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)

- **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)

Preprocessing

The data cleanup and preliminary feature engineering was done in Python.

Methodology and steps

1 - Prior to importing there was a search within the *.csv* file with the following regex: `\"\s\[[\w\S]{1,}\]\"`