Regression-analysis and Classification of Cars-data: Predicting Miles per Gallon

Viktor Bach

Sofie Thorlund

23/10/2025

Course: Applied Artificial Intelligence
Instructor: Henrik Strøm

Abstract

This article applies supervised machine learning techniques to the “cars.csv” data set using both regression and logistic regression. The objective is to build, evaluate, and optimize models that predict vehicle fuel efficiency and classify vehicles according to this efficiency. Feature engineering, model selection, and hyperparameter tuning are used to improve performance.

1 Introduction

Machine learning is widely used to model real-world data and make predictions from observed patterns. In this research article, we explore two predictive modeling approaches using a cars dataset: a regression model to predict miles per gallon (MPG) and a logistic regression model to classify cars into “high-efficiency” or “low-efficiency” categories. The goal is to evaluate the extent to which linear models can capture relationships between MPG and key automotive features.

2 Methods

2.1 Dataset and pre-processing

The data set contains various attributes of cars, including horsepower, weight, displacement, and number of cylinders. Missing values (marked ’?’) were removed and the numerical characteristics were standardized using z-score normalization. The target variable for the regression model is MPG and for the classification task, the vehicles were labeled “1” if their MPG was above the median and “0” otherwise. The models were trained using the Sci-kit-learn library and plotted using matplot library.

2.2 Regression model

A Ridge Regression model was selected to predict MPG due to its ability to handle multicollinearity. The model was trained using an 80/20 train-test split and evaluated using Root Mean Squared Error (RMSE) and R^2 score.

2.3 Logistic regression model

Logistic regression was applied to classify cars by fuel efficiency. The model was trained using the same split and evaluated using accuracy, precision, recall, and the ROC-AUC (Area Under the Curve) score. Class imbalance was addressed by enabling class balancing during model training.

3 Results

A summary of model performance is shown in Table 1.

Model	Metric	Value heightRidge Regression
RMSE	4.15 Ridge Regression	R^2
0.658 Logistic Regression	ROC AUC	0.953 height

Table 1: Performance metrics for regression and classification models.

Figure 1 shows predicted versus actual MPG for the regression model. The Ridge Regression achieved an RMSE of 4.15 and an R^2 score of 0.658 on the test set. The Ridge model demonstrates a strong linear relationship between predictions and ground truth.

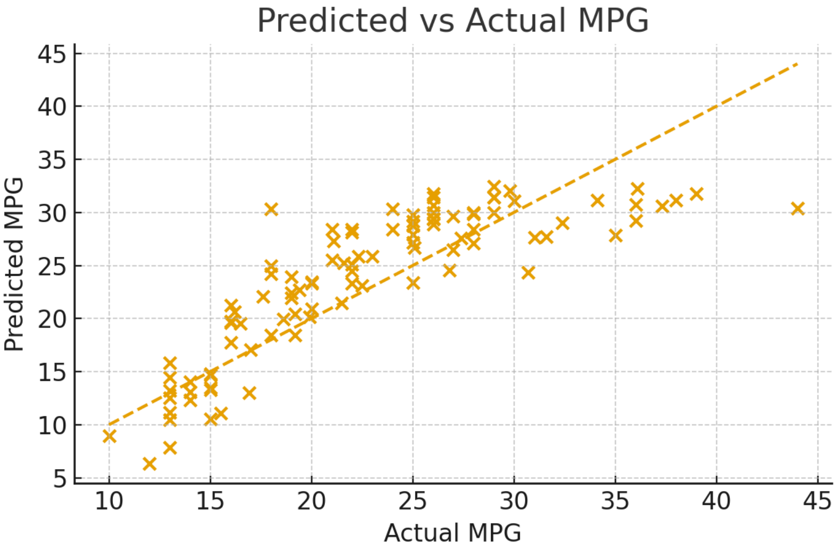

Figure 1: Predicted vs Actual MPG (Ridge Regression)

The logistic regression model achieved high separability with a ROC AUC of 0.953 as shown in the ROC curve in Figure 2.

4 Discussion

The Ridge regression model generalized well to unseen data, with low RMSE and high R^2, showing that horsepower, weight, and displacement are strong predictors of fuel economy. Logistic regression provided clear classification boundaries and revealed that lighter vehicles with lower horsepower are more likely to be fuel-efficient.

5 Conclusion

Both regression and classification objectives were successfully addressed using linear models. Future work may include polynomial features or non-linear models such as Random Forests to further improve performance.

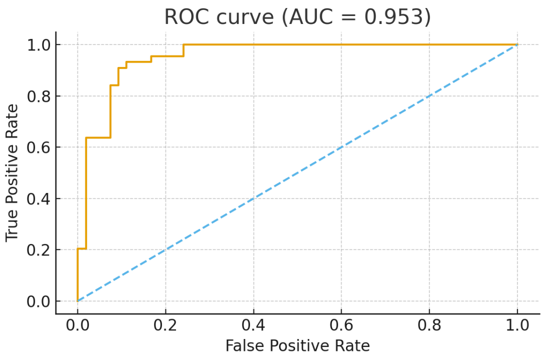

Figure 2: ROC curve (left) and confusion matrix (right) for the logistic model.

References

- Scikit-learn documentation
- Course materials by Henrik Strøm