Geely Auto case study

Assumptions:

1) P-value < 0.05 and VIF < 5 are used for manual feature selection.
   Once the RFE filters out the insignificant independent variables, I used P-value and VIF for selecting the variables for the model. To do this, I considered variables with > 0.05 and/or variables with VIF > 5 for dropping from the model.

   However, this was not always the case. Based on some domain knowledge, I did decide to retain some variables in the model and investigate further. For example, while creating model_8, carlength had a high VIF value of 13.86. But it was also highly correlated with 'price' at 0.68. Dropping this variable only based on VIF will bring down the R squared and Adjusted R squared values significantly. So I ***checked the correlation matrix to see if carlength is highly correlated with any other variable and then dropped the carWLratio*** (width to length ratio).

2) P-value over VIF.
   While dropping the variables. a variable with higher P-value with low VIF takes precedence over a variable with lower P-value but high VIF. In model_4, I dropped 'hpengsizeratio' with P-value of 0.33 and VIF of 4.82 although there were other variables with VIF > 25 and but their P-value has < 0.05. I followed the same reasoning in model_6, where I ***dropped CarName_saab with P-value of 0.078 but VIF of 1.45.***

3) Impact of variable on price.
   Throughout the analysis, carlength and enginesize consistently had high correlation with 'price'. So even when their VIF values were high, I retained them and looked for variables that are correlated with them and have lesser relation with price. I dropped such variables from the model (example – model_8)

4) Correlation with one or more variables.

After using RFE to drop variables the first time, I noticed that the VIF values for 2 of the variables was infinity. This happens when they are highly correlated with each other. On using df['cylindernumber_two'].corr(df['enginetype_rotor'), it was found that their correlation is 1. So, I ***dropped enginetype_rotor and the VIF values were no longer infinity.***

5) Derive metrics with ratio for highly correlated variables.

During Exploratory Data Analysis, the heatmap revealed the variables that are highly correlated. When I dropped some of them, I noticed that the models had only car name related variables. So I guessed I was missing out on something important if I drop these variables. Instead I created ***derived metrics using ratios of these highly correlated variables.***

6) In case of ambiguity, use RFE.

In model_6, all variables had P-value = 0.00 and some had high VIF***.*** I was not sure how to proceed, so I ***used RFE second time to drop more variables*** and then continued with the analysis.

7) RMSE is not used for model evaluation.

Although I have noted the RMSE value for test data and training data, I am not using it as a parameter for model evaluation after Prof.Neelam mentioned in the live session that RMSE value is not a good indicator to evaluate the model.

Model evaluation metrics:

For the final model I built, the following are the evaluation metrics:

1) R-squared

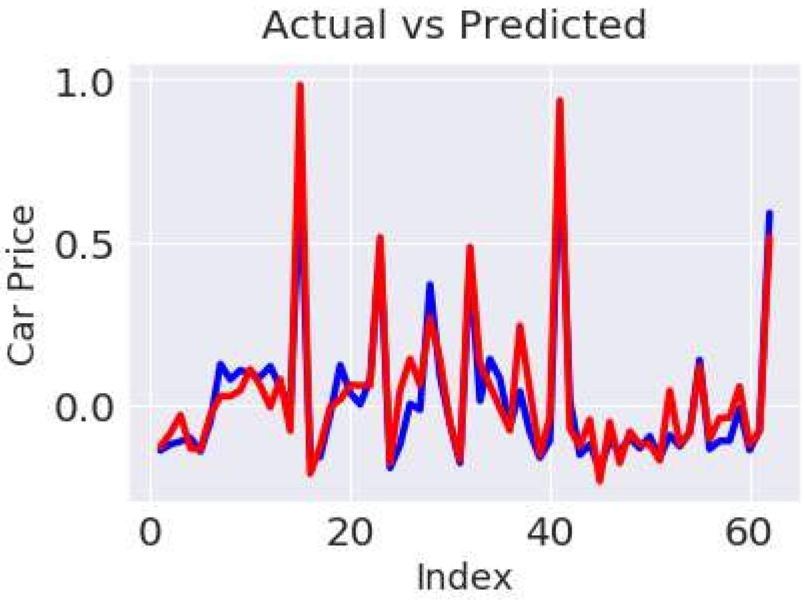
R squared = Explained variation/ Total variation

R squared = 1 – RSS/TSS

RSS = Residual sum of squares

TSS = Sum of errors of the data from the mean

R squared value determines the amount of variance that the model can explain. For my final model, this value is **<u>0.890</u>**. The higher the variance is accounted for by the model, closer are the data points to the fitted regression line. When I look at the plot with Actual car price Vs Predicted car price, the graphs match very closely.



Limitations of R squared value:

A high R squared value does not always guarantee a best fit line.

2) Adjusted R-squared

Adjusted R squared = 1 – (1- R squared) (N -1)/(N-p-1)

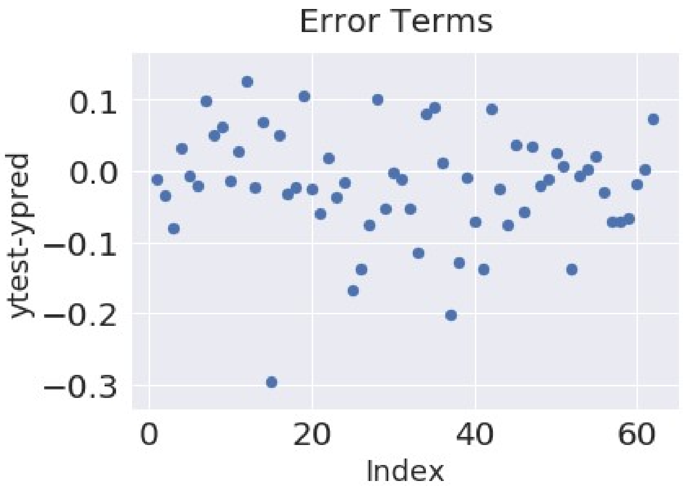
N = number of variables

p = Total sample size

For my final model, adjusted R squared value is **<u>0.883</u>** which is very close to R squared value. This value considers the number of variables in the model. Since I have 8 variables in my final model, the slightly lower value of the adjusted R squared can be attributed to that. However, if I drop more variables from the final model, the P-values go up and adjusted R squared value dips.

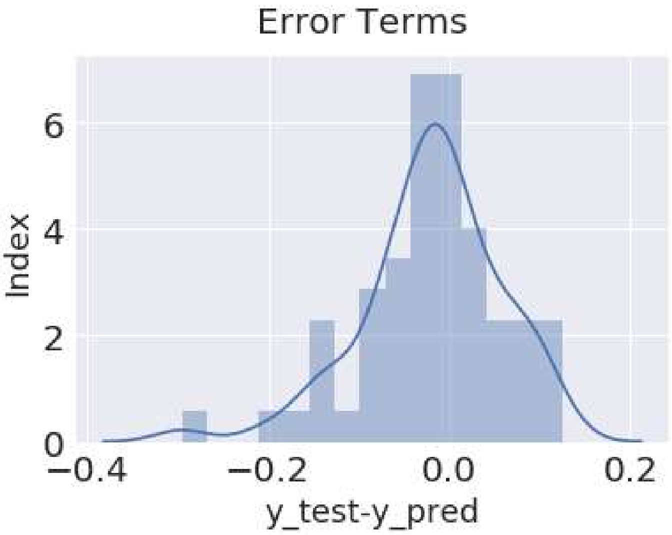
3) Residual plots and error terms

The error term for my final model is randomly distributed.



This means that the residuals are neither systematically high nor low. So, the residuals should be centered on zero throughout the range of fitted values. In other words, the model is correct on average for all fitted values.

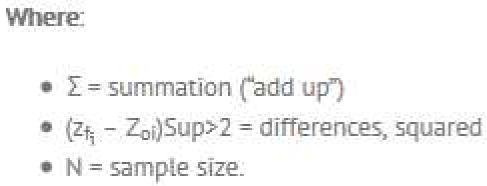
The residual plot for my final model is a normal distribution as shown:



Random errors are assumed to produce residuals that are normally distributed. Therefore, the residuals should fall in a symmetrical pattern and have a constant spread throughout the range. Therefore, the normal distribution of error term is correct.

4) RMSE

$$\mathrm{RMSE}_{fo} = [\sum_{i=1}^{N} (z_{f_i} - z_{o_i})^2 / N]^{1/2}$$



Where:

- Σ = summation ("add up")
- (z_{f_i} – Z_{oi})Sup>2 = differences, squared
- N = sample size.

I noticed that RMSE for test and training data match closely. The RMSE value shoots to more than 2000 when the 'price' (dependent variable) is not normalized. When it is normalized, the value comes down to less than 1.