Impact of Demographics and Behaviors on Diabetes and Cardiovascular Diseases (CVD)

Team members: Yu Fu, Hao Li, Nina Nguyen

Problem statement

Chronic disease, especially diabetes and CVD, is a considerable burden for patients and health insurance payers in the United States. Diabetes and CVD can lead to high medical bills, impaired physical mobility, and increased risk of sudden death, ultimately causing a decline in the quality of life of patients. The factors that cause these chronic diseases are the complex effects of genes and acquired environments. Demographics and daily human behaviors, including diet, sleep, and exercise, are important environmental factors for chronic diseases. The goal of this project is to study the correlations between chronic diseases and demographics plus daily behaviors and identify important variables in predicting indicators of chronic diseases. These important variables could be valuable in further studies on causal relations.

Data sources

- The Behavioral Risk Factor Surveillance System (BRFSS)
    - The primary source for demographics, behavior and health condition variables
    - 2018 BRFSS data https://www.cdc.gov/brfss/annual_data/annual_2018.html
- Data EDA and cleaning are finished. There are **437436 records and 188 variables** in total.
    - EDA Notebook: https://github.com/yf23/FromBehaviorToHealth/blob/master/EDA_All_Variables.ipynb
    - EDA Report (PDF): https://github.com/yf23/FromBehaviorToHealth/blob/master/EDA_Reports_All_Variables.pdf
- **Target variables (y)**: binary diabetes indicator and CVD indicator
- **Feature variables (X)** of the following categories
    - *Diabetes-Related*: These variables are highly related to or depended on the diabetes indicator. They might be too strong predictors for diabetes indicators. We will see whether these variables overshadow the importance of other variables in our model, in order to determine whether to use these variables as features.
    - *Survey Related*: Year, month and day of the survey
    - *Demographics*: Variables including location, residence, cell phone usage, age, gender, ethnicity/race, education, marital status, employment, income.
    - *Health Status and Conditions*: Current health conditions and diseases. The target variables of diabetes and CVD are selected from this category.
    - *Cancer*: Cancer is a special module in the survey. It covers variables of conditions and treatments related to cancer.
    - *Oral Health*: Visit frequency and number of teeth removed.

    - *Behavior*: Multiple categories including *Smoking, Alcohol Consumption, Sleep, Exercise, Drive* and *Sun Exposure*
    - *Healthcare Access*: Mostly related to the frequency of doctor visits and health insurance
    - *Health Check, Test, and Treatments*: The name and frequency of tests or treatments that have been applied to a respondent person.
    - *Children*: Demographics and health conditions of the children of respondent persons.
- Sample weights: There is a variable _LLCPWT representing the sample weight of each record in the BRFSS dataset. The sample weights can be used as an input in predictive models.
- Missing Value Analysis
    - Number of variables with NA percentage 90% or higher: 52/188
    - Number of variables with NA percentage 80% or higher: 80/188
    - Number of variables with NA percentage 70% or higher: 85/188
    - Number of variables with NA percentage 60% or higher: 96/188
    - Number of variables with NA percentage 50% or higher: 112/188

Algorithms / solution technologies

1. Apply EDA and cleaning for all variables. Boxplots and distribution plots are used for continuous variables, and count plots (with percentages) are used for categorical variables. Plots and cleaning steps can be viewed in the EDA report (see link above). During cleaning, binary categorical variables are converted to 0 and 1 for easier data preprocessing of models. For some variables, outliers are identified using [Q1 – 1.5IQR, Q3 + 1.5IQR] range, then clipped or marked as NA.
2. Build predictive classification models on the diabetes indicator and the CVD indicator using all features. The predictive model that will be used is Random Forest, which is a tree-based ensemble model using bagging.
3. The dataset will be divided into training and test in a ratio of 8:2. The model will be evaluated with F1 score and/or AUC ROC score (Area Under Curve of Receiver Operating Characteristic). Accuracy is not a good metric here because the indicators are skewed with a lower percentage of 1's. A grid search of hyperparameters of the random forest will be performed using the cross-validation method.
4. The random forest will output the feature importance scores. If some score is overwhelming high, we need to check whether there is already a dependency relationship in the survey. Finally, we can get a list of important features variables that affect diabetes and CVD indicators the most. A sensitivity analysis can be performed (Maybe? Not sure how to do it on tree-based models).

Risks

1. There is a risk that none of the behavior variables are strong predictors of target variables, and the overall model is not a good predictor of target variables. That means the demographics and behavior data does not show correlations with CVD and diabetes indicator with the BRFSS data.
2. A large amount of NA values in data. 112 out of 188 variables have 50% or higher percentage of NA values. This might negatively affect the model, and we need to try different models before deciding whether to throw out some features with extremely high NA percentages.

3. Another issue from the survey is that the sample size is always limited. Some minority groups might not be well represented in the survey. The minority groups are the groups with low prevalence rates in the combination of various demographic variables. For example, we might only have one record represents the group "native American + 70-79 years old + <15000 annual income + high school education". As a result, we will either completely ignore these minority groups or have a very high bias on them in our solution.
4. The dataset is large, and the hyperparameter grid search space is also large, but laptops are slow. There is a time risk to tune the model finely.

Challenge

1. One of the challenges is to Identify which variables can be dropped due to high NA percentage or due to minimal importance score. We will ultimately need to decide what will variables will have a direct effect on our target variables.
2. Trying out different models requires lots of experiments, which is time-consuming. We need a good design for the experiment to pick the right models.
3. The interpretation of feature importance scores in the final presentation is challenging. Only posting the scores seems not enough.

Plan for Completion

- *Where we are now:* We just finished EDA and data cleaning for this large dataset. The workload of EDA and cleaning is large due to the number of variables. Now we have a cleaned dataset that is ready for applying predictive models.
- *The following steps:*
    a. (Week of Nov 18, 3 hours/person) Setup code for the pipeline: preprocessing -> train model -> evaluate on test set. Data preprocessing includes the one-hot encoding for multi-categorical variables and the train-test split of the dataset. Use the pipeline to train models with default random forest hyperparameters and see whether there is a need to modify the features used for each target variable.
    b. (Week of Nov 25, 3-6 hours/person, but expect a long code running time) Set up code to perform the grid search. Find the model with hyperparameters that give the best evaluation metrics and use the model as the final result.
    c. (Week of Dec 2, 4 hours/person) Interpret the results and prepare the presentation

Citation

This project is inspired by Behavior Predictor from PricewaterhouseCoopers DoubleJump Health.

https://www.pwc.com/us/en/industries/health-industries/library/doublejump/behavior-predictor.html

https://www.pwc.com/us/en/industries/assets/pwc-2019-us-hi-behavior-predictors-placemat-rev6.pdf