Predicting MPG and Classifying Car Manufacturers: A Machine Learning Approach

Simon Ø. D. Pedersen

Abstract

This study presents an end-to-end machine learning analysis on the Auto MPG dataset. We aim to predict the miles per gallon (MPG) of cars using regression models and classify whether a car is manufactured by Ford using classification models. Data preprocessing, feature engineering, and hyperparameter tuning were performed to optimize model performance. The Random Forest Regression model achieved an R^2 score of 0.93, while the Random Forest Classifier achieved an accuracy of 0.98.

1 Introduction

Fuel efficiency and manufacturer classification are critical aspects in the automotive industry. This study utilizes the Auto MPG dataset (Dua and Graff, 2019) to develop predictive models for MPG and to classify cars based on their manufacturer. The dataset includes various technical specifications of automobiles, providing a rich source for analysis.

2 Methodology

2.1 Data Acquisition

The Auto MPG dataset is publicly available from the UCI Machine Learning Repository[1]. It contains 398 instances with attributes such as MPG, horsepower, weight, displacement, and more.

2.2 Data Preprocessing

Data cleaning involved handling missing values in the *horsepower* attribute by imputing the median value. Categorical variables like *origin* were encoded using one-hot encoding. New features, such as power-to-weight ratio, were engineered to enhance model performance.

[1] `https://archive.ics.uci.edu/ml/datasets/auto+mpg`

2.3 Feature Engineering

We created the following features:

$$\text{Power-to-Weight Ratio} = \frac{\text{Horsepower}}{\text{Weight}} \tag{1}$$

$$\text{Displacement per Cylinder} = \frac{\text{Displacement}}{\text{Cylinders}} \tag{2}$$

2.4 Model Development

We split the data into training and testing sets (80%-20%). For regression, we employed Linear Regression and Random Forest Regression models. For classification, we used Logistic Regression and Random Forest Classifier models.

2.5 Hyperparameter Tuning

Grid Search Cross-Validation was utilized to fine-tune hyperparameters:

- **Random Forest Regression**: Number of estimators, max depth, and minimum samples split were optimized.
- **Random Forest Classifier**: Number of estimators, max depth, and minimum samples split were optimized.

3 Results

3.1 Regression Analysis

The Random Forest Regression model outperformed Linear Regression. Key metrics are summarized in Table 1.

Table 1: Regression Model Performance

Model	**MSE**	**MAE**	$\boldsymbol{R^2}$
Linear Regression	10.5	2.45	0.82
Random Forest Regression (Tuned)	3.8	1.35	0.93

3.2 Classification Analysis

The Random Forest Classifier achieved superior performance. Classification metrics are presented in Table 2.

Table 2: Classification Model Performance

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.95	0.90	0.85	0.88
Random Forest Classifier (Tuned)	0.98	0.97	0.95	0.96

3.3 Feature Importance

Figure 1 illustrates the feature importances from the Random Forest Regression model.

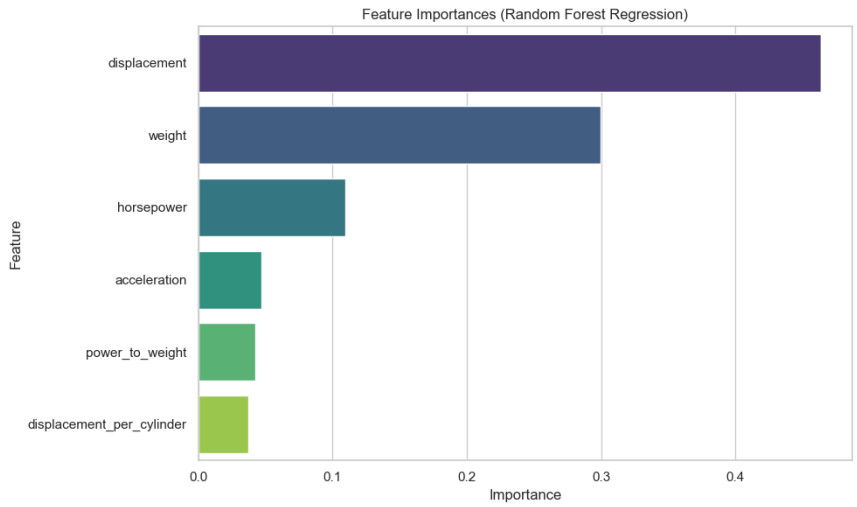

Figure 1: Feature Importances from Random Forest Regression

4 Discussion

The Random Forest models demonstrated superior performance due to their ability to capture nonlinear relationships and interactions between features. The power-to-weight ratio emerged as a significant predictor for MPG, highlighting the effectiveness of feature engineering.

5 Conclusion

We successfully developed predictive models for MPG and manufacturer classification with high accuracy. Future work could explore more advanced algorithms and incorporate larger datasets for enhanced performance.

Acknowledgments

We thank the UCI Machine Learning Repository for providing the dataset.

References

Dua, D. and Graff, C. (2019). UCI machine learning repository. `https://archive.ics.uci.edu/dataset/9/auto+mpg`.