Data Science: Machine Learning

Samantha Scott[a]

[a]*Stellenbosch University, Cape Town, South Africa*

Keywords: Machine Learning, Heart Disease Prediction, Random Forests

Email address: 20945043@sun.ac.za (Samantha Scott)

Table of Contents

```
## Recipe
##
## Inputs:
##
##       role #variables
##    outcome          1
##  predictor         13
##
## Operations:
##
## Dummy variables from all_nominal()
```

```
##
##         0         1
## 0.5555556 0.4444444
```

1. Introduction

The following paper is a comparison between two Machine Learning algorithms, namely Random Forests and Support Vector Machines, as prediction tools. Using a Linear Regression model as a baseline, the RMSE scores are compared.

2. Research Question

Problem type: supervised binomial classification

"Much like EDA, the ML process is very iterative and heurstic-based. With minimal knowledge of the problem or data at hand, it is difficult to know which ML method will perform best. This is known as the no free lunch theorem for ML (Wolpert 1996). Consequently, it is common for many ML approaches to be applied, evaluated, and modified before a final, optimal model can be determined. Performing this process correctly provides great confidence in our outcomes. If not, the results will be useless and, potentially, damaging.1"

"RMSE: Root mean squared error. This simply takes the square root of the MSE metric so that your error is in the same units as your response variable. If your response variable units are dollars, the units of MSE are dollars-squared, but the RMSE will be in dollars. Objective: minimize"

3. Data and Methodology

The data used in this investigation is heart disease data from Kaggle.

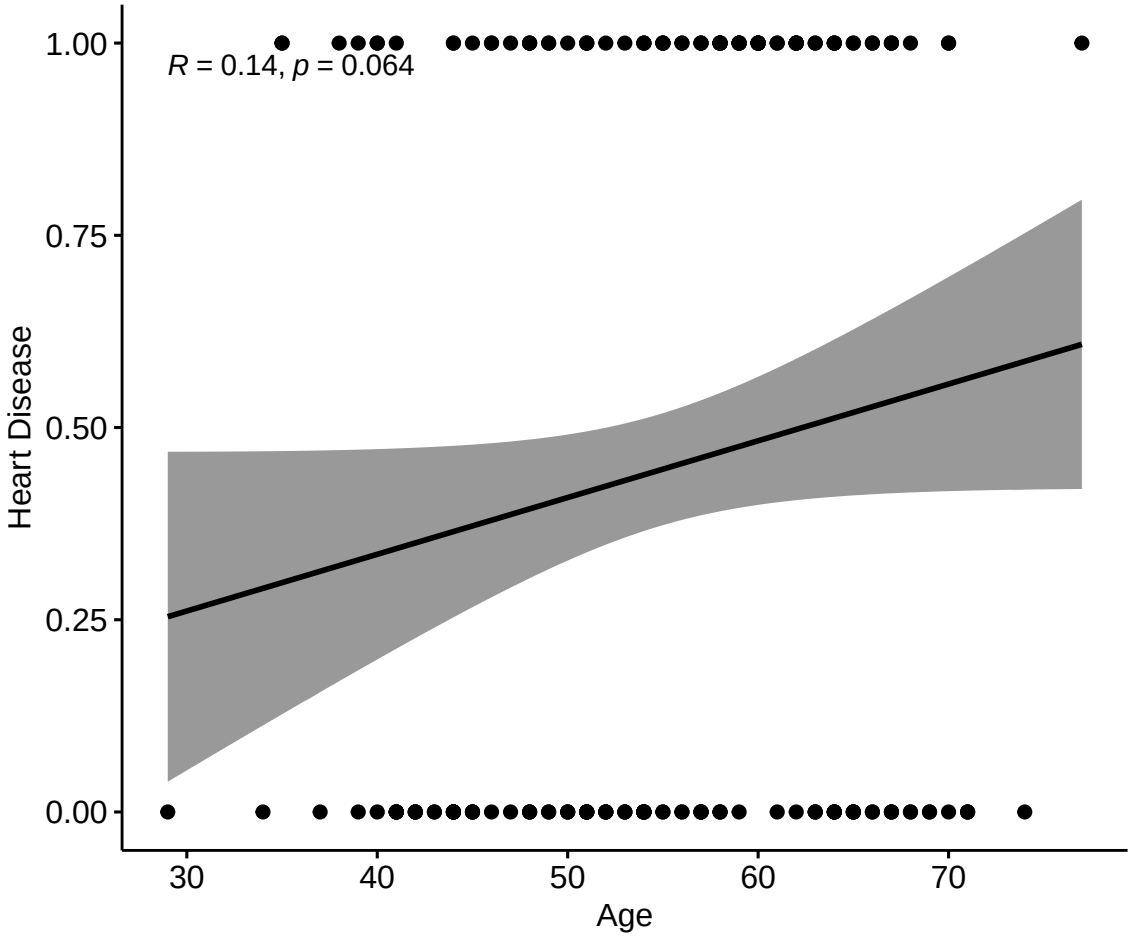
4. Results and Discussion

4.1. Linear Regression

```
##
## Call:
## lm(formula = heart_disease_present ~ ., data = heart_d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.87287 -0.22568 -0.04891  0.20849  0.93636
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     -0.1608082  0.4533297  -0.355  0.72325
## slope_of_peak_exercise_st_segment  0.0830549  0.0594343   1.397  0.16416
## thalnormal                      -0.0801911  0.1419747  -0.565  0.57296
## thalreversible_defect            0.1745812  0.1394239   1.252  0.21228
## resting_blood_pressure           0.0006047  0.0017496   0.346  0.73005
## chest_pain_type                  0.0885153  0.0324033   2.732  0.00699 **
## num_major_vessels                0.1346337  0.0318310   4.230 3.87e-05 ***
## fasting_blood_sugar_gt_120_mg_per_dl -0.0414389  0.0778126  -0.533  0.59506
## resting_ekg_results              0.0378482  0.0279049   1.356  0.17685
## serum_cholesterol_mg_per_dl      0.0004803  0.0005405   0.889  0.37551
## oldpeak_eq_st_depression         0.0421798  0.0327468   1.288  0.19953
## sex                              0.1591962  0.0667508   2.385  0.01822 *
## age                             -0.0024920  0.0035588  -0.700  0.48476
## max_heart_rate_achieved         -0.0014613  0.0015861  -0.921  0.35825
## exercise_induced_angina          0.1458715  0.0688974   2.117  0.03574 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3559 on 165 degrees of freedom
## Multiple R-squared:  0.5298, Adjusted R-squared:   0.49
```

```
## F-statistic: 13.28 on 14 and 165 DF,  p-value: < 2.2e-16
```

```
## [1] 0.3558662
```



4.2. Random Forests

```
## [1] 0.3391407
```

4.3. Support Vector Machine

5. Conclusion

6. Reference List