Car Price Prediction

By: Juan Herrera

Abstract

The goal of this project was to create a regression model that accurately predicts the price of a car in order to help car dealerships around the Alabama area. I scraped the data from cars.com to end up with a dataset with 10 different features and 1183 rows of data. I had to clean the data, visualize the data to understand what features were going to be the most helpful, and create new features to improve my model's performance.

Design

I obtained this data through the web scraper that I built. Trying to build regression models to predict the price of a car is a problem that many people have tried to tackle but in this case I wanted to see what car's features were the most important when it came to building an accurate model.

Data

The dataset contains 1183 different cars with 10 features for each, 8 of which are categorical. A few feature highlights include Mileage, MPG, fuel type, year of the car. Many of these features could be bucketed into more general categories. Other features were created through feature engineering to boost the model's performance.

Algorithms

Feature engineering

1. Creating a "years_old" column by subtracting the year of the car from 2022 which helped reduce multicollinearity
2. Converting categorical features to binary dummy variables
3. Removing duplicate values
4. Bucketing a column of data into more general categories
5. Creating a "luxury" column that included the cars that were the most expensive

Models

Validation, Ridge, Random Forest were used before using Lasso as my final model. Lasso was not probably the model that best performed but I liked how the coefficients of the irrelevant features were brought down to zero.

The official metrics to evaluate the performance of the model were R^2 and RMSE.

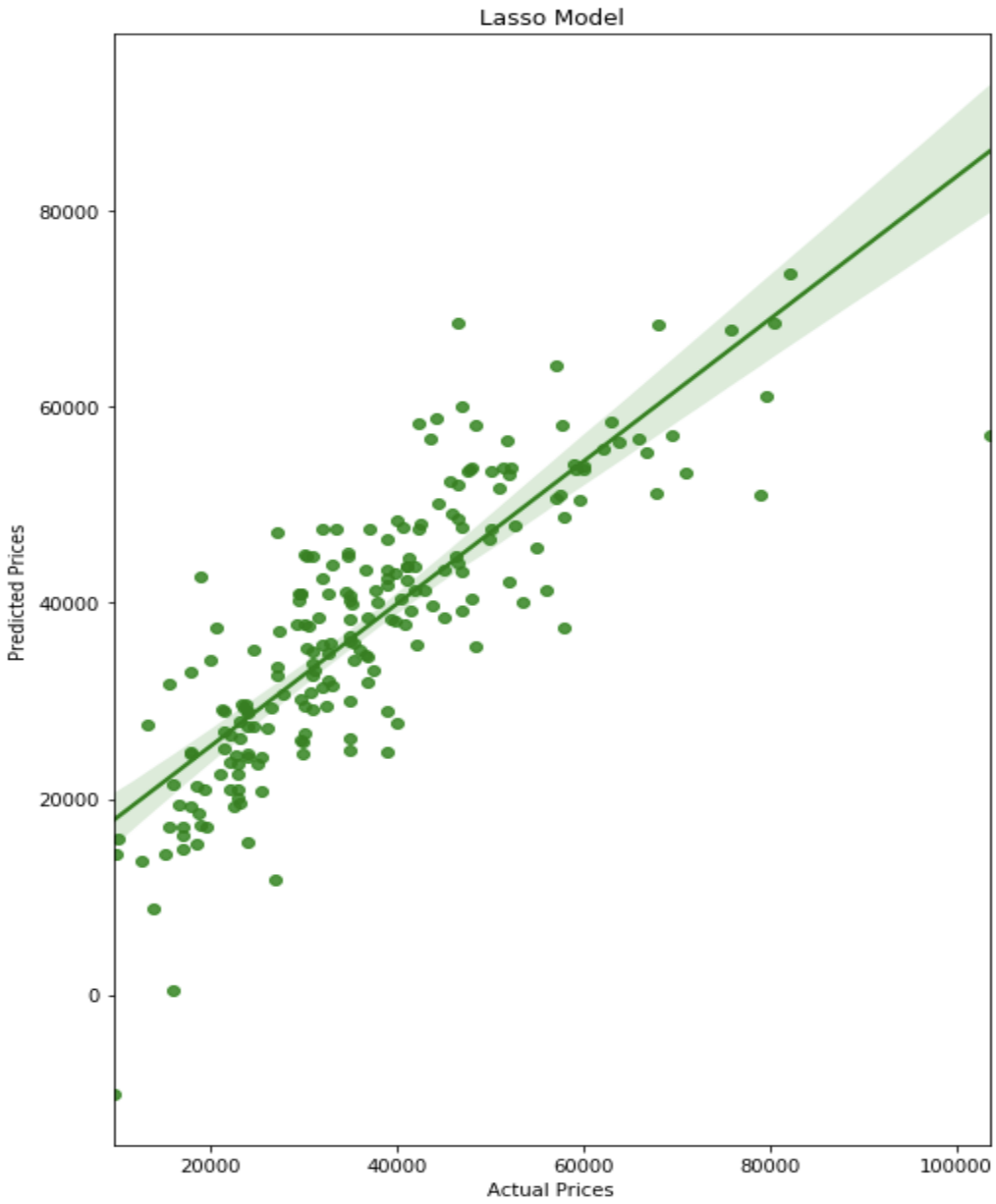
When training my model there were 110 features, including all the dummy variables, in my data set:

- R^2 = 69.05
- RMSE = $93.10

Tools

- Pandas for data manipulation
- Seaborn and Matplotlib for data visualization
- Scikit-learn and statsmodels for modeling

Communication


Lasso Model
Predicted Prices
Actual Prices
0
20000
40000
60000
80000
100000