First Home Assignment - Parkinsons Telemonitoring Data Set

Group 20:

- João Santos – nº 55789 – 18h

- Rodrigo Santos – nº 60552 – 18h

- Tomás Santos – nº 43566 – 18h

Introduction

For this assignment, we worked on a dataset composed of a range of biomedical voice measurements from 42 people with early-stage Parkinson's disease recruited to a six-month trial of a telemonitoring device for remote symptom progression monitoring.

Using this dataset, we created several linear and tree machine learning models with two objectives in mind. The first one was to create several and choose the best regression model that predicted the Motor Unified Parkinson Disease Rating Scale (UPDRS) based on given features. The second was to create several and choose the best binary classification model that would classify as positive all instances with a total UPRDS above 40 and as negative all instances equal or below 40.

Data Pre-Processing

After visually analyzing the data, we noticed that the data entries were grouped by person, and to avoid unnecessary correlation, the data was shuffled.

After that, and since we didn't have the possibility to validate the models with real world data, we split and saved 20% of the data as an Independent Validation Set (IVS).

The final dataset used for training and validating the models had 4700 data entries.

For the validation of each model, we choose to do a k-Fold cross validation, with k=5.

Objective 1 - Produce the best regression model for 'motor UPDRS'.

For this objective we created models using regression decision tree, multiple linear regression, ridge regression, least absolute shrinkage and selection operator (Lasso) regression and elastic nets.

After creating the models, the hyperparameters were varied in search of the best set that would give us a good model without overfitting it.

For the linear models we did a variable selection based on the value coefficient and the p-value. We selected the variables that had a large coefficient and at the same time a low p-value. There were some exceptions with some cases that we tried to remove the variables and the model improved, even though they had large coefficient and low p-value, and variables that had high coefficient and high p-value that improved the model.

We summarized the regression metrics used for each model in a table:

Motor_UPDRS					
	Ridge Model	**Lasso Model**	**Elastic Net**	**Linear Regression**	**Decision Tree Regressor**
RVE	0.1468	0.0946	0.0956	0.1468	0.6538
RMSE	7.4612	7.6861	7.6818	7.4612	4.7531
Correlation Score	0.3838	0.3076	0.3092	0.3838	0.8086
Maximum Error	39.7555	18.8383	18.8744	39.7555	18.4239
Mean Absolute Error	6.2985	6.5994	6.5929	6.2985	3.5793

Table 1: Relevant statistics for the trained linear and decision tree models; Ratio of the Variance Explained (RVE; Root Mean Squared Error (RMSE); Pearson Correlation (Correlation Score).

Objective 2 - Produce the best binary classification model assuming as positive all instances with values of total_UPRDS > 40 and as negatives all remaining cases.

For this objective we created models using classification decision tree and logistic regression.

After creating the models, the hyperparameters were varied in search of the best set that would give us a good model without overfitting it. We summarized the essential statistics in a table:

Total_UPDRS							
	Decision Tree Classifier						**Logistic Regression**
criterion	**entropy**			**gini**			
max_depth	**3**	**4**	**5**	**3**	**4**	**5**	
Accuracy	0.8515	0.8749	0.9006	0.8600	0.8902	0.9287	0.8285
Precision	1	0.7263	0.9794	0.7922	0.8454	0.9421	0.3415
Recall	0.1198	0.4149	0.4199	0.2308	0.4275	0.6154	0.0177
F1 Score	0.2140	0.5281	0.5878	0.3574	0.5678	0.7445	0.0336
Matthews correlation coefficient	0.3188	0.4862	0.6045	0.3785	0.5518	0.7268	0.0433

Table 2: Relevant statistics for the trained linear and decision tree models.

Discussion

Objective 1 - Produce the best regression model for 'motor UPDRS'

Analyzing the values in Table 1, the regression decision tree showed to be better in the overall statistical indicators.

In the Ratio of the Variance Explained (RVE), the best regressor is the one that have a value closest to 1.0 and the regression decision tree have the highest with 0.65. Since the RVE has the best value, it is normal that the regression decision tree also has Root Mean Squared Error (RMSE) with a good value, considering that both are related.

Regarding the Pearson Correlation (Correlation Score), a good value should be closest to 1 or to -1 because is when they are more correlated, either positively or negatively, the regression decision tree has the closest to 1, with 0.8.

Everything stated supports the regression decision tree as a good model for predicting the motor UPRDS and that is why we choose it to be validated with the IVS.

Objective 2 - Produce the best binary classification model assuming as positive all instances with values of total_UPRDS > 40 and as negatives all remaining cases.

Considering the values in Table 2, the decision tree classifier using the Gini criterion, instead of Entropy, and max depth of 5, showed to be the one that performs better compared to the other models.

When choosing which models perform the best, we want to look first for the Matthews Correlation Coefficient and for a value closest to one. After this we want to look for the other parameters, such as the accuracy and the F1-score, which also show better values compared to the others.

Everything stated supports the decision tree classifier using the Gini criterion as a good model for classifying the total UPRDS and that is why we choose it to be validated with the IVS.

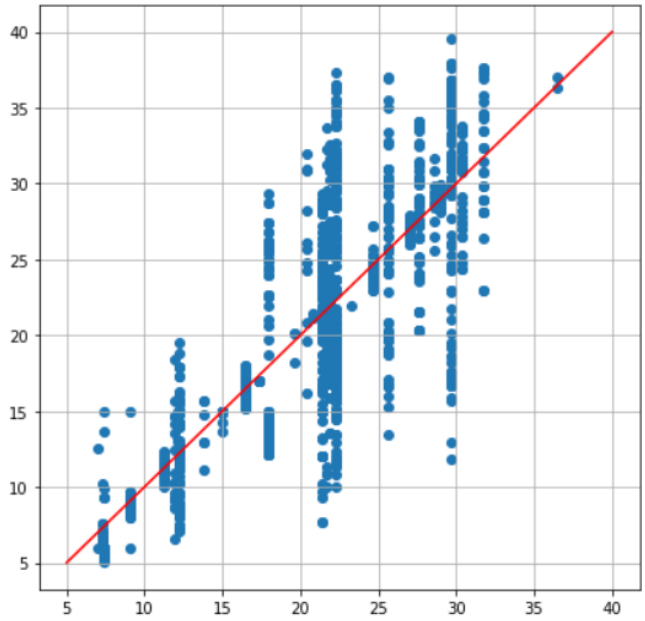
Conclusion

Validating the regression model (**Objective 1**) with the IVS, we get the following values for the relevant statistics.

Motor_UPDRS					
	RVE	**RMSE**	**Correlation Score**	**Maximum Error**	**Mean Absolute Error**
IVS validation	0.635	5.0327	0.7969	17.8944	3.787

Table 3: Relevant statistics for the regressor tree model validate with the Independent Validation Set (IVS); Ratio of the Variance Explained (RVE; Root Mean Squared Error (RMSE); Pearson Correlation (Correlation Score).

Plotting the predictions of the model against the ground truth for the IVS we get:



Graph 1: Predicted-Truth plot for the regressor tree model validated with the Independent Validation Set (IVS)

Therefore, we can conclude that the regressor decision tree is a good model considering the objective we had.

Validating the classification model (**Objective 2**) with the IVS, we get the following values for the relevant statistics.

Total_UPDRS					
	Accuracy	**Precision**	**Recall**	**F1 Score**	**Matthews correlation coefficient**
IVS validation	0.9268	0.9774	0.6103	0.7514	0.7383

Table 4: Relevant statistics for the regressor tree model validate with the Independent Validation Set (IVS).

We can also summarize the essential statistics with a Confusion Matrix:

Confusion Matrix			
Total Population = 1175		**Predicted Condition**	
		Below 40	**Above 40**
Actual Condition	**Below 40**	130	83
	Above 40	3	959

Table 5: Confusion Matrix for the decision tree classifier validated with the Independent Validation Set (IVs)

Consequently, we conclude that the decision tree classifier is a good model for classification of the positive/negative for the total UPRDS.

Overall and for this dataset we think that the decision tree models performed better. However, the models are not achieving values in the parameters around 0.8, so there might be better models for this dataset.