Project 1 Part 1: Regression

4375 Machine Learning with Dr. Mazidi

Ali Hilal

March 31, 2022

For the first part of this project, I am analyzing data in a data set using regression algorithms. This is the link where I got the data from, titled: Stellar Classification Data set - SDSS17

https://www.kaggle.com/datasets/fedesoriano/stellar-classification-dataset-sdss17

In the original data set, there are 18 columns and 100,000 rows.

##Reading in data:

```
sdf <- read.csv("star_classification.csv")
str(sdf)
```

```
## 'data.frame':    100000 obs. of  18 variables:
##  $ obj_ID     : num  1.24e+18 1.24e+18 1.24e+18 1.24e+18 1.24e+18 ...
##  $ alpha      : num  136 145 142 339 345 ...
##  $ delta      : num  32.495 31.274 35.582 -0.403 21.184 ...
##  $ u          : num  23.9 24.8 25.3 22.1 19.4 ...
##  $ g          : num  22.3 22.8 22.7 23.8 17.6 ...
##  $ r          : num  20.4 22.6 20.6 21.6 16.5 ...
##  $ i          : num  19.2 21.2 19.3 20.5 16 ...
##  $ z          : num  18.8 21.6 18.9 19.3 15.5 ...
##  $ run_ID     : int  3606 4518 3606 4192 8102 8102 7773 7773 3716 5934 ...
##  $ rerun_ID   : int  301 301 301 301 301 301 301 301 301 301 ...
##  $ cam_col    : int  2 5 2 3 3 3 2 2 5 4 ...
##  $ field_ID   : int  79 119 120 214 137 110 462 346 108 122 ...
##  $ spec_obj_ID: num  6.54e+18 1.18e+19 5.15e+18 1.03e+19 6.89e+18 ...
##  $ class      : chr  "GALAXY" "GALAXY" "GALAXY" "GALAXY" ...
##  $ redshift   : num  0.635 0.779 0.644 0.932 0.116 ...
##  $ plate      : int  5812 10445 4576 9149 6121 5026 11069 6183 6625 2444 ...
##  $ MJD        : int  56354 58158 55592 58039 56187 55855 58456 56210 56386 54082 ...
##  $ fiber_ID   : int  171 427 299 775 842 741 113 15 719 232 ...
```

##Data cleaning:

I begin cleaning this data by removing any irrelevant information from the data set. This primarily consists of the numerous ID numbers as well as a column for the date of the discovery of the star. Secondly, I rename the remaining columns for easier understanding of the data. I also changed the Class coolumn to a factor for easier data analysis. Finally, I removed all instances with incomplete data.

```
sdf <- subset(sdf, select = -c(obj_ID, run_ID, rerun_ID, field_ID, spec_obj_ID, MJD, fiber_ID))
names(sdf) <- c("Alpha", "Delta", "Ultraviolet", "Green", "Red", "Near-Infrared", "Infrared", "Camera",
sdf$Class <- as.factor(sdf$Class)
sdf <- sdf[complete.cases(sdf), ]
```

##Data Exploration:

A small introduction to the data set we are working with. Here there is the first and last 6 elements in the dataset are displayed, a summary of the data, and every column with the name and the data type.

```
head(sdf)
```

```
##      Alpha      Delta Ultraviolet    Green      Red Near-Infrared Infrared
## 1 135.6891 32.4946318    23.87882 22.27530 20.39501      19.16573 18.79371
## 2 144.8261 31.2741849    24.77759 22.83188 22.58444      21.16812 21.61427
## 3 142.1888 35.5824442    25.26307 22.66389 20.60976      19.34857 18.94827
## 4 338.7410 -0.4028276    22.13682 23.77656 21.61162      20.50454 19.25010
## 5 345.2826 21.1838656    19.43718 17.58028 16.49747      15.97711 15.54461
## 6 340.9951 20.5894763    23.48827 23.33776 21.32195      20.25615 19.54544
##   Camera  Class  Redshift Plate
## 1      2 GALAXY 0.6347936  5812
## 2      5 GALAXY 0.7791360 10445
## 3      2 GALAXY 0.6441945  4576
## 4      3 GALAXY 0.9323456  9149
## 5      3 GALAXY 0.1161227  6121
## 6      3    QSO 1.4246590  5026
```

```
tail(sdf)
```

```
##            Alpha      Delta Ultraviolet    Green      Red Near-Infrared
## 99995  317.24700 -0.6822542    20.96526 19.81625 19.34186      19.14711
## 99996   39.62071 -2.5940737    22.16759 22.97586 21.90404      21.30548
## 99997   29.49382 19.7988744    22.69118 22.38628 20.45003      19.75759
## 99998  224.58741 15.7007074    21.16916 19.26997 18.20428      17.69034
## 99999  212.26862 46.6603653    25.35039 21.63757 19.91386      19.07254
## 100000 196.89605 49.4646428    22.62171 21.79745 20.60115      20.00959
##        Infrared Camera  Class  Redshift Plate
## 99995  19.05790      2 GALAXY 0.1752061  1025
## 99996  20.73569      2 GALAXY 0.0000000  9374
## 99997  19.41526      1 GALAXY 0.4048950  7626
## 99998  17.35221      4 GALAXY 0.1433656  2764
## 99999  18.62482      4 GALAXY 0.4550396  6751
## 100000 19.28075      4 GALAXY 0.5429442  7410
```

```
summary(sdf)
```

```
##      Alpha              Delta          Ultraviolet          Green
##  Min.   :  0.0055   Min.   :-18.785   Min.   :-9999.00   Min.   :-9999.00
##  1st Qu.:127.5182   1st Qu.:  5.147   1st Qu.:   20.35   1st Qu.:   18.96
##  Median :180.9007   Median : 23.646   Median :   22.18   Median :   21.10
##  Mean   :177.6291   Mean   : 24.135   Mean   :   21.98   Mean   :   20.53
##  3rd Qu.:233.8950   3rd Qu.: 39.902   3rd Qu.:   23.69   3rd Qu.:   22.12
##  Max.   :359.9998   Max.   : 83.001   Max.   :   32.78   Max.   :   31.60
##       Red         Near-Infrared      Infrared            Camera
##  Min.   : 9.822   Min.   : 9.47   Min.   :-9999.00   Min.   :1.000
##  1st Qu.:18.136   1st Qu.:17.73   1st Qu.:   17.46   1st Qu.:2.000
##  Median :20.125   Median :19.41   Median :   19.00   Median :4.000
##  Mean   :19.646   Mean   :19.08   Mean   :   18.67   Mean   :3.512
```

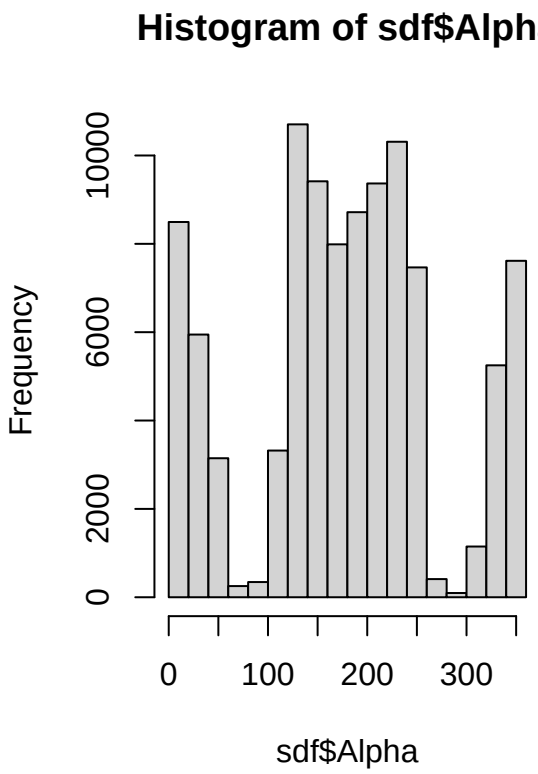
```
##  3rd Qu.:21.045   3rd Qu.:20.40   3rd Qu.:   19.92   3rd Qu.:5.000
##  Max.   :29.572   Max.   :32.14   Max.   :   29.38   Max.   :6.000
##     Class          Redshift             Plate
##  GALAXY:59445   Min.   :-0.009971   Min.   :  266
##  QSO   :18961   1st Qu.: 0.054517   1st Qu.: 2526
##  STAR  :21594   Median : 0.424173   Median : 4987
##                 Mean   : 0.576661   Mean   : 5137
##                 3rd Qu.: 0.704154   3rd Qu.: 7400
##                 Max.   : 7.011245   Max.   :12547
```

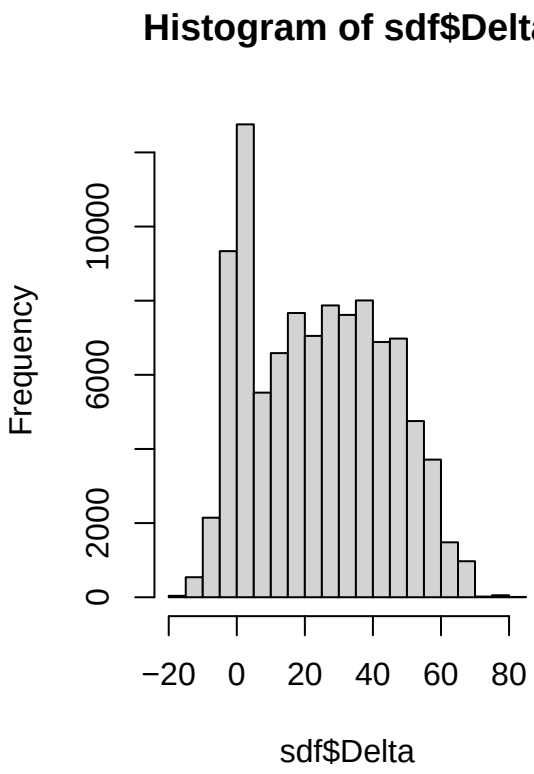
```
str(sdf)
```

```
## 'data.frame':    100000 obs. of  11 variables:
##  $ Alpha        : num  136 145 142 339 345 ...
##  $ Delta        : num  32.495 31.274 35.582 -0.403 21.184 ...
##  $ Ultraviolet  : num  23.9 24.8 25.3 22.1 19.4 ...
##  $ Green        : num  22.3 22.8 22.7 23.8 17.6 ...
##  $ Red          : num  20.4 22.6 20.6 21.6 16.5 ...
##  $ Near-Infrared: num  19.2 21.2 19.3 20.5 16 ...
##  $ Infrared     : num  18.8 21.6 18.9 19.3 15.5 ...
##  $ Camera       : int  2 5 2 3 3 3 2 2 5 4 ...
##  $ Class        : Factor w/ 3 levels "GALAXY","QSO",..: 1 1 1 1 1 2 2 1 1 3 ...
##  $ Redshift     : num  0.635 0.779 0.644 0.932 0.116 ...
##  $ Plate        : int  5812 10445 4576 9149 6121 5026 11069 6183 6625 2444 ...
```

These are two histograms comparing the Alpha and the Delta values to map and relationship between the two.

```
par(mfrow=c(1,2))
hist(sdf$Alpha)
hist(sdf$Delta)
```





##Linear Regression:

First, I created a Linear Regression Model based on all the predictors.

```
split1 <- sample(1:nrow(sdf), nrow(sdf)*0.75, replace=FALSE)
train <- sdf[split1, ]
test  <- sdf[-split1, ]
slm <- lm(Alpha~., data = train)
summary(slm)
```

```
##
## Call:
## lm(formula = Alpha ~ ., data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -197.83  -56.97   -2.77   51.55  214.26
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.787e+02  5.107e+00  35.001  < 2e-16 ***
## Delta            6.506e-01  1.800e-02  36.143  < 2e-16 ***
## Ultraviolet      2.342e+00  3.268e-01   7.166 7.82e-13 ***
## Green            1.813e-02  5.984e-01   0.030 0.975831
## Red             -2.636e+00  1.108e+00  -2.380 0.017315 *
## `Near-Infrared`  1.265e+00  1.147e+00   1.103 0.270127
```

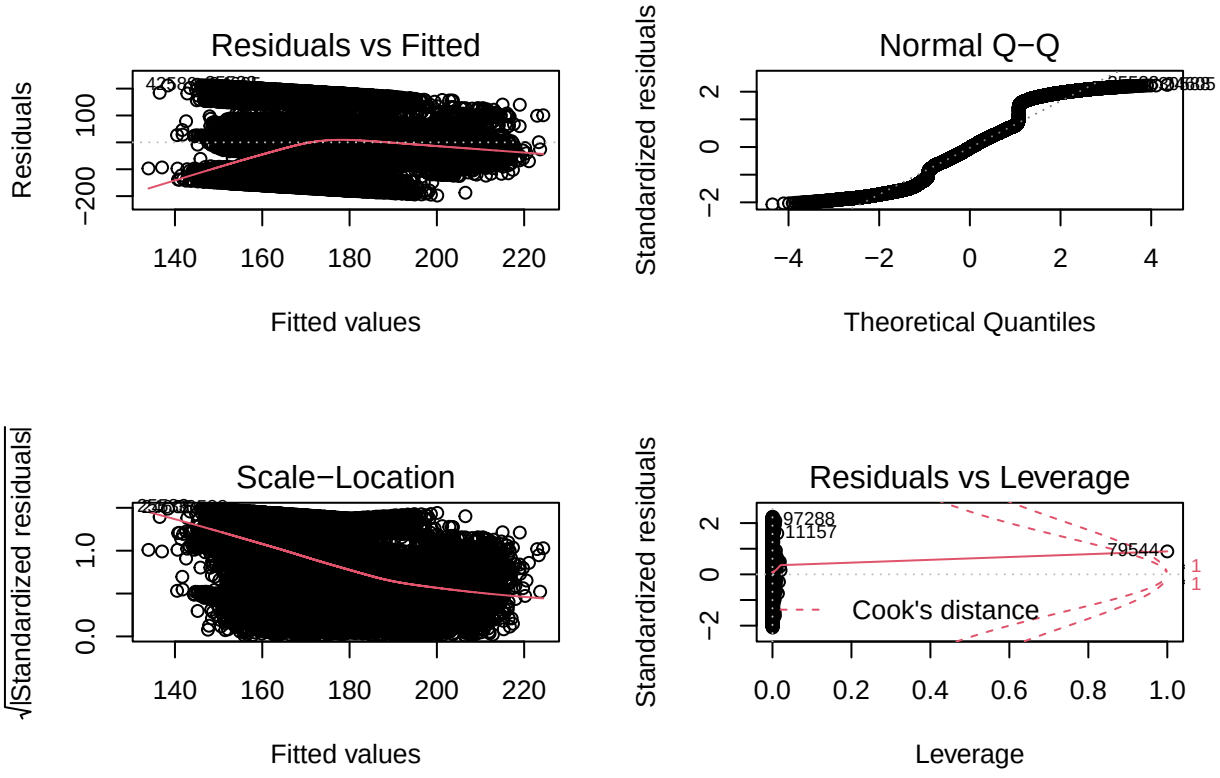
```
## Infrared        -2.367e+00  4.665e-01  -5.074 3.90e-07 ***
## Camera           1.073e+00  2.200e-01   4.875 1.09e-06 ***
## ClassQSO         1.175e+01  1.591e+00   7.384 1.55e-13 ***
## ClassSTAR       -2.883e+00  1.000e+00  -2.882 0.003953 **
## Redshift        -2.785e+00  8.343e-01  -3.338 0.000845 ***
## Plate           -4.505e-05  1.621e-04  -0.278 0.781034
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 95.38 on 74988 degrees of freedom
## Multiple R-squared:  0.02095,    Adjusted R-squared:  0.0208
## F-statistic: 145.9 on 11 and 74988 DF,  p-value: < 2.2e-16
```

Based on the summary, we can see that the R-squared value is 0.01 (really bad) and the F-statistic value is 47.8 (really good).

```
par(mfrow=c(2,2))
plot(slm)
```

```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```



These are graphs made using the Linear Regression Model Residuals VS Fitted: We want to see a straight, red line, however there is a lot of bend suggesting heavy variation in our data Normal Q-Q: We want to see

a straight line that sits right along the dotted line, however we do not see that. Scale- Location: We want to see a horizontal line with evenly spaced data on both sides, however we do not see that. Residuals VS Leverage: This graphs says what leverage points are influencing the line of regression.

Using the summary of the data above, I selected the best predictors, Delta, Ultraviolet, and Plate, and created a Linear Model using only those three to test if I will yield better results

```
bslm <- lm(Alpha~Delta+Ultraviolet+Plate, data = train)
summary(bslm)
```

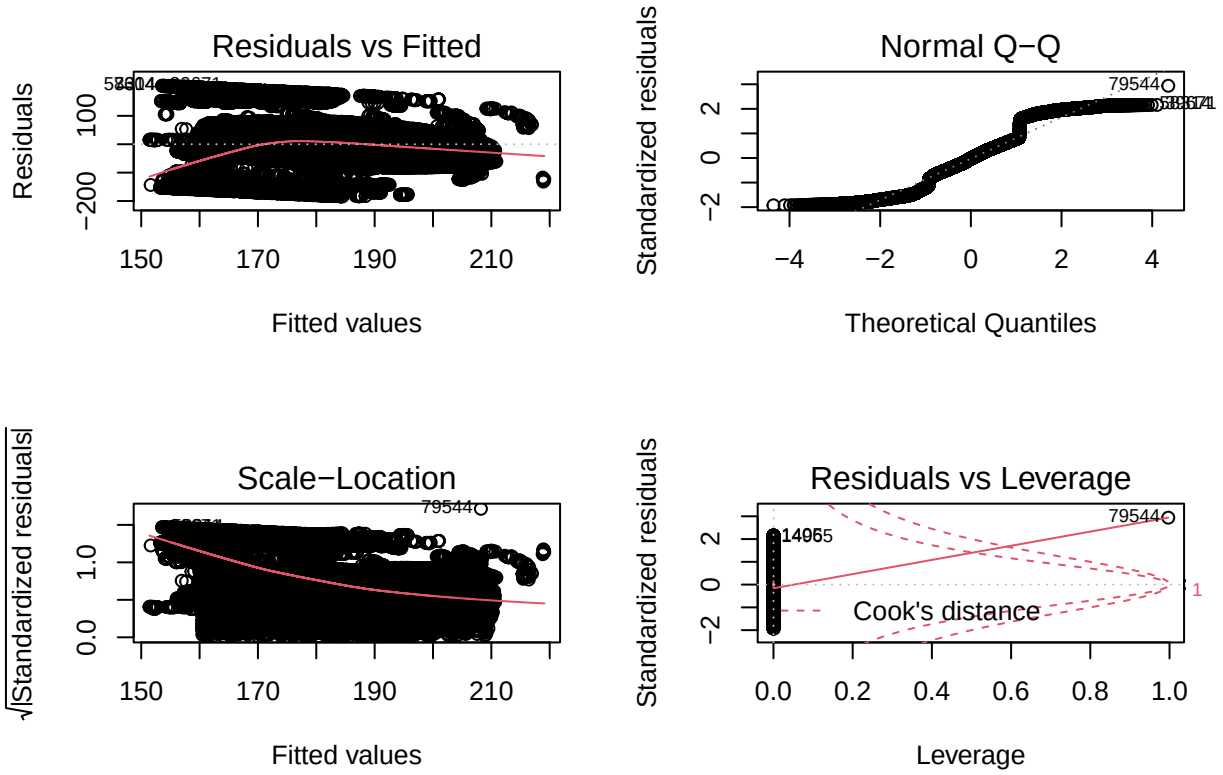
```
##
## Call:
## lm(formula = Alpha ~ Delta + Ultraviolet + Plate, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -183.867  -56.701   -2.556   51.661  206.284
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.650e+02  8.118e-01 203.215  < 2e-16 ***
## Delta        6.714e-01  1.786e-02  37.593  < 2e-16 ***
## Ultraviolet -4.566e-03  9.515e-03  -0.480    0.631
## Plate       -5.980e-04  1.191e-04  -5.023 5.11e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 95.5 on 74996 degrees of freedom
## Multiple R-squared:  0.01851,    Adjusted R-squared:  0.01847
## F-statistic: 471.4 on 3 and 74996 DF,  p-value: < 2.2e-16
```

As shown above, the R-squared value stayed relatively the same, however, the F-statistic value increased tremendously to 151.5, showing some improvement.

```
par(mfrow=c(2,2))
plot(bslm)
```

```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```



When compairing the two sets of graphs, there is very little to no improvement at all between the two models.

Metrics from Linear Reg Model with all data:

```
pred <- predict(slm, newdata = test)
cor <- cor(pred, test$Alpha)
mse <- mean((pred - test$Alpha)^2)
rmse <- sqrt(mse)
print(paste("Correlation: ", cor))
```

```
## [1] "Correlation:  0.153419689549609"
```

```
print(paste("MSE: ", mse))
```

```
## [1] "MSE:  9158.23149605451"
```

```
print(paste("RMSE: ", rmse))
```

```
## [1] "RMSE:  95.6986493951431"
```

Metrics from Linear Reg model with *** data:

```
pred <- predict(bslm, newdata = test)
cor <- cor(pred, test$Alpha)
mse <- mean((pred - test$Alpha)^2)
rmse <- sqrt(mse)
print(paste("Correlation: ", cor))
```

```
## [1] "Correlation:  0.151282390888226"
```

```
print(paste("MSE: ", mse))
```

```
## [1] "MSE:  9166.01821331926"
```

```
print(paste("RMSE: ", rmse))
```

```
## [1] "RMSE:  95.7393242785809"
```

Based on the numbers shown above, the Model made with only the best predictors performed slightly better than the model will all predictors. However, with a Correlation of 13% and a RSME of 95.8, this model is not a reliable model to make predictions on.

##kNN:

My second algorithm is the kNN algorithm. Before building the models, I first test to see which value of K yields the best results. Any value after 11 crashed, so I capped the value to 11.

You will also notice that I re-split the data into train/test. That is because I needed to change the Class variable to numeric for the kNN model. So after I made that change, I re-split the data into new train/test models using the new variable type.

```
library(caret)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```
set.seed(1234)
sdf$Class <- as.numeric(sdf$Class)
split <- sample(1:nrow(sdf), nrow(sdf)*0.75, replace=FALSE)
train <- sdf[split1, ]
test  <- sdf[-split1, ]
cor <- rep(0, 11)
mse <- rep(0, 11)
i <- 1
for (k in seq(1, 21, 2)){
  fit <- knnreg(train[,2:11], train[,1], k=k)
  pred <- predict(fit, test[,2:11])
  cor[i] <- cor(pred, test$Alpha)
  mse[i] <- mean((pred-test$Alpha)^2)
  print(paste("k=", k, cor[i], mse[i]))
  i <- i + 1
}
```

```
## [1] "k= 1 0.968775866279156 585.47098418187"
## [1] "k= 3 0.976654203766022 433.218028859147"
## [1] "k= 5 0.97584648265685 447.577036051279"
## [1] "k= 7 0.974944997075302 463.966573919218"
## [1] "k= 9 0.973342643545149 493.29558017041"
## [1] "k= 11 0.971630823859196 524.591887536437"
## [1] "k= 13 0.970975116495592 536.72351622205"
## [1] "k= 15 0.97000146156705 554.619713741977"
## [1] "k= 17 0.968270209120035 586.260274653491"
## [1] "k= 19 0.966999835219897 609.485780142016"
## [1] "k= 21 0.965804263802345 631.391462838067"
```

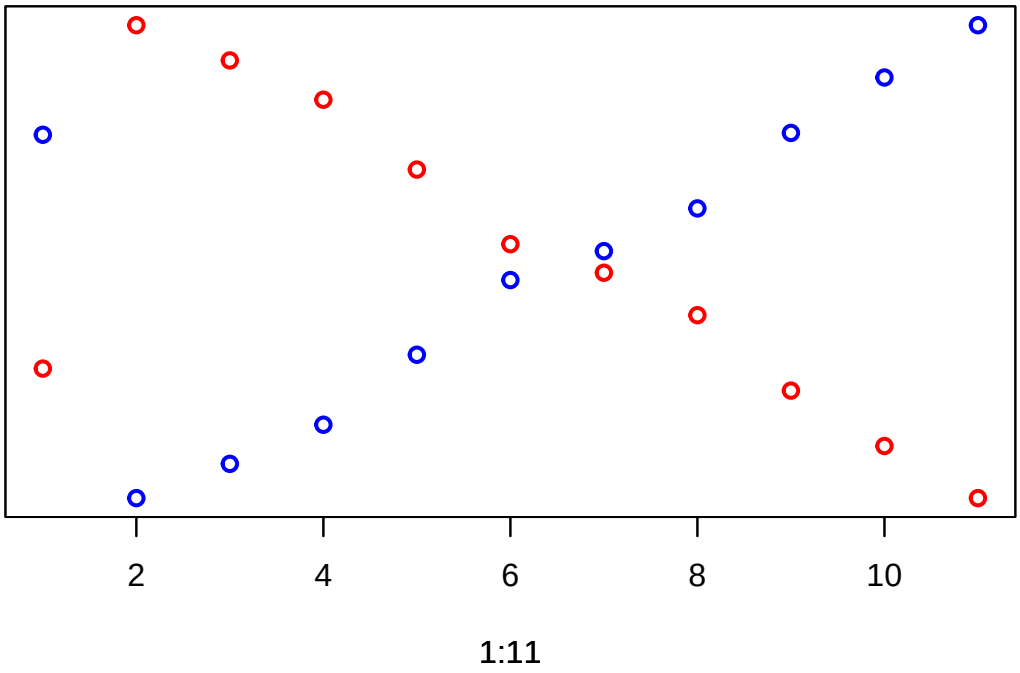
```
plot(1:11, cor, lwd=2, col='red', ylab="",yaxt='n')
par (new=TRUE)
plot(1:11, mse, lwd=2, col='blue', labels=FALSE, ylab="",yaxt='n')
```

```
## Warning in plot.window(...): "labels" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "labels" is not a graphical parameter

## Warning in box(...): "labels" is not a graphical parameter

## Warning in title(...): "labels" is not a graphical parameter
```



As shown in the graph above, the k value that yielded the best results was 2, which is uncommon. I now will use this to build a kNN model on my dataset.

```
fit <- knnreg(train[,2:11], train[,1], k=2)
pred <- predict(fit, test[,2:11])
cor <- cor(pred, test$Alpha)
mse <- mean((pred-test$Alpha)^2)
rmse <- sqrt(mse)
print(paste("Correlation: ", cor))
```

```
## [1] "Correlation:  0.976569908603225"
```

```
print(paste("MSE: ", mse))
```

```
## [1] "MSE:  435.674711145312"
```

```
print(paste("RMSE: ", rmse))
```

```
## [1] "RMSE:  20.8728223090533"
```

The results above shows that the kNN model is a very good model to base predictions on. Even though the RSME value is still very high, it is significantly lower than the linear regression model. In addition, the correlation is at a 93%, which is extremely well.

Next, I scaled the data set and built a kNN model using the scaled data.

Scaling data:

```
trainscaled <- train[,2:11]
testscaled <- test[,2:11]
means <- sapply(trainscaled, mean)
stddev <- sapply(trainscaled, sd)
trainscaled <- scale(trainscaled, center = means, scale = stddev)
testscaled <- scale(testscaled, center = means, scale = stddev)
```

Applying knn to scaled data:

```
fit <- knnreg(trainscaled, train[,1], k=2)
pred <- predict(fit, testscaled)
cor <- cor(pred, test$Alpha)
mse <- mean((pred-test$Alpha)^2)
rmse <- sqrt(mse)
print(paste("Correlation: ", cor))
```

```
## [1] "Correlation:  0.498311142985604"
```

```
print(paste("MSE: ", mse))
```

```
## [1] "MSE:  8069.84078681173"
```

```
print(paste("RMSE: ", rmse))
```

```
## [1] "RMSE:  89.8322925612596"
```

The results of this is unconventional since scaling the dataset usually yields better results. However, it is very clear that the un-scaled version of the kNN model returned much better results than the scaled version.

##Decision Trees:

The last algorithm I used is the decision trees algorithm. Since I already split the data into train/test, I do not need to repeat that step and can just go straight into calculation.

```
library(rpart)
set.seed(1234)
stree <- rpart(Alpha~., method = "anova", data = train)
summary(stree)
```

```
## Call:
## rpart(formula = Alpha ~ ., data = train, method = "anova")
##   n= 75000
##
##           CP nsplit rel error    xerror        xstd
## 1 0.05287387      0 1.0000000 1.0000218 0.004422121
## 2 0.01420353      1 0.9471261 0.9486779 0.004561499
## 3 0.01212187      5 0.8903120 0.8902083 0.004535789
## 4 0.01204679      6 0.8781901 0.8842505 0.004555919
## 5 0.01175630      7 0.8661433 0.8762015 0.004563842
## 6 0.01100279     11 0.8191182 0.8481828 0.004623417
## 7 0.01000000     22 0.6620667 0.6769108 0.004481608
##
## Variable importance
##         Plate         Delta           Red         Green Near-Infrared
##            56            32             3             3             2
##      Redshift      Infrared
##             2             2
##
## Node number 1: 75000 observations,    complexity param=0.05287387
##   mean=178.0332, MSE=9290.782
##   left son=2 (8992 obs) right son=3 (66008 obs)
##   Primary splits:
##       Delta         < -0.0239901    to the left,  improve=0.052873870, (0 missing)
##       Plate         < 9275.5        to the right, improve=0.014863220, (0 missing)
##       Redshift      < -0.0001664985 to the right, improve=0.004405489, (0 missing)
##       Red           < 21.50824      to the right, improve=0.003055258, (0 missing)
##       Near-Infrared < 21.1707       to the right, improve=0.002024371, (0 missing)
##   Surrogate splits:
##       Plate < 348.5         to the left,  agree=0.884, adj=0.029, (0 split)
##
## Node number 2: 8992 observations,    complexity param=0.0117563
##   mean=117.9827, MSE=14230.12
##   left son=4 (4061 obs) right son=5 (4931 obs)
##   Primary splits:
##       Plate  < 4216.5        to the right, improve=0.059868370, (0 missing)
##       Delta  < -2.634855     to the left,  improve=0.031372360, (0 missing)
##       Camera < 4.5           to the left,  improve=0.021197760, (0 missing)
##       Red    < 21.42276      to the right, improve=0.009886309, (0 missing)
##       Green  < 18.767        to the right, improve=0.006267073, (0 missing)
##   Surrogate splits:
##       Red           < 20.40628      to the right, agree=0.735, adj=0.412, (0 split)
```

```
##       Redshift      < 0.467282      to the right, agree=0.731, adj=0.404, (0 split)
##       Green         < 21.31324      to the right, agree=0.728, adj=0.399, (0 split)
##       Near-Infrared < 19.50901      to the right, agree=0.725, adj=0.391, (0 split)
##       Infrared      < 19.1423       to the right, agree=0.715, adj=0.368, (0 split)
##
## Node number 3: 66008 observations,    complexity param=0.01420353
##   mean=186.2136, MSE=8059.757
##   left son=6 (64913 obs) right son=7 (1095 obs)
##   Primary splits:
##       Delta    < 0.2080372     to the right, improve=0.018234580, (0 missing)
##       Plate    < 9297          to the right, improve=0.011351220, (0 missing)
##       Redshift < -0.0001667998 to the right, improve=0.005295516, (0 missing)
##       Camera   < 5.5           to the right, improve=0.004370719, (0 missing)
##       Red      < 21.47954      to the right, improve=0.001562400, (0 missing)
##
## Node number 4: 4061 observations,    complexity param=0.0117563
##   mean=85.81984, MSE=15594.39
##   left son=8 (2535 obs) right son=9 (1526 obs)
##   Primary splits:
##       Delta       < -1.980944     to the left,  improve=0.089148060, (0 missing)
##       Plate       < 9274.5        to the right, improve=0.054408190, (0 missing)
##       Camera      < 5.5           to the right, improve=0.008629235, (0 missing)
##       Ultraviolet < 26.98728      to the right, improve=0.001808622, (0 missing)
##       Class       < 1.5           to the left,  improve=0.001637850, (0 missing)
##   Surrogate splits:
##       Camera        < 2.5           to the right, agree=0.689, adj=0.171, (0 split)
##       Plate         < 8559.5        to the left,  agree=0.687, adj=0.166, (0 split)
##       Red           < 22.03914      to the left,  agree=0.644, adj=0.054, (0 split)
##       Near-Infrared < 21.92872      to the left,  agree=0.639, adj=0.040, (0 split)
##       Infrared      < 21.1791       to the left,  agree=0.636, adj=0.030, (0 split)
##
## Node number 5: 4931 observations
##   mean=144.4708, MSE=11552.99
##
## Node number 6: 64913 observations,    complexity param=0.01420353
##   mean=184.6391, MSE=7849.983
##   left son=12 (1010 obs) right son=13 (63903 obs)
##   Primary splits:
##       Delta    < 0.6308733     to the left,  improve=0.017227000, (0 missing)
##       Plate    < 9297          to the right, improve=0.010081090, (0 missing)
##       Redshift < -0.000162424  to the right, improve=0.005790316, (0 missing)
##       Camera   < 5.5           to the right, improve=0.003668100, (0 missing)
##       Red      < 21.47954      to the right, improve=0.002218836, (0 missing)
##   Surrogate splits:
##       Plate < 295           to the left,  agree=0.985, adj=0.014, (0 split)
##
## Node number 7: 1095 observations
##   mean=279.5536, MSE=11636.12
##
## Node number 8: 2535 observations
##   mean=56.89117, MSE=10340.01
##
## Node number 9: 1526 observations,    complexity param=0.0117563
##   mean=133.8763, MSE=20623.37
```

```
##   left son=18 (348 obs) right son=19 (1178 obs)
##   Primary splits:
##       Plate         < 9274.5        to the right, improve=0.138103000, (0 missing)
##       Delta         < -1.519187     to the right, improve=0.101073900, (0 missing)
##       Camera        < 4.5           to the left,  improve=0.079676730, (0 missing)
##       Red           < 20.2323       to the left,  improve=0.012245360, (0 missing)
##       Near-Infrared < 19.82714      to the left,  improve=0.009429237, (0 missing)
##   Surrogate splits:
##       Near-Infrared < 23.50616      to the right, agree=0.773, adj=0.003, (0 split)
##
## Node number 12: 1010 observations
##   mean=92.13964, MSE=9457.472
##
## Node number 13: 63903 observations,    complexity param=0.01420353
##   mean=186.1011, MSE=7687.208
##   left son=26 (6220 obs) right son=27 (57683 obs)
##   Primary splits:
##       Plate    < 9297          to the right, improve=0.010333370, (0 missing)
##       Redshift < -0.0003252258 to the right, improve=0.005524837, (0 missing)
##       Camera   < 5.5           to the right, improve=0.004696723, (0 missing)
##       Delta    < 1.055036      to the left,  improve=0.003559109, (0 missing)
##       Red      < 21.47954      to the right, improve=0.001974414, (0 missing)
##   Surrogate splits:
##       Red < 22.32801      to the right, agree=0.903, adj=0.002, (0 split)
##
## Node number 18: 348 observations
##   mean=35.68697, MSE=928.2623
##
## Node number 19: 1178 observations,    complexity param=0.0117563
##   mean=162.883, MSE=22752.08
##   left son=38 (841 obs) right son=39 (337 obs)
##   Primary splits:
##       Plate         < 8966          to the left,  improve=0.56395340, (0 missing)
##       Delta         < -1.518679     to the right, improve=0.12532230, (0 missing)
##       Camera        < 2.5           to the left,  improve=0.07999260, (0 missing)
##       Near-Infrared < 20.68429      to the left,  improve=0.07521627, (0 missing)
##       Infrared      < 19.94214      to the left,  improve=0.06973311, (0 missing)
##   Surrogate splits:
##       Red      < 22.05202      to the left,  agree=0.724, adj=0.036, (0 split)
##       Green    < 23.69742      to the left,  agree=0.717, adj=0.012, (0 split)
##       Infrared < 23.80331      to the left,  agree=0.716, adj=0.006, (0 split)
##
## Node number 26: 6220 observations,    complexity param=0.01212187
##   mean=158.9595, MSE=8142.737
##   left son=52 (592 obs) right son=53 (5628 obs)
##   Primary splits:
##       Plate < 9496.5        to the left,  improve=0.16677170, (0 missing)
##       Delta < 31.89203      to the left,  improve=0.10600550, (0 missing)
##       Class < 1.5           to the left,  improve=0.03449313, (0 missing)
##       Green < 22.07614      to the right, improve=0.02588146, (0 missing)
##       Red   < 21.91066      to the right, improve=0.02145581, (0 missing)
##   Surrogate splits:
##       Delta < 2.036113      to the left,  agree=0.919, adj=0.149, (0 split)
##
```

```
## Node number 27: 57683 observations,    complexity param=0.01420353
##   mean=189.0277, MSE=7550.087
##   left son=54 (56974 obs) right son=55 (709 obs)
##   Primary splits:
##       Plate    < 9078          to the left,  improve=0.036814660, (0 missing)
##       Camera   < 5.5           to the right, improve=0.004967925, (0 missing)
##       Redshift < -0.0003249548 to the right, improve=0.004923932, (0 missing)
##       Delta    < 1.048903      to the left,  improve=0.003801004, (0 missing)
##       Class    < 2.5           to the right, improve=0.001736548, (0 missing)
##
## Node number 38: 841 observations
##   mean=91.17807, MSE=13842.82
##
## Node number 39: 337 observations
##   mean=341.8262, MSE=133.7804
##
## Node number 52: 592 observations
##   mean=45.33752, MSE=2028.646
##
## Node number 53: 5628 observations,    complexity param=0.01204679
##   mean=170.9113, MSE=7285.046
##   left son=106 (5207 obs) right son=107 (421 obs)
##   Primary splits:
##       Delta    < 6.074283      to the right, improve=0.20473800, (0 missing)
##       Plate    < 11277.5       to the left,  improve=0.13927550, (0 missing)
##       Camera   < 3.5           to the left,  improve=0.01814684, (0 missing)
##       Class    < 1.5           to the left,  improve=0.01664732, (0 missing)
##       Redshift < -8.185505e-06 to the right, improve=0.01442778, (0 missing)
##   Surrogate splits:
##       Plate < 12115.5       to the left,  agree=0.937, adj=0.154, (0 split)
##
## Node number 54: 56974 observations,    complexity param=0.01100279
##   mean=187.1679, MSE=7360.942
##   left son=108 (3038 obs) right son=109 (53936 obs)
##   Primary splits:
##       Delta    < 2.23694       to the left,  improve=0.015580640, (0 missing)
##       Plate    < 8637          to the right, improve=0.007647595, (0 missing)
##       Redshift < -0.0003249548 to the right, improve=0.005402204, (0 missing)
##       Camera   < 5.5           to the right, improve=0.005086036, (0 missing)
##       Class    < 1.5           to the right, improve=0.001538073, (0 missing)
##   Surrogate splits:
##       Plate < 416.5         to the left,  agree=0.95, adj=0.053, (0 split)
##
## Node number 55: 709 observations
##   mean=338.4797, MSE=135.6289
##
## Node number 106: 5207 observations
##   mean=159.9297, MSE=5795.599
##
## Node number 107: 421 observations
##   mean=306.7326, MSE=5767.812
##
## Node number 108: 3038 observations
##   mean=142.0442, MSE=16974.81
```

```
##
## Node number 109: 53936 observations,    complexity param=0.01100279
##   mean=189.7096, MSE=6698.283
##   left son=218 (1271 obs) right son=219 (52665 obs)
##   Primary splits:
##       Plate       < 8637          to the right, improve=0.0109982000, (0 missing)
##       Redshift    < -0.0003335784 to the right, improve=0.0066746220, (0 missing)
##       Delta       < 27.89527      to the right, improve=0.0026312040, (0 missing)
##       Camera      < 5.5           to the right, improve=0.0020412730, (0 missing)
##       Ultraviolet < 21.51199      to the left,  improve=0.0006469446, (0 missing)
##
## Node number 218: 1271 observations
##   mean=134.4598, MSE=4729.529
##
## Node number 219: 52665 observations,    complexity param=0.01100279
##   mean=191.0429, MSE=6670.349
##   left son=438 (27309 obs) right son=439 (25356 obs)
##   Primary splits:
##       Plate       < 4876.5        to the left,  improve=0.010800500, (0 missing)
##       Redshift    < -0.0003335784 to the right, improve=0.006593982, (0 missing)
##       Delta       < 27.89527      to the right, improve=0.003769587, (0 missing)
##       Camera      < 5.5           to the right, improve=0.002116614, (0 missing)
##       Ultraviolet < 21.51199      to the left,  improve=0.000886308, (0 missing)
##   Surrogate splits:
##       Red           < 19.2868       to the left,  agree=0.743, adj=0.466, (0 split)
##       Green         < 19.72585      to the left,  agree=0.735, adj=0.450, (0 split)
##       Near-Infrared < 18.68144      to the left,  agree=0.734, adj=0.447, (0 split)
##       Infrared      < 18.23401      to the left,  agree=0.726, adj=0.432, (0 split)
##       Redshift      < 0.2494968     to the left,  agree=0.709, adj=0.396, (0 split)
##
## Node number 438: 27309 observations,    complexity param=0.01100279
##   mean=182.8643, MSE=4973.26
##   left son=876 (4759 obs) right son=877 (22550 obs)
##   Primary splits:
##       Plate         < 4414.5        to the right, improve=0.072540230, (0 missing)
##       Redshift      < -0.000173024  to the right, improve=0.008904983, (0 missing)
##       Infrared      < 17.69939      to the right, improve=0.004635950, (0 missing)
##       Near-Infrared < 18.02545      to the right, improve=0.004399151, (0 missing)
##       Delta         < 68.60949      to the left,  improve=0.003970373, (0 missing)
##   Surrogate splits:
##       Redshift    < 2.195179      to the right, agree=0.826, adj=0.004, (0 split)
##       Ultraviolet < 27.6967       to the right, agree=0.826, adj=0.000, (0 split)
##
## Node number 439: 25356 observations,    complexity param=0.01100279
##   mean=199.8516, MSE=8348.519
##   left son=878 (23499 obs) right son=879 (1857 obs)
##   Primary splits:
##       Plate    < 5087.5        to the right, improve=0.066987920, (0 missing)
##       Delta    < 41.85683      to the right, improve=0.010180560, (0 missing)
##       Redshift < -0.0003276062 to the right, improve=0.006048975, (0 missing)
##       Camera   < 4.5           to the right, improve=0.002310603, (0 missing)
##       Class    < 2.5           to the left,  improve=0.001025087, (0 missing)
##
## Node number 876: 4759 observations
```

```
##   mean=141.519, MSE=3830.043
##
## Node number 877: 22550 observations,    complexity param=0.01100279
##   mean=191.5898, MSE=4777.629
##   left son=1754 (18887 obs) right son=1755 (3663 obs)
##   Primary splits:
##       Plate       < 3835.5        to the left,  improve=0.08583288, (0 missing)
##       Delta       < 24.07322      to the right, improve=0.02176410, (0 missing)
##       Green       < 20.90887      to the left,  improve=0.01319352, (0 missing)
##       Ultraviolet < 21.47942      to the left,  improve=0.01262570, (0 missing)
##       Redshift    < 0.263021      to the left,  improve=0.01187345, (0 missing)
##   Surrogate splits:
##       Green         < 21.3053       to the left,  agree=0.893, adj=0.340, (0 split)
##       Red           < 20.04266      to the left,  agree=0.884, adj=0.283, (0 split)
##       Redshift      < 0.4370557     to the left,  agree=0.866, adj=0.174, (0 split)
##       Near-Infrared < 19.3131       to the left,  agree=0.864, adj=0.162, (0 split)
##       Ultraviolet   < 23.14172      to the left,  agree=0.855, adj=0.106, (0 split)
##
## Node number 878: 23499 observations,    complexity param=0.01100279
##   mean=193.2037, MSE=8235.552
##   left son=1756 (255 obs) right son=1757 (23244 obs)
##   Primary splits:
##       Plate    < 5140.5        to the left,  improve=0.038639420, (0 missing)
##       Redshift < -0.0002221646 to the right, improve=0.005988348, (0 missing)
##       Delta    < 19.70695      to the left,  improve=0.005406339, (0 missing)
##       Class    < 1.5           to the left,  improve=0.001803195, (0 missing)
##       Camera   < 4.5           to the right, improve=0.001389908, (0 missing)
##
## Node number 879: 1857 observations
##   mean=283.9759, MSE=2141.877
##
## Node number 1754: 18887 observations
##   mean=182.6718, MSE=3736.097
##
## Node number 1755: 3663 observations
##   mean=237.5727, MSE=7623.427
##
## Node number 1756: 255 observations
##   mean=22.891, MSE=22.20983
##
## Node number 1757: 23244 observations,    complexity param=0.01100279
##   mean=195.0721, MSE=8003.949
##   left son=3514 (4842 obs) right son=3515 (18402 obs)
##   Primary splits:
##       Plate    < 5893.5        to the left,  improve=0.018465030, (0 missing)
##       Redshift < -0.0002221646 to the right, improve=0.006110257, (0 missing)
##       Delta    < 27.88457      to the right, improve=0.005472111, (0 missing)
##       Camera   < 4.5           to the right, improve=0.001829647, (0 missing)
##       Class    < 2.5           to the left,  improve=0.001298454, (0 missing)
##   Surrogate splits:
##       Delta    < 17.73964      to the left,  agree=0.844, adj=0.249, (0 split)
##       Redshift < -0.003611535  to the left,  agree=0.792, adj=0.000, (0 split)
##
## Node number 3514: 4842 observations
```

```
##   mean=171.3722, MSE=2441.785
##
## Node number 3515: 18402 observations,    complexity param=0.01100279
##   mean=201.3081, MSE=9280.804
##   left son=7030 (15701 obs) right son=7031 (2701 obs)
##   Primary splits:
##       Plate    < 6169.5        to the right, improve=0.086453690, (0 missing)
##       Delta    < 27.88457      to the right, improve=0.020643460, (0 missing)
##       Redshift < -0.0002550971 to the right, improve=0.004896388, (0 missing)
##       Camera   < 4.5           to the right, improve=0.002782063, (0 missing)
##       Red      < 21.50792      to the right, improve=0.001394068, (0 missing)
##   Surrogate splits:
##       Delta         < 17.21303      to the right, agree=0.866, adj=0.090, (0 split)
##       Red           < 13.99958      to the right, agree=0.853, adj=0.001, (0 split)
##       Green         < 14.50382      to the right, agree=0.853, adj=0.001, (0 split)
##       Near-Infrared < 13.34717      to the right, agree=0.853, adj=0.001, (0 split)
##       Infrared      < 13.00675      to the right, agree=0.853, adj=0.001, (0 split)
##
## Node number 7030: 15701 observations,    complexity param=0.01100279
##   mean=189.5596, MSE=8943.602
##   left son=14060 (831 obs) right son=14061 (14870 obs)
##   Primary splits:
##       Plate    < 6288.5        to the left,  improve=0.190668100, (0 missing)
##       Delta    < 17.32698      to the left,  improve=0.083408280, (0 missing)
##       Redshift < -0.000207865  to the right, improve=0.004620279, (0 missing)
##       Camera   < 5.5           to the right, improve=0.003952893, (0 missing)
##       Class    < 1.5           to the left,  improve=0.003111757, (0 missing)
##   Surrogate splits:
##       Delta < 17.50367      to the left,  agree=0.961, adj=0.262, (0 split)
##
## Node number 7031: 2701 observations,    complexity param=0.01100279
##   mean=269.6027, MSE=5774.472
##   left son=14062 (1733 obs) right son=14063 (968 obs)
##   Primary splits:
##       Plate         < 6113.5        to the left,  improve=0.60274810, (0 missing)
##       Delta         < 21.57582      to the right, improve=0.31773540, (0 missing)
##       Infrared      < 16.61185      to the right, improve=0.01390892, (0 missing)
##       Near-Infrared < 16.95518      to the right, improve=0.01301732, (0 missing)
##       Red           < 17.33136      to the right, improve=0.01198228, (0 missing)
##   Surrogate splits:
##       Delta         < 21.18674      to the right, agree=0.913, adj=0.757, (0 split)
##       Red           < 17.25861      to the right, agree=0.650, adj=0.023, (0 split)
##       Infrared      < 16.35632      to the right, agree=0.650, adj=0.023, (0 split)
##       Near-Infrared < 16.73591      to the right, agree=0.649, adj=0.021, (0 split)
##       Green         < 18.71616      to the right, agree=0.648, adj=0.019, (0 split)
##
## Node number 14060: 831 observations
##   mean=14.87679, MSE=1935.396
##
## Node number 14061: 14870 observations
##   mean=199.3216, MSE=7534.694
##
## Node number 14062: 1733 observations
##   mean=225.5104, MSE=2995.002
```

```
##
## Node number 14063: 968 observations
##   mean=348.5406, MSE=1038.782
```

```
pred <- predict(stree, newdata = test)
cor <- cor(pred, test$Alpha)
mse <- mean((pred-test$Alpha)^2)
rmse <- sqrt(mse)
print(paste("Correlation: ", cor))
```

```
## [1] "Correlation:  0.582666345465184"
```

```
print(paste("MSE: ", mse))
```

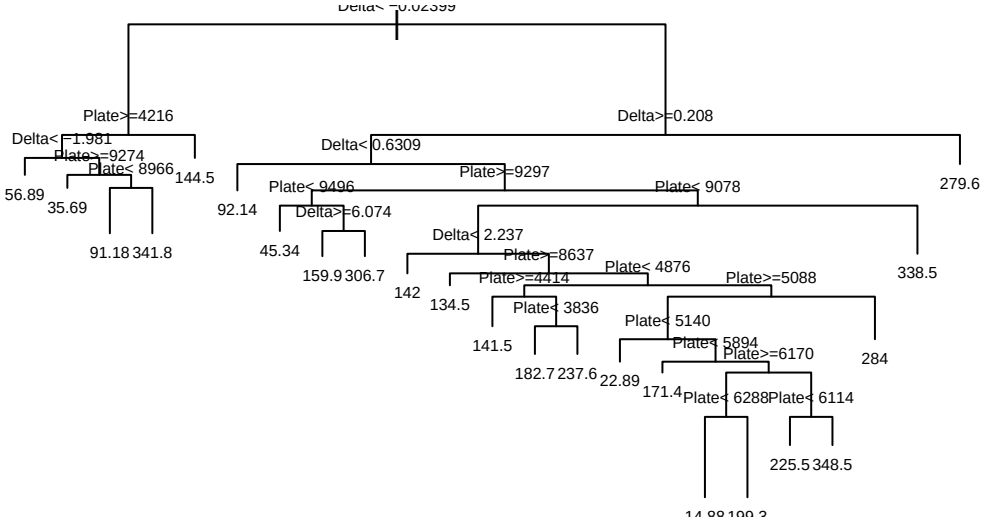
```
## [1] "MSE:  6194.90314796585"
```

```
print(paste("RMSE: ", rmse))
```

```
## [1] "RMSE:  78.7077070429945"
```

Even though this model outperformed the linear regression model, it still fell short to the kNN.

```
plot(stree)
text(stree, cex=0.5, pretty = 0)
```



Results:

The ranking for the best to worst algorithms for this particular data set is as follows:

1. kNN with Unscaled Data
2. Decision Trees
3. Modified Linear Regression

kNN with unscaled data performed the best for my dataset. With a correlation of 93%, it out performed its counterpart - kNN with scaled data with a correlation of 20%. It also outperformed Decision trees, which had a correlation of 48% and the both Linear regression models, the non-optimized model with a correlation of 5% and the optimized version of the model with a correlation of 13%.