Car Price Prediction Model Report

Car Price Prediction Report Using CarDekho Dataset

1. Introduction

Predicting the price of used cars is a valuable task in the automotive market, benefiting both buyers and sellers in making informed decisions. This report explores the creation of a machine learning model to predict car prices using the CarDekho dataset. The dataset includes features such as car type, kilometers driven, ownership history, brand, model, manufacturing year, fuel type, and more, which are all critical in determining the resale value of a vehicle.

Additionally, a **Streamlit app** was created as a user-friendly interface to enable users to easily input car details and receive price predictions using the machine learning model.

2. Data Overview

2.1 Dataset Description

The dataset provided contains 6,171 records with the following attributes:

- **bt**: Body type of the car (e.g., Hatchback, SUV, Sedan).
- **km**: Kilometers driven.
- **owner**: Ownership history (e.g., 1st Owner, 2nd Owner, 3rd Owner).
- **oem**: Original Equipment Manufacturer (brand of the car).
- **model**: Car model.
- **modelyear**: Year of manufacture.
- **price**: Selling price of the car.
- **insurance_validity**: Type of insurance (e.g., Comprehensive, Third Party).
- **fuel_type**: Type of fuel used (Petrol/Diesel).

- **seats**: Number of seats in the car.
- **rto**: Regional Transport Office code.
- **engine_displacement**: Engine size in cubic centimeters.
- **mileage**: Fuel efficiency in kilometers per liter (kmpl).
- **state**: State where the car is being sold.

2.2 Data Cleaning and Preprocessing

Before building a machine learning model, the data was cleaned and preprocessed. The steps included:

- **Handling Missing Values**: Ensured that there were no missing values in the dataset.
- **Encoding Categorical Variables**: Converted categorical variables such as `bt`, `owner`, `oem`, `model`, `insurance_validity`, `fuel_type`, `rto`, and `state` into numerical values using label encoding.
- **Feature Engineering**:
  - **Car_Age**: A new feature was created by subtracting `modelyear` from the current year to represent the car's age.
  - **Insurance_Binary**: The `insurance_validity` feature was converted into a binary variable (1 for Comprehensive, 0 for Third Party insurance).
- **Feature Scaling**: Features like `km`, `engine_displacement`, and `mileage` were scaled to normalize the data.

3. Exploratory Data Analysis (EDA)

3.1 Correlation Analysis

A correlation matrix was computed to identify relationships between the numerical features and the target variable `price`. The following observations were made:

- **Positive Correlations**: `engine_displacement` and `mileage` showed a moderate positive correlation with the car's selling price.
- **Negative Correlations**: `km` (kilometers driven) had a negative correlation with `price`, indicating that cars with higher mileage tend to have lower resale

values.

3.2 Distribution Analysis

- **Price Distribution**: The price distribution was skewed to the right, indicating that most cars in the dataset have a lower selling price.
- **Ownership Analysis**: Cars with fewer previous owners generally had higher resale values.
- **Fuel Type**: Diesel cars, particularly in the SUV category, tend to have a higher resale value compared to petrol cars.

3.3 Categorical Feature Analysis

- **Body Type**: SUVs and Sedans generally have higher resale prices compared to Hatchbacks.
- **State-wise Analysis**: Cars sold in metropolitan areas like Bangalore and Hyderabad fetched higher prices compared to smaller cities.

4. Model Building

4.1 Model Selection

Several regression models were considered for predicting the car price, including:

- **Linear Regression**: Simple baseline model to understand linear relationships.
- **Decision Tree Regressor**: Captures non-linear relationships in the data.
- **Random Forest Regressor**: An ensemble method to improve prediction accuracy by combining multiple decision trees.
- **Gradient Boosting Regressor**: Another ensemble method that builds trees sequentially to reduce errors.

4.2 Model Training and Evaluation

- **Train-Test Split**: The dataset was split into 80% training and 20% testing data.
- **Evaluation Metrics**: Models were evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2) score.

4.3 Model Performance

- **Linear Regression**:

- MAE: 127666.76682834557
- MSE: 32919566948.194935
- R^2: 0.6957906968876024

- **Decision Tree Regressor**:
  - MSE: 18563055524.62411
  - R^2: 0.82845903794332
- **Random Forest Regressor**:
  - MSE: 11678605427.403215
  - R^2: 0.8920781544913426
- **Gradient Boosting Regressor**:
  - MSE: 8842713567.999805
  - R^2: 0.9182845954086506

The Gradient Boosting Regressor outperformed other models in terms of prediction accuracy, making it the best model for this dataset.

5. Streamlit App for Car Price Prediction

To make the car price prediction model more accessible and user-friendly, a **Streamlit app** was developed. The app allows users to input details about a car (such as kilometers driven, year of manufacture, engine size, and other features) and receive a price prediction based on the trained machine learning model.

Features of the Streamlit App:

- **User Input**: The app provides dropdown menus and sliders for users to input details like body type, car age, fuel type, and more.
- **Real-time Price Prediction**: Once the user inputs the car details, the app uses the Gradient Boosting Regressor model to predict the price instantly.
- **Simple Interface**: The app was designed with a simple, intuitive interface to cater to a wide audience, even those without technical expertise.

This Streamlit app serves as a practical tool for both car buyers and sellers, helping them estimate the price of used vehicles quickly and accurately.

6. Conclusion

The Gradient Boosting Regressor was identified as the best model for predicting car prices in the CarDekho dataset, achieving an R^2 score of 0.91. Key factors influencing car prices include kilometers driven, engine displacement, body type, fuel type, and ownership history. The development of the **Streamlit app** further enhanced the accessibility of the model, providing users with an easy-to-use platform for car price predictions. The insights and model developed can be applied to predict the prices of used cars accurately, aiding buyers and sellers in making informed decisions.