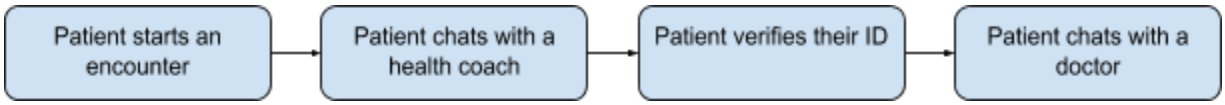
Data Scientist Intern Take Home Assignment

Background

Curai provides a platform for patients to chat with medical professionals and get the medical advice and help they need through a text based platform. A typical flow of a patient's visit is depicted below.



The difference between a health coach and a doctor is that a health coach's role is to educate a patient on their symptoms and provide additional resources needed to help the patient. A health coach cannot provide a diagnosis nor a prescription. A doctor is a California licensed medical doctor who is authorized to provide diagnosis, prescriptions, and lab orders.

In this assignment, we will focus on the patient ID verification step and try to understand it more. Attached are 3 tables with data on ID verification event, doctor entering the encounter, and first patient response. Each record in the id_verification table is a full ID verification funnel for a patient's visit. A patient can repeat an action within the ID verification flow multiple times.

Data definition

There are 3 tables: id_verification, doctor_enters, first_patient_response

1. Id_verification

   encounter_id
   - Unique identifier for a patient's visit

   user_id
   - unique identifier for a user. For the purpose of this assignment, assume user and patient has a 1 to 1 relationship

   starts_at
   - encounter start timestamp

   ends_at
   - encounter_end timestamp

   platform
   - Platform user used during the ID verification flow
   - One of (ios, android, web)

   saw_verification_timestamp
   - when the patient saw the ID verification request

start_verification_timestamp
- when the patient started the ID verification flow
- Null if patient did not start the ID verification

confirmed_selfie_timestamp
- when the user confirmed the submission of their selfie to match their ID
- Null if patient did not confirm their selfie

verification_confirmed_timestamp
- when the ID verification was confirmed successful
- Null if verification was not confirmed

2. Doctor_enters

encounter_id
- Unique identifier for a patient's visit

dr_enter_timestamp
- when the doctor enters the visit
- Null if doctor did not enter the chat

3. first_patient_response

encounter_id
- Unique identifier for a patient's visit

patient_ressponse_timestamp
- first patient message after the doctor has entered
- Null if patient did not send a message after doctor enters

Data Problem

Our goal is to try and understand what is happening during the ID verification process and extract any actionable insights to convert as many users as possible from when they see the ID verification request to chatting with the doctor. Here are the questions we are trying to answer. **Please complete the assignment in SQL and include your query to each question as part of the response.**

1. Data summary
    a. How many **unique** encounters are there in the data?
    b. How many **unique** users are there in the data?
2. What is the drop-off rate between each step from saw_verification to patient_response on a **per encounter basis?**
    a. Further breakdown by platform.
3. What is the median time between each step from saw_verification to patient_response on a **per user basis?**
    a. Further breakdown by platform.
4. Are there any other additional valuable insights from the data?

Evaluation

To give you a sense of how we will assess your analysis, our evaluation criteria looks at:

- How did the candidate make decisions on defining ambiguous metrics?
- Did they explore the pros and cons of their choice and articulate it in their report.
- Did they look back at the results and see whether they passed the smell test? If not, did they find out what the issue is or at least point out their concern in the report?

By design, the assignment is ambiguous in parts, as in many cases it's not always clear how to define a metric, and we do not look for the candidate to approach the problem as we would have, but rather to make a case for their approach, and to be their own devil's advocate.