Data Scientist (PS+ML) Homework

Disease Modeling:

Here at Clover Health we pride ourselves on considering candidates with non-traditional resumes. To make sure that you can showcase your full range, we ask you to complete a take home exercise. We know we're asking for you to volunteer your time, so we want to be sure that we explain why we do this.

It gives you the opportunity to learn about us. Did we choose a thoughtful and challenging exercise? Did we give you the opportunity to get excited about ways you could improve a complicated system with analytics? Did you feel like you were able to show yourself off as ”more than a resume”? If so, you'll be more excited about working for us. (If not, we may not be right for you - and we don't want to waste your time.)

It gives us the opportunity to learn about you. Were you able to clearly communicate your ideas? Were you able to find a good balance between completeness and time to ship? Can what you did be leveraged by the rest of your team? This is your time to shine, and show us what we can't see on your resume.

Our expectations: We don't expect you to spend days on this – likely, a few hours will suffice. We don't expect you to deliver production-ready models. We know you have limited information; it's okay to make some assumptions and base your work off those. (But if you have questions, don't hesitate to ask!) We look forward to getting a glimpse into your thinking process and creativity!

Complete the exercise below in whatever format you prefer. For some loose requirements, imagine yourself as a member of our team. So:

- Your results will be used by others.
- Others will need to read your code.
- They may want to reuse some of your components.
- They (or you) might need to do something similar in the near future.
- They will certainly want to make use of the same data.

The intention of this assignment is to learn more about candidates as part of Clover Health's interview process. It is not to acquire confidential information. Contact Peter at peter.loscutoff@cloverhealth.com you have any questions.

The Assignment

At Clover Health, one of the primary responsibilities of the data science team is to build the products that give needed context to the person seeing one of our members. We want whoever is interacting with the patient to know who they are talking to, and also to take the opportunity to nudge them towards behaviors that will benefit them. Examples are:

- A nurse on a home visit to check on a member who has been recently discharged from the hospital should be able to check for new diagnoses we suspect that member to have.
- A diabetic patient calls in to our call center with a membership question but hasn't seen their doctor in the last six months. Our customer service agent should be able to make them an appointment.

A critical piece of this is knowing what chronic conditions affect the member. Unfortunately, we sometimes don't have all of the information we need to say with certainty what conditions or ailments someone may have - in particular if a patient has just joined our plan, we won't have any clinical diagnoses. In lieu of direct information, we might use other information such as demographics, medical history, and medications to build a picture of a member's conditions.

Given the meps_base_data & meps_meds data sets:

- What are the most common medications for each disease in the base file?
- What medications are most indicative of each disease?
- Choose ONE of the diseases and build a model to infer whether that disease is present from the medications.
- Demonstrate that the end user should be confident in the result.
- Can you find any evidence that for the disease you've modeled, a certain drug is preferred by a certain demographic subgroup?

Some assumptions to keep in mind:

- The "real" datasets will be very very large.
- They may contain duplicates or malformed records.

Constraints:

- You must use Python.
- If you work in a Jupyter Notebook, please submit an html version of the notebook with the notebook itself (this makes it easier for us to evaluate).

Please email us:

- Your source code, in a runnable manner.
- A brief summary of methods and results.