# Analysis of Chicago Public High Schools' Graduation Rate

Author: Jacynda Alatoma

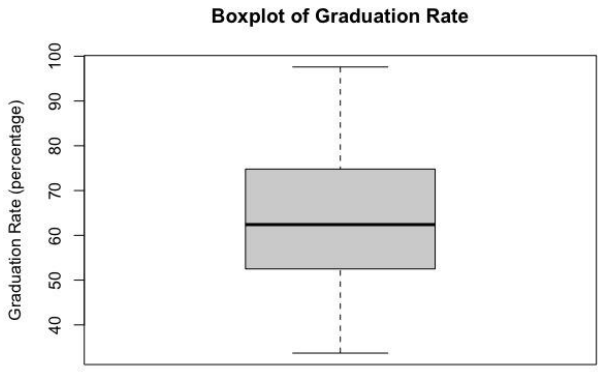
## Abstract

Graduation rate is one of the most important factors to measure the success of a school as well as the success of a student in future careers. The goal of the research was to see which factors affected graduation rate in Chicago public high schools in the years 2011-2012. Multiple linear regression and random forest regression were both used to analyze the data. Linear regression was found to be more accurate and reliable when predicting the graduation rate. It was found that the factors that affected graduation rate were college eligibility, student attendance rate, and rate of misconducts per 100 students. Both college eligibility and student attendance rate had a positive relationship with graduation rate while misconducts had a negative relationship with graduation rate.

## Introduction and Background

The success of students in school is very important to families and educators. In the years 2011-2012 over one-third of the public schools in Chicago, IL, were reported to be on "probation." This means the schools were lacking in areas including ACT scores, student retention rate, graduation rate, and more (Cora 2012). Graduation rates are all over the board, referring to Figure 1, it is seen that graduation rates range from 33.7% to 97.6%, with the middle 50% of schools having graduation rates between 52.5% to 74.8% (see Appendix A).

*Figure 1: Boxplot of Graduation Rate*




It is important to understand what factors are affecting graduation rate to hopefully increase the rate and bring some schools out of the "probation" status. Some factors that might affect graduation rate are attendance (of both teachers and students), college eligibility, and home environment scores. It is intuitive to expect graduation rates to be affected positively when attendance rates are high, as well as high eligibility rates and high environment scores. However, multiple factors will be analyzed to see if they affect graduation rate in Chicago public high schools.

**Methods and Materials**

The dataset was pre-researched and contains 566 observations of 81 predictors. It includes both public high schools and elementary schools located in Chicago. The data was collected not by random, as data was collected for every elementary and high school in Chicago was reported for the 2011-2012 school year. The response variable of interest is "Graduation Rate" which is measured on a scale from 0-100%, 100% being that everyone in the class graduates. Besides the variable of interest, there are 80 other factors in the dataset. This is an observational study since we cannot choose how each school performs in certain areas. The

variables pertaining to location, elementary schools, test scores, or ID numbers have all been removed to simplify the dataset and restrict the view to only high schools' graduating class. Rows missing values for graduation rate have also all been removed for an easier analysis. Other missing cells were imputed using the "mice" package in R and one extreme outlier point from the response was also removed (see Appendix A). There are 18 variables to consider after deleting all unnecessary columns.

Before any analysis were run, multicollinearity was checked among the dataset using VIF. Two variables contained VIF values larger than 10, so they were removed before the analysis continued (see Appendix A). The data was split in a 70% training and 30% test set split (see Appendix A). The first analysis was multiple linear regression. Both a forward and backwards stepwise regression was performed to reduce the number of predictors in the final model. This used AIC to select a model that achieved a minimum. The model was reduced from 18 predictors to 3. A pseudo-final model is listed below:

$$GradRate = \beta_0 + \beta_1 X_1 + \cdots + \beta_3 X_3 + \epsilon \quad (1)$$

Where $\beta_1 - \beta_3$ are estimates of the predictors and $X_1 - X_3$ are the predictors for average student attendance, rate of misconducts per 100 students, and college eligibility rate, respectively. The new stepwise model was fit on the training set and then tested for predictive accuracy on the test set (see Appendix A). Assumptions including normality, constant variance, Cook's distance, and BoxCox plots were all checked, and no transformations nor violations were present in the dataset (See Appendix B). Finally, a type 1 ANOVA was used to analyze the data and find the most significant predictors in affecting graduation rate.

Random forest regression was also used to analyze the data. Random forest algorithms create multiple random decision trees and uses bootstrap aggregation/bagging to come up with a

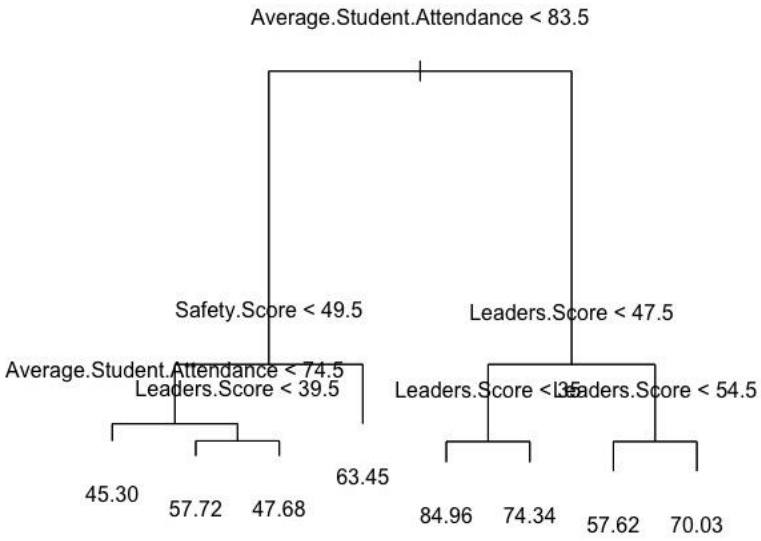
singular outcome. This method was chosen to analyze this dataset because it is less prone to overfitting data than regular decision trees, works well with very large data sets, and increases diversity and reduces dimensionality since not every predictor is used in every tree created (R 2022). The way random forest prevents overfitting is by not using all predictors in each decision tree created. That way, the trees don't overlap or have the exact same predictors and risk overfitting certain predictors. Averaging over all the trees reduces variance as well as reduces correlation between trees. Bootstrap aggregation/bagging means to create many trees from different training sets, selecting with replacement, get their outcome, and then choose the best outcome based off the majority (Genuer 2020). The dataset is split into a training set and a test set. The training set will create the specific individual decision trees, and the test set will test how accurate the trees made are at predicting their responses. The reason why this statistical technique is useful for large datasets, like the one in this research, is because each decision tree made does not use all the predictors (only about a third are used in each tree), so it's not as computationally expensive to generate one tree, but there are multiple trees made in total. Each tree made has a root node, internal nodes, and leaves. A root node is the first predictor split in the tree. Then the random forest algorithm will split into internal nodes based on an equation which minimizes residual of sum squares (RSS), (2) contains the formula for RSS.

$$RSS = \sum_{m=1}^{M} \sum_{i \in R_m} (y_j - \hat{y}_{R_m})^2 \quad (2)$$

Where $\hat{y}_{R_m}$ is the predicted value from the tree with a given input, and M is the number of partitions in a feature space (Faraway 283). An example of one decision tree for this dataset can be seen in Figure 2, however, it is important to note that the random forest regression run for this dataset will produce multiple of these trees and does not include the same predictors in each tree.

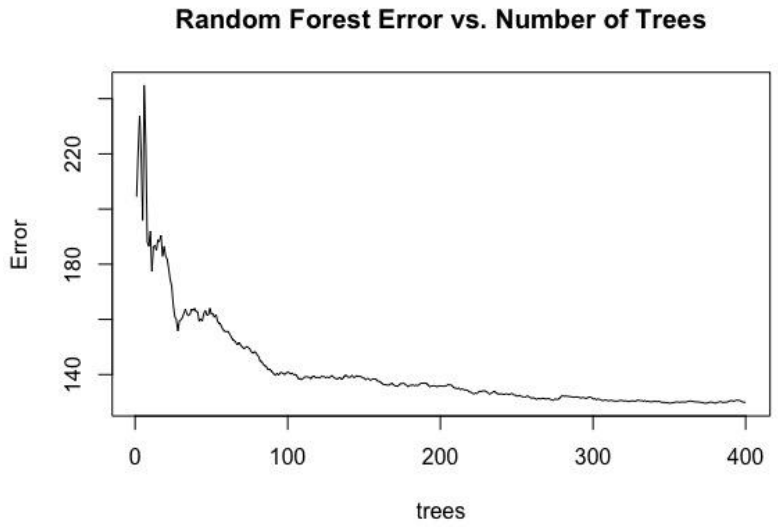
Each decision tree is also unpruned in random forest, like the one below, which means no branches are removed, even if they don't affect predictive performance if they were there or not.

*Figure 2: Unpruned Decision Tree*




There are no formal assumptions for random forest, so none were checked on the data. Each tree considered 4 predictors at each split, a common choice for regression being the square root of the number of predictors in the dataset. A total of 400 decision trees were created to ensure that the error would be at a minimum after all trees were created. Once the random forest algorithm was run on the training set decision trees, the trees were tested for their predictive accuracy using the test set of data (see Appendix A). The random forest regression model was plotted to make sure that the number of trees were appropriate to reduce the amount of error as much as possible, and as in Figure 3, it passed this test.

*Figure 3: Random Forest Error per Amount of Trees*




Finally, a variable importance plot was made along with a data frame comparing the accuracy of the regression model versus the random forest algorithm, and final significant/important predictors were drawn from the outputs.

**Results**

For the multiple linear regression model, a final condensed ANOVA Table 1 is seen below (see Appendix B).

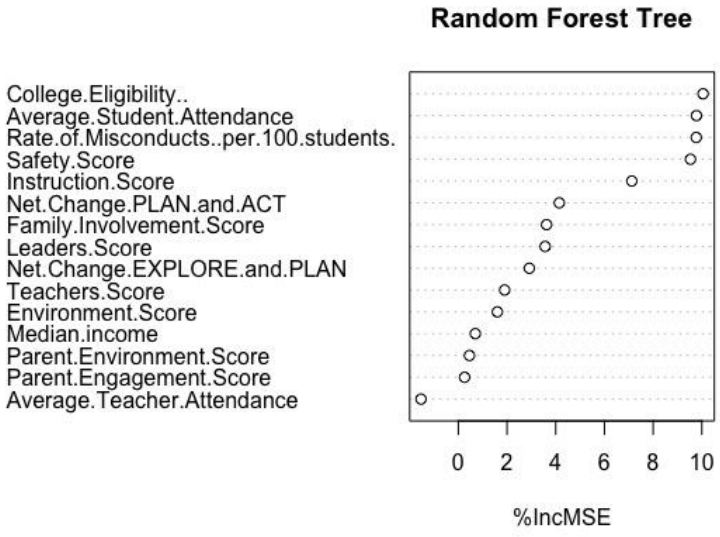
*Table 1: Condensed ANOVA table*

| | F Value | Pr(>F) | Significance |
|---|---|---|---|
| Average Student Attendance | 12.00 | 0.0011 | ** |
| Rate of Misconducts / 100 Students | 6.89 | 0.012 | * |
| College Eligibility | 71.09 | 4.21e-11 | *** |

In the significance column, "***," "**," and "*" denote significance of a factor. All three factors in the model were found to be significant. From the summary of the linear model it shows that the higher student attendance and college eligibility, the higher the predicted graduation rate. Conversely, the more misconducts per 100 students, the less the graduation rate will be. Specifically, there is a 0.54 and 0.3 increase in graduation rate per each 1 unit increase in student attendance and college eligibility, respectively while keeping all other variables fixed. There is a -0.3 percent decrease in graduation rate for every 1 unit increase in misconducts per 100 students while keeping all other variables fixed (see Appendix A).

After running the random forest regression, a plot was generated that orders the most important predictors in the dataset. Random forest orders these predictors by seeing how much each one contributes to the MSE, as seen on the x-axis the label reads "%IncMSE," so these variables have the most influence on the MSE. Figure 4 illustrates this plot.

*Figure 4: Random Forest Variable Importance Plot*



According to the random forest model, the top 3 important factors were college eligibility, average student attendance, and rate of misconducts per 100 students, which agrees with our results from our linear regression model. However, in comparison with the linear regression model, safety score is also another important factor. The multiple linear regression model and the random forest tree model were compared in predictability when predicting scores in the test data set. To check which model fit the dataset better, mean squared error (MSE) was used to compare the fit of the two models. The MSE was lower for the regression model with an MSE of 84.9 compared to 127.1 in the random forest model, meaning that there was less variance in the regression model, and it fit the data better (see Appendix A). The reason why regression outperformed random forest is because random forest doesn't require linearity as one of the assumptions. Therefore, in a linear model determined by the horizontal line at 0 in a residual plot (see Appendix B), regression would outperform predictive performance of random forest. Random forest would outperform regression is in a almost non-parametric setting where linearity assumption is violated.

**Discussion and Summary**

Graduation rate is one of the most important metrics that schools look at to assess their performance. Increasing this metric is ideal and knowing which factors allow that to happen is very important. According to the research found in this paper for the 2011-2012 school year, the factors that affected graduation rates in Chicago Public Schools were college eligibility, average student attendance, and rate of misconducts per 100 students. As one might assume, the more a student shows up to class and engages, the more they will learn. Making sure kids can get to school and learn will positively impact schools' graduation rate. College eligibility, as it states, measures how likely students can attend college. Being eligible for college takes into

consideration standardized testing, grades, and more. A student that is excelling in this is probably a good indicator of teachers being helpful, good resources at school, etc. that will ultimately lead to more students graduating or being able to attend college. If a student knows they are eligible for college, they will be more likely to want to graduate and receive their high school diploma. Finally, the more misconducts a school has, the more chances there is for students to be expelled, suspended, etc. and ultimately lowering the schools' graduation rate.

As stated earlier, many observations, rows, and columns were removed from the dataset because of missing values. This was a big limitation when analyzing the data. Even after removing the major rows with missing values, many cells were missing values within what was left. A suggestion for this would be to run further statistical tests on the data, impute the missing values more efficiently or with a more advanced method. Many of the columns that were deleted because of collinearity should be reincorporated into the model somehow and used to predict the response. Random forest also only fit the data to a ~52% accuracy, further investigation would be required to increase the variation explained. Another suggestion would be to get newer data of the schools now, this data might not be useful anymore as it has been a decade since it was collected, and some schools might have changed the way they run the school. It would also be helpful to collect data earlier on and consider elementary and middle school in a repeated measures cohort study, so even in those early years certain things can be changed to promote high school graduation among the students.

The random forest technique is very useful and easy to understand. One of the main advantages is that it can be used for both problems of regression and classification, so many questions can be applied to this method. Decision trees in themselves are very useful but the way random forest created multiple of them while also not using all the data or predictors to increase

variety makes it less prone to overfitting and reduces variance. This is a very useful technique for a problem like this one as it is most useful for large datasets with many predictors, however perhaps a linear classifier may be preferred due to the poor accuracy of the model and the linearity of the original dataset.

## References

Cora, C. (2012, December 21). *More Than One-third of CPS Schools on Probation*. DNAinfo Chicago. Retrieved November 14, 2022, from https://www.dnainfo.com/chicago/20121221/chicago/more-than-one-third-of-cps-schools-on-probation/

Faraway, J. J. (2004). *Extending linear model with R*. Chapman & Hall/CRC.

Genuer, R., Poggi, JM. (2020). Random Forests. In: Random Forests with R. Use R!. Springer, Cham. https://doi-org.ezp1.lib.umn.edu/10.1007/978-3-030-56485-8_3

R, S. E. (2022, June 21). *Random Forest: Introduction to random forest algorithm*. Analytics Vidhya. Retrieved November 14, 2022, from https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

## Appendix A

```
> summary(CPS$Graduation.Rate..)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  10.30   51.45   62.25   62.46   74.53   97.60
```

Figure A1: Summary statistics of the response, Graduation Rate

```
# MICE imputing missing data
library(mice)
set.seed(4893)
tempData <- mice(CPS,m=5,maxit=50,meth=c("pmm"))

CPS <- complete(tempData, 1)

# remove influential point
outliers <- boxplot(CPS$Graduation.Rate.., plot=FALSE)$out
CPS<- CPS[-which(CPS$Graduation.Rate.. %in% outliers),]
```

Figure A2: Imputing missing data with the ‘mice’ package in R, also removing the extreme outlier point in Graduation Rate

```
> mod <- lm(Graduation.Rate..~., data = CPS)
> DAAG::vif(mod) # ASK ABOUT THIS remove highly correlated values VIF>=10
                     Track.ScheduleStandard                      Track.ScheduleTrack_E
                                    23.4220                                    21.5360
CPS.Performance.Policy.StatusNot on Probation     CPS.Performance.Policy.StatusProbation
                                    29.4650                                    32.2540
                             Safety.Score                  Family.Involvement.Score
                                    9.1595                                     5.3820
                        Environment.Score                       Instruction.Score
                                    9.7388                                     8.8271
                           Leaders.Score                          Teachers.Score
                                    6.4579                                     4.3683
                  Parent.Engagement.Score                 Parent.Environment.Score
                                    4.3679                                     5.8268
                Average.Student.Attendance      Rate.of.Misconducts..per.100.students.
                                    4.6403                                     2.2251
                Average.Teacher.Attendance            Net.Change.EXPLORE.and.PLAN
                                    2.4459                                     3.5663
                   Net.Change.PLAN.and.ACT                    College.Eligibility..
                                    4.0219                                     7.9201
                             Median.income
                                    4.4415
> CPSnew = CPS[, -c(1,2)]
```

Figure A3: Checking VIF and removing 2 variables

```r
# split the data
split <- sample.split(CPSnew$Graduation.Rate.., SplitRatio = 0.7)

train <- subset(CPSnew, split == "TRUE")
test <- subset(CPSnew, split == "FALSE")
```

Figure A4: Splitting the data

```r
train.model = lm(Graduation.Rate..~College.Eligibility..+Average.Student.Attendance+
                   Rate.of.Misconducts..per.100.students., data = train)
# predicting
predreg = predict(train.model, newdata = test)
```

Figure A5: Stepwise model fit on training and predicting on test

```r
tree.rf <- randomForest(Graduation.Rate..~., mtry = 4, ntree = 400, data = train, importance = TRUE)
pred = predict(tree.rf, newdata = test)
```

Figure A6: Random Forest fit on training and predicted on test

```
Call:
lm(formula = Graduation.Rate.. ~ College.Eligibility.. + Average.Student.Attendance +
    Rate.of.Misconducts..per.100.students., data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-21.3382  -6.4856   0.1334   5.1154  31.4969

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                         16.46108   19.93635   0.826 0.412987
College.Eligibility..                0.30494    0.08542   3.570 0.000811 ***
Average.Student.Attendance           0.54752    0.24296   2.254 0.028736 *
Rate.of.Misconducts..per.100.students. -0.31868    0.12168  -2.619 0.011705 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.24 on 49 degrees of freedom
Multiple R-squared:  0.6474,    Adjusted R-squared:  0.6258
F-statistic: 29.99 on 3 and 49 DF,  p-value: 3.741e-11
```

Figure A7: Regression model coefficients

```
> # MSE regression
> mean((predreg - test$Graduation.Rate..)^2)
[1] 84.8686
> # MSE Random Forest
> mean((pred-test$Graduation.Rate..)^2)
[1] 127.138
```

Figure A8: Calculating MSE for both regression model and random forest

## Appendix B

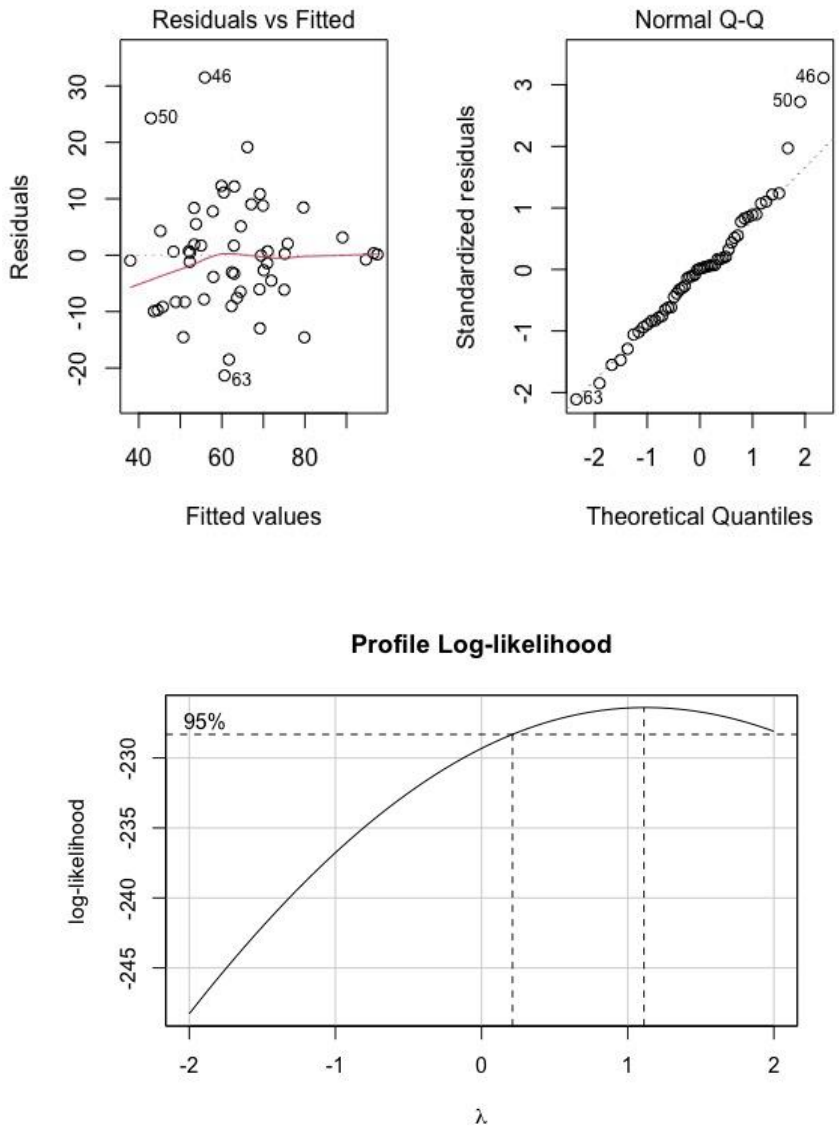



Figure B1: Assumptions and BoxCox for multiple linear regression model

```
> anova(train.model)
Analysis of Variance Table

Response: Graduation.Rate..
                                    Df Sum Sq Mean Sq F value    Pr(>F)
College.Eligibility..                1 7455.4  7455.4  71.092 4.212e-11 ***
Average.Student.Attendance           1 1259.1  1259.1  12.007  0.001112 **
Rate.of.Misconducts..per.100.students.  1  719.3   719.3   6.859  0.011705 *
Residuals                           49 5138.6   104.9
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table B3: ANOVA table for the stepwise reduced multiple linear regression model