CSE602 Machine Learning – 1

Homework # 1

Issue Date: January 25, 2022

Deadline (via LMS): January 30, 2022 (11:55 PM)

Submitted by Muhammad Amin Ghias

ERP ID : 25366

A file *ToyotaCorolla.xls* has been uploaded on the LMS (Datasets section) that contains data of used cars (Toyota Corolla) on sale during 2004 in the Netherlands. It has 1436 records and contains attributes like Price, Age, Kilometers, HP, and other specifications. The goal is to predict the price of a used Toyota Corolla based on its specifications.

1. Run a multiple linear regression model using (a) the complete data set and (b) after train-test split (70-30%) and report the R2 and adjusted R2. Please note that you don't need to do average when running on full data as the answer would be similar in each run. However, when you split, it is important that you do the experiment at least 5 times and report the average value.

Data	Average R2 (Average of 5 runs)	Average of Adjusted R2 (Average of 5 runs)
Full Data (without splitting)	0.908355	0.906131
Train Data (after splitting)	0.912223	0.910092
Test Data (after splitting)	0.888785	0.886086

2. Using statsmodel API, identify the variables that have p-values higher than 0.05. Report those variables and their p-values. [Note: You can use the full data without splitting]

S.No	Features	P_Values
1	Met_Color	0.56249
2	cc	0.18775
3	ABS	0.13051
4	Airbag_1	0.25558
5	Airbag_2	0.74108
6	Central_Lock	0.29529
7	Power_Steering	0.19132

8	Radio	0.2923
9	Mistlamps	0.92071
10	Backseat_Divider	0.45047
11	Metallic_Rim	0.13437
12	Radio_cassette	0.2713

3. **After discarding the variables identified above, run a multiple linear regression model using (a) the complete data set and (b) after train-test split (70-30%) and report the R2 and adjusted R2.**

Data	Average R2 (Average of 5 runs)	Average of Adjusted R2 (Average of 5 runs)
Full Data (without splitting)	0.907394	0.905953
Train Data (after splitting)	0.907472	0.906032
Test Data (after splitting)	0.902969	0.901458

4. **Compare the results of 1 and 3 and write your observations/findings:**

After comparing the results of part 1 and 3 it shows that after discarding variables with higher p-values the R2 value on test data after splitting have become much better where is there is negligible affect on R2 values on other cases. Hence discarding the variables seems to be beneficial

5. **Keeping the dataset of Step 4, apply Serial FeatureSelection routine in both backward and forward direction for 8 and 4 variables. Run multiple linear regression model in each case using (a) the complete data set and (b) after train-test split (70-30%) and report the R2 and adjusted R2.**

# of Variables	Direction	Data	Average R2 (Average of 5 runs)	Average of Adjusted R2 (Average of 5 runs)
8	Forward	Full Data	0.884173	0.883523
		Train Data	0.882468	0.881809
		Test Data	0.884867	0.884221
8	Backward	Full Data	0.870332	0.869605
		Train Data	0.872439	0.871724
		Test Data	0.861923	0.861149
4	Forward	Full Data	0.861914	0.861528
		Train Data	0.866296	0.865922
		Test Data	0.84975	0.84933
4	Backward	Full Data	0.857154	0.856754
		Train Data	0.856109	0.855707
		Test Data	0.85785	0.857453

6. What are your observations when you compare the results of 5 with that of 1 and 3?

On comparing results of part 1, 3 and 5 it can be seen that R2 values of all full data test data and train data is best all above (0.90) for part 3 with features removed on basis of high p-value. Using Feature selection with 8 features forward selection gives better result of R2 values around (0.88) while 4 features results are less than (0.87). This shows so far part 3 seems better.

7. Run a kNN regression model using (a) the complete data set and (b) after train-test split (70-30%) and report the R2 and adjusted R2. Try k=3 and k=5. Use the same set of variables as used in Step 3 above.

K=	Data	Average R2 (Average of 5 runs)	Average of Adjusted R2 (Average of 5 runs)
3	Full Data	0.853746	0.850197
	Train Data	0.834609	0.830596
	Test Data	0.636641	0.627823
5	Full Data	0.789026	0.783906
	Train Data	0.769269	0.76367
	Test Data	0.620309	0.611095

8. What are your observations/findings after this experiment?

For this part we have used all the features and the results obtained in this part are surprisingly bad good with R2 value of test around (0.62) for both k = 3,5.

9. Based on the complete experiment, which particular model (along with set of variables) you would suggest for deployment and why?

Based on the complete experience of part 1, 3, 5, 7 we can conclude that part 3 yields the best results and discarding values based on higher p=value for our given dataset and business problem. Part 3 has the highest accuracy

A summary of results shown in table below:

		Train Data		Test Data		Full Data	
	Case	**Average R2_Train**	**Adj_r2_train_avg**	**Average R2_Test**	**Adj_r2_test_avg**	**R2**	**Adj_r2**
0	**All features**	0.912223	0.910092	0.888785	0.886086	0.908355	0.906131
1	**Removed feature with P > 0.05**	0.907472	0.906032	0.902969	0.901458	0.907394	0.905953
2	**Serial feature selection forward 8 features**	0.882468	0.881809	0.884867	0.884221	0.884173	0.883523
3	**Serial feature selection backward 8 features**	0.872439	0.871724	0.861923	0.861149	0.870332	0.869605
4	**Serial feature selection Forward 4 features**	0.866296	0.865922	0.84975	0.84933	0.861914	0.861528
5	**Serial feature selection backward 4 features**	0.856109	0.855707	0.85785	0.857453	0.857154	0.856754
6	**knn k =3**	0.834609	0.830596	0.636641	0.627823	0.853746	0.850197
7	**knn k =5**	0.769269	0.76367	0.620309	0.611095	0.789026	0.783906

This table also concludes that part 3 has best results and more accuracy of results and predictions.