Hi! Thanks for investing the time in our Oscar interview process.

Our data science team is looking to evaluate how well you understand a problem and data set quickly and then prototype ideas. We want to see your code and some of its output, but it doesn't need to be production-level. We encourage you to think creatively and outside the box here - this problem is open ended by design!

Data Science Case Study

Classifying health status from medical and prescription drug usage

Determine a way of inferring information about members' health issues using their medical diagnosis data and prescription drug utilization. This is an important component in any real-world member profiling model. It can help us to detect members that are not receiving their required medical care as well as to identify members with highly unusual care patterns. We can also use it to anticipate future care needs and be proactive about managing them.

There are three provided data sets that simulate medical utilization for 200,000 members over a period of 3 years:

1. Claim lines with diagnosis codes: This data set models the diagnosis codes that are present on claims for medical procedures and services. Every row lists one diagnosis given to a member on a certain day. To keep things simple, we have excluded the procedural and financial information. The diagnosis codes are in a format called ICD-10.

2. A mapping of diagnosis codes to clinical categories (CCS): Diagnosis codes found on claim lines are mapped to higher level clinical categories. We use multi-level CCS as described here. The diagnosis code formatting is slightly different in the outpatient claim lines and in this mapping, so just correct for that. Not all diagnosis codes have a matching CCS code.

3. Prescription drug data: Every entry in this data set corresponds to a drug prescription filled by a member. The drugs are identified by their National Drug Code (NDC). The table provides additional information about which drug category, drug group and drug class a specific drug belongs within.

Requirements

1. **Build a characterization of a member's health status based on their outpatient data.** The CCS mapping will be useful for that. Feel free to define whatever you would consider a useful high-level description of a member's health status - e.g. "Member xyz had cancer and a broken leg." This does not need to be overly complicated - the challenge lies in finding a robust and useful characterization.

2. **Investigate how to infer this health status from the prescription drug data alone.** By the end of this exercise, you should have a simple model that could be applied to data for additional members to try to predict their status. It's up to you as to how you want to build this model. If there is not enough time to build an actual working model, include a plan explaining what you have learned from the data and how you would build a model given additional time.

****Estimated completion time: 4 hours within a 72 hour time frame. We highly recommend not exceeding 4 hours as we want to be respectful of your overall time investment.****