

Predict Car Prices

for Scraping and Regression

Abstract:

Most people have the problem of not knowing the price of a car when selling it. So, we will use data from cars.com to help them determine the expected price of their car based on their car specifications and build a linear regression model to predict car prices.

Using data scraped from car.com, apply EDA process, feature engineering, training, and validation to choose the best model after that we build models that address a useful prediction, after refining a model, we built an interactive dashboard to visualize and communicate the results using the seaborn library.

Design:

The goal of this project is to predict car prices by using Cars.com dataset through Regression Algorithms.

Data:

Using cars.com dataset. The focus is on Audi cars. Using Python library Beautiful Soup to scrape the cars information.

The dataset contains 797 rows with 8 features for each, 2 of which are categorical.

A few feature highlights include the price of the car, the rate, number of reviews, miles driven. Also, there are some features we did not include which are the car name, the dealer, and the model.

Algorithms:

- We use beautiful soap to web scaping.
- Feature Engineering Preprocessing:
Check if there are any missing values - check if there are any duplicate values - Create Dummy Variables - Convert the Datatype of the MODEL.

- Visualizations:

Histogram plot: to show the Count of Cars by Price.

Pair plot: to show the relationship for (X, Y) combination of features in a Data Frame.

Heatmap: to show the correlation between the features.

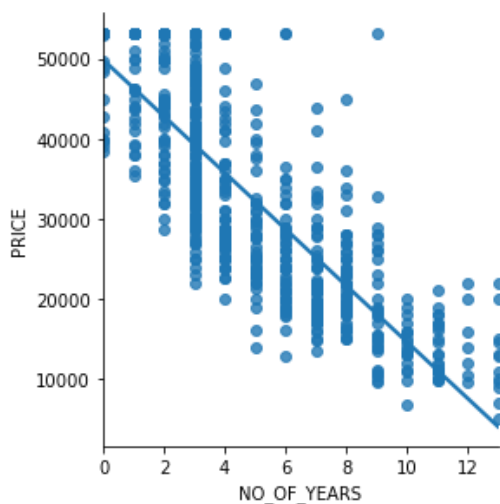
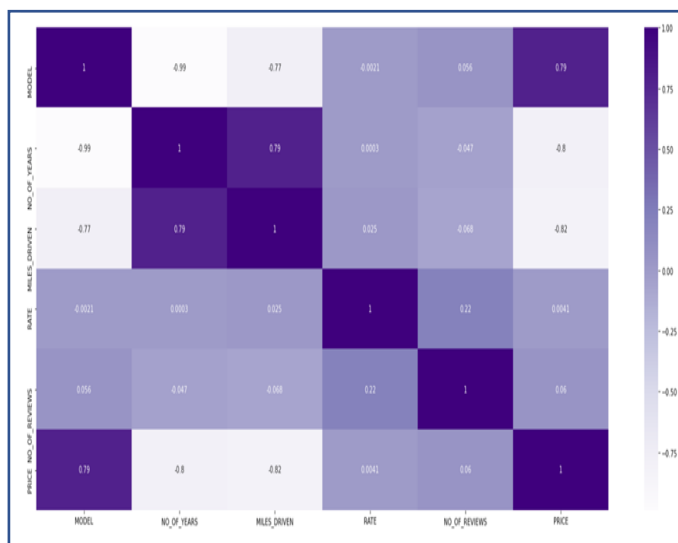
- Model Evaluation and Selection:

Build model The entire training dataset was split into 80/20 train vs. holdout, and all scores reported below were calculated with 5-fold cross-validation on the training portion only. Predictions on the 20% holdout.

Tools:

The tool that we need is Pandas packages to manipulate data, BeautifulSoup to scrape data, LinearRegression, Redge and Lasso from the sklearn.linear_model class of the sklearn module to perform the linear regression then predicted the car prices, Scikit Learn Framework and visualization library (such as Seaborn and Matplotlib), Metrics library to see the model performance on data, and Jupyter notebook to execute the code.

Communication:



By Ghadah Alharbi and Rahaf Alyousef