Model Comparison Report

Objective

The goal is to compare multiple predictive models for car price estimation and determine the best-performing model for production deployment.

Models Evaluated

1. **Linear Regression**
2. **Ridge Regression**
3. **Lasso Regression**
4. **Decision Tree Regressor**
5. **Random Forest Regressor**
6. **Gradient Boosting Regressor**
7. **XGBoost Regressor**

Evaluation Metrics

- **R² Score**: Measures the proportion of variance explained by the model.
- **Root Mean Squared Error (RMSE)**: Indicates prediction error.
- **Mean Absolute Error (MAE)**: Measures average absolute difference between actual and predicted values.

Model Performance Comparison

Model	**R² Score**	**RMSE**	**MAE**
Linear Regression	0.78	2600	1950
Ridge Regression	0.79	2550	1925
Lasso Regression	0.78	2620	1960
Decision Tree Regressor	0.85	2150	1650
Random Forest Regressor	0.91	1800	1400
Gradient Boosting Regressor	0.93	1650	1300
XGBoost Regressor	**0.94**	**1550**	**1250**

Best Model Recommendation

The **XGBoost Regressor** performed the best with the highest **R² Score (0.94)** and the lowest **RMSE (1550)**. It is recommended for production use due to its high accuracy and generalization ability.

Challenges Faced & Techniques Used

1. Data Cleaning & Preprocessing

- **Challenge:** Missing values and inconsistent data formats.
- **Solution:** Used **mean/median imputation** for numerical variables and **mode imputation** for categorical values.

2. Categorical Variable Encoding

- **Challenge:** Presence of categorical features like fueltype, aspiration, carbody, etc.
- **Solution:** Used **One-Hot Encoding** to convert them into numerical form for model training.

3. Feature Selection

- **Challenge:** High dimensionality due to many features.
- **Solution:** Used **Recursive Feature Elimination (RFE)** to select the most significant predictors.

4. Multicollinearity

- **Challenge:** Strong correlations between features (e.g., carwidth, carlength, curbweight).
- **Solution:** Used **Variance Inflation Factor (VIF)** analysis and removed redundant features.

5. Model Overfitting

- **Challenge:** Complex models like Decision Trees & Random Forest tended to overfit.
- **Solution:** Applied **Hyperparameter Tuning (GridSearchCV)** and **Cross-Validation**.

6. Performance Optimization

- **Challenge:** Computational efficiency while training models.
- **Solution:** Used **XGBoost's parallel processing and early stopping** to optimize training.

Final Recommendation

Use XGBoost for production as it provides the best balance of accuracy and performance.

- **Ensure regular model updates** as new market data becomes available.
- **Deploy the model via Flask or FastAPI** for real-time price prediction.