Linear Regression Project Report

Data Science T5 Boot camp

Name: Faisal Mirza

Gmail: faisal.a.mirza11@gmail.com

GitHub: faisalmirza11

Abstract:

In this linear regression project, my goal was to predict the price of a used SUV/ Crossover posted on Cargurus.com, or any car selling website. I got the data from Kaggle.com and I prepared the data for modeling by choosing only the features related to the cars and then I started to look for the best model, some of the models that I tried are: linear regression baseline model, lasso, polynomial, cross validation.

Framing question: Does a new post of a used SUV for sale have a fair price or not.?

Who benefits: Cargurus.com, or a website for selling used cars, and also people who are in the market to buy used cars.

Design:

Cargurus.com is a website that sells all kinds of new and used cars, it also allows owners of cars to post and sell their cars through the website, and due to the increased sales and demand of SUV's and crossovers over sedans and other types of vehicles in the recent years, Cargurus wants to know if a used SUV posted for sale on their website has a fair price or not. So by using the current used SUV's prices on the website (assuming all have a fair price) we will predict if a new post has a fair price or not, which will allow Cargurus to clean their search results for users of the website

Data:

The dataset has been obtained from Kaggle, and it contains 3 million rows of new and used cars each with 66 columns, I only used cars of type SUV/ Crossover and cars that were used. And some of these columns include:

- The car's brand name, model, year of production.
- The mileage of the car.
- The price of the car (target variable).

Scope:

- data of roughly 550,000 SUV/ Crossovers was used (rows)
- Each row will have about 50 columns with dummy variables of car make

Algorithms:

I started by viewing the dataset to understand it clearly, then after looking at the data and understanding it, I started to look for issues and problems in the data that required cleaning, and some of the cleaning that I did includes:

- Removing null values
- Removing new cars and keeping only used cars
- Keeping only the SUV/ Crossover body type cars

I made a subset of only the numerical columns and I found out that many of the columns were not related to the cars features, so I dropped them.

Feature engineering:

- Adding dummy variables for the car make

I split the data to 60% training, 20% validation, 20% testing, then after cleaning the dataset and performing the feature engineering I started to experiment using linear regression models to find the best score, and some of the models I used include:

- Baseline model without dummy variables
- Baseline model with dummy variables
- Polynomial without dummy variables
- Polynomial with dummy variables
- Cross validation with dummy variables
- Lasso on polynomial with dummy variables
  In the end the best model was polynomial with dummy variables

Tools:

Technologies: Python, Jupiter notebook.

Libraries: Sklearn, Statsmodels, Numpy, Pandas, Matplotlib, Seaborn.

Communication:

Everything related to this project will be available on my GitHub at faisalmirza11 in the repository DS-T5-projects/Linear Regrassion - Scraping project.