**Used Car Price Prediction: Methodology and Model Evaluation**

**1. Introduction**

This document details the methodology, models used, and evaluation results for the used car price prediction project. The goal is to develop a predictive model that accurately estimates the price of a used car based on various features.

**2. Methodology**

**2.1 Data Preprocessing**

* The dataset was preprocessed, ensuring missing values were handled and categorical variables were encoded.
* Outliers were treated using the IQR method.
* A train-test split was performed, with 80% of data used for training and 20% for testing.

**2.2 Feature Selection**

* Important features were identified using Random Forest feature importance and correlation analysis.
* The most relevant features were used for training to improve model efficiency.

**2.3 Models Used**

Six regression models were trained and evaluated:

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

**3. Model Training and Evaluation**

**3.1 Cross-Validation Performance**

Each model was evaluated using 5-fold cross-validation, and the performance was measured using Mean Absolute Error (MAE), Mean Squared Error (MSE), and R² score.

**Cross-validation summary:**

* **Linear Regression:** Moderate performance, R² = 0.5503.
* **Decision Tree:** Better than Linear Regression, R² = 0.7335.
* **Random Forest:** Strong performer, R² = 0.8655.
* **Gradient Boosting:** Competitive, R² = 0.8345.
* **Lasso:** Underperformed, R² = 0.4713.
* **Ridge:** Similar to Linear Regression, R² = 0.5503.

**3.2 Hyperparameter Tuning**

Grid search with cross-validation was applied to optimize each model:

* **Decision Tree:** Best parameters - max\_depth=None, min\_samples\_leaf=4, min\_samples\_split=2.
* **Random Forest:** Best parameters - max\_depth=20, min\_samples\_split=2, n\_estimators=200.
* **Gradient Boosting:** Best parameters - learning\_rate=0.1, max\_depth=5, n\_estimators=200.
* **Lasso:** Best parameter - alpha=0.1.
* **Ridge:** Best parameter - alpha=1.

**3.3 Final Model Evaluation on Test Set**

Models were evaluated on the test data:

* **Linear Regression:** R² = 0.5164.
* **Decision Tree:** R² = 0.8294.
* **Random Forest:** R² = 0.8830.
* **Gradient Boosting:** R² = 0.8857.
* **Lasso:** R² = 0.5109.
* **Ridge:** R² = 0.5164.

**3.4 Best Model Selection**

* The **Gradient Boosting Regressor** had the highest R² score (0.8857), making it the best-performing model.
* The model was saved using pickle for future use.

**4. Conclusion**

* Gradient Boosting was the best model, showing strong predictive power.
* The model is now ready for deployment in the Streamlit app for user interaction and price estimation.

s