Predicting used car prices in the UK market

Author: Kostas Batsis

Date: 3/12/2020

This project concerns the construction of a linear regression model for the prediction of used car prices in the UK market. It is based on datasets of 108,540 scraped used car listings, split into car brand, recovered from kaggle [1]. These include information on prices, mileage, miles per gallon, engine size, year of production, transmission type and fuel type. We merged the sets into a single dataset and added the car brand as an extra variable. Price will serve as the dependent variable and the rest of the mentioned variables will serve as our independent variables. The analysis was performed with Anaconda Python 3.8 and the statsmodels module.

We begun by looking for missing data. We found 9,353 missing values in the miles per gallon variable (8.6 percent of the total variable data, 4,947 cars using petrol, 4,249 using diesel, 151 hybrid ones and 6 of other fuel type). We produced a visual missing data matrix to examine the distribution of the missingness (see figure 1). All the missing data are concentrated in one continuous section of the dataset indicating the website of origin as a possible source for the missingness. In any case the data are of the missing not at random type and therefore we excluded the miles per gallon variable from the analysis.

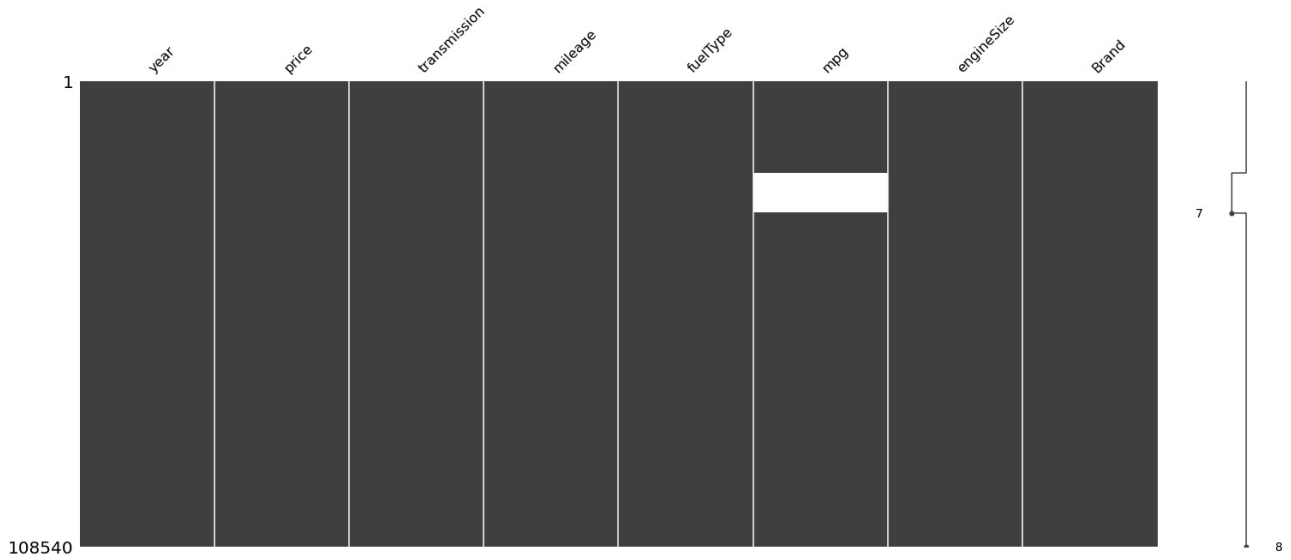

Figure 1: Visual missing data matrix, the missing data are represented by white lines. On the right a line plot indicating the amount of missing data in each row.

We then converted our categorical variables (car brand, transmission type and fuel type) into dummy variables. Table 1 presents counts for these variables. We note that the other transmission and the electric car categories have low counts (10 and 6 respectively).

Car brand		Transmission		Fuel type	
VW	15157	Semi-Auto	24903	Diesel	45177
Vauxhall	13632	Automatic	22319	Hybrid	3229
Mercedes	13119	Other	10	Other	253
BMW	10781			Electric	6
Audi	10668				
Toyota	6738				
Skoda	6267				
Focus	5454				
Hyundi	4860				
Cclass	3899				

Table 1: Counts for the categorical independent variables.

Figure 2 presents histograms of our numerical variables: price, mileage, engine size and model year. We observe that the price, mileage and engine size all have a positive skew. We observe one case larger than 2020 in the year variable which is clearly erroneous and thus removed. We also found 286 cases of a nonsensical value of 0 litres engine size (172 petrol using cars, 73 diesel using cars, 38 hybrid ones, 2 electric cars and 1 using an other fuel type). We considered these as artifacts of the scraping process although the electric cars might be an exception. Since for this latter category the engine size feature is not particularly meaningful and the sample size is small (6, see table 1) we decided to exclude the category from the model. We also performed a natural logarithm transformation of price, in order to make our dependent variable more normal since a preliminary residual plot (see figure 5) revealed serious problems with non-linearity and heteroscedasticity as well as negative predicted prices. Figure 3 presents all the modified variables, price has now become approximately normal while year has a negative skew.

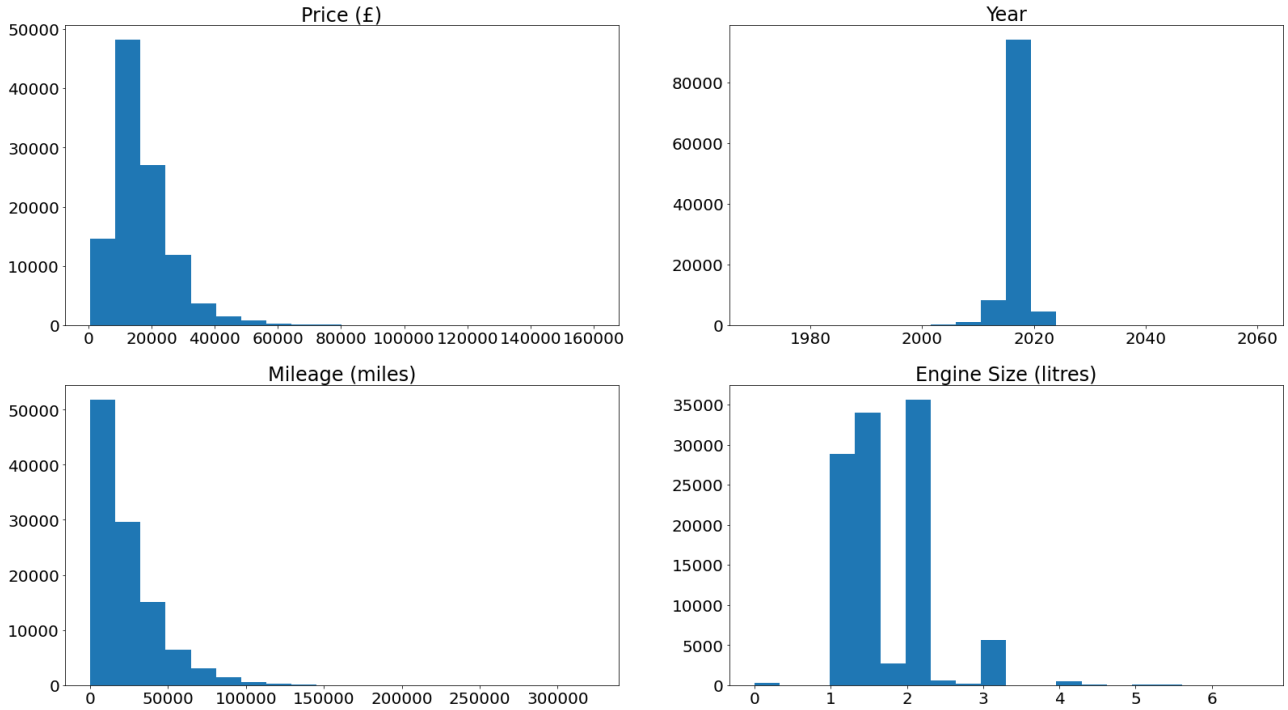

Figure 2: Histograms of the numerical variables.

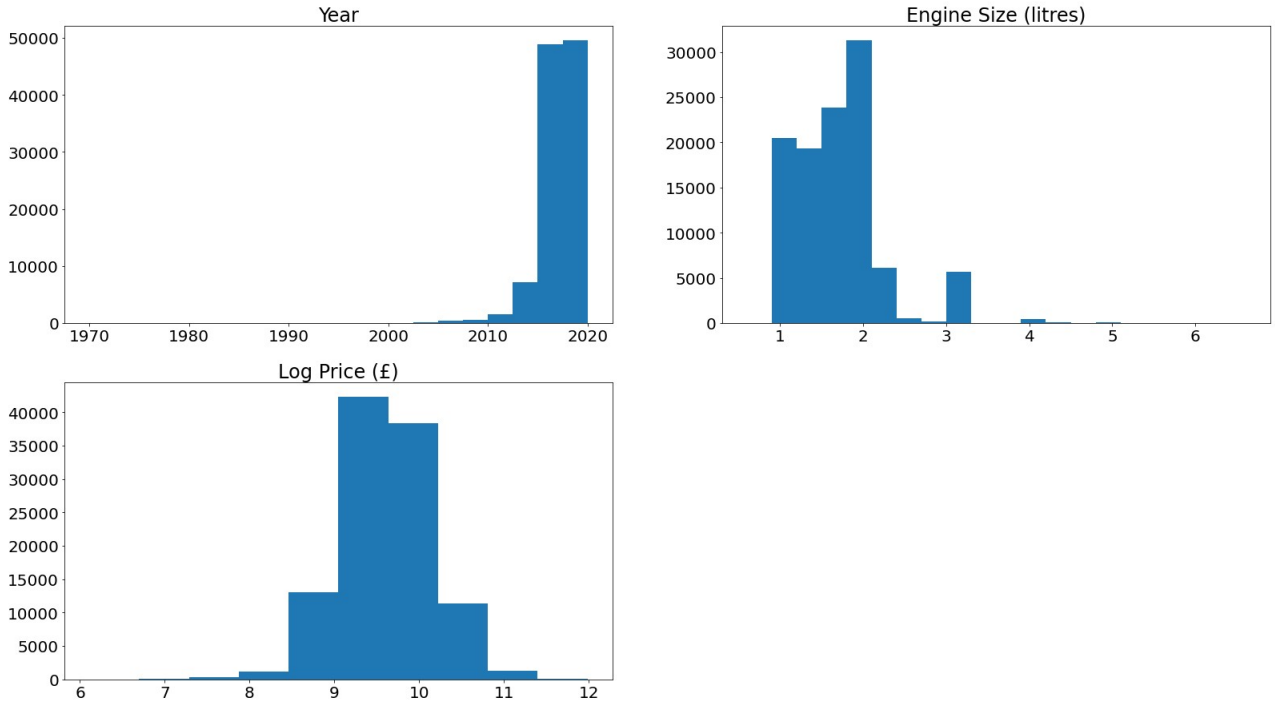

Figure 3: Histograms of the modified numerical variables.

We then searched for possible multivariate outliers with a 3D scatterplot of the IVs (figure 4). All points had logical values with no abnormalities present. There was one extreme case consisting of a car with a model year value of 1970 but since there is nothing unreasonable for such old cars to be in the market we let it in the analysis for the time being.

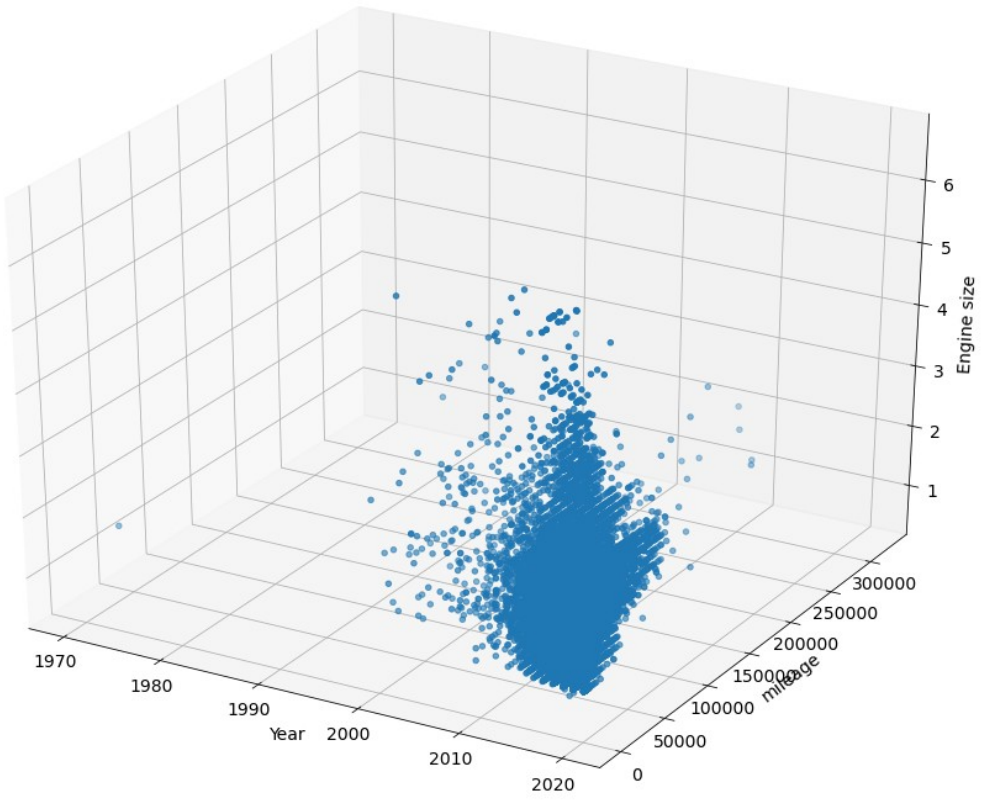

Figure 4: Scatterplot of all the numerical IVs.

We also screened for multicollinearity among the IVs using the Variance Inflation Factor (VIF, see table 2). The year and engine size variables had very high VIF values but since multicollinearity doesn't affect the overall fit of the model or produce bad predictions but only affects coefficient estimates [2] we proceeded with the analysis as is.

Regressor	VIF
year	48.592735
mileage	2.424685
engineSize	19.49333
bmw	2.04829
cclass	1.396727
focus	1.629351
ford	3.074147
hyundi	1.535457
merc	2.319353
skoda	1.66473
toyota	2.010253
vauxhall	2.560968
vw	2.527612
Manual	5.660133
Other	1.000638
Semi-Auto	2.29209
Hybrid	1.442368
Other	1.026311
Petrol	3.336715

Table 2: Variance Inflation Factors for the IVs.

Subsequently we fitted the regression equation using the untransformed price and as noted before this revealed serious issues with non-linearity, heteroscedasticity and negative predictions (figure 5). We fitted the equation again with the log transformed price and although the residual plot improved we found an extreme outlier (figure 6). This outlier was the case with the car made in 1970 that we noted before, we deleted the case keeping in mind that our model might fail for very old cars. We ran the regression again and got a residual plot that satisfied the normality, linearity and homoscedasticity assumptions to a reasonable degree.

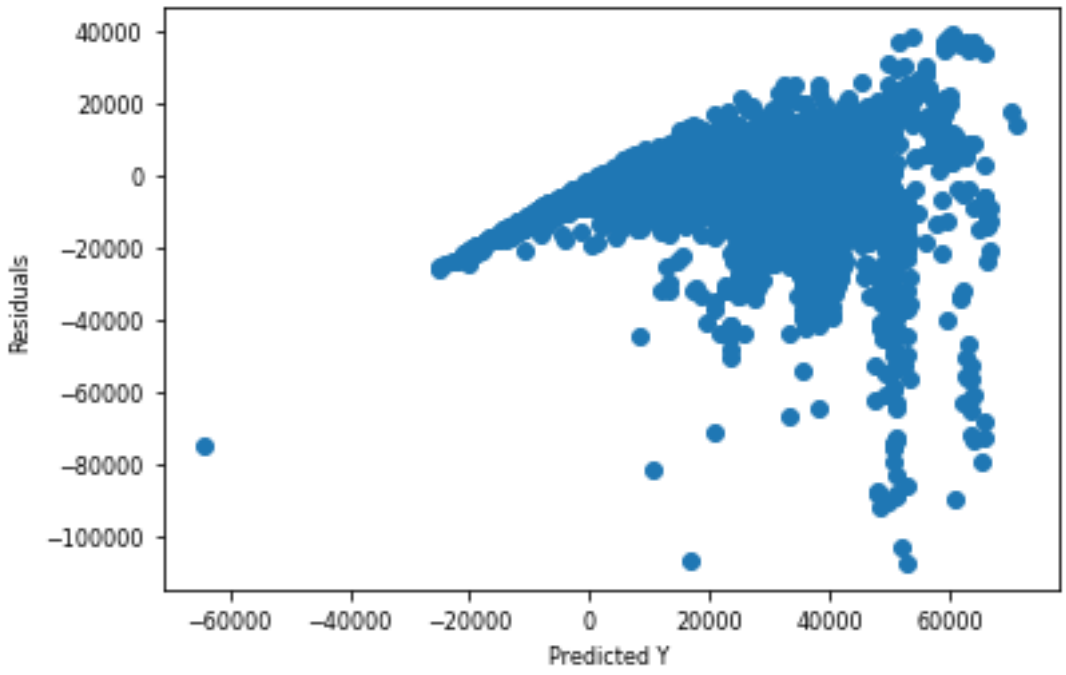

Figure 5: Residual plot for regression run with the untransformed price variable (DV).

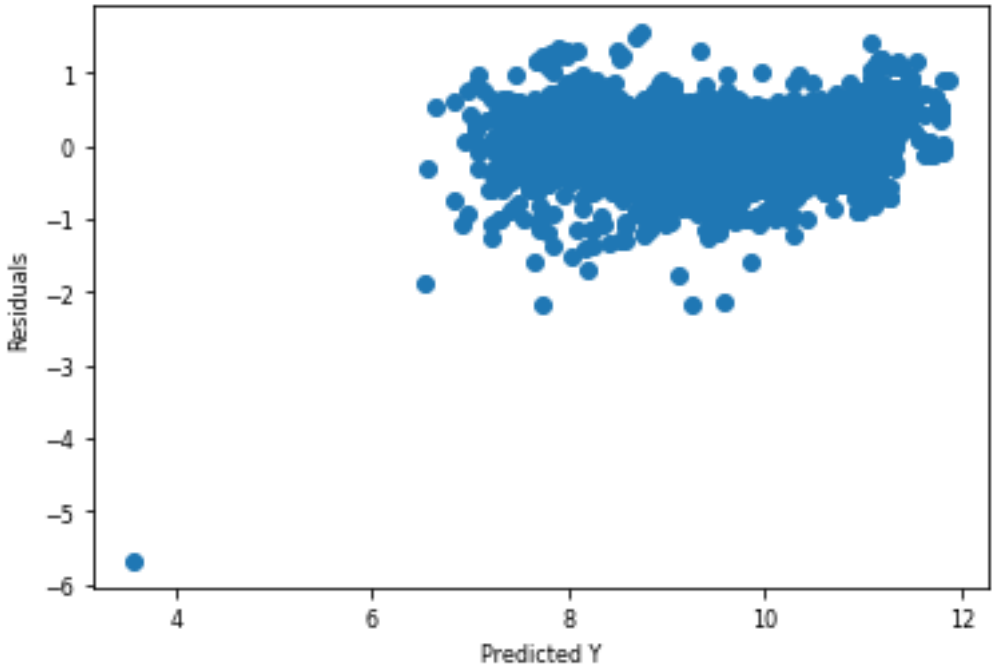

Figure 6: Residual plot for regression run with the transformed price variable (DV).

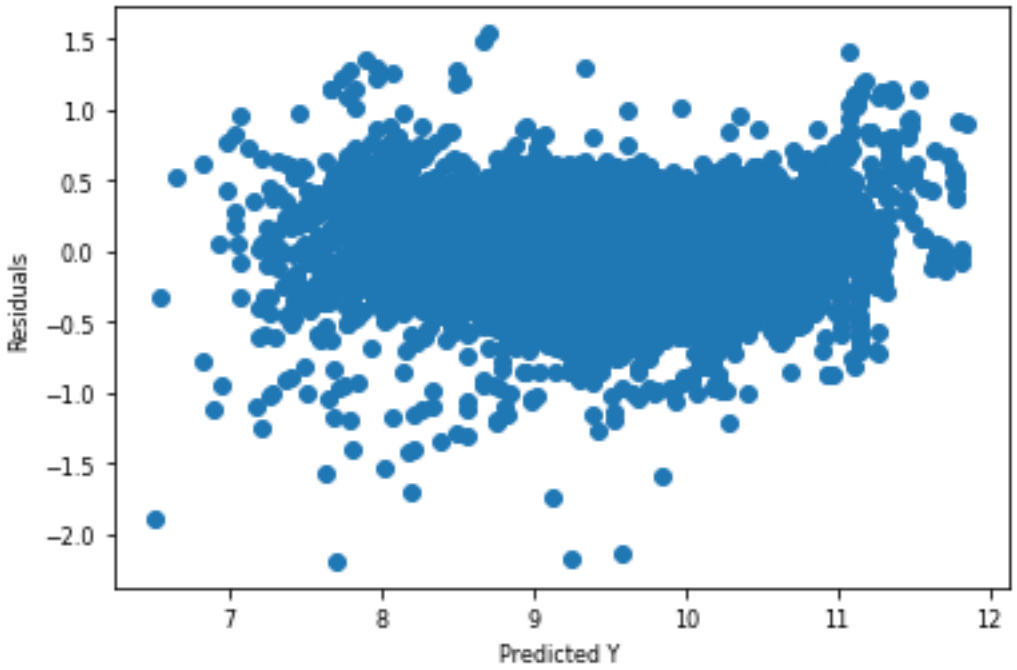

Figure 7: Residual plot for regression run with the transformed price variable (DV) and one extreme outlier removed.

Table 3 presents the statsmodels output for this final regression. We first notice that the Durbin-Watson test had a value of 1.5 indicating that there were no problems with non-independence of errors. Most importantly the adjusted R-squared equaled 0.880 which means that our model accounted for 88 percent of the variance in car prices and therefore is a useful and valid tool for predicting these values.

OLS Regression Results

Dep. Variable:	priceln	R-squared:	0.880
Model:	OLS	Adj. R-squared:	0.880
Method:	Least Squares	F-statistic:	4.193e+04
Date:	Wed, 02 Dec 2020	Prob (F-statistic):	0.00
Time:	15:14:20	Log-Likelihood:	29665.
No. Observations:	108252	AIC:	-5.929e+04
Df Residuals:	108232	BIC:	-5.910e+04
Df Model:	19		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-229.7708	0.820	-280.257	0.000	-231.378	-228.164
year	0.1185	0.000	291.960	0.000	0.118	0.119
mileage	-5.226e-06	4.15e-08	-125.867	0.000	-5.31e-06	-5.14e-06
engineSize	0.4166	0.001	295.555	0.000	0.414	0.419
bmw	-0.1328	0.003	-52.253	0.000	-0.138	-0.128
cclass	-0.0713	0.003	-20.460	0.000	-0.078	-0.064
focus	-0.1910	0.003	-60.003	0.000	-0.197	-0.185
ford	-0.2561	0.002	-106.099	0.000	-0.261	-0.251
hyundi	-0.3653	0.003	-111.111	0.000	-0.372	-0.359
merc	-0.0474	0.002	-19.375	0.000	-0.052	-0.043
skoda	-0.2919	0.003	-97.308	0.000	-0.298	-0.286
toyota	-0.4145	0.003	-130.071	0.000	-0.421	-0.408
vauxhall	-0.4468	0.003	-176.915	0.000	-0.452	-0.442
vw	-0.1623	0.002	-68.258	0.000	-0.167	-0.158
Manual	-0.1457	0.002	-82.272	0.000	-0.149	-0.142
Other	0.0117	0.061	0.190	0.849	-0.109	0.132
Semi-Auto	0.0030	0.002	1.701	0.089	-0.000	0.006
Hybrid	0.2007	0.004	51.304	0.000	0.193	0.208
Other	0.0917	0.012	7.808	0.000	0.069	0.115
Petrol	0.0017	0.001	1.234	0.217	-0.001	0.004

Omnibus:	6860.779	Durbin-Watson:	1.501
Prob(Omnibus):	0.000	Jarque-Bera (JB):	32185.440
Skew:	0.060	Prob(JB):	0.00
Kurtosis:	5.669	Cond. No.	4.59e+07

Table 3: Results of the final regression run.

[1] Aditya, 100,000 UK Used Car Data set, https://www.kaggle.com/adityadesai13/used-car-dataset-ford-and-mercedes/tasks?taskId=1258

[2] Michael Kutner, William Wasserman, Christopher Nachtsheim, John Neter, Applied Linear Statistical Models, 4th edition