Vehicle Fuel Economy Modeling Results

This write-up summarizes the notebook results for 49,846 vehicles spanning model years 1984 to 2026. The workflow combined data cleansing, feature engineering, classification, regression, and deployment-focused analysis on the FuelEconomy.gov dataset. Key preprocessing steps standardized drivetrain, transmission, and fuel labels; extracted gear count; engineered turbo, supercharger, start-stop, electrified, AWD, and vehicle-age features; and excluded leakage-prone variables such as fuel cost, petroleum barrels, and emissions from the MPG prediction inputs.

Figure 1 shows the main structure in the data. Electric vehicles occupy the highest combined MPG / MPGe range, efficiency generally improves in newer model years, and larger engines correspond to lower efficiency. These visual patterns explain why year, fuel group, displacement, cylinders, drivetrain, and electrification all appear among the strongest predictors later in the analysis.

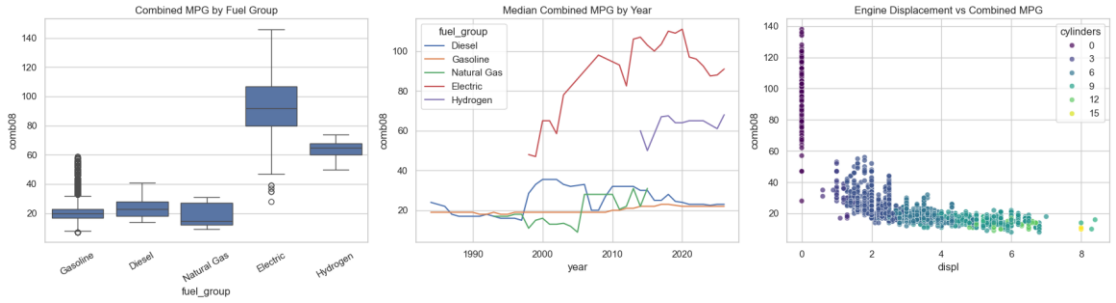

Figure 1. Exploratory notebook plots showing combined MPG by fuel group, median MPG trend by model year, and the inverse relationship between engine displacement and fuel economy.

For classification, the random forest clearly outperformed logistic regression with 96.75% accuracy and an F1 score of 0.941, versus 91.25% accuracy and an F1 score of 0.854 for logistic regression. The confusion matrix in Figure 2 shows strong performance on both classes, including 94.8% recall for the high-efficiency class and 97.5% recall for the lower-efficiency class. SelectKBest ranked engine displacement, cylinders, front-wheel drive, start-stop, electrification, vehicle age, and year among the most informative classification predictors, which is consistent with engineering intuition.

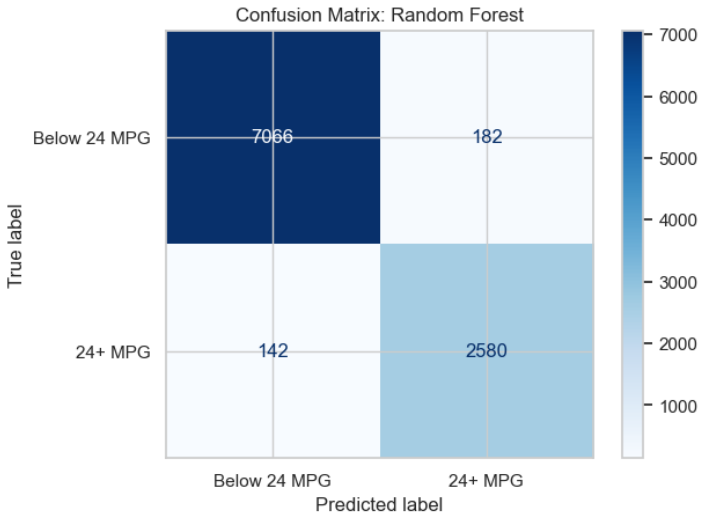

Figure 2. Random forest confusion matrix for classifying high-efficiency vehicles, where the target label is combined MPG greater than or equal to 24.

The most important regression issue was that a single pooled MPG model mixed conventional MPG with electric MPGe and treated missing electric-vehicle engine fields as ordinary missing data. That setup can create unrealistic predictions near zero or far outside the observed range. The corrected workflow cleansed the regression labels to realistic values (`comb08` 8 to 150 and `co2TailpipeGpm` 0 to 900), assigned sensible electric defaults for cylinders, displacement, and gear count, and trained separate regressors for electrified and non-electrified vehicles. The segment summary confirms why this helped: 46,392 non-electrified vehicles had a median combined MPG of 20, while 3,449 electrified vehicles had a median of 42 and extended up to 146.

This segmented strategy substantially improved regression performance. For combined MPG, the random forest achieved MAE 0.775 MPG, RMSE 1.607 MPG, and R^2 0.986, compared with MAE 1.784 MPG and R? 0.956 for linear regression. For tailpipe CO2, the random forest achieved MAE 14.70 g/mi and R^2 0.969. SelectKBest reinforced the physical interpretation of the model: the strongest MPG predictors included electric fuel group, cylinders, electrified status, gasoline fuel group, displacement, and EV-heavy makes such as Tesla and Lucid. In the Streamlit dashboard, the measured-versus-predicted MPG delta is now displayed for each selected vehicle, making it easier to detect whether one segment is still biased.

Figure 3 confirms that the final segmented MPG regressor tracks both electrified and non-electrified vehicles closely instead of collapsing one group toward the other. Together, the notebook and app form a practical workflow: computer vision verifies that the uploaded image contains a car, the regression models estimate efficiency and emissions, and the local RAG/Ollama assistant explains the selected vehicle using retrieved context from the dataset.

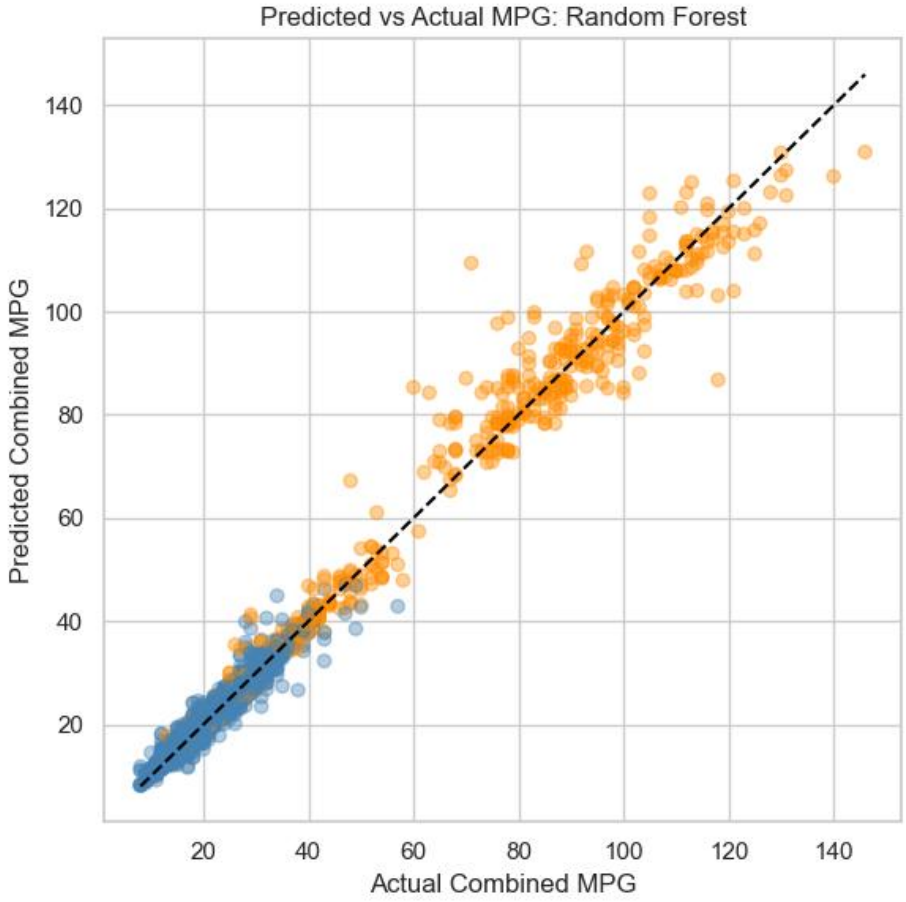

Figure 3. Predicted versus actual combined MPG for the segmented random forest regressor. Blue points are non-electrified vehicles and orange points are electrified vehicles.