Project Report

Analyzing Biomedical Voice Measurements to Predict the Parkinson Disease Rating Scale Scores

Introduction

Neurological disorders affect people profoundly and claim lives at an epidemic rate worldwide. One of such disorder is Parkinson’s Disease, the second most common neurodegenerative disorder after Alzheimer’s, and it is estimated that more than one million people in North America alone are affected.
The cause of Parkinson's disease is generally unknown, but believed to involve both genetic and environmental factors. In 2015, PD affected 6.2 million people and resulted in about 117,400 deaths globally.
The Parkinson’s disease has 5 stages having different levels of symptoms, whose rating system has been largely supplanted by the Unified Parkinson's Disease Rating Scale (UPDRS). A total of 199 points are possible. 199 represents the worst (total disability), 0--no disability.
Providing computational tools for Parkinson disease using a set of data that contains medical information is very desirable for alleviating the symptoms that can help the amount of people who want to discover the risk of disease at an early stage.
We aim to analyze the Biomedical Voice Measurements to predict the Parkinson’s Disease Rating Scale Scores (for the prediction of PD progression).

Dataset Information

This dataset is 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. The recordings were automatically captured in the patient's homes.
Columns in the table contain subject number, subject age, subject gender, time interval from baseline recruitment date, motor UPDRS, total UPDRS, and 16 biomedical voice measures. Each row corresponds to one of 5,875 voice recording from these individuals. The main aim of the data is to predict the motor and total UPDRS scores ('motor_UPDRS' and 'total_UPDRS') from the 16 voice measures.
The data is in ASCII CSV format. The rows of the CSV file contain an instance corresponding to one voice recording. There are around 200 recordings per patient, the subject number of the patient is identified in the first column.

Data exploration and Insights

- We trained our predictive model on 3000 voice recording instances and test our model on the remaining 2875 instances.
- We treated each instance as a unique voice recording (not in respect to a particular individual, the dataset had 42 subject id’s and each individual has 200 voice recordings) hence removing the subject id from the dataset.

- We inspected the data for outliers, typos, missing values by plotting histograms of each variables.
- We generated the below plot for the correlation matrix to identify correlations among variables:

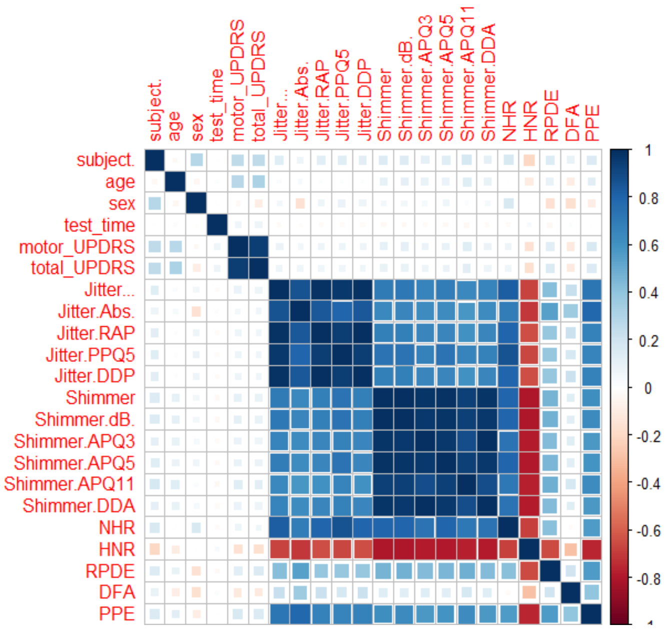


Figure1: Correlation Matrix

- We observed that the two output variables Total_UPDRS and Motor_UPDRS are highly correlated so we decided to calculate the average of the two as Avg_UPDRS and treat this as our output variable.
- Also, observing high correlations and similar metrics among the shimmer variables such as Shimmer:APQ3, Shimmer:APQ5 and Shimmer_APQ11, we decided to take average as Shimmer_avg.
- During interactions, we have categorized the 'Age' variable as Age less than 60 to be considered as '0' and Age greater than 60 to be considered as '1'.

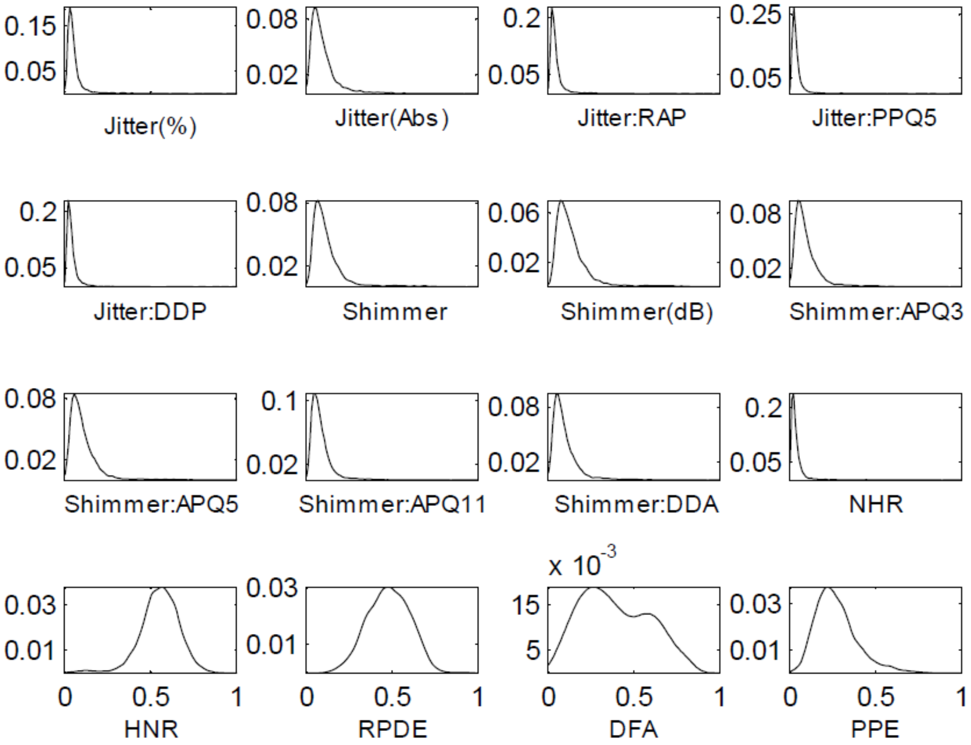
Table1: VIF

	age	sex	test_time	Jitter..	Jitter.Abs.	Jitter.RAP	Jitter.PPQ5	Jitter.DDP
VIF	1.096538978	1.348692878	1.010601737	89.02572552	7.85305124	1324029.777	31.00871506	1324266.011

	Shimmer	Shimmer.dB.	Shimmer.APQ3	Shimmer.APQ5	Shimmer.APQ11	Shimmer.DDA
VIF	173.4509304	76.87073687	23989466.55	52.56724518	15.28205117	23989420.97

	NHR	HNR	RPDE	DFA	PPE
VIF	8.586972847	5.420735976	2.100950436	1.66173679	4.440222648

- Performing the exploratory analysis we identified various skewness in the scatterplots.



We decided to go with below transformations on the predicting variables:

1. All Shimmer and Jitter values: log(1/sqrt(x))
2. NHR: log(1/sqrt(x))
3. PPE: sqrt(x)
4. Test_time: log(x)

We performed Regression Analysis on the Parkinson’s dataset to predict the Avg_UPDRS value for biomedical voice measurements variables.

Measurement of Variable Importance:

After performing the Stepwise regression and plotting the variable importance plot from Random Forest analysis, we decided to choose the following 8 variables from a set of 18 predictor variables: **DFA, test_time, RPDE, Jitter.Abs., HNR, PPE, NHR, Shimmer.DDA.**

Data Analysis Methods

We used various regression analysis methods to measure and compare the performance of different models and identify the importance of the predictor variables. We have used splines to determine the non-linearities between each variables. Following algorithms and regression techniques were implemented in the process of determining the best RMSE value:

Multiple Linear Regression
Ridge Regression
Lasso Regression
Random Forest
Boosted Trees_____
Splines Using GAM:

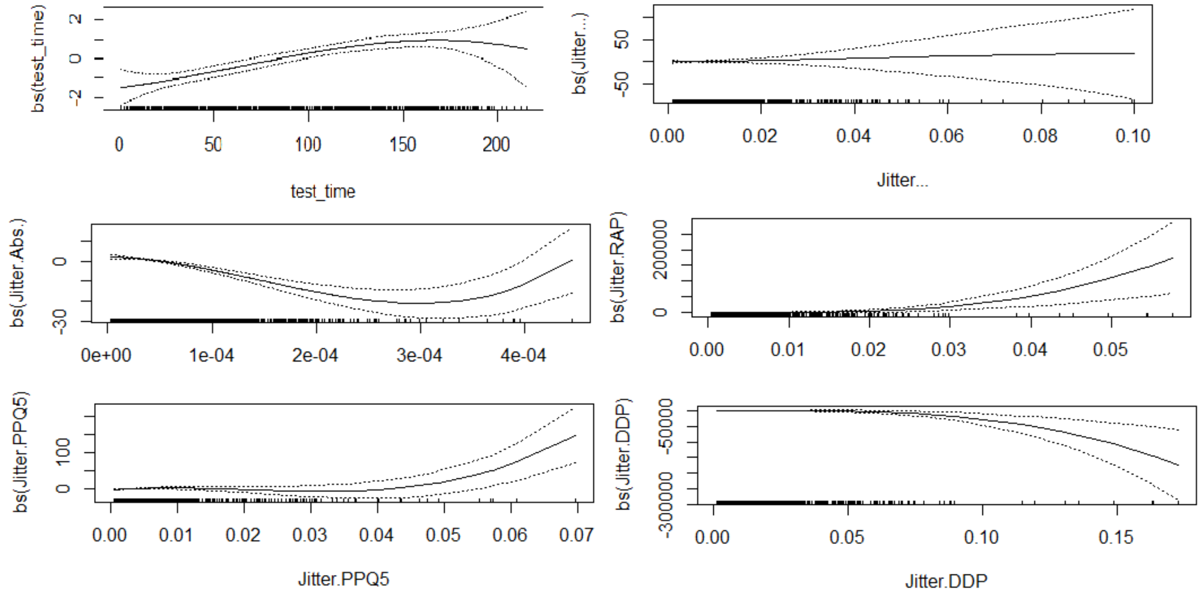


Fig 2. Splines on Individual Jitter Values

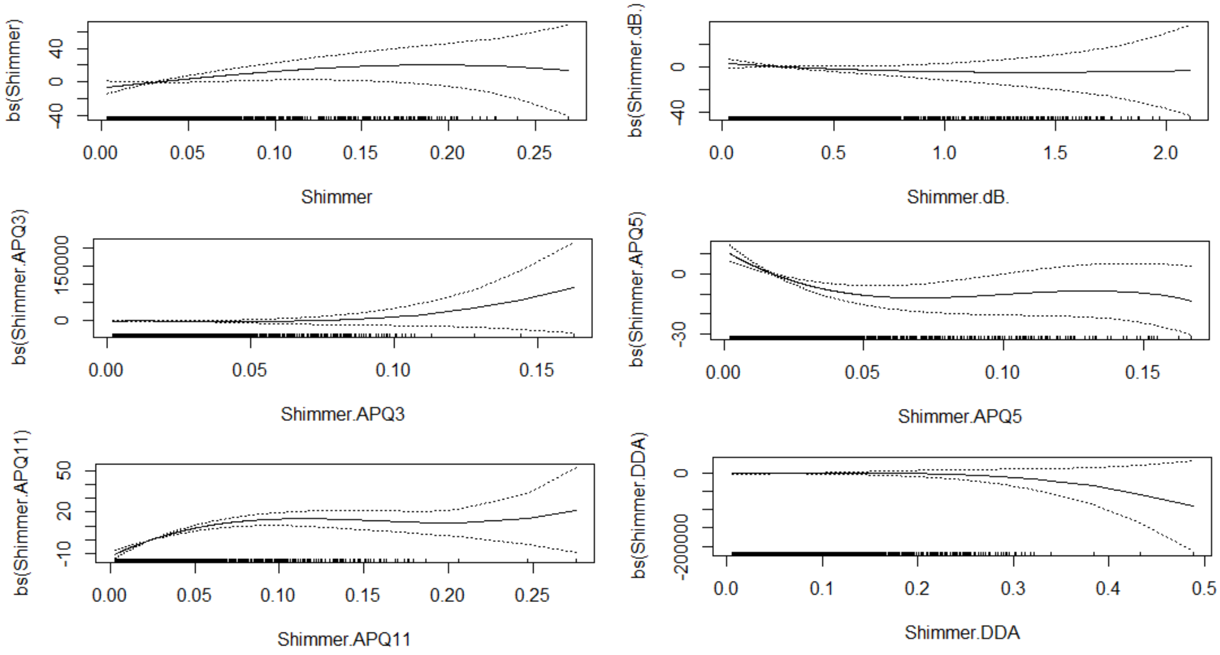


Fig 3. Splines on Individual Shimmer Values

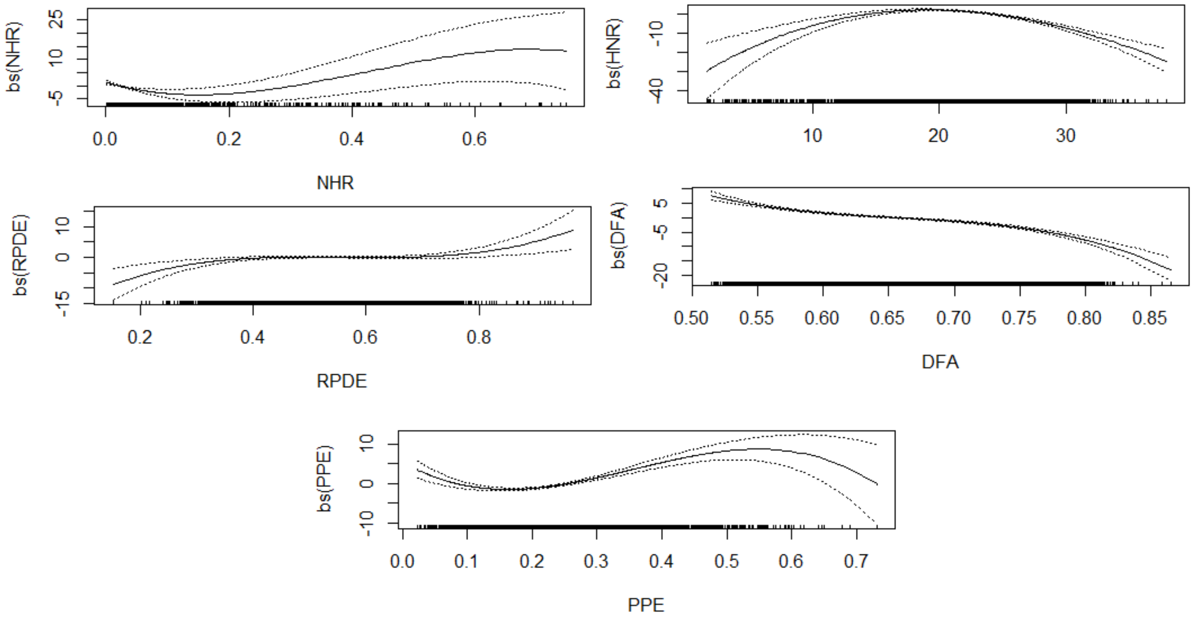


Fig 4. Splines on NHR, HNR, RPDE, DFA, PPE

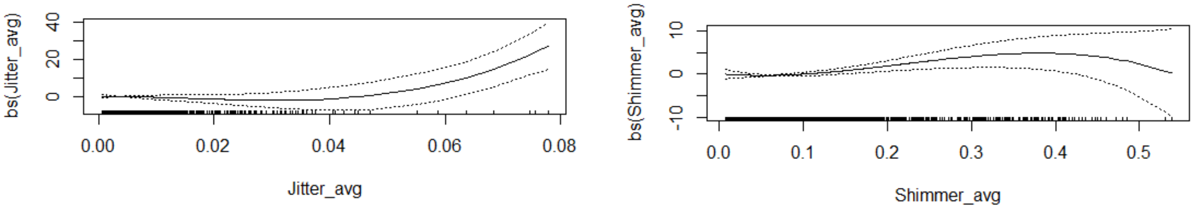


Figure 5. Splines on Jitter_avg & Shimmer_avg

Results

- All basic models has 18 predictor variables.
- After Stepwise and varImpPlot, only below mentioned 8 predictor variables were used: DFA, test_time, RPDE, Jitter.Abs., HNR, PPE, NHR, Shimmer.DDA

Table 3: RMSE values without any transformation

Algorithm	**RMSE**
Basic Linear Regression	8.85387
Stepwise	8.853026
Ridge Regression	8.812963
Lasso Regression	8.843341
Random Forest	6.546595
Boosted Trees	6.82942
Boosted Trees(Shrinkage = 0.01)	10.02562

Table 4: RMSE values after transformation:

Algorithm	**RMSE**
Basic Linear Regression	7.642006
Stepwise	7.649297
Ridge Regression	7.741428
Lasso Regression	7.702846

Boosted Trees	6.82942
Random Forest	**6.531224**

Table 5: Interaction of sex and age with all other values

Algorithm	**RMSE**
Basic Linear Regression	7.08058
Stepwise	7.096714
Ridge Regression	7.456866
Lasso Regression	7.163373
Boosted Trees	6.621275

Conclusion

After applying different regression algorithms to predict the Average_UPDRS value from the different biomedical voice measurements, we decided to conclude the estimated best model based on the best RMSE value. In our analysis, Boosted Trees gave a relatively good RMSE value of 6.82942 but we decided to conclude with Random Forest with a better RMSE value of **6.531224** as the best proposed model obtained after implementing transformations on the predicting variables.

Appendix/Source

https://archive.ics.uci.edu/ml/datasets/Parkinsons+Telemonitoring