# Math Scores in Secondary Schools

Meredith Lou

# Data Overview

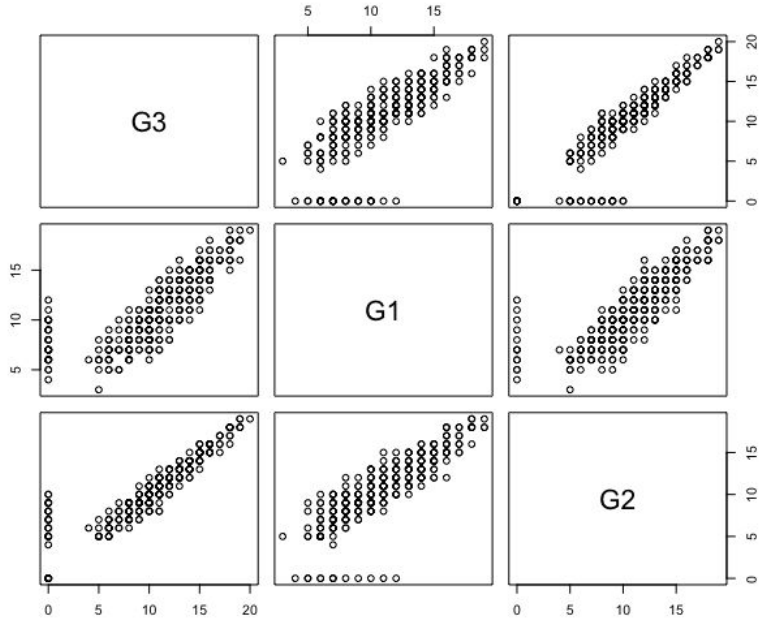
- Data taken from UCI Machine Learning Repository
- Data looks at student math performance from two Portuguese secondary education schools
  - Collected using school reports and questionnaires
- Response variable is “G3,” measures final grade as integer values from 0-20
- Data contain 395 rows with 32 predictors, original dataset contains no missing values

1 school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
2 sex - student's sex (binary: 'F' - female or 'M' - male)
3 age - student's age (numeric: from 15 to 22)
4 address - student's home address type (binary: 'U' - urban or 'R' - rural)
5 famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
6 Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
7 Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary education or 4 â€“ higher education)
8 Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary education or 4 â€“ higher education)
9 Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
10 Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
11 reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
12 guardian - student's guardian (nominal: 'mother', 'father' or 'other')
13 traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
14 studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
15 failures - number of past class failures (numeric: n if 1<=n<3, else 4)
16 schoolsup - extra educational support (binary: yes or no)
17 famsup - family educational support (binary: yes or no)
18 paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
19 activities - extra-curricular activities (binary: yes or no)
20 nursery - attended nursery school (binary: yes or no)
21 higher - wants to take higher education (binary: yes or no)
22 internet - Internet access at home (binary: yes or no)
23 romantic - with a romantic relationship (binary: yes or no)
24 famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
25 freetime - free time after school (numeric: from 1 - very low to 5 - very high)
26 goout - going out with friends (numeric: from 1 - very low to 5 - very high)
27 Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
28 Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
29 health - current health status (numeric: from 1 - very bad to 5 - very good)
30 absences - number of school absences (numeric: from 0 to 93)

# these grades are related with the course subject, Math or Portuguese:
31 G1 - first period grade (numeric: from 0 to 20)
31 G2 - second period grade (numeric: from 0 to 20)
32 G3 - final grade (numeric: from 0 to 20, output target)

# Data Cleaning

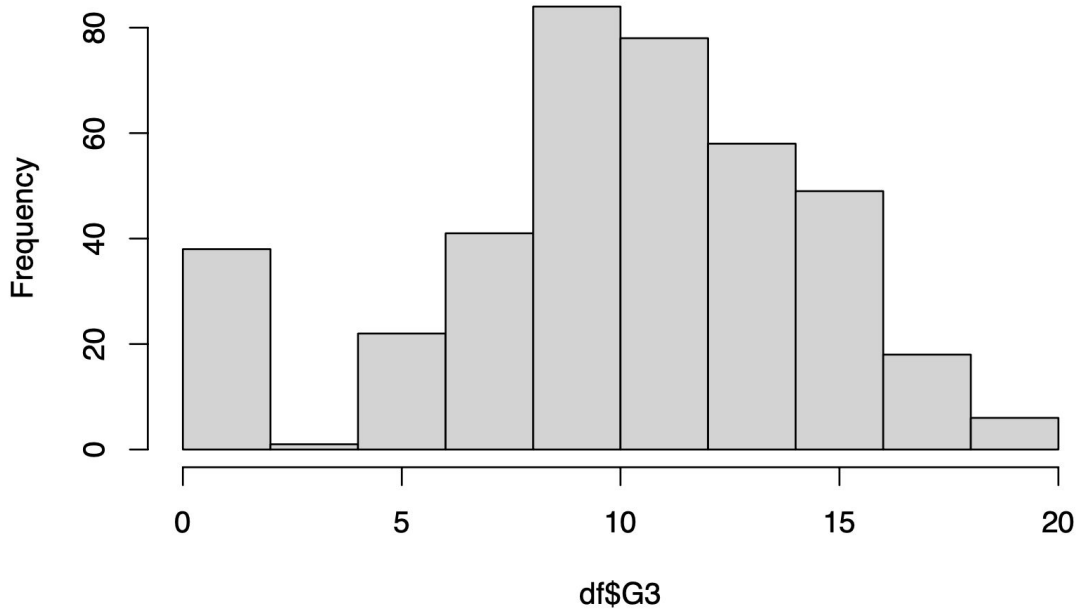
- Removed G1 and G2 (first and second period grades, respectively) as predictors due to their high correlation with final grade




| | G3 | G1 | G2 |
|---|---|---|---|
| G3 | 1.0000000 | 0.8014679 | 0.9048680 |
| G1 | 0.8014679 | 1.0000000 | 0.8521181 |
| G2 | 0.9048680 | 0.8521181 | 1.0000000 |

# Data Cleaning

- Removed all rows where G3 = 0 to fit model assumptions in linear regression



# Overview of Machine Learning Algorithms

- For each respective model, I chose to be consistent with a 5-fold cross-validated model using set.seed(1) for repeatability
- Models used
  - Penalized linear regression
  - k-nearest neighbor
  - Decision tree
  - Random forest

# Penalized Linear Regression

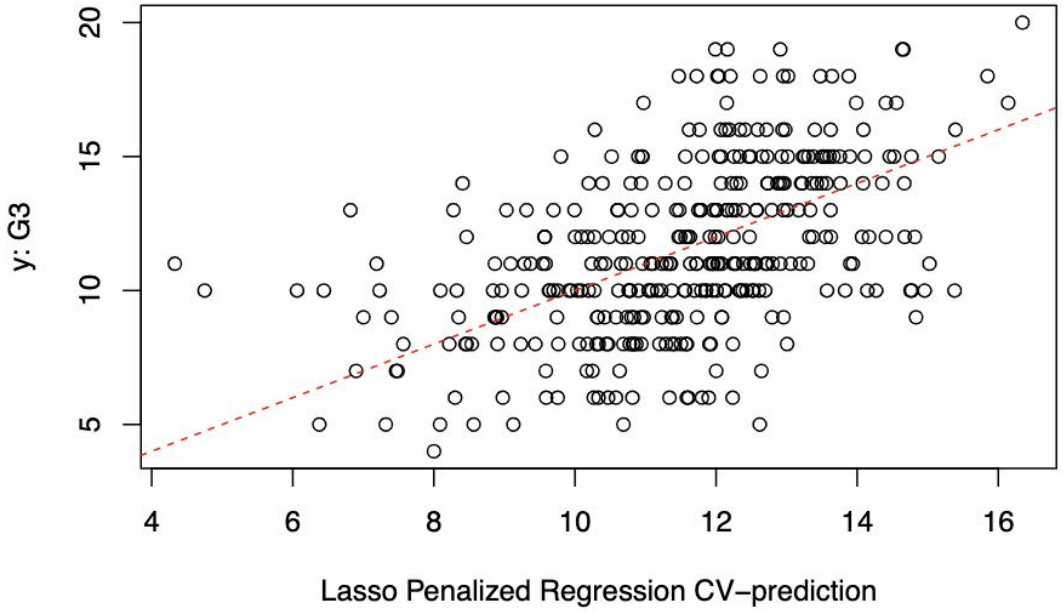
- The lasso-penalized variables were selected using lambda.min
- The penalized linear regression model showed studytime, failures, schoolsup, goout, health, and absences as statistically significant coefficients for predicting final grades
- Passes model assumptions and global test

```
"schoolMS"     "sexM"         "age"          "addressU"     "famsizeLE3"
"Medu"         "Fedu"         "Mjobhealth"   "Mjobother"    "Mjobservices"
"Fjobteacher"  "studytime"    "failures"     "schoolsupyes" "famsupyes"
"paidyes"      "internetyes"  "goout"        "Walc"         "health"
"absences"
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  15.19944    2.61269   5.818 1.41e-08 ***
schoolMS     -0.60552    0.53110  -1.140 0.255060
sexM          0.64148    0.33194   1.932 0.054151 .
age          -0.12963    0.14083  -0.920 0.358009
addressU      0.32648    0.38545   0.847 0.397608
famsizeLE3    0.28593    0.32855   0.870 0.384771
Medu          0.16076    0.21685   0.741 0.459011
Fedu          0.14496    0.18756   0.773 0.440149
Mjobhealth    1.21985    0.76648   1.591 0.112455
Mjobother    -0.38813    0.50154  -0.774 0.439561
Mjobservices  0.91767    0.56585   1.622 0.105811
Mjobteacher  -0.58429    0.71795  -0.814 0.416328
Fjobhealth   -0.73965    0.98033  -0.754 0.451096
Fjobother    -0.67090    0.71735  -0.935 0.350340
Fjobservices -0.75592    0.74498  -1.015 0.310997
Fjobteacher   1.08971    0.91019   1.197 0.232072
studytime     0.50166    0.19864   2.526 0.012018 *
failures     -0.87584    0.24562  -3.566 0.000416 ***
schoolsupyes -2.37121    0.45511  -5.210 3.32e-07 ***
famsupyes    -0.64747    0.33049  -1.959 0.050938 .
paidyes      -0.42824    0.32054  -1.336 0.182462
internetyes   0.60131    0.43177   1.393 0.164654
goout        -0.42895    0.15468  -2.773 0.005866 **
Walc         -0.14890    0.14138  -1.053 0.293015
health       -0.22631    0.10912  -2.074 0.038850 *
absences     -0.05928    0.01962  -3.022 0.002707 **
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 2.744 on 331 degrees of freedom
Multiple R-squared:  0.3283, Adjusted R-squared:  0.2775
F-statistic:  6.47 on 25 and 331 DF,  p-value: < 2.2e-16
```

# Penalized Linear Regression (CV)

- The CV predicted performance saw observed vs. predicted values to have a Pearson correlation coefficient of 0.4977297
- The CV penalized linear regression model has a RMSE of 2.809316



# Penalized Linear Regression (CV)

- The 10 most important variables using the varImp() function are shown
- Schoolsupyes is the top most important variable, which indicates students who received extra educational support
- Top 6 important variables same as the 6 significant coefficients in regression output

| | Overall |
|---|---|
| schoolsupyes | 100.000 |
| failures | 63.203 |
| absences | 51.035 |
| goout | 45.466 |
| studytime | 39.925 |
| health | 29.821 |
| famsupyes | 27.250 |
| sexM | 26.654 |
| Mjobservices | 19.701 |
| Mjobhealth | 19.024 |

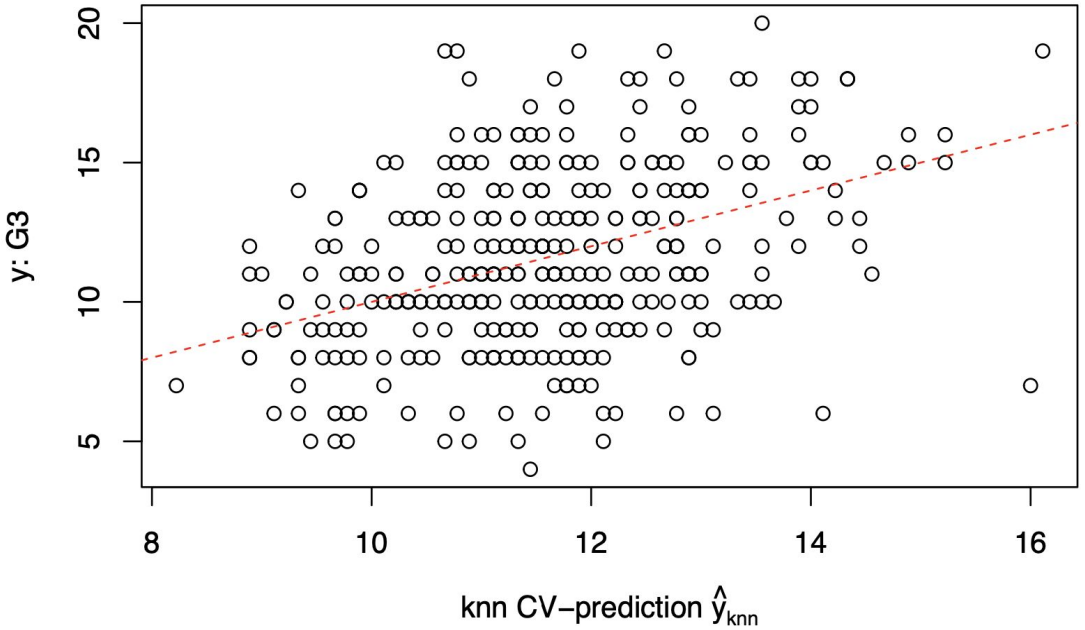
# k-Nearest Neighbors

- I used the PreProcess() function to center and normalize the data before using train() function to run the model.
- I trained data using 5 fold cross validation
- Lowest RMSE was achieved at k = 9

```
k-Nearest Neighbors

357 samples
 30 predictor

Pre-processing: centered (39), scaled (39)
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 285, 286, 284, 286, 287
Resampling results across tuning parameters:

  k   RMSE      Rsquared    MAE
   1  3.891856  0.08312940  3.096005
   2  3.394838  0.09408289  2.711922
   3  3.219741  0.10435458  2.561031
   4  3.104950  0.11570968  2.508100
   5  3.080317  0.12138164  2.465107
   6  2.973049  0.16408221  2.372265
   7  2.985415  0.15582263  2.384613
   8  2.972727  0.15680386  2.375232
   9  2.965655  0.15874539  2.377958
  10  2.977210  0.15005004  2.409359
  11  2.999241  0.13826554  2.433580
  12  2.989427  0.14273722  2.438748
  13  3.008080  0.13268017  2.452846
  14  3.015575  0.12914812  2.459827
  15  2.995364  0.14299073  2.449252
  16  2.999452  0.14182627  2.452620
  17  3.010196  0.13592402  2.465615
  18  3.018780  0.13137892  2.475261
  19  3.018649  0.13266998  2.482379
  20  3.011002  0.13857752  2.480723
  21  3.010958  0.13997861  2.473962
  22  3.011320  0.14261028  2.483350
  23  3.025980  0.13352844  2.492934
  24  3.034070  0.12891936  2.500922
  25  3.036237  0.12923077  2.505449
  26  3.024078  0.14141448  2.499806
  27  3.019808  0.14657399  2.498936
  28  3.017655  0.14814866  2.495506
  29  3.021116  0.14924982  2.497827
  30  3.026519  0.14500094  2.499342

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was k = 9.
```

# k-Nearest Neighbor

- The CV predicted performance saw observed vs. predicted values to have a Pearson correlation coefficient of 0.3891399
- The CV kNN model has a optimal RMSE of 2.965655



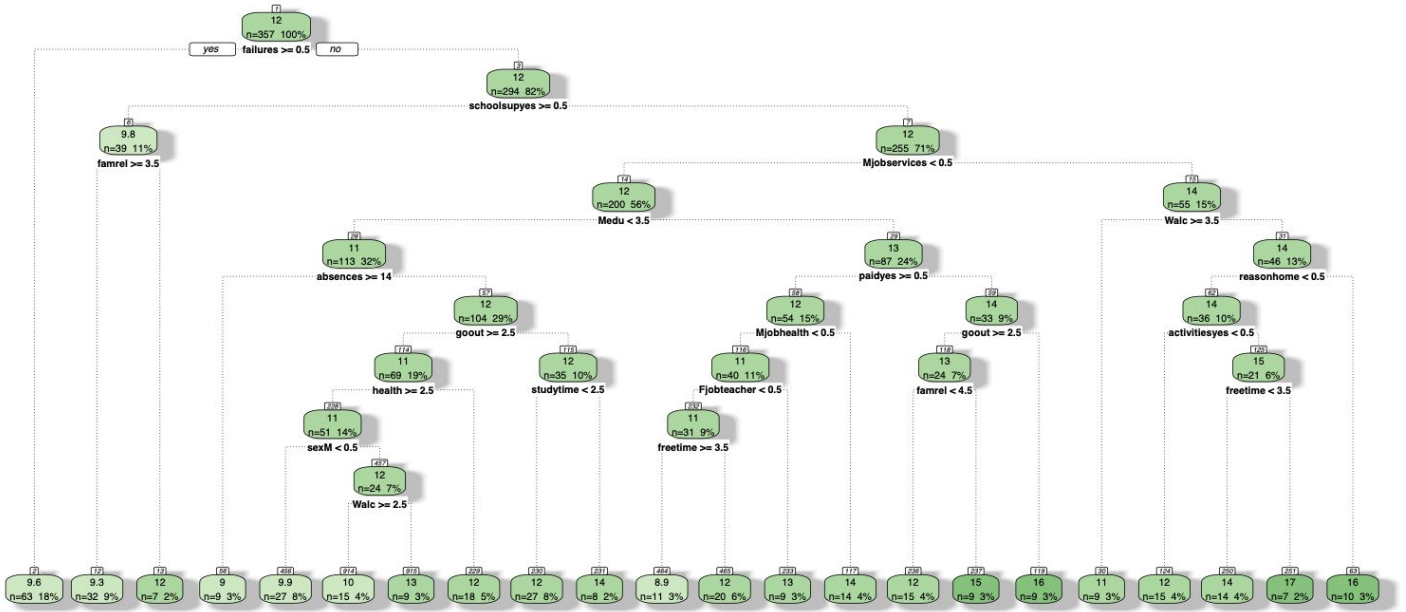
# k-Nearest Neighbor

- The 10 most important variables using the varImp() function are shown
- Absences were a major predictor along with failures, schoolsup, mother's education, mother's job, and weekend alcohol consumption

| | Overall |
|---|---|
| absences | 100.000 |
| failures | 63.288 |
| schoolsup | 40.223 |
| Medu | 34.151 |
| Mjob | 32.522 |
| Walc | 29.548 |
| Fedu | 26.266 |
| goout | 25.713 |
| studytime | 25.211 |
| age | 25.165 |

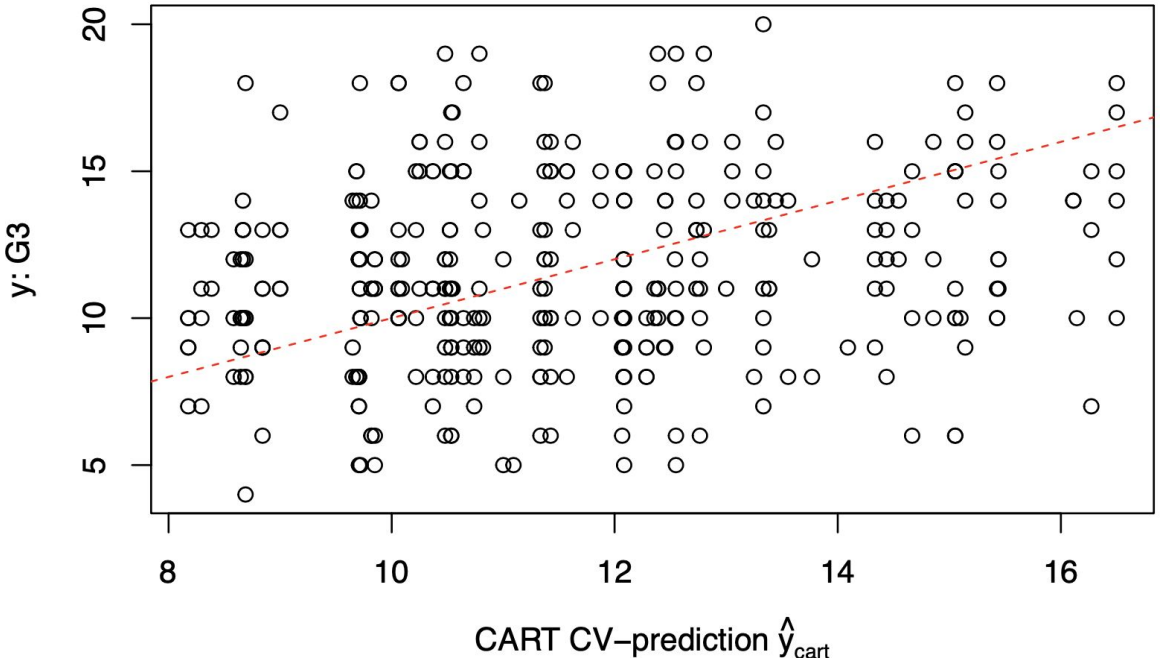
# Decision Tree (Regression Tree)

- My decision tree utilizes the rpart1SE method in R, which does not require tuning parameters and computes the complexity parameter internally using the one-standard error rule



# Decision Tree (Regression Tree)

- The observed vs. predicted values have a Pearson correlation coefficient of 0.204
- The decision tree model has a RMSE of 3.467



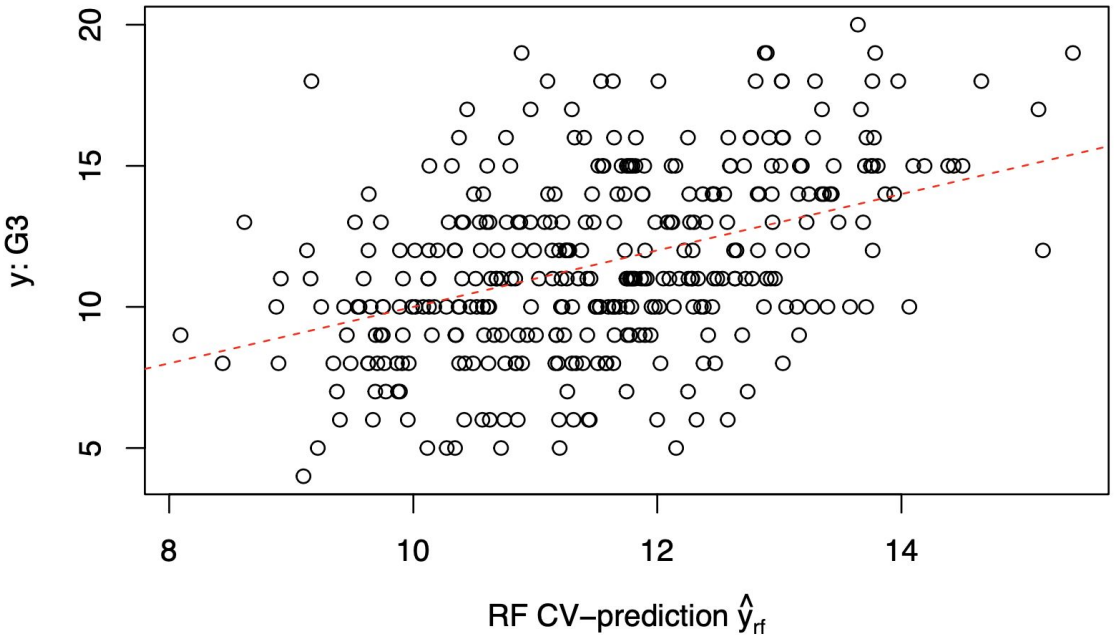
# Decision Tree (Regression Tree)

- The 10 most important variables using the varImp() function are shown
- Compared to kNN and linear regression models, absences is the only variable to appear in the top 10 for each in terms of variable importance

| | Overall |
|---|---|
| absences | 100.00 |
| freetime | 73.05 |
| Walc | 52.34 |
| famrel | 44.69 |
| Medu | 31.58 |
| studytime | 27.58 |
| Fedu | 26.65 |
| age | 24.85 |
| goout | 14.99 |
| romanticyes | 12.95 |

# Random Forest

- Using a 5-fold CV random forest supervised machine learning model
- Final value for mtry was 20
- The observed vs. predicted values have a Pearson correlation coefficient of 0.46
- The random forest model has a RMSE of 2.862

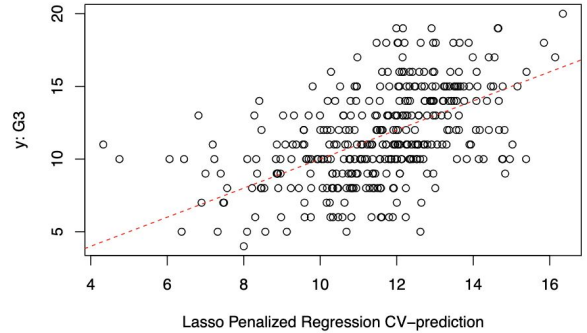


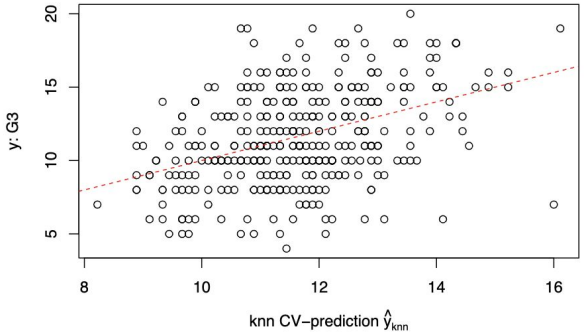
# Random Forest

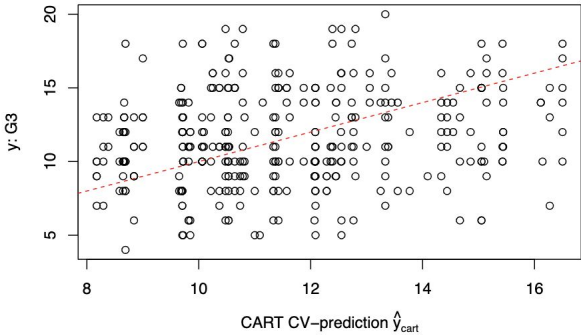
- The 10 more important variables using the varImp() function are shown
- Compared to decision tree, they have an identical most important variable (absences), share several other similarities such as Medu and studytime

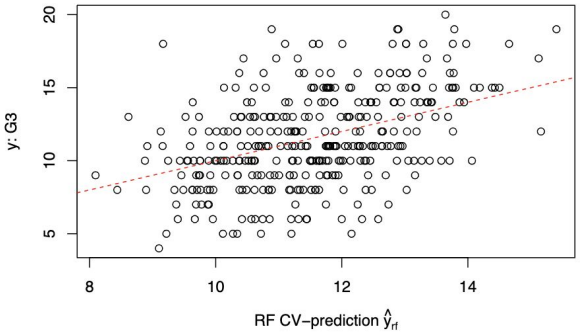
| | Overall |
|---|---|
| absences | 100.00 |
| failures | 63.59 |
| schoolsupyes | 47.79 |
| Medu | 46.78 |
| Walc | 46.09 |
| health | 45.94 |
| studytime | 45.42 |
| goout | 43.67 |
| freetime | 42.74 |
| age | 39.30 |

# Model Comparison




| Method | RMSE | Correlation |
|---|---|---|
| Penalized linear regression | 2.809316 | 0.4977297 |
| k-nearest neighbor | 2.965655 | 0.3891399 |
| Decision tree | 3.46698 | 0.2037499 |
| Random forest | 2.861684 | 0.4600431 |

# Model Comparison

| Algorithm Name(s) | Shared variable(s) in top 10 variable importance using varImp() function |
|---|---|
| LM, RF | schoolsupyes, health |
| kNN, CART | Fedu |
| CART, RF | freetime |
| LM, kNN, RF | failures |
| kNN, CART, RF | Medu, Walc, age |
| LM, kNN, CART, RF | absences, goout, studytime |

- Absences is the top variable for all but linear regression (3rd most important)
- Failures is the second most important variable for all but decision tree algorithm (not present in top 20 using varImp() function)
- Goout and studytime are present in top 10 for all models, but not top 3 for any

# Conclusion

- Among all models, penalized linear regression performed best in terms of RMSE and correlation of predicted and observed
  - Still a moderately weak correlation coefficient
- Predictors of final math grades in this dataset were thus weekly study time, number of failures, extra educational support, time spent with friends, health status of the student, and number of absences
  - Increase weekly study time, decrease prior class failures, limit time spent with friends, and decrease number of absences
  - Better health and extra educational support decrease final math grades?