Advanced Regression and Prediction Final Project: Predicting access to and reliance on technology

Daniel Alonso

May 9th, 2021

Contents

Dataset of choice

For this project I picked a custom-built dataset obtained from the World bank Databank, specifically the World Development Indicators database. This is the "primary World Bank collection of development indicators" as stated on the database description. It is a database that especifically lists development variables, many of which also are related to health, economics, the environment and human development in a broad sense.

A second helper dataset that I also used and merged with the one obtained through the World Bank is a COVID-19 dataset constructed by Our World in Data and obtained from their GitHub repository, or directly from the direct link in the README.

The amount of historical data on both databases is massive, and the amount of possible predictors is immense. I have decided to stick to an amount of variables that is also quite large, but will be very interesting to reduce with the techniques learned in class.

Variables

I have made a really large selection of variables (over 45) initially, however, as many included NAs and a significant amount of countries and dependencies lacked a large amount of metrics, the amount of columns and countries has been reduced to 38 and 173 respectively. Nonetheless, this is still a reasonable amount of data, and also enough to make a reasonably acceptable prediction.

NOTE:

- **blue** = used for training/predicting
- **red** = target variable
- **green** = ID variables

Variables in the original dataset as constructed using the World Bank Databank tool and the COVID-19 dataset (all variables were renamed):

- **year**: year the data was obtained in
- **year_code**: code for the year as the world bank databank sets it
- **country_name**: name of the country
- **country_code**: alpha-3 ISO 3166 code for the country
- **access_to_electricity**: Access to electricity (% of population)
- **birth_rate**: Birth rate, crude (per 1,000 people)
- **life_expectancy**: Life expectancy at birth, total (years)
- **exports_perc_gdp**: Exports of goods and services (% of GDP)
- **education_years**: Compulsory education, duration (years)
- **gdp_per_capita_ppp**: GDP per capita, PPP (current international USD)
- **perc_internet_users**: Individuals using the Internet (% of population)
- **infant_mort_rate**: Mortality rate, infant (per 1,000 live births)
- **inflation_perc**: Inflation, consumer prices (annual %)
- **consumer_price_index**: Consumer price index (2010 = 100)
- **crop_production_index**: Crop production index (2014-2016 = 100)
- **goods_imports**: Goods imports (BoP, current USD)
- **prevalence_of_anemia_pregnancy**: Prevalence of anemia among pregnant women (%)
- **diabetes_prevalence**: Diabetes prevalence (% of population ages 20 to 79)
- **human_capital_index**: Human capital index (HCI) (scale 0-1)
- **net_enrollment_rate**: Adjusted net enrollment rate, primary (% of primary school age children)
- **measles_immunization**: Immunization, measles (% of children ages 12-23 months)
- **co2_emissions_gaseous_fuel**: CO2 emissions from gaseous fuel consumption (% of total)
- **fossil_fuel_energy_consumption**: Fossil fuel energy consumption (% of total)
- **fuel_exports**: Fuel exports (% of merchandise exports)
- **fuel_imports**: Fuel imports (% of merchandise imports)
- **investment_inflows**: Foreign direct investment, net inflows (% of GDP)
- **investment_outflows**: Foreign direct investment, net outflows (% of GDP)
- **mobile_subscriptions**: Mobile cellular subscriptions (per 100 people)
- **agricultural_land**: Agricultural land (% of land area)
- **greenhouse_gas_em**: Total greenhouse gas emissions (% change from 1990)
- **age_dependency_ratio**: Age dependency ratio (% of working-age population)
- **imports_annual_growth**: Imports of goods and services (annual % growth)
- **int_tourism_arrivals**: International tourism, number of arrivals
- **total_covid_cases_per_million**: Total COVID-19 cases per 1 million inhabitants
- **total_covid_deaths_per_million**: Total COVID-19 deaths per 1 million inhabitants
- **population_density**: Population density

- **median_age**: Median Age
- **human_development_index**: Human Development Index (HDI, scale 0-1)
- **ar_tech**: Access to and reliance on technology (scale 0-1)

The target variable

From a very roughly inferential standpoint and looking at the data, we can somewhat see that clearly, both developed and developing countries were hit very hard. However, there's a pattern, where a large chunk of developing countries have barely felt the effects of the pandemic (either due to isolation or the fact that there's just other more serious issues, like large-scale wars or poverty).

Mixing development metrics and COVID data might help us see whether COVID data does indeed contribute to this, and also, how well it can perform when compared to those development metrics when it comes to predicting a particularly good metric to assess development of a nation.

In my statistical learning project, I predicted the HDI Group of a country based on many of these development metrics (and a few other differing ones). One of those metrics, and one that was particularly good at assessing a country's development were **access_to_electricity**: percentage of population with access to electricity,**perc_internet_users**: percentage of population which use the internet and **mobile_subscriptions**: mobile cellular subscriptions per 100 people.

These three were incredibly good at predicting HDI and could be a fantastic substitute for it. And reasonably so, we know that the more developed a country is, the more modern its infrastructure for electricity distribution is and the more it eventually relies on the internet to perform certain acitivities.

Having lived in both a developing and developed country myself, I can say that the consistency of services like electricity supply and how common the internet was used to provide services to the population (to internet users) is quite significant.

Therefore, either of these could be used as target variable, but, as this would be somewhat repetitive, I've decided to create a weighted metric to contain all three of them. This metric will be called: **ar_tech**.

This measure will represent the following: *"Access to and reliance on technology"*, or *ART* for short.

The measure will be calculated as a weighted measure as follows:

$$\texttt{ar\_tech} = \frac{\texttt{mobile\_subscriptions} * 10 + \texttt{perc\_internet\_users} * 35 + \texttt{access\_to\_electricity} * 55}{100}$$

This highlights the importance of electricity, as none of the other two would be possible without it, and that clearly, electricity is significantly more important than use of the internet or mobile subscriptions. However, more people using the internet is a huge sign of development, this is also true for mobile subscriptions, but perhaps to a lesser extent.

Data preprocessing

The data cleanup and very basic feature selection (only removing those with too many NAs) was performed in Python. Then a small imputation was then performed in R (only for 5 remaining missing values) within the same Jupyter Notebook (called *preprocessing.ipynb*).

Methodology and steps

1. Prior to importing the data, I performed a find and replace within the .csv file with the following regex:

   ```
   \s\[[\w\S]{1,}\]
   ```

   to remove every instance of it. This regex matches a metadata tag that the world bank uses in their dataset. Following this step, the data was imported.

2. Metadata at the end of the dataset was removed. This data corresponded with indexes and regions, whereas the rest of the dataset corresponded with data specific to the countries conforming these indexes/regions. We only keep the countries.

3. The year column was converted to integer and the '..' values (which represent NAs in the World bank databank) were replaced for numpy NaNs and then the values were sorted by year and country name.
4. Data missing in later years (2021, 2020) was backfilled with data from previous years (2000-2019), as it still fits our modelling purposes, many of the backfilled data points still correspond to metrics that fit our criteria. The oldest possible data point would go as far back as the year 2000, however, it's unlikely we should have data this old, most should have been backfilled.
5. The dataset which contains the COVID-19 cases, deaths, population density, population, median age, etc.. is imported, we make a short variable selection within it, only keeping values which correspond to May 1st, 2021. Then, we merge this dataset with the previous dataset utilizing ISO code and using an inner join, only keeping countries in common in both datasets.
6. Columns with more than 45 NA values were removed as this represents around 25
7. Numerical columns were converted into floats and column names were simplified for easier future manipulation and to call the column names in a more practical way.
8. We take the **mobile_subscriptions** variable, and convert it to a percentage of the population, in this case, we leave it in ratio form and do not multiply by 100, as we prefer a 0-1 scale.
9. The categorical target variable is constructed as explained previously to this step.
10. The data is then exported as a csv (the dataset itself, called *data.csv*), along with a json (*columns.json*) file which contains the renaming used for the columns in the originally preprocessed dataset up to that step.
11. The data is then imported in an R cell and a *mice* imputation is performed using the 'cart' method with m = 5 in order to impute the very few missing values remaining (about 5 missing values).

Exploratory Data Analysis

Correlation

Here we obtain the correlation matrix maximized per correlation coefficient (using all three: Kendall, Pearson and Spearman) for our dataset.

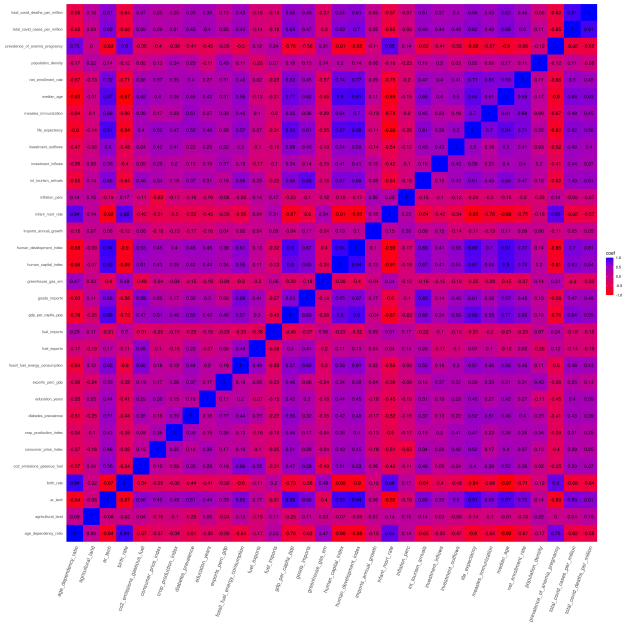


Figure 1: Correlation matrix maximizing correlation between variables

There's several things we can notice here that could prelimarily tell us how some of these variables could be excellent predictors.

The most relevant values to compare with our target variable are:

Very high correlation

Negative

- *age_dependency_ratio*: -0.84
- *birth_rate*: -0.87
- *infant_mort_rate*: -0.92
- *prevalence_of_anemia_pregnancy*: -0.93

Positive

- *gdp_per_capita_ppp*: 0.88
- *human_capital_index*: 0.92
- *human_development_index*: 0.94
- *life_expectancy*: 0.91
- *median_age*: 0.87

All these metrics are clearly all significantly descriptive of how developed a country is, those positive ones tell us that the higher their value, the more developed a country is, and the opposite for the negative ones.

A reasonable hypothesis can be proposed, that our constructed target variable is a reasonably good descriptor of development as a whole. These high correlation variables could be instrumental to our predictions

Moderately correlated

Positive

- *total_covid_cases_per_million* and *total_covid_deaths_per_million* at 0.69 and 0.61 respectively
- *net_enrollment_rate*: 0.72
- *int_tourism_arrivals* and *fossil_fuel_energy_consumption* at 0.66 and 0.65 respectively
- *goods_imports*: 0.68
- *investment_outflows*: 0.6

These moderately correlated variables could also be interesting at predicting, perhaps for better or for worse depending on how we treat them. But it's interesting to see how correlated COVID-19 cases and deaths are to our constructed target index. We could go back to a previous inference we made where we especulated that the most developed and some developing countries were hit the hardest when it comes to the pandemic, while other developing countries, while beign hit hard economically, also handled the pandemic much better than most developed countries.

The rest of the metrics in this category we can infer that are related to each country's development measure as well, but maybe not as much as those previous ones we saw.

Not correlated almost *at all*

- *agricultural_land*
- *imports_annual_growth*
- *fuel_exports*
- *population_density*
- *inflation_perc*

These variables maybe relate with development, maybe don't, but for the most part, we can't comfortably say they're very significantly related to it.

They will be used, but I suspect these will be the first ones to be ditched by variable selection algorithms or models themselves.

Modelling: statistical tools

Onto the modelling part. We will test several models, simple and robust ones, however, all will be done utilizing 5 repeats of 10-fold cross validation. Defined as follows using the *caret* package:

```
ctrl <- trainControl(method = "repeatedcv",
                     number = 10, repeats = 5)
```

Simple models

Simple linear regression (univariate):

We will test simple and multiple regression model for each set of variables, of those that have the highest correlations with our target variable.

Prevalence of anemia in pregnancy to predict ART index

Utilizing prevalence of anemia in pregnancy to predict ART index, we obtain an Rsquared of ~0.69 and a relatively low MAE of ~0.099. It's a very simple model, but we can say we're off to a good start.

Results table

Table 1: ART index ~ Prevalence of anemia in pregnancy

RMSE	MAE	Rsquared
0.127912	0.0989445	0.6855365

Infant mortality rate to predict ART index

Testing with a different strongly negatively correlated variable, in this case infant mortality rate, a metric that's usually quite high in countries with low ART index, we can see that the model is reasonably better than that of prevalence of anemia in pregnancy ~ ART index.

We obtain an Rsquared of ~0.79 and a MAE of ~0.079, which is in itself an acceptable model. Perhaps this would be one of those variables that contributes to more robust and multivariate models.

Results table

Table 2: ART index ~ Infant mortality rate

RMSE	MAE	Rsquared
0.1060661	0.078138	0.7915996

Human development index to predict ART index

As an analogous index to HDI, it is reasonable to check whether it is acceptable at predicting it in a simple linear regression model.

And we find that yes!, it's actually quite good, with an Rsquared of ~0.88 and MAE of ~0.06. This variable is probably going to contribute somewhat well in more robust and multiple models.

Results table

Table 3: ART index ~ HDI

RMSE	MAE	Rsquared
0.0798913	0.0605907	0.8795126

Simple linear regression (multivariate):

For multivariate linear regression we first test out with all the variables, utilizing the same trainControl defined as before:

Using all variables to predict ART index

Utilizing all the variables yields a reasonably good model.

The Rsquared is of ~0.89 and the MAE is of ~0.061. This is only slightly better than when using only HDI.

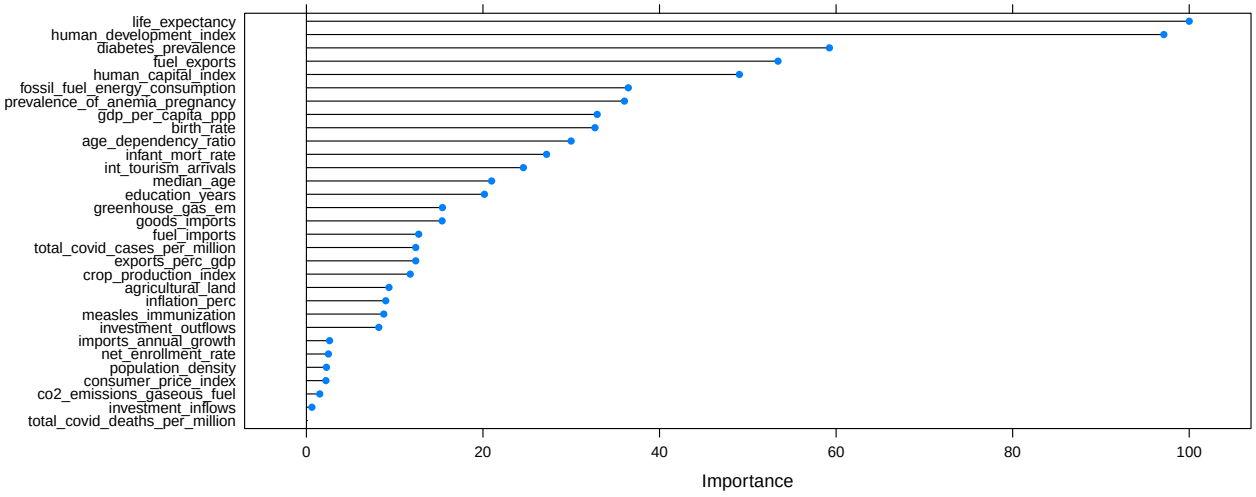
Results table

Table 4: ART index ~ .

RMSE	MAE	Rsquared
0.078643	0.0621862	0.8895499

Variable importance

As we suspected, we can see that the most important variables are headed by *life expectancy*, *HDI*, and a few others like *diabetes prevalence*, *fuel exports* and *human capital index*. Interestingly, some of these we did not expect to see up here, but seem to yield decent results along with the rest.



From our variable importance plot, we can then test a model where the variable importance only exceeds 25% importance.

Using only variables with 20%+ importance

We filter the previously shown data used to plot variable importance, and we only keep those with 20% or more importance filtering the table.

And looking at our results, we can see there's been a reasonable improvement in model quality, these apparently small incremental improvements when we have a very high Rsquared are quite welcome, and as we have done utilizing the following set of variables:

The model yields an Rsquared of ~0.915 and a MAE of ~0.054, both improving on previous model attempts.

Unfortunately, the COVID-19 metrics do not really contribute as much as we had expected before, yielding a quite low variable importance in the previous all variable model.

Results table

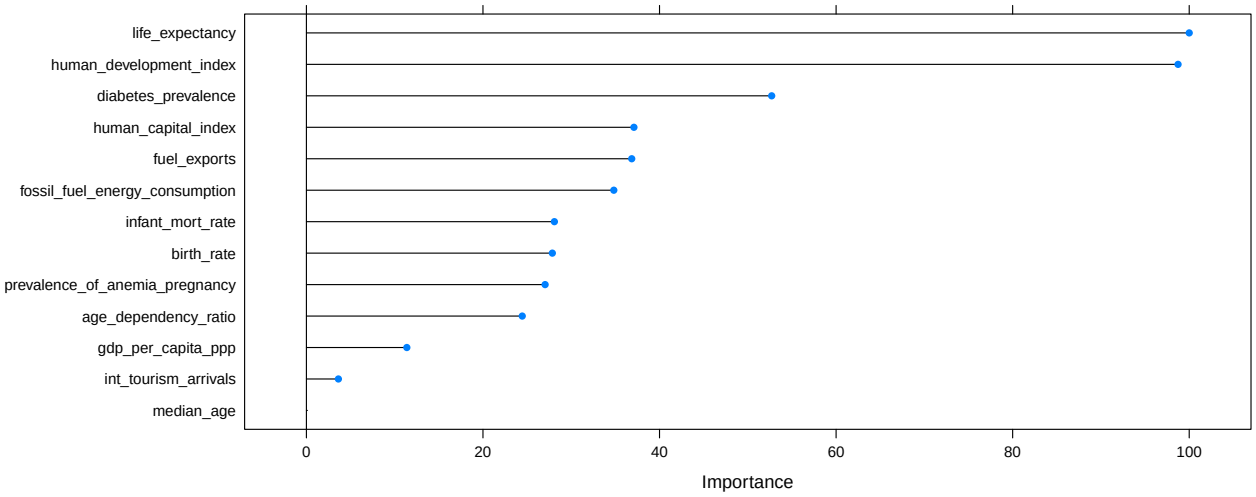
Table 5: ART index ~ 20%+ importance variables from all variable model

RMSE	MAE	Rsquared
0.0682417	0.0533993	0.9165328

Variable importance

We also want to see variable importance after clearing up variables that are less important, and we see that the top variables remain the same with a few others going down in importance, much less so than in the previous models.

Before removing these, we want to test a model with all possible interactions, just to determine whether these serve a good purpose in the model.



Including all 2-way interactions between previously selected variables

Robust models