BNEJSS | Balkan and Near Eastern Journal of Social Sciences
Balkan ve Yakın Doğu Sosyal Bilimler Dergisi
Parlar, 2025: 11 (Special Issue)

Predicting CO_2 Emissions using Machine Learning Methods: A Comparative Regression and Feature Importance Analysis

Tuba PARLAR[1]

[1] Hatay Mustafa Kemal Üniversitesi, tparlar@mku.edu.tr, ORCID: 0000-0002-8004-6150

Abstract: Road transportation is a major source of atmospheric carbon dioxide (CO_2), posing a significant threat to global climate stability and environmental sustainability. According to the International Energy Agency, the transport sector is responsible for approximately one-quarter of worldwide energy-derived CO_2 emissions, underscoring its role in the global carbon footprint. The continuous rise in the number of passenger vehicles has intensified concerns regarding environmental degradation, deteriorating air quality, and associated public health risks. This study presents a machine learning based framework for predicting vehicle CO_2 emissions using a publicly available dataset. Three regression models were implemented to provide a comparative performance evaluation. To improve model interpretability, feature importance analysis was applied using the Shapley Additive explanations method. The experimental findings indicate that ensemble-based models outperform the linear regression approach, achieving superior predictive accuracy across all evaluation metrics. Overall, the proposed framework improves reliability in CO_2 emission prediction, while offering actionable insights for policymakers, manufacturers, and consumers aiming to support low-carbon transportation.

Key Words: CO_2 Emissions, Feature Importance, Machine Learning, Regression Models, SHAP.

1. INTRODUCTION

The transportation sector remains one of the most critical contributors to global greenhouse gas (GHG) emissions, with road transport accounting for a substantial share of worldwide energy-related carbon dioxide (CO_2) emissions. According to international assessments, the steady growth in private vehicle ownership, coupled with increasing travel demand, has intensified concerns regarding climate change, air pollution, and associated public health impacts. As governments and industries pursue ambitious decarbonization targets, the need for accurate and transparent tools to estimate and monitor vehicle emissions has become more pressing than ever.

Carbon dioxide (CO_2) emissions from the transportation sector represent a critical challenge in the global fight against climate change. According to the international Energy Agency (IEA) global transport CO_2 emissions increased by 38% from 2000 to 2021, from nearly 4250 Mt CO_2, to 5860 Mt CO_2 adding more than 16100 Mt CO_2 to the atmosphere (Marzouk, 2025). The IEA's Net Zero Emissions (NZE) scenario emphasizes the urgency of this challenge, requiring transport emissions to fall by approximately 25% to around 6 Gt CO_2 by 2030 to align with climate goals. This ambitious target necessitates rapid electrification of road vehicles, enhanced energy efficiency measures, and the development of low-emission fuels. With global passenger vehicle registrations continuing to grow, vehicular emissions have become a primary driver of environmental degradation, urban air pollution, and associated public health concerns. This escalating challenge necessitates the development of accurate and interpretable predictive models that can quantify vehicle environmental impact and inform evidence-based policy decisions.

The continuous increase in passenger vehicle registrations worldwide has made car-sourced emissions a principal driver of environmental degradation, urban air pollution, and associated public health risks. This issue is particularly pronounced in developing economies, where total direct CO_2 emissions surged by 180% between 2000 and 2021. The primary contributors to this dramatic increase are concentrated in energy generation, heavy industry, and road transportation.

Recent advances in machine learning have demonstrated considerable potential for modeling complex systems. Previous studies have shown the advantages of machine learning methods over traditional statistical approaches for CO2 emission prediction (Li andSun, 2021; Zhao et al., 2021). Ensemble-based models such as Random Forest and gradient boosting algorithms have shown significant predictive performance in various energy related applications.

This study presents a framework that combines machine learning models with explainable artificial intelligence techniques. Random Forest, CatBoost, and Bayesian Ridge regression models are employed to predict CO_2 emissions using CatBoost-based Shapley Additive exPlanations (SHAP) method to identify the most important emission drivers. The remainder of the study is organized as follows. Section 2 describes the methodology and evaluation metrics. Section 3 presents the

experimental results and Section 4 concludes the study with key insights.

2. MATERIAL AND METHODS

The study addresses a supervised regression problem aimed at predicting CO2 emissions based on a set of features related to fuel consumption, vehicle characteristics.

2.2. Machine Learning Methods

The predictive modeling is executed using three regression models to capture both linear and nonlinear relationship between predictors and CO_2 emissions: namely Random Forest Regressor, CatBoost Regressor, and Bayesian Ridge Regression.

Random Forest is an ensemble learning method that constructs multiple decision trees using bootstrap sampling and aggregates their predictions through averaging (Cutler et al., 2012). The approach reduces variance and improves generalization compared to single decision trees. Random Forest is particularly effective in handling nonlinear relationships and feature interactions.

CatBoost is a gradient boosting algorithm that builds an ensemble of decision trees sequentially, where each new tree attempts to correct the residual errors of the previous ensemble. A key advantage of CatBoost lies in its ability to handle categorical variables efficiently and to reduce prediction bias through ordered boosting. Due to its strong performance on structured data and its compatibility with advanced interpretability techniques such as SHAP, Cat Boost is adopted in many studies (Prokhorenkova et al., 2018).

Bayesian Ridge Regression is a probabilistic linear regression model that incorporates Bayesian inference by placing prior distributions on the regression coefficients. Unlike traditional ridge regression, Bayesian Ridge estimates regularization parameters directly from the data. Bayesian Ridge was retained as a linear baseline model, enabling comparison between classical statistical approaches and more complex nonlinear machine learning methods (Tipping, 2001).

2.3. Feature Importance and Explainability

To identify the most influential predictors of CO_2 emissions, a two-stage feature importance analysis is conducted. CatBoost's built-in feature importance mechanism is used to quantify the contribution of each variable to model performance. This approach evaluates the impact of feature permutations on prediction accuracy. To enhance interpretability, Shapley Additive exPlanations (SHAP) method is applied. SHAP method is based on cooperative game theory and assigns each feature a contribution value representing its impact on the prediction (Lundberg andLee, 2017).

2.4. Evaluation Metrics

Model performance was evaluated using 5-fold cross validation. The following metrics are employed: coefficient of determination (R^2), mean squared error (MSE), root mean squared error (RMSE), and mean absolute error (MAE) defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^{T}(y_t - \widehat{y_t})^2}{\sum_{i=1}^{T}(y_t - \bar{y})^2} \quad (1)$$

$$MSE = \frac{1}{T}\sum_{t=1}^{T}(y_t - \widehat{y_t})^2 \quad (2)$$

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(y_t - \widehat{y_t})^2} \quad (3)$$

$$MAE = \frac{1}{T}\sum_{t=1}^{T}|y_t - \widehat{y_t}| \quad (4)$$

where T denotes the number of predictions, y_t represents the actual value of the t-th observation, and $\hat{y}_t$ denotes the predicted value of the t-th observation, and $\bar{y}$ is the mean of the observed values.

3. RESULTS AND DISCUSSION

The experiments were conducted on a MacBook equipped with a 2.5GHz Core i7 and 16GB RAM and developed using Python scikit-learn library (Pedregosa et al., 2011).

3.1. Dataset

In this study, we use the Fuel Consumption Ratings (Canada, 2023) dataset, which provides model-specific fuel consumption ratings and estimated CO_2 emissions for a wide range of vehicles. The dataset is obtained using standardized fuel consumption testing procedures. Table 1 presents the attributes of the dataset with their descriptions and measurement units.

Table 1: Description of the features, definitions, and measurement units of the dataset

Features	Definitions	Measurement Units
Make	Manufacturer	Brand name
Model	Vehicle model	Alphanumeric

Vehicle Class	Classification	Category (Compact, Sedan, SUV, etc.)
Engine Size (L)	Engine volume	Liters (L)
Cylinders	Number of engine cylinders	Integer
Transmission	Type of transmission system	Category (**A**utomatic, **M**anual, AM, AS, AV, etc.)
Fuel Type	Type of Fuel	Category (Reg. Gasoline(**X**), **D**iesel, **N**atural gas, etc.)
Fuel Consumption City	Liters per km in city	L/100 km
Fuel Consumption Hwy	Highway fuel consumption	L/100 km
Fuel Consumption Comb	Combined fuel consumption	L/100 km
Fuel Consumption Comb	Combined fuel economy	Miles per gallon (MPG)
CO2 Emissions	$C0_2$ released per km	g/km

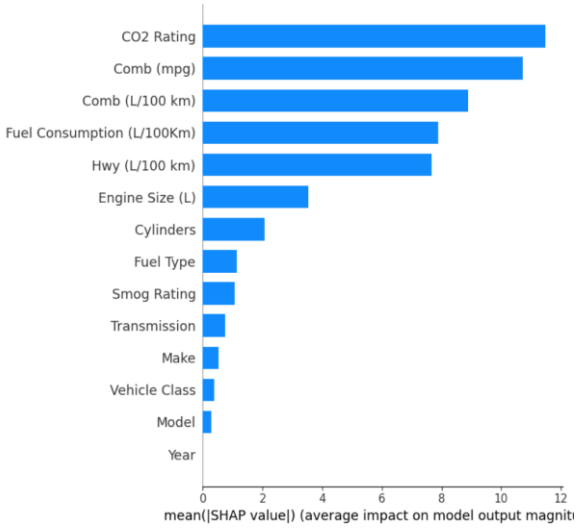
The predictive performance of three regression models; Random Forest, CatBoost, and Bayesian Ridge was first evaluated using all features in the dataset. The results are shown in Table 2. Table 2 shows that the Random Forest model achieved the highest predictive accuracy, with an R^2 value of 0.9844 and lowest error metrics (RMSE=7.98, MAE=2.42). CatBoost also demonstrated strong performance (R^2=0.9728), although with higher error values compared to Random Forest. In contrast, Bayesian Ridge regression exhibited substantially lower predictive accuracy, indicating limited capability in capturing the complex relationship governing CO_2 emissions. These results confirm the superiority of nonlinear ensemble-based models over linear regression approaches for CO_2 emission prediction and justify the selection of Random Forest as the primary model for subsequent analysis.

Table 2: The performances of models using all features

Model	**R^2**	**MSE**	**RMSE**	**MAE**
RandomForest	0.9844	63.62	7.98	2.42
CatBoost	0.9728	105.88	10.29	3.50
BayesianRidge	0.9199	248.69	15.77	8.73

To identify the most influential predictors of CO_2 emissions, a feature importance analysis was conducted using CatBoostRegressor in combination with SHAP. The SHAP summary plot is presented in Figure 1, where features are ranked according to their mean absolute SHAP values. As illustrated in Figure 1, CO_2 Rating, combined fuel consumption (mpg and L/100 km), and fuel consumption metrics dominate the importance ranking, indicating their strong contribution to the model output. Highway fuel consumption and engine size also exhibit noticeable influence, while variables such as vehicle make, model year, transmission type, and fuel type contribute relatively low to prediction accuracy. The SHAP results reveal that a small subset of fuel efficiency-related variables captures the majority of the information required for accurate CO_2 emission prediction. This observation motivated the construction of reduced feature sets based on the top-ranked features.

Figure 1. SHAP Summary Bar Plot of Feature Importances for CO_2 Emissions Prediction using CatBoostRegressor



Following the feature importance analysis, the Random Forest model was retrained using the top 4, 5, and 6 features identified by CatBoost and SHAP method. The performance of these reduced models was compared with the baseline Random Forest model trained on all features. The results are summarized in Table 3.

Table 3: RandomForestRegressor performance comparison using all features and selected top features

Feature set	R^2	MSE	RMSE	MAE
All features	0.9844	63.62	7.98	2.42
Top 4 features	**0.9874**	51.33	7.16	2.34
Top 5 features	0.9864	54.28	7.37	2.36
Top 6 features	0.9864	55.55	7.45	2.33

The results demonstrate that feature selection leads to improved or comparable performance relative to the baseline. Notably, the model trained with the top 4 features achieves the highest R^2 (0.9874) and the lowest RMSE (7.16), outperforming the baseline model despite using significantly fewer input variables. Increasing the number of features from four to five and six does not yield further improvements and results in a slight increase in error metrics. This suggest diminishing returns from additional features, likely due to redundancy among highly correlated fuel consumption variables.

4. CONCLUSIONS

This study investigated the effectiveness of machine learning models for predicting CO_2 emissions using the fuel consumption dataset with feature importance analysis. Three regression approaches; Random Forest, CatBoost, and Bayesian Ridge were evaluated using cross validation. The results demonstrated that nonlinear ensemble models substantially outperform linear regression, with Random Forest achieving the highest predictive accuracy when all features were considered. Feature importance analysis was conducted using CatBoost in combination with SHAP method. The analysis revealed that a limited number of fuel efficiency related features, particularly CO_2 rating and combined fuel consumption metrics, dominate the prediction of CO_2 emissions.

REFERENCES

Canada, N. R. (2023). Fuel Consumption Ratings 2023. Retrieved from https://open.canada.ca/data/en/dataset/98f1a129-f628-4ce4-b24d-6f16bf24dd64

Cutler, A., Cutler, D. R., andStevens, J. R. (2012). Random forests. In Ensemble machine learning (pp. 157-175). Springer.

Li, Y., andSun, Y. (2021). Modeling and predicting city-level CO2 emissions using open access data and machine learning. Environmental Science and Pollution Research, 28(15), 19260-19271.

Lundberg, S. M., andLee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

Marzouk, O. A. (2025). Summary of the 2023 report of TCEP (tracking clean energy progress) by the International Energy Agency (IEA), and proposed process for computing a single aggregate rating. E3S Web of Conferences,

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., andDubourg, V. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A., andGulin, A. (2018). Catboost: Unbiased boosting with categorical features. Advances in Neural Information Processing Systems,

Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. Journal of machine learning research, 1(Jun), 211-244.

Zhao, H.-X., He, R.-C., andYin, N. (2021). Modeling of vehicle CO2 emissions and signal timing analysis at a signalized intersection considering fuel vehicles and electric vehicles. European Transport Research Review, 13(1), 5.