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

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

3. Data and Methodology

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

4. Results

4.1. Linear Regression

```
##
## Call:
## lm(formula = heart_disease_present ~ ., data = heart_d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.81385 -0.23535 -0.07213  0.25418  0.90884
##
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          -0.3435048  0.4431192  -0.775 0.439319
## slope_of_peak_exercise_st_segment     0.0995764  0.0609650   1.633 0.104282
## resting_blood_pressure                0.0008023  0.0017905   0.448 0.654683
## chest_pain_type                       0.1120314  0.0329010   3.405 0.000828 ***
## num_major_vessels                     0.1438576  0.0328175   4.384 2.06e-05 ***
## fasting_blood_sugar_gt_120_mg_per_dl -0.0479164  0.0801823  -0.598 0.550921
## resting_ekg_results                   0.0244777  0.0286543   0.854 0.394196
## serum_cholesterol_mg_per_dl           0.0005870  0.0005571   1.054 0.293524
## oldpeak_eq_st_depression              0.0616716  0.0334386   1.844 0.066907 .
## sex                                   0.2413338  0.0638961   3.777 0.000220 ***
## age                                  -0.0028883  0.0036829  -0.784 0.434008
## max_heart_rate_achieved              -0.0014835  0.0016266  -0.912 0.363063
## exercise_induced_angina               0.1960958  0.0695805   2.818 0.005412 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3685 on 167 degrees of freedom
## Multiple R-squared:  0.4899, Adjusted R-squared:  0.4532
## F-statistic: 13.36 on 12 and 167 DF,  p-value: < 2.2e-16
```

```
## [1] 0.3684575
```

4.2. Random Forests

```
## [1] 0.4178554
```

```
##
##      0  1
##   0 16  8
##   1  7 23
```

```
## [1] 0.7222222
```

4.3. Support Vector Machine

5. Conclusion

6. Reference List