Greenhouse Gas Emission Prediction using Industry & Commodity Supply Chain Data (2010–2016)

Presented By: Sitesh Virju Gupta
Organization: Edunet Foundation × Shell AI/ML Internship Program
Duration :4 week
AICTE Student ID:STU67e55b65b2b2be1743084389

Learning Objectives

What I Set Out to Learn Through This Project:

- Understand how machine learning models can be used to **predict greenhouse gas emissions** based on supply chain data.
- Learn how to **preprocess real-world environmental datasets** cleaning, handling missing values, and preparing for modeling.
- Train and evaluate a **regression model using scikit-learn**, and understand how performance metrics like R^2 and RMSE reflect accuracy.
- Explore the role of **data quality indicators** (like temporal, geographical, and technological correlation) in model performance.
- Get comfortable with key data tools like **Pandas, Seaborn, and Matplotlib** for exploratory analysis.
- Build a **user-friendly web interface using Flask** to make the model accessible and interactive.
- Most importantly, apply what I've learned in a way that supports **sustainable decision-making and real-world impact**.

Source : www.freepik.com/

Tools and Technology used:

Languages & Libraries:

- **Python 3.10** – The core language behind both the model and the web app.
- **pandas & openpyxl** – For reading Excel files and cleaning/manipulating real-world data.
- **scikit-learn** – To handle data preprocessing, train the regression model, and evaluate results.
- **joblib** – For saving/loading the trained model and scaler efficiently without retraining.

Web Development:

- **Flask** – To build a lightweight, easy-to-use web interface for user input and model output.

Data Analysis & Visualization:

- **Matplotlib & Seaborn** – To explore patterns, trends, and correlations within the dataset.

Development & Deployment:

- **Visual Studio Code** – For exploratory analysis and model development.
- **GitHub** – To track changes and share the entire project repo.

Methodology:

Data Collection & Understanding:

- Started with a government-published dataset on U.S. industry and commodity-based GHG emissions (2010–2016).
- Focused on key features like **Substance**, **Source**, **Supply Chain Industry**, and multiple data Quality(DQ indicator.

Data Cleaning & Preprocessing:

- Handled missing values and outliers.
- Encoded categorical variables [**Substance, Unit, Source**].
- Scaled features using **Standard Scaler.**
- Selected only the most relevant input for modelling.

Exploratory Data Analysis (EDA):

- Visualized trends and patterns in emission factors.
- Checked feature correlations to guide model building.

Model Building & Evaluation:

- Trained a linear **Regression model** using **sckit-learn.**
- Evaluated using **R^2 Score ,MSE,RMSE.**
- Saved the model with **joblib** for deployment.
- Built a simple ,responsive **Flask Interface.**

Problem Statement:

What Was the Challenge?

In today's world, industries are under pressure to reduce their carbon footprint — but many companies still struggle to **measure or predict their greenhouse gas (GHG) emissions**, especially across complex supply chains. The challenge was simple but impactful.

We had access to a rich dataset of **U.S. supply chain emissions (2010-2016)**,categorized by:

- **Industries & commodities**
- **Emission Sources**
- **Dats Quality Indicator** (like reliability, geography, time, and technology relevance).

The goal was to turn this into something useful:

- A working prediction model.
- A web app anyone could use to estimate emissions from key input values.

Solution:

To Make GHG Prediction simple and usable, I build complete end to end solution.

Step 1:Cleaned & Prepared the Data

- Removed missing values, fixed inconsistencies.
- Selected the most useful features.
- Scaled and encoded the data for machine learning.

Step 2:Trained a Regression Model

- Used **Linear Regression** to estimate GHG emissions based on input features like supply chain, margin, and reliability metrics.
- Evaluated it using **R^2 score**, **MSE,** and **RMSE** to ensure accuracy.

Step 3:Build a Web App:

- Designed a simple **Flask Interface** where users can enter values like Substance. Source, and DQ score.
- The app instantly shows the **predicted emissions.**

Exploratory Data Analysis (EDA):

What the Data Told Me Before Modeling

1.Target Variable Distribution:

- Shows most emission values are highly skewed toward the lower end.
- Most industries emit low amount but a few outliers have very high emissions.

2.Distribution of Sources(Industry vs Commodity):

- Confirms data is balanced the three main substances.
- Emissions were recorded almost equally from industry and commodity sources.

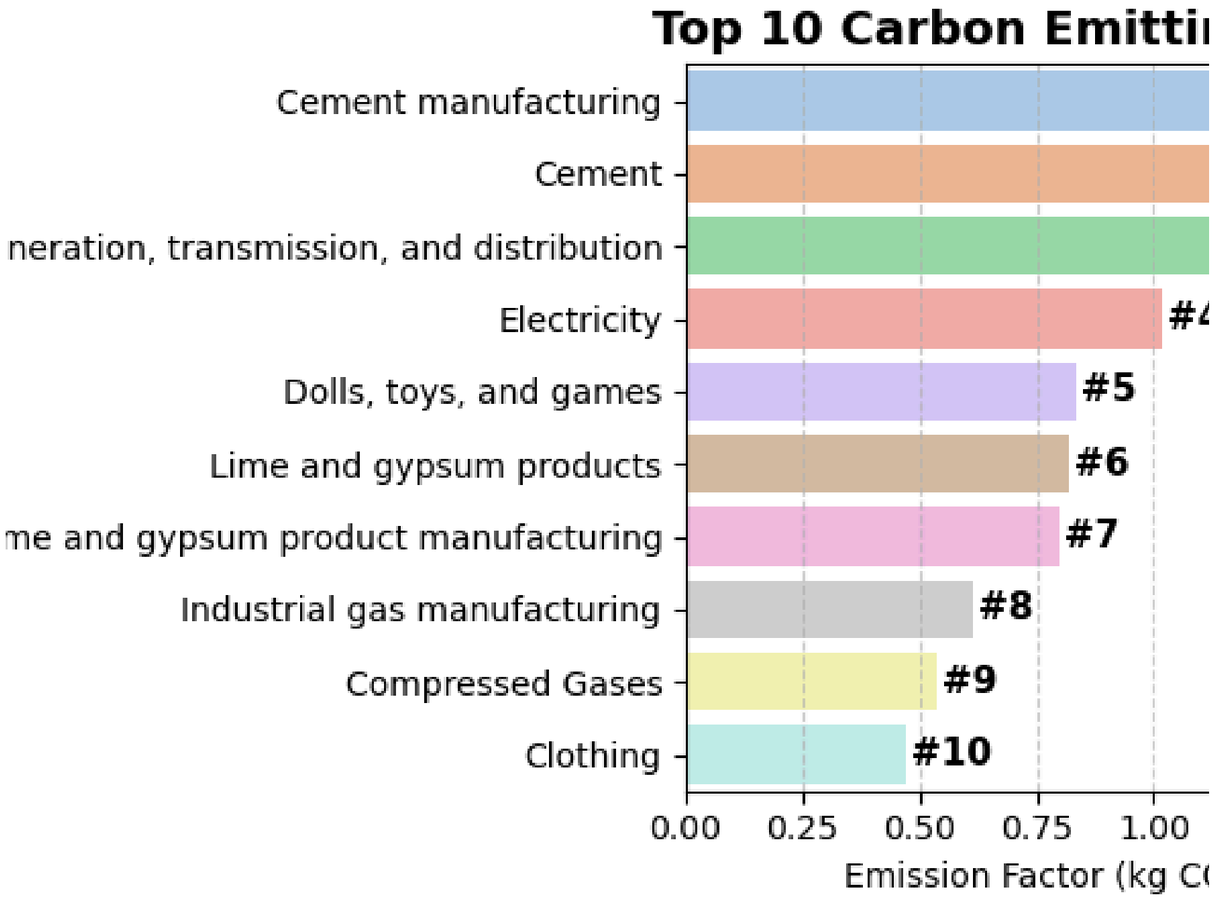
3.Distribution of Substances:

- Even spread among the three main substances.
- No dominants bias-all key GHG substances were well represented.

Top 10 Carbon Emitting Industries

Graphs Shows:

- **Cement Manufacturing** had the Highest emission factor, followed by **power generation** and **electricity.**
- Many of the **top emitters are heavy industries** energy-intensive sectors with large-scale operation.
- Interesting to see **non-obvious contributors** like toys and game also appear showing that emissions aren't just a big industry problems.



Correlation Heatmap of Numerical Features

What Affects Emissions the Most?

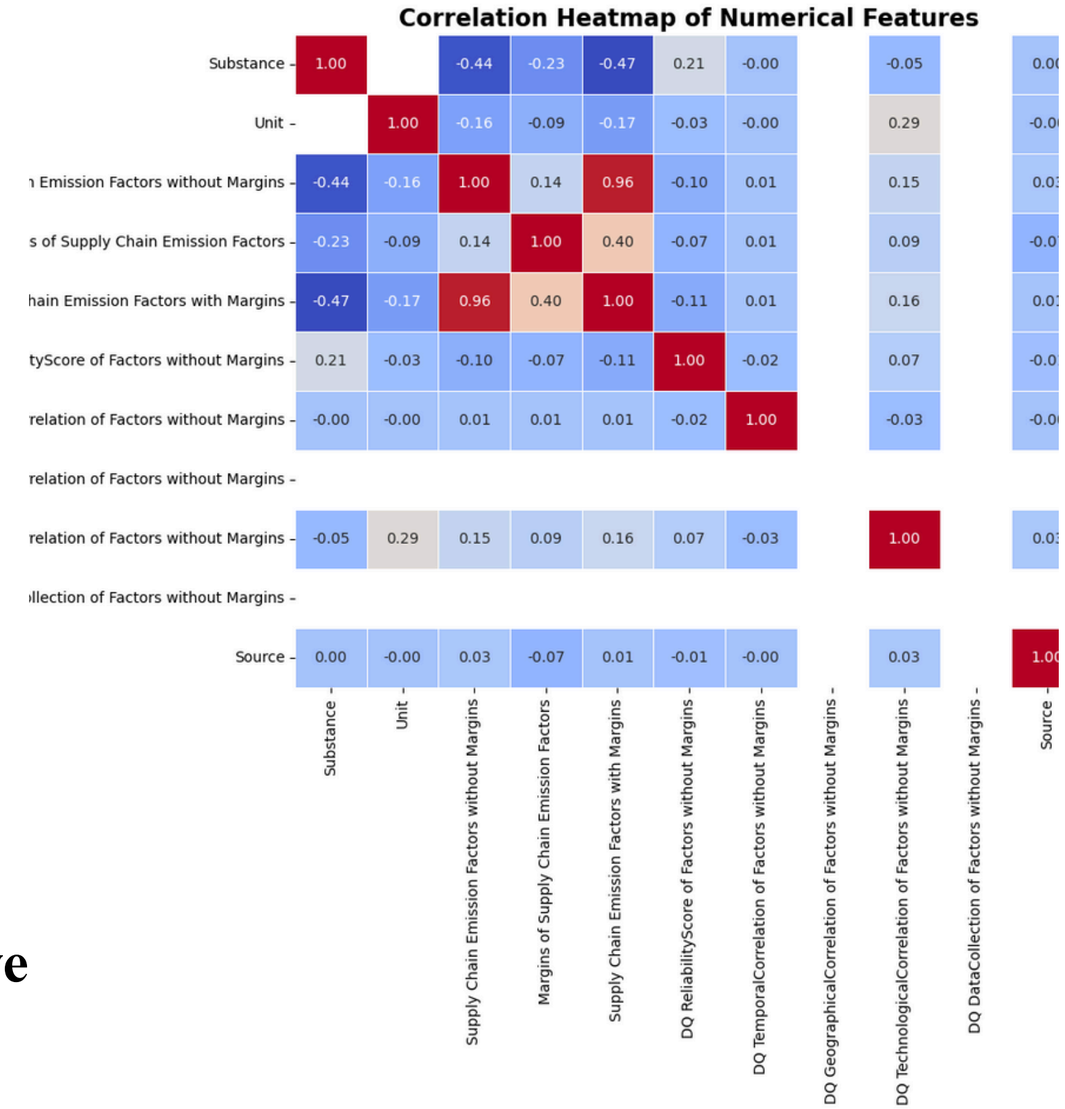
Graphs Shows:

- Shows the relationships between features like DQ indicators, Substances, Margins, etc.

- **The Target variables –Supply Chain Emission with Margins**-has the **Strong correlation(0.96)** with the version without margins. Makes sense, since they're tightly linked.

- **Data Quality features** like Technological and Geographical Correlation show moderate but noticeable influence.

- Some Variables like **Substance** and **Unit** have **negative correlations**, which helped me understand which features might not help the model.



Model Performance:

To find the most accurate and reliable model for predicting Greenhouse gas emissions.

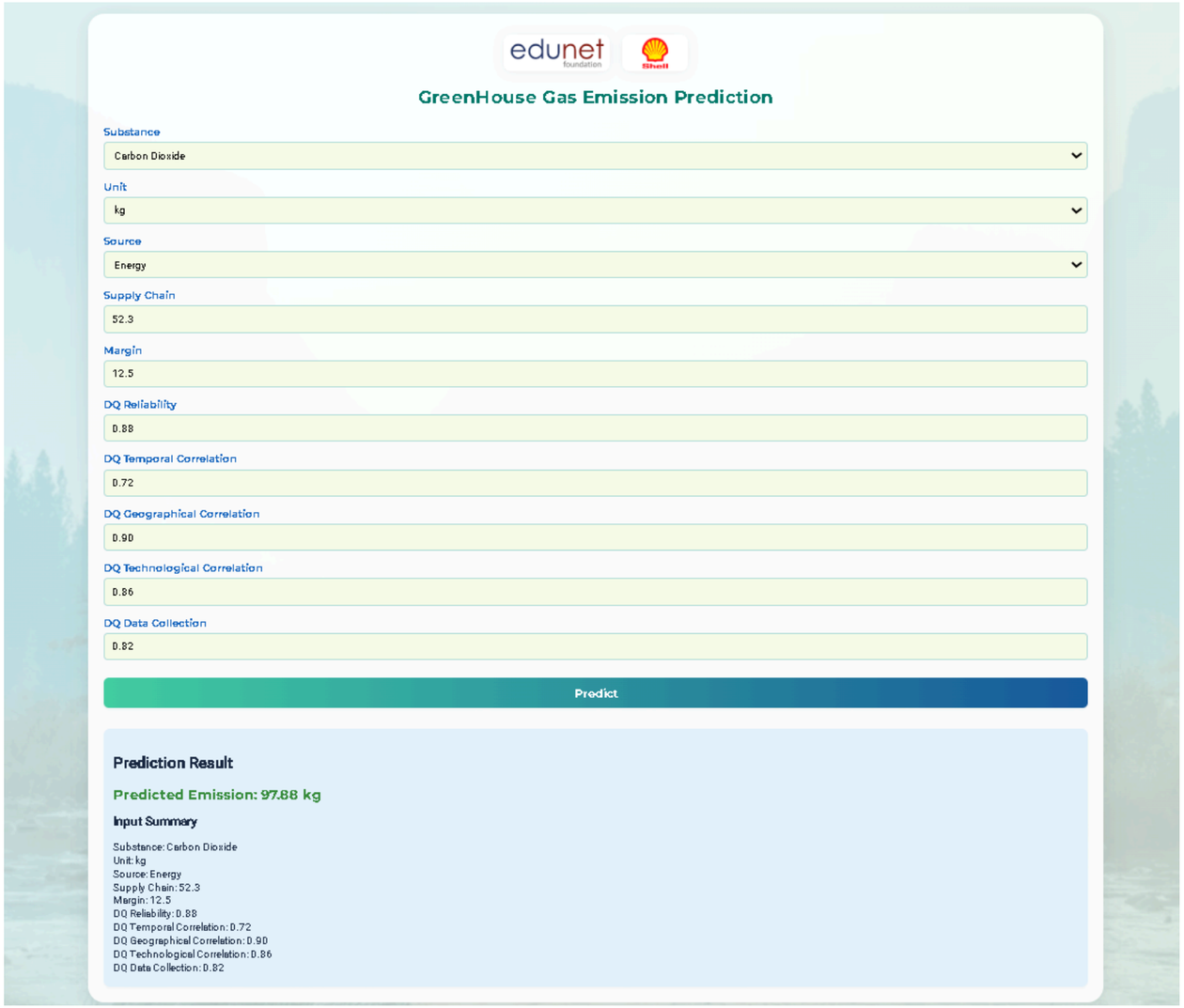
Model Evaluated:

Model Name	R^2	MSE	RMSE	Remark
Linear Regression	1.00	0.00000078	0.00281	Very Accurate on the dataset-almost perfect fit
Random Forest	0.9993	0.000036	0.0059	Strong Performance, Handled data variations well
Random Forest(Tuned)	0.9995	0.0054	0.074	Still Performed great, though RMSE increased slightly after tuning

Conclusion From Testing:

Even though all models performed well, **Linear Regression(0.99999)** gave the best overall accuracy with the **lowest error(0.00000078)** – So, I selected it for the final Model.

Project Output:



Conclusion:

- Gained hands-on experience with real-world GHG emissions data from U.S. industries and commodities.
- Learned how to clean, preprocess, and prepare raw datasets for machine learning.
- Explored multiple regression models and selected the best-performing one based on accuracy and error.
- Built a simple, interactive web application using Flask and deployed it live via Render.
- Understood how data quality impacts model prediction and decision-making.
- Realized the importance of combining data skills with sustainability goals to build something meaningful.

GitHub Link: https://github.com/Siteshgupta123/Edunet-Shell-Internship/tree/main