Modeling Fuel Efficiency: A Comparative Analysis of Linear and Ensemble Regression Techniques

Douglas Fulcher

Bellevue University

Applied Data Science DSC680 T301 - Fall 2025

Instructor: Amirfarrokh Iranitalab

Abstract

This study examines the impact of vehicle characteristics on fuel efficiency (miles per gallon, MPG) and compares the predictive performance of linear and ensemble regression techniques. Using the Auto MPG dataset from the UCI Machine Learning Repository, eight vehicle attributes—such as weight, horsepower, and engine displacement—were analyzed after preprocessing that included mean imputation and one-hot encoding. Four regression models were developed: Linear Regression, Decision Tree, Random Forest, and Gradient Boosting Regressor. A key differentiator of this research is the creation of a simplified, dictionary-based Python framework that automates the execution of multiple regression models, supports hyperparameter tuning, and logs all model configurations for transparent, reproducible benchmarking. Models were trained using an 80/20 split and evaluated using five-fold cross-validation, with R^2, RMSE, and MAE metrics. The Random Forest achieved the highest test accuracy ($R^2 = 0.915$, RMSE = 2.14), offering the best balance between bias and variance. SHAP analysis highlighted weight, horsepower, and engine displacement as the most influential predictors of MPG. The findings confirm that ensemble regression methods—particularly Random Forest—offer superior predictive power for fuel efficiency modeling, while the flexible code framework streamlines experimentation across modeling approaches.

Modeling Fuel Efficiency: A Comparative Analysis of Linear and Ensemble Regression Techniques

Business Problem Statement

The global automotive industry is facing pressure from rising fuel costs, stringent environmental standards, and sustainability goals. This study examines which vehicle characteristics most affect fuel efficiency (miles per gallon, MPG). It compares how different regression modeling approaches—from traditional linear models to modern ensemble methods—can more effectively predict fuel efficiency. Key questions include:

- Which type of regression model predicts vehicle MPG most accurately?
- How well do these models perform on new, unseen data (and how do they avoid overfitting)?
- Which vehicle traits—like weight or horsepower—have the most substantial impact on fuel efficiency?
- How does model explainability change as we move from simple linear models to powerful ensemble methods?
- What are the trade-offs between accuracy, interpretability, and computational effort?

Background and History of GLMs in Auto Insurance Pricing

Traditional linear regression has long been used to model relationships between physical variables, particularly in studies of fuel economy. However, these models assume linear relationships between predictors and outcomes, which often oversimplify real-world interactions between engine size, power, weight, and efficiency.

Modern ensemble methods, such as Random Forests and Gradient Boosting, have emerged as leading techniques to overcome these limitations by capturing nonlinear relationships and complex variable interactions (Friedman, 2001; Breiman, 2001). These methods aggregate many “weak learners” into stronger predictors, producing more robust and generalizable results without requiring strict parametric assumptions.

Data Explanation

The dataset used, Auto MPG (UCI Machine Learning Repository), includes 398 vehicles manufactured between 1970 and 1982. It records nine attributes associated with automotive design and performance. After preprocessing, eight predictive features were selected for retention. The complete list of features and their definitions is provided in Table 1 (see Appendix A)

Missing values in horsepower were imputed with the mean, and categorical variables were converted using one-hot encoding. The data revealed a strong negative correlation (r = -0.83) between vehicle weight and fuel efficiency, suggesting weight reduction as a major driver of higher MPG. Visualizations from Exploratory Data Analysis can be found in Appendix B.

Methods

This section provides a concise overview of the modeling approach, tools, and evaluation metrics used in this effort.

Modeling Approach

For this project, we utilized scikit-learn's regression modeling toolkit to develop multiple predictive models for vehicle fuel efficiency (MPG). In this case, we developed Linear Regression, Decision Tree Regressor, Random Forest Regressor, and Gradient Boosting Regressor Models.

The suite of models allows direct performance comparison between traditional and ensemble-based methods. Scikit-learn was selected as the core library due to its robust API support, interpretability tools, and flexibility for integrating both classical statistical and machine learning techniques within a unified framework.

Each model was trained using an 80/20 train-test split and evaluated with five-fold cross-validation to ensure stability and reproducibility. For the ensemble models, we conducted hyperparameter tuning via grid search, optimizing parameters such as the number of estimators, maximum depth, learning rate, and sample size. These adjustments were designed to minimize overfitting, reduce model variance, and ensure fair cross-model benchmarking. Details on the model's purpose and hyperparameters are provided in Table 2 (Appendix A).

Model Evaluation Metrics

For this effort, our code was implemented to automate iterative experiments across multiple model types and configurations. The current version of the code accepts an input dictionary that controls key parameters—such as which modeling algorithms to run, which feature engineering variants to evaluate, and which cross-validation metrics to store—thereby allowing batch execution of comparable model scenarios. All results were programmatically recorded, enabling subsequent visual and quantitative comparisons of R^2, RMSE, and MAE outcomes.

Table 3: Evaluation Metrics

Metric	Description	Goal
R^2	Percentage of variance in MPG explained	Higher = better
RMSE	Root mean squared error	Lower = better
MAE	Mean absolute error	Lower = better

Metric	Description	Goal
Fit Time	Total time to train the model	Lower = faster
Predict Time	Time to produce predictions	Lower = faster

Analysis

Baseline Results and Comparative Performance

Results indicate a transparent performance gradient from simpler models to complex ensembles. Random Forest achieved the best test R^2 (0.914), lowest RMSE (2.16), and effective generalization with minimal overfitting.

Table 4: Modeling Results Comparison

Model	CV R^2 Mean	Test R^2	Test RMSE	Test MAE	Fit Time (s)
Linear Regression	0.804	0.845	2.89	2.29	**0.012**
Decision Tree	0.798	0.876	2.58	1.88	0.310
Random Forest	0.873	**0.915**	**2.14**	**1.60**	9.121
Gradient Boosting	**0.860**	0.895	2.37	1.76	4.426

Model Comparison and Evaluation

Linear Regression provided a quick, interpretable benchmark, but did not capture the effects of nonlinear relationships. Decision Trees improved local precision but introduced instability across splits. Ensemble methods (Random Forest, Gradient Boosting) significantly enhanced predictive accuracy, with Random Forest delivering the best balance between bias and variance.

SHAP analysis confirmed the key feature drivers of fuel efficiency (MPG) were weight, horsepower, engine displacement, and model year, reinforcing engineering expectations that lighter, newer, and lower-displacement vehicles yield higher efficiency. Residual Plots, Feature Importance, and SHAP visualizations are presented in Appendix B.

Conclusion

Ensemble regression models—particularly Random Forest—demonstrate the strongest accuracy, robustness, and practical value for predicting automobile fuel efficiency. Simpler methods still offer transparency and quick interpretability during exploratory stages, but ensemble models significantly outperform them when relationships are nonlinear.

Research Questions

1. Random Forest yielded the highest predictive accuracy and best generalization.
2. Model complexity improved predictive performance but reduced interpretability.
3. Weight, horsepower, and engine displacement were the dominant factors influencing MPG.
4. Gradient Boosting achieved stable fold-wise performance with less overfitting.
5. Linear Regression remained computationally efficient for diagnostic use.

Limitations

The dataset used in this analysis was relatively small, containing 398 vehicles manufactured between 1970 and 1982. This limited and dated sample restricts the generalizability of the results to modern vehicles. Additionally, key variables such as aerodynamics, hybrid systems, and drivetrain type were missing, limiting the model's coverage of newer automotive technologies. While ensemble models improved prediction accuracy, their complexity made interpretation more difficult for nontechnical audiences. Future studies should utilize more recent datasets and consider approaches such as regularized polynomial regression or SHAP-based explanations to enhance transparency and interpretability.

Challenges

Balancing interpretability and accuracy presented the primary challenge. Ensemble approaches delivered strong generalization but required parameter tuning and technical skill to avoid overfitting. Interpreting variable importance across tree-based models demanded advanced visualization via SHAP analysis.

Future Uses/Applications

This project introduced a flexible framework using Python dictionaries to automate model and hyperparameter configuration. This design enabled quick, side-by-side comparisons between simpler and ensemble regression methods, helping determine when added model complexity yields meaningful performance gains. The approach allows future teams to easily replicate or extend experiments for other predictive modeling tasks, streamlining evaluation while maintaining transparency and reproducibility.

Recommendations

Random Forest should be the primary model for structured, low-dimensional data due to its strong accuracy and reliability. Gradient Boosting is a solid choice for faster, scalable deployment when runtime is a critical factor. Linear Regression remains beneficial for transparency and stakeholder trust. All models should be retrained regularly with newer vehicle data to maintain relevance and accuracy.

This effort could also benefit from additional feature engineering to improve the results of linear regression sufficiently, thereby reducing the necessity of more complex model methodologies on this dataset. For example, leveraging Ridge or Lasso Regression or engineering polynomial features may provide better results from models. However, this was not pursued for this effort.

Implementation Plan

The deployment of this model could involve automotive analytics dashboards capable of evaluating design attributes in the early stages of the engineering process. Integrating ensemble models into simulation pipelines would enhance iterative vehicle testing and optimize emissions compliance.

Ethical Assessment

The dataset used contained no personally identifiable data, posing minimal ethical risk. However, future applications integrating telematics or vehicle telemetry must ensure strict data governance, anonymization, and transparency regarding modeling assumptions and outcomes.

Another ethical consideration here is the use of Generative AI in peer reviewing and editing this paper. To fit the content within the desired length, we utilized Generative AI to reduce verbosity while retaining the critical observations of the research. All Generative AI content is derived from guidance and constraints provided by the researcher(s).

Appendix A

Tables

Table 1: Auto MPG Data Dictionary

Feature	Description	Data Type	Role
mpg	Miles per gallon (target variable)	Float	Target
cylinders	Number of cylinders	Integer	Predictor
displacement	Engine displacement (cubic inches)	Float	Predictor
horsepower	Engine power	Float	Predictor
weight	Vehicle weight	Integer	Predictor
acceleration	0–60 mph acceleration time	Float	Predictor
model year	Year of manufacture	Integer	Predictor
origin	Region of manufacture (USA, Europe, Japan)	Categorical	Predictor
car name	Vehicle Make/Model descriptions (unique for each record)	Text	Dropped

Table 2: Summary of Regression Models Used

Model	Type	Key Parameters	Purpose
Linear Regression	Statistical	Standard least squares	Establishes baseline interpretability
Decision Tree Regressor	Nonlinear, interpretable ML	Tuned `max_depth`, `min_samples_leaf`	Captures simple nonlinearities with explicit thresholds.
Random Forest Regressor	Ensemble (bagging)	Tuned `max_depth`, `max_features`, `min_samples_leaf`	An ensemble of trees that reduces variance and overfitting.

Model	Type	Key Parameters	Purpose
Gradient Boosting Regressor	Ensemble (boosting)	Tuned `learning_rate`, `n_estimators`, `max_depth`, `subsample`	sequentially corrects residual errors for higher precision.

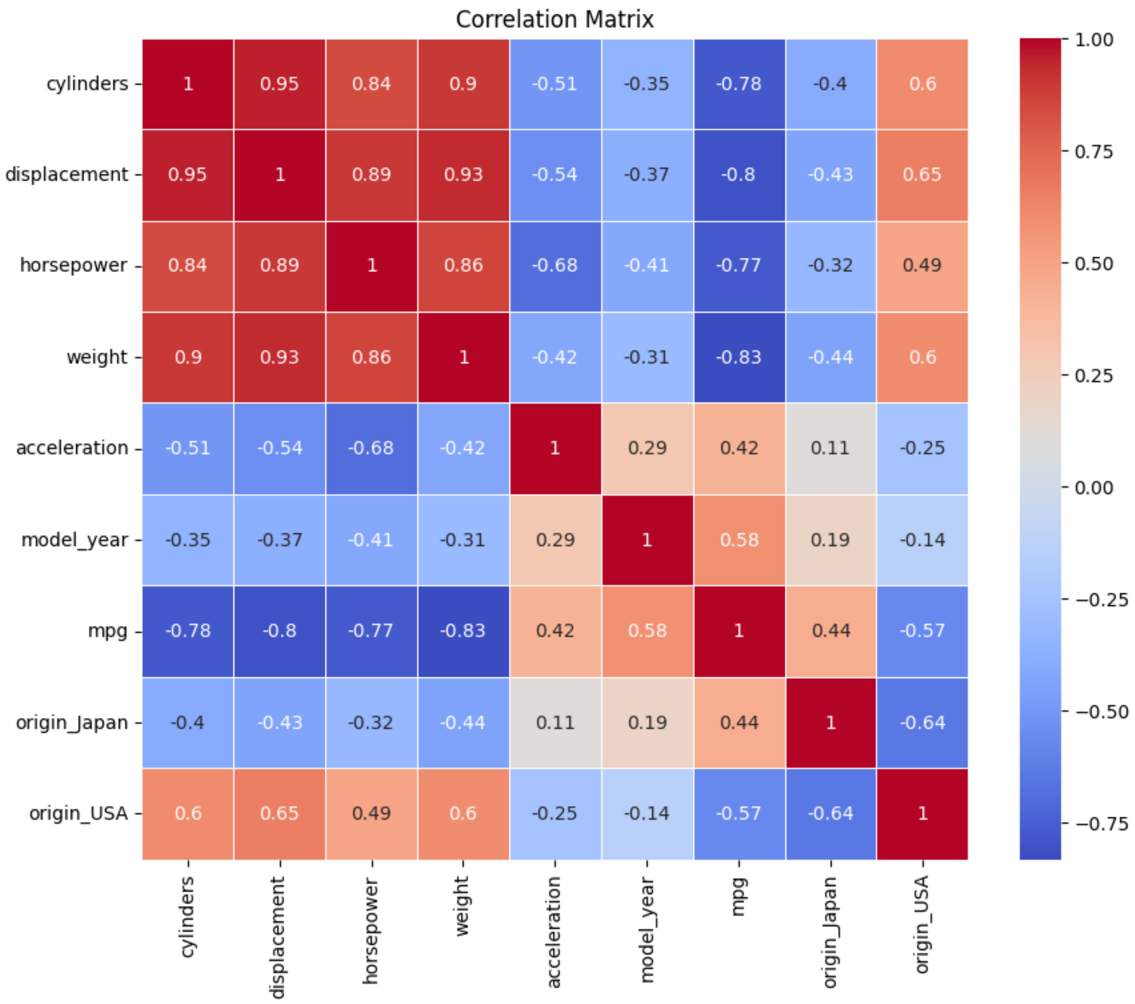
Table 5: Summary of Key Model Insights

Dimension	Best Performer	Insight
Accuracy (R^2, RMSE, MAE)	Random Forest	Most accurate and generalizable model
Consistency (Cross-Validation)	Gradient Boosting	Most stable performance across folds
Interpretability	Linear Regression / Decision Tree	Simplest to explain and visualize
Computation Speed	Linear Regression	Fastest by orders of magnitude
Bias-Variance Trade-off	Random Forest	Ideal balance of complexity and robustness

Appendix B

Figures

Figure 1: Correlation Matrix Heatmap



	cylinders	displacement	horsepower	weight	acceleration	model_year	mpg	origin_Japan	origin_USA
cylinders	1	0.95	0.84	0.9	-0.51	-0.35	-0.78	-0.4	0.6
displacement	0.95	1	0.89	0.93	-0.54	-0.37	-0.8	-0.43	0.65
horsepower	0.84	0.89	1	0.86	-0.68	-0.41	-0.77	-0.32	0.49
weight	0.9	0.93	0.86	1	-0.42	-0.31	-0.83	-0.44	0.6
acceleration	-0.51	-0.54	-0.68	-0.42	1	0.29	0.42	0.11	-0.25
model_year	-0.35	-0.37	-0.41	-0.31	0.29	1	0.58	0.19	-0.14
mpg	-0.78	-0.8	-0.77	-0.83	0.42	0.58	1	0.44	-0.57
origin_Japan	-0.4	-0.43	-0.32	-0.44	0.11	0.19	0.44	1	-0.64
origin_USA	0.6	0.65	0.49	0.6	-0.25	-0.14	-0.57	-0.64	1

Figure 2: Pairplot of Selected Features

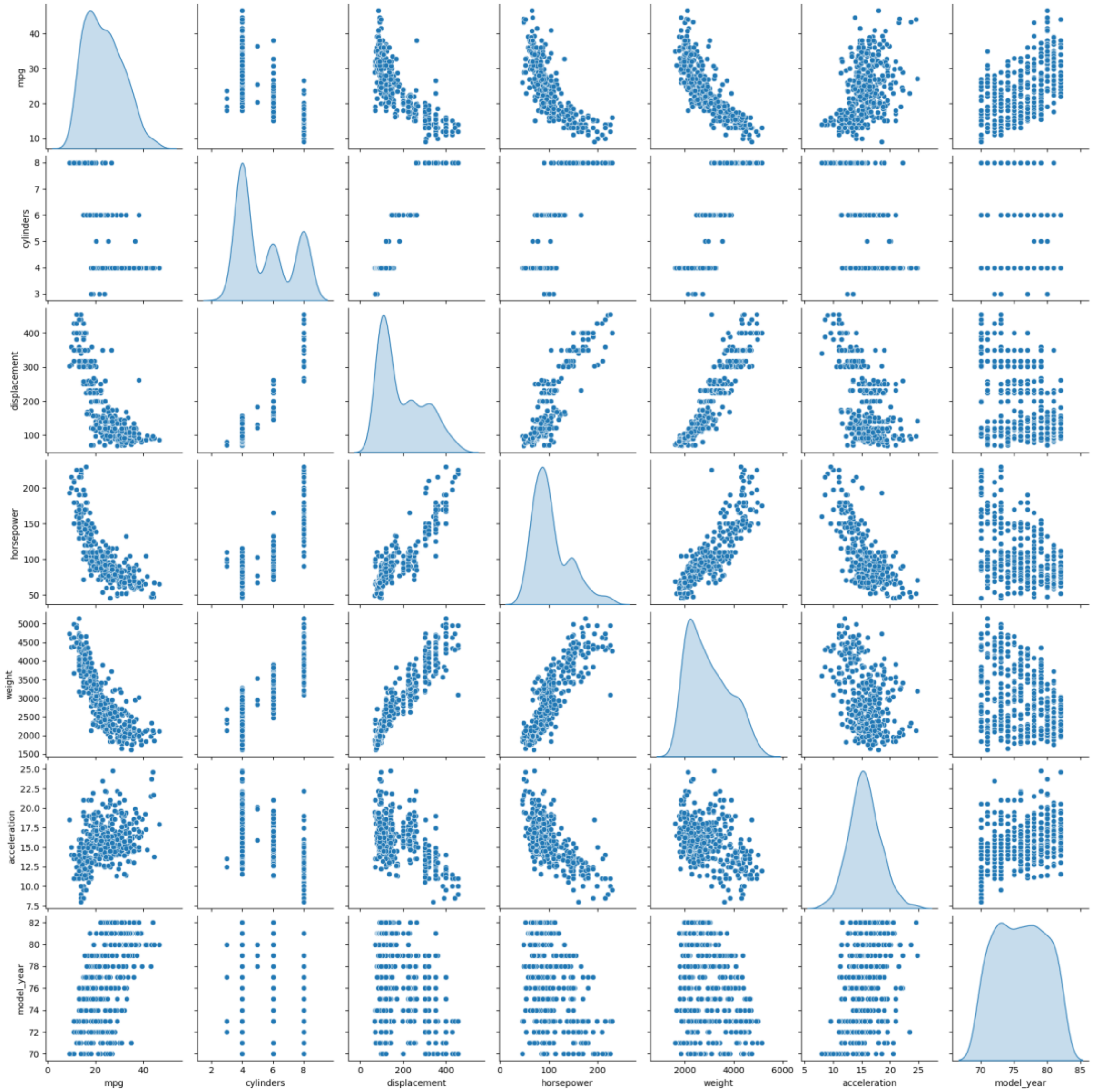


Figure 3: Residual Plots for Linear Regression – Train

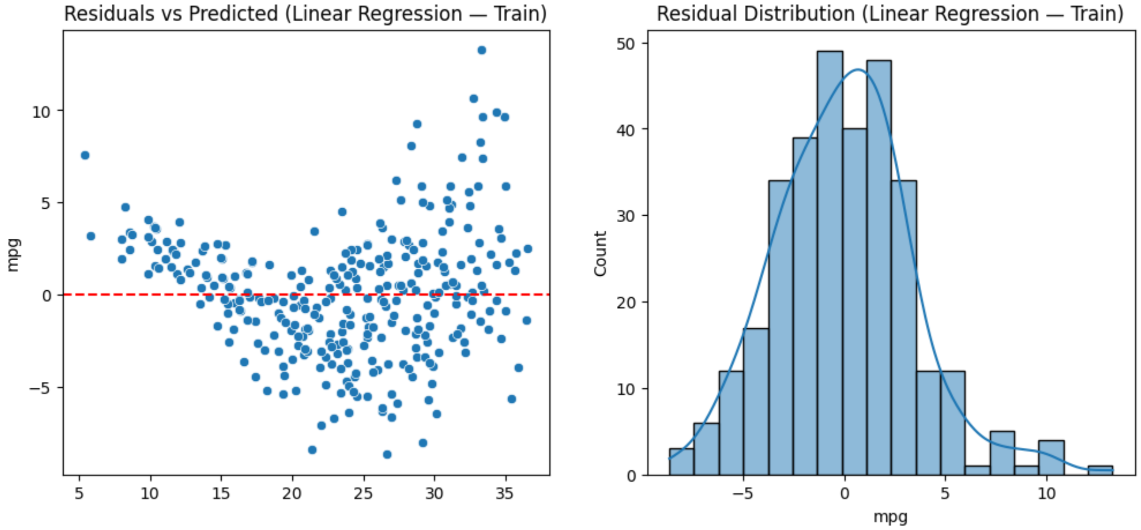


Figure 4: Residual Plots for Linear Regression – Test

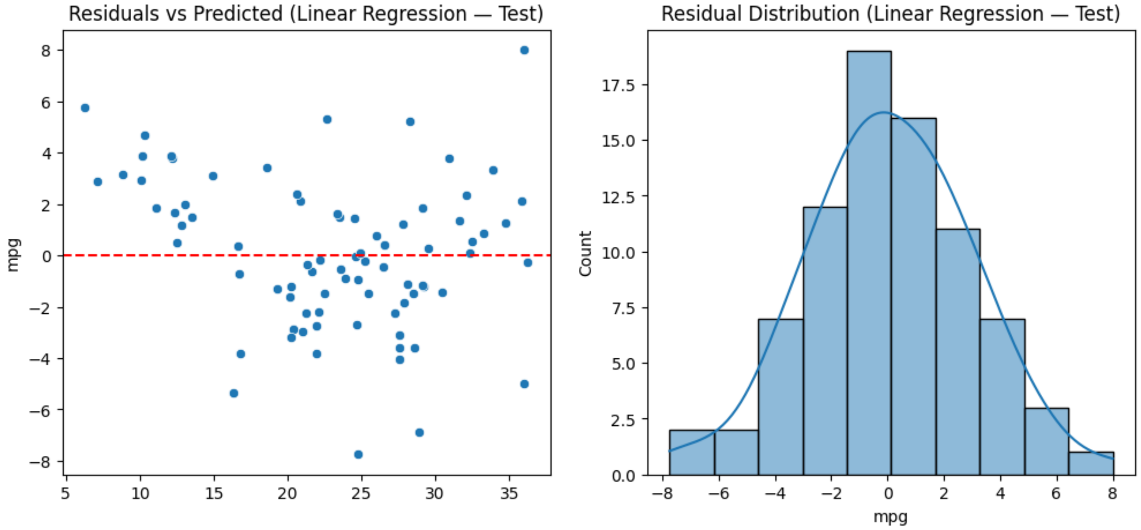
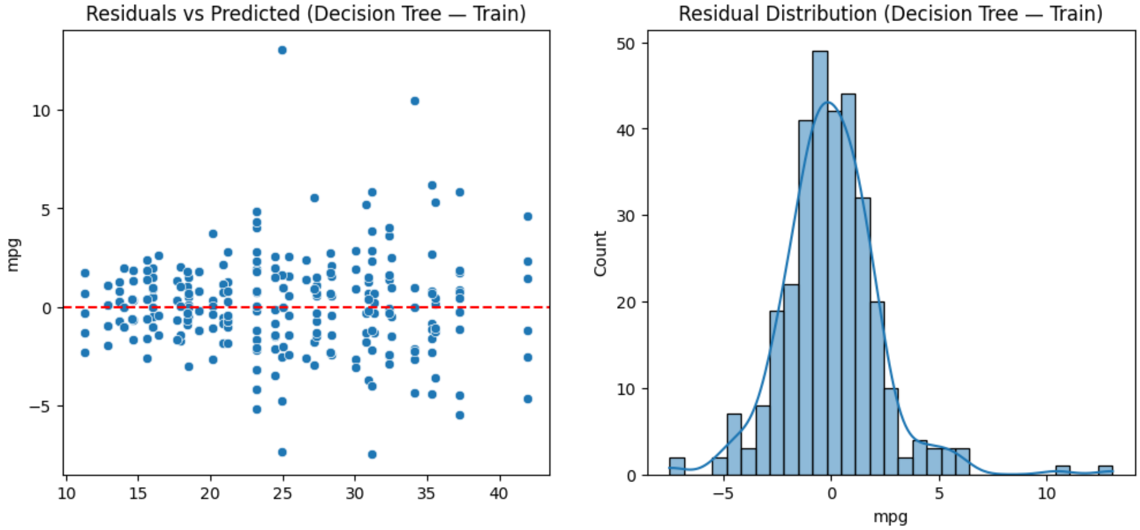


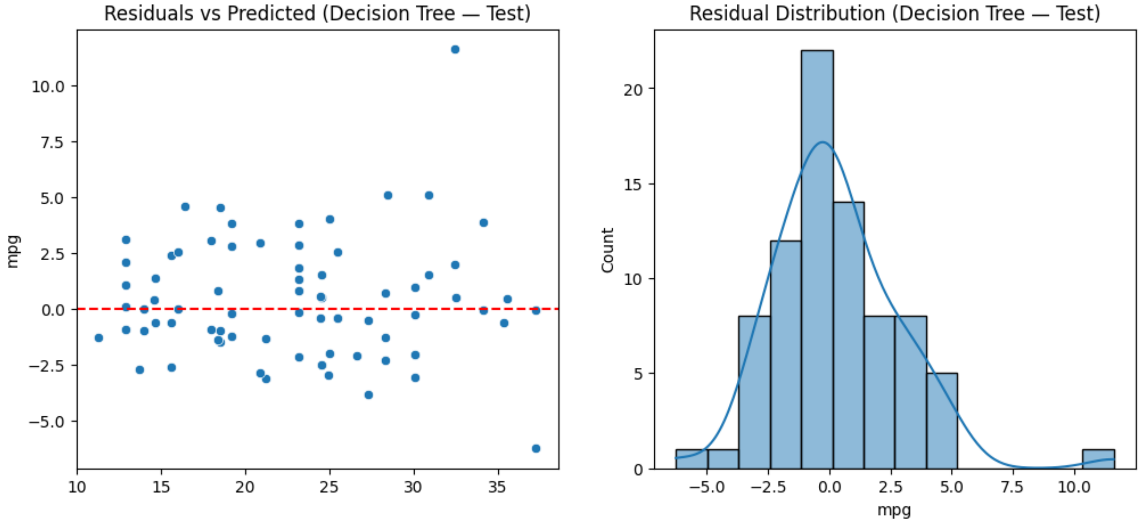
Figure 5: Residual Plots for Decision Tree - Train



Residuals vs Predicted (Decision Tree — Train)

Residual Distribution (Decision Tree — Train)

Figure 6: Residual Plots for Decision Tree - Test



Residuals vs Predicted (Decision Tree — Test)

Residual Distribution (Decision Tree — Test)

Figure 7: Residual Plots for Random Forest - Train

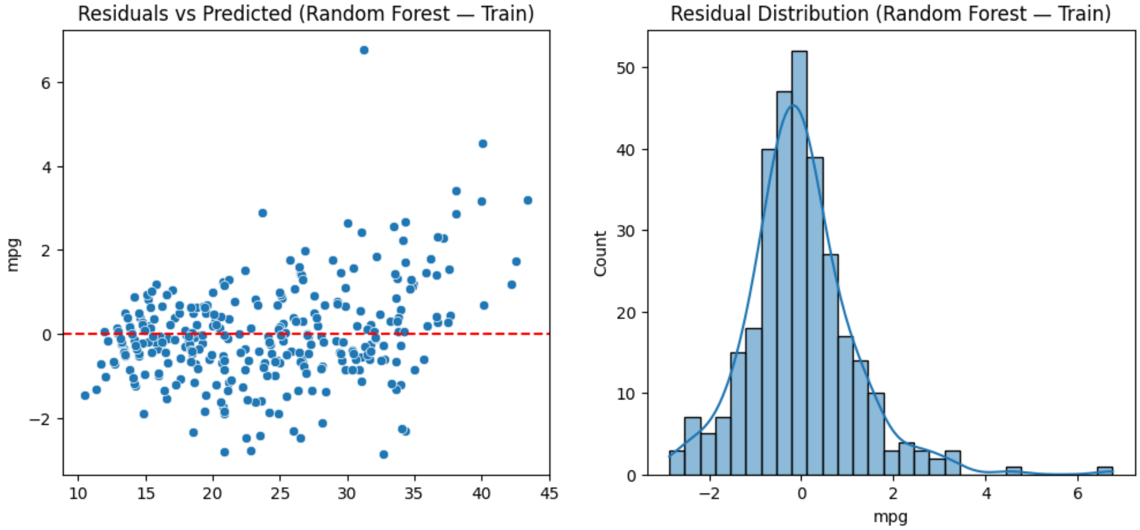


Figure 8: Residual Plots for Random Forest - Test

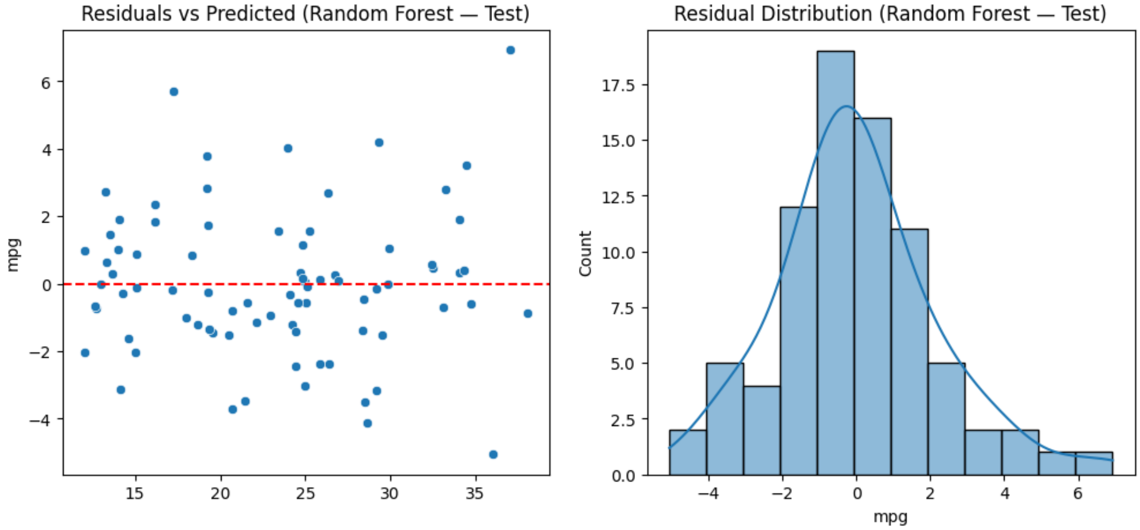


Figure 9: Residual Plots for Gradient Boosting - Train

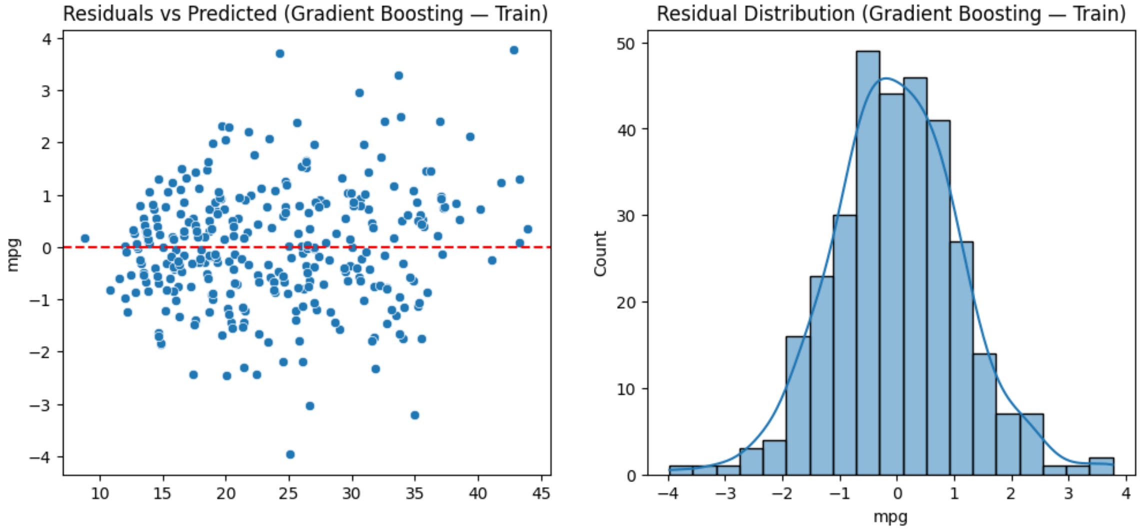


Figure 10: Residual Plots for Gradient Boosting - Test

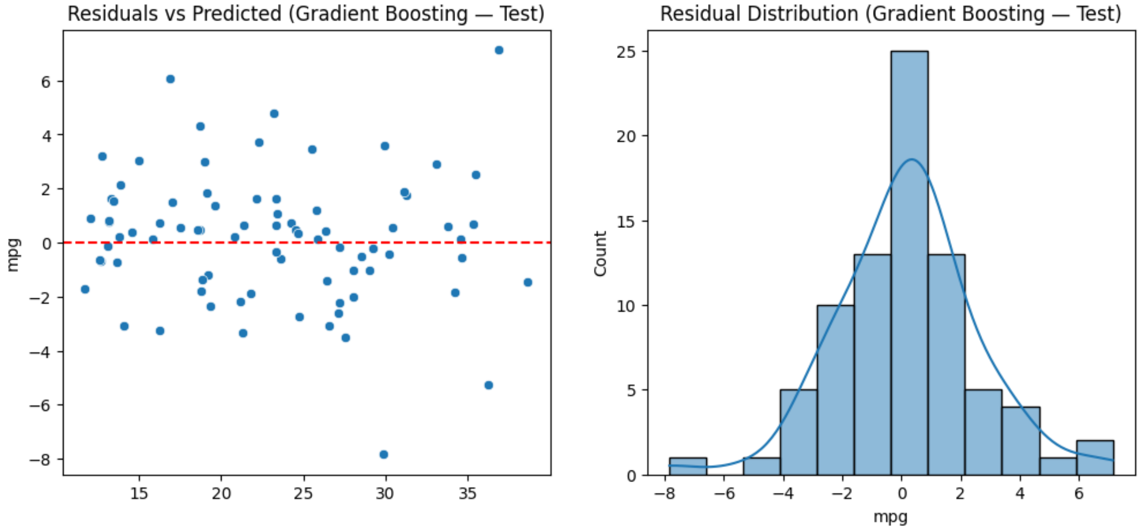


Figure 11: Top Feature Importances for Decision Tree

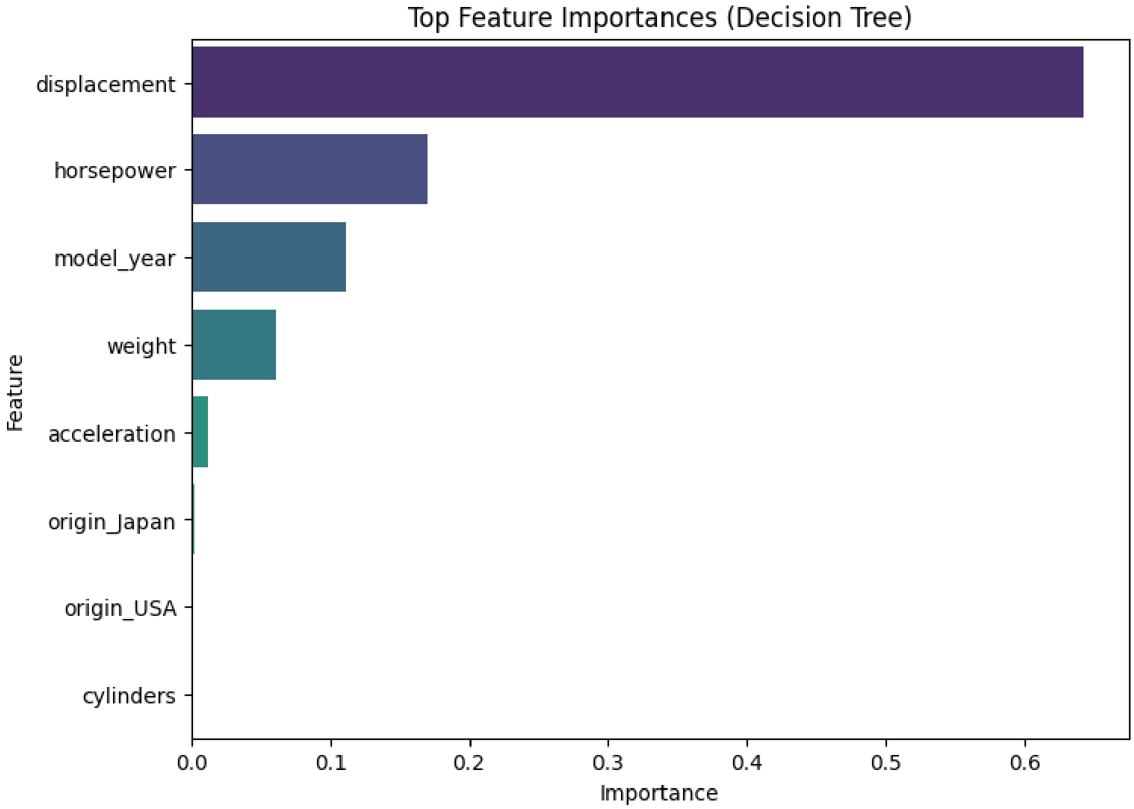


Figure 12: Top Feature Importances for Random Forest

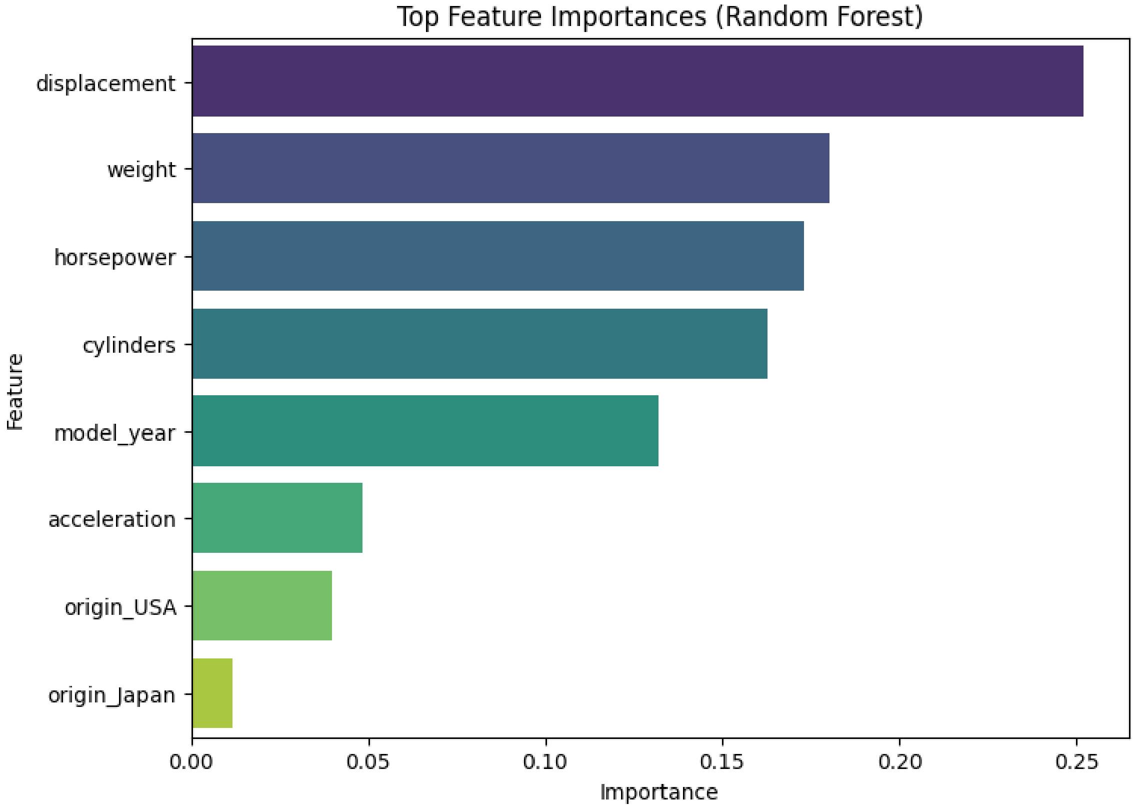


Figure 13: Top Feature Importances for Gradient Boosting

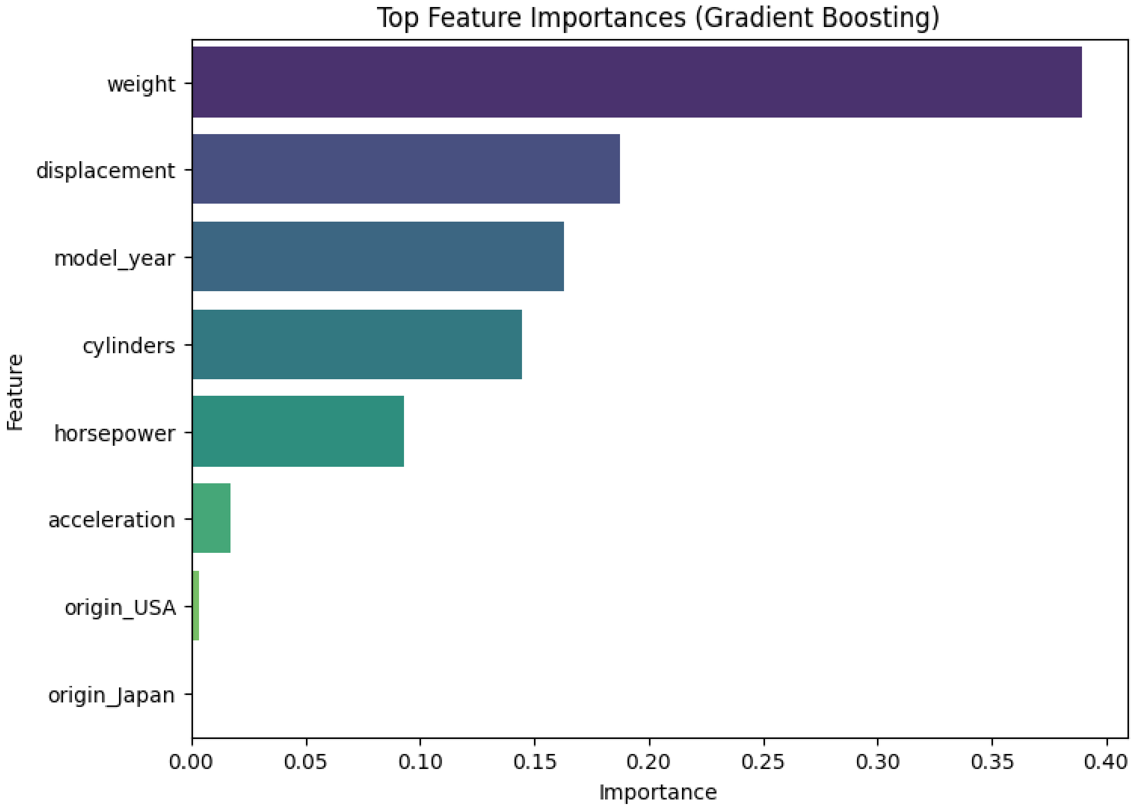


Figure 14: Linear Regression – Global Feature Importance

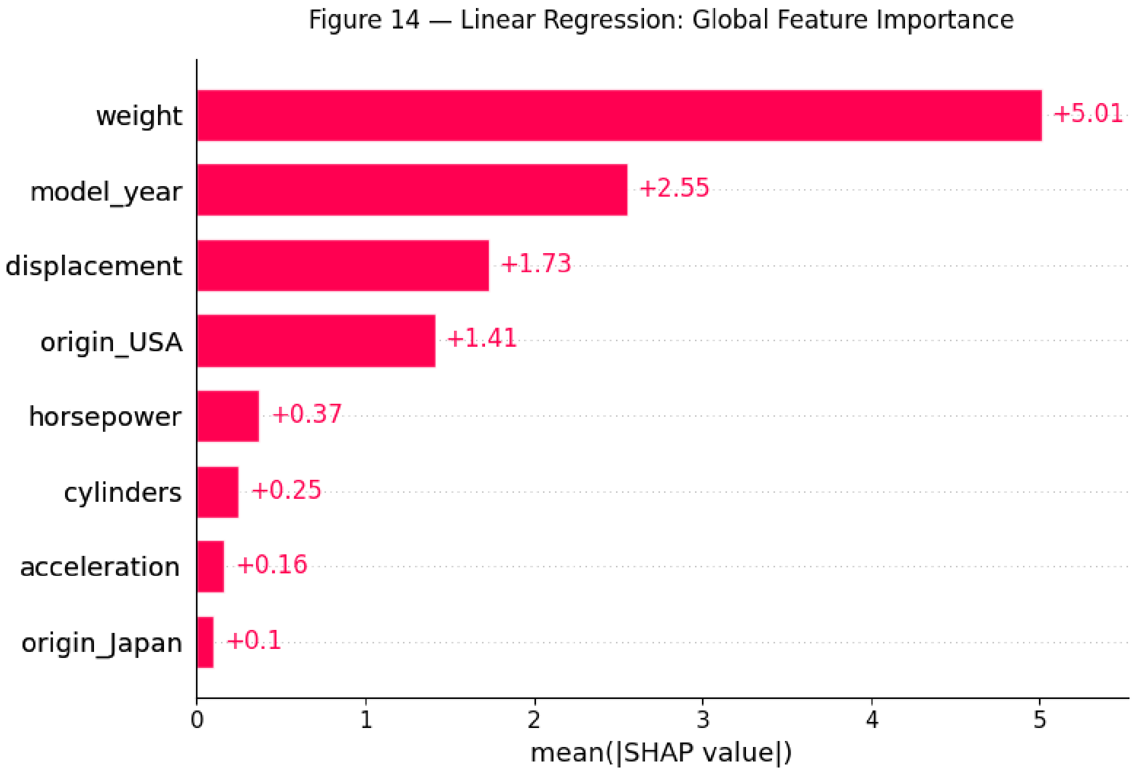


Figure 15 — Linear Regression: Feature Effects and Direction

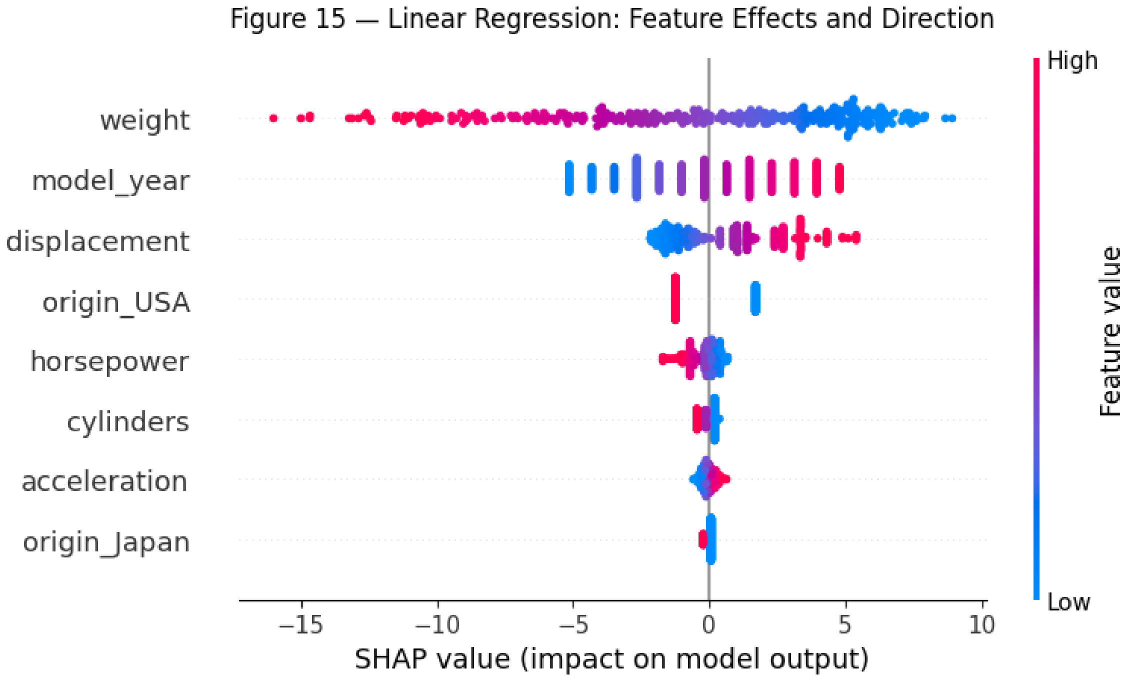


Figure 16 — Linear Regression: Dependence Plot for Top Feature

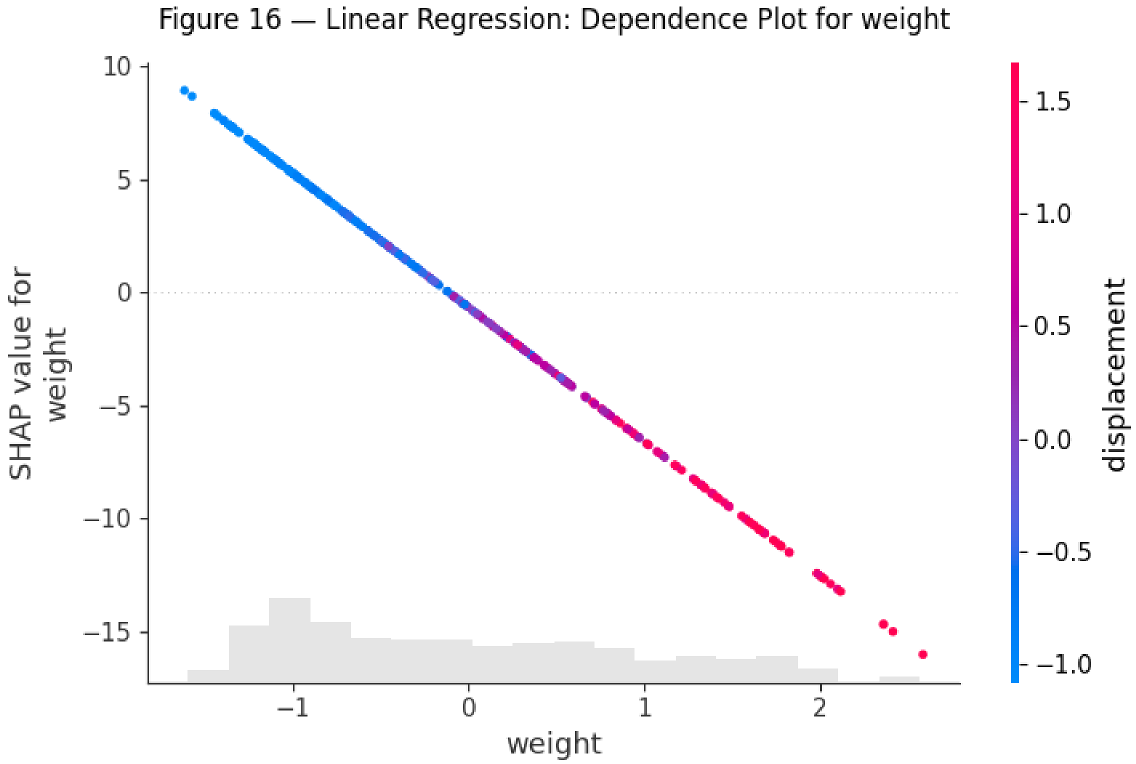


Figure 17 — Decision Tree: Global Feature Importance

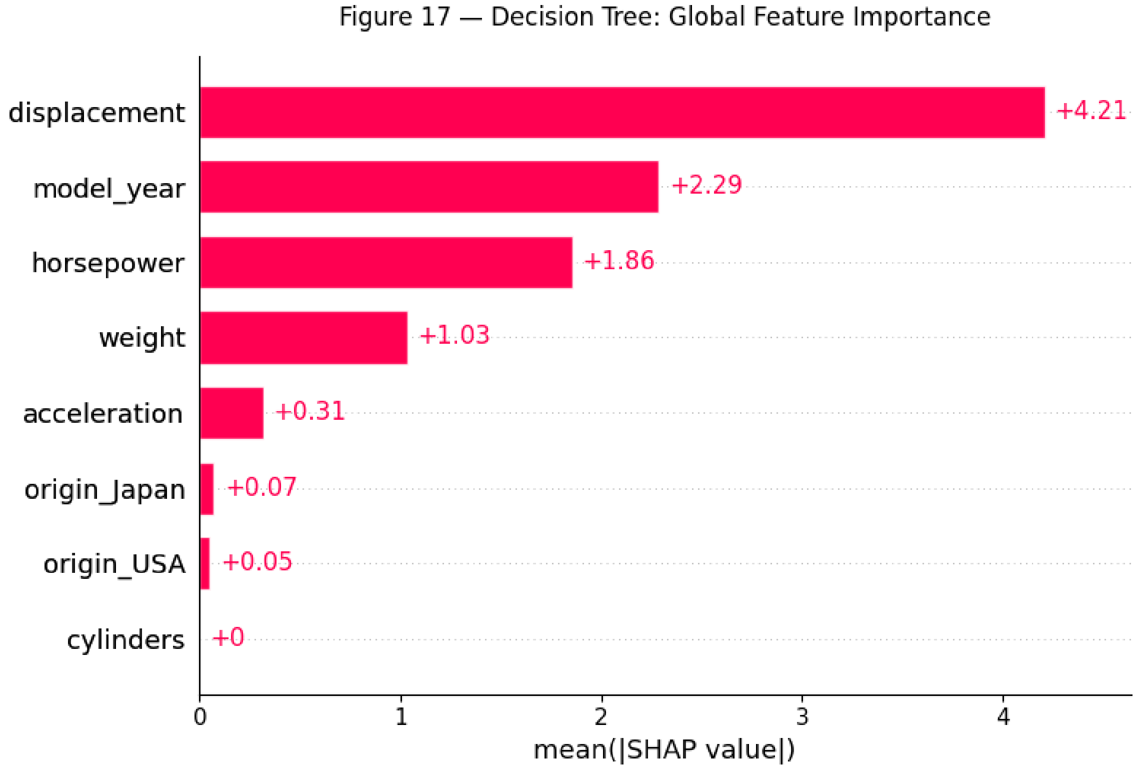


Figure 18 — Decision Tree: Feature Effects and Direction

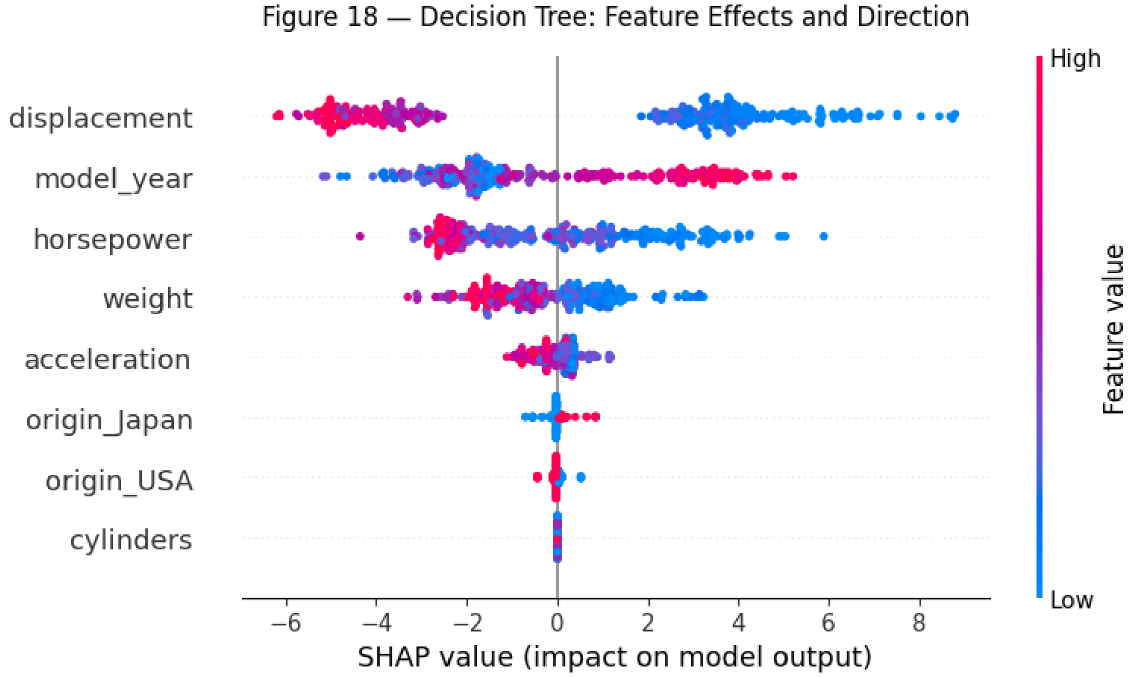


Figure 19 — Decision Tree: Dependence Plot for Top Feature

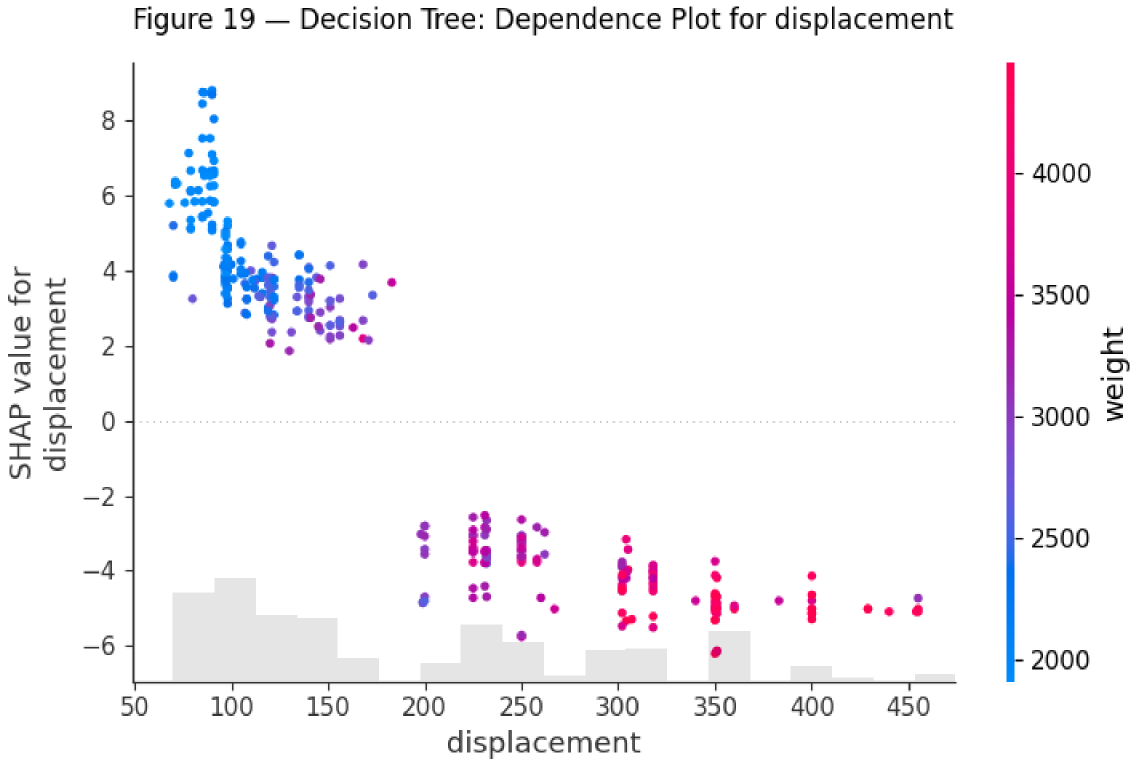


Figure 20 — Random Forest: Global Feature Importance

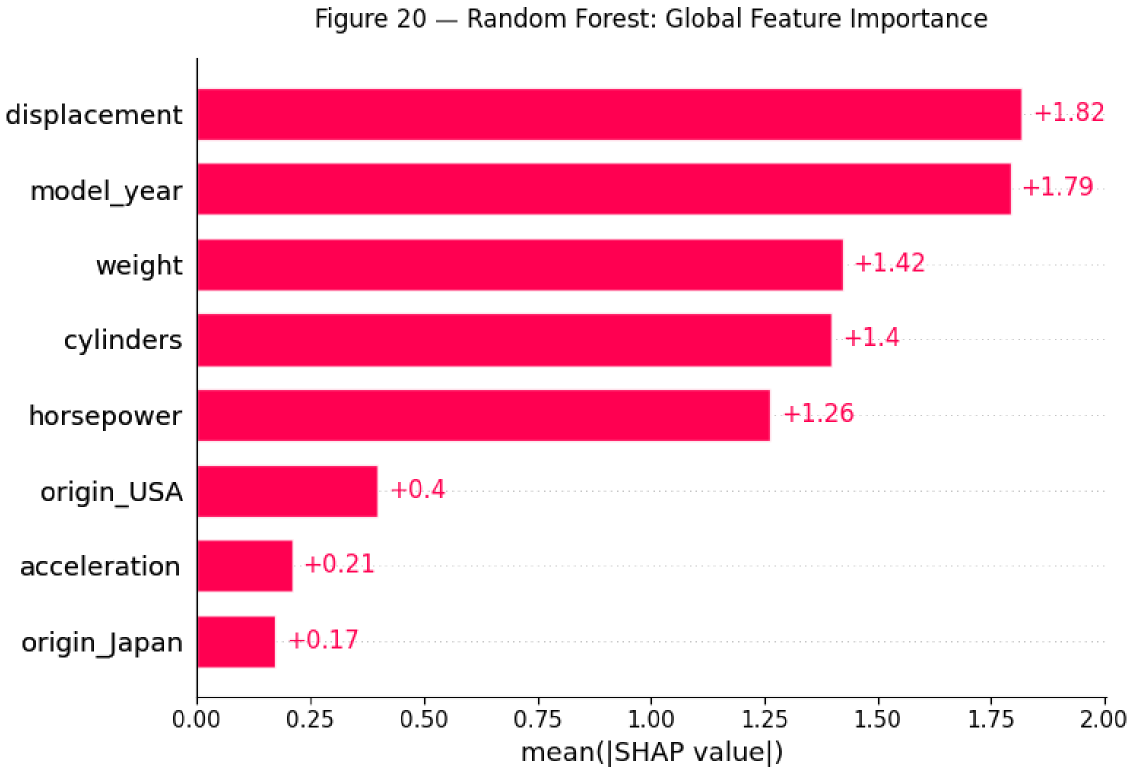


Figure 21 — Random Forest: Feature Effects and Direction

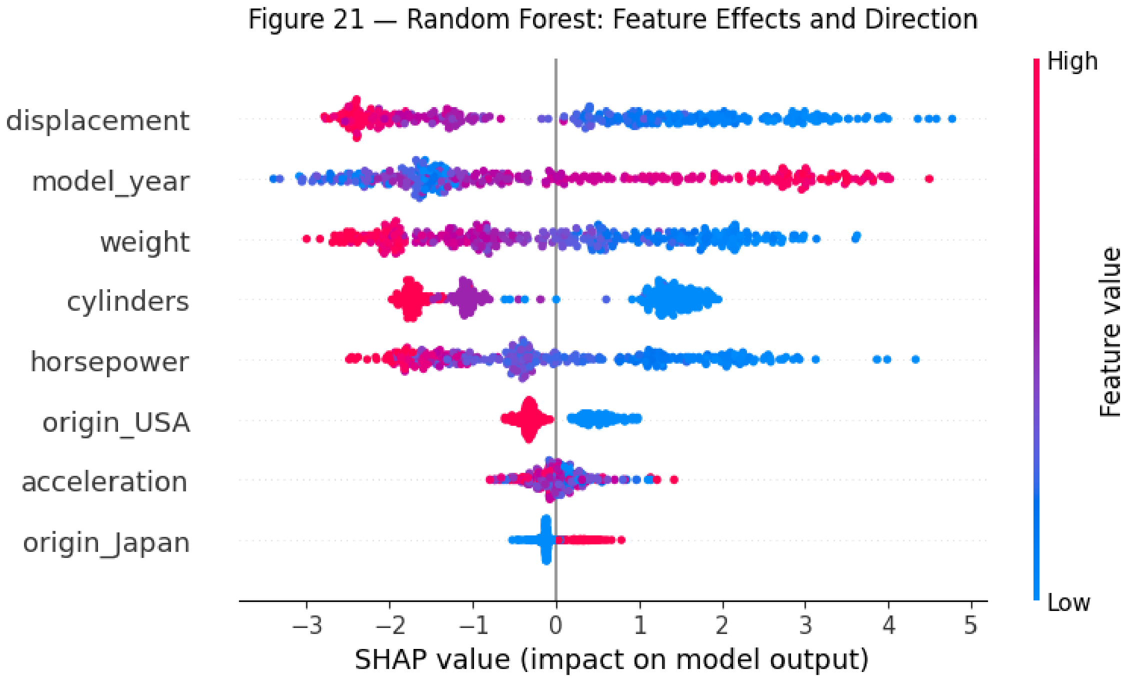


Figure 22 — Random Forest: Dependence Plot for Top Feature

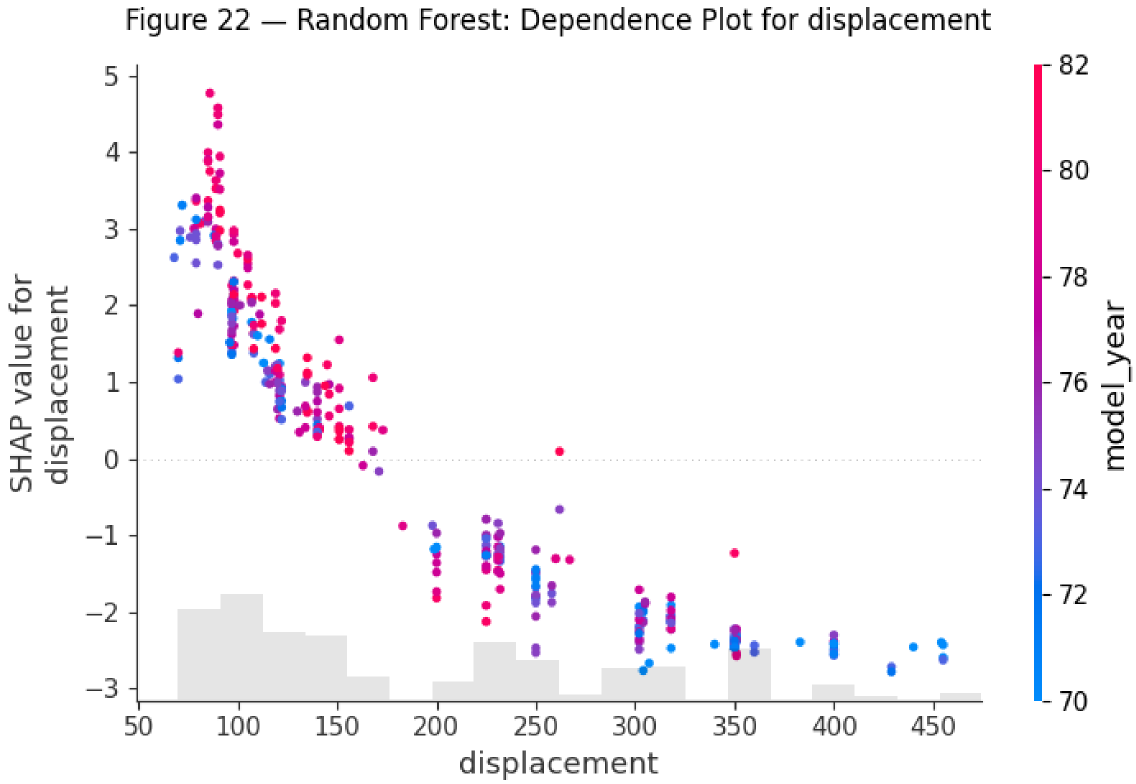


Figure 23 — Gradient Boosting: Global Feature Importance

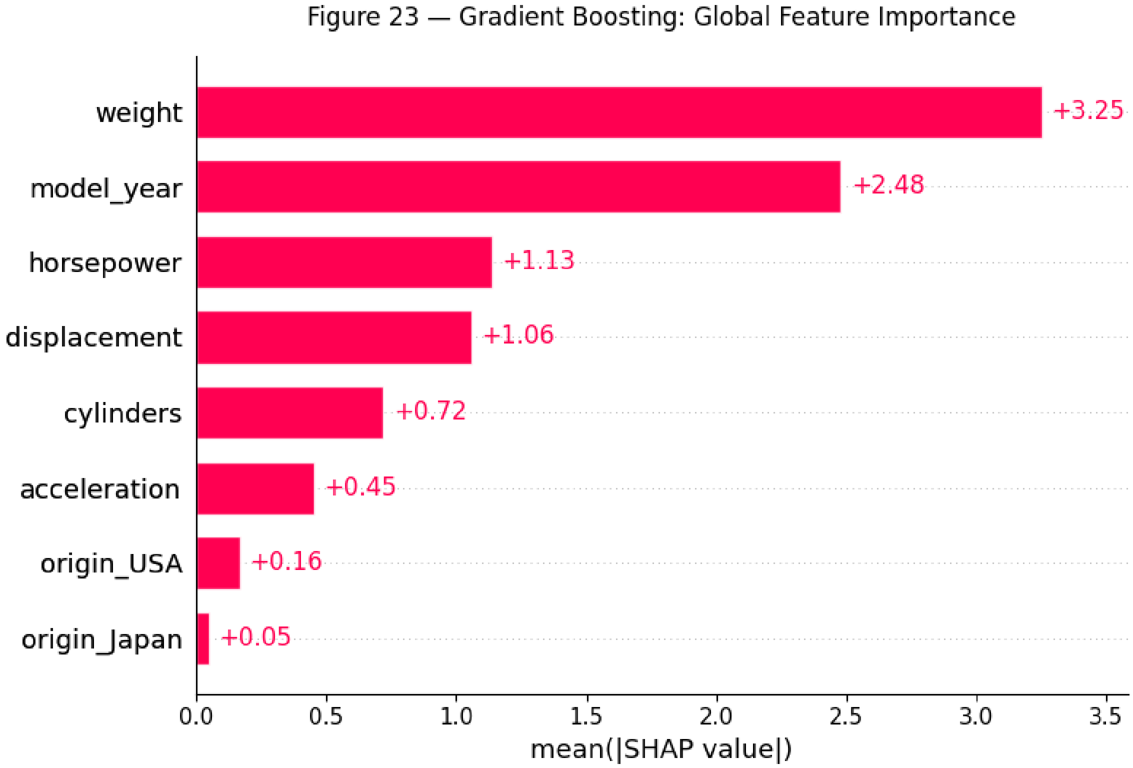


Figure 24 — Gradient Boosting: Feature Effects and Direction

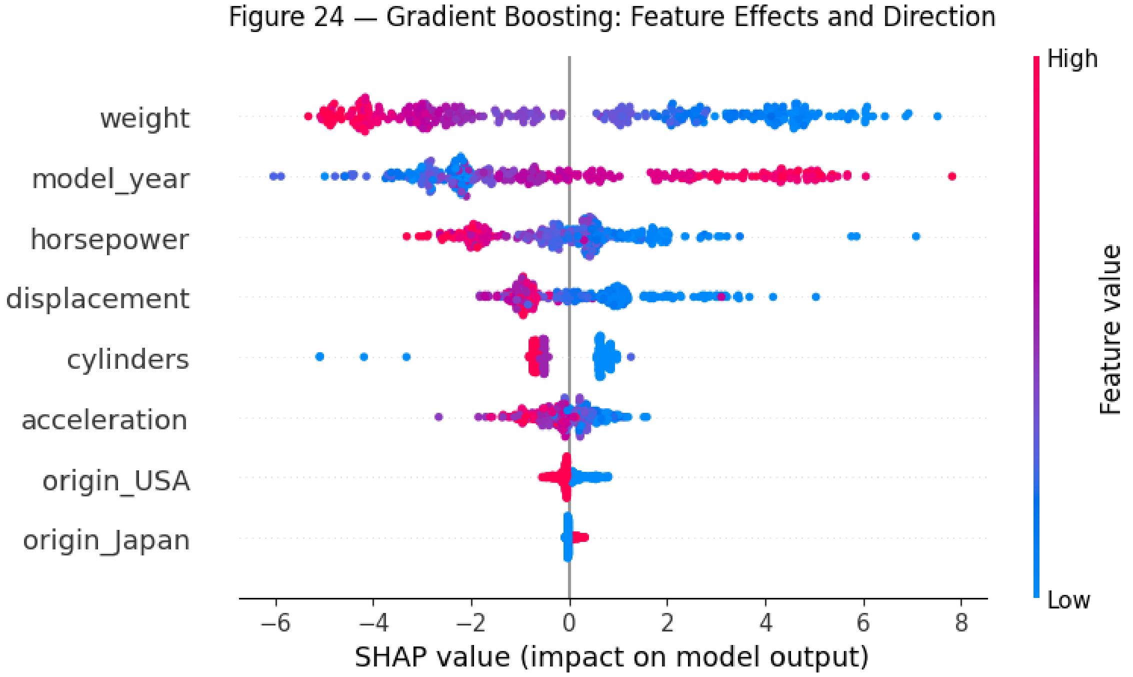
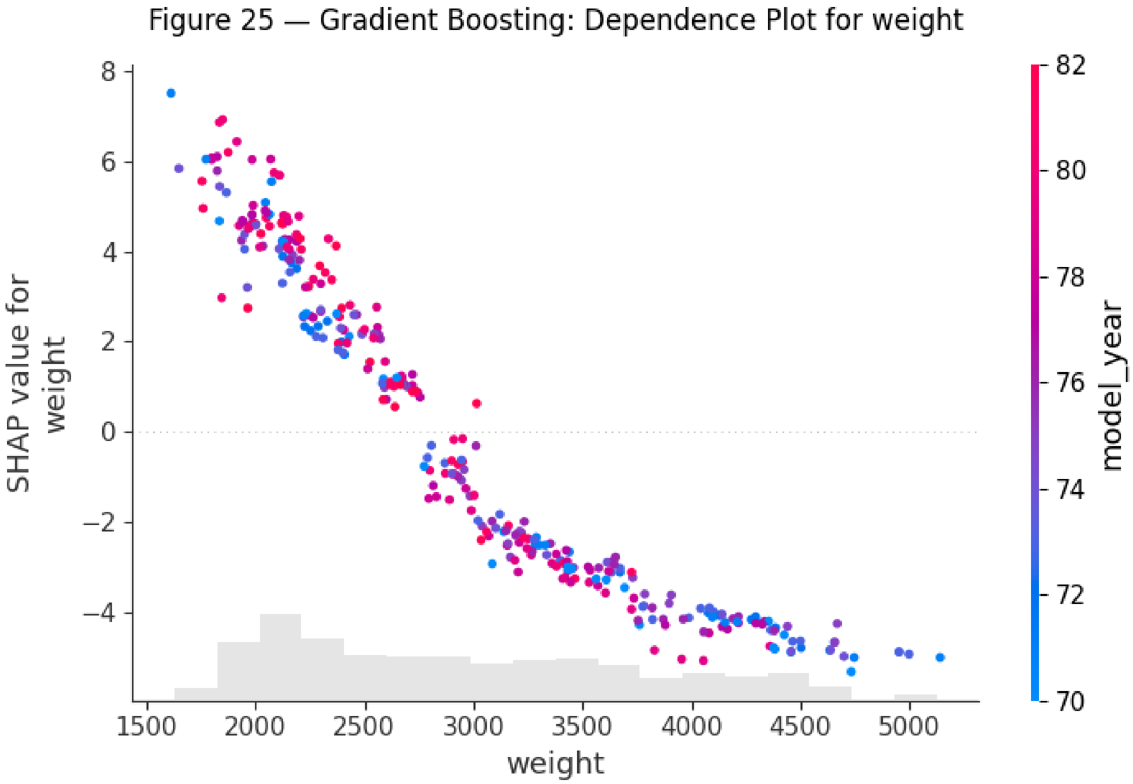


Figure 25 — Gradient Boosting: Dependence Plot for Top Feature



Appendix C

Potential Audience Questions

1. How did you ensure that the models were not overfitting, given the relatively small dataset size?

2. Why did you choose Random Forest as the preferred model over Gradient Boosting, despite similar performance metrics?

3. How would the model performance change if modern vehicle data (e.g., hybrid or electric cars) were included?

4. Can the ensemble modeling framework used here generalize to other energy efficiency domains, such as building or appliance energy prediction?

5. What hyperparameter tuning methods were applied for the ensemble models, and how sensitive were the results to those settings?

6. How might you enhance the interpretability of the ensemble models for use by policymakers or nontechnical stakeholders?

7. What are the environmental or economic implications of the model findings for automotive manufacturers?

8. Did you test polynomial or interaction terms within the linear model to help it capture some nonlinear relationships?

9. How do you plan to handle potential dataset bias due to the model year range (1970–1982)?

10. What steps would you take to deploy this model in a production environment for continuous learning and retraining using live data?

References

- Breiman, L. (2001). *Random forests*. Machine Learning, 45(1), 5–32.

- Dua, D., & Graff, C. (2019). *UCI Machine Learning Repository: Auto MPG Dataset*.

- Friedman, J. H. (2001). *Greedy function approximation: A gradient boosting machine*. *Annals of Statistics, 29(5)*.
- Pedregosa, F., Varoquaux, G., et al. (2011). *Scikit-learn: Machine learning in Python*. *Journal of Machine Learning Research, 12*, 2825-2830.
- McKinney, W. (2010). *Data Structures for Statistical Computing in Python*. SciPy 2010 Proceedings.
- OpenAI. (2025). *ChatGPT 5* (ChatGPT 5) [Auto]. https://chatgpt.com/