Forecasting School Budgets
Predicated on Expected Student Attendance

Andrew Harris, Elihu Whitney, Priyanka Biswas, Sarvari Ventrapragada

May 1, 2016

Contents

1 Abstract

Using multiple machine learning algorithms (linear regression, elastic net techniques, support vector machines, neural nets, etc.) we attempted to demonstrate a relationship between student demographics and number of absences for a sample data set obtained from the UCI machine learning repository [9].

2 Objective

Schools are typically paid, and hence determine their budgets, based on the attendance rate of their students. To accurately forecast yearly budgets it would be beneficial if there were a way to determine if there is a relationship between student demographics and expected absences.

3 Data

For the purposes of our analysis, we used a data set of math students at one particular high school in Portugal [10]. Please see methodology section for additional information.

4 Methodology

The data set contains two distinct csv files, one for the math and one for the Portuguese classes. For the sake of convenience, we limited our analysis to just one file. Given that Portuguese is only taught in a few countries, Brazil & Portugal specifically, we chose the math data set as it will be more telling as to the rates of absence worldwide.

Given that there are two distinct high schools being represented in the sample, we want to limit our scope to just one high school as there may be unique conditions specific to each that would influence our analysis.

We also need to remove the grade variables "G1, G2, and G3" for the sake of code simplicity. Clearly there is some kind of relationship between final grades and the number of absences in the year, but the converse is not true. It is not the case that the final grades from the current period cause the number of absences. Therefore, as at least linear regression is predicated on a one way causality from x to y, an attempt to reverse it would violate the assumptions of regression.

At this point, we explore the numerical factors and the associated principal components to get a sense of the data and what predictors will be important in the following analysis, noting that we scaled the principal components to prevent variable magnitude domination.

From the corrplot, we note that there are few issues with multicollinearity, and hence we can proceed without worrying too much about that assumption. We also note that the performance of the PCA was rather poor, with the first two principal components accounting for only 36% of the information. However, we can still use the biplot to determine which of the numeric variables are relatively important and can then use them in our subjective models.

Based on the biplot, we note that there are three sets of variables highlighted by the visualization, age, Walc, and Medu/Fedu; where Walc is weekly alcohol consumption and Medu and Fedu represent mother and father education respectively. As this information is coming from student surveys, there is likely to be under-reporting in the Walc category, so for our subjective model we will exclude Walc. Given the high correlation between mother and father education, we will solely use father's education as it has more influence on the principal components.

Using the information derived above, we subjectively looked through the variables and judgmentally selected the variables we felt would be most appropriate for the regression. Giving us a regression model best guess, which can be expressed as follows:

$$\text{Absences} = \beta_0 + \beta_1 T + \beta_2 A + \beta_3 S + \beta_4 P + \beta_5 H + \beta_6 F + \epsilon$$

Where T, A, S, P, H, and F represent travel time to school, age of the student, sex of the student, status of parental separation, desire for a higher education, and educational attainment of the father respectively. These variables were, at least partially, selected due to that fact that they would be relatively easy to retrieve from district records or parent questionnaires upon student enrollment.

To test our model, we will be using cross-validation techniques from various R packages, with Leave-one-out Cross Validation (LOOCV) being preferred and 10 fold CV being considered a reasonable estimate of LOOCV. To test the normality of the residuals and the presence of heteroskedasticity in the model, we use the Shapiro and Breusch-Pagan tests respectively. Based on the results of the hypothesis test, we note that the model rejects both null hypotheses, that the residuals are normally distributed and that the distribution is homoskedastic, and the note that the model has multiple OLS violations. However, the ultimate test in this case is predictive power, so we reserved judgment until we reviewed the CV error of all models at the end of our script.

In addition to our best guess model, we also considered a regression on all regressors and a step-wise backwards regression as potential model candidates.

Given the poor predictive power of the linear model, we expanded the scope of our model search to shrinkage methods such as elastic net regression and partial least squares, and tree based methods; where we considered a random forest and boosting model. Finally, we considered a support vector machine (SVM) method with the kernel and parameters chosen automatically by algorithm, and several neural net models with varying degrees of hidden layers. Please note that for the purposes of calculating the elastic-net regression, we had to create a "model-matrix" that had appropriate dummy variables for the categorical variables.

Collecting all the CV errors together, we made a table of our results, expressed in MSE, which we converted to root mean square error for comparative purposes to the mean absences noted. RMSE error and mean observed absences are both raised to unitary powers, whereas MSE is raised to the second power. We then compared our error rates relative to the observed average per-student number of absences to get a sense of the percentage error of our model.

5 Statistical Insight

Based on our cross-validation testing, we determined that the percentage error is too high for any of these methods to be reliably used to predict absences.

Given our current variables and the bias introduced by limiting scope to one class subject at one school, additional observations or reconsideration of variables would be the most appropriate step at this point.

6 Executive Summary

Based on our model, there is no way to accurately predict number of student absences. The model should be fundamentally reconsidered as, in its current state, it has little predictive power.

A Associated R Code

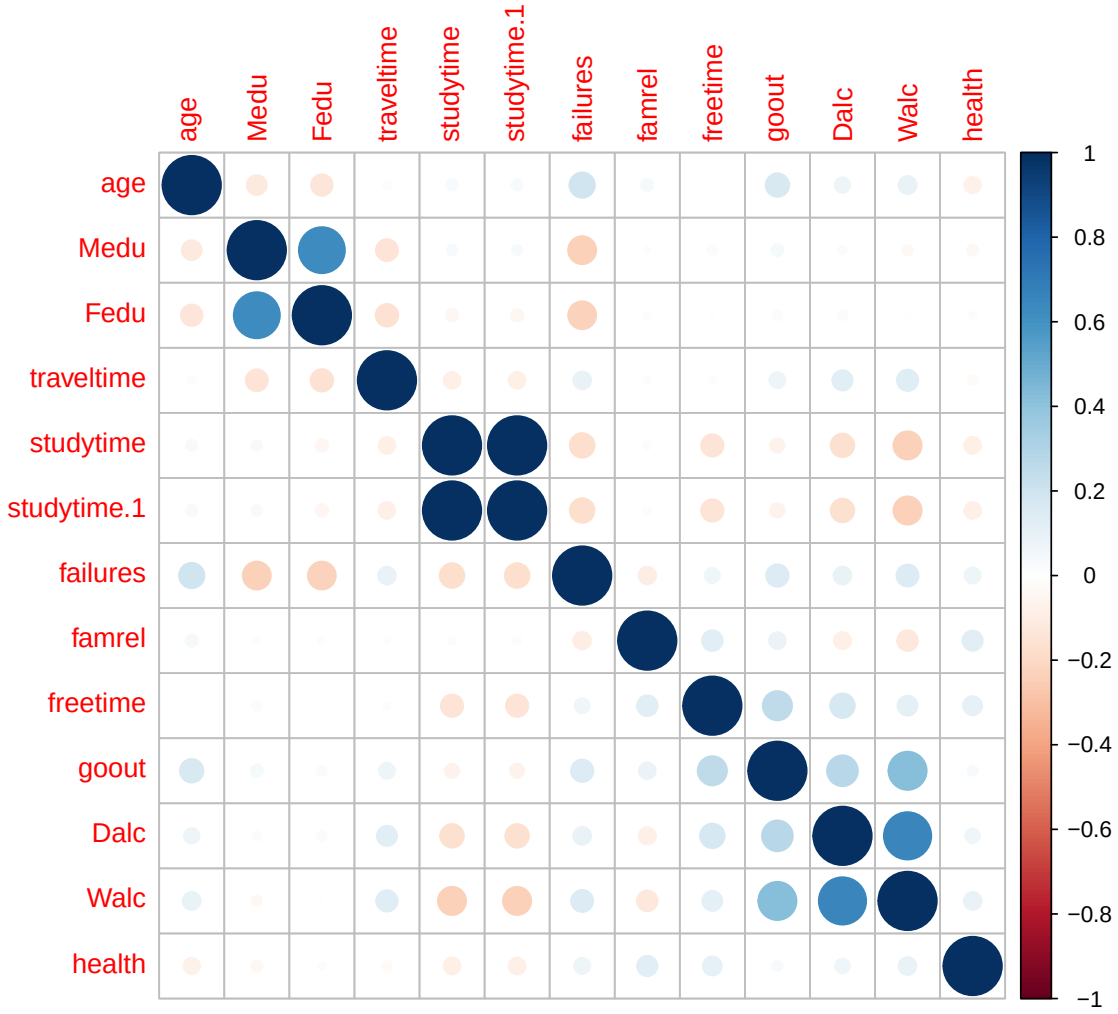
```
# Load required packages ===============================================

suppressMessages(require(boot))  # CV for General Linearized Models
suppressMessages(require(corrplot))  # Corrplot functions
suppressMessages(require(lmtest))  # To get BP test for heteroskedasticity
suppressMessages(require(MASS))  # Stepwise Function
suppressMessages(require(glmnet)) # Elastic net packages
suppressMessages(require(pls))  # Partial Least Squares package
suppressMessages(require(randomForest))  # Random forest package
suppressMessages(require(gbm))  # Boosting
suppressMessages(require(e1071))  # SVM package
suppressMessages(require(nnet))  # Neural net package

# Load Data ============================================================
math <- read.csv2("student-mat.csv")  # To deal with period separated csv
math <- na.omit(math) # To remove missing observations
math <- math[math$school == "GP", ]  # To filter out only GP schools
math <- math[-c(1, 31, 32, 33)]   # To remove school and grade categories
attach(math)

# Exploring the data ===================================================
numeric_preds <- math[, c("age", "Medu", "Fedu", "traveltime",
                          "studytime", "studytime", "failures",
                          "famrel", "freetime", "goout",
                          "Dalc", "Walc", "health")]

corrplot::corrplot(cor(numeric_preds))  # Corrplot to explore relationships
```

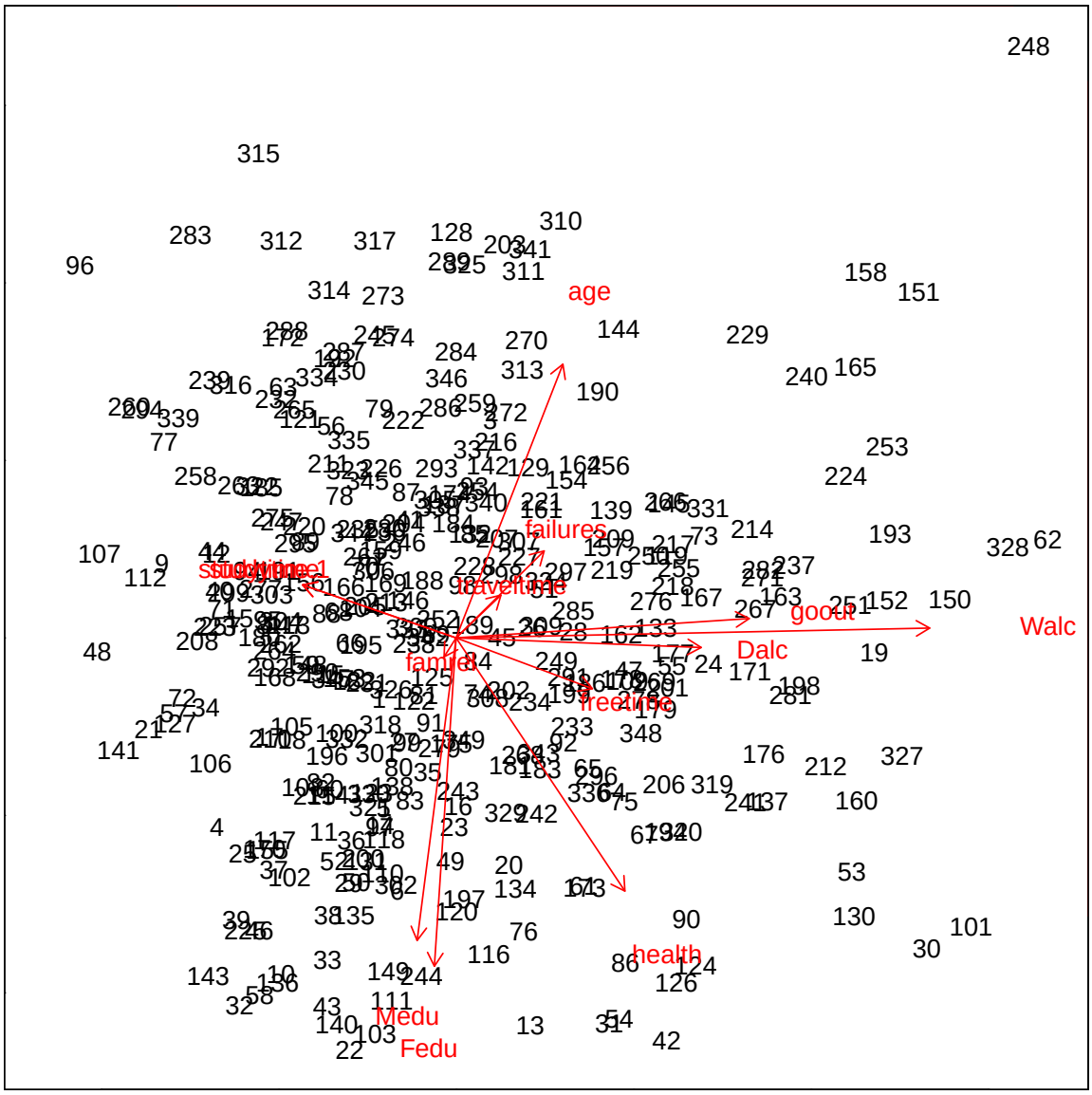



```
pca_numeric <- princomp(numeric_preds, scale = TRUE)
summary(pca_numeric)
```

```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3    Comp.4     Comp.5
## Standard deviation     1.6809870 1.4375313 1.3876761 1.2064252 1.10755290
## Proportion of Variance 0.2086818 0.1526126 0.1422107 0.1074872 0.09059097
## Cumulative Proportion  0.2086818 0.3612945 0.5035051 0.6109923 0.70158331
##                           Comp.6     Comp.7     Comp.8     Comp.9
## Standard deviation     1.0423678 0.85189072 0.81538922 0.66128584
```

```
## Proportion of Variance 0.0802413 0.05359493 0.04910049 0.03229493
## Cumulative Proportion  0.7818246 0.83541954 0.88452003 0.91681497
##                           Comp.10    Comp.11    Comp.12      Comp.13
## Standard deviation     0.64346148 0.63182001 0.55959988 1.490115e-08
## Proportion of Variance 0.03057743 0.02948103 0.02312657 1.639818e-17
## Cumulative Proportion  0.94739240 0.97687343 1.00000000 1.000000e+00

biplot(pca_numeric)
```



```r
# Linear Regression ============================================================

# Estimation based on PCA analysis and judgement
guess <- glm(absences ~ traveltime + age + sex + Pstatus + higher + Fedu,
             data = math)
summary(guess)
```

```
##
## Call:
## glm(formula = absences ~ traveltime + age + sex + Pstatus + higher +
##     Fedu, data = math)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -13.687   -4.487   -1.663    2.060   61.560
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -18.1159     7.2722  -2.491  0.01321 *
## traveltime    0.5189     0.6574   0.789  0.43046
## age           1.6391     0.3684   4.450 1.17e-05 ***
## sexM         -1.3605     0.8781  -1.549  0.12225
## PstatusT     -3.6696     1.3874  -2.645  0.00855 **
## higheryes    -1.1515     2.1204  -0.543  0.58745
## Fedu          0.5068     0.4156   1.220  0.22347
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 64.63888)
##
##     Null deviance: 24216  on 348  degrees of freedom
## Residual deviance: 22106  on 342  degrees of freedom
## AIC: 2454.3
##
## Number of Fisher Scoring iterations: 2
```

```r
shapiro.test(resid(guess))  # To test normality assumption
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(guess)
## W = 0.75224, p-value < 2.2e-16
```

```r
bptest(guess)  # To test for heteroskedasiticty
```

```
##
```

```
##  studentized Breusch-Pagan test
##
## data:  guess
## BP = 21.948, df = 6, p-value = 0.001237

cv_guess <- cv.glm(math, guess)$delta[2]


# Full Linear Regression
linreg <- glm(absences ~ ., data = math)
summary(linreg)

##
## Call:
## glm(formula = absences ~ ., data = math)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -12.705   -4.207   -1.037    2.561   55.750
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -22.3376     8.9937  -2.484 0.013529 *
## sexM              -2.0738     1.0023  -2.069 0.039369 *
## age                1.4471     0.4191   3.453 0.000631 ***
## addressU          -1.2271     1.2242  -1.002 0.316925
## famsizeLE3         0.9011     0.9966   0.904 0.366563
## PstatusT          -2.4200     1.4248  -1.698 0.090414 .
## Medu               1.4634     0.6636   2.205 0.028167 *
## Fedu              -0.2134     0.5731  -0.372 0.709867
## Mjobhealth        -3.4778     2.2964  -1.514 0.130935
## Mjobother          0.5189     1.5016   0.346 0.729890
## Mjobservices      -0.1344     1.6517  -0.081 0.935185
## Mjobteacher       -0.8845     2.1518  -0.411 0.681300
## Fjobhealth         2.7559     2.8958   0.952 0.341999
## Fjobother          1.4791     2.1598   0.685 0.493978
## Fjobservices       2.4660     2.2406   1.101 0.271925
## Fjobteacher        0.1647     2.6825   0.061 0.951074
## reasonhome         2.9784     1.1218   2.655 0.008337 **
## reasonother        2.4869     1.8213   1.365 0.173086
## reasonreputation   2.4983     1.1438   2.184 0.029700 *
## guardianmother     1.7052     1.1141   1.531 0.126873
## guardianother      3.4801     2.0296   1.715 0.087404 .
## traveltime         0.5676     0.7097   0.800 0.424417
## studytime         -1.1541     0.5817  -1.984 0.048118 *
## failures          -0.3514     0.6753  -0.520 0.603214
```

```
## schoolsupyes       1.5240     1.2723   1.198 0.231905
## famsupyes          0.3344     0.9638   0.347 0.728852
## paidyes           -1.2133     0.9595  -1.265 0.206993
## activitiesyes     -0.3011     0.9008  -0.334 0.738416
## nurseryyes         0.6716     1.1468   0.586 0.558518
## higheryes         -0.3223     2.2496  -0.143 0.886151
## internetyes        2.9704     1.3014   2.282 0.023140 *
## romanticyes        1.9831     0.9581   2.070 0.039290 *
## famrel            -0.2455     0.5131  -0.478 0.632657
## freetime          -0.5155     0.4820  -1.070 0.285638
## goout             -0.4823     0.4548  -1.060 0.289792
## Dalc              -0.4121     0.6957  -0.592 0.554018
## Walc               1.1100     0.4931   2.251 0.025064 *
## health             0.2820     0.3305   0.853 0.394113
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 60.89168)
##
##     Null deviance: 24216  on 348  degrees of freedom
## Residual deviance: 18937  on 311  degrees of freedom
## AIC: 2462.3
##
## Number of Fisher Scoring iterations: 2

shapiro.test(resid(linreg))  # To test normality assumption

##
##  Shapiro-Wilk normality test
##
## data:  resid(linreg)
## W = 0.80462, p-value < 2.2e-16

bptest(linreg)  # To test for heteroskedasiticty

##
##  studentized Breusch-Pagan test
##
## data:  linreg
## BP = 57.651, df = 37, p-value = 0.01642

cv_lin <- cv.glm(math, linreg)$delta[2]


# Stepwise Linear Regression
step <- stepAIC(linreg, direction = "backward")
```

```
## Start:  AIC=2462.26
## absences ~ sex + age + address + famsize + Pstatus + Medu + Fedu +
##     Mjob + Fjob + reason + guardian + traveltime + studytime +
##     failures + schoolsup + famsup + paid + activities + nursery +
##     higher + internet + romantic + famrel + freetime + goout +
##     Dalc + Walc + health
##
##              Df Deviance    AIC
## - Fjob        4    19101 2457.3
## - Mjob        4    19236 2459.7
## - higher      1    18939 2460.3
## - activities  1    18944 2460.4
## - famsup      1    18945 2460.4
## - Fedu        1    18946 2460.4
## - famrel      1    18951 2460.5
## - failures    1    18954 2460.6
## - nursery     1    18958 2460.7
## - Dalc        1    18959 2460.7
## - traveltime  1    18976 2461.0
## - health      1    18982 2461.1
## - famsize     1    18987 2461.2
## - address     1    18998 2461.4
## - goout       1    19006 2461.5
## - freetime    1    19007 2461.5
## - schoolsup   1    19025 2461.9
## - paid        1    19035 2462.1
## <none>              18937 2462.3
## - guardian    2    19166 2462.4
## - Pstatus     1    19113 2463.5
## - studytime   1    19177 2464.7
## - sex         1    19198 2465.0
## - romantic    1    19198 2465.0
## - reason      3    19449 2465.6
## - Medu        1    19233 2465.7
## - Walc        1    19246 2465.9
## - internet    1    19254 2466.1
## - age         1    19663 2473.4
##
## Step:  AIC=2457.26
## absences ~ sex + age + address + famsize + Pstatus + Medu + Fedu +
##     Mjob + reason + guardian + traveltime + studytime + failures +
##     schoolsup + famsup + paid + activities + nursery + higher +
##     internet + romantic + famrel + freetime + goout + Dalc +
##     Walc + health
##
##              Df Deviance    AIC
```

```
## - Mjob        4    19347 2453.7
## - activities  1    19103 2455.3
## - famsup      1    19104 2455.3
## - higher      1    19105 2455.3
## - famrel      1    19107 2455.4
## - Fedu        1    19111 2455.4
## - failures    1    19113 2455.5
## - Dalc        1    19116 2455.6
## - nursery     1    19117 2455.6
## - traveltime  1    19133 2455.8
## - health      1    19142 2456.0
## - famsize     1    19144 2456.1
## - address     1    19157 2456.3
## - goout       1    19173 2456.6
## - paid        1    19181 2456.7
## - schoolsup   1    19184 2456.8
## - freetime    1    19204 2457.1
## <none>             19101 2457.3
## - guardian    2    19331 2457.4
## - Pstatus     1    19262 2458.2
## - studytime   1    19309 2459.1
## - romantic    1    19356 2459.9
## - Medu        1    19366 2460.1
## - sex         1    19377 2460.3
## - reason      3    19608 2460.4
## - internet    1    19432 2461.3
## - Walc        1    19471 2462.0
## - age         1    19770 2467.3
##
## Step:  AIC=2453.72
## absences ~ sex + age + address + famsize + Pstatus + Medu + Fedu +
##     reason + guardian + traveltime + studytime + failures + schoolsup +
##     famsup + paid + activities + nursery + higher + internet +
##     romantic + famrel + freetime + goout + Dalc + Walc + health
##
##              Df Deviance    AIC
## - famrel      1    19347 2451.7
## - famsup      1    19347 2451.7
## - Dalc        1    19350 2451.8
## - activities  1    19350 2451.8
## - higher      1    19353 2451.8
## - Fedu        1    19353 2451.8
## - failures    1    19358 2451.9
## - nursery     1    19359 2451.9
## - health      1    19373 2452.2
## - famsize     1    19376 2452.2
```

```
## - traveltime  1    19387 2452.5
## - address     1    19401 2452.7
## - goout       1    19429 2453.2
## - paid        1    19434 2453.3
## - freetime    1    19445 2453.5
## <none>              19347 2453.7
## - schoolsup   1    19460 2453.8
## - guardian    2    19598 2454.2
## - studytime   1    19526 2454.9
## - Pstatus     1    19548 2455.3
## - Medu        1    19556 2455.5
## - reason      3    19833 2456.4
## - sex         1    19610 2456.4
## - romantic    1    19616 2456.5
## - Walc        1    19671 2457.5
## - internet    1    19676 2457.6
## - age         1    20051 2464.2
##
## Step:  AIC=2451.74
## absences ~ sex + age + address + famsize + Pstatus + Medu + Fedu +
##     reason + guardian + traveltime + studytime + failures + schoolsup +
##     famsup + paid + activities + nursery + higher + internet +
##     romantic + freetime + goout + Dalc + Walc + health
##
##              Df Deviance    AIC
## - famsup      1    19348 2449.8
## - Dalc        1    19350 2449.8
## - activities  1    19351 2449.8
## - higher      1    19353 2449.8
## - Fedu        1    19354 2449.8
## - failures    1    19359 2449.9
## - nursery     1    19360 2450.0
## - health      1    19373 2450.2
## - famsize     1    19376 2450.3
## - traveltime  1    19388 2450.5
## - address     1    19402 2450.7
## - goout       1    19433 2451.3
## - paid        1    19435 2451.3
## - freetime    1    19448 2451.5
## <none>              19347 2451.7
## - schoolsup   1    19460 2451.8
## - guardian    2    19598 2452.2
## - studytime   1    19527 2453.0
## - Pstatus     1    19549 2453.3
## - Medu        1    19558 2453.5
## - reason      3    19833 2454.4
```

```
## - sex         1    19614 2454.5
## - romantic    1    19619 2454.6
## - internet    1    19676 2455.6
## - Walc        1    19683 2455.7
## - age         1    20052 2462.2
##
## Step:  AIC=2449.75
## absences ~ sex + age + address + famsize + Pstatus + Medu + Fedu +
##     reason + guardian + traveltime + studytime + failures + schoolsup +
##     paid + activities + nursery + higher + internet + romantic +
##     freetime + goout + Dalc + Walc + health
##
##              Df Deviance    AIC
## - Dalc        1    19351 2447.8
## - activities  1    19352 2447.8
## - Fedu        1    19354 2447.9
## - higher      1    19354 2447.9
## - failures    1    19359 2447.9
## - nursery     1    19361 2448.0
## - health      1    19375 2448.2
## - famsize     1    19377 2448.3
## - traveltime  1    19390 2448.5
## - address     1    19404 2448.8
## - goout       1    19434 2449.3
## - paid        1    19436 2449.3
## - freetime    1    19448 2449.5
## <none>             19348 2449.8
## - schoolsup   1    19463 2449.8
## - guardian    2    19599 2450.2
## - studytime   1    19527 2451.0
## - Pstatus     1    19549 2451.3
## - Medu        1    19561 2451.6
## - reason      3    19838 2452.5
## - romantic    1    19619 2452.6
## - sex         1    19622 2452.7
## - internet    1    19678 2453.7
## - Walc        1    19683 2453.7
## - age         1    20053 2460.2
##
## Step:  AIC=2447.81
## absences ~ sex + age + address + famsize + Pstatus + Medu + Fedu +
##     reason + guardian + traveltime + studytime + failures + schoolsup +
##     paid + activities + nursery + higher + internet + romantic +
##     freetime + goout + Walc + health
##
##              Df Deviance    AIC
```

```
## - activities  1    19355 2445.9
## - Fedu        1    19357 2445.9
## - higher      1    19357 2445.9
## - failures    1    19362 2446.0
## - nursery     1    19365 2446.1
## - health      1    19378 2446.3
## - famsize     1    19379 2446.3
## - traveltime  1    19392 2446.5
## - address     1    19406 2446.8
## - goout       1    19437 2447.3
## - paid        1    19440 2447.4
## - freetime    1    19457 2447.7
## <none>              19351 2447.8
## - schoolsup   1    19463 2447.8
## - guardian    2    19602 2448.3
## - studytime   1    19532 2449.1
## - Pstatus     1    19550 2449.4
## - Medu        1    19563 2449.6
## - reason      3    19839 2450.5
## - romantic    1    19620 2450.6
## - sex         1    19634 2450.9
## - internet    1    19679 2451.7
## - Walc        1    19799 2453.8
## - age         1    20053 2458.2
##
## Step:  AIC=2445.87
## absences ~ sex + age + address + famsize + Pstatus + Medu + Fedu +
##     reason + guardian + traveltime + studytime + failures + schoolsup +
##     paid + nursery + higher + internet + romantic + freetime +
##     goout + Walc + health
##
##              Df Deviance    AIC
## - Fedu        1    19362 2444.0
## - higher      1    19363 2444.0
## - failures    1    19366 2444.1
## - nursery     1    19369 2444.1
## - health      1    19381 2444.3
## - famsize     1    19382 2444.4
## - traveltime  1    19394 2444.6
## - address     1    19408 2444.8
## - goout       1    19442 2445.4
## - paid        1    19442 2445.4
## - freetime    1    19464 2445.8
## - schoolsup   1    19465 2445.9
## <none>              19355 2445.9
## - guardian    2    19605 2446.3
```

```
## - studytime   1    19542 2447.2
## - Pstatus     1    19561 2447.6
## - Medu        1    19566 2447.7
## - reason      3    19839 2448.5
## - romantic    1    19621 2448.6
## - sex         1    19654 2449.2
## - internet    1    19681 2449.7
## - Walc        1    19808 2451.9
## - age         1    20056 2456.3
##
## Step:  AIC=2443.99
## absences ~ sex + age + address + famsize + Pstatus + Medu + reason +
##     guardian + traveltime + studytime + failures + schoolsup +
##     paid + nursery + higher + internet + romantic + freetime +
##     goout + Walc + health
##
##              Df Deviance    AIC
## - higher      1    19371 2442.2
## - failures    1    19371 2442.2
## - nursery     1    19375 2442.2
## - health      1    19387 2442.4
## - famsize     1    19391 2442.5
## - traveltime  1    19405 2442.8
## - address     1    19412 2442.9
## - paid        1    19447 2443.5
## - goout       1    19452 2443.6
## - freetime    1    19468 2443.9
## - schoolsup   1    19469 2443.9
## <none>             19362 2444.0
## - guardian    2    19624 2444.7
## - studytime   1    19543 2445.2
## - Pstatus     1    19568 2445.7
## - Medu        1    19619 2446.6
## - reason      3    19851 2446.7
## - romantic    1    19628 2446.8
## - sex         1    19658 2447.3
## - internet    1    19686 2447.8
## - Walc        1    19813 2450.0
## - age         1    20070 2454.5
##
## Step:  AIC=2442.16
## absences ~ sex + age + address + famsize + Pstatus + Medu + reason +
##     guardian + traveltime + studytime + failures + schoolsup +
##     paid + nursery + internet + romantic + freetime + goout +
##     Walc + health
##
```

```
##               Df Deviance    AIC
## - failures     1    19376 2440.2
## - nursery      1    19384 2440.4
## - health       1    19395 2440.6
## - famsize      1    19397 2440.6
## - traveltime   1    19417 2441.0
## - address      1    19420 2441.1
## - goout        1    19465 2441.8
## - paid         1    19465 2441.8
## - freetime     1    19476 2442.1
## - schoolsup    1    19478 2442.1
## <none>              19371 2442.2
## - guardian     2    19630 2442.8
## - studytime    1    19558 2443.5
## - Pstatus      1    19575 2443.8
## - Medu         1    19623 2444.7
## - reason       3    19854 2444.8
## - romantic     1    19644 2445.1
## - sex          1    19660 2445.3
## - internet     1    19699 2446.0
## - Walc         1    19824 2448.2
## - age          1    20154 2454.0
##
## Step:  AIC=2440.25
## absences ~ sex + age + address + famsize + Pstatus + Medu + reason +
##     guardian + traveltime + studytime + schoolsup + paid + nursery +
##     internet + romantic + freetime + goout + Walc + health
##
##               Df Deviance    AIC
## - nursery      1    19389 2438.5
## - health       1    19399 2438.7
## - famsize      1    19404 2438.8
## - traveltime   1    19421 2439.1
## - address      1    19424 2439.1
## - paid         1    19466 2439.9
## - goout        1    19476 2440.0
## - schoolsup    1    19482 2440.2
## - freetime     1    19482 2440.2
## <none>              19376 2440.2
## - guardian     2    19632 2440.8
## - studytime    1    19558 2441.5
## - Pstatus      1    19579 2441.9
## - reason       3    19856 2442.8
## - romantic     1    19645 2443.1
## - Medu         1    19652 2443.2
## - sex          1    19663 2443.4
```

```
## - internet    1    19701 2444.1
## - Walc        1    19825 2446.2
## - age         1    20154 2452.0
##
## Step:  AIC=2438.49
## absences ~ sex + age + address + famsize + Pstatus + Medu + reason +
##     guardian + traveltime + studytime + schoolsup + paid + internet +
##     romantic + freetime + goout + Walc + health
##
##              Df Deviance    AIC
## - health      1    19412 2436.9
## - famsize     1    19422 2437.1
## - traveltime  1    19436 2437.3
## - address     1    19438 2437.4
## - paid        1    19474 2438.0
## - goout       1    19487 2438.2
## - freetime    1    19493 2438.3
## - schoolsup   1    19499 2438.5
## <none>              19389 2438.5
## - guardian    2    19636 2438.9
## - studytime   1    19567 2439.7
## - Pstatus     1    19598 2440.2
## - reason      3    19877 2441.2
## - romantic    1    19664 2441.4
## - sex         1    19673 2441.6
## - Medu        1    19703 2442.1
## - internet    1    19711 2442.2
## - Walc        1    19827 2444.3
## - age         1    20180 2450.4
##
## Step:  AIC=2436.91
## absences ~ sex + age + address + famsize + Pstatus + Medu + reason +
##     guardian + traveltime + studytime + schoolsup + paid + internet +
##     romantic + freetime + goout + Walc
##
##              Df Deviance    AIC
## - famsize     1    19443 2435.5
## - traveltime  1    19453 2435.6
## - address     1    19466 2435.9
## - paid        1    19497 2436.4
## - goout       1    19510 2436.7
## - freetime    1    19510 2436.7
## - schoolsup   1    19518 2436.8
## <none>              19412 2436.9
## - guardian    2    19655 2437.2
## - studytime   1    19593 2438.1
```

```
## - Pstatus     1    19616 2438.5
## - reason      3    19883 2439.3
## - sex         1    19684 2439.8
## - romantic    1    19698 2440.0
## - Medu        1    19720 2440.4
## - internet    1    19725 2440.5
## - Walc        1    19865 2442.9
## - age         1    20184 2448.5
##
## Step:  AIC=2435.46
## absences ~ sex + age + address + Pstatus + Medu + reason + guardian +
##     traveltime + studytime + schoolsup + paid + internet + romantic +
##     freetime + goout + Walc
##
##              Df Deviance    AIC
## - traveltime  1    19490 2434.3
## - address     1    19490 2434.3
## - paid        1    19525 2434.9
## - freetime    1    19538 2435.2
## - goout       1    19544 2435.3
## - schoolsup   1    19548 2435.3
## <none>              19443 2435.5
## - guardian    2    19684 2435.8
## - studytime   1    19628 2436.8
## - Pstatus     1    19675 2437.6
## - reason      3    19920 2437.9
## - sex         1    19705 2438.1
## - romantic    1    19735 2438.7
## - Medu        1    19738 2438.7
## - internet    1    19757 2439.0
## - Walc        1    19916 2441.8
## - age         1    20222 2447.2
##
## Step:  AIC=2434.3
## absences ~ sex + age + address + Pstatus + Medu + reason + guardian +
##     studytime + schoolsup + paid + internet + romantic + freetime +
##     goout + Walc
##
##             Df Deviance    AIC
## - address    1    19563 2433.6
## - paid       1    19571 2433.7
## - goout      1    19579 2433.9
## - freetime   1    19592 2434.1
## - schoolsup  1    19597 2434.2
## <none>             19490 2434.3
## - guardian   2    19732 2434.6
```

```
## - studytime  1    19690 2435.9
## - reason     3    19936 2436.2
## - Pstatus    1    19721 2436.4
## - sex        1    19743 2436.8
## - Medu       1    19765 2437.2
## - romantic   1    19786 2437.6
## - internet   1    19791 2437.6
## - Walc       1    19987 2441.1
## - age        1    20242 2445.5
##
## Step:  AIC=2433.61
## absences ~ sex + age + Pstatus + Medu + reason + guardian + studytime +
##     schoolsup + paid + internet + romantic + freetime + goout +
##     Walc
##
##             Df Deviance    AIC
## - paid       1    19650 2433.2
## - goout      1    19671 2433.5
## <none>            19563 2433.6
## - schoolsup  1    19677 2433.6
## - freetime   1    19686 2433.8
## - guardian   2    19820 2434.2
## - studytime  1    19740 2434.8
## - reason     3    19987 2435.1
## - Pstatus    1    19790 2435.6
## - sex        1    19794 2435.7
## - internet   1    19818 2436.1
## - Medu       1    19818 2436.1
## - romantic   1    19880 2437.2
## - Walc       1    20111 2441.3
## - age        1    20339 2445.2
##
## Step:  AIC=2433.15
## absences ~ sex + age + Pstatus + Medu + reason + guardian + studytime +
##     schoolsup + internet + romantic + freetime + goout + Walc
##
##             Df Deviance    AIC
## - goout      1    19747 2432.9
## <none>            19650 2433.2
## - guardian   2    19880 2433.2
## - freetime   1    19772 2433.3
## - schoolsup  1    19772 2433.3
## - reason     3    20026 2433.8
## - sex        1    19849 2434.7
## - internet   1    19872 2435.1
## - studytime  1    19874 2435.1
```

```
## - Medu       1    19880 2435.2
## - Pstatus    1    19897 2435.5
## - romantic   1    19988 2437.1
## - Walc       1    20147 2439.9
## - age        1    20448 2445.0
##
## Step:  AIC=2432.87
## absences ~ sex + age + Pstatus + Medu + reason + guardian + studytime +
##     schoolsup + internet + romantic + freetime + Walc
##
##             Df Deviance    AIC
## - guardian   2    19964 2432.7
## <none>            19747 2432.9
## - schoolsup  1    19862 2432.9
## - reason     3    20133 2433.6
## - sex        1    19922 2433.9
## - freetime   1    19942 2434.3
## - internet   1    19952 2434.5
## - Medu       1    19957 2434.6
## - studytime  1    19985 2435.1
## - Pstatus    1    19997 2435.3
## - romantic   1    20094 2436.9
## - Walc       1    20147 2437.9
## - age        1    20477 2443.5
##
## Step:  AIC=2432.7
## absences ~ sex + age + Pstatus + Medu + reason + studytime +
##     schoolsup + internet + romantic + freetime + Walc
##
##             Df Deviance    AIC
## - schoolsup  1    20074 2432.6
## <none>            19964 2432.7
## - reason     3    20360 2433.6
## - freetime   1    20154 2434.0
## - internet   1    20160 2434.1
## - sex        1    20166 2434.2
## - Medu       1    20182 2434.5
## - studytime  1    20221 2435.2
## - Pstatus    1    20266 2435.9
## - Walc       1    20315 2436.8
## - romantic   1    20348 2437.3
## - age        1    21068 2449.5
##
## Step:  AIC=2432.6
## absences ~ sex + age + Pstatus + Medu + reason + studytime +
##     internet + romantic + freetime + Walc
```

```
##
##             Df Deviance    AIC
## <none>           20074 2432.6
## - reason     3    20475 2433.5
## - freetime   1    20264 2433.9
## - internet   1    20264 2433.9
## - Medu       1    20271 2434.0
## - sex        1    20324 2434.9
## - studytime  1    20333 2435.1
## - Pstatus    1    20391 2436.1
## - Walc       1    20415 2436.5
## - romantic   1    20439 2436.9
## - age        1    21077 2447.6
```

```
summary(step)
```

```
##
## Call:
## glm(formula = absences ~ sex + age + Pstatus + Medu + reason +
##     studytime + internet + romantic + freetime + Walc, data = math)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -12.594   -4.437   -0.906    2.112   59.463
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -17.3135     6.4088  -2.702  0.00725 **
## sexM              -1.8719     0.9139  -2.048  0.04132 *
## age                1.4458     0.3528   4.098 5.24e-05 ***
## PstatusT          -3.1114     1.3511  -2.303  0.02190 *
## Medu               0.7383     0.4066   1.816  0.07031 .
## reasonhome         2.4102     1.0566   2.281  0.02317 *
## reasonother        1.3696     1.6594   0.825  0.40974
## reasonreputation   2.2437     1.0632   2.110  0.03557 *
## studytime         -1.1083     0.5325  -2.081  0.03815 *
## internetyes        2.1331     1.1966   1.783  0.07556 .
## romanticyes        2.2605     0.9143   2.472  0.01392 *
## freetime          -0.7813     0.4384  -1.782  0.07563 .
## Walc               0.8042     0.3363   2.391  0.01735 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 59.7433)
##
##     Null deviance: 24216  on 348  degrees of freedom
```

```
## Residual deviance: 20074  on 336  degrees of freedom
## AIC: 2432.6
##
## Number of Fisher Scoring iterations: 2

shapiro.test(resid(step))  # To test normality assumption

##
##  Shapiro-Wilk normality test
##
## data:  resid(step)
## W = 0.77798, p-value < 2.2e-16

bptest(step)  # To test for heteroskedasiticty

##
##  studentized Breusch-Pagan test
##
## data:  step
## BP = 23.891, df = 12, p-value = 0.02104

cv_step <- cv.glm(math, step)$delta[2]

# Shrinkage Methods  ============================================================
# Create a model matrix
modmat <- model.matrix(~ sex + age + address + famsize + Pstatus +
                         Medu + Fedu +  Mjob + Fjob + reason + guardian +
                         traveltime + studytime + failures + schoolsup +
                         famsup + paid + activities + nursery + higher +
                         internet + romantic + famrel + freetime + goout +
                         Dalc + Walc + health)

# Elastic Net
elastic <- cv.glmnet(x = modmat, y = math[, 29])
cv_elas <- mean(elastic$cvm)

# Partial Least Squares
pls <- plsr(absences ~ ., data = math, scale = TRUE, validation = "CV")
cv_pls <- mean(pls$validation$pred ^ 2)

# Tree methods ==================================================================

# Random Forest
randomFor <-randomForest(absences ~ ., data = math)
cross_valuation <- rfcv(trainx = modmat, trainy = math[, 29])
cv_rf <- mean(cross_valuation$error.cv)
```

```r
# Boosting
boosting <- gbm(absences ~ ., data = math, cv.folds = 10)

## Distribution not specified, assuming gaussian ...

cv_boosting <- mean(boosting$cv.error)

# SVM =========================================================================
(svmtune <- tune(svm, absences ~ ., data = math))

##
## Error estimation of 'svm' using 10-fold cross validation: 65.42312

svmtune$best.model

##
## Call:
## best.tune(method = svm, train.x = absences ~ ., data = math)
##
##
## Parameters:
##    SVM-Type:  eps-regression
##  SVM-Kernel:  radial
##        cost:  1
##       gamma:  0.02631579
##     epsilon:  0.1
##
##
## Number of Support Vectors:  300

cv_svm <- svmtune$best.performance

# Neural Network ==============================================================
(nntune <- tune(nnet, absences ~ ., data = math, size = 1))

## # weights:  40
## initial  value 33091.445685
## final  value 31353.000000
## converged
## # weights:  40
## initial  value 32576.864575
## final  value 31263.000000
## converged
## # weights:  40
## initial  value 34205.423913
## final  value 31639.000000
## converged
```

```
## # weights:  40
## initial  value 31234.587204
## final  value 29156.000000
## converged
## # weights:  40
## initial  value 32927.636585
## final  value 31246.000000
## converged
## # weights:  40
## initial  value 27487.440788
## final  value 25431.000000
## converged
## # weights:  40
## initial  value 32448.453865
## final  value 30298.000000
## converged
## # weights:  40
## initial  value 32386.761803
## final  value 31147.000000
## converged
## # weights:  40
## initial  value 30721.431135
## final  value 28380.000000
## converged
## # weights:  40
## initial  value 27047.110520
## final  value 25476.000000
## converged
## # weights:  40
## initial  value 34382.330453
## final  value 32821.000000
## converged
##
## Error estimation of 'nnet' using 10-fold cross validation: 94.39529

(nntune5 <- tune(nnet, absences ~ ., data = math, size = 5))

## # weights:  196
## initial  value 30803.915734
## final  value 28633.000000
## converged
## # weights:  196
## initial  value 31378.540377
## final  value 30727.000000
## converged
## # weights:  196
```

```
## initial  value 21644.942283
## final  value 19891.000000
## converged
## # weights:  196
## initial  value 33943.994599
## final  value 31463.000000
## converged
## # weights:  196
## initial  value 31246.932347
## final  value 30173.000000
## converged
## # weights:  196
## initial  value 33897.250822
## final  value 31697.000000
## converged
## # weights:  196
## initial  value 33295.906766
## final  value 32059.000000
## converged
## # weights:  196
## initial  value 31696.931604
## final  value 29973.000000
## converged
## # weights:  196
## initial  value 33442.668095
## final  value 31739.000000
## converged
## # weights:  196
## initial  value 30474.979365
## final  value 29034.000000
## converged
## # weights:  196
## initial  value 33653.464491
## final  value 32821.000000
## converged
##
## Error estimation of 'nnet' using 10-fold cross validation: 93.86874

(nntune10 <- tune(nnet, absences ~ ., data = math, size = 10))

## # weights:  391
## initial  value 34019.382610
## final  value 31936.000000
## converged
## # weights:  391
## initial  value 33249.505343
```

```
## final  value 30281.000000
## converged
## # weights:  391
## initial  value 32430.396837
## final  value 30996.000000
## converged
## # weights:  391
## initial  value 26683.893351
## final  value 24930.000000
## converged
## # weights:  391
## initial  value 31828.992871
## final  value 30051.000000
## converged
## # weights:  391
## initial  value 31984.961033
## final  value 30194.000000
## converged
## # weights:  391
## initial  value 29131.562032
## final  value 27447.000000
## converged
## # weights:  391
## initial  value 27956.974420
## final  value 26668.000000
## converged
## # weights:  391
## initial  value 33003.258458
## final  value 31589.000000
## converged
## # weights:  391
## initial  value 32319.294595
## final  value 31297.000000
## converged
## # weights:  391
## initial  value 34452.595857
## final  value 32821.000000
## converged
##
## Error estimation of 'nnet' using 10-fold cross validation: 93.99504

(nntune20 <- tune(nnet, absences ~ ., data = math, size = 20))

## # weights:  781
## initial  value 31642.881636
## final  value 30444.000000
```

```
## converged
## # weights:  781
## initial  value 33675.880302
## final  value 31548.000000
## converged
## # weights:  781
## initial  value 32942.087593
## final  value 31878.000000
## converged
## # weights:  781
## initial  value 31494.053095
## final  value 31274.000000
## converged
## # weights:  781
## initial  value 24636.396795
## final  value 24145.000000
## converged
## # weights:  781
## initial  value 31759.513684
## final  value 29799.000000
## converged
## # weights:  781
## initial  value 26827.941418
## final  value 25034.000000
## converged
## # weights:  781
## initial  value 29705.385541
## final  value 29207.000000
## converged
## # weights:  781
## initial  value 34501.384266
## final  value 31179.000000
## converged
## # weights:  781
## initial  value 33141.001314
## final  value 30881.000000
## converged
## # weights:  781
## initial  value 35397.134440
## final  value 32821.000000
## converged
##
## Error estimation of 'nnet' using 10-fold cross validation: 94.02824

cv_nn1 <- nntune$best.performance
cv_nn5 <- nntune5$best.performance
```

```r
cv_nn10 <- nntune10$best.performance
cv_nn20 <- nntune20$best.performance

# Cross Validation Table ========================================================
CV_table <- list("Guess" = cv_guess,
                 "Linear" = cv_lin,
                 "Adjusted-Step" = cv_step,
                 "Elastic Net" = cv_elas,
                 "Partial Least Squares" = cv_pls,
                 "Random Forest" = cv_rf,
                 "Boosting" = cv_boosting,
                 "Support Vector Machine" = cv_svm,
                 "Neural Net 1" = cv_nn1,
                 "Neural Net 5" = cv_nn5,
                 "Neural Net 10" = cv_nn10,
                 "Neural Net 20" = cv_nn20)

# Model Assessment ==============================================================
average <- mean(math$absences)
lapply(CV_table, function(x) sqrt(x) / average * 100)  # Percentage of error

## $Guess
## [1] 137.9752
##
## $Linear
## [1] 140.5685
##
## $`Adjusted-Step`
## [1] 132.1765
##
## $`Elastic Net`
## [1] 139.5125
##
## $`Partial Least Squares`
## [1] 120.8201
##
## $`Random Forest`
## [1] 141.2936
##
## $Boosting
## [1] 139.8646
##
## $`Support Vector Machine`
## [1] 135.5846
##
## $`Neural Net 1`
```

```
## [1] 162.862
##
## $`Neural Net 5`
## [1] 162.4072
##
## $`Neural Net 10`
## [1] 162.5164
##
## $`Neural Net 20`
## [1] 162.5451
```

References

[1] A. Liaw and M. Wiener (2002). Classification and Regression by randomForest. R News 2(3), 18–22.

[2] Achim Zeileis, Torsten Hothorn (2002). Diagnostic Checking in Regression Relationships. R News 2(3), 7-10. URL http://CRAN.R-project.org/doc/Rnews/

[3] Angelo Canty and Brian Ripley (2016). boot: Bootstrap R (S-Plus) Functions. R package version 1.3-18.

Davison, A. C. & Hinkley, D. V. (1997) Bootstrap Methods and Their Applications. Cambridge University Press, Cambridge. ISBN 0-521-57391-2

[4] Bjrn-Helge Mevik, Ron Wehrens and Kristian Hovde Liland (2015). pls: Partial Least Squares and Principal Component Regression. R package version 2.5-0. https://CRAN.R-project.org/package=pls

[5] David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel and Friedrich Leisch (2015). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.6-7. https://CRAN.R-project.org/package=e1071

[6] Gareth Hames, Daniela Witten, Trevor Hastie, Robert Tibshirani (2013). An Introduction to Statisical Learning. Volume 103 2013. ISBN: 978-1-4614-7137-0 (Print) 978-1-4614-7138-7 (Online).

[7] Greg Ridgeway with contributions from others (2015). gbm: Generalized Boosted Regression Models. R package version 2.1.1. https://CRAN.R-project.org/package=gbm

[8] Jerome Friedman, Trevor Hastie, Robert Tibshirani (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software, 33(1), 1-22. URL http://www.jstatsoft.org/v33/i01/.

[9] Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[10] P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7

[11] R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[12] Robertas Gabrys. DSO 530: Applied Modern Statistical Learning Methods. Final Project Instructions. Retrieved on 30 April 2016.

[13] Taiyun Wei (2013). corrplot: Visualization of a correlation matrix. R package version 0.73. https://CRAN.R-project.org/package=corrplot

[14] Venables, W. N. & Ripley, B. D. (2002). MASS package. Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

[15] Venables, W. N. & Ripley, B. D. (2002). nnet package. Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

[16] Yihui Xie (2016). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.12.3.

Yihui Xie (2015) Dynamic Documents with R and knitr. 2nd edition. Chapman and Hall/CRC. ISBN 978-1498716963

Yihui Xie (2014) knitr: A Comprehensive Tool for Reproducible Research in R. In Victoria Stodden, Friedrich Leisch and Roger D. Peng, editors, Implementing Reproducible Computational Research. Chapman and Hall/CRC. ISBN 978-1466561595