Nirmal Balachundhar, Biraj Parikh, Nibir Pathak
EECS 349 Fall 2018 Final Project

Machine Learning for Public Health

1. Problem Statement

Our task is to determine the most important economic and healthcare traits of a country that have the largest impact on the average citizen's general health. While certain traits such as the health of the economy and hospital spending appear obvious, a deeper investigation is required to both reveal new characteristics that may be just as important. This can be achieved by learning the factors that are pivotal to different public health systems in different countries. We hope that our project sheds light on the best ways a country can allocate its investments in different sectors of public health to maximize the well-being of its population.

2. Dataset

Our dataset consists of various public health related attributes. These attributes can be seen in Figure 1 of the appendix[1]. Some of the attributes track the progress of various SDGs (Sustainable Development Goals), a set of goals set out by the UN in 2015 that aim to create a world where all people enjoy a free and adequate standard of living. We believe these to be reliable attributes to track as their monitoring and reporting is done by a reputable independent international organization.

We attempt to balance our attributes by including certain 'good' attributes as well as 'bad' attributes. For example, our 'good' attributes include the total amount of physicians and nurses per every thousand and the universal health coverage index which includes essential health services, tracer indicators, and service coverage inequalities. Some of the 'bad' features would include the neonatal, under-5, air pollution, and sanitation mortality rates. We believe that a combination of these attributes represents a holistic view of a given country's public health system.

The output values we will test against utilize the Health Access and Quality Index (HAQ), which summarizes healthcare access and quality for a given location. HAQ is a known scientific standard that represents a range of health service areas. These include, but are not limited to, vaccine-preventable diseases, infectious diseases, maternal and child health, and non-communicable diseases such as cancer, cardiovascular diseases and diabetes.

3. Note

To preface our results, we'd like to provide context to the data we collected.

Some of the data from sources collected information from the countries themselves. These countries may have altered these values for political reasons. We can't account for these outliers as these governments as government transparency is outside the scope of this project.

The number of countries, fragmented recording of data over the years prior 2016, and the strong requirement for recent data impeded us from training and testing on larger datasets. The approximately 200 instances, each of which corresponded to one country, were not enough for the model to find a strong reproducible correlation. This can be seen in Figure 2. Within the figure, two datasets have been created: the first, our original dataset, uses information from the years 2015 and 2016. The second dataset uses information from 2010 and 2014. All the attributes are constant. We independently trained two linear regression models on this data, and found that the weights were extremely different for either model. This should not be the case for data that is only a few years apart. That is, the output variable's construction should not change so drastically such that the weights between each model have this much variability. This implies that we don't have enough data to make any verifiable claims. Furthermore, Figure 3 shows the p-values for a select number of correlations. The majority of these values are less than 0.05 and therefore show that the correlations are statistically significant. However, because of the dearth of data, these values can largely vary. As a result, our correlations suffer from data scarcity as well.

4. Methods & Results

We use multivariate linear regression models and feed-forward neural networks to analyze our data. Both of these would be most appropriate for our task since they are advantageous when attempting to answer mysterious regression problems. Polynomial regression isn't used as we found it difficult to quantify how well the model performed.

Our linear regression model predicts HAQ index values with an average R2 score of 0.9219, which suggests that a strong linear correlation between our input variables and the HAQ index. This average score is calculated with 10-fold cross validation. For our dataset, 10-fold cross validation is preferable to a held out test set because we can generate more models with the same amount of data, and therefore more accurately test these models.

We find that the UHC Index has one of the greatest corresponding weights in the regression. This suggests that countries should aim towards meeting the UHC goals set out by the WHO and World Bank[2]. These include maximizing coverage over essential health services such as doctors' services, inpatient and outpatient hospital care, prescription drug coverage, pregnancy and childbirth. All of these factors are components of a national health insurance, which implies that a national health service would be beneficial to public health.

In addition, prepaid private spending per capita has one of the highest weights but has a negative effect on the efficacy of a public health system. This occurs because high private spending indicates that the richer portion of the population doesn't believe the government provided healthcare to be sufficient for a good quality of life. Not only does this indicate that measures taken to improve/maintain public health may be inadequate, but also that there is a large imbalance in the public insurance pool. That is, those who cannot afford private healthcare are the only citizens to remain within the pool. These individuals, because of their lower income, probably lead a less healthier lifestyle. This increases the government's risk, and thereby raises their health insurance costs. In the end, this money could be spent elsewhere. Sanitation and under-5 mortality were the next largest predictors for an HAQ index and they both suggest that it might be effective to redistribute resources within the public health system with an increased awareness on sanitation in public areas.

With all of this being said, the fact that our weights vary between datasets so largely (see section 3) means that while there is a strong linearity in the data, the components of this linearity have not been assigned accurate weights.

Layer #	# Nodes	Activation Function	Layer #	# Nodes	Activation Function
Input	128	ReLU	Hidden 2	256	ReLU
Hidden 1	256	ReLU	Output	1	Linear

The above table summarizes the feed-forward neural network regressor that we implemented. In order to measure the accuracy of our model, we evaluated the mean absolute error (MAE) of our predicted values versus the actual values in the test set. Ten different models were trained over 500 epochs with stochastic gradient descent. There are ten models due to our use of 10-fold cross validation. Such a low MAE suggests that the neural network learns an accurate function to predict the HAQ index given our set of attributes. A nonlinear activation function (sigmoid) was also tested in the output layer. However, the MAE at each epoch for all the models in the cross-validation did not change, suggesting that the function cost had reached an unsurpassable minima. With so little instances, it is less likely that the data can be modeled by a complicated function. As a result, the linear regressor does much better than a nonlinear regressor. The final average MAE across all 10 models was 2.5358.

While there would be more research required to finalize a specific NN architecture, we believe that such a model presents a lot of opportunities for governments since they can test different input values for certain attributes and understand its direct quantitative effects on the HAQ index.

Throughout the study, we found some interesting correlations between some of our attributes which may have impacted their assigned weights in both regression problems. Our explanations for these correlations are mere hypotheticals that could greatly benefit from further research as well as domain knowledge. First, we found that developmental assistance towards the country had a positive correlation with an increase in health spending as a percentage of their GDP. At first, this correlation did not make much sense since countries that spend more on health seemingly wouldn't need foreign assistance. After further evaluation, we think that this may be due to the growth induced nature of some developing countries. As a country increases spending as a percentage of its GDP on healthcare, it reaches a limit in the total value provided. However, foreign countries may perceive this maximum level of healthcare as insufficient and would continue to aid developing countries. This would contribute to the positive correlation between both variables.

Second, we found that both neonatal mortality and under-5 mortality were highly correlated with UHC. This seems extremely unlikely because if a country has good health coverage for its citizens, common sense tells us that there should be less death. However, the fact that developed countries have easier access to medical staff and the proper medical facilities during and after pregnancy gives a larger chance that the child is born alive in the first place. We believe that because there are more live births in developed countries, there will be greater neonatal and under-5 mortality per 1000 persons. In developing countries with less access to the proper medical facilities, there may be less live births and therefore less chance of such death occurring.

5. Future work

We think that there is a lot of potential for future work in this area. First, public health systems would benefit from a machine learner that evaluated the effect of prepaid private insurance on the inequality between access to healthcare for the rich and the poor. We think that it may be important to evaluate this difference because it may shed light on the difference in the quality of healthcare that both groups receive based on the specific amount of prepaid private health care. Second, we think that public health systems could benefit from understanding how they can better optimize their systems under given budget constraints. In order to truly create an optimal system, one must further study the relationships between investments and allocation of resources in each of those categories.

Given more time and resources, this project could greatly benefit from a larger amount of data. This could include some of the following: more nuanced features identified by someone with greater domain knowledge, incorporating the time of data into the models, or a combination of both. An accurate estimator of public health and what influences a good public health system would be an extremely powerful tool.

All the members in our group decided that we wanted to best understand the entire project from a holistic standpoint, so we all contributed to the project an equal amount. This included initial research, data mining, model research and evaluation, and the actual project itself.

Appendix:

1. **List of Features**

Literacy rate	Doctors per 1000	Nurses per 1000	Health spending as % of GDP	Total Health Spending per Capita	Gov. Spending per Capita	Prepaid Private Spending per Capita
Out of pocket Spending per Capita	Development Assistance for Health Spending per Capita	Neonatal Mortality per 1000	Under-5 Mortality per 1000	Universal Health Coverage (UHC) Index	Air Pollution Mortality per 1000	Mortality due to Sanitation per 1000 (WaSH)

2. **Weights of each Feature in multivariate linear regression (Average across 10 folds)**

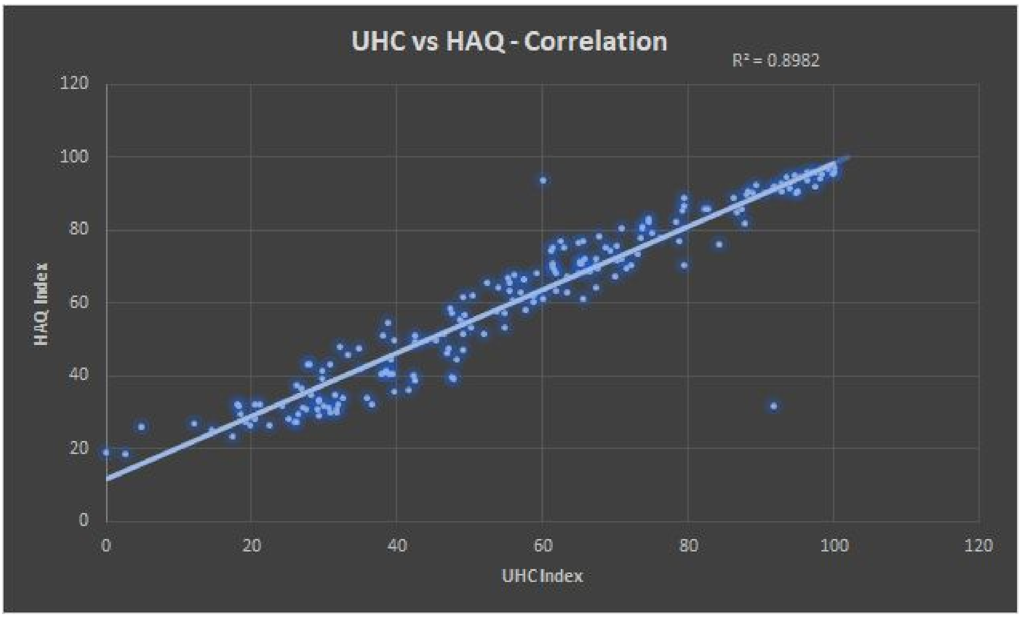
Attributes	Trial 1 Weights	Trial 2 Weights
Literacy Rate	2.6949	6.1144
Doctors	4.1033	4.1665
Nurses	-0.33709	1.2665
Health Spending	-12.902	-24.557
Total Health Spending	4092.8	3205.5
Government Spending	-2490	1890
Prepaid Private Spending	-1568.9	1233.6
Out of Pocket Spending	-1368.5	1089
Development Assistance for Health	-184.98	135.24
Neonatal Mortality	-1.9007	-5.08
Under-5 Mortality	12.323	12.927
Universal Health Coverage	45.587	48.23
Air Pollution Mortality	2.4137	-0.66729
Mortality due to Sanitation	21.446	19.3

3. **Correlation Index (Select Few Correlations)**

	Spending(%GDP)	UHC	Gov Health Spending (per capita)
Neonatal Mortality	0.1464, p=0.045	0.912317, $p=4.81^{-74}$	0.7636, $p=3.54^{-37}$
DAH	0.6636, $p=3.14^{-25}$	-0.3934, $p=2.33^{-8}$	-0.2904, $p=5.28^{-5}$

Doctors (per 1000)	0.1710, p=0.0189	0.7549, p=6.46^{-36}	0.6109, p=1.28^{-20}

4. UHC vs HAQ Correlation



Resources:

[1] Global Health Data Exchange, 2018, Institute for Health Metrics & Evaluation, http://ghdx.healthdata.org/gbd-results-tool

[2] World Health Organization, “Tracking Universal Health Coverage”, 2018 https://www.who.int/health_financing/topics/financial-protection/qa-tracking-uhc/en/