Project Report

Predicting car prices

Predict the costs of used cars given the data collected from various sources and distributed across various locations in India.

List of features and their description

* **Name:**The brand and model of the car.
* **Location:** The location in which the car is being sold or is available for purchase.
* **Year:**The year or edition of the model.
* **Kilometres Driven:**The total kilometres driven in the car by the previous owner(s) in KM.
* **Fuel Type:**The type of fuel used by the car.
* **Transmission:**The type of transmission used by the car.
* **Owner Type:** Whether the ownership is First-hand, Second hand or other.
* **Mileage:**The standard mileage offered by the car company in kmpl or km/kg
* **Engine:**The displacement volume of the engine in cc.
* **Power:**The maximum power of the engine in bhp.
* **Seats:**The number of seats in the car.
* **New Price:** The price of a new car of the same model.
* **Price:**The price of the used car in INR Lakhs.

Python Modules Used

* Pandas
* Numpy
* Matplotlib
* Seaborn
* Scikit
* Turicreate

Approach

Following is the approach for the project

1. Cleaning and organizing data.
2. Exploratory Data Analysis
3. Developing a Model

Implementing the Model

One hot encoding is done for categorical features. The categories in each feature are replaced by a numeric value. This one hot encoding is done for both testing and training set.

A correlational matrix was plotted to analyse the multicollinearity between different features and price of the car and features were modified accordingly.

Features having near zero correlation are excluded from the dataset as they won’t affect the price of car as such and would help us save time in training the dataset.

The model is then implemented using Lasso (or L-1) regularisation

Through the process mentioned above, 88.59 was obtained as the best penalty value according to the RMSE score. So, the final model is trained using this value and car prices are predicted.

Accuracy of model is obtained via R2 score.

For the validation set, R2 score is 84.53.

For entire training set, R2 score is 93.72.