**ECOPREDICT: MODELING CO2 EMISSIONS FOR SUSTAINABLE SOLUTIONS**

**Objective:** The objective of this project is to analyze the predominant environmental pollutant, CO2, and explore its underlying causes and sources. Additionally, we aim to develop a robust machine-learning model to assist stakeholders in mitigating its impact. Our focus lies in creating a predictive model that empowers automobile designers to forecast CO2 emissions resulting from feature modifications. Furthermore, we aim to provide consumers with insights into the environmental consequences of their vehicle choices.

**Problem Definition and Data Selection:**

* Identify the greenhouse gas that will be the major cause of concern for global warming.

The **GWP** is a measure of how much energy the emissions of 1 ton of a gas will absorb over a given period relative to the emissions of 1 ton of carbon dioxide (CO2). The standard time frame used is 100 years.

A graph with different colored rectangles

Description automatically generated

Data Source: <https://www.ipcc.ch/assessment-report/ar4/>

Here's an overview of the GWP for each of the main greenhouse gases over 100 years, as defined by the Intergovernmental Panel on Climate Change (IPCC):

Carbon dioxide (CO2): GWP = 1 (baseline)

Methane (CH4): GWP = 25 (This means methane is 25 times more potent than CO2 at trapping heat in the atmosphere over 100 years.)

Nitrous oxide (N2O): GWP = 298 (Nitrous oxide is 298 times more effective than CO2 at trapping heat over 100 years.)

Fluorinated gases (F-gases): These vary widely in GWP, often being several thousand times more potent than CO2. For instance, some hydrofluorocarbons (HFCs) have GWPs as high as 14,800.

* The Global Warming Potentials (GWP) of Nitrous Oxide (N2O) and Methane (CH4) are significant concerns due to the very high heat-trapping capacity of these gases as observed above. However, the sheer volume of Carbon Dioxide (CO2) emissions annually, makes the impact of CO2 emissions the primary concern. Ideally, our focus should be on developing systems and methods to reduce and control CO2 emissions as much as possible. CO2 is a by-product of various human activities, such as fossil fuel generation and electricity production, as well as N2O emissions. There is ample evidence to suggest that efforts to reduce CO2 emissions will have the most significant impact on reducing other greenhouse gases as well.

Article: <https://news.mit.edu/2021/reducing-emissions-decarbonizing-industry-0721>

A graph of gas emissions

Description automatically generated “Carbon dioxide is the GHG emitted in the largest quantities. The 2.47 billion metric tons of CO2 reported for 2022 represent 91.9% of the GHGs reported in 2022a. Methane emissions represent 6.9% of reported 2021 GHG emissions, N2O represents 0.7%, and fluorinated gases (HFCs, PFCs, SF6, NF3, Other Fully Fluorinated GHGs, HFEs, Very Short Lived Compounds, Other) represent 0.5%”

Data Source: [GHGRP Emissions by GHG | US EPA](https://www.epa.gov/ghgreporting/ghgrp-emissions-ghg)

A graph showing a green and blue graph

Description automatically generated

* The CO2 Emissions percentage increased from 79.7% of total greenhouse gas emissions from 2021 to 91.9% in 2022.

Global Trend:

A graph showing the cost of carbon dioxide

Description automatically generated

A graph of a graph showing the number of countries/regions by average emissions per capita

Description automatically generated

A graph of the number of emitting countries/regions

Description automatically generated

Data Source: <https://worldpopulationreview.com/>

Among the top nations with the highest CO2 emissions, the USA has the highest per capita CO2 emissions.

**Data Preparation:**

We have now chosen two data sources with API access. We'll use these to analyze the primary contributors, sectors, and regions associated with CO2 emissions in the USA.

1. <https://www.eia.gov/opendata/browser/> - **The U.S. Energy Information Administration (EIA)** collects, analyzes, and disseminates independent and impartial energy information to promote sound policymaking, efficient markets, and public understanding of energy and its interaction with the economy and the environment.
2. <https://open.canada.ca/> - **Open Canada**, through its Open Government initiative, is dedicated to promoting transparency and accountability in government operations. The portal allows public access to a wide array of data and information that the Canadian government has committed to making open to the public.

* First, we will access the data which with **CO2 emissions from each state in the US, overall emissions for each year and emissions from each category or sector.**
* Then using the similar steps we used to create the visuals for the first data set, we will access the data containing **CO2 and NOx emissions for Electricity generation from different 'fuels' in each 'state' in the USA from '2005-2022'**
* **Next, we will develop a predictive model to predict the CO2 emissions for an automobile using its most important features.**

**Steps:**

We will be using the Google Collaboratory environment for the entire project:

Packages to import :

A screenshot of a computer

Description automatically generated

Accessing the data using an API request:  


**Extracting the JSON format data and then converting it to a pandas data frame :**

A screen shot of a computer code

Description automatically generated

**Data frame Information:**

A screenshot of a computer code

Description automatically generated

**Converting the ‘Value’ and ‘period’ columns to numeric for the sake of analysis**:  
A close-up of a computer code

Description automatically generated

**Separating the Overall US info rows and others rows for the sake of analysis:**

A screenshot of a computer code

Description automatically generated

**Data Modelling & Visualizations:**

A graph of different colored bars

Description automatically generated

A graph of carbon dioxide emissions

Description automatically generated

**Overall in the US, the Top Contributor to CO2 Emissions is the “Transportation Sector", followed closely by the "Electric Power Sector"**

**Top 7 states by CO2 Emissions:** ( value represents CO22 emissions)

A screenshot of a computer screen

Description automatically generated

A graph of the same number of the same number of the same number of the same number of the same number of the same number of the same number of the same number of the same number of

Description automatically generated

A graph of the same number of the same number of the same number of the same number of the same number of the same number of the same number of the same number of the same number of

Description automatically generated

**Top CO2 Emission sectors in the top 4 states :**

A graph of blue rectangles with text

Description automatically generatedA graph of blue rectangles

Description automatically generated

A graph of gas emissions

Description automatically generatedA graph of blue squares

Description automatically generated**­**

Transportation and Electric power sectors are the top CO2 emission sectors in the top 4 states as well.

A graph with blue lines

Description automatically generated

A graph of gas emissions

Description automatically generated

Electricity resulting from burning Coal contributes highest to the CO2 emissions in the Electric Power Sector.

**In the U.S. in 2022, natural gas was the largest source of electricity generation with 1,687,067 million kWh, but coal was the most carbon-intensive per unit of electricity generated. Despite generating less than half the electricity of natural gas at 831,512 million kWh, coal produced 2.30 pounds of CO2 per kWh, compared to natural gas, which produced 0.97 pounds of CO2 per kWh. This made coal more carbon intensive.**

Data Source: [US Electricity Profile 2022 - U.S. Energy Information Administration (EIA)](https://www.eia.gov/electricity/state/)

**Now, we will look at the transportation sector, especially automobiles that contribute to CO2 emissions and we will model a predictive model that will take in inputs such as vehicle Mileage, model, brand, Engine Size, number of cylinders, type of fuel used and will return a predicted output for the CO2 emissions.**

**Data Description:** [**https://natural-resources.canada.ca/energy-efficiency/transportation-alternative-fuels/personal-vehicles/choosing-right-vehicle/buying-electric-vehicle/understanding-the-tables/21383**](https://natural-resources.canada.ca/energy-efficiency/transportation-alternative-fuels/personal-vehicles/choosing-right-vehicle/buying-electric-vehicle/understanding-the-tables/21383)

A screenshot of a car model

Description automatically generated

A white sheet with black text

Description automatically generated

A screenshot of a graph

Description automatically generated

Let us look at each of the following visualizations separately:

A graph of a graph

Description automatically generated

CO2 Ratings are generally given for a vehicle based on its “CO2 Emissions Performance”.  
The vehicle's tailpipe emissions of carbon dioxide are rated on a scale from 1 (worst) to 10 (best)

A graph of a graph showing the cost of carbon emissions

Description automatically generated

A graph of a number of vehicles

Description automatically generated with medium confidence

Passenger Vans have the highest average CO2 emissions whereas Compact Cars have the lowest average emissions.

A graph of blue bars

Description automatically generated with medium confidence A graph of a graph

Description automatically generated

**Fuel Types:**

X = Regular gasoline;

Z = Premium gasoline;

D = Diesel;

E = E85;

N = Natural Gas

E85, a fuel type containing 85% ethanol, is renewable but emits higher CO2 due to energy-intensive production and lower energy density. Despite its renewable nature, combustion and production processes contribute to its higher CO2 emissions than gasoline.

A graph of a graph showing the number of emissions

Description automatically generated

A = Automatic; AM = Automated manual; AS = Automatic with select shift; AV = Continuously variable; M = Manual;

Number of gears/speeds (1–10)

Automatic Transmissions have the highest emissions while Automated-Manual transmissions have the lowest emissions on average.

**Hypothesis Tests:**

A screenshot of a computer code

Description automatically generated

A screenshot of a computer code

Description automatically generated

Next, we performed target encoding on the categorical variables and then visualized a co-relation matrix,

Co-relation Matrix:  
A screen shot of a chart

Description automatically generated

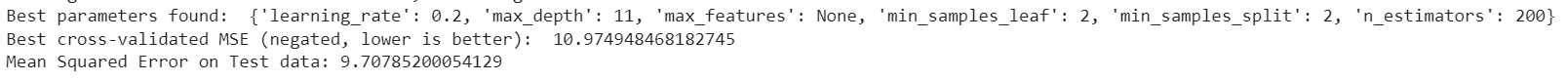
**Machine Learning Modelling:**

Here are the results for the Machine learning methods we tried:

A screenshot of a computer

Description automatically generated

1. After modeling various machine learning techniques, it was observed that the Random Forest regressor provided the least error-prone predictions on the test data. However, it exhibited signs of overfitting. Conversely, both Gradient Boosting and Bi-directional LSTM models demonstrated better generalization to the test data.
2. **When comparing Gradient Boosting and Bi-directional LSTM, although Gradient Boosting yielded slightly higher error rates, it was preferred due to its interpretability.**

Now, **Fine-tuning** the Gradient boost model to improve the performance:

**After the hyperparameter tuning the model's error rate on the test data set was reduced from 18.24 to 9.70, which is a significant improvement.**

**SHAP Visualizations:**A bar graph with text

Description automatically generated

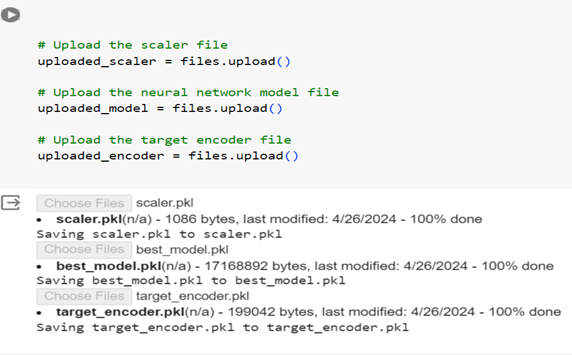
**The “Combined Mileage” ( in city and on highways) is the major contributing factor that is effecting the CO2 emission values in the gradient boost model prediction, followed by Engine Size of the Vehicle.**

The target encoder, scaler and the final fine-tuned model are saved and downloaded onto the pc.  
Now we create a predictor.ipynb python file which will load the said models and then run a function to return the prediction of CO2 emissions for the given input data.  
  
**Example Usage:**

1. Install the encoder package:  
   A close-up of a web page

   Description automatically generated
2. Import the required modules:  
   A white background with black text

   Description automatically generated
3. Upload the downloaded encoder, scaler and model files:



1. Load the files:  
   A computer code with text

   Description automatically generated with medium confidence
2. Enter the input values, Run the scoring function and get the output prediction as follows:  
   A screenshot of a computer

   Description automatically generated
3. Interpret the model prediction using SHAP values for that prediction.  
   A graph with text and numbers

   Description automatically generated

(**SHAP values** are a way to explain prediction models by quantifying the impact each feature has on the prediction. Each SHAP value is a number that represents how much a given feature contributes to pushing the model's prediction away from the baseline, or expected value, for that prediction,   
**Baseline (Expected Value):** This is the average prediction across all data points in the training set or another reference point determined by the model. It's what the model would predict if it didn't know any specific information about the data point in question)

This predictive model, as showcased in the provided interface, can be highly beneficial for vehicle manufacturers during the initial stages of ideation and design. Here's how manufacturers can leverage this model effectively:

**1. Early Design Adjustments:**

Optimization for Emissions: By inputting preliminary specifications into the model, designers can see the projected CO2 emissions per kilometre for a vehicle still in the conceptual or early design phase. This allows them to adjust engine size, fuel type, or other variables to meet specific emissions targets before finalizing the design.

**2. Regulatory Compliance:**

Meeting Emission Standards: Different markets have varying emissions regulations. Using the model, manufacturers can predict whether a new vehicle design will comply with these regulations in different regions and make necessary modifications to ensure compliance.

**3. Cost Efficiency:**

Reducing Prototype Costs: By predicting emissions outcomes based on the vehicle specifications, manufacturers can minimize the number of prototypes built and tested, thus reducing costs. Early virtual adjustments to the design based on the model’s feedback can eliminate the need for extensive physical modifications later.

**4. Target Market Optimization:**

Aligning with Market Demands: Understanding the emissions profile of a vehicle helps manufacturers tailor their products to the preferences of specific markets that prioritize low emissions, enhancing marketability.

**5. Integrated Development Process:**

Feedback Loop Integration: Incorporating this predictive model into the iterative design process ensures continuous feedback on emissions implications, allowing for agile responses to design outcomes. This can be part of a broader strategy to integrate simulation tools with engineering workflows.

**6. Sustainability Goals:**

Supporting Environmental Objectives: By using the model to predict and minimize CO2 emissions, manufacturers contribute to broader sustainability goals, potentially lowering the environmental impact of new vehicles.

**How to Implement:**

Integration into CAD Tools: Manufacturers can integrate this model directly into the software tools used by designers (such as CAD systems), allowing for real-time emissions predictions as different design parameters are adjusted.

Training and Protocols: Educate design and engineering teams on utilizing the model effectively, establishing standard operating procedures for assessing every new vehicle design concerning predicted emissions.

By embedding this predictive model into the early stages of vehicle design and development, manufacturers not only enhance their capacity to produce environmentally friendly vehicles but also streamline their design processes and align closely with global trends toward emission reductions.