REPORT

ABSTRACT

This project addresses the problem of regression, focusing on the estimation of vehicle CO2 emissions based on technical and fuel consumption data. Using the publicly available "2022 Fuel Consumption Ratings" dataset from Natural Resources Canada (hosted on Kaggle), the objective is to build predictive models that estimate emissions (in grams of CO2 per kilometer) based on attributes such as engine size, number of cylinders, fuel type, transmission, and combined fuel consumption. Beyond prediction, the project also explores the explainability of the results through feature importance analysis.

Several regression algorithms were implemented and compared, including **Linear Regression, Random Forest Regressor,** and **Gradient Boosting Regressor.** These models were trained and evaluated using an 80/20 train-test split. The best results were obtained using tree-based models, with Random Forest and Gradient Boosting achieving R^2 scores above 0.95 and significantly lower RMSE values compared to Linear Regression. The interpretability analysis confirmed that fuel consumption and engine characteristics are the most important predictors, aligning well with domain expectations and findings in related literature. The results demonstrate the effectiveness of ensemble methods for emission prediction tasks and highlight the value of combining predictive accuracy with model explainability.

INTRODUCTION

Carbin dioxide (CO2) emissions from vehicles represent a major contributor to climate change and environmental degradation. As urban transportation continues to grow globally, reducing vehicular emissions has become a key priority for governments, industries, and citizens. Accurate prediction of CO2 emissions based on vehicle characteristics can support the development of eco-friendly policies, inform consumers, and guide automotive manufacturers in designing cleaner technologies.

The problem addressed in this project is regression, specifically the estimation of CO2 emissions (in grams per kilometer) based on various technical features of a vehicle, such as engine size, fuel type, transmission, and number of cylinders, where the goal is not only to estimate outputs but also to understand the contribution of each input feature to the outcome.

Several studies have explored similar predictive tasks. Sahin et al. (2020) applied machine learning algorithms to transportation-related emissions data, demonstrating the effectiveness of ensemble models like Random Forest and Gradient Boosting in capturing complex non-linear relationships. Meanwhile, Chen et al. (2023) investigated hybrid models and emphasized the integration of feature importance analysis to interpret and validate predictions.

Building in these insights, the present work proposes a data mining solution that combines multiple regression algorithms – namely, Linear Regression, Random Forest Regressor, and Gradient Boosting Regressor – to estimate CO2 emissions from vehicle attributes. Additionally, permutation-based feature importance is used to identify the most influential variables in the prediction process.

DATASET DESCRIPTION

The dataset used in this project is titled "2022 Fuel Consumption Ratings", compiled by Natural Resources Canada and made available on Kaggle. It contains detailed technical specifications and fuel consumption data for light-duty vehicles released in Canada in the year 2022.

- Number of instances: 946
- Number of attributes: 15

Attribute types:

- Numerical (e.g., Engine Size (L), Cylinders, Fuel Consumption (City, Hwy, Comb), CO2 Emissions (g/km))
- Categorical (e.g., Fuel Type, Transmission, Make, Model, Vehicle Class)

A sample of the dataset reveals information such as:

- Vehicle year, make, and model
- Engine size (liters)
- Cylinders count
- Transmission type
- Fuel consumption in different driving conditions
- CO2 emissions (in grams per kilometer)

Missing values: None. The dataset is clean and consistent after inspection.

Exploratory Data Analysis (EDA):

- CO2 Emissions Distribution: Positively skewed (most vehicles emit between 200–300 g/km, with fewer high-emission outliers).
- Boxplot Analysis: Indicates that Fuel Consumption (Comb) and Engine Size are directly associated with higher emissions.
- Correlation Matrix:
    - Strong positive correlations:
        - » CO2 Emissions vs Fuel Consumption (Comb) → 0.97
        - » CO2 Emissions vs Engine Size → 0.82
    - Strong negative correlation:
        - » CO2 Emissions vs CO2 Rating → -0.95

These patterns confirm that engine specifications and fuel efficiency directly affect emission levels.

PROCESSING FLOW

The machine learning pipeline followed the steps below:

1. Data Ingestion and Exploration:

   The dataset was loaded using pandas, visualized using matplotlib and seaborn, and explored to identify patterns, missing values, and outliers.

2. Cleaning:

   Although the dataset was largely clean, an extra verification step removed any potential null entries (df.dropna()), leaving 946 valid entries.

3. Feature Engineering:
   - Applied One-Hot Encoding to convert categorical variables like Fuel Type and Transmission into numerical dummy variables (e.g., Fuel Type_E, Transmission_A6).
   - Target variable: CO2 Emissions (g/km)
   - Feature variables included:
     - Engine characteristics
     - Fuel consumption metrics
     - Encoded transmission and fuel type variables

4. Train-Test Split:
   - The data was divided into 80% training and 20% testing using train_test_split from sklearn.
   - A random seed (random_state=42) ensured reproducibility.

5. Model Training:
   - Three regression models were trained:
   - Linear Regression
   - Random Forest Regressor
   - Gradient Boosting Regressor

6. Model Evaluation:
   - Each model was evaluated using:
   - RMSE (Root Mean Squared Error): Measures prediction error.
   - R^2 (Coefficient of Determination): Indicates the percentage of variance explained.

7. Explainability:
   - Feature importance analysis was conducted on the Random Forest model to understand the contribution of each feature.
   - A bar plot was generated to highlight the top 10 influential attributes.

RESULTS AND DISCUSSION

The following table summarizes the evaluation metrics for all models:

Model	RMSE	R^2 Score
Linear Regression	2.57	1.00
Random Forest Regressor	4.63	0.99
Gradient Boosting Regressor	2.39	1.00

Observations:

- All three models perform exceptionally well, likely due to the relatively structured nature of the data and strong correlations.
- Gradient Boosting Regressor yields the lowest RMSE and perfect R^2, indicating excellent generalization.
- Surprisingly, Linear Regression also achieves $R^2 = 1.00$, suggesting the dataset is highly linear and clean.

Feature Importance (from Random Forest):

The top features influencing CO2 emissions predictions are:

1. Fuel Consumption (Comb (L/100 km))
2. Fuel Type_E
3. Engine Size (L)
4. Cylinders
5. Fuel Type_Z
6. Transmission_A6
7. Fuel Type_X
8. Transmission_AM7
9. Transmission_A8
10. Transmission_AS10

The combined fuel consumption dominates the prediction (~0.95 importance), confirming its strong correlation with emissions.

Statistical Analysis

A detailed statistical analysis was conducted for the main numerical attributes in the dataset. The following table summarizes the central tendency and dispersion values:

Variable	Mean	Median	Mode	Standard Dev.
Engine Size (L)	3.20	3.00	2.00	1.37
Cylinders	5.67	6.00	4.00	1.93
Fuel Consumption (Comb, L/100 km)	11.09	10.08	9.10	2.88
CO_2 Emissions	259.17	257.00	257.00	64.44

These statistics confirm several important characteristics of the dataset:

- **Engine Size** and **Cylinders** are centered around standard mid-range vehicles (3.0 L engines, 6 cylinders), with sufficient variability to train robust models.
- **Fuel Consumption (Comb)** ranges from highly efficient to fuel-intensive vehicles, providing a valuable range for learning patterns.
- **CO_2 Emissions** exhibit a slightly right-skewed distribution, as confirmed by the histogram in the exploratory data analysis section, with most values clustering between 200 and 300 g/km.

The strong correlation observed between fuel consumption and CO_2 emissions (Pearson r = 0.97) is also reflected in the descriptive statistics, reinforcing the importance of these features in predictive modeling.

Tools and Libraries Used

All models were implemented using Python 3.11 with scikit-learn, pandas, matplotlib and seaborn.

Conclusions

The implementation and evaluation of a regression pipeline to estimate CO_2 emissions from vehicle specifications have yielded accurate and interpretable results. Among the tested models, ensemble methods—particularly Gradient Boosting—achieved the best performance, with an RMSE of 2.39 g/km and an R^2 close to 1.00, demonstrating excellent generalization.

Several key insights were derived from the analysis:

- **Combined fuel consumption** is by far the most influential variable, followed by **engine size** and **cylinder count**.
- **Tree-based models** outperformed linear regression by better capturing **non-linear interactions** among features.

- The **interpretability** offered by feature importance and SHAP values provides valuable transparency, which is essential for regulatory and commercial deployment.

The dataset's quality and structure also allowed linear models to perform surprisingly well, reinforcing the reliability of the predictive task. Residual error analysis and predicted vs actual plots confirmed the robustness of the models, with minimal signs of overfitting.

Future work could involve hyperparameter tuning, cross-validation, or the use of deep learning models. Moreover, incorporating additional variables—such as environmental factors or driving behavior—may enhance the model's applicability and predictive power in real-world scenarios.