LIFE EXPECTANCY PREDICTION

By Trent Casillas

OUTLINE

- **Generalized Capstone Timeline:**
- Motivation
- Dataset Introduction
- Exploratory Data Analysis
- Feature Correlation
  - Top Positive and Negative correlated features to Life Expectancy and correlated pairs overall
- Regression Methods and Evaluation
  - Models include: Ridge Regression, Gradient Boosting , Robust (Thiel-sen) Regression
- Clustering and Evaluation
  - K-means, Mean-Shift
- Mixed Effect Model Evaluation
  - Check ICC
  - Run Random Intercepts, Random Slopes, Random Intercepts and Slopes Models
  - Compare best model to Regression Methods
- Conclusions
  - Takeaways and further studies

MOTIVATION AND PURPOSE

- Predict life expectancy by looking at the positive and negatively correlated factors to improve life quality
- Try to discern any differences between countries groupings
- Serves as an example for countries to assess to improve life expectancy for their citizens.

DATA SET

DATASET INFORMATION AND CLEANING

- My data is from The World Health Organization and United Nations website, Life Expectancy expresses the results of 193 countries for life expectancy with 2938 rows for 20 different features columns spanning from 2000-2015 with 2563 missing values.
- The features vary from whether country demographics such as Population, GDP, Total Expenditure spend on health to population statistics such immunizations and mortality rates along with BMI, Alcohol Consumption, and years of schooling.
- The data cleaning included removed 1100 removed between GDP(448) and Population,(652),
- 818 values replaced with missing respective country means.
- 645 replaced by the mean related to the status of the country.

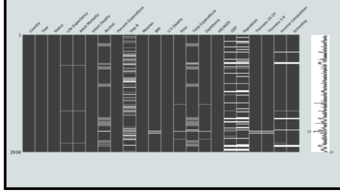
	Country	Year	Status	Life Expectancy	Adult Mortality	Infant Deaths	Alcohol	Percent Expenditure	Hep B	Measles	BMI	U-5 Deaths	Polio	Total Expenditure	Diphtheria	HIV/AIDS	GDP	Population	Thinness 10-19	Thinness 5-9	Income Composition	Schooling
0	Afghanistan	2015	Developing	65	263	62	0.01	71.279624	65	1154	19.1	83	6	8.16	65	0.1	584.25921	33736494	17.2	17.3	0.479	10.1
1	Afghanistan	2014	Developing	59.9	271	64	0.01	73.523582	62	492	18.6	86	58	8.18	62	0.1	612.696514	327582	17.5	17.5	0.476	10
2	Afghanistan	2013	Developing	59.9	268	66	0.01	73.219243	64	430	18.1	89	62	8.13	64	0.1	631.744976	31731688	17.7	17.7	0.47	9.9
3	Afghanistan	2012	Developing	59.5	272	69	0.01	78.184215	67	2787	17.6	93	67	8.52	67	0.1	669.959	3696958	17.9	18	0.463	9.8
4	Afghanistan	2011	Developing	59.2	275	71	0.01	7.097109	68	3013	17.2	97	68	7.87	68	0.1	63.537231	2978599	18.2	18.2	0.454	9.5
5	Afghanistan	2010	Developing	58.8	279	74	0.01	79.679367	66	1989	16.7	102	66	9.2	66	0.1	553.32894	2883167	18.4	18.4	0.448	9.2
6	Afghanistan	2009	Developing	58.6	281	77	0.01	56.762217	63	2861	16.2	106	63	9.42	63	0.1	445.893298	284331	18.6	18.7	0.434	8.9
7	Afghanistan	2008	Developing	58.1	287	80	0.03	25.873925	64	1599	15.7	110	64	8.33	64	0.1	373.361116	2729431	18.8	18.9	0.433	8.7
8	Afghanistan	2007	Developing	57.5	295	82	0.02	10.910156	63	1141	15.2	113	63	6.73	63	0.1	369.835796	26616792	19	19.1	0.415	8.4
9	Afghanistan	2006	Developing	57.3	295	84	0.03	17.171518	64	1990	14.7	116	58	7.43	58	0.1	272.56377	2589345	19.2	19.3	0.405	8.1

FEATURE LIST

- **Country- Country**
  Year- Year
  Status- Developed or Developing status
  Life Expectancy- Age(years)
  Adult Mortality- Adult Mortality Rates of both sexes(probability of dying between 15&60 years per 1000 population)
  Infant Deaths- Number of Infant Deaths per 1000 population
  Alcohol- Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol)
  Percent Expenditure- Expenditure on health as a percentage of Gross Domestic Product per capita(%)
  Hep B- Hepatitis B (HepB) immunization coverage among 1-year-olds(%)
  Measles- number of reported measles cases per 1000 population
  BMI- Average Body Mass Index of entire population
  U-5 Deaths- Number of under-five deaths per 1000 population
  Polio- Polio(Pol3) immunization coverage among 1-year-olds(%)
  Total Expenditure- General government expenditure on health as a percentage of total government expenditure(%)
  Diphtheria- Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds(%)
  HIV/AIDS- Deaths per 1000 live births HIV/AIDS(0-4 years)
  GDP- Gross Domestic Product per capita(in USD)
  Population- Population
  Thinness 10-19- Prevalence of thinness among children and adolescents for Age 10 to 19
  Thinness 5-9(%)- Prevalence of thinness among children for Age 5 to 9(%)
  Income Composition-Human Development Index in terms of income composition of resources(0-1)
  Schooling- Number of years of Schooling

DATASET

Feature	Year	Life Expectancy	Adult Mortality	Infant Deaths	Alcohol	Percent Expenditure	Hep B	Measles	BMI	U-5 Deaths	Polio	Total Expenditure	Diphtheria	HIV/AIDS	GDP	Population	Thinness 10-19	Thinness 5-9	Income Composition	Schooling
count	2938	2928	2928	2938	2744	2938	2385	2938	2904	2938	2919	2712	2919	2938	2490	2.29E+03	2904	2904	2771	2775
mean	2007.51872	69.224932	164.796448	30.303948	4.602861	738.251295	80.940461	2419.59224	38.321247	42.035739	82.550188	5.93819	82.324084	1.742103	7483.15847	1.28E+07	4.839704	4.870317	0.627551	11.992793
std	4.613841	9.523867	124.292079	117.926501	4.052413	1987.914858	25.070016	11467.2725	20.044034	160.445548	23.428046	2.49832	23.716912	5.077785	14270.1693	6.10E+07	4.420195	4.508882	0.210904	3.35892
min	2000	36.3	1	0	0.01	0	1	0	1	0	3	0.37	2	0.1	1.68135	3.40E+01	0.1	0.1	0	0
25%	2004	63.1	74	0	0.8775	4.685343	77	0	19.3	0	78	4.26	78	0.1	463.935626	1.96E+05	1.6	1.5	0.493	10.1
50%	2008	72.1	144	3	3.755	64.912906	92	17	43.5	4	93	5.755	93	0.1	1766.9476	1.39E+06	3.3	3.3	0.677	12.3
75%	2012	75.7	228	22	7.7025	441.534144	97	360.25	56.2	28	97	7.4925	97	0.8	5910.80634	7.42E+06	7.2	7.2	0.779	14.3
max	2015	89	723	1800	17.87	19479.91161	99	212183	87.3	2500	99	17.6	99	50.6	119172.742	1.29E+09	27.7	28.6	0.948	20.7



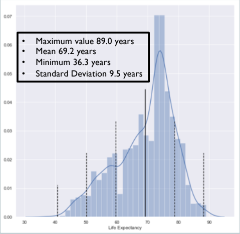
- Population and GDP will be dropped due to discrepancies with the data.
- Fill in remaining missing with country mean where applicable
- In other cases, fill in data with associated country status mean.

3 SIGMA OUTLIERS

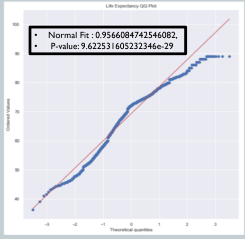
- Adult Mortality 99 Percentile 675.2 Max 723.0
- Infant Deaths 99th Percentile 1307.0 Max 1800
- Alcohol 99th Percentile 15.07 Max 17.87
- Percent Expenditure 99th Percentile 15357.4 Max 19479.9
- Percent Expenditure Measles 99th Percentile 111472.7 Max 212183
- BMI 99th Percentile 76.7 Max 87.3
- U-5 Deaths 99th Percentile 1807.0 Max 2500
- Total Expenditure 99th Percentile 16.2 Max 17.6
- HIV/AIDS 99th Percentile 42.1 Max 50.6
- Thinness 10-19 99th Percentile 27.0 Max 27.7
- Thinness 5-9 99th Percentile 27.9 Max 28.6
- Schooling 99th Percentile 20.3 Max 20.7

EXPLORATORY DATA ANALYSIS

LIFE EXPECTANCY

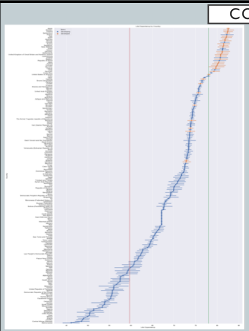


- Maximum value 89.0 years
- Mean 69.2 years
- Minimum 36.3 years
- Standard Deviation 9.5 years



- Normal Fit : 0.95660847425460 82,
- P-value: 9.6225316052323 46e-29

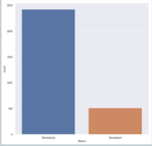
COUNTRY CATEGORIES

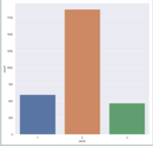


Country Status:
Developing: 82.6% Developed: 17.4%

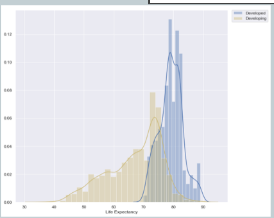
World Feature:
3rd World:16.0% 2nd World:63.8% 1st World: 20.3%

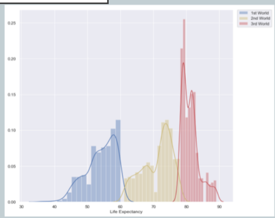
3rd World is below 60 and 1st world is above 78





STATUS VS WORLD FEATURE

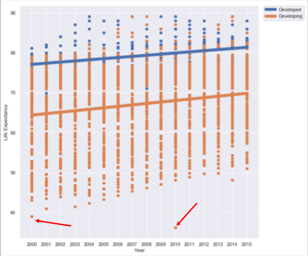




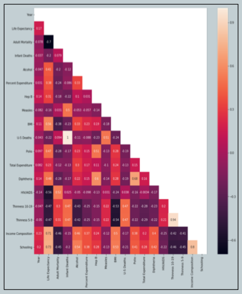
Status	count	mean	std	min	25%	50%	75%	max
Developed	512	79.2	3.9	69.9	76.8	79.25	81.7	89
Developing	2426	67.1	9.0	36.3	61.1	69	74	89

world	count	mean	std	min	25%	50%	75%	max
1	594	54.0	4.5	36.3	51.425	54.8	57.9	60
2	1877	71.0	4.6	61	67.5	72.5	74.6	78
3	467	81.2	2.6	78.1	79.1	81	82.3	89

LIFE EXPECTANCY VERSUS YEAR



LIFE EXPECTANCY CORRELATION AND HEAT MAP



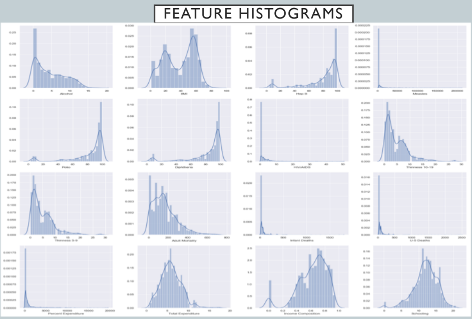
Schooling	0.732084
Income Composition	0.708651
BMI	0.560105
Diphtheria	0.483574
Polio	0.470396
Alcohol	0.407103
Percent Expenditure	0.38199
Hep B	0.314744
Total Expenditure	0.226319
Year	0.168709
Measles	-0.157401
Infant Deaths	-0.196324
U-5 Deaths	-0.222286
Thinness 5-9	-0.468218
Thinness 10-19	-0.473757
HIV/AIDS	-0.556165
Adult Mortality	-0.696386

Top 10 Positive and Negative Correlated Pairs

Pair		Value
Infant Deaths	U-5 Deaths	0.996629
Thinness 10-19	Thinness 5-9	0.939122
Income Composition	Schooling	0.801269
Life Expectancy	Schooling	0.732084
	Income Composition	0.708651
Polio	Diphtheria	0.676641
Hep B	Diphtheria	0.600456
Life Expectancy	BMI	0.560105
Alcohol	Schooling	0.541402
BMI	Schooling	0.529271

Pair		Value
Adult Mortality	Schooling	-0.451296
Thinness 5-9	Schooling	-0.452244
Adult Mortality	Income Composition	-0.455119
Thinness 10-19	Schooling	-0.462581
Life Expectancy	Thinness 5-9	-0.468218
	Thinness 10-19	-0.473757
BMI	Thinness 10-19	-0.532155
	Thinness 5-9	-0.539039
Life Expectancy	HIV/AIDS	-0.556165
	Adult Mortality	-0.696386

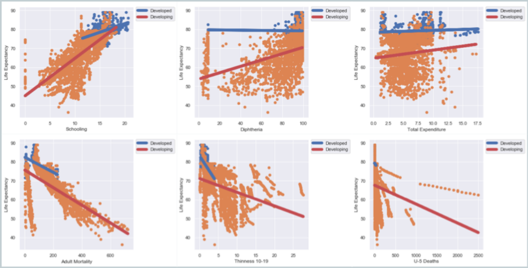
FEATURE HISTOGRAMS



LIFE EXPECTANCY TRENDS

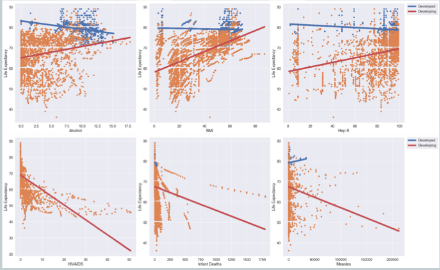
+Correlation

-Correlation

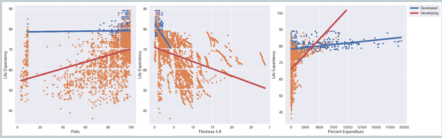


Highest Tier Correlation (~.7) Mid Tier Correlation(~.5) Low Tier Correlation(~.2)

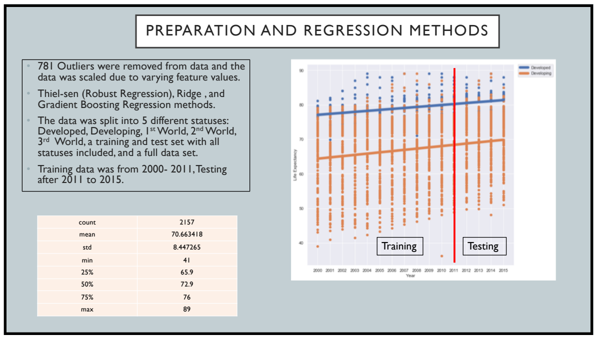
LIFE EXPECTANCY TRENDS



LIFE EXPECTANCY TRENDS



REGRESSION

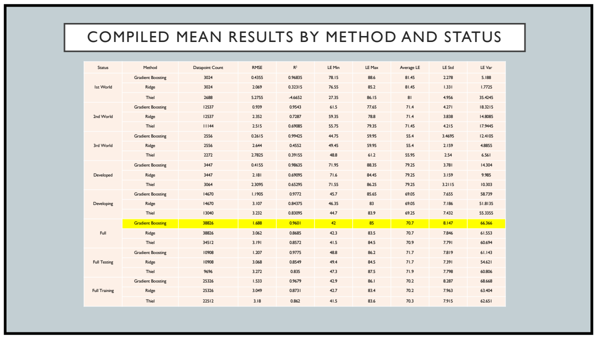


PREPARATION AND REGRESSION METHODS

- 781 Outliers were removed from data and the data was scaled due to varying feature values.
- Thiel-sen (Robust Regression), Ridge , and Gradient Boosting Regression methods.
- The data was split into 5 different statuses: Developed, Developing, 1st World, 2nd World, 3rd World, a training and test set with all statuses included, and a full data set.
- Training data was from 2000- 2011, Testing after 2011 to 2015.

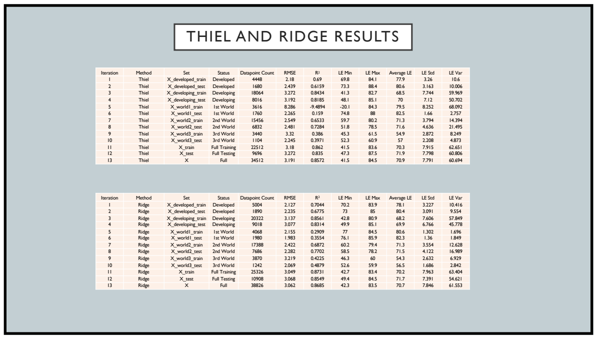
count	2157
mean	70.663418
std	8.447265
min	41
25%	65.9
50%	72.9
75%	76
max	89

Developed
Developing
Life Expectancy
90
80
70
60
50
40
Training
Testing
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
Year



COMPILED MEAN RESULTS BY METHOD AND STATUS

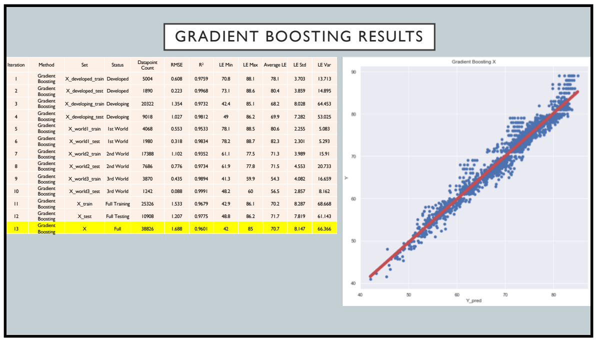
Status	Method	Datapoint Count	RMSE	R²	LE Min	LE Max	Average LE	LE Std	LE Var
1st World	Gradient Boosting	3024	0.4355	0.96835	78.15	88.6	81.45	2.278	5.188
	Ridge	3024	2.069	0.32315	76.55	85.2	81.45	1.331	1.7725
	Thiel	2688	5.2755	-4.6652	27.35	86.15	81	4.956	35.4245
2nd World	Gradient Boosting	12537	0.939	0.9543	61.5	77.65	71.4	4.271	18.3215
	Ridge	12537	2.352	0.7287	59.35	78.8	71.4	3.838	14.8085
	Thiel	11144	2.515	0.69085	55.75	79.35	71.45	4.215	17.9445
3rd World	Gradient Boosting	2556	0.2615	0.99425	44.75	59.95	55.4	3.4695	12.4105
	Ridge	2556	2.644	0.4552	49.45	59.95	55.4	2.159	4.8855
	Thiel	2272	2.7825	0.39155	48.8	61.2	55.95	2.54	6.561
Developed	Gradient Boosting	3447	0.4155	0.98635	71.95	88.35	79.25	3.781	14.304
	Ridge	3447	2.181	0.69095	71.6	84.45	79.25	3.159	9.985
	Thiel	3064	2.3095	0.65295	71.55	86.25	79.25	3.2115	10.303
Developing	Gradient Boosting	14670	1.1905	0.9772	45.7	85.65	69.05	7.655	58.739
	Ridge	14670	3.107	0.84375	46.35	83	69.05	7.186	51.8135
	Thiel	13040	3.232	0.83095	44.7	83.9	69.25	7.432	55.3355
Full	Gradient Boosting	38826	1.688	0.9601	42	85	70.7	8.147	66.366
	Ridge	38826	3.062	0.8685	42.3	83.5	70.7	7.846	61.553
	Thiel	34512	3.191	0.8572	41.5	84.5	70.9	7.791	60.694
Full Testing	Gradient Boosting	10908	1.207	0.9775	48.8	86.2	71.7	7.819	61.143
	Ridge	10908	3.068	0.8549	49.4	84.5	71.7	7.391	54.621
	Thiel	9696	3.272	0.835	47.3	87.5	71.9	7.798	60.806
Full Training	Gradient Boosting	25326	1.533	0.9679	42.9	86.1	70.2	8.287	68.668
	Ridge	25326	3.049	0.8731	42.7	83.4	70.2	7.963	63.404
	Thiel	22512	3.18	0.862	41.5	83.6	70.3	7.915	62.651



THIEL AND RIDGE RESULTS

Iteration	Method	Set	Status	Datapoint Count	RMSE	R²	LE Min	LE Max	Average LE	LE Std	LE Var
1	Thiel	X_developed_train	Developed	4448	2.18	0.69	69.8	84.1	77.9	3.26	10.6
2	Thiel	X_developed_test	Developed	1680	2.439	0.6159	73.3	88.4	80.6	3.163	10.006
3	Thiel	X_developing_train	Developing	18064	3.272	0.8434	41.3	82.7	68.5	7.744	59.969
4	Thiel	X_developing_test	Developing	8016	3.192	0.8185	48.1	85.1	70	7.12	50.702
5	Thiel	X_world1_train	1st World	3616	8.286	-9.4894	-20.1	84.3	79.5	8.252	68.092
6	Thiel	X_world1_test	1st World	1760	2.265	0.159	74.8	88	82.5	1.66	2.757
7	Thiel	X_world2_train	2nd World	15456	2.549	0.6533	59.7	80.2	71.3	3.794	14.394
8	Thiel	X_world2_test	2nd World	6832	2.481	0.7284	51.8	78.5	71.6	4.636	21.495
9	Thiel	X_world3_train	3rd World	3440	3.32	0.386	45.3	61.5	54.9	2.872	8.249
10	Thiel	X_world3_test	3rd World	1104	2.245	0.3971	52.3	60.9	57	2.208	4.873
11	Thiel	X_train	Full Training	22512	3.18	0.862	41.5	83.6	70.3	7.915	62.651
12	Thiel	X_test	Full Testing	9696	3.272	0.835	47.3	87.5	71.9	7.798	60.806
13	Thiel	X	Full	34512	3.191	0.8572	41.5	84.5	70.9	7.791	60.694

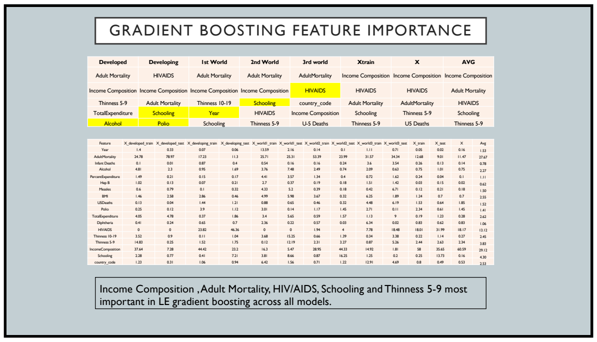
Iteration	Method	Set	Status	Datapoint Count	RMSE	R²	LE Min	LE Max	Average LE	LE Std	LE Var
1	Ridge	X_developed_train	Developed	5004	2.127	0.7044	70.2	83.9	78.1	3.227	10.416
2	Ridge	X_developed_test	Developed	1890	2.235	0.6775	73	85	80.4	3.091	9.554
3	Ridge	X_developing_train	Developing	20322	3.137	0.8561	42.8	80.9	68.2	7.606	57.849
4	Ridge	X_developing_test	Developing	9018	3.077	0.8314	49.9	85.1	69.9	6.766	45.778
5	Ridge	X_world1_train	1st World	4068	2.155	0.2909	77	84.5	80.6	1.302	1.696
6	Ridge	X_world1_test	1st World	1980	1.983	0.3554	76.1	85.9	82.3	1.36	1.849
7	Ridge	X_world2_train	2nd World	17388	2.422	0.6872	60.2	79.4	71.3	3.554	12.628
8	Ridge	X_world2_test	2nd World	7686	2.282	0.7702	58.5	78.2	71.5	4.122	16.989
9	Ridge	X_world3_train	3rd World	3870	3.219	0.4225	46.3	60	54.3	2.632	6.929
10	Ridge	X_world3_test	3rd World	1242	2.069	0.4879	52.6	59.9	56.5	1.686	2.842
11	Ridge	X_train	Full Training	25326	3.049	0.8731	42.7	83.4	70.2	7.963	63.404
12	Ridge	X_test	Full Testing	10908	3.068	0.8549	49.4	84.5	71.7	7.391	54.621
13	Ridge	X	Full	38826	3.062	0.8685	42.3	83.5	70.7	7.846	61.553



GRADIENT BOOSTING RESULTS

Iteration	Method	Set	Status	Datapoint Count	RMSE	R²	LE Min	LE Max	Average LE	LE Std	LE Var
1	Gradient Boosting	X_developed_train	Developed	5004	0.608	0.9759	70.8	88.1	78.1	3.703	13.713
2	Gradient Boosting	X_developed_test	Developed	1890	0.223	0.9968	73.1	88.6	80.4	3.859	14.895
3	Gradient Boosting	X_developing_train	Developing	20322	1.354	0.9732	42.4	85.1	68.2	8.028	64.453
4	Gradient Boosting	X_developing_test	Developing	9018	1.027	0.9812	49	86.2	69.9	7.282	53.025
5	Gradient Boosting	X_world1_train	1st World	4068	0.553	0.9533	78.1	88.5	80.6	2.255	5.083
6	Gradient Boosting	X_world1_test	1st World	1980	0.318	0.9834	78.2	88.7	82.3	2.301	5.293
7	Gradient Boosting	X_world2_train	2nd World	17388	1.102	0.9352	61.1	77.5	71.3	3.989	15.91
8	Gradient Boosting	X_world2_test	2nd World	7686	0.776	0.9734	61.9	77.8	71.5	4.553	20.733
9	Gradient Boosting	X_world3_train	3rd World	3870	0.435	0.9894	41.3	59.9	54.3	4.082	16.659
10	Gradient Boosting	X_world3_test	3rd World	1242	0.088	0.9991	48.2	60	56.5	2.857	8.162
11	Gradient Boosting	X_train	Full Training	25326	1.533	0.9679	42.9	86.1	70.2	8.287	68.668
12	Gradient Boosting	X_test	Full Testing	10908	1.207	0.9775	48.8	86.2	71.7	7.819	61.143
13	Gradient Boosting	X	Full	38826	1.688	0.9601	42	85	70.7	8.147	66.366

Gradient Boosting X
90
80
70
60
50
40
Y
40
50
60
70
80
Y_pred



GRADIENT BOOSTING FEATURE IMPORTANCE

Developed	**Developing**	**1st World**	**2nd World**	**3rd world**	**Xtrain**	**X**	**AVG**
Adult Mortality	HIVAIDS	Adult Mortality	Adult Mortality	AdultMortality	Income Composition	Income Composition	Income Composition
Income Composition	Income Composition	Income Composition	Income Composition	HIVAIDS	HIVAIDS	HIVAIDS	Adult Mortality
Thinness 5-9	Adult Mortality	Thinness 10-19	Schooling	country_code	Adult Mortality	AdultMortality	HIVAIDS
TotalExpenditure	Schooling	Year	HIVAIDS	Income Composition	Schooling	Thinness 5-9	Schooling
Alcohol	Polio	Schooling	Thinness 5-9	U-5 Deaths	Thinness 5-9	U5 Deaths	Thinness 5-9

Feature	X_developed_train	X_developed_test	X_developing_train	X_developing_test	X_world1_train	X_world1_test	X_world2_train	X_world2_test	X_world3_train	X_world3_test	X_train	X_test	X	Avg
Year	1.4	0.33	0.07	0.06	13.59	2.16	0.14	0.1	1.11	0.71	0.05	0.02	0.16	1.53
AdultMortality	24.78	78.97	17.23	11.3	25.71	25.31	53.39	23.99	31.57	34.34	12.68	9.01	11.47	27.67
Infant Deaths	0.1	0.01	0.87	0.4	0.54	0.16	0.16	0.24	3.6	3.54	0.26	0.13	0.14	0.78
Alcohol	4.81	2.3	0.95	1.69	3.76	7.48	2.49	0.74	2.09	0.63	0.75	1.01	0.75	2.27
PercentExpenditure	1.49	0.21	0.15	0.17	4.41	3.57	1.34	0.4	0.72	1.62	0.24	0.04	0.1	1.11
Hep B	1.02	0.13	0.07	0.21	2.7	0.37	0.19	0.18	1.51	1.42	0.03	0.15	0.02	0.62
Measles	0.6	0.79	0.1	0.32	4.33	5.2	0.39	0.18	0.42	6.71	0.12	0.21	0.18	1.50
BMI	1.46	2.58	2.86	0.46	4.99	5.98	3.67	0.32	6.25	1.89	1.24	0.7	0.7	2.55
U5Deaths	0.13	0.04	1.44	1.21	0.88	0.65	0.46	0.32	4.48	6.19	1.53	0.64	1.85	1.52
Polio	0.25	0.12	3.9	1.12	3.01	0.14	1.17	1.45	2.71	0.11	2.34	0.61	1.45	1.41
TotalExpenditure	4.05	4.78	0.37	1.86	3.4	5.65	0.59	1.57	1.13	9	0.19	1.23	0.28	2.62
Diphtheria	0.41	0.24	0.65	0.7	2.36	0.22	0.57	0.03	6.34	0.02	0.83	0.62	0.83	1.06
HIVAIDS	0	0	23.82	46.36	0	0	1.94	4	7.78	18.48	18.01	31.99	18.17	13.12
Thinness 10-19	3.52	0.9	0.11	1.04	3.68	15.25	0.66	1.39	0.34	3.38	0.22	1.14	0.27	2.45
Thinness 5-9	14.83	0.25	1.52	1.75	0.12	12.19	2.31	3.27	0.87	5.26	2.44	2.63	2.34	3.83
IncomeComposition	37.64	7.28	44.42	23.2	16.3	5.47	28.95	44.33	14.92	1.81	58	35.65	60.59	29.12
Schooling	2.28	0.77	0.41	7.21	3.81	8.66	0.87	16.25	1.25	0.2	0.25	13.73	0.16	4.30
country_code	1.23	0.31	1.06	0.94	6.42	1.56	0.71	1.22	12.91	4.69	0.8	0.49	0.53	2.53

Income Composition ,Adult Mortality, HIV/AIDS, Schooling and Thinness 5-9 most important in LE gradient boosting across all models.

CLUSTERING

CLUSTERING METHODS AND EVALUATION

- Mean-Shift and K-Means Clustering with scaling of the full data set.
- K-Means cluster chosen by comparing scores calculated below.
- Silhouette and Calinski-Harabaz scores were used as non-ground truth scores.
- Completeness, Homogeneity, and ARI scores were used as ground truth scores related to the world feature.
- Percentages of each cluster were calculated to choose the appropriate clustering methods.

CLUSTERING EVALUATION SCORES

Calinski-Harabaz Index-

- Additionally known as the Variance Ratio Criterion where a higher score means a better defined cluster.
- It compares the ratio of the between-clusters dispersion mean and the within-cluster dispersion.
- Scores are higher when dense and separated from other clusters.
- The score is normalized with respect to the others scores for comparison in one chart.

Silhouette Score-

- the ratio of difference between the mean nearest-cluster distance and mean intra-cluster distance over the maximum between both scores.
- +1 indicates a highly dense cluster while scores close to 0 indicates overlapping clusters and -1 indicates incorrect clustering.
  s=(b-a)/max(b,a)
  mean intra-cluster distance (a)
  mean nearest-cluster distance (b)

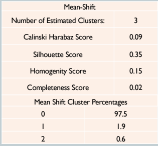
Source: Caliński, T., & Harabasz, J. (1974). "A dendrite method for cluster analysis". Communications in Statistics-theory and Methods 3: 1-27. doi:10.1080/03610926.2011.560741.
Source: Peter J. Rousseeuw (1987). "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis". Computational and Applied Mathematics 20: 53–65. doi:10.1016/0377-0427(87)90125-7.

CLUSTERING EVALUATION SCORES

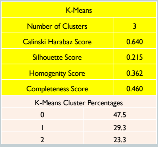
- Homogeneity Score- Each cluster contains only members of a single class.
- Completeness Score-All members of a given class are assigned to the same cluster.

- Adjusted Rand Score-The Rand Index compares how pairs of datapoints relate in the ground truth and in the post-clustering assignment. There are four possible types of pair relationships:
  a=Members of the same cluster in the ground truth match same cluster in the new solution.
  b=Members of the same cluster in the ground truth match different clusters in the new solution.
  c=Members of different clusters in the ground truth match the same cluster in the new solution.
  d=Members of different clusters in the ground truth match different clusters in the new solution.
  E(RI)- expected RI
  RI=(a+c)/sum(a,b,c,d)
  ARI=(RI-E(RI)/(max(RI)-E(RI))

COMPARISON OF K-MEANS AND MEAN SHIFT

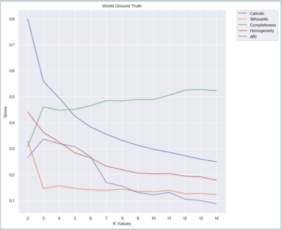


Mean-Shift	
Number of Estimated Clusters:	3
Calinski Harabaz Score	0.09
Silhouette Score	0.35
Homogenity Score	0.15
Completeness Score	0.02
Mean Shift Cluster Percentages	
0	97.5
1	1.9
2	0.6

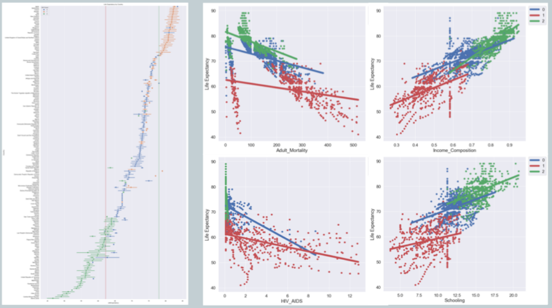


K-Means	
Number of Clusters	3
Calinski Harabaz Score	0.640
Silhouette Score	0.215
Homogenity Score	0.362
Completeness Score	0.460
K-Means Cluster Percentages	
0	47.5
1	29.3
2	23.3

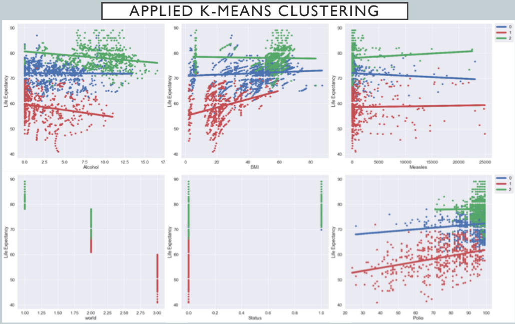
K-CLUSTER DETERMINATION



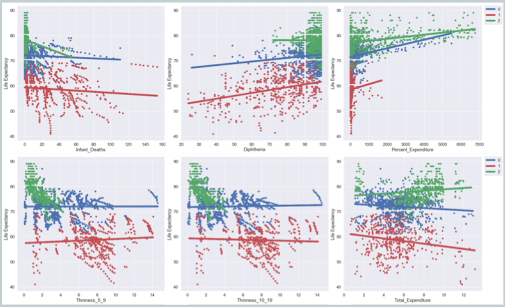
APPLIED K-MEANS CLUSTERING



APPLIED K-MEANS CLUSTERING



APPLIED K-MEANS CLUSTERING

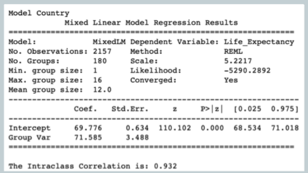


MIXED EFFECT MODELS

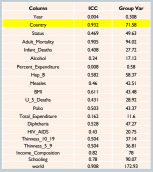
MIXED EFFECT MODEL ASSUMPTIONS

- Mixed Effect Models have the following assumptions:
- The features are related linearly to the outcome.
- The errors have constant variance: In other words, the model fits equally well for all values of the outcome and features within a level.
- The errors are independent: The fit of the model within a group(level 1) is uncorrelated with the fit of the model at between a group (level 2).
- The errors are normally distributed.
- Observations within a person/group are correlated with one another(high ICC).

ICC AND GROUP FEATURE SELECTION



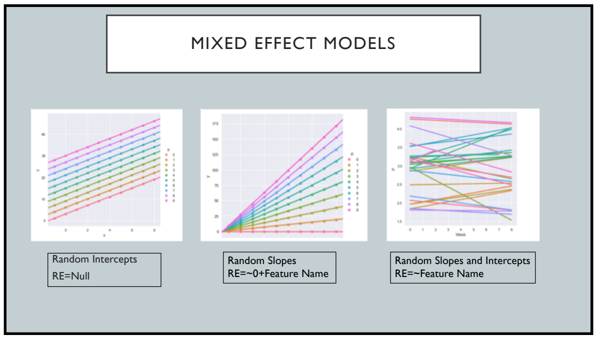
```
Model Country
            Mixed Linear Model Regression Results
==============================================================
Model:              MixedLM Dependent Variable: Life_Expectancy
No. Observations: 2157    Method:             REML
No. Groups:       180     Scale:              5.2217
Min. group size:  1       Likelihood:         -5290.2892
Max. group size:  16      Converged:          Yes
Mean group size:  12.0
--------------------------------------------------------------
              Coef.  Std.Err.    z    P>|z| [0.025  0.975]
--------------------------------------------------------------
Intercept    69.776    0.634  110.102 0.000 68.534  71.018
Group Var    71.585    3.488
==============================================================

The Intraclass Correlation is: 0.932
```



Column	ICC	Group Var
Year	0.004	0.308
Country	0.932	71.58
Status	0.469	49.63
Adult_Mortality	0.905	94.02
Infant_Deaths	0.408	27.72
Alcohol	0.24	17.12
Percent_Expenditure	0.008	0.58
Hep_B	0.582	58.37
Measles	0.46	42.51
BMI	0.611	43.48
U_5_Deaths	0.431	28.92
Polio	0.503	43.37
Total_Expenditure	0.162	11.6
Diphtheria	0.528	47.27
HIV_AIDS	0.43	20.75
Thinness_10_19	0.504	37.14
Thinness_5_9	0.504	36.81
Income_Composition	0.82	78
Schooling	0.78	90.07
world	0.908	172.93

Sample Calculation
ICC=71.585/71.585+5.22=.932

group var: between-group variance
resid: within-group variance
ICC= group var/(group var +resid)

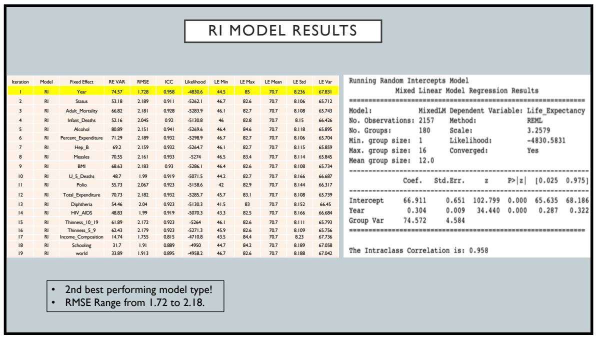


MIXED EFFECT MODELS

Random Intercepts
RE=Null

Random Slopes
RE=~0+Feature Name

Random Slopes and Intercepts
RE=~Feature Name

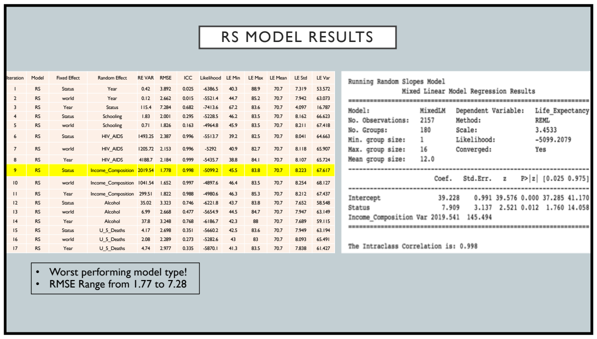


RI MODEL RESULTS

Iteration	Model	Fixed Effect	RE VAR	RMSE	ICC	Likelihood	LE Min	LE Max	LE Mean	LE Std	LE Var
1	RI	Year	74.57	1.728	0.958	-4830.6	44.5	85	70.7	8.236	67.831
2	RI	Status	53.18	2.189	0.911	-5262.1	46.7	82.6	70.7	8.106	65.712
3	RI	Adult_Mortality	66.82	2.181	0.928	-5283.9	46.1	82.7	70.7	8.108	65.743
4	RI	Infant_Deaths	52.16	2.045	0.92	-5130.8	46	82.8	70.7	8.15	66.426
5	RI	Alcohol	80.89	2.151	0.941	-5269.6	46.4	84.6	70.7	8.118	65.895
6	RI	Percent_Expenditure	71.29	2.189	0.932	-5298.9	46.7	82.7	70.7	8.106	65.704
7	RI	Hep_B	69.2	2.159	0.932	-5264.7	46.1	82.7	70.7	8.115	65.859
8	RI	Measles	70.55	2.161	0.933	-5274	46.5	83.4	70.7	8.114	65.845
9	RI	BMI	68.63	2.183	0.93	-5286.1	46.4	82.6	70.7	8.108	65.734
10	RI	U_5_Deaths	48.7	1.99	0.919	-5071.5	44.2	82.7	70.7	8.166	66.687
11	RI	Polio	55.73	2.067	0.923	-5158.6	42	82.9	70.7	8.144	66.317
12	RI	Total_Expenditure	70.73	2.182	0.932	-5285.7	45.7	83.1	70.7	8.108	65.739
13	RI	Diphtheria	54.46	2.04	0.923	-5130.3	41.5	83	70.7	8.152	66.45
14	RI	HIV_AIDS	48.83	1.99	0.919	-5070.3	43.3	82.5	70.7	8.166	66.684
15	RI	Thinness_10_19	61.89	2.172	0.923	-5264	46.1	82.6	70.7	8.111	65.793
16	RI	Thinness_5_9	62.43	2.179	0.923	-5271.3	45.9	82.6	70.7	8.109	65.756
17	RI	Income_Composition	14.74	1.755	0.815	-4710.8	43.5	84.4	70.7	8.23	67.736
18	RI	Schooling	31.7	1.91	0.889	-4950	44.7	84.2	70.7	8.189	67.058
19	RI	world	33.89	1.913	0.895	-4958.2	46.7	82.6	70.7	8.188	67.042

```
Running Random Intercepts Model
          Mixed Linear Model Regression Results
==============================================================
Model:              MixedLM Dependent Variable: Life_Expectancy
No. Observations: 2157    Method:             REML
No. Groups:         180     Scale:              3.2579
Min. group size:   1       Likelihood:         -4830.5831
Max. group size:   16      Converged:          Yes
Mean group size:   12.0
--------------------------------------------------------------
            Coef.  Std.Err.    z     P>|z| [0.025  0.975]
--------------------------------------------------------------
Intercept   66.911    0.651 102.799 0.000 65.635 68.186
Year         0.304    0.009  34.440 0.000  0.287  0.322
Group Var   74.572    4.584
==============================================================

The Intraclass Correlation is: 0.958
```

- 2nd best performing model type!
- RMSE Range from 1.72 to 2.18.

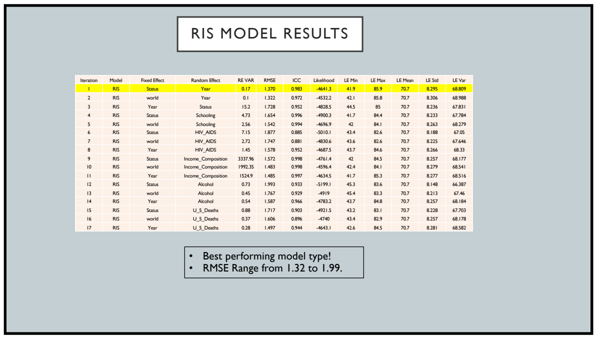


RS MODEL RESULTS

Iteration	Model	Fixed Effect	Random Effect	RE VAR	RMSE	ICC	Likelihood	LE Min	LE Max	LE Mean	LE Std	LE Var
1	RS	Status	Year	0.42	3.892	0.025	-6386.5	40.3	88.9	70.7	7.319	53.572
2	RS	world	Year	0.12	2.662	0.015	-5521.4	44.7	85.2	70.7	7.942	63.073
3	RS	Year	Status	115.4	7.284	0.682	-7413.6	67.2	83.6	70.7	4.097	16.787
4	RS	Status	Schooling	1.83	2.001	0.295	-5228.5	46.2	83.5	70.7	8.162	66.623
5	RS	world	Schooling	0.71	1.826	0.163	-4964.8	45.9	83.5	70.7	8.211	67.418
6	RS	Status	HIV_AIDS	1493.25	2.387	0.996	-5513.7	39.2	82.5	70.7	8.041	64.663
7	RS	world	HIV_AIDS	1205.72	2.153	0.996	-5292	40.9	82.7	70.7	8.118	65.907
8	RS	Year	HIV_AIDS	4188.7	2.184	0.999	-5435.7	38.8	84.1	70.7	8.107	65.724
9	RS	Status	Income_Composition	2019.54	1.778	0.998	-5099.2	45.5	83.8	70.7	8.223	67.617
10	RS	world	Income_Composition	1041.54	1.652	0.997	-4897.6	46.4	83.5	70.7	8.254	68.127
11	RS	Year	Income_Composition	299.51	1.822	0.988	-4980.6	46.3	85.3	70.7	8.212	67.437
12	RS	Status	Alcohol	35.02	3.323	0.746	-6221.8	43.7	83.8	70.7	7.652	58.548
13	RS	world	Alcohol	6.99	2.668	0.477	-5654.9	44.5	84.7	70.7	7.947	63.149
14	RS	Year	Alcohol	37.8	3.248	0.768	-6186.7	42.3	88	70.7	7.689	59.115
15	RS	Status	U_5_Deaths	4.17	2.698	0.351	-5660.2	42.5	83.6	70.7	7.949	63.194
16	RS	world	U_5_Deaths	2.08	2.289	0.273	-5282.6	43	83	70.7	8.093	65.491
17	RS	Year	U_5_Deaths	4.74	2.977	0.335	-5870.1	41.3	83.5	70.7	7.838	61.427

```
Running Random Slopes Model
              Mixed Linear Model Regression Results
=================================================================
Model:               MixedLM   Dependent Variable:   Life_Expectancy
No. Observations:    2157      Method:               REML
No. Groups:          180       Scale:                3.4533
Min. group size:     1         Likelihood:           -5099.2079
Max. group size:     16        Converged:            Yes
Mean group size:     12.0
-----------------------------------------------------------------
                           Coef.  Std.Err.   z    P>|z| [0.025 0.975]
-----------------------------------------------------------------
Intercept                  39.228    0.991 39.576 0.000 37.285 41.170
Status                      7.909    3.137  2.521 0.012  1.760 14.058
Income_Composition Var 2019.541  145.494
=================================================================

The Intraclass Correlation is: 0.998
```

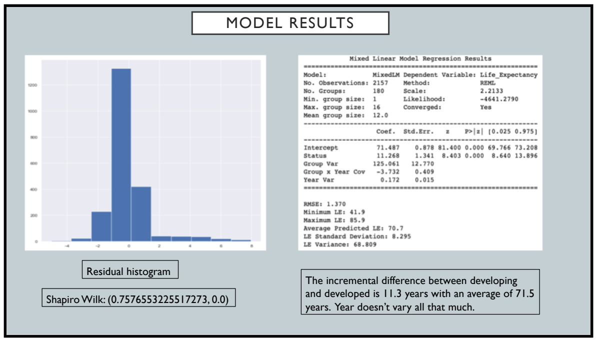
- Worst performing model type!
- RMSE Range from 1.77 to 7.28



RIS MODEL RESULTS

Iteration	Model	Fixed Effect	Random Effect	RE VAR	RMSE	ICC	Likelihood	LE Min	LE Max	LE Mean	LE Std	LE Var
1	RIS	Status	Year	0.17	1.370	0.983	-4641.3	41.9	85.9	70.7	8.295	68.809
2	RIS	world	Year	0.1	1.322	0.972	-4532.2	42.1	85.8	70.7	8.306	68.988
3	RIS	Year	Status	15.2	1.728	0.952	-4828.5	44.5	85	70.7	8.236	67.831
4	RIS	Status	Schooling	4.73	1.654	0.996	-4900.3	41.7	84.4	70.7	8.233	67.784
5	RIS	world	Schooling	2.56	1.542	0.994	-4696.9	42	84.1	70.7	8.263	68.279
6	RIS	Status	HIV_AIDS	7.15	1.877	0.885	-5010.1	43.4	82.6	70.7	8.188	67.05
7	RIS	world	HIV_AIDS	2.72	1.747	0.881	-4830.6	43.6	82.6	70.7	8.225	67.646
8	RIS	Year	HIV_AIDS	1.45	1.578	0.952	-4687.5	43.7	84.6	70.7	8.266	68.33
9	RIS	Status	Income_Composition	3337.96	1.572	0.998	-4761.4	42	84.5	70.7	8.257	68.177
10	RIS	world	Income_Composition	1992.35	1.483	0.998	-4596.4	42.4	84.1	70.7	8.279	68.541
11	RIS	Year	Income_Composition	1524.9	1.485	0.997	-4634.5	41.7	85.3	70.7	8.277	68.516
12	RIS	Status	Alcohol	0.73	1.993	0.933	-5199.1	45.3	83.6	70.7	8.148	66.387
13	RIS	world	Alcohol	0.45	1.767	0.929	-4919	45.4	83.3	70.7	8.213	67.46
14	RIS	Year	Alcohol	0.54	1.587	0.966	-4783.2	43.7	84.8	70.7	8.257	68.184
15	RIS	Status	U_5_Deaths	0.88	1.717	0.903	-4921.5	43.2	83.1	70.7	8.228	67.703
16	RIS	world	U_5_Deaths	0.37	1.606	0.896	-4740	43.4	82.9	70.7	8.257	68.178
17	RIS	Year	U_5_Deaths	0.28	1.497	0.944	-4643.1	42.6	84.5	70.7	8.281	68.582

- Best performing model type!
- RMSE Range from 1.32 to 1.99.



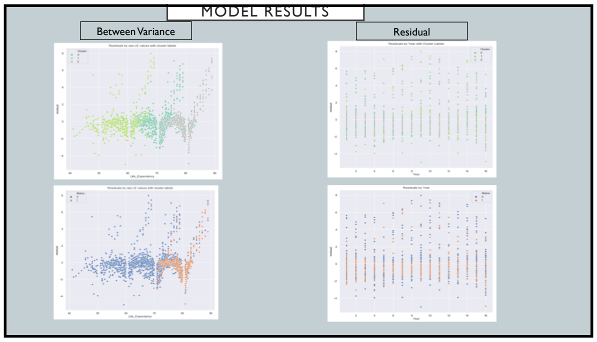
MODEL RESULTS

```
              Mixed Linear Model Regression Results
===============================================================
Model:              MixedLM Dependent Variable: Life_Expectancy
No. Observations: 2157    Method:             REML
No. Groups:         180     Scale:              2.2133
Min. group size:   1       Likelihood:         -4641.2790
Max. group size:   16      Converged:          Yes
Mean group size:   12.0
---------------------------------------------------------------
                 Coef. Std.Err.   z    P>|z| [0.025 0.975]
---------------------------------------------------------------
Intercept        71.487    0.878 81.400 0.000 69.766 73.208
Status           11.268    1.341  8.403 0.000  8.640 13.896
Group Var       125.061   12.770
Group x Year Cov -3.732    0.409
Year Var          0.172    0.015
===============================================================

RMSE: 1.370
Minimum LE: 41.9
Maximum LE: 85.9
Average Predicted LE: 70.7
LE Standard Deviation: 8.295
LE Variance: 68.809
```

Residual histogram

Shapiro Wilk: (0.7576553225517273, 0.0)

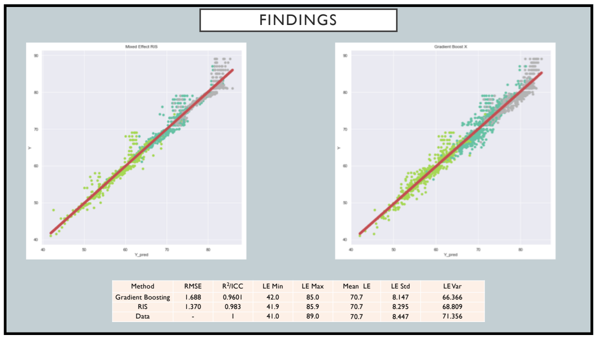
The incremental difference between developing and developed is 11.3 years with an average of 71.5 years. Year doesn't vary all that much.



MODEL RESULTS

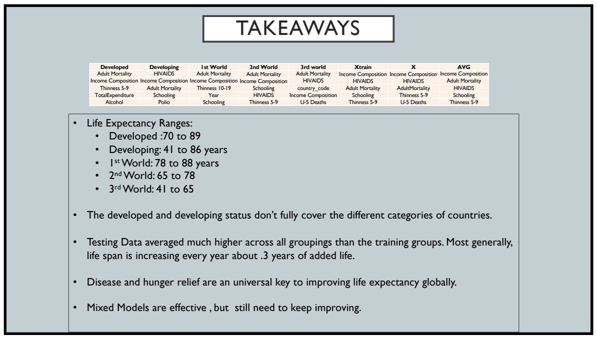
Between Variance

Residual



FINDINGS

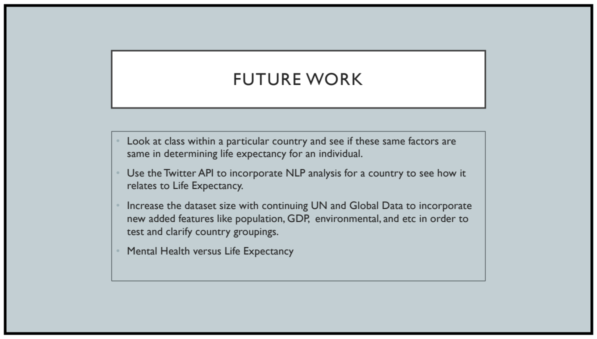
Method	RMSE	R^2/ICC	LE Min	LE Max	Mean LE	LE Std	LE Var
Gradient Boosting	1.688	0.9601	42.0	85.0	70.7	8.147	66.366
RIS	1.370	0.983	41.9	85.9	70.7	8.295	68.809
Data	-	1	41.0	89.0	70.7	8.447	71.356



TAKEAWAYS

Developed	Developing	1st World	2nd World	3rd world	Xtrain	X	AVG
Adult Mortality	HIVAIDS	Adult Mortality	Adult Mortality	Adult Mortality	Income Composition	Income Composition	Income Composition
Income Composition	Income Composition	Income Composition	Income Composition	HIVAIDS	HIVAIDS	HIVAIDS	Adult Mortality
Thinness 5-9	Adult Mortality	Thinness 10-19	Schooling	country_code	Adult Mortality	AdultMortality	HIVAIDS
TotalExpenditure	Schooling	Year	HIVAIDS	Income Composition	Schooling	Thinness 5-9	Schooling
Alcohol	Polio	Schooling	Thinness 5-9	U-5 Deaths	Thinness 5-9	U-5 Deaths	Thinness 5-9

- Life Expectancy Ranges:
  - Developed :70 to 89
  - Developing: 41 to 86 years
  - 1st World: 78 to 88 years
  - 2nd World: 65 to 78
  - 3rd World: 41 to 65
- The developed and developing status don't fully cover the different categories of countries.
- Testing Data averaged much higher across all groupings than the training groups. Most generally, life span is increasing every year about .3 years of added life.
- Disease and hunger relief are an universal key to improving life expectancy globally.
- Mixed Models are effective , but still need to keep improving.



FUTURE WORK

- Look at class within a particular country and see if these same factors are same in determining life expectancy for an individual.
- Use the Twitter API to incorporate NLP analysis for a country to see how it relates to Life Expectancy.
- Increase the dataset size with continuing UN and Global Data to incorporate new added features like population, GDP, environmental, and etc in order to test and clarify country groupings.
- Mental Health versus Life Expectancy