Data Science Recommendations

A comprehensive analysis and recommendations report

Your Report

Based on the dataset summary, I provide the following recommendations for the data science workflow:

1. Handling Missing Values

* For columns with a small number of missing values (less than 5%), I recommend filling them with the median or mean for numerical columns (e.g., `bore`, `stroke`, `horsepower`, `peak-rpm`, and `price`) and the mode for categorical columns (e.g., `num-of-doors`).
* For columns with a larger number of missing values (e.g., `normalized-losses`), I recommend creating a 'Missing' category, as this might be a relevant feature in the analysis.
* Dropping rows or columns with missing values might not be the best approach, as it could lead to loss of valuable information and biased results.

2. Recommended Visualizations

* Histograms for numerical columns (e.g., `engine-size`, `horsepower`, `city-mpg`, `highway-mpg`, and `price`) to understand the distribution of values.
* Scatter plots for numerical columns (e.g., `wheel-base` vs. `length`, `width` vs. `height`, and `horsepower` vs. `peak-rpm`) to identify correlations and relationships between variables.
* Box plots for categorical columns (e.g., `make`, `fuel-type`, and `body-style`) to compare

distributions across categories.

* Correlation heatmaps for all numerical columns to identify strong correlations and potentially reduce dimensionality.

3. Machine Learning Model Recommendations

* Based on the dataset, I recommend a regression model, as the target variable `price` is continuous.

* Suitable models include:

 + Linear Regression: a simple and interpretable model for predicting `price` based on other numerical features.

 + Decision Trees: a robust model that can handle non-linear relationships and interactions between features.

 + Random Forest: an ensemble method that can improve the accuracy and robustness of Decision Trees.

 + Gradient Boosting: another ensemble method that can handle complex relationships and interactions between features.

4. Model Evaluation Techniques

* For regression models, I recommend evaluating models using:

 + Root Mean Squared Error (RMSE): measures the average distance between predicted and actual

values.

+ Mean Absolute Error (MAE): measures the average absolute difference between predicted and actual values.

+ R^2-score (Coefficient of Determination): measures the proportion of variance in the target variable explained by the model.

* These metrics provide a comprehensive understanding of the model's performance, including accuracy, bias, and variance.

By following these recommendations, you can develop a robust data science workflow that effectively handles missing values, explores the dataset through informative visualizations, and selects suitable machine learning models with appropriate evaluation metrics.