Analyzing Global Development Using Health and Education

Jimmy Chen, Data Analytics
Spring 2025

Problem Area

BACKGROUND

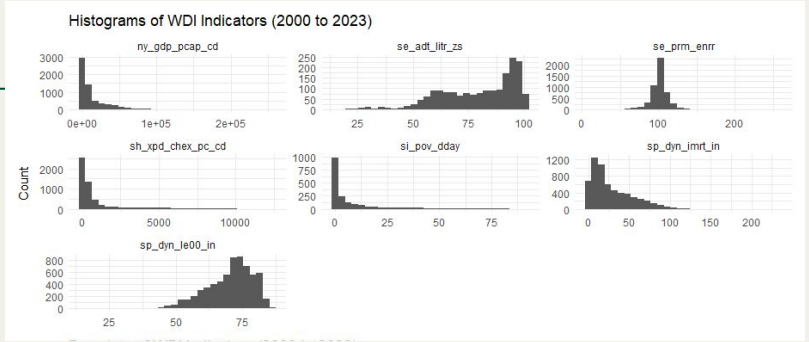
Why am I interested?

- To figure out how economic performance relates to health and education.

What do you want to predict/hypothesize?

- I believe that higher health-expenditure per capita and better education outcomes (adult literacy & primary-school enrolment) lead to higher GDP per capita.

Dataset Source



Source

World Development Indicators (WDI) from the World Bank, accessible via the World Bank DataBank.

[https://databank.worldbank.org/source/world-development-indicators]

Metrics/Applications

Economic: GDP per capita, Poverty Headcount Ratio

Health: Life Expectancy at Birth, Infant Mortality Rate, Health Expenditure per Capita

Education: Adult Literacy Rate, Primary School Enrollment Rate

Preliminary Assessments

Comprehensive time-series data across 200+ economies spanning several decades. (1960 - 2023)

Strong skew in monetary indicators (GDP per capita and health expenditure per capita)

Exploratory Data Analysis (EDA)

Data Exploration & Analytics:

- Used R to clean and filter the bulk WDI CSV. (only took data from 2000 - 2023)
- Used skimr to get counts, means, SDs, and missing-value rates
- Applied transformations to address skewness
- Created correlation matrices that showed revealed very strong positive correlation between GDP and health expenditures, and adult literacy rates. Least correlated was primary school.

Model Construction:

- Trained both a linear regression and a random forest tree
- Tuned the forest by grid-searching its variables per split and leaf size to minimize validation error

Techniques Used/Not Used

- **Used:** interpolation for missing data; log transforms; standardization; cross-validation
- **Not used:** Aggressive outlier trimming because it already gave good, stable results within the data's time and scope

What Worked/What Didn't Work

Worked:

- Filling gaps by country median and smoothing by year kept our data complete without big biases.
- Log-transforming skewed indicators and scaling everything gave stable, comparable inputs.

Didn't Work:

- Primary School Enrollment as a indicator didn't really affect anything with less than 5% importance
- Didn't remove outliers because the data was large enough for it to not affect anything

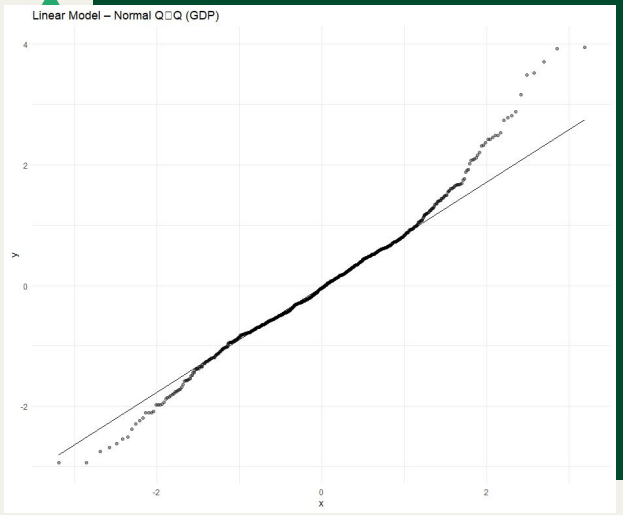
Optimizations and Uncertainties

- For the random forest model I did a small grid search over variables per split and leaf size inside the cross validation loop to pick the best settings.
- Used the spread of CV scores (mean ± standard error) to know how much performance might vary.

- Primary school rates were very lowly correlated to the overall GDP, so it's true effect is uncertain

model	rmse	rsq	mae
<chr>	<dbl>	<dbl>	<dbl>
Linear	0.230	0.917	0.171

Prediction Results




```
> print(cm_lin)
          Truth
Prediction Low High
      Low  338   24
      High  13  326
> print(cm_rf)
          Truth
Prediction Low High
      Low  338   18
      High  13  332
>
```

Predicted:

Each country's GDP per capita (on a log scale) based on its health spending, education rates, life expectancy, infant mortality, and poverty levels.

Decisions/Prescriptions:

We can boost health investment as health spending was the strongest predictor for increasing GDP

Focus on policies that lengthen healthy lifespans as it also supports higher GDP

Reduce poverty and infant mortality as it correlates with both health and economic performance

Outcomes:

Linear regression was 92.7% accurate, Random forest was 94.3%

Health expenditure per capita showed up as the single most powerful lever. A one-SD increase in health spending corresponded to roughly a 0.85-SD rise in GDP.

I concluded that directing resources toward health systems yields strong economic returns, while education and social programs were less contributive but still offered additional, supportive benefits.

Thank You!