Project Proposal: Predicting Death

Kenneth Lipke (kel89), Charlotte Wang (xw476)

September 22, 2017

We want to predict why and when someone will die. This is would be of incredible value to really any kind of insurer. The most obvious beneficiary would be a life insurer attempting to calculate premiums, but other insurers would benefit as well. For example, a car insurer may want to know for how much longer they can expect an elderly customer to be paying premiums, and if they will be alive long enough after an accident to pay premiums sufficient to cover the cost of the accident. These are all very morbid, and arguably sinister uses. One could, however, use this tool to decide how much they need to save for retirement, or when they need to begin estate planning.

In order to tackle this question, we plan to use the Center for Disease Control's Mortality Database, which can be found at `https://www.cdc.gov/nchs/nvss/deaths.htm`. This database contains individual files for the years 2005 to 2015. Each one contains information on the deaths of every person in the United States (and territories) for that year. As a result, each one is approximately 2.5 million observations. There is the obvious data one would expected, date of death, age upon death, gender, education, marriage status. Beyond these basic variables, however, there are the more interesting race and ethnicity, along with geographic location, and past injury data.

This vast data presents a number of opportunities for interesting analysis. As we already mentioned, our initial plan is to attempt to predict the age of death. As we have so much data, we have the freedom to concoct very complicated models without having to worry too much about over fitting. Therefore, we can push "linear" models to their limits: looking at countless interactions, and non-linearities. Other models that we are currently considering are tree based models. We have a hunch that this data is going to be very fragmented which may respond particularly well to these methods.

We realize this is a very ambitious and difficult question to attempt to answer, however, we feel we stand a good chance of succeeding. Firstly, the data is disturbingly vast. Over the ten year period we will have over twenty millions observations. Even if after cleaning we only have ten percent of observations left, we will still have over two millions observations. Further, there is a wide variety of possible predictors, as already discussed. The data is also very well documented. A significant number of the variables are encoded (which will likely be a headache during cleaning), but the CDC documentation for each year spends considerable time decoding and explaining the meaning of each variable. Again, as mentioned above, the large number of variables gives us confidence that we will be able to find at least model that makes good predictions.

In summary, we think this has the potential to be an incredibly useful tool, both from the perspective of companies (mostly insurance companies of all kinds), as well as for individuals. There is tremendous data, that is vast and well documented, and we feel that we have the tools required to make sense of it.