Car Price Prediction (ML (SLR + SLC + USL))

Presented By

Anoosh Kumar Nidhi Jaiswal Vignesh Bhatt Sai Charan Sylesh JL

BUSINESS PRBOLEM:

Geely Auto, a Chinese automobile company, has plans to expand into the American market by establishing a local manufacturing unit and producing cars that will rival their American and European counterparts. In preparation for this move, they have enlisted the services of an automobile consulting company to identify the key factors that determine car prices in the American market. This is crucial as there may be significant differences in pricing factors between the American and Chinese markets.

OBJECTIVE

Our aim is to develop a model that can accurately predict car prices based on the available independent variables. This model will provide valuable insights to management, allowing them to understand how prices vary based on these independent variables. By having a deep understanding of these pricing dynamics, management can make informed decisions about the design of the cars, business strategies, and other factors to achieve their desired price levels. Additionally, this model will be a useful to

EXPLORATORY DATA ANALYSIS

Descriptive Statistics

	count	mean	std	min	25%	50%	75%	max
car_ID	205.0	103.000000	59.322565	1.00	52.00	103.00	154.00	205.00
symboling	205.0	0.834146	1.245307	-2.00	0.00	1.00	2.00	3.00
wheelbase	205.0	98.756585	6.021776	86.60	94.50	97.00	102.40	120.90
carlength	205.0	174.049268	12.337289	141.10	166.30	173.20	183.10	208.10
carwidth	205.0	65.907805	2.145204	60.30	64.10	65.50	66.90	72.30
carheight	205.0	53.724878	2.443522	47.80	52.00	54.10	55.50	59.80
curbweight	205.0	2555.565854	520.680204	1488.00	2145.00	2414.00	2935.00	4066.00
enginesize	205.0	126.907317	41.642693	61.00	97.00	120.00	141.00	326.00
boreratio	205.0	3.329756	0.270844	2.54	3.15	3.31	3.58	3.94
stroke	205.0	3.255415	0.313597	2.07	3.11	3.29	3.41	4.17
compressionratio	205.0	10.142537	3.972040	7.00	8.60	9.00	9.40	23.00
horsepower	205.0	104.117073	39.544167	48.00	70.00	95.00	116.00	288.00
peakrpm	205.0	5125.121951	476.985643	4150.00	4800.00	5200.00	5500.00	6600.00
citympg	205.0	25.219512	6.542142	13.00	19.00	24.00	30.00	49.00
highwaympg	205.0	30.751220	6.886443	16.00	25.00	30.00	34.00	54.00
price	205.0	13276.710571	7988.852332	5118.00	7788.00	10295.00	16503.00	45400.00

Data Preprocessing & Cleaning

Company Name Spelling Correction:

Before:

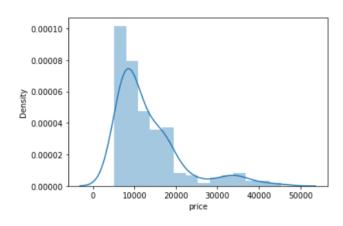
```
array(['alfa-romero', 'audi', 'bmw', 'chevrolet', 'dodge', 'honda', 'isuzu', 'jaguar', 'maxda', 'mazda', 'buick', 'mercury', 'mitsubishi', 'Nissan', 'nissan', 'peugeot', 'plymouth', 'porsche', 'porcshce', 'renault', 'saab', 'subaru', 'toyota', 'toyouta', 'vokswagen', 'volkswagen', 'vw', 'volvo'], dtype=object)
```

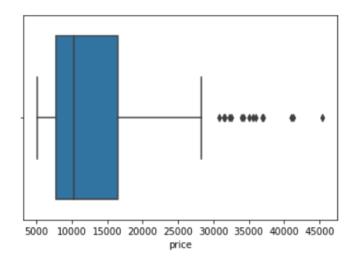
After:

```
array(['alfa-romero', 'audi', 'bmw', 'chevrolet', 'dodge', 'honda', 'isuzu', 'jaguar', 'mazda', 'buick', 'mercury', 'mitsubishi', 'Nissan', 'peugeot', 'plymouth', 'porsche', 'renault', 'saab', 'subaru', 'toyota', 'volkswagen', 'volvo'], dtype=object)
```

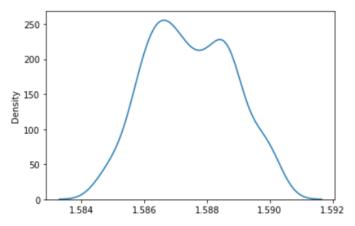
After name correction no records were duplicate

Target variable transformation

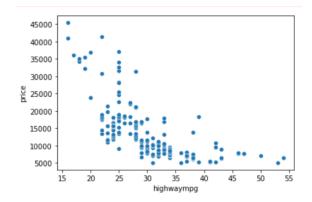


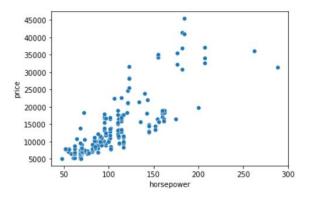


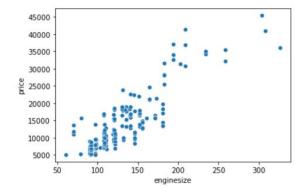
After Transformation



CORRELATION





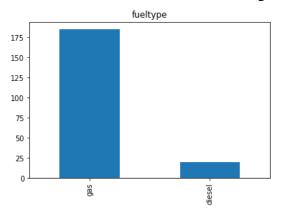


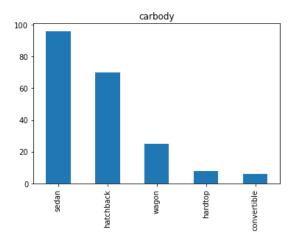
PEARSON COEFFICIENT INDEPENDEPENDENT VARIABLES CORRELATION

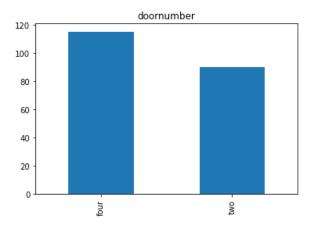
symboling -	- 1	-0.53	-0.36	-0.23	-0.54	-0.23	-0.11	-0.13	-0.0087	-0.18	0.071	0.27	-0.036	0.035	-0.08
wheelbase -	-0.53	1	0.87	0.8	0.59	0.78	0.57	0.49	0.16	0.25	0.35	-0.36	-0.47	-0.54	0.58
carlength -	-0.36	0.87	1	0.84		0.88	0.68	0.61	0.13	0.16		-0.29	-0.67	-0.7	0.68
carwidth -	-0.23	0.8	0.84	1	0.28	0.87	0.74		0.18	0.18	0.64	-0.22	-0.64	-0.68	0.76
carheight -	-0.54		0.49	0.28	1	0.3	0.067	0.17	-0.055	0.26	-0.11	-0.32	-0.049	-0.11	0.12
curbweight -	-0.23	0.78	0.88	0.87	0.3	1	0.85	0.65	0.17	0.15	0.75	-0.27	-0.76	-0.8	0.84
enginesize -	-0.11		0.68	0.74	0.067	0.85	1	0.58	0.2	0.029	0.81	-0.24	-0.65	-0.68	0.87
boreratio -	-0.13		0.61		0.17	0.65		1	-0.056	0.0052		-0.25	-0.58	-0.59	
stroke -	-0.0087	0.16	0.13	0.18	-0.055	0.17	0.2	-0.056	1	0.19	0.081	-0.068	-0.042	-0.044	0.079
compressionratio -	-0.18	0.25	0.16	0.18	0.26	0.15	0.029	0.0052	0.19	1	-0.2	-0.44	0.32	0.27	0.068
horsepower -	0.071	0.35	0.55	0.64	-0.11	0.75	0.81	0.57	0.081	-0.2	1	0.13	-0.8	-0.77	0.81
peakrpm -	0.27	-0.36	-0.29	-0.22	-0.32	-0.27	-0.24	-0.25	-0.068	-0.44	0.13	1	-0.11	-0.054	-0.085
citympg -	-0.036	-0.47	-0.67	-0.64	-0.049	-0.76	-0.65	-0.58	-0.042	0.32	-0.8	-0.11	1	0.97	-0.69
highwaympg -	0.035	-0.54	-0.7	-0.68	-0.11	-0.8	-0.68	-0.59	-0.044	0.27	-0.77	-0.054	0.97	1	-0.7
price -	-0.08	0.58	0.68	0.76	0.12	0.84	0.87	0.55	0.079	0.068	0.81	-0.085	-0.69	-0.7	1
	symboling -	wheelbase -	carlength -	carwidth -	carheight -	aurbweight -	enginesize –	boreratio -	stroke -	compressionratio -	horsepower -	peakrpm -	atympg -	highwaympg -	price -

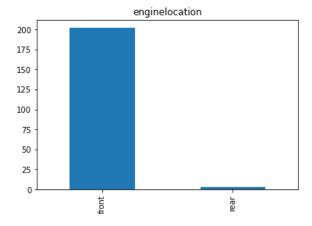
- 1.00 - 0.75 - 0.50 - 0.25 - 0.00 - -0.25 - -0.50 - -0.75

Univariate Analysis:

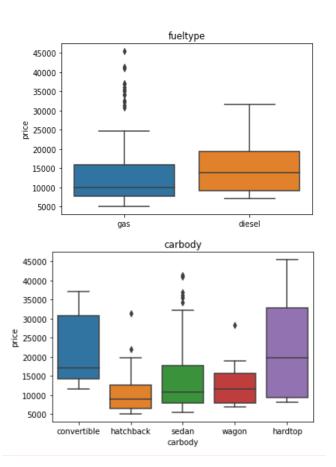


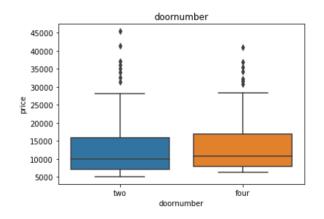


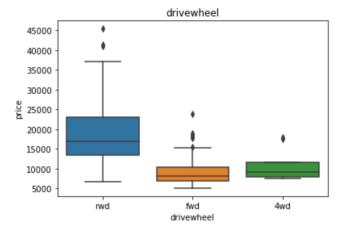




Bivariate Analysis







Feature Engineering

- df['mpg']=(df['citympg']+df['highwaympg'])/2
- df['Square footage']=round((df['carlength']*df['carwidth']),2)

Scaling and encoding has been done

Model Building:

OLS Regression Results

Dep. Variable:	price	R-squared:	0.897
Model:	OLS	Adj. R-squared:	0.874
Method:	Least Squares	F-statistic:	38.69
Date:	Tue, 07 Feb 2023	Prob (F-statistic):	1.66e-48
Time:	19:32:59	Log-Likelihood:	200.89
No. Observations:	153	AIC:	-343.8
Df Residuals:	124	BIC:	-255.9
Df Model:	28		
Covariance Type:	nonrobust		

2.169	Durbin-Watson:	42.650	Omnibus:
214.656	Jarque-Bera (JB):	0.000	Prob(Omnibus):
2.44e-47	Prob(JB):	0.858	Skew:
188.	Cond. No.	8.543	Kurtosis:

train score : 0.8418807629352103 test score : 0.8332429736237275

Final Model:

OLS Regression Results

Kurtosis: 6.953

Dep. Variable	e:		price	R-	squared	i: 0.	863
Mode	l:		OLS	Adj. R-	Adj. R-squared		857
Method	i:	Least So	st Squares		F-statistic:		31.0
Date	: Tu	ie, 07 Feb	2023	Prob (F-	statistic	2.016	e-59
Time	: :	19	33:42	Log-Lil	kelihood	i: 179	9.12
No. Observations	s:		153		AIC	:: -34	12.2
Df Residuals	s:		145		BIC	: -3	18.0
Df Mode	l:		7				
Covariance Type	: :	non	robust				
		coef	std err	t	P> t	[0.025	0.975]
				_	- 1-1	•	-
C	onst	0.0694	0.053	1.320	0.189	-0.034	0.173
enginet	type	-0.0754	0.038	-2.006	0.047	-0.150	-0.001
engine	size	0.6283	0.073	8.574	0.000	0.483	0.773
str	roke	-0.0791	0.044	-1.778	0.078	-0.167	0.009
Square foot	tage	0.3294	0.056	5.884	0.000	0.219	0.440
enginelocation_	rear	0.3530	0.052	6.838	0.000	0.251	0.455
cylindernumber_	four	-0.1219	0.024	-5.105	0.000	-0.169	-0.075
cylindernumber	_six	-0.0530	0.028	-1.898	0.060	-0.108	0.002
Omnibus:	35.0	89 D u	ırbin-Wa	tson:	2.156		
Prob(Omnibus):	0.0	00 Jarq	ue-Bera	(JB):	116.564		
Skew:	0.8	15	Prob	(JB): 4	.88e-26		

22.9

Cond. No.

r2_score(ypred,ytrain)

0.8855335646553271

r2_score(ypred1,ytest)

0.8956859473729187

THANK YOU