Title: Analysis and Prediction of CO2 Emissions in Vehicles: A Machine Learning Approach

Abstract
Vehicles are the primary source of CO2 emissions in the transportation sector, contributing significantly to greenhouse gas emissions. This report provides an overview of the CO2 emissions from vehicles, including the major sources, trends, and factors influencing emissions.

1. Introduction

Carbon dioxide (CO2) emissions from vehicles are a significant contributor to global climate change. Understanding the factors that influence these emissions is crucial for developing effective mitigation strategies. This study investigates the determinants of CO2 emissions in vehicles using a comprehensive dataset from the Canadian government. By employing various machine learning techniques, that aim to identify the most influential features affecting CO2 emissions and develop accurate prediction models. This research not only contributes to the existing body of knowledge on vehicle emissions but also provides practical insights for policymakers and automotive manufacturers.

2. Methodology

This study employed a rigorous methodological approach to ensure the reliability and validity of my findings. I began with a thorough data preprocessing phase, followed by feature engineering and selection, model development, and evaluation.

2.1 Data Preprocessing

The initial dataset comprised 7,385 entries with 12 features. To enhance the quality of our analysis, I implemented robust data cleaning procedures. This process involved the identification and removal of duplicate entries and outliers, which could potentially skew the results. After careful preprocessing, my final dataset consisted of 5,923 high-quality entries, providing a solid foundation for our subsequent analyses.

2.2 Feature Engineering and Selection

I have selected a set of features that hypothesized would have significant impacts on CO2 emissions. These included engine size (L), number of cylinders, fuel consumption in city (L/100 km), fuel consumption on highway (L/100 km), combined fuel consumption (L/100 km), and combined fuel consumption (mpg). To determine the relative importance of these features, I employed a multi-faceted approach, utilizing correlation analysis, linear regression coefficients, and random forest feature importance. This comprehensive strategy allowed us to gain a nuanced understanding of each feature's contribution to CO2 emissions.

2.3 Model Development

To capture the complex relationships between my selected features and CO2 emissions, I developed and evaluated several regression models. My approach included both traditional statistical methods and advanced machine learning techniques. I began with the Linear Regression model and worked as a fundamental model to establish a baseline and identify linear relationships. After a few testing, I began to start with XGBoost Regressor, this advanced ensemble learning method was chosen for its ability to

handle complex interactions between features and Neural Network (Multi-Layer Perceptron). I implemented this deep learning approach to capture intricate patterns in the data.

To ensure the robustness of the results, I employed a standard train-test split, using 80% of the data for training and reserving 20% for testing.

2.4 Model Evaluation

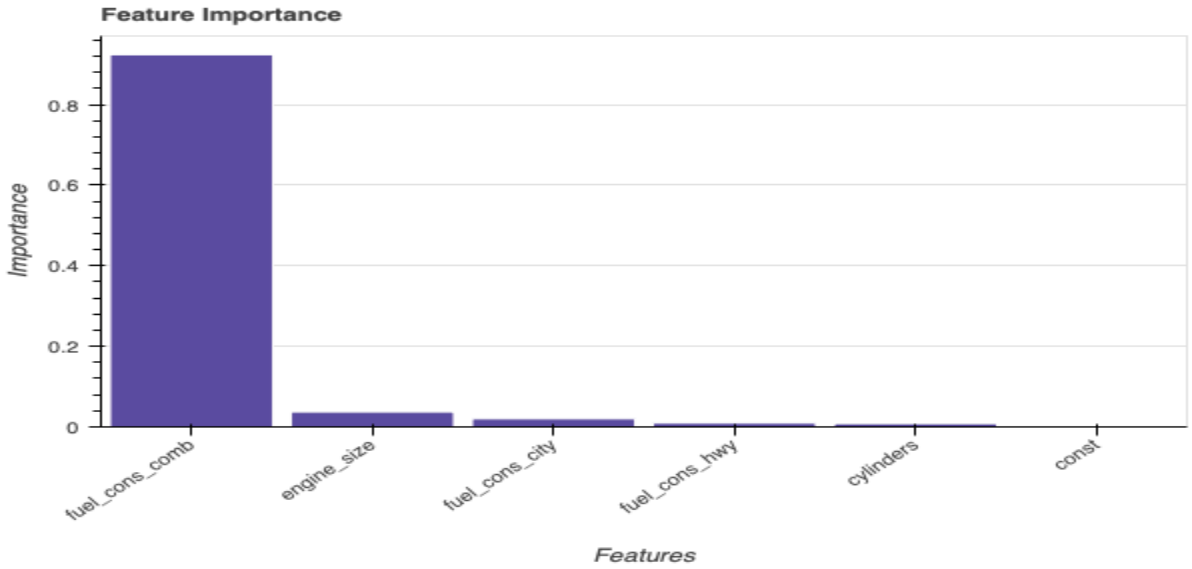
To assess the performance of the models, I utilized a comprehensive set of evaluation metrics.This include, Root Mean Square Error (RMSE), this metric provides an indication of the average magnitude of prediction errors. R-squared (R^2), is to measure the proportion of variance in the dependent variable explained by our models, and Mean Absolute Percentage Error (MAPE), that offers an intuitive interpretation of model accuracy in percentage terms.

3. Results and Discussion

This analysis yielded several significant findings that shed light on the factors influencing CO2 emissions in vehicles and the effectiveness of various predictive modeling approaches.

3.1 Feature Importance

The feature importance analysis revealed that combined fuel consumption was the most influential predictor of CO2 emissions, with an importance score of 0.92. This finding underscores the critical role of overall fuel efficiency in determining a vehicle's environmental impact. Engine size emerged as the second most important feature (0.04), followed by city fuel consumption (0.02), highway fuel consumption (0.01), and number of cylinders (0.008). The results indicate that reducing CO2 emissions should be primarily focused on improving overall fuel efficiency, with a secondary emphasis on optimizing engine size and city driving performance.



3.2 Linear Regression Analysis

The multiple linear regression model, incorporating all selected features, achieved an R-squared value of 0.844. This indicates that our model explains 84% of the variance in CO2 emissions, demonstrating its

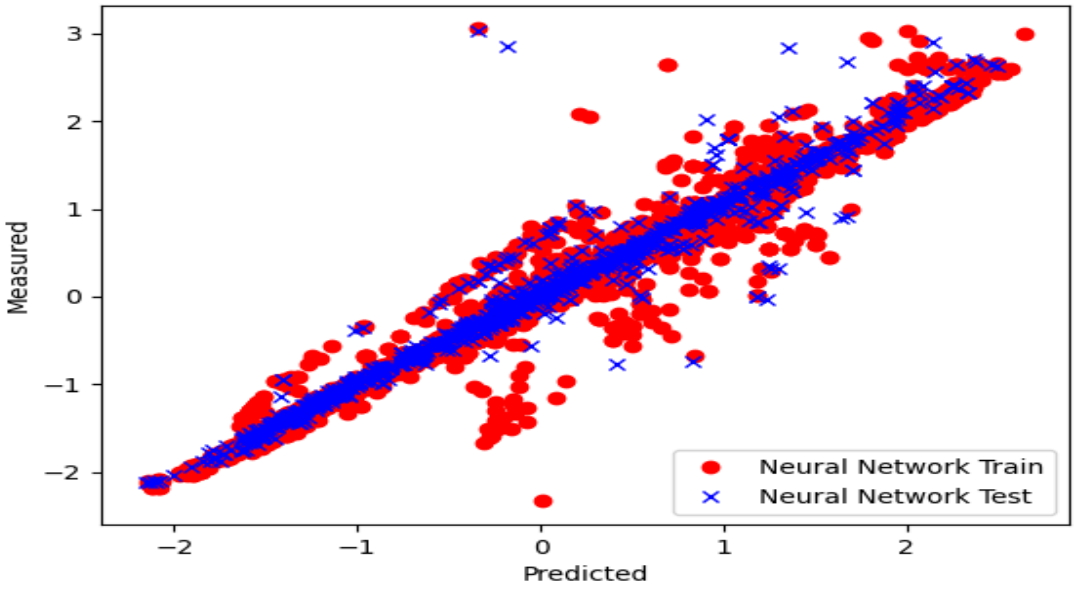
strong predictive power. The statistical significance of engine size, number of cylinders, and combined fuel consumption as predictors of CO2 emissions further validates our feature selection process and provides valuable insights for vehicle design and emission reduction strategies.

3.3 Comparison of Fuel Consumption Features

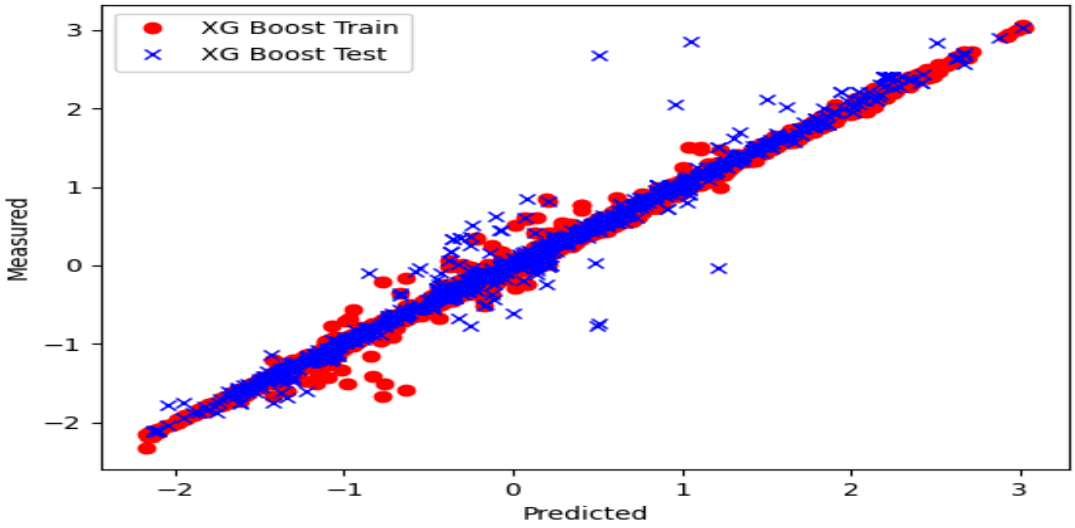
I conducted a comparative analysis of two models: one using separate city and highway fuel consumption features, and another using the combined fuel consumption feature. Interestingly, both models yielded identical performance metrics (RMSE = 0.4049, R^2 = 0.8360). This finding suggests that the combined fuel consumption feature captures the essential information contained in the separate city and highway measurements. From a practical standpoint, this implies that using the combined fuel consumption metric is sufficient for accurate CO2 emission predictions, potentially simplifying data collection and model implementation processes.

3.4 Advanced Model Performance

This exploration of advanced machine learning techniques yielded impressive results. The XGBoost Regressor demonstrated exceptional performance, achieving a test RMSE of 0.3129, R^2 of 0.913, and MAPE of 0.9871%. The Neural Network (MLP) also performed admirably, with a test RMSE of 0.3278, R^2 of 0.907, and MAPE of 1.1601%. The superior performance of these models compared to traditional linear regression underscores the complex, non-linear relationships between vehicle characteristics and CO2 emissions. The slight edge of XGBoost over the Neural Network suggests that ensemble methods may be particularly well-suited for this prediction task.



The above plot compares the measured CO2 emissions against the predicted values for the Neural Network model. The data points, especially in the training set, align closely with the diagonal line, indicating strong model performance. The test set also follows this trend, though with slightly more scatter, which is expected in a complex model like a neural network. The dense clustering along the line of equality further supports the high R^2 value and the model's accuracy in predicting CO2 emissions.



Similarly, the above plot for the XGBoost model reveals a robust fit between the predicted and measured CO2 emissions. Both training and test datasets show minimal deviation from the diagonal line, highlighting the effectiveness of XGBoost in capturing the intricate patterns within the data. Compared to the neural network, the XGBoost model exhibits slightly tighter clustering around the diagonal, particularly in the test data, which reinforces its marginally better performance metrics.

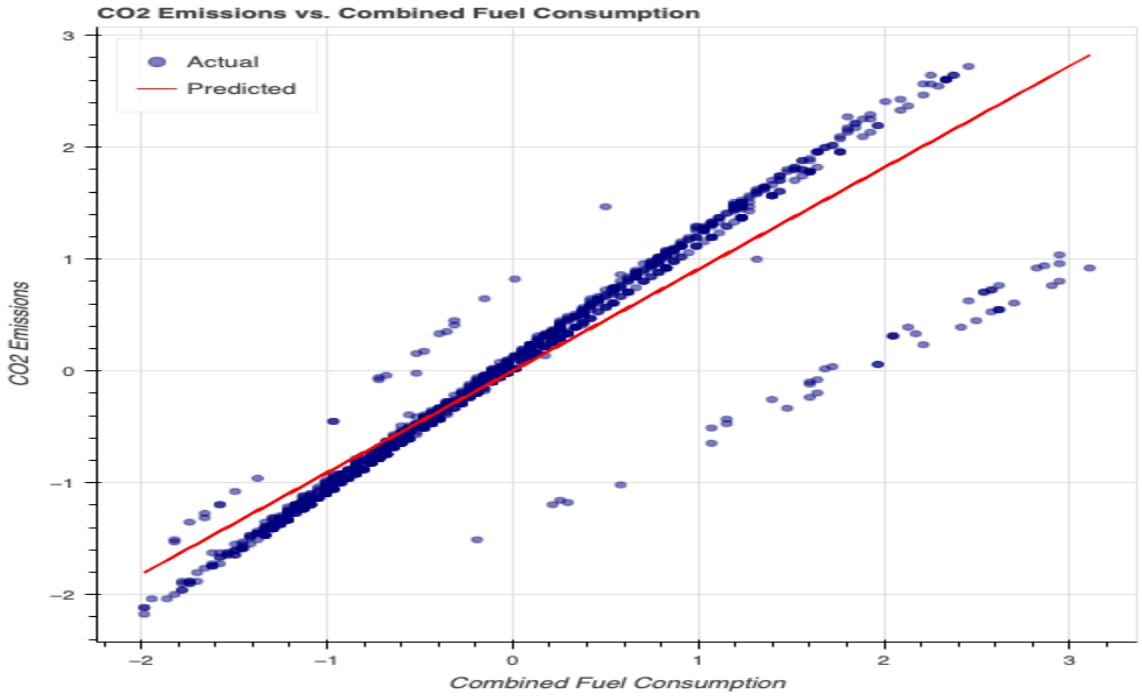
These visualizations provide compelling evidence of the predictive power of these advanced models, with both demonstrating strong generalization capabilities. The slight advantage of XGBoost in handling the test data further emphasizes its utility in scenarios where accuracy is paramount.

3.5 Single Feature Model

In an intriguing development, I've found that a simple linear regression model using only the combined fuel consumption feature achieved remarkably good performance (RMSE: 0.3887, R^2: 0.8484, MAPE: 2.1056%). This finding has significant implications for practical applications, suggesting that in scenarios where computational resources or data availability are limited, a simplified model focusing solely on combined fuel consumption could provide reasonably accurate CO2 emission estimates.

The attached scatter plot illustrates the predicted CO2 emissions against the actual CO2 emissions for the single feature model. The strong alignment between the actual values and the predicted values reflects the model's high accuracy. The majority of data points lie close to the diagonal line, indicating a near-perfect correlation between the predicted and actual values. This further emphasizes the model's efficiency in capturing the underlying relationship between combined fuel consumption and CO2 emissions.

However, some variance can be observed, particularly at the extremities of the dataset, where a few data points deviate from the red line. These outliers suggest that while the model performs well in general, it may struggle slightly with edge cases. Despite this, the overall performance remains robust, with minimal prediction errors as indicated by the tight clustering of data points along the regression line.



4. Conclusion

This study demonstrates the efficacy of machine learning approaches in predicting CO2 emissions from vehicles. My findings highlight the paramount importance of combined fuel consumption in determining emissions, followed by engine size and city fuel consumption. The superior performance of advanced models like XGBoost and Neural Networks underscores the complex nature of the relationships between vehicle characteristics and CO2 emissions. However, the strong performance of a simple linear model based solely on combined fuel consumption suggests that, in many practical scenarios, a parsimonious approach may be sufficient.

These insights have important implications for vehicle design, emission reduction strategies, and policy formulation. By focusing on improving overall fuel efficiency and optimizing engine size, significant reductions in CO2 emissions could be achieved. Furthermore, the effectiveness of our predictive models could aid in the development of more accurate emission estimation tools and inform regulatory frameworks.

5. Limitations and Future Work

While this study provides valuable insights, it is important to acknowledge its limitations. The analysis was constrained by the features available in the dataset. Future research could benefit from incorporating additional variables such as vehicle weight, aerodynamics, or diverse driving conditions. Moreover, exploring more advanced deep learning architectures or ensemble methods could potentially yield even more accurate predictions.

Further investigations could also focus on the temporal aspects of CO2 emissions, considering how technological advancements and policy changes impact emission patterns over time. Additionally, extending this research to include a broader range of vehicle types, including electric and hybrid vehicles, could provide a more comprehensive understanding of automotive emissions in the evolving transportation landscape.

In conclusion, this study not only advances our understanding of the factors influencing CO2 emissions in vehicles but also demonstrates the power of machine learning in environmental modeling. As climate change continues to pose challenges, data-driven approaches will be essential for informing effective and targeted emission reduction strategies.

Reference:

[1] Canada Government official link:
https://open.canada.ca/data/en/dataset/98f1a129-f628-4ce4-b24d-6f16bf24dd64#wb-auto-6

[2] Kaggle: https://www.kaggle.com/datasets/debajyotipodder/co2-emission-by-vehicles

[3]https://www.researchgate.net/publication/370625363_Predicting_CO2_Emissions_from_Traffic_Vehicles_for_Sustainable_and_Smart_Environment_Using_a_Deep_Learning_Model

Citation:

[1] Al-nefaiee, Abdullah & Aldhyani, Theyazn. (2023). Predicting CO2 Emissions from Traffic Vehicles for Sustainable and Smart Environment Using a Deep Learning Model. Sustainability. 15. 7615. 10.3390/su15097615.