Car Selling Price Prediction – Final Report

1. Objective

The goal was to build a predictive model that accurately estimates the selling price of cars based on their attributes such as name, km_driven, year, fuel type, transmission, owner, max_power, etc.

2. Process Overview

a. Data Exploration & Preprocessing (EDA)

- Performed EDA using visual and statistical summaries.
- Identified and removed irrelevant rows (e.g., test drives with no owners).
- Detected outliers and skewed distributions (e.g., selling price).
- Log-transformed target variable to normalize distribution.

b. Feature Engineering

- Categorized features into numerical and categorical.
- Applied label encoding and one-hot encoding where appropriate.
- Scaled numerical features using standardization.

c. Model Building

Multiple regression models were implemented and compared:

- Linear Regression
- Decision Tree Regressor
- Random Forest Regressor
- XGBoost
- Support Vector Regressor (SVR)
- LightGBM
- CatBoost

d. Hyperparameter Tuning

Used RandomizedSearchCV for:

- Random Forest Regressor
- CatBoost
- LightGBM
- XGBoost

3. Evaluation Metrics Used

- R^2 Score
- MAE (Mean Absolute Error)
- MSE (Mean Squared Error)
- RMSE (Root Mean Squared Error)

4. Best Model & Accuracy

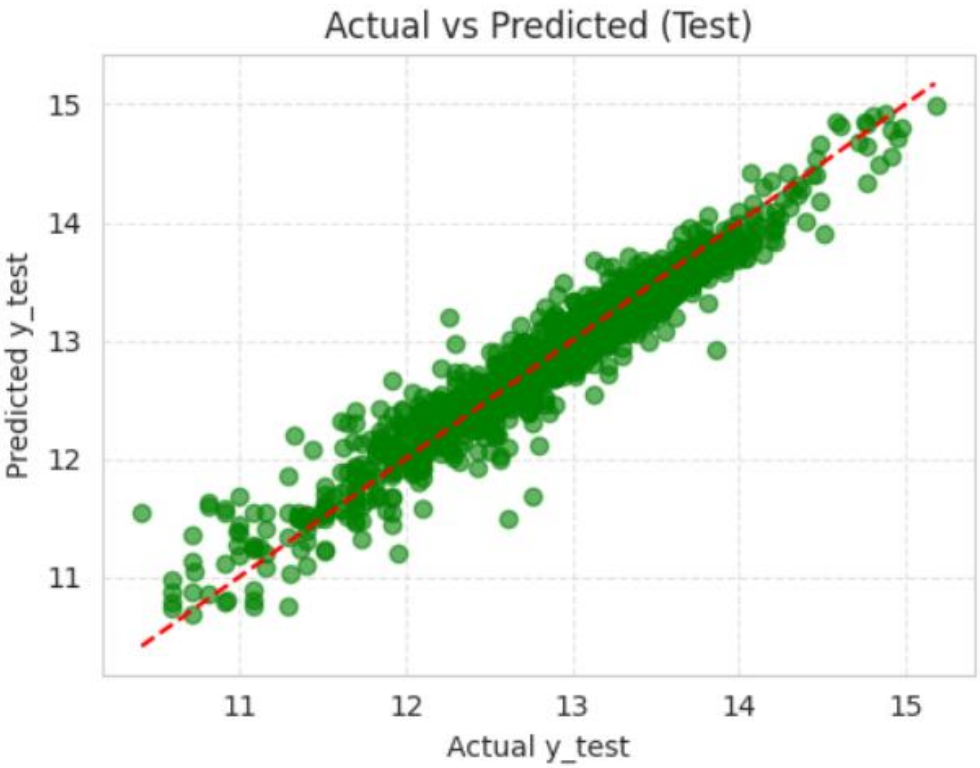
The best-performing model was CatBoost, achieving the highest R^2 and lowest error metrics after tuning.

```
--- Model Evaluation Summary ---
                      Model  R2 Score (Train)  R2 Score (Test)  Mean CV Score (Train)
0         Linear Regression              0.88             0.88                   0.88
1             Decision Tree              1.00             0.85                   0.84
2   Random Forest (Untuned)              0.99             0.91                   0.90
3         XGBoost (Untuned)              0.97             0.92                   0.91
4                       SVR              0.26             0.27                   0.25
5        LightGBM (Untuned)              0.95             0.92                   0.91
6        CatBoost (Untuned)              0.95             0.93                   0.92
7     Random Forest (Tuned)              0.95             0.91                   0.91
8           XGBoost (Tuned)              0.94             0.92                   0.91
9          CatBoost (Tuned)              0.95             0.92                   0.92
10         LightGBM (Tuned)              0.95             0.92                   0.92

--- Conclusion ---
Based on the R2 scores and Cross-Validation scores, we can draw the following conclusions:

Overall best performing model(s) based on R2 Score (Test):
- CatBoost (Untuned) with R2 Score (Test) = 0.93
```

5. Role of LLMs in the Workflow

- Exploratory Data Analysis (EDA): Used LLM to interpret patterns, suggest visuals, and guide transformation (e.g., log-transform for skewed targets).
- Feature Engineering: LLM helped identify which variables needed encoding/scaling and created reusable code blocks for transformation pipelines.
- Reusable Code Generation: Quickly generated loops, pipelines, and visualizations using prompts.
- Model Evaluation: LLM helped write reusable evaluation functions and interpret results in a concise manner.

6. Results and Prediction



🔍 Evaluation for Test Sample Index: 421
📉 Actual Selling Price (Log): 12.5602
📈 Actual Selling Price : $285,000.00

Model Comparisons:
🟦 XGBoost Prediction (Log): 12.5477
🟦 XGBoost Prediction : $281,455.47
🟥 CatBoost Prediction (Log): 12.6069
🟥 CatBoost Prediction : $298,620.31