**Car Sales Forecasting Models**

our target is the car price (a regression problem). Below are the performance metrics for each model:

**1. Decision Tree**

**Performance Metrics:**

* **Train Results:**
  + **R²:** 0.9998  
    *Almost perfect fit on training data, indicating severe overfitting.*
  + **Mean Absolute Error (MAE):** $2.17  
    *Extremely low error on train – a sign the tree is memorizing the training data.*
  + **Mean Squared Error (MSE):** 37,303.56
  + **Root Mean Squared Error (RMSE):** $193.14
  + **Accuracy-like Measure (within 10%):** 99.99%
* **Test Results:**
  + **R²:** 0.4717  
    *Only about 47% of the variance is explained on unseen data.*
  + **MAE:** $4,422.45
  + **MSE:** 116,816,936.69
  + **RMSE:** $10,808.19
  + **Accuracy-like Measure (within 10%):** 72.63%

**Insights:**

* **Overfitting:** The decision tree almost perfectly fits the training set but generalizes only moderately on the test set.
* **Reliability:** Its extreme training performance signals that the tree has learned noise and specific patterns from training data that do not transfer to new data.

**2. Random Forest**

**Performance Metrics:**

* **Train Results:**
  + **R²:** 0.9525  
    *Very high, indicating the forest fits the training data well without perfect memorization.*
  + **MAE:** $1,694.64
  + **MSE:** 10,548,415.31
  + **RMSE:** $3,247.83
  + **Accuracy-like Measure (within 10%):** 78.09%
* **Test Results:**
  + **R²:** 0.6627  
    *About 66% of the variance is explained on the test set – a solid performance improvement over the decision tree.*
  + **MAE:** $4,435.82
  + **MSE:** 68,502,619.31
  + **RMSE:** $8,276.63
  + **Accuracy-like Measure (within 10%):** 57.30%
* **Cross-Validation:**
  + **CV Accuracy Scores:** [0.6623, 0.6258, 0.6892, 0.5845, 0.6108]
  + **Mean CV Accuracy:** ~63.45% with a low standard deviation (0.0372)
  + This suggests stable performance across folds.

**Insights:**

* **Balanced Performance:** Random Forest provides a good compromise between fitting the data and generalizing to new data.
* **Improved Generalization:** With a higher test R² (0.6627) compared to the decision tree, it demonstrates better predictive performance.
* **Reliability:** Although training metrics are very high, the test set performance confirms that the ensemble averages out the noise.

**3. XGBoost (Randomized Search Tuned)**

**Performance Metrics:**

* **Train Results:**
  + **R²:** 0.7747  
    *Indicates a moderate fit – not as high as Random Forest, suggesting the model is not overfitting much but might be underfitting as well.*
  + **MAE:** $8,344.52
  + **MSE:** 124,741,944.00
  + **RMSE:** $11,168.79
  + **Accuracy-like Measure (within 10%):** 19.30%
* **Test Results:**
  + **R²:** 0.3948  
    *Explains only about 39% of the variance on the test set, which is lower than both Decision Tree and Random Forest.*
  + **MAE:** $8,533.40
  + **MSE:** 133,821,816.00
  + **RMSE:** $11,568.14
  + **Accuracy-like Measure (within 10%):** 19.25%

**Insights:**

* **Underfitting Issues:** Both the training and test performance metrics indicate that XGBoost is underfitting the data (low R², high error metrics, and a very low accuracy-like measure).
* **Potential Tuning Required:** Despite performing hyperparameter tuning with randomized search, these metrics suggest that either the chosen parameter grid or feature set may need further improvement.
* **Comparative Performance:** XGBoost currently underperforms relative to Random Forest, which might be due to insufficient model complexity, inadequate parameter tuning, or a need for additional feature engineering.

**Overall Comparison & Recommendations**

* **Decision Tree**:
  + *Pros:* Extremely high training accuracy, high accuracy-like measure on train data.
  + *Cons:* Severe overfitting with much lower test performance.
* **Random Forest**:
  + *Pros:* Provides the best balance between training and test performance (test R² ~66%). Stable cross-validation results.
  + *Cons:* Slight drop in accuracy-like measure on test vs. train.
* **XGBoost** (after randomized search tuning):
  + *Pros:* Potential power when well-tuned; may capture complex nonlinearities with proper adjustments.
  + *Cons:* Currently underfitting; metrics indicate it is not effectively capturing patterns in the dataset.