Springboard Data Science Career Track - Capstone 1

Chronic Health Conditions at The City Level

Goals

* To predict negative & positive health outcomes in the United States

Research Objectives

* Identify behaviors that increase and/or decrease the risk of negative health outcomes
* Understand how different populations negative health outcomes are affected by different behaviors

Who This Wil Serve

* Who: Government Health Officials, Health Insurance Companies
* Actions based on outcomes:
  + By understanding which behaviors better predict positive or negative health outcomes, Government health officials can create incentives and other nudges to take part in behaviors that lead to optimum health.
  + This will be similar for Health Insurance companies but more in the scope of lowering their costs of coverage due to less hospitalizations/doctor visits

Data Set

I’ll be using the data.gov 500 city data which is a model based, small area estimate for 27 measures of chronic disease related to unhealthy behaviors (5), health outcomes (13), and use of preventive services (9).

Where is the data from? From the data abstract:

* *Data were provided by the Centers for Disease Control and Prevention (CDC), Division of Population Health, Epidemiology and Surveillance Branch. The project was funded by the Robert Wood Johnson Foundation (RWJF) in conjunction with the CDC Foundation.*

Note of caution from the data supplier:

* *Because the small area model cannot detect effects due to local interventions, users are cautioned against using these estimates for program or policy evaluations.*

Dataset Link: [Data.gov 500 city local data for better health](https://catalog.data.gov/dataset/500-cities-local-data-for-better-health-b32fd/resource/8a49a1f7-4fcc-49a6-acb5-fcd3c0796782)

Approach to Solving Problem

I’ll be exploring the data for this problem in a Jupyter iPython notebook. I’ll start with visual data exploration using seaborn/matplotlib/pandas along with using statistical data exploration. This will start on the high level (aggregate of all measures) and will move to individual cities & populations measured in the data.

I’ll try to implement different split train/test models such as logistical regression and decision tree analysis.

I’ll be looking at negative health behaviors and how they link to health outcomes, along with positive health behaviors and how they link to outcomes as well.

Deliverables

The insights and findings will shared in the following formats:

* High level insights will be shared in a powerpoint/pdf slide deck
* In depth analysis will be shared in a paper
* Code will be shared in a Python Jupyter Notebook