

# Hx Data Science Capstone - Choose Your Own Project

Tim Bishop

07 January 2021

## Executive Summary

This report describes work performed for the “Choose Your Own Project” which is part of the Hx Data Science Capstone Module. This Capstone Module is the last of 9 modules which form the Hx Data Science certificate program.

I selected a dataset from the UCI Machine Learning Depository created for the goal of identifying people with Parkinson’s Disease (“PD”) based on voice measurement data. The resulting objective of this project is then to build a machine learning model to predict Parkinson’s disease patients, while demonstrating good modeling practice and good knowledge of the Hx Data Science series content.

The data provided consisted of a range of biomedical voice measurements from 31 people, 23 with Parkinson’s disease. Each person had multiple measurement with 195 voice recording in total. The average incidence of Parkinson’s in the data file is 75.4%. The data appeared quite clean with no missing values or obvious outliers. No modifications were made to the data. The data was split into training, validation and a final independent test set for model training, assessment and final testing respectively.

In total 5 model forms were tested for this binary classification problem. In some cases, variants of each model form were tested as well. Models tested included a decision tree, logistic regression, principle component analysis transformed data with logistic regression, XGBoost and random forest.

To assess what model was best, I set three main objectives which included model accuracy, tendency for false positive or false negative errors and how interpretable the model is and its linkage from predictors to predictions. False negatives were considered to be most costly, in a possible future model use setting, as these predictions could potentially prolong actual Parkinson’s patients from getting early medical attention.

In the end, a simple classification tree was the best performing model. When used on the independent test set, it had a prediction accuracy of 90%, which is much higher than the 70% Parkinson’s rate in the test set. Further, this model had the lowest false negative prediction rate of all models when compared on the validation data set.

It should always be noted that future performance of this model on other datasets depends on many factors such as the consistency of data, for example in voice recording measurement approach or quality.

## Problem Statement & Objective

The dataset selected was produced for the goal of identifying people with Parkinson’s Disease (“PD”) based on voice measurement data. Following this, the objective of this analysis is to build a machine learning model that best predicts the presence of PD given audio voice measurement information. I will define “best” to mean a model that balances various perspectives such as:

1. Predictive accuracy of the model.
2. The models behaviour with false positive or false negative predictions.
3. How interpretability is the model and the relationship of predictions to predictor variables in the model.
4. More broadly, potential use of this model in future.

In order to assess and balance in the objectives above, I have made the following assumptions:

1. Model accuracy will be tracked as the percentage (in decimal) of total correct predictions (ie True Positive and True Negative),
2. False positive and false negative prediction performance will be measured by specificity and sensitivity scores. The specificity score measures the rate of negative predictions given a person is well. The sensitivity score measures the rate of positive predictions given the person has Parkinson's. I will also use the Area Under Curve ("AUC") measures that is a combination of specificity and sensitivity. All these numbers are between 0 and 1 with higher being better.
3. I'm assuming the PD onset predictions from this model will be used as an indicative, early warning measure as opposed to a real diagnosis tool. I'm assuming this would be used as a screening tool or as an inexpensive way to flag early onset. For this reason:
  - a) I will give more weight to sensitivity than specificity. Said another way, false positive predictions are generally more favourable than false negatives. The rationale is that a false negative could allow someone with PD to progress longer without the benefit of early treatment, which could be damaging. However, a false positive may cause some worry initially but will get a person in front of a doctor for a more thorough exam that will ultimately rule out the disease.
  - b) I will give predictive accuracy more weight than model interpretability. More complex, less interpretable models will be given priority if their predictive results are much better.
4. I do not have access to a subject matter expert ("SME"). Further work would best include consultation or collaboration with a SME in this subject. This would be very helpful for feature selection or considering interaction effects. However, for this project I assume I have no access to a SME. I will not use Google to do become a "15 minute expert" and act as if I know something about the subject which I do not.

Of course, the broader objective of this work is to demonstrate a strong understanding of the machine learning modeling process, strengths, limitations, etc. For context, there are several reasons why I chose this dataset:

1. My personal interest in machine learning is in health and insurance contexts.
2. This is a classification problem, which provides a different challenge than the regression modeling work done on the MovieLens data.
3. The dataset is very manageable in size. I wanted to work with a dataset where I could run any model on my computer without crashing or waiting a very long time. While using a smaller dataset would impact the results of the model in use on other data outside of this project, it does not impact my ability to demonstrate good machine learning modeling practice for the purpose of assessment/grading. Also, not all data science work will be done on big data sets and smaller dataset can introduce their own challenges.
4. This was a context where I knew little or nothing about Parkinson's disease, its early signals or the voice measurements provided in the data. I was intrigued by an example where I had to really rely on data discovery and modeling as opposed to having some intuition about what could influence predictions (like in MovieLens).
5. The data is from an excellent source, high quality and there were no prior papers on this dataset to influence my work.

## Analysis and Methods

### Data Provided

Data for this project was obtained from the UCI Machine Learning Depository website.

Quoting from the data description file included with the dataset:

This dataset is composed of a range of biomedical voice measurements from 31 people, 23 with Parkinson's disease (PD). Each column in the table is a particular voice measure, and each row corresponds one of 195 voice recordings from these individuals ("name" column). The main aim of the data is to discriminate healthy people from those with PD, according to "status" column

which is set to 0 for healthy and 1 for PD.

Refer to Appendix A for the full data description provided with the dataset.

The first field in this dataset is the composite field “name” which is comprised of a patient identifier appended with a test number identifier. This was excluded from analysis for three reasons.

1. The patient identifier has a one to one relationship with the patient Parkinson’s diagnosis. Including an independent variable that is equivalent to the dependent variable is not appropriate.
2. This is a character field which, as non-numeric, will not work in many models.
3. The name field varies also with a voice recording test number. It could be considered to split this test identifier out as a new feature. While it is possible there are measurement error differences that vary with each voice recording test, these would not be differences in the indicative voice measures per se but measurement error. I am assuming that measurement error is not significant and that measurement is consistent accross tests. Further I have no knowledge of if measurements were performed with the same or multiple measuring devices, etc. In future work, impact of measurement number or measurement patient combinations could be tested if considered appropriate.

### **Initial Data Validation & Cleaning**

The “summary” function was used to perform initial data quality assessment. Appendix B contains the summary output on the full dataset. My observations and comments are as follows:

1. There are no missing values in any data fields.
2. There is a modest imbalance issue in the dependent variable field “status” (ie where PD is coded with a value of 1). There are more measurements from PD positive people than PD negative, with 75.4% PD positive.

To look for outlier and further insights, further visualizations and analysis will follow

### **Data Validation Exploration**

As will be explained further in the Modeling Approach section of this report, all data exploration was performed on the training dataset only. In this way, the test dataset remains completely independent of any feature selection or modeling judgement.

The following graphs show the distribution of each variable independently.

Observations are as follows:

1. Most measurements start from a zero origin.
2. Most distributions are right skewed with most of the observations closer to zero than not.
3. Given the similar patterns noted above, there is likely a strong positive correlation among many independent variables. This could cause problems in fitting certain models.
4. Other variables appear to have single mode distributions which generally appear non-normal in shape.

It is difficult to assess for outliers. A subject matter expert would have appropriate opinion on the reasonable range for these sorts of voice measurements and would be a valuable reference for making this assessment. From the graphs above, no observations appear to be very extreme. Without much of a basis of assessment, I have left the data as is and not considered any potential adjustments for outliers.

To address the imbalance issue with the number of Parkinson’s vs Healthy data subjects, I used up sampling to create a second training dataset for testing models. This will be discussed more below.

Visualizations are limited when the dependent variable is binary and the predictive variables are real or integer. The following graph shows the relationship between the dependent variable “status” and each individual predictive variable.

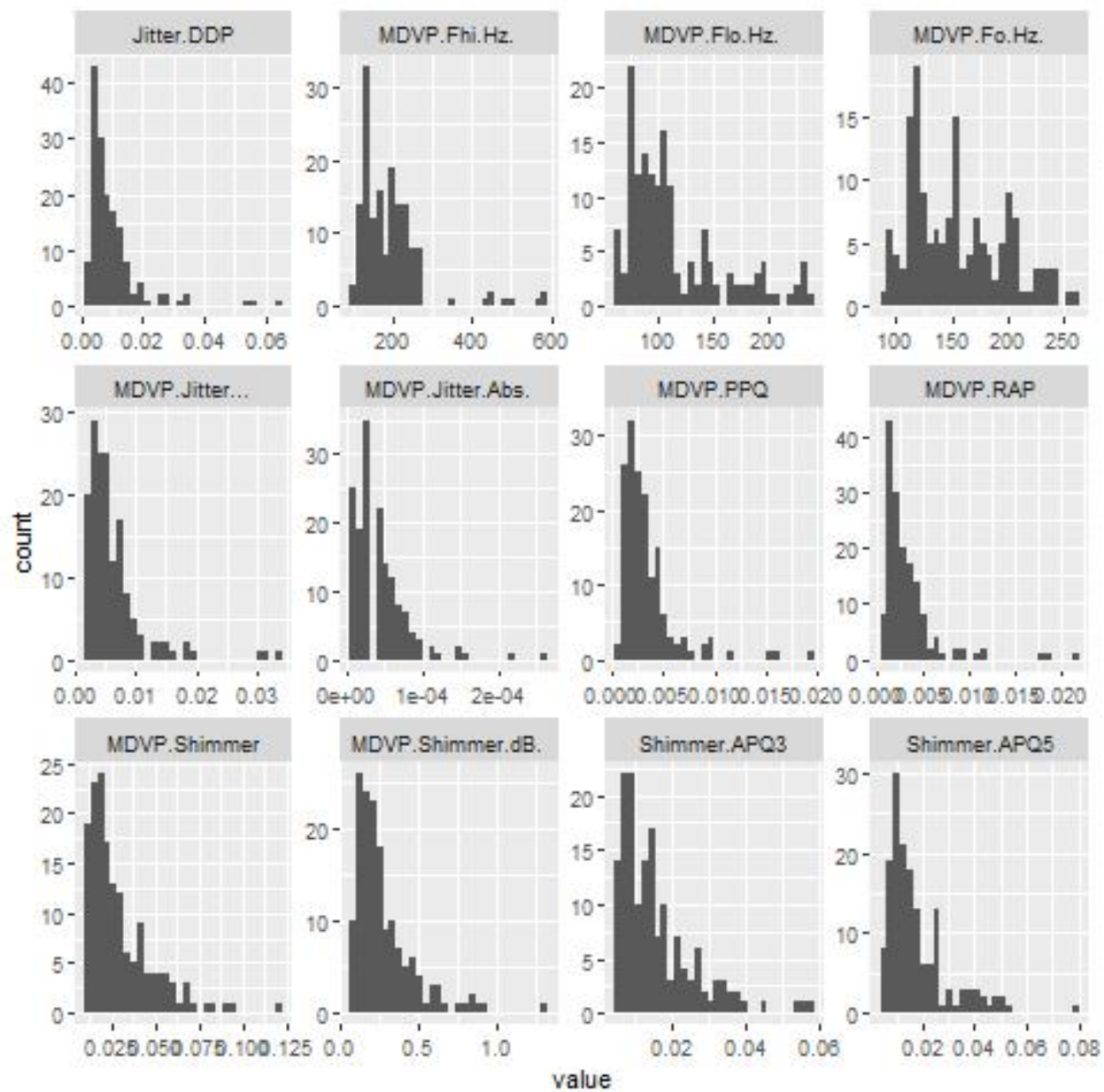


Figure 1: Data Variable Distribution Plots 1

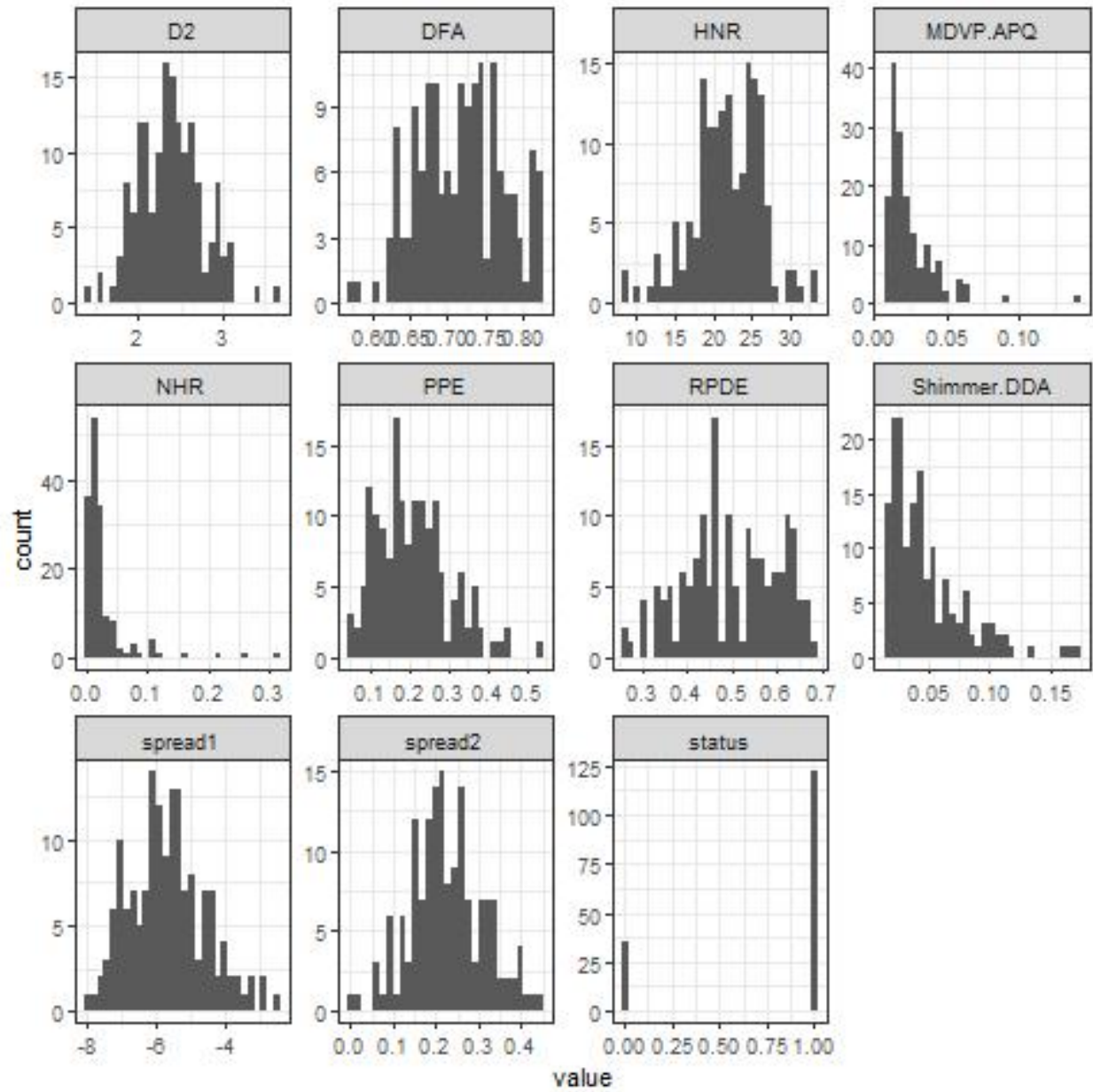


Figure 2: Data Variable Distribution Plots 2

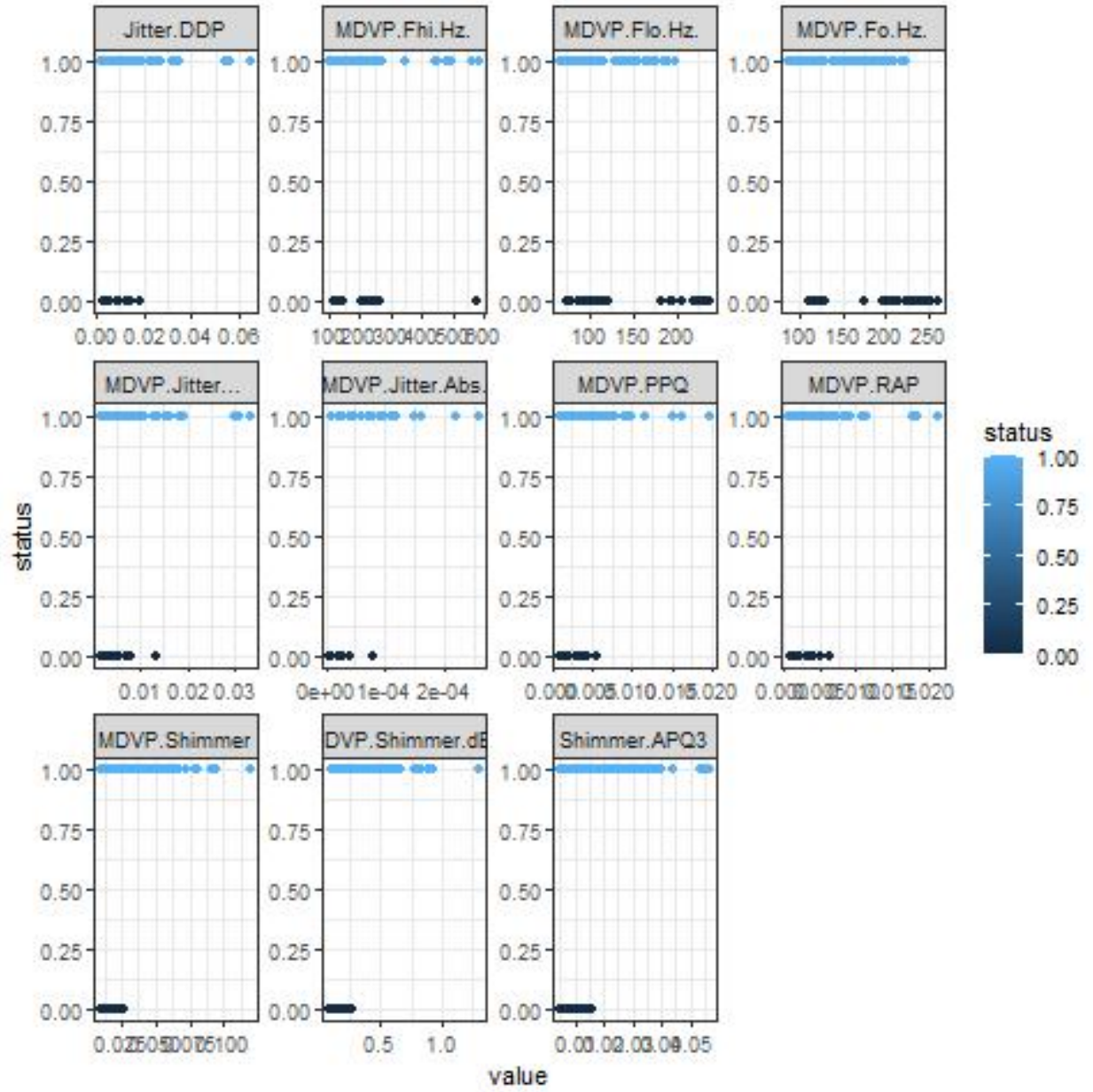
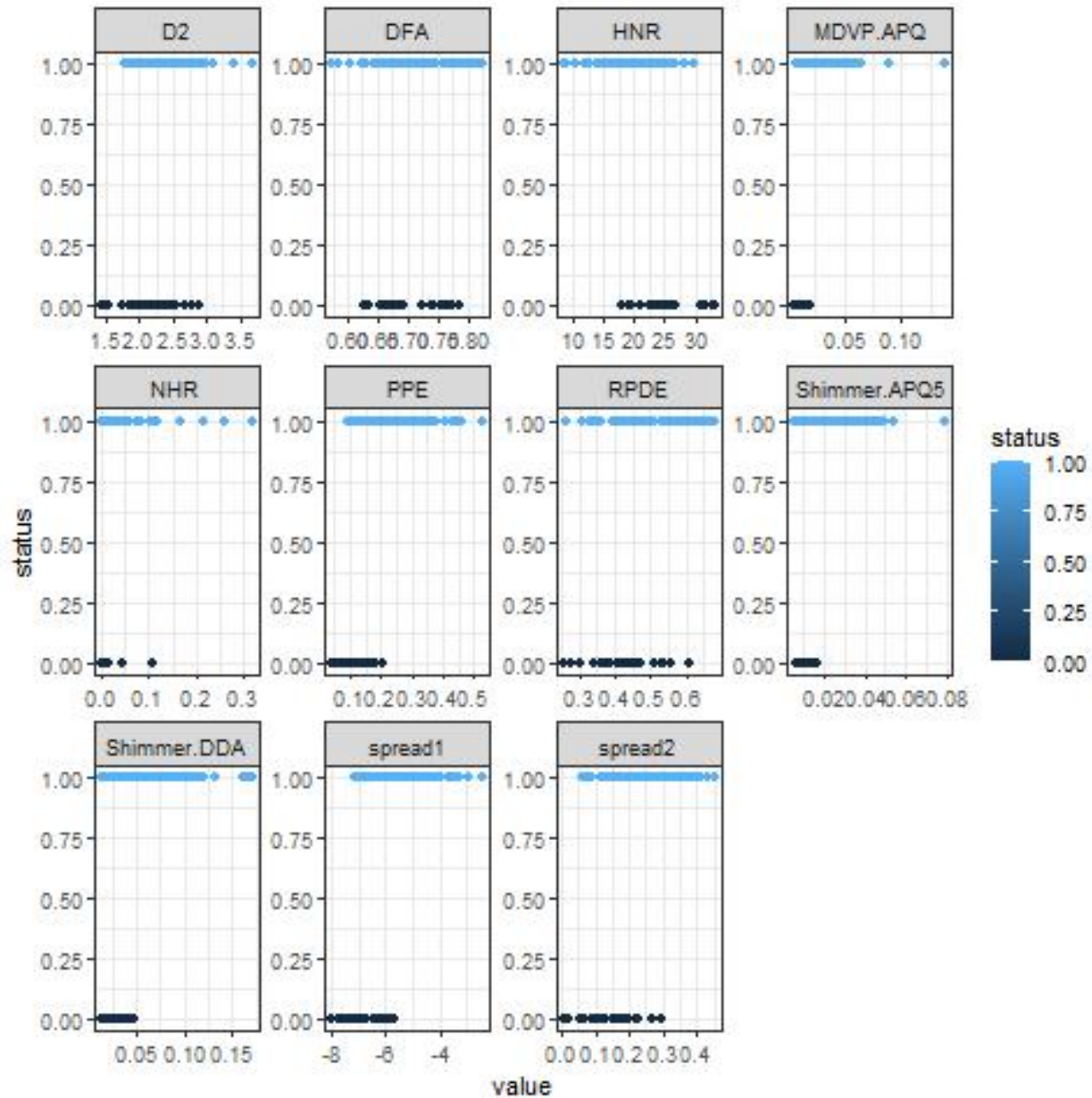


Figure 3: status Indicator VS Independent Variable Plots 1



There are a number of variables that, for their larger values, show a strong difference between Parkinson's status values of 0 and 1. However, there are no variables that show a strong difference over lower values (near zero). This is expected to limit the power of a predictive model. For most independent variables, when the variable value is close to zero, it is a toss up as to whether status is 0 or 1. Perhaps there is a combination of variables that will compensate for this.

As previously noted, given the patterns seen in the data variable distributions, there are likely correlations among predictors. The following is a correlation matrix among the data variables.

The following observations are noted:

1. The HNR variable has a high negative correlation with most other variable. This can be seen in the stand out white lines.
2. There are grouping of variables that have a high positive correlations. This is seen in the darker red boxes and line.
3. The dependent variable status has a low to moderate correlation with most all independent variables.







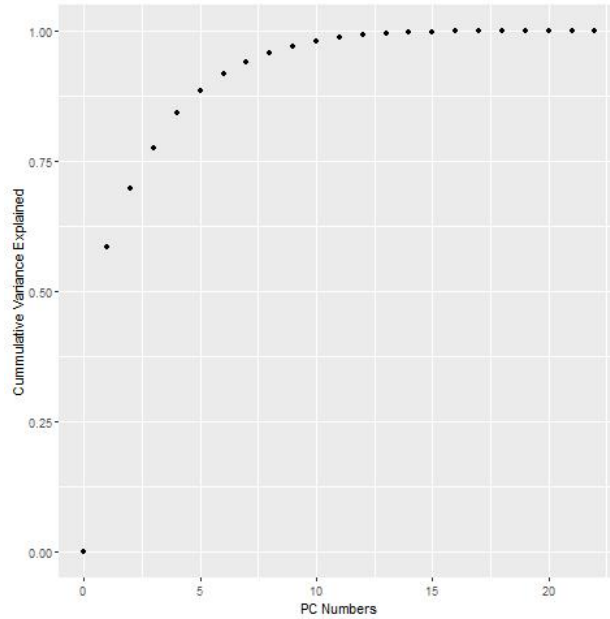


Figure 5: PCA - Cumulative Variance Explained Plot

2. `data_validation`: Comprised of approximately 10% of the total data, this dataset will be used for initial testing of models fitted on the training set. The prediction results on this dataset will be used to compare the performance of models for selecting the final model recommendation.
3. `data_train`: Comprised of approximately 80% of the total data, this set will be used for training models and tuning parameters. This was the dataset used for all data exploration work.

Further a second up sampled training dataset was derived from the original training set. To create a dataset for model fit testing which had an equal number of positive and negative Parkinson's patients, the dataset `data_up_train` was created from `data_test` by repeated random resampling of the negative status records until there was an equal proportion of both.

The percentage of total data in each dataset is open to the judgement of the modeler. I will note that a common split for train/validation/test is 70/20/10. I have used 80/10/10 given the smaller size of the total dataset. The cost of using a higher percentage of data in the training set is a higher tendency for model over fitting. I will rely on the results from the validation and test sets to look for signals of over fitting, such as lower performance on the final test set than on the validation set.

To check if the ratio of positives is relatively consistent across the dataset, the follow table was produced. The ratio of positives is relatively consistent given the size of the datasets. The ratio is 50% as expected for the up sampled train set.

Group	Status_Mean
Full Data Set	0.7538462
Train Set	0.7770701
Model Val Set	0.6111111
Test Set	0.7000000
Up Sampled Train Set	0.5000000

**First Model Form - Classification Tree** To start the modeling process, a simple classification tree was fit to the data. A picture of the resulting decision tree is below.

I did not see a need to consider pruning the tree given it's general simplicity as is. This model performed well. The performance measures were:

9				
Method	Accuracy	Sensitivity	Specificity	AUC
Simple Classification Tree	0.83	1.00	0.57	0.79

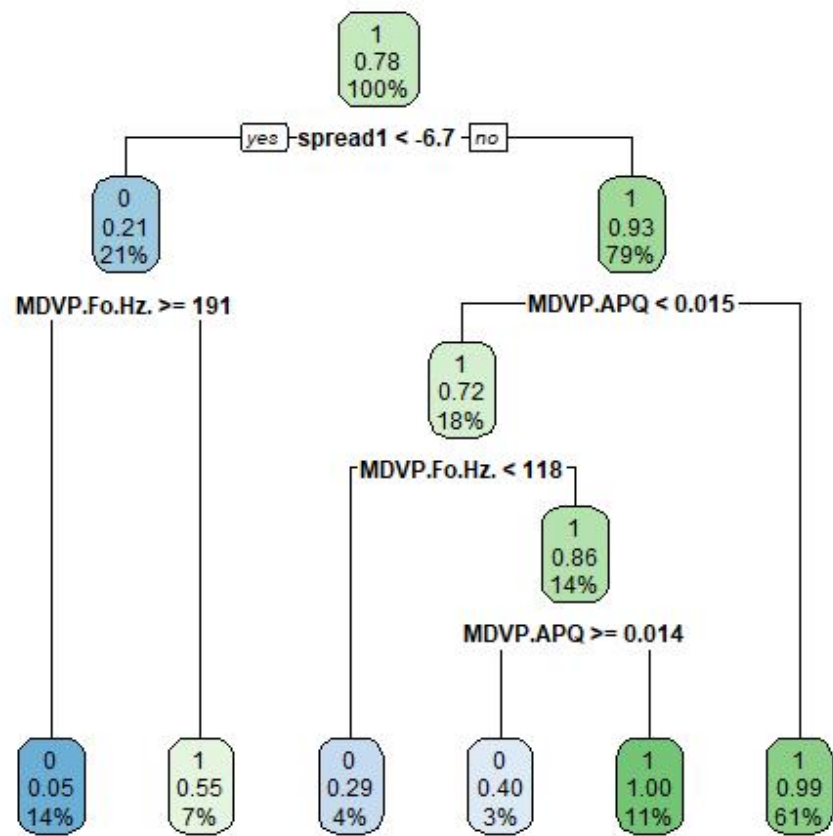


Figure 6: Classification Tree Plot

Forward stepwise regression was then used to fit a logistic regression model. Forward stepwise regression adds available independent variables one at a time, using looking to find the most impactful variable, until new variables no longer have a coefficient that is statistically significant (or meets some p-value threshold as given).

The final model form was as follows:

status ~ spread1 + MDVP.APQ + Shimmer.DDA + MDVP.Flo.Hz. + D2 + MDVP.Fhi.Hz. + DFA + HNR + PPE + Shimmer.APQ3 + RPDE + spread2

Performance was as follows:

Method	Accuracy	Sensitivity	Specificity	AUC
Logistic Regression	0.83	0.79	1.00	0.79

The accuracy of this model was also good. However, as the sensitivity rate was below 1.0, the model did produce some false negative predictions.

As a test, the up sampled dataset was used with forward stepwise regression to fit a logistic regression model. This model appeared promising when reviewing the model form and coefficients, but the performance did not change from the first logistic regression model using the original training set.

**Third Model Form - PCA with Logistic Regression** Given the multicollinearity issue noted with the logistic regression model fitting, I used the transformed data matrix from the PCA as independent variables in a logistic regression model. That is, I used the principle components as the independent variables. The first 10 PC's were used as these explained 98% of the total variance in the data. Since the PCs are orthogonal, the model converged without issue. Stepwise backward regression was performed on this model, which was then reduced to using 6 PCs from the original 10. The stepwise backward regression approach remove weaker variables from a more complex model one at a time until removing further variables is too punishing on model performance according to some measure.

The final formula of the model was:

status ~ PC1 + PC2 + PC3 + PC5 + PC9 + PC10

Before testing this model on the validation set, the data from the validation dataset was standardized and rotated based on the weights from the PCA performed on the training data. When used on the transformed validation dataset, the predictions provided from this model were as follows:

Method	Accuracy	Sensitivity	Specificity	AUC
PCA with Logistic Regression	0.50	0.62	0.40	0.77

This model did not perform well and was rejected.

**Fourth Model Form - XGBoost** An XGBoost model was tested as this is a very popular model which reports to be powerful. The results were as follows:

Method	Accuracy	Sensitivity	Specificity	AUC
XGBoost Model	0.83	1.00	0.57	0.79

The performance same as the simple classification tree. This is not entirely surprising as this model uses an iterative tree fitting algorithm. Quoting Machinelearningmastery.com/:

Boosting is an ensemble technique where new models are added to correct the errors made by existing models. Models are added sequentially until no further improvements can be made.

XGBoost also allows a “DART” based modeling approach, as opposed to tree based. The DART parameter was tested and made no change to results.

Given that there was no prediction performance improvement, this model was rejected given the significant additional complexity and reduction in interpretability over the classification tree.

**Fifth Model Form - Random Forest** The final model model considered was the random forest. Using the caret package train function, K-fold cross validation on the training set was used to tune the mtry parameter, which controls the number of independent variables randomly sampled as for potential use in splitting the tree at each decision point.

Method	Accuracy	Sensitivity	Specificity	AUC
Random Forest	0.72	1.00	0.29	0.50

Performance was not as good as classification tree.

## Final Model Selection and Testing on Validation Set

In the Objectives section I set out the following model characteristics for assessing what model is “best”:

1. Predictive accuracy of the model.
2. Interpretability of both the model and the relationship of predictions to predictor variables.
3. Relative risk or cost of false positive or false negative predictions.
4. Potential use of this model in future.

The results of the models considered are in the following table:

Method	Accuracy	Sensitivity	Specificity	AUC
Mean Of Validation set status	0.61	1.00	0.00	0.00
Simple Classification Tree	0.83	1.00	0.57	0.79
Classification Tree Upsampled Data	0.83	1.00	0.57	0.79
Logistic Regression	0.83	0.79	1.00	0.79
Logistic Regression & Up Sampling	0.83	0.79	1.00	0.79
PCA with Logistic Regression	0.50	0.62	0.40	0.79
XGBoost Model	0.83	1.00	0.57	0.79
Random Forest	0.72	1.00	0.29	0.50

The final model chosen was the classification tree as it had the best accuracy, is simple to interpret and has the most favourable sensitivity and specificity according to the objectives I have outlined previously.

This final model type was then trained on the combined data from the training and validation datasets, This is done to utilize the most data outside of the final test set for training the final model in an effort to increase the tuning quality of the final model.

The shape of the decision tree changed slightly after re-training on the train and validation data. In fact, the tree became more simple. Pruning this final tree was not considered given its natural simplicity.

A plot of the final classification tree is as follows.

The final model fit was then used to predict Parkinson’s disease on the final independent test set. The results are as follows.

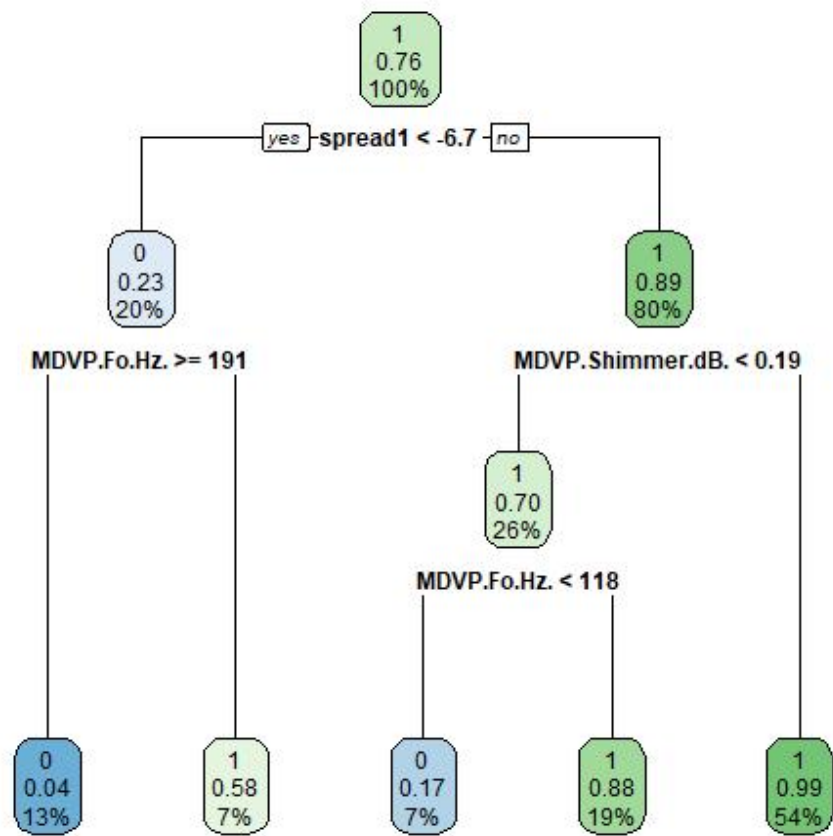


Figure 7: Final Model - Classification Tree Plot

Method	Accuracy	Sensitivity	Specificity	AUC
Final Model - Classification Tree	0.9	1	0.6666667	0.8333333

The results on the final test set have actually improved from the those on the validation set. This is evidence that the model generalizes well to new data and the model was not over fit to the training data. The actual positive rate in the test set is 0.7, so the model has improved well over guessing.

## Conclusions

The objective best measures model chosen results

As always, future performance of this model depends on user behaving similar to past users in their rating practices as seen in the data used in this project. All data provided was used, even though there is a clearly a change in ratings methodology after 2002, when half star ratings first appeared in the data.

Future work to consider would include:

1. Given computing capacity, mdels more complicated and computationally demanding could be tested.
2. Interaction effects between features could be tested. The current work did not explicitly look for interaction effects between features.
3. Further work could be performed to see if the 2003 change in ratings collected could be better utilized or reflected. One possibility would be to reperform the analysis using only data from 2003 onward, was this would presumably better match the user behaviour under a whole and half star rating scale now used.

# Appendices

## Appendix A - Data Description file provided with the dataset

Below is the data description file provided for the dataset used. It can be accessed at the following link.

Title: Parkinsons Disease Data Set

Abstract: Oxford Parkinson's Disease Detection Dataset

---

Data Set Characteristics: Multivariate Number of Instances: 197 Area: Life Attribute Characteristics: Real  
Number of Attributes: 23 Date Donated: 2008-06-26 Associated Tasks: Classification Missing Values? N/A

---

Source:

The dataset was created by Max Little of the University of Oxford, in collaboration with the National Centre for Voice and Speech, Denver, Colorado, who recorded the speech signals. The original study published the feature extraction methods for general voice disorders.

---

Data Set Information:

This dataset is composed of a range of biomedical voice measurements from 31 people, 23 with Parkinson's disease (PD). Each column in the table is a particular voice measure, and each row corresponds one of 195 voice recording from these individuals ("name" column). The main aim of the data is to discriminate healthy people from those with PD, according to "status" column which is set to 0 for healthy and 1 for PD.

The data is in ASCII CSV format. The rows of the CSV file contain an instance corresponding to one voice recording. There are around six recordings per patient, the name of the patient is identified in the first column. For further information or to pass on comments, please contact Max Little (littlem '@' robots.ox.ac.uk).

Further details are contained in the following reference – if you use this dataset, please cite: Max A. Little, Patrick E. McSharry, Eric J. Hunter, Lorraine O. Ramig (2008), 'Suitability of dysphonia measurements for telemonitoring of Parkinson's disease', IEEE Transactions on Biomedical Engineering (to appear).

---

Attribute Information:

Matrix column entries (attributes):

- name - ASCII subject name and recording number
  - MDVP:F0(Hz) - Average vocal fundamental frequency
  - MDVP:F1(Hz) - Maximum vocal fundamental frequency
  - MDVP:F2(Hz) - Minimum vocal fundamental frequency
  - MDVP:F3(Hz) - Jitter(%) , MDVP:F3(Hz) - Jitter(Abs) , MDVP:F3(Hz) - RAP, MDVP:F3(Hz) - PPQ, Jitter:DDP - Several measures of variation in fundamental frequency
  - MDVP:F3(Hz) - Shimmer, MDVP:F3(Hz) - Shimmer(dB), Shimmer:APQ3, Shimmer:APQ5, MDVP:APQ, Shimmer:DDA - Several measures of variation in amplitude
  - NHR, HNR - Two measures of ratio of noise to tonal components in the voice
  - status - Health status of the subject (one) - Parkinson's, (zero) - healthy
  - RPDE, D2 - Two nonlinear dynamical complexity measures
  - DFA - Signal fractal scaling exponent
  - spread1, spread2, PPE - Three nonlinear measures of fundamental frequency variation
-



Citation Request:

If you use this dataset, please cite the following paper: 'Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection', Little MA, McSharry PE, Roberts SJ, Costello DAE, Moroz IM. BioMedical Engineering OnLine 2007, 6:23 (26 June 2007)

## Appendix B - Output from summary function on Dataset

```
##      name      MDVP.Fo.Hz.      MDVP.Fhi.Hz.      MDVP.Flo.Hz.
## Length:195      Min.      : 88.33      Min.      :102.1      Min.      : 65.48
## Class :character 1st Qu.:117.57      1st Qu.:134.9      1st Qu.: 84.29
## Mode  :character Median :148.79      Median :175.8      Median :104.31
##                      Mean  :154.23      Mean  :197.1      Mean  :116.32
##                      3rd Qu.:182.77      3rd Qu.:224.2      3rd Qu.:140.02
##                      Max.   :260.11      Max.   :592.0      Max.   :239.17
## MDVP.Jitter... MDVP.Jitter.Abs.      MDVP.RAP      MDVP.PPQ
## Min.      :0.001680      Min.      :7.000e-06      Min.      :0.000680      Min.      :0.000920
## 1st Qu.:0.003460      1st Qu.:2.000e-05      1st Qu.:0.001660      1st Qu.:0.001860
## Median :0.004940      Median :3.000e-05      Median :0.002500      Median :0.002690
## Mean  :0.006220      Mean  :4.396e-05      Mean  :0.003306      Mean  :0.003446
## 3rd Qu.:0.007365      3rd Qu.:6.000e-05      3rd Qu.:0.003835      3rd Qu.:0.003955
## Max.   :0.033160      Max.   :2.600e-04      Max.   :0.021440      Max.   :0.019580
## Jitter.DDP      MDVP.Shimmer      MDVP.Shimmer.dB.      Shimmer.APQ3
## Min.      :0.002040      Min.      :0.00954      Min.      :0.0850      Min.      :0.004550
## 1st Qu.:0.004985      1st Qu.:0.01650      1st Qu.:0.1485      1st Qu.:0.008245
## Median :0.007490      Median :0.02297      Median :0.2210      Median :0.012790
## Mean  :0.009920      Mean  :0.02971      Mean  :0.2823      Mean  :0.015664
## 3rd Qu.:0.011505      3rd Qu.:0.03789      3rd Qu.:0.3500      3rd Qu.:0.020265
## Max.   :0.064330      Max.   :0.11908      Max.   :1.3020      Max.   :0.056470
## Shimmer.APQ5      MDVP.APQ      Shimmer.DDA      NHR
## Min.      :0.00570      Min.      :0.00719      Min.      :0.01364      Min.      :0.000650
## 1st Qu.:0.00958      1st Qu.:0.01308      1st Qu.:0.02474      1st Qu.:0.005925
## Median :0.01347      Median :0.01826      Median :0.03836      Median :0.011660
## Mean  :0.01788      Mean  :0.02408      Mean  :0.04699      Mean  :0.024847
## 3rd Qu.:0.02238      3rd Qu.:0.02940      3rd Qu.:0.06080      3rd Qu.:0.025640
## Max.   :0.07940      Max.   :0.13778      Max.   :0.16942      Max.   :0.314820
## HNR      status      RPDE      DFA
## Min.      : 8.441      Min.      :0.0000      Min.      :0.2566      Min.      :0.5743
## 1st Qu.:19.198      1st Qu.:1.0000      1st Qu.:0.4213      1st Qu.:0.6748
## Median :22.085      Median :1.0000      Median :0.4960      Median :0.7223
## Mean  :21.886      Mean  :0.7538      Mean  :0.4985      Mean  :0.7181
## 3rd Qu.:25.076      3rd Qu.:1.0000      3rd Qu.:0.5876      3rd Qu.:0.7619
## Max.   :33.047      Max.   :1.0000      Max.   :0.6852      Max.   :0.8253
## spread1      spread2      D2      PPE
## Min.      : -7.965      Min.      :0.006274      Min.      :1.423      Min.      :0.04454
## 1st Qu.: -6.450      1st Qu.:0.174350      1st Qu.:2.099      1st Qu.:0.13745
## Median : -5.721      Median :0.218885      Median :2.362      Median :0.19405
## Mean  : -5.684      Mean  :0.226510      Mean  :2.382      Mean  :0.20655
## 3rd Qu.: -5.046      3rd Qu.:0.279234      3rd Qu.:2.636      3rd Qu.:0.25298
## Max.   : -2.434      Max.   :0.450493      Max.   :3.671      Max.   :0.52737
```

## Appendix C - Sample Principle Component Rotations

The following show the PCA rotational weights for the first five PCs.

	PC1	PC2	PC3	PC4	PC5
MDVP.Fo.Hz.	0.0762501	-0.5319502	-0.1098997	0.1663821	0.0785128
MDVP.Fhi.Hz.	-0.0055656	-0.3899382	-0.2892906	-0.1391814	0.2009665
MDVP.Flo.Hz.	0.0840633	-0.3593485	0.2500613	0.1776042	0.3243199
MDVP.Jitter. . .	-0.2554905	-0.0916632	0.0913139	-0.2739731	0.0683285
MDVP.Jitter.Abs.	-0.2455755	0.0551001	0.1217917	-0.3275099	0.0126448
MDVP.RAP	-0.2500076	-0.1249302	0.1149515	-0.2842329	0.0170492
MDVP.PPQ	-0.2580603	-0.0706535	0.1319110	-0.1813786	0.1559706
Jitter.DDP	-0.2500013	-0.1249433	0.1149863	-0.2842729	0.0170531
MDVP.Shimmer	-0.2598035	-0.0552833	0.0764070	0.2506901	-0.0987407
MDVP.Shimmer.dB.	-0.2605909	-0.0776361	0.0817294	0.2212474	-0.0386027
Shimmer.APQ3	-0.2537809	-0.0593947	0.1051443	0.2389753	-0.1627567
Shimmer.APQ5	-0.2506664	-0.0463139	0.0909405	0.3152863	-0.0226518
MDVP.APQ	-0.2507142	-0.0508115	0.0242438	0.2629758	0.0080562
Shimmer.DDA	-0.2537823	-0.0593968	0.1051474	0.2389677	-0.1627569
NHR	-0.2328798	-0.1944594	0.0143900	-0.2786402	-0.0971977
HNR	0.2404895	0.0204078	0.1339871	-0.1072189	0.1830361
RPDE	-0.1558544	0.2392783	-0.3099646	0.0140713	-0.3985240
DFA	-0.0346417	0.3359834	0.4527806	0.1566990	0.4544930
spread1	-0.2236087	0.2344404	-0.1939371	-0.0732396	0.1945062
spread2	-0.1418965	0.2010549	-0.3810231	0.1284856	0.4155225
D2	-0.1396283	-0.1303784	-0.4595585	0.1146075	0.2660367
PPE	-0.2298016	0.2158181	-0.1340731	-0.0333455	0.2532358