Important Factors in Human Development

Katie Harris

Intro to Artificial Intelligence (95-891)

# Problem Context

There is obvious inequity in today’s world. Development organizations, both national and international, try to address this problem through various initiatives and projects. The goal of this project is to aid decision makers in those organizations to better allocate resources and funding. Human Development has been measured by the Human Development Index (HDI) since 1990. This index was developed by a Pakistani Economist to “shift the focus of development economics from national income accounting to people-centered policies”.[[1]](#endnote-1) It is a geometric mean of normalized indices for three dimensions: a long and healthy life, being knowledgeable, and having a decent standard of living. These are boiled down into three separate indexes.

1. Life Expectancy Index (LEI)

LEI is 1 when life expectancy at birth is 85 and 0 when it is 20.

1. Education Index (EI)

Mean Years of Schooling Index (MYSI)

MYS: Years that a person aged 25 or older has spent in formal education.

Fifteen is the projected maximum for this indicator for 2025.

Expected Years of Schooling Index (EYSI)

Eighteen is equivalent to a master’s degree in most countries.

EYS: Total expected years of schooling for children under 18 years of age

1. Income Index (II)

The HDI uses logarithm of income, to reflect the diminishing importance of income with increasing Gross National Income.

II is 1 when GNI per capita is $75,000 and 0 when GNI per capita is $100.

GNIpc: Gross national income at purchasing power parity per capita

This measure heavily weighs education, health, and income as metrics of development. While these are certainly important, it does not reflect all of what human development entails. For instance, it ignores factors like inequity, poverty, and gender disparity.

To better understand what factors are important for human development as measured by the HDI, I looked at financial data from the International Monetary Fund ([IMF](https://data.imf.org/?sk=4C514D48-B6BA-49ED-8AB9-52B0C1A0179B&sId=1409151240976)) and more general data from the CIA world [factbook](https://www.cia.gov/the-world-factbook/).

# Data

I used several data sets from the IMF which look at individual metrics over years. Specifically, I looked at the Consumer Price Index, Government Finance, Nominal GDP, Exchange Rate, and Trade Value from before 2019 and how well they related to HDI in 2019. This data was easily processed in python given that it all came in the format of excel or csv files. The main problem with this data was that it was often incomplete, missing values for different years or for certain countries all together. For example, Government Finance, which I was particularly interested to see its impact on HDI had many missing data points and only 47 countries.

Table

Description automatically generated

The CIA world book data had the opposite problem. It was so dense and detailed that I could not use a library function to import it but rather ended up parsing the data myself to pull out relevant information. The CIA world factbook is essentially an online encyclopedia of all the data publicly available that the CIA has collected on 266 countries. It includes information in various categories including history, economy, geography, military, transnational issues and more. I was able to access this data in JSON format via a GitHub [API](https://github.com/iancoleman/cia_world_factbook_api/blob/23920af57109e4205112239ff0468bb7c665e584/data/factbook.json). This API included 162 countries.

Text

Description automatically generated

I was able to parse this data into a Pandas data frame with 38 fields. The code is provided in the attached Juypter notebook.

The HDI data set was directly available from the [UN](http://hdr.undp.org/en/indicators/137506). This data set contains the HDI for 189 countries from 1995 to 2019. It also contains aggregate data for different countries based off their geographic location and development status.

From this chart we can see that overall HDI is trending upwards, and the world has made significant gains in the last 20 years.

# Approach

To better understand how these factors impact human development I planned to run different regression models on both the world factbook data and the financial data. I would then evaluate the best models using R-squared as a metric. Once I found the best model, I would use the feature importance to answer my question of which factors are most important for development. I also wanted to look further in the data and see if we saw different factors being important in different regions.

# Implementation

Before running the regressions though I created a confusion matrix to eliminate some covariates.

Graphical user interface

Description automatically generated with medium confidence

Here we can see those features that are highly correlated with each other i.e., have |r| > .8:

* Median age is highly correlated with birth rate, mother’s mean age at first birth, infant mortality rate, life expectancy at birth, and physician density.
* Birth rate is highly correlated with median age, mother’s mean age at first birth, infant mortality rate, life expectancy, and total electrification.
* Maternal mortality rate is highly correlated with total electrification and infant mortality rate.
* Mother’s mean age at birth is highly correlated with median age.
* Infant mortality rate is highly correlated with median age, birth rate, and life expectancy
* Life expectancy is highly correlated with median age, birth rate, and infant mortality rate.
* Total electrification is highly correlated with birth rate, maternal mortality rate, and the infant mortality rate.

From here I eliminated life expectancy, median age, birth rate, maternal mortality rate and the infant mortality rate.

I chose to implement 4 different regression models: a basic linear model, a lasso model, a random forest model, and a gradient boost model (See attached notebook file for more details). I used the basic linear regression model to establish a baseline. I chose the 5 features with the highest correlation coefficients: total electrification, mother’s mean age at first birth, drinking water source, physician density, literacy rate, and population growth rate to run the regression.

Chart, histogram

Description automatically generated

I evaluated the models using adjusted R-squared as a metric, with a 70/30 train/test split. Gradient boost did the best with an adjusted test r-squared score of .91, followed by random forest, followed by the basic linear model, and lastly the lasso model had an adjusted test r-squared of .69.

Because the gradient boost model did the best, I chose to use it as my model for the rest of the regressions I ran when I separated the dataset by geographic region.

# Results

Looking at random forest feature importance we can see these results.Chart, histogram

Description automatically generated

Here we see that the percent of people with access to electricity as well as the percent of people with access to drinking water come in far ahead the rest. This makes sense given how important electricity is to productivity and how essential drinking water is to life. The next few factors also make sense based on how the HDI is calculated. We saw earlier how strongly correlated mother’s mean age at first birth was to median age which in turn was highly correlated with life expectancy. Similarly, literacy rate as well as physician density are likely strongly correlated with the education domain and life expectancy domain of the HDI. The other factors make considerably less impact but also make a reasonable amount of sense. I’d assume a country with high net migration, more people coming than going, would be sign of opportunity in that country. Gross national savings is also important for economic growth. Population growth rate is strongly correlated with the birth rate which is strongly correlated with life expectancy. Adult obesity and the death rate I’d also expect to be highly related to life expectancy.

I decided to look at the SHAP values for the gradient boost model to see if there were any major differences.

A picture containing graphical user interface

Description automatically generated

While the first 6 are the same, the next 4 are a bit different. Days since independence, population growth rate, death rate, and the percentage of unpaved road are next. Days since independence would be important, given how many of the developing countries were previously colonized. Colonization, through the stealing of resources and sowing of political discord, has been devastating some countries’ development.

Here is a map of HDI for reference. The darker the green the higher human development. Map

Description automatically generated[[2]](#endnote-2)

Here we can see that those with lower HDI are in South and Central America, Africa, and South East Asia. To compare these regions, I created separate subsets of each of these regions. Here are those SHAP values.

Graphical user interface

Description automatically generated

Africa

Table

Description automatically generated with medium confidence

South and Central America

Chart, scatter chart

Description automatically generated

Southeast Asia

The important features for each region are relatively similar. Physician density and drinking water availability rank high in the top ten for all of them. Adult obesity also ranks higher in each of these regions than in world as a whole. In Africa, we see that agriculture as economic sector and days since independence are particularly important. Independence from Spain and military expenditures are both important features in South and Central America. In South East Asia, it is difficult to say anything with confidence since the sample is so small.

From here I would recommend we look specifically in different countries for opportunities to invest in projects related to health care and drinking water in developing countries. Specifically in Africa, it also seems beneficial to invest in the agriculture sector. This also likely comes from the fact that many sub-Saharan countries are not yet industrialized so having a economy heavily reliant on agriculture is much more common. In South America, where the average HDI is higher it seems beneficial to invest in improving the literacy rate. Given a higher HDI, the need for a more skilled labor force is reasonable. These findings are preliminary. Also, noticeably there are features high on the list of importance that are not changeable, like independence from Spain and area of a given country.

While I did run regressions on the financial data too, it was clear the models were not correct. The resulting R-squared coefficients were non sensical; some were above 1. See the jupyter notebook for more details.

## Limitations

These are regressions with advanced models. They are not experiments or pseudo-experiments using instrument variables or fixed effects. I therefore **cannot** **conclude** **causality** within any of my findings. I do think, however, the findings still show us important factors related to human development and would aid decision makers in the development space.

It should also be mentioned how susceptible these models are to overfitting given the small sample size.

# Next Steps

Panel regression should be applied to the financial data. Panel regression is designed for longitudinal data (also known as panel data) and would likely present more reasonable and accurate results.

Also, I would like to reemphasize these are preliminary results. It would be necessary to look at programs or initiatives that invested in the same features that are deemed important by the regressions and see how effective they were. Also, countries are unique in their history, culture, economic resources, etc. These factors need to be considered when thinking about development.

# Lessons Learned

Given the nature of economic data, it is much more suited to econometric analysis like Panel regression. While python has the capabilities to run such regressions there are better tools for that. Traditional AI supervised learning methods are not suited for this data given the small size and rampant missing values. Missing or inaccurate data is a common problem in applications of AI in development. Because of this, one of the most common uses of AI in development is to fill in this missing data.[[3]](#endnote-3)

If I were to redo this project, I think I would try to focus more on the CIA world factbook data and not include the IMF data. I would have liked to focus in on a case study or regional study instead.

1. https://en.wikipedia.org/wiki/Human\_Development\_Index [↑](#endnote-ref-1)
2. Ibid [↑](#endnote-ref-2)
3. https://mdp.berkeley.edu/an-overview-of-the-application-of-ai-in-development-practice/ [↑](#endnote-ref-3)