Parkinson's Disease Symptom Severity Prediction based on Patient Age, Sex, and Vocal Features

Authors: Sze Pui Tsang, Haolin Zhong

1 Introduction

Parkinson's Disease has become increasingly common with more than 10 million patients affected worldwide. People aged 50 and above are especially vulnerable to Parkinson's disease as the risk of Parkinson's disease is proportional to age. However, there is no treatment for Parkinson's disease till today - patients can simply rely on medication to relieve their symptoms. Vocal impairment is considered one of the earliest indicators of Parkinson's disease; therefore, scientists are developing an automatic diagnosis Parkinson's test that is much simpler, less costly and non-invasive: a speech test.

1.1 Aim of the project

Based on the Oxford Parkinson's Disease Detection Dataset, we attempted to build regression models that map patients' age, sex and vocal features to the severity of their symptoms. We would also compare the performance of different regression models and select the best model.

1.2 Dataset and Variable description

The data was collected from 42 subjects with early-stage Parkinson's disease and includes information on subjects' age, gender, UPDRS score, and vocal features extracted from recordings of identical telemonitoring devices in a six-month trial. The data consist of 5875 rows and 19 columns, with each row corresponding to one recording.

Predictors are vocal features which includes `jitter` (the extent of variation in speech frequency from vocal cycle to vocal cycle), `shimmer` (the extent of variation in speech amplitude from cycle to cycle), noise-harmonics ratios (`NHR` or `HNR`) (the amplitude of noise relative to tonal components in the speech), and some other variables measure vocal vibration (`RPDE`), the extent of turbulent noise (`DFA`) and vocal pitch instability (`PPE`). (See Appendix section 5.3 Variable Description section for detailed descriptions.)

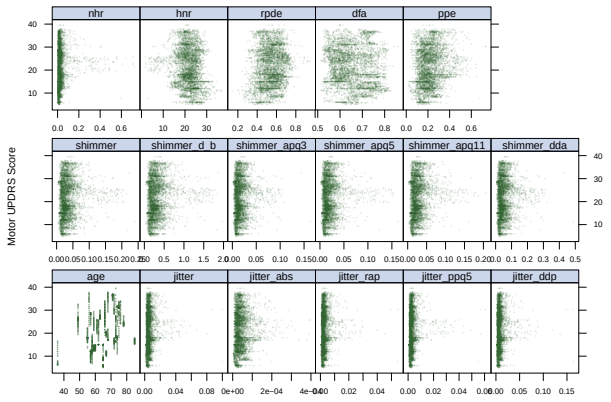
The response is `UPDRS` (Unified Parkinson's diseases rating scale), which reveals the presence and the severity of symptoms.

1.3 Data cleaning process

There is a total of 5875 voice recordings with no missing data. We have removed several variables in the original data, including patient id, recorded time since they should have no contribution in UPDRS score prediction.The responses are motor_UPDRS and total_UPDRS. We selected motor_UPDRS as our response variable and discarded total UPDRS score for the conciseness of our analysis. The score of motor_UPDRS motor UPDRS score ranges from 0 to 108, 0 indicates disease-free while 108 denotes severe motor disabilities in facial expression, action and speech.

2 Exploratory Data Analysis

Before diving into modeling, relationships between the response variable and each predictor, and pairwise correlation among variables were analyzed. Except for a vague trend of increased UDPRS score corresponding to increasing age, no obvious trend was found in the relationship analysis. The correlation analysis suggests that strong correlations exist among various predictors, thus in the future modeling process, methods should be taken to avoid variance inflation.



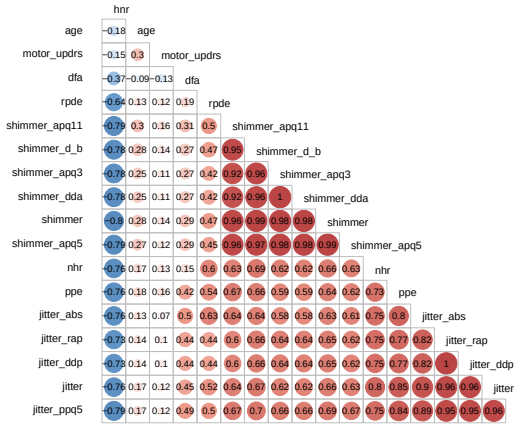


Figure 1 (left). The scatterplot showing the predictors against motor_updrs.

Figure 2 (right). The correlation plot showing the collinearity between predictors.

3 Models

Although we observed severe multicollinearity among variables, in the model training process, we still use all the variables including age, sex and all vocal features as input, as our methods allow automatic feature selection and coefficient regularization.

According to characteristics of the data - a relatively small number of features, high pairwise correlations between variables and low correlations between response and each predictor, a possible heterogeneity among subjects, Linear Regression (LS) model, Elastic Net (ENET) model, Generalized Additive Model (GAM), Multivariate Adaptive Regression Splines (MARS) model, Support Vector Regression Model(SVR), Random Forest (RF) model, Gradient-boosted Tree Model (GBM) and Linear Mixed Effects (LMM) Model were trained to predict patient motor UPDRS score. In models with tuning parameters, a range of parameters was searched and the best value for each parameter was chosen according to 10-fold cross-validated RMSE of resulting models. See Appendix section 5.1 for model performance under sequences of tuning parameters, and Appendix section 5.2 for model coefficients/terms.

3.1 Model Training Methods

- *Linear Regression*: LS assumes that the residuals are independent and identically distributed Gaussian random variables, which may not always be the truth. Therefore, collinearity between predictors would result in poor coefficient estimates. To avoid this problem, `regsubsets` was used to search through the parameter space and find the best combination of features that leads to the model with the smallest BIC, Then, we use these selected features to build our linear model and estimate coefficient based on cross-validated RMSE.
- *Elastic Net*: As a shrinkage method, ENET combines L1 and L2 regularization terms as penalty and can be regarded as a combination of LASSO and Ridge. It assumes:

$$\text{loss}(\alpha, \boldsymbol{\beta}, \lambda) = \text{RSS} + \lambda \left[(\frac{1-\alpha}{2}) \|\boldsymbol{\beta}\|_2^2 + \alpha \|\boldsymbol{\beta}\|_1 \right]$$

ENET with automatic parameter search will help to decide to use LASSO or Ridge or their combination. The best tune of this model is $\alpha = 0$, $\lambda = 0.30$, suggesting that the model is actually a Ridge regression model, although it still reduced some terms' (`jitter_ppq5`, `shimmer`, `shimmer_d_b`) coefficients to 0.

- *Generalized Additive Model*: GAM enables automatical capture of non-linear relationships between predictors and response, while maintaining the addictive structure of predictors. The fitting result displayed complex relationships between each predictor and the response, except for `jitter_rap` and `jitter_ppq5`, whose exponential were set to be 0 and, therefore, removed from the model.
- *Multivariate Adaptive Regression Splines*: MARS generate a piecewise linear model by generating hinge functions of original predictors as new predictors. The algorithm of MARS automatically choose cut points for predictors. The best tune of the MARS model was achieved when degree equals 2 (the product of at most 2 hinge functions was allowed to be a predictor) and 21 terms, generated from 4 original predictors, `age`, `sex`, `hnr` and `dfa`, were included in the model.
- *Support Vector Regression* : Support vector regression shares similar principles with SVM, but it is applied for solving the regression. It decides a decision boundary at a distance from the original hyperplane ensuring the data points are closest to the hyperplane. Unlike other Regression models, SVR tries to minimize the coefficients and fit the best line within a threshold value instead of minimizing the squared error. A hyperparameter of SVR is cost C: when the C increases the penalty term in the model's loss function decreases. Here we decided to use the radial kernel, which provides non-linear decision boundaries, and therefore introduced another parameter sigma, which controls the degree of curviness of the decision boundary. In our case, the best parameter is obtained with sigma = 0.51, C = 4.24. It is also noted that there are several limitations to this model: (1) it is not suitable for large datasets, (2) it doesn't perform well when there is noise.
- *Random Forest*: As a bagging model, Random Forest models achieve more accurate predictions by averaging the outputs from numerous trees which are built on bootstrapped data sets. Instead of including lots of correlated data, random forest builds decorrelating trees by randomly selecting predictors as split candidates. In this case, the best random forest model under 10-fold cross-validation is obtained with a minimum node size of 2 and a mtry equals to 18 (the numbers of predictors that were selected as split candidates). The weakness of this model lies in the long training time, the requirement for computational resources and the difficulty in interpreting it.
- *Gradient-boosted Trees*: In the GBM model, new trees are grown sequentially by using the information from the previously grown tree. GBM is highly flexible as it contains several parameters, which are the number of trees, the depth of trees, the shrinkage and the minobsinnode respectively. The number of trees refers to the optimal number of trees needed that minimizes the loss function of interest under cross-validation. Depth of trees refers to the number d of splits in each tree, controlling the complexity of the boosted ensemble. Shrinkage regulates the speed of the algorithm proceeding down the gradient descent. A small value of shrinkage reduces the chance of overfitting. However, it also increases the time to search for the best fit. The parameter minobsinnode refers to the minimum observations of the data contained in tree nodes. In our case, the best combination of parameters are tree size = 3000, depth = 18 and shrinkage = 0.012. Since this model is also an ensemble model like RF, it shares the same weakness with RF.
- *Linear Mixed Effects Model*: Linear Mixed Effects Model (LMM), which is often used in longitudinal analysis, is an extension of simple linear model to allow the modeling of both fixed effects and random effects. Because our data is composed of thousands of recordings collected from 42 subjects, subject heterogeneity may exist and affect the response variable and the modeling outcome. Therefore, a subject-specific intercept model was trained under the assumption that subjects have different baseline motor UPDRS scores. Features used in this model are those selected in the linear regression section.

3.2 Train/Test Error

The figure below displayed the 10-fold cross-validated RMSE of the 8 models.

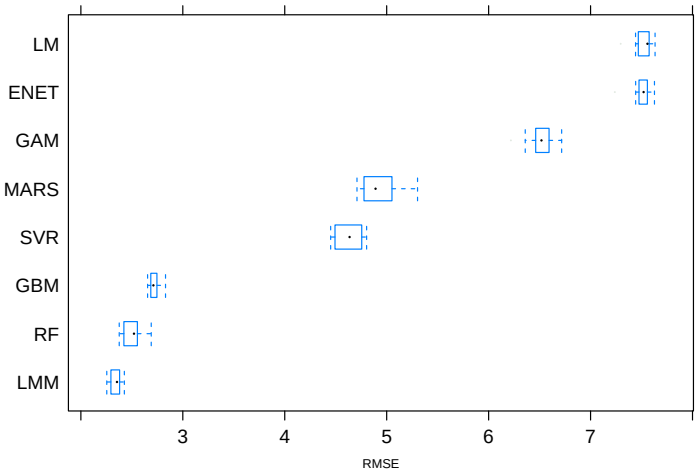


Figure 3. 10-fold CV RMSE of models.(LMM model showed the lowest RMSE.)

Model RMSEs on the test dataset were also in line with the result obtained in CV: LM: 7.58; ENET: 7.59; GAM: 6.62; MARS: 4.92; SVR: 4.62; GBM: 2.67; RF: 2.46; LMM: 2.36.

Among the 8 models, the LMM model displayed the best performance, followed by the RF and the GBM model. The result suggests that heterogeneity among the 42 subjects in the data significantly affected the prediction accuracy of models. Therefore, the LMM model, which has a structure much simpler than ensemble models but takes the heterogeneity into account by setting subject-specific intercepts, captured the underlying truth in the data and achieved the best performance. Two ensemble models, i.e. RF and GBM, displayed slightly larger RMSE with extremely complex model structures. (Best tunes of these models suggest interaction terms involving various original predictors were included.) Such complexity in model structure might be a compensation for ignoring subject heterogeneity in their model assumptions. Other models were obviously outperformed by these 3 models, as they neither considered heterogeneity nor complex enough to compensate for it.

3.3 Variable Importance

We choose the LMM model as our final model because of its smallest RMSE and its simple linear structure. The RF model, although reached a similar RMSE, are unnecessarily complicated compared to the LMM model and thus discarded. (Individual Conditional Expectation Curves were additionally used to interpret the Random Forest model (Appendix 5.4)). The table below displayed coefficients of predictors in the model. For the same subject, with 1 unit increasing in a predictor, the subject's estimated UPDRS score will increase by the value of the coefficient of this predictor correspondingly. In addition, under a significance level of 90%, we found that `age`, `jitter_ddp`, `hnr` and `dfa` are significant predictors.

The inclusion of `age` is consistent with the phenomenon that the incidence of Parkinson's disease increases with age. `jitter_ddp` measures vocal tremors that are possibly induced by trembling or shaking of one or more of the muscles of the larynx (the voice box), and `dfa`, `hnr` are measurements of noises, which are often caused by turbulent airflow in the glottis (a specific position in the throat), and resulted from incomplete closure of the vocal folds. As Parkinson's disease indeed has an impact on patients' vocal-related muscles, it is plausible that these terms were significant.

Table 1: LMM Model Coefficients

	Value	Std.Error	p-value
(Intercept)	4.9980230	8.4422059	0.5538605
age	0.2549588	0.1288578	0.0547749
jitter_ddp	15.8507121	6.5905915	0.0162090
shimmer_apq5	-6.4191655	8.1973478	0.4336210
shimmer_apq11	-2.2803617	6.1600732	0.7112621
hnr	0.0499741	0.0215786	0.0206068
dfa	-2.3650499	1.0695908	0.0270723
ppe	0.2604039	0.7972065	0.7439507

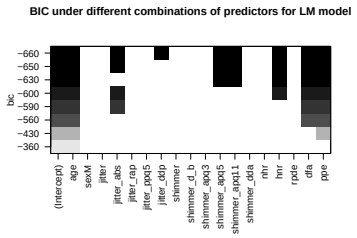
4 Conclusions

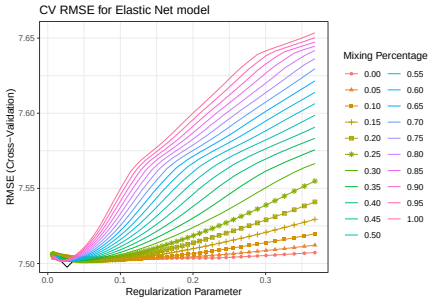
In this modelling analysis, we trained models with Linear Regression, Elastic Net, Generalized Additive Model, Multivariate Adaptive Regression Splines, Support Vector Regression, Random Forest, Gradient-boosted Tree Model and Linear Mixed Effects Model based on Parkinson's disease patients' age, sex, and vocal features. Due to the heterogeneity in subject baseline UPDRS scores, the LMM model best captured the underlying structure of the data. In the model, patients' age and features related to vocal noise and tremor are significant predictors for motor UPDRS score prediction, which reflects the severity of symptoms. The model has achieved a mean 10-fold cross-validated RMSE of 2.34 on the training dataset and an RMSE of 2.36 on the test dataset. Considering the scale of motor UPDRS score, which ranges from 0 to 108, the model gives highly accurate predictions and reached our expectations.

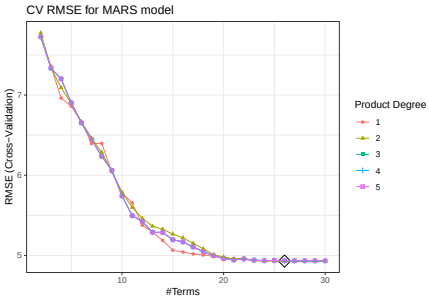
The LMM model provides information on both patterns of change in the mean UPDRS score associated with changes in the features in the population of this study (coefficients) and each subject's deviation from the population mean intercept (subject-specific intercept). However, there are still some limitations of this model. At first, for a new subject, repeated measuring is needed to estimate the subject's specific intercept in the model, which may bring inconvenience. Besides, the study population in the data only contains 42 patients, making it a relatively small sample. The small sample size may not honestly present the baseline vocal characteristics distributions of the patient population or the overall population, leading to bias in applying this model to a more general population. To fix this issue, more vocal data from patients and healthy individuals should be collected and used in training to achieve a more generalizable model.

5 Appendix

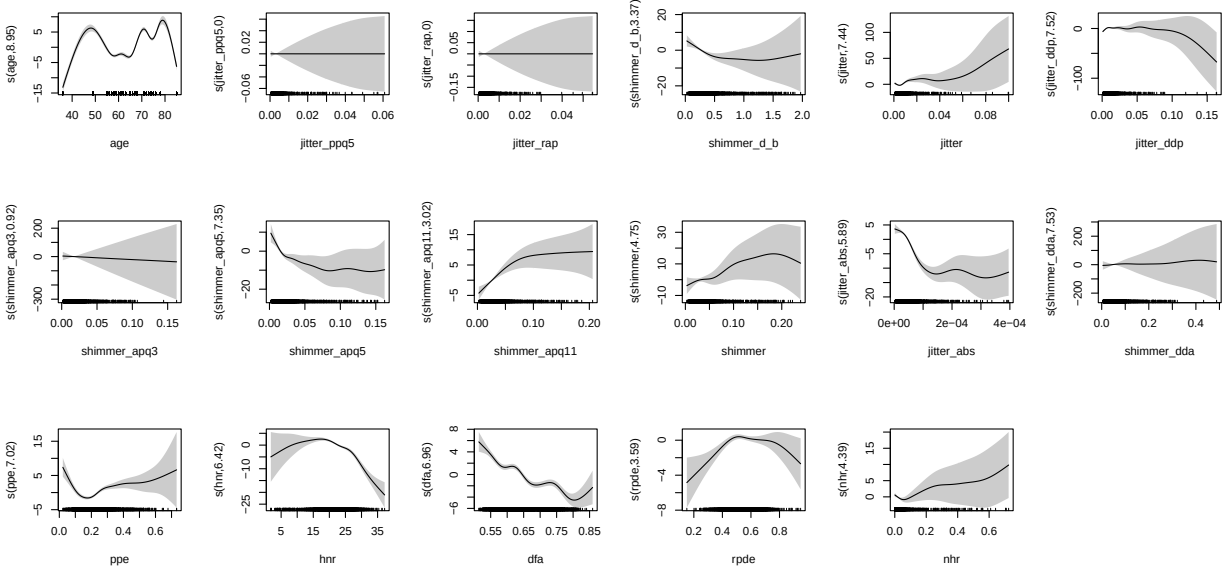
5.1 Parameter Selection

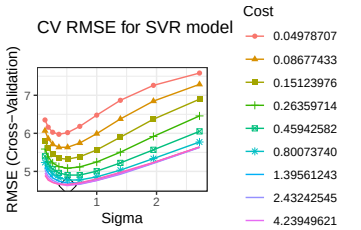


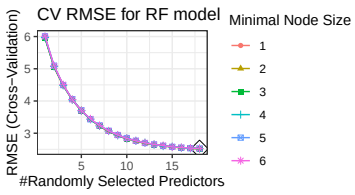


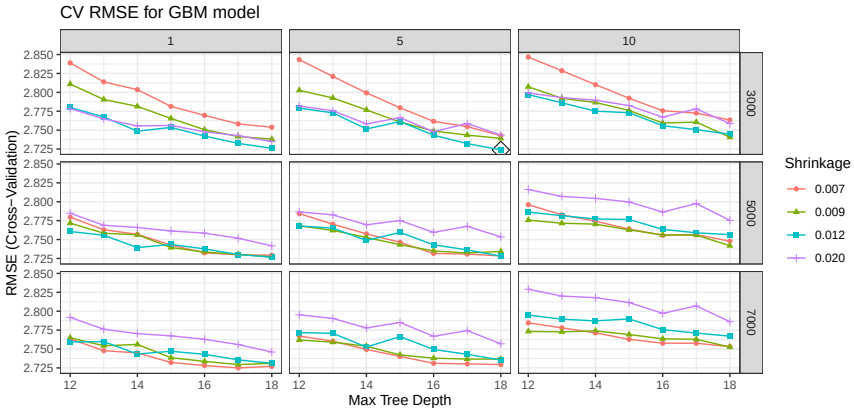


Terms used in GAM









5.2 Model Coefficients/Terms/Interpretation

5.2.1 Coefficients for Linear Regression Model and Elastic Net Model

term	LM coefficient	ENET coefficient	term	LM coefficient	ENET coefficient
(Intercept)	28.27	28.59	jitter	0.00	11.22
age	0.20	0.19	jitter_rap	0.00	306.5
jitter_abs	-48592.04	-55347.32	jitter_ppq5	0.00	0
jitter_ddp	111.34	52.98	shimmer	0.00	0
shimmer_apq5	-211.27	-113.42	shimmer_d_b	0.00	0
shimmer_apq11	131.04	98.29	shimmer_apq3	0.00	-24.31
hnr	-0.37	-0.36	shimmer_dda	0.00	-9.07
dfa	-21.23	-22.68	nhr	0.00	-10.33
ppe	17.56	16.70	rpde	0.00	2.23
sexM	0.00	-1.11			

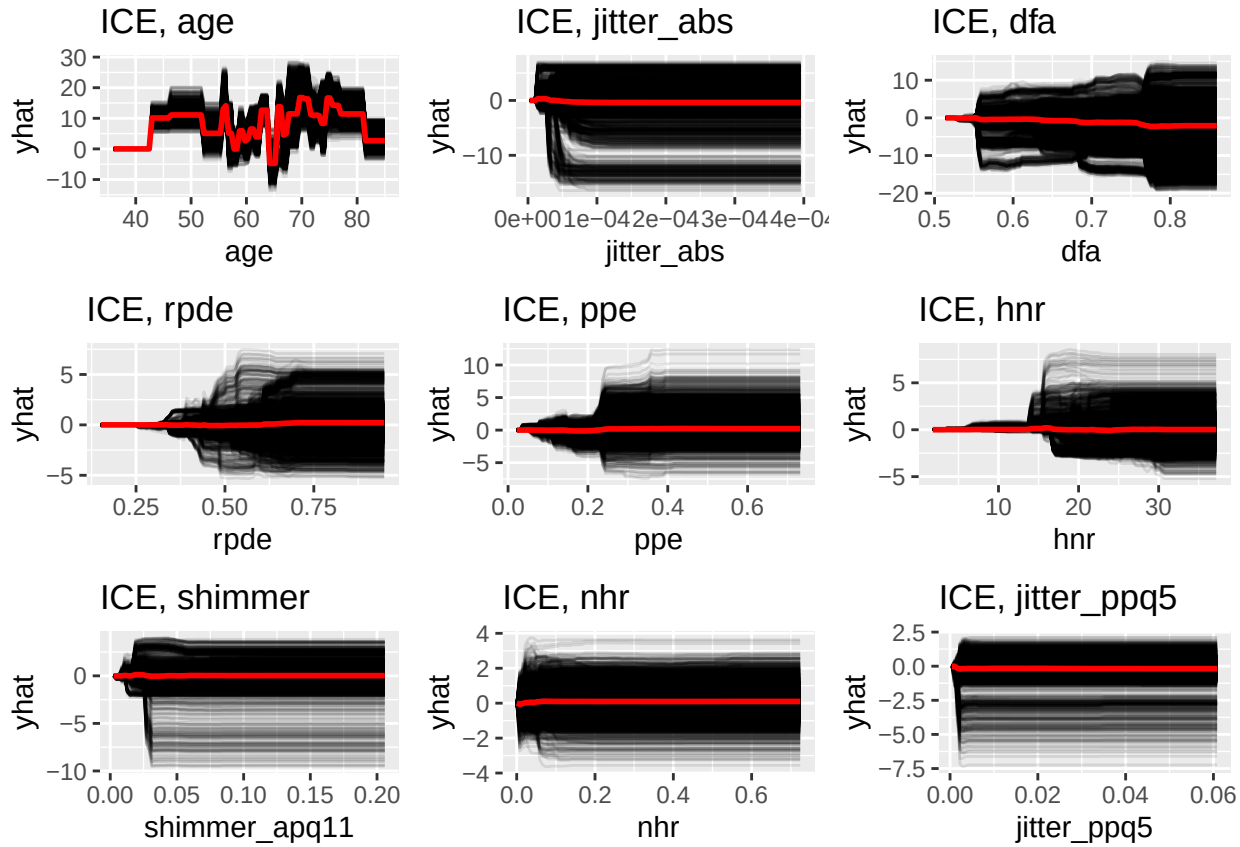
5.2.2 Terms and Coefficients for MARS Model

term	coefficient	term	coefficient
(Intercept)	74.60	h(age-63)	-8.05
h(age-76)	2.16	h(age-58)	2.67
h(76-age)	-1.60	h(age-65)*h(dfa-0.7473)	64.54
h(age-65)	30.22	h(age-65)*h(0.7473-dfa)	-44.72
h(age-49)	-3.03	h(age-71)	-17.65
h(age-68)	-24.40	h(age-72)	16.28
h(76-age)*h(dfa-0.58972)	-1.86	h(age-68)*h(dfa-0.57801)	111.47
h(76-age)*h(0.58972-dfa)	2.43	h(age-68)*h(0.57801-dfa)	-277.84
h(76-age)*h(hnr-26.707)	-0.06	h(age-74)*h(dfa-0.7185)	-348.48
h(76-age)*h(26.707-hnr)	0.00	h(age-74)*h(0.7185-dfa)	40.18
h(age-74)	8.09	h(age-66)*h(dfa-0.5788)	-125.42
h(age-66)	-21.90	h(age-66)*h(0.5788-dfa)	266.74
h(age-67)	29.18	h(age-75)	-12.72

5.2.3 Subject-specific Intercept for LMM Model

subject	intercept	subject	intercept	subject	intercept	subject	intercept	subject	intercept	subject	intercept
1	8.36	8	-3.56	15	-7.56	22	-9.67	29	0.01	36	3.71
2	-5.22	9	-3.38	16	-12.21	23	-6.12	30	8.89	37	12.83
3	7.69	10	-6.10	17	4.81	24	-5.46	31	2.97	38	-1.86
4	-7.29	11	0.28	18	-15.46	25	4.82	32	-3.84	39	8.31
5	8.03	12	-3.48	19	-0.89	26	8.42	33	5.01	40	-9.70
6	6.97	13	-3.51	20	-10.23	27	-8.23	34	5.31	41	12.52
7	-6.97	14	-6.54	21	6.11	28	5.84	35	13.70	42	2.72

5.3 Black Box Model Interpretation



Individual Conditional Expectation curves of the top 10 important variables were ploted for the Random Forest model. Each plot shows the relationship between one feature and predicted motor UPDRS score for each observation separately. For the most important variable, `age`, the curves of individual observations fluctuate around the average line, which is due to the small sample size of subjects. Other variables have negative and positive responses of various different scales, which implies high correlations among variables and the existence of complex interaction term in the model.

5.4 Vocal Feature Description

Feature	Description
jitter	the extent of variation in speech frequency from vocal cycle to vocal cycle
jitter_abs	absolute jitter in microseconds
jitter_rap	relative amplitude pertubation of jitter
jitter_ppq	5-point period perturbation quotient of jitter
jitter_ddp	average absolute difference of differences between cycles, divided by the average period
shimmer	the extent of variation in speech amplitude from cycle to cycle
shimmer_db	shimmer in decibels
shimmer_apq3	3-point amplitude perturbation quotient of shimmer
shimmer_apq5	5-point amplitude perturbation quotient of shimmer
shimmer_apq11	11-point amplitude perturbation quotient of shimmer
shimmer_dda	average absolute difference between consecutive differences between the amplitudes of consecutive periods
nhr	noise-to-harmonics ratio
hnr	harmoincs-to-noise ratio
rpde	recurrence period density entropy, measuring vocal vibration
dfa	detrended fluctuation analysis, measuring the extent of turbulent noise
ppe	pitch period entropy, measuring vocal pitch instability

The above table clearly describes all the predictors (vocal features) that we included in the model. There are a total of 16 predictors included to predict for the response motor_updrs.