Final Capstone – The Prevalence of Diabetes

Springboard Data Science Career Track

Aisling Casey, June 21st , 2021

Problem Statement:

What was the county-level prevalence of diabetes among adults in the United States in 2018 and 2019?

Context:

The rate of Diagnosed Diabetes in American adults has increased more than 50% from 2000 and 2018, from a prevalence of 6.0% to 9.1%[i]. In 2017, about 1 in 7 dollars spent on healthcare in America was spent on treating Diabetes and its complications[ii]. Knowing the state of diabetes at the local level informs public policy and can improve the efficacy of public health outreach.

The Centers for Disease Control (CDC) periodically releases the National Diabetes Statistics Report, which provides county level information on various diabetes related data, such as prevalence. While a report on the prevalence of diabetes at the national level was released for 2020, the last report with county level data was released for 2017[i].

Though it can afflict people of all backgrounds, diabetes is known to be tied to socioeconomic factors. By modeling local trends and considering new socioeconomic data for the years 2018 and 2019, it should be possible to estimate the local prevalence of Diabetes for these years.

Scope of Solution Space:

A model for the county-level prediction of diabetes prevalence based on American Community Survey (ACS) 1-year estimate data.

Criteria for Success:

For each county on average, the model must yield a lower loss in predicting prevalence based on the last supervised test sample (the year 2017) than simply predicting the previous year's prevalence value for that county.

Data Sources:

All data will be taken from the county level.

- **Dataset 1:** National Diabetes Statistics Report - Centers for Disease Control (CDC)
- **Dataset 2:** American Community Survey (ACS) - US Census Bureau

Feature Category	Detail	Type	Source
Diabetes	% Prevalence 20+ years	Target	CDC
Population	# OR % 21+ years	Explanatory	ACS
Sex	# OR %	Explanatory	ACS
Race	# OR %	Explanatory	ACS
Age	# in 10 year bins	Explanatory	ACS

Income	$ Median household income	Explanatory	ACS
Education I	% Highschool or >	Explanatory	ACS
Education II	% Bachelor's or >	Explanatory	ACS

Constraints:

- For valid yearly estimates of population demographics, yearly estimates only available for counties with populations >65,000
- Overlapping data from both sources available yearly from 2006-2017 only; additionally, have ACS data for 2018 and 2019.
- Need to be wary of high correlation between variables and the risk of simply predicting prevalence based on nationwide average

Stakeholders:

- **Federal Government.** Diabetes puts a major strain on Medicare and Medicaid.
- **Public Health Organizations.** Knowing which communities will face a greater burden of diabetes, public health organizations can direct their resources accordingly.

Deliverables:

- A predictive model that uses demographic data to predict diabetes prevalence.
    - Will be documented with a cookie-cutter GitHub repository of Jupyter Notebooks, model performance reports, and the final model in pickled form.
- Predictions based on the 2018 and 2019 1-Year ACS data.
- An interactive dashboard to visualize trends, model results and explore relationships between variables.

Citations

[i] "U.S. Diabetes Surveillance System." Centers for Disease Control and Prevention. Centers for Disease Control and Prevention. Accessed June 19, 2021. https://gis.cdc.gov/grasp/diabetes/DiabetesAtlas.html.

[ii] "The Cost of Diabetes." The Cost of Diabetes | ADA. Accessed June 19, 2021. https://www.diabetes.org/resources/statistics/cost-diabetes.