Fuel Efficiency Tracker: A Machine Learning Approach for Vehicle Fuel Consumption Prediction

AL Basel Waleed, Asem Eliwa, Mahmoud Mohamed
Nile University
Emails: {a.waleed2227, a.eliwa2248, m.elmahdy2179}@nu.edu.eg

***Abstract*—This paper presents a machine learning-based approach for predicting vehicle fuel efficiency (MPG) using various regression models including Random Forest, XGBoost, and Neural Networks. We use a cleaned and processed car dataset, employing feature encoding and data normalization to improve model performance. The models are evaluated using Mean Squared Error (MSE) and R-squared (R^2) metrics. Random Forest demonstrated the highest accuracy, followed by XGBoost and Neural Networks. Our approach can be integrated into a Fuel Efficiency Tracker tool for real-time vehicle fuel consumption monitoring.**

I. Introduction

Fuel efficiency has become a critical concern in the automotive industry due to increasing environmental regulations, rising fuel costs, and the global push towards sustainability. Accurate prediction of fuel consumption enables consumers and fleet operators to optimize vehicle usage, reduce carbon footprint, and lower operational costs. Traditional methods of assessing fuel efficiency rely heavily on standardized testing, which may not fully capture real-world driving conditions. Machine learning (ML) techniques offer a promising alternative by leveraging historical vehicle data and complex feature interactions to provide more personalized and precise predictions.

This paper presents an approach to fuel efficiency prediction using advanced machine learning models, including ensemble methods such as Random Forest and XGBoost, as well as deep learning models like Neural Networks. We explore data preprocessing techniques, model tuning, and performance evaluation metrics to build robust predictive models. The goal is to develop a reliable Fuel Efficiency Tracker tool that can assist users in understanding and monitoring their vehicle's fuel consumption based on measurable vehicle attributes.

II. Related Work

Over the past decade, numerous studies have investigated the application of machine learning algorithms for fuel consumption prediction. Early research predominantly employed linear regression models due to their simplicity and interpretability; however, these models often fall short when capturing the non-linear and complex relationships inherent in automotive datasets.

Ensemble learning techniques, such as Random Forest and Gradient Boosting Machines (GBMs), have shown superior predictive performance in recent years. For instance, Breiman's Random Forest algorithm aggregates multiple decision trees to improve generalization and reduce overfitting, making it effective for regression tasks with heterogeneous data. Similarly, XGBoost, introduced by Chen and Guestrin, enhances traditional boosting methods by incorporating regularization, parallel processing, and handling missing data, thus achieving state-of-the-art results in many machine learning competitions and industrial applications.

Deep learning approaches have also gained traction due to their ability to model high-dimensional and non-linear feature interactions. Neural networks, with multiple hidden layers, can automatically learn feature representations and have been successfully applied in various domains, including vehicle performance prediction. However, they typically require larger datasets and careful tuning to prevent overfitting and ensure stable convergence.

Several hybrid approaches combining ensemble and deep learning models have been proposed to harness the strengths of both methods, aiming to further boost prediction accuracy and robustness. This body of work motivates the current study's choice of models and methodology.

III. Dataset Description

We used a comprehensive car dataset containing features such as model, year, transmission type, mileage, fuel type, engine size, and manufacturer, alongside the target variable MPG. The dataset was cleaned to remove outliers and categorical features were encoded using Label Encoding.

IV. Methodology

A. Data Preprocessing

Outliers in MPG values were removed (retaining values between 1 and 150). Categorical variables were encoded into numerical values using Label Encoding. The data was split into training and testing sets (80% train, 20% test).

B. Models Used

- **Random Forest Regressor:** An ensemble method that builds multiple decision trees and averages their predictions to reduce overfitting and improve accuracy. Parameters used included 200 trees with max depth 15.
- **XGBoost Regressor:** An advanced gradient boosting technique that sequentially builds trees to correct previous errors. Hyperparameters included 200 estimators, learning rate of 0.05, and max depth 6.

- **Neural Network:** A deep learning model constructed with two hidden layers (128 and 64 neurons respectively) using ReLU activation, Batch Normalization, Dropout regularization, and trained with Adam optimizer.

V. Results and Discussion

TABLE I
Model Performance Comparison

Model	MSE	R^2 Score
Random Forest	12.3303	0.9184
XGBoost	16.2165	0.8927
Neural Network	31.7604	0.7899

Random Forest outperformed other models, achieving the lowest Mean Squared Error and highest R^2 score, indicating superior accuracy and strong generalization ability on the dataset. XGBoost also delivered competitive results with slightly higher error, demonstrating its effectiveness in capturing complex relationships. The Neural Network showed reasonable predictive capability but was less accurate than the ensemble methods, likely due to model complexity and training constraints.

Scatter plots comparing actual versus predicted MPG illustrate that Random Forest's predictions are tightly clustered along the ideal prediction line, confirming its robustness. XGBoost also shows a good fit, while Neural Network predictions exhibit wider deviations, reflecting its comparatively lower performance.

VI. Conclusion

This study explored several machine learning approaches to predict vehicle fuel efficiency using a comprehensive car dataset. Our experiments demonstrated that ensemble methods, particularly Random Forest, provide superior predictive accuracy compared to XGBoost and Neural Networks. The proposed models effectively capture complex relationships between vehicle features and fuel consumption, enabling the development of reliable Fuel Efficiency Tracker tools. Future work could focus on incorporating larger datasets and exploring hybrid models to further enhance prediction performance.

References

[1] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2016.

[2] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[3] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.