Phreesia Data Science: Data Scientist Data Challenge

The following semi-fictional scenario is meant to be representative of the typical challenges that you will face in this position. The business problem, data, and other details have been tailored to be suitable for a take-home challenge.
You will have 3 days to address the following business problem. We want to be respectful of your time and to see what decisions and assumptions you chose to make in order to complete within the given time frame.

Honor Code

Please only submit your own work and do not share the content of this exam or your submission with any outside party.

Business Problem

In the wake of the COVID epidemic, practices across the Phreesia network have seen major disruptions to typical visit patterns. Also, an increasing number of practices now offer COVID screening. The resulting surge in screening visits, while great for public health, creates many technical and practical challenges for Phreesia and our clients.

The DS team has been tasked with predicting the volume of screening visits in the upcoming month. If we expect more than 300,000 screening visits on a given day, then we will need to alert the engineering team that more resources will need to be dedicated to handle the load.

Please **predict the number of COVID screenings we expect to see in July** given the overall visit volume and the increase of screening capabilities. Specifically, your code should 1) answer whether we should expect to exceed 300k screening visits *on any day*, and 2) give the likelihood of that prediction.

Details

Operate under the assumption that the code you develop for this prediction will get integrated with our existing forecasting code base, and thus needs to be polished – notebooks and spaghetti code are not going to cut it. Other DS team members will need to interact with and build upon your code in the future, so it needs to be clear, concise, and well documented with doc strings and code comments.

Submit a Python module that takes as input:

- Visit volume
- Screening volume
- The screening threshold (300k)

and returns as output:

- True/False of whether we expect to cross the screening threshold in the next month on any given day
- The likelihood of the prediction (metric of your choice)
- A visualization of the prediction

You may tackle this problem through a purely statistical approach or a ML approach, whichever you feel is best suited.

Data

Use the included csv files for this analysis:

- Total daily visit volume: visit_data.csv
- COVID daily screening volume: screening_data.csv

Questions

Also include a word document with your submission addressing each of the following:

- Why did you choose your prediction methodology?
- What are the strengths and weakness of your approach?
- Important caveats or assumptions
- What did you provide as your likelihood metric and why?
- How could you improve the quality of your prediction?
- What questions would you ask the stakeholders if you were tasked with this problem?
- Given the output of your prediction, what recommendation would you make to the stakeholders?