ASSIGNMENT 2

STQD6024: MACHINE LEARNING

SEMESTER 2 SESSION 2022/2023

NAME: NUR MARDHIAH BT. ZULKHARI

Question : Find out the best MLR model to predict the UPDRS scores based on the dataset.

The objectives of this case study are to identify and predict the target variable based on a few sets of preditors/independent variables by the best MLR model in the dataset of UPDRS scores. This dataset contains 20 persons with Parkinson's (6 female, 14 male) and 20 healthy individuals (10 female, 10 male) with information on their voice recording and clinical background.

For this case study, the method that being applied to get the best models in MLR which is the forward selection by using cross-validation with the k of 10. The dataset is being divided into two set which are training set and testing set. The training set will hold a bit larger portion of the data than the testing set. From the training set is being trained and evaluated using the testing which this is used to generate the predictions based on the training set. These prediction will be compared to the actual target values which can access the model performance. The model selection process involved some steps which are data preprocessing that included checking on the null value and renaming the columns name with the suitable features listed, selecting the best subset of features which in this case is implementing the forward selection which the data is split into train and test data, fit the MLR model with the number of features, evaluate each model's performance using the evaluation metrics.

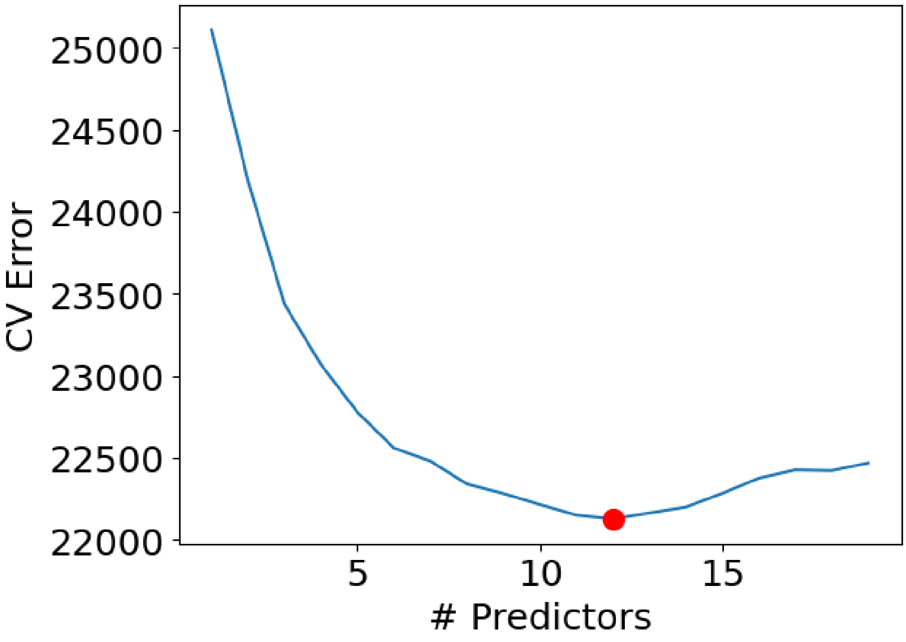
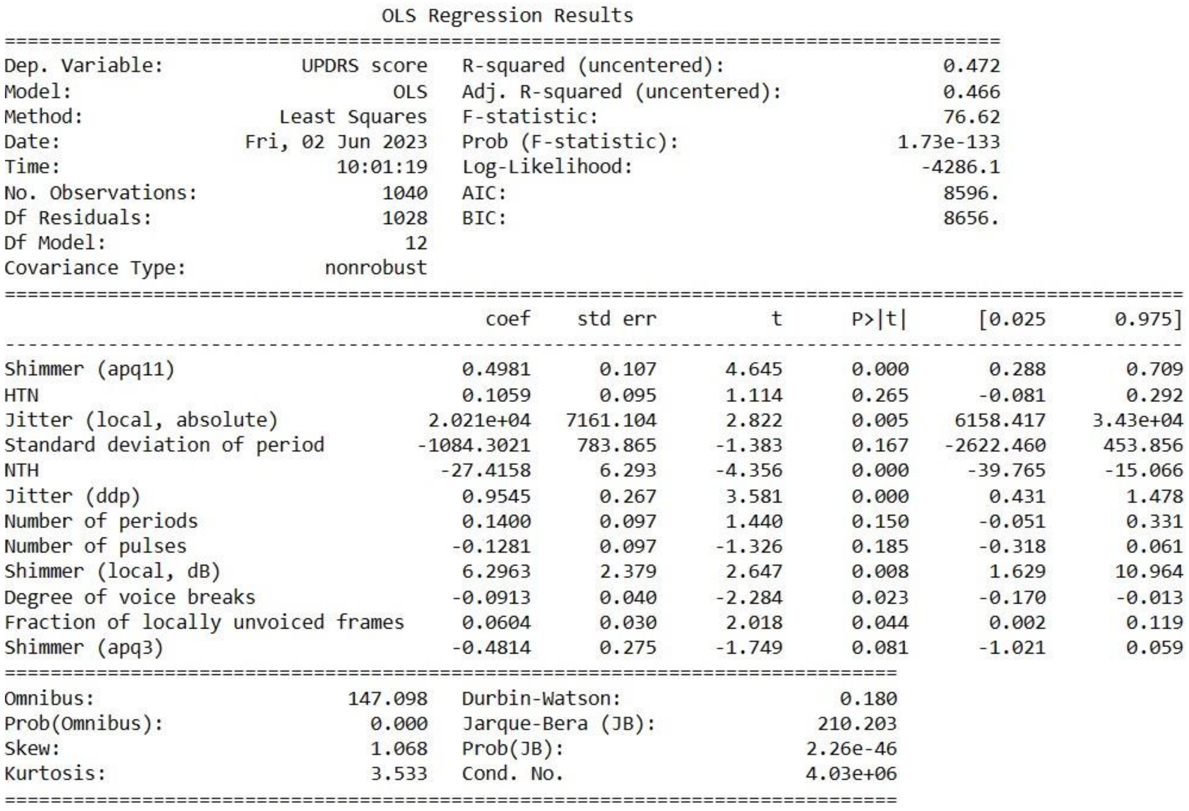
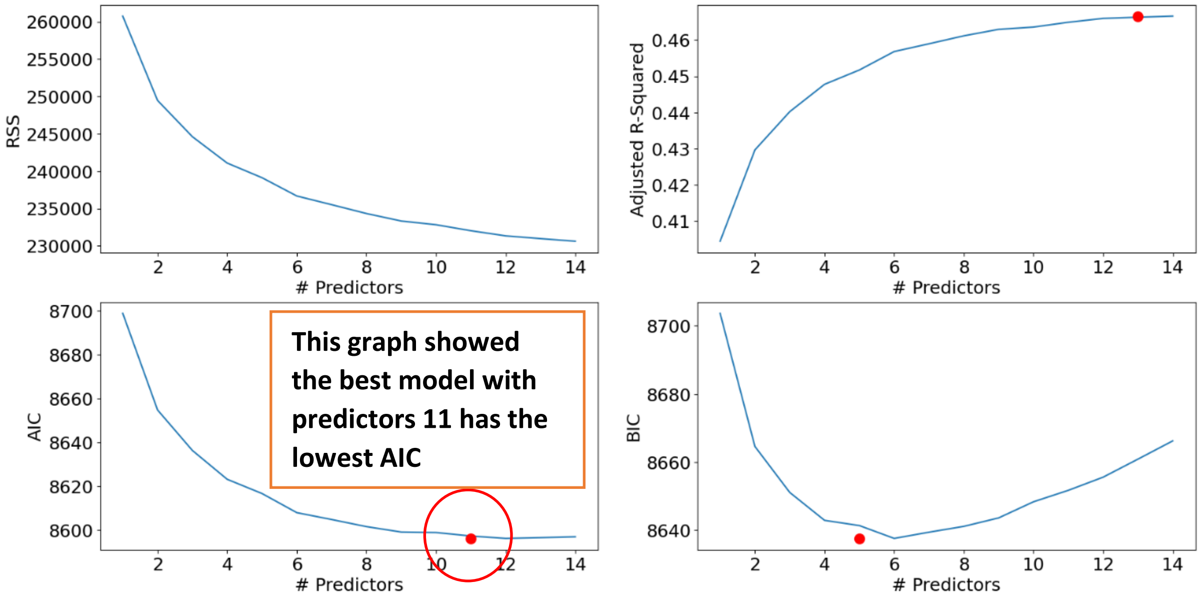


Figure 1 : Cross Validation by using Forward



```
                           OLS Regression Results
=======================================================================================
Dep. Variable:            UPDRS score   R-squared (uncentered):                   0.472
Model:                            OLS   Adj. R-squared (uncentered):              0.466
Method:                 Least Squares   F-statistic:                              76.62
Date:                Fri, 02 Jun 2023   Prob (F-statistic):                   1.73e-133
Time:                        10:01:19   Log-Likelihood:                         -4286.1
No. Observations:                1040   AIC:                                      8596.
Df Residuals:                    1028   BIC:                                      8656.
Df Model:                          12
Covariance Type:            nonrobust
=======================================================================================================
                                         coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------------
Shimmer (apq11)                        0.4981      0.107      4.645      0.000       0.288       0.709
HTN                                    0.1059      0.095      1.114      0.265      -0.081       0.292
Jitter (local, absolute)            2.021e+04   7161.104      2.822      0.005    6158.417    3.43e+04
Standard deviation of period       -1084.3021    783.865     -1.383      0.167   -2622.460     453.856
NTH                                  -27.4158      6.293     -4.356      0.000     -39.765     -15.066
Jitter (ddp)                           0.9545      0.267      3.581      0.000       0.431       1.478
Number of periods                      0.1400      0.097      1.440      0.150      -0.051       0.331
Number of pulses                      -0.1281      0.097     -1.326      0.185      -0.318       0.061
Shimmer (local, dB)                    6.2963      2.379      2.647      0.008       1.629      10.964
Degree of voice breaks                -0.0913      0.040     -2.284      0.023      -0.170      -0.013
Fraction of locally unvoiced frames    0.0604      0.030      2.018      0.044       0.002       0.119
Shimmer (apq3)                        -0.4814      0.275     -1.749      0.081      -1.021       0.059
==============================================================================
Omnibus:                      147.098   Durbin-Watson:                   0.180
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              210.203
Skew:                           1.068   Prob(JB):                     2.26e-46
Kurtosis:                       3.533   Cond. No.                     4.03e+06
==============================================================================
```

Figure 2 : OLS Regression Results : Forward selection



This graph showed the best model with predictors 11 has the lowest AIC

Figure 3 : Overall graph on AIC, BIC, R-Squared and RSS

By applying the Cross validation approached in the full model with the number of folds is 10, in **Figures 1 and 2**, it showed the summarized graph by using Forward Selection with the result of AIC, BIC, Adjusted R-Squared. This showed that the best model based on the criteria is predictors 11 with AIC of 8596, BIC of 8656 and Adjusted R-Squared is 46.6%. Based on **Figure 3**, the figure showed that respective curves where the preditor is 11 is the best model due to AIC is is showed the lowest at 11. However, the RSS and R-Squared showed the best model is at predictors between 13-14, and BIC showed that model with 5 predictors is the best due to the lowest BIC.

In conclusion, based on the graph and OLS summary above, we can be conclued that the best MLR model is with a preditor of 11 as the best model overall. This model can help to identify the most relevant features and most reliable estimates of UPDRS scores for each individual with Parkison’s disease.