**Modeling the Parkinson’s Disease Telemonitoring Data**

**Abstract**

I have started with doing the basic data check. Later, I visualized the data to see if the patterns and correlations. After this, I created a multiple regression model which is the statistical method to predict the numerical data. For the dataset I have chosen, the total UPDRS score is the numerical feature that is to be predicted. The feature set that has the best performance on predicting is considered as the final model.

**Introduction**

Neurological disorders largely affect the lives of patients and their families. These also include Parkinson’s disease, Alzheimer’s, and epilepsy. Over one million people are affected by Parkinson’s disease in North America [2]. For many people with Parkinson’s disease, physical visits to the clinic are difficult. The remote monitoring of symptoms looks like a feasible option in this case. The dataset that I chose contains various biomedical voice measurements from 42 people having early-stage Parkinson's disease. These people were recruited to a six-month trial of a telemonitoring device for monitoring symptom progression. This monitoring was done in a remote manner to reduce cost and ensure physical convenience. Parkinson’s disease symptom progression is tracked using the Unified Parkinson’s Disease Rating Scale (UPDRS) [1].

**Research questions**

* We are supposed to predict the total UPDRS score. For this, we will create a multiple linear regression model. We need to check which features are correlated with the total UPDRS score. This means we check whether there is any feature that increases or decreases along with an increase or decrease in the value of the total UPDRS score.
* We need to check which features out of a total of 16 features contribute to the prediction of this feature (total UPDRS). They are also called significant features. For this, we will do exploratory data analysis along with building multiple regression models with different combinations of features. We will use r-squared and MSE which are statistical terms to determine the best model.

**Methods**

The methodology is broken down into dataset description, data analysis, modeling, validation.

**Dataset description:**

* Columns in the table contain subject number, subject age, subject gender, time interval, motor UPDRS, total UPDRS, and 16 biomedical voice measures.
* Rows represent 5,875 voice recordings (vowel phonations). The phonations were recorded using head-mounted microphones [2].
* The data file can be found at: <https://archive.ics.uci.edu/ml/machine-learning-databases/parkinsons/telemonitoring/parkinsons_updrs.data>
* Information on the columns [3]:

1. subject# - Number that identifies each subject uniquely [3]
2. age – Age of subject [3]
3. sex - Subject gender '0' for male, '1' for female [3]
4. test time - The integer part is the number of days since recruitment [3]
5. motor UPDRS - Clinician's motor UPDRS score [3]
6. total UPDRS - Clinician's total UPDRS score [3]
7. Jitter (%), Jitter (Abs), Jitter: RAP, Jitter: PPQ5, Jitter: DDP - Several measures of variation in fundamental frequency [3]
8. Shimmer, Shimmer(dB), Shimmer: APQ3, Shimmer:APQ5, Shimmer: APQ11, Shimmer: DDA - Several measures of variation in amplitude [3]
9. NHR, HNR - Two measures of the ratio of noise to tonal components in the voice [3]
10. RPDE - A nonlinear dynamical complexity measure [3]
11. DFA - Signal fractal scaling exponent [3]
12. PPE - A nonlinear measure of fundamental frequency variation [3]

**Data analysis, modeling**

Part 1: Data analysis

Doing the basic data check:

1. Data has 5875 rows and 22 columns.
2. It has only numerical data.
3. The quantile values are close to each other this means the spread of the data is less, so the variance is less.

Sequences of features for pairplot is as follows:

'subject', 'age', 'sex', 'test\_time', 'motor\_UPDRS', 'Jitter\_percnt', 'Jitter\_Abs', 'Jitter\_RAP', 'Jitter\_PPQ5', 'Jitter\_DDP', 'Shimmer', 'Shimmer\_dB', 'Shimmer\_APQ3', 'Shimmer\_APQ5', 'Shimmer\_APQ11', 'Shimmer\_DDA', 'NHR', 'HNR', 'RPDE', 'DFA', 'PPE'

1. The pairplot of all variables plotted against total\_UPDRS indicate that similar pattern of relation is followed by few features.
2. The first two features do not have any particular pattern.

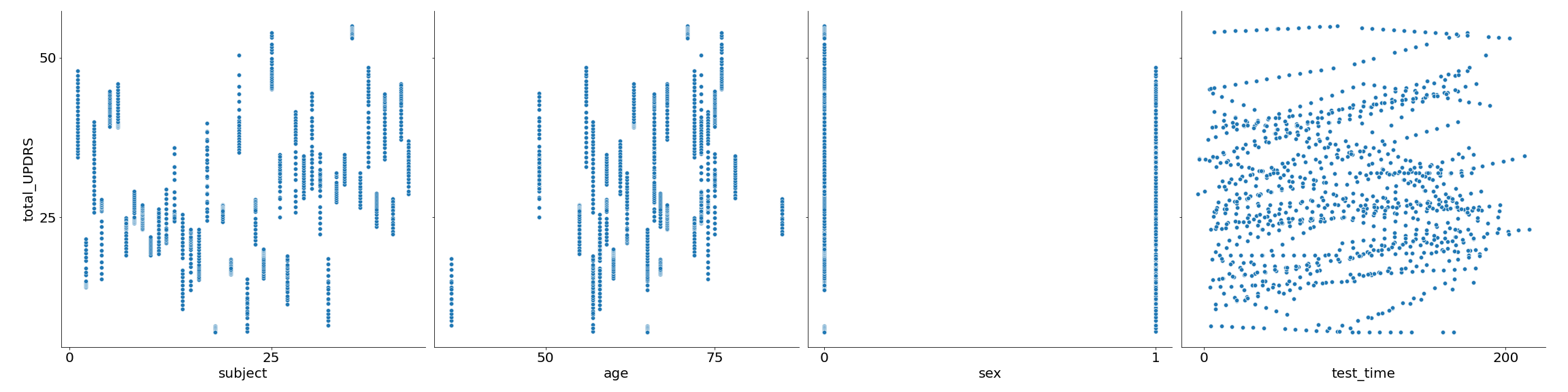
A picture containing calendar

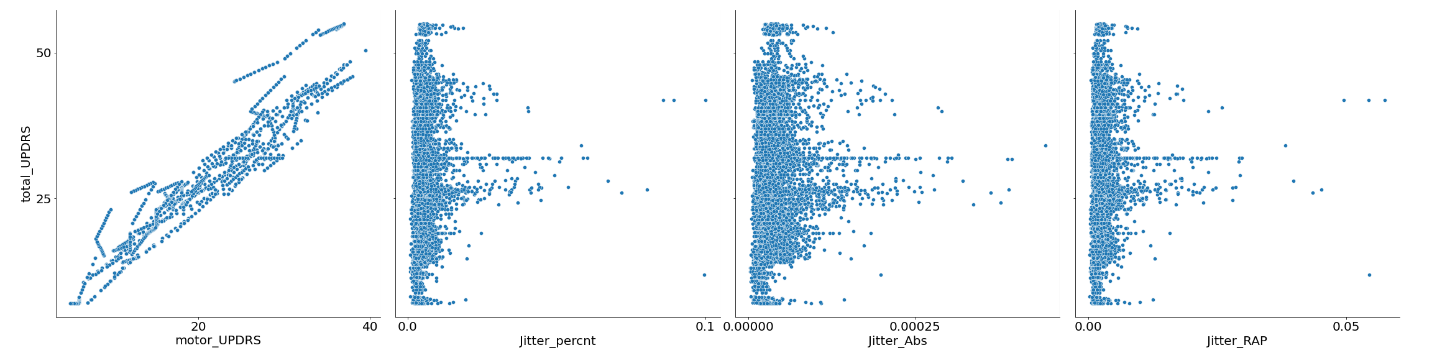
Description automatically generated

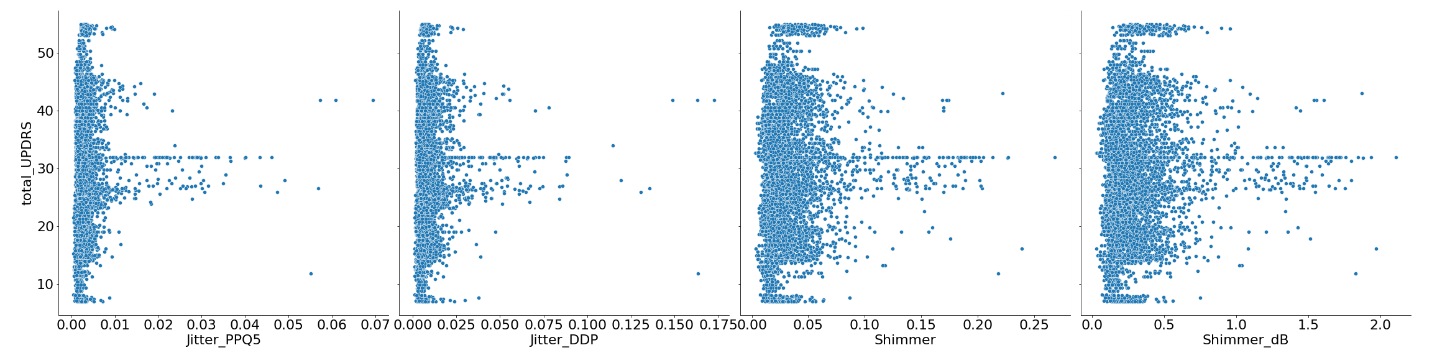
Figure: Heatmap showing correlation between all the features

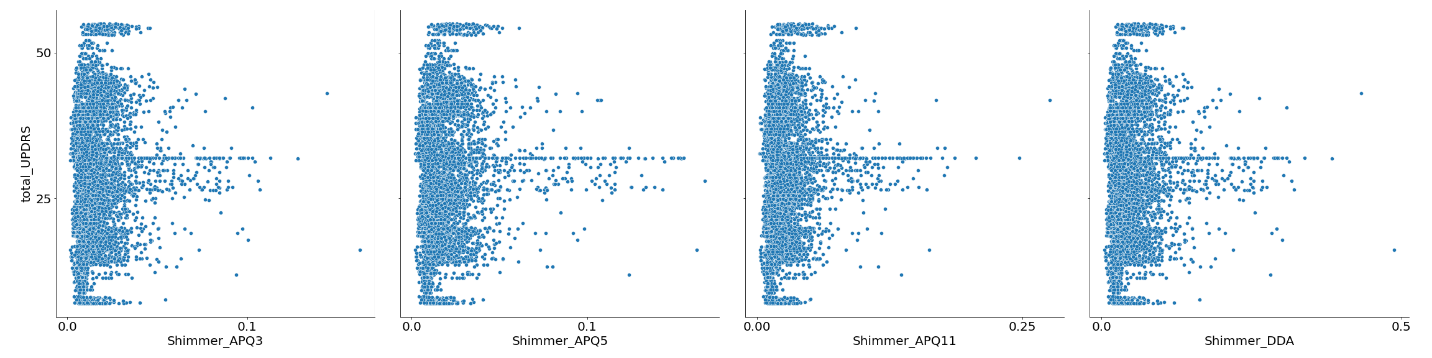
* We see the motor UPDRS is highly correlated with total UPDRS (dependent variable).
* All the variants of the jitter and shimmer feature except “Shimmer\_APQ11” have a relatively weaker correlation with total UPDRS.
* HNR, DFA, sex has a negative correlation with total UPDRS.

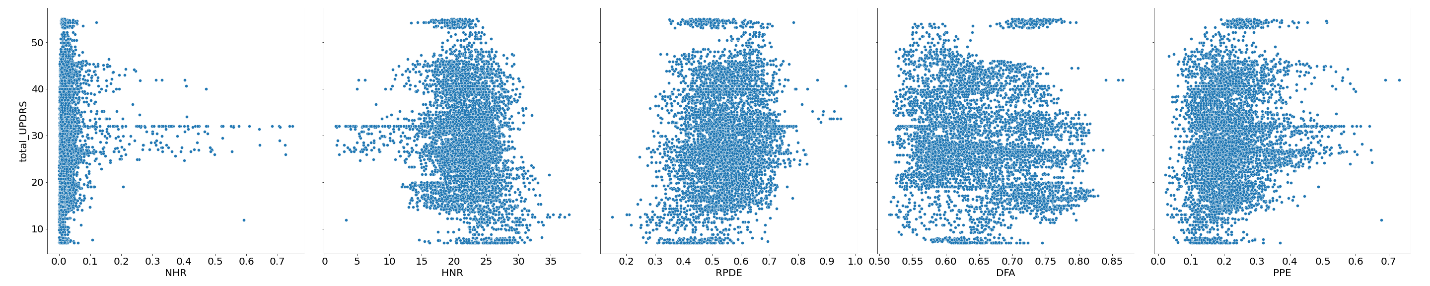
Given below are the pair plots of total UPDRS with all the features.











* Motor UPDRS has a linear relationship with total UPDRS.
* Other features have relationships that cannot be seen visually.

**Modeling and validation**

Since the data is numerical, we will use a multiple linear regression model.

First, we will start with considering all the features and building the model. After building the model we get to see how many features are insignificant.

The significant features left are subject, age, sex, test time, motor UPDRS, Jitter Abs, Shimmer, Shimmer\_APQ5, Shimmer\_APQ11, NHR, HNR, RPDE, DFA, PPE

After removing insignificant features, we again proceed with creating a model with the remaining features.

We incorporate interaction terms to see the significance of features concerning one another. This confirmed that our features are significant.

To build a model we take 70 percent of the data and use it for the training of the model. We later select 30 percent of the total data to check if the model is performing well.

We perform cross-validation which is the concept of considering different subsets of data so that we do not end up modeling on the biased set of data. When there is bias, our model tends to overfit or underfit the test set.

Talking about the features, for training the multiple regression model along with cross-validation, we select a set of features and see if there is an increase or decrease in the performance and if there is no change in performance after removing a feature, it means that the particular feature does not contribute much in predicting the total UPDRS. Else, if there is an increase in model performance after removing a particular feature, we will discard the feature.

Following the above steps, the number of features came down to 12. These are the features that help in predicting the total UPDRS score better.

These features are:

'subject', 'age', 'sex', 'motor UPDRS', 'Shimmer', 'Shimmer\_APQ5', 'Shimmer\_APQ11', 'NHR', 'HNR', 'RPDE', 'DFA', 'PPE'

The model does not overfit. This means we avoided bias.

The r-squared value which is the metric to evaluate the performance is 81 percent which is better out of all the other models.

**Discussions**

In this, we have established a relation between the total UPDRS and other important features. To get a better result it is suggested that we use CART (classification and regression tree) and IRLS (iteratively re-weighted least squares) [1].

**Conclusion**

After getting the final model, we see that for each unit change in total UPDRS, motor UPDRS increases by 1.2. Shimmer, Shimmer\_APQ11, NHR, HNR, DFA, PPE witness a negative change with an increase in total UPDRS. We got the model r-squared of 91 on the test set and 90 on the train set.

**References**

1. Athanasios Tsanas, Max A. Little, Patrick E. McSharry, Lorraine O. Ramig (2009),  
   'Accurate telemonitoring of Parkinson’s disease progression by non-invasive speech tests',  
   IEEE Transactions on Biomedical Engineering (to appear).
2. Max A. Little, Patrick E. McSharry, Eric J. Hunter, Lorraine O. Ramig (2009),  
   'Suitability of dysphonia measurements for telemonitoring of Parkinson's disease',  
   IEEE Transactions on Biomedical Engineering, 56(4):1015-1022
3. <https://archive.ics.uci.edu/ml/machine-learning-databases/parkinsons/telemonitoring/parkinsons_updrs.data>