LEAL Carbon Machine Learning Engineer Case Study Exercise

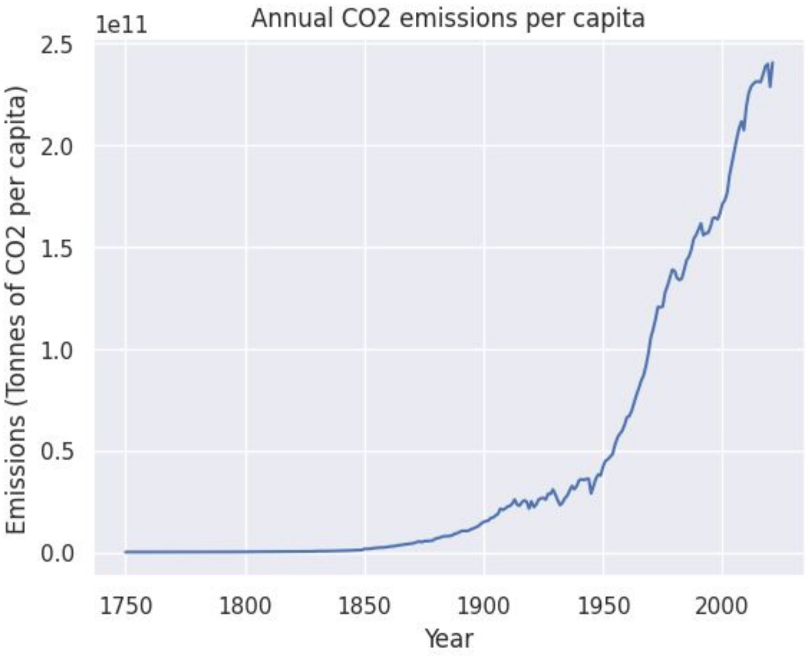
Dalia Andrea Rodriguez
August 10, 2023

Table of Contents

Introduction

Greenhouse gas emissions

- Gases in the atmosphere trap heat from the Sun to keep our planet warm enough to support life, a process known as the greenhouse effect[1].
- In the last 150 years, human activity has corrupted this naturally-occurring phenomenon.



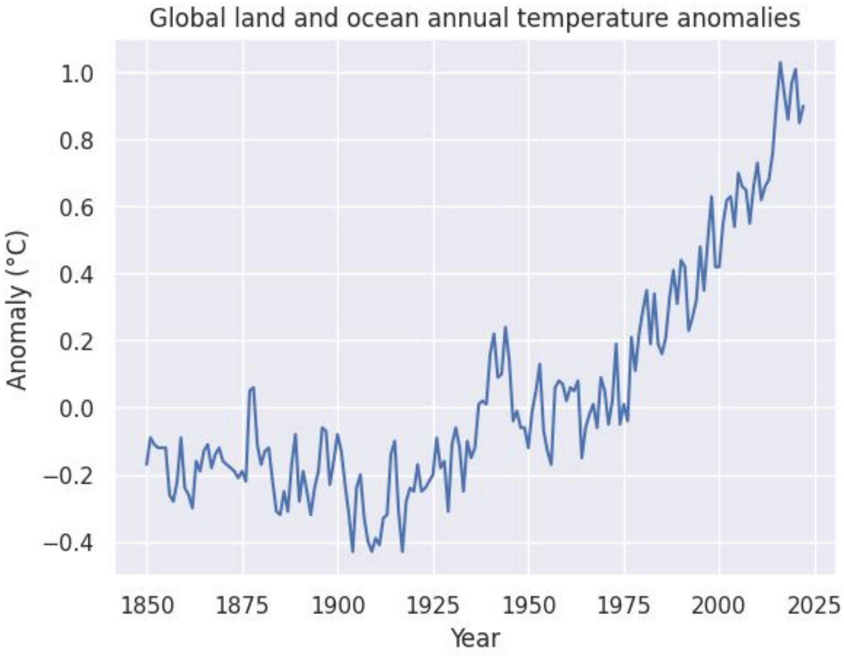
[1] https://www.epa.gov/ghgemissions/global-greenhouse-gas-emissions-data

Global warming

Abnormally large concentrations of

- carbon dioxide (CO_2) from fossil fuel burning for transportation and electricity;
- methane (CH_4) and nitrous oxide (N_2O) from agricultural activities and waste;
- Fluorinated gases (F-gases) from industrial processes and refrigeration.

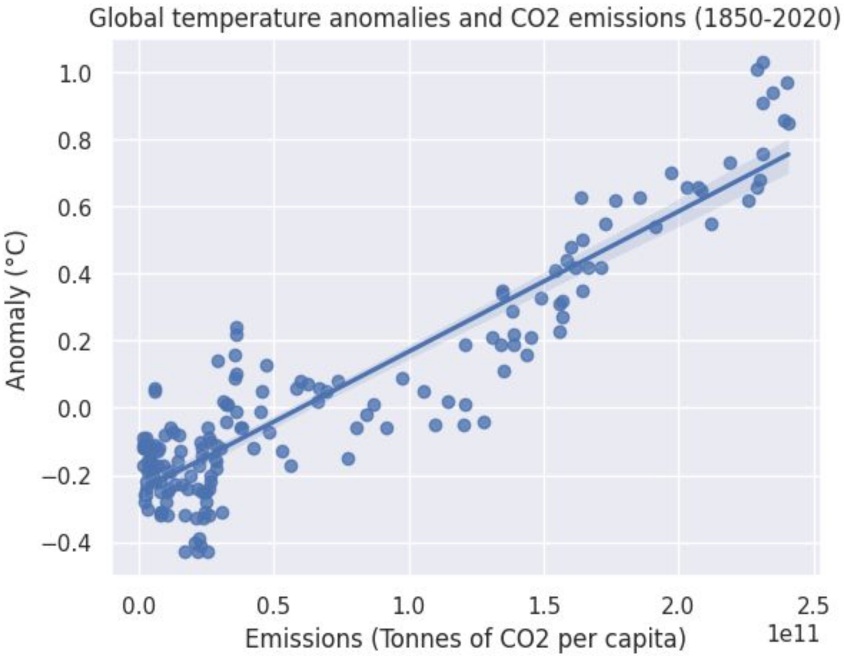
are being trapped in the atmosphere, warming up the Earth to unsustainable temperatures [1].



[1] https://www.epa.gov/ghgemissions/global-greenhouse-gas-emissions-data

Consequences of climate change

- Since 1850, global temperatures have increased by 1.1 °C[2].
- This minuscule change in temperature has led to more extreme weather events: flooding, droughts, melting ice caps, a rise in sea levels, and changes in habitat ranges for plants and animals[3].
- July 2023 was the hottest month ever recorded[4].



[2] https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/global/time-series/globe/land_ocean/ytd/12/1850-2022
[3] https://www.climate.gov/news-features/understanding-climate/climate-change-global-temperature
[4] https://public.wmo.int/en/media/news/copernicus-confirms-july-2023-was-hottest-month-ever-recorded/

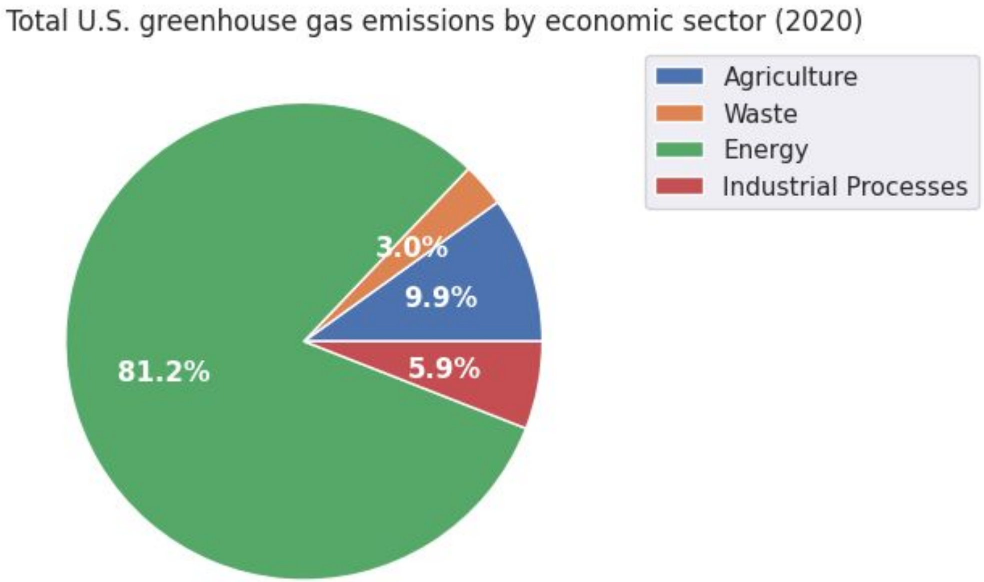
The immediate future

- The Paris Agreement, signed by 175 countries, seeks to maintain global warming under 1.5 °C[5].
- Despite efforts to stabilize temperatures, it is estimated that temperatures will rise beyond 1.5 °C in the next one to five years[6].
- Higher temperatures mean we will see harsher and more frequent severe extreme weather events.

[5] https://unfccc.int/process-and-meetings/the-paris-agreement
[6] https://unfccc.int/process-and-meetings/the-paris-agreement

Sources of Greenhouse Gas Emissions

Emissions by economic sector



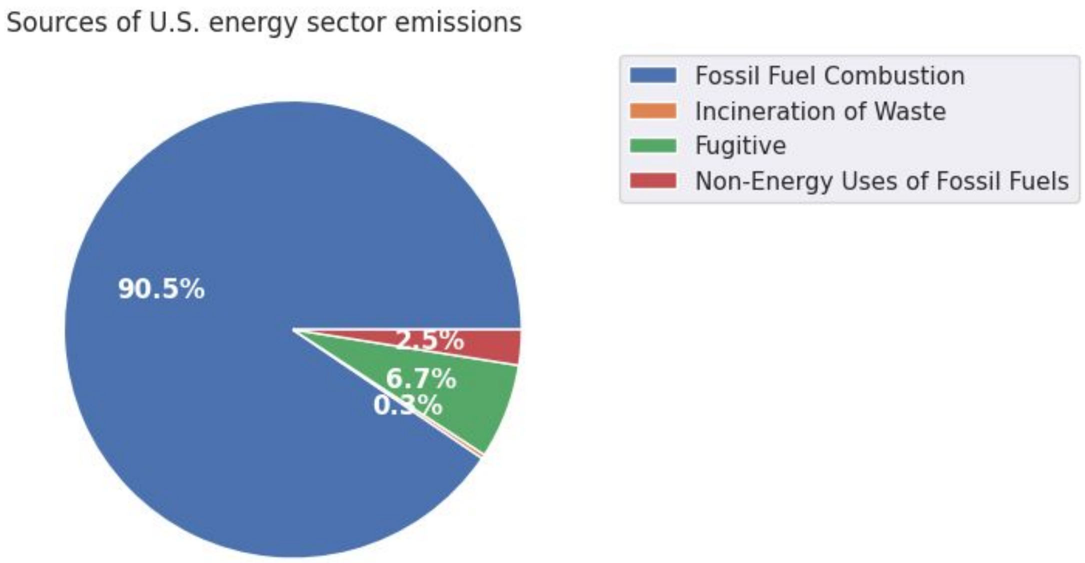
(*)

The U.S. is the largest contributor of CO_2 emissions in the world[7],

(*) The Industrial Processes sector emissions percentage does not include energy-related emissions

[1] https://www.epa.gov/ghgemissions/global-greenhouse-gas-emissions-data
[7] https://ourworldindata.org/contributed-most-global-co2

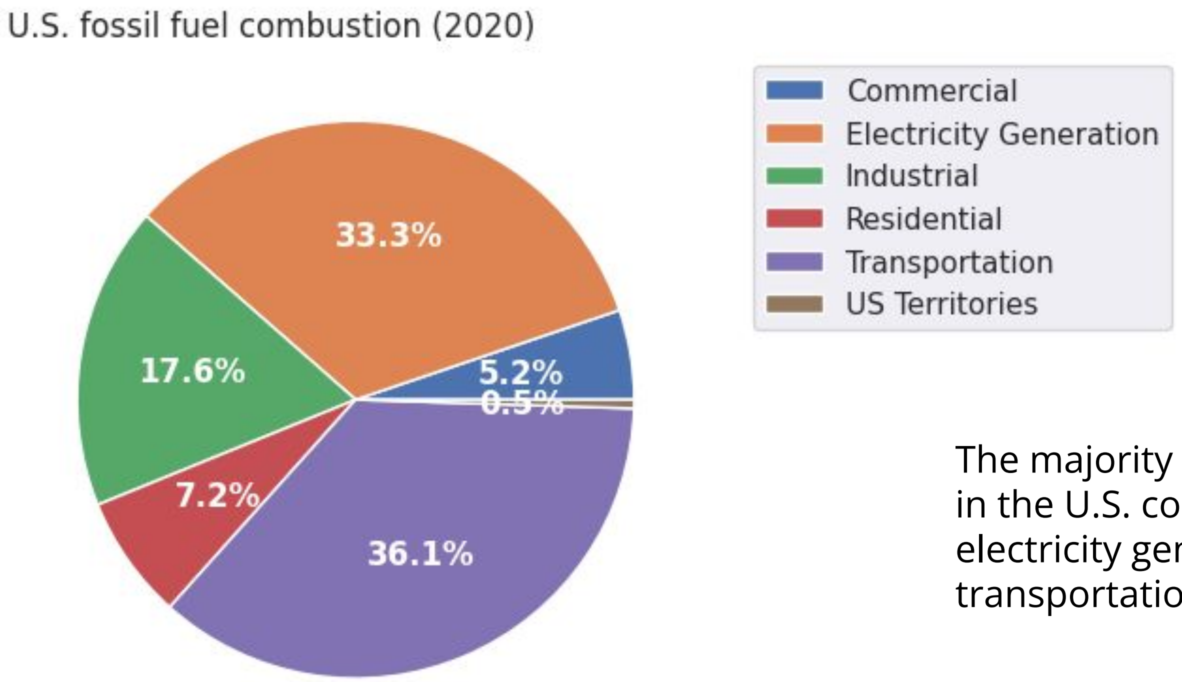
Emissions from energy



Most energy-based emissions in the U.S. are from the burning of fossil fuels like gasoline[1].

[1] https://www.epa.gov/ghgemissions/global-greenhouse-gas-emissions-data

Emissions from fossil fuel combustion



The majority of fossil fuel burning in the U.S. comes as a result of electricity generation and transportation.

[1] https://www.epa.gov/ghgemissions/global-greenhouse-gas-emissions-data

Kaya identity

- The Kaya identity is a framework that decomposes greenhouse gas emissions for a given population into the product of its size, GDP, energy intensity, and carbon intensity.

$$\text{total } CO_2 \text{emissions} = \text{population} * \text{GDP} * \text{energy intensity} * \text{carbon intensity}$$

- Energy intensity: energy consumption per unit of GDP.
- Carbon intensity: CO2 emissions from each unit of energy.

[9] https://www.sciencedirect.com/science/article/abs/pii/0196890495000259
[10] https://ourworldindata.org/emissions-drivers

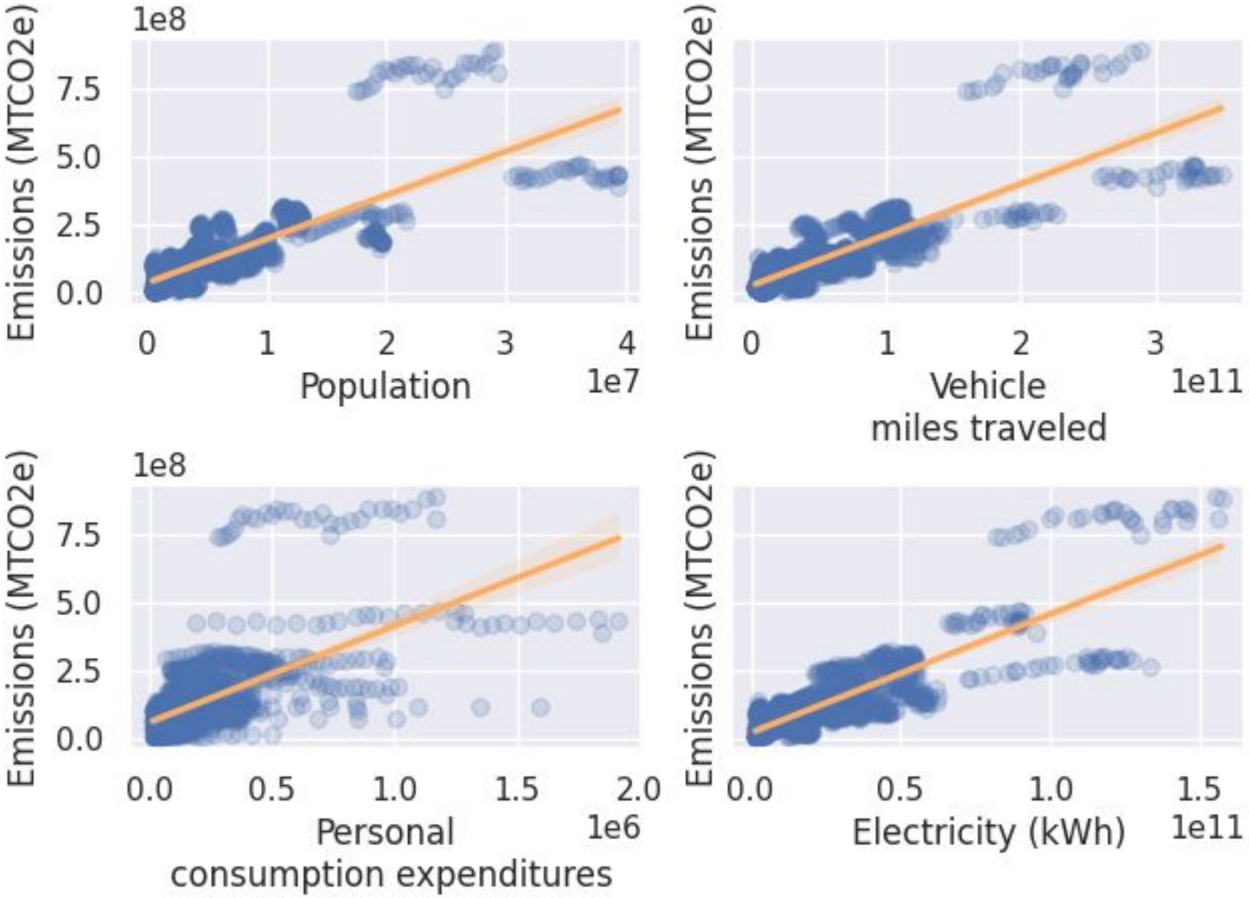
Kaya identity

$$\text{total } CO_2 \text{emissions} = \text{population} * \frac{\text{GDP}}{\text{population}} * \frac{\text{energy}}{\text{GDP}} * \frac{CO_2}{\text{energy}}$$

- The more money a population has, the more goods and services they have access to.
- Consumption requires energy for manufacture and transport of goods and services.
- Energy and transportation are powered by burning fossil fuels, the largest sources of greenhouse gases.

[9] https://www.sciencedirect.com/science/article/abs/pii/0196890495000259
[10] https://ourworldindata.org/emissions-drivers

Key drivers of CO2 emissions in the U.S.



[1] https://www.epa.gov/ghgemissions/global-greenhouse-gas-emissions-data

Food

- A kilogram of beef releases 60 kilograms of CO2 equivalents (kg CO2e) into the atmosphere per kilogram of meat[11].
- A kilogram of poultry releases 6 kg CO2e into the atmosphere per kilogram of meat[11].
- Wheat, corn, tomatoes, soy milk, and other vegetable products have less emit 2 kg CO2e or less into the atmosphere per kg of food[11].

[11] https://ourworldindata.org/food-choice-vs-eating-local

Mitigating Climate Change

What can we do?

We may be able to prevent the most catastrophic effects of global warming if we take measures to reduce our carbon footprint[8].

- Consumption and lifestyle habits influence individual carbon footprints.
- Driving less, recycling, eating less meat, are all ways to offset our carbon footprint.
- **An AI-powered carbon footprint calculator** may facilitate reduction of personal greenhouse gas emissions by providing us with data-driven insight about the sources of our emissions[12][13].

[8] https://climate.nasa.gov/faq/16/is-it-too-late-to-prevent-climate-change/
[12] https://coolclimate.berkeley.edu/calculator
[13] https://www.sciencedirect.com/science/article/pii/S0959652620304431

LEAL Carbon Case Study Exercise: Machine Learning Engineer Role

Overview

- In this section, I will provide insight on how I built a personal carbon footprint estimation machine learning model for LEAL Carbon.
- The section is organized as follows:
  - a discussion of related works that inspired my model;
  - methodology for creating an appropriate dataset for the task at hand;
  - model selection;
  - experiments;
  - results;
  - discussion and limitations;
  - future work.

Related work

- Existing literature on personal carbon footprint calculators is limited[13].
- I could not find any machine learning publications on personal carbon footprint estimation.
- Birnik developed an evidence-based set of principles for estimating personalized carbon footprints[14]:
  - Estimate at the minimum emissions relating to CO2, CH4, and N2O.
  - Allocate income and household size into the equation.
  - Allow users to model household consumption in detail (electricity, gas, rent, furniture,...) and by household size.
  - Let users model their food, transportation, and consumption habits in detail (dietary choices, miles of travel, clothing, entertainment services, etc.).
- Anthony et al.'s CarbonTracker estimates the carbon footprint of deep learning models through handcrafted features engineering and a simple linear model[15].

[13] https://www.sciencedirect.com/science/article/pii/S0959652620304431
[14] https://www.sciencedirect.com/science/article/abs/pii/S1750583613002168
[15] https://arxiv.org/abs/2007.03051

Dataset: co2variables

- Using Birnik's[14] principles as guidance, I curated the dataset **co2variables**:
  - 1530 rows and 32 columns.
  - Annual state-level data about emissions, income, energy use, and consumption habits related to housing, transportation, food, entertainment, and other goods and services for the years 1991-2020.
  - Data was downloaded by hand and by using an API.
  - Sources include the U.S. Department of Transportation and U.S. Census Bureau[16-21].
- Let *s* be a state of population size *N* in your dataset of states S*.* The personal carbon footprint *f(r)* of a resident *r* in s is approximately equal to the carbon footprint *f(s)/N.*

[14] https://www.sciencedirect.com/science/article/abs/pii/S1750583613002168
[16] https://www.fhwa.dot.gov/policyinformation/statistics/2020/
[17] https://www.census.gov/data/tables/time-series/demo/popest/intercensal-2000-2010-state.html
[18] https://www.bea.gov/data/consumer-spending/state
[19] https://www.epa.gov/ghgemissions/state-ghg-emissions-and-removals
[20] https://www.eia.gov/electricity/data/state/
[21]https://www.eia.gov/opendata/browser/natural-gas/sum/lsum

Model

- Anthony et al.'s CarbonTracker estimates the carbon footprint of deep learning models through handcrafted features engineering and a simple linear model[14].
  - Training deep learning models is energy-intensive.
  - Linear models require little computational power and are easy to interpret.
- Inspired by Anthony et al.'s work, I trained three different linear models for the task of carbon footprint estimation:
  - **Linear Regression**;
  - **Ridge Regression**: Linear Regression with L2-regularization;
  - **Huber Regression**: L2-regularized linear Regression model that is robust to outliers.

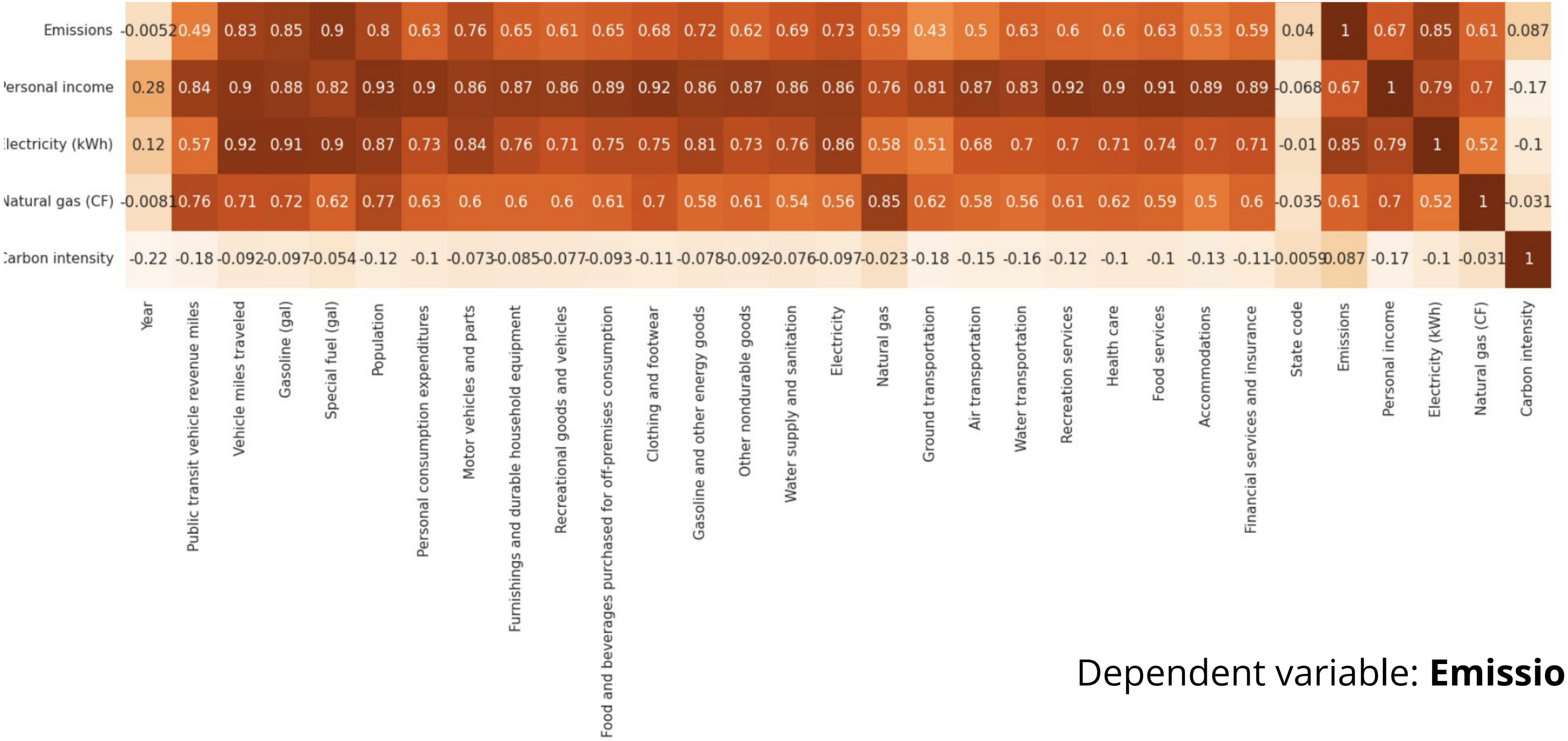
[15] https://arxiv.org/abs/2007.03051
[22]https://scikit-learn.org/

Experiments

- **Hyperparameter tuning**:
  - ‘alpha’ on Ridge and Huber;
  - ‘solver’ on Ridge.
- **Performance metrics**:
  - **R-squared**: in [0,1]; describes how well your Regression line approximates the data[23].
  - **Mean absolute error** (MAE): in [0,∞); penalizes outliers less.
  - **Mean absolute percentage error** (MAPE): in [0,∞); how far your predicted value falls from the real value in terms of percentages.
- Features were standardized by removing the mean and scaling to unit variance.
  - I experimented training the models with different feature sets.
- Experiments were performed on Jupyter Notebook using the sklearn library on Python.

[23] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8279135/

Correlation matrix of columns in co2variables

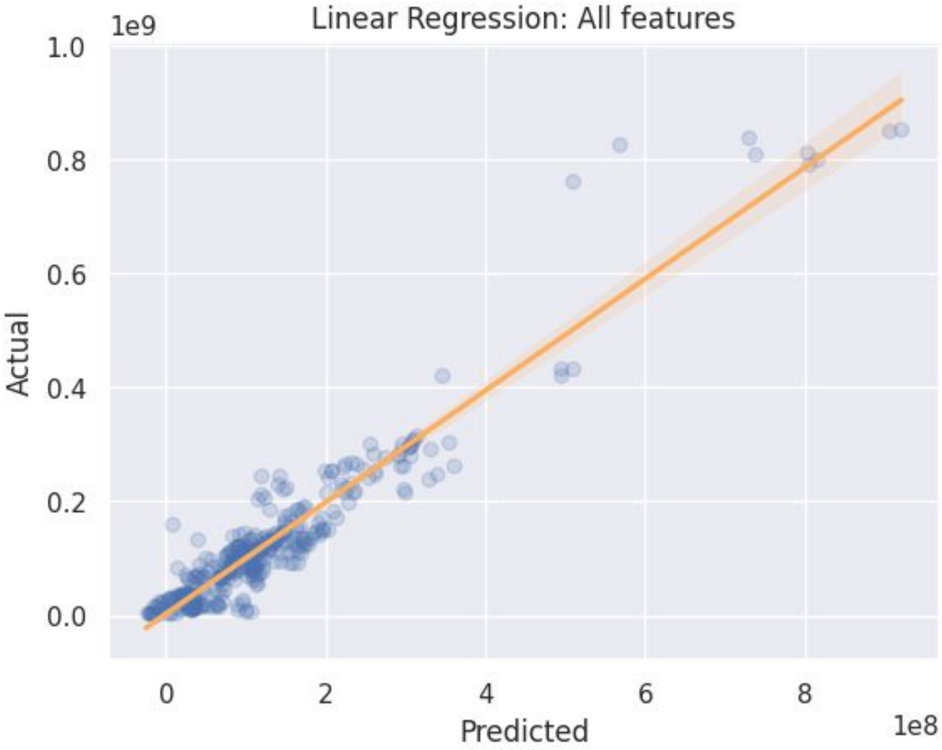


Dependent variable: **Emissions**

Results

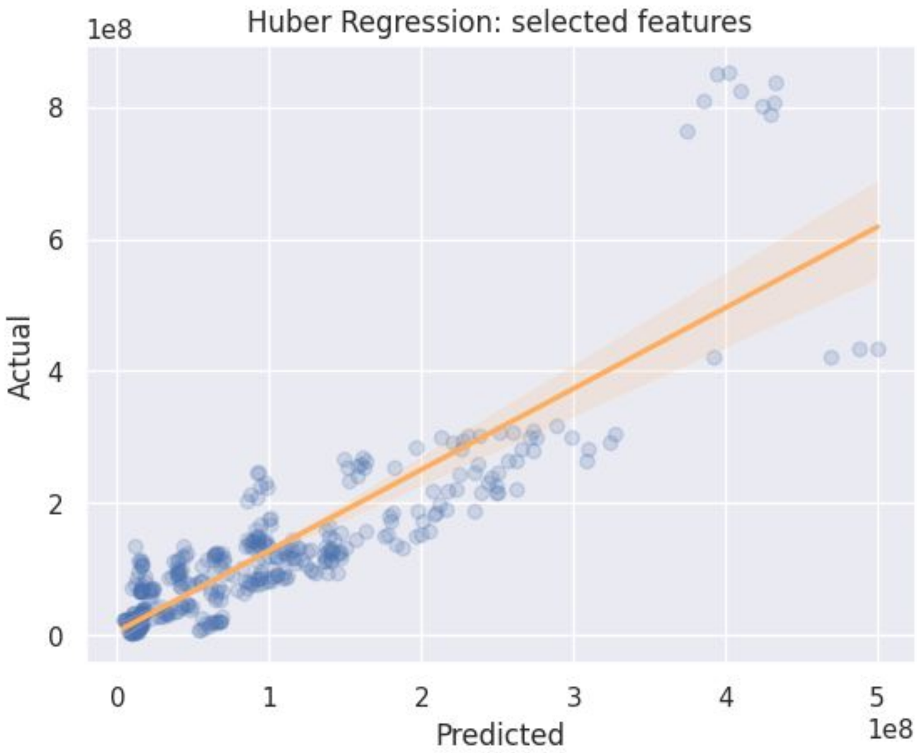
Metric	Linear Regression	Ridge Regression	Huber Regression
R-squared	91.63	88.70	68.97
MAE	26802936.34	31375611.37	41154502.89
MAPE	66.08	52.90	49.32

Actual vs predicted: Linear Regression



Actual vs predicted: Ridge Regression

Actual vs predicted: Huber Regression



Discussion and limitations

- **Discussion**:
  - Results are similar across feature sets and models.
  - High R-squared scores indicate that the data seems to be a good fit for the model.
  - We need to improve the mean absolute percentage error.
  - Highly correlated feature set makes it hard to assess which features are most significant.
- **Limitations**:
  - There is limited literature on personal carbon footprint estimators.
  - State-level data is not freely available for every desired category. For example, food availability data is only available at the national level.
  - Time constraints.

Future work

- Mitigate collinearity in the feature set.
- Experiment with hybrid ensemble models that integrate rule-based and machine learning methods of carbon footprint estimation.
- Decompose and incorporate Kaya identity into model.
- Incorporate more information on waste and agricultural emissions to increase feature robustness.
- Explore different machine learning models for carbon footprint estimation.
    - Neural Ordinary Differential Equations[24], Random Forests, Decision Trees[25], and Gaussian Process Regression[26].

[24] https://arxiv.org/pdf/2201.02433.pdf
[25] https://www.sciencedirect.com/science/article/pii/S2352550922001737
[26] https://www.frontiersin.org/articles/10.3389/fenrg.2021.756311/full

Conclusion

Conclusion

- I presented the co2variables dataset and a linear model for personal carbon footprint estimation that aligns with **LEAL Carbon**'s mission to facilitate positive environmental change through data-driven insight on personal greenhouse gas emissions.
- Linear Regression model achieved an R-squared score of 91.63, indicating that our data is a good fit for the task at hand.
- For future work, I plan to expand our feature set and experiment with different machine learning models for carbon footprint analysis.

Climate change is happening, but we can mitigate its most severe consequences if we hold ourselves accountable for our carbon footprint and take measures to reduce it.