Predicting the Number of OFF Periods Per Week for Parkinson's Disease Patients Leveraging Statistical Learning

Adil Baran Narin[1], Rahul Kamal Pandey[1], Abhishek Shirsat[1], Sayanti Mukherjee[2]

[1]Graduate Students; Department of Industrial and Systems Engineering; University at Buffalo, SUNY Buffalo, NY 14260, USA

[2]Assistant Professor; Department of Industrial and Systems Engineering; University at Buffalo, SUNY Buffalo, NY 14260, USA

Introduction

National Institute of Neurological Disorder and Strokes states that Parkinson's disease (PD) belongs to a group of conditions called motor system disorders, which are the result of the loss of dopamine-producing brain cells. Machine learning is making huge strides in understanding various aspects of PD, whose causes and symptoms remained mystery. In this study, we will discuss the initial exploratory data analysis, required data transformation and various regression models that can be used to make requisite predictions.

Research Needs and Objective

- Currently, the diagnosis made by doctors to assess the severity level of PD is conducted using various methods based on several research domains, including cognitive deficits, speech disorders, human stability, gait cycle and others [1][2].
- More than a million people in US are living with the symptoms of PD, which get worse overtime. At present, there is no definitive cure for PD but a variety of medications provide dramatic relief from the symptoms.

It is evident that the lack of specific test for Parkinson's disease makes it challenging to diagnose PD subjects.

The objective of this study is to use supervised learning methods to predict the number of OFF periods for Parkinson's Disease patients based on symptoms and level of impact of PD on their lifestyle. This will help determine the level of severity of PD in a particular patient.

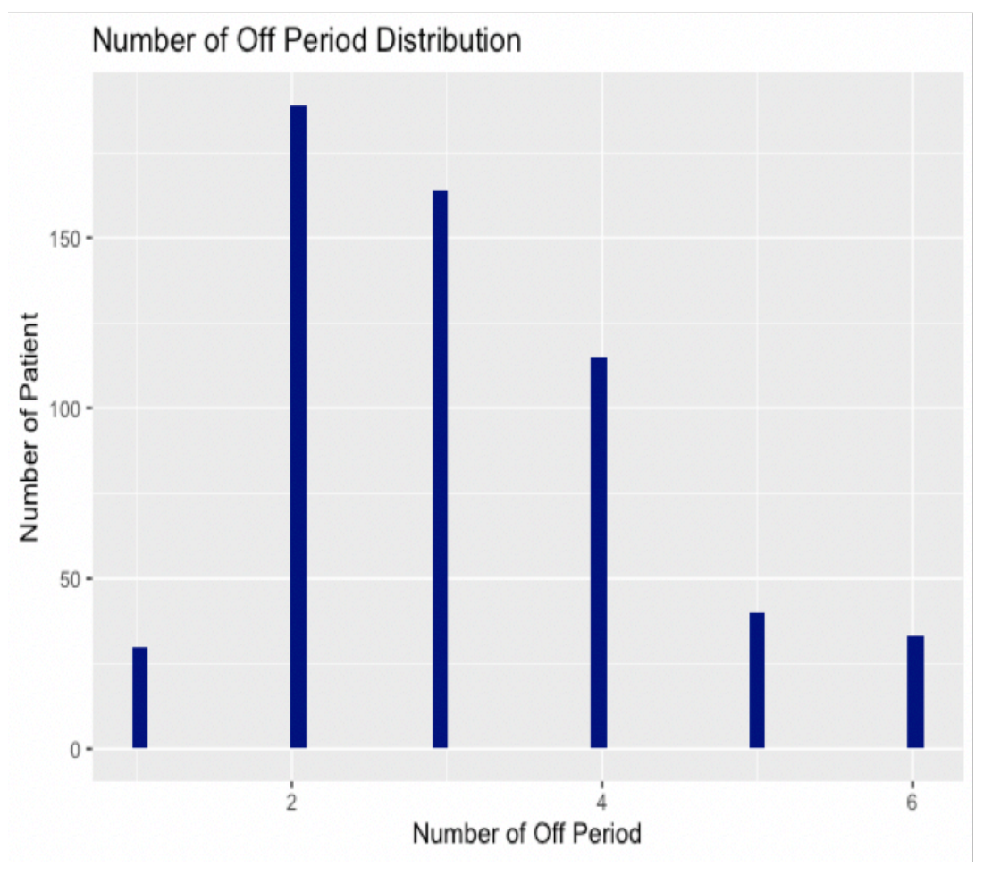
Data Analysis

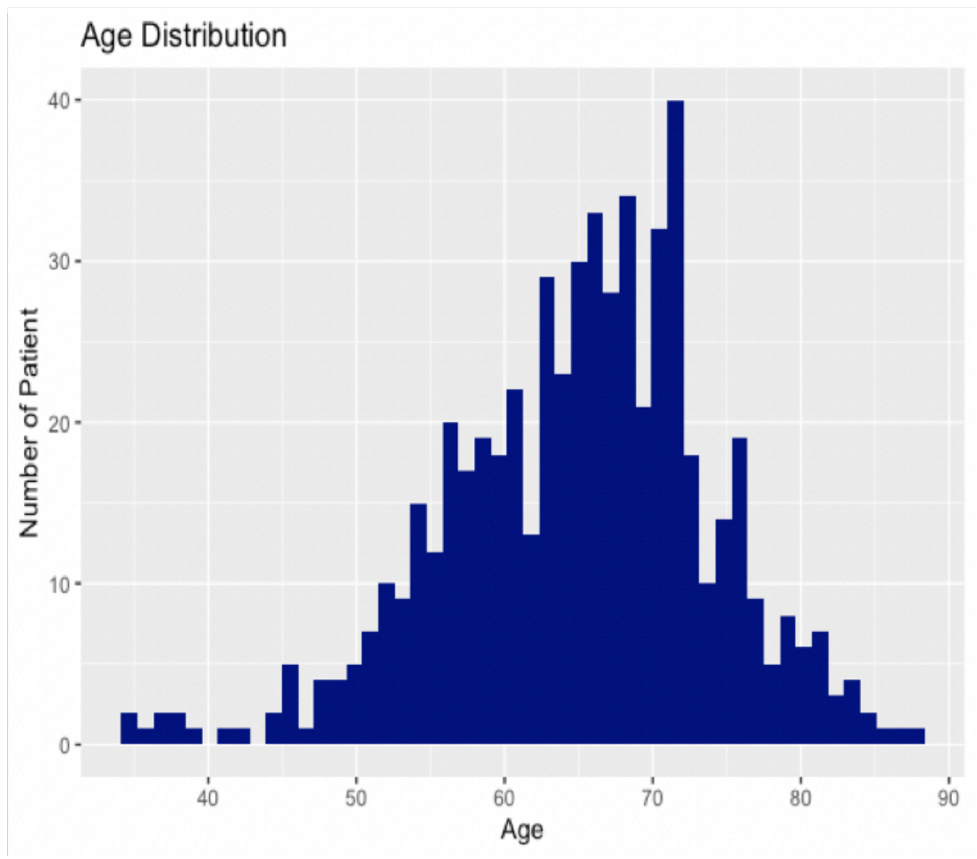
NGO and Government foundations support PD research by making the patient's data publicly available. Data analytics has opened a new realm od research of answer the previously unanswered questions pertaining to PD. We use supervised learning methods to develop various predictive models to predict the number of OFF periods for PD patients. An OFF period is defined as the period when a the PD symptoms return for a patient on CL medication. CL is considered to the gold standard of prescribed medication for PD. The initial exploratory data analysis shows us that most of the patients in study are 65 to 70 years old and they experienced 2 to 3 OFF periods/week.

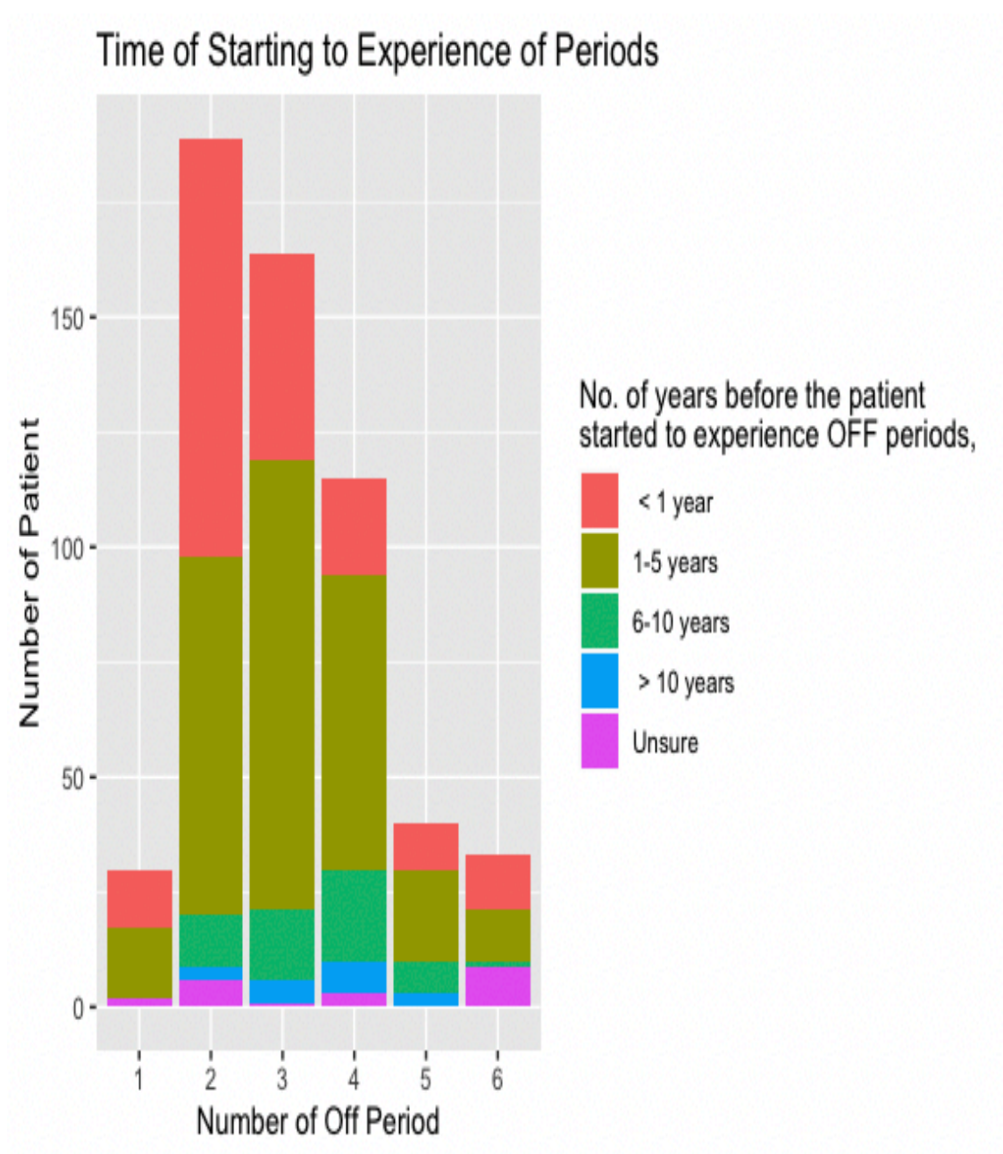
- The dataset is divided with 80% randomly chosen observations in training data set and remaining in test dataset.
- We train the training data set with a range of predictive models to predict our said response.
- The trained model is then tested on our test data set and model accuracy is assessed.
- MSE is the baseline for comparison since it tells us how close the regression line is to the points.

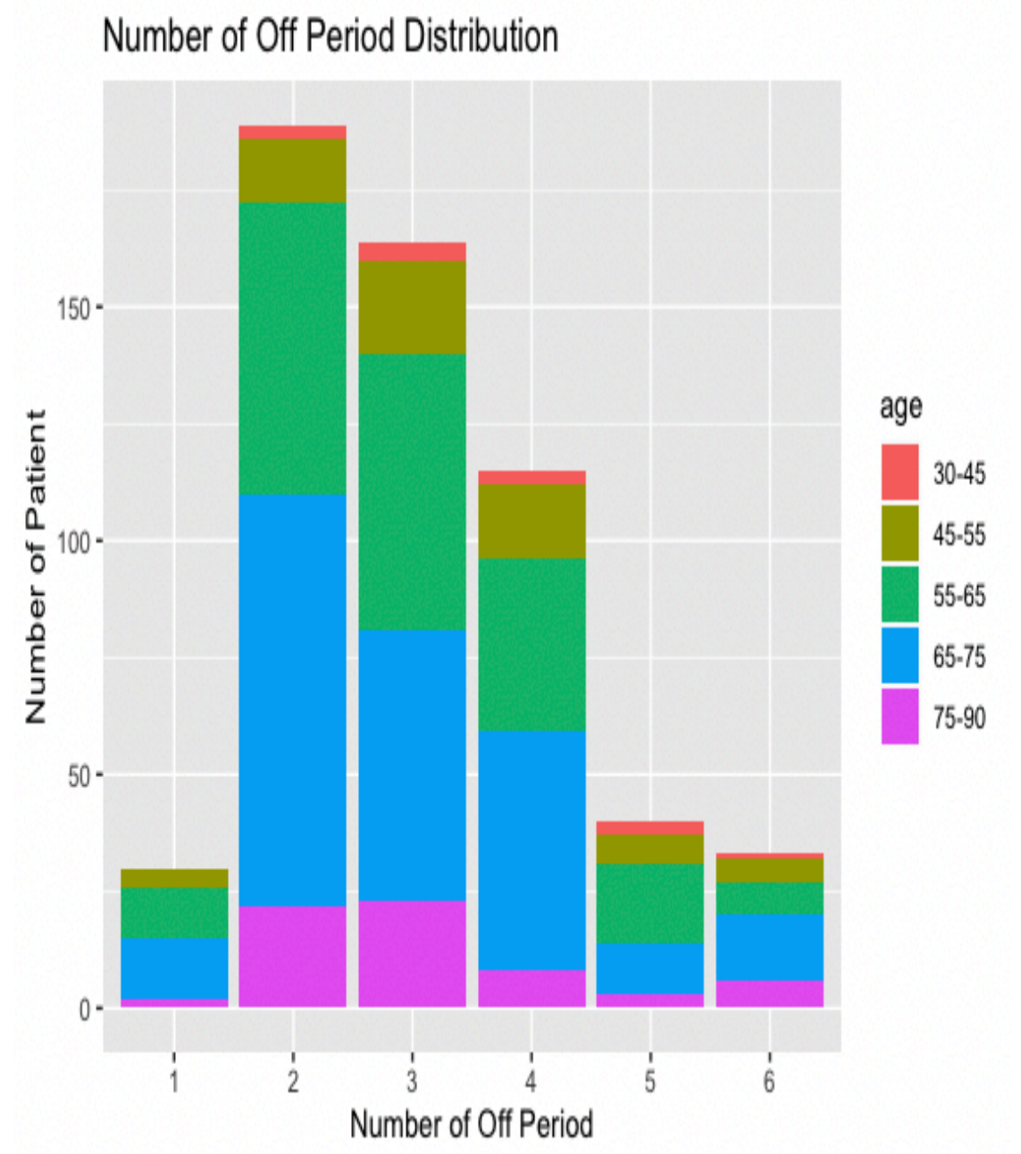
The different parametric and non-parametric predictive models developed were, Multiple Linear Regressions (MLR), Multivariate Adaptive Regression Splines (MARS), Forward, Backward and Best Subset Selection, Ridge and Lasso Regressions, GAM, Bagging, Random Forest and Boosting. The least MSE value obtained is for Bagging.

- In the bagging model, we start with 500 trees. After determining the optimal number of trees (170), the model is improved. Totally 37 variables are used in the model.
- The most significant variables are obtained. The first time of experiencing off periods has the most effect on MSE. Moreover, proportion of the unpredictable off periods follows as an important variable. The impact on physical activity is another factor for increasing MSE value.





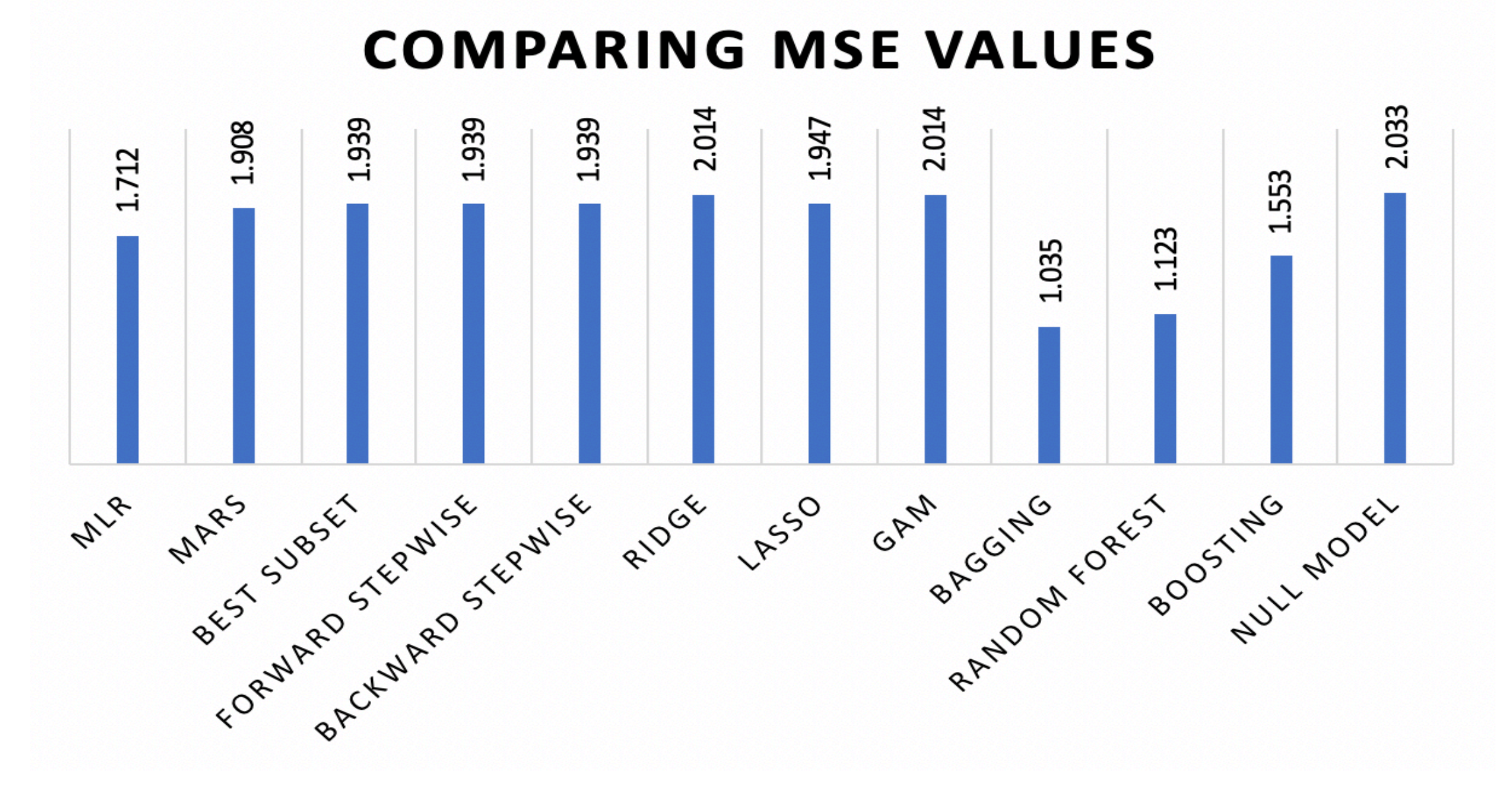




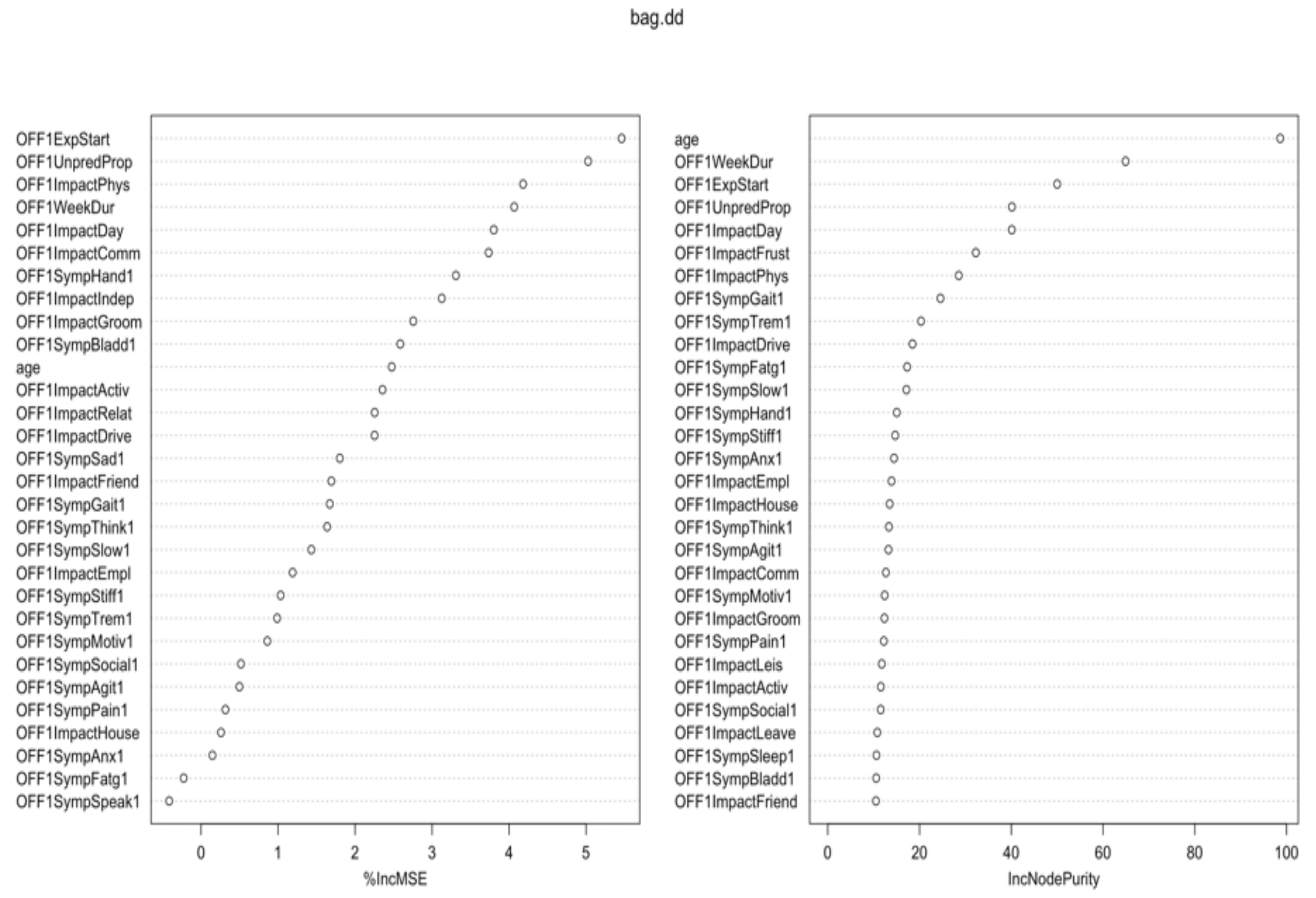
Bagging

Tree based models basically segment the predictor space into a number of simpler rectangular or box shaped regions for easier interpretation. We make the predictions for observations in each region which is the mean of response values in that particular region. Bagging is a general procedure to reduce the variance of a statistical learning method. Number of trees are built on bootstrapped training samples. A random sample of 'm' predictors are chosen as split candidates from a full set of 'p' predictors.[3][4] Bootstrapping is done by taking repeated samples from training set. 'B' different bootstrapped datasets are generated, model is trained on Bth bootstrapped dataset and an average of all predictions is done to get:

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{*b}(x)$$



Based on the above mentioned values, Bagging is the best fit predictive model for our data set since it has the least MSE value. There is almost 50 percent reduction in the MSE value as compared to Null model. Further, we analyze the important variables obtained from Bagging. The important variables as obtained from the model are as depicted below.



Some of the most important variables are described below:

Variables	Variable Description
OFF1ExpStart	How many years ago did you begin to experience OFF periods?
OFF1UnpredProp	What proportion of your OFF periods come at unpredictable times?
OFF1ImpactPhys	Impact on physical activity
OFF1WeekDur	What is the average duration of OFF periods over the last week?
OFF1ImpactComm	Impact on communication
OFF1SympHand1	Difficulty with hand co-ordination
OFF1ImpactIndep	Impact on independence in daily routine

Conclusion

- We used parametric, semi-parametric and non-parametric predictive models to find the best fit for our data set.
- The predictive model with least MSE value, our defining parameter, is the best fit.
- The difference in Test MSE and Train MSE for bagging obtained is acceptable and shows us that model was not overfit.
- In the end, we determined the significant variables for predicting OFF periods. Many studies suggest that age is an important parameter for Parkinson's Disease [5]. However, in this study, we found that age is not that significant. One of the reasons for this we believe is that PD commonly arises in old age. Since the disease is progressive, the symptoms go on becoming severe with age. Most of the patients that showed severe symptoms were above the age of 60 years.
- The number of years before the patient starts experiencing OFF periods is the most significant factor.
- Apart from this, the patients should note how difficult it is for them to Drive, do regular scheduled activities, be independent and communicate. They should consult their medical advisor if they frequently feel depressed/sad, have difficulty in thinking, have difficulty in hand coordination and difficulty in Bladder control.

Future scope

In this study, we developed a model to predict the number of OFF periods the patient will experience in a week, given the severity of symptoms. In future, with detailed literature survey, we can study the kind of medication prescribed for PD and then consequently, try to match the proper medication and it's required doses based on the severity. [6] provides a good read for importance of medications in Parkinson's Disease treatment.

References

[1] UPDRS Development Committee et al. Recent developments in parkinson's disease. 1987.

[2] Matthew B Stern, Anthony Lang, and Werner Poewe. Toward a redefinition of parkinson's disease. Movement disorders, 27(1):54–60, 2012.

[3] Towards data science, howpublished = https://towardsdatascience.com/@tonester524. Accessed: 2019-11-15.

[4] Analytics Vidhya, howpublished = https://www.analyticsvidhya.com/myfeed/?utm-source=blogutm-medium=top-iconAccessed: 2019-11-15.

[5] Amy Reeve, Eve Simcox, and Doug Turnbull. Ageing and parkinson's disease: why is advancing age the biggest risk factor? Ageing research reviews, 14:19–30, 2014.

[6] J Gert van Dijk, J Haan, Koos Zwinderman, Berry Kremer, BJ Van Hilten, and RA Roos. Autonomic nervous system dysfunction in parkinson's disease: relationships with age, medication, duration, and severity. Journal of Neurology, Neurosurgery & Psychiatry, 56(10):1090{1095, 1993.