ASSIGNMENT-3

AKSHAY SRIRANGAM (AXS156630)

VISHNU KUMAR REDDY CHINAREDDY(VXC151130)

ANUPAMA VIJAYAN (AXV146030)

1. Predicting Car Trade In Value:

Conclusions Drawn:

* A linear model is built initially with all the features included.
* The MSE for the model is observed to be **6083.794 (with 12 attributes).**
* The not so important features are removed on the basis of R2 and RSE values (Residual standard error).
* The method of backward selection method is used for filtering the unimportant features.
* The MSE after taking out some features is observed to be **6084.498 (with 6 attributes)**
* So, the removal of the features namely Cylinder, Liter, Cruise, Type, Doors, Make is justified as the MSE is almost similar even after removing the above said features.
* The summary function gives other observations and can be seen by running the code.
* VIF did not work as there was collinearity between the predictors.
* A final residual plot is also plotted which shows the probability of a non-linear function.
* There was no effect found even on using the cross terms.
* The outliers found after using the studentized residuals are the following data points:

a)105 b)446 c)359 d)644

* The outliers found again after removing the above data points:

a)20 (although this is an outlier according to the plot, this point cannot be excluded)

* After checking the leverage using the hat values, we observe the following data points:

a)435 (by removing this point, might lead to overfitting as the MSE raises from 6083 to 8000 )

2.Predicting Carbon Monoxide Levels in Cigarettes:

* Brand is not a predictor, it’s an identifier as its impact on building the model is zero.
* The data is split into 75% training and 25% test data.
* The MSE is observed to be 0.3799.
* The attribute “weight” can be dropped from table as it has collinearity and provides no real improvement towards R2 and RSS.
* The final set of predictors are tar and nicotine [even nicotine can be removed since there is collinearity between tar and nicotine because R2 > 0.8].
* The outliers are the following:

a)13 [but, excluding this data point from the set, will lead to an over-fitted model]

* Thus, there are hardly any outliers which can be observed from the studentized residual plot.

3.Predicting Low Birth Weight**:**

* Initially, the data is split into 75% training and 25% test data.
* Applying logistic function and using the binomial family, the accuracy was found to be 1.00
* After checking for the collinearity, the following attributes were found essential:

1. Id and PTL as predictors
2. LOW as a response variable.

* All other predictors are found to be less important.
* After removing the other attributes, the accuracy is found to be the same as earlier.