Linear Regression Assignment – Post Graduate Diploma in ML and AI – 2019

A Chinese automobile company Geely Auto aspires to enter the US market by setting up their manufacturing unit there and producing cars locally to give competition to their US and European counterparts.

The company wants to know:

* Which variables are significant in predicting the price of a car
* How well those variables describe the price of a car

Some of the data cleanups done involved making all string columns to lower case, changing car-names to have only the company name and correcting incorrect spellings and changing number-of-cylinders and number-of-doors to numerical variables. Highly correlated variables were removed from the dataset so that multi-collinearity does not affect the model.

Three models viz. without car company in the independent variables, car companies categorized based on price and with car companies were experimented with.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model Info** | **R-squared** | **Adj. R-squared** | **F-statistic** | **Prob (F-statistic)** | **R-2 score of Actual Vs Predicted Test Price** |
| With Car Company | 0.912218 | 0.905568 | 137.172187 | 1.174994e-64 | 0.903537 |
| With Company Groups | 0.912159 | 0.908284 | 235.374936 | 2.990401e-69 | 0.873528 |
| Without Company | 0.839564 | 0.831245 | 100.922365 | 1.780241e-50 | 0.730047 |

Statistical values and R-2 Score of Actual Vs Predicted Test Values for the model with companies are high. Since the actual company is observed to be affecting the price, knowing these companies will help. Hence, Model with Companies, rather than the company groups is recommended.

**Engine Size, Aspiration, Car Height, Peak RPM** have an effect on the price of a car. Additionally, price also is getting affected based on whether the engine type is a **Rotor Engine or not** and whether the car is a **Convertible or not**. Car companies having an effect on the price are **BMW, Porsche, Buick, Mitsubishi**. The business factors like margin, luxury premium, etc which has not been included in the data set could be a reason for this.