# TinkerWeek.py

# Machine Learning

# **Project Bash**

By

Lasitha E

**Email** : lasithaeaswaran@gmail.com | **Github link**(to this project)

**Problem Statement**

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.

They have contracted an automobile consulting company to understand the factors on which the pricing of cars depends. Specifically, they want to understand the factors affecting the pricing of cars in the American market, since those may be very different from the Chinese market. 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

Based on various market surveys, the consulting firm has gathered a large data set of different types of cars across the American market.

**Business Goal**

We are required to model the price of cars with the available independent variables. It will be used by the management to understand how exactly the prices vary with the independent variables. They can accordingly manipulate the design of the cars, the business strategy, etc. to meet certain price levels. Further, the model will be a good way for management to understand the pricing dynamics of a new market.

**Steps involved in experimental evaluation**

1. Reading and Understanding the Data
2. Data Cleaning and Preparation
3. Visualising the data(Numerical & Categorical)
4. Deriving new features
5. Bivariate Analysis
6. Dummy Variables
7. Train-Test Split and feature scaling
8. Model Building
9. Residual Analysis of Model
10. Prediction and Evaluation

# Summary

According to the initial analysis of the given dataset I came to a conclusion that there were 24 independent variables and the dependent variable, price.The data was all-in-all clean,except for some cleaning needed in the CarName.I could gather preliminary insights and get the hang of data by visualising it. I used matplotlib and seaborn for it. I drew inferences by visualising the relations of numerical data('symboling', 'wheelbase', 'carlength', 'car width', 'carheight', 'curb weight', 'enginesize', 'boreratio', 'stroke', 'compression ratio', 'horsepower', 'peak rpm', 'city mpg', 'highway mpg') with price with the help of pairplot from seaborn and visualised the relations of Categorical data('CarName', 'fuel type', aspiration', carbody', 'drive wheel', 'engine location', 'engine type', 'fuel system', 'door number', 'cylinder number') using box plot and distplot from the same.

Some Inferences after visualising the data:

1. The plot seemed to be right-skewed, meaning that the most prices in the dataset are low(Below 15,000).
2. There is a significant difference between the mean and the median of the price distribution.
3. The data points are far spread out from the mean, which indicates a high variance in the car prices.(85% of the prices are below 18,500, whereas the remaining 15% are between 18,500 and 45,400.)
4. Toyota seemed to be favored car company.
5. Toyota seemed to be favored car company.
6. sedan is the top car type prefered.
7. It seems that the symboling with 0 and 1 values have high number of rows (i.e. They are most sold.)
8. The cars with -1 symboling seems to be high priced (as it makes sense too, insurance risk rating -1 is quite good). But it seems that symboling with 3 value has the price range similar to -2 value. There is a dip in price at symboling 1.

9. ohc Engine type seems to be most favored type.
10. ohcv has the highest price range (While dohcv has only one row), ohc and ohcf have the low price range.
11. Jaguar and Buick seem to have highest average price.
12. diesel has higher average price than gas.
13. hardtop and convertible have higher average price.
14. doornumber variable is not affecting the price much. There is no significant difference between the categories in it.
15. It seems aspiration with turbo have higher price range than the std(though it has some high values outside the whiskers.)
16. Very few datapoints for enginelocation categories to make an inference.
17. Most common number of cylinders are four, six and five. Though eight cylinders have the highest price range.
18. mpfi and 2bbl are most common type of fuel systems. mpfi and idi having the highest price range. But there are few data for other categories to derive any meaningful inference
19. A very significant difference in drivewheel category. Most high ranged cars seem to prefer rwd drivewheel.
20. carwidth, carlength and curbweight seems to have a positive correlation with price.
21. carheight doesn't show any significant trend with price.
22. enginesize, boreratio, horsepower, wheelbase - seem to have a significant positive correlation with price.
23. citympg, highwaympg - seem to have a significant negative correlation with price.
24. High ranged cars prefer rwd drivewheel with idi or mpfi fuelsystem.

**List of significant variables after Visual analysis :**

- Car Range
- Engine Type
- Fuel type
- Car Body
- Aspiration
- Cylinder Number
- Drivewheel
- Curbweight
- Car Length
- Car width
- Engine Size
- Boreratio
- Horse Power
- Wheel base
- Fuel Economy

Using the heatmaps to understand the correlations, we can come to a conclusion that , the highly correlated variables to price are - curbweight, enginesize, horsepower, carwidth and highend.

In the model building stage, I tried random forest , decision tree and OLS Regression, but found that OLS Regression is the best fit , as it has the highest r2 score among these three.

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | price | R-squared: | 0.912 |
| Model: | OLS | Adj. R-squared: | 0.909 |
| Method: | Least Squares | F-statistic: | 284.8 |
| Date: | Fri, 15 Jan 2021 | Prob (F-statistic): | 1.57e-70 |
| Time: | 11:29:54 | Log-Likelihood: | 190.93 |
| No. Observations: | 143 | AIC: | -369.9 |
| Df Residuals: | 137 | BIC: | -352.1 |
| Df Model: | 5 | | |
| Covariance Type: | nonrobust | | |

## End analysis

I continued the OLS Regression by dropping the columns with high p-value(>0.05) which are least significant . The dropped columns being : ‘twelve’,’fueleconomy’ ,’curbweight’, ‘sedan’ and ‘wagon’. I further did not drop the ‘hatchback’ column with a p-value of 0.04(which indicates that it is statistically significant ) as the accuracy(r2_score) was getting below 0.9.Hence the most important variables are horsepower, carwidth, hatchback, dochv, highend.

Final result :

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | price | R-squared: | 0.912 |
| Model: | OLS | Adj. R-squared: | 0.909 |
| Method: | Least Squares | F-statistic: | 284.8 |
| Date: | Fri, 15 Jan 2021 | Prob (F-statistic): | 1.57e-70 |
| Time: | 11:29:54 | Log-Likelihood: | 190.93 |
| No. Observations: | 143 | AIC: | -369.9 |
| Df Residuals: | 137 | BIC: | -352.1 |
| Df Model: | 5 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | -0.0970 | 0.018 | -5.530 | 0.000 | -0.132 | -0.062 |
| horsepower | 0.5013 | 0.051 | 9.832 | 0.000 | 0.401 | 0.602 |
| carwidth | 0.3952 | 0.043 | 9.252 | 0.000 | 0.311 | 0.480 |
| hatchback | -0.0336 | 0.012 | -2.764 | 0.006 | -0.058 | -0.010 |
| dohcv | -0.3231 | 0.072 | -4.502 | 0.000 | -0.465 | -0.181 |
| Highend | 0.2833 | 0.021 | 13.615 | 0.000 | 0.242 | 0.324 |

| | | | |
|---|---|---|---|
| Omnibus: | 36.097 | Durbin-Watson: | 2.028 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 78.717 |
| Skew: | 1.067 | Prob(JB): | 8.07e-18 |
| Kurtosis: | 5.943 | Cond. No. | 16.3 |