Analysis of the Parkinson's Study of Voice Data and Problem Statement

The problem statement I have is to take a set of data describing an individual and determine whether or not, based the data recorded, this individual has Parkinson's disease.

There are a multitude of datasets available on Parkinson's. I have selected one data set of subjects voice recordings, where the observations are of individuals speaking. Some of these recordings are of individuals with Parkinson’s. Based on these recordings, my hypothesis is that a diagnosis / determination can be made as to whether or not a subject has Parkinson’s.

My interest in this area is due to the fact that my mother started exhibiting symptoms of cognitive decline. After much testing, and many visits to various health care providers and medical specialists, the diagnosis was inconclusive, until she visited a neurologist at the Mayo Clinic in Jacksonville, Florida. This doctor observed her and provided a diagnosis of "atypical" Parkinson's.

Atypical Parkinson’s essentially indicates that an individual has a series of observed symptoms which, when taken together, have the appearance of Parkinson’s, but the root cause of these symptoms is not determined. So my thought is to determine if a more accurate diagnosis can be made to determine if an individual has Parkinson's, specifically, by using a set of data of individuals known to have it.

There are also other data sets using brain imaging, handwriting, demographic details, lifestyle details and clinical measurements which can be used to determine if an individual has Parkinson's. Also, sensor data can be pulled from a patient who has a device such as an apple watch, and data sets exist which can be used for this type of analysis and modeling.

Later, I intend to explore some of these other data sets.

The techniques I will explore are going to be those which fall into classification, since this is essentially a classification problem. I will use GridSearch to determine which modeling technique provides the best result, and may include variance SVM models, decision trees, random forests, K Nearest Neighbors and logistic regression (Sigmoid). Additionally, I may use numerical encoding and categorical encoding, as well as explore standardization. I will use a confusion matrix, looking to reduce the number of false negatives, to the highest extent possible. I will look at Recall, the F-1 score, ROC AUC, and the False Negative Rate Difference Metric to evaluate models.

I expect to be able to determine with an 85% or more accuracy whether or not a given individual, with a given set of observations, can be classified as having Parkinson's disease or not.

This question is important because determining the correct and accurate diagnosis for a patient. Accurate diagnosis is the first step in treatment.

Conclusions from Initial Model

For my first trial run, I used PCA analysis on the data set, and chose logistic regression as the modelig technique. I ran through the model with 1-20 components, and identified after 8 components additioanal cumulative gains were small, I decided to use 10 components.

Applying a standard scaler and grid search with best params identified, allowed for a model which showed a test accuracy of 83.33% showss the model has a reasonably good overall ability to distinguish between individuals with and without Parkinson's based on the voice data and the chosen approach (scaling, PCA, and Logistic Regression with tuned hyperparameters), but is slightly lower than the average cross-validation score on training data with 10 PCA components at 86.23%, but understandable as models often perform slightly better on the data they were trained on compared to completely new data. This does not meet my target.

However, after threshold tuning, the model surpassed the target. # An analysis of different probability thresholds shows setting the classification threshold to 0.28 for the model yields the highest F1-score on the test data, achieving a score of approximately 0.9365 (93.65%).

With the default threshold of 0.5, the model had a test Accuracy of approximately 0.8333 (83.33%) and the confusion matrix showed 7 false negatives. By tuning the threshold to 0.28, the F1-score went to 0.9365. And the number of false negatives was reduced to 0.

From a medical perspective however, relying on voice data may be problematic, because of the need for specialized equipment and a standardized environment to capture the voice data. Various factors may impact voice data recordings, such as the microphone used, the room in which the data was recorded, the analog to digital conversion of the data, the storage compression of the data, etc... Unless the exact same set of variables can be controlled for, for each patient's set of voice data captured, the test results and model performance may vary and not be repeatable. For this reason, I will explore using other types of data for the next analysis.