Diagnostic Plot
Ref : http://data.library.virginia.edu/diagnostic-plots

Author : Pramod
Date : 17-06-2017

Residual vs Fitted plots.

This plot shows if residuals have non-linear patterns. There could be a non-linear relationship between predictor variables and an outcome variable and the pattern could show
up in this plot if the model doesn't capture the non-linear relationship. If you find equally spread residuals around a horizontal line without distinct patterns,
that is a good indication you don't have non-linear relationships.
Let's look at residual plots from a 'good' model and a 'bad' model.
The good model data are simulated in a way that meets the regression assumptions very well, while the bad model data are not.

Examples : s1_lm1 Res ~ Fit , s1_lm2 Res~ Fit and train_price Res~Fit

Normal Q-Q Plots.

This plot shows if residuals are normally distributed. Do residuals follow a straight line well or do they deviate severely?
It's good if residuals are lined well on the straight dashed line.
Examples : s1_lm1 QQ, s1_lm2QQ, train_priceQQ

Scale-Location.

It's also called Spread-Location plot. This plot shows if residuals are spread equally along the ranges of predictors. This is how you can check the assumption of equal variance (homoscedasticity). It's good if you see a horizontal line with equally (randomly) spread points.
Examples : s1_lm1 Scale~loca, s2_lm2 Scale~loca, train_price Scale~loca

# If the the residuals begin to spread wider along the x-axis as it passes around, its because of wide spread of the residual value.
the red smooth line will not be horizontal and shows a steep angle.

Cooks Distance.

#Gives out points which when not consider might have significant effect on the linear model or the relationship between the variables.

This plot helps us to find influential cases (i.e., subjects) if any.
Not all outliers are influential in linear regression analysis (whatever outliers mean).
Even though data have extreme values, they might not be influential to determine a regression line.
That means, the results wouldn't be much different if we either include or exclude them from analysis.
They follow the trend in the majority of cases and they don't really matter; they are not influential.
On the other hand, some cases could be very influential even if they look to be within a reasonable range of the values.
They could be extreme cases against a regression line and can alter the results if we exclude them from analysis.
Another way to put it is that they don't get along with the trend in the majority of the cases.
Unlike the other plots, this time patterns are not relevant. We watch out for outlying values at the upper right

corner or at the lower right corner.
Those spots are the places where cases can be influential against a regression line. Look for cases outside of a dashed line, Cook's distance.
When cases are outside of the Cook's distance (meaning they have high Cook's distance scores), the cases are influential to the regression results.
The regression results will be altered if we exclude those cases.

Example
s1_lm1 Cook
In this if we remove the points 1183 (at the top left) and 1300 at the bottom right, we can have a significant effect on the Rsquare value of the model, which will decrease, the
efficiency of it. Like wise in others too.

Conclusion.

Over all we can conclude that, our hypothesis that Price or Sale price is dependent on the variables and some of them have a positive cor and some negative and
some have no correlation, I have created 3 prediction models

Model train_price :
Consider all the variables, which gave good prediction results on the test data set

Model price_pred2 :
Consider all positive correlated and variables which have less that 0.5 but more than 0 , which also yielded good results with residuals falling under normal dist
their QQ plots s1_lm2 had slight tailing which can be controlled by taking log values.

Model price-pred1 :
Consider only positive correlated var and also which have significant effect on the sale price,
This yielded the best, with high accuracy, mean value close to the given data and better plots.