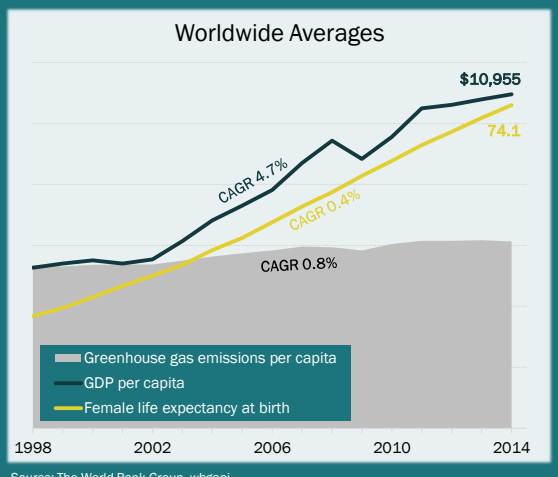
1

An amazing time to be alive!



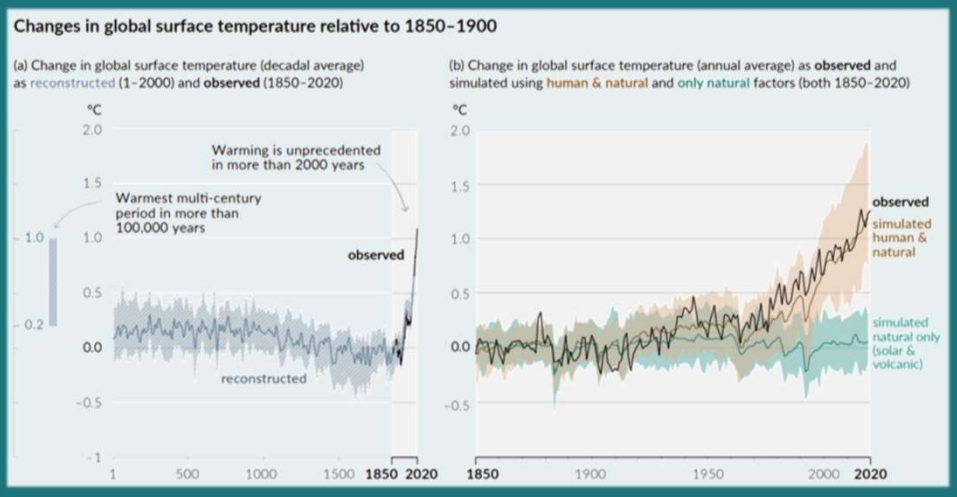
Source: The World Bank Group, wbgapi

"If you had to choose a moment in history to be born, and you did not know ahead of time who you would be – you didn't know whether you were going to be born into a wealthy family or a poor family, what country you'd be born in, whether you were going to be a man or a woman – if you had to choose blindly what moment you'd want to be born, you'd chose now."

Barack Obama, 2016

2

Prosperity comes at a cost…



Source: IPCC Summary for Policymakers, 6th Assessment Report, 2021

3

Can we prosper and reduce our impact on the climate?

- Data source
  - The World Bank Group, wbgapi
- Measure of prosperity: GDP per capita
  - GDP / Total population, sum of economic value per person
- Measure of Climate Impact: Greenhouse gas emissions per capita
  - Total greenhouse gas emissions / Total population
- Other Measures
  - Access to electricity (% of population), Electricity production from ____ sources (coal, oil, natural gas, nuclear, hydroelectric, renewables, all as % of total), Electric power transmission and distribution losses (% of output), Exports of goods and services (% of GDP), Imports of goods and services (% of GDP), ____ value added (Agriculture, Industry, Services, all as % of GDP), Female life expectancy at birth (years), Urban population (% of total), Fertility rate (births per woman)
- Data granularity
  - Year, country

4

Analysis Description

1) GPD per Capita Model

Type: Continuous

Question: What factors drive quality of life?

1. Unsupervised
   - PCA
   - k-means
     - PCA components, country
2. Supervised
   - Linear regression
   - Random Forest

2) GHGE per Capita Model

Type: Continuous

Question: What factors drive greenhouse gas emissions?

1. Unsupervised
   - PCA
   - k-means
     - PCA components, country
2. Supervised
   - Linear regression
   - Random Forest

3) Combination of 1 & 2

Type: Categorical

Question: What can we do to continue to improve human flourishing while reducing greenhouse gas emissions?

1. Unsupervised
   - k-means
     - GDP pe capita, GHGE per capita
2. Supervised
   - Random Forest
   - kNN

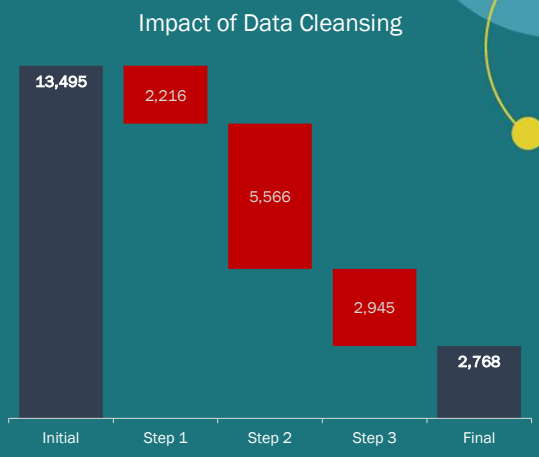
Disclaimer!!

Unfortunately, the data set does not contain sufficient granularity to truly diagnose the core issues or provide a roadmap of technically feasible corrective actions. The right answer is of course to grow economic activities that contribute to human prosperity while generating energy to fuel that economic growth with low carbon fuels. But how? Sorry, I can't answer that here…

5

Data Cleansing

- Raw data set
  - 29 columns, 13495 records
- Step 1
  - Drop 3 columns due to incompleteness and remove records without values for the targets
- Step 2
  - Exclude records from countries with significant incompleteness including values for 2019 and 2020 for all countries
- Step 3
  - Impute values in the access to electricity field, add 4 calculated columns, drop 10 additional columns, then remove records with remaining NULL values
- Final data set
  - 20 columns, 2768 records



6

1) What factors drive quality of life?

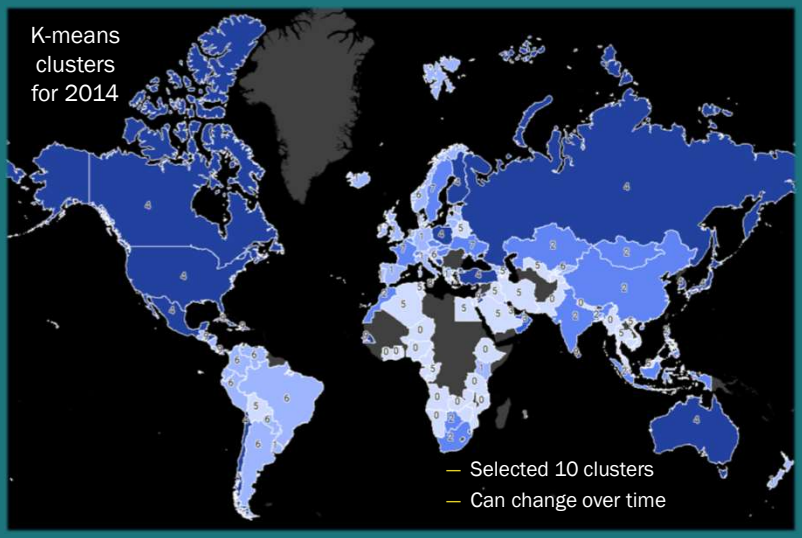
- PC1
  - \+ Electric transmission and distribution losses, agricultural economic activities, and increased fertility drive PC1 up
  - – Access to electricity, increased life expectancy, more urban
- PC2
  - \+ Electricity from natural gas, exports, GDP contribution from the industrial sector, more greenhouse gas emissions
  - – GDP contribution from the service sector
- PC3
  - \+ Electricity from oil, exports, imports
- PC4
  - \+ Electricity from coal
  - – Electricity from hydroelectric
- PC5
  - \+ Electricity from hydroelectric, exports, imports
  - – Electricity from oil
- PC6
  - \+ Electricity from natural gas, electricity from renewables
  - – Electricity from renewables

	PC1	PC2	PC3	PC4	PC5	PC6
EG.ELC.ACCS.ZS	-0.358509	0.023526	-0.092499	-0.130759	-0.176933	-0.043863
EG.ELC.COAL.ZS	-0.082952	-0.110751	-0.210929	0.711069	0.166622	0.160493
EG.ELC.HYRO.ZS	0.189721	-0.179774	-0.113694	-0.606026	0.259375	-0.055726
EG.ELC.LOSS.ZS	0.375125	-0.020363	0.103546	-0.066080	-0.055392	-0.051698
EG.ELC.NGAS.ZS	-0.088053	0.441073	-0.028588	-0.086830	0.131903	0.269162
EG.ELC.NUCL.ZS	-0.166671	-0.161599	-0.150397	0.085632	0.206640	-0.570111
EG.ELC.PETR.ZS	0.070894	0.005494	0.412703	0.126620	-0.698385	-0.195021
EG.ELC.RNWX.ZS	-0.089288	-0.228786	0.121251	-0.114570	-0.120399	0.686577
NE.EXP.GNFS.ZS	-0.195706	0.266799	0.491553	0.002519	0.313671	-0.033401
NE.IMP.GNFS.ZS	-0.139858	0.092515	0.612272	0.039917	0.318214	-0.070839
NV.AGR.TOTL.ZS	0.373028	-0.027450	0.058174	0.054206	0.130250	0.042473
NV.IND.TOTL.ZS	-0.037800	0.534596	-0.188748	-0.061248	-0.179308	-0.079739
NV.SRV.TOTL.ZS	-0.241121	-0.409280	0.125980	-0.021286	-0.005320	0.059687
SP.DYN.LE00.FE.IN	-0.362621	-0.142543	-0.005065	-0.133192	-0.069714	0.010133
SP.DYN.TFRT.IN	0.357653	0.150016	0.029927	-0.006214	-0.104796	0.022918
ghge_per_capita	-0.227389	0.322979	-0.195763	0.012578	0.043727	0.128470
urb_pop_pct	-0.348737	0.022459	-0.064873	-0.186916	-0.204216	-0.089603

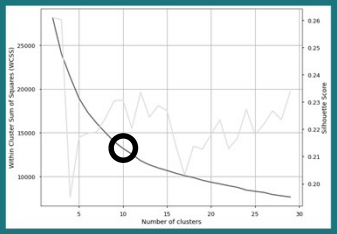
Factor loadings

7

1) How can we group economies?



Elbow Method Chart

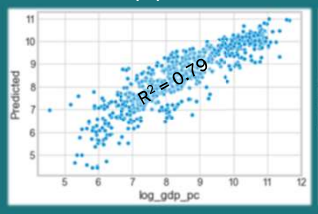


8

1) Regression on GDP per capita

1. Target transformed by natural log (np.log)
2. Residuals normally distributed
3. Adj R^2 of the regression = 0.802 on training
4. R^2 on test data = 0.792
5. R^2 on untransformed test data = 0.717

Transformed (ln)



No transformation

Good model but hard to interpret, good to see negative coefficient on PC1

OLS Regression Results

Dep. Variable:	np.log(gdp_per_capita)	R-squared:	0.803
Model:	OLS	Adj. R-squared:	0.802
Method:	Least Squares	F-statistic:	557.7
Date:	Fri, 12 Nov 2021	Prob (F-statistic):	0.00
Time:	08:52:22	Log-Likelihood:	-2063.5
No. Observations:	2063	AIC:	4159.
Df Residuals:	2047	BIC:	4249.
Df Model:	15		
Covariance Type:	nonrobust		

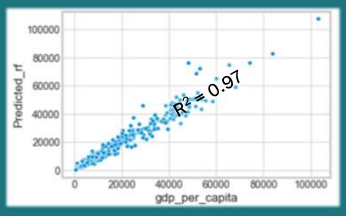
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	8.9295	0.070	126.710	0.000	8.791	9.068
C(Cluster_ID)[T.1]	-0.4332	0.122	-3.562	0.000	-0.672	-0.195
C(Cluster_ID)[T.2]	-0.5976	0.094	-6.332	0.000	-0.783	-0.413
C(Cluster_ID)[T.3]	-0.3981	0.177	-2.249	0.025	-0.745	-0.051
C(Cluster_ID)[T.4]	-0.1789	0.112	-1.603	0.109	-0.398	0.040
C(Cluster_ID)[T.5]	-0.8961	0.092	-9.690	0.000	-1.077	-0.715
C(Cluster_ID)[T.6]	-0.6324	0.086	-7.372	0.000	-0.801	-0.464
C(Cluster_ID)[T.7]	-0.3721	0.125	-2.982	0.003	-0.617	-0.127
C(Cluster_ID)[T.8]	-0.2214	0.180	-1.232	0.218	-0.574	0.131
C(Cluster_ID)[T.9]	-0.8898	0.095	-9.411	0.000	-1.075	-0.704
PC1	-0.5945	0.015	-38.973	0.000	-0.624	-0.565
PC2	-0.0184	0.019	-0.987	0.324	-0.055	0.018
PC3	-0.0945	0.019	-5.092	0.000	-0.131	-0.058
PC4	-0.1756	0.020	-8.701	0.000	-0.215	-0.136
PC5	-0.0963	0.019	-5.179	0.000	-0.133	-0.060
PC6	0.0749	0.025	2.991	0.003	0.026	0.124

Omnibus:	38.189	Durbin-Watson:	2.043
Prob(Omnibus):	0.000	Jarque-Bera (JB):	73.736
Skew:	-0.062	Prob(JB):	9.74e-17
Kurtosis:	3.918	Cond. No.	45.1

9

1) Random Forest Modeling

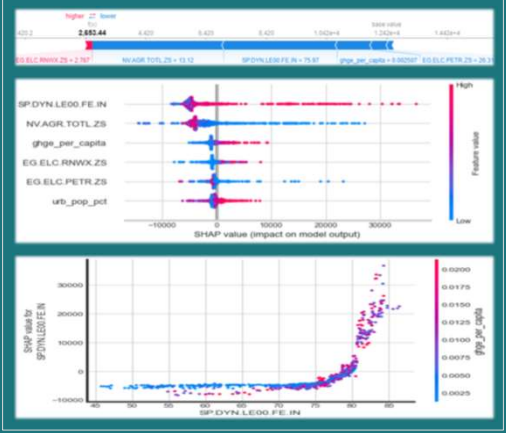
1. Ran 2 models, first to determine feature importance, 2nd only includes "important" features
2. R^2 on test data = 0.965



3. Features by order of importance
   - Female life expectancy, contribution of agriculture to GDP, electricity generated from oil, electricity generated from renewables, greenhouse gas emissions per capita, percent of population in urban areas

Shapley Additive Explanations

- The average marginal contribution of an instance of a feature among all possible coalitions



Source: https://towardsdatascience.com/shap-shapley-additive-explanations-5a2a271ed9c3

10

2) What factors drive emissions?

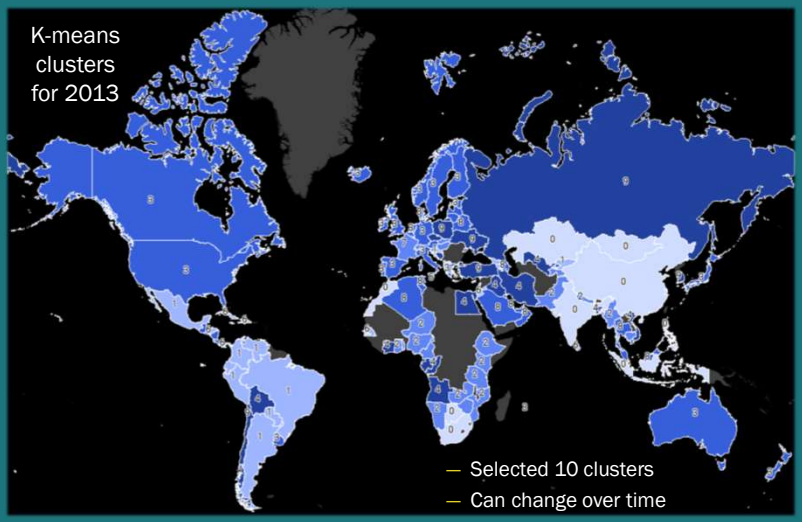
- PC1
  - \+ Access to electricity, service based economic activities, increased life expectancy, higher GDP per capita, more urban
  - – Electric transmission and distribution losses, agricultural economic activities, and increased fertility
- PC2
  - \+ Electricity from natural gas, exports, imports, GDP contribution from the industrial sector
  - – Service based economic activities
- PC3
  - \+ Electricity from oil, exports, imports
  - – GDP contribution from the industrial sector
- PC4
  - \+ Electricity from hydroelectric
  - – Electricity from coal and oil
- PC5
  - \+ Electricity from coal, exports, imports
  - – Access to electricity, electricity from oil
- PC6
  - \+ Electricity from natural gas and renewables
  - – Electricity from nuclear

	PC1	PC2	PC3	PC4	PC5	PC6
EG.ELC.ACCS.ZS	0.339965	0.065525	-0.161647	0.019417	-0.257646	-0.065084
EG.ELC.COAL.ZS	0.075724	-0.155820	-0.175683	-0.640594	0.346839	0.129881
EG.ELC.HYRO.ZS	-0.170438	-0.241456	-0.023756	0.616232	0.035796	-0.218105
EG.ELC.LOSS.ZS	-0.264792	-0.027999	0.132976	0.057745	-0.072876	-0.053616
EG.ELC.NGAS.ZS	0.066276	0.448780	-0.170237	0.154426	0.173468	0.337609
EG.ELC.NUCL.ZS	0.172988	-0.166984	-0.132988	-0.056189	0.196285	-0.493584
EG.ELC.PETR.ZS	-0.071492	0.108264	0.391475	-0.292817	-0.642392	-0.080036
EG.ELC.RNWX.ZS	0.107874	-0.204045	0.199162	0.127024	-0.084727	0.685909
NE.EXP.GNFS.ZS	0.189238	0.396140	0.403230	0.058378	0.279357	-0.101457
NE.IMP.GNFS.ZS	0.142806	0.251957	0.583111	-0.010895	0.251271	-0.175623
NV.AGR.TOTL.ZS	-0.360023	-0.058168	0.101244	-0.006304	0.144405	0.027497
NV.IND.TOTL.ZS	-0.000158	0.502099	-0.356540	0.032257	-0.155469	-0.021810
NV.SRV.TOTL.ZS	0.264194	-0.359284	0.232075	0.008539	-0.024647	0.047934
SP.DYN.LE00.FE.IN	0.381606	-0.089601	-0.012715	0.063592	-0.144818	-0.023748
SP.DYN.TFRT.IN	-0.352320	0.100199	0.039193	0.053756	-0.029372	0.099570
gdp_per_capita	0.297428	-0.048189	0.041781	0.227203	0.216614	0.174260
urb_pop_pct	0.338564	0.047903	-0.103959	0.118618	-0.248811	-0.065817

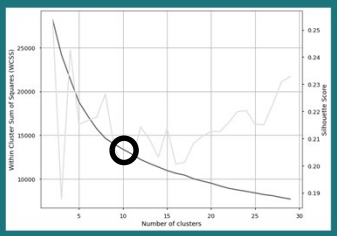
Factor loadings

11

2) How can we group economies?



Elbow Method Chart

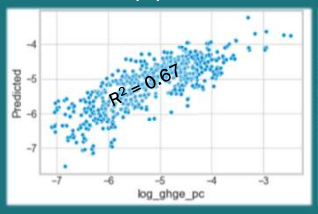


12

2) Regression on GHGE per capita

1. Target transformed by natural log (np.log)
2. Residuals normally distributed
3. Adj R^2 of the regression = 0.675 on training
4. R^2 on test data = 0.672
5. R^2 on untransformed test data = 0.519

Transformed (ln)



No transformation

Good model but hard to interpret. Appears that variables that lead to high GDP per capita, also lead to higher GHGE per capita

OLS Regression Results

Dep. Variable:	np.log(ghge_per_capita)	R-squared:	0.677
Model:	OLS	Adj. R-squared:	0.675
Method:	Least Squares	F-statistic:	286.2
Date:	Fri, 12 Nov 2021	Prob (F-statistic):	0.00
Time:	14:02:13	Log-Likelihood:	-1484.3
No. Observations:	2063	AIC:	3001.
Df Residuals:	2047	BIC:	3091.
Df Model:	15		
Covariance Type:	nonrobust		

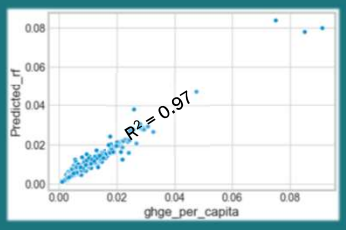
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-5.0632	0.055	-92.001	0.000	-5.171	-4.955
C(Cluster_ID)[T.1]	-0.1119	0.072	-1.549	0.122	-0.254	0.030
C(Cluster_ID)[T.2]	0.0432	0.077	0.560	0.576	-0.108	0.194
C(Cluster_ID)[T.3]	0.0836	0.084	0.998	0.318	-0.081	0.248
C(Cluster_ID)[T.4]	-0.3439	0.073	-4.689	0.000	-0.488	-0.200
C(Cluster_ID)[T.5]	0.0051	0.142	0.036	0.971	-0.274	0.284
C(Cluster_ID)[T.6]	-0.3194	0.078	-4.088	0.000	-0.473	-0.166
C(Cluster_ID)[T.7]	-0.2293	0.083	-2.777	0.006	-0.391	-0.067
C(Cluster_ID)[T.8]	-0.2546	0.086	-2.952	0.003	-0.424	-0.085
C(Cluster_ID)[T.9]	0.1552	0.061	2.564	0.010	0.036	0.274
PC1	0.2651	0.012	22.102	0.000	0.242	0.289
PC2	0.1749	0.014	12.519	0.000	0.148	0.202
PC3	-0.1954	0.014	-13.883	0.000	-0.223	-0.168
PC4	0.0152	0.015	1.041	0.298	-0.013	0.044
PC5	-0.0352	0.015	-2.421	0.016	-0.064	-0.007
PC6	-0.0361	0.016	-2.315	0.021	-0.067	-0.006

Omnibus:	11.785	Durbin-Watson:	2.023
Prob(Omnibus):	0.003	Jarque-Bera (JB):	11.949
Skew:	0.185	Prob(JB):	0.00254
Kurtosis:	2.953	Cond. No.	43.5

13

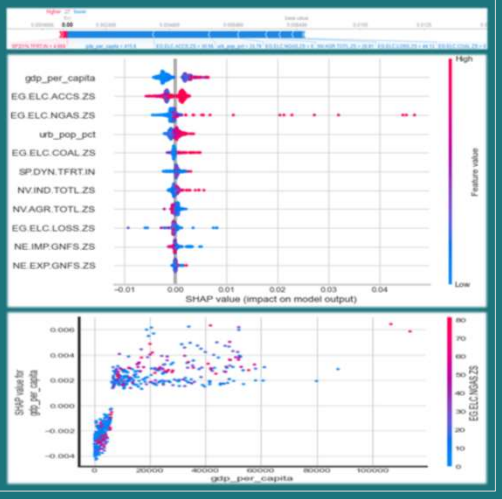
2) Random Forest Modeling

1. Ran 2 models, first to determine feature importance, 2nd only includes "important" features
2. R^2 on test data = 0.966



3. Features by order of importance
   - Electricity generated from natural gas, GDP per capita, electric transmission and distribution losses, access to electricity, percent of population in urban areas, GDP contribution from industrial sector, fertility rate, imports, electricity from coal, agricultural economic activities, exports

Shapley Additive Explanations



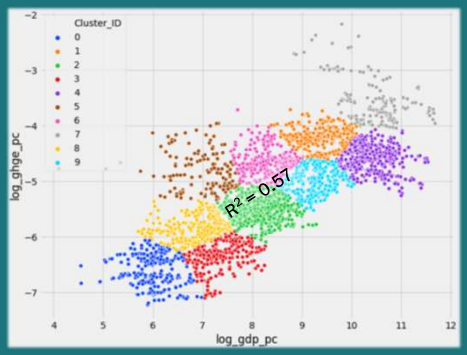
Source: https://towardsdatascience.com/shap-shapley-additive-explanations-5a2a271ed9c3

14

3) What can we do to prosper and reduce greenhouse gas emissions?

1. Use k-means to create a categorical target variable that represents groupings of GDP per capita and GHGE per capita
2. Create 10 clusters
3. Run supervised models to predict the "cluster ID"
4. Review feature importance and contrast feature responses from cluster to cluster

GDP per capita is correlated with greenhouse gas emissions per capita



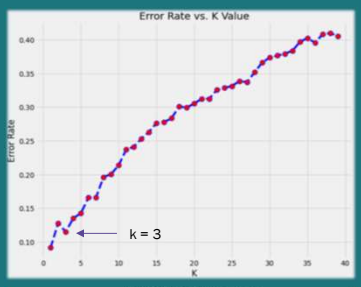
15

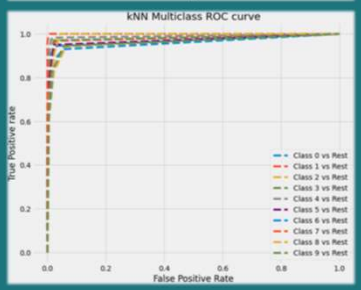
3) kNN Modeling

1. Ran 2 models, first to determine appropriate k-value, 2nd with k = 3
   - k = 3 lowest error rate on test data
2. Accuracy = 89%, AUC = 0.994
3. All independent variables considered

Confusion Matrix

	0	1	2	3	4	5	6	7	8	9
0	50	0	0	4	0	0	0	0	1	0
1	0	53	0	0	5	0	3	0	0	1
2	0	0	88	0	0	0	0	0	12	4
3	6	0	2	56	0	0	0	0	2	0
4	0	0	0	0	113	0	0	1	0	2
5	0	0	6	0	0	31	3	0	0	0
6	0	1	3	0	0	2	50	0	0	1
7	0	1	0	0	0	0	0	44	0	0
8	1	0	2	2	0	1	0	0	65	0
9	0	1	3	0	2	0	7	0	0	59





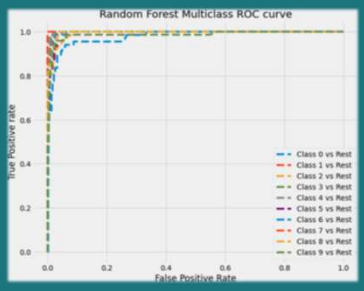
16

3) Random Forest Modeling

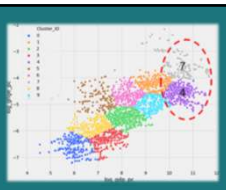
1. Ran 2 models, first to determine feature importance, 2nd only includes "important" features
2. Accuracy = 89%, AUC = 0.976
3. Features by order of importance
   - Agricultural economic activities, female life expectancy, percent of population in urban areas , access to electricity , electric transmission and distribution losses , GDP contribution from service sector, fertility rate, electricity generated from hydroelectric, electricity generated from renewables, GDP contribution from industrial sector, electricity from oil, electricity from coal, electricity from natural gas, exports, imports, electricity from nuclear

Confusion Matrix

	0	1	2	3	4	5	6	7	8	9
0	67	0	0	1	0	0	0	0	1	0
1	0	50	0	0	2	0	3	0	0	0
2	0	0	96	2	0	2	2	0	3	5
3	6	0	3	61	0	0	0	0	2	0
4	0	2	0	0	116	0	0	0	0	0
5	0	0	3	0	0	25	2	0	0	0
6	0	2	4	0	0	5	51	0	0	5
7	0	1	0	0	0	0	0	38	0	0
8	5	0	3	1	0	0	0	0	46	0
9	0	1	2	0	3	0	2	0	0	65

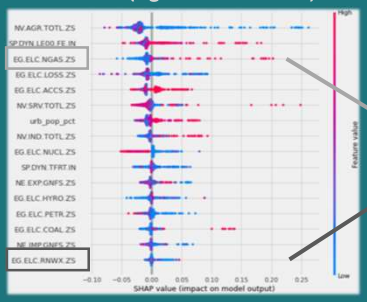


17

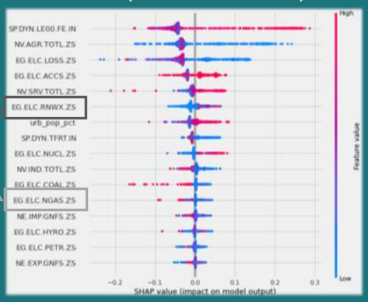


3) Cluster Comparison

Cluster ID = 7 (higher GHG emissions)



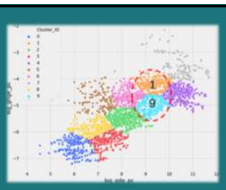
Cluster ID = 4 (lower GHG emissions)



Features are sorted by order of importance within category
SHAP values represent change in probability of prediction

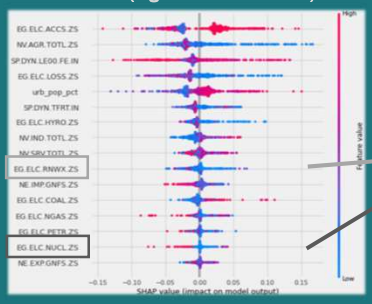
Clusters 4 and 7 have similar GDP per capita, but different GHG emissions per capita

18

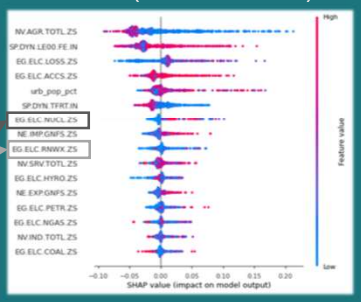


3) Cluster Comparison

Cluster ID = 1 (higher GHG emissions)



Cluster ID = 9 (lower GHG emissions)



Features are sorted by order of importance within category
SHAP values represent change in probability of prediction

Clusters 1 and 9 have similar GDP per capita, but different GHG emissions per capita

19

Implications and Conclusions

1. Increased prosperity results in higher greenhouse gas emissions
2. Marginal improvements are possible with reduced dependence on fossil fuels and increased reliance on low carbon alternatives such as renewables and nuclear.
3. Model 1: More renewables = higher GDP per capita
   - Synergistic opportunity?
   - Correlation or coincidence?
4. A net zero carbon future will require active sequestration and technological advancement

20