

Part 1

0:00 (Host): All done. So, guys, once again, good afternoon. Uh, I'd like to first thank everyone for agreeing to voluntarily participate in this stage, which is very important for my master's thesis, that I've been working on, applying it here within PicPay as well. Uh, and so, just for bureaucratic reasons, I just wanted everyone to raise their hand to confirm that they agree with the recording and that they also answered that consent form. Done. Perfect. Just to have it on record. And so, how the dynamic will work here, okay? Uh, I added you guys to this Miro group here. I'll also send the link here in the comments. You can log in with your PicPay account itself, right? Because I put that account there. Confirm if you can access it. Easy. Everyone joined already.

1:04 (P2 - Purple): I got in.

1:12 (Host): Can you guys, like, click on an element, uh, can you edit? Ready, you can click, right, and edit, write, like, inside the post-its. Ready, perfect.

1:28 (P1 - Red): Entering. Yes.

1:36 (Host): Uh, just waiting for Cala who said she had to step out. I'll explain the dynamic just once.

4:16 (Host): Well, to save time, I'll explain the dynamic. Then if anything, when P2 returns, I'll explain it again. Uh, here in this group session we'll have two stages, okay? Then I'll explain the first part, and when we enter the second, I'll explain it there. Well, first we're going to talk about four general topics, right? They are affirmative statements involving the central theme of my dissertation, which is precisely the process of continuous experimentation to compare the performance of supervised machine learning models, uh, besides during the process of model deployment, of evaluating if ah, uh, there are these two model versions, there are these two models or more, right, different ones, which of them will I deploy to production, right? So, the use of an experimentation process with the application of statistical tests to make this comparison. And so there are four affirmative statements involving this area, this concept as a whole. Uh, and the idea of this first stage is to collect the perceptions of each of you regarding these specific topics, right? So, if you notice, each topic is a square, we will go through one at a time. And then in that square we have the affirmative statement, okay? Which is in English, but I will read it in Portuguese and you can write the post-it in Portuguese as well, okay? No problem. And the idea is for you to allocate the points, right, the little circles, how much you agree or disagree with this statement and allocate the post-its writing your perceptions regarding points in favor or against this statement in question. Uh, so like, there's not really a right or wrong here, okay? The idea is that they are general topics and it's really to collect everyone's perception around this. And then I will give about a time of three to 5 minutes, right, for I will read the topic, then I will give this time for you to vote and write the post-its with pros and cons, right? And a detail is that if you agree with the statement, but you see, for example, some scenario, some point against it, you can allocate a post-it in the "against" and the same otherwise, okay? It's not like, ah, just because I agree, I can't write any point against or if I disagree, I can't place a point in favor. No, it's very open, right? The idea is that you really place your perception as much as possible, both in favor and against, right? Regarding the topics you will see. Then,

the only thing I'm going to ask is that you choose a color, right? I put five colors here, so choose a color and keep it from beginning to end, right? Everything will be anonymous, right? At the end of the research, like, it won't be "ah, (P1 - Red) did this and that", it's just so we can understand that participant one, right, and understand their whole perception regarding all topics, okay? Uh, everyone OK? Did you understand? P2 arrived around the time I was explaining, but if it wasn't clear I can repeat some part.

7:20 (P2 - Purple): Uh, we have to agree or disagree with the topics, right? And then the second part that I didn't -

7:28 (Host): P2, yes. Ready. Uh, we are going to go through each topic. I'll speak the affirmative statement in Portuguese. And then you are going to vote regarding how much you disagree and agree and will be able to allocate post-its in the pros and cons that are below the statement, right? Points in favor, points against, which would be more to describe your perspective regarding the topic.

7:44 (P2 - Purple): I understand. Uh-huh. Beauty.

7:52 (P1 - Red): I was going to say that. I think it's good to choose.

7:52 (P2 - Purple): Beauty, I'll take the purple. Don't throw me under the bus.

7:52 (Host): Everyone already chose a color, uh, yeah, everyone uh took a color, if you could write, like, um, just the, which then I'll omit later, right? Just so we don't risk someone taking the same color as the other and mixing them up. But then I'll anonymize it later, okay? Ready. Um, let me see. P4 took one already too, right? Ah, okay, no worries.

8:28 (P4 - Blue): Wait, I'm choosing here because I don't think I'm able to write. Let me see here. Yeah, but I think I showed here that I wasn't logged into the app. Wait. Now it begins.

8:44 (Host): Ah, no worries. See if you can do it now so we can begin. Ah, ready, now it appeared. Ready, perfect. Okay, so let's go, right? Topic A here, uh, let's focus on it for now. Uh, topic A, it says the following, right, that the use of statistical tests to detect significant differences in performance data and a direct comparison using a measure of central tendency like the mean of performance data, can lead to similar conclusions. If you didn't understand something about the statement, I can explain it better, right? In case it wasn't clear.

9:36 (P2 - Purple): Ah, I think I'd like you to say it again, because the second part wasn't very clear.

9:44 (Host): It wasn't very clear. Ready. Ready. Uh, regarding two ways of evaluating model performance. You take, for example, the average uh of accuracy, for example, between two models and compare them directly, or you run a battery of statistical tests involving groups uh of accuracy data. If using these two methods will reach similar results, right, similar conclusions regarding model performance. Did you understand? It's like this, imagine I have two machine learning models. Then I'm going to take and calculate the accuracy average, right? So during the test there, I took the metric, the accuracy and compared them directly. "Ah, this method, this model here has a better average accuracy than this other one here,"

for example. Uh, this is one, is a, is a way. Then the other is I have a group, for example, I took different test groups and generated 100 accuracy data points. And then I will take each model and apply statistical tests to see if there is a significant difference regarding accuracy. So it's like, will these two methodologies, will they reach similar results, similar conclusions regarding model performance? Is it better? Uh, (P3 - Yellow) can comment, you raised your hand.

10:56 (P3 - Yellow): Yeah, no, I was just going to ask if, like, when you uh, it's that I don't know if I understood very well, but for example, when you speak of this separation into test groups, if you think of comparing tests of the same audiences or kind of random like that?

11:12 (Host): Can you repeat the very end because it glitched out for me, but it was a problem on my end.

11:20 (P3 - Yellow): No, for example, you talked about comparing the accuracy of two models in a general way and or separating into several test groups and seeing their accuracy uh specifically comparing, right?

11:28 (Host): Ready. Uh, it would be in general really, like, I'm going to take different groups of data groups, right? Different test data. I will collect the accuracy from each scenario. Then I will have the group, right? Like, ah, I will have, for example, 100 different test data points. Then each one will generate an accuracy, then I will have 100 accuracy data points, you understand? For each model. Then I will apply the statistical tests.

11:44 (P3 - Yellow): Ah, okay. Understood, understood, understood. Thank you.

11:52 (Host): You're welcome. So, ready. Then, feel free to vote and place points against and in favor. Again, there's no right or wrong here, because it really is a very broad discussion in the literature. Uh, the goal really is to collect everyone's perception. Ah, P2, just mentioning that if you want to write in Portuguese, there's no problem, okay? Ah, don't worry. But you can write in Portuguese, if it's better that way, no problem.

12:36 (P2 - Purple): It's just that I saw it in English, I... Okay, okay. I think that's it for me. Are you going to do each one with us, we can go answering the others, okay? Done.

14:48 (Host): I'll do one at a time like this, just waiting for the others, since I'm timing here more or less 4 to 5 minutes for each. I think everyone has finished. Is anyone going to write any other point against or in favor? So, I think everyone is done, right? Right. Um, let me just go over it in a general way. It would be necessary to guarantee the isonomy of the experiment comparing the same groups. Uh, separate test groups are not comparable to the whole in a general way. Uh, the averages, um, just zooming in here for me, the averages might even be identical, but if one group has a huge standard deviation, the other doesn't, the test can take this uncertainty into account. And are they, are not directly comparable. I think I said it right. Uh, does anyone want to add anything, explain some point better?

16:12 (P2 - Purple): I think the part about comparing the averages is easier, right? But I think it's the point P4 made, right? It was P4, right? The blue uh you disregard, right? like, the behavior, right? Sometimes it's bimodal and you just put an average there and anyway.

16:36 (Host): Understood. Right. Um, does anyone else want to add anything or is that generally it? Is it perfect? So, let's just go to the second topic. Uh, topic B, it says the following, right? AB tests are a suitable approach to compare two versions of machine learning models in a production environment. Uh, did you catch the statement? Does anyone have any doubts? Don't worry. So, feel free to allocate and place pro and con points, okay? Remembering that if you agree or disagree, you can place post-its on both sides, okay? It's not mandatory to be just on one. So, feel free.

22:00 (Host): Everyone finished, is anyone still writing, no? Okay. So, I believe everyone finished. So, I'll do a general reading. It's going to be cool that there were pro and con points. Uh, I think it'll be cool, that will catch different views. Uh, so, starting with the pro points, right, around this topic, uh, you can automate the model update decision or not. And models can degrade over time and updating them based on an AB test makes the process more agile. Uh, then another one here, conceptually it's the most appropriate way of comparison. Uh, AB testing is the gold standard, enables capturing behavior with real users and more live data. AB tests are more reliable ways to compare and evaluate models, so they are exposed to the same audience profiles and conditions, being one of the main parameters used for decision making. Regarding the pros, does anyone want to make any addition?

23:08 (P2 - Purple): No. M.

23:16 (Host): I think it's cool. So, I'm just giving it a second read to see if there's any other point... over time. Updating makes the process more agile. Then I'll give a... ready. Uh, just one point from this yellow one, just to see if I understood the statement correctly, right? When it says: "So that they are exposed to the same audience profiles and conditions." Uh, what does that mean? Uh, would it be regarding being able to better capture differences or other parameters due to the same profiles? Just to better understand what was meant by the statement.

23:56 (P3 - Yellow): Yeah, exactly. Like, you apply the model, you choose the model, for example, for the same profile, same audience. Exactly. So, like, you can compare uh if there's any different performance of the models. I use this quite a lot, like in fraud, at least.

24:12 (Host): Understood. Understood. No, beauty, alright. Ah, perfect. Okay. Going to the con points, uh, automatic model updating, for example, based solely on the AB test, may not take into account risks that the business can afford to take. Example, old anti-fraud model approves more, but lets a little more fraud pass than the new one, but the tradeoff would be worth it. Uh, then the red one here, let me just get this to zoom in. Ready. Uh, we don't always deal with the ideal scenario and depending on the volume required for the test, it's possible to generate a negative impact on business metrics and cost/complexity, okay? Uh, does anyone want to add anything? I will stress some points here, but I wanted to know if anyone wants to make any comments, okay?

25:16 (Host): Uh, just so I understand, right? Ah, old anti-fraud model approves more, but lets a little more fraud pass than the new one. But in the tradeoff it would be worth it. Just a little bit...

25:32 (P2 - Purple): It's this one I wrote. It's like this, the model degrades, right, usually Y of anti-fraud, right? Uh, but sometimes, for example, the new model even catches more fraud,

for example, ah, a better recall, right, a better F1 score and such. But then the business says: "Man, I prefer to have 99% approval like this previous one is giving me and have like I don't know, besides 40% recall, have 30% because it's compensating, you know, in the tradeoff." So leave that one. Sometimes these are things we don't put... we can even put it in the flow, but just by itself, right? Like AB test normally, without some optimization, you understand? Without an optimization beyond that, it wouldn't solve it.

26:28 (Host): Understood, understood. It became clear. Uh, and then here from the red, right, we don't always deal with the ideal scenario. Uh, and depending on the volume required for the test, what would this ideal scenario be? Would it be tied to volume or another characteristic? Ah, okay.

26:44 (P1 - Red): What I put is that I placed it sort of as if it were a single text in the pro and the con, okay? Uh, it's because it's as if the positive side is that it's the ideal scenario for us to use the AB test, but that we are always dealing with the ideal scenario. What does that mean? It's that, for example, people gave the example of fraud, I have some models that are propensity models, uh, and they directly deal with money, for example. So, in this case, we need, in some scenarios, a large volume of people inside the experiment to be able to validate the experiment, because the conversion is low, for example, of a given product. But then if we take too large a volume and the model that is testing it doesn't perform well, this means a loss for the business that is testing it, you understand? So there's this negative side of the AB test in this kind of scenario.

27:40 (Host): Understood, understood. Cool. And here cost slash complexity and would it be related to a little of these points or would it also encompass another aspect of what wasn't commented on?

27:56 (P4 - Blue): No, I think it has a good connection with what (P1 - Red) commented. Here in PJ (Corporate accounts), at least, we live in a somewhat opposite scenario, right? We don't really have a very large base to be able to test. And the channels we have to use for carrying out these tests, they are very complex, because everything is still very manual. So, for you to have an idea, like, we've been trying to address a price elasticity test here within the CRM channels of PJ for a few months and it doesn't go, man, it's not possible to do the test because we can't, we can't, for example, parallelize the rate. Uh, it's not possible to put, for example, two models to the test against each other, because our whole flow, our pipeline here is still very manual. So, we deal with both this complexity issue and also the cost issue of knowing how much we can afford to err. Anyway, there are few people we can test with, there are many variables, but it's more in that sense.

29:00 (Host): Understood. Perfect. Um, great. Um, anyone else want to make a general addition or can we move on to the next topic? Okay? So, let's go to topic C. Uh, topic C says the following, right, that there is a challenge in adopting statistical testing practices in data teams due to the learning curve that professionals face to use them correctly. Yeah.

32:56 (P1 - Red): I put it in the middle of this axis division here because I couldn't, I didn't feel that this topic fit very well into pro and con, okay?

33:04 (P3 - Yellow): I was in doubt too.

33:04 (Host): No, it's alright. No, no worries. If anything during the discussion we can capture, right, the parts without a problem.

33:12 (P4 - Blue): I wrote two cards that if you read will be the same thing.

33:12 (Host): Ah, alright.

33:52 (Host): Is anyone still writing or has everyone finished? Just to know, okay? I think everyone has finished, okay? So I'm going to, since it was very pros/cons, uh, I will do a general reading of the pros and cons, uh, and then we'll do the general discussion. Uh, so, PROS: We are often inserted into the data field without prior knowledge, existing the need to study complex concepts on our own. Uh, data profiles with a good background in probability and statistics topics would have an easier time. Uh, the concepts are not trivial and with the tools/libraries we have today, we can even reach a result baseline, but it's necessary for the analyst, the scientist to understand the test premises well, if the test premises were met. Data area is very multidisciplinary, depends a lot on the person's background to have more or less ease, but the market ends up influencing more practical development with less basic learning, quote unquote. Uh, it depends a lot on the data person's profile, there are people from different knowledge areas acting in this context and we have ML tools and experimentation platforms that automate a large part of the process, not requiring as much need to be an expert in the calculations. Uh, does anyone want to make a comment before I go deeper?

35:20 (P4 - Blue): I think... you can speak.

35:28 (P4 - Blue): I was just going to give an overview of my two cards, because as I said, it's almost the same thing. I think in short everyone said that the data field is populated by many people coming from different contexts. For example, on my team there's a girl who is a biologist, so she wasn't necessarily born here, she didn't arrive here fully understanding all these concepts. So, uh, it's necessary that you keep studying there, evolving to be able to grasp everything. But, on the other hand, we also already have a lot available there, libraries, tools, platforms, which somewhat minimizes the need for us to know the math behind things so deeply, right? So it's kind of that balance like, man, it's like us using generative AI today, you kind of have to know how to validate well, man, uh, I don't know, folks are struggling with this nitro because they're trusting that it'll run everything, know everything, but they're forgetting to validate. I think that's sort of the key point. Yeah.

36:32 (P3 - Yellow): About mine, I even wanted to use myself as an example, because I got into data science, I started in an internship program and I got in without knowing how to program, without knowing absolutely anything and then I was inserted into the data field without knowing anything about statistics, without knowing anything about anything. I was learning everything in practice and, like, I evolved a lot over time, but I realized I had many limitations, because within the daily work routine, people rarely talk about it, you end up doing what people tell you a lot and, like, you can't often get deep into the technical knowledge and everything. So you do what you're told and that's it. But you don't understand why you're doing that, how it works. And that happened a lot to me. Uh, it's been a short time since I decided to really study, to understand the statistical part more, understand more of the data science concepts themselves, because I realized it wasn't cool, you know? Because

I, I was really very limited within data science. I realized I was a good scientist, but I didn't understand the technical concepts very well.

37:44 (Host): Understood, understood. Yeah, I think it very well complements the vision that basically everyone commented on, right, regarding I think even the red point here that ended up pretty central, right, which is about this issue of basic knowledge, right, basic learning, that the market ends up, right, in a way devaluing a little of that. Uh, and then it pairs with the tools issue, right, that, uh, as you guys commented, it already automates, facilitates, right, something more practical. Eh, and then in your view, for example, uh, is there something, even taking this market point, right, regarding maybe not valuing this. Do you see that there is this devaluation or is it more like something that isn't a priority, but if the person has it, they can generate value, how do you see it that way? Did you get what I'm exploring?

38:48 (P4 - Blue): I think I understood, but I don't really agree. I believe that this area we are in here is a highly specialized area. So, I think like, further back, the folks who managed maybe to get by without much, I don't know, like, without getting too deep into this technical line, maybe came like (P3 - Yellow) did, like, from an internship and kept evolving there inside the company itself, but I think today, for example, you don't see as many junior, mid-level vacancies, they already demand a much more specialized crowd. So I strongly believe that they are still very worthwhile knowledge for us to study, to deepen, even thinking that right now almost everything boils down to LLMs, right? Like you being an AI engineer... But uh I don't know how long this will last and I don't know, in the end we'll end up realizing we aren't generating value with anything and that we'll continue investing in traditional modeling anyway and faithfully, you know? Uh, I don't know. I think I doubt a bit that, like, we don't need to know these things. I think we do have to know.

40:00 (Host): Cool. Uh, does anyone want to add anything to this point? No, cool. Uh, another thing, ah, was someone going to say something? Ah, don't worry.

40:16 (P1 - Red): Yeah, I echo the sentiment here because it's beautiful. Yeah.

40:24 (Host): Uh, then also just one last point I wanted to uh explore with you guys before we go to the last topic of this stage. Uh, do you guys feel how do you feel regarding, for example, uh, statistical testing practices, thinking of a large company like PicPay, which has different data teams, uh, do you see that there is also this some challenge precisely due to having different teams with different models, managing, for example, to have a standardization of how this process is or like wow, there's no way to standardize. What is your view regarding this? Which is a bit tied to the learning curve, but thinking more in the macro there, uh, of the company, right, with several data teams. I don't know if you could understand my point.

41:12 (P4 - Blue): Understood. But this question is quite difficult, because I think it goes beyond our scope here of like modeling, data science. I think there are some peculiarities of the business itself, as I mentioned, right? Here in PJ it's a very different universe than PF (Personal accounts). We're kind of still like, you guys ran so we could walk. So, there's a lot we can't do yet. So, like, uh, for you to have an idea and continuing a bit that chat about the price optimization model that I mentioned to you guys, our alternative, like, is now to be able to execute some outline of a price elasticity test is using the history. So, we leave the price running for a few months, build a history, change the price, leave it running a little longer and

compare uh one period with another. It's not the best way for you to execute an AB test, but given the circumstances we have here, like all the problems we have to get it to run, it's what's showing itself to be feasible. So I think it's a problem that goes a bit past us wanting to like know a lot about techniques and building models. I think it's like much more basic than that, you know? Like the difficulty. So I think it's hard to answer this question.

42:32 (P3 - Yellow): Yeah, I think they are really very different contexts, but even bringing it a bit closer, I don't know, thinking a bit about the PF world, even so the people's modeling ways are very different. Like, you take even within one team, we are trying to standardize the modeling way now, you know? Like, we never managed to standardize and to this day the modeling way is not standardized. I think that uh there is also a lot of individuality in the modeling way. I think that yes, there has to be a more basic standardization like that to unify the understanding of things, but I don't think it's at that level yet. Like, I think there's a long way to go still, as she said. Like, we still worry about databases not being updated, thousands of things before thinking about standardizing the modeling way, you know?

43:36 (Host): Right. Understood. Well, thanks for the points. Uh, I think now we can go to the last topic, right, of this stage. Uh, well, topic D, right, which is the last one, states the following, right? It is safe and recommended to completely automate the promotion of a candidate model to 100% of the user base if it wins the hypothesis test with statistical significance compared to the previous version. If there's any doubt about the statement, I can clarify.

44:08 (P2 - Purple): I was answering, I already ended up answering based on what we were talking about before, you understand? That's related, right?

44:24 (Host): Yeah, yes, yes. There were some comments in past topics about this automation part, so it's very tied in indeed.

46:20 (P1 - Red): I read your comments and changed my answer slightly. I'm influenced.

46:28 (P2 - Purple): Me too.

47:28 (Host): Everyone already wrote awesome. Uh, the points here are all somewhat disagree, just to confirm there is a yellow one here that is floating, okay? I'll just drag it closer here so when I analyze it I remember. Ready. I'll read it generally.

47:44 (P1 - Red): Yeah, hey, it's because I tried to put it on a scale, okay? I partially disagree, but I'm leaning toward disagreeing altogether.

47:52 (Host): Ah, yes, understood. No, alright. Uh, then I'll read the pros and cons here. Um, well, the pros, right? If it were only from the technical side, I would agree, but that's not all that drives the decisions. Uh, cons very mature, business, technical team are on the same page, uh, I don't believe it's ideal to say it's 100% safe. There are other necessary steps before pushing the model to 100% of the base, like being sure that the production variables work well and the model stands up. migrating 100% of the base makes it impossible to monitor this model over time. Uh, promoting 100% immediately is quite dangerous, even more based on just one hypothesis test. Uh, does anyone want to add something? (P1 -

Red) there who was kind of disagree, part disagree, if you want to add, take the chance to add.

48:48 (P1 - Red): I just want to see. Uh, it's because I put it thinking a lot about my case and the type of model I make there. Then I saw the example especially of P2 there. And it's more or less the following, uh, in the scenario, for example, of a propensity model or you're going to send a message, going to see the guy to to hire something, you will never migrate 100% of the base and you need to have some control group that is comparable so you can know, even for you to measure if the model isn't degrading over time. Uh, and and see this kind of thing. There are other issues too, but this one especially. But, for example, if we're looking at fraud, for example, it makes perfect sense. You migrate 100% because you don't want to let any fraud slip. you don't, the comparison becomes less important in this case. I just wanted to make this differentiation.

49:44 (P3 - Yellow): But I think even in fraud it's very dangerous. We don't deploy to 100% right away no. At least here on the team we don't. Like, for example, we have a model in production already and then we want to push a new one that, anyway, the old one is losing performance, and all that, we want to push a new one. we don't push to 100% straight away, no. We go pushing it up slowly, like 20%, 40%, and so on, and then we go evaluating, because then, for example, you have an attack or something like that, we're going to see how this model will behave too, you know? If we're not going to take, if it doesn't end up being too permissive as well, you know? It's true that there are many policies behind it too that will help us make a decision or not. It's not just the raw model that we use here either, right? But the policies are very important for us here. But that's it. I'm saying this because we've already done this in the past of pushing both policy rule and model to 100% and we did poorly.

50:52 (Host): Ah, understood, understood. Yeah, but then just to explore your point further. Yeah, supposing that, for example, it wasn't 100% of the userbase, right, it was another percentage, still the automated process would be uh safe and recommended in your view or no would it still need to be handled differently?

51:16 (P4 - Blue): No. And I think this is a bit of the complement of my answer. I put that I partially disagreed because in the sentence the automation part, I think it is the end goal of our whole pipeline here alongside MLOps. But when you go to the rest of the sentence where you have this idea of promoting 100% to the base all at once and relying solely on a hypothesis test, things get kind of weird, because I think even if we did, I don't know, a partial deployment there and we push it little by little to the whole base, uh in production it's a little bit more complex, right? So, if we rely here uh on a hypothesis test, we would only be considering maybe here or focusing on average or an aggregated view. You might be missing a specific niche, for example, in (P1 - Red)'s propensity models. There might be some group for him that is much more interesting that uh it won't be able to capture. So, I think this hypothesis test thing is also something we would have to pay attention to, you know?

52:28 (Host): Understood. Does anyone want to add any other comment? No. Alright.

Part 2

0:00 (Host): Awesome. Uh, and well, what's going to happen now, right? In the first part we had more of a discussion of topics about the experimentation process, right? gathering your perspective regarding the process, challenges that you notice, right, existing in the organization. And now, right, in this second part, I will show uh the candidate solution that I made to help in this investigative process, right, of the experimentation process, which was applied, right, in one already in a context here inside PicPay, uh, which was to do a small valuation, right, which was it was a V1 MVT, right, let's say, a prototype. And then the idea is for me to show you how this tool works, uh, and then subsequently collect your perspective, right, regarding, like, ah, what do I feel, right, that will add, right, in what does the tool add, what do I feel it hinders more than it helps or in the sense of like scenarios that you see where it fits, that it is useful, right? So, anyway, uh, that's the idea, right? I'll do this presentation and then there'll be this part that's similar to what we did in part one, right? There will be three squares, right? Each one focused on one aspect of efficiency, usability and how easy it is to use. And then it will be the same stage, there will be the voting, right, of how much you believe it meets, right, and pros and cons, okay? And like, uh, the idea here isn't for you guys to feel cornered like, ah, criticizing the tool. In reality, the ideal is that you really are very transparent on the points, uh, because it really is part of the scientific method, right? The ideal is here's this tool, right? What is it, what is it good for, what does it not serve, what doesn't make sense from your perspective. So, the idea is this moment, okay? I will share my screen there, so, I will explain the tool. If you have any specific doubt, you can uh ask me on the spot like that, raise your hand and ask, okay? Uh, but like, before I speak, show in code, right, its usage, that I will even show the small evaluation, right, that we applied, uh, what's the idea of this tool, right? It was a Python library, okay? That we call MLXP, right? Uh, initially it was XML, but then I inverted it because it's better, right? Machine learning experimentation, right, the acronym. So, what's the idea behind it? The idea is for it to be a facilitation of automating the process of applying statistical tests for you to compare performance, right, of supervised machine learning models. Uh, so this flow here that is on Miro as well and available for you guys, what's its objective? It will uh work as follows, you will have the supervised models, right, which can be different models. Like, look, I have a classification model that I used a different algorithm, right, compared to this other one I have. Or they can be models that, like, ah, they are from the same base, but I use different features, right? They possess uh the use of different features on input, right, to produce the same output, anyway. Or it can literally be the same model, but in an improved version, like, ah, I have this model, V1, I retrained it, I have a V2, then I want to make the comparison. So, these cases, the tool covers, right? Because how does it work? Uh, first you will map with it what we call contexts, which is like you saying, look, these are my test data and then you can indicate both by file as well as by dataframe that's inside there, like in the notebook, right, at memory level. So you, look, I have this test base that I'll call test one, for example, right? N, so you manage to map all your test bases, uh, and then afterwards you will say which model goes to which test base, right? Uh, so, like, ah, I have this model uses column A B C. So, this is a context, right? I have this model to apply to this test base of mine that I mapped, that it, right, is for this context. Then, ah, I have this other model that uses column A B C D. Then you will have this other separate base and will say: "Look, this model here is for this base here, okay?" in the when I show the code it will help a bit to maybe convey this better, but basically you can direct which models go to which test bases, right? And like, if there's the case of being the same model, same model no, sorry, different models, but using the same test base, you can also do that, okay? And the models you can pass both by file, right? It accepts pickle format as well as ONNX compression, which is more general. And

also it accepts you passing loaded inside it the cases from SK Learn, right, which are what is implemented in the tool so far. Then after you map the contexts, then you can hit run, right, trigger it to run. Then what does it do? Then it will be these steps here that are numbered, right? Uh, for each context, that is, for model and test base, it will take and generate k-folds, right, of test data, which by default is 100, right? So, it will take and generate from this base for each context, right? It will take the test base and generate those folds, right? Uh, in case of classification, it uses a specific one, I think it's the stratified, if I remember the nomenclature correctly, that it tries to leave it proportional, right, the amount of of classes there, uh, in the same fold, right, so as not to have much unbalance. Then it creates, right, these folds. So, let's consider it's 100, just to make it easy to understand, right? So, there are 100. So, for each context, that is, each test base, it will generate 100 groups, right, of these test data. Then it will run, it will apply your model, right, to make the prediction and will compare, right, then make the comparison uh with the expected, right, to get the performance metric. So, let's suppose a scenario that I am talking about classification, I want to know the accuracy. Then it will generate, right, let's say I have, I'm comparing two models only. Let's take a simple scenario, it's two two models that have different features, so they give two contexts. Then it will take, generate 100, right, data of groups, right, there from all the test data, will apply the prediction. So in the end here, for each context I will have 100 performance metrics. Could you follow? Because I will have one metric regarding each fold. Is it understood up to this part?

6:56 (Host): Awesome. Uh, then, great. So, in the end, I'll have uh as if I had two two lists, right, with with 100 values, right, which is even what I say here, look, I'll have n performance data groups with values, right? n is the number of contexts, so if I have two contexts, I will have 100, assuming K is equal to 100, right, used K and 100 folds, I'll have 100 and performance data points, right, in each of these groups, right? Uh, and then after that, in step two, it will apply these lists, right, of performance metrics, of values, right, of performance metric inside this flow, which is what is here on the right side, right, which is this flowchart, that basically it will apply a series of statistical tests according to the scenario, right? So, first it applies Shapiro Wilk and Levene to understand characteristics of normality, homoscedasticity of the groups, right, of of performance data. And based on whether you have just two groups or more, it directs. That in this case, the group I speak of here is these performance data groups that are tied to the amount of contexts, right? That is, ah, right, how many models here I'm comparing, how many scenarios I'm comparing. If you're comparing two, it'll direct here to the left. If it's three, four, five scenarios, right, five contexts there, different models, it'll already go here below, right? Above two, it comes here to the right part, right? Considering the context of two, right? It will, if the, if there's an 'and', right, between normality and homoscedasticity, it applies the Student, right, which then is the indicated one when you have this characteristic. And if not, right, if you don't have normality or don't have homoscedasticity, then it applies Mann-Whitney, right, that the idea here is to detect if these two groups, right, of performance data they possess significant differences, right, and then basically it closes this statistical flow, right? I'll say what happens when it closes the flow, right? But this is the context of two, right? Ah, if I have three models or more, then it's the same thing, it will check if ah, based on Shapiro and Levene is there normality and homoscedasticity among these groups. There is. Then it applies ANOVA to detect in a general way if there's a significant difference. If it detects that there's a significant difference in these groups as a whole, it sets up the pairs, right, between these groups to apply Tukey, right, to detect exactly which pairs possess significant differences, right? Uh, and then it

closes, right, after it applies this test. And for the case of not having normality or not having homoscedasticity, it goes to Kruskal-Wallis, right, to detect significant difference. If it detects that generally there's some significant difference, it does the same strategy, right? It sets up the pairs, right, of the groups applying a Mann-Whitney to find which are the specific groups that have significant difference, okay? And then it closes, okay? Then, basically what happens? Ah, just moving forward. Was the test flow understandable? Anyone have any doubts? Cool.

10:08 (Host): Uh, then what happens when it closes the flow? It will uh sort of generate a report, right, which brings all the statistical test results, right, brings an HTML, right, assembled, uh, exactly bringing these statistics. In the end, it tells which are the groups, right, the scenarios that were significantly different. It also brings some statistics of uh central tendency there, right, like ah, accuracy, ah, it will show like what was the standard deviation mean, right, regarding the groups. So you also manage to have this view and automatically it does a check there which is like this, it takes all the groups that had significant differences and it takes the median comparing to say which is the best context, right? In this initial version it does that. So, supposing there were four four models, four contexts, right, different. Ah, let's say that in the end it detected that there were only two there two contexts, that between them there is a significant difference, the others didn't have significant difference. Then what does it do? It takes the median of these groups and makes a direct comparison. And then depending on the metric, right, if higher is better or lower is better, right, it depends on the metric, but assuming it's accuracy, right, it will take the highest accuracy, say, look, the model, right, that had the best accuracy median was this one here, based on significant differences, right, it does this automatically. Uh, but anyway, right, uh there is it brings all the rest of the report too, in case you want to do a slightly more specific, detailed analysis, right? Then, beauty, this is the most general explanation of the tool's functioning. Uh, the repository is already here on Git available, including, right, after this, if you eventually want to play a bit, mess around, uh, it's available. Uh, but showing you the code here, right, on how it's used. I took a scenario that was Pix fraud detection, okay? the context. Uh, and then I replicated more or less what had been done, right, previously, which was with four models, right, and basically I went loading these models and the test bases, right, replicating the same processing, right, in this case, the test bases I pulled really from the tables, right, but the models I pulled from the ML Flow, right? So, just showing you guys here, a good part of the code is very uh like this, ah there are installations here, the library, which I sort of separated a bit in markdown, just to make it easy to understand, right? So here I load the model, right, respective which was which is the first case. Then these lines below here I preparing, doing the same data processing, right, that the data scientist did. It was practically Ctrl C, Ctrl V, right? You have your same data frame in the end. Uh, and then when it arrives here it's this whole stage, right? It arrives at this part here where I will have the test base of a test base, right? Then what did I do here? It's the part where MLXP enters, right? It's going to do the following, I'm going to ah detail I forgot to mention, right? You can place more than one metric if you want to analyze. Of course it will run that whole flow for each performance metric you include, right? Uh, so here I will have accuracy and ROC AUC, right? And here you can choose the report name, right? That it puts the default name, but if you want, you can also customize this, including the location too, okay? Uh, so I left just the custom name and here is the part where I map this test base. So, look, I have this test base that I'll call Pix CC1. Uh, I have this X test here and I have the Y test, right? So I passed exactly the X and the Y here from the test base, right? Um, to it. And here later I already add the context,

right? Which basically this base I'm going to use just for this model that I just loaded in this context. So, right? Then I used the same name here, right? The context name I put the base's name just to, right? Speak, because really for each base I am loading a single model, right? There isn't a case where the same base I will use on different models. So that's why I left the same name, just explaining to you guys. So, here I put the context name, say what the trained model is and I indicate what the name of the test base is, right, that is that I will apply this model. Then, in this case, you use the same name that you mapped here, understand? Then ready, added the context, right? But again, right, it will only run that flow when I hit run, right, which will take all the contexts you added and run the logic there, right? Then here is the same thing, I do the second Pix case which I load, right, the model, load the base too, replicate, right, algorithmic processing and such. And here is where I do, right, add the test base, same thing, right, pix2, I pass the X test to a respective test, then add the context, right, passing the loaded model in this other model and indicating this other test base. The third Pix case, same thing, right? Prepares, prepares. Then here is add test data add context, see? It's the same little thing. It working the same way. And finally, the fourth case, right? There were four models here that I took. So it's the same process. Then I do the addition of the fourth model, okay? Then when you have everything, you simply give it point run, right? Then it will execute and then it generates the reports folder, right? By default, if you don't inform the location, it will generate there here is the Pix Case folder, right? the name there. Here it has, anyway, there I already ran it several times, right? It creates a folder for each execution, okay? Inside here, that then you can kinda have a little bit of this control. Uh, then I'll take here, for example, then how does it bring the data, right? It brings the JSON of that it generated of the data, right? So, it brings a summary here. You can format it here. Let me see if it'll be better or not? And there. Doesn't give formatting here, no. Okay. Uh, anyway. So, it will bring the accuracy, right? Then it will show here, look, the model, right? In that it isn't necessarily a model, it's more context, right? But anyway, the context here, Pix CC1, right? The average was this here, standard deviation was this here, median was this here, minimum was this here, maximum was this here. Anyway, it sort of brings the measures of central tendency of the groups, right, that was generated. And it also brings the JSON with the AB test results. So, if you want, you can read, right, in a way, like if you want to interact with this JSON, via code and such, you manage to get everything it generates, right, of intelligence there, of statistics via JSON, right, and it does this for each metric. So, it generated one for accuracy and another for ROC AUC, okay? Uh, and there's an HTML that here it probably won't uh won't show. So, let me download it. I'll open it here just for you to see. Ready, it comes like this, right? So it will bring first separating for CAD score. So it's, in this case here, this version I ran, I ran it with just three, but it's okay, no problem. Uh, it will bring here organized, right? It will first bring the contexts' statistics, right? So it will show those same uh measures, right? Of central tendency. It will show how that flow went, right? So it'll say: "Ah, arrived at normality with Shapiro". It saw that there were more than three uh three or more models, right? It saw the homoscedasticity with Levene as well. And then it saw, right, that it was false, right? That is, the data weren't normal or weren't homoeda-homoscedastic, that's how you say it. Then it performed Kruskal-Wallis and it simply gave done. Probably it gave done here because in Kruskal-Wallis it didn't detect significant difference. Probably it was due to that. Then look, Shapiro it will show here, look, Shapiro how the result went. Levene Kruskal-Wallis, right? So it will show for each one. Then really look on on Kruskal-Wallis gave none, right? The significant there ROC AUC ROC AUC it showed it the following way too. Same thing the tests here, right? It was very similar, right? Kruskal-Wallis too, Done. And then in the end it gave not significant. Then it will show the

significance conclusions, right? It didn't detect any significant difference, right? And here it won't point to the best model, okay? Then I'll grab now just to show a case that has the four models. Let me see here if this one has all four. Here is PIC C1, C3. Um, C3. This one I think not yet. Wait this one has four. Yeah, this one already has four, I think. So, let me download it here. Download open here. Let me see if Ready, that's it indeed. That's what I was wanting to show too, look. Then, look how interesting, in that, in that one there where I took just PC, these three here, they didn't yield significant difference neither in accuracy, nor in in which is the if, right? When we take all four, right, which is the drop fit resum, then you see, right, that really, like, the drop fit resum it eh it appears here, right, again, right, one one more line here that will appear. Then the flow in the accuracy case is very similar. In Kruskal-Wallis it will give no without significant differences for accuracy, right? So practically very similar to the past result. Only that when we go to ROC AUC what will happen? it will detect significant difference, because it will pass in Shapiro, will pass in pass in Levene, will see that there are three or more models, will eh will detect that eh it doesn't exist, right? I mean, it has either a normalization problem or issue of homoscedasticity of the data, right? it will enter in false, then it will check Kruskal-Wallis, then it will detect that there is significant difference in Kruskal-Wallis and will perform the Mann-Whitney to detect which groups, right, which pairs eh really have the significant difference. Then it brings all the results here. Then in Mann-Whitney it shows like this, let me take this out of the way, eh it shows here, look, Mann-Whitney between model such and model such, right? Then it shows the result here and if there was significant difference. Then, anyway, it sets up all the pairs here, right? Then what it will detect between the one, the first model and the fourth is basically the one, two and three. They don't have significant difference among them, but each of them has significant difference with the the this fourth one, right, which is the drop features, right? Then here in the conclusion it shows like this, look, that there wasn't in the accuracy, but it detected, right, between the models with Kruskal in ROC AUC. And then, eh, the best context, right, it won't show anything for accuracy, right, because there was no significance, but the best model based on the median, it will say that it's Pix CCC2, right, with median is around ROC AUC, okay? Anyway, it will basically compare, right, the ones that had significant difference. So, it takes, look, this one compares with this, then this compares with this, with this, right, the ones that have significant difference to take from this context, from those contexts, right, which ones had the highest median, right? Uh, but anyway, this here is more of an automated conclusion that it brings, right? But you have here as well the transparency, right, of of these statistical results. Uh, when I did, right, this here in the context there, it matched the conclusion too that the data science had made, right, that she really had detected that the fourth model it, this fourth here, it was more different, right, than these other three here. And she ended up choosing, right, from these three here. I don't remember now off the top of my head which was the one she chose, right, but she ended up choosing within these three, because this here has a significant difference, right, but it was worse. And this here, these other three here among themselves, they were similar, but then due to an a business decision, ended up choosing the other one, right? A specific one. Then just to close the explanation, this tool also has a command line, right? Okay? That I show even here in the README, right? That the command line, basically you can pass the paths, right? Of of the test bases, right? And also the contexts, right? So you go kind of passing in pairs there, right? Uh, kinda all this we did via code, right? In the notebook, you pass through here, right? The limitation of the command line is because it has to be via file, right? There in the code you can pass the loaded object, right? Uh, here no, you have to pass the model directly, right? And the test bases in file format, okay? The data included, right? Uh, I even

forgot to mention, it supports the same formats that pandas supports, okay? which is CSV, parquet anyway, everything that Pandas loads it accepts, okay? Uh, that I had forgotten to mention that. What happens? When it runs this here via CLI, right? You manage it will generate the report the same way, but you can also set it up to return which is the best model, right? So, like, it returns in the console output uh the name, right? So, ah, PixC2 is the best, understood? It's that this here is more aimed at, for example, ah, I want to run and sort of put it in an automated way for it to do this check, compare and see if this this new, this new model version, for example, is the most appropriate aiming at this mechanism of it, right? So that's it, right? Regarding the tool's general functioning, the main functionalities. Uh, I don't know if anyone had any doubts or if everything is OK. Could you understand? Awesome. Uh, I will send the Git link here, right, just for you guys. Later I'll send it in the group too. I'll send it here. Uh, then here what are we going to do now in this second stage? Uh, same thing we did previously, right, in part one. Uh, so in topic e here of the board, right, we are going to, in this case, I want you guys to bring your perspectives based on what I presented, how much you believe the tool, right, it is useful to support the the experimentation process, right, with machine learning, right, here it says during deployment, right, which is during implantation. If you see that, ah, I don't see it so much for implantation see it for another scenario, you can also complement it this way, okay? It doesn't need to be restricted to just what's written here. Whatever you feel in your heart that meets, doesn't meet, you can eh vote, right, and put the pros and cons, that the idea here is exactly that, to understand based on this structure what it is that in your perspective makes some difference or doesn't make much difference, right? Anyway, that's the idea, okay? Then, guys, feel free to vote here in disagree, agree, right, when you agree with this. And pros and cons, right? Same mechanism we did in part one.

25:40 (P1 - Red): Host, just a doubt. Uh, did you say to put it only in e or put it in all? Okay.

25:48 (Host): Hum. Go ahead. Put it for now just in E. Will like for each square at a time.

29:12 (P1 - Red): My block, it froze on the cons, but it's not on the con no my little block here, it got stuck, I'm not able to move it around, but it's... I did it. I did it. Calm down. It's weird, but anyway, fixed it.

29:20 (Host): No, I didn't understand. Want to throw it where... it was able to be on top, I don't know. No worries, no worries. Um, everyone wrote already or is someone still going to write more? Fine?

29:56 (P2 - Purple): Yep.

30:04 (Host): Okay, so, four photos in agree. Then I'll start with the pros. Uh, I can't say in the deployment process, because I've never deployed models. However, in the general experimentation part, I believe that the use of the tool is quite positive to facilitate and assist in our analyses. Uh, in the context of models retrained weekly, it could be capable of helping in the decision of whether the update makes sense. Uh, the tool has a very robust structure for experimentation, uh, structured pipeline, supports a series of contexts, collects metrics, uh, already used in the anti-fraud payment link model update, has yielded good results. Uh, and I'll read the cons too because then I think a general discussion fits here. Uh, I don't know if it would work 100% in the online experimentation process. Is online experimentation like doing an AB test or something in that context only so I understand, understood, understood.

30:52 (P4 - Blue): Yes, but I speak of a model in production itself. I think the library applies very well there for us to manage to compare, right, the models before or after or before or during the deploy, but I think in production with it totally live, I don't know if the dynamics would be so equal, right? I don't know, it was more a doubt like that, it's not as if it were a negative point, it's more a, I don't know, I don't know if it would work.

31:24 (P1 - Red): Oh, Alene, I was the one who was in doubt with what you said. Uh, what you're talking about an online model, is it like a real-time model, for example, or not? Okay, okay.

31:32 (P4 - Blue): Hi. Yeah, it... I speak more of experimentation in production indeed, like the model there completely live using to not go...

31:48 (Host): Does anyone want to make any more comment? Any counter point, pro point? Ready.

31:56 (P3 - Yellow): I wanted to ask if there is a limit on the amount of models that we can use to analyze.

32:04 (Host): Uh, regarding the amount of models, no, there isn't a limitation. Uh, I would say that for a limitation like that, of size, it's the data volume issue, right? because how it works it, the tool is implemented at the pandas level, right? So, depending on the data size there and such, the maybe the tool has performance issues, right? Which really is more geared toward dealing with samples that aren't so huge, right? Wait someone commented.

32:40 (P1 - Red): Do you, do you have any, man, any plan to put it for larger volumes or understanding that you don't need to?

32:48 (Host): Ah, yeah, no, like, I believe that since this tool it was more a prototype, right, initially, but it has the possibility of extension to, for example, ah put it within the spark context, right, for example, to manage to run using spark and such. Uh, I believe it's something that that would be possible to make this increment, right? But really today it's not something that it has this integration like that, right? Because, for example, from ML Flow, ML Flow I had to load, right? The model to to reference, right? And the same thing for for the data, right? Of there, like, it supports you indicating files, right? Uh, but it has to be at the system level, right? There accessible, right?

33:36 (P1 - Red): Good. Go ahead.

33:44 (Host): Any other point? Okay, fine? Then the the second here, right, which is topic F, is how much you believe that it is easy to use, right? how intuitive the usage is, right, for for the experimentation process in deploying new models or anyway.

36:40 (Host): Anyone writing anything else or did everyone go? Beauty. Uh, so let me do a general read. Two partially agree and two agree. So next the tool's development was done with Python language super used by the teams, which facilitates attendance of the involved processes. Usage seems simple for a separate analysis, from what was shown in the presentation, seems simple to be applied and utilized in the code generally realized. Uh, despite containing a well-structured pipeline, it is still necessary to have good

statistical/experimental knowledge to manage to deal with all the chained stages, CFolds, metrics and statistical tests. And probably would need some adaptations to add in the automated pipeline, right? Uh, does anyone want to make any general addition? Okay, I think what I'll do here is just press to understand. Let me just see here the one from the blue wants to make an addition regarding the statistical slash experimental knowledge issue, that you see that, for example, it could be a point of concern here, just to understand, okay? It's not criticizing the post-it, it's just to understand really.

38:12 (P4 - Blue): No, I think it's simple like this. I think we talked even in the other stage about some challenges we have adopting statistical tests even in our area, right? So, we have, for example, a series of people who are making a career transition, who are here today exercising these functions of, I don't know, data science, data analytics and everything. And I think that, although the library has a very interesting structure on top of the steps, uh, you still need to know what you are messing with, right? It has a certain operational complexity, let's say. So it's more in that sense like, it doesn't have it's not a point, ah, my God, super critical. I think there is already a simplification like that in what we're doing, but it's more an attention point indeed.

38:52 (Host): No. Yes, don't worry. Uh-huh. Understood. No, don't worry. Uh, and here from the adaptations to add to the automated pipeline, what adaptations? Uh, I believe like that that it would make a difference in this part here.

TXT

39:16 (P1 - Red): Yeah, it's because I found that by the initial explanation, okay, that it is very complete, it's because the two little cards they are complementary, uh, that it's very complete for us to do an analysis when we already have a separate model and already have, like, things running and do a separate analysis. However, in the, I thought a bit about my use case. Uh, I I do retrains weekly to like try to have the most the most updated model possible. And then a use case that would be interesting for me would be exactly being able to do this comparison between the current model and the new one and be able based on this comparison to update or not. And then I think for that it would need to adapt a little bit, at least from what I understood, it would need to have some some changes there in the in the output and such to be able to to make this decision making, because in this case it's automated, you understand? It's not a thing where I push the button.

40:20 (Host): Uh-huh. Understood. Understood. No. Perfect, perfect. Just to understand in more detail indeed. Awesome. Um, does anyone want to make one last addition or no worries? Right? So, to close with a golden key, uh, how much intention would you have to utilize the tool to support the experimentation process in comparing versions during deployment?

43:08 (Host): Awesome, everyone already wrote. Beauty, I will do a general read, then. Uh, squat agree there, link the implementation utilize the model pipeline. Oh, cool. I'll even see later calmly. It's cool.

43:16 (P2 - Purple): I've already used all of them, including there's one for PF, one for PJ, all of them.

43:24 (Host): Uh, cool. Uh, again, I can't say during deployment, but have the intention of using the tool yes in the deployment process. Believe it will be very positive. I already want to be able to start testing to see what would need to be adapted in my pipeline. I suggested something about the LIB. Cool. Uh, the library's diagram shows that through it it's possible to perform a well-done comparison among the model versions. So, beauty. Does anyone want to make any addition? So, no worries, no worries.

TXT

44:00 (P1 - Red): I think only those who aren't using it are curious to use it. M.

TXT

44:08 (Host): H, no, then, awesome. Uh, so I think that's it. Thank you, uh, again guys, for your availability, right, because I know everyone's time here is very rushed, so it will help a lot, it's already helping a lot in my dissertation process, okay? All these perspective points that you guys brought, right, with the pros, right, everything will add to the dissertation, right, which is part of the process and, anyway, eventually you guys using the the tool, when you start using it, feel free. If you notice that there's some, some problem, some limitation, uh, anyway, that you notice during usage or something that isn't intuitive. Yeah, feel free to send me a message directly giving feedback, okay? There's no problem regarding this. The idea really is to improve the tool and all this from there has already been some some perspectives that Luana who used it in Myvaluation brought me, right, of improvement for the tool and all this is being documented in the dissertation, because really the intention is that. Uh, so that's it, right? I thank you again there and also if you have any doubt there in the library usage you can also send me a message too.

45:20 (P4 - Blue): Well, congratulations on the work.