Now, let’s practice applying counterfactual evaluation in a scenario that could occur in the real world. Suppose you’re working on a team responsible for maintaining an informational chatbot for a large banking company’s website. When a customer clicks on the chat icon, your chatbot must to choose how to start the conversation. Based on previous UX research, your team currently uses the three ‘best’ introductions, measured by percentage of customers who respond: (0) “Hi there, how can we help you today?”; (1) “Nice to see you again, <customer name>! How can we assist you today?”; (2) “Hi <customer name>, do you have questions about navigating our website?”. The current chatbot system naively selects (0) 50% of the time and (1) and (2) 25% of the time each. You believe that this system could be improved by using a customer’s age and the number of times they’ve logged in to the banking website to determine the introduction they receive. Based on your intuition, you’ve created six customer categories by partitioning age into three groups and number of logins into two groups. Below is the system you propose.

|  |  |  |  |
| --- | --- | --- | --- |
|  | 0-25 years old | 26-50 years old | 51 or more years old |
| 0-5 logins | (0) | (2) | (2) |
| 6 or more logins | (0) | (1) | (2) |

Unfortunately, intuition alone isn’t enough to justify an A/B test between the current system and your proposed system. Given the above information and the provided log data from the currrent chatbot system, your job is to use counterfactual evaluation to show that your proposed system will likely yield a better response rate than the current system.

Suggested Workflow:

1. Calculate the response rate for the current system
2. Add a column to the current system logs for propensities (probability that the current system displayed the introduction shown in the logs given the user context)
3. Use IPS to estimate the response rate for your new system
   1. For each log in the current system’s logs
      1. If the introduction your system *would have* given based on customer context matches the introduction actually given in the logs, add the inversely weighted response to your estimate
      2. Otherwise, continue
4. Compare your IPS estimate to the observed response rate for the current system—which is better? Does your estimate make sense?