Abstract

The second project objective requires that your team predict the motor and the total UPDRS scores assigned by a physician to people with Parkinson’s Disease

HIT140

FOUNDATIONS OF DATA SCIENCE

Contents

[Introduction 2](#_Toc147345112)

[Task 1 2](#_Toc147345113)

[Task 2 4](#_Toc147345114)

[Task 3 5](#_Toc147345115)

[Task 4 6](#_Toc147345116)

[Visualizations 7](#_Toc147345117)

[Line Graphs 7](#_Toc147345118)

[Box Plots 7](#_Toc147345119)

[Scatter Plot Total UPDRS 8](#_Toc147345120)

[Scatter Plot Mobile UPDRS 9](#_Toc147345121)

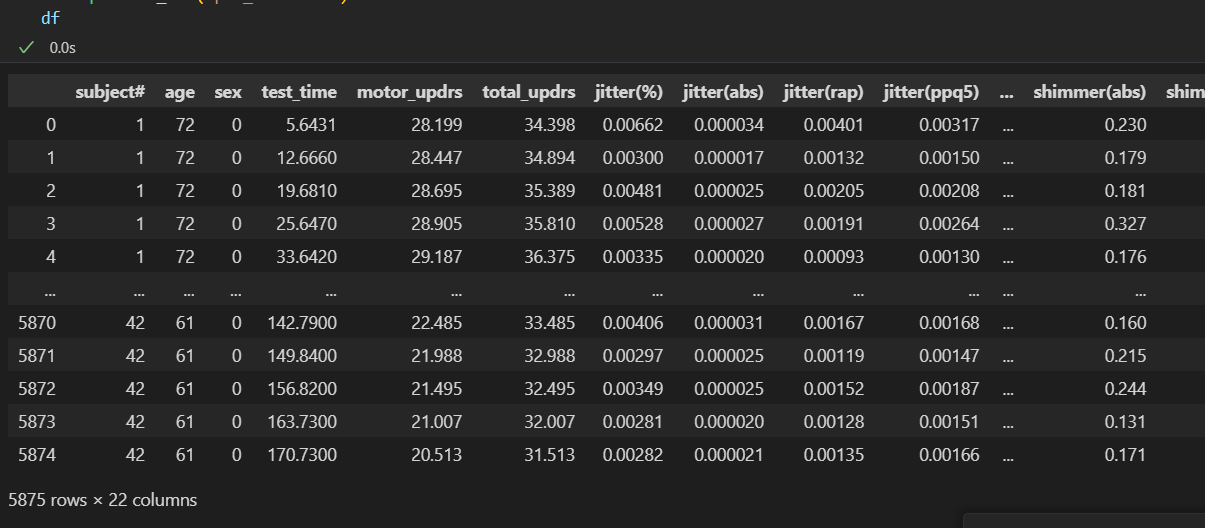
[Discussions & Limitations 10](#_Toc147345122)

[Conclusion 10](#_Toc147345123)

[Work Division 10](#_Toc147345124)

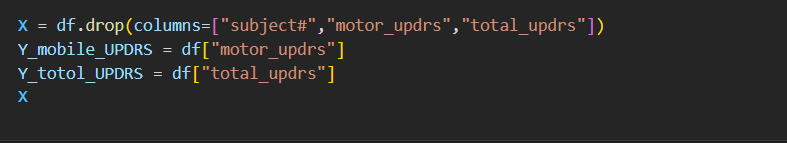
# Introduction

The objective of this assessment was to study and analysis the data of 42 subjects. Assignment 3 contained 22 \* 5875 records. It means 22 distinct features for estimation, approximation and prediction. Out of those 22, 1 is subject identifier that can not be used in any calculation. We had to create a linear regression to predict motor\_updrs, total\_updrs using factors such as age, sex, test\_time, jitter(%), jitter(abs), jitter(rap), jitter(ppq5), jitter(ddp), shimmer(%), shimmer(abs), shimmer(apa3), shimmer(apq5), shimmer(apq11), shimmer(dda), nhr, rpde, dfa, ppe.

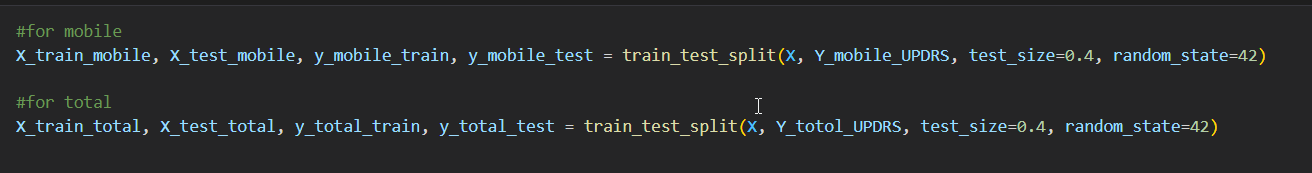


# Task 1

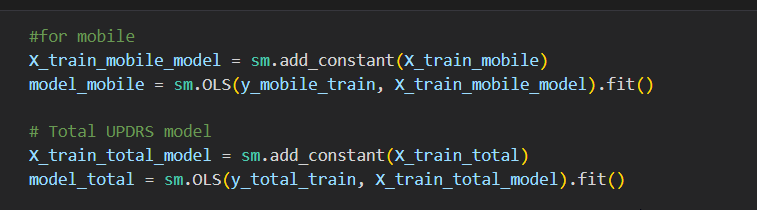
Before starting with task one. We had to separate the columns into 3 parts. Independent, Dependent, and columns that we were not going to use. Motor\_updrs, and total\_updrs were our two dependent columns. Subject# was the column that we were not going to use. Rest of the columns were independent.

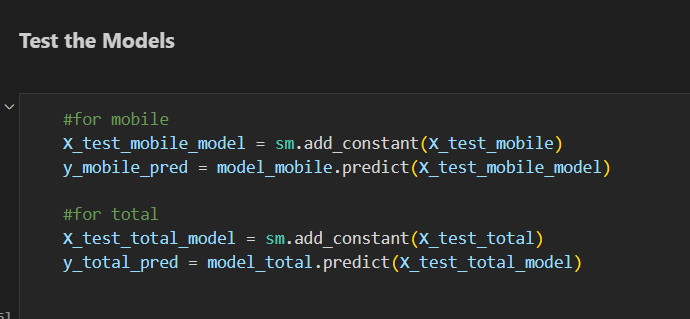


Then we had to create a train test ratio. For the first task, we chose the test size to be 40% of the train size. After doing train test split for motor\_updrs and total\_updrs.

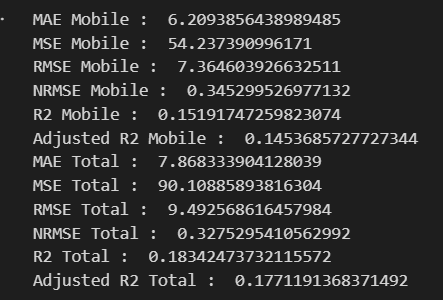


We have stated random state 42 to ensure data is randomly split between train and test set. Then we trained the model using that train-test data, and performed the prediction



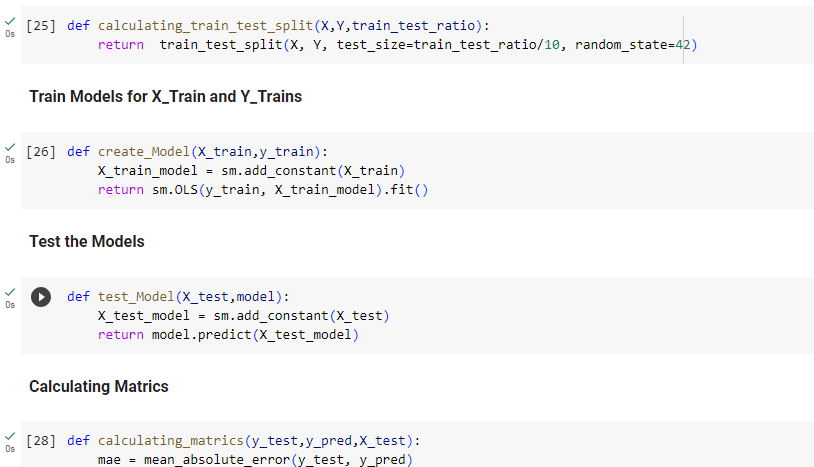


After the models had been tested, we printed the matrices. The results were not good mse, mae were very high. On the other hand, R^2 values were very low.

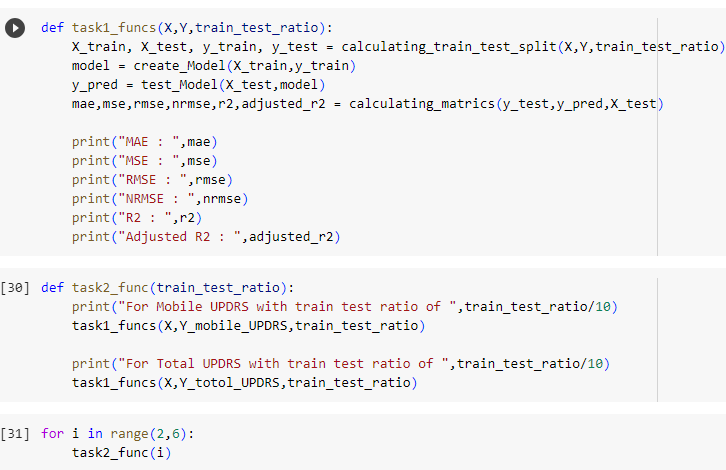


# Task 2

For Task 2 we had to use the pieces of code we wrote for task 1. There was slight modification. First, we convert all the pieces of codes into functions/modules for reusability.



Once the entire task 1 code base had been converted into modular/reusable code, we created another function to call them with different train test rations and reviewed the results.



There was slight change in performance metrics with different train/test splits, but overall the performance remained same as the improvement/deterioration was not enough to make the dent.



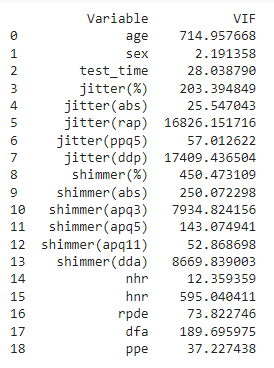




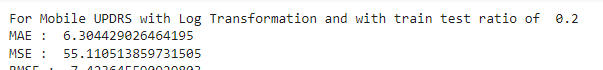


# Task 3

For task 3 we had to recreate the linear regression using log linear transform and do a collinearity analysis to identify which columns affected the total and mobile updrs most. Most of the columns had a very high VIF. This mean they played little to no effect in the prediction of the updrs.



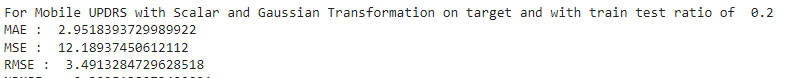
This indicates that majority of columns provide little to no effect. The higher the VIF the lesser the predictability using those columns.

Moreover, using Log linear transformation did very little to improve our results. 

These results are almost the same as those with normal linear regression.

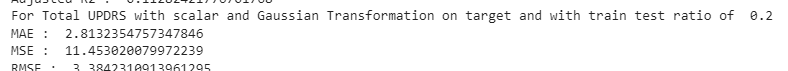
# Task 4

For task 4 we had to perform Gaussian transformation on the target and Standardization on the depending fields that are used for training the prediction model. This approach significantly increased the performance of the system.



It can be seen that mae, mse, rmse have significantly decreased. A decrease in them means an increase in the correct predictions.

As stated in the conclussion of task 3 Lower MAE, MSE, RMSE, and higher R-squared values indicate better performance.



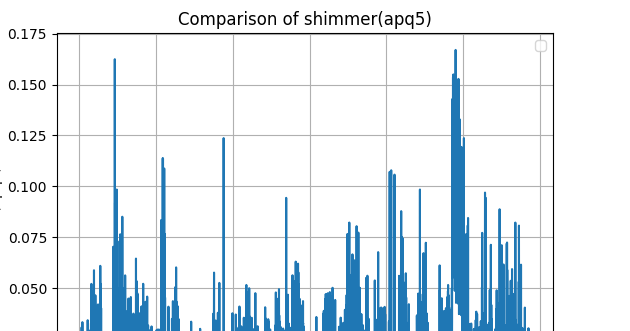
It can be seen there is a significant performance gain. when standardize the X and gaussian transform the target(y\_mobile\_updrs and y\_total\_updrs)

# Visualizations

We performed some visualizations just to get a clear idea of data division that is happening in the fields. For identification, we used four basic visualizations. First, we plot line graphs of each column. For second, we visualization them into boxplots. For 3rd we chose to draw scatter plot of each column against total updrs to identify its relationship. For 4th we chose to draw scatter plot of each column against Mobile updrs to identify its relationship.

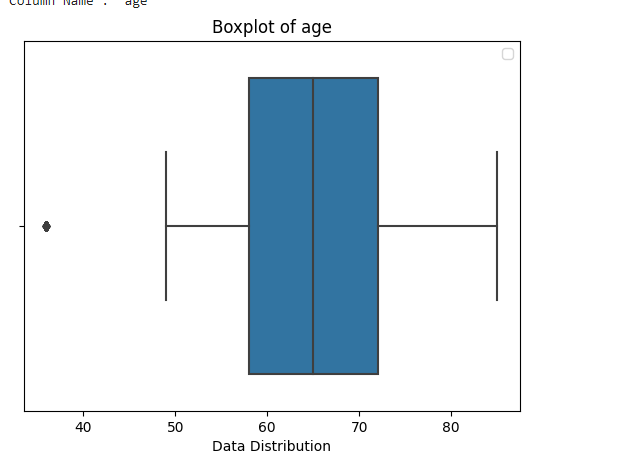
## Line Graphs

We performed this visualization to visualize the shape and spread of the columns. We observed that jitter and shimmer have a lot of variances, as some values are very high and some are very low. This can happen due to a lot of reasons, but these can possibly be used to predict the mobile and total updrs.



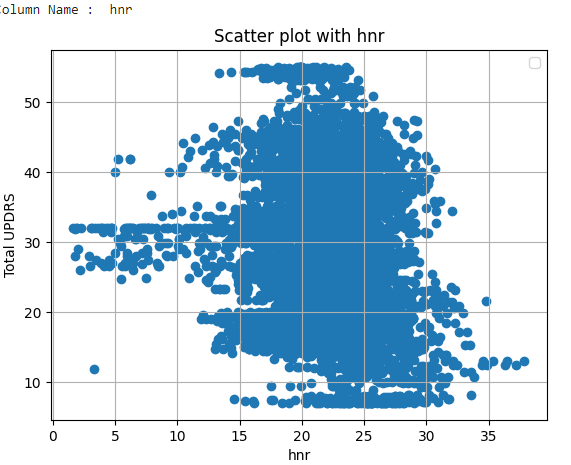
## Box Plots

The reason for performing a box plot visualization was the same as that of line graphs to identify the outliers. It is easier to identify an outlier using box plots. Box plot showed that jitter and shimmer had a lot of outliers. Fields such as age had only one outlier. Thus age can be used for the prediction of total and mobile updrs.



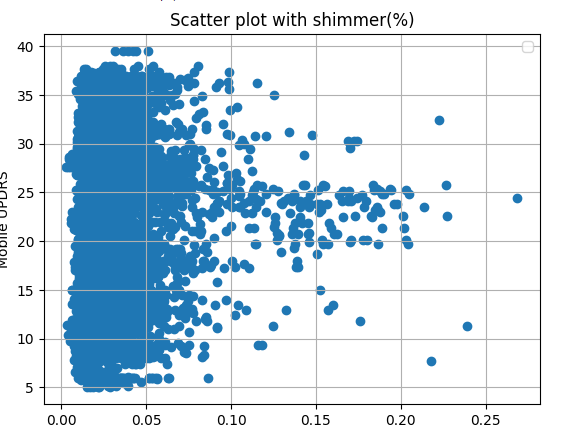
## Scatter Plot Total UPDRS

We built this to represent the interdependency between each column and Total updrs column. Results are some what messy and hard to interpret, but they can be used to identify field that give the most gain in the result.



## Scatter Plot Mobile UPDRS

We built this to represent the interdependency between each column and Mobile updrs column. Results are somewhat messy and hard to interpret, but they can be used to identify field that give the most gain in the result.



# Discussions & Limitations

The time and the information provided for the research were limited. There must be a lot feature that can be used for the prediction of Mobile and Total Updrs. Parkinsons Disease can be cause due to various genetic and environmental markers. Those markers were not mentioned in the research or in the building of linear regression.

# Conclusion

In conclusion, the provided fields do yield some result, but according to the performance metrics the errors are very high to make the field of much use. It is concluded that more parameters to be identified that yield a more gain in the result.

# Work Division

The importing libraries part was done by everyone. Whenever, someone performed their tasks, they inserted their used libraries. All of us were present in every meeting. We all are friends, so we have each other’s cell numbers and we don’t hesitate to call each other for help. Chander took the role of the leader and noted down the meeting minutes and created github repository for code collaboration. Chander imported the data for and printed it for verification. Then we had a meeting to discussion how we were going to approach the Assignment. Chander performed the task 1 and Nawal reviewed it. Nawal refactored Chander’s code and performed task 2. Vinay and Wahab worked closely together having constant discussions on task 3, and task 4. Vinay plotted box plots and line graphs and Wahab verified them. Wahab built the scatter plot for total and mobile updrs. Nawal and Chander worked closely on the report, fixing and correcting and building upon each other’s work. Wahab and Vinay then reviewed it, and made some last minute fixes.