PRESENTED BY
Warren Morelli

USED CAR VALUATION PREDICTION:

A Linear Regression Model for Estimating Used Car Sale Prices

21 FEB, 2023

AGENDA OUTLINE

1. BUSINESS PROBLEM
2. OBTAINING DATA
3. DATA PREPARATION
4. MODELLING
5. DEPLOYMENT - PREDICTIVE MODELING

BUSINESS PROBLEM

A used car enterprise seeks a precise approach for predicting used car prices to assist in their purchasing of undervalued vehicles and to accurately set prices to enable profitability upon the sale.

MY GOAL: Create a model to predict the sale price of a used car based on its various features such as model, year, mileage, engine power, etc.

OBTAINING DATA

The data utilized in this project was sourced from the Used-Car-data CSV file on the Kaggle website. The data set is comprised of 7906 rows and 18 columns, providing an ample amount of information for linear regression modeling purposes. This comprehensive data set is well-suited to support the modeling process and can be expected to yield reliable results.

DATA PREPARATION

Upon initial inspection, I noticed there were a couple of steps required to prepare my data, including:

- Drop variables that have no significance on price
- Group data by vehicle brand values
- Convert selling price to USD
- Drop outliers
- Rename variables for improved clarity
- Check for Variables that have a high degree of correlation (Multicollinearity)

MODELLING ITERATION 1

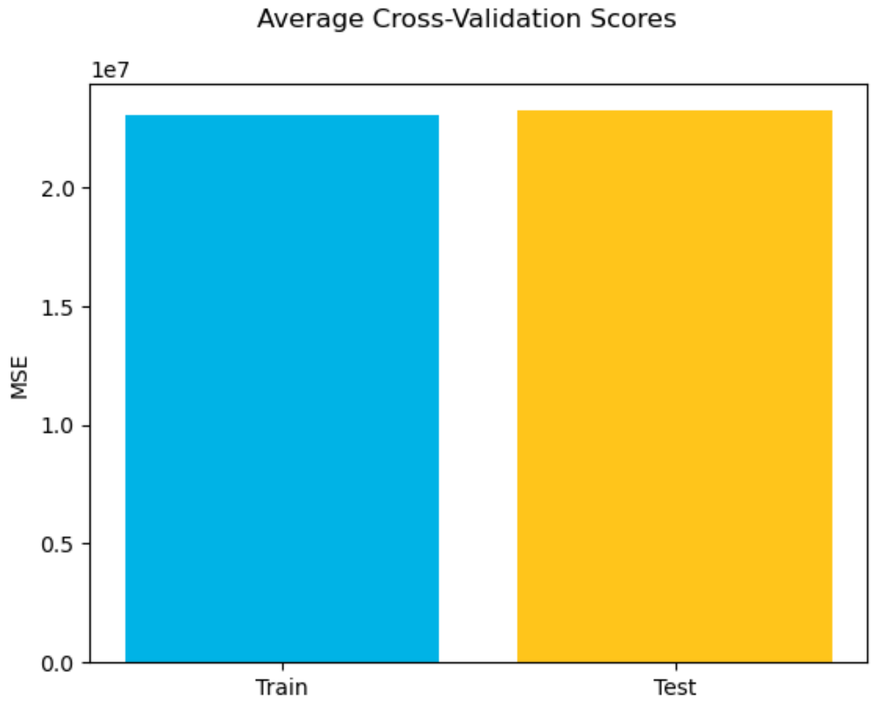
STEP 1: TESTING & TRAINING DATA

I divided the available data into two sets: a training set and a testing set. The training set is used to develop and train the model to make predictions. Once the model is developed, it is tested on the testing set to evaluate how well it can perform to new, unseen data.

STEP 2: CHECK MSE

I checked the mean square error on both testing and training date to see if they are similar. This is to assess the predictive performance of a model by estimating how well it will perform on independent datasets.

As the histogram suggests, the model is accurate.

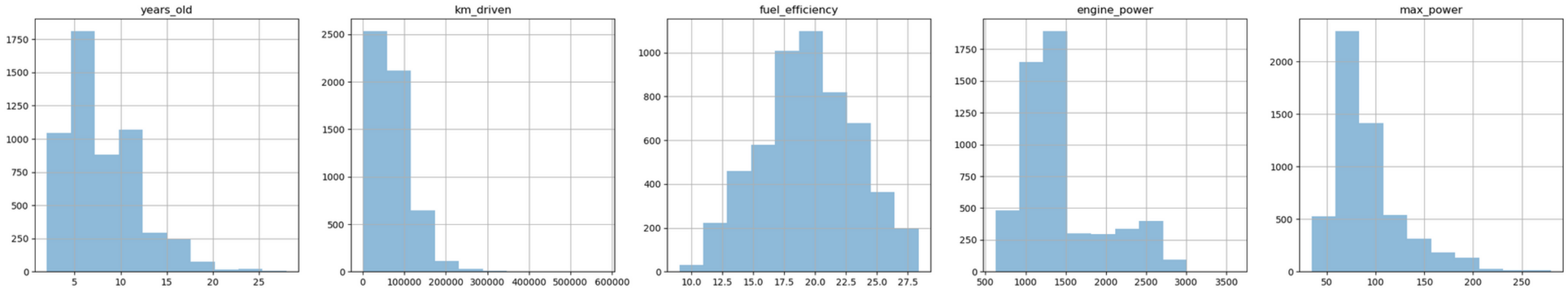



```
Cross Validation Test Mean Squared Error: 23215120.779188894
Cross Validation Train Mean Squared Error: 23036614.88218996
```

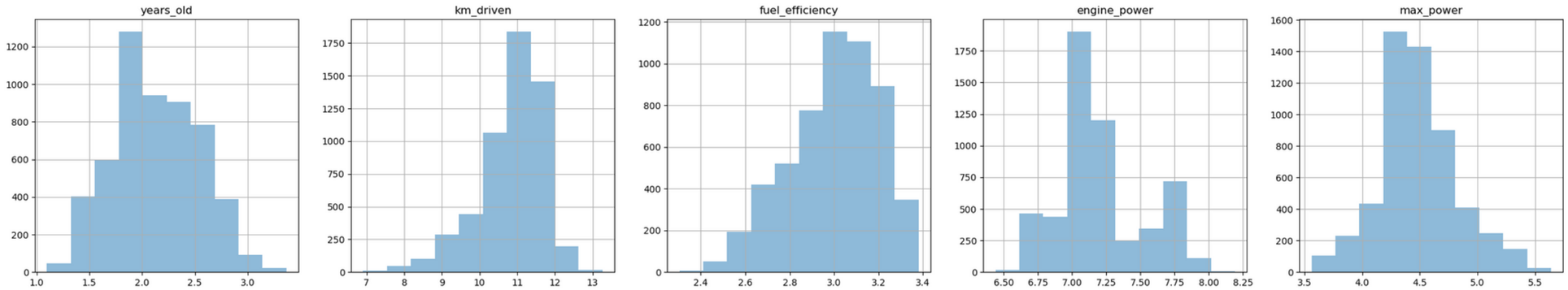
MODELLING ITERATION 2

SKEWNESS

I noticed that the data had a skewness, which could potentially impact the accuracy of the predictive model. To address this issue, I utilized a mathematical operation called a log transformation, which converts the data values into their logarithms. This transformation has the added benefit of improving the skewness of the data, thereby ensuring that the model is more accurate and reliable.



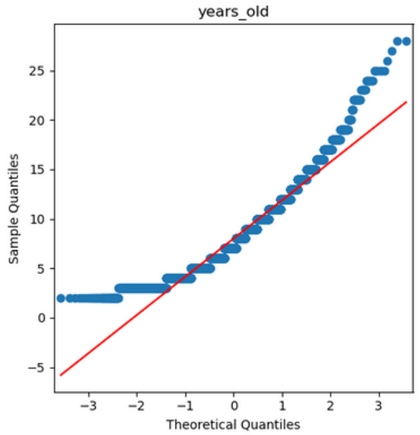
Before Log Transformation

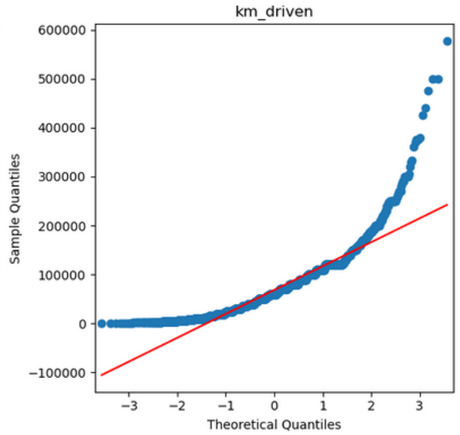


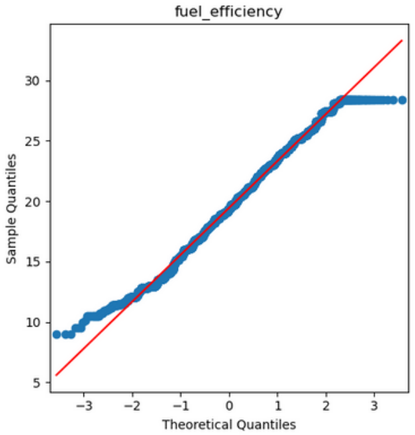
After Log Transformation

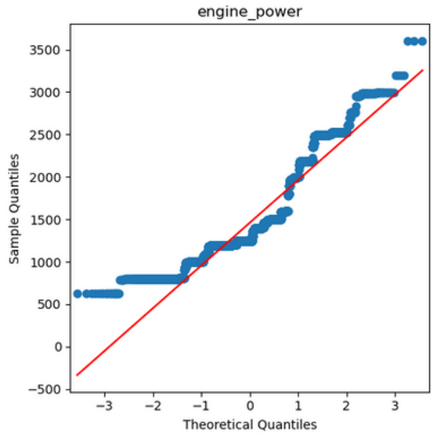
SKEWNESS

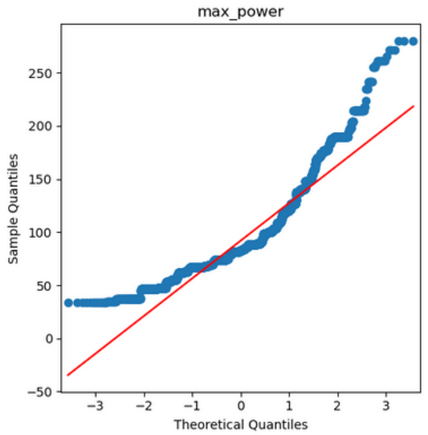
Before Log Transformation



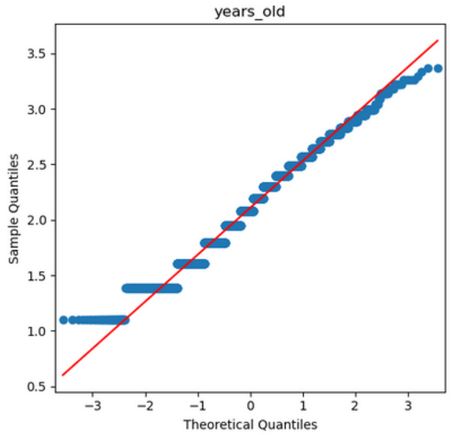


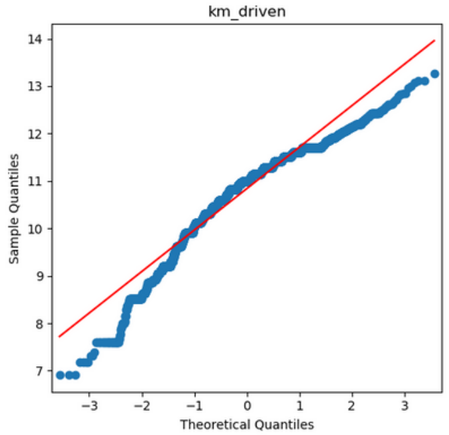


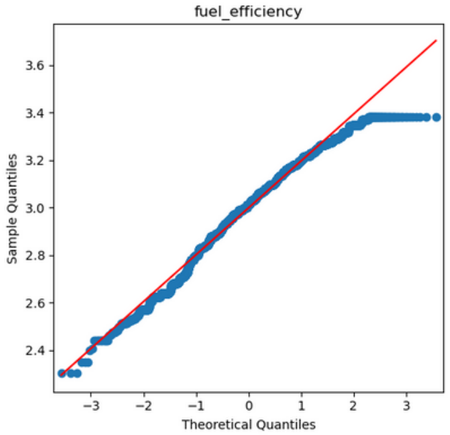


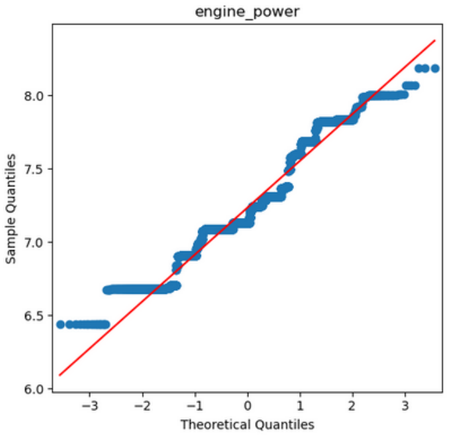


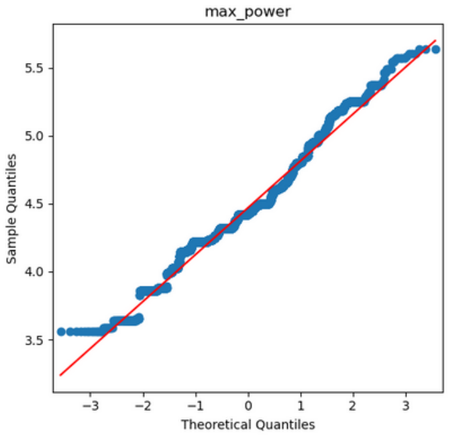
After Log Transformation











OLS Regression Results

Dep. Variable:	selling_price	R-squared:	0.868
Model:	OLS	Adj. R-squared:	0.868
Method:	Least Squares	F-statistic:	3584.
Date:	Wed, 22 Feb 2023	Prob (F-statistic):	0.00
Time:	13:11:33	Log-Likelihood:	-1198.3
No. Observations:	5459	AIC:	2419.
Df Residuals:	5448	BIC:	2491.
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.0477	0.102	10.231	0.000	0.847	1.249
years_old	-1.0296	0.015	-70.075	0.000	-1.058	-1.001
km_driven	-0.0065	0.007	-0.950	0.342	-0.020	0.007
fuel_efficiency	0.3072	0.035	8.858	0.000	0.238	0.377
engine_power	0.5185	0.032	16.024	0.000	0.455	0.582
max_power	0.7699	0.023	33.746	0.000	0.725	0.815
low_budget_brand	0.0857	0.035	2.480	0.013	0.018	0.153
mid_budget_brand	0.1768	0.036	4.935	0.000	0.106	0.247
high_budget_brand	0.7854	0.038	20.448	0.000	0.710	0.861
seats_0-4	0.4834	0.052	9.358	0.000	0.382	0.585
seats_5+	0.5643	0.056	10.052	0.000	0.454	0.674
fuel_Diesel	0.5840	0.056	10.422	0.000	0.474	0.694
fuel_Petrol	0.4837	0.047	9.915	0.000	0.372	0.555
transmission_Automatic	0.5806	0.052	11.103	0.000	0.478	0.683
transmission_Manual	0.4872	0.051	9.100	0.000	0.367	0.568

Omnibus:	345.167	Durbin-Watson:	2.047
Prob(Omnibus):	0.000	Jarque-Bera (JB):	496.045
Skew:	-0.546	Prob(JB):	1.93e-108
Kurtosis:	3.995	Cond. No.	1.99e+17

CHECK OLS REGRESSION RESULTS

Prior to developing my predictive model, I reviewed the OLS (Ordinary Least Squares) results to gain insights into the potential accuracy of my model.

FINDINGS:

1) According to the OLS analysis, the p-values of all variables are deemed acceptable except for 'km_driven'. This has a weak statistical effect on the dependent variable.

2) The adjusted R-squared indicates that there has been an improvement, making the goodness of fit even better.

3) The 'max_power' and 'engine_power' variables have high multicollinearity.

4) The skewness had improved from 0.92 to 0.54

OLS Regression Results

Dep. Variable:	selling_price	R-squared:	0.840
Model:	OLS	Adj. R-squared:	0.840
Method:	Least Squares	F-statistic:	3589.
Date:	Wed, 22 Feb 2023	Prob (F-statistic):	0.00
Time:	13:11:33	Log-Likelihood:	-1716.4
No. Observations:	5459	AIC:	3451.
Df Residuals:	5450	BIC:	3510.
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.7246	0.127	5.702	0.000	0.475	0.974
years_old	-1.1100	0.013	-86.351	0.000	-1.135	-1.085
fuel_efficiency	0.3912	0.039	10.111	0.000	0.315	0.467
engine_power	1.1019	0.030	37.048	0.000	1.044	1.160
low_budget_brand	-0.1221	0.010	-11.723	0.000	-0.143	-0.102
high_budget_brand	0.8766	0.025	35.062	0.000	0.828	0.926
seats_0-4	0.2316	0.063	3.652	0.000	0.107	0.356
seats_5+	0.4930	0.069	7.167	0.000	0.358	0.628
fuel_Diesel	0.3964	0.069	5.768	0.000	0.262	0.531
fuel_Petrol	0.3282	0.059	5.589	0.000	0.213	0.443
transmission_Automatic	0.4635	0.065	7.140	0.000	0.336	0.591
transmission_Manual	0.2611	0.063	4.124	0.000	0.137	0.385

Omnibus:	275.412	Durbin-Watson:	2.027
Prob(Omnibus):	0.000	Jarque-Bera (JB):	358.267
Skew:	-0.498	Prob(JB):	1.60e-78
Kurtosis:	3.764	Cond. No.	6.44e+17

MODELLING ITERATION 3

Following the removal of variables that had weak statistical significance on the price of a vehicle and high multicollinearity, I re-examined the OLS analysis.

FINDINGS:

1) The OLS suggests the p-values of all variables are acceptable.

2) The Adj. R-squared suggests there is acceptable goodness of fit.

3) Skewness had improved even more.

4) The coefficients suggest that engine power, years old and the vehicle brand have the biggest impact on the overall cost of a vehicle.

5) Ready to create a Predictive Model.

PREDICTIVE MODEL

```
new_row = pd.concat([new_row, pd.DataFrame({
                            'years_old': 6,
                            'fuel_efficiency': 19,
                            'engine_power': 1500,
                            'low_budget_brand': 1,
                            'high_budget_brand': 0,
                            'seats_0-4': 1,
                            'seats_5+': 0,
                            'fuel_Diesel': 1,
                            'fuel_Petrol': 0,
                            'transmission_Automatic': 1,
                            'transmission_Manual': 0
}, index=[0])])
```

Now, I can finally make my predictive model.

To generate a price estimate for a vehicle, I simply need to input its specific values into the predictive model. An example of this process is demonstrated on the left-hand side.

PREDICTIVE MODEL

By applying my predictive model, I can determine the value of a vehicle, as demonstrated below.

This information can be useful for assessing whether a vehicle is undervalued when purchasing, and for setting a selling price after purchase to maximize profits.

```
# prediction needs to be scaled and exponentiated
predicted_price = np.exp(new_row_pred_log)

predicted_price = int(predicted_price)
print("The predicted vehicle price is ${:.2f}".format(predicted_price))
```

```
The predicted vehicle price is $7457.00
```

PRESENTED BY

WARREN MORELLI

warren@momo-mktg.com
GitHub: @Warren-Morelli
LinkedIn:
https://www.linkedin.com/in/warren-morelli/