Car Price Prediction Model: Project Report

Project Objective: To develop a machine learning model using Linear Regression to accurately predict the selling price of used cars based on a given set of features.

1. Data Understanding and Exploration

The project began by loading the CarPricePrediction.csv dataset. An initial inspection revealed the following:

- **Dataset Shape:** The dataset contains 4,340 entries and 8 columns.
- **Columns:** name, year, selling price, km driven, fuel, seller type, transmission, and owner.
- **Data Types:** The data consists of a mix of numerical (year, selling price, km driven) and categorical (name, fuel, seller type, transmission, owner) data types.
- **Missing Values:** The dataset was found to be clean, with no missing values, which simplified the preprocessing stage.
- **Feature Engineering:** A new feature, car age, was created by subtracting the year of manufacture from the year (2024). This provides a more intuitive measure for the model than the manufacturing year itself. The original year and name columns were dropped as name had too many unique values to be practical for one-hot encoding.

2. Exploratory Data Analysis (EDA)

EDA was performed to uncover patterns and insights from the data:

- **Selling Price Distribution:** The distribution of selling price is right-skewed, indicating that most cars have a lower price, with a few high-priced outliers.
- **Price vs. Car Age:** A clear negative correlation was observed; as the age of the car increases, its selling price tends to decrease significantly.
- **Price vs Km Driven:** Similarly, cars with higher km driven generally have lower selling prices.
- **Categorical Features:**
  - **Fuel Type:** Diesel cars, on average, command higher prices than petrol cars.
  - **Seller Type:** Cars sold by a Dealer have a higher median price than those sold by an Individual.

    - **Transmission:** Automatic cars tend to be more expensive than manual cars.

- **Correlation Heatmap:** The heatmap of numerical features confirmed a strong negative correlation between selling price and car age, and a moderate negative correlation with km driven.

3. Data Preprocessing

To prepare the data for the linear regression model, the following steps were taken:

- **Categorical Encoding:** Categorical features (fuel, seller type, transmission, owner) were converted into numerical format using **One-Hot Encoding**. This method was chosen to prevent the model from assuming any ordinal relationship between categories.
- **Data Splitting:** The dataset was split into training (80%) and testing (20%) sets to train the model and evaluate its performance on unseen data.

4. Model Development and Evaluation

A **Linear Regression** model was chosen for its simplicity and interpretability.

- **Training:** The model was trained on the preprocessing training data.
- **Prediction:** Predictions were made on the test set.
- **Evaluation:** The model's performance was assessed using standard regression metrics:
    - **Mean Absolute Error (MAE):** 1.25
    - **Mean Squared Error (MSE):** 4.07
    - **Root Mean Squared Error (RMSE):** 2.02
    - **R-squared** (R2) **Score:** 0.40

The R2 **score of 0.40** indicates that approximately **40%** of the variability in the selling price can be explained by the features included in our model. This is a respectable result for a baseline linear model. The RMSE of 2.02 suggests that, on average, the model's predictions are off by about 2.02 lakhs.

5. Model Interpretation and Conclusion

- **Key Feature Impacts:**
    - **Car Age:** The most significant factor, with each additional year of age decreasing the price.
    - **Km Driven:** Higher mileage leads to a lower price.

    - **Fuel Type (Diesel):** Being a diesel car significantly increases the predicted price compared to a petrol car.
    - **Seller Type (Dealer):** Cars sold by dealers are predicted to be more expensive.
- **Model Strengths and Limitations:**
    - **Strength:** The model is simple to understand and its predictions are easily interpretable from the feature coefficients.
    - **Limitation:** The model's assumption of linearity may not capture more complex, non-linear relationships in the data. The 40% R-squared value(R2 Score) suggests there is room for improvement. The model is likely **underfitting**, meaning it's too simple to capture the underlying structure of the data fully.

Conclusion

The Linear Regression model provides a solid baseline for predicting used car prices, achieving an accuracy (R-squared) of 40.31%. The analysis confirms that a car's age, km driven, fuel type, and seller type are significant predictors of its price. To further improve accuracy, future work could involve exploring more complex models (e.g., Random Forest, Gradient Boosting), performing more advanced feature engineering, and addressing potential non-linear relationships.