

Predicting Income Groups Using World Development Indicators

Dametreus Vincent

Capstone Project

Springboard Data Science

August 2019



WDI & Income Groups

- Indicators help us understand a country's development level
- Income groups tell us the class a country falls into economically
- Gives insight on nonmonetary measures of the quality of life
- Can help us in our quest for true globalization

Example China:

2nd largest economy in the world

Lower middle and upper middle income categories

Disparities in living standards throughout the country

Income group gives us a broad understanding of the development, living standards and quality of life



Approach

- **Identify Indicator Categories**
 - Economic Policy & Debt
 - Education and Gender Issues
 - Access to Advanced Communication
 - Health
- **Look for correlation with GNI per capita**
 - Main contributing factor to income category
- **Predict Class**
 - Anova Testing
 - Machine Learning
 - Feature Engineering



Client

- **Government**
 - Understand developmental comparisons to those in the same class and to the world
- **Non-Government Organization (NGO)**
 - Specialize in certain aspects that may help with the development of a country of interest
 - Send volunteers to countries in need of certain aid (education)
- **International Funds, Banks and Financial Institutions**
 - Financing countries
 - Direct funding for development projects



Dataset

- **World Bank's World Development Indicators Database**
 - 5 datasets featuring indicators, countries, regions, change over time, definitions, etc.
 - <https://datacatalog.worldbank.org/dataset/world-development-indicators>
- **World Bank Help Desk**
 - Historical information on income groups and GNI range for classes by year
 - <https://datahelpdesk.worldbank.org/knowledgebase/articles/378834-how-does-the-world-bank-classify-countries>



Data Wrangling

- Merge datasets
- Categorize indicators
- Drop insufficient information
- Drop unrelated columns
- Standardize column names
- Extract countries from noise
- Drop columns with too much missing data
- Drop indicators with too much missing data
- Find the income group of each country for each year
- Drop incorrect data
- Pivot the dataset for analysis and modeling



Cleaned Dataset

	country	Access to clean fuels and technologies for cooking (% of population)	Access to electricity (% of population)	Access to electricity, rural (% of rural population)	Access to electricity, urban (% of urban population)	Adjusted savings: carbon dioxide damage (% of GNI)	Adjusted savings: carbon dioxide damage (current US\$)	Adjusted savings: consumption of fixed capital (% of GNI)	Adjusted savings: consumption of fixed capital (current US\$)	Adjusted savings: education expenditure (% of GNI)	...	Urban population growth (annual %)	Vulnerable employment, female (% of female employment) (modeled ILO estimate)
14	Albania	58.14	100.0	100.0	100.0	0.753375	8.267836e+07	9.078691	9.963315e+08	3.045469	...	1.492215	57.132000
15	Albania	60.75	100.0	100.0	100.0	0.747768	9.680885e+07	9.286781	1.202302e+09	3.067987	...	1.435124	66.658001
16	Albania	63.24	100.0	100.0	100.0	0.847386	1.004384e+08	10.167240	1.205096e+09	3.090506	...	1.473288	64.900002
17	Albania	65.23	100.0	100.0	100.0	0.932935	1.101591e+08	11.951006	1.4111151e+09	3.113024	...	1.609373	62.600000
18	Albania	67.81	100.0	100.0	100.0	1.022553	1.318501e+08	11.355192	1.464162e+09	3.135542	...	1.787784	64.118002
19	Albania	69.96	100.0	100.0	100.0	1.059772	1.295891e+08	11.503503	1.406650e+09	3.158061	...	1.848379	67.010000

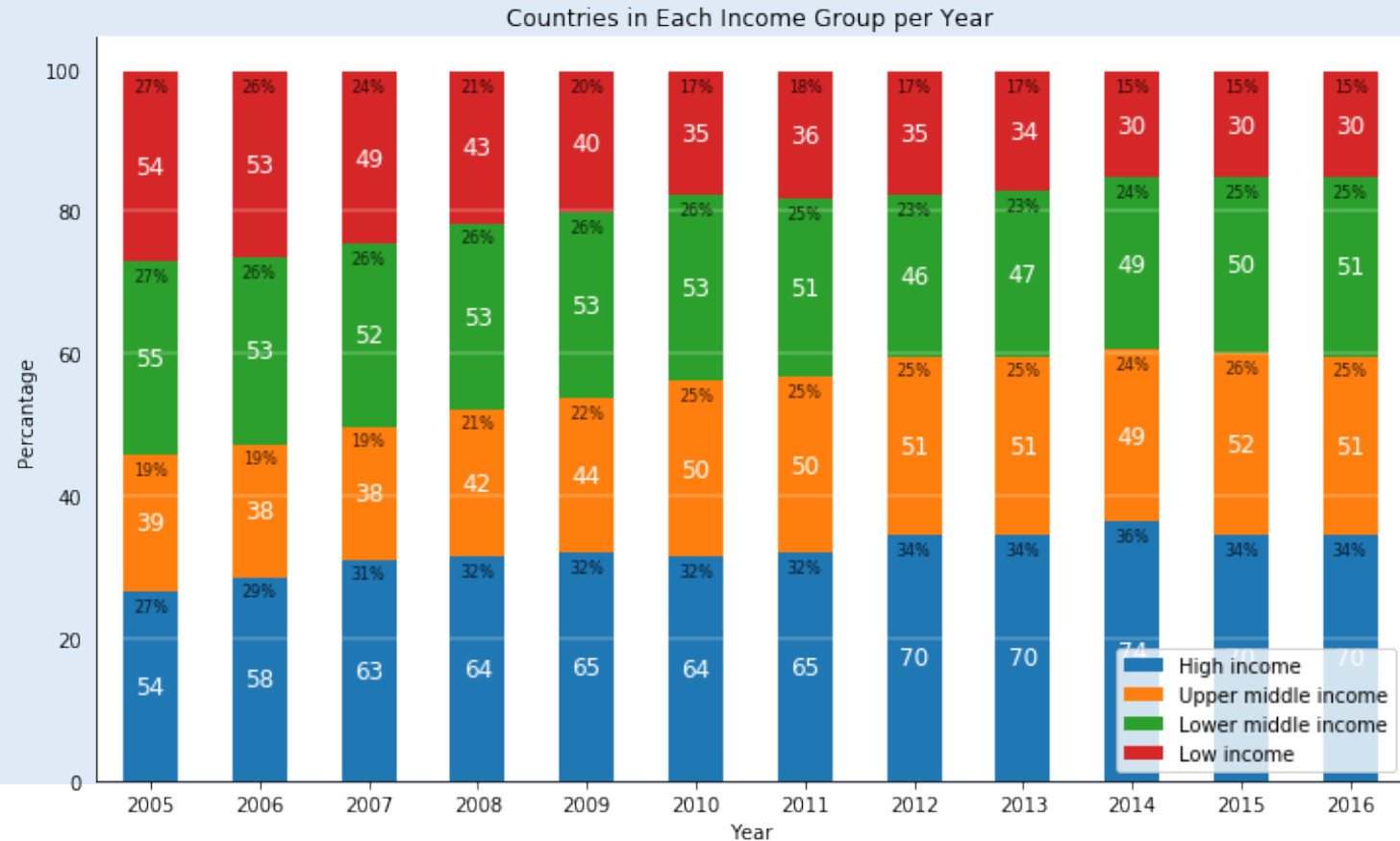
6 rows × 345 columns

Adjusted savings: education expenditure (% of GNI)	...	Urban population growth (annual %)	Vulnerable employment, female (% of female employment) (modeled ILO estimate)	Vulnerable employment, male (% of male employment) (modeled ILO estimate)	Vulnerable employment, total (% of total employment) (modeled ILO estimate)	Wage and salaried workers, female (% of female employment) (modeled ILO estimate)	Wage and salaried workers, male (% of male employment) (modeled ILO estimate)	Wage and salaried workers, total (% of total employment) (modeled ILO estimate)	year	region	income_group
3.045469	...	1.492215	57.132000	53.249002	54.881001	41.918999	41.893002	41.903999	2007	Europe & Central Asia	Lower middle income
3.067987	...	1.435124	66.658001	52.441001	58.424000	32.766998	44.959999	39.828999	2008	Europe & Central Asia	Lower middle income
3.090506	...	1.473288	64.900002	50.718999	56.473001	34.757000	47.074001	42.076000	2009	Europe & Central Asia	Upper middle income
3.113024	...	1.609373	62.600000	49.828002	55.132999	36.946999	47.812000	43.298000	2010	Europe & Central Asia	Upper middle income
3.135542	...	1.787784	64.118002	55.577000	59.194000	34.728001	42.000999	38.922001	2011	Europe & Central Asia	Lower middle income
3.158061	...	1.848379	67.010000	57.317001	61.534998	32.007000	40.738998	36.939999	2012	Europe & Central Asia	Upper middle income

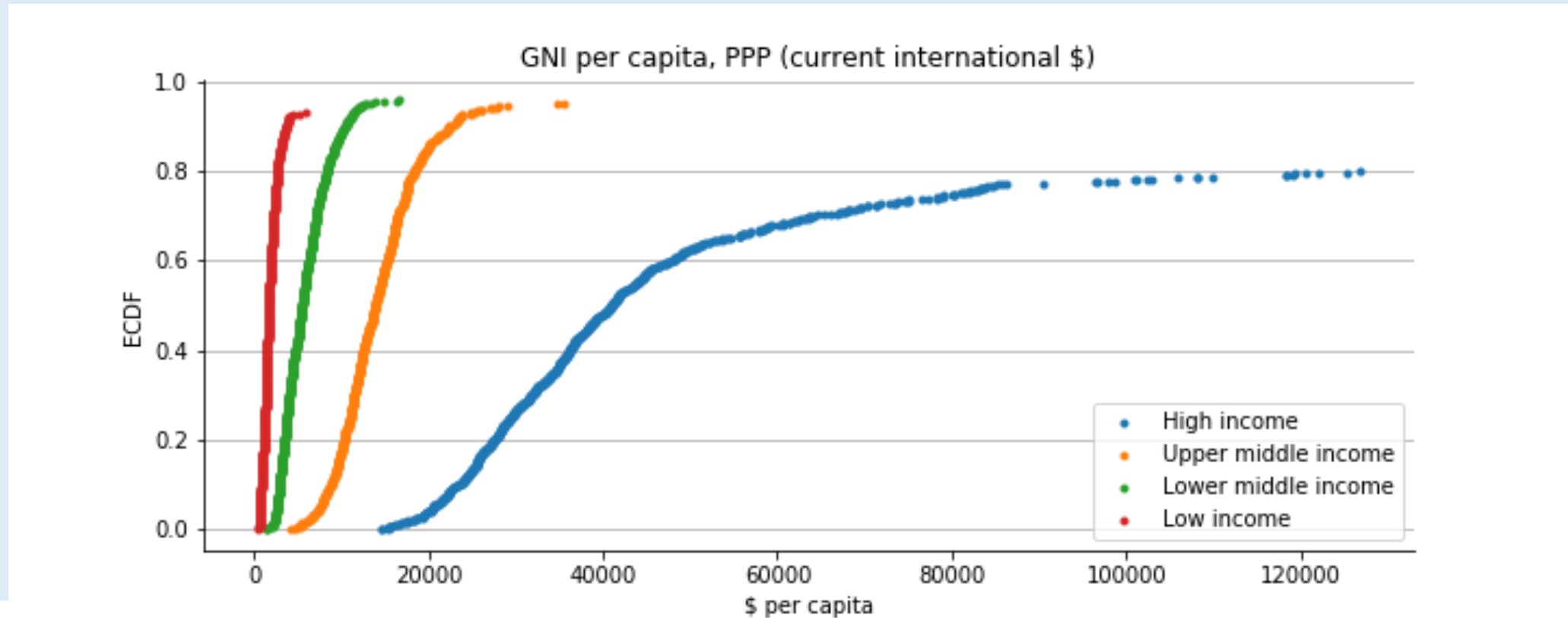


EDA – Income groups change over time

- Countries change income groups from year to year
- Higher income groups are growing while lower income groups are shrinking from year to year.
- The shift happening between income groups is showing that the world is becoming more developed.



EDA – GNI per cap direct correlation to income groups



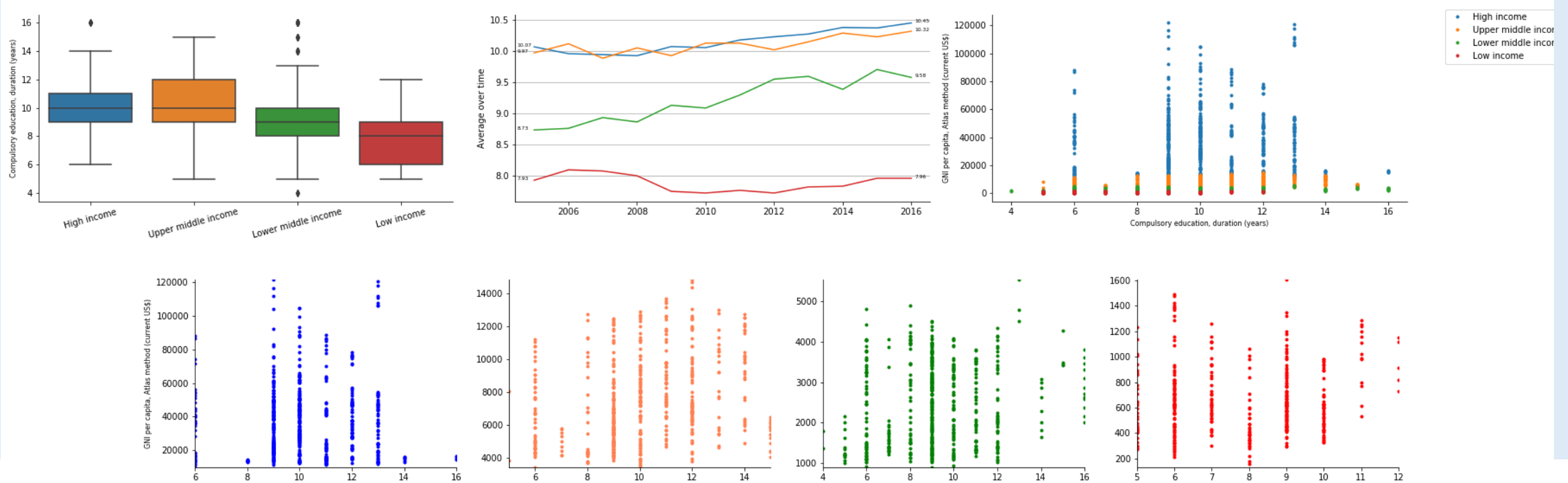
EDA – Economic Policy & Debt

- Annual GDP growth is larger among lower income countries than higher income countries.



EDA – Education & Gender Issues

- Average years for schooling between 8 and 10
- Proportion of seats held by women in government between 21.8% and 17.5% for all income groups



EDA – Access to Advanced Communication

- Income groups have seen a rise in internet usage over time
- Mobile cellular subscriptions increase with the income level



EDA – Environment, Resources & Population

- Access to clean fuels and technologies for cooking relates to access to electricity
- Total natural resources rents see a few countries, regardless of income group, produce most of the natural resources
- There is a significant difference in means of urban population to total population from income group to income group



EDA – Social Protection & Labor

- Labor force participation the same for high income countries and low income
- Low income work to survive, while high income work for comfort
- Employment in services is extremely high among high income countries



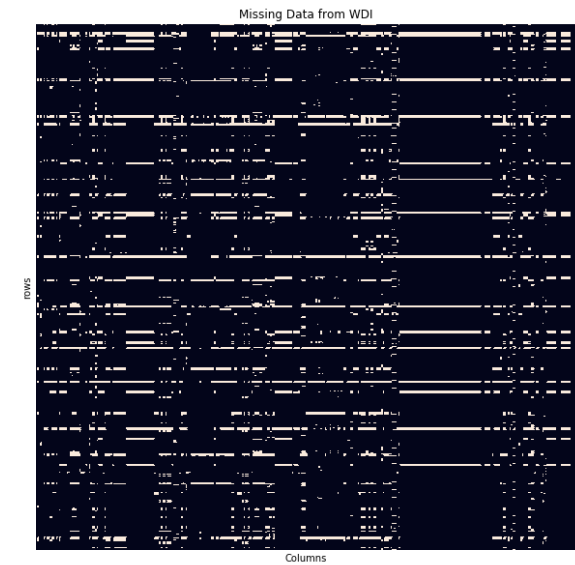
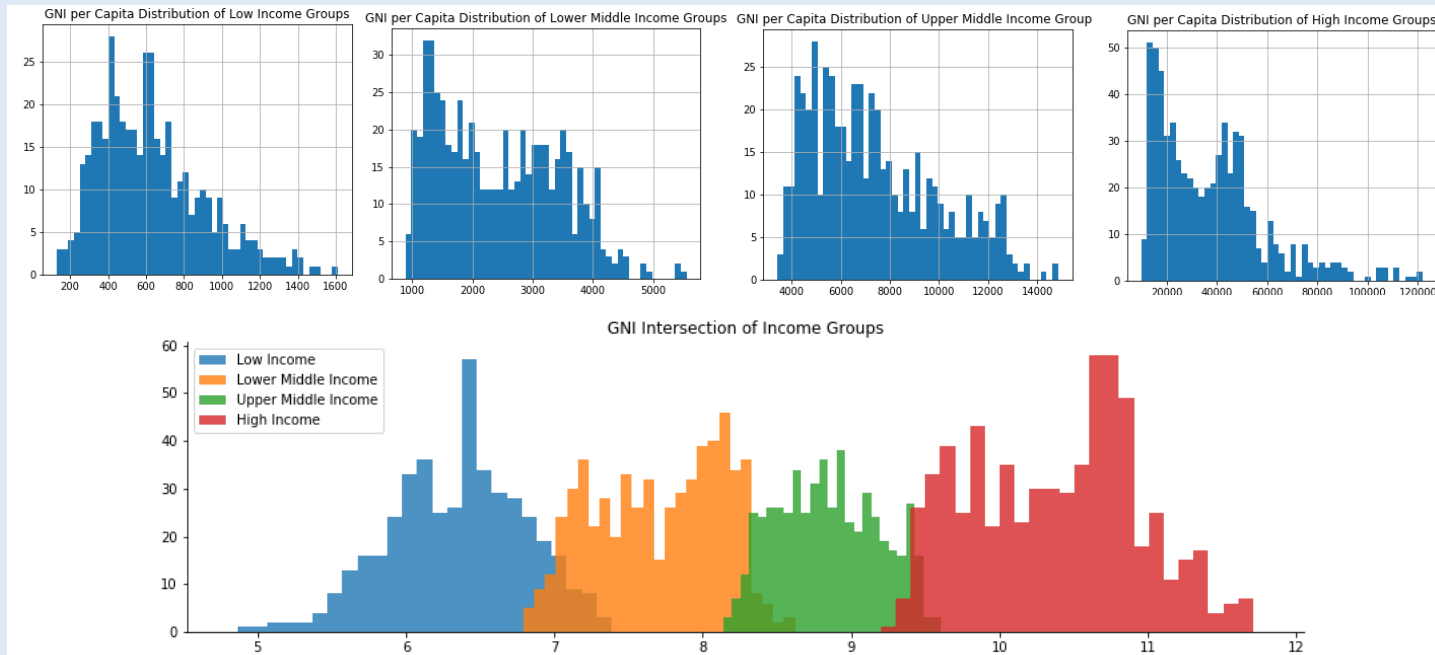
EDA – Health

- Access to basic drinking water is major in determining income groups
- Average life expectancy is much higher for high income countries
- Minimum life expectancy on the rise for all income groups



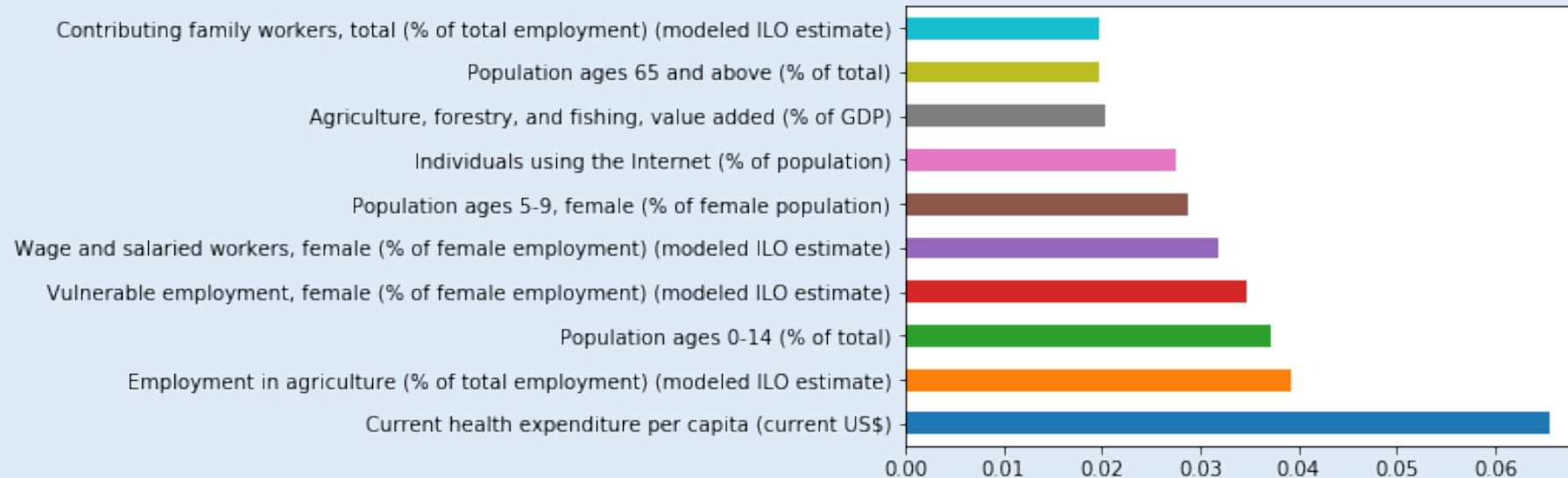
Model Preperation - Data

- NaN values
- GNI, GDP & World Development



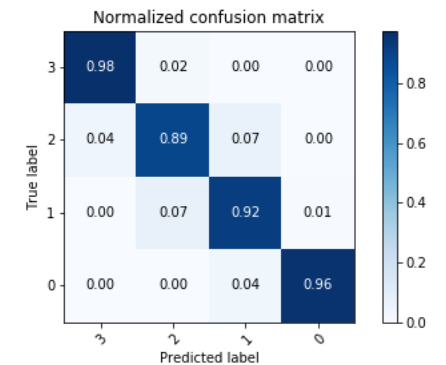
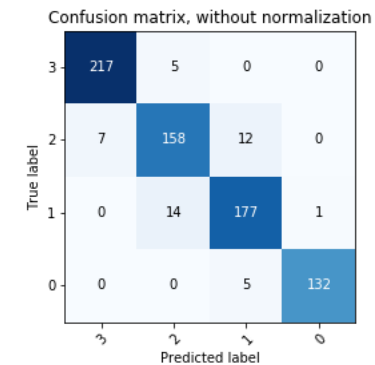
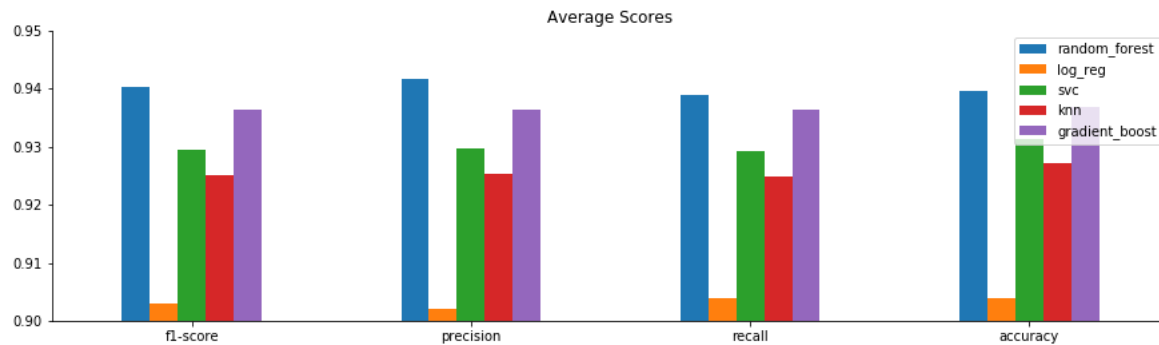
Model Preperation - Features

- Normalize data – subtract by mean and divide by standard deviation
- Feature Selection – univariate anova, tree-based, mix & match



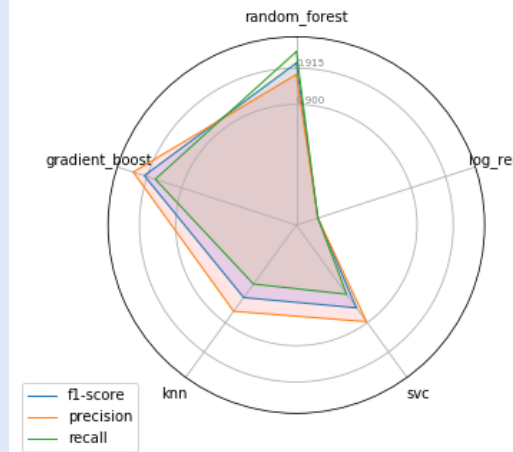
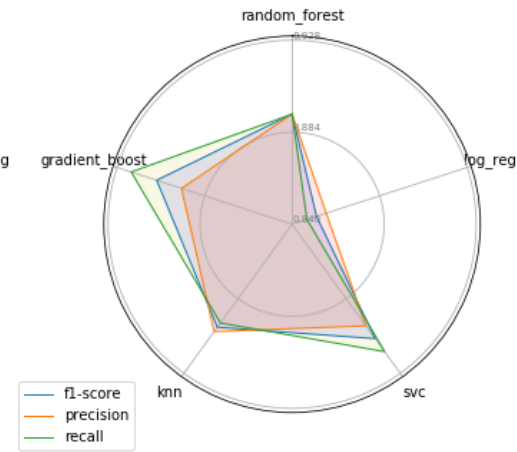
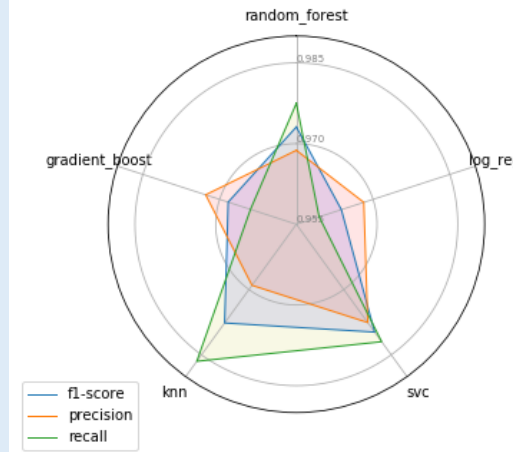
Model Testing & Tuning

- Pipeline to test each feature & hyperparameter combination on a 5 fold cv set
- Get best output, test on unseen data, repeat
- 5 models tested & tuned



Model Performance

- **Best Model – Random Forest**
 - Only 17 features needed
 - Highest accuracy 93.9%
 - Best at precision for low income groups 99%
- All models had over 95% accuracy & precision on high income groups
- All models had some issue with middle income groups



Limitation & Future Work

- A lot of missing data in WDI database
- Many indicators could not be used because of missing information
- May have affected more accurate predictions
- Try using one vs. one or one vs all approach (took very long on tuning models)



Questions?

