Linear Regression of Used Toyota Cars in UK Project

Authors: Ahmad, Faisal, Yahya

Abstract:

The goal of this project is to determine which factor has a strong effect on car price. In addition to that, applying a regression model on car cost to estimate the price based on mileage and engine size. We worked as a team to collect the data from Kaggle website using beautiful soup library to parse and extract all related features of used cars. We preprocess the data, include handling the missing values, removing duplicate observations, and encoding categorical features. We achieved promising results by implementing Lasso regression model to this dataset.

Data Description:

This data set was collected from kaggle.com website which contains information of price, transmission, mileage, fuel type, road tax, miles per gallon (mpg), and engine size of used Toyota cars in United Kingdom.

Algorithms:

- Feature engineering:
  - Converting categorical variables to binary dummy variables
  - Handling outliers using IQR method

- Models
  - Linear regression
  - Ridge regression
  - Lasso regression
  - Polynomial regression

- Model Evaluation and Selection

The entire dataset of (~5000) observations was split into 80/20 for training and testing. Through the use of cross validation function from sklearn library with 5 folds, we achieved an average of R^2 score of 0.9454. We apply this evaluation for all four models and based on the R^2 we choose Lasso regression model as the best fit for this dataset.

Tools:

- For data manipulation: Numpy and Pandas
- For data visualization: Matplotlib and Seaborn
- For machine learning algorithm: Scikit-learn