Proceedings of CONF-APMM 2026 Symposium: Applied Mathematics and Quantitative Modeling
DOI: 10.54254/2753-8818/2026.PU37781

Predicting New-Vehicle CO_2 Emissions from Manufacturer Specifications: A Feature-Tier Ladder and Fuel-Type Conditional Study on Canadian EnerGuide Ratings

Jiaxin Fei

School of Mathematical and Physical, Xi'an Jiaotong-Liverpool University, Suzhou, China
Jiaxin.fei23@student.xjtlu.edu.cn

Abstract. Public vehicle-rating data can appear to support nearly perfect carbon-dioxide-emission prediction, although the apparent accuracy may depend more on the information supplied than on the selected algorithm. This study analysed 10,059 Canadian EnerGuide light-duty vehicle configurations from model years 2015–2024. Predictors were organized into three availability tiers: engine size and cylinders; complete manufacturer-sheet specifications; and certified city, highway, and combined fuel consumption. Ordinary least squares, random forest, and gradient boosting were compared on identical shuffled five-fold splits. The Tier 2 random forest was the strongest specification-only model, with mean R^2 of 0.898, mean absolute error of 13.38 g/km, and root mean squared error of 19.29 g/km. Tier 3 raised R^2 to 0.996–0.998, but fuel-conditional regressions produced slopes of 23.41, 23.33, 16.67, and 26.82 g/km per L/100 km for regular gasoline, premium gasoline, 85% ethanol fuel (E85), and diesel. These values closely matched official combustion-factor slopes, identifying certified fuel consumption as a target proxy rather than an ordinary pre-certification predictor. Permutation importance and vehicle-bucket analyses further showed that specification effects varied across cars, sport utility vehicles, pickups, and other vehicles. Separating transmission type from gear count allowed both components to enter the models. The results demonstrate that feature provenance matters more than small differences among common regression algorithms.

Keywords: vehicle CO_2 emissions, feature tiers, regression, target leakage, EnerGuide ratings

1. Introduction

Certified fuel-economy labels connect engineering specifications with standardized estimates of fuel use and tailpipe carbon dioxide (CO_2) emissions. Natural Resources Canada (NRCan) publishes model-specific ratings for new light-duty vehicles sold in Canada, allowing both consumers and researchers to compare vehicles under controlled conditions [1]. The same type of tabular data has become a common machine-learning benchmark because it combines a continuous target, physical engine variables, categorical specifications, and fuel-consumption measurements.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2026.PU37781

Recent studies have reported strong performance when predicting vehicle CO_2 emissions from Canadian rating data. Natarajan et al. compared several algorithms on light-duty vehicles [2], Udoh et al. evaluated machine-learning models for the same broad task [3], and Vu and Ko combined feature selection with linear and nonlinear models [4]. Guo et al. also analysed Canadian vehicle emissions in relation to carbon-neutrality objectives [5]. Transportation studies in China have extended machine learning to emission forecasting and large-scale private-car estimation [6, 7].

A central limitation of benchmark-style comparisons is that all available columns are often treated as equivalent. Certified fuel consumption, however, is generated by the same laboratory procedure as certified tailpipe CO_2 and is combined with the fuel carbon content in the reporting process. Including it may therefore answer a post-certification reconstruction question rather than the practical question of what can be inferred from a manufacturer specification sheet. Leakage research emphasizes that strong cross-validation scores do not repair a mismatch between predictor availability and the intended deployment setting [8].

This study addresses that gap in four ways. First, predictors are divided into three tiers according to when and where they become available. Second, fuel-specific regression slopes are compared with official fuel carbon factors to identify the status of certified fuel consumption. Third, the fifteen vehicle classes are consolidated into four interpretable buckets and evaluated separately. Fourth, the compact transmission codes are decomposed into transmission type and gear count. The objective is to distinguish gains produced by model choice from gains produced by access to a different class of information.

2. Literature review

Linear regression remains useful in vehicle-emission studies because its coefficients provide a transparent baseline, whereas tree ensembles accommodate nonlinearities and interactions. Random forests average randomized trees [9], and gradient boosting sequentially improves residual fit [10]. These complementary structures explain their recurring use in vehicle-emission comparisons [2-5]. The present analysis does not treat a winning algorithm as sufficient evidence. The models are held to an identical evaluation protocol so that changes in feature availability can be separated from changes in functional form.

Two interpretive issues are especially relevant. First, permutation importance measures how much a fitted model's validation performance deteriorates after a predictor is disrupted; it describes model reliance and is not a causal effect [11]. Correlated predictors may share or redistribute this reliance. Second, coefficient interpretation can become unstable when predictors are collinear. Variance inflation factors (VIFs) should therefore be reported with the Tier 1 regression rather than applying an unqualified cutoff rule [12]. These considerations motivate the combination of predictive evaluation, leakage diagnosis, permutation importance, and standardized linear coefficients used below.

3. Methodology

3.1. Data source and cleaning

The source was the official NRCan Fuel Consumption Ratings file "my2015-2024-fuel-consumption-ratings.csv", downloaded from the Government of Canada Open Data Portal on 5 August 2026 [1]. It covers model years 2015–2024 and contained 10,060 rows and 15 columns. The response was certified tailpipe CO_2 emissions in grams per kilometre. This file is not the commonly

DOI: 10.54254/2753-8818/2026.PU37781

reused 7,385-record derivative; no Kaggle copy was used. The CSV was read with encoding="utf-8-sig", and the printed headers confirmed that no byte-order-mark prefix remained.

Manufacturer and model were removed as identifiers. CO_2 rating and smog rating were removed because they were ordinal scores and contained 1,128 (11.21%) and 2,234 (22.21%) missing values, respectively. "Combined (mpg)" was removed as a unit-converted duplicate of combined L/100 km. The single natural-gas observation was excluded because it could not support fuel-conditional analysis. The resulting modelling table had 10,059 observations, 11 retained or derived columns, and no missing predictor or target values.

3.2. Feature engineering and tiers

Transmission codes were parsed into an alphabetic type and an optional numeric gear count. A, AM, AS, AV, and M were mapped to automatic, automated manual, automatic with select shift, continuously variable, and manual. The 410 pure AV records had no numeric suffix and were assigned zero gears because the type variable already retained their continuously variable identity. The fifteen vehicle classes were mapped a priori to the four groups in Table 1.

Table 1. Mapping of the fifteen original vehicle classes to four analysis buckets

Bucket	Original vehicle classes	n
Car	Compact; full-size; mid-size; minicompact; subcompact; two-seater; station wagon (small and mid-size)	5,619
SUV	Sport utility vehicle (small and standard)	3,099
Pickup	Pickup truck (small and standard)	1,098
Other	Minivan; passenger van; special purpose vehicle	243

The nested feature tiers in Table 2 represent increasing information availability. Categorical variables were one-hot encoded within each training fold. Tier 3 was prespecified as a diagnostic upper bound because it includes certified test outputs that are not available in an ordinary pre-certification specification sheet.

Table 2. Predictor availability tiers used in the modelling comparison

Tier	Predictors	Interpretation
1	Engine size; cylinders	Minimal mechanical specifications
2	Tier 1 + model year; gear count; fuel type; transmission type; vehicle bucket	Complete manufacturer-sheet specifications
3	Tier 2 + city, highway, and combined fuel consumption	Post-test diagnostic upper bound

3.3. Models, validation, and diagnostics

Ordinary least squares (OLS) supplied an interpretable baseline. The random forest used 300 trees, random_state=42, and otherwise default scikit-learn settings. Gradient boosting used 100 trees, learning rate 0.1, maximum depth 3, squared-error loss, and the same seed. Numeric scaling and categorical encoding were contained in pipelines and refitted inside every training fold. All nine tier-

DOI: 10.54254/2753-8818/2026.PU37781

by-model combinations reused KFold(n_splits=5, shuffle=True, random_state=42). Cross-validation is a standard approach to out-of-sample model comparison [13]. Each fold recorded R^2, mean absolute error (MAE), and root mean squared error (RMSE). The implementation used Python 3.12, scikit-learn 1.9.0 [14], and statsmodels 0.14.6.

Tier 2 random-forest importance was evaluated on validation folds by repeating raw-column permutation ten times per fold and recording the decrease in R^2 [11]. The Tier 2 protocol was then repeated separately within each vehicle bucket. For the leakage diagnostic, heteroskedasticity-consistent type 3 (HC3) robust fuel-specific OLS regressions related combined L/100 km to CO_2 g/km. Observed slopes were compared with NRCan factors of 2.3 kg CO_2/L for gasoline, 2.7 for diesel, and 1.6 for 85% ethanol fuel (E85) [15]. Multiplication by ten converts these factors to expected g/km per L/100 km. Finally, Tier 1 VIFs and standardized HC3 coefficient intervals were reported because engine size and cylinder count were strongly correlated.

4. Results

4.1. Descriptive patterns

The cleaned sample had mean CO_2 emissions of 253.7 g/km, median 250 g/km, an interquartile range of 210–292 g/km, and a range of 94–608 g/km. Figure 1 shows a right tail and visible fuel-group differences. Median emissions were 230 g/km for regular gasoline, 262 for premium gasoline, 282 for E85, and 263 for diesel. These unadjusted contrasts also reflect group differences in engine size, vehicle class, and market position.

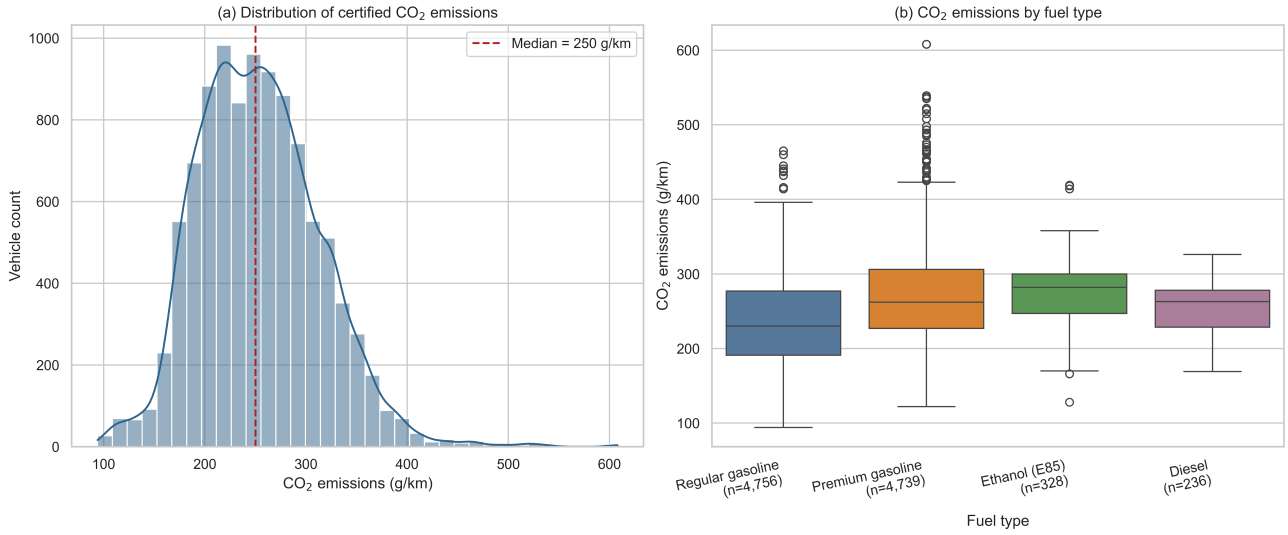


Figure 1. (a) Distribution of certified CO_2 emissions; (b) emissions by fuel type

Data from: [1].

The correlation heatmap in Figure 2 identified substantial information overlap. Combined, city, and highway fuel consumption correlated with emissions at 0.945, 0.940, and 0.911. The three fuel-consumption measures correlated with one another from 0.931 to 0.992. Engine size and cylinders correlated at 0.923, while their correlations with emissions were 0.828 and 0.824.

DOI: 10.54254/2753-8818/2026.PU37781

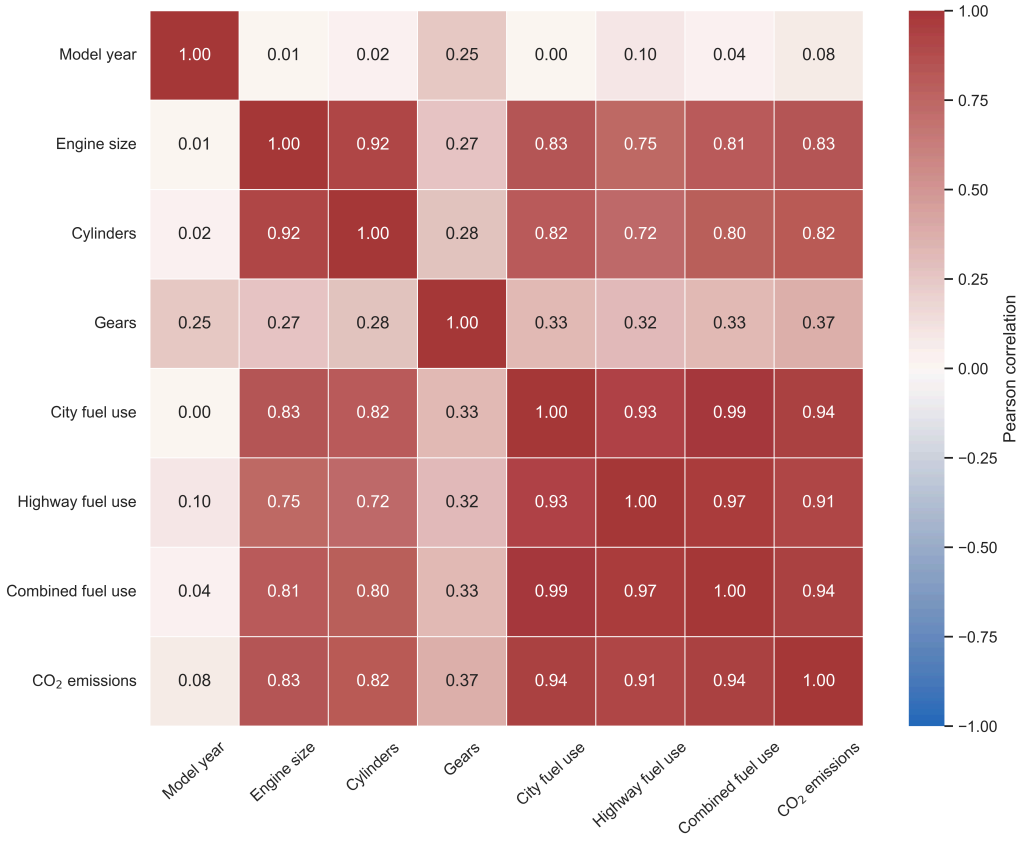


Figure 2. Pearson correlations among the retained numeric variables

Data from: [1].

4.2. Feature-tier performance

Table 3 and Figure 3 show the complete performance ladder. Within Tier 1, random forest achieved R^2=0.776 and MAE=21.56 g/km, modestly outperforming gradient boosting and more clearly outperforming OLS. Adding the full specification set raised random-forest R^2 by 0.122 to 0.898 and reduced MAE by 8.18 g/km. Moving from Tier 2 to Tier 3 raised its R^2 by a further 0.100 to 0.998 and reduced MAE by 12.32 g/km. The information-tier changes were therefore at least as consequential as the choice among the three algorithms.

Table 3. Five-fold performance of the nine feature-tier and model combinations

Tier	Model	R^2	MAE (g/km)	RMSE (g/km)
1	OLS	0.710 ± 0.005	24.74 ± 0.35	32.55 ± 0.35
1	Random forest	0.776 ± 0.007	21.56 ± 0.44	28.59 ± 0.48
1	Gradient boosting	0.774 ± 0.007	21.73 ± 0.42	28.74 ± 0.46
2	OLS	0.800 ± 0.004	20.27 ± 0.35	27.03 ± 0.41
2	Random forest	0.898 ± 0.002	13.38 ± 0.32	19.29 ± 0.19
2	Gradient boosting	0.854 ± 0.005	17.37 ± 0.27	23.08 ± 0.44

DOI: 10.54254/2753-8818/2026.PU37781

Table 3. (continued)

3	OLS	0.996 ± 0.001	2.05 ± 0.07	3.92 ± 0.39
3	Random forest	0.998 ± 0.000	1.06 ± 0.06	2.81 ± 0.28
3	Gradient boosting	0.997 ± 0.001	2.01 ± 0.13	3.30 ± 0.33

Values are fold mean ± sample standard deviation. All combinations used identical fold assignments.

In Figure 3, error bars represent one sample standard deviation across the five folds.

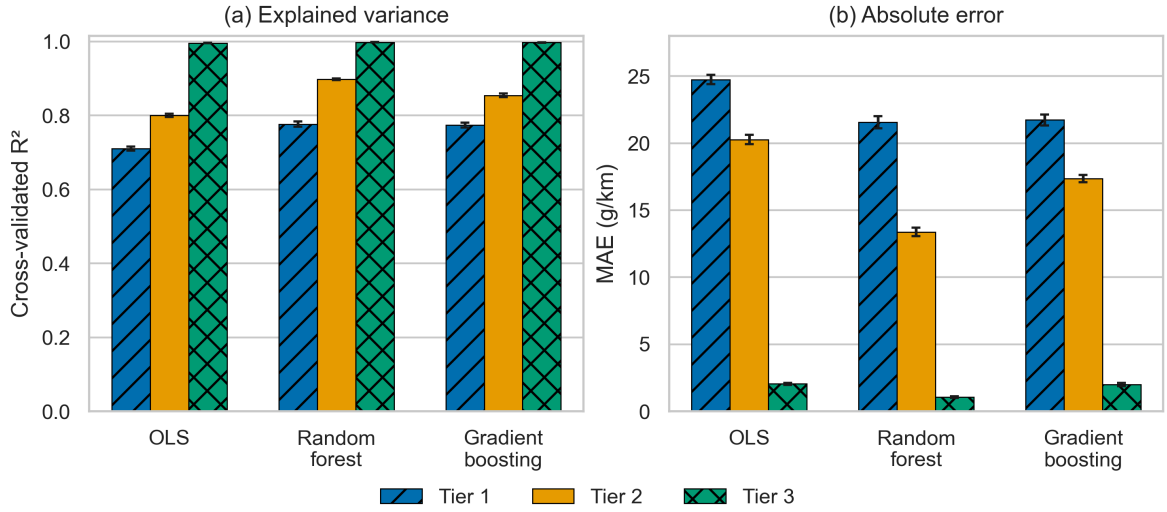


Figure 3. Cross-validated performance across feature tiers and models

Data from: [1].

4.3. Specification importance

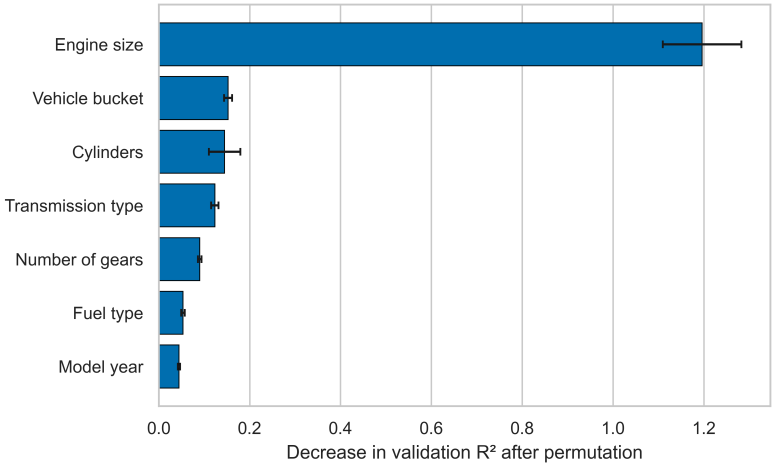


Figure 4. Tier 2 random-forest permutation importance

Data from: [1].

DOI: 10.54254/2753-8818/2026.PU37781

Validation-fold permutation importance for the Tier 2 random forest is shown in Figure 4. Engine size produced the largest mean decrease in R^2 (1.196), followed by vehicle bucket (0.152), cylinders (0.145), transmission type (0.123), gear count (0.090), fuel type (0.053), and model year (0.044). A decrease greater than one is possible because permutation can make validation R^2 negative. These values describe dependence of this fitted model, not causal effects, and the strong engine size–cylinder correlation means their separate magnitudes should not be interpreted in isolation. In Figure 4, error bars represent the standard deviation across the five fold-level permutation estimates.

4.4. Fuel-conditional slopes and leakage

Within each fuel type, combined fuel consumption almost completely determined certified CO_2 emissions (Figure 5). The observed slopes in Table 4 differed from the official theoretical slopes by only −0.67% to 4.17%, and within-fuel R^2 ranged from 0.981 to 0.999. Small deviations are consistent with displayed fuel consumption being rounded to one decimal place, emissions being reported as integer g/km, and details of the certification calculations.

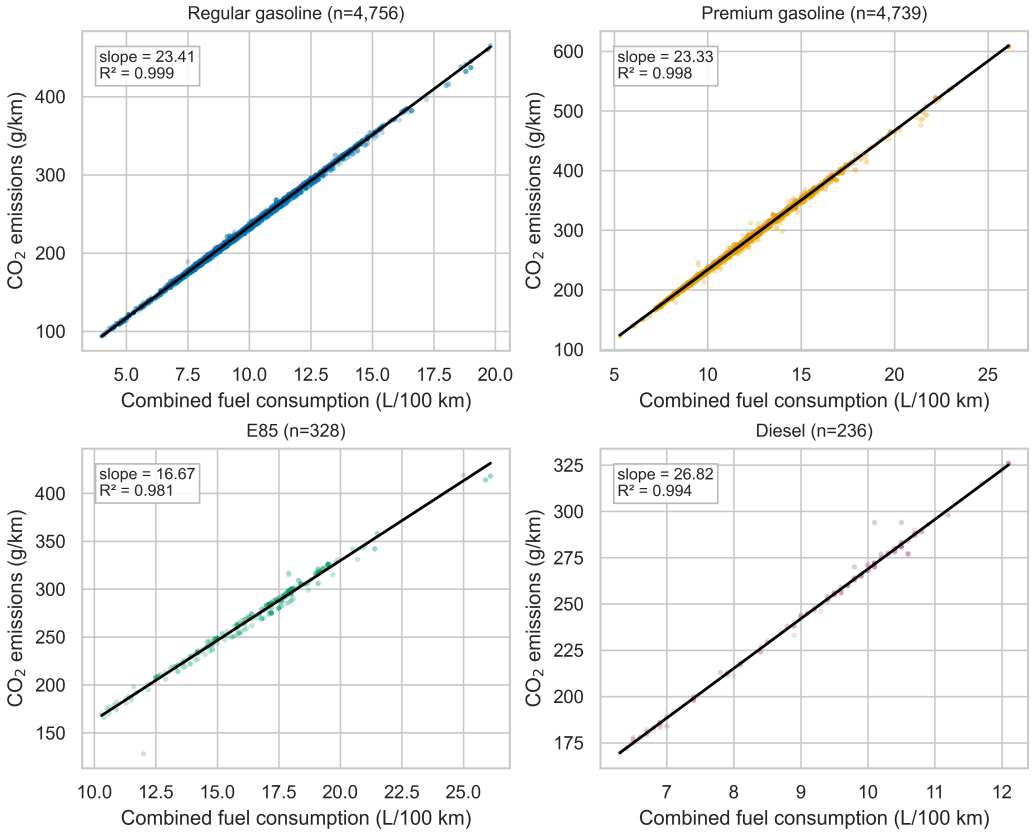


Figure 5. Fuel-specific relationships between combined fuel consumption and certified CO_2 emissions

Data from: [1].

Table 4. Observed fuel-specific slopes compared with NRCan combustion factors

Fuel	n	Slope	HC3 95% CI	Official	Difference	R^2
Regular gasoline	4,756	23.413	[23.387, 23.439]	23.0	+1.79%	0.999

DOI: 10.54254/2753-8818/2026.PU37781

Table 4. (continued)

Premium gasoline	4,739	23.334	[23.293, 23.374]	23.0	+1.45%	0.998
E85	328	16.667	[16.253, 17.081]	16.0	+4.17%	0.981
Diesel	236	26.818	[26.622, 27.014]	27.0	-0.67%	0.994

The Tier 3 result is therefore not conventional preprocessing leakage: validation rows did not influence training-fold transformations. Instead, it is feature–target definition leakage. Once fuel type is known, certified combined fuel consumption is a near-deterministic proxy for the target. Tier 3 should be interpreted as a post-certification reconstruction ceiling, while Tier 1 and Tier 2 address the intended pre-certification prediction question.

4.5. Vehicle-bucket comparison

The Tier 2 specification model did not perform uniformly across vehicle groups (Figure 6), and Table 5 compares the best model in each group with the group's observed CO_2 spread. Random forest was best for cars and SUVs, whereas gradient boosting gave the highest mean R^2 for pickups and the small Other group. The pickup and car RMSEs were similar (20.14 and 19.25 g/km), but the pickup CO_2 standard deviation was only 35.10 g/km compared with 62.72 g/km for cars. Thus, the same absolute error consumed a much larger share of the variation available to explain in pickups, producing their lower best R^2 of 0.664; it does not show that pickups had substantially larger absolute residuals. The Other group reached R^2=0.929, but its five-fold standard deviation was 0.042 and its error bars were comparatively wide because only 243 vehicles were available. In Figure 6, error bars represent one sample standard deviation across folds.

Table 5. Best Tier 2 model within each vehicle bucket

Bucket	n	Model	R^2	MAE (g/km)	RMSE (g/km)	CO_2 SD (g/km)
Car	5,619	Random forest	0.905 ± 0.010	13.16 ± 0.48	19.25 ± 1.06	62.72
SUV	3,099	Random forest	0.866 ± 0.017	13.84 ± 0.72	19.51 ± 1.11	53.47
Pickup	1,098	Gradient boosting	0.664 ± 0.085	14.29 ± 0.67	20.14 ± 2.64	35.10
Other	243	Gradient boosting	0.929 ± 0.042	9.98 ± 1.50	15.91 ± 2.69	65.48

Importance patterns also differed (Figure 7). Engine size ranked first for cars, pickups, and Other vehicles, but cylinder count ranked first for SUVs. Model year and gear count mattered more for pickups than for the other buckets. Vehicle bucket itself had zero within-bucket importance because it was constant after stratification. Thus, the aggregate importance ranking does not represent every vehicle category equally.

DOI: 10.54254/2753-8818/2026.PU37781

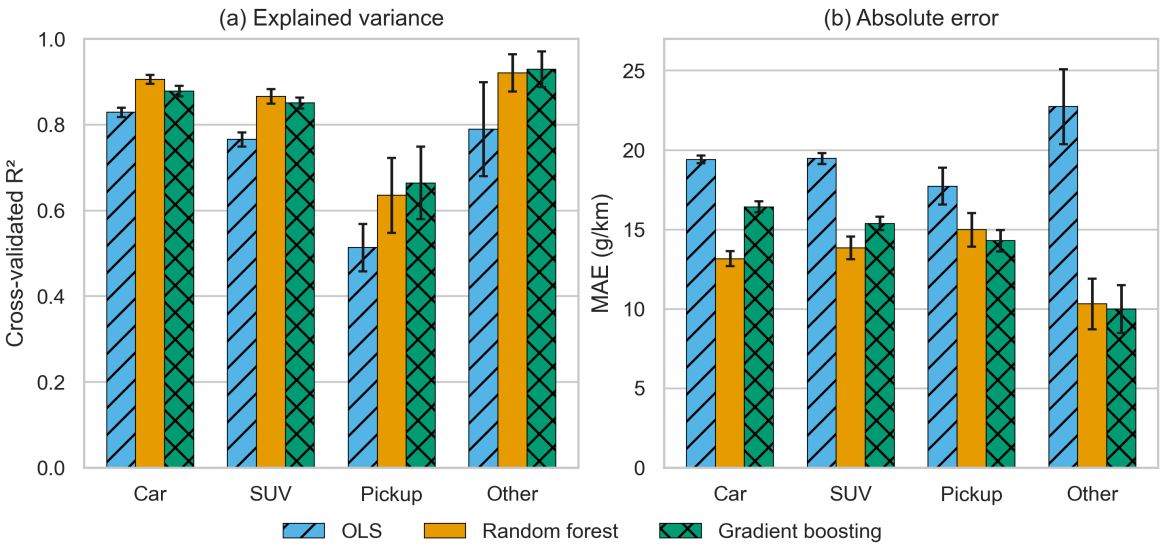


Figure 6. Tier 2 performance within vehicle buckets

Data from: [1].

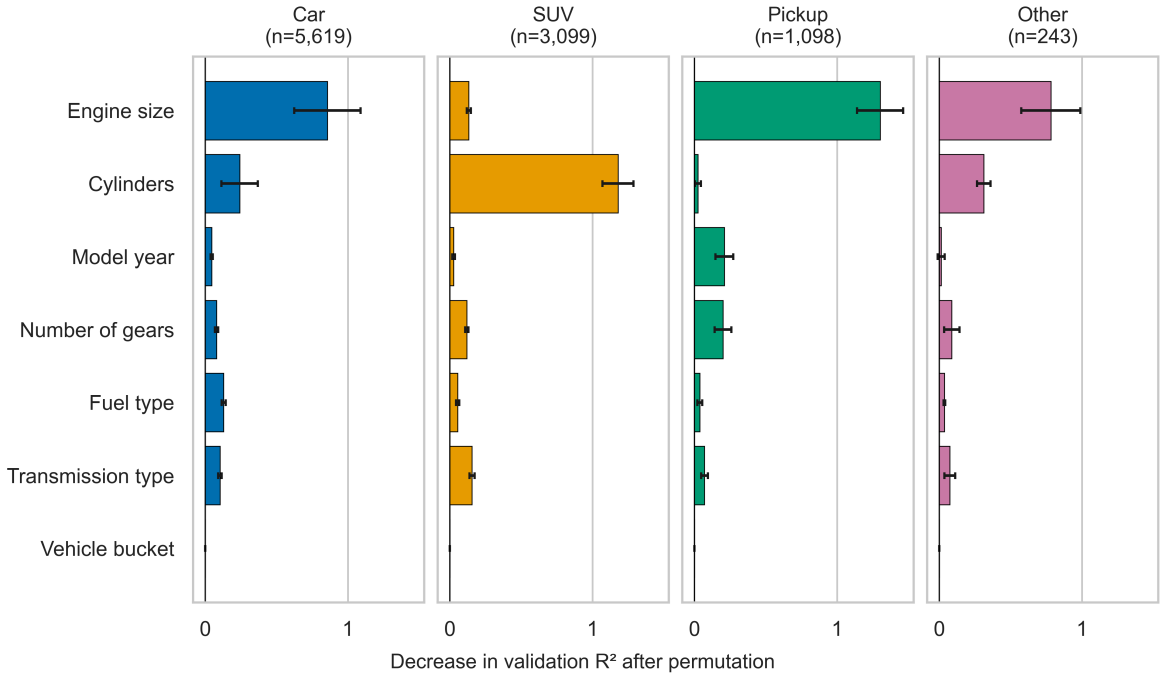


Figure 7. Tier 2 random-forest permutation importance within each vehicle bucket

Data from: [1].

4.6. Tier 1 coefficient stability

Engine size and cylinders had Pearson r=0.923 and VIF=6.76 for each predictor. Rather than interpreting raw coefficients as independent engineering effects, Table 6 reports standardized full-sample associations with HC3 uncertainty. Engine size had β=0.455 and cylinders β=0.404; both intervals excluded zero. The similar magnitudes, combined with elevated VIFs, show that the two

DOI: 10.54254/2753-8818/2026.PU37781

variables carry substantial overlapping information. These coefficients are descriptive and are distinct from the cross-validated prediction results.

Table 6. Tier 1 multicollinearity diagnostics and standardized OLS coefficients

Predictor	VIF	Standardized β	HC3 95% CI
Engine size (L)	6.76	0.455	[0.422, 0.488]
Cylinders	6.76	0.404	[0.369, 0.440]

5. Discussion

5.1. Interpretation

The feature-tier design changes the interpretation of model accuracy. Moving from two mechanical variables to the full manufacturer-sheet set produced a meaningful gain: the Tier 2 random forest explained about 90% of the variance and reduced MAE to 13.38 g/km. The extra fields—vehicle group, transmission type, gear count, fuel type, and model year—are therefore worthwhile when they are available before certification. The decomposed transmission variables were both used by the fitted model, with larger aggregate permutation reliance on transmission type than on gear count.

The Tier 2-to-Tier 3 improvement is different in kind. It is numerically larger than many between-model differences, but the fuel-specific slopes show that the added measurement nearly reconstructs the target using fuel chemistry. The result therefore illustrates why a high R^2 cannot be judged without predictor provenance. For a user who already possesses certified fuel consumption, Tier 3 is a highly accurate emissions conversion. For a designer estimating emissions before testing, it is not a legitimate forecasting benchmark.

The subgroup results further caution against a single global explanation. Although the specification-only R^2 was weakest for pickups, their absolute RMSE was close to that for cars; the lower R^2 primarily reflects the much narrower pickup CO_2 distribution, against which a similar-sized residual represents a larger fraction of total variance. The leading importance also shifted from engine size to cylinders in SUVs. Differences in powertrain packaging, vehicle mass, towing capability, and other omitted variables may still contribute, but the present comparison does not identify them as the cause. The high score in the Other bucket should not be overgeneralized because its small sample produced visibly greater fold uncertainty.

5.2. Limitations and future work

The data represent the Canadian market and model years 2015–2024. Ratings were generated under standardized laboratory cycles, not real-road driving, so driver behaviour, weather, load, and traffic are absent. Battery-electric vehicles were not part of this combustion-fuel table. Manufacturer was excluded, so brand-level structure was not modelled, and the dataset lacked vehicle mass, horsepower, aerodynamics, and hybrid-system detail. Random shuffled folds also evaluate interpolation within the pooled period rather than forward prediction to unseen model years. Future research should use year-blocked and manufacturer-grouped validation, add mass and power variables, and replicate the feature-tier analysis with comparable United States or European certification data. Conditional importance methods could also examine reliance when correlated engine variables are permuted jointly.

DOI: 10.54254/2753-8818/2026.PU37781

6. Conclusion

A feature's source was more consequential than small differences among the three regression algorithms. Minimal engine specifications supported useful prediction, while the complete manufacturer-sheet tier raised the best mean R^2 to 0.898 and reduced MAE to 13.38 g/km. Adding certified fuel consumption pushed R^2 above 0.995, but fuel-conditional slopes matched official combustion factors closely, demonstrating target-proxy leakage rather than a realistic pre-certification gain. Vehicle-bucket analyses showed that both accuracy and predictor reliance varied by category: pickups were less predictable, and SUVs relied most strongly on cylinder count rather than engine size. Splitting transmission codes successfully retained both type and gear information. For practical emissions estimation from public specifications, the Tier 2 random forest is the most defensible overall model in this comparison, whereas Tier 3 should be presented only as a post-test reconstruction ceiling.

References

[1] Natural Resources Canada. (2026). Fuel consumption ratings. *Government of Canada Open Data Portal*. https://open.canada.ca/data/en/dataset/98f1a129-f628-4ce4-b24d-6f16bf24dd64

[2] Natarajan, Y., Wadhwa, G., Sri Preethaa, K. R., & Paul, A. (2023). Forecasting carbon dioxide emissions of light-duty vehicles with different machine learning algorithms. *Electronics, 12*(10), 2288. https://doi.org/10.3390/electronics12102288

[3] Udoh, J., Lu, J., & Xu, Q. (2024). Application of machine learning to predict CO_2 emissions in light-duty vehicles. *Sensors, 24*(24), 8219. https://doi.org/10.3390/s24248219

[4] Vu, H. T. T., & Ko, J. (2024). Effective modeling of CO_2 emissions for light-duty vehicles: Linear and non-linear models with feature selection. *Energies, 17*(7), 1655. https://doi.org/10.3390/en17071655

[5] Guo, X., Kou, R., & He, X. (2024). Towards carbon neutrality: Machine learning analysis of vehicle emissions in Canada. *Sustainability, 16*(23), 10526. https://doi.org/10.3390/su162310526

[6] Liu, H., & Hu, D. (2024). Construction and analysis of a transportation carbon-emission prediction model based on machine learning [translated from Chinese]. *Environmental Science, 45*(6), 3421–3432. https://www.hjkx.ac.cn/hjkx/ch/reader/view_abstract.aspx?file_no=20240627&flag=1

[7] Chen, X., Li, J., Yang, Y., & Hu, D. (2024). Method for estimating carbon emissions from fuel-powered private cars in China based on large-scale user data [translated from Chinese]. *Resources Science, 46*(4), 717–727. https://doi.org/10.18402/resci.2024.04.05

[8] Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns, 4*, 100804. https://doi.org/10.1016/j.patter.2023.100804

[9] Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32. https://doi.org/10.1023/A:1010933404324

[10] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

[11] Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research, 20*(177), 1–81.

[12] O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity, 41*(5), 673–690. https://doi.org/10.1007/s11135-006-9018-6

[13] Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th International Joint Conference on Artificial Intelligence, 2*, 1137–1143.

[14] Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*(85), 2825–2830.

[15] Natural Resources Canada. (2015). 2015 fuel consumption guide. *Government of Canada*. https://natural-resources.canada.ca/sites/www.nrcan.gc.ca/files/oee/pdf/transportation/tools/fuelratings/Model%20Year%202015%20Vehicle%20Tables.pdf