Sapience Edu Connect Pvt Ltd

Data Science Internship

Week 1 Task

Observations about Data set

- The dataset contains **8128 rows** and **12 columns**.
- Columns such as `mileage(km/ltr/kg)`, `engine`, `max_power`, and `seats` have missing values.
- The column data types are appropriate for their contents:
    - Categorical: `name`, `fuel`, `seller_type`, `transmission`, `owner`
    - Numeric: `year`, `selling_price`, `km_driven`, `mileage(km/ltr/kg)`, `engine`, `seats`
- Some columns like `max_power` might require cleaning if inconsistencies exist.

Key Insights from the Dataset

1. Price Distribution

- Most cars in the dataset have selling prices below **₹10 lakhs**.
- A small fraction of cars fall into the luxury segment with prices exceeding ₹10 lakhs.

2. Fuel Type Trends

- **Diesel** and **Petrol** are the most common fuel types.
- **CNG** and **LPG** cars are less frequent, possibly due to niche demand or regional availability.

3. Transmission Insights

- **Manual transmission** cars dominate lower price ranges, indicating their popularity in budget-friendly segments.
- **Automatic transmission** cars are more common in higher price ranges, typically in premium models.

4. Age of Cars

- Newer cars (recent manufacturing years) generally have higher selling prices.
- This suggests a clear depreciation trend as cars age, impacting their resale value.

5. Engine Size

- Cars with larger engines (measured in cc) tend to have higher prices.
- This could be attributed to higher performance, premium features, or luxury branding associated with larger engines.

6. Mileage and Fuel Type

- Diesel cars typically have higher mileage, making them a popular choice for long-distance drivers.
- Petrol cars offer moderate mileage, while CNG/LPG cars cater to cost-conscious users seeking fuel efficiency.

Patterns Observed

- **Price-Engine Relationship**: A direct correlation exists where larger engine capacities drive higher prices.
- **Price Depreciation**: The resale price of cars sharply decreases as they age, highlighting the importance of a car's year in determining its market value.
- **Fuel Type Preferences**: Diesel cars dominate due to better mileage, but petrol cars remain popular for their availability and lower initial cost.

Determination of Machine Learning Feasibility

It's a **Regression Problem** ,Because

- The target variable in this dataset appears to be `selling_price`, which is a **continuous numerical variable**.
- The goal would likely be to **predict the selling price** of a car based on various features such as `year`, `fuel`, `engine`, `mileage`, etc.
- Continuous numerical targets naturally point to a regression problem.

We can Use **Supervised Learning approach** here,Because

  - ○ In supervised learning, we use labeled data where the target variable (`selling_price`) is known.

- Since the dataset includes both the features (e.g., `year`, `engine`, etc.) and the target variable (`selling_price`), this problem falls under supervised learning.
- If we aim to predict or model the relationship between features and the target, supervised learning methods such as Linear Regression, Decision Trees, or Random Forests are suitable.