Jacob Perrone

Graded Assignment 1 Report

\* on the excel workbook the LInRegOutput6 has the best model and the tab is colored yellow.

\*\* the tab top5 and 95% variance correspond to the tables containing the summary statistics of the models for the respective PCA method.

1. Data handling/ Missing data/ 11a

Going through each column of predictors and sorting in descending order to find invalid or missing cells. There were none. Then I created dummy values for the features fuel\_type and color. Once that was done the following screenshots are of the scatterplot matrices. The first is using the first five predictors and the second uses predictors I thought could have lead to dimension reduction as they are highly correlated. In the second screenshot, those variables seem correlated but PCA did not have any of them selected as a top predictor (up to top 15).

Graphical user interface, diagram

Description automatically generated

Calendar

Description automatically generated

11b.

The top 5 predictors were:

|  |
| --- |
| Age\_08\_04 |
| fuel\_type\_petrol |
| backseat divider |
| mistlamps |
| radio |

These 5 covered 36% of the variance.

Graphical user interface, table

Description automatically generated

The top 15 predictors were:

|  |
| --- |
| Age\_08\_04 |
| fuel\_type\_petrol |
| backseat divider |
| mistlamps |
| radio |
| gears |
| color red |
| color blue |
| color grey |
| color green |
| color black |
| color white |
| color violet |

\* two of them were used for multiple components

The top 15 covered 64% of the variance.

Table

Description automatically generated

2. Partitioning and PCA

For PCA, I first ran using the top 5 predictors which ended up explaining 36% of the variance. I highlighted the weights in the PCAOutput tab and recorded the components to use for partitioning. I used partitioning with 80/20, 75/25, 70/30 splits and recorded the R2 and RMSE values (in file “Tables for graded assignment 1” as to avoid clutter in the main workbook). Once the results were tabulated I then went back to the encoding tab and ran PCA for 95% of the variance which gave 33 predictors. Since that was a lot I took the top 15 as instructed in the assignment instruction document. As before I highlighted the components weights and recorded the top 15 to pass to partitioning. The model 4 in this table had an extra predictor (doors) which led to better validation R2. As a note, when attempting to run PCA with the cylinders predictor it would not allow me to add it, upon looking at the column all the values were the same so I decided to omit that predictor from PCA.

Top 5 table:

Table

Description automatically generated

95% variance table:

Graphical user interface, application, table, Excel

Description automatically generated

3. Scoring

For choosing the best model out of the 6, first I looked at if the R2 value for both training and validation were > 70% (they all were) and if the two R2 values were close to each other. Then the model with the highest validation R2 value was chosen as the best model for each (top 5 and 95% variation). Then out of the two best I again chose the one with the highest R2 to deploy on the new data and predict the price (model 4). Below is a screenshot of the predicted values on the new data, which I believe to be accurate as the predicted values fall in the range of the prices of the original data.

Graphical user interface, text, application

Description automatically generated