Predicting Car Prices in the American Market

A Comprehensive Regression Analysis

Description of the Analysis

This project aims to model car prices in the American market using a comprehensive dataset of automobile features. The analysis includes the following steps:

1. Loading and Preprocessing: The dataset was cleaned by handling missing values, encoding categorical variables, and scaling numerical features to prepare it for machine learning models.

2. Model Implementation: Five regression algorithms were implemented:

    ○ Linear Regression
    ○ Decision Tree Regressor
    ○ Random Forest Regressor
    ○ Gradient Boosting Regressor
    ○ Support Vector Regressor

3. Model Evaluation: The models were evaluated using metrics like R-squared, Mean Squared Error (MSE), and Mean Absolute Error (MAE). This helped identify the most effective model for predicting car prices.

4. Feature Importance Analysis: Using the Random Forest Regressor, the most significant factors influencing car prices were identified and visualized, providing insights into the critical variables.

5. Hyperparameter Tuning: The Random Forest model was fine-tuned using GridSearchCV to optimize performance, resulting in improved predictive accuracy.

This analysis not only identified key variables affecting car prices but also provided a robust and well-evaluated model to predict prices in the American automobile market.

Loading and Preprocessing:

○ Load the dataset into a pandas DataFrame to inspect its structure.
○ View column names, data types, and the presence of missing or null values.
○ Summarize the data using `.info()` and `.describe()`.

- Identify and treat missing or null values, if any, using appropriate methods (imputation or removal).
- Remove duplicate rows if they exist.
- Convert categorical features into numerical ones using techniques like one-hot encoding or label encoding.
- Drop unnecessary columns like IDs or any irrelevant attributes based on domain knowledge.
- Normalize or scale numerical features to ensure compatibility with regression models.
- Split the dataset into training and testing sets (e.g., 80%-20% split) for model evaluation.

Regression Modeling:

- Import and set up the algorithms with default parameters.
- Train each algorithm on the training dataset.
- Use the trained models to predict values on the test dataset.
- Evaluate models using R-squared, Mean Squared Error (MSE), and Mean Absolute Error (MAE).
- Compare the results of all models to determine the best-performing algorithm.

Model Evaluation:

To evaluate the models and compare their performance, we'll calculate the following metrics for each regression model:

1. R-squared (R^2): Measures the proportion of the variance in the dependent variable that is predictable from the independent variables.
2. Mean Squared Error (MSE): The average of the squared differences between predicted and actual values.
3. Mean Absolute Error (MAE): The average of the absolute differences between predicted and actual values.

Feature Importance Analysis:

1. Trained a model that provides feature importance (e.g., Random Forest or Gradient Boosting).
2. Extracted feature importance scores.
3. Visualized the top contributing features using a bar chart.

Hyperparameter Tuning:

1. Taken a model like Random Forest or Gradient Boosting for tuning.
2. Specified the range of hyperparameters to test.
3. Performed cross-validation on the dataset to find the optimal hyperparameters.
4. Trained the model with the best parameters and compared its performance.