Cause of Mortality and Medical Transcript Analysis

October 6, 2020

1 Background

Every year, centers for disease control and prevention (CDC) provides detailed statistics of deaths and their underlying causes in the United States. The CDC mortality data is used by various industries like medical, health and insurance to provide better services. It provides the basis of numerous researches and is widely cited in public papers.

Because of your outstanding knowledge, you are hired as a data scientist by a prestigeous insurance company. The VP of your department wants to launch a life insurance product but would like to do an analysis on the cause of death in the US population first. The following questions needs to be answered:

2 Problems

2.1 Cause of Death in the U.S.

- What are the major causes of death in the US (show your evidence)?
- Are the major causes of death changing over time? Are there any significant increasing or decreasing trends? Can you explain the trends (using references, other sources of data etc.)? How can you address the significance of the trends using statistical modeling method?
- For different ages what are the main causes of death? What are the main causes of death among young people? What are the main cause of death among old people (can you design a way to visualize your findings)?
- Pick a few causes of death and estimate the probability of death due to those causes (For example: build a binary classification model to understand the impacts of age, gender, calendar year, education ... For example, estimate the death caued by heart disease v.s. other diseases).
- Explore the data creatively and see if you can get any insights beyond your VP's question list (OPTIONAL).

2.2 Medical Transcript

You have sample medical transcripts of 5000 patients."**medicaltranscriptions.csv**". You want to determine if any patients have medical conditions that are associated with ICD codes of major causes of death.

- Design a similarity measure metric
- Calculate the similarity measure between ICD code descritption and medical transcirpts
- Assign ICD code to a medical transcripts only if the similarity score is above certain threshold.

3 Data

- Download the CDC mortality data between 2005 and 2015 from `https://www.kaggle.com/cdc/mortality` the data files include "**2005_data.csv**", "**2006_data.csv**" **...** "**2015_data.csv**". Each file contains 77 columns and the fields are selectively described as below:
  - current_data_year: the year of the statistics
  - detail_age: death age
  - icd_code_10th_revision: ICD10 code, a medical classification list by the World Health Organization (WHO). It contains codes for diseases, signs and symptoms, abnormal findings, complaints, social circumstances, and external causes of injury or diseases. A detailed descriptions of the code can be found in **allvalid2011 (detailed titles headings).xls**"
- Medical transcripts of 5000 patients: "**medicaltranscriptions.csv**".

4 Project Requirements

Please use the data provided and do the analysis that are needed to answer the questions listed above. Put together a report and present it to the VP of your department. Specifically,

- Form a team of 3-4 people and work together to accomplish the tasks.
- Prepare a report that describes the project's background, your methods, results (including graphs) and conclusions.
- Put together a powerpoint and present your results.
- Submit the report together with the codes.