Car Price Prediction Project

Abstract:

Determining whether the listed price of a used car is a challenging task, due to the many factors that drive a used vehicle's price on the market. The focus of this project is developing machine learning models that can accurately predict the price of a used car based on its features, in order to make informed purchases. We implement and evaluate various learning methods on a dataset consisting of the sale prices of different makes and models. I opted for linear regression model due to the advantage that it takes lower training time compared to other methods and also produces best results.

Dataset:

I obtained the dataset from Kaggle. You can access the dataset from;
"https://www.kaggle.com/datasets/erolmasimov/price-prediction-multiple-linear-regression"

Pre-processing:

I decided to use two features, which is engine size and horsepower which I believed in most it contributes a lot in determining the price of a car. Further, I plotted scatter plots to see if there is any relationship between the two aforementioned features and price.

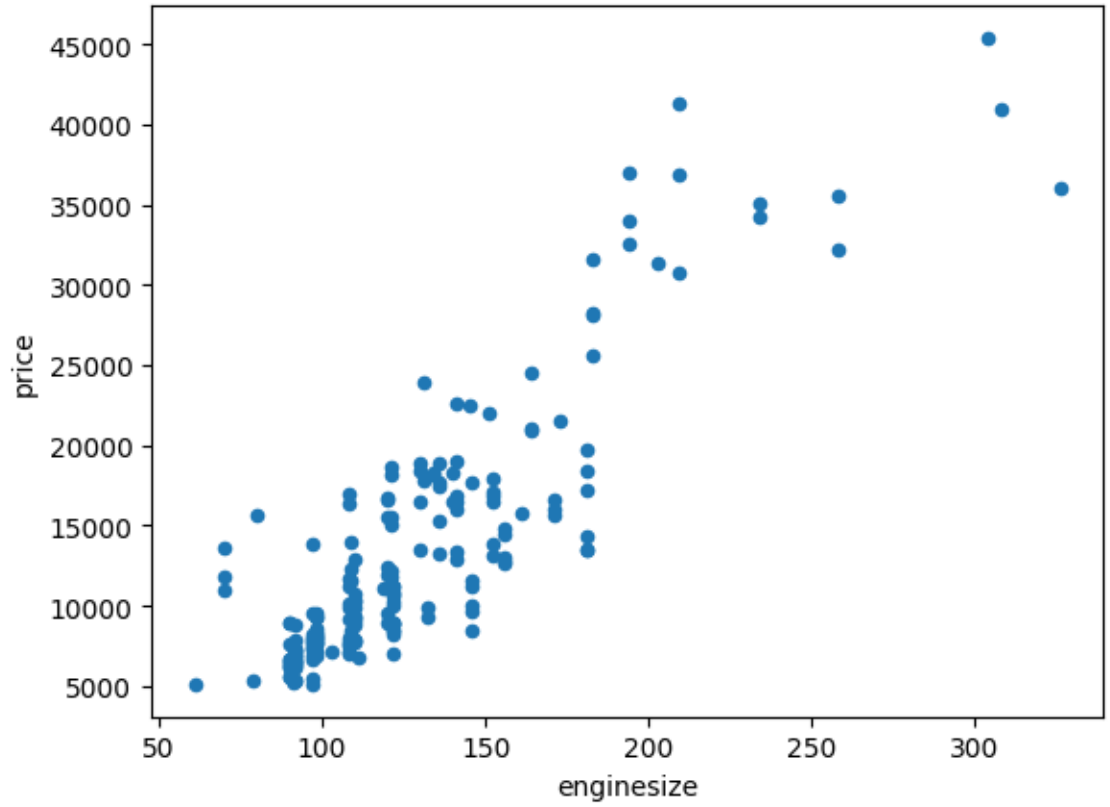

Fig1: scatter plot of engine size and price

This shows that there is a linear relationship between engine size and price.

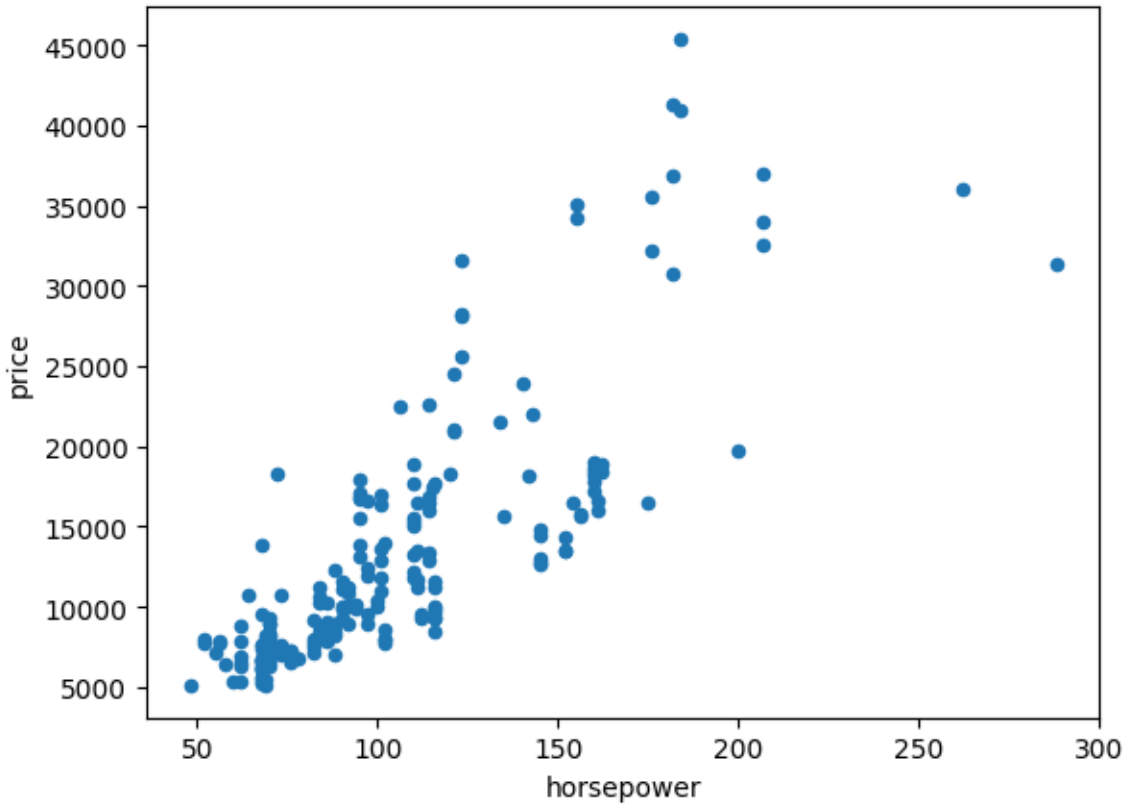

Fig2: scatter plot of horse power and price

This also indicates shows the linear relationship between horse power and price.

Comparison with other features:

Comparing with other features, I decided to check if fuel types can determine the price of a car. Below is the scatter plot of fuel types vs price.

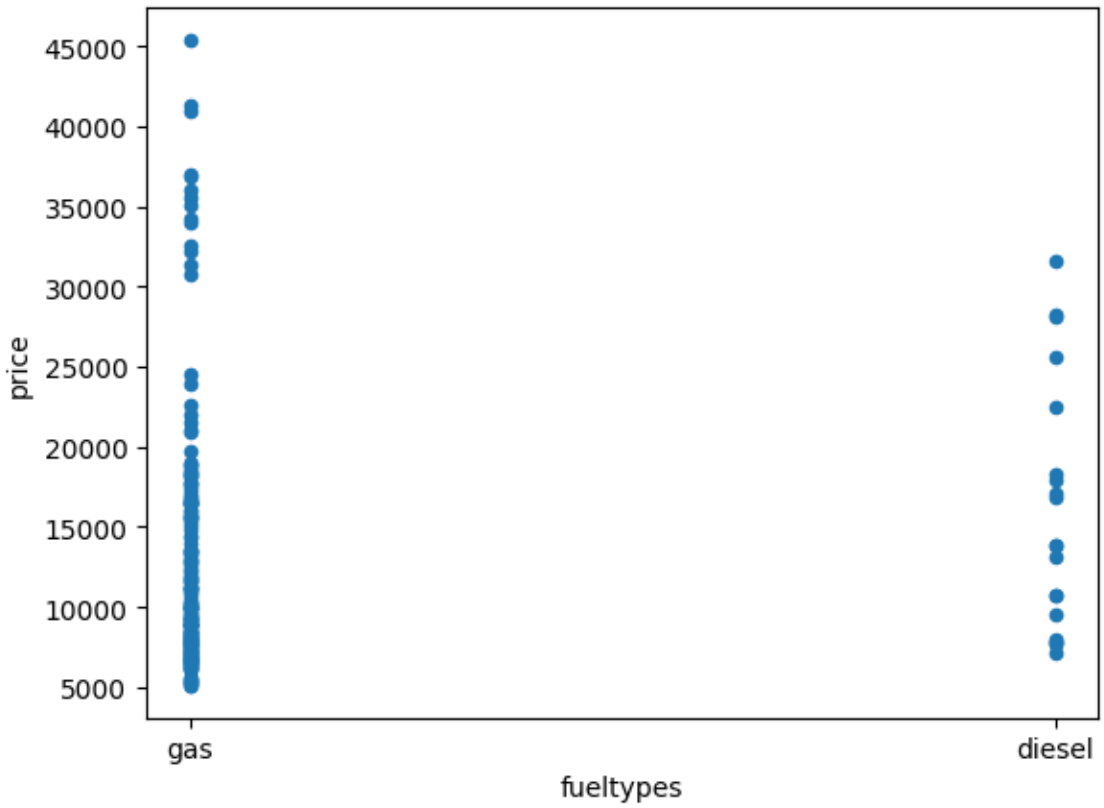

Fig3: scatter plot of fuel types and price

From the above, it is clear that fuel types don't necessarily determine the price of a car.

Methodology:

Linear Regression

Linear regression model was chosen due to its simplicity and it takes lower training time.

Results:

I utilized the following metrics in order to achieve better results;

1. **Mean Squared Error(MSE)**
   MSE measures the amount of error in statistical models. It assesses the average squared distance between the actual and predicted values.
2. **Root Mean Squared Error(RMSE)**
   This is the square root of MSE.

3. Mean Absolute Error(MAE)

This measures how close the predictions are to the actual model.

4. R^2

R^2 score was used to measure the percentage of accuracy of the model from a scale of 0-100. The higher the percentage the accurate the model

Below is the scatter plot between the actual and predicted values

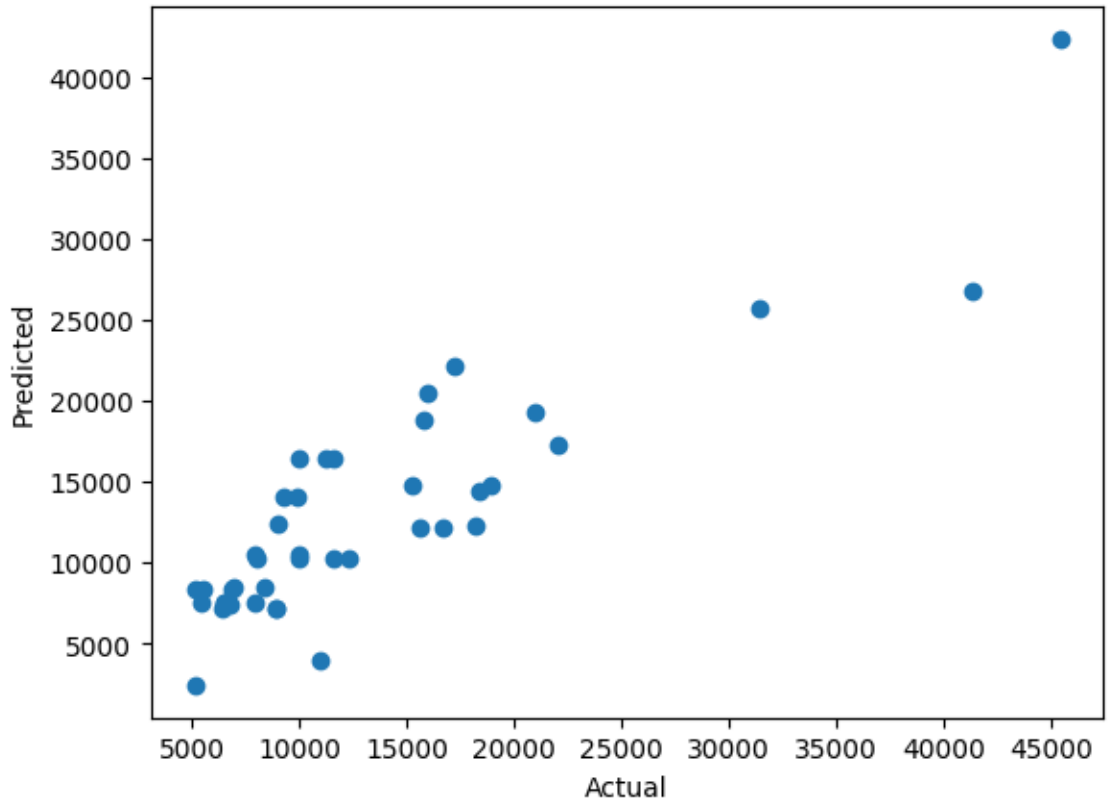

Fig4: scatter plot of actual and predicted values

From the scatter plot above, its clear that there is no much difference between the actual and predicted values.

Linear regression lines for both models:

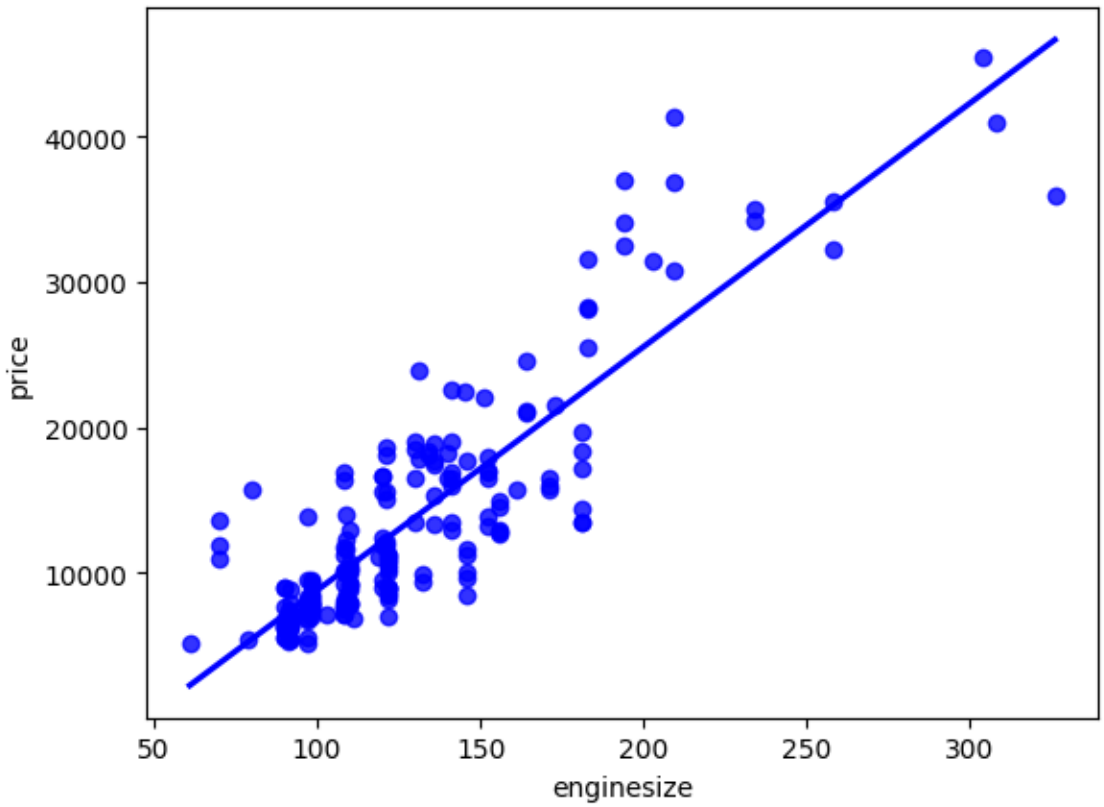

Fig5: model regression line of engine size and price

It indicates the positive linear relationship

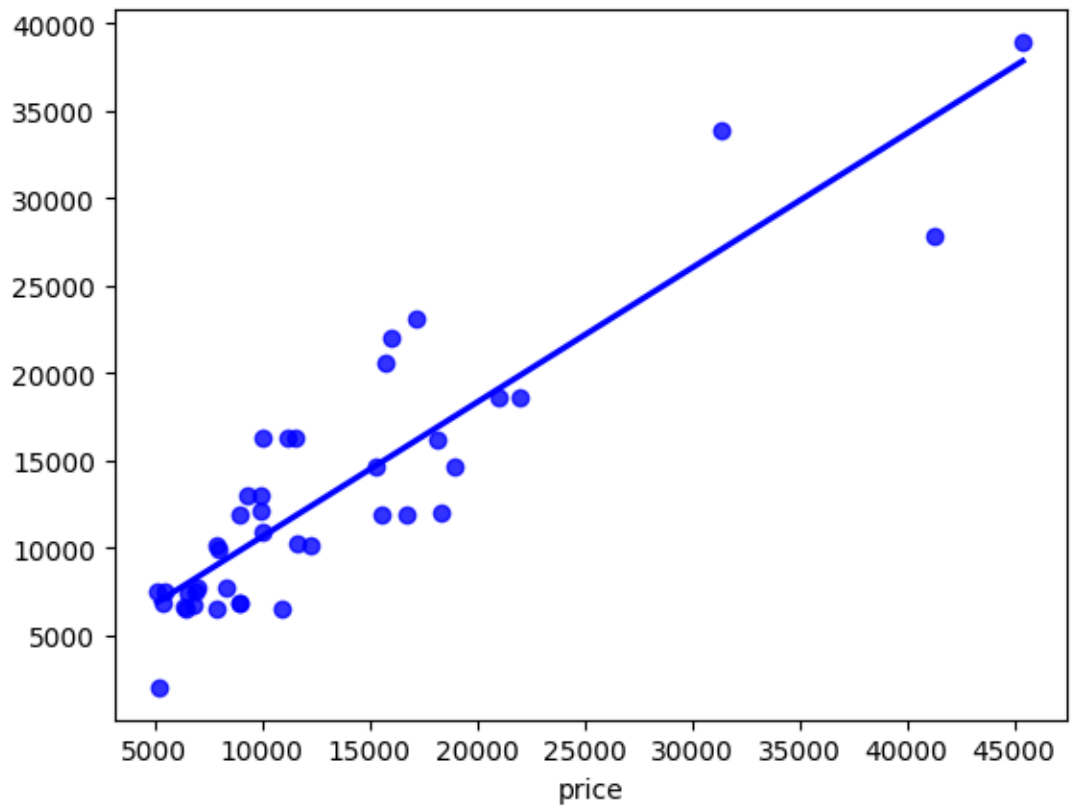

Fig6: model regression line of actual price and predicted values

Generally, the R^2 score was high on both selected features;

Feature	**R^2 Score**
Engine size	78%
Horse power	79%

Conclusion:

In conclusion, the price of a car is highly dependent on the size of the engine and the horse power.