Understanding Depression through measures of Heart Beat

Finding the relationship between heart rate variability (HRV) and depression from an NIH validated dataset

UCB Data Analytics BootCamp
Loveleen Bangar
Fatema Nourzaie

Heart Rate Variability (HRV): Understanding its importance

Previous studies have linked depression to heart rate variability, a noninvasive index of cardiac autonomic nervous system regulation.

Depression is a disease with high prevalence.

- **It is source of burden of disease and contributes significantly to years lived with disability.**

What is Heart Rate Variability?

Heart rate variability can be traced back to our autonomic nervous system (ANS).

The autonomic nervous system regulates systems in our body, including heart and respiration rate and digestion.

- The autonomic nervous system has a *parasympathetic* (rest) and a *sympathetic* (activation) branch.
    - See appendix a

Formula to explain HRV and ANS:

Parasympathetic Active (rest) + Sympathetic Inactive (activation) **= lower heart rate + high HRV**

Problem Statement

HRV has been shown to be linked to depression. Given that we have a dataset that is a nationally representative sample:

- How can we understand the role of heart rate variability to predict depression levels?

Feeling Depressed?

You know your Heart Rate

Variability?

Let’s use HRV to predict

Depression level

using machine learning.

Data Cleaning and Validation

Dataset

- NIH compiled dataset with various measure on heart rate, demographics, mortality, and survey responses for measuring depression
  - Variables are combination of categorical, and numeric
- 45 columns, 972 rows
- Excluded rows of data with missing values, dropped columns not related to analysis

Data Cleaning

Initial Dataset

gender_x	race_x	age_s1_x	ang_date	chd_dthdt	chf_date	cvd_dthdt	mi_date	stk_date	...	ihr	NN_RF
2	3	56	NaN	NaN	NaN	NaN	NaN	NaN	...	75.112951	0.9920
2	3	56	NaN	NaN	NaN	NaN	NaN	NaN	...	70.633025	0.9688
1	1	40	NaN	NaN	NaN	NaN	NaN	NaN	...	59.355400	0.9966
1	1	40	NaN	NaN	NaN	NaN	NaN	NaN	...	56.666320	0.9964
2	1	60	NaN	NaN	NaN	NaN	NaN	NaN	...	64.954397	0.9907

Cleaning Dataset: removing NaNs and columns not relevant to analysis

Gender	Race	Age	BLUE25	ihr	NN_RR	AVNN	SDNN	VLF	LF	HF	HF_n
2	3	56	5.0	75.112951	0.992000	798.797	46.21180	695.6390	285.53000	827.04600	0.7433
1	1	40	6.0	59.355400	0.996610	1010.860	39.43320	739.2121	531.02400	242.83100	0.3137
2	1	60	5.0	64.954397	0.990741	923.725	27.72130	532.0790	168.40600	81.90150	0.3272
2	1	71	6.0	76.865428	0.997389	780.585	8.41453	55.3561	8.87934	4.38036	0.3303
1	1	48	5.0	62.348286	0.996785	962.336	88.53440	3987.3524	3269.69000	292.34900	0.0820

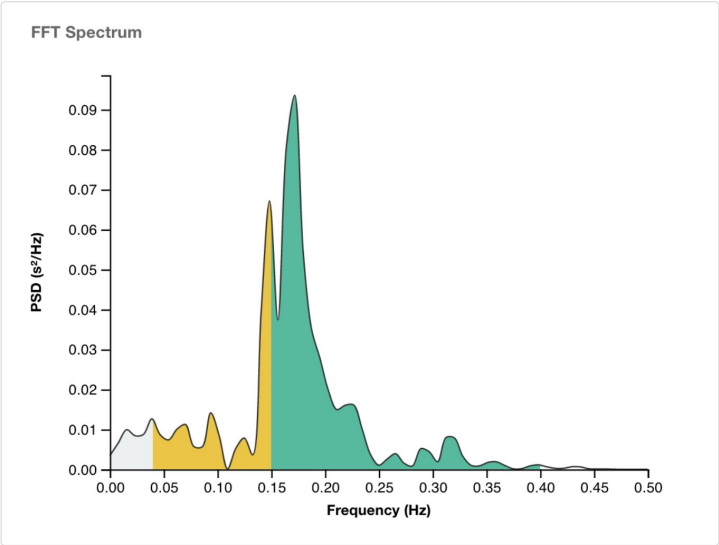
Data Cleaning

Creating dummy variables for “depression” parameter

Gender	Race	Age	BLUE25	ihr	NN_RR	AVNN	SDNN	VLF	LF	HF	HF_n
2	3	56	1.0	75.112951	0.992000	798.797	46.21180	695.6390	285.53000	827.04600	0.7433
1	1	40	1.0	59.355400	0.996610	1010.860	39.43320	739.2121	531.02400	242.83100	0.3137
2	1	60	1.0	64.954397	0.990741	923.725	27.72130	532.0790	168.40600	81.90150	0.3272
2	1	71	1.0	76.865428	0.997389	780.585	8.41453	55.3561	8.87934	4.38036	0.3303
1	1	48	1.0	62.348286	0.996785	962.336	88.53440	3987.3524	3269.69000	292.34900	0.0820

A Closer Look at the Data

What are HRV Frequency Measurements (LF, HF, LF/HF)



SDNN is the gold standard measure for HRV.

- Measure used for Apple Watch to find HRV

In addition to SDNN we will be looking at other parameters in the dataset that can affect depression

Machine Learning Algorithm

Supervised Learning: Classification

Predicting the level of depression (dependent variable) with HRV and other independent variables

Logistic Regression

- Create model
- Fit and Train model
- Validate using Test data
- Make Predictions
- Confusion matrix model

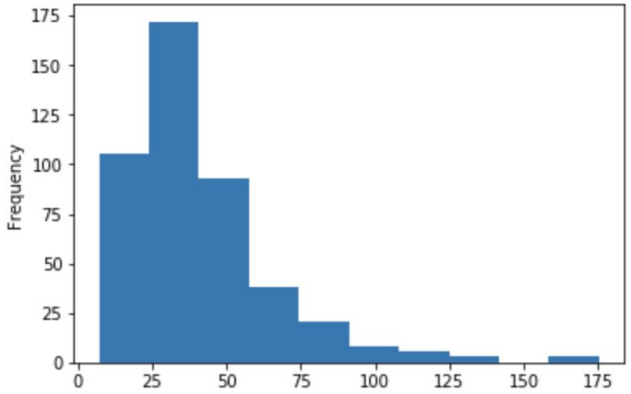
Random Forest Classification

Initial Analysis

To gain a more clear understanding of what is occurring with the parameters provided in the data set...

```
dep_data['SDNN'].plot.hist(bins=10)
```

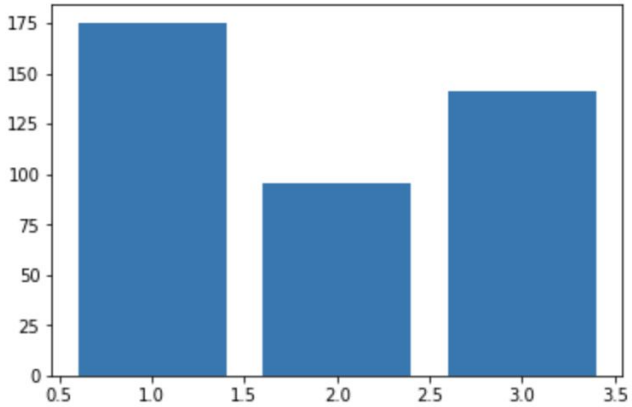
```
<matplotlib.axes._subplots.AxesSubplot at 0x1a1e9f6860>
```



Mapped out independent variable (SDNN), left skewed histogram

```
plt.bar(test_data['Race'], dep_data['SDNN'])
```

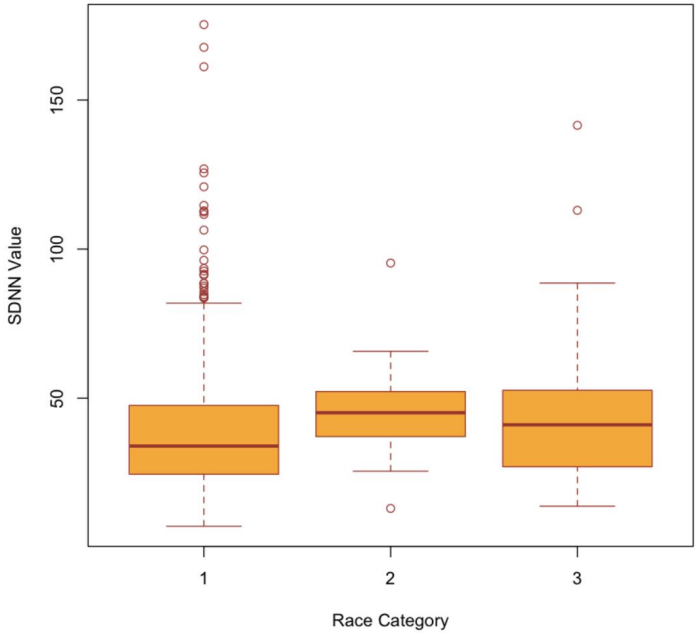
```
<BarContainer object of 449 artists>
```

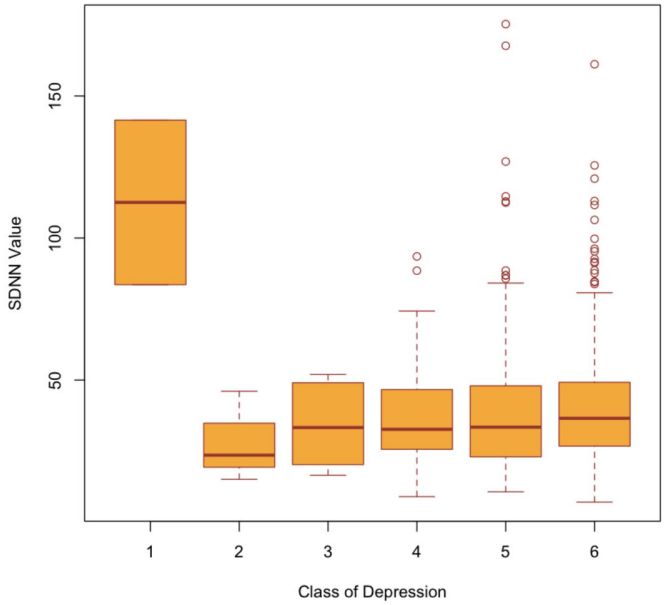


Race against SDNN

Initial Analysis

Mapping Boxplots...





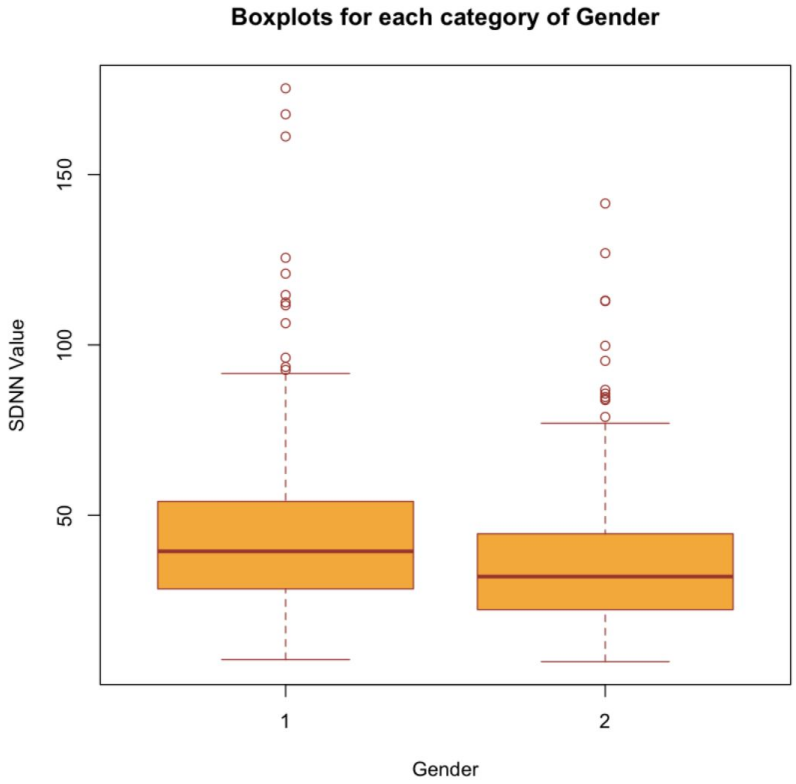
Initial Analysis

Boxplot analysis

Shows outliers

Shows distribution of data points across the strata

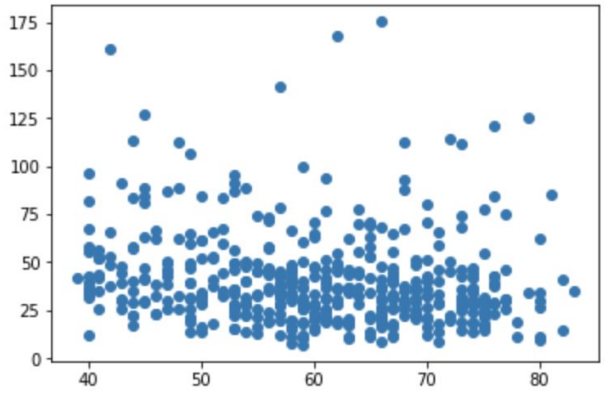
What have we learned from the data so far?



Initial Analysis

```
plt.scatter(dep_data['Age'], dep_data['SDNN'])
```

```
<matplotlib.collections.PathCollection at 0x1a1ea52080>
```



Spread of Age against SDNN

Age (yr)	SDNN (ms)
10–19	176 ± 38
20–29	153 ± 44
30–39	143 ± 32[a]
40–49	132 ± 30[a]
50–59	121 ± 27[a]
60–69	121 ± 32[a]
70–79	124 ± 22[a]
80–99	106 ± 23[abc]

Normal range for Age groups

What do we know so far?

Dataset, although validated and nationally representative is not providing a clear picture for what roles the dependent variables are playing yet.

When looking at the SDNN ranges compared to standard ranges, we see that our dataset is overall lower, possibly indicating a population that has lower SDNN than the actual US population.

Logistic Regression

1. Data cleaning step completed
2. Created dummy variables for depression score
3. Selected dataset for machine learning
4. Split data for train and test
5. Fitted data into Logistic Regression Model
6. See jupyter notebook for details

Dependent variable:

BLUE25 = Quality of Life (QOL) (Sleep Heart Health Study Visit One (SHHS1)): Felt downhearted and blue.

1: All of the time
2: Most of the time
3: A good bit of the time
4: Some of the time
5: A little of time
6: None of the time

Training Data Score: 0.9821428571428571
Testing Data Score: 0.9557522123893806

Logistic Regression

We did a few...using R and Python

Result of data analysis.

The parameter race is statistically significant.

Independent variable, Race, may play a role in influencing the dependent variable (depression)

```
logitMod <- glm(depression ~ SDNN + race_x + gender_x + VLF + HF + LF
summary(logitMod)
```

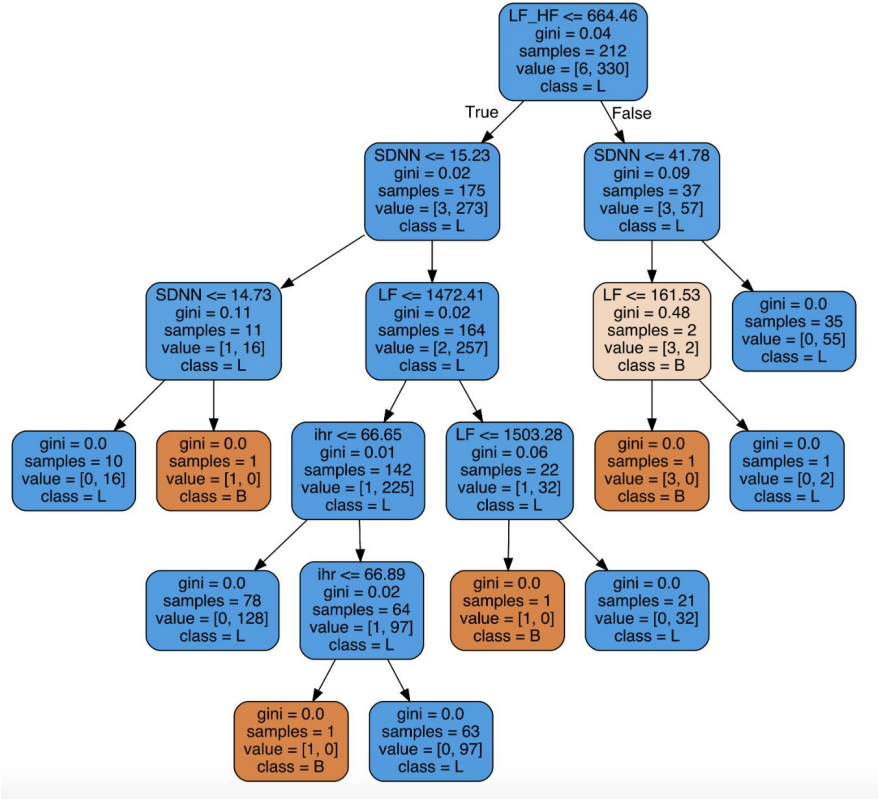
```
Call:
glm(formula = depression ~ SDNN + race_x + gender_x + VLF + HF +
    LF + LF_HF, family = binomial(link = "logit"), data = DF2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3214   0.4197   0.4571   0.4847   0.8369

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.248e+00  8.223e-01   2.734  0.00626 **
SDNN         8.705e-03  2.027e-02   0.429  0.66757
race_x      -5.426e-01  1.963e-01  -2.764  0.00570 **
gender_x     8.290e-02  3.128e-01   0.265  0.79100
VLF         -7.065e-05  1.144e-04  -0.618  0.53685
HF          -6.996e-05  2.603e-04  -0.269  0.78812
LF           1.075e-04  3.501e-04   0.307  0.75875
LF_HF        4.603e-02  7.834e-02   0.588  0.55679
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 317.84  on 448  degrees of freedom
Residual deviance: 309.11  on 441  degrees of freedom
AIC: 325.11

Number of Fisher Scoring iterations: 5
```

Random Forest Classification

Decision tree for Random Forest:

Test and train data, accuracy is 100%

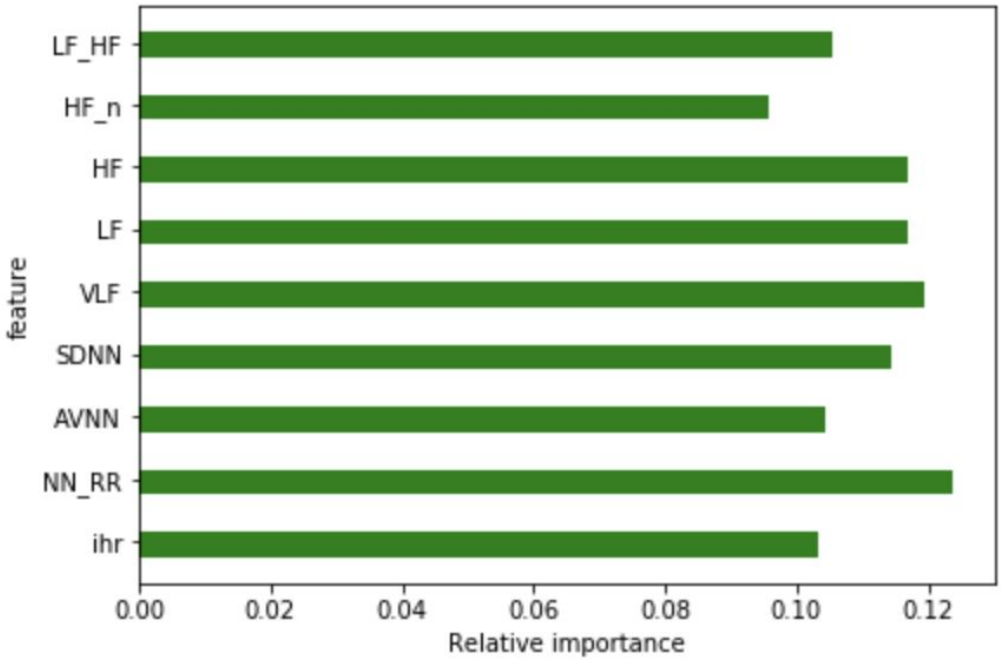
For 3 and 9 attributes



Random Forest Classification

Importance Graph:

Shows the relative Importance of all variables in our analysis



Conclusions

After using a variety of analytical methods, our results do not reproduce what we have seen in the academic literature that shows support for HRV in predicting depression levels.

There is significance in racial category and level of SDNN.

This has been indicated in scientific literature that specific populations may be prone to or exposed to more stressors in the environment which can lead to more depressive health outcomes.

Requirements Satisfied

Scikit Learn

Python Pandas

Python Matplotlib

Machine learning: logistic regression, random forest

Created an analysis of existing data to make a prediction and used classification

Jupyter Notebook

Data Cleaning and Analysis:

https://github.com/loveleenb/Depression_Analysis

Please note that data is not hosted on the Github channel due to privacy restrictions

Appendix A

- Research indicates a reduced HRV indexed by lower values of SDNN, RMSSD and HF power and increased values of LF power for patients with depression in comparison to healthy controls.
- LF/HF Ratio has repeatedly been found to be decreased in depression
- HRV has been associated with severity of depression and parameters derived from HRV have been applied to delineate the severity of depression or even changes in symptom severity
- The Sympathetic Nervous System controls your body's "fight or flight" reactions in response to internal or external stressors. It stimulates blood glucose (to fuel your muscles), pupil dilation (to see tigers better), slows digestion/peristalsis (to focus energy on the present danger), and **increases heart rate** (to ensure adequate blood circulation to run or fight). The SNS is ideally activated to overcome short term stress situations such as running from a tiger or fighting an intruder.
- The Parasympathetic Nervous System controls your body's "rest and digest" responses and is associated with recovery. Parasympathetic activation conserves energy, constricts pupils, aids digestion, and **slows heart rate**. The PSNS is meant to help build for the long term and is needed to grow faster, stronger, and healthier.
- The SNS and PSNS control the same organs with opposite effects. Both branches are always working and both are needed to maintain homeostasis (balance or equilibrium) in your body.
- Apple watch uses SDNN value for HRV calculation

Questions?

Thanks to

Minda for providing dataset and guidance

Gracias

Sash, Ryan, Minda, and Beheshteh

For so much help to reach the finish line.

And

Congratulations to all of us, we made it.