Hani Shakrah

15 January 2024

Project Design and Requirements

**Constraints:**

1. Time

I have roughly 2.5 weeks to complete this project. This is further broken down by the specific steps below.

1. Resources

With ~300k subjects for each year, my computer may not have the capacity to efficiently process all the data. I will attempt to utilize four years of data into my modeling, but decisions may have to be made to accommodate for the space limitations.

**Steps:**

1. Understand data and define objectives (1 day)
2. Extract from CDC.gov for features initially deemed somewhat useful
3. Compile preliminary features into one dataset (2018-2022)
4. Simple exploration
5. Define more nuanced objective based on constraints
6. Model and evaluate performance
7. **Understand data and define objectives:**

*Read about the census:*

* Are there limitations to the way the surveys were conducted?

Survey is conducted only in the United States and surrounding islands. There may be underlying phenomena that may have to do with specific geographic or cultural conditions that affect the results. Caution should be taken before generalizing to other countries and continents.

The only method of contact for interviewers is through a phone call (landlines and cell phones) using Random Digit Dialing (RDD) techniques. Certain demographics may be hard to contact and introduce bias into the results.

* Is the information collected rational to be used for diabetes prediction based on current science?

Yes, there seems to be an abundance of information retrieved from these surveys, some useless to the task at hand, but some that seem relevant. Current science points to blood pressure and increasing age as important factors, both of which are included in the survey. <https://www.cdc.gov/pcd/issues/2017/16_0244.htm>

* Look for potential biases

As noted, lack of geographic diversity and sampling bias from utilizing one method of contact.

Recall bias – difficulty remembering, inaccuracy in responses due to forgetting of details

Selection bias – only English and Spanish speakers

**Can the same person participate in multiple years?**

* What work has already been done on this census

<https://www.cdc.gov/pcd/issues/2019/19_0109.htm>

*Define success:*

* Building a model that can generalize well to new patients. Features need to be numerous and important enough to allow for accurate flagging of diabetes but must not be over complicated for the practitioner to implement.

**Steps 2-6:**

See code notebook

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Attribute Definitions** | | | | | | | |
| Feature | Description | 2018 Code | 2019 Code | 2020 Code | 2021 Code | 2022 Code | Issue |
| Sex | What is your sex? | SEX1 | SEXVAR | SEXVAR | SEXVAR | SEXVAR | Remove 2018 “7” and “9” |
| Age | Six-level imputed age category | \_AGE\_G | \_AGE\_G | \_AGE\_G | \_AGE\_G | \_AGE\_G |  |
| Elder | Two-level age category | \_AGE65YR | \_AGE65YR | \_AGE65YR | \_AGE65YR | \_AGE65YR |  |
| Race | Race/ethnicity categories | \_RACE | \_RACE | \_RACE | \_RACE | \_RACE1 | 2022 – no “6” |
| Urban | Urban/Rural Status | \_URBSTAT | \_URBSTAT | \_URBSTAT | \_URBSTAT | \_URBSTAT |  |
| Education | Level of education completed | \_EDUCAG | \_EDUCAG | \_EDUCAG | \_EDUCAG | \_EDUCAG |  |
| Marital | Marital status | MARITAL | MARITAL | MARITAL | MARITAL | MARITAL |  |
| Veteran | Ever served in the US Armed Forces? | VETERAN3 | VETERAN3 | VETERAN3 | VETERAN3 | VETERAN3 |  |
| Employment | Employment status | EMPLOY1 | EMPLOY1 | EMPLOY1 | EMPLOY1 | EMPLOY1 |  |
| Kids Household | Number of children in household | \_CHLDCNT | \_CHLDCNT | \_CHLDCNT | \_CHLDCNT | \_CHLDCNT |  |
| Height | Reported height in inches | HTIN4 | HTIN4 | HTIN4 | HTIN4 | HTIN4 |  |
| Deaf | Deaf or have serious difficulty hearing? | DEAF | DEAF | DEAF | DEAF | DEAF |  |
| Blind | Blind or have serious difficulty seeing, even when wearing glasses? | BLIND | BLIND | BLIND | BLIND | BLIND |  |
| Healthcare | Some form of health coverage? | HLTHPLN1 | HLTHPLN1 | HLTHPLN1 | \_HLTHPLN | \_HLTHPLN | 2018-2020 – remove “7”, “9”, “BLANK” |
| No Doc Cost | A time in the past 12 months when you needed to see a doctor but could not because you could not afford it? | MEDCOST | MEDCOST | MEDCOST | MEDCOST1 | MEDCOST1 |  |
| Check Up | About how long has it been since you last visited a doctor for a routine checkup? | CHECKUP1 | CHECKUP1 | CHECKUP1 | CHECKUP1 | CHECKUP1 |  |
| Alcohol | Had at least one drink of alcohol in the past 30 days? | DRNKANY5 | DRNKANY5 | DRNKANY5 | DRNKANY5 | DRNKANY6 |  |
| Marijuana | # Days used marijuana or cannabis in past 30 days | MARIJAN1 | MARIJAN1 | MARIJAN1 | MARIJAN1 | MARIJAN1 |  |
| Cigarettes | Smoked at least 100 cigarettes (5 packs) in your entire life? | SMOKE100 | SMOKE100 | SMOKE100 | SMOKE100 | SMOKE100 |  |
| Physical Activity | Physical activity or exercise during the past 30 days other than their regular job? | \_TOTINDA | \_TOTINDA | \_TOTINDA | \_TOTINDA | \_TOTINDA |  |
| Physical Health | # days in past 30 days physical health not good? (includes physical illness and injury) | PHYSHLTH | PHYSHLTH | PHYSHLTH | PHYSHLTH | PHYSHLTH |  |
| Difficulty Walking | Serious difficulty walking or climbing stairs? | DIFFWALK | DIFFWALK | DIFFWALK | DIFFWALK | DIFFWALK |  |
| BMI | Four-categories of Body Mass Index | \_BMI5CAT | \_BMI5CAT | \_BMI5CAT | \_BMI5CAT | \_BMI5CAT |  |
| Weight | Reported weight in kilograms | WTKG3 | WTKG3 | WTKG3 | WTKG3 | WTKG3 |  |
| Asthma | Asthma status | \_ASTHMS1 | \_ASTHMS1 | \_ASTHMS1 | \_ASTHMS1 | \_ASTHMS1 |  |
| Depression | Ever told you had a depressive disorder? | ADDEPEV2 | ADDEPEV3 | ADDEPEV3 | ADDEPEV3 | ADDEPEV3 |  |
| General Health | Good or better health? | \_RFHLTH | \_RFHLTH | \_RFHLTH | \_RFHLTH | \_RFHLTH |  |
| Mental Health | 3 level not good mental health status | \_MENT14D | \_MENT14D | \_MENT14D | \_MENT14D | \_MENT14D |  |
| CHD/MI | Ever reported having coronary heart disease (CHD) or myocardial infarction (MI)? | \_MICHD | \_MICHD | \_MICHD | \_MICHD | \_MICHD |  |
| Stroke | Ever told you had a stroke? | CVDSTRK3 | CVDSTRK3 | CVDSTRK3 | CVDSTRK3 | CVDSTRK3 |  |
| Skin Cancer | Ever told you had skin cancer? | CHCSCNCR | CHCSCNCR | CHCSCNCR | CHCSCNCR | CHCSCNC1 |  |
| COPD | Ever told you have chronic obstructive pulmonary disease, C.O.P.D., emphysema or chronic bronchitis? | CHCCOPD1 | CHCCOPD2 | CHCCOPD2 | CHCCOPD3 | CHCCOPD3 |  |
| Kidney Disease | Not including kidney stones, bladder infection or incontinence, were you ever told you had kidney disease? | CHCKDNY1 | CHCKDNY2 | CHCKDNY2 | CHCKDNY2 | CHCKDNY2 |  |
| Arthritis | Had a doctor diagnose them as having some form of arthritis | \_DRDXAR1 | \_DRDXAR2 | \_DRDXAR2 | \_DRDXAR3 | \_DRDXAR2 |  |
| Diabetes | Ever told you have diabetes? | DIABETE3 | DIABETE4 | DIABETE4 | DIABETE4 | DIABETE4 |  |