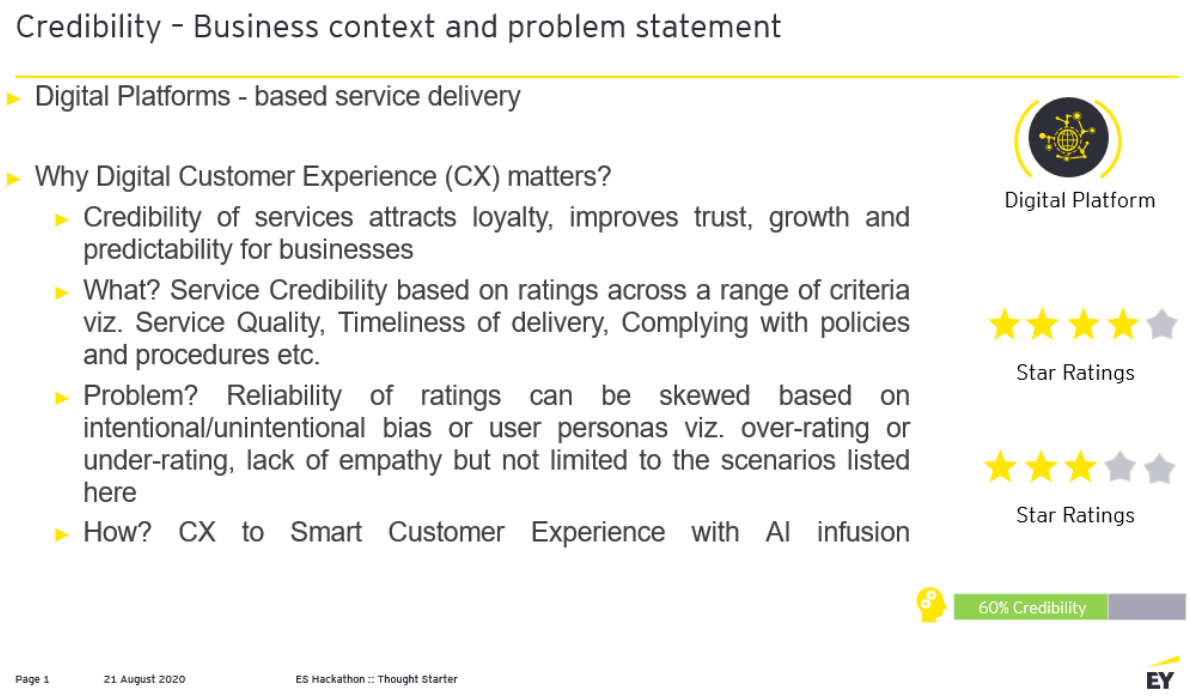
Theme 3: Credibility Calculator

Introduction

Thank you for joining the Credibility Calculator webinar, being organized as part of the Radio – EY GDS Hackathon. In this session, we give you a quick overview about the theme, the expectations and answer some of the common questions that you may have.

Business overview

Most of us are already aware of the digital platforms-based service delivery organizations such as Ola and Uber. Digital customer experience is an important aspect and it is significant to maintain the credibility of these services, as mentioned in the below slide.



We have to attract loyalty, improve trust, growth and predictability for the respective businesses. Currently, what we have noticed across multiple additional platforms is that *service credibility* is captured in the form of ratings. Many of these ratings will have certain key criteria, but not limited to:

- What is the service quality?
- What are the timelines of the deliverables?
- Are they complying with the policies or procedures or the key deliverables that have been committed on those platforms?

This translates in terms of the star ratings given by the users who consume those services. Most importantly, it's a problem from the perspective of how do we rely on these ratings? That's what we're going to focus on in today's webinar.

Reliability of the ratings may have some skew in terms of the feedback that the user would have given. It could be an intentional or unintentional bias or depending upon certain behavioral attributes of the individuals who may give a rating that's either overrated or underrated, or they may not even give the ratings. These are some of the common issues that we witness in the industry, not limited to our theme. That's the broader scenario, so that you can correlate it with this theme, not only from this challenge perspective, but also to look at some of the commonly used digital platforms in the market. You'll get an idea as to what this theme is about. Now, the key expectation out of this challenge is that we want to move more to the next level of moderation, which is from customer service, customer experience to smart customer experience, with some of the next generation technologies, like artificial intelligence. We're looking for you all to participate and contribute in terms of design, coming up with an algorithm or a calculator for looking at the data points, for some of the use cases that we're going to present and be able to look at some of these biases, not limited to, but, different personas than the users. Look at how artificial intelligence can help us give more predictability, more reliability of the credibility score.

Technical implementation

For a person or a company, when there is a service done, there is a feedback survey that goes out, capturing different criteria for a particular service rate. It could be a service, task, or any deliverable, which the person has availed. The customers will be giving you feedback for each of those criteria. The criteria score can range from 1 to 5, 1 being the lowest score and 5 being the highest score.

Feedback providers will have their own thought process based on which they will give their inputs. Some may think that 'Ok, it's a good job' and may write it as 5. Some may like something in that person, say, the way he interacts with the customer. And they may also give 5 for 'quality' as a criteria, while it was not a quality delivery. This means that they may have a bias towards the person. Similarly, there could be different anomalies in such cases. If we are arriving at a normal feedback score based on the feedbacks provided, then the genuineness of that score will not be there. So, what we're looking for is a solution which can do a credibility calculation of the feedback score given.

I may have an average feedback score of 5 out of 5, but the credibility should be calculated based on the past data, which is historical data, based on the different feedbacks received. Also, for the same task, there could be different people who have the same feedback provider and may have different feedback. All these factors have to be considered for arriving at a credibility score. There should be an AI implementation, where, based on the past data, the credibility score has to be calculated. You can also think about other ways to enhance that capturing mechanism, to make it in a full proof way.

The main crux of the solution needs to revolve around the fact that based on the historical data, arrive at a credibility calculation such that the rating, which is given by the system automatically is based on the feedback provider rating. Post an additional calculation it should show that, 4.5 out of 5 was the rating given by a person 'A' for a particular task, for a particular person. However as per the system record, the credibility score for such a person for such a task could be 3.7. That is what we want to arrive at, so that we can give a value-added indication to the person who is using this system by showing that the score is a better score, without any bias.

We will provide a few attributes where some feedback would have already been captured by us, on a scale of 1 to 5, across quality, timelines, overall effectiveness, and so on, and so forth. We have also given you samples of datasets which you can simulate and extrapolate, by artificially inducing a bias into the dataset, that would be either skewed towards overrating or under rating, or individuals who are giving the ratings in the way they are selecting a score across the range of transactions.

How the solution can look into the anomalies, or the contradictions in terms of the feedback the same individual would have given, across different data points, is the kind of simulation that we are expecting. We will give you an opportunity to run through a sample sheet, giving you a smaller dataset, but we're expecting that collectively, you should go and extrapolate based on the high-level guidance that we have given. Create those artificial datasets, use cases, scenarios, know how you may want to call it and tag them saying, this is for artificial simulation of a bias case number 1.

You will have few other use cases, and then you write your algorithm, that can pass through these datasets, to show the moderation saying that, the actual scores reported was X but the intelligence that this algorithm is going to give an insight, is going to say X minus Y, or X plus Y.

Looking forward to any questions or clarifications on the themes and the approach that we are expecting you to work on this challenge.

Let's go to the dataset, which we have projected, for this particular use case.

Dataset overview

Sample Input

					Quality	Quantum of Work	Schedule	Performance	
				Scoring Range	1-5 High	1-5	1-5	1-5	Overall Score/Star Rating
		Nature of Work	Performer	Reviewer					
Task 1	All	Consulting	Anand	Prashant(101)	2	3	2.5	5	4

				Scoring Range	1-5 High	1-5	1-5	1-5	Overall Score/Star Rating
Task 2	Phase 1	Consulting	Anand	Prashant(101)	2	3	2.5	5	4.5
	Phase 2	Presenation	Anand	Vivek (102)	2	2	2.5	5	5
	Phase 3	Documentation	Anand	Shiela (103)	3	3	4	5	3
	Phase 4	Presenation	Anand	Mary (104)	3	2	2.5	5	4
	Phase 5	Documentation	Anand	Joseph (105)	2	3	2.5	5	2
	Phase 6	Presenation	Anand	ABC (106)	5	3	2.5	5	3.5

As you can see on the slide here, 'Task 1' and 'Task 2' are nothing but another representation of some of the work or the service. Let us say you take an Uber ride and then you can put whatever name (we have given a few names here). You can add more names for the purposes of extrapolating the data.

For example: There's an activity or a service that was given by an individual who performed it. The corresponding recipient of that service has given a rating on a scale of 1 to 5, where 5 is on the higher side, 1 being basic. The criteria is cutting across quality, quantum of work, schedule and performance. So, the aggregation of all these feedback translates into an overall average score, which is a star rating - an equivalent of this you would have seen in Uber or Ola or any of these digital platforms we commonly use. Once the service is over, it prompts, "can you give your experience overall" and you give 2 stars or 5 stars. So, that's basically the dataset.

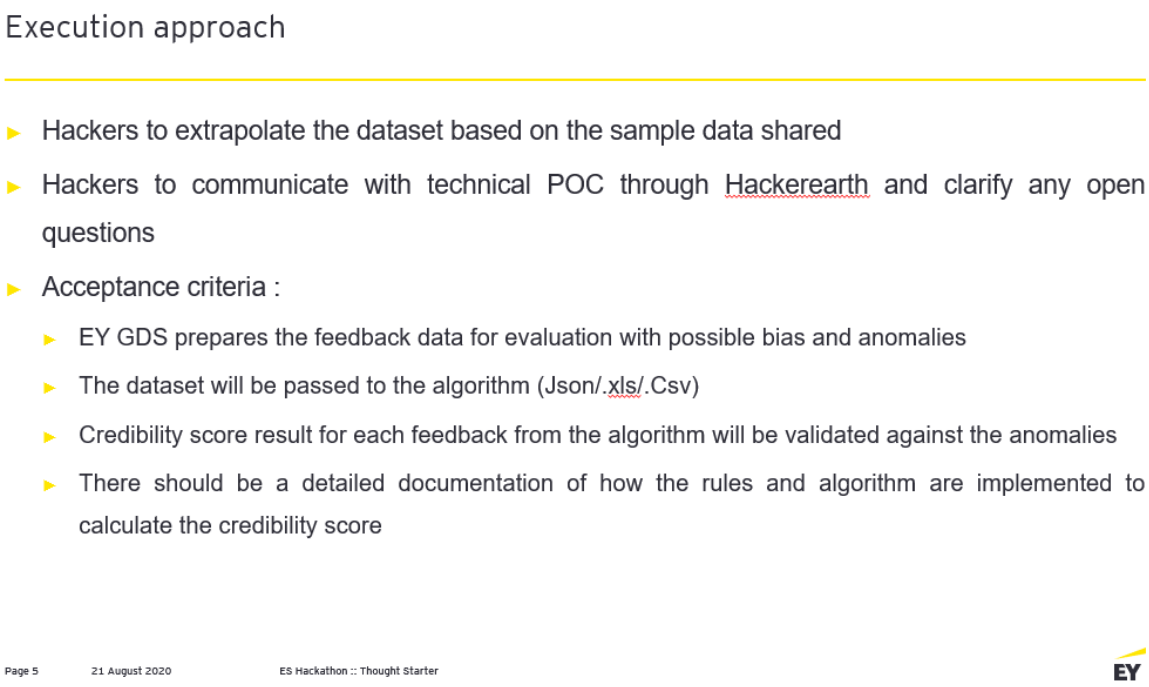
What we are expecting is that you will create, and now, this is only a dataset that we're not telling you, what is the bias here. But you may want to extrapolate this data. This means that you will need to create certain artificially infused personas and you want to tag those personas, let's say, Ram, Shyam or Joseph. Then, you will need to create certain patterns in terms of how you can artificially induce for those respective people who would have performed the services and the reviewer is giving the feedback.

And when the reviewer is giving the feedback, there's a pattern and how their behaviors are reflecting in terms of your algorithm spotting those anomalies and spotting out, saying that when Prashant does a scoring of 1 to 5, one quality, or quantum of work or whatever, we find inconsistencies.

Even though this individual Prashant has given 4.5, the outcome that we are expecting is, it's not 4.5, it's actually 4.2. Not that we want to scrub, you scrub out the 4.5 below the 4.5. We want you to show, it is 4.2 or 4.6 based on the algorithm that you like.

Technical expectation

This slide is talking about the execution approach, and the acceptance criteria. Some of them are already answered through the constants. So, I'll just go through to make sure that everybody's clear.



The execution approach was for the team who is going to be interacting with this.

- To extrapolate the dataset based on the sample data shared.

- It can be a collaborative approach in which you can take together and arrive on that. Then a company gets back to our team, the technical POC who digs deep through the platform. And then have any clarification on that and make sure that the dataset is extrapolated. Get that feedback, and then take the input from us, and then go for developing the algorithm.

What we expect is that, the algorithm which is developed should be integrated with any solution as an API and hosted on Azure platform. So, that it is aligning with the general standards which are shared. And, whatever data, like the feedback score, which you are publishing, or posting to that API, should be able to give the corresponding score back for that particular feedback. That should be a machine learning model, which originated based on the dataset. After the dataset is prepared, that is going to be the scenario for which you may have to work, based on your teams. And as an acceptance criteria, based on the solution which we are going to deliver, we will have a dataset from our side to validate that data, and particular response. And then, it will be evaluated based on the results. We are expecting that we could pass the data in CSV. or JSON in an Excel format [JSON would be better for that execution].

And the credibility score result for each feedback from the algorithm will be evaluated against the anomalies. So, we know what anomalies, we are bringing it up on the dataset and will look forward to what is going to be the dataset. What is going to be the result set coming in? There should be a detailed documentation of how the rules and algorithm are implemented internally to arrive at the credibility score. And that logic will be evaluated to see that will also be on the evaluation criteria.

In phase 1, we are expecting all participants to come up with the dataset and the different artificial anomalies and use cases. And say that this dataset is designed for inducing artificial bias and give it a name - what type of bias, what type of persona. We want to hear that back, because once you give that back to us, then we can establish the criteria against which we can test your algorithm. You can come up with a dataset together and submit for review before starting with development of algorithm. This will provide a common ground for building the algorithm individually.

Q&A

1. In simple words – what's the input and output?

Ans. The input will be a set of feedback scores based on different criteria defined by the developer for the solution. You can refer to the sample dataset provided with the theme. This could be a batch input (like a multiple row JSON dataset) or it could be a single row. The output is expected as a credibility score for each row provided in the input dataset.

2. Do we have any existing calculation logic?

Ans. The basic calculation will be an average of your rating and there could be a weightage also for each of the criteria. If you are asking for a normal calculation, it will be based on the weightages given for each of these criteria, and an average of those weightage taken.

We recommend you to collectively come up with the weightages that you would want to consider. For example, the industry typically has more focus in terms of quality and schedule. And also, quantum of work is a relative term. For example, you give the quality, 20% the quantum of work schedule, you have 30%, and so on. If you can come up with a commonality that you would want to take a weightage, I think we are good. But right now, what we are giving you is an open-ended approach, where we have given an average score based on the feedback rating given per each of these.

3. If they are giving the rating based on the behavior, will that rating be counted?

Ans. Yes, the rating based on behavior could be included, however, it depends on you if this needs to be included as one of the feedback criteria.

4. What are the KPIs on which this task, or jobs otherwise are measured upon?

Ans. The KPI's measured on the task or jobs are given in the example dataset. While you prepare your dataset, you can give more or any other relevant KPI's. One of the key criteria is going to be based on how much of the dataset you can simulate and:

- What are the 4 or 5 simulations of the bias categories that you're going to have. Bias is either intentional or unintentional or you'll want to give a name to that.
- We have only taken the average here, but you may want to come up with a weightage for each of these parameters.

So how is your code parsing across these parameters, and how is it becoming smarter every time you run? Basically, what we want to see is you put a dataset of 30,000 records for a range of performers and the range of reviewers and a range of scenarios. Then when you run your algorithm, you will get a certain moderation score with respect to the actual scores given by the reviewer. When you use 10,000 more records that you would have simulated, but you're not being applied to the algorithm during training, then your code is expected to adapt based on the prior learnings. So that is going to be another KPI.

1. Creating the use case scenarios for inducing the artificial bias
2. Your algorithm meeting the criteria of those use case scenarios that you'd be able to demonstrate
    - given the details, the mathematical formula that you're applying on your overall algorithm you have to provide as the calculation engine
3. How is it able to adapt when you infuse more datasets to it?
    - From the prior inference that you have, how does the algorithm become much more intelligent.

So, in which case, the expectation is when your AI algorithm is running and passing through 30,000 records, their intelligence has to be stored somewhere. We're expecting you to create that reference point. The next iteration of the 10,000 records that you put into play for similar performers, and reviewers, and the personas that you put in, your engine is able to adapt.

We would want to see that the first time around 5 has become 3.5, then over a period of time, it's become intelligent saying that the moment some data point comes up for Vivek, it just knows that, it's not 5. It's going to be 3.5 forever. We need to see that intelligence and would categorize the KPI's in those categories.

5. Is there any weightages between the attributes? (Like quality, quantity of work, time, duration, etc.)

Ans. We may use your own weightages in our business context, but we want to keep it open to know what are the best practices that are out there in the commonly found commercial digital platforms. You could have your own point of view, but then what we do want to have is, let's say, a group of 20 or 30, to come together and brainstorm. Come up with a common weightage that you are applying for the whole dataset, prepare the dataset comprehensively for all the use case scenarios collectively, because creating the dataset is not the competition or the challenge, but that's something that you want to establish. And when you establish that, we want you to come up with the weightage that you are going to apply to the photo as per the criteria we have listed. From thereon, it's your calculation engine that's going to moderate by looking into their datasets, which is a distinctive element that each one of you is going to be judged upon.

In summary, we are keeping it open-ended for you to come up with what weightages you would want to pay. The view of the dataset that you would have prepared along with the weightages that you are considering, along with all the use cases that you are considering for the datasets in iteration 1 and iteration 2. Iteration one being the first time it runs, the second time it runs, its adapting. We'll want to see that adaptive intelligence. Each one of you has the opportunity to have your programming style, you know, you could do a Python, you could use your technology of choice. You'll have your dataset structured in a way that it takes in as an input and it has a moderated score.

6. Would you give the feedback columns also that the customer has provided?

Ans. This is a sample data that we have provided. But you can extrapolate all the data and use it to design, develop the logic that you are building and for the evaluations.