Comprehensive Report: Car Price Prediction Model

1. Problem Specification

The automotive industry and potential car buyers need reliable estimates of car prices based on vehicle attributes to make informed decisions. Currently, price prediction can be complex due to many interrelated features like engine size, weight, fuel efficiency, and more. Accurately predicting car prices from these diverse features is a challenging problem that requires effective data analysis and modeling.

2. Aim and Objectives

Aim:
To develop a predictive model that accurately estimates car prices using a range of vehicle features including specifications, performance metrics, and categorical attributes.

Objectives:

- Analyze a dataset of 205 cars with 26 features to understand the relationships between car attributes and price.
- Clean and preprocess the dataset, including handling inconsistencies in categorical variables and encoding where necessary.
- Explore and quantify the correlation between features and the car price.
- Build and train a regression model to predict car prices.
- Evaluate the model's performance using key metrics such as RMSE, MAPE, and R-squared.

3. Methodology

Data Description

- The dataset contains **205 rows** (cars) and **26 columns** (features).
- No missing or duplicate values were found, ensuring data integrity.

- Features include numeric attributes (e.g., wheelbase, horsepower, curbweight) and categorical variables (e.g., CarName, fueltype, drivewheel).

Data Summary

- **Numerical Features:**
    - Dimensions: wheelbase (mean ~98.8 in), car length (~174 in), width (~65.9 in), height (~53.7 in)
    - Weight: curbweight ranges from 1,488 to 4,066 lbs (mean ~2,556 lbs)
    - Engine: size (61 to 326), horsepower (48 to 288), cylinders (mostly 4)
    - Fuel economy: city mpg (13–49), highway mpg (16–54)
    - Price ranges widely from $5,118 to $45,400 (mean ~$13,277)
- **Categorical Features:**
    - 147 unique car names; *peugeot 504* is the most frequent (6 occurrences).
    - Fuel type predominantly gas (185 out of 205).
    - Majority are 4-door sedans with front-engine, front-wheel drive layouts.
    - Engine types mostly OHV/OHC and 4 cylinders.
    - Aspiration mainly standard, some turbocharged.

Data Cleaning and Encoding

- Inconsistencies in car brand/model names were corrected via mapping dictionary (e.g., "maxda" → "mazda").
- Door numbers converted from text to numeric (two → 2, four → 4).

Feature Correlation with Price

- Strong positive correlations (>0.6) with price found for:

    - **enginesize (0.87), curbweight (0.84), horsepower (0.81), carwidth (0.76), cylindernumber (0.72), carlength (0.68)**

- Moderate positive correlations (~0.3 to 0.5) for drivewheel, wheelbase, boreratio, fuelsystem, enginelocation.

- Negative correlations observed for fuel efficiency metrics: citympg (-0.69) and highwaympg (-0.70), indicating more expensive cars tend to consume more fuel.

Modeling Approach

- The dataset was split into training and testing sets using an 80:20 ratio.

- Linear Regression was chosen as the baseline predictive model to capture linear relationships between features and price.

- The model was trained on the training set and tested on the unseen test set.

Model Evaluation

- **Mean Absolute Percentage Error (MAPE):** 20.29% — on average, predictions are within ~20% of actual prices.

- **Root Mean Squared Error (RMSE):** 2601.62 — average prediction error in dollars.

- **R-squared (R^2):** 0.78 — the model explains about 78% of the variance in car prices, indicating good fit.

Summary

This study successfully developed a linear regression model to predict car prices based on vehicle attributes with strong predictive performance. Features like engine size, weight, and horsepower are key price drivers, while fuel economy negatively correlates with price. Data cleaning and proper feature encoding were crucial steps before modeling. Future work could explore nonlinear models or ensemble methods to potentially improve accuracy.

If you want, I can help you prepare this in a formatted document or extend with further analysis or visualizations.