Project Demonstration

Statistics for Data Science

Contents

Introduction

Cars24 is a leading AutoTech company focused on the sale, purchase, and financing of pre-owned cars.

The company offers an online marketplace for buying and selling used cars, complemented by a suite of services including car financing, quality checks, warranties, and seamless documentation for transactions.

Cars24 primarily serves the automotive industry with a customer base looking for pre-owned vehicle solutions.

Problem Statement

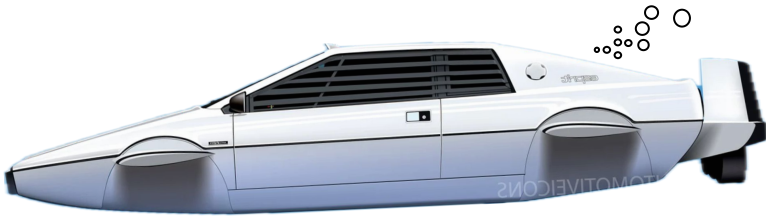
The main idea I had behind using this dataset was to try and find some way to predict the selling price of a used car based on brand, model, age, no of previous owners, fuel type, kilometers driven and transmission type.

Consider a scenario like while adding a new record of a used car data, someone should make an assessment of the car and figure out what the selling price should be.

The goal of this project is to automate this task using linear regression.

Initial EDA

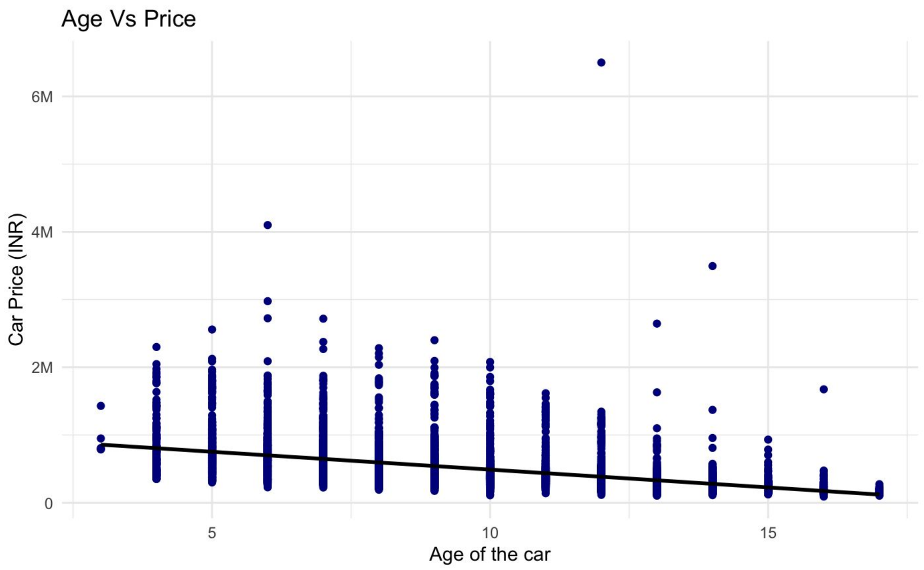
- Removed rows that contains null values
- Explored each column
    - car_brand: 26 different brands
    - age(from year) : [3, 17]
    - fuel : Petrol, Diesel, LPG, CNG, Electric
    - km_driven : [179, 912380]
    - gear : Manual & Automatic
    - ownership : 1, 2, 3, 4
    - price : [91000, 6500000]



Hypothesis Testing

Tested hypothesis : Does price of the car decreases as the age increase ?

Result: Yes it does !



Preprocessing the data

Converting categorical variables into numerical

- One hot encoding
- Target encoding

Scaling the data

This is essential for maintaining consistent relationship between the features and improving model performance. Consider the below example

- age ranges between 3 to 17
- km_driven ranges between 179 to 912380

Since the features have a massive difference on their range, it is better to have these values in scale.

Before preprocessing

car_brand	model	price	year	location	fuel	km_driven	gear	ownership	emi
Hyundai	EonERA PLUS	330399	2016	Hyderabad	Petrol	10674	Manual	2	7350
Maruti	Wagon R 1.0LXI	350199	2011	Hyderabad	Petrol	20979	Manual	1	7790
Maruti	Alto K10LXI	229199	2011	Hyderabad	Petrol	47330	Manual	2	5098
Maruti	RitzVXI BS IV	306399	2011	Hyderabad	Petrol	19662	Manual	1	6816
Tata	NanoTWIST XTA	208699	2015	Hyderabad	Petrol	11256	Automatic	1	4642
Maruti	AltoLXI	249699	2012	Hyderabad	Petrol	28434	Manual	1	5554
Maruti	AltoLXI	240599	2011	Hyderabad	Petrol	31119	Manual	1	5352
Maruti	Alto K10LXI	191999	2010	Hyderabad	Petrol	10910	Manual	1	4271
Honda	Brio1.2 S MT I VTEC	362299	2013	Hyderabad	Petrol	40362	Manual	2	8059
Maruti	Wagon R 1.0VXI	385799	2013	yderabad	Petrol	15673	Manual	2	8582

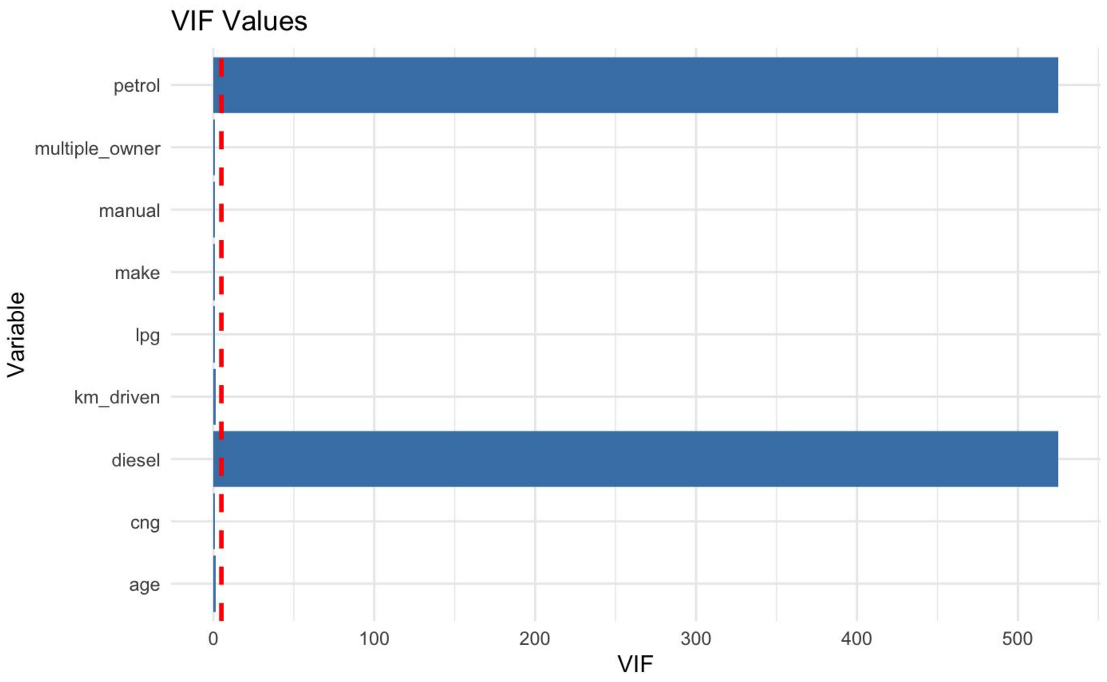
After preprocessing

	car_brand	brand_new	petrol	diesel	lpg	cng	age	km_driven	manual	multiple_owner	price
1	Hyundai	0.1526935	1	0	0	0	0.3571429	0.011505140	1	1	0.03735357
2	Maruti	0.1257020	1	0	0	0	0.7142857	0.022801992	1	0	0.04044297
3	Maruti	0.1257020	1	0	0	0	0.7142857	0.051689266	1	1	0.02156327
4	Maruti	0.1257020	1	0	0	0	0.7142857	0.021358231	1	0	0.03360883
5	Tata	0.2437464	1	0	0	0	0.4285714	0.012143157	0	0	0.01836464
6	Maruti	0.1257020	1	0	0	0	0.6428571	0.030974533	1	0	0.02476190
7	Maruti	0.1257020	1	0	0	0	0.7142857	0.033917963	1	0	0.02334202
8	Maruti	0.1257020	1	0	0	0	0.7857143	0.011763855	1	0	0.01575893
9	Honda	0.1647287	1	0	0	0	0.5714286	0.044050598	1	1	0.04233094
10	Maruti	0.1257020	1	0	0	0	0.5000000	0.016985292	1	1	0.04599766

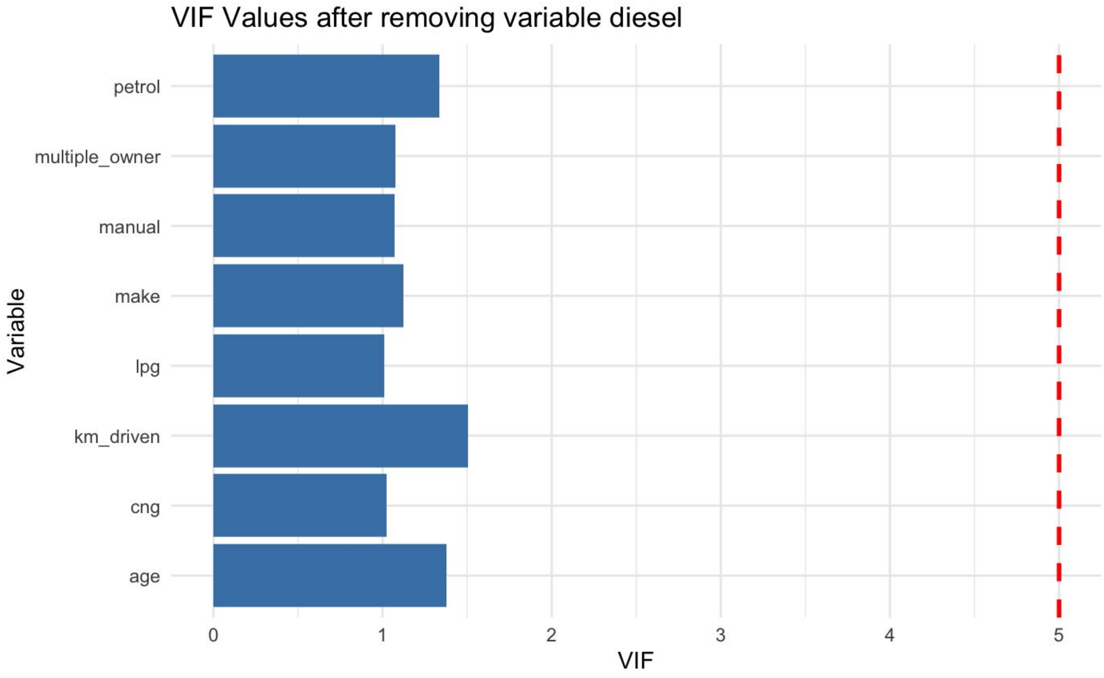
Linear Regression Model

$$
\begin{aligned}
Price = 0.025 &+ 0.0240 \times \text{brand} - 0.100 \times \text{age} + 0.068 \times \text{petrol} \\
&+ 0.090 \times \text{diesel} + 0.009 \times \text{lpg} - 0.006 \times \text{cng} \\
&- 0.038 \times \text{km\_driven} - 0.027 \times \text{manual} - 0.001 \times \text{multiple\_owner}
\end{aligned}
$$

Multicollinearity check



Multicollinearity check



New Linear Regression Model

$$\begin{aligned}Price = 0.116 &+ 0.0240 \times \text{brand} - 0.100 \times \text{age} + 0.022 \times \text{petrol} \\ &+ 0.009 \times \text{lpg} - 0.006 \times \text{cng} - 0.038 \times \text{km\_driven} \\ &- 0.027 \times \text{manual} - 0.001 \times \text{multiple\_owner}\end{aligned}$$

R - Squared value

Insight gathered:

This value implies that approximately 64% of the variance in car price can be explained by the independent variables (e.g., age, km_driven, make(brand), etc.) in your regression model.

This value makes sense, because we haven't considered some variables like model, location and emi.

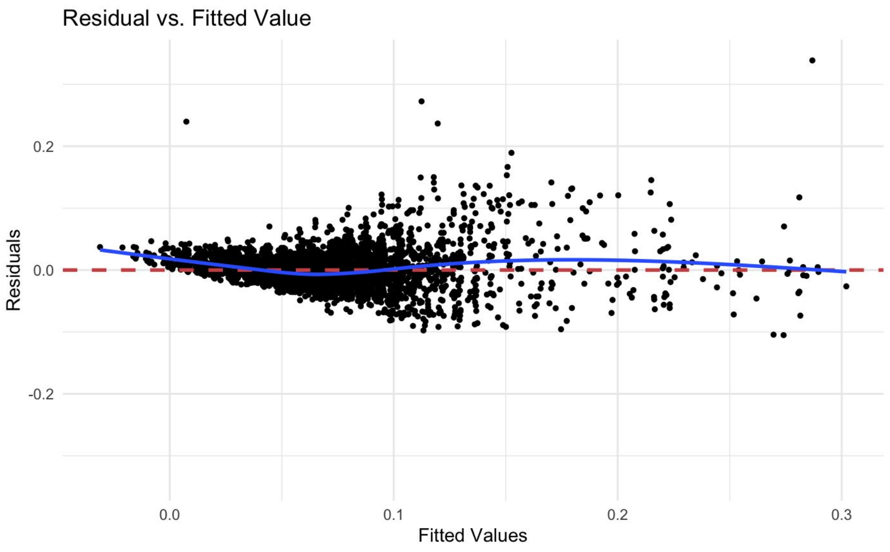
```
{r}
summary(model_new)$r.squared

 [1] 0.6416685
```

Model Evaluation

Residual vs Fitted value

Insight gathered:

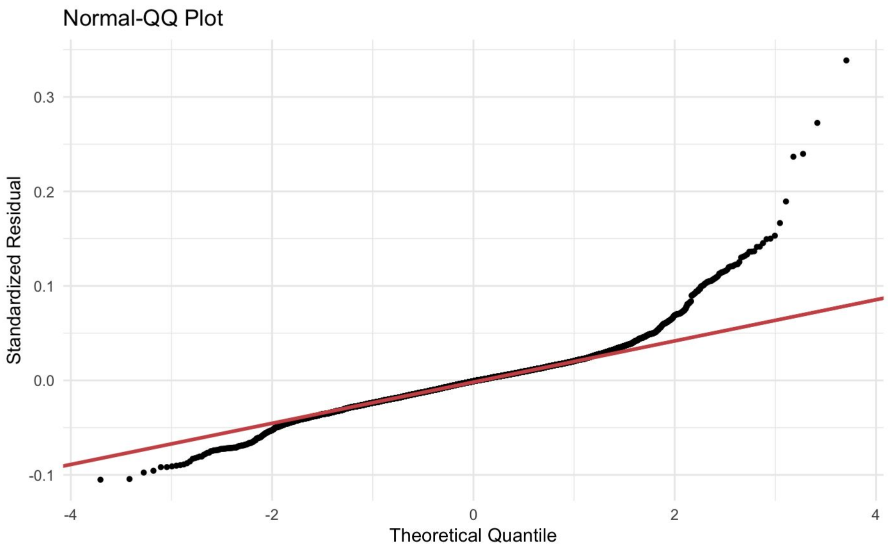
- Curvature suggests possible nonlinearity issues.
- Spread indicates heteroscedasticity in residuals.
- Outliers may heavily influence model results



QQ-Plots

Insight gathered:

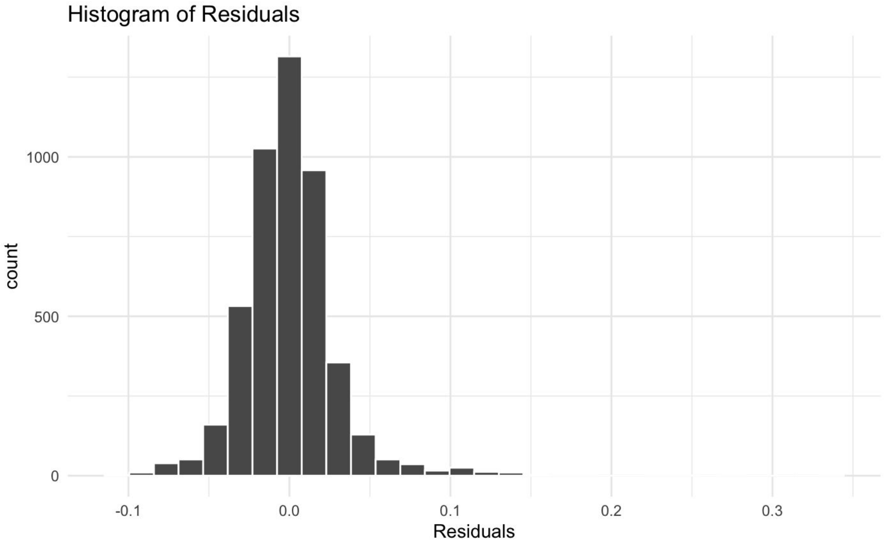
- Deviations occur at extreme quantiles.
- Skewness observed in tails.
- Potential outliers at upper quantile extremes.



Residual Histogram

Insight gathered:

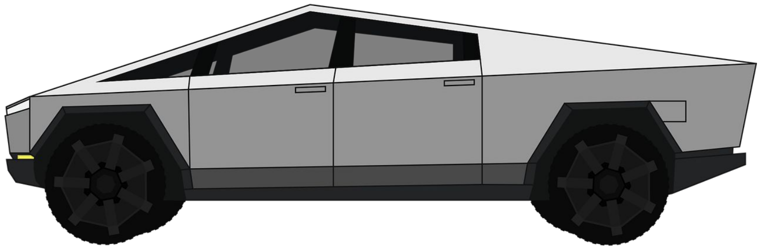
- Residuals show approximate normal distribution shape.
- Outliers appear at the distribution's tails.
- Minor skewness suggests slight model deviation.



Conclusion

- Older cars significantly lower the predicted price.
- Residuals are nearly normal, validating assumptions.
- QQ plot shows slight deviations at extremes.
- Petrol and manual reduce car price substantially.
- Model explains 64.17% variation in car price.

Out of Curiosity



Root Mean Square Error

Insight gathered:

- If RMSE is 0, the model's predictions are perfect.
- A lower RMSE indicates better model performance, but it should be compared to other models or benchmarks for context.

```r
mse <- mean(model_new$residuals ^ 2)

# root mean squared error
rmse <- sqrt(mse)

print(rmse)
```

```
[1] 0.02893458
```

Applying the model to test set

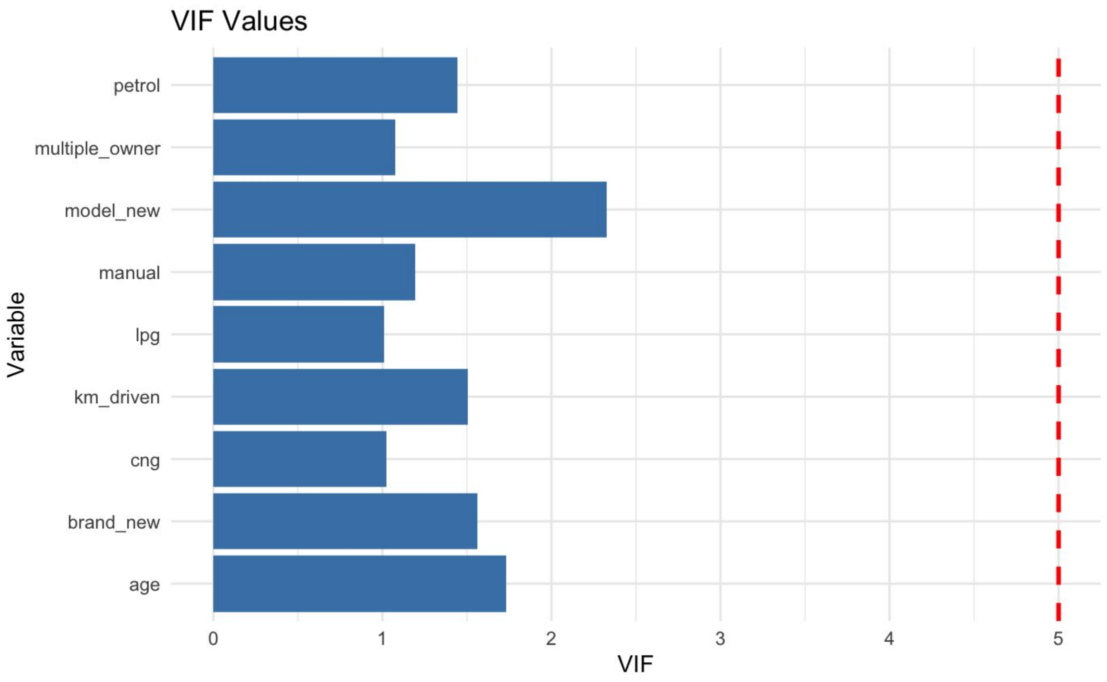
car_brand <chr>	price_ <dbl>	predicted_price_ <dbl>
Maruti	249699	284226.4
Maruti	240599	237681.0
Hyundai	401599	453509.5
Maruti	241599	238834.0
Hyundai	535699	557352.5
Maruti	375299	610155.6
Maruti	410699	609495.9
Hyundai	401699	417560.1
Hyundai	357399	318969.7
Maruti	238099	282509.4

Linear Regression Model

I have created a another linear regression model incorporating car's model and below is the equation of the model.

$$
\begin{aligned}
Price = {} & 0.028 + 0.082 \times \text{brand} - 0.033 \times \text{age} + 0.0068 \times \text{petrol} \\
& + 0.420 \times \text{model} + 0.009 \times \text{lpg} - 0.002 \times \text{cng} \\
& - 0.050 \times \text{km\_driven} - 0.0018 \times \text{manual} - 0.002 \times \text{multiple\_owner}
\end{aligned}
$$

Multicollinearity check



RMSE comparison

First model, after removing the variable diesel.

```
mse <- mean(model_new$residuals ^ 2)

# root mean squared error
rmse <- sqrt(mse)

print(rmse)

[1] 0.02893458
```

Final model, this includes the variable model.

```
mse <- mean(cars24_model$residuals ^ 2)

# root mean squared error
rmse <- sqrt(mse)

print(rmse)

[1] 0.01731665
```

R - squared comparison

First model, after removing the variable diesel.

```
{r}
summary(model_new)$r.squared

 [1] 0.6416685
```

Final model, this includes the variable model.

```
{r}
summary(cars24_model)$r.squared

 [1] 0.8716553
```

Questions ?

Thank you !