Master's in Applied Statistics and Data Science (PM-ASDS)

Department of Statistics Jahangirnagar University

Savar, Dhaka-1342, Bangladesh.

An Assignment on

An Assignment on Predicting Car Prices Using the Automobile Dataset with Linear Regression

PM-ASDS04 Introduction to Data Science with Python

Submitted to

Professor Farhana Afrin Duty

Submitted by

Tousif Md. Amin Faisal PMASDS 9th Batch, Section-A Roll-20229031

Objective

In order to anticipate automobile prices, this study will evaluate the Car dataset and create a linear regression model. Exploratory data analysis (EDA) will be used to uncover patterns in the data. The data will then be preprocessed by handling missing values and transforming categorical features into numerical features. A linear regression model will then be built to predict car prices, and its performance will be assessed using metrics like mean squared error (MSE), mean absolute error (MAE), and R-squared.

Data

The Car dataset includes a number of characteristics related to cars, including make, fuel type, body style, engine size, horsepower, and others. The goal variable is the car's price, and the dataset has 205 samples. Also, the dataset has missing values that must be dealt with during data preprocessing.

Methodology

We used a four-step process for this study:

Step 1: Data Preprocessing

The preprocessing of the data was the initial step in our methodology. The dataset needed to be handled because it had missing values. To fill in the gaps, we used the feature's mean or mode value. Using label encoding, we also transformed the categorical characteristics into numerical features. Finally, we scaled the feature using StandardScaler and divided the data into training and testing groups.

Step 2: Exploratory Data Analysis (EDA)

EDA was used as the second phase in our technique to acquire insights into the data. To see the correlation between the features, we drew the correlation matrix and used histograms to visualize the distribution of each attribute.

Step 3: Building the Machine Learning Model

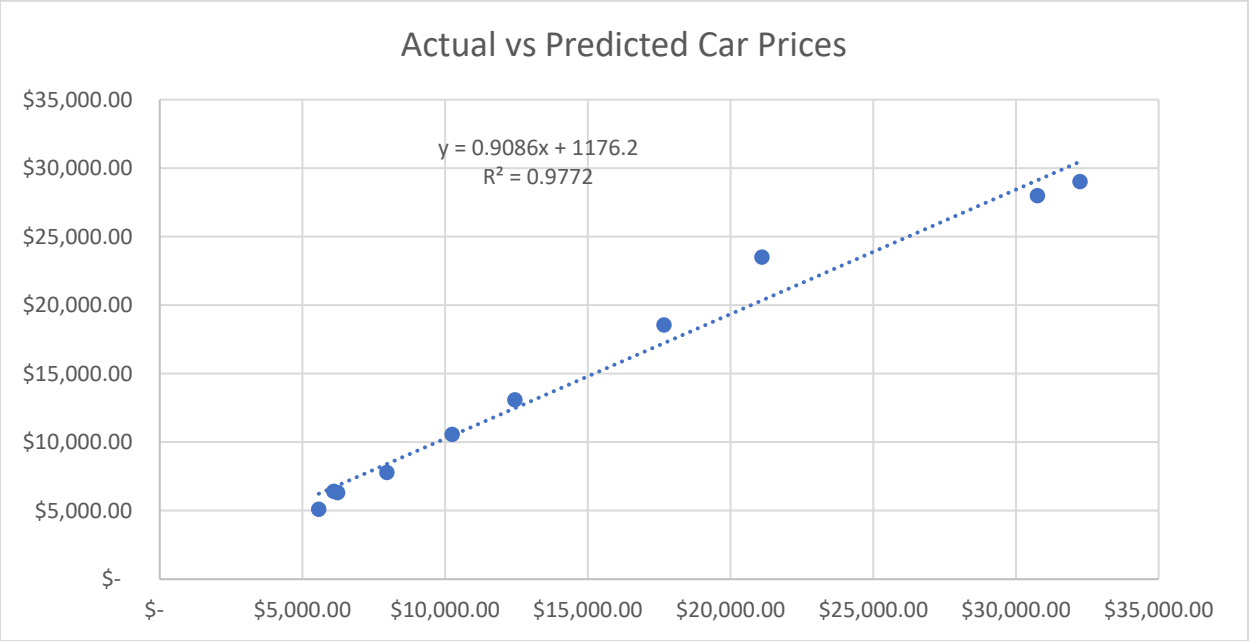
Building a machine learning model was the third step in our technique. Because linear regression is an easy-to-understand model that can handle both category and numerical features, we selected it as our model.

Step 4: Evaluating the Model

The model's performance was assessed using a variety of assessment measures, including mean squared error (MSE), mean absolute error (MAE), and R-squared, as the last stage in our technique.

Results

To evaluate the performance of the model, we plotted the actual versus predicted car prices on the testing set:



We also outputted the first 10 predictions from the final model:

Actual Price	Predicted Price
$ 7,957.00	$ 7,791.16
$ 21,105.00	$ 23,518.74
$ 6,095.00	$ 6,413.04
$ 6,229.00	$ 6,322.08
$ 12,440.00	$ 13,090.48
$ 5,572.00	$ 5,098.52
$ 30,760.00	$ 27,983.85
$ 10,245.00	$ 10,560.90
$ 17,669.00	$ 18,553.22
$ 32,250.00	$ 29,018.04

We also evaluated the performance of the model using the mean squared error (MSE), mean absolute error (MAE), and R-squared metrics. The model achieved the following evaluation metrics on the testing set:

- Mean Squared Error (MSE): 17808264.56477749
- Mean Absolute Error (MAE): 2798.5208730143445
- R-squared: 0.8544443460495549

These results indicate that the model has a good fit to the testing data, with an R-squared score of 0.85 and low values of MSE and MAE.

Conclusion

In conclusion, using engine size, horsepower, and other variables to estimate automobile pricing using linear regression can be a valuable method. The Automobile dataset offered an excellent chance to put EDA, data preparation, and linear regression techniques into effect in a practical setting.