**ITCS 6100: HW2 Answers - Group 21**

**Answer 1**:

1. Training set: This is the sample of the data (the subset of the whole dataset) which is used to fit our model. This is the part of the dataset we use to train the model. This essentially means that the model learns from this data. The training datasets consist of pairs of input vectors and corresponding output vector which is called as a label.
2. Validation set: Validation set is
3. a subset of the dataset which evaluates fitting of the model by hyperparameter tuning. This avoids overfitting or underfitting of the data. As the values are sampled for every set, the validation set provides an unbiased or fair judgment about the fit of the model. This means that the model does not actually learn from validation data but helps judge the model to tune the hyperparameters/weights/biases.
4. Test set: Test data is a subset of the data which is used to get a final evaluation of the model. Test data is only used when the whole model is completely trained.

The partition is being performed and is present in the Excel sheet under STDPartion name.

**Answer 2:**

Consider there are 3 dummy variables. We can select 2 out of those 3 variables to build a regression model. This is because if 1 and 2 are classified as negative, the third has to positive, considering that all of them are mutually exclusive and exhaustive events. This means that either of them taking place will block the occurrence of any other and one of the must occur. So if we have N dummy variables, we can consider the maximum. (N-1) variables to perform regression.

**Answer 3:**

Multiple Linear Regression is the type of regression which is used to map the relationship between a continuous and one or more than one independent variables. A pattern or prediction is made based on this relationship between the continuous and independent variables.

Refer **LineReg\_Output** sheet and the links are present for all training, validation and test data set

After applying the Multiple Linear Regression model, we got below equation.

In this, we have three dummy variables for Fuel Type, but during the model generation, we will consider only two as the third variable can be from the first two, reason of this given in Answer 2.

Linear Regression Expression with all features:

y(Price)= 9658.29213 + (-106.7156099\*Age\_08\_04) + (-0.020894995\*KM) + (37.47969865\*HP) + (438.6578275\*Automatic) + (104.5171611\*Doors) + (15.76205668\*Quarterly\_Tax) + (194.6413277\*Mfr\_Guarantee) + (63.85650868\*Guarantee\_Peroid) + (133.6504502\*Airco) + (3049.588727\*Automatic\_airco) + (233.8708788\*CD\_Player) + (395.0633454\*Powered\_Windows) + (415.7743878\*Sport\_Model) + (-268.148481\*Tow\_Bar) + (2511.749109\*Fuel\_Type\_Diesel) + (1977.800463\*Fuel\_Type\_Petrol)

Training Score:

**Refer LineReg\_TrainingScore**

|  |  |
| --- | --- |
| **Metric** | **Value** |
| SSE | 1.07E+09 |
| MSE | 1486720 |
| RMSE | 1219.311 |
| MAD | 909.2012 |
| R2 | 0.89009 |

We can see a good R2 (R-Squared) score, which means that the model fits good for the given data at given parameters. As far as RMSE is considered, we will compare it further as we see validation and testing scores.

Validation Score:

Refer **LineReg\_ValidationScore**

|  |  |
| --- | --- |
| **Metric** | **Value** |
| SSE | 638881325.5 |
| MSE | 1482323.261 |
| RMSE | 1217.506986 |
| MAD | 918.7466752 |
| R2 | 0.870954665 |

The R2 value seems to have decreased for the validation which could indicate that the model is generalizing either too well (underfitting), or is too rigid (overfitting). Also, RMSE has decreased slightly which could indicate towards underfitting of the data, meaning the noise in the data is not being captured as much as needed. But again, since the shift is very low we cannot infer for sure whether underfitting has really occurred.

Testing Score:

Refer **LineReg\_TestScore**

|  |  |
| --- | --- |
| **Metric** | **Value** |
| SSE | 419812806.2 |
| MSE | 1462762.391 |
| RMSE | 1209.447143 |
| MAD | 934.5427253 |
| R2 | 0.899804479 |

Here, the R2 value has increased and is even more than that of the training set. So, we can say that the model is fitting well on the test data. Here we can also see that the RMSE is lower than that of the training set. This indicates a slight underfitting of the data but not so much to invalidate the goodness of the model

The below is the final output and as we can see from below the value of R2 is pretty much good this shows the predictors we select for this are correct choice and model is performing well.

**Answer 4**

Regression output after running Stepwise Regression**.**

After stepwise regression is executed we got 15 subsets and out of these 15 subsets we have selected, Subset as our best model and generated the output. (Refer LineReg\_FS1 sheet for all the subsets).

If we see subset 15 we can observe that the two variables which are less significant and not adding much value to mode.

The Predicator dropped are “CD\_Player” and “Airco”

The value of R2 we got is higher which means the model is performing well and the choice of predictors, therefore, is correct.

**Linear Regression Expression with stepwise regression feature:**

y(Price)= 9900.030816 + (-110.7816853 \* Age\_08\_04) + (-0.020653045 \* KM) + (37.67943417 \* HP) + (425.2329172 \* Automatic) + (107.672169 \* Doors) + (15.67312131 \* Quarterly\_Tax) + (204.2092266 \* Mfr\_Guarantee) + (60.34685136 \* Guarantee\_Period) + (3076.800973 \* Automatic\_airco) + (458.3169486 \* Powered\_Windows) + (416.9424331 \* Sport\_Model) + (-264.8019505 \* Tow\_Bar) + (2550.23462 \* Fuel\_Type\_Diesel) + (2013.947078 \* Fuel\_Type\_Petrol)

**Refer LinReg\_Ouput2**

**Training Score:**

**Refer LineReg\_TrainingScore2**

|  |  |
| --- | --- |
| **Metric** | **Value** |
| **SSE** | **1074579145** |
| **MSE** | **1496628.336** |
| **RMSE** | **1223.367621** |
| **MAD** | **915.8753848** |
| **R2** | **0.889357716** |

For the training data, we have achieved a good R2 score which means that the model learned well on the test set. We can safely assume that the model can also predict well as we have a good R2 value. Although, we cannot judge the goodness of the model solely based upon the Test R2 as RMSE and R2 of validation and test set also play a crucial part in it.

**Validation Score**

**Refer LineReg\_ValidationScore2**

|  |  |
| --- | --- |
| **Metric** | **Value** |
| **SSE** | **653566962.7** |
| **MSE** | **1516396.665** |
| **RMSE** | **1231.420588** |
| **MAD** | **927.9715874** |
| **R2** | **0.867988366** |

As expected, the R2 has decreased indicating that the model is adapting to the new values. The RMSE has increased due to the addition of new data points indicating that it is trying to closely fit the new values of validation set. Could indicate slight overfitting of data.

**Testing Score**

**LineReg\_TestScore2**

|  |  |
| --- | --- |
| **Metric** | **Value** |
| **SSE** | **421513636.4** |
| **MSE** | **1468688.628** |
| **RMSE** | **1211.894644** |
| **MAD** | **939.8237343** |
| **R2** | **0.899398547** |

We have a great R2 value for the test set which is indicated that the model fit well on unknown data. This indicates that the model is indeed a good predictor. Here, the RMSE test is less than RMSE train but not by a very significant value. It could indicate underfitting, but the shift is not large enough. We can say that the model is a good fit.

**Refer LineReg\_ LinReg\_Ouput2**

|  |  |
| --- | --- |
| **Metric** | **Value** |
| **Residual DF** | **703** |
| **R2** | **0.889357716** |
| **Adjusted R2** | **0.887154314** |
| **Std. Error Estimate** | **1236.350308** |
| **RSS** | **1074579145** |

We can conclude that having a good R2 and Adjusted R2 scores indicate that the model predicts well.

**Answer 5**

In the regression model which is generated using exhaustive search and stepwise are below

**Stepwise**

|  |  |
| --- | --- |
| **Metric** | **Value** |
| **Residual DF** | **703** |
| **R2** | **0.889357716** |
| **Adjusted R2** | **0.887154314** |
| **Std. Error Estimate** | **1236.350308** |
| **RSS** | **1074579145** |

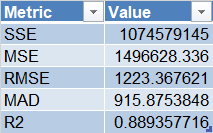
**Exhaustive**

|  |  |
| --- | --- |
| **Metric** | **Value** |
| **Residual DF** | **701** |
| **R2** | **0.890090244** |
| **Adjusted R2** | **0.887581604** |
| **Std. Error Estimate** | **1234.007365** |
| **RSS** | **1067464698** |

From above comparisons we can see both models have fair value of R2 and both are acceptable. As we can see there is a slight difference in the value of R2 in both the models, but in Stepwise model has stopped at subset 15 and removed the less significant predictors which were not adding much value in the model.

Below are the details of each training, validation, test data for model 3 and 4

**Training score of model 4**



**Training Score of model 3**

|  |  |
| --- | --- |
| **Metric** | **Value** |
| SSE | 1067464698 |
| MSE | 1486719.635 |
| RMSE | 1219.311131 |
| MAD | 909.2011524 |
| R2 | 0.890090244 |

**Validation Score of stepwise**

|  |  |
| --- | --- |
| **Metric** | **Value** |
| SSE | 653566962.7 |
| MSE | 1516396.665 |
| RMSE | 1231.420588 |
| MAD | 927.9715874 |
| R2 | 0.867988366 |

**Validation Score of Exhaustive**

|  |  |
| --- | --- |
| **Metric** | **Value** |
| SSE | 638881325.5 |
| MSE | 1482323.261 |
| RMSE | 1217.506986 |
| MAD | 918.7466752 |
| R2 | 0.870954665 |

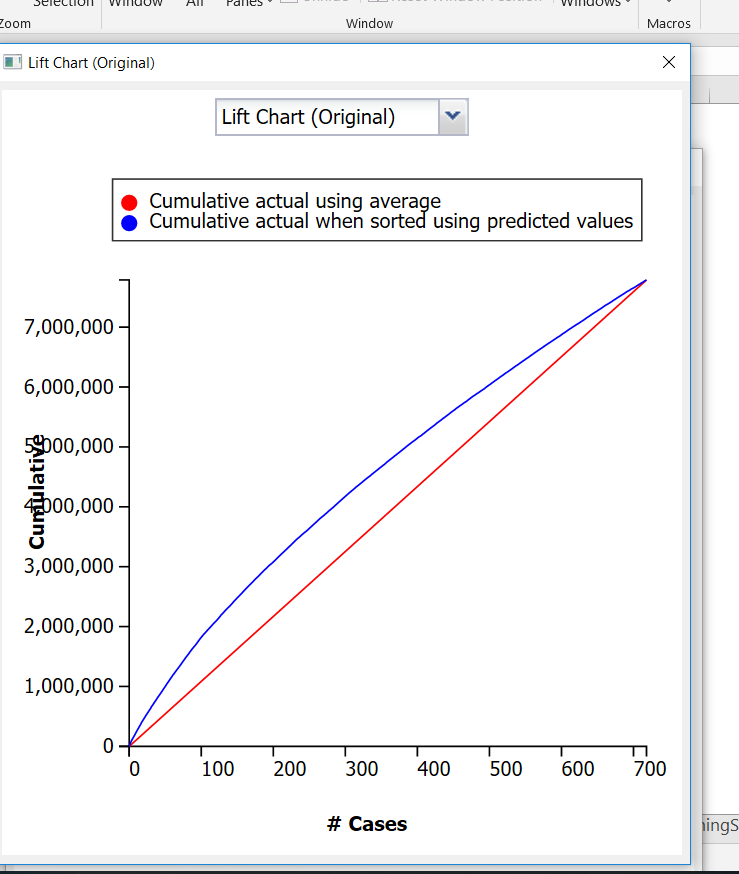
**Testing Score of Exhaustive**

|  |  |
| --- | --- |
| **Metric** | **Value** |
| SSE | 419812806.2 |
| MSE | 1462762.391 |
| RMSE | 1209.447143 |
| MAD | 934.5427253 |
| R2 | 0.899804479 |

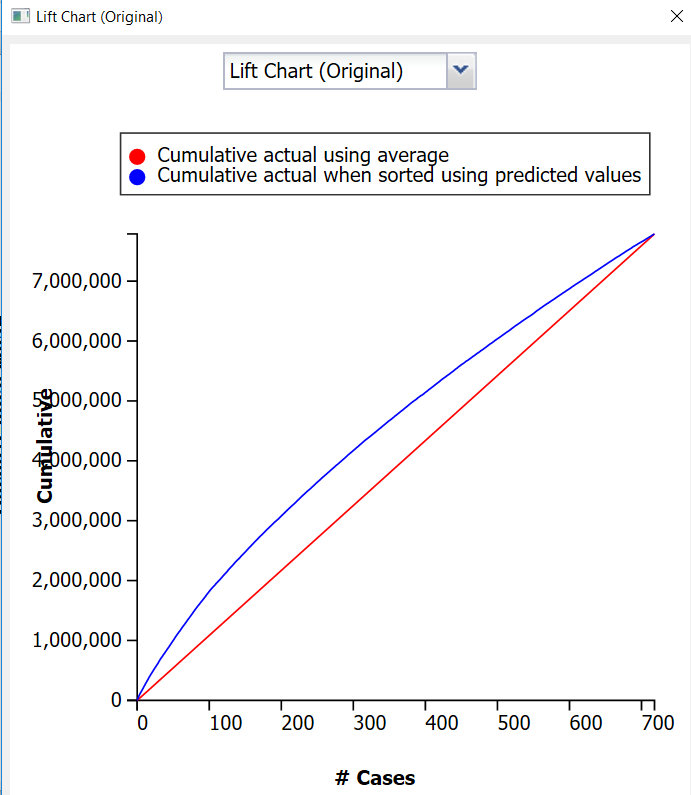
**Testing Score of Stepwise**

|  |  |
| --- | --- |
| **Metric** | **Value** |
| SSE | 421513636.4 |
| MSE | 1468688.628 |
| RMSE | 1211.894644 |
| MAD | 939.8237343 |
| R2 | 0.899398547 |

**Exhaustive Training Lift Chart**

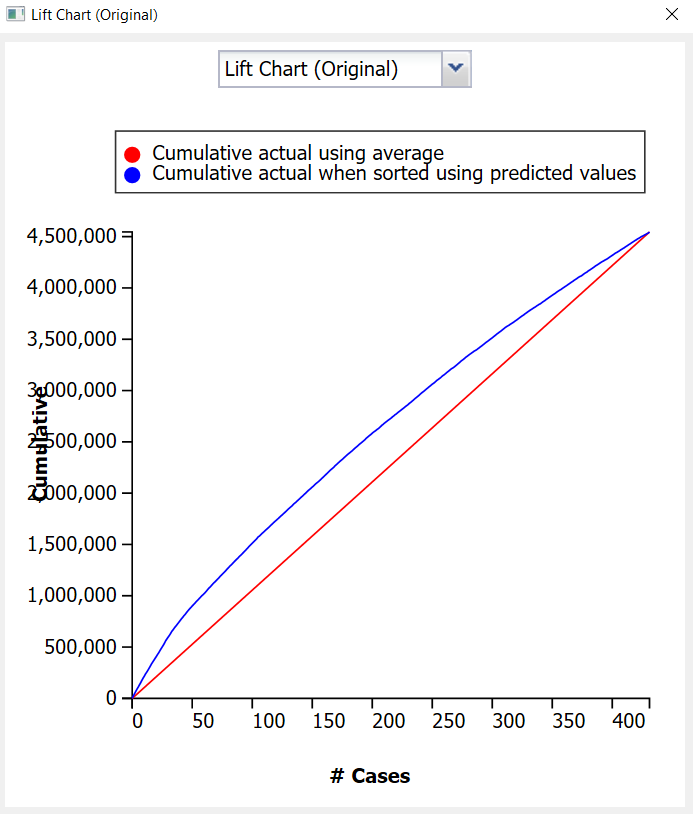


**Stepwise Training Lift Chart**

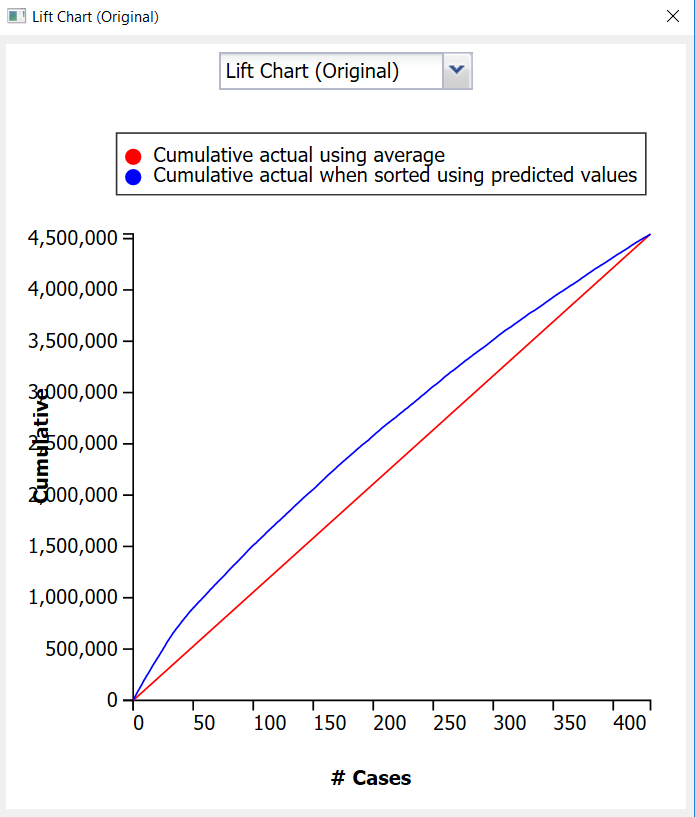


By comparing the lift chart of exhaustive training and stepwise training, we observe that there is no significant difference between the charts as the number of cases used and the area between the curve and baseline is approximately same. For a better model, the area between curve and baseline should greater, greater the area better the model.

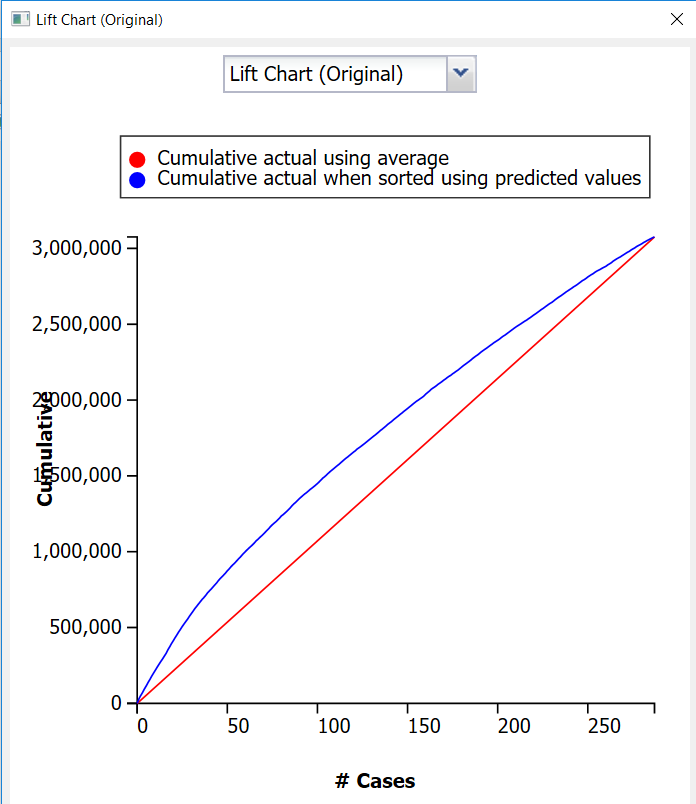
**Exhaustive Validation Lift Chart**

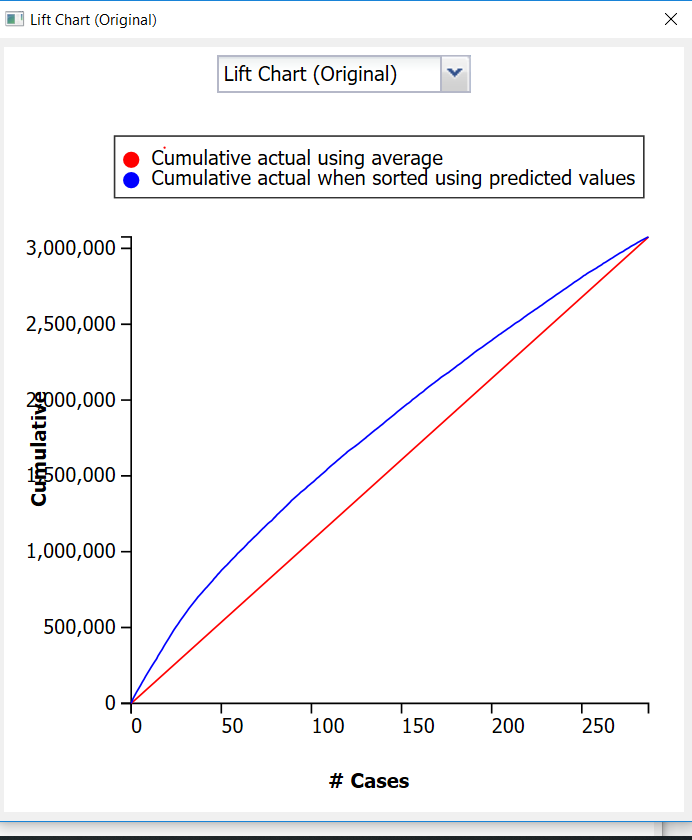


**Stepwise Validation Lift Chart**

****

By comparing the lift chart of exhaustive validation and stepwise validation, we observe that there is no significant difference between the charts as the number of cases used and the area between the curve and baseline is approximately same. For a better model, the area between curve and baseline should greater, greater the area better the model.

**Exhaustive Test Lift ChartA stepwise Testing Lift chart**



By comparing the lift chart of exhaustive testing and stepwise testing, we observe that there is no significant difference between the charts as the number of cases used and the area between the curve and baseline is approximately same. For a better model, the area between curve and baseline should greater, greater the area better the model.

**Answer 6:**

Referring to the sheet LinReg\_FS1, we can notice that the R2 value starts stabilizing after Subset 6 which means that the first five variables added become the most significant ones.

We can also justify our answer by looking p values and the same answer is coming.

**Age** is the most significant feature as it significantly shoots the R2 value from 0 to around 76.

**Automatic\_airco** comes next as it also boosts R2 value.

**KM** is the next feature selected.

**Quarterly tax** is the next feature as it also adds to the increment of the R2 value.

**HP** is the last feature among the top 5

**Answer 7**

If we put the values in Answer 2’s equation the predicted price value we get is below

Predicted Price value is

**18671.22189**