Statistical Learning Final Project Plots

Daniel Alonso

December 19th, 2020

Contents

Dataset of choice

For this project I decided to pick 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 has lots of economic, education, energy use, and population specific metrics.

I find demographic data fascinating, and I think this dataset will be quite good for predicting country development measures along with providing quite interesting and relevant information.

Variables

NOTE:

- **blue** = used for training/predicting
- **red** = target variable
- **green** = ID variables
- **purple** = variables excluded as either they were *components* of HDI or they were 100% correlated to another variable (like GNI/GDP, which are both 100% correlated and GNI is a component of HDI)

Variables in the original dataset as constructed using the World Bank Databank tool (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
- **foreign_inv_inflows**: Foreign direct investment, net inflows (BoP, current US$)
- **exports_perc_gdp**: Exports of goods and services (as a % of GDP)
- **inflation_perc**: Inflation, consumer prices (annual %)
- **education_years**: Compulsory education, duration (years)
- **education_perc_gdp**: Government expenditure on education, total (as a % of GDP)
- **gds_perc_gdp**: Gross domestic savings (as a % of GDP)
- **gross_savings_perc_gdp**: Gross savings (as a % of GDP)
- **int_tourism_arrivals**: International tourism, number of arrivals
- **int_tourism_receipts**: International tourism, receipts (in current US$)
- **perc_internet_users**: Individuals using the Internet (as a % of population)
- **access_to_electricity**: Access to electricity (% of population)
- **agricultural_land**: Agricultural land (% of land area)
- **birth_rate**: Birth rate, crude (per 1,000 people)
- **gne**: Gross national expenditure (% of GDP)
- **mobile_subscriptions**: Mobile cellular subscriptions (per 100 people)
- **infant_mort_rate**: Mortality rate, infant (per 1,000 live births)
- **sex_ratio**: Sex ratio at birth (male births per female births)
- **greenhouse_gas_em**: Total greenhouse gas emissions (kt of CO2 equivalent)
- **urban_pop_perc**: Urban population (% of total population)
- **hdi**: human development index
- **hdi_cat**: Human development index as a category
- **life_exp**: Life expectancy at birth, total (years)
- **gdp**: GDP (current US$)
- **gni**: GNI (current US$)
- **fertility_rate**: Fertility rate, total (births per woman)

The target variable

As all these variables could perhaps tell us how developed a country is, we used a constructed categorized Human development index variable in order to classify the countries using the above variables (unless stated otherwise by their colour) as training variables.

The criteria for constructing the categorical variable *hdi_cat* was the following:

- Very high: HDI above 0.8
- High: HDI between 0.7 and 0.799
- Medium: HDI between 0.55 and 0.699
- Low: HDI under 0.55

This categorization is emulated from Wikipedia's construction and uses the same ranges as used in every Wikipedia article referencing HDI.

Data preprocessing

The data cleanup and preliminary feature engineering was done in Python and the imputation was then done in R within the same Jupyter Notebook (called *preprocessing.ipynb*).

Methodology and steps

1. Prior to importing there was a search within the .csv file with the following regex:

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

   as a find and replace and removing 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, the dataset was filtered excluding these region codes (the codes were excluded as they were country aggregates, and we're only interested in the countries themselves).

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 (2020, 2019) was backfilled with data from previous years, as it still fits our modelling purposes. The data that goes the furthest back is from 18 years prior, so 2002.

5. Columns with more than 45 NA values were removed as this represents roughly 25

6. The life expectancy, GDP, GNI and fertility rate columns were removed as these are either components of HDI or (in the case of fertility rate) are 100

7. We scrape Wikipedia in order to obtain an updated metric on HDI estimates for each country. We also scrape it to obtain the alpha-3 3166 codes for each country. The data was obtained with the purpose of making it easier to join with the dataset obtained from the World Bank, as joining by country names is not reliable enough (misses 20 countries which the alpha-3 codes do not).

8. We add HDI to the main dataframe

9. Numerical columns were converted into floats and column names were simplified for easier future manipulation.

10. The categorical target variable is constructed as explained previously in this step.

11. The data is then exported as a csv, along with a json file that shows how the columns were renamed

12. 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.