was, yeah.

95 **Dan:** Okay. Which I think that's, I think that's an interesting idea. So maybe that would be like another way that the filtering could work.

96 **Participant:** I'm trying to think about what that would even do though. So that would like give you only the features that shared that relationship.

97 **Dan:** Mm hmm.

98 **Participant:** Yeah.

99 **Dan:** Or it'd be only the features where like the highlighted instances show like a decreasing trend or something.

100 **Particpant:** Right.

101 **Dan:** Yeah, because right now it's just based on the PDP.

102 **Participant:** Right. Okay. Yeah. And you said that. Right. I definitely thought it was for the highlighted. Okay.

103 **Dan:** Mm hmm. Okay. So apart from that, do you think are there any other ways you think it'd be useful to filter? Or are there ways to make the filtering by shape like more useful because it seemed like it wasn't like a, apart from when you like highlighted and then filtered it, I don't think you really like tried to filter by shape. So there are like reasons for for not doing that or ways that that could be made more useful?

104 **Participant:** Mmmmm. I don't think so. I think that it I think that it is useful, but I think that maybe it's also like. In this particular data set, it seemed like there were like six features that were kind of driving most of the predictive variance. And so I could see the shapes of those already so like

105 **Dan:** Got it.

106 **Participant:** the the feature importance rank sort was already kind of like performing that task for me.

107 **Dan:** Got it.

108 **Participant:** Yeah. But but if it's like, especially in a super high dimen in like a higher dimensional data set where there, where there's like more complexities. That would probably be really nice. Or more more useful. Right. And like,

109 **Dan:** yeah

110 **Participant:** my background is to is in medium sized. So definitely no bigger than like a hundred potential features. And in the end. I'm always going after highly explainable simple models.

111 **Dan:** Yeah.

112 **Participant:** Yeah. So different than, I mean, other people do different things.

113 **Dan:** Mm hmm. Okay. So next, how old the tool enable you to analyze feature interactions?

114 **Participant:** Pretty well, I think that the. I I think I really used it once to look at the interaction between basement area and above ground area and that just I guess sort of like reiterated the core the the pattern in the ice plots.

115 **Dan:** Okay. So what would you say the tool's, biggest weaknesses or limitations are, and are there any improvements or additional capabilities you would want PDPilot to have?

116 **Participant:** So I don't think it's lacking. I don't think it's lacking any features. I think that. It is first for somebody that mostly would do this analysis. Like in an notebook, selecting a subset of graphs to create. It's overwhelming at first to have so much visual visual visualizable at once. But I think that that is a small learning curve. So I think that given like more time. That's like not that isn't a reason that I wouldn't use the tool. Like

117 **Dan:** Right

118 **Participant:** Yeah, I will totally use this. Like, it's, it seems super useful to what I, what I do every day. So I don't think that like that, that isn't a weakness because ultimately. Like with any tool, like there has to be learning and investment at the front end in order to maximize its helpfulness.

119 **Dan:** Okay.

120 **Participant:** I don't think I could give you feedback to say like, Hey, you should make it simpler in this way because that would just like probably be taking away some of the functionality.

121 **Dan:** Right.

122 **Participant:** It's a good balance. Anymore. If like there was another like serious, if there was like another. Panel. So there's one way, two way detailed view. If there was like something else, I think it would be that would make it. Mm. Unnecessarily more complex.

123 **Dan:** Got it.

124 **Participant:** Maybe the SHAP values. A lot of people use SHAP values too. But. Yeah.

125 **Dan:** Okay. And then lastly, this is kind of like a catch all question for any other feedback that you wanted to add, but didn't necessarily fit into any of the previous questions.

126 **Participant:** I don't think so.

127 **Dan:** Okay. Great. So that's it. So thank you very much. I really appreciate you taking two hours out of your day to help me with this research. I greatly appreciate it. I know that's a big time commitment and I'm very grateful. So today I'll send you your compensation. So it'll be $100 Amazon gift card, which will go to the email address you used to sign up for the study.

128 **Participant:** Cool.

129 **Dan:** So PDPilot is open source and is installable through PIP. So let me know if you're interested in using it for your work or with your own data and model. I'm happy to answer any questions or to help you get set up. I'm also happy to make any changes to the tool to support your needs and I'm always interested in hearing about your experience, any findings and any feedback.

130 **Participant:** Cool, yeah.

131 **Dan:** Thank you so much, XXXX. I really appreciate it.

132 **Participant:** For sure. Yeah, I'll definitely I'm excited to touse it in my work. It's it's a really cool project. I think it'll be helpful for a lot of people.

133 **Dan:** That's great to hear. All right. Hope you have a great day.

134 **Participant:** You too. See you.

135 **Dan:** Thank you.