Health Estimator

Team Data Dogs

Chris Heng, Stephanie Loomer, Alex Tsai, Alexis Rangel, Saidy Estudillo, Anthony Lopez

Agenda

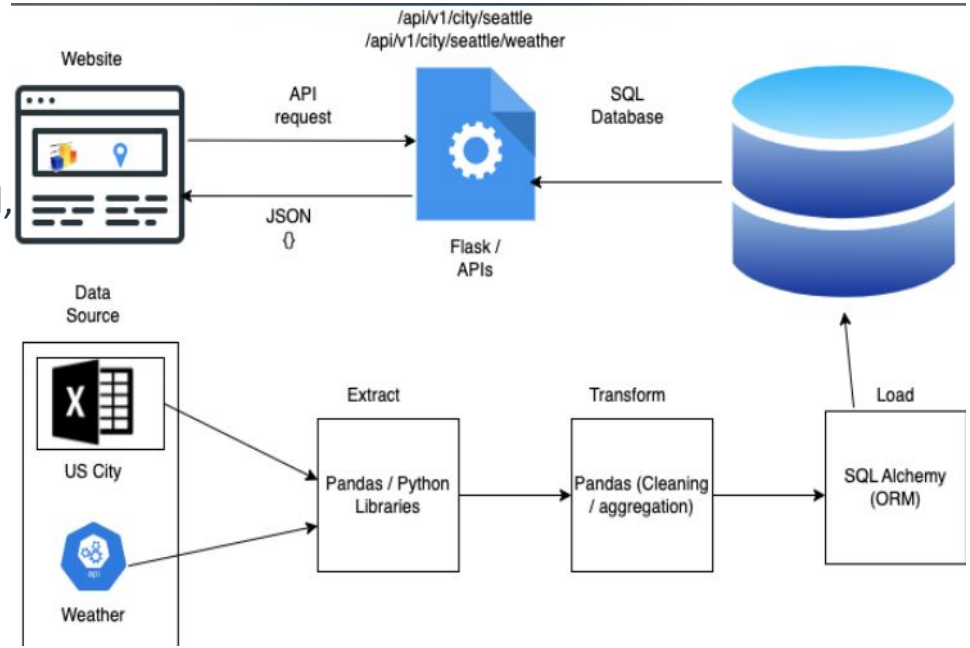
1. Motivation
    a. Questions to answer
2. Data Prep and Loading
    a. How we loaded into Database
    b. Data transformation (what features did we remove)
3. Exploratory Data Analysis
    a. Unsupervised Learning: K-Means & PCA
    b. Birch Model and Agglomerative Clustering
4. Feature Selection
5. Model Selection and Insights
    a. Supervised Learning: Linear Regression
6. Demo
7. Limitations of Model and Data
8. Q&A

Motivation

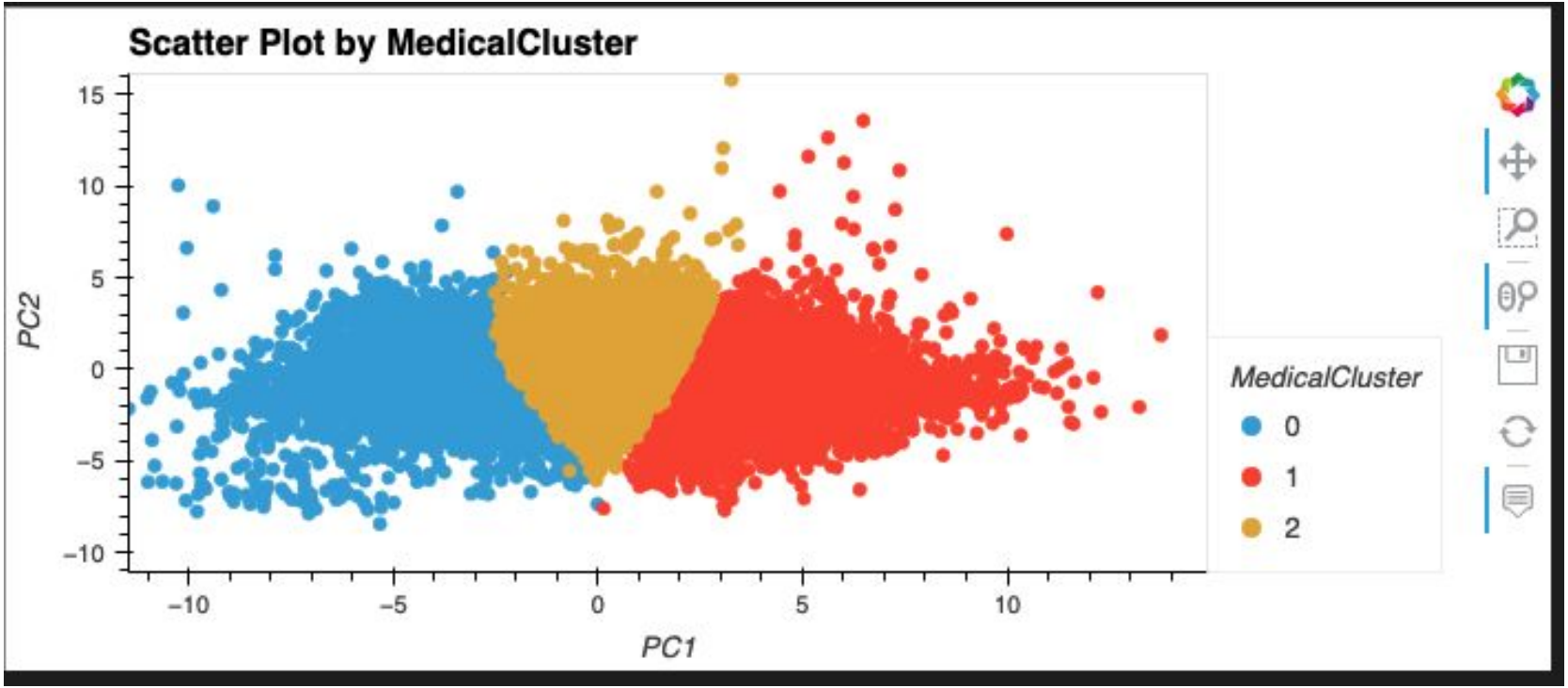
- Data from reliable government sources, enough data to use a deep neural net
- Data Sources:
  - https://data.cdc.gov/500-Cities-Places/PLACES-Local-Data-for-Better-Health-ZCTA-Data-2023/qnzd-25i4
  - https://www.irs.gov/statistics/soi-tax-stats-individual-income-tax-statistics-2020-zip-code-data-soi
- Questions:
  - Does a certain lifestyle feature correlate with a specific health disease?
  - Does the more income a individual has contribute to a healthier life?
  - Can we predict the probability of individuals with a certain health problem? (Cancer, Diabetes, or Obesity.)
  - Can we use a health dataset to find a prevalence of a health condition in a certain area

Data Prep

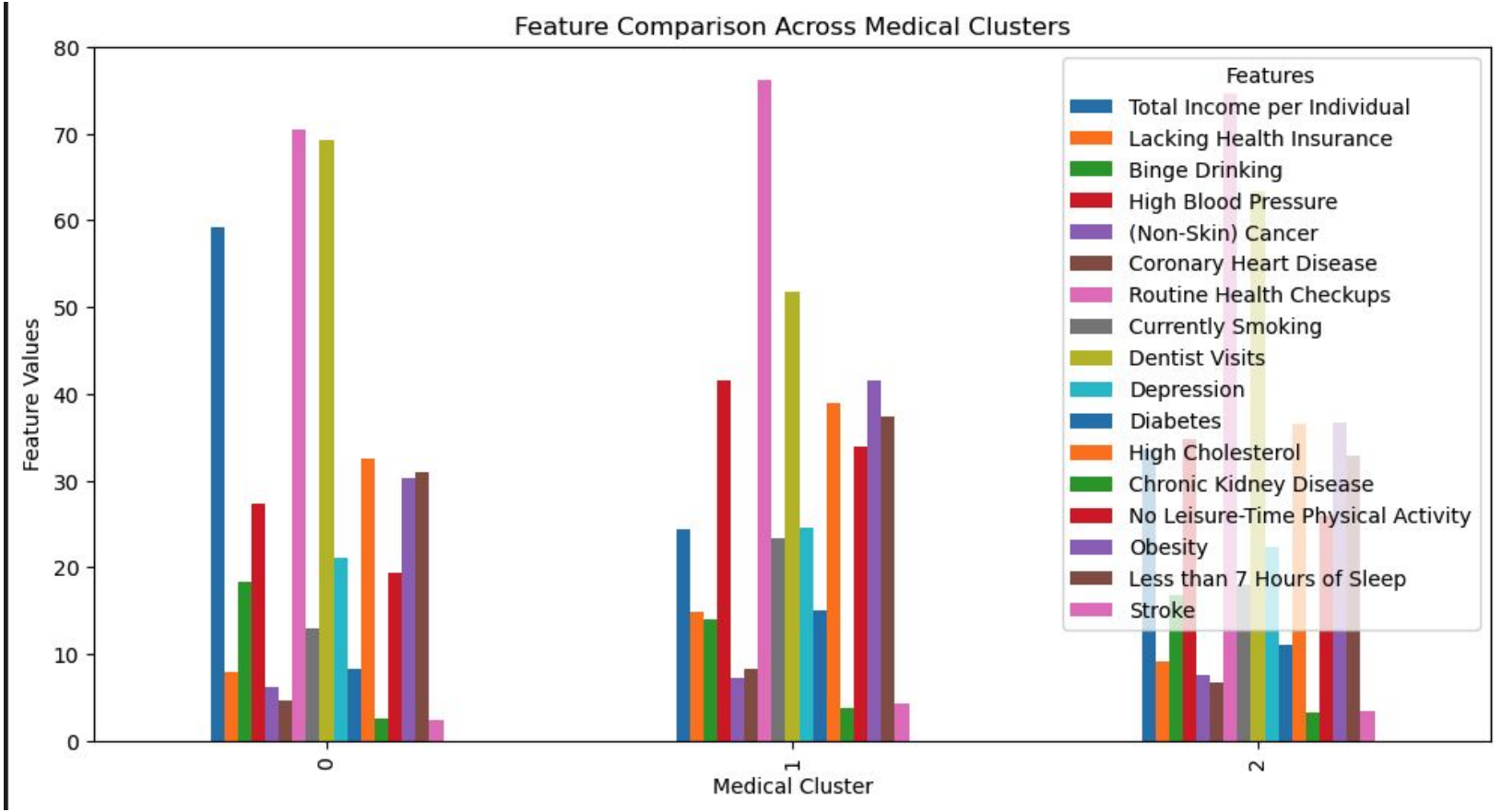
- CDC medical data & IRS CSVs
- Cleaned using Pandas (dropped columns/merged DataFrames)
- Aggregated data (train data, clustered, regress, DNN, unsupervised)
- Sql
- Flask, Java Script
- Html



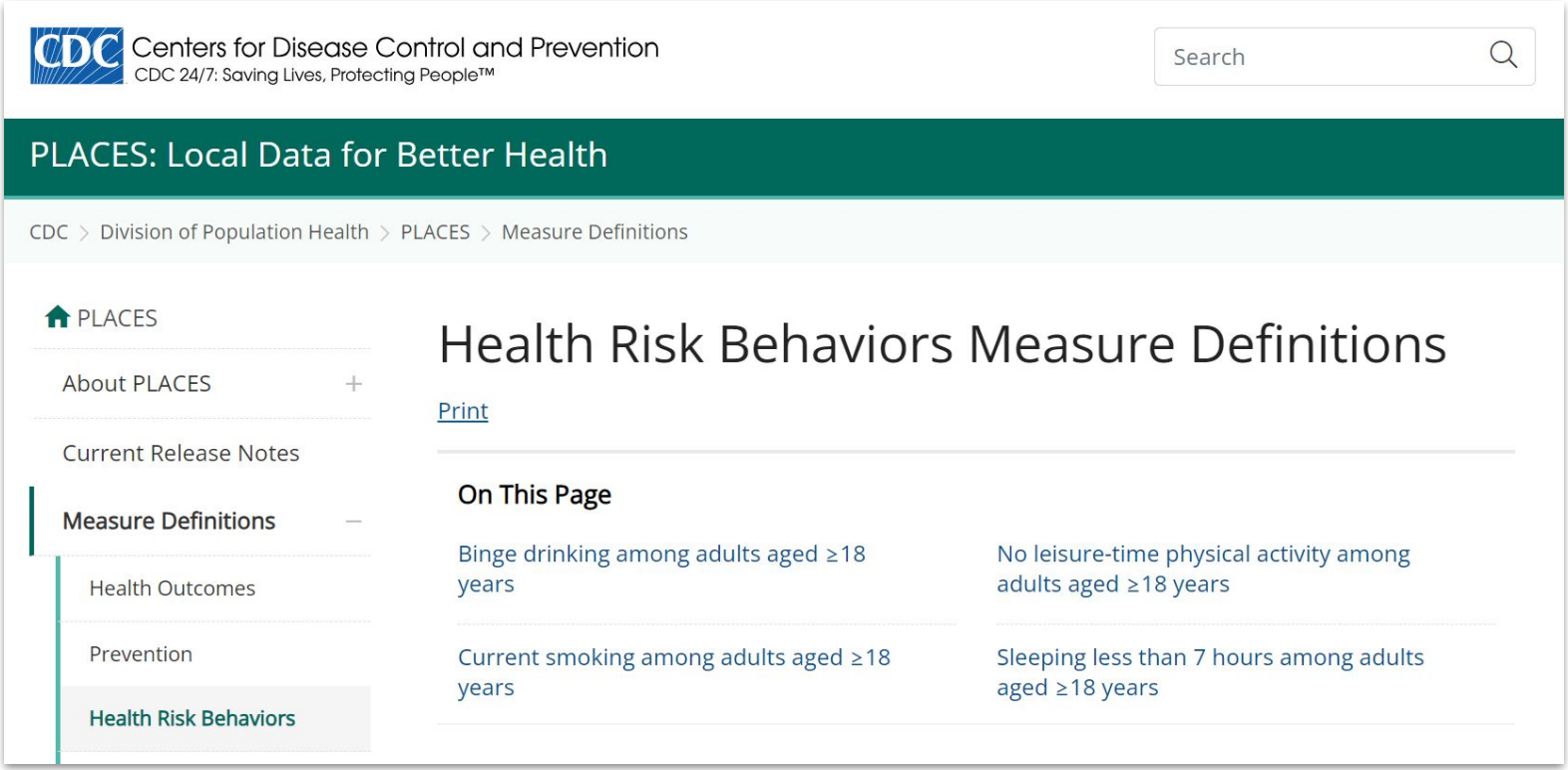
EDA: Unsupervised Learning Clustering



EDA (Clustering) cont.



Feature Selection



Feature Selection

A	B	C
Binge Drinking	Binge Drinking	1
Stroke	Stroke	1
	Chronic Kidney Disease	0.967688121
Chronic Kidney Disease	Stroke	0.967688121
Diabetes	Stroke	0.926398436
Stroke	Diabetes	0.926398436
High Blood Pressure	Stroke	0.910545319
Stroke	High Blood Pressure	0.910545319
Diabetes	Chronic Kidney Disease	0.88758517
Chronic Kidney Disease	Diabetes	0.88758517
High Blood Pressure	Diabetes	0.879662578
Diabetes	High Blood Pressure	0.879662578
Chronic Kidney Disease	High Blood Pressure	0.867419762
High Blood Pressure	Chronic Kidney Disease	0.867419762
No Leisure-Time Physical Activity	Diabetes	0.85168081
Diabetes	No Leisure-Time Physical Activity	0.85168081
Currently Smoking	No Leisure-Time Physical Activity	0.85033218
No Leisure-Time Physical Activity	Currently Smoking	0.85033218
Dentist Visits	No Leisure-Time Physical Activity	0.846921405
No Leisure-Time Physical Activity	Dentist Visits	0.846921405
Obesity	No Leisure-Time Physical Activity	0.79919
No Leisure-Time Physical Activity	Obesity	0.79919
Obesity	Currently Smoking	0.799082002
Currently Smoking	Obesity	0.799082002
	Dentist Visits	0.783568862
Dentist Visits	Currently Smoking	0.783568862

9 Feature Columns:

- Income Per Capita
- Prevalence of Lacking Health Insurance
- Prevalence of Binge Drinking
- Prevalence of High Blood Pressure
- Prevalence of Routine Health Checkups
- Prevalence of Currently Smoking
- Prevalence of Depression
- Prevalence of No Leisure Time Physical Activity
- Prevalence of Less Than 7 Hours of Sleep

3 Target Columns:

- Prevalence of Obesity
- Prevalence of Cancer
- Prevalence of Diabetes

Model Selection & Insights

- Our goal was to predict cancer, obesity, and diabetes based on certain lifestyles within a zip code
- We used tensorflow to build a regression model with a deep neural network with a total of 3 layers
- Our model had a total 9 features that we used to try and predict prevalence for each of our selected targets.
- Initially normalization of data occurred but used original data after comparing to simplify frontend process

Results

ZIPCODE	Diabetes	Predicted_Diabetes	Total Income per Individual	Lacking Health Insurance	Binge Drinking	High Blood Pressure	Routine Health Checkups	Currently Smoking	Depression	No Leisure-Time Physical Activity	Less than 7 Hours of Sleep
44702	28.7	26.827890	91.070001	15.8	10.0	57.5	79.7	36.6	23.9	48.4	44.4
44510	28.7	25.454704	11.967296	18.7	9.7	51.8	83.0	37.1	25.3	55.4	50.5
79901	28.1	26.921309	13.840397	55.0	10.7	46.6	71.4	26.3	25.9	51.1	37.2
36612	27.0	24.857916	13.887329	21.0	10.3	55.0	81.4	30.3	21.9	54.2	50.3
78353	26.9	25.847067	20.426667	46.2	11.9	46.6	72.8	21.3	21.4	47.6	36.2

Demo

Website

Limitations of Model & Data

- Raw data is organized by whole areas (zip codes) that are unevenly distributed (zip code population is extremely varied), instead of individual patients
- Model is only taking into account features/columns we deemed were more valuable and meaningful → could lead to unintentional bias

Mean absolute error

Obesity	0.403693
Cancer	0.47445
Diabetes	1.815076

Q&A

Work Cited

- CDC. “Places: Local Data for Better Health, ZCTA Data 2023 Release.” *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 25 Aug. 2023, data.cdc.gov/500-Cities-Places/PLACES-Local-Data-for-Better-Health-ZCTA-Data-2023/qnzd-25i4.