Car Price Prediction using Machine Learning

1. Introduction

Car price prediction is an important application of machine learning in the automobile industry. Accurate price estimation helps buyers make informed decisions and assists sellers in setting fair prices. Since car prices depend on multiple factors, machine learning models can be used to identify patterns and predict prices effectively.

This project focuses on predicting the selling price of used cars using regression techniques.

---

2. Problem Statement

The objective of this project is to build a machine learning model that can predict the selling price of a car based on its features such as year of manufacture, fuel type, transmission type, kilometers driven, and ownership details.

---

3. Dataset Description

The dataset consists of approximately 300 records with the following attributes:

- Car_Name
- Year
- Selling_Price (Target Variable)
- Present_Price
- Kms_Driven
- Fuel_Type
- Seller_Type
- Transmission
- Owner

The dataset includes both numerical and categorical features.

---

4. Data Preprocessing

The following preprocessing steps are applied:

- Removal of the Car_Name column as it does not contribute to prediction

- Creation of a new feature Car_Age from the year of manufacture

- Encoding categorical variables such as fuel type, seller type, and transmission

- Splitting the dataset into training and testing sets

These steps help improve model performance and interpretability.

---

5. Exploratory Data Analysis (EDA)

Exploratory analysis is performed to understand the relationship between different features and the selling price. Visualizations are used to analyze:

- Distribution of selling prices

- Relationship between car age and selling price

- Impact of fuel type and transmission on car prices

The analysis shows that older cars generally have lower selling prices, while fuel type and present price significantly influence resale value.

---

6. Model Selection

The Random Forest Regressor algorithm is used for this project. It is chosen because:

- It handles non-linear relationships effectively

- It performs well on small to medium-sized datasets

- It reduces overfitting through ensemble learning

---

7. Model Training

The Random Forest model is trained using the training dataset. Multiple decision trees are built internally, and the final prediction is obtained by averaging predictions from all trees.

---

8. Model Evaluation

The model's performance is evaluated using regression metrics such as:

- Mean Absolute Error (MAE)

- Root Mean Squared Error (RMSE)

- R^2 Score

The evaluation results indicate that the model predicts car prices with good accuracy and minimal error.

---

9. Feature Importance

Feature importance analysis reveals that:

- Present price

- Car age

- Kilometers driven

are the most influential factors in determining a car's selling price.

---

10. Conclusion

In this project, a machine learning model was developed to predict car prices using a Random Forest regression approach. Feature engineering and exploratory data analysis played a crucial role in improving prediction accuracy. The model can be useful for estimating fair resale prices of used cars.

---

11. Future Scope

The project can be extended by:

- Using larger and more diverse datasets

- Applying advanced regression models such as XGBoost

- Deploying the model as a web application for real-time predictions