Machine Learning Midterm Project

Kevin Jeswani, Laiba Shah and Junaid Zafar

Agenda

- Industry Research/Motivation
- Data Problem and Description
- Data Understanding
  - Data Exploration
  - Data Visualization
- Linear Regression
- Machine learning model
  - Approach
  - Ensemble Methods
  - Interpreting Results
- Business Application

Industry Research - Effects of Indoor Air pollution

- Around a third of the global population is affected by harmful household air pollution.
- Household air pollution was responsible for an estimated 3.2 million deaths per year in 2020
- Household air pollution exposure leads to noncommunicable diseases e.g. stroke, COPD, Lung cancer
- Women and children bear the greatest health burden from the indoor pollution
- It is essential to identify these household air pollution factors and expand use of clean fuels and technologies[1]

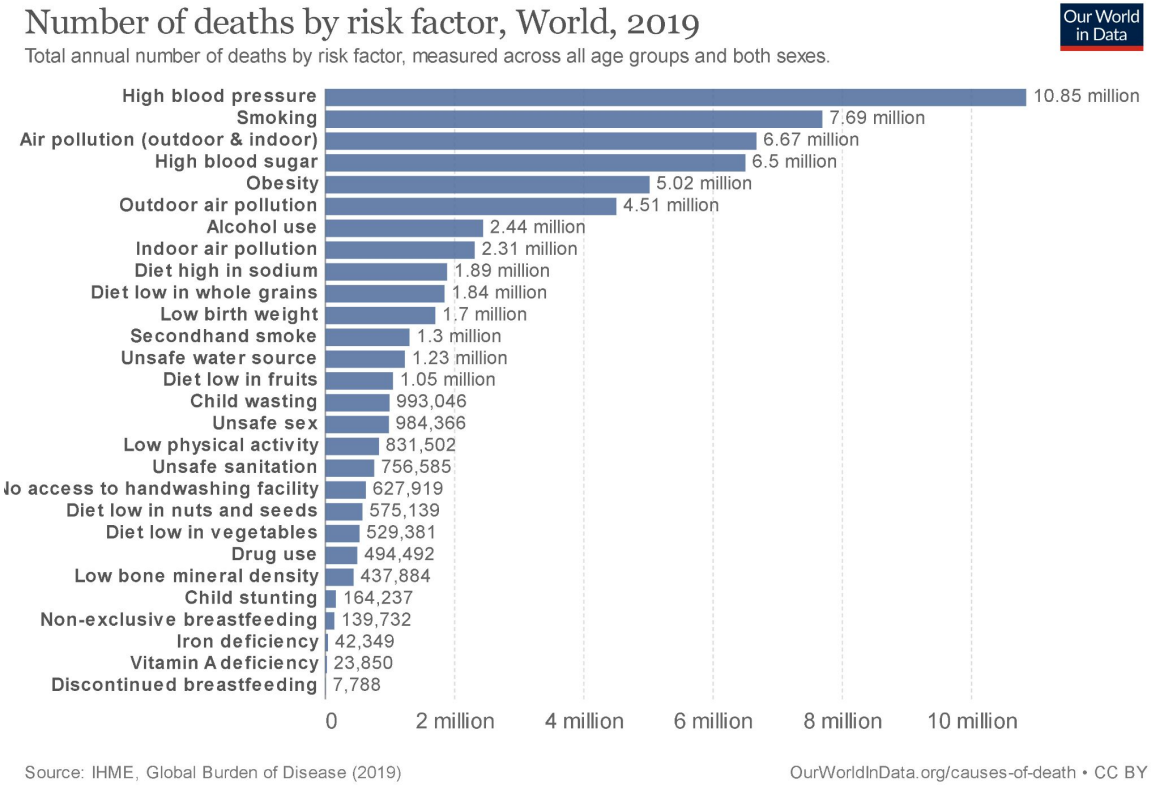
Industry Research - Effects of Outdoor Air pollution

- The lower the levels of outdoor air pollution affects both long- and short-term.
- Air pollution is still significant since recently in 2019, 99% of the world population was living in places where the WHO air quality guidelines levels were not met.
- Outdoor air pollution in both cities and rural areas was estimated to cause 4.2 million premature deaths worldwide in 2016.
- Some 91% of those premature deaths occurred in low- and middle-income countries, and the greatest number in the WHO South-East Asia and Western Pacific regions.
- Policies and investments supporting cleaner and efficient technologies are key to improving air and life quality across the world. [2]

Industry Research

Other prominent factors

- Unsafe Water
- Unsafe Sanitation
- Handwashing[3]



Data Description

- Death rate due to indoor air pollution
- Death rate due to outdoor particulate matter (PM2.5) air pollution
- Death rate due to outdoor ozone air pollution
- Social demographic/Internal factors

Data Understanding

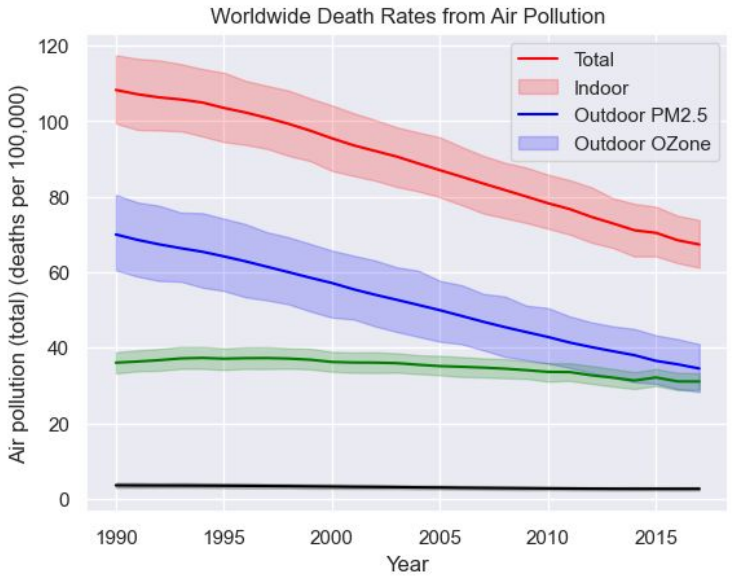
- Data cleaning was performed to remove unnecessary column(s)
- Additional data columns were added from OECD and World bank - other env. factors
  - Socio-Demographic Index, GDP/capita, Population, WB Income Class
- A challenge was to merge data sets with the corresponding country and year

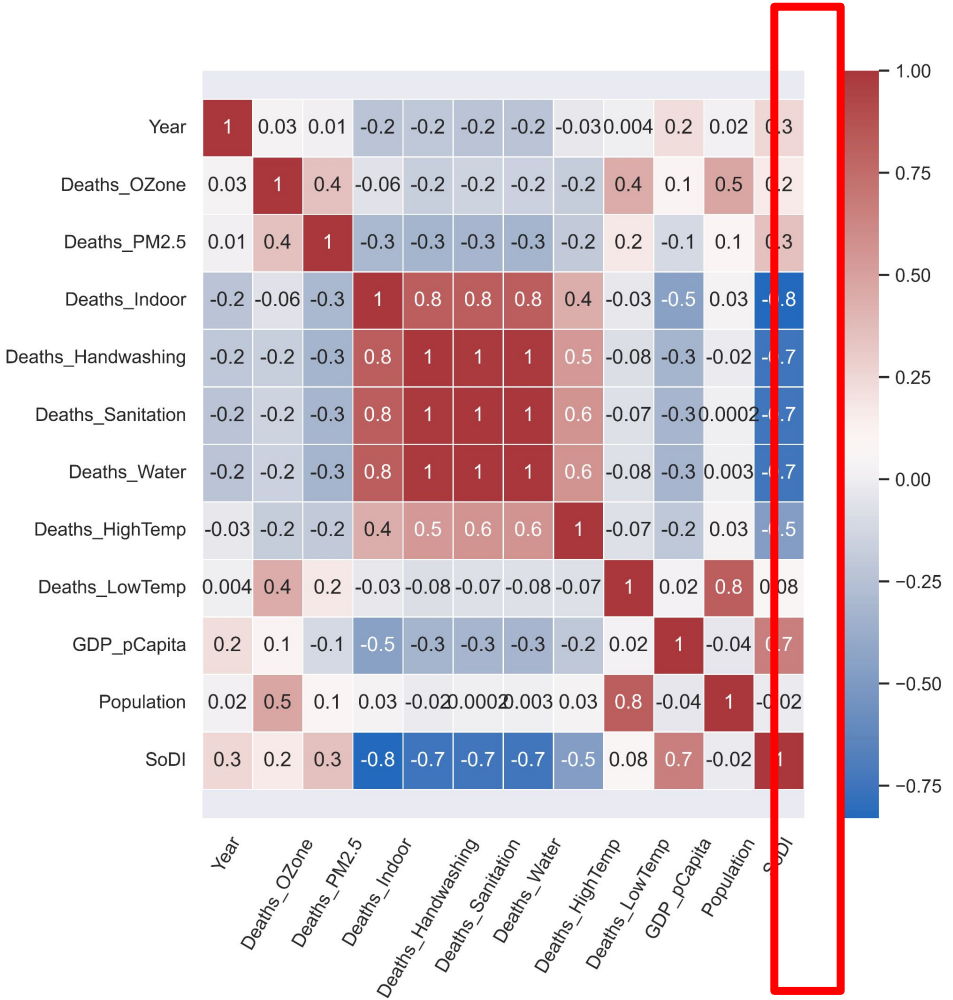
Country	Year	Deaths_OZone	Deaths_PM2.5	Deaths_Indoor	Deaths_Handwashing	Deaths_Sanitation	Deaths_Water	Deaths_HighTemp	Deaths_LowTemp	GDP_pCapita	Population	Income_Class	SoDI
Afghanistan	1990	387.0	2782.0	34372.0	4825.0	2798.0	3702.0	1085.0	7076.0	..	12412311	L	0.187
Afghanistan	1991	376.0	2846.0	35392.0	5127.0	3254.0	4309.0	925.0	7610.0	..	13299016	L	0.191
Afghanistan	1992	364.0	3031.0	38065.0	5889.0	4042.0	5356.0	908.0	8255.0	..	14485543	L	0.195
Afghanistan	1993	367.0	3256.0	41154.0	7007.0	5392.0	7152.0	1159.0	8430.0	..	15816601	L	0.196
Afghanistan	1994	387.0	3401.0	43153.0	7421.0	5418.0	7192.0	1398.0	8659.0	..	17075728	L	0.194
...	...	...	...	...	...	...	...	...	...	...	...	...	...
Zimbabwe	2015	50.0	2785.0	10435.0	4328.0	2879.0	4336.0	389.0	742.0	1445.069702	13814642	L	0.452
Zimbabwe	2016	58.0	2723.0	10365.0	4295.0	2798.0	4244.0	464.0	762.0	1464.588957	14030338	L	0.459
Zimbabwe	2017	70.0	2630.0	10257.0	4251.0	2744.0	4193.0	180.0	891.0	1235.189032	14236599	L	0.465
Zimbabwe	2018	69.0	2600.0	10113.0	4153.0	2608.0	4013.0	282.0	767.0	1254.642265	14438812	LM	0.471
Zimbabwe	2019	73.0	2607.0	10019.0	4113.0	2531.0	3914.0	326.0	744.0	1316.740657	14645473	LM	0.476

Data Understanding

Visualizations

- Most deaths are due to Indoor Air pollution
- Overall Death Rates due to Air pollution is going down in the world

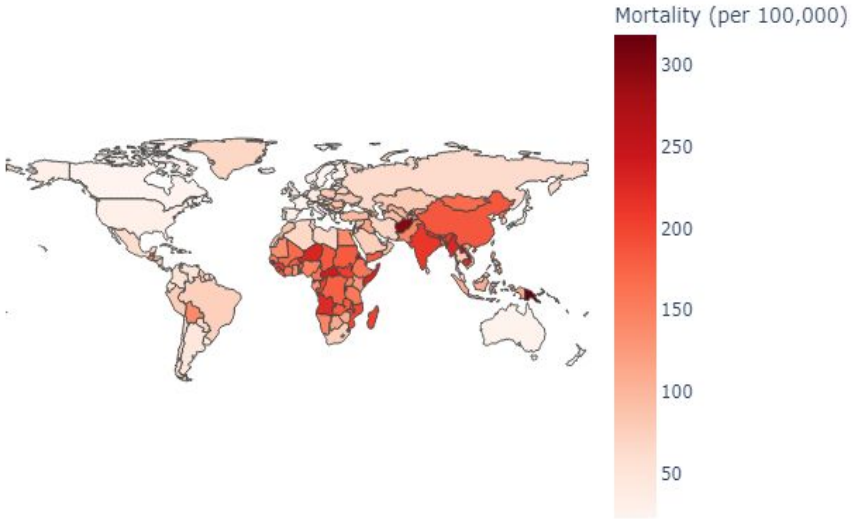




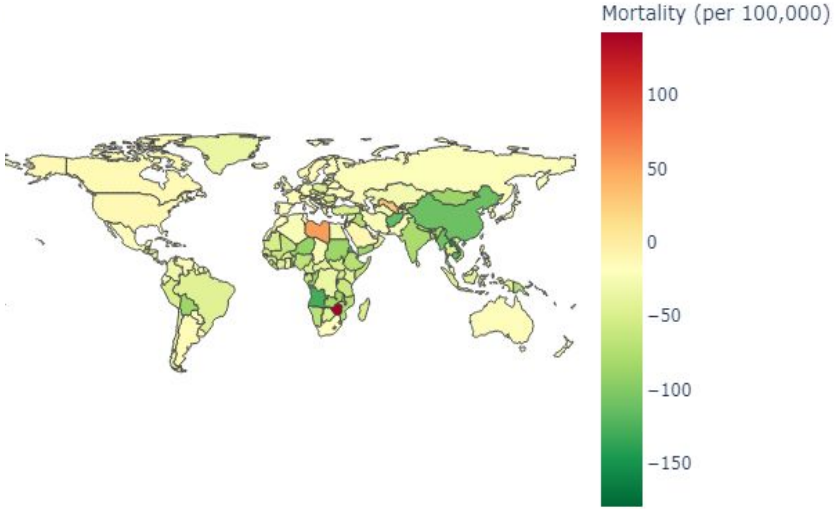
Data Understanding

Heat Maps

Heat Map - Total Air Pollution Deaths in 1990

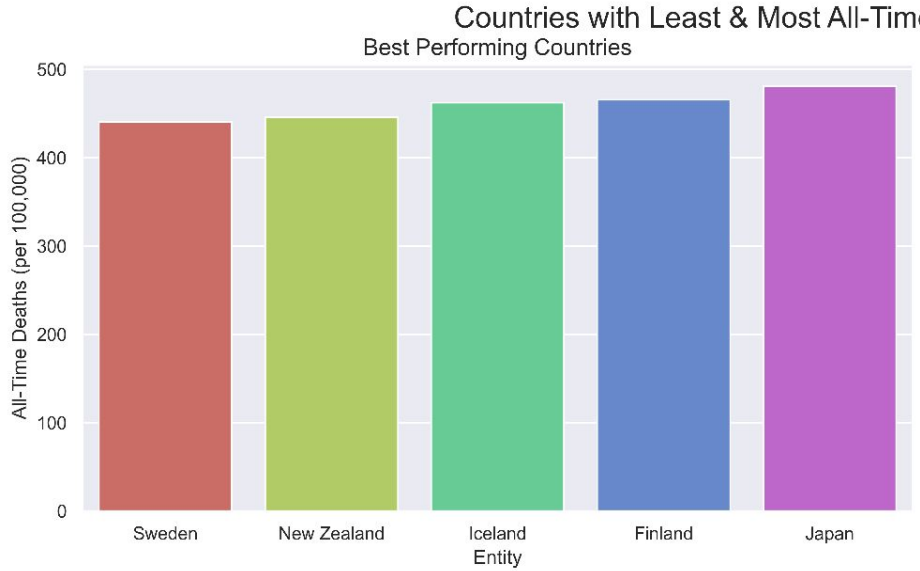


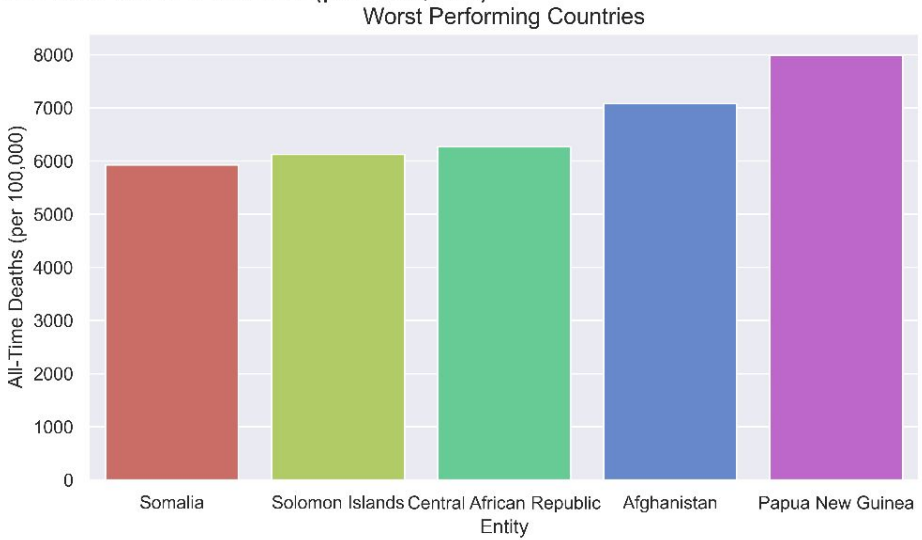
Map - Difference in Total Air Pollution Deaths in Year 1990 & 2017



Data Understanding

Barplots

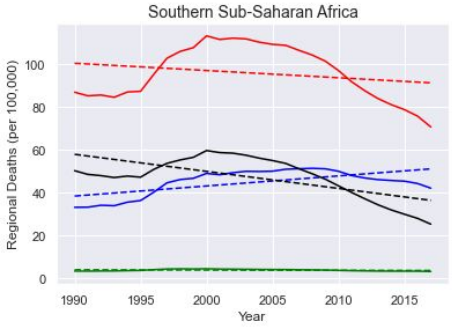


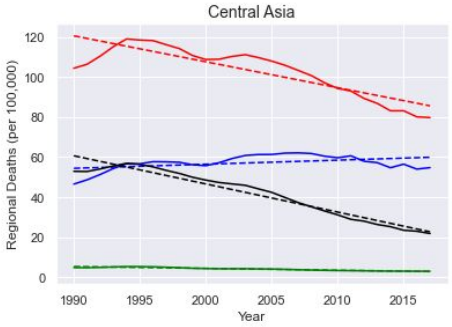


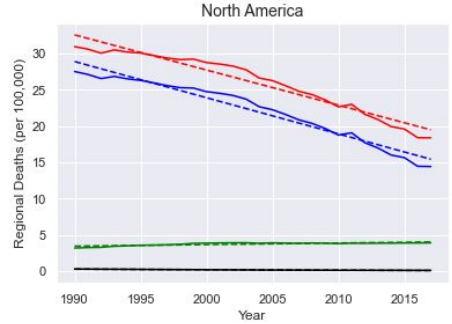
Linear Regression

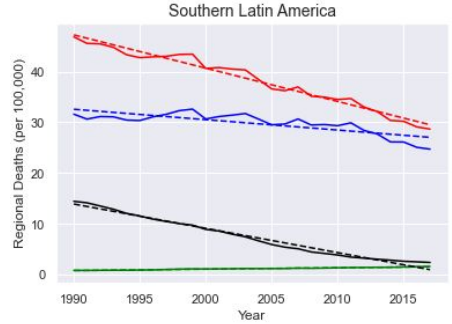
Regional, SoDI & National-level

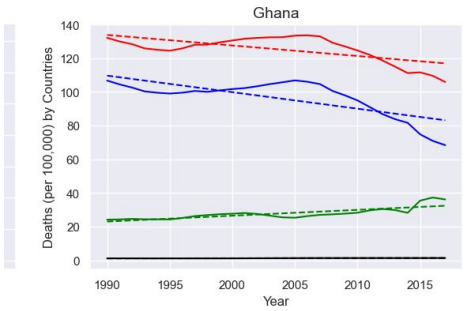
R^2 for **All** Models	Mean	Std
Indoor .	0.89	0.15
PM2.5	0.61	0.34
OZone	0.67	0.31

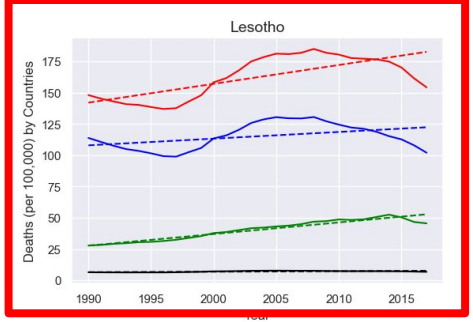


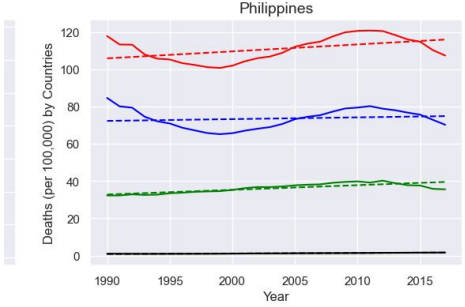


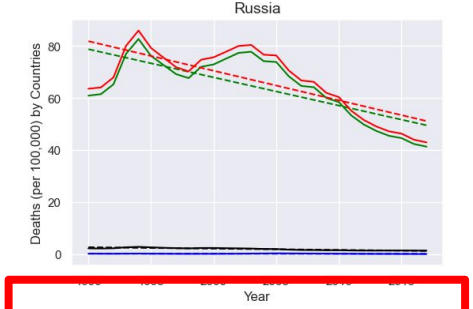


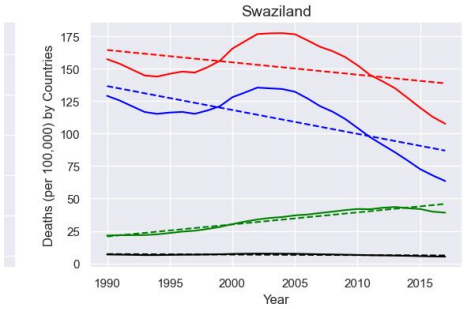


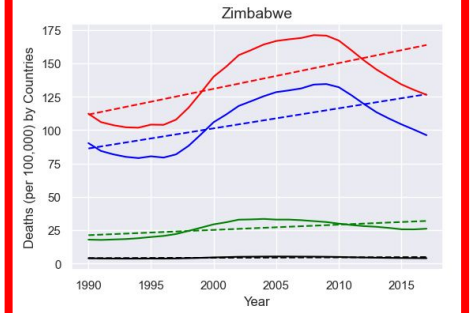


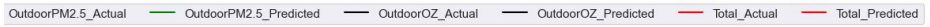




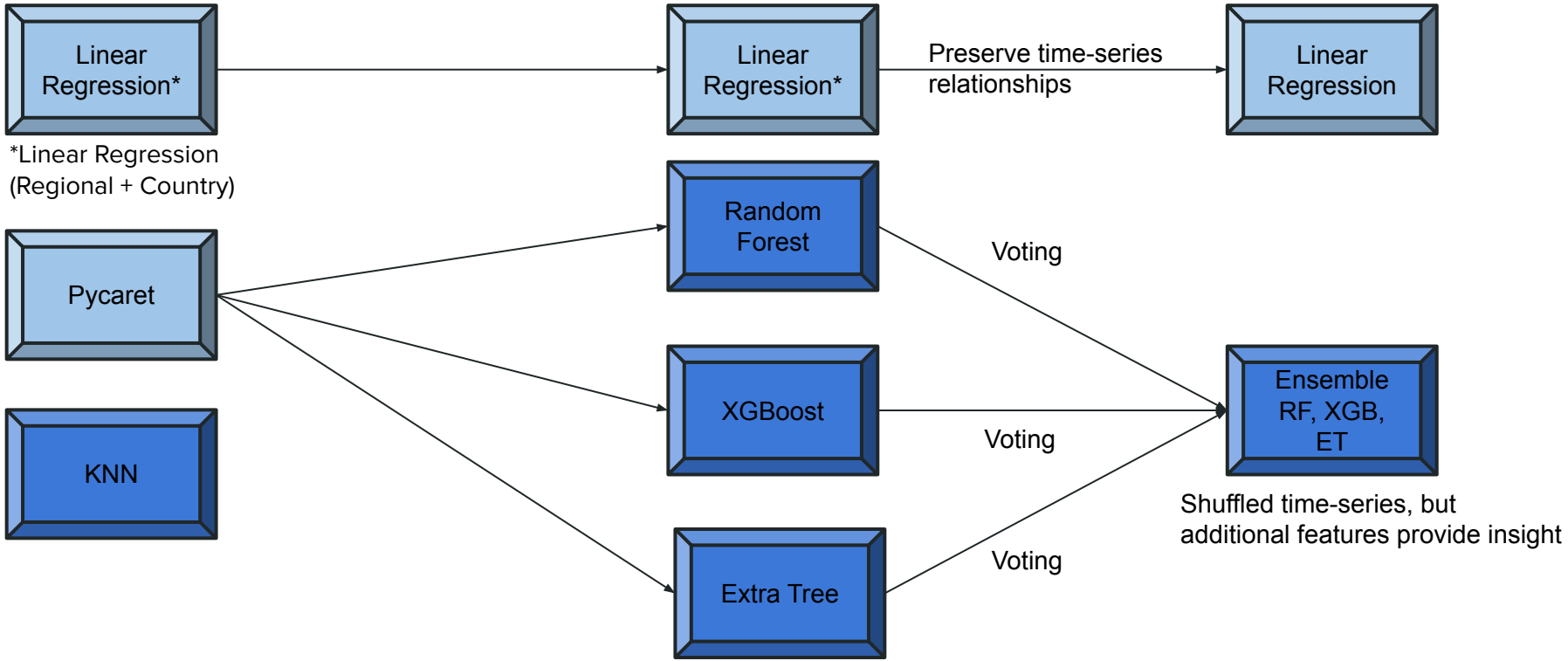






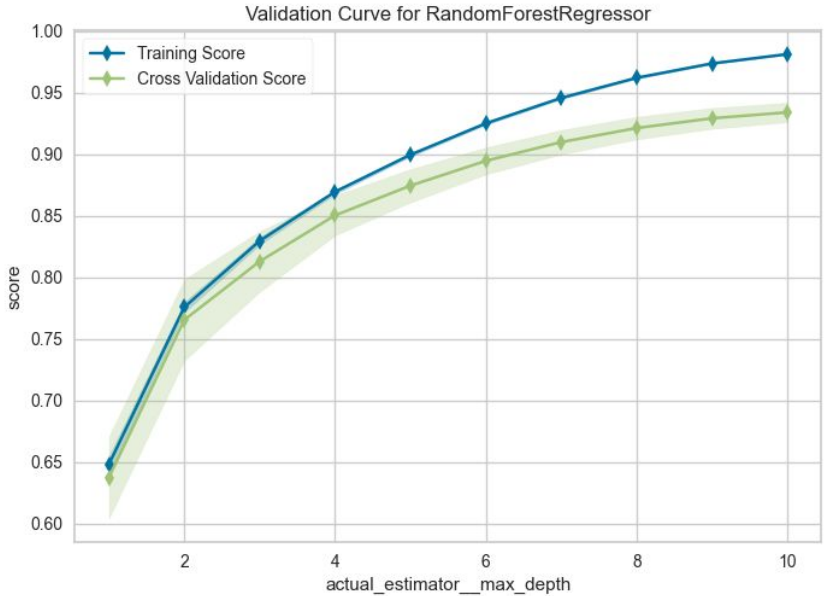


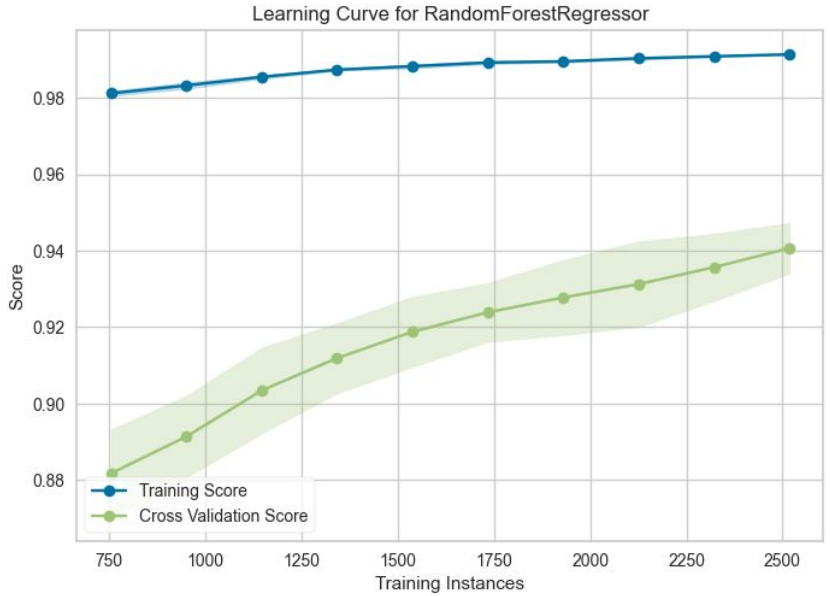
Machine Learning Model



Machine Learning Model

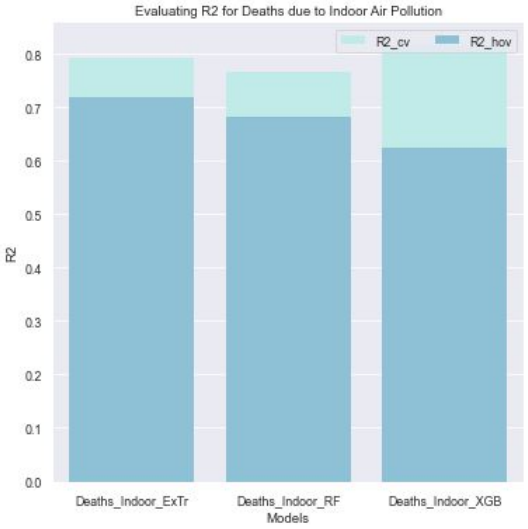
- Pre-processing: Ordinal label encoding, removing entries with missing feature values, tested the use of feature and target scaling
- Narrowed Target: Air Pollution (Indoor, PM2.5, & OZone) Deaths
- Features: Income_Class (encoded), Population, GDP/Capita, SoDI, Year,
- PyCaret selection: Indoor & PM2.5 =Random Forest; OZone = Extra Tree, 2nd/3rd option Gradient Boosting

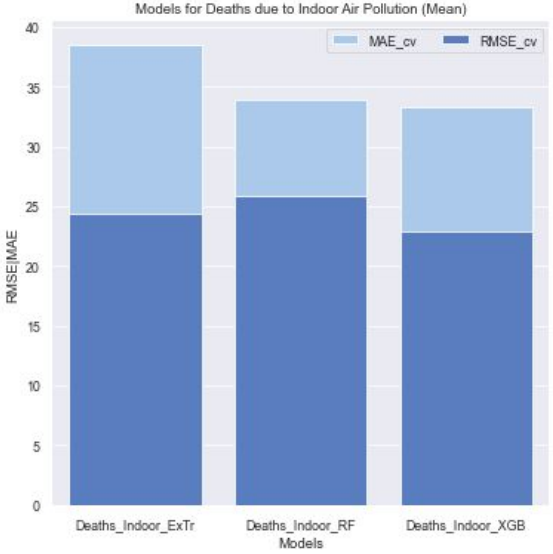


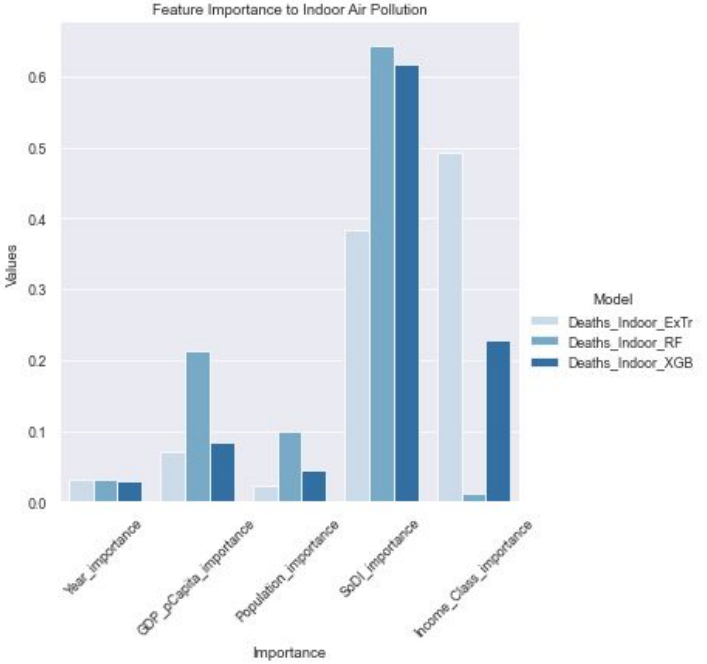


XGB, Random Forest and Extra Trees Models

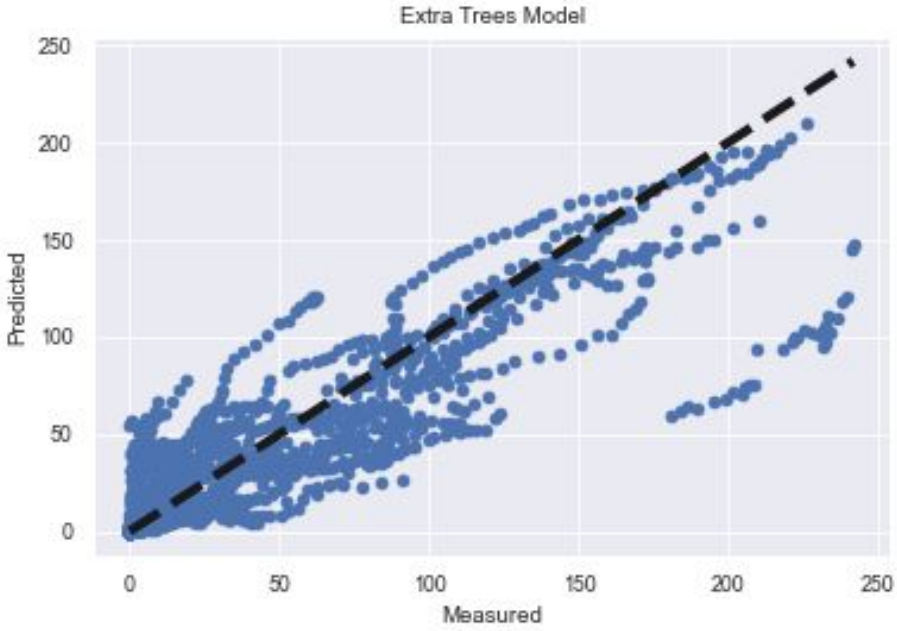
- Max depth = 7
- Number of estimators (trees) = 1000 to 2000
- 65 Train - 20 Test - 15 Hold-Out (attempt to preserve timeseries, but challenging)
- Cross-Validation with 10-kfolds
- Hold-out Validation
- MAE (less bias for outliers)
- Created ensemble model with voting of RF, ET, XGB

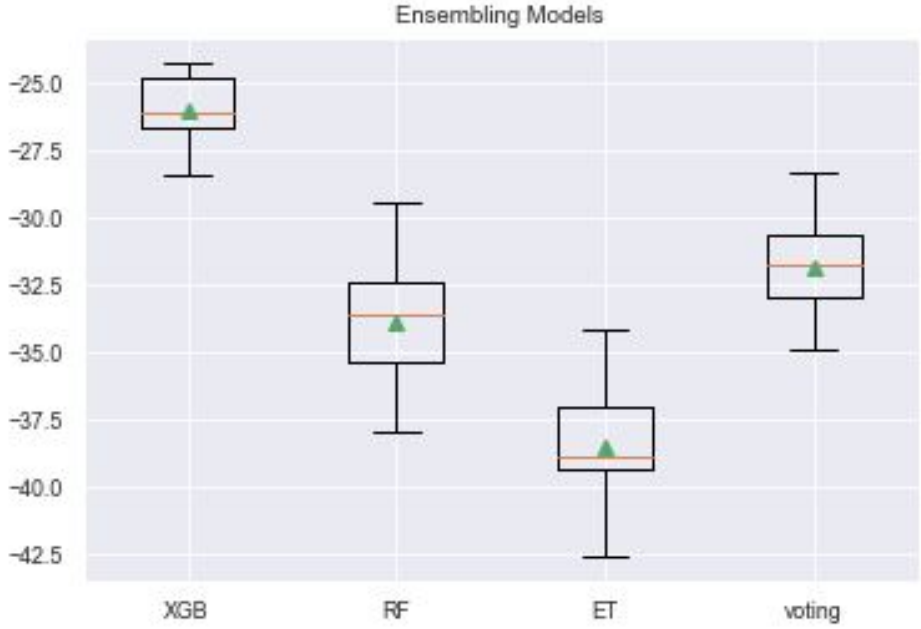






Ensemble Method





Business Applications

- Air pollution can be estimated and concerned organizations can focus on the local political policies to improve life quality in certain countries and regions
- This model can be further trained with population and migration data to assess migration patterns across the world
- This model can be further trained at city level in a country to focus on the industries that are responsible for the highest pollution rate.
- We gathered data on PM2.5 and OZone exposure that could be used to better model those deaths & gain insight.

Conclusions

- Predicting indoor pollution deaths (linearity)
- PM2.5 and OZone deaths are more challenging (non-linear)
- Air pollution deaths are highly dependent on Socio-Demographic Index & moderately dependent on GDP/Capita & Population
- LR is useful for predicting general trends per nation (preserved timeseries) vs. Tree-based Regression (predictions on socio-demographic/economic factors).
- **Future works:**
  - Integrate PM2.5/OZone exposure, and other socio-economic predictors
  - Explore climate-related risks and historical temperature data
  - Look at historical spending on environmental protection
  - Polynomial regressions

References

[1] Household air pollution and health, https://www.who.int/news-room/fact-sheets/detail/household-air-pollution-and-health, Data accessed: October 18, 2022

[2] Ambient (outdoor) air pollution, https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health, Data accessed: October 18, 2022

[3] Hygiene, https://ourworldindata.org/hygiene, Data accessed: October 18, 2022