Cause of Mortality and Medical Transcript Analysis

September 11, 2021

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 prestigious 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.

Understand the cause of death and analyze the trends. Pick a few causes of death and do deep analysis (e.g. death caused by heart disease v.s. other diseases). You may want to analyze the cause of death by using different data science techniques that you learned in class, for example, visualize the data, statistical modeling. Answer the following questions:

- What are the major causes of death in the US? Do the causes vary by year, age, gender or other variables? Do you see any interesting trends? Visualize the trend by variables.
- Can you design statistical models to address the trends' significance against age/gender/year/seasons/locations etc.? You can look at one cause at a time. How can you address non-linear trends using feature engineering? Did you do any other types of feature engineers?
- Can you verify the trends using ways other than statistic modeling (citing references, other sources of data etc.)?
- Explain the design of your model, the process of your variable selection. How do you prevent overfitting? Is the size of the data causing issues for your analysis and how do you handle the issues?
- What are the major contributing factors/variables to the cause of death that you picked.

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.
- How would you assess the accuracy of your similarity score design? Are they making sense?

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.