PRINCIPAL COMPONENT ANALYSIS-CLUSTERING ASSIGNMENT

Analysis on key factors in predicting
Unified Parkinson's Disease Rating Scale (UPDRS)

Prepared by:

Team Data Insighters

Sreekar Bethu

Balacoumarane Vetrivel

Meghna Vinay Amin

Mohammed Ismail Khan

Table of Contents

1. INTRODUCTION

Parkinson's disease (PD) is a neurodegenerative disorder that affects predominately dopamine-producing ("dopaminergic") neurons in a specific area of the brain called '*substantia nigra*' and the central nervous system affecting the motor systems. The disease is named after an English doctor, James Parkinson, who published the first detailed description. There is no cure for this disease, but medications, surgery and multidisciplinary management can provide relief from the symptoms [1]. The researchers at the University of Oxford gathered the dataset based on the information obtained through telemonitoring devices of patients who were part of the study.

1.1 Data Source

The data was collected by the researchers at the University of Oxford in collaboration with 10 medical centers in the US and Intel Corporation who developed the telemonitoring device to record the speech signals. The dataset is a collection of biomedical voice measurements of 42 patients with early-stage Parkinson's disease who had a six-month trial of a telemonitoring device for remote symptom progression monitoring. The variables and their metadata are given below:

Variable	Description	Type
Subject	Integer that uniquely identifies each person who undergoes the telemonitoring test for PD	Integer
Age	Age of patient	Continuous
Sex	Gender of the person True-male and False- Female	Categorical
test_time	Time since the person is recruited into the test. It is represented in number of days	Numeric/ Continuous
motor_UPDRS	This Variable gives understanding of how much the person is affected by the disease.	Numeric/ Continuous
Jitter	It is the measure of cycle to cycle variations of fundamental Frequency.	Numeric/ Continuous
Jitter_abs	This is the average absolute difference between consecutive periods, in seconds. jitter(abs)=$\sum_{i=2}^{N}\|T_i-T_{i-1}\|/(N-1)$ Where Ti is the duration of the i-th interval and N is the no. of intervals	Numeric/ Continuous
Jitter_rap	This is the Relative Average Perturbation, the average absolute difference between a period and the average of it and its two neighbours, divided by the average period Average Perturbation = $\sum_{i=2}^{N-1}\|T_i - (T_{i-1} + T_i + T_{i+1}) / 3\| / (N - 2)$	Numeric/ Continuous
Jitter_ppq5	This is the five-point Period Perturbation Quotient, the average absolute difference between a period and the average of it and its four closest neighbours, divided by the average period. Five-point Period Perturbation Quotient = $\sum_{i=3}^{N-2}\|T_i - (T_{i-2} + T_{i-1} + T_i + T_{i+1} + T_{i+2}) / 5\| / (N - 4)$	Numeric/ Continuous
Jitter_ddp	This is the average absolute difference between consecutive differences between consecutive periods, divided by the average period. Abs_ddp = $\sum_{i=2}^{N-1}\|(T_{i+1} - T_i) - (T_i - T_{i-1})\| / (N - 2)$	Numeric/ Continuous
Shimmer	It is the measure of cycle to cycle variations of Amplitude.	Numeric/ Continuous

Shimmer_db	This is the average absolute base-10 logarithm of the difference between the amplitudes of consecutive periods, multiplied by 20	Numeric/ Continuous
Shimmer_apq3	This is the three-point Amplitude Perturbation Quotient, the average absolute difference between the amplitude of a period and the average of the amplitudes of its neighbours, divided by the average amplitude.	Numeric/ Continuous
Shimmer_apq5	This is the five-point Amplitude Perturbation Quotient, the average absolute difference between the amplitude of a period and the average of the amplitudes of it and its four closest neighbours, divided by the average amplitude	Numeric/ Continuous
Shimmer_apq11	This is the 11-point Amplitude Perturbation Quotient, the average absolute difference between the amplitude of a period and the average of the amplitudes of it and its ten closest neighbours, divided by the average amplitude	Numeric/ Continuous
Shimmer_dda	This is the average absolute difference between consecutive differences between the amplitudes of consecutive periods	Numeric/ Continuous
NHR	It is the ratio of the amplitude of noise relative to tonal components in the speech.	Numeric/ Continuous
HNR	It is the ratio of tonal components in the speech to the noise in speech. It helps to separate healthy from PD subjects	Numeric/ Continuous
RPDE	A nonlinear dynamical complexity measure. It is the periodicity and repetitiveness of a signal	Numeric/ Continuous
DFA	Signal fractal scaling exponent. It is the measure of the extent of the noise in the speech signal	Numeric/ Continuous
PPE	A nonlinear measure of fundamental frequency variation. It is calculated on the probability distribution of occurrence of relative semitone variations	Numeric/ Continuous
total_UPDRS **(target variable)**	This is the final score given by clinician considering various factors. If the score is high, then the person has very high probability of getting Parkinson disease. Threshold value of score is 15. [2]	Numeric/ Continuous

1.2 Objective

To build a model which shall help predict the 'total_UPDRS' score (Target Variable) based on prominent contributing factors from the dataset provided using the method of Principal Component Analysis and Clustering.

2. EXPLORATORY DATA ANALYSIS

There are no duplicates or missing values in the data.

This can be inferred from the summary of the data given below:

```
    subject            age            sex          test_time         motor_updrs
 Min.   :  1.00   Min.   :36.0   false:4008   Min.   : -4.263   Min.   : 5.038
 1st Qu.:10.00   1st Qu.:58.0   true :1867   1st Qu.: 46.847   1st Qu.:15.000
 Median :22.00   Median :65.0                Median : 91.523   Median :20.871
 Mean   :21.49   Mean   :64.8                Mean   : 92.864   Mean   :21.296
 3rd Qu.:33.00   3rd Qu.:72.0                3rd Qu.:138.445   3rd Qu.:27.596
 Max.   :42.00   Max.   :85.0                Max.   :215.490   Max.   :39.511
  total_updrs        jitter            jitter_abs          jitter_rap
 Min.   : 7.00   Min.   :0.000830   Min.   :2.250e-06   Min.   :0.000330
 1st Qu.:21.37   1st Qu.:0.003580   1st Qu.:2.244e-05   1st Qu.:0.001580
 Median :27.58   Median :0.004900   Median :3.453e-05   Median :0.002250
 Mean   :29.02   Mean   :0.006154   Mean   :4.403e-05   Mean   :0.002987
 3rd Qu.:36.40   3rd Qu.:0.006800   3rd Qu.:5.333e-05   3rd Qu.:0.003290
 Max.   :54.99   Max.   :0.099990   Max.   :4.456e-04   Max.   :0.057540
  jitter_ppq5          jitter_ddp          shimmer          shimmer_db      shimmer_apq3
 Min.   :0.000430   Min.   :0.000980   Min.   :0.00306   Min.   :0.026   Min.   :0.00161
 1st Qu.:0.001820   1st Qu.:0.004730   1st Qu.:0.01912   1st Qu.:0.175   1st Qu.:0.00928
 Median :0.002490   Median :0.006750   Median :0.02751   Median :0.253   Median :0.01370
 Mean   :0.003277   Mean   :0.008962   Mean   :0.03404   Mean   :0.311   Mean   :0.01716
 3rd Qu.:0.003460   3rd Qu.:0.009870   3rd Qu.:0.03975   3rd Qu.:0.365   3rd Qu.:0.02057
 Max.   :0.069560   Max.   :0.172630   Max.   :0.26863   Max.   :2.107   Max.   :0.16267
  shimmer_apq5       shimmer_apq11       shimmer_dda            nhr                hnr
 Min.   :0.00194   Min.   :0.00249   Min.   :0.00484   Min.   :0.000286   Min.   : 1.659
 1st Qu.:0.01079   1st Qu.:0.01566   1st Qu.:0.02783   1st Qu.:0.010955   1st Qu.:19.406
 Median :0.01594   Median :0.02271   Median :0.04111   Median :0.018448   Median :21.920
 Mean   :0.02014   Mean   :0.02748   Mean   :0.05147   Mean   :0.032120   Mean   :21.680
 3rd Qu.:0.02375   3rd Qu.:0.03272   3rd Qu.:0.06173   3rd Qu.:0.031463   3rd Qu.:24.444
 Max.   :0.16702   Max.   :0.27546   Max.   :0.48802   Max.   :0.748260   Max.   :37.875
      rpde              dfa               ppe
 Min.   :0.1510   Min.   :0.5140   Min.   :0.02198
 1st Qu.:0.4698   1st Qu.:0.5962   1st Qu.:0.15634
 Median :0.5423   Median :0.6436   Median :0.20550
 Mean   :0.5415   Mean   :0.6532   Mean   :0.21959
 3rd Qu.:0.6140   3rd Qu.:0.7113   3rd Qu.:0.26449
 Max.   :0.9661   Max.   :0.8656   Max.   :0.73173
```

Linear regression model can be used for predicting the target from given independent variables.

2.1 Outlier Treatment

Each numeric variable is analyzed for outliers. It is observed that there are 224 observations in the data as outliers.

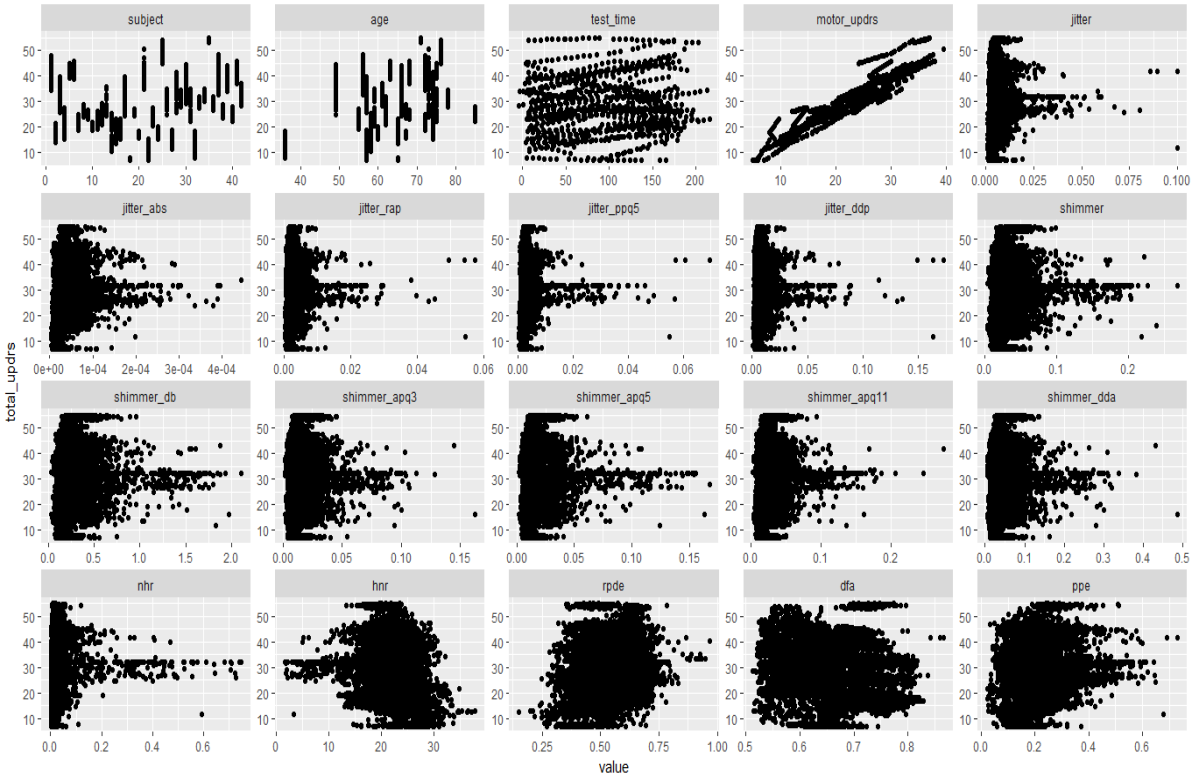
Process used to analyze outliers is given below:

- Quantile is computed for variable for the probabilities 0.01 & 0.99
- Inter quartile range is calculated for the variable and is multiplied by 1.5
- Data subset has been done from the variable value greater than (Quantile[1] – range) and less than (Quantile[2] + range)

Reformed final dataset was generated by eliminating the outliers.

2.2 Target VS Predictor variables

Below plot represents the target VS predictor variables to identify the most influencing factors:

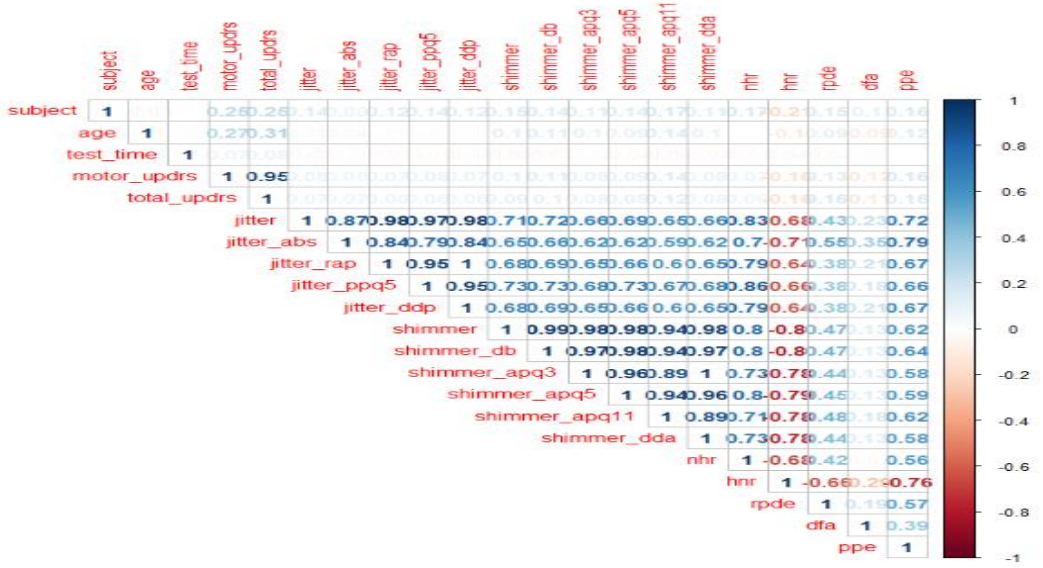


From the plot, it can be inferred that

- 'motor_updrs' variable is highly related to target variable.
- 'test_time' variable does not have much impact on the target variable.
- Plots of target vs other predictors are not clear to infer the relation, so all other variables are considered for further analysis.

2.3 Correlation between Numeric variables

As all the independent variables are numeric continuous except 'Sex', there is a high chance of correlation between the variables. Below is the plot representing the correlation among the independent variables:



From the above correlation values, it can be inferred that

- Jitter variables are highly correlated among each other.
- Shimmer variables are highly correlated among each other.
- 'nhr' & 'hnr' are highly correlated with all 'Jitter' & 'Shimmer' variables.
- 'jitter_rap' & 'jitter_ddp' are correlated with value 1. Hence, we are considering only 'jitter_rap' for further analysis.
- 'shimmer_apq3' & 'shimmer_dda' are correlated with value 1. Hence, we are considering only 'shimmer_apq3' for further analysis.

As the variables are highly correlated in the form of groups, Principal Component Analysis can be used to generate Principal Components with appropriate loadings of original variables.

3. PRINCIPAL COMPONENT ANALYSIS

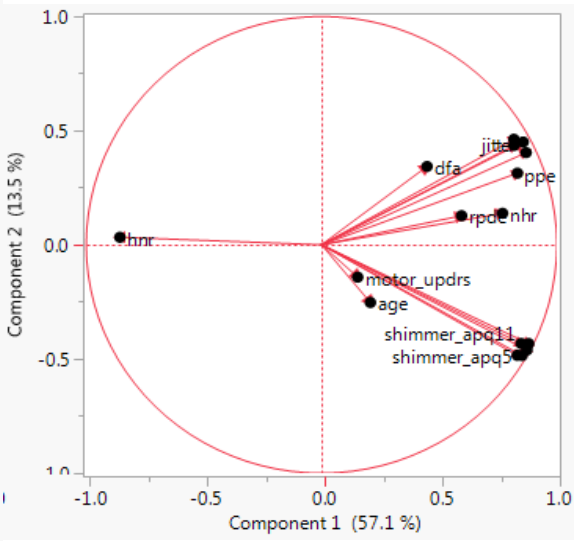
Below are the variables that are not considered while generating Principal Components.

- 'Subject '– this variable signifies the identity of the patients
- 'test_time' – as mentioned in EDA
- 'Sex' – this variable is categorical and hence not considered for PCA
- 'jitter_ddp' & 'shimmer_dda' – These variables are not considered as they exactly correlated with 'jitter_rap' & 'shimmer_apq3' respectively.

Below are the Eigen values obtained and Summary plot of original variables:

Eigenvalues

Number	Eigenvalue	Percent	Cum Percent
1	9.1425	57.141	57.141
2	2.1599	13.500	70.640
3	1.3132	8.208	78.848
4	0.8461	5.288	84.136
5	0.7645	4.778	88.914
6	0.7048	4.405	93.319
7	0.2882	1.802	95.120
8	0.2384	1.490	96.610
9	0.2159	1.349	97.959
10	0.1266	0.791	98.750
11	0.0969	0.605	99.356
12	0.0421	0.263	99.618
13	0.0266	0.166	99.785
14	0.0179	0.112	99.897
15	0.0125	0.078	99.975
16	0.0040	0.025	100.000



As per the Eigen Values and Cumulative percentage, first three Principal Components provide 78.8% of information of data.

Below is the Loading matrix of three Principal Components:

Formatted Loading Matrix

	Prin1	Prin2	Prin3
shimmer_db	0.880167	-0.432646	-0.093203
shimmer	0.873015	-0.460378	-0.122404
jitter_ppq5	0.869247	0.403551	0.043906
jitter	0.856614	0.449888	0.073871
shimmer_apq5	0.852313	-0.484822	-0.148281
shimmer_apq11	0.848305	-0.432220	-0.070763
ppe	0.832586	0.313778	0.099864
shimmer_apq3	0.831595	-0.483772	-0.155268
jitter_rap	0.818449	0.435463	0.046233
jitter_abs	0.817234	0.462172	-0.000560
nhr	0.769341	0.138104	0.158482
rpde	0.594927	0.127353	0.151179
motor_updrs	0.153565	-0.141563	0.753943
age	0.207435	-0.252579	0.647536
dfa	0.448225	0.343145	-0.427835
hnr	-0.856986	0.033036	-0.014298

Prin1 component shows high loadings of original variables and Prin2 does not show high loadings. So, to get a better understanding of the components, Varimax rotation is performed and rotated factors are as below:

Rotated Factor Loading

	Factor 1	Factor 2	Factor 3
age	0.057713	0.168449	0.703152
motor_updrs	0.105313	0.029982	0.774637
jitter	0.937461	0.249618	0.022388
jitter_abs	0.908211	0.231420	-0.055426
jitter_rap	0.896624	0.240160	-0.003524
jitter_ppq5	0.912074	0.297419	0.005797
shimmer	0.315054	0.941472	0.058795
shimmer_db	0.342327	0.920254	0.080680
shimmer_apq3	0.265201	0.937269	0.029814
shimmer_apq5	0.280467	0.950445	0.038328
shimmer_apq11	0.321927	0.893206	0.099946
nhr	0.674062	0.388776	0.174839
hnr	-0.606367	-0.600798	-0.084153
rpde	0.538402	0.279713	0.157784
dfa	0.508408	0.162407	-0.465662
ppe	0.831488	0.322419	0.079386

From the rotated factors, it can be inferred that.

- Factor 1: It is highly loaded with all ‘jitter’ variables, ‘hnr’, ‘nhr’, ‘rpde’, ‘dfa’ & ‘ppe’.
  These original variables are related to frequency/time-period.
  Hence, this component can be termed as ‘Frequency factors of voice’.
- Factor 2: It is highly loaded with all ‘shimmer’ variables.
  ‘Shimmer’ variable denoted the Amplitude.
  Hence, this component can be termed as ‘Amplitude of voice’.
- Factor 3: It is highly loaded with ’age’ & ‘motor_updrs’
  ‘motor_udrs’ denotes the likeliness of a person having Parkinson’s disease.
  Hence, this component can be termed as ‘Disease rating as per the age’

The three-rotated components generated from highly correlated original variables are independent of each other and it can be inferred from the below matrix:

Correlations

	Frequency factors of voice	Amplitude of voice	Disease rating as per the age
Frequency factors of voice	1.0000	-0.0000	-0.0000
Amplitude of voice	-0.0000	1.0000	0.0000
Disease rating as per the age	-0.0000	0.0000	1.0000

From the above values it can be inferred that no correlation is present between the Factor components and that these Factors are used to cluster the data into groups.

4. CLUSTER ANALYSIS

4.1 K-Means Clustering

K-means clustering is performed for this dataset as the factors used to cluster are numerical.
K-means clustering analysis is done using different number of cluster groups, i.e. 3 group clustering, 4 group clustering, 5 group clustering and 6 group clustering.
On analyzing the result of K-means clustering, it can be inferred that 3-group clustering provides better inference.

The factors taken into consideration for selecting n=3 are:

- Minimum overlap between the clusters
- Ability to profile the clusters based on the rotated components
- Number of observation in each cluster is significant to distinguish the three clusters.

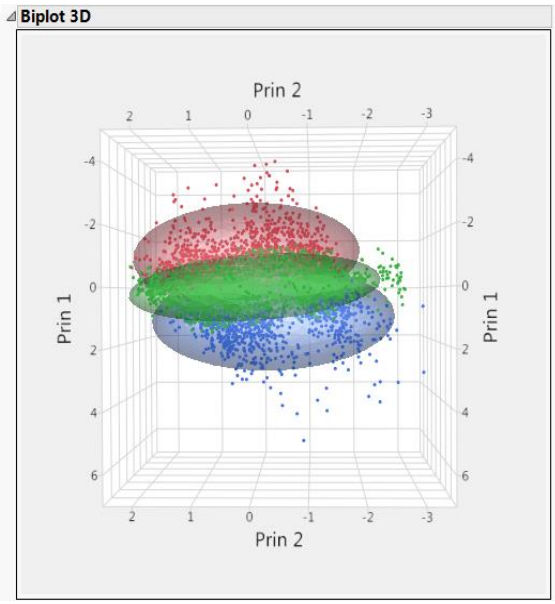
Below are the clusters formed for n=3:

Cluster Summary

Cluster	Count	Step	Criterion
1	1094	35	0
2	3423		
3	1134		

Cluster Means

Cluster	Frequency factors of voice	Amplitude of voice	Disease rating as per the age
1	1.52212313	-0.2722729	0.10671411
2	-0.438658	-0.4296565	-0.0157954
3	-0.1443355	1.55959508	-0.0552713



4.2 Profiling of clusters

CLUSTER1 (RED): - The Cluster 1 is aggregated with the observations which have very high magnitude in Frequency related variables such as Jitter, hnr, nhr, rpde, dfa and PPE. As per the data set and clusters formed, Probability of person having Parkinson's disease is higher in this cluster amongst the three.

CLUSTER2 (GREEN): - The Cluster 2 is aggregated with the observations which have less magnitude in all the variables. As per the data set and clusters formed, Probability of person having Parkinson's disease is least in this cluster.

CLUSTER3 (BLUE): - The Cluster 3 is aggregated with the observations which have very high magnitude in Amplitude, that is, 'Shimmer' variables. As per the data set and clusters formed, Probability of person having Parkinson's disease is moderate in this cluster among all other clusters.

Cluster	Inference
Cluster 1-Red	High chances of encountering Parkinson's disease
Cluster 2-Green	Low chances of encountering Parkinson's disease
Cluster 3-Blue	Moderate chances of encountering Parkinson's disease

5. REGRESSION

Linear regression model is built by considering the variable 'Sex' and three rotated components as below:

Effect Summary

Source	LogWorth	PValue
Disease rating as per the age	1112.298	0.00000
Frequency factors of voice	30.447	0.00000
sex	23.976	0.00000
Amplitude of voice	3.262	0.00055

Remove Add Edit FDR

Summary of Fit

RSquare	0.604467
RSquare Adj	0.604187
Root Mean Square Error	6.814937
Mean of Response	28.96062
Observations (or Sum Wgts)	5651

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	400732.15	100183	2157.101
Error	5646	262219.22	46	Prob > F
C. Total	5650	662951.36		<.0001*

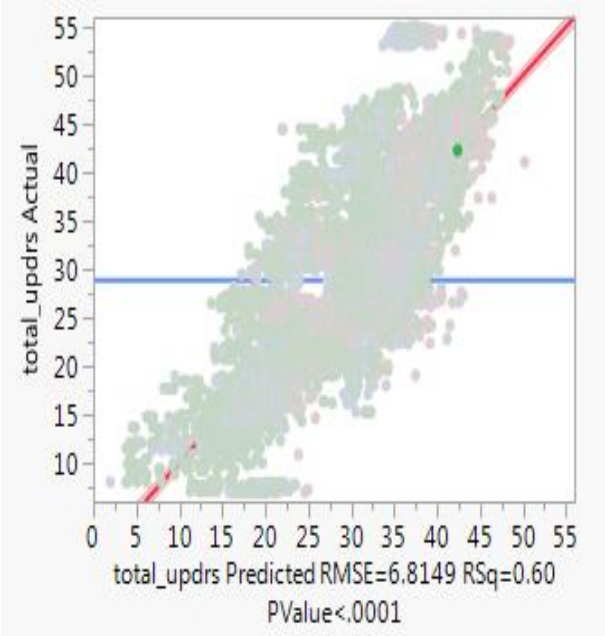
Parameter Estimates

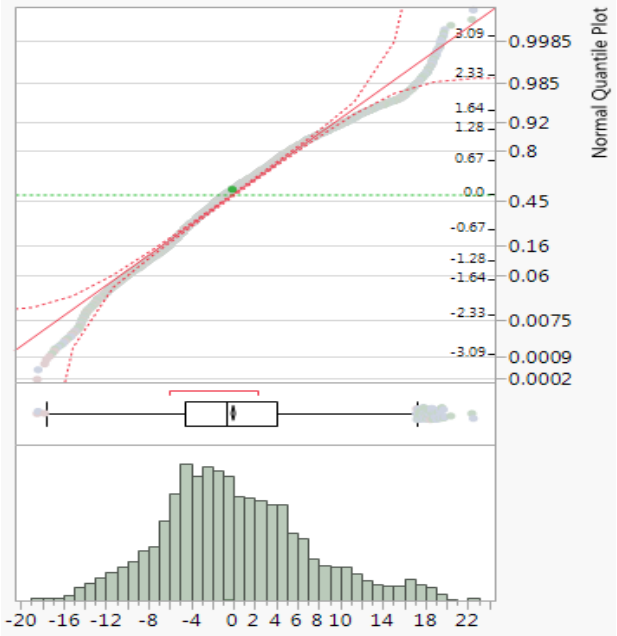
Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	28.551961	0.098944	288.57	<.0001*
sex[false]	1.0360345	0.100494	10.31	<.0001*
Frequency factors of voice	1.0732061	0.091865	11.68	<.0001*
Amplitude of voice	0.3152874	0.09116	3.46	0.0005*
Disease rating as per the age	8.2702397	0.090665	91.22	<.0001*

As per the above output, it can be inferred that all the rotated components and Sex are significant in predicting the target variable 'total_updrs'.

The Linear Regression model has an Adj. RSquare of 0.6, which can be accepted as good. Parameter estimate of each variable is shown in the above output.

Actual VS Predicted Plot shows that actual values are almost distributed around predicted values. Residual qqplot also shows that model built is strong enough to predict the values of target 'total_updrs'.





Below is the prediction expression of linear regression model built:

Prediction Expression

28.551960735

$$+\text{Match}\left(\text{sex},\begin{pmatrix}\text{"false"} \Rightarrow 1.0360345382\\ \text{"true"} \Rightarrow -1.036034538\\ \text{else} \Rightarrow .\end{pmatrix}\right)$$

$+ 1.0732061085 \cdot \text{Frequency factors of voice}$

$+ 0.3152874345 \cdot \text{Amplitude of voice}$

$+ 8.2702397075 \cdot \text{Disease rating as per the age}$

6. CONCLUSION

Therefore, based on the study it can be concluded that the possibility of a person having Parkinson's disease can be inferred from predictor variables. The order of level of significance of original variables are as below:

1)Disease rating as per the age
   - Age
   - Motor_updrs

2)Frequency variables of voice
   - Jitter
   - Jitter_abs
   - Jitter_rap
   - Jitter_ppq5
   - Jitter_ddp
   - Hnr
   - Nhr
   - Jitter_ddp
   - Rpde
   - Dfa
   - Ppe

3)Amplitude
   - Shimmer
   - Shimmer_db
   - Shimmer_apq3
   - Shimmer_apq5
   - Shimmer_apq11

4)Sex

7. LIMITATIONS

1. Outliers are removed from the dataset for this study. In future, if any data is received as an outlier then the model may not predict the accurate score.
2. Significant Principal Components in the study are 3 that provide 78.8% of information about data. A cumulative percent greater than 80% would have been even better for analysis.
3. Data is limited to parameters concerning only voice.

8. REFERRENCE

[1] N.E. Piro, L. Baumann, M. Tengler, L. Piro, R. Blechschmidt-Trapp, " Telemonitoring of Patients with Parkinson's Disease Using Inertia Sensors", Sensors (Basel) 2016 Jun; 16(6): 930. Published online 2016 Jun 21. doi: 10.3390/s16060930

[2] B. E. Sakar, C. Sakar, G. Serbes, O. Kursun, "Determination of the optimal threshold value that can be discriminated by dysphonia measurements for unified Parkinson's disease rating scale", Bioinformatics and Bioengineering (BIBE) 2015 IEEE 15th International Conference in Nov 2015, pp. 1-4.