Best Machine Learning on Predict UPDRS Score

Introduction

The Parkinson data set contains 20 persons with Parkinson's (6 female and 14 male) and 20 persons with good healthy (10 female and 10), including multiple types of sound recordings. The main objective is to find out the best machine learning model to predict the UPDRS scores based on the dataset. A standardized rating system called the UPDRS (Unified Parkinson's Disease Rating System) is used to evaluate the severity and course of Parkinson's disease. Researchers and medical professionals often use it to assess the motor symptoms and activities of daily living (ADL) of Parkinson's disease patients. Plus, the subject id patient needs to remove because it has meaningless to the model.

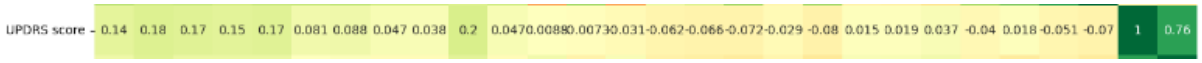
Data splitting

For the Parkinson's data set, it has been decided that 80% of the dataset is for training data, and the remaining is used for test data.

Features selection techniques

Correlation Matrix with Heatmap

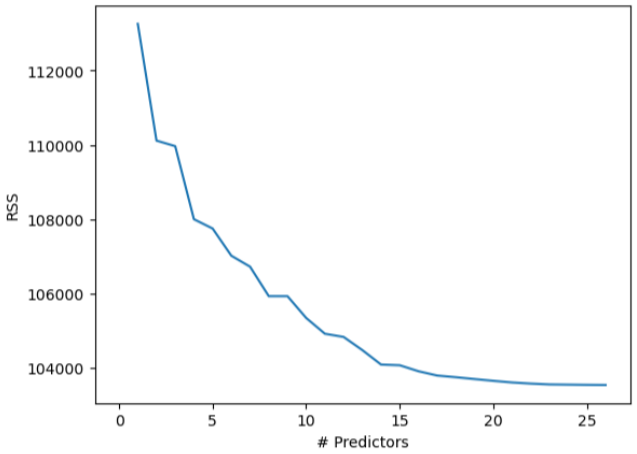
Correlation Analysis is based on how strong the correlation between the target variables with other variables correlates with each other. The selection variables depend on how much the correlation coefficient value of different variables on the UPDRS score(Target Variables). As we can see from the heatmaps, only a few variables that have a weak correlation and also a strong correlation in this feature selection. Therefore, this feature selection model represents 7 variables only which are Jitter(local, absolute), Jitter(rap), Jitter(ppq5), Jitter(ddp), 'Shimmer (apq11), and class information, the remaining variables will be eliminated due to no correlation with UPDRS score.



The picture above shows the heat maps correlation on the UPDRS score

Forward and Backward Elimination

Forward selection elimination is feature selection that begins with an empty set of variables and starts by adding one by one to find out the best model for performance meanwhile backward selection is the opposite of forward selection, beginning with full variables and then gradually removing the least relevant ones. The selection number of variables for this model depends on how good the model's performance is by determining how many variables have been selected. Therefore, forward selection is selected in around 15 variables and backward selection in 9 variables based on RSS.



The picture above shows the backward elimination of the graph on RSS with the predictors of the number

Unvariate selection

Unvariate selection is a feature selection technique used in machine learning. It is a feature selection method that assesses each characteristic separately from the other features according to how closely it relates to the target variable. The selection variables depend on the score which refers to a statistical measure that expresses the strength of each feature's association with the objective variable the properties are ranked using this score, and the most pertinent ones are chosen for additional research or model training. For this model, we choose 11 variables that have above or equivalent to 500.0 scores.

	Specs	Score
20	Number of periods	8973.072021
19	Number of pulses	8473.153445
18	Maximum pitch	4151.806007
16	Standard deviation	3457.264591
14	Median pitch	3368.703523
15	Mean pitch	2987.568298
17	Minimum pitch	2449.519851
23	Fraction of locally unvoiced frames	1579.525238
25	Degree of voice breaks	647.788016
10	Shimmer (dda)	534.279155
26	class information	520.000000

The picture above shows the model that has been selected based on a score of 500 above or equivalent.

Result and Discussion

From the score, we can see that the best model that would predict the UPDRS score is the univariate selection model which consists of 11 variables with around 0.769 accuracies. This shows that this model is very good compared to the other 4 models. The correlation matrix selection is not suitable for this kind of type because it has a weak correlation and only one has a strong correlation. Another reason some models cannot get high frequency is because of the multicollinearity in their models.

Model	Accuracy	MAE
Full Model	0.524	10.707
Correlation Model(Filter Method)	0.486	11.178
Backward Elimination Model	0.514	10.668
Forward Elimination Model	0.519	10.197
Unvariate Model	0.769	4.457

Classification Report

From the classification report from the best model that we selected, we can see that 4 common metrics that we can explain through this model which is precision, recall, F1-score, and support:

The precision depends on how the model can predict based on the data that has been trained. As we can see that the model can predict the UPDRS score which 1 and 55 have the highest percentage score meanwhile 31 has a zero score. UPDRS score that scores 1 has shown the best possible value indicating perfect precision and recall. A low F1 score denotes an imbalance or subpar performance in either accuracy or recall, and a high F1 score indicates a healthy balance between the two. The higher the F1 score, the best model can predict. The distribution of samples among various classes by examining the support value. When working with datasets that are unbalanced, it specifically helps you evaluate the validity of the performance measures for each class. As we can see, based on the classification report, those who have higher support values have a great number of instances means good for predicting the model.

	precision	recall	f1-score	support
1	1.00	1.00	1.00	109
5	0.61	0.73	0.67	15
8	0.43	0.43	0.43	7
11	0.50	0.60	0.55	5
12	0.50	0.67	0.57	3
16	0.50	0.33	0.40	6
20	0.75	0.50	0.60	6
23	0.44	0.67	0.53	6
24	0.50	0.25	0.33	4
26	0.62	0.62	0.62	8
31	0.00	0.00	0.00	2
32	0.55	0.60	0.57	10
40	0.56	0.38	0.45	13
46	0.31	0.40	0.35	10
55	1.00	0.50	0.67	4
accuracy			0.77	208
macro avg	0.55	0.51	0.52	208
weighted avg	0.78	0.77	0.77	208

Conclusion

In conclusion from this topic, it is found that not all variables can be used to predict the UPDRS score due to multicollinearity from models and figured out by using the multiple features technique to select the best model.

For the best variables in the model to predict the UPDRS score are:

1. Number of periods
2. Number of pulses
3. Maximum pitch
4. Standard deviation
5. Median pitch
6. Mean pitch
7. Minimum pitch
8. Fraction of locally unvoiced frames
9. Degree of voice breaks
10. Shimmer (dda)
11. class information