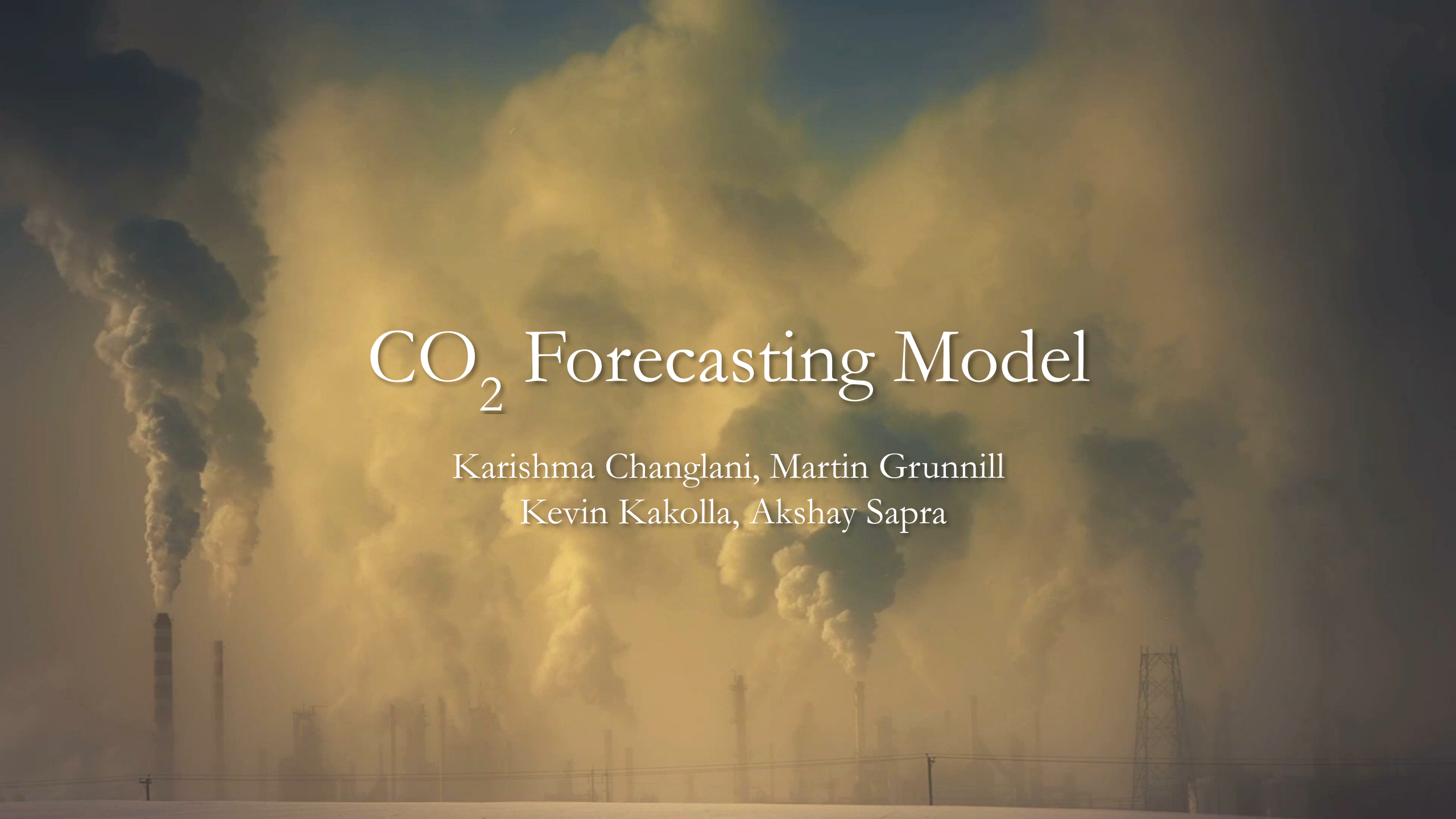


CO_2 Forecasting Model

Karishma Changlani, Martin Grunnill
Kevin Kakolla, Akshay Sapra

Problem Statement

- Climate change is ones of the biggest challenges of our time
- A lot of policy decisions are often based around cutting back CO_2
- If we can predict CO_2 using information like imports and exports we can mitigate matters.

Methodology-Data Sources

Sources (merged on year & iso code):

- Our World in Data:
  - CO2: https://github.com/owid/co2-data
  - Energy: https://github.com/owid/energy-data
  - Land use, Median Age & Military Spending: https://ourworldindata.org/
- World Bank Development Indexes:
  - Demographics, Imports & Exports: https://databank.worldbank.org/reports.aspx?source=world-development-indicators

Methodology-Features: Engineering

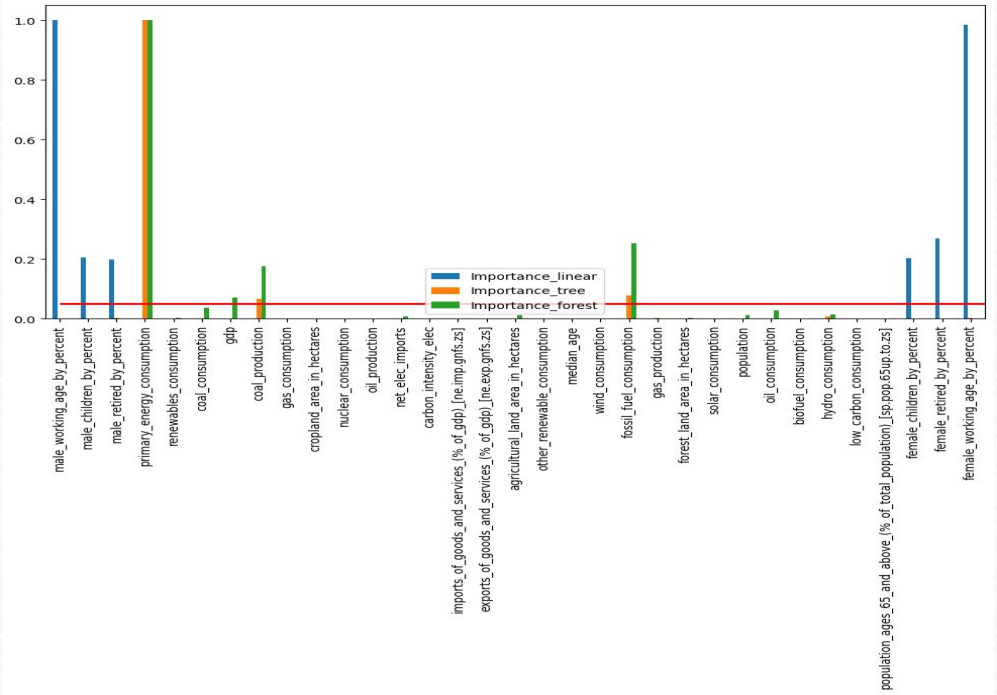
- % Male out of all males in 5 yearly age groups to:
  - % Male: Children (0-14)
  - % Female Working age Adults (15-64)
  - % Female Retired (>64)
- % Female out of all females in 5 yearly age groups to:
  - % Female Children (0-14)
  - % Female Working age Adults (15-64)
  - % Female Retired (>64)

Methodology-Features: Removing Data Leaks, Spurious

- Data leaks
  - CO_2 from particular sources or changes from previous years.
  - Whole Greenhouse Gasses (e.g. total ghg, ghg per capita).
  - Other Greenhouse Gasses (nitrous oxide and methane), often same industrial source .
- Overlapping energy consumption listing:
  - e.g. x_consumption, x_share, x_per_gdp, x_per_capita & x_electricity.
  - Removed but for x_consumption.
- Spurious
  - Year: why build a model to look at past when you can look it up.
  - Country:
    - Could be dependent on government policies that change.
    - What about a new country that broke away from an old one.

Methodology-Feature Selection

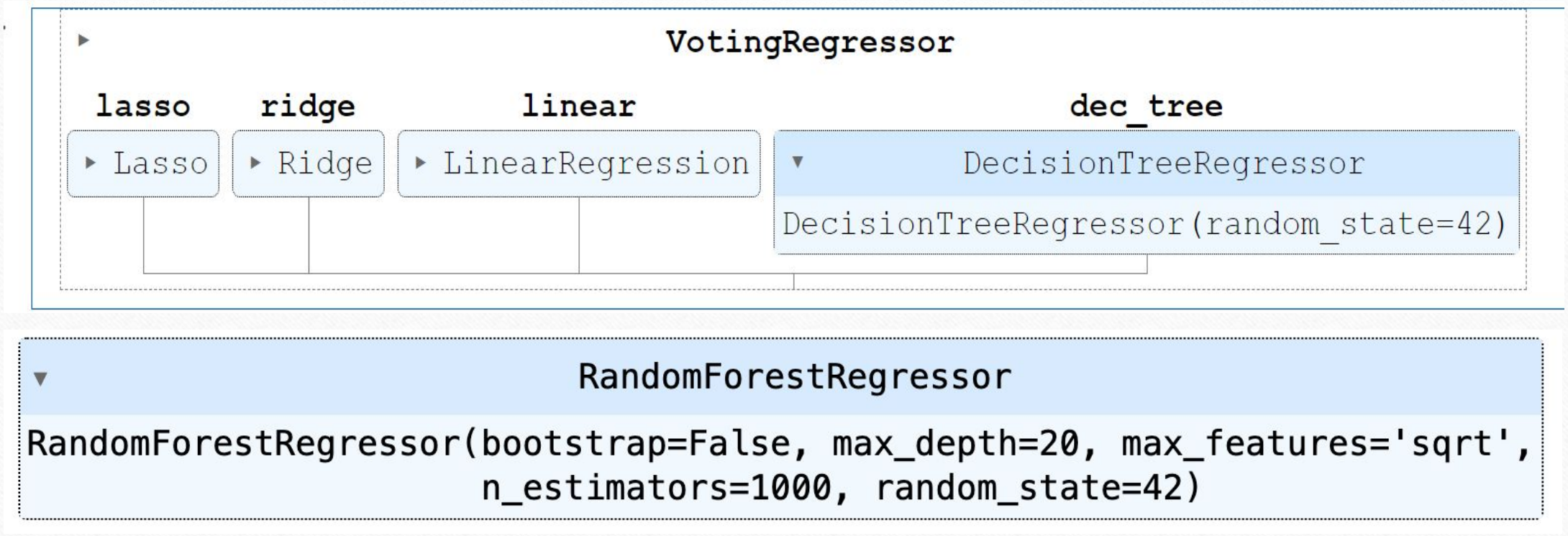
- Fitted linear, decision tree & random forest regressor(ion) models to training data (default hyper parameters).
- Obtained feature importance (or coefficients) for each model and scaled:
  - Converting to absolute value
  - Min-Max scaled
- Plotted and chose a cut off value to select a set of top 10-20 features.
  - 0.05
  - 10 features

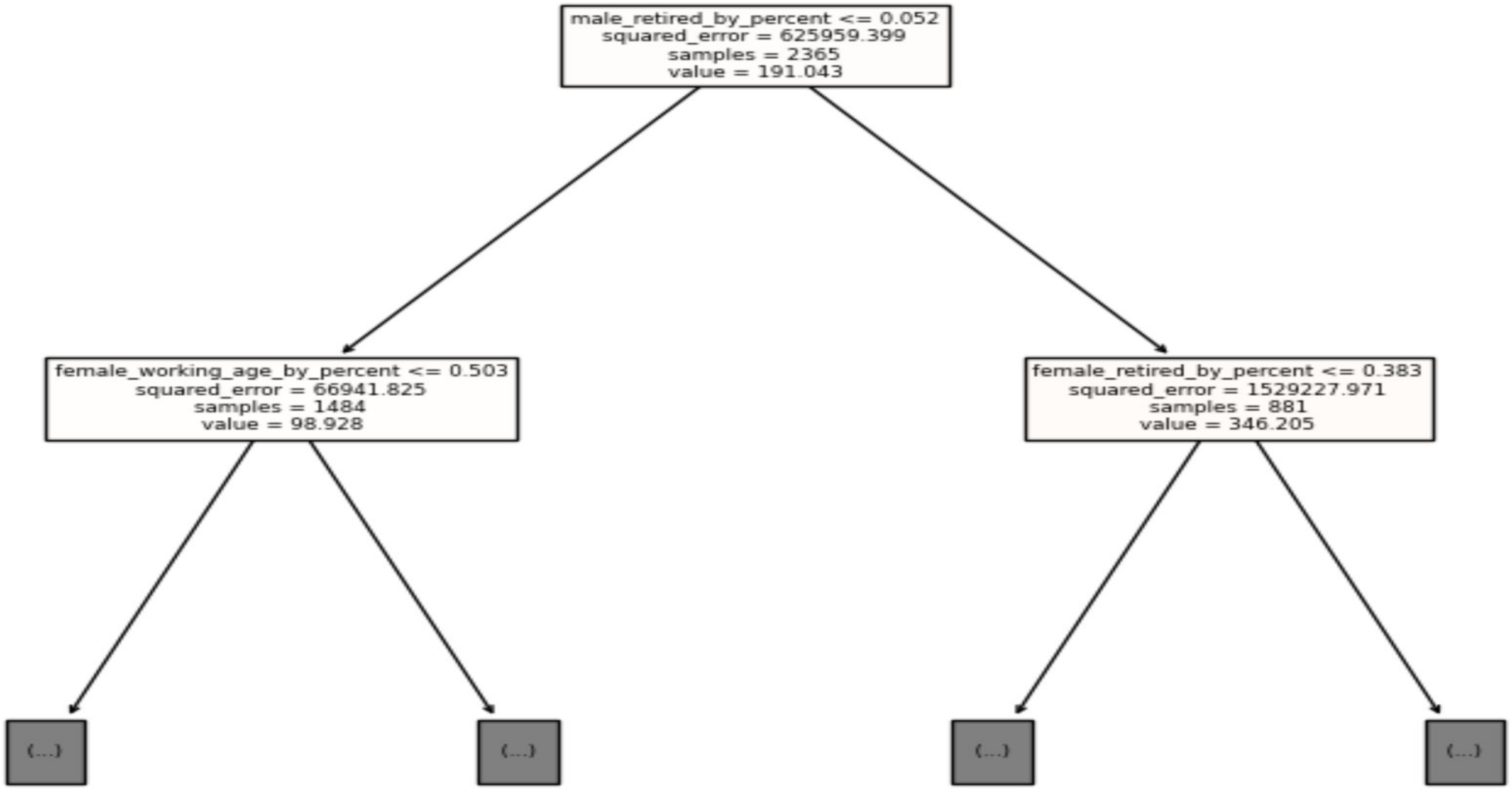


Methodology-
Hyper-Parameter Tuning (Cross Validation)

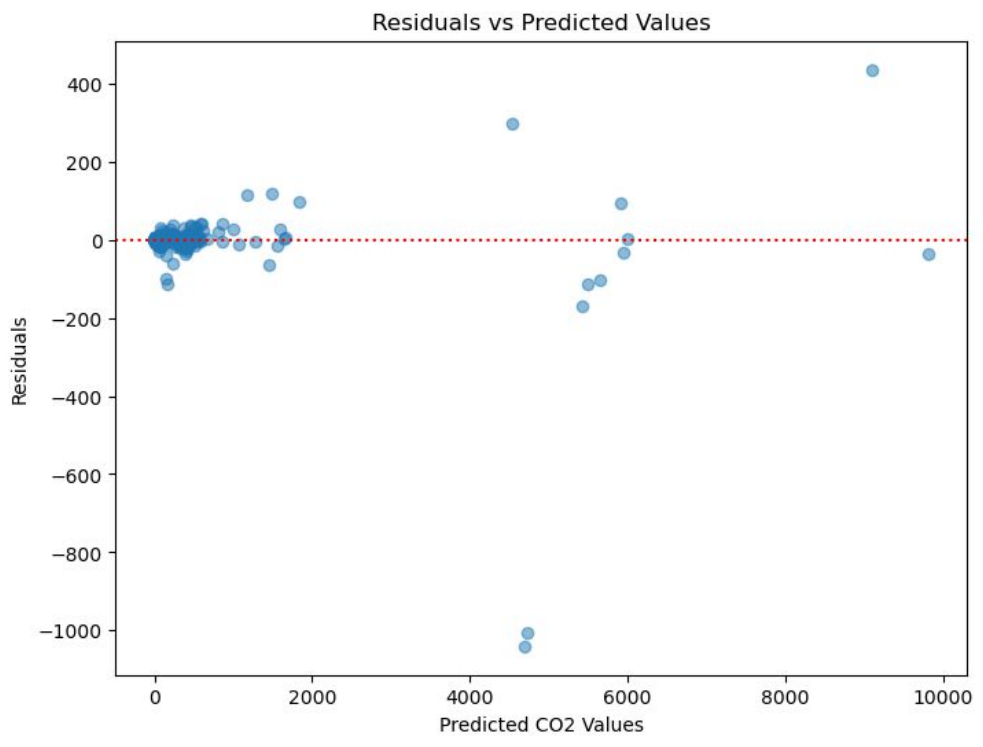
Model	**Grid used**	**Optimal value for hyper-parameters**
Linear Regression	Fit intercept True or False	Fit_intercept: True
Lasso Regression	Alpha: 0.1, 2; Max iterations: 100 2000 tolerance: 0.01, 0.0001	Alpha :0.1, max Iterations 2000 Tolerance 0.0001
Ridge Regression	Alpha: 0.1, 2; Max iterations: 100 2000 tolerance: 0.01, 0.0001	Alpha :0.1, max Iterations 100 Tolerance 0.01
Decision Tree	Max Depth: 10, 20, 50, 100; Max Features: sqrt, log2 Min Samples Split 2, 5, 10; Min samples leaf: 1, 2, 4	Max Depth: 50; Max Features: sqrt Min Samples Split 2; Min samples leaf: 1
Random Forest	Max Depth: 10, 20, 50, 100; Max Features: sqrt, log2 Min Samples Split 2, 5, 10; Min samples leaf: 1, 2, 4 Bootstrap True/False;	Max Depth: 20; Max Features: sqrt Min Samples Split: 1; Min samples leaf: 2

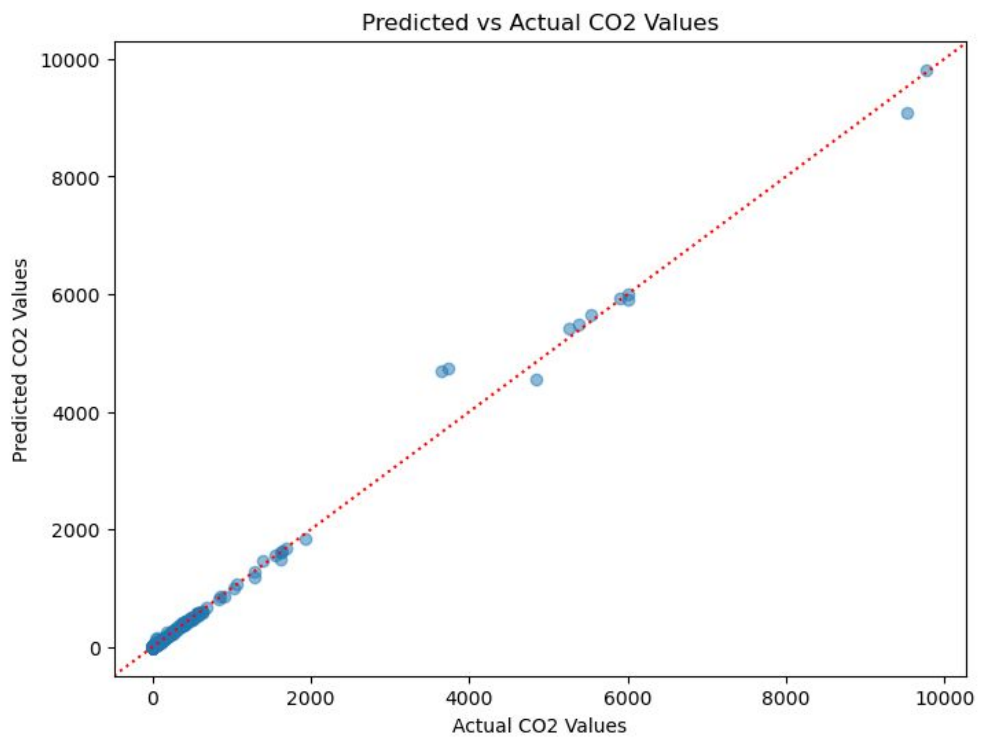
Methodology - Ensemble Model




male_retired_by_percent <= 0.052
squared_error = 625959.399
samples = 2365
value = 191.043
female_working_age_by_percent <= 0.503
squared_error = 66941.825
samples = 1484
value = 98.928
female_retired_by_percent <= 0.383
squared_error = 1529227.971
samples = 881
value = 346.205
(...)
(...)
(...)
(...)

Results from Random Forest





Conclusion - Decision Tree Overfit Random Forest Better

Fit to Training data (80%)

Lasso	MAE: 26.7
Ridge	MAE: 26.9
Linear Regression	MAE: 26.9
DecisionTreeRegressor	MAE: 0.05
VotingRegressor	MAE: 20.0
RandomForestRegressor	MAE: 2.2

Fit to Test data (20%)

Lasso	MAE: 31.9
Ridge	MAE: 31.9
Linear Regression	MAE: 31.9
DecisionTreeRegressor	MAE: 18.1
VotingRegressor	MAE: 25.9
RandomForestRegressor	MAE: 9.7

Challenges/Lessons (potential solutions)

Challenges	Lessons (potential solutions)
Locating data sources, limited years 2000-2018	Merging data from different sources.
Limiting data leakage with so many features.	This could be mitigated by learning about the data and its sources.
Differences between people running the same code.	Listing package requirements, google colab, conda environments or docker.