Yerzen Arman AAI-2502M

Car Price Prediction Project Report

Link:
https://github.com/zenu9/car-price-prediction.git

1. Introduction

The primary objective of this project is to predict the selling price of cars using historical data and machine learning techniques. Accurate car price prediction can assist buyers, sellers, and dealerships in making informed decisions and optimize the buying and selling process.

This project leverages a combination of traditional regression models, ensemble methods, and deep learning techniques to model the relationship between car features and their selling price.

2. Data Description

The dataset used contains the following features:

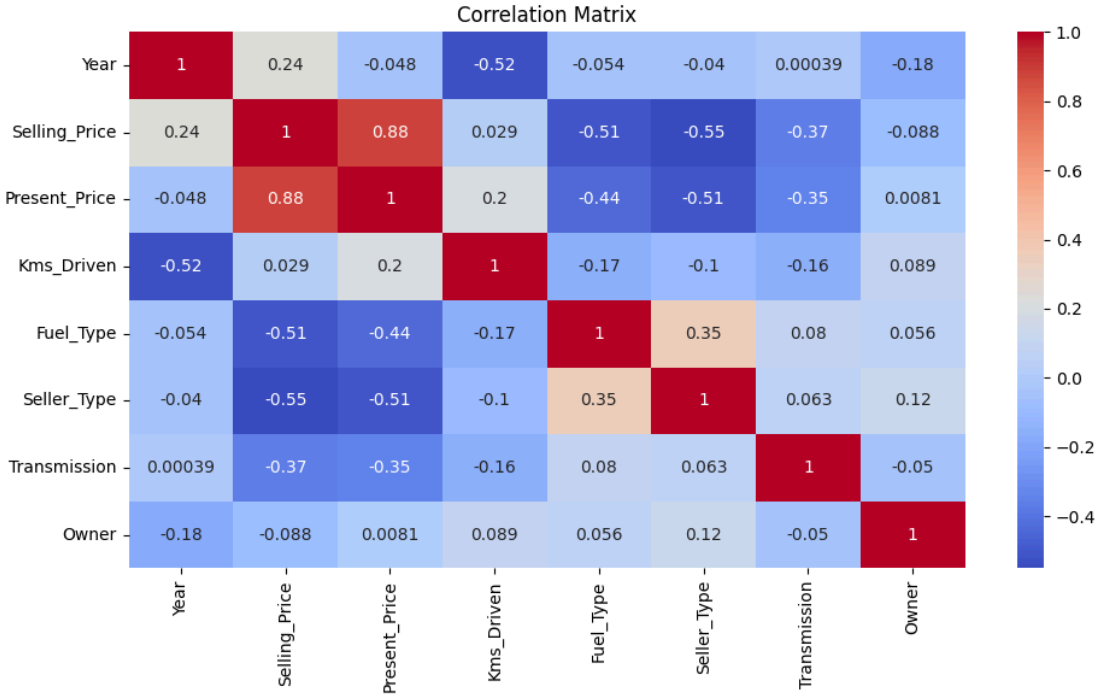
Feature	Description
Year	Year of manufacture
Present_Price	Current price of the car (in lakhs)
Kms_Driven	Kilometers driven by the car
Fuel_Type	Type of fuel (Petrol, Diesel, CNG)
Seller_Type	Type of seller (Dealer, Individual)
Transmission	Type of transmission (Manual, Automatic)
Owner	Number of previous owners
Selling_Price	Target variable (selling price in lakhs)

Categorical features (Fuel_Type, Seller_Type, Transmission) were encoded using label encoding. All features were standardized using StandardScaler to improve model performance, particularly for models sensitive to feature scales such as Support Vector Regressor (SVR) and Multi-Layer Perceptron (MLP).

3. Exploratory Data Analysis (EDA)

3.1 Correlation Analysis

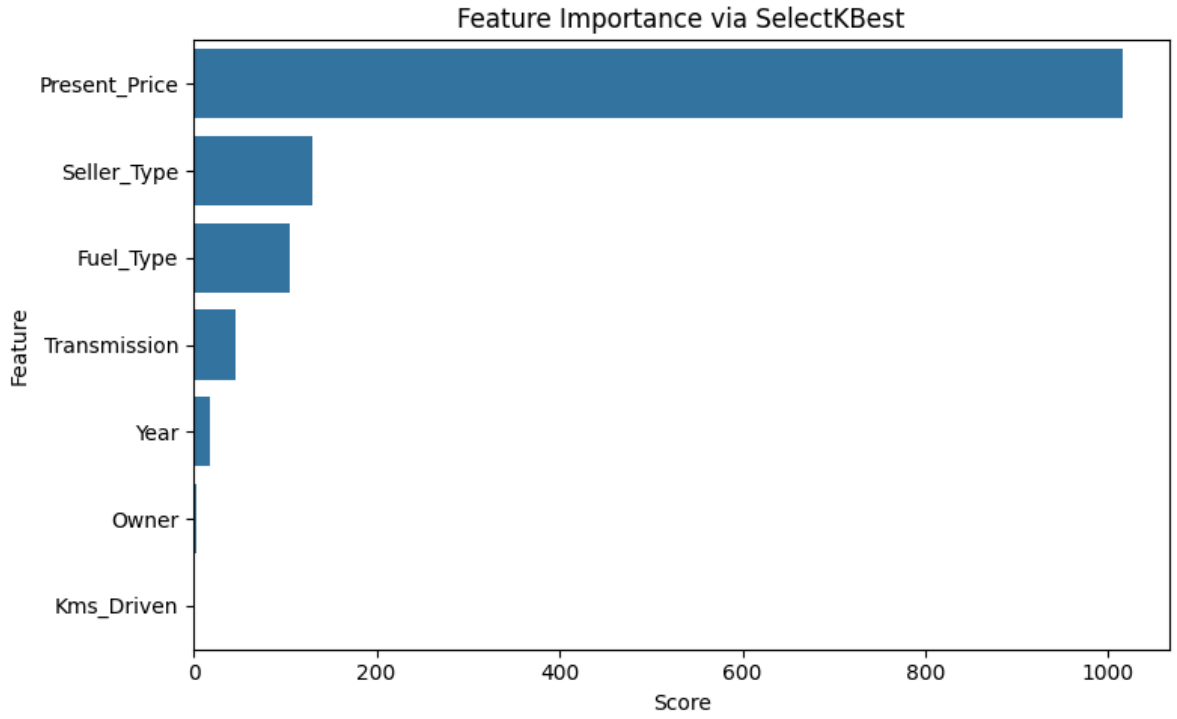
A correlation matrix was generated to understand relationships between features and the target variable:
Present_Price showed the strongest positive correlation with Selling_Price.



(Correlation heatmap saved in plots/correlation/correlation_matrix.png)

3.2 Feature Selection

Two feature selection methods were applied:
SelectKBest (F-regression)



(Bar chart saved in plots/feature_selection/feature_selection_selectkbest.png)
Recursive Feature Elimination (RFE) with Linear Regression

```
=== Features Selected by RFE ===
         Feature  Selected
0           Year      True
1  Present_Price      True
2     Kms_Driven     False
3      Fuel_Type      True
4    Seller_Type      True
5   Transmission      True
6          Owner     False
```

4. Modeling Approaches

Model	**Description**
Dummy Regressor	Baseline model predicting the mean price
Linear Regression	Traditional regression model
Random Forest Regressor	Ensemble model leveraging decision trees
XGBoost Regressor	Gradient boosting tree-based model
Support Vector Regressor (SVR)	Kernel-based regression for non-linear relationships
MLP (PyTorch)	Deep learning neural network with batch normalization and ReLU activations

Hyperparameter tuning was performed using GridSearchCV for Random Forest, XGBoost, and SVR. The MLP model was trained for 200 epochs with the Adam optimizer and MSE loss.

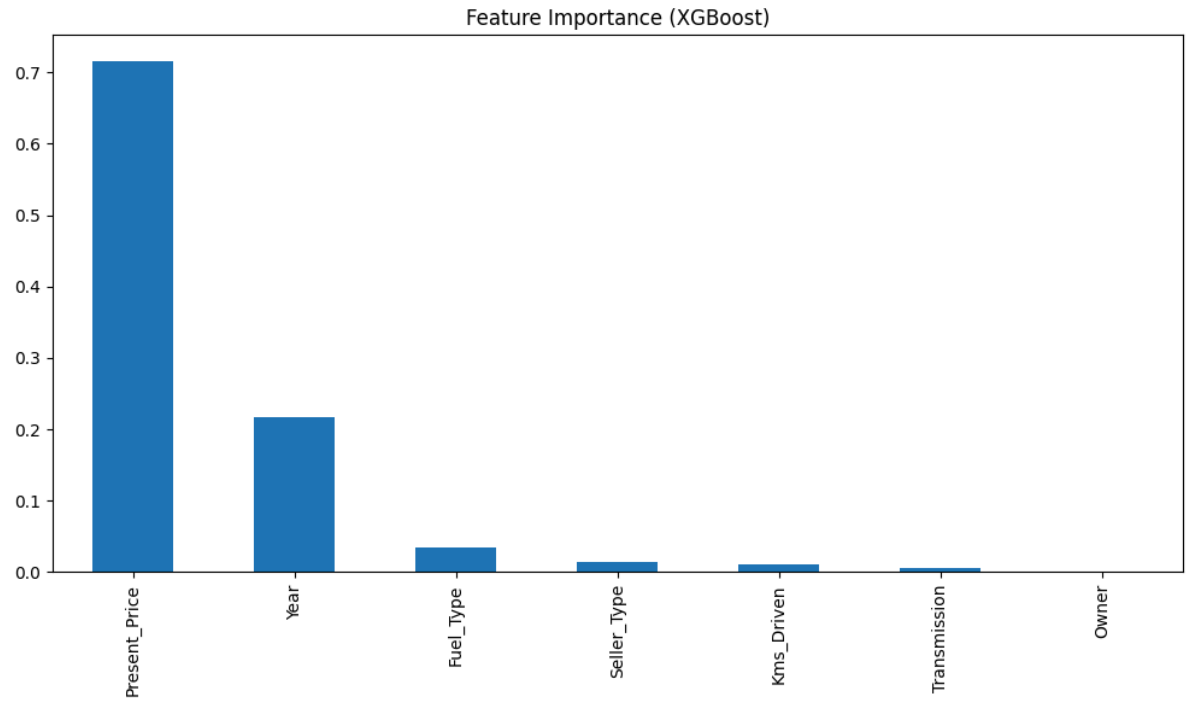
5. Model Evaluation

Models were evaluated on the test set using:
RMSE (Root Mean Squared Error)
MAE (Mean Absolute Error)
R^2 Score

```
=== Model Performance ===
                    RMSE       MAE        R2
Dummy           4.800456  3.385837 -0.000382
LinearRegression 1.878544  1.221762  0.846805
RandomForest    0.942832  0.622749  0.961410
XGBoost         1.155550  0.662600  0.942033
SVR             0.779138  0.465496  0.973647
MLP_Torch       0.918819  0.616547  0.963351
```

6. XGBoost feature importance

XGBoost feature importance identified Present_Price and Year as most critical.



(Chart saved in plots/feature_importance/feature_importance_xgb.png)

7. Error Analysis

Error distributions and actual vs predicted plots were generated for all models.
(Plots saved under plots/errors/ and plots/actual_vs_pred/)