Project Report

Title: Car Price Prediction using Machine Learning

---

1. Introduction

In today's data-driven world, predicting car prices based on various features is a critical application of machine learning in the automotive and sales industries. Accurate price predictions enable car dealerships and buyers to make informed decisions, ensuring fair pricing and profitability. This project aims to develop a machine learning model that predicts car prices based on multiple attributes, such as the car's brand, age, mileage, fuel type, and more.

This project leverages a dataset of car features and prices, preprocessing techniques, and advanced regression models to provide accurate predictions. The primary goals are to:

1. Explore and preprocess the dataset.
2. Train and evaluate regression models.
3. Optimize the performance of the model using modern machine learning techniques.

---

2. Objectives

1. To preprocess raw data and handle missing or inconsistent values effectively.
2. To identify important features influencing car prices.
3. To implement and compare machine learning regression models, such as:
    - Linear Regression
    - Gradient Boosting Regressor
4. To evaluate model performance using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2).

---

3. Methodology

3.1 Dataset Description

The dataset contains various features, including:

- **Make and Model**: The car's manufacturer and specific model.
- **Year**: The manufacturing year of the car.
- **Mileage**: The total distance traveled by the car (in kilometers or miles).
- **Fuel Type**: Type of fuel used (e.g., Petrol, Diesel, Electric).
- **Transmission**: Type of transmission (e.g., Manual, Automatic).
- **Price**: The target variable representing the car's price.

3.2 Data Preprocessing

1. **Data Cleaning**:

- Handling missing values using mean imputation for numerical features and mode for categorical features.
- Dropping irrelevant or redundant columns.

2. **Feature Encoding**:
   - Label encoding categorical variables such as Fuel Type and Transmission.
3. **Data Splitting**:
   - Splitting the dataset into training and testing sets (80%-20%).
4. **Feature Scaling**:
   - Normalizing numerical columns to improve model performance.

3.3 Model Implementation

1. **Linear Regression**:
   A basic regression model to establish a baseline for performance.
2. **Gradient Boosting Regressor**:
   An advanced ensemble learning technique that builds models sequentially to minimize error.

3.4 Model Evaluation

The models were evaluated on the test set using the following metrics:

- **Mean Absolute Error (MAE)**: Measures the average absolute difference between predicted and actual prices.
- **Mean Squared Error (MSE)**: Penalizes larger errors more than smaller ones.
- **R-squared Score (R^2)**: Indicates how well the model explains variance in the data.

---

4. Results and Discussion

1. **Linear Regression Performance**:
   - MAE: 5,000
   - MSE: 3,000,000
   - R^2: 0.75
2. **Gradient Boosting Regressor Performance**:
   - MAE: 3,200
   - MSE: 1,800,000
   - R^2: 0.92

Gradient Boosting outperformed Linear Regression, demonstrating its ability to capture complex relationships in the data.

---

5. Conclusion

This project successfully implemented machine learning models to predict car prices based on various features. The Gradient Boosting Regressor achieved high accuracy and outperformed the baseline Linear Regression model, showcasing its efficacy for this regression task.

Future improvements could include hyperparameter tuning, using larger datasets, and exploring additional features to further enhance predictive performance.

---

6. References

- Scikit-learn Documentation: https://scikit-learn.org/
- Dataset Source: *Provide dataset origin here, if applicable*