Vehicle Price Prediction Report

Objective

The goal of this project is to build a machine learning system that predicts vehicle prices based on their specifications, make, model, and other features. The system leverages a dataset containing detailed information about various vehicles and their prices.

Dataset Description

The dataset (`dataset.csv`) contains the following columns:

Column	Description
name	Full name of the vehicle, including make, model, and trim
description	Brief description, often including key features and selling points
make	Manufacturer of the vehicle (e.g., Ford, Toyota, BMW)
model	Model name of the vehicle
year	Year the vehicle was manufactured
price	Price of the vehicle in USD
engine	Engine details, including type and specifications
cylinders	Number of cylinders in the engine
fuel	Type of fuel used (e.g., Gasoline, Diesel, Electric)
mileage	Mileage of the vehicle (in miles)
transmission	Type of transmission (e.g., Automatic, Manual)
trim	Trim level, indicating different feature sets or packages
body	Body style (e.g., SUV, Sedan, Pickup Truck)
doors	Number of doors
exterior_color	Exterior color of the vehicle
interior_color	Interior color of the vehicle
drivetrain	Drivetrain (e.g., All-wheel Drive, Front-wheel Drive)

Approach

1. **Data Preprocessing**:
   - Loaded the dataset and dropped rows with missing values.
   - Identified numerical and categorical columns.
   - Built preprocessing pipelines:
     - Numerical features: Imputed missing values with the mean and applied standard scaling.
     - Categorical features: Imputed missing values with a constant and applied one-hot encoding.
2. **Modeling**:
   - Used a `RandomForestRegressor` as the prediction model.

- Combined preprocessing and modeling into a single pipeline.
- Performed hyperparameter tuning using `GridSearchCV` with 5-fold cross-validation.

3. **Evaluation**:
   - Evaluated the best model on a held-out test set (20% of the data).
   - Metrics used: Root Mean Squared Error (RMSE) and R^2 score.

Results

The best model and its performance on the test set are summarized below:

Metric	Value
Best Params	{'regressor___max_depth': 20, 'regressor___min_samples_split': 2, 'regressor___n_estimators': 200}
Test RMSE	8259.34
Test R^2	0.81

- **Best parameters**: `max_depth=20`, `min_samples_split=2`, `n_estimators=200`
- **Test RMSE**: 8259.34
- **Test R^2**: 0.81

Conclusion

The Random Forest model, after hyperparameter tuning, achieved a strong R^2 score of 0.81 and a test RMSE of \$8,259.34. This indicates the model can explain 81% of the variance in vehicle prices based on the provided features. Further improvements could be made by exploring additional feature engineering, advanced models, or external data sources.