CS445 NPM Package Modeling

Matthew MacNeil

Selection/Linear Regression

PCA:

- finalScore
- numDownloads
- releases
- testsSize

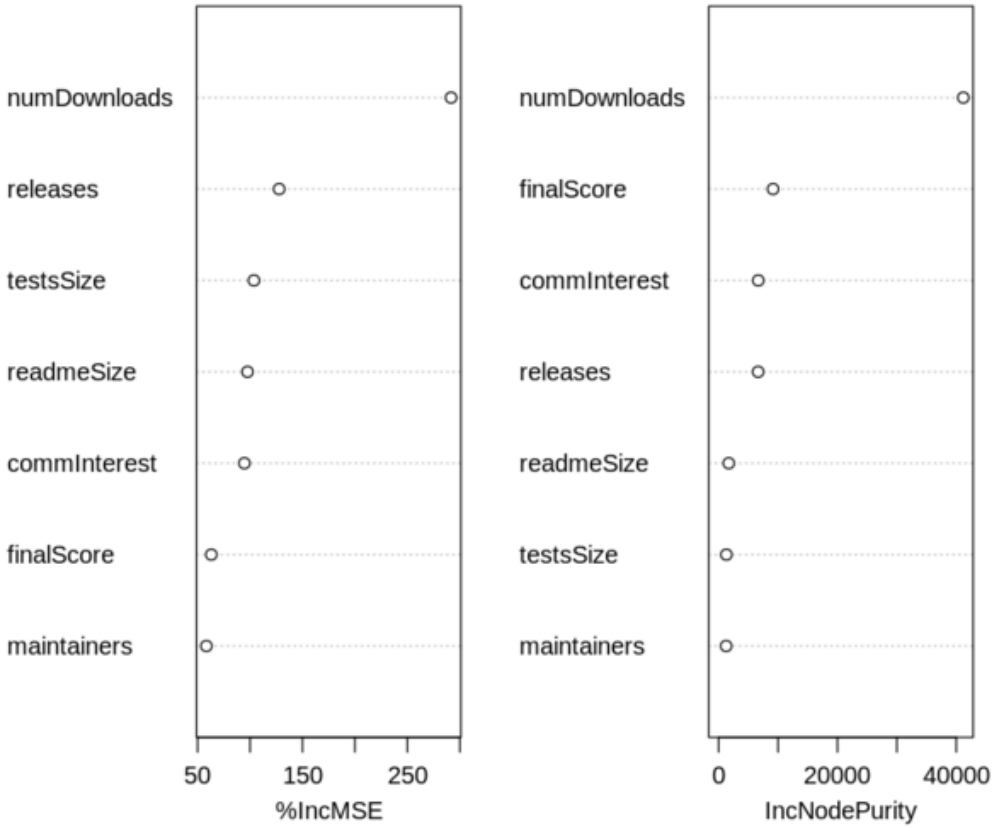
Regression

- commInterest
- numCommits

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.988472   0.026282  113.71  < 2e-16 ***
finalScore   -0.805860   0.045231  -17.82  < 2e-16 ***
numDownloads  0.923898   0.002055  449.56  < 2e-16 ***
devDepends   -0.003197   0.002014   -1.59     0.11
releases     -0.042762   0.002031  -21.06  < 2e-16 ***
testsSize     0.016267   0.000803   20.25  < 2e-16 ***
maintainers   0.034586   0.004130    8.37  < 2e-16 ***
readmeSize    0.009159   0.001016    9.02  < 2e-16 ***
commInterest  0.054198   0.001985   27.30  < 2e-16 ***
numCommits   -0.037362   0.006325   -5.91  3.5e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.47 on 20037 degrees of freedom
Multiple R-squared:  0.937,     Adjusted R-squared:  0.937
F-statistic: 3.31e+04 on 9 and 20037 DF,  p-value: <2e-16
```

Random Forest



Number of downloads has biggest impact on both MSE and IncNodePurity

Random Forest model much more able to predict number of downloads than the linear model