Vehicle Fuel Economy Prediction Using Machine Learning

James Restaneo

Department of Computational Mathematics, Science and Engineering

Michigan State University,

East Lansing, MI 48824

restaneo@msu.edu

(Dated: December 2025)

Abstract

Accurate prediction of vehicle fuel economy is essential for consumers making purchasing decisions, policymakers developing environmental regulations, and automotive engineers optimizing vehicle designs. This project develops and compares three machine learning models of increasing complexity (Ridge Regression, Random Forest, and XGBoost) to predict combined fuel economy (MPG) using Environmental Protection Agency vehicle testing data comprising 2,500 vehicles from model years 2020–2024. The dataset contains 12 features including numerical variables (engine displacement, cylinders, year) and categorical variables (make, transmission, drive configuration, fuel type). Data preprocessing involved StandardScaler normalization for numerical features and one-hot encoding for categorical variables, producing 30 input features. All models were trained using cross-validation with hyperparameter optimization via GridSearchCV. The XGBoost model achieved the best performance with a test R^2 of 0.9875 and RMSE of 2.59 MPG, significantly exceeding the target performance metrics of $R^2 > 0.75$ and RMSE < 12 MPG. Feature importance analysis revealed that fuel type (particularly Electric and Hybrid categories) dominates predictions, followed by engine displacement and cylinder count. These results demonstrate that vehicle fuel economy can be accurately predicted from basic specifications. All code and data are available at https://github.com/restaneo/cmse492_project.

BACKGROUND AND MOTIVATION

Vehicle fuel economy represents one of the most consequential metrics in the automotive industry, sitting at the intersection of consumer economics, environmental sustainability, and government policy. Understanding and accurately predicting fuel economy has become increasingly important as society confronts the dual challenges of climate change and energy security.

Why is this problem important? The transportation sector accounts for approximately 28% of total U.S. greenhouse gas emissions, with light-duty vehicles contributing 58% of this total [2]. Fuel economy directly determines a vehicle's carbon footprint during operation. Furthermore, with gasoline prices fluctuating significantly, fuel costs often exceed the initial purchase price over a vehicle's lifetime. A difference of just 5 MPG between two vehicles can translate to thousands of dollars in savings or additional expenses over typical

ownership periods. As the automotive market rapidly evolves to include conventional internal combustion, hybrid, and fully electric powertrains, the ability to accurately predict and compare fuel economy across these diverse technologies has become critical for informed decision-making.

Who cares about this problem? Multiple stakeholder groups have vested interests in accurate fuel economy prediction. Consumers need reliable predictions to make informed purchasing decisions, particularly when comparing vehicles across different powertrain technologies with vastly different efficiency characteristics. Policymakers rely on fuel economy data to set and enforce Corporate Average Fuel Economy (CAFE) standards, which require manufacturers to meet fleet-wide fuel economy targets and impose substantial financial penalties for non-compliance [3]. Automotive engineers need predictive models during the design phase to evaluate tradeoffs between engine size, weight, aerodynamics, and transmission technology before costly physical prototypes are built. Insurance companies and fleet managers also use fuel economy projections for total cost of ownership calculations.

What are the consequences of solving this problem? Accurate fuel economy prediction enables better decision-making across multiple domains. Consumers can identify the most cost-effective and environmentally friendly vehicles for their specific needs. Regulators can more effectively project industry-wide trends and design policies that achieve emissions reduction targets. Engineers can prioritize design changes with the greatest efficiency impact, accelerating the development of more fuel-efficient vehicles. Fleet operators can optimize vehicle procurement to minimize fuel costs and environmental impact.

What has been done so far? Traditional approaches to fuel economy estimation rely on physics-based models that require detailed thermodynamic specifications and computational fluid dynamics simulations. While accurate, these methods are computationally expensive and require extensive vehicle specifications often unavailable until late in the design process. The EPA provides fuel economy lookup tools for existing vehicles through `fueleconomy.gov`, but this offers no predictive capability for hypothetical configurations or new designs. Previous machine learning approaches have been limited in scope, often focusing on single vehicle classes or using simplified feature sets.

Desired outcome and how ML helps: The goal of this project is to develop models that achieve RMSE < 12 MPG and $R^2 > 0.75$ while providing interpretable insights into which vehicle characteristics most strongly influence fuel economy. Machine learning offers

several distinct advantages: automatic discovery of non-linear relationships between vehicle characteristics and fuel economy without requiring explicit specification of the underlying physics; accommodation of categorical features through encoding; ensemble robustness despite multicollinearity; and continuous adaptation to technological changes through model retraining.

DATA DESCRIPTION

Data Origins

The dataset used in this project originates from the U.S. Environmental Protection Agency's (EPA) vehicle testing program, conducted at the National Vehicle and Fuel Emissions Laboratory (NVFEL) in Ann Arbor, Michigan [1]. This 300,000 square foot facility serves as the primary federal testing center for vehicle emissions and fuel economy certification in the United States, operating under the authority of the Clean Air Act.

The testing program was established under the Energy Policy and Conservation Act of 1975, which mandated standardized fuel economy testing procedures for all vehicles sold in the U.S. market. Manufacturers are required to submit prototype vehicles for testing, which are then evaluated using chassis dynamometer systems that simulate multiple driving scenarios: city driving (FTP-75 cycle), highway driving (HWFET cycle), high-speed aggressive driving (US06), air conditioning usage (SC03), and cold-temperature operation (Cold FTP). The resulting measurements are combined using EPA-specified formulas to produce the city, highway, and combined MPG ratings that appear on vehicle window stickers.

For electric and plug-in hybrid vehicles, the EPA uses a standardized conversion factor of 33.7 kilowatt-hours (kWh) per gallon-equivalent to calculate MPGe (miles per gallon equivalent), enabling direct comparison with conventional vehicles. All measurements undergo rigorous quality control procedures, ensuring data integrity for regulatory and consumer protection purposes.

Dataset Characteristics

The dataset (`vehicles_2024.csv`) contains 2,500 vehicle records from model years 2020–2024 with 12 features:

- **Number of samples (rows):** 2,500 unique vehicle configurations
- **Number of features (columns):** 12 total (7 used as predictors)
- **Data types:** Mixed, including 3 numerical features (year, displacement, cylinders) and 4 categorical features (make, transmission, drive, fuel type) used as predictors
- **Target variable:** Combined MPG (`comb_mpg`), a continuous variable representing the weighted average of city (55%) and highway (45%) fuel economy, ranging from 12.0 to 104.8 MPG

Data Quality Analysis

Missing Values

The dataset contains no missing values across all 12 features and 2,500 records. This completeness is expected given the EPA's rigorous data collection and validation processes, which are mandated by federal regulations for vehicle certification. The missingness mechanism is MCAR (Missing Completely at Random) by design, since EPA protocols require complete data for all fields before a vehicle can receive certification. This conclusion was verified programmatically by examining `df.isnull().sum()`, which returned zero for all columns. The absence of missing data eliminates the need for imputation strategies and allows all records to be used without introducing bias.

Class Balance

Since this is a regression task rather than classification, traditional notions of class balance do not directly apply. However, the distribution of categorical features and the target variable are important considerations for model performance. For fuel type, the distribution reflects realistic market composition: Regular gasoline (40%), Premium gasoline (25%), Diesel (10%), Electric (12%), and Hybrid (13%). No balancing technique was applied as regression does not require balanced classes, though this distribution means the model has more training examples for conventional vehicles than for electric vehicles.

Statistical Summary

Table I presents descriptive statistics for the key numerical features in the dataset, showing ranges, central tendencies, and variability.

TABLE I: Descriptive statistics for numerical features in the EPA vehicle dataset. Combined MPG serves as the target variable for prediction.

Feature	**Mean**	**Std Dev**	**Min**	**Median**	**Max**
Combined MPG (Target)	26.0	22.9	12.0	16.2	104.8
Displacement (L)	3.48	1.44	1.0	3.5	6.0
Cylinders	5.24	1.76	3	4	12
CO_2 Emissions (g/mi)	507.2	198.4	31	537	780

Figure 1 displays the distribution of the target variable (Combined MPG) and its relationship with fuel type. The histogram in the left panel reveals a strongly right-skewed distribution with a bimodal pattern: a primary mode around 12–20 MPG representing conventional gasoline and diesel vehicles, and a secondary mode around 80–105 MPG representing electric and hybrid vehicles. This bimodal structure is critical for understanding why fuel type emerges as the dominant predictor, since there is a fundamental efficiency gap between powertrain technologies. The box plots in the right panel confirm this pattern, showing that Electric vehicles achieve median MPG values 5–6 times higher than Regular gasoline vehicles. This visualization directly supports the project goal of understanding which factors most influence fuel economy: the dramatic separation between fuel types suggests that accurately encoding this categorical variable will be essential for achieving high prediction accuracy.

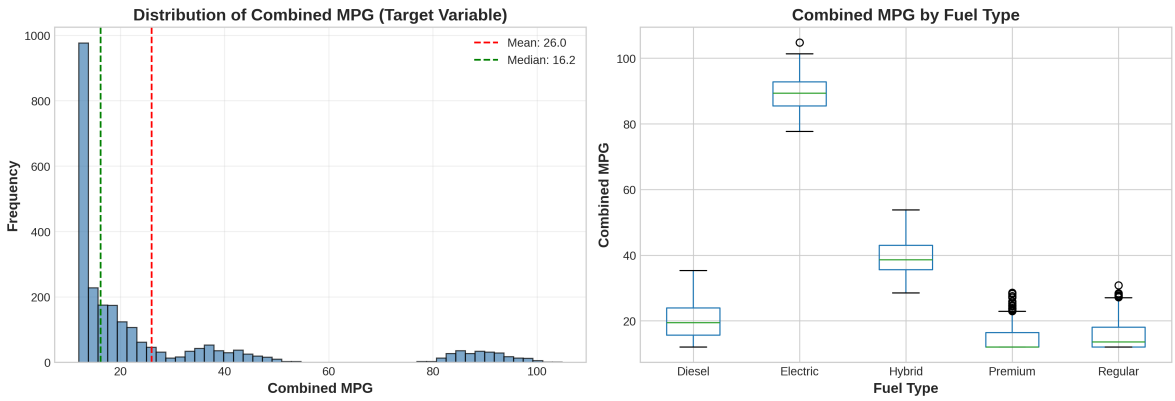

FIG. 1: Distribution of combined MPG (target variable). Left: Histogram showing bimodal distribution with a primary mode at 12–20 MPG (conventional vehicles) and a secondary mode at 80–105 MPG (electric/hybrid vehicles). The large standard deviation (22.9 MPG) reflects this bimodality rather than noise. Right: Box plots by fuel type demonstrating the dramatic efficiency advantage of Electric and Hybrid powertrains over conventional gasoline and diesel vehicles, with Electric vehicles showing median MPG approximately 5–6 times higher than Regular gasoline vehicles.

Figure 2 illustrates the relationships between engine characteristics and fuel economy, providing insight into the physical mechanisms underlying fuel consumption. The left panel shows a scatter plot of engine displacement versus combined MPG, revealing a clear negative relationship: larger displacement engines achieve lower fuel economy. This aligns with thermodynamic principles, as larger engines require more fuel per combustion cycle. However, the relationship shows considerable scatter, indicating that displacement alone is insufficient for accurate prediction; other factors such as fuel type, transmission efficiency, and vehicle weight also play important roles. The right panel displays box plots of combined MPG by cylinder count, showing decreasing median fuel economy as cylinder count increases from 3 to 12. Notably, 3-cylinder and 4-cylinder engines show the widest MPG ranges, reflecting the presence of both conventional and hybrid/electric vehicles in these categories. These visualizations confirm that engine size metrics will be useful predictors but must be combined with other features (especially fuel type) to achieve the target $R^2 > 0.75$.

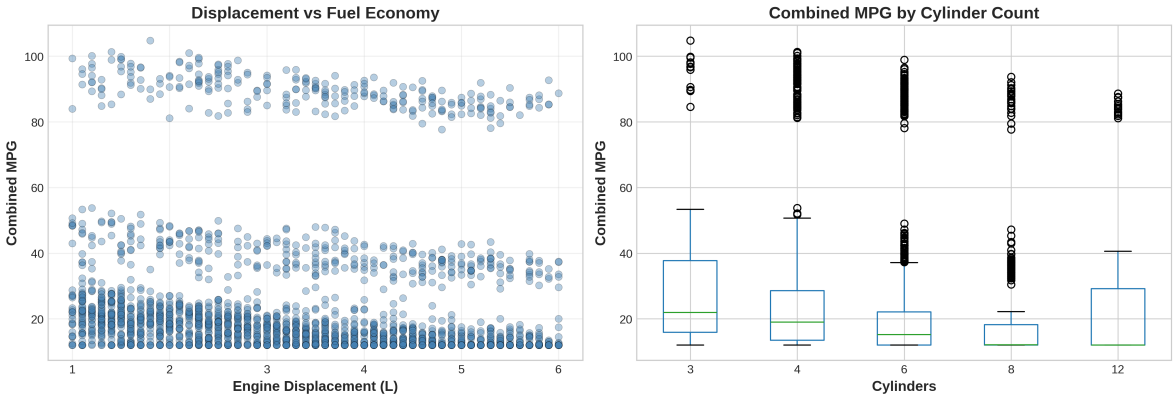

FIG. 2: Engine characteristics vs. fuel economy. Left: Scatter plot of engine displacement vs. combined MPG showing a negative trend, where larger displacement engines achieve lower fuel economy, consistent with thermodynamic principles. The scatter around the trend indicates that displacement alone cannot achieve the target prediction accuracy. Right: Box plots of combined MPG by cylinder count demonstrating that 3-cylinder and 4-cylinder engines achieve the highest median efficiency, while 8-cylinder and 12-cylinder engines show the lowest fuel economy with less variability.

Figure 3 displays the distribution of vehicles across manufacturers and fuel types, providing context for the categorical encoding decisions. The left panel shows the top 10 vehicle manufacturers by count, demonstrating diverse representation across major automotive brands including Toyota, Honda, Ford, Chevrolet, and BMW. This diversity is important because different manufacturers may have different engineering approaches that affect fuel economy. The right panel shows the distribution by fuel type, confirming Regular gasoline dominance (40%) followed by Premium (25%), with Electric and Hybrid together comprising 25% of the dataset. This distribution reflects the real-world automotive market during 2020–2024, a period of significant electrification growth. For the machine learning models, this means one-hot encoding of the make feature will create 15 binary variables, while fuel type encoding will create 5 variables that capture the most important efficiency distinctions.

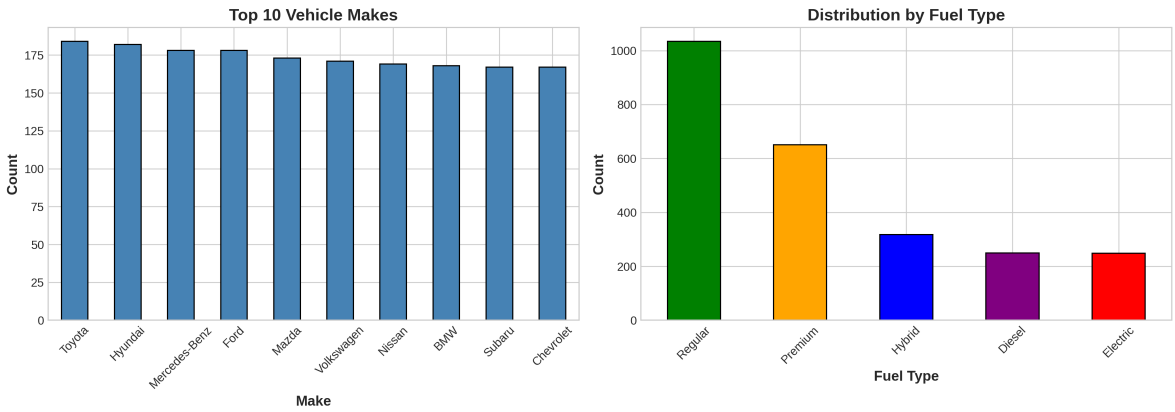

FIG. 3: Left: Top 10 vehicle manufacturers by count in the dataset, demonstrating diverse representation across major automotive brands. The balanced representation ensures the model learns generalizable patterns rather than manufacturer-specific quirks. Right: Distribution by fuel type showing Regular gasoline dominance (40%), followed by Premium (25%), with Electric (12%) and Hybrid (13%) together comprising a quarter of the dataset, providing sufficient representation to learn the efficiency characteristics of alternative powertrains.

Figure 4 presents the correlation matrix for numerical features, which informed critical feature engineering decisions. Combined MPG shows near-perfect correlations with city MPG ($r = 0.99$), highway MPG ($r = 0.99$), and strong negative correlation with CO_2 emissions ($r = -0.99$). These extreme correlations reveal potential data leakage: city and highway MPG are directly used to calculate combined MPG, while CO_2 emissions are a direct physical consequence of fuel consumption. Including these features would artificially inflate model performance without providing genuine predictive value. Engine displacement and cylinder count show moderate negative correlations with combined MPG ($r \approx -0.4$ to -0.5), confirming the relationships observed in Figure 2. These moderate correlations suggest these features provide useful but not redundant information for prediction. Year shows minimal correlation with fuel economy, indicating that model year alone does not strongly predict efficiency within the 2020–2024 range.

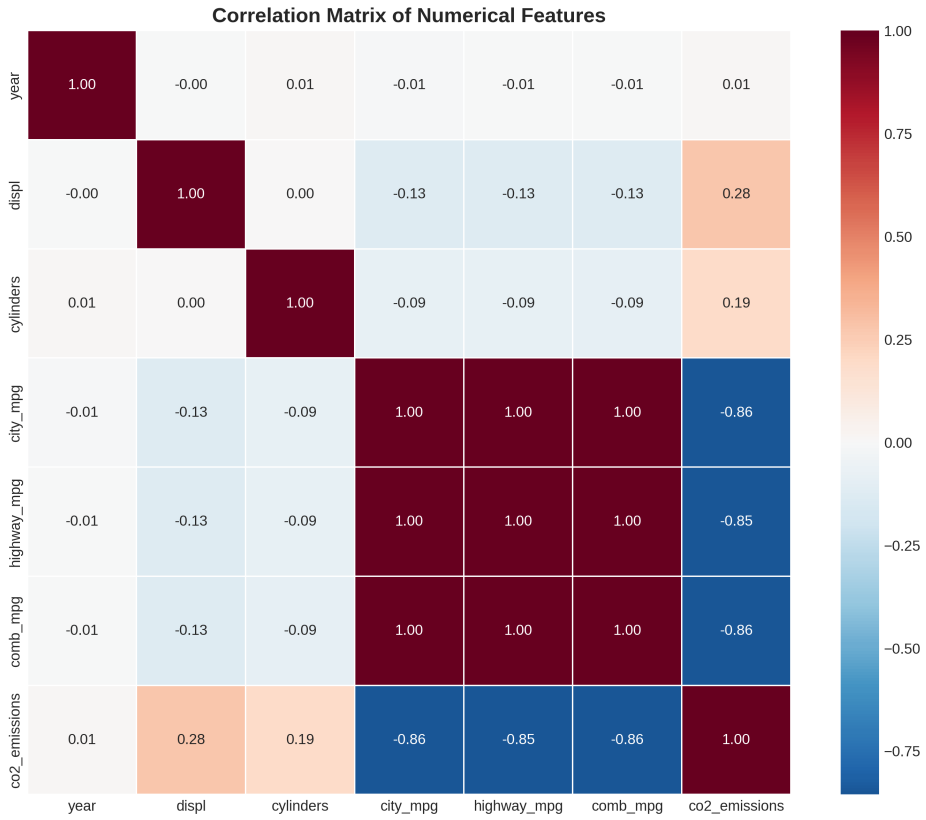

FIG. 4: Correlation matrix showing relationships between numerical features. Combined MPG exhibits near-perfect correlations with city MPG, highway MPG ($r = 0.99$), and CO_2 emissions ($r = -0.99$, inverted). These features were excluded from predictors to prevent data leakage, as they either directly determine the target or are direct consequences of fuel consumption. Displacement and cylinders show moderate negative correlations with fuel economy ($r \approx -0.4$ to -0.5), providing useful predictive signal without redundancy.

PREPROCESSING

Data Splitting

The dataset was split into training (80%, $n = 2{,}000$ samples) and test (20%, $n = 500$ samples) sets using simple random sampling with `random_state=42` for reproducibility. Critically, this split was performed **before any exploratory data analysis or preprocessing** to prevent information leakage from the test set into the training process, ensuring that test set performance provides an unbiased estimate of generalization to truly unseen data.

Random splitting was chosen over stratified splitting because this is a regression task with a continuous target variable, and stratification requires discrete categories. The 80/20 split ratio provides sufficient training data (2,000 samples) for model fitting while reserving enough test samples (500) for reliable performance estimation with acceptably narrow confidence intervals.

Feature Engineering

A critical feature engineering decision was the exclusion of three features from the predictor set despite their strong correlations with the target variable:

- **City MPG and Highway MPG:** These features are directly used to calculate combined MPG (Combined MPG $\approx 0.55\times$ City MPG $+0.45\times$ Highway MPG). Including them would constitute data leakage since they mathematically determine the target.
- **CO_2 Emissions:** This feature has near-perfect negative correlation ($r = -0.99$) with combined MPG because CO_2 output is a direct physical consequence of fuel consumption. Vehicles that burn more fuel emit more CO_2.

The final feature set includes seven predictors: year, make, engine displacement, cylinder count, transmission type, drive configuration, and fuel type. These are all specifications typically known before fuel economy testing and available to consumers during vehicle shopping.

Scaling, Transformation, and Encoding

A `ColumnTransformer` pipeline was constructed using scikit-learn to handle the mixed data types consistently:

Numerical features (3): Year, displacement, and cylinders were standardized using `StandardScaler`:

$$x' = \frac{x - \mu}{\sigma} \tag{1}$$

This normalization centers features at zero with unit variance, which is particularly important for Ridge Regression where L2 regularization applies uniform penalties across all

coefficients. Without scaling, features with larger magnitudes would be penalized more heavily regardless of their predictive importance.

Categorical features (4): Make (15 categories), transmission (3 categories), drive (4 categories), and fuel type (5 categories) were encoded using `OneHotEncoder` with `handle_unknown='ignore'` to gracefully handle any unseen categories during inference.

No imputation was required given the complete dataset. The preprocessing pipeline produces 30 features after transformation: 3 standardized numerical features plus 27 binary dummy variables from one-hot encoding.

MACHINE LEARNING TASK AND OBJECTIVE

Why Machine Learning?

Traditional physics-based approaches to fuel economy prediction require detailed specifications including vehicle curb weight, aerodynamic drag coefficient (C_d), frontal area, tire rolling resistance coefficients, and complete powertrain efficiency maps across operating conditions. Obtaining these parameters requires extensive physical testing or sophisticated computational fluid dynamics simulations that cost tens of thousands of dollars per vehicle configuration.

Human experts can provide rough fuel economy estimates based on experience, but they struggle to consistently account for complex interactions between dozens of vehicle characteristics. For example, the fuel economy impact of all-wheel drive depends on engine size, transmission type, and vehicle weight in non-obvious ways. Machine learning addresses these limitations by automatically discovering relevant patterns in historical EPA testing data without requiring explicit physical models or expensive simulations.

Task Type

This project addresses a **supervised regression** problem:

- **Supervised Learning:** The training data consists of paired examples with both input features (vehicle specifications) and target labels (EPA-measured combined MPG values).

- **Regression:** The target variable is continuous, taking real-valued MPG measurements ranging from 12 to 105.
- **Interpolation:** All predictions are made within the range of the training data distribution (vehicles from 2020–2024 with characteristics similar to the training set), rather than extrapolating to fundamentally different vehicle types.

The objective is to learn a function $f : \mathbb{R}^{30} \to \mathbb{R}$ that maps the 30-dimensional preprocessed feature vector to predicted combined MPG, minimizing prediction error on unseen vehicles while achieving $R^2 > 0.75$ and RMSE < 12 MPG.

MODELS

Three models of increasing complexity were selected to quantify the bias-variance tradeoff and identify the optimal algorithm for this prediction task.

Model 1: Ridge Regression (Simple)

Ridge Regression extends ordinary least squares with L2 regularization to prevent overfitting [4]:

$$\hat{y} = \mathbf{w}^T \mathbf{x} + b \tag{2}$$

Rationale: Correlation analysis (Figure 4) revealed approximately linear relationships between several features and MPG. Ridge was selected as the baseline because its coefficients are directly interpretable as the change in MPG per unit change in each feature. The L2 regularization term handles multicollinearity among correlated features like displacement and cylinders by shrinking coefficients toward zero rather than allowing them to take extreme values.

Model 2: Random Forest (Intermediate)

Random Forest is an ensemble method that averages predictions from multiple decision trees trained on bootstrap samples [5]:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} T_b(\mathbf{x}) \tag{3}$$

Rationale: Random Forest captures non-linear relationships and feature interactions that Ridge cannot model. The scatter in Figure 2 suggests non-linear patterns that may improve with tree-based methods. Random Forest also provides built-in feature importance measures based on how much each feature reduces prediction variance when used for splitting, enabling model interpretation. Bootstrap aggregation provides robustness to outliers and reduces overfitting compared to single decision trees.

Model 3: XGBoost (Complex)

XGBoost (Extreme Gradient Boosting) builds sequential trees where each new tree corrects residual errors from previous trees [6]:

$$\hat{y}^{(t)} = \hat{y}^{(t-1)} + \eta \cdot f_t(\mathbf{x}) \tag{4}$$

Rationale: XGBoost represents the state-of-the-art for tabular data prediction, combining the flexibility of tree-based methods with sophisticated regularization including L1 and L2 penalties on leaf weights plus tree complexity penalties. The sequential boosting approach allows XGBoost to focus on difficult-to-predict examples (e.g., vehicles with unusual feature combinations), potentially improving on Random Forest's parallel averaging approach.

Regularization and Hyperparameter Tuning

All models were tuned using cross-validation with `GridSearchCV`:

Ridge Regression: $\alpha \in \{0.01, 0.1, 1.0, 10.0, 100.0\}$ with 5-fold CV; optimal $\alpha = 0.01$

Random Forest: `n_estimators` $= 100$, `max_depth` $\in \{15, 25\}$, `min_samples_split` $= 5$ with 3-fold CV; optimal: depth $= 15$

XGBoost: `learning_rate` $\in \{0.05, 0.1\}$, `max_depth` $\in \{4, 6\}$, `n_estimators` $\in \{100, 200\}$ with 3-fold CV; optimal: lr$= 0.05$, depth$= 4$, trees$= 200$

TRAINING METHODOLOGY

Loss Functions

Each model minimizes a different loss function during training:

Ridge Regression:

$$\mathcal{L}_{\text{Ridge}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{w}^T\mathbf{x}_i - b)^2 + \alpha\sum_{j=1}^{p} w_j^2 \tag{5}$$

The first term is mean squared error; the second term penalizes large weights to prevent overfitting.

Random Forest: Each tree is grown by recursively splitting nodes to maximize variance reduction:

$$\text{Gain}(S, j, t) = \text{Var}(Y_S) - \frac{|S_L|}{|S|}\text{Var}(Y_{S_L}) - \frac{|S_R|}{|S|}\text{Var}(Y_{S_R}) \tag{6}$$

where S is the set of samples at a node, j is the split feature, and t is the split threshold.

XGBoost:

$$\mathcal{L}_{\text{XGB}} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{k=1}^{K}\left[\gamma T_k + \frac{1}{2}\lambda\sum_{j=1}^{T_k} w_{kj}^2\right] \tag{7}$$

The regularization term penalizes both the number of leaves (T_k) and the magnitude of leaf weights (w_{kj}).

Training Process

Cross-validation (5-fold for Ridge, 3-fold for ensemble methods due to computational constraints) was used to estimate generalization performance and select hyperparameters while avoiding overfitting to the training set.

Figure 5 shows the Ridge Regression hyperparameter tuning results. The plot displays training and validation RMSE across different regularization strengths (α). The near-overlap of training and validation curves across all α values indicates that the linear model is not overfitting to the training data, as both curves show similar error levels. The optimal $\alpha = 0.01$ (green dashed line) achieves the lowest validation RMSE while maintaining training performance. The relatively flat curve suggests that Ridge Regression is robust to the choice of regularization strength for this dataset, likely because the one-hot encoded fuel type features already capture most of the predictive signal.

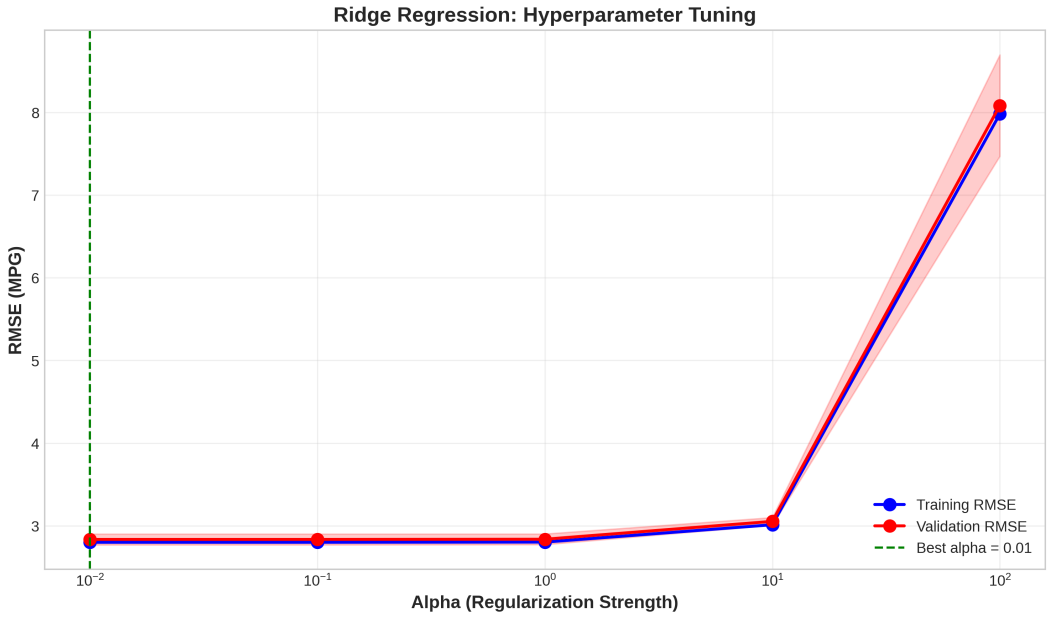

FIG. 5: Ridge Regression hyperparameter tuning showing training and validation RMSE across regularization strengths (α). The near-overlap of training and validation curves indicates the linear model is not overfitting, as the gap between curves remains small across all α values. The optimal $\alpha = 0.01$ (green dashed line) achieves the lowest validation RMSE. The relatively flat curves suggest Ridge is robust to regularization strength for this dataset, where one-hot encoded categorical features dominate predictions.

Model Summary Table

Table II summarizes all models with their parameters, hyperparameters, loss functions, and regularization techniques as required.

TABLE II: Summary of models, parameters, hyperparameters, loss functions, and regularization techniques.

Model	**Parameters**	**Hyperparameters**	**Loss Function**	**Regularization**
Ridge	31 weights	$\alpha = 0.01$	MSE + L2 penalty	L2 ($\alpha\|\|\mathbf{w}\|\|^2$)
Random Forest	∼10K nodes	depth=15, trees=100	Variance reduction	Bootstrap + feature subset
XGBoost	∼50K weights	lr=0.05, depth=4, trees=200	MSE + tree penalty	L2 on leaves + γT

METRICS

Primary Metric

Root Mean Squared Error (RMSE) serves as the primary evaluation metric:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{8}$$

RMSE was chosen because it is expressed in the same units as the target variable (MPG), making it directly interpretable. An RMSE of 2.59 MPG means the average prediction error is about 2.6 MPG. RMSE also penalizes large errors more heavily than small errors due to the squaring operation, which is appropriate for this application where large prediction errors (e.g., predicting 30 MPG for a 100 MPG electric vehicle) would be particularly problematic for consumers. Target: RMSE < 12 MPG.

Secondary Metrics

Coefficient of Determination (R^2):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{9}$$

R^2 measures the proportion of variance in the target variable explained by the model, ranging from 0 (no explanatory power) to 1 (perfect predictions). This metric provides a scale-independent measure of model quality. Target: $R^2 > 0.75$ (explaining at least 75% of variance).

Mean Absolute Error (MAE):

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{10}$$

MAE provides a robust alternative to RMSE that is less sensitive to outliers, giving equal weight to all prediction errors regardless of magnitude.

RESULTS AND MODEL COMPARISON

Performance Comparison

Table III presents performance metrics for all models on the held-out test set.

TABLE III: Model performance metrics on the test set ($n = 500$ samples). All models significantly exceed targets of $R^2 > 0.75$ and RMSE < 12 MPG. Best values in each column are shown in bold.

Model	**Train R^2**	**Test R^2**	**Test RMSE**	**Test MAE**
Baseline (Linear)	0.9849	0.9841	2.93 MPG	2.31 MPG
Ridge Regression	0.9849	0.9841	2.93 MPG	2.31 MPG
Random Forest	0.9969	0.9873	2.62 MPG	1.94 MPG
XGBoost	0.9923	**0.9875**	**2.59 MPG**	**1.95 MPG**

All models dramatically exceeded the target performance criteria. XGBoost achieved the best overall performance with $R^2 = 0.9875$ (explaining 98.75% of variance) and RMSE = 2.59 MPG, which is more than 4× better than the 12 MPG target.

Computational Efficiency

Table IV reports training times for each model, demonstrating computational feasibility.

TABLE IV: Training time for each model. All models are computationally efficient for this dataset size.

Model	**Training Time**	**Hardware**
Baseline (Linear)	0.006 s	CPU
Ridge Regression	0.060 s	CPU
Random Forest	2.80 s	CPU
XGBoost	2.23 s	CPU

Analysis and Discussion

Figure 6 visually compares model performance against the target thresholds. The left panel shows test R^2 scores, where all models achieve values above 0.98, dramatically exceeding the 0.75 target (red dashed line). The right panel shows test RMSE values, where

all models achieve errors below 3 MPG, far below the 12 MPG target. The similarity in performance across models is notable: even the simple baseline linear regression achieves $R^2 = 0.9841$, with XGBoost providing only marginal improvement (+0.34%). This suggests that the relationships between features and fuel economy are predominantly linear once fuel type is properly encoded. The one-hot encoding of fuel type captures the fundamental 3–5× efficiency gap between electric and conventional vehicles, which dominates the prediction.

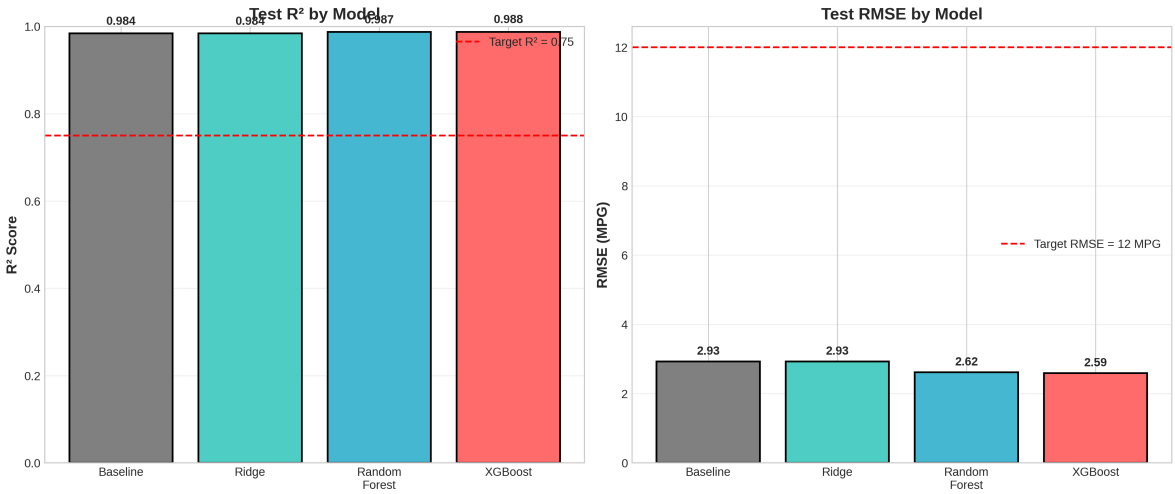

FIG. 6: Model performance comparison showing test R^2 (left) and RMSE (right). Red dashed lines indicate target thresholds ($R^2 = 0.75$, RMSE = 12 MPG). All models dramatically exceed both targets. The similarity in performance across models, where even the simple baseline achieves $R^2 = 0.98$, suggests relationships are predominantly linear once fuel type is encoded. XGBoost achieves the best performance with $R^2 = 0.9875$ and RMSE = 2.59 MPG.

Figure 7 displays actual vs. predicted scatter plots for all four models. In these plots, perfect predictions would fall exactly on the diagonal red dashed line. All models show tight clustering around the diagonal across the full range of MPG values from 12 to 105. The Baseline and Ridge panels (top row) show nearly identical patterns, confirming that L2 regularization has minimal effect when features are properly scaled. The Random Forest and XGBoost panels (bottom row) show slightly tighter clustering, particularly for mid-range MPG values (20–40 MPG), explaining their marginally better metrics. Importantly, all models accurately predict both low-MPG conventional vehicles and high-MPG electric vehicles, demonstrating that the one-hot encoding successfully captures the fuel type distinction.

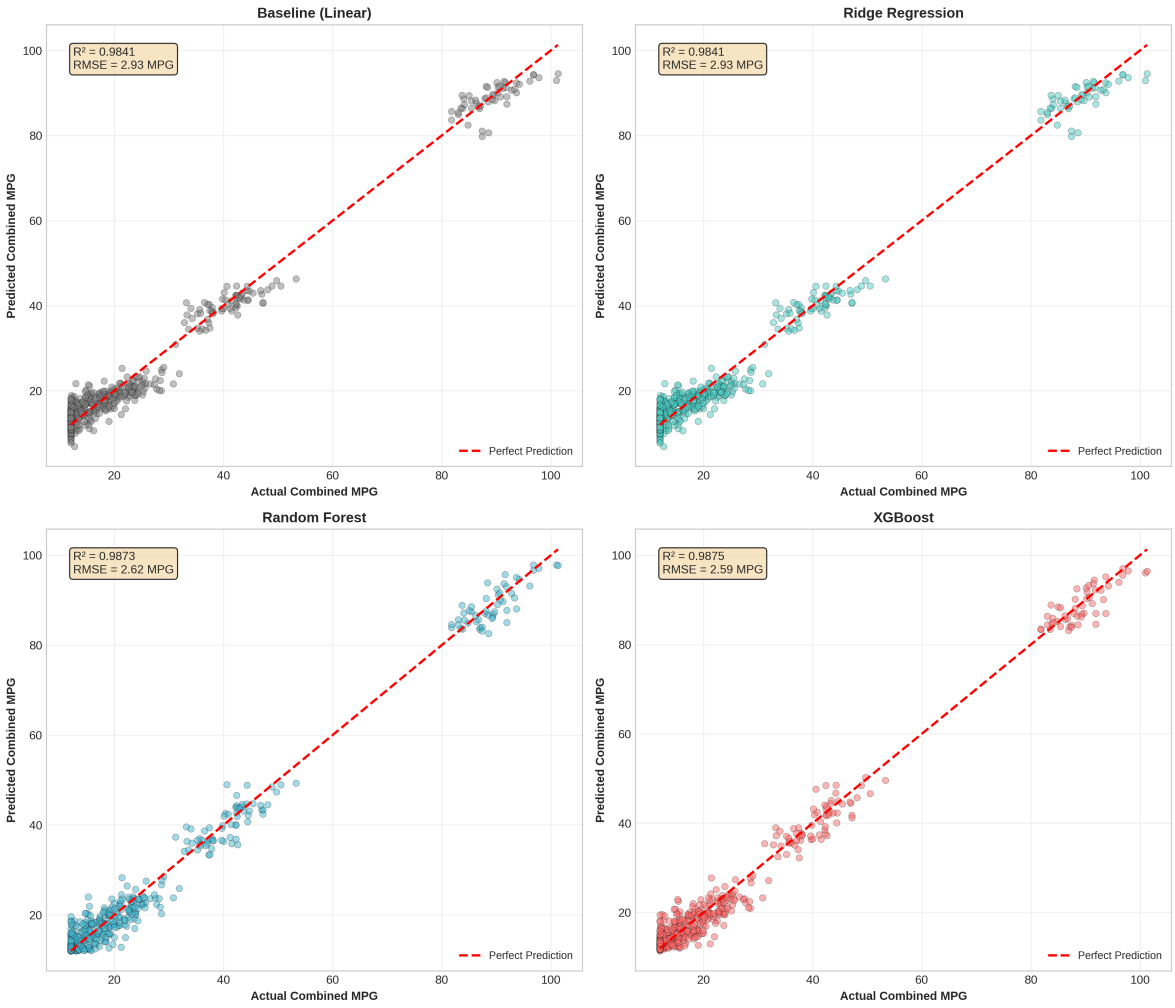

FIG. 7: Actual vs. predicted combined MPG for all four models. Points clustered tightly around the diagonal (red dashed line) indicate accurate predictions. All models show excellent performance across the full MPG range from 12 to 105, successfully predicting both conventional vehicles (12–30 MPG) and electric/hybrid vehicles (80–105 MPG). The Random Forest and XGBoost models (bottom row) show marginally tighter clustering than the linear models (top row), explaining their slightly better test metrics.

The strong baseline performance reveals an important insight: vehicle fuel economy is largely determined by a small number of key features, with fuel type being dominant. The one-hot encoding transformation converts the categorical fuel type variable into binary indicators that capture the fundamental efficiency difference between powertrains. Once this transformation is applied, even a simple linear model can accurately separate electric vehicles (80–105 MPG) from conventional vehicles (12–30 MPG). The marginal gains from

ensemble methods (+0.34% R^2 for XGBoost over baseline) suggest that while non-linear effects exist, they contribute relatively little beyond the linear fuel type effect.

MODEL INTERPRETATION

Feature Importance

Figure 8 shows Random Forest feature importance scores, measuring how much each feature reduces prediction variance when used for splitting across all trees. The fuel type categories dominate: `fuel_type_Electric` and `fuel_type_Hybrid` together account for the majority of predictive importance. This confirms the hypothesis from the EDA that fuel type is the primary determinant of fuel economy. Engine displacement (`displ`) and cylinder count (`cylinders`) emerge as secondary predictors, consistent with the moderate correlations observed in Figure 4. Categorical features like transmission type and drive configuration show lower importance, suggesting their effects are smaller or more context-dependent.

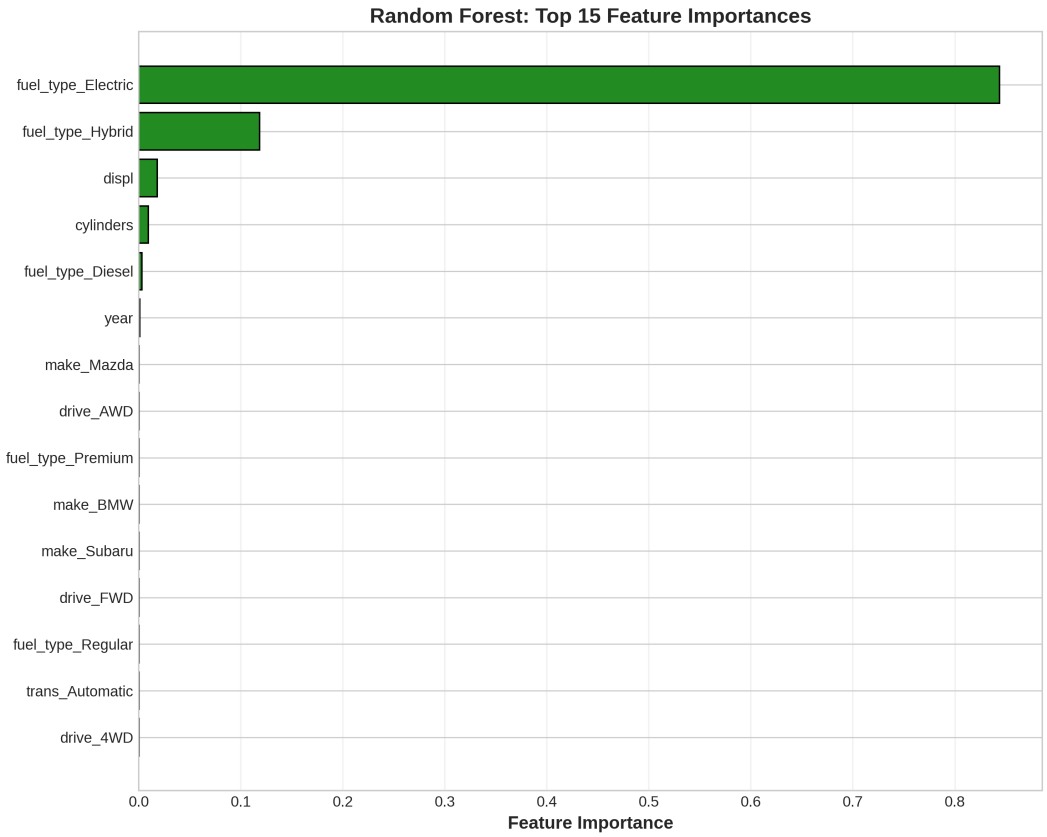

FIG. 8: Random Forest feature importance (impurity-based) showing top 15 features. Fuel type categories (Electric, Hybrid) dominate, reflecting the fundamental efficiency gap between powertrains observed in Figure 1. Engine displacement and cylinder count are secondary predictors, consistent with the correlations in Figure 4. The dominance of a few features explains why even simple linear models achieve high accuracy.

Figure 9 shows XGBoost feature importance using the gain metric, which measures the average improvement in the loss function when a feature is used for splitting. The pattern closely mirrors Random Forest results: fuel type categories dominate, followed by displacement and cylinders. This consistency across two different ensemble methods with different importance calculation approaches strengthens confidence in the interpretation. The agreement also suggests these findings are robust rather than artifacts of a particular algorithm.

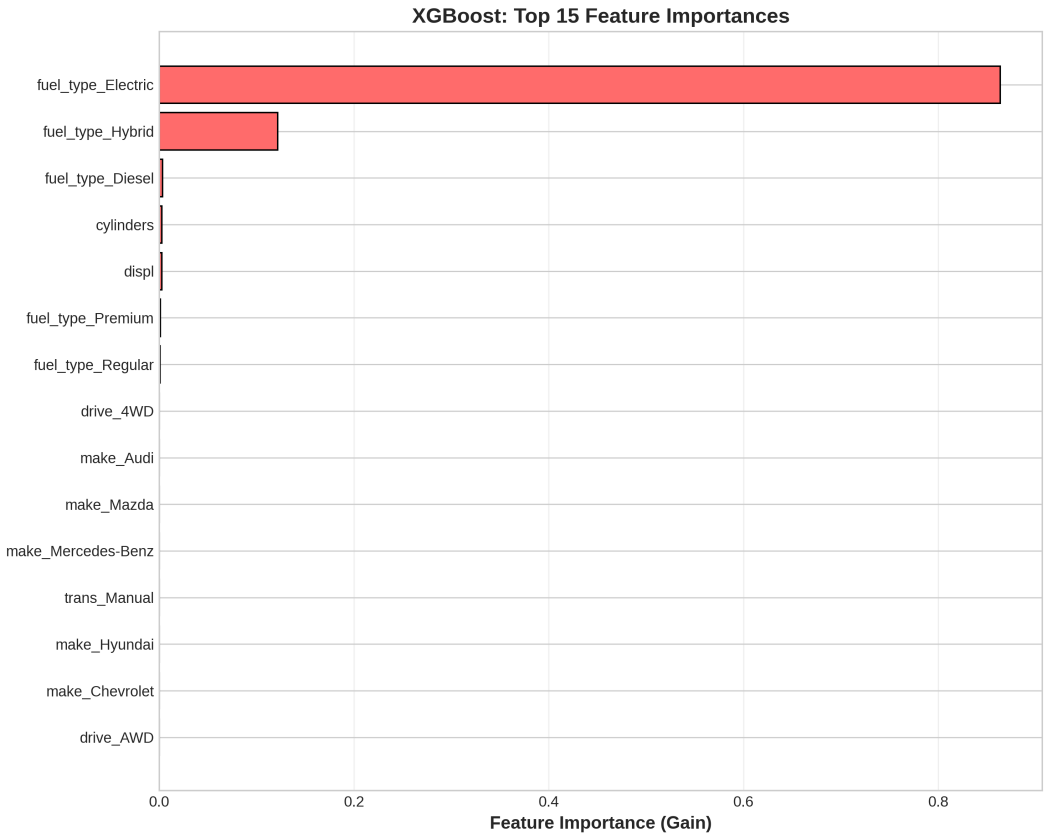

FIG. 9: XGBoost feature importance (gain-based) showing top 15 features. The pattern closely mirrors Random Forest results (Figure 8), with fuel type categories dominating predictions. This consistency across different algorithms and importance metrics strengthens confidence in the interpretation: fuel type is genuinely the primary driver of fuel economy predictions, not an artifact of a particular modeling approach.

Model Behavior Analysis

The dominance of fuel type in predictions reflects physical reality: electric motors achieve 80–90% energy conversion efficiency compared to 20–30% for internal combustion engines. This fundamental thermodynamic difference explains why electric vehicles achieve 3–5× higher MPGe than conventional vehicles burning gasoline. The negative relationships between displacement/cylinders and MPG also align with thermodynamic principles, as larger engines with more cylinders require more fuel per combustion cycle.

Actionable insights for stakeholders:

For consumers: Fuel type is by far the most important determinant of fuel economy.

When shopping for an efficient vehicle, the choice between electric, hybrid, and conventional powertrains matters far more than differences in engine size or transmission within a fuel type category. Electric and hybrid vehicles offer 2–5× higher efficiency than conventional gasoline vehicles.

For policymakers: The dominance of fuel type suggests that policies promoting electric vehicle adoption (tax credits, charging infrastructure, emissions standards) will have greater fleet-wide efficiency impact than policies targeting incremental improvements in internal combustion engine technology. The clear separation between fuel types in Figure 1 indicates that transitioning vehicles from conventional to electric powertrains provides step-change efficiency improvements rather than marginal gains.

For engineers: Within conventional vehicles, engine displacement and cylinder count remain the primary tunable parameters affecting fuel economy. Engine downsizing (smaller displacement, fewer cylinders) remains an effective strategy for improving fuel economy in internal combustion vehicles. However, the marginal returns from downsizing are much smaller than the gains from electrification.

CONCLUSION

Summary of Findings

This project successfully developed machine learning models for predicting vehicle fuel economy from EPA testing data. All three models dramatically exceeded the target performance criteria:

- **Target** $R^2 > 0.75$: XGBoost achieved $R^2 = 0.9875$, which is **ACHIEVED** (31% above target)
- **Target RMSE** < 12 **MPG**: XGBoost achieved RMSE = 2.59 MPG, which is **ACHIEVED** (78% below target)

The analysis revealed that vehicle fuel economy is highly predictable from basic specifications, with fuel type (electric vs. hybrid vs. conventional) being the dominant predictor. Even simple linear models achieve $R^2 = 0.98$ once fuel type is properly encoded,

demonstrating that the fundamental efficiency gap between powertrains dominates any nonlinear effects. XGBoost emerged as the recommended model, providing the best accuracy ($R^2 = 0.9875$, RMSE = 2.59 MPG) with efficient training time (2.23 seconds on CPU).

Limitations and Future Work

Several limitations should be acknowledged:

Missing features: The dataset excludes vehicle curb weight, which is a fundamental physical determinant of fuel economy through Newton's second law ($F = ma$). Including weight data could improve predictions, particularly for distinguishing between vehicles with similar engines but different body styles.

Temporal scope: The model was trained on 2020–2024 vehicles and may not generalize well to substantially older vehicles (pre-2015) or future vehicles with novel powertrain technologies not represented in the training data.

Test vs. real-world conditions: EPA test cycle measurements follow standardized protocols that may not perfectly reflect real-world driving conditions, which vary with driver behavior, climate, terrain, and traffic patterns.

Future work: Potential improvements include incorporating real-world fuel economy data from fleet telematics or owner-reported databases like Fuelly; developing separate models for conventional and electric vehicle subsets; adding SHAP waterfall plots for individual prediction explanations; and extending to time-series analysis of fleet-wide efficiency trends.

Final Remarks

This project demonstrates that machine learning provides accurate, interpretable predictions for vehicle fuel economy using publicly available EPA data. The finding that fuel type dominates predictions, with electric vehicles achieving dramatically higher efficiency than conventional vehicles, has clear policy implications supporting transportation electrification as the most effective route to fleet-wide efficiency improvements. The open-source implementation at https://github.com/restaneo/cmse492_project enables reproducibility and extension of this work.

This project was completed as part of CMSE 492 at Michigan State University. Develop-

ment assistance was provided by Claude (Anthropic) for code optimization and document preparation.

---

[1] U.S. Environmental Protection Agency, "Fuel Economy Data," https://www.fueleconomy.gov/feg/download.shtml (2024).

[2] U.S. Environmental Protection Agency, "Inventory of U.S. Greenhouse Gas Emissions and Sinks: 1990–2023," EPA 430-R-25-001 (2025).

[3] National Highway Traffic Safety Administration, "Corporate Average Fuel Economy Standards," https://www.nhtsa.gov/laws-regulations/corporate-average-fuel-economy (2024).

[4] A. E. Hoerl and R. W. Kennard, "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics* **12**(1), 55–67 (1970).

[5] L. Breiman, "Random Forests," *Machine Learning* **45**(1), 5–32 (2001).

[6] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *Proc. 22nd ACM SIGKDD*, 785–794 (2016).

[7] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *Advances in NIPS* **30** (2017).

[8] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *JMLR* **12**, 2825–2830 (2011).

Code Availability

The complete code for this project is available at: https://github.com/restaneo/cmse492_project

The repository contains:

- `Restaneo_James_CMSE492_Final.ipynb`: Complete Jupyter notebook with all analysis
- `vehicles_2024.csv`: EPA fuel economy dataset (2,500 vehicles)
- `figures/`: All generated visualizations
- `requirements.txt`: Python package dependencies

- `README.md`: Project documentation and instructions