**Data:** This dataset is related to some measurements with the people having Parkisons Disesase. The objective of the solution is to come up with models from the data to discriminate healthy people from those with PD, based on the "status" column which is set to 0 for healthy and 1 for PD.

**Size :** The data set size is about 195 rows and 23 columns

**Exploratory Data Analysis:**

**Structure of the data:**

'data.frame': 195 obs. of 24 variables:

$ name : Factor w/ 195 levels "phon\_R01\_S01\_1",..: 1 2 3 4 5 6 7 8 9 10 ...

$ MDVP.Fo.Hz. : num 120 122 117 117 116 ...

$ MDVP.Fhi.Hz. : num 157 149 131 138 142 ...

$ MDVP.Flo.Hz. : num 75 114 112 111 111 ...

$ MDVP.Jitter... : num 0.00784 0.00968 0.0105 0.00997 0.01284 ...

$ MDVP.Jitter.Abs.: num 0.00007 0.00008 0.00009 0.00009 0.00011 0.00008 0.00003 0.00003 0.00006 0.00006 ...

$ MDVP.RAP : num 0.0037 0.00465 0.00544 0.00502 0.00655 0.00463 0.00155 0.00144 0.00293 0.00268 ...

$ MDVP.PPQ : num 0.00554 0.00696 0.00781 0.00698 0.00908 0.0075 0.00202 0.00182 0.00332 0.00332 ...

$ Jitter.DDP : num 0.0111 0.0139 0.0163 0.015 0.0197 ...

$ MDVP.Shimmer : num 0.0437 0.0613 0.0523 0.0549 0.0643 ...

$ MDVP.Shimmer.dB.: num 0.426 0.626 0.482 0.517 0.584 0.456 0.14 0.134 0.191 0.255 ...

$ Shimmer.APQ3 : num 0.0218 0.0313 0.0276 0.0292 0.0349 ...

$ Shimmer.APQ5 : num 0.0313 0.0452 0.0386 0.0401 0.0483 ...

$ MDVP.APQ : num 0.0297 0.0437 0.0359 0.0377 0.0447 ...

$ Shimmer.DDA : num 0.0654 0.094 0.0827 0.0877 0.1047 ...

$ NHR : num 0.0221 0.0193 0.0131 0.0135 0.0177 ...

$ HNR : num 21 19.1 20.7 20.6 19.6 ...

$ status : int 1 1 1 1 1 1 1 1 1 1 ...

$ RPDE : num 0.415 0.458 0.43 0.435 0.417 ...

$ DFA : num 0.815 0.82 0.825 0.819 0.823 ...

$ spread1 : num -4.81 -4.08 -4.44 -4.12 -3.75 ...

$ spread2 : num 0.266 0.336 0.311 0.334 0.235 ...

$ D2 : num 2.3 2.49 2.34 2.41 2.33 ...

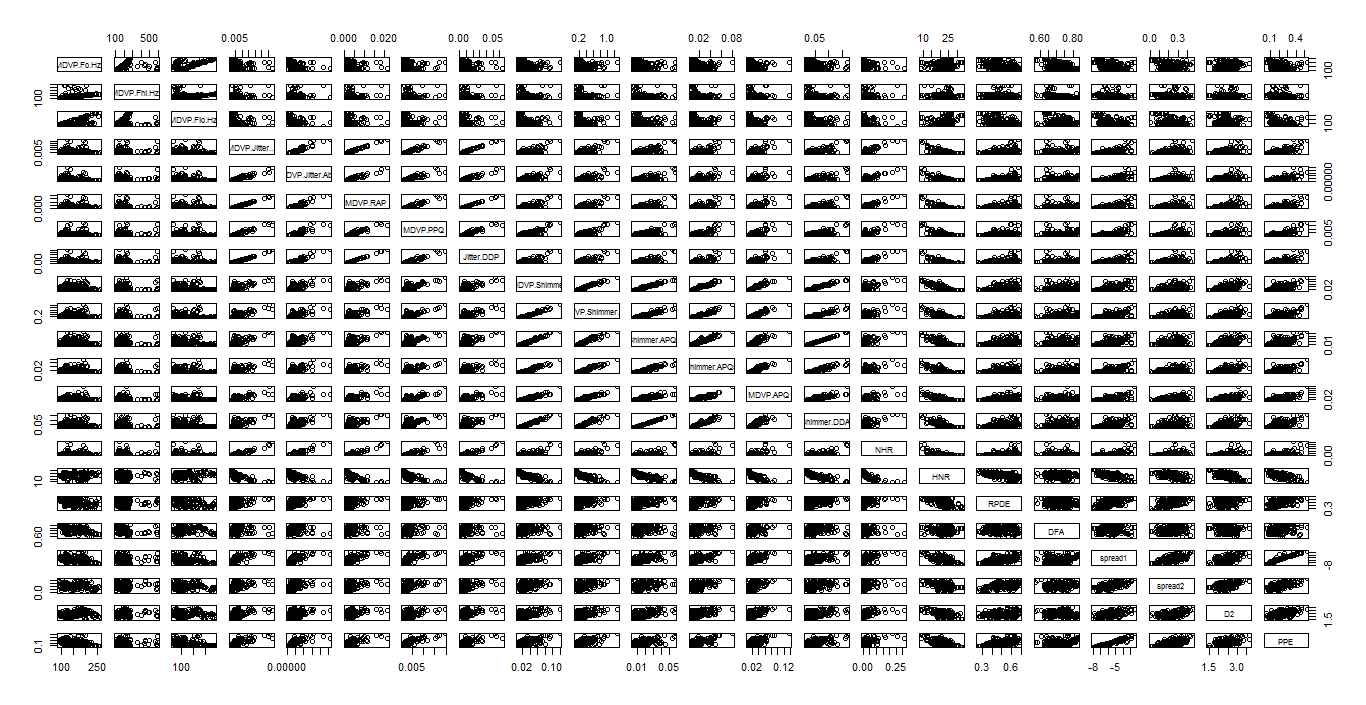
$ PPE : num 0.285 0.369 0.333 0.369 0.41 ...

Except for the name variable, all the variables are numeric. The status variable has two values ‘0’ and ‘1’.

**Missing Values:** There are no missing values in the data

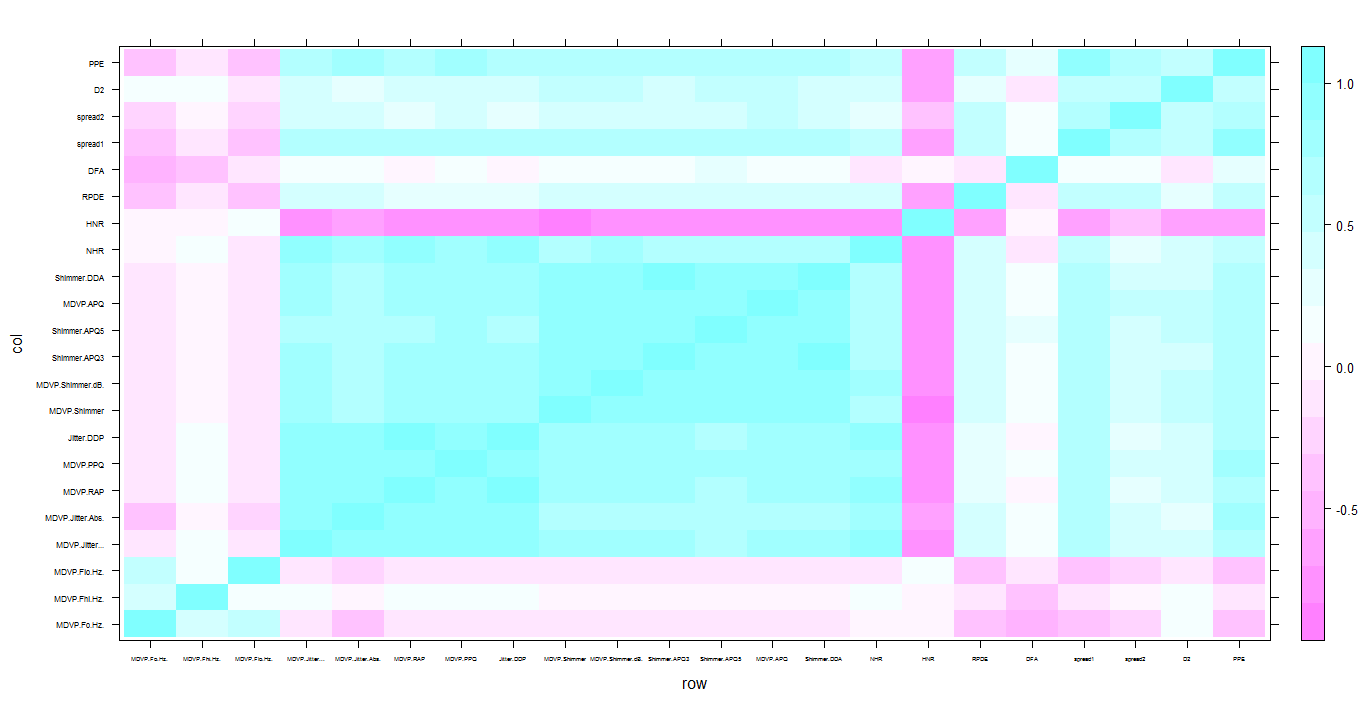
**Correlation among Predictors:** The correlation values for all the variables except for the ‘name’ and ‘status’ are obtained to see if there is any multicollinearity.

Below are the scatter plots and level plots for the correlation values.



There appears to be a strong positive correlation between the variables which are highlighted in the plot.

The below level plot describes the strength of correlation among predictors. Clearly there is lot of area which is blue and purple. This indicates presence of highly correlated variables.



The highly correlated variables where correlation > 0.7 are below:

highcorvar

[1] "MDVP.Shimmer" "Jitter.DDP" "MDVP.RAP" "MDVP.Flo.Hz." "Shimmer.APQ5" "MDVP.APQ"

[7] "MDVP.Shimmer.dB." "Shimmer.APQ3" "MDVP.Jitter..." "MDVP.Jitter.Abs." "MDVP.PPQ" "spread2"

[13] "NHR"

**Outliers:** Boxplots have been plotted to identify variables with outlier values. But most of these variables are multicollinear as well. These will be handled appropriately while building the model

**Modelling:**

**Train vs Test Split:** Since the data set size is small, I will use a 80:20 ratio between the train and test datasets

**Modelling Strategies:**

**Logistic Regression:** First I plan to build a logistic model with all the features. Then variable selection will be done using stepwise regression in backward direction. In this way, all the variables which are collinear and not so useful for prediction will be weeded out.

Summary of the model when all the variables are used for building the model with the training data set. Good number of variables didn’t come to be significant

> summary(fullmod)

Call:

glm(formula = status ~ ., family = binomial, data = data\_train[-1])

Deviance Residuals:

Min 1Q Median 3Q Max

-2.39658 0.00000 0.02593 0.32764 1.90252

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -18.547953 22.039911 -0.842 0.4000

MDVP.Fo.Hz. -0.010897 0.023915 -0.456 0.6486

MDVP.Fhi.Hz. -0.007390 0.005707 -1.295 0.1953

MDVP.Flo.Hz. -0.007497 0.012391 -0.605 0.5452

MDVP.Jitter... -2833.311186 1635.791440 -1.732 0.0833 .

MDVP.Jitter.Abs. -11233.248657 90115.065910 -0.125 0.9008

MDVP.RAP -141716.385893 167219.195820 -0.847 0.3967

MDVP.PPQ -3151.982614 2291.112354 -1.376 0.1689

Jitter.DDP 49192.557723 55902.133110 0.880 0.3789

MDVP.Shimmer -1998.831978 1404.730980 -1.423 0.1548

MDVP.Shimmer.dB. 179.300894 82.454480 2.175 0.0297 \*

Shimmer.APQ3 -45065.241256 129008.515155 -0.349 0.7268

Shimmer.APQ5 6.683631 540.690633 0.012 0.9901

MDVP.APQ 481.321341 486.182198 0.990 0.3222

Shimmer.DDA 15197.954345 43007.215378 0.353 0.7238

NHR -32.514681 75.672860 -0.430 0.6674

HNR 0.201250 0.277780 0.724 0.4688

RPDE 1.925158 5.784390 0.333 0.7393

DFA 0.113489 10.606834 0.011 0.9915

spread1 -0.403521 2.122172 -0.190 0.8492

spread2 5.700999 7.164123 0.796 0.4262

D2 3.474723 2.073074 1.676 0.0937 .

PPE 49.956526 32.972426 1.515 0.1297

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 176.022 on 156 degrees of freedom

Residual deviance: 66.177 on 134 degrees of freedom

AIC: 112.18

Number of Fisher Scoring iterations: 9

**Backward Regression:** Once the logistic regression with all the variables is built, I have tried to build the backward regression model. Only 9 variables came to significant.

> summary(backwards)

Call:

glm(formula = status ~ MDVP.Fo.Hz. + MDVP.Fhi.Hz. + MDVP.Jitter... +

MDVP.PPQ + Jitter.DDP + MDVP.Shimmer + MDVP.Shimmer.dB. +

D2 + PPE, family = binomial, data = data\_train[-1])

Deviance Residuals:

Min 1Q Median 3Q Max

-2.23126 0.00000 0.02555 0.34470 1.74934

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -6.787695 3.445909 -1.970 0.048863 \*

MDVP.Fo.Hz. -0.020001 0.011406 -1.754 0.079499 .

MDVP.Fhi.Hz. -0.004939 0.003572 -1.382 0.166837

MDVP.Jitter... -2148.123004 859.751489 -2.499 0.012471 \*

MDVP.PPQ -3920.808862 1489.312736 -2.633 0.008473 \*\*

Jitter.DDP 1623.245340 455.755545 3.562 0.000369 \*\*\*

MDVP.Shimmer -1290.896159 519.133227 -2.487 0.012896 \*

MDVP.Shimmer.dB. 166.799651 61.026030 2.733 0.006271 \*\*

D2 2.980269 1.538838 1.937 0.052782 .

PPE 53.322832 14.252239 3.741 0.000183 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 176.022 on 156 degrees of freedom

Residual deviance: 70.929 on 147 degrees of freedom

AIC: 90.929

Number of Fisher Scoring iterations: 8

**Measurement:** Since this is a problem where we need to predict if a person has a disease or not, the cost of misclassification are not same across both the classes. By making a prediction that a person has disease, even though in reality he doesn’t is not of much worry compared to the other case where a person actually has a disease and we predict him as no.

The below matrix tells about the costs involved with this problem. Here “Recall” is very very important compared to “Precision”.

|  |  |  |
| --- | --- | --- |
| **Actual/Predicted** | **No** | **Yes** |
| **No** | NoCost | LowCost |
| **Yes** | HighCost | NoCost |

**Train Performance:** At a cutoff of 0.3 (Since class distribution is not balanced, I chose 0.3)

|  |  |  |
| --- | --- | --- |
| **Actual/Predicted** | **No** | **Yes** |
| **No** | 29 | 10 |
| **Yes** | 5 | 113 |

Precision = 92% and Recall = 96%

**Test Performance:** At 0.3 cutoff

|  |  |  |
| --- | --- | --- |
| **Actual/Predicted** | **No** | **Yes** |
| **No** | 6 | 3 |
| **Yes** | 2 | 27 |

Precision = 0.9 and Recall = 93%

**Model 2:** Build Decision tree models based on CART technique. Tree based models are generally robust to multicollinear variables since at the time of splitting it will only choose the best variable.

Below is the final tree which is obtained after pruning:

****

**Performance:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Actual/Predicted** | **No** | **Yes** |  |
| **No** | 32 | 7 |  |
| **Yes** | 2 | 116 |  |
|  |  | P | 94% |
|  |  | R | 98% |
|  |  |  |  |
| **Actual/Predicted** | **No** | **Yes** |  |
| **No** | 4 | 5 |  |
| **Yes** | 0 | 29 |  |
|  |  | P | 85% |
|  |  | R | 100% |

**RandomForest Models:** Finally I have tried random forest model. Tuning is done for one variable mtry i.e. the number of variables to available randomly at each node for splitting. I have used caret package.

Top 10 important variables from the rf model:

Overall

PPE 100.000

spread1 86.260

MDVP.Fo.Hz. 57.032

MDVP.Flo.Hz. 32.534

MDVP.APQ 27.446

NHR 26.002

Shimmer.APQ5 22.347

Jitter.DDP 22.286

spread2 22.129

MDVP.RAP 14.568

**Performance:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Actual/Predicted** | **No** | **Yes** |  |
| **No** | 39 | 0 |  |
| **Yes** | 0 | 118 |  |
|  |  | Precision | 100% |
|  |  | Recall | 100% |
|  |  |  |  |
| **Actual/Predicted** | **No** | **Yes** |  |
| **No** | 6 | 3 |  |
| **Yes** | 0 | 29 |  |
|  |  | Precision | 91% |
|  |  | Recall | 100% |

**Comparison of 3 models:**

Among all the 3 models, random forest did the better job. It had highest recall among all the 3 models. Training performance for Random Forest model is also good compared to the other 2 models.

**Key Inferences:**

1. PPE variable has come important across all the three models
2. Recall is important since we can’t predict a guy with a disease as non-disease person
3. Multicollinearity is present in the data. So, I have tried Trees and RF which are robust to these problems