Car Price Prediction

1. Objective:

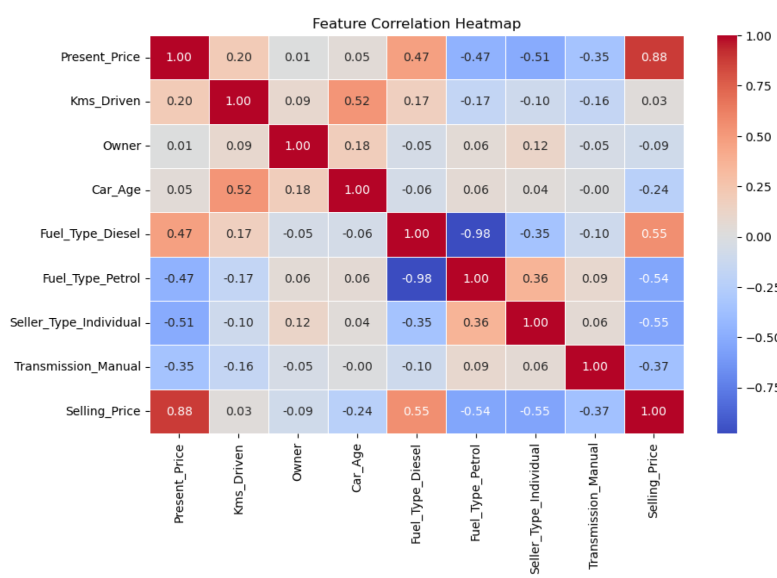
The primary goal of this project is to develop a machine learning model that accurately predicts the resale value of used cars. By analyzing key car attributes, the project aims to identify critical factors that influence price variations and provide insights into the used car market.

2. Dataset Overview:

- The dataset consists of records related to used cars, including specifications, seller types, and selling prices.
- Key features include **Car Age, Kms Driven, Fuel Type, Transmission, and Seller Type**.
- The dataset is source: https://github.com/dishant21001/Car-Price-Prediction

Visualization:

- **Feature Correlation Heatmap** → Helps understand relationships between numerical features.



3. Data Preprocessing

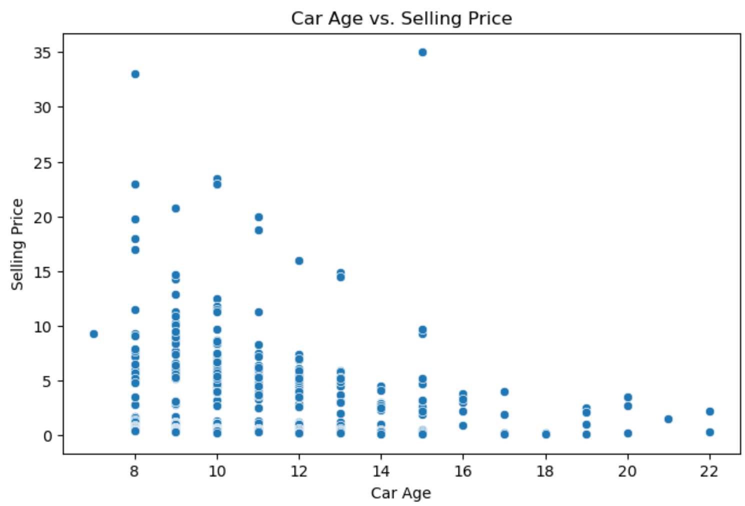
Data Cleaning & Feature Engineering:

1. **Irrelevant Feature Removal**:
    - The **Car Name** column was removed as it does not contribute to price prediction.
2. **Feature Transformation:**
    - A new feature, **Car Age**, was created by subtracting the manufacturing year from the current year.
    - The **Year** column was then removed.
3. **Encoding Categorical Variables:**
    - Categorical features (**Fuel Type, Seller Type, Transmission**) were converted into numerical values using **one-hot encoding**.
4. **Data Splitting:**
    - The dataset was split into **80% training** and **20% testing** for model evaluation.

Visualizations:

- **Price Distribution Histogram** → Displays selling price distribution before training.

- **Car Age vs. Selling Price Scatter Plot** → Highlights depreciation trends.



Model Selection & Evaluation

Machine Learning Models Used:

1. **Linear Regression** – Evaluates the linear relationship between car attributes and selling price.
2. **K-Nearest Neighbors (KNN)** – Uses neighboring data points to predict car prices.
3. **Decision Tree Regressor** – Captures non-linear relationships in the dataset.

Model Performance Metrics:

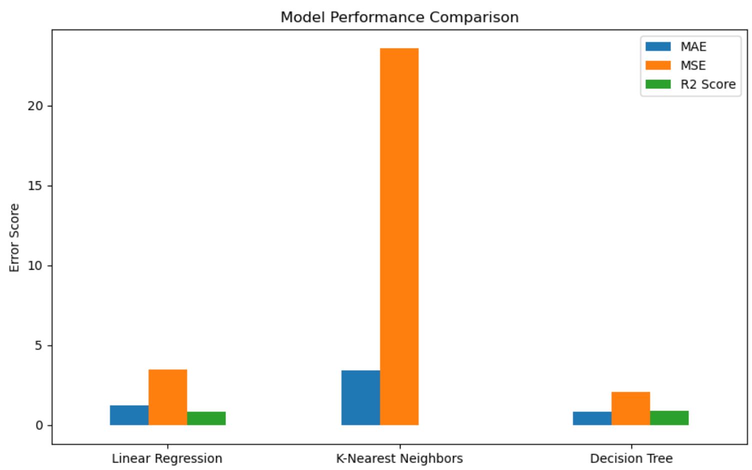
- **Mean Absolute Error (MAE):** Measures average prediction error.
- **Mean Squared Error (MSE):** Penalizes larger errors more heavily.
- **R^2 Score:** Reflects the proportion of variance explained by the model.

Initial Model Evaluation:

Model	MAE	MSE	R^2 Score
Linear Regression	1.21	3.78	0.84
K-Nearest Neighbors	3.41	23.57	0.55
Decision Tree Regressor	**0.78**	**1.46**	**0.93**

Visualization:

- **Model Performance Comparison Bar Chart** → Compares MAE, MSE, and R^2 scores across models.



4. Hyperparameter Tuning (Decision Tree Optimization)

Optimization with GridSearchCV:

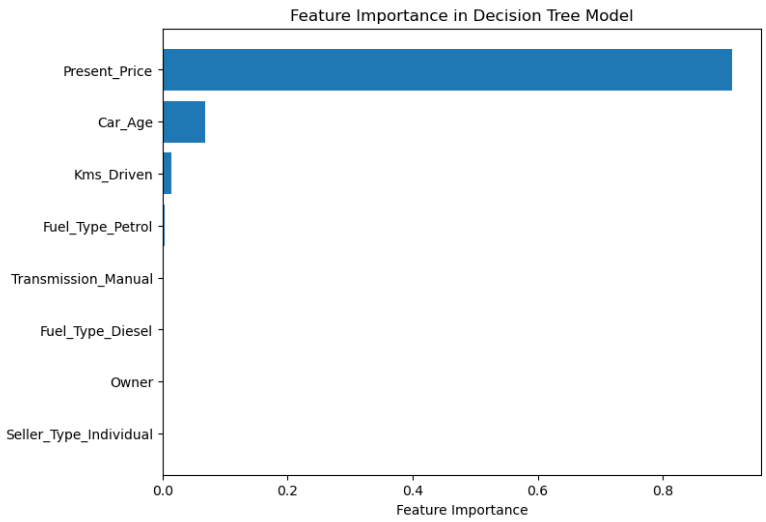
- Tuned parameters: **max_depth, min_samples_split, min_samples_leaf**.
- The best parameters obtained:
  - max_depth=None
  - min_samples_split=2
  - min_samples_leaf=1

Final Model Evaluation:

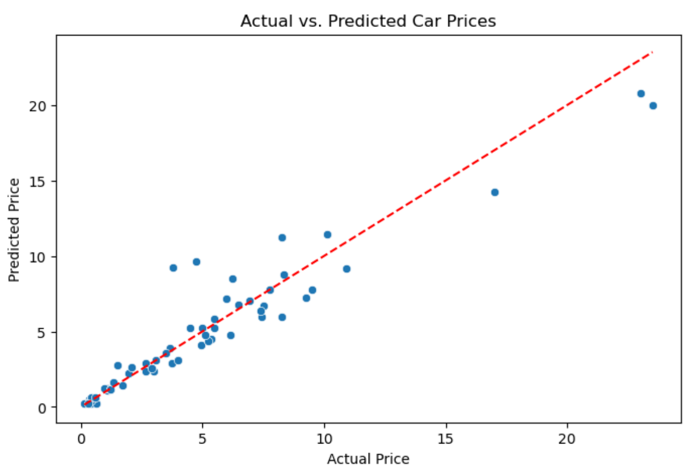
- **Test MAE:** 0.744
- **Test MSE:** 1.46
- **Test R^2 Score:** 0.93

Visualization:

- **Feature Importance Plot** → Displays the impact of each feature on price prediction.



- **Actual vs. Predicted Price Scatter Plot** → Measures how well the model predicts car prices.



This confirms that the **Decision Tree Regressor**, after tuning, provides highly accurate car price predictions.

5. Visual Analysis & Insights

Key Visualizations Used:

1. **Feature Correlation Heatmap** – Highlights relationships between features.
2. **Price Distribution Histogram** – Shows how car prices are distributed.
3. **Car Age vs. Selling Price Scatter Plot** – Displays depreciation trends.
4. **Actual vs. Predicted Prices** – Measures model accuracy.
5. **Feature Importance (Decision Tree)** – Identifies the most impactful features.

6. **Model Performance Comparison** – Compares error metrics across models.

6. Conclusion & Future Work

Key Takeaways:

- The **Decision Tree Regressor** outperformed other models due to its ability to capture non-linear relationships.
- Feature engineering (e.g., **Car Age**) significantly improved model performance.
- Hyperparameter tuning further enhanced prediction accuracy.

Potential Improvements:

- Implement **ensemble methods** (Random Forest, Gradient Boosting) for further performance gains.
- Perform **feature importance analysis** to identify the most influential attributes.
- Explore additional data sources to improve prediction reliability.