Car Dheko - Used Car Price Prediction

1. Introduction

This project focuses on predicting the prices of used cars based on various features using machine learning techniques. The prediction model aims to provide a reliable estimate for car prices based on parameters such as the car's mileage, brand, model year, body type, fuel type, and more.

Objective

The goal of this project is to build a machine learning model capable of accurately predicting the price of used cars based on historical data. Additionally, a web application has been developed using Streamlit to make predictions based on user input.

---

2. Dataset Overview

The dataset used in this project contains various features about used cars, such as:

- **Categorical Variables**:
    - body_type, brand, model, fuel_type, transmission, insurance_validity(type of insurance available), color, city
- **Numerical Variables**:
    - km (distance driven), modelyear (year of manufacture), price_in_lakh (target variable), mileage, seats, top_features_count, comfort_count, safety_count, displacement(engine cc)

Data Cleaning and Preprocessing

- **Handling Missing Values**:
    - Missing values in the seats, safety_count, and mileage columns were filled using available information and interpolation where necessary.
- **Data Type Conversions**:
    - Columns like price_in_lakh and km were converted to float64 after handling their non-numeric values (e.g., unit conversions).
- **Feature Encoding**:
    - One-hot encoding was applied to all categorical columns, particularly body_type, fuel_type, brand, and city, to convert them into numerical form.
- **Feature Scaling**:
    - StandardScaler was used to normalize all numerical features to ensure all variables have a consistent scale.
- **Handling Outliers**:
    - IQR and Z-score methods were applied to remove outliers from critical columns such as price_in_lakh and km.

---

3. Exploratory Data Analysis (EDA)

Key Insights:

- **Distribution of Price**:
    - Prices are right-skewed, with most cars falling within the lower price range.
- **Correlations**:
    - Features like km (distance driven), modelyear, and mileage have a significant correlation with the price.
- **Feature Importance**:
    - The feature importance analysis revealed that modelyear, km, and brand were the top contributors to the model's prediction power.

Visualizations:

- Seaborn was used for generating count plots and histograms.
- Box plots were used to visualize outliers, and correlation heatmaps showed relationships between numerical features.

---

4. Model Development

Model Selection and Evaluation:

The following models were evaluated to predict the car prices:

1. **Linear Regression**:
    - MSE: 1.0900223133557542e+20
    - Poor performance due to non-linearity in the data.
2. **Decision Tree Regressor**:
    - MSE: 0.0001369263608648713
    - Slight overfitting noticed.
3. **Random Forest Regressor** (Best Model):
    - MSE: 9.093812459470057e-05
    - **R^2 Score**: 0.996431460977345 on the test set.
    - **Mean Absolute Error**: 0.003980087969238656 on the test set.
    - **Cross-Validation** (K-fold) yielded an average score of 0.9983.
4. **Gradient Boosting Regressor**:
    - MSE: 9.278345359790541e-05

- Similar performance to Random Forest but slightly worse.

Regularization:

- **Lasso Regression** (L1 Regularization): MSE = 0.041908693050106194.
- **Ridge Regression** (L2 Regularization): MSE = 0.004987053227973024.

Hyperparameter Tuning:

Random Forest hyperparameters were optimized using GridSearchCV. The best parameters obtained were:

- bootstrap: True
- max_depth: 30
- min_samples_leaf: 1
- min_samples_split: 2
- n_estimators: 300

After tuning:

- **R^2 score on the test set**: 0.9964201870711981
- **Mean Absolute Error**: 0.003920658680211264

---

5. Model Deployment

Streamlit Application:

A Streamlit application has been developed to allow users to input the car's features and get an estimated price instantly. The app is easy to use and provides predictions based on the Random Forest model.

User Guide:

- **Input Fields**: Users can enter details such as car brand, model year, mileage, and other relevant features.
- **Output**: The predicted car price is displayed based on the user input.

---

6. Justification for Approach

Model Selection:

- **Random Forest** was chosen as the final model due to its ability to handle non-linear relationships and provide robust predictions.
- Linear models like **Linear Regression** and **Ridge Regression** did not perform well due to the complexity of the data.

Feature Engineering:

- Interaction features between important variables (like modelyear and km) were considered to improve the model's ability to capture deeper relationships between features.

Hyperparameter Tuning:

Hyperparameter tuning was carried out to improve the model's performance, reducing overfitting and ensuring the best possible predictive power.

---

7. Conclusion

The project successfully demonstrates how machine learning can be applied to predict used car prices. The final model, deployed via a user-friendly Streamlit application, provides accurate predictions based on car characteristics. Random Forest was selected as the best-performing model, with thorough tuning and validation ensuring a high level of accuracy.
