AI-Powered Socioeconomic Prediction of Lifespan

Rustom Ichhaporia [rustomi2@illinois.edu][*]

2020-12-12

Abstract

How long will you live? This age-old question has extensive implications in the billions of risk estimations made by individuals planning for the future every day. Although never certain, a stronger approximation of an indidual's lifespan can enable more reliable future planning and a greater sense of stability than none at all. We reviewed publicly available datasets containing socioeconomic information about U.S. citizens to create a naïve model that predicts the likelihood of a person's death at different ages given characteristics such as location, income, place of birth, and more. The results are explained and visualized in this report. While more work must be done to achieve a more accurate predictor, this work provides a baseline for lifespan prediction in coordination with other financial models to aid financial planning.

1 Background

1.1 Desired Features

Lifespan prediction is important to a broad variety of scientific and industrial fields. Each field in which it is relevant requires a different scope, accuracy, and set of input features. For example, the medical field might look at a patient's weight and blood pressure to determine their risk of passing away from heart disease. Millions of specific, dynamic variables affect an individual's lifespan, ranging from minute physical details to sociological environments to occupational conditions. Naturally, it is currently impossible to accurately measure all of these variables for an individual, let alone every individual. Thus, we must let the domain of our application dictate which variables we use, as well as the size and diversity of the dataset we use to predict lifespan.

If we were to make a perfect predictor of mortality, some of the high-level features we might consider include:

- General biographical information (e.g. age, sex)
- Location
- Occupational and residential environment conditions
- Income
- Medical information

In particular, the most likely available features would fall into the groups of socioeconomic and medical. Unfortunately, in our dataset search, finding a dataset that combined socioeconomic status, medical information, and lifespan information was difficult to come by. There are datasets linking two of those three features, but a comprehensive, large scale study documenting all three with useful sample sizes was not found. As a result, some compromises had to be made in favoring the socioeconomic data over medical data, as that is more relevant and available in actuarial settings.

[*] This research project was completed during my time as a research intern at the Illinois Risk Lab (https://irisklabuiuc.wixsite.com/) during the Fall of 2020. My research was a part of the AI-Powered Lifecycle Financial Planning project, which is still under development. I appreciate the help of Dr. Runhuan Feng, Dr. Frank Quan, Dr. Yong Xie, Dr. Linfeng Zhang, and my fellow interns throughout the process. Thank you!

1.2 Dataset Selection

1.2.1 CDC Dataset

The first dataset that we attempted to use was the Mortality Multiple Cause-of-Death dataset created by the U.S. Center for Disease Control (CDC)[1]. While this dataset contained several of the features that we wanted to include in our analysis, it was still missing a lot of the socioeconomic factors that we were looking for and the medical information it contained was difficult to parse. After a few weeks of attempting to work with this data, we decided to search for a new dataset that better matched the needs of the research.

1.2.2 NLMS Dataset

The best dataset that we found within our timeframe was from the National Longitudinal Mortality Study (NLMS) created by the United States Census Bureau[2]. The NLMS

2 Preprocessing

The appendix of this report contains the code for training the model and saving the results in a file. It does not include the code for statistical plots.

3 Modeling

4 Results

Last week, I registered for HAL access so that I could run the hyperparameter optimization script remotely because my computer overheated and could not run it for the appropriate number of trials. I was able to upload my files and run the first half of the script, but unfortunately when running the `fmin` optimization function, the program crashes after the first of 150 loops with the error:

4.1 Limitations

One significant drawback of the approach taken to estimating

[1] https://www.cdc.gov/nchs/nvss/mortality_public_use_data.htm

[2] https://www.census.gov/topics/research/nlms.html

5 Appendix