95-851 Making Products Count: Data Science for Product Managers
Homework 2: Clustering and Prediction
Fall 2022

Due 11:59 PM EST November 18, 2022

Overview

On https://www.countyhealthrankings.org/ , the University of Wisconsin's Population Health Institute has published rankings of all 3142 counties in the US by their health outcomes and behaviors. In class, we described some exploratory data analysis on the underlying data from 2017. In this assignment, you will apply clustering and prediction techniques to the data from 2022 to determine if there are interesting groupings of counties with similar health outcomes and behaviors, and develop a predictive model to see which factors influence health outcomes. Individuals and governments can use such a model segment the population and improve public health outcomes.

Tasks

The data file for you to analyze is in the assignment in Canvas and also available at https://www.countyhealthrankings.org/sites/default/files/media/document/analytic_data2022.csv . Documentation of the measures in the data file is in https://www.countyhealthrankings.org/sites/default/files/media/document/2022%20Analytic%20Documentation.pdf .

Part of the goal of this homework is to give you practice in choosing features when many are available. You should select from among those features mentioned on p. 3 of the measure documentation (the "ranked measures," not the "additional measures").

- Use only the columns with the words "rawvalue" (e.g. v001_rawvalue) for your features; you don't need to analyze the numerators, denominators, confidence intervals for this assignment.
- Input features for your models should include only those listed under "HEALTH FACTORS" (e.g. smoking) in the documentation and not the "HEALTH OUTCOMES" variables (e.g. premature death, poor or fair health, poor mental health or low birthweight) – including such outcome variables would lead to data leakage.
- It is OK to leave out features that could not reasonably be changed by public policy. You should just explain what you are doing and note your assumptions.

The remainder of https://www.countyhealthrankings.org/ is worth exploring to obtain more context about how this data is collected and used to improve public health.

You should turn in a Jupyter notebook with code and results including visualizations to answer the following questions (20 points total for the assignment):

1) What steps did you use for exploratory data analysis and subsequent data preparation on this data set? This should address summary statistics and handling of missing data and outliers (5 points)
2) Are there any noteworthy groupings of counties that have similar health outcomes and behaviors? You should use an unsupervised learning technique like clustering and show how you decided on the number of clusters. (5 points)
3) What are the five most important factors predicting premature death as shown by this data? In this data set, premature death is defined as the number of years of potential life lost before age 75 per 100,000 population. Develop two models (using different supervised learning approaches) to answer this question. (5 points)
4) Which of the two models do you believe is more accurate and how can you tell? Where would you focus public health efforts to reduce premature deaths in Allegheny County? (5 points)

Rubric Points Distribution:

Item	Points
Exploratory data analysis and data preparation	5
Development of clustering model	3
Determination of number of clusters	2
Development of first supervised learning model predicting premature death	2
Development of second supervised learning model predicting premature death	2
List of 5 most important factors influencing premature death	1
Evaluating accuracy of the two supervised learning models	3
Recommendations for reducing premature death in Allegheny County	2
Total for full credit	20

Hints/Suggestions:

1. Describe how you have chosen to handle outliers and missing values
2. Include visualizations such as histograms and boxplots where appropriate in describing the results of EDA
3. Describe how the findings from EDA influence your choice of data preparation and choice of modeling techniques
4. For the supervised learning models, consider how you will select the features and measure accuracy of the resulting models. How can you avoid overfitting?
5. Do not rename datafiles
6. Use relative path while importing the data

Submission:

The submission will be a Jupyter Notebook with the name *DSPM_HW2_<your Andrew id>.ipynb*, and an html export of the notebook.