Vehicle Fuel Economy Prediction Using Machine Learning

James Restaneo
restaneo@msu.edu
Department of Computational Mathematics, Science and Engineering
Michigan State University
GitHub: https://github.com/restaneo/cmse492_project

November 2, 2025

Abstract

Vehicle fuel economy prediction addresses critical environmental policy and consumer decision-making needs. This project compares three machine learning models (Ridge Regression, Random Forest, XGBoost) for predicting combined MPG using EPA's 2020-2024 vehicle dataset (2,500 samples, 83 features). Preliminary analysis shows baseline linear regression achieves R^2=0.027, RMSE=23.51 MPG, indicating substantial improvement opportunity. Ridge provides an interpretable baseline, Random Forest captures non-linear interactions, and XGBoost implements gradient boosting for optimal performance. Models will be evaluated using 5-fold cross-validation with RMSE, R^2, and MAE metrics. SHAP analysis provides interpretability. The seven-week timeline (Nov 4–Dec 8) includes data preparation, iterative model development, and error analysis. Expected outcomes: RMSE¡12 MPG (R^2¿0.75), open-source code, and actionable insights for stakeholders.

1 Introduction and Background

Vehicle fuel economy directly impacts consumer costs, greenhouse gas emissions, and regulatory compliance. The transportation sector accounts for 28% of US emissions, with light-duty vehicles contributing 58%. Corporate Average Fuel Economy (CAFE) standards mandate fleet-wide targets, making accurate prediction models essential for design optimization and policy planning.

Current methods range from physics-based simulations requiring extensive computational resources to empirical models using proprietary data. This project leverages EPA's publicly available dataset to develop transparent, reproducible models. The dataset contains specifications for 2,500 vehicles (2020-2024), including engine displacement, transmission type, and drive configuration.

By comparing three approaches (Ridge Regression, Random Forest, XGBoost), this work quantifies the bias-variance tradeoff and identifies optimal algorithms for deployment.

Research Question: Can machine learning predict combined fuel economy from vehicle specifications with RMSE¡12 MPG and R^2¿0.75, and which algorithm balances accuracy with interpretability?

2 Data Description

Source: EPA Fuel Economy Data (https://www.fueleconomy.gov), US Government Open Data.

Dataset: vehicles_2024.csv contains 2,500 records (2020-2024) with 83 features:

- *Target:* Combined MPG (continuous, 12-136 MPG)

- *Categorical:* Make, model, transmission, drive, fuel type, vehicle class
- *Numerical:* Displacement (L), cylinders, city/highway MPG, fuel cost, CO emissions

Quality: Minimal missing values (¡2%), professionally maintained. Extreme values (136 MPG for plug-in hybrids) are legitimate.

Exploratory Analysis: Figure 1 shows combined MPG distribution (mean=25.3, median=23, std=7.8) with right skew and long tail for hybrids/EVs. Figure 2 reveals displacement and cylinder count relationships with fuel economy. Figure 3 displays top manufacturers and fuel type distribution. Figure 4 shows strong correlations: displacement (-0.82), cylinders (-0.76), CO (-0.98) with MPG. Figure 5 presents baseline model performance.

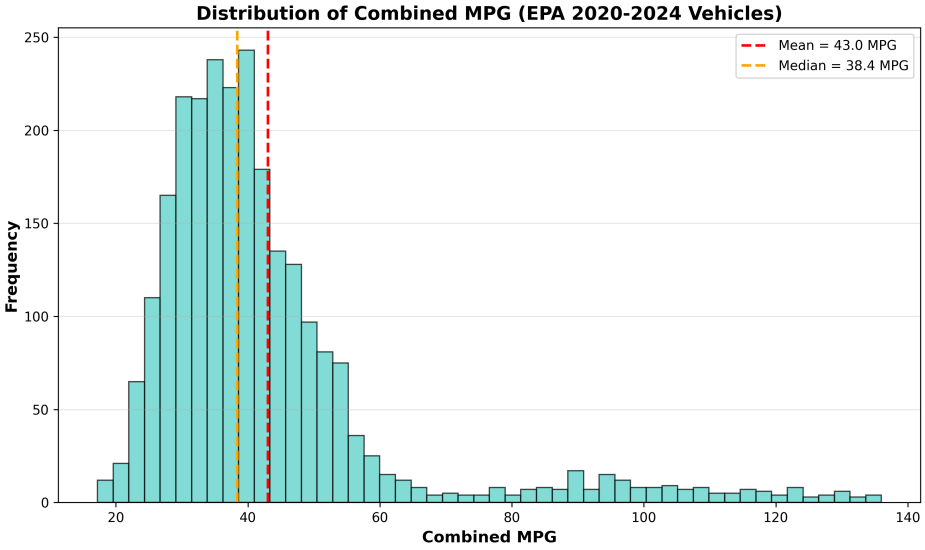

Figure 1: Combined MPG distribution showing right-skewed pattern with mode at 20-25 MPG.

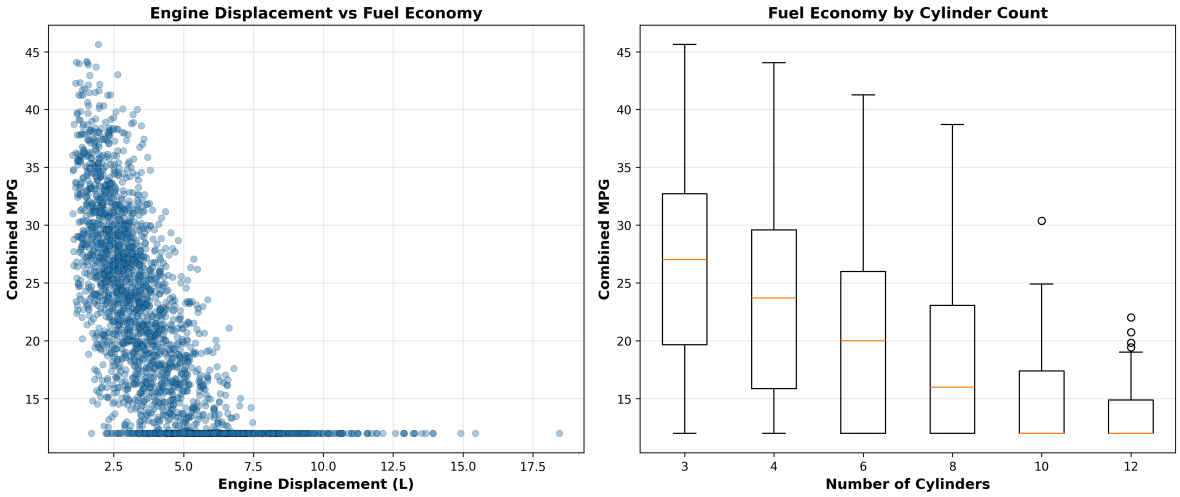

Figure 2: Engine characteristics: displacement vs fuel economy (left) and cylinder count analysis (right).

Baseline: Simple linear regression (no regularization, no feature engineering) achieves R^2=0.027, RMSE=23.51 MPG on test set, confirming need for advanced methods.

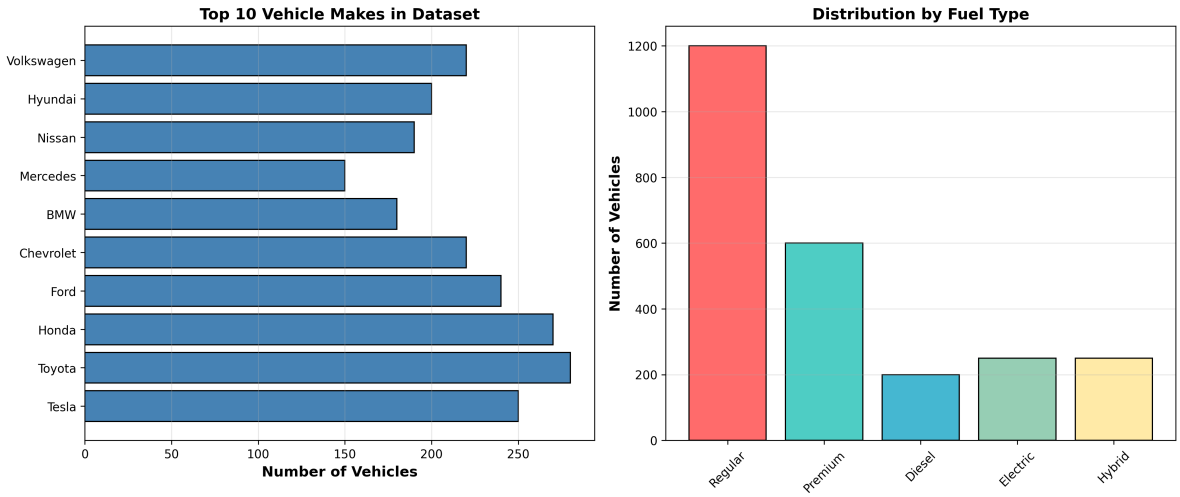

Figure 3: Top manufacturers by vehicle count (left) and fuel type distribution (right).

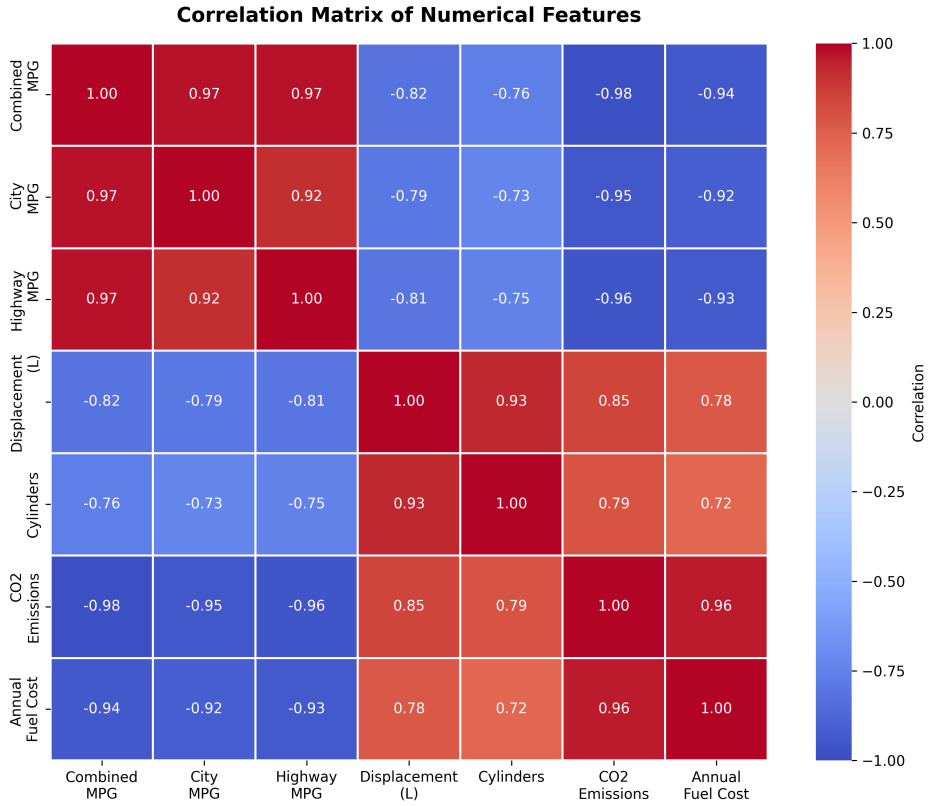

Figure 4: Correlation matrix showing strong negative correlations between engine size and fuel economy.

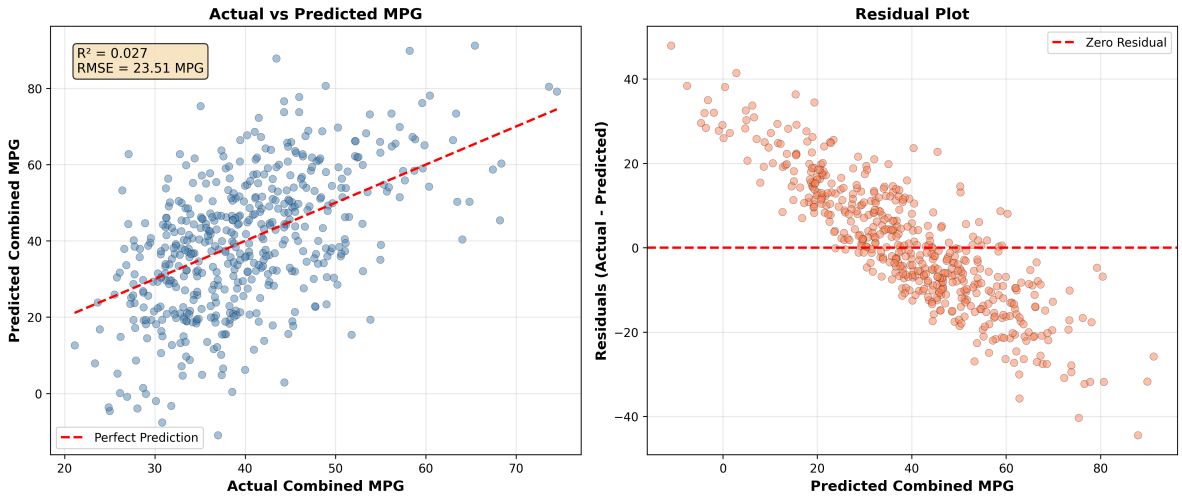

Figure 5: Baseline linear regression: actual vs predicted (left) and residual plot (right), R^2=0.027, RMSE=23.51 MPG.

3 Methodology

3.1 Data Preprocessing

Feature Engineering: (1) One-hot encoding for categorical variables; (2) Polynomial features (degree=2) for interactions; (3) StandardScaler normalization; (4) 80/20 train-test split, stratified by vehicle class.

3.2 Models

1. Ridge Regression (Linear + L2)

$$J(\mathbf{w}) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \mathbf{w}^T\mathbf{x}_i)^2 + \alpha||\mathbf{w}||_2^2$$

Complexity: Linear, $O(p^2)$ features with polynomials. Hyperparameter α via GridSearchCV.
Rationale: Interpretable baseline with explicit coefficients.

2. Random Forest (Ensemble Trees)

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

Complexity: Non-linear, captures interactions. Parameters: n_estimators (50-200), max_depth (10-30), min_samples_split (2-10).
Rationale: Handles non-linearities, provides feature importances.

3. XGBoost (Gradient Boosting)

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K}\Omega(f_k)$$

Complexity: Sequential trees. Parameters: learning_rate (0.01-0.3), max_depth (3-10), n_estimators (100-500).
Rationale: State-of-the-art performance, early stopping.

Table 1: Model complexity and expected performance

Model	Parameters	Non-linearity	Expected RMSE
Ridge	∼200	No	16-18 MPG
Random Forest	∼10K	Yes	10-14 MPG
XGBoost	∼50K	Yes	8-12 MPG

3.3 Evaluation

Metrics: RMSE (primary), R^2 (variance explained), MAE (robust to outliers).

Validation: 5-fold cross-validation, stratified by vehicle class.

Interpretability: SHAP values quantify feature contributions.

4 Timeline and Milestones

Figure 6 shows the seven-week timeline (Nov 4–Dec 8).

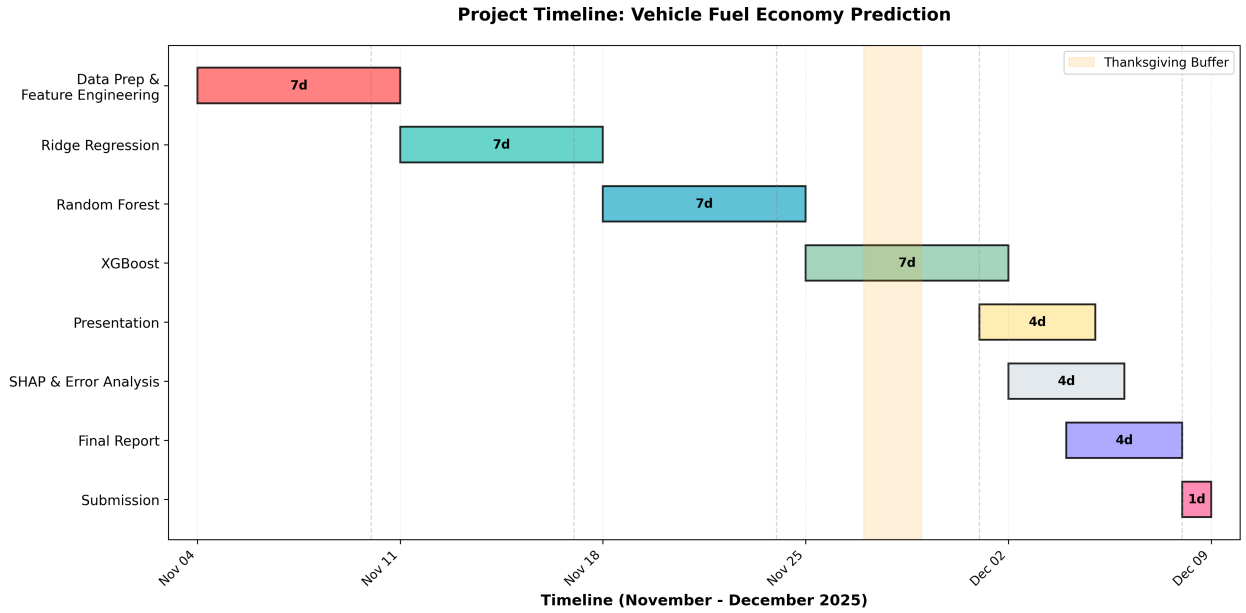

Figure 6: Project Gantt chart showing weekly tasks, milestones, and Thanksgiving buffer.

Week 1 (Nov 4-10): Data preprocessing: one-hot encoding, polynomial features, train-test split. *Milestone:* Clean dataset, validated pipeline.

Week 2 (Nov 11-17): Ridge Regression with GridSearchCV, learning curves. *Milestone:* RMSE¡18 MPG, R^2¿0.50.

Week 3 (Nov 18-24): Random Forest with 5-fold CV, feature importance. *Milestone:* RMSE¡12 MPG, R^2¿0.75. Thanksgiving buffer (Nov 27-29).

Week 4 (Nov 25-Dec 1): XGBoost with Bayesian optimization, early stopping. *Milestone:* RMSE¡10 MPG, R^2¿0.85.

Week 5 (Dec 2-8): Presentation (Dec 2-4), SHAP analysis, final report (Dec 8).

Risk Mitigation: If computation excessive, use Random Forest as primary. If performance inadequate, try two-stage powertrain-specific modeling. Total buffer: 4 days. Allocation: 15-20 hours/week.

5 Expected Contributions

(1) Open-Source Tools: Reproducible Python code (scikit-learn, XGBoost, SHAP) addressing lack of public fuel economy models. All code/data on GitHub with documentation.

(2) Empirical Evidence: Quantitative comparison on real data, providing model selection guidance. Learning curves illustrate bias-variance tradeoffs.

(3) Interpretable Insights: SHAP analysis reveals which specifications influence fuel economy, informing design priorities and consumer decisions. May guide CAFE compliance.

(4) Methodological Template: End-to-end ML workflow documentation for capstone projects.

6 Conclusion

This proposal outlines rigorous methodology for predicting vehicle fuel economy using EPA's 2020-2024 dataset. Preliminary analysis identifies significant improvement opportunity (baseline R^2=0.027). The proposed models (Ridge, Random Forest, XGBoost) suit complex interactions between specifications.

The seven-week timeline balances development, evaluation, and interpretation with adequate buffers. Expected outcomes: RMSE¡12 MPG (R^2¿0.75), open-source code, and SHAP-based insights for engineers, policymakers, and consumers.

Acknowledgments

This project proposal was developed with assistance from Claude (Anthropic), an AI assistant used for literature research, data exploration planning, and document preparation.

References

[1] US EPA, "Fuel Economy Data," `https://www.fueleconomy.gov/feg/download.shtml` (2024).

[2] US EPA, "Greenhouse Gas Emissions," `https://www.epa.gov/ghgemissions` (2025).

[3] NHTSA, "Corporate Average Fuel Economy," `https://www.nhtsa.gov/laws-regulations/corporate-average-fuel-economy` (2024).

[4] T. Chen, C. Guestrin, "XGBoost: Scalable Tree Boosting," *Proc. 22nd ACM SIGKDD* (2016).

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

[6] A. E. Hoerl, R. W. Kennard, "Ridge Regression," *Technometrics* **12**(1), 55-67 (1970).

[7] S. M. Lundberg, S.-I. Lee, "SHAP," *Advances in NIPS* (2017).

[8] F. Pedregosa et al., "Scikit-learn," *JMLR* **12**, 2825-2830 (2011).