Regression Model Communication

Interview Study with Data Scientists & Subject Matter Experts

PI: *redacted*

PI: *redacted*

Interview Agenda

- The interview will not go past the hour we have
- 3 parts in the interview, about 15 minutes for each part
- This interview will be transcribed and possibly published in research
  - Feel free at any time to speak in abstractions
  - *redacted* and I will remove all sensitive information and anonymize all quotes
  - If you'd like to provide further comments or remove any comments after the interview, please let us know at any time
- Any questions before we continue?

Refresher on Regression Models

- For the purpose of this interview, when we talk about a regression model, we refer to any predictive model that takes in data and outputs a number
    - Unlike classification models, which output a yes or no decision, regression models output a predicted number (e.g., ‘7’ or ‘0.023’)
- A classic example of a regression model is one used at a weather station to predict the daily temperature based on a set of weather conditions as input
- I won’t ask you anything technical about regression models – this is not a test, and there are no right or wrong answers. I just want to understand how you assess regression models and/or use them at work
- Any questions about regression models before we continue?

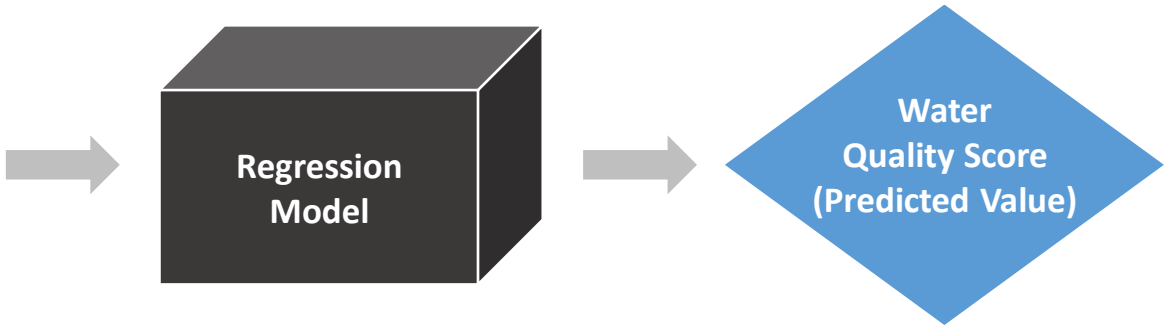
Part 1:
Theoretical scenario about how you would assess a new regression model.

Part 1: *A theoretical regression model that you are assessing*

In this theoretical scenario, you are presented with a complex regression model that is under evaluation. You are told this regression model predicts a water quality score that represents how potable (i.e., drinkable) the water is, given a set of input predictor values.

pH value	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic Carbon	Water quality score
3.7160	204.8904	20791.3189	7.3002	356.8861	564.3087	10.3797	2.9631
8.0991	129.4229	18630.0578	6.6352	310.1357	592.8854	15.1800	4.5006
8.3167	224.2362	19909.5417	9.2758	326.6784	418.6062	16.8686	3.0559
9.0922	214.3733	22018.4174	8.0593	393.6634	363.2665	18.4365	4.6287
5.5840	181.1015	17978.9863	6.5466	303.3098	398.4108	11.5582	4.0750
10.2238	188.3133	28748.6877	7.5448	268.6469	280.4679	8.3997	2.5597
8.6358	248.0717	28749.7165	7.5134	404.0416	283.6516	13.7897	2.6729
7.3606	203.3615	13672.0917	4.5630	326.6244	564.3087	10.3797	2.9631

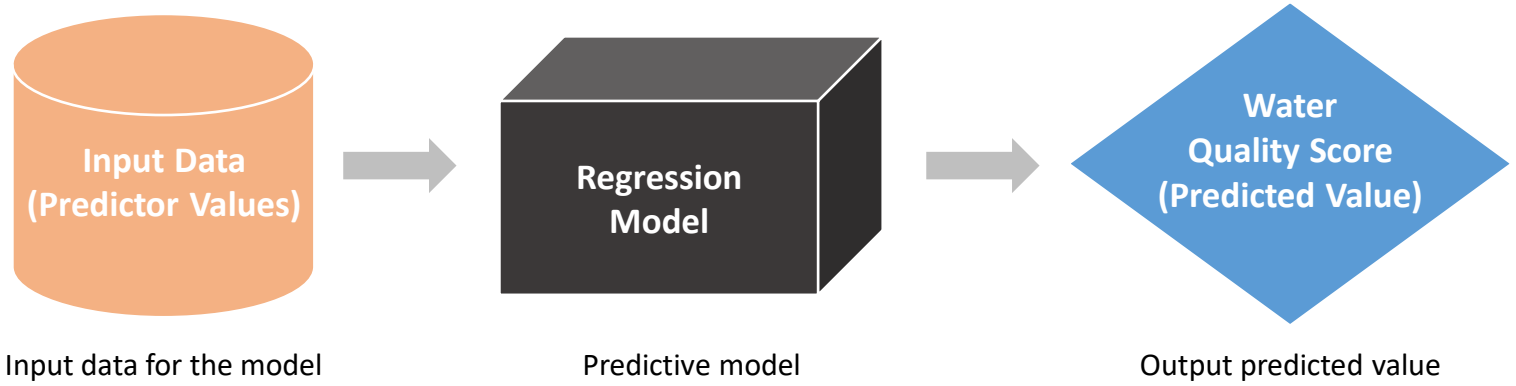
Input predictor values (orange variables) are used to predict the target variable (in blue)



Model that you are being presented

Output predicted value: water quality score

Part 1: *A theoretical regression model that you are assessing*



Your task is to figure out whether you and your team might want to use this model for a project you are on.

I am going to ask you some questions to understand how you might perform this task.

Any questions about the model, or this theoretical scenario, before we move on to the actual interview questions?

Part 1: *A theoretical regression model that you are assessing*

Generally speaking, what would you want to know about this model to recommend its use to your team?

Part 1: *A theoretical regression model that you are assessing*

Suppose you are being given one of the following presentations on this model to help you assess whether you'd want to use it:

- Single (PowerPoint-style) slide with an executive summary of the model
- 20 minute in-depth presentation on the strengths and weaknesses of the model
- Interactive exploration of the model that is facilitated by a software tool or application

Any questions about these presentation styles before I move on to the next set of questions?

Part 1: *A theoretical regression model that you are assessing*

Let's first consider this scenario:

- Single (PowerPoint-style) slide with an executive summary of the model

What information would you prefer to see in a single slide that would help you recommend the model's use to your team?

Part 1: *A theoretical regression model that you are assessing*

Now, let's consider this scenario:

- 20 minute in-depth presentation on the strengths and weaknesses of the model

What information would you prefer to see in a 20 minute in-depth presentation that would help you recommend the model's use to your team?

Part 1: *A theoretical regression model that you are assessing*

Finally, let's consider this scenario:

- Interactive exploration of the model that is facilitated by a software tool or application

What information would you prefer to see in an interactive, explorative presentation that would help you recommend the model's use to your team?

Part 1: *A theoretical regression model that you are assessing*

- Single (PowerPoint-style) slide with an executive summary of the model
- 20 minute in-depth presentation on the strengths and weaknesses of the model
- Interactive exploration of the model that is facilitated by a software tool or application

In your opinion, are there any advantages / disadvantages to these presentation styles?

Part 1: *A theoretical regression model that you are assessing*

- Single (PowerPoint-style) slide with an executive summary of the model
- 20 minute in-depth presentation on the strengths and weaknesses of the model
- Interactive exploration of the model that is facilitated by a software tool or application

In your opinion, is there a scenario in which one of these presentation styles is preferable / not preferable?

Part 2:
Real-life scenario in which a regression model was introduced to your daily work.

Please note: Feel free to speak in abstractions to protect sensitive information. As a reminder, this interview will be transcribed and potentially published, so please do not divulge any information that you do not feel comfortable making public. We will make every effort to anonymize quotes and information when transcribing.

Part 2: *A real-life scenario in which a regression model was introduced to your daily work*

Recall that a regression model is any predictive model that takes in data and outputs a numerical value. For example, a regression model might be used to predict the cost of running a clinical trial, or to predict the cost of a stock's price in the market, or to predict the number of adversarial effects from a drug.

Try to think about a time when a (completed or in-progress) regression model was introduced to your daily work. Please feel free to take your time. When you're ready, we'll move on to the interview questions.

Part 2: *A real-life scenario in which a regression model was introduced to your daily work*

Talk me through the model in this scenario.

Part 2: *A real-life scenario in which a regression model was introduced to your daily work*

How did you assess the performance of the model?

Part 2: *A real-life scenario in which a regression model was introduced to your daily work*

Were there any methods you used to communicate the performance of the model? If so, were any of them more or less effective?

Part 2: *A real-life scenario in which a regression model was introduced to your daily work*

Were there any costs or tradeoffs to using the model? If so, how were they identified / quantified?

Part 2: *A real-life scenario in which a regression model was introduced to your daily work*

When you were trying to assess the model's use with your co-workers, were there any options available to you based on the model's performance?

Examples of options you might have had:

- Option to send it back based on the model's performance
- Option to have more time for the model to be worked on / improved
- Option to use or not use the model at all

Part 2: *A real-life scenario in which a regression model was introduced to your daily work*

When evaluating whether to use the model or not, would you say your final assessment was a distinct ‘yes’ or ‘no’ answer – or was it something between those two?

Part 2: *A real-life scenario in which a regression model was introduced to your daily work*

Did you encounter any issues with the model once it was deployed / in use? If so, could these issues have been better understood when the model was being assessed?

Part 3:
Understanding the different roles in the modelling process.

Please note: In some of our upcoming questions, you will see "(data scientists / SMEs)" in the text – please use the *opposite role* to answer the question. For example, if you are a subject matter expert (SME), the question will be about data scientists. If you are a data scientist, the question will be about SMEs.

When presenting or assessing a model's performance, is there anything you wish (data scientists / SMEs) already knew about predictive models, such as their performance or their use?

Part 3: *Understanding the different roles in the modelling process*

What are some things that you wish (data scientists / SMEs) could communicate about, or give feedback about, in regards to a predictive model?

Is there anything that (data scientists / SMEs) could do to help you build trust and confidence in a predictive model?

Part 3: *Understanding the different roles in the modelling process*

Is there any particular language (e.g., keywords, phrases, ideas, etc.) typically used by your team when working with predictive models?

If so, would any of them be useful for (data scientists / SMEs) to also know when communicating predictive models with you?

Generally, is there anything that works well for you when communicating a model's performance, quality, and tradeoffs to (data scientists / SMEs)?

Have you seen or used any metrics or visualizations that have been helpful or not helpful in your experience of working with and assessing predictive models?

If so, have you run into any common misconceptions when showing or talking about these metrics or visualizations to (data scientists / SMEs)?

Have you run into any common challenges when communicating a model's performance, quality, or tradeoffs with (data scientists / SMEs)? If so, can you speculate about why these are common challenges?

Is there anything else you would like to point out, discuss, or clarify regarding what we've talked about in the past hour?

END OF THE INTERVIEW

Thank you again for participating and giving us your time today!

If at any time you would like to add further comments or remove comments from our interview, do not hesitate to reach out to *redacted* or me via email, teams chat, or call.

- *redacted* email: *redacted*
- *redacted* email: *redacted*

redacted and I will remove any sensitive information that might have come up during the interview. All information and quotes will be anonymized.

Do you have any other questions about our study or our interview?

Again, if you know anyone else that might be interested in participating in our study, please let us know so we may contact them.