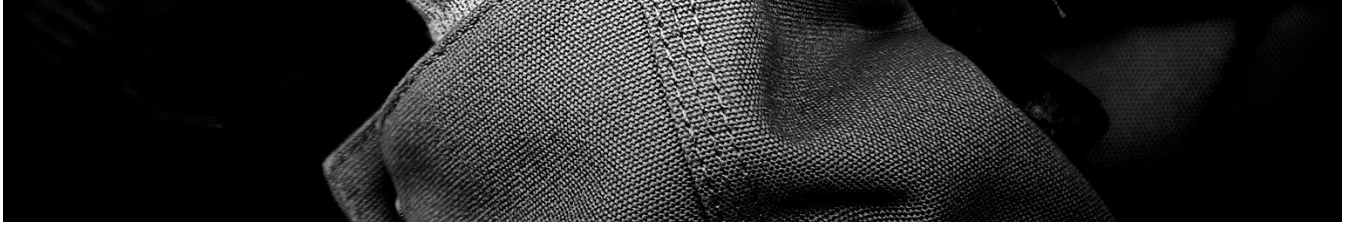
Machine learning to predict life expectancy

Caitlin McDonnell Follow

Feb 4, 2018 · 4 min read



We're in an unprecedented era where humans are living longer and longer (although 2015 & 2016 were a scary blip in the otherwise upward trend). It's no secret, though, that life expectancy varies widely across the globe.

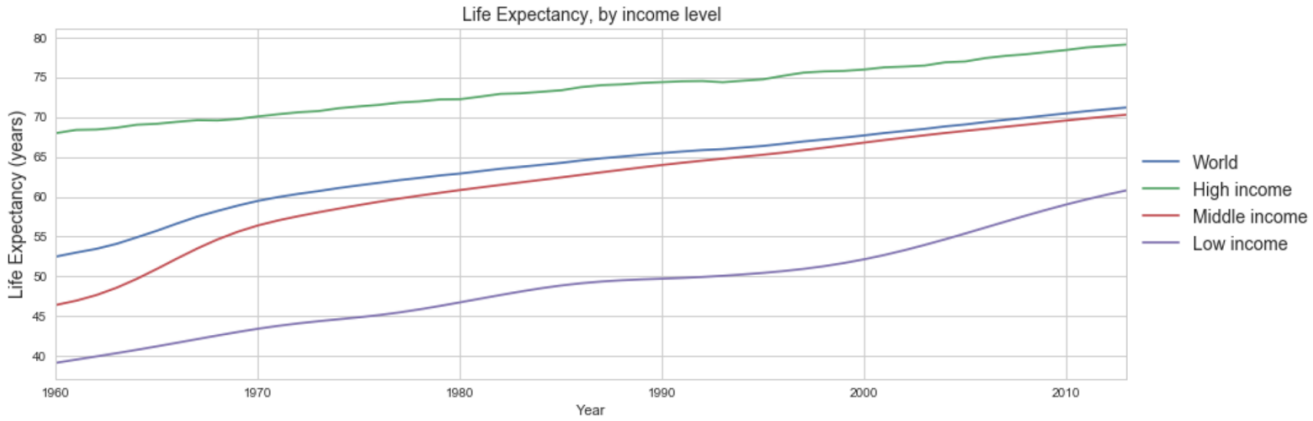
As a biomedical engineer, I've always attributed living longer with increased access to modern science and healthcare. To test this assumption and understand what actually best predicts life expectancy, I explored country-level data across a subset of the World Health Organizations's 1,345 development indicators (from 1960- 2015).

I pared the data down to a subset of indicators that were not only reasonably related to life expectancy, but also representative of the main indicator categories (Economic Policy & Debt; Health; Infrastructure; Poverty; Private Sector & Trade; Public Sector; Social Protection & Labor).

Of these, the most statistically significant indicators of life expectancy was access to sanitation. Using sanitation alone to predict life expectancy had 77% accuracy and showed that a a one-year increase in life expectancy is associated with every 0.25% of the population with access to modern sanitation facilities.

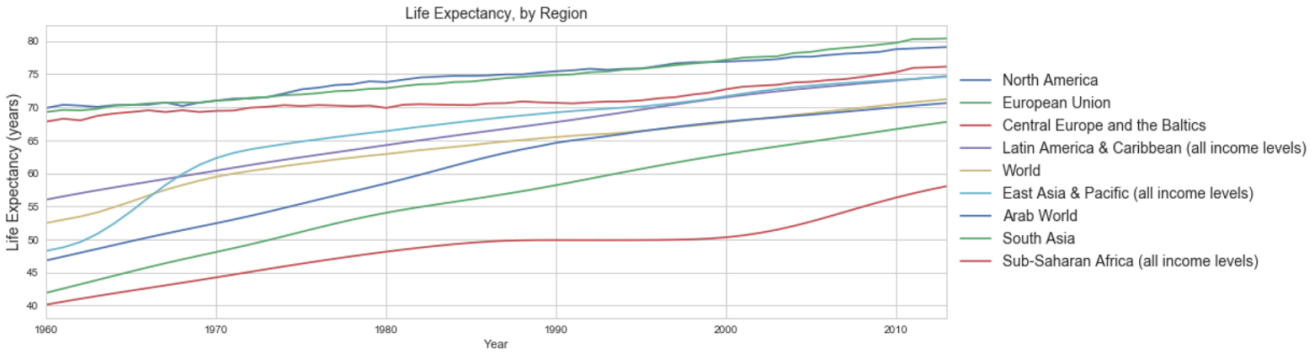
"Up and to the right"

Pretty much no matter how the data is cut, life expectancy trends between 1960 and 2015 look like sales projections in a startup pitch deck: up and to the right.



Life span by country income levels: The gap between high and low income countries is closing

Classifying countries based on income levels, we can see how economics impact health: a higher income is associated with a longer life expectancy, continuously between 1960 and 2015. High income countries see a steady, but slower, increase in life expectancy; middle income countries experienced large growth in the 1960s and saw a corresponding up-tick in life expectancy. In low income countries, life expectancy gains all but stopped between mid-80s — mid-90s, followed by a strong increase in life expectancy in recent years. Regional groupings (bottom graph) reveal which geographies drove the changes in the 1960 and 1980s.



Life expectancy by region: Sub-saharan Africa (red, bottom) has a noticeable plateau in the mid-80s, coinciding with the HIV/AIDs epidemic.

. . .

I pared down the 1.3k development indicators to a representative subset of 21. Running a correlation on these metrics vs. life expectancy provides a quick assessment for relative importance.

Life expectancy has the highest positive correlation with: sanitation (r = 0.88), GDP per capita (r = 0.698), and health spend per capita (r = 0.63). There's a strong negative correlation with: % of population in rural areas (r = -0.80) and adolescent fertility (r = -0.77).

In case you're new to reading heat maps: A dark red square indicates a high positive correlation between the column & row header; whereas a blue square represents a negative relationship.

. . .

Machine learning to predict life expectancy

To determine which indicators are statistically significant, and to predict life expectancy, I ran a few different models with various combinations of features. This application is a perfect use case for regression, which determines the relationship between one dependent variable (life expectancy) and a number of independent variables (development indicators).

Actual vs. predicted (model) life expectancy: The model tends to better predict higher life expectancy, and does not preform as well for lower life expectancies (as indicated by wide-spread data in the bottom left).

Using the entire subset of 21 features was 89.7% accurate, and showed that sanitation and rural population (%) were both statistically significant.

Even with just these two variables, this model predicts life expectancy with 80.4% accuracy. On average life expectancy is 63 years with a one-year increase for every 0.18% increase in access to sanitation and 0.11% decrease in the percentage of population living in rural areas. Adding healthcare spend to the model only slightly increases accuracy (81%).

Probing into sanitation: To dig deeper into how sanitation specifically, I evaluated access to sanitation against statistics on child mortality rates. Not surprisingly, there's a strong negative correlation between sanitation and mortality: as sanitation improves, mortality rates decrease for neonates, infants, and children under 5.

. . .

Sanitation as the key predictor of life expectancy — specifically child mortality — was unexpected, yet logical. Poor sanitation conditions increase the risk of sharing life-threatening contagions.

> According to the World Health Organization, diarrheal disease, the leading cause of death for children under five, is spread by poor sanitation conditions:
>
> *'Diarrhoea is a symptom of infections caused by a host of bacterial, viral and parasitic organisms, most of which are spread by faeces-contaminated water.'*

Once sanitation has been improved in communities, initiatives could address other key indicators of life expectancy (e.g. education, health care, adolescent fertility). And, because sanitation is highly-correlated with childhood mortality, addressing on sanitation first should result in more children who would benefit from additional initiatives.

gussie/Springboard_data_science

For data and source code (Jupyter notebook), check out my github.

github.com

Machine Learning Life Expectancy Development Indicators Who Health