Analysis of Death in the United States

Bryan Brent, Nathan Lile, Alex Ray, Austin Pearman

Project Description

This project is about gaining a better understanding of mortality, particularly with respect to the interrelationships between features such as race, sex, day and month of death, age, education, and circumstance of death in the United states from 2005 to 2015.

Research Questions

To what extent can we predict manner of death from day-of-death features such as age, education level, race, and marital status?

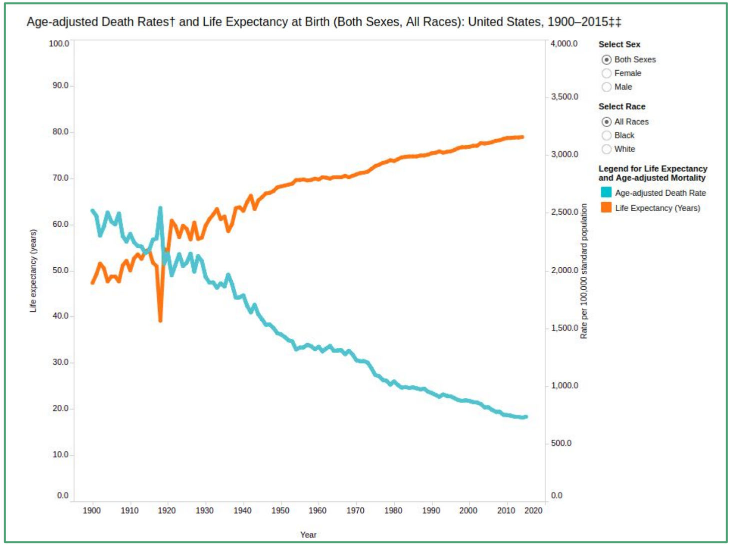
To what extent can we predict a general cause of death from day-of-death features such as age, education level, race, and marital status as well as historical features and death specifics such as activity during death, place of injury, and descendant status?

To what extent can we extract interesting information regarding mortality trends and patterns from a cluster analysis of mortality features?

To what extent are we able automatically “roll up” ~500 causes of death into meaningful supergroups?

Prior Work

- Lessons learned from data mining who mortality database
  - Useful conclusions: “Classification tools produced the poorest results in predicting cause of death. Given the inadequacy of variables in the WHO database, creation of a classification model to predict specific cause of death was impossible. Clustering and association tools yielded interesting results that could be used to identify new areas of interest in mortality data analysis. This can be used in data mining analysis to help solve some quality problems in mortality data.”
- Graphs of trends of deaths in the United states from 1900 until 2015 NCHS data visualization
  - Useful conclusions: Graph displaying a century long longitudinal study showing trends in life expectancy and age adjusted death rates.



Age-adjusted Death Rates† and Life Expectancy at Birth (Both Sexes, All Races): United States, 1900–2015‡‡

Select Sex
◉ Both Sexes
○ Female
○ Male

Select Race
◉ All Races
○ Black
○ White

Legend for Life Expectancy and Age-adjusted Mortality
🟦 Age-adjusted Death Rate
🟧 Life Expectancy (Years)

Prior Work (cont)

- Heart Diseases and Mortality data among adults ages 35 and older
- CDC NVSS (National Vital Statistics System) Publication Page
    - Contains list of publications on the CDC mortality data (our dataset), spanning everything from simple trend analyses to specific classification and clustering problems
- Applying data mining in medical data with focus on mortality related to accident in children
    - Uses data mining techniques in with goals related to ours; achieves results with both decision trees and Bayes’ theorem which gives insight into possible techniques in our analysis

Datasets

- Death in the United States
    - https://www.kaggle.com/cdc/mortality
    - Data from 2005-2015 of death in the United States from the National Vital Statistics System
        - Every death includes a wide array of attributes including age, race, cause of death, decedent status, education, marital status, month, day (among others)
    - Provided by the US Center for Disease Control and Prevention via Kaggle
    - Downloaded by *Bryan Brent*
- Temporality in the CDC dataset opens the door for integration of other data to add additional features (weather, stock market, political events, etc)

Proposed Work

- Data cleaning
  - Several areas in the json files are of an incorrect syntax, essentially missing quotation marks.
  - Converting cause-of-death and age “recodes” to a more usable format
  - Ensuring continuity of feature representation and format across the entire dataset
- Data preprocessing
  - Possibly “rolling up” cause-of-death
  - Determining features that can be dropped to reduce dimensionality
  - Exploring any features that can be joined for simplicity, continuity
- Data integration
  - Possible to integrate more data into existing set
    - Weather
    - Sporting events
    - Political events
    - Financial markets

Tools

- Pandas
  - Data conglomeration, manipulation
  - Provides helper functions for data cleaning
- Scikit-learn
  - Models for ML algorithms (classification, clustering, etc)
- Matplotlib
  - For simple visualizations
  - Utilized by Pandas
- D3.js
  - For more advanced visualizations, infographics, etc

Evaluation

We plan on evaluating results for each of our research questions, *individually*.

Classification questions can be evaluated numerically with measures such as accuracy and precision

The “interestingness” of clustering depends, to some extent, on the predictability of the results as well as various measures of closeness

The success of “rolling up” cause-of-death features depends on how well we’re able to semi-autonomously create a meaningful spanning set of causes to aid in other data mining tasks

References

- Lessons learned from data mining of WHO mortality database
  - https://www.ncbi.nlm.nih.gov/pubmed/21691674
- NCHS Data Visualization
  - https://www.cdc.gov/nchs/data-visualization/mortality-trends/
- Applying data mining in medical data with focus on mortality related to accident in children
  - http://ieeexplore.ieee.org/document/4777728/?reload=true
- Heart Diseases and Mortality data among adults ages 35 and older
  - https://data.cdc.gov/Heart-Disease-Stroke-Prevention/Heart-Disease-Mortality-Data-Among-US-Adults-35-by/r35g-znws/about
- CDC NVSS (National Vital Statistics System) Publication Page
  - https://www.cdc.gov/nchs/nvss/deaths.htm

References (cont.)

- Applying data mining in medical data with focus on mortality related to accident in children
  - http://ieeexplore.ieee.org/document/4777728/