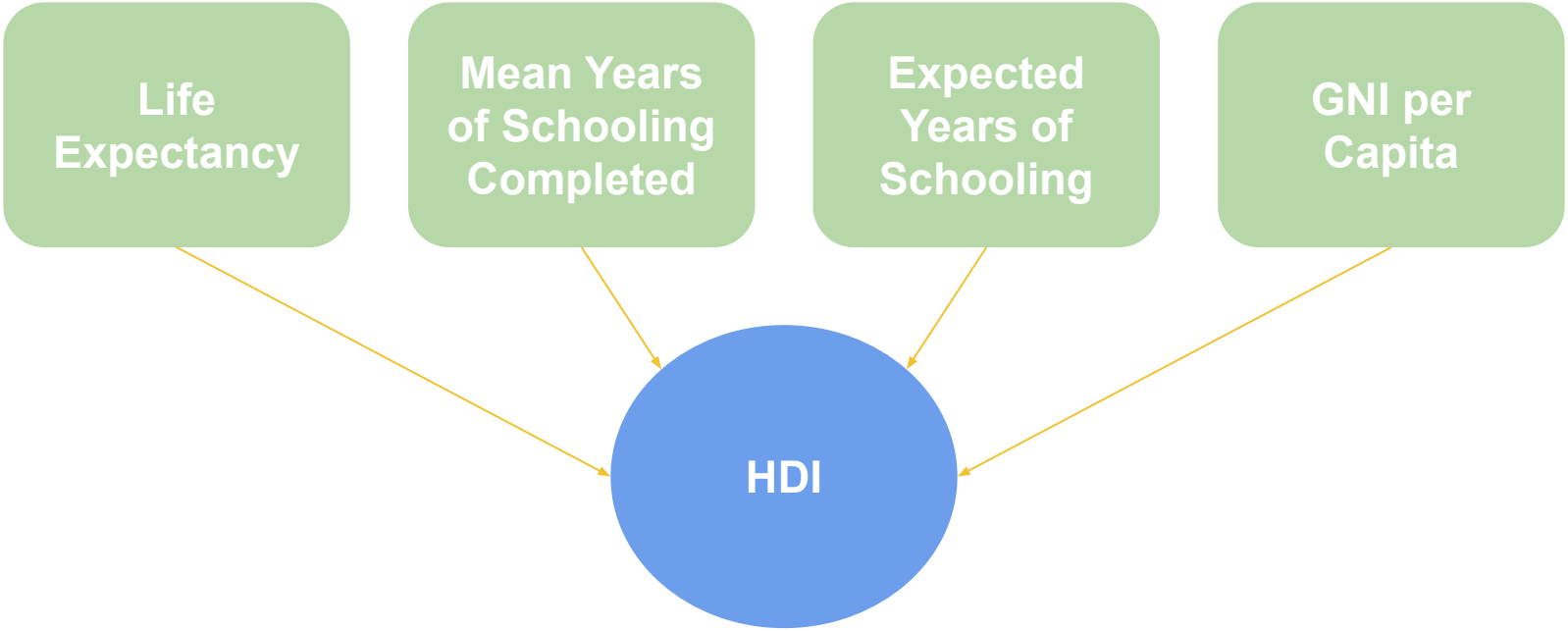
Predicting HDI Using Alternate Development Data

Project 5 | GA DSI 523 |
Group 4 - Giovanni Ceretti, Peter Murphy & Roman Tedeschi

An Introduction to the Human Development Index

HDI is a Static Composite of 4 Components



Our Mission

To create a model which uses ‘soft’ quality of life factors to predict human development

Our Data

Data Collection

- We collected ‘soft’ human development indicators from The World Bank & Gapminder
- Illustrative Examples of ‘soft’ Statistics:
  - Rural Population %
  - Under 5 Mortality Rate
  - % Population with Access to Electricity
  - Female Labor Participation Rate
  - Fertility Rate

Data Cleaning

- Many World Bank datasets include null values
- These null values were replaced by assigning each country a value in line with either:
  - Median
  - 25th Percentile Value
  - 75th Percentile Value
- Values were replaced using both our estimations and the results of their geographical and developmental peers

Number of Omissions by Metric

rural_pop_percent_17	1
food_production_index_17	2
ag_land_area_17	1
arable_land_percent_2017	2
net_migration_2017	6
hiv_prevalence_2017	55
mat_mortality_ratio_2017	7
under5_mortality_ratio_2017	2
tubercul_incidence_2017	1
elec_access_2017	1
ren_energy_percent_2017	0
ffuel_energy_percent_2017	187
co2_emissions_2017	2
pop_air_pollution_2017	4
foreign_dir_inv_2017	4
atm_access_2017	17
adol_fertility_rate_2017	6
fem_labor_part_rate_2017	11
male_labor_part_rate_2017	11
fertility_rate_2017	5
dpt_immuniz_rate_2017	2
undernourished_rate_2017	33
cell_subscriptions_per100	1
internet_per_mil_2017	0
military_exp_2017	37
women_seats_percent_2017	5
male_bus_start_2017	4
female_bus_start_2017	4
patent_apps_2017	72
sci_articles_2017	1
pop_density_2017	0
HDI	0

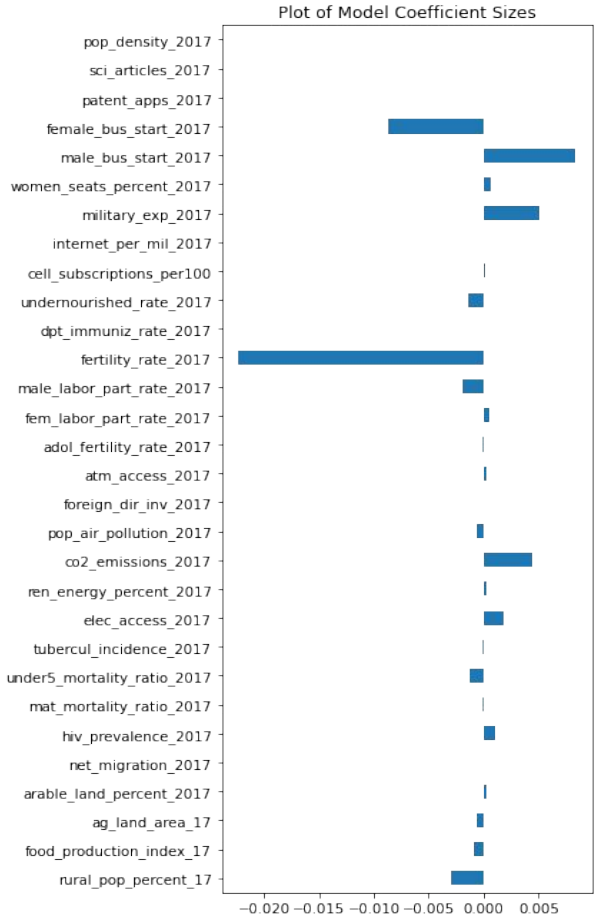
*A Note on Train/Test Split

- Because our model is based on an entire population (all countries on Earth) and not a sample,
- And because it is being used for interpretive insights rather than prediction a train/test split is probably not necessary
- However, we have left the split in to keep this project consistent with the standard methods we have learned in this course
- The models we chose to keep largely were not overfit, and as such, the difference in coefficients was small regardless of whether the split was kept or not

Linear Regression

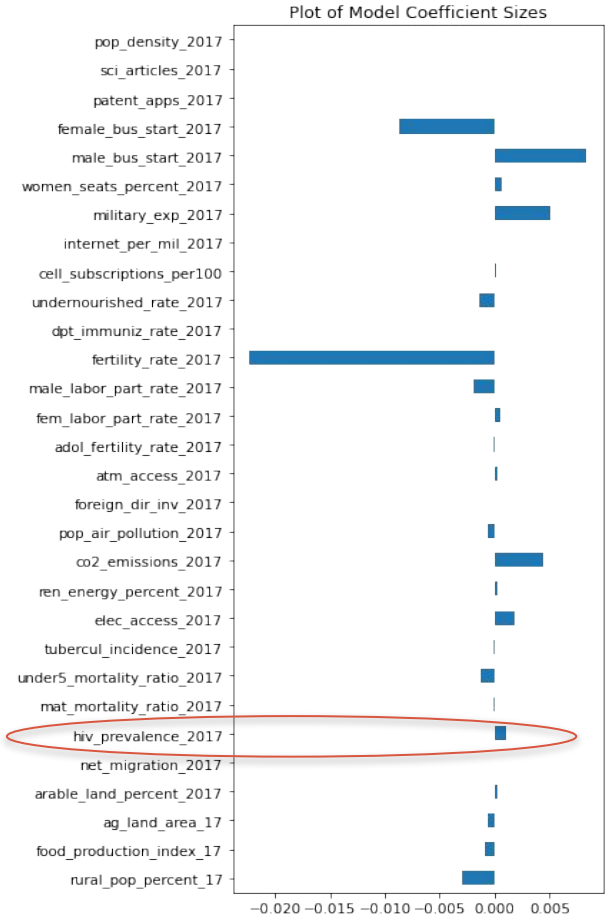
First, All Features

- To get a baseline model a linear regression was performed using all of our data categories as features
- This initial model had an r-squared of 0.89…
- But it also came with some strange results:



All Features?

- The coefficient produced by our model for the HIV prevalence had a value of 0.00105
- This means that for every % increase in HIV prevalence, HDI goes up by 0.00105
- It turns out that that data was simply missing for 55 countries and had been replaced by the median



Regularization

	Training R-Squared:	Test R-Squared:
Original OLS	0.94	0.89
Ridge	0.929	0.924
Lasso	0.915	0.901

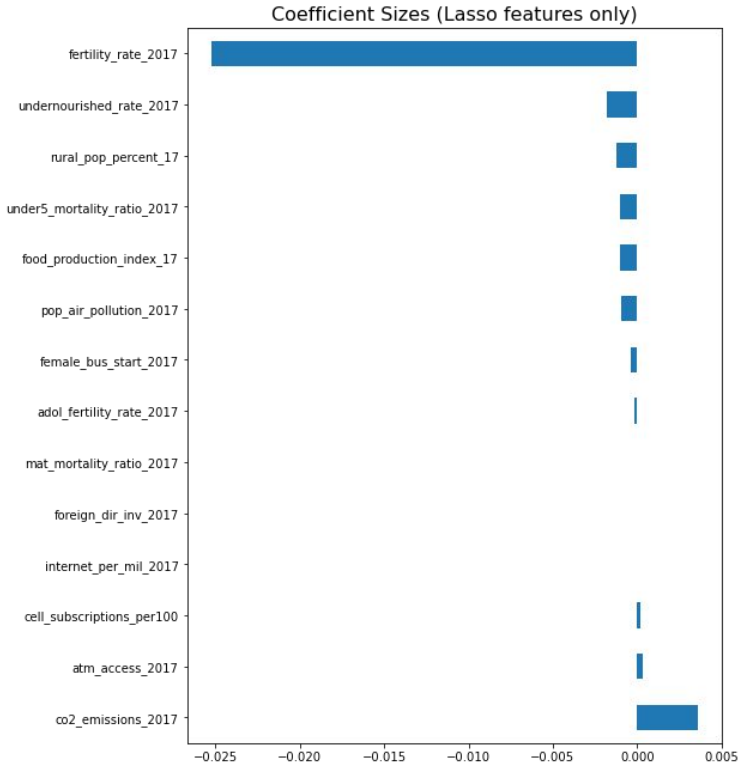
Non-Zero Lasso Coefficients

- Co2 Emissions
- ATM Access
- Cell Phone Subscriptions (per 100 people)
- Secure Internet (per million people)
- Foreign Direct Investment
- Adolescent Fertility Rate (age 15-19)
- Food Production Index
- Business Start Time (for females)
- Undernourished Rate
- Maternal Mortality Rate
- Population % Exposed to Air Pollution
- Rural Population Percent
- Fertility Rate
- Children Mortality Rate

Linear (with Lasso Features)

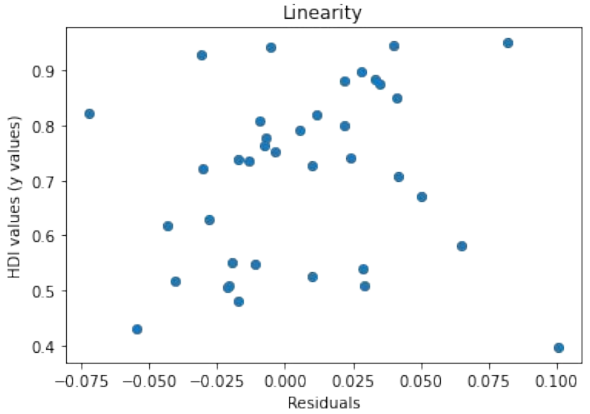
	Training R-Squared:	Test R-Squared:
Original OLS	0.941	0.892
Ridge	0.929	0.924
Lasso:	0.915	0.901
OLS (with Lasso Features	0.920	0.945

Null MSE (mean) = 0.023
Model MSE = 0.001

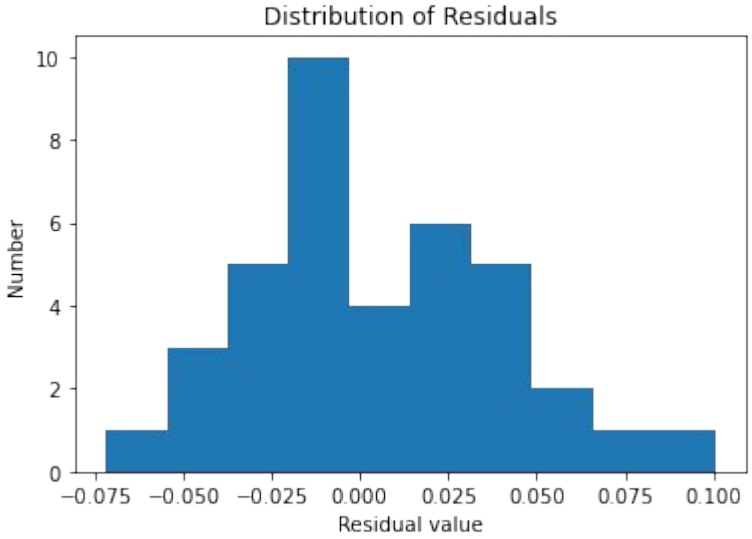


LINE Assumptions

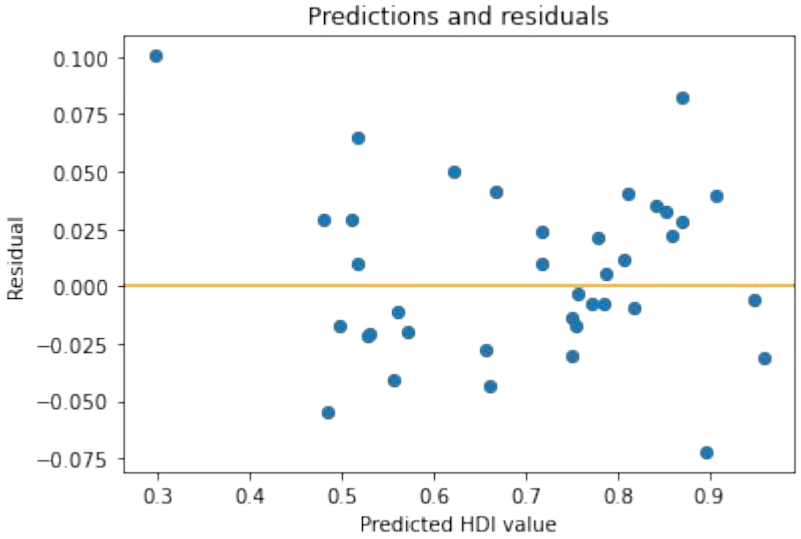
- Since our goal here is inference, the LINE assumptions are critical to accuracy
- **Linearity** - the relationship between our features and the HDI value are linear
- **Independence** - there is nothing to indicate that observations are not independent, and this can be assumed



- **Normality** - the distribution of our residuals is approximately normally distributed

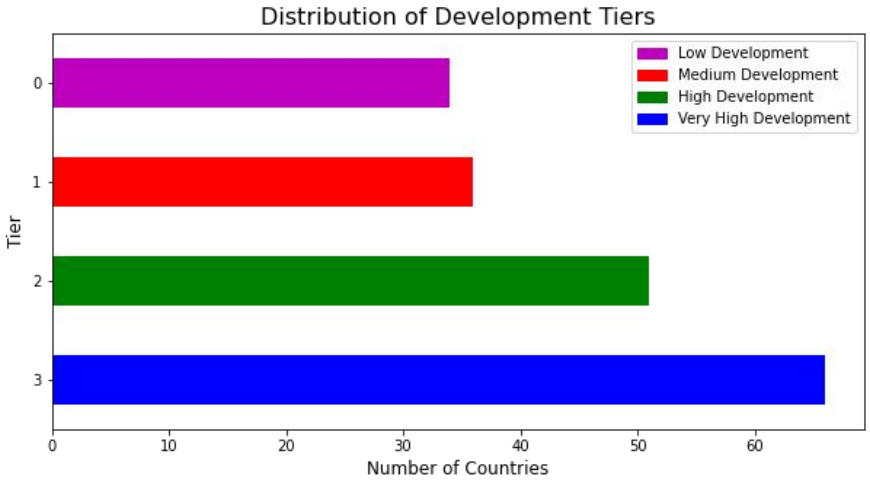


- **Equal Variance** - our residuals also display homoscedasticity (equal variances)

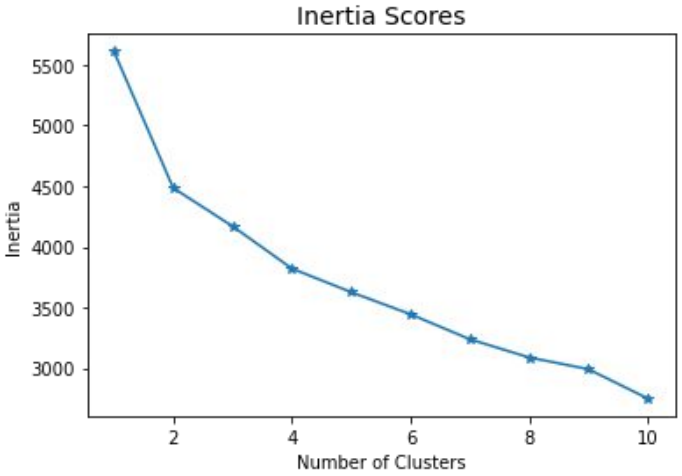


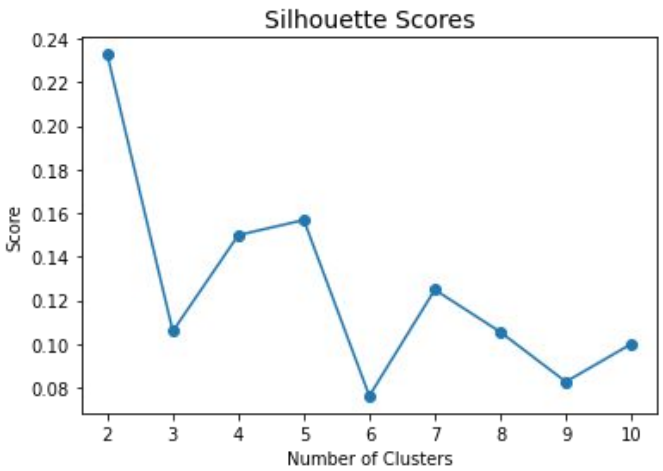
HDI Classifications

Category	Range
low human development	(less than 0.550)
medium human development	(0.550–0.699)
high human development	(0.700–0.799)
very high human development	(0.800 or greater)



Clustering





Kmeans silhouette score: 0.23

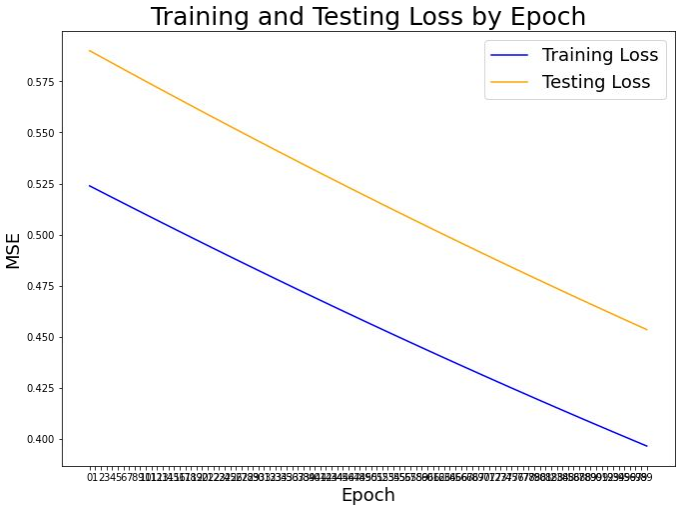
DBSCAN silhouette: -0.19

Random Forest & Neural Nets

Random Forest Regressor:
Training R-Squared: 0.9876
Test R-Squared: 0.9138

Random Forest overfit compared to the OLS (with select Lasso features).

Neural Net produced a R-Squared of -22.45%



Takeaways

Top 5 Coefficient Values	
Coefficient	**Value**
Fertility Rate (births per woman)	-0.0252
CO2 emissions in metric tons per capita	0.0036
Percent of the population that is undernourished	-0.0018
Percent of population living rurally	-0.0012
Mortality Rate for Children under 5 (per 1,000 live births)	-0.0010

1. It's possible to predict HDI, while excluding education and wealth-related features
2. Out of the "soft" features, fertility rate, CO2 emissions, undernourishment, rural population, and children mortality tend to have the greatest effect on HDI
3. Interpretation: For every 1 birth increase per woman (fertility rate), the HDI index is expected to be -0.025 lower. (Ex: Egypt and China)

Thank You!
Questions?