Nirmal Balachundhar, Biraj Parekh, Nibir Pathak

Machine Learning for Public Health- Status Report

Our task is to determine the most important economic and healthcare traits of a country that can help the country control a public health crisis. This can be achieved if we get to know which factors are pivotal in preventing the effects of a crisis from escalating to alarming levels. With this knowledge, public health organizations can effectively optimize their investments in different sectors to maximize wellbeing for their population. The factors could very well include investment in medical research, number of vaccinations programs, number of health personal per capita, literacy rates etc.

Our dataset consists of various attributes that are related to public health for almost every country. These attributes include the number of health professionals per every thousand people, percentage of government spending on healthcare, percentage of external aid to that country, health expenditures as a percentage of GDP, both out-of-pocket healthcare spending and prepaid private insurance spending as a percentage of total government spending. Our output values that we will be testing with utilizes the Health Access and Quality Index which summarizes healthcare access and quality for a given location. This is a known scientific standard that accounts for different factors and standardizes healthcare access measures across different countries. In addition, it is commonly used in many research publications. We utilize about 186 examples which cover almost every single country in the world. We plan on using 70% of our data for training and 30% for testing. In order to validate and improve various machine learning models, we will conduct 10-fold cross validation.

We have used two regression models to analyze our data. These include a linear regression model and a polynomial regression model. Our linear regression model predicted HAQ index values with an R2 score of 0.81. This suggests that a linear correlation is present between our input variables and the HAQ index. Our polynomial regression model predicted HAQ index values with an R2 score of 0.54. This suggests that a polynomial model is most likely not a good fit between the input variables and the HAQ index. We used the scikit-learn implementations of the polynomial and linear regression models to calculate these values.

For the remainder of the quarter, we would like to test different regression models such as logistic regression and neural networks. These models would potentially increase the accuracy of the numbers because they have the capability to learn different linear and logistic relationships that may be present between the input variables and the output variables. Specifically, neural networks utilize hidden layers to assist the model in learning to a better degree. Currently, our input variables have different orders of magnitude. As a result, the weights from our linear and polynomial models were not scaled relatively to each other. Therefore, we plan on scaling our features so that we can obtain correct representations of how important our attributes are.

We currently have about eight distinct attributes, but a lot of them are correlated amongst each other. We would like to find new attributes that may be able to shed more light on the nature of public health in different countries. Lastly, we would like to perform dimension reduction on our attributes. We would understand which attributes have the largest impact on the

regression and create the most correlation. Furthermore, by using dimension reduction, we can get to the most valuable investments that a public health department in a given country can make.