KMVC_Capstone_Final_Report

KMVC

January 21, 2019

Table of Contents

Table of Figures

3 - KMVC_Capstone_Final_Report

Introduction

This report provides a thorough summary of my Capstone Project: an introduction to the problem the project examined, a look at the data set(s) utilized as well as the kinds of data wrangling that was required, and the findings from preliminary data exploration and data visualisations. This report also provides the results of training random forest and gradient boosting regression models and recommendations based on the project's results.

The Problem

It is hard to understate the value accorded by many to primary and secondary education. Providing access to good education, having "good" schoools, is one indicator by which some measure the overall functionality of a state.

This has raised the question: "what is a good school?" The answers to this are myriad. Is the purpose of school to facilitate students' ability to achieve high academic grades? Is the purpose of school to help raise children, to impart social values and teach a child how to be a citizen? Where many educational systems, at least in North America seems to have landed, when it comes to evaluating schools, are measures that involve both the environmental and social factors surrounding and cultivated by the school, as well as students' academic outcomes.

Chicago Public Schools (CPS), the third largest school district in the United States, seems to have adopted the approach outlined above. In their annual progress report cards, schools within CPS' district are scored on factors such as school safety and health, but the progress report cads also include each school's percentile rank for each grade in math and reading, so one can see how a school is faring academically (or in these subjects at least) compared with the national average.

The information collected for these progress reports are part of a initiative associated with UChicago Impact, a non-profit "dedicated to creating an evidence-based profession for K-12 education"(https://uchicagoimpact.org/about). In gathering data on Chicago Public Schools using surveys filled out by full-time faculty and students from grade six to grade twelve, Uchicago Impact employs the 5Essentials paradigm. The 5Essentials are: Effective Leaders, Collaborative Teachers, Involved Families, Supportive Environment, and Ambitious Instruction (https://uchicagoimpact.org/our-offerings/5essentials). The progress report cards therefore also provide every school a score for each of these 5Essentials. The schools are scored on these Essentials from a range of five possible scores: "very weak", "weak", "neutral", "strong" and "very strong."

Importantly, UChicago Impact states that "Researchers have found that schools strong on at least three of the five essentials were 10 times more likely to show

substantial gains in student learning over time than schools weak on three or more of the five essentials. Research also shows that a persistently low score in even just one of the five essentials reduced the likelihood of improvement to less than 10 percent"(https://uchicagoimpact.org/our-offerings/5essentials). In reading the research upon which these statements are based, it seems as though Uchicago Impact is adopting a long-term perspective - when providing evidence for the effectiveness of the 5Essentials paradigm , they cite schools' performance over a ten year period.

As the progress reports have been built with a prediction in mind (strength in 3 or more of the 5Essentials leads to improved learning outcomes over time), my project 's problem was to take on the task of assessing the strength of the predicted relationship between learning and the 5Essentials. Or to put it another way, my project took on the task of assessing the extent to which social/environmental factors such as family involvement, instruction, environment, teacher collaboration and effective leadership can be used to predict academic outcomes.

The hope at the start of the project, was that if academic performance can be reliably predicted, schools predicted to have poorer results in the future could be given more consideration, resources and attention in the present day, thereby (hopefully) thwarting the poor predicted outcomes. The question I also hoped to explore was whether the 5Essentials was sufficient, or if other demographic information was necessary to predict academic outcomes.

The Data Sets

Three main data sets were employed for this project: progress reports for CPS schools from 2014, 2015 and 2016. These progress reports/data sets can be found here;

2014 progress report: https://data.cityofchicago.org/Education/Chicago-Public-Schools-Elementary-School-Progress-/tj8h-mnuv

2015 progress report: https://data.cityofchicago.org/Education/Chicago-Public-Schools-School-Progress-Reports-SY1/fvrx-esxp

2016 progress report: https://data.cityofchicago.org/Education/Chicago-Public-Schools-School-Progress-Reports-SY1/cp7s-7gxg

All data sets were available in .csv format.

What Important Fields and Information Do The Data Sets Contain?

The data sets contain data on over 400 public schools in Chicago (elementary, middle and high schools). They also contain the 5Essential scores on Leadership, Collabortive Teaching, Ambitious Instruction, Involved Families, and Supportive Environment (an explanation of each of the 5Essentials can be found on pg.6 of this report:https://consortium.uchicago.edu/sites/default/files/publications/Statewide%205E%20Report.pdf). Percentile ranks for schools by grade in reading and math, as well as the locations of the schools, safety ratings, health ratings and student attendance numbers are also included in these progress reports.

What Are The Data Sets' Limitations?

There are a number of limitations with the data sets. Some information about the schools was missing, information that could be helpful, such as the ratio of students to teachers and rates of teacher turn-over. These kinds of statistics would provide some insight into the learning environment of each school, and might provide further context for academic outcomes.

As well, one could argue that there is a paucity of demographic data in the data sets. Beyond each school's location, there is no information in the data sets about the gender or race of the students, nor the economic circumstances that the students are living in, the employment status of their parents, the education levels of their parents, or the languages spoken by the students. This kind of demographic information seems like it would be useful for contextualizing (and/or predicting) the schools's academic outcomes and their 5Essential scores.

However, the biggest limitation of these data sets is linked to politics. The kinds of information collected by the surveys that form the backbone of these data sets was altered in the early 2010s because of changes to education policy. So though I found a data set from 2012, I could not use it with the data sets from 2014, 2015 and 2016, because the metrics collected for 2012 were different from those collected for 2014, 2015 and 2016.

In 2012, the metric for academic performance had to do with the schools meeting state requirements for a particular subject. But from 2014-2016, the metrics for academic performance are the schools' percentile attainment in reading or math compared with the national average.

This is problematic because to really assess the value of the 5Essentials paradigm according to Uchicago Impact's terms (which were discussed above), one would need to examine schools' performances over at least a ten year period. These alterations in policy, and therefore metrics, make evaluating the 5Essentials approach over a ten year period not possible.

I made the decision to simply use the data sets from 2014, 2015 and 2016, but this means that my assessment of the value of the 5Essentials paradigm cannot be considered fulsome, as data over a longer period would be required for that. As outcome variables, reading and math percentiles from 2016 were chosen. However, I decided to primarily use the data from the 2014 data set, to see if 5Essential scores from 2014 could help predict academic outcomes a few years later (in 2016), as the 5Essential scores are meant to work over time (strong scores in a particular year is supposed to lead to higher academic outcomes years later.)

Data Wrangling

What Kind of Wrangling Was Done?

The major steps taken during the data wrangling process involved tidying up the data, pulling demographic data into the data sets, and joining the three original data sets together to form one new, cleaned-up data set.

One of the first steps in the data wrangling process was to eliminate the middle schools and high schools from the 2015 and 2016 datasets. The main reason for eliminating the middle schools was the scarcity of these schools within the data sets(only 10 middle schools in the 2015 data set, and 12 in 2016). As for high schools, there are more of these within the data, with 188 high schools in the 2015 set and 181 high schools in 2016.

However, as the 2014 data sets contains no data on middle schools or high schools, and all three data sets contain much more data on elementary schools (483 elementary schools in 2014, 472 in 2015, and 470 in 2016), it seemed that the elementary schools would provide a richer source from which to create a predictive algorithm and to mine patterns. The filter function from dplyr was employed to remove non-elementary schools from the 2015 and 2016 datasets.

```
#eliminating middle and elementary schools from the 2015 and 2016 data
sets
schools2015 <-  filter(schools2015, Primary_Category == "ES")
schools2016 <-  filter(schools2016, Primary_Category == "ES")
```

The final data set for this project is focused on data from 2014, only taking from the 2015 and 2016 data sets the percentile scores for schools on grade 3 math and grade 3 reading. The rationale for this choice was discussed in greater detail earlier in this report. The 2015 and 2016 data sets were individually culled of many of their variables, and then joined together, using the inner_join function from dplyr. I chose to use inner_join because it would ensure that all observations from 2015 found in 2016 would be returned; having previously used set_diff, I knew that there were no observations in 2016 that were not in 2015, but that there were 2 schools in 2015 not in 2016 which would be dropped from the newly created data set.

Some relevant socio-economic demographic information was located on the American Fact Finder site (https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml). A data set was created from this information, which contains, by zip code, the median income for each zip code, as well as the percentage of people in that zip code who have graduated from high school. This data set was joined to the joined-together 2015/2016 data set.

As well, I employed the tidycensus package, which lets it users "interface with the US Census Bureau's decennial Census and five-year American Community APIs and return tidyverse-ready data frames" (https://walkerke.github.io/tidycensus) to pull demographic data on race, employment, age, gender, language fluency, home ownership, earnings and poverty from the American Community Survey.

```
#getting total population by zip code from the ACS survey
total_pop <- get_acs(geography = "zcta", variables =  "B00001_001")
Getting data from the 2012-2016 5-year ACS
```

However, as the margins of error for this survey are quite large, a fair number of the variables that were pulled from the survey ended up being un-usable.

The last major step taken, after tidying the 2014 dataset, was using inner_join to bring together the now-tidied 2014 dataset to the 2015/2016 dataset.The 2014 dataset had 21 schools that differed from the 2015/2016 dataset, and those were removed from the dataset by the use of the inner_join function, leaving the new dataset with 462 observations.

```
#join the datasets together
schools14_16 <- inner_join(schools2014, schools201516, by =
c("School.ID" = "School_ID"))
```

Exploratory Data Analysis

Here are some statistics about the data set, as well as some statistics about the set gathered from using the summary function the data set as a whole:

- Total Number of Schools: 462
- Median for Median Income : $37,574 (USD)
- Mean Median Income: $45,443 (USD)
- Median High School Graduation Rate in Schools' Zip Codes: 83.10%
- Mean High School Graduation Rate in Schools' Zip Codes: 80.18%
- Mean Percentile for 2016 Reading: 48.94
- Mean Percentile for 2016 Math : 47.43

As can be seen, on average, CPS schools were slightly below the national average when it came to grade 3 reading and grade 3 math in 2016.

In terms of preliminary exploration, I approached the data set in three ways: I explored the data set as a whole, and then subsetted the data set, first by 5Essentials and then by percentile ranks.

The Data Set as a Whole

Plotting each of the 5Essentials scores (Family.Rating, Collaboration.Rating, Leadership.Rating, Instruction.Rating and Environment.Rating) against the outcome variables (using ggplot) did not indicate strong predictive connections between the 5Essentials and academic performance. The bulk of the Essentials only seemed to show a bit of a correlation at the extreme of the data - schools very weak in Essentials such as Ambitious Instruction, Collaborative Teachers and Effective Leaders seemed to also have poorer outcomes in math and reading, and vice versa for schools scoring very strong in the Essentials. Particularly for the Family Rating, schools scoring "very weak" for that Essential seemed to have poorer outcomes in math and reading.

When plotted against the outcome variables, median income and high school graduation rates had some correlation with the outcome variables (median income had stronger correlations than high school graduation), while percentiles ranks from 2014 and 2015 had the strongest correlations with the outcome variables (which are from 2016). These findings on correlations between the variables were re-confirmed by using the corrplot function on the data set (from the corrplot package).

Throughout the data set, the 5Essential to be cited most frequently as strong or very strong was Instruction, followed by Family, then Environment, then Collaboration and in last place, Leadership. In terms of being cited as weak or very weak, Collaboration was most frequently scored as weak or very weak, followed by Leadership, then Family, then Environment. Instruction was the Essential least commonly scored as weak or very weak in the dataset.

Sub-setting the Data by 5Essentials Scores

I also subsetted the data so that I could examine schools that scored very strong in all 5Essentials, strong or very strong in all 5Essentials, strong in all 5Essentials, neutral in all 5Essentials, weak in all 5Essentials and very weak in all Essential categories. I thought this might provide a clearer way of seeing the impact of the 5Essentials on math and reading percentiles. Zero schools scored very weak in all categories.

This subset seemed to show some correlations between the 5Essentials and outcome variables. In examining the subsets from very strong to weak, one could see in most cases, that as 5Essential scores went down, the mean and median percentiles for reading and math decreased (with the odd exception of math, where the neutral subset had a higher mean percentile than the subset that contained schools that had scored strong in every essential), median/mean income decreased, and mean/median high school graduation rates decreased as well. The schools scoring very strong or strong or neutral for all 5Essentials fell within a higher income bracket than the schools scoring weak for all 5Essentials.

Sub-setting the Data by Percentiles

For the next phase of statistical analysis, I decided to subset the data by schools that scored over the 50th percentile in math and reading. I therefore created two subsets, one for above average math percentiles and the other for above average reading percentiles. 204 schools (out of 462) scored above the 50th percentile for math and 209 schools scored above the 50th percentile for reading.

Scoring above average in math or reading seemed to correlate with scoring above average in the other subject: schools that scored above the 50th percentile in math, had a mean reading percentile of 72.66, while schools that scored above the 50th percentile in reading had a mean math percentile of 71.13. Most interesting was seeing which of the 5Essentials was cited most frequently as strong/very strong for these above average schools. For above average reading, Family Scores were the most frequent, followed by Instruction, and then Collaboration, Environment and Leadership. For above average math, Family also came first, then Instruction, Collaboration and Leadership were tied for third with Environment coming last. The data therefore suggests that strong/very strong Family Involvement and Ambitious Instruction have some correlation with scoring above the 50th percentile in math and reading.

In short, what can be seen from this statistical analysis is that:

- Family and Instruction Rating are the two Essentials most strongly correlated with achieving above the 50th percentile in both math and reading
- While there are some oddities in the data, scoring weak or very weak on multiple Essentials seems to be slightly correlated with scoring a lower percentile rank in math and reading.
- Scoring strong/very strong /neutral seems to lead to high percentile attainment in reading and math as compared with scoring weak in all the Essentials
- Schools scoring very strong/strong/neutral in all 5Essentials fall into a higher income bracket than schools scoring weak on all 5Essentials.
- However, overall, correlation between the 5Essentials and academic outcomes seems to be quite low

The point about schools scoring very strong/strong/neutral falling into a higher income bracket suggests that economic circumstances might be linked to a school's ability to provide a supportive environment, ambitious instruction, to have a high level of family involvement, to have engaged and collaborative teachers and to have effective leadership.

Data Visualizations

In this section, some visualizations created through the process of conducting Exploratory Data Analysis (discussed in the previous section) are presented.

Here is a plot of median income by zip code and reading 2016 (an outcome variable), that shows a somewhat linear correlation between income and the outcome variable:

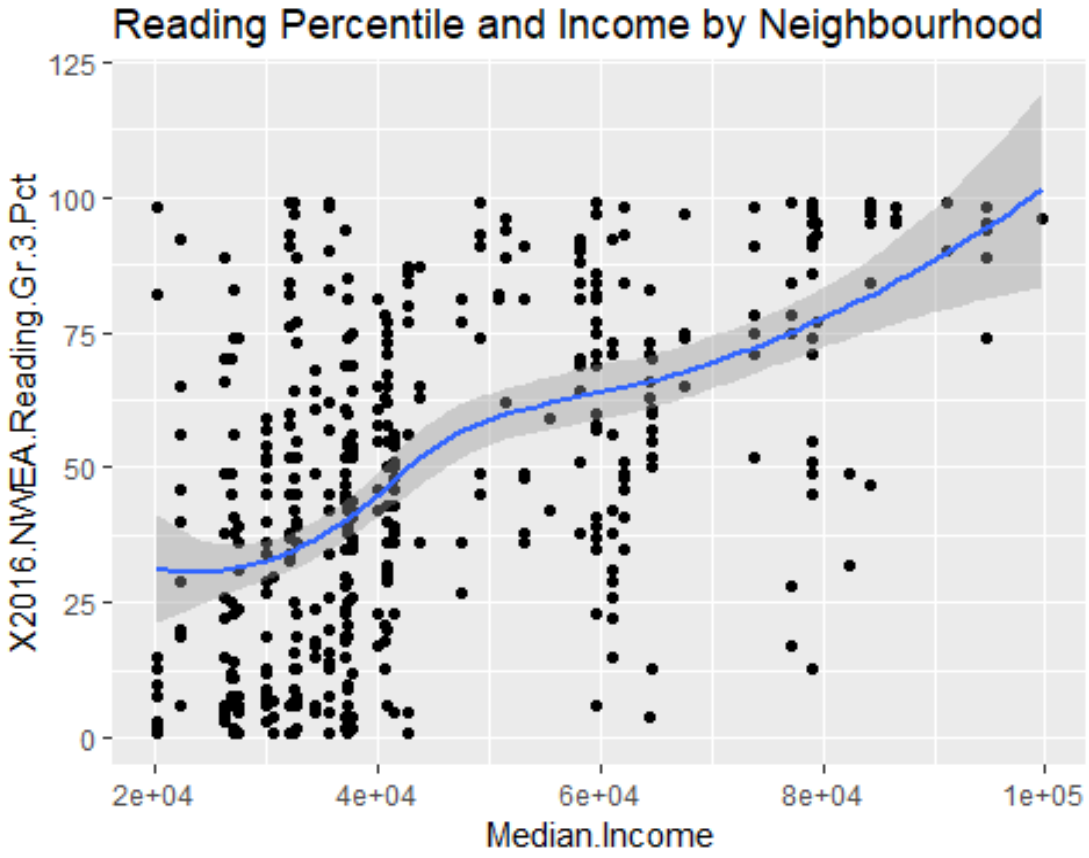

Figure 1: Scatter plot of Median Income by Zip Code and Reading 2016 Percentiles

Here is a plot of median income by zip code and math 2016 (an outcome variable)that shows a somewhat linear correlation between income and the outcome variable:

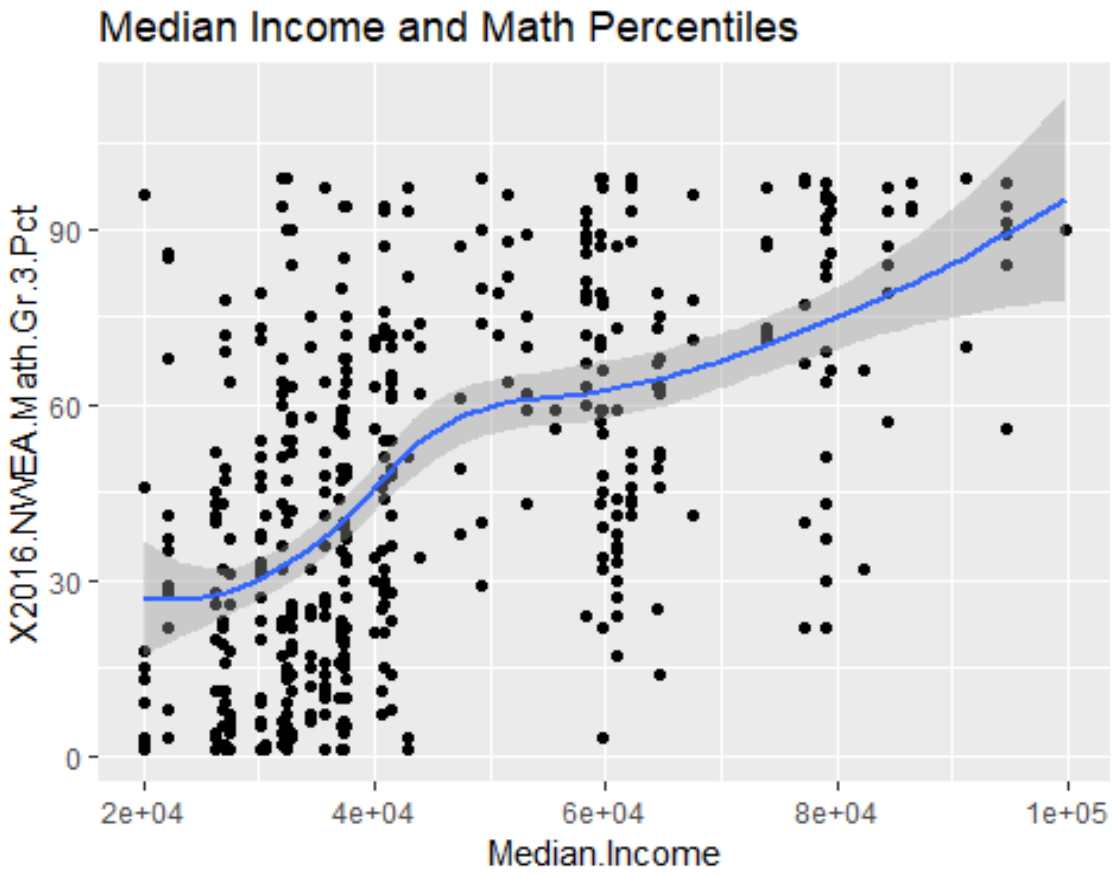

Figure 2: Scatter Plot of Median Income by Zip Code and Math 2016 Percentiles

Here is a correlation plot, showing the levels of correlations between the variables in the data set (the darker the colour and the bigger the circle, the higher the level of correlation). One can see that for the outcome variables (2016 Reading and 2016 Math), the strongest correlated variables are percentiles scores from previous years, with income, safety ratings and the family score having weaker correlations.

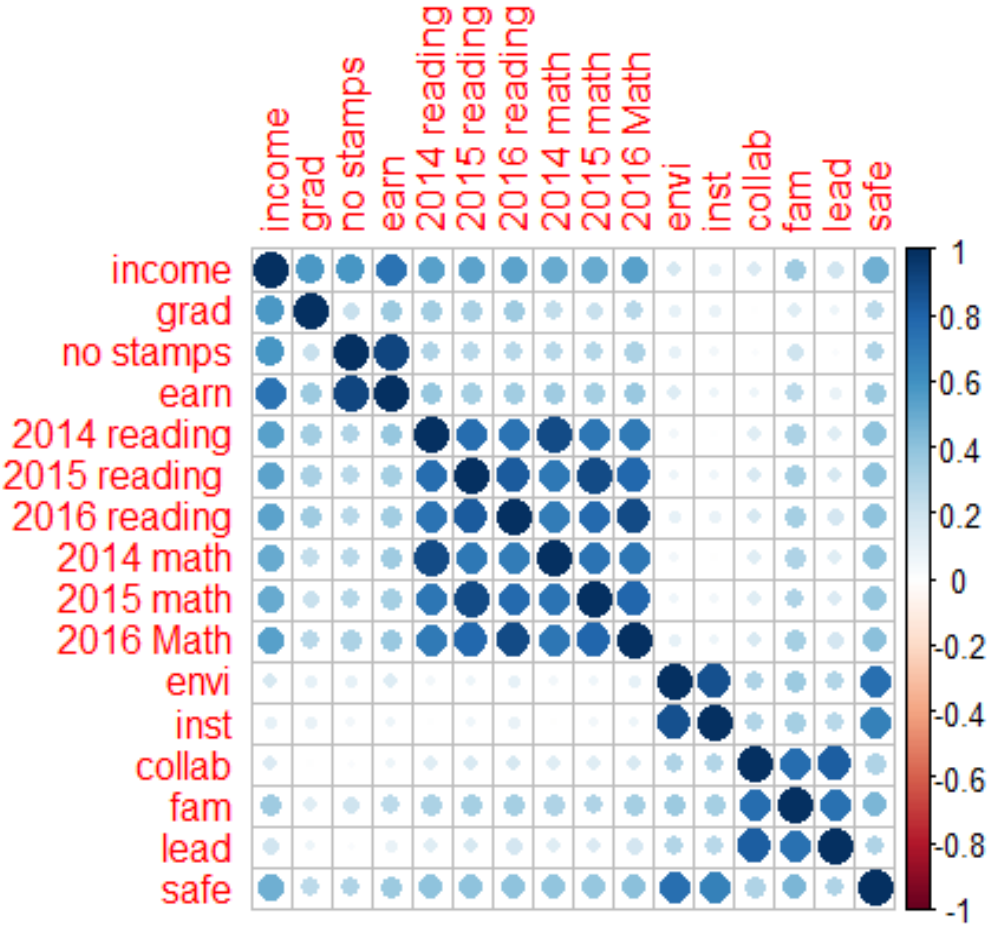

Figure 3: A Correlation Colour Plot of the Data Set

Here is a plot showing schools that scored “strong” on all 5Essentials by income, safety their 2016 reading percentiles:

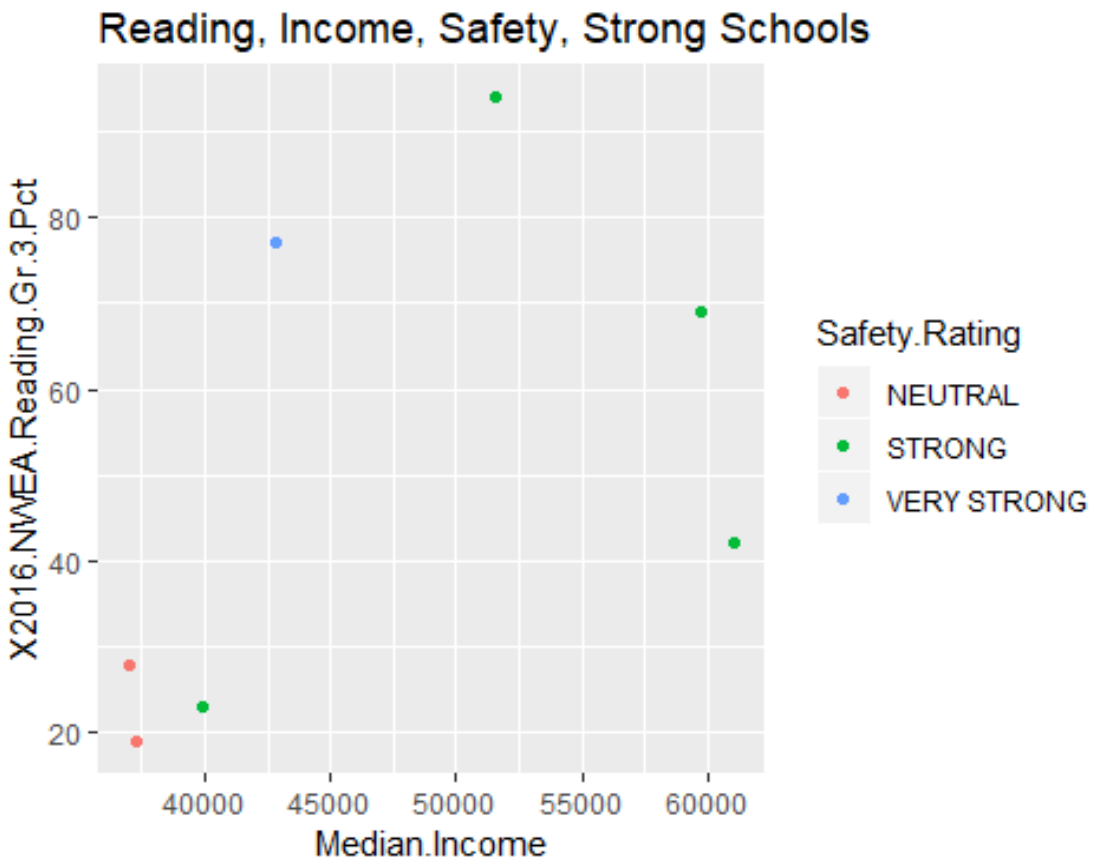

Figure 4: A Scatter Plot of Median Income and Reading 2016 Percentiles for Strong Schools, with Safety Rating

Here is a plot showing schools that scored “weak” on all 5Essentials by their income, safety rating and their 2016 reading percentiles. It can be seen that compared with the previous plot on strong schools, that incomes fall into a lower bracket here and safety ratings are also tending to be weaker:

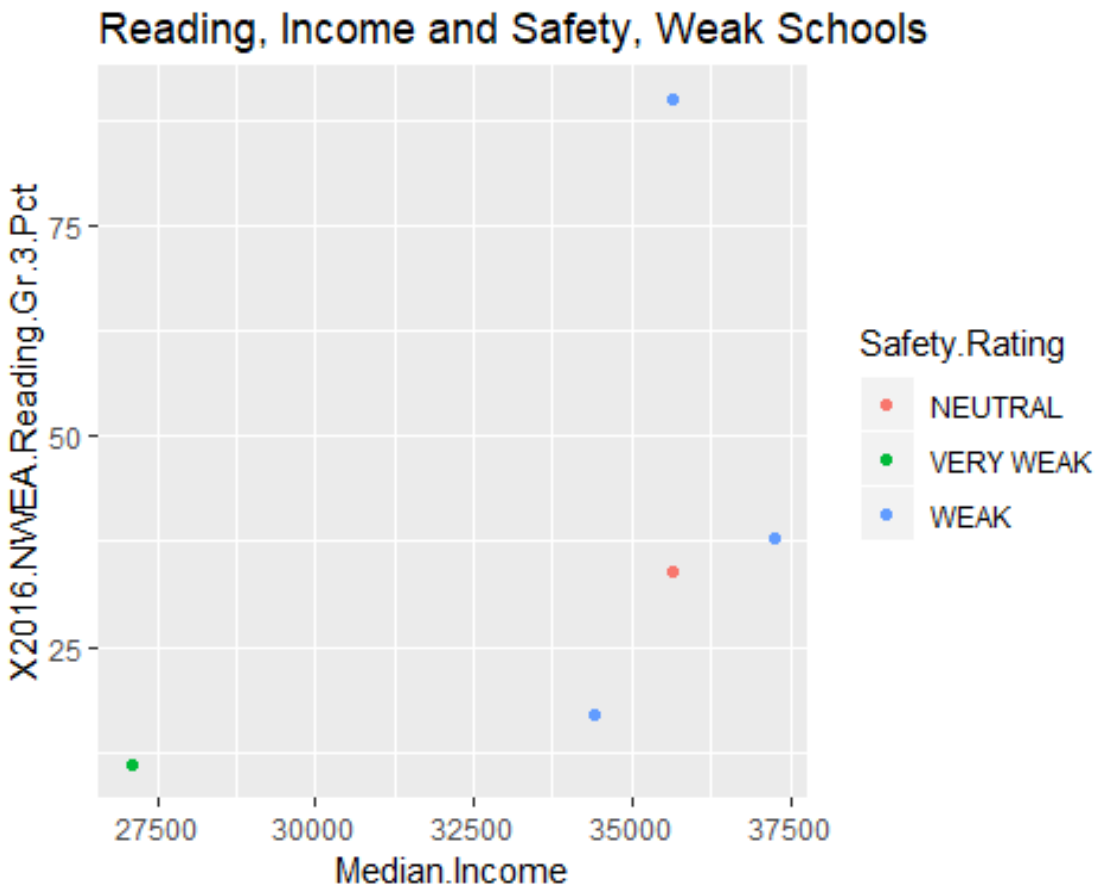

Figure 5: A Scatter Plot of Median Income and Reading 2016 Percentiles, with Safety Rating For Weak Schools

Here is a plot showing schools that scored “strong” on all 5Essentials by income, safety their 2016 math percentiles:

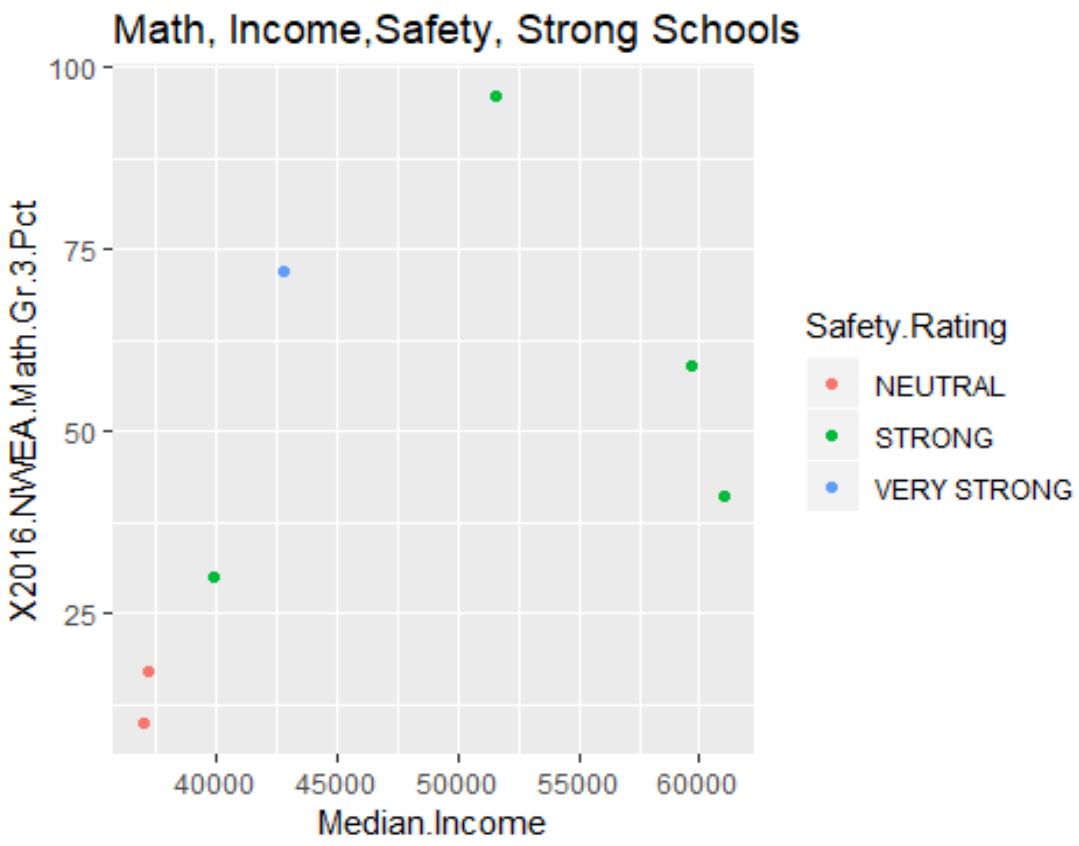

Figure 6: A Scatter Plot of Median Income and Math 2016 Percentiles, with Safety Rating For Strong Schools

Here is a plot showing schools that scored “weak” on all 5Essentials by their income, safety rating and their 2016 math percentiles. It can be seen that compared with the plot above on strong schools, that incomes fall into a lower bracket here and safety ratings are also tending to be weaker. The percentiles scores are also lower compared with percentiles attained by the strong schools above:

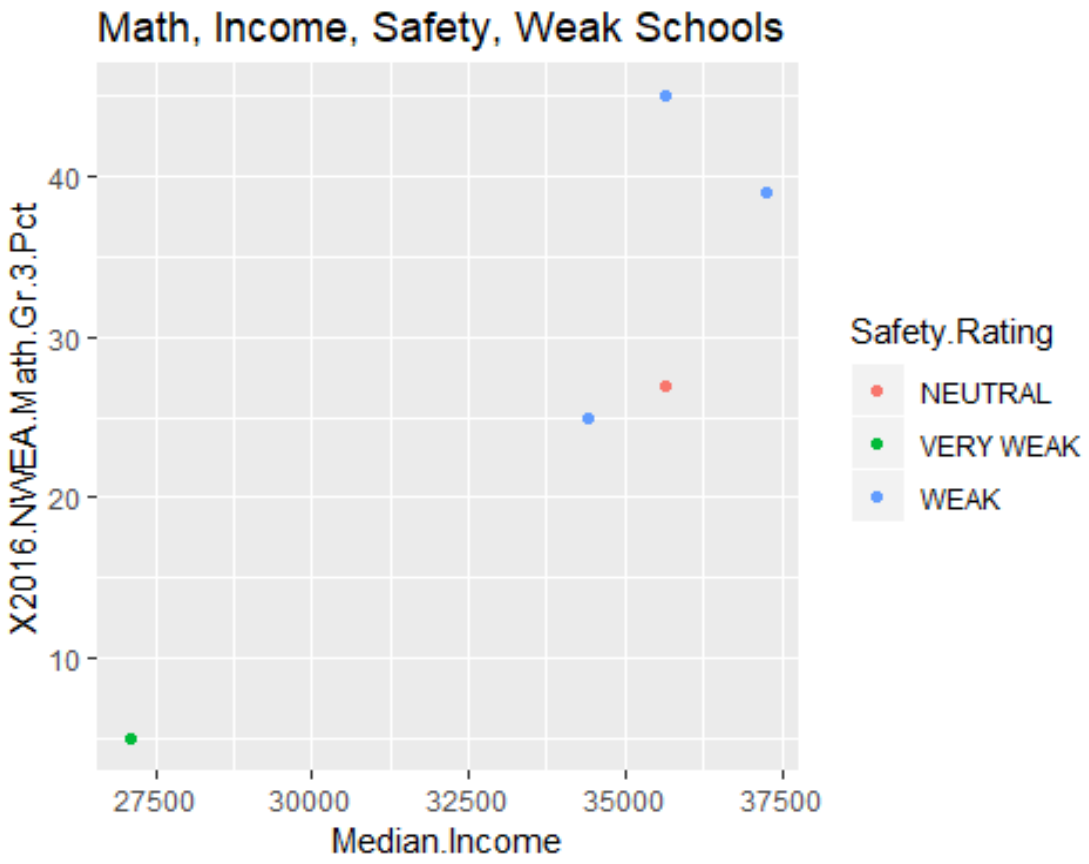

Figure 7: A Scatter Plot of Median Income and Math 2016 Percentiles, with Safety Rating For Weak Schools

Here is a plot showing the frequency which each of the 5Essentials is scored Strong/Very Strong for schools that achieved above the 50th percentile for math 2016. As can be seen below, Involved Families is most frequently scored strong/very strong:

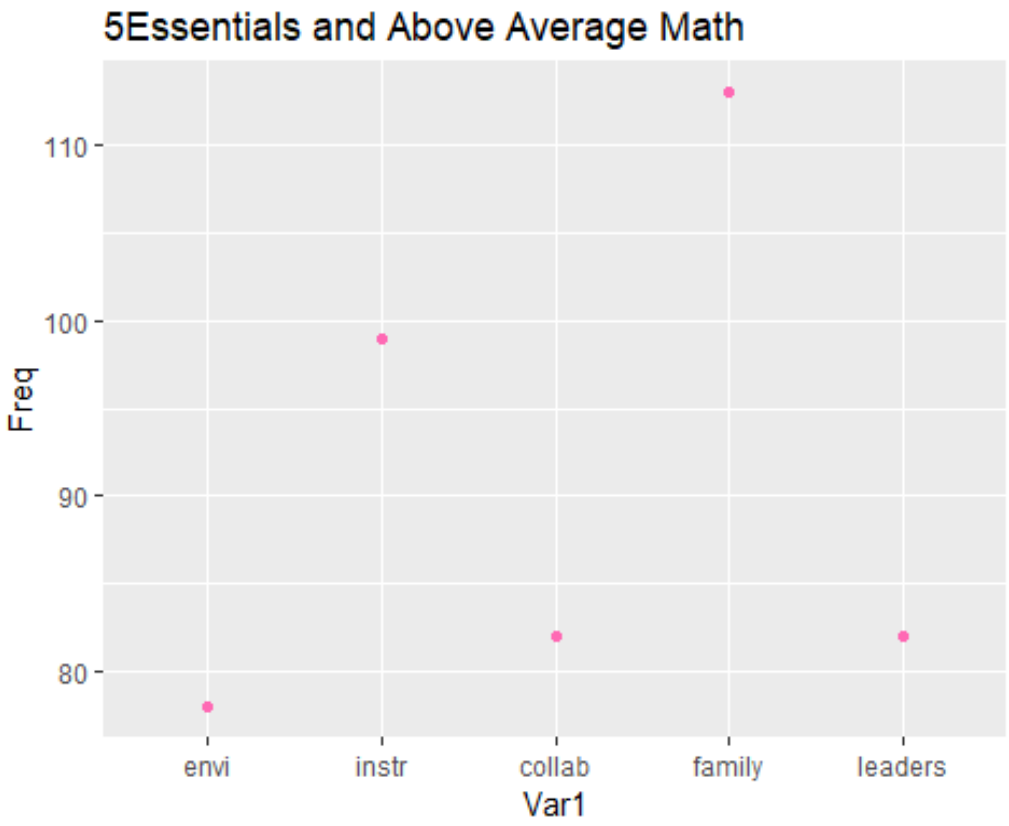

Figure 8: A Scatter Plot Showing the Frequency a 5Essential was Scored Strong/Very Strong for Schools Scoring for schools scoring above the 50th Percentile in Math

Here is a plot showing the frequency which each of the 5Essentials is scored Strong/Very Strong for schools that achieved above the 50th percentile for reading 2016. As can be seen below, Involved Families is most frequently scored strong/very strong :

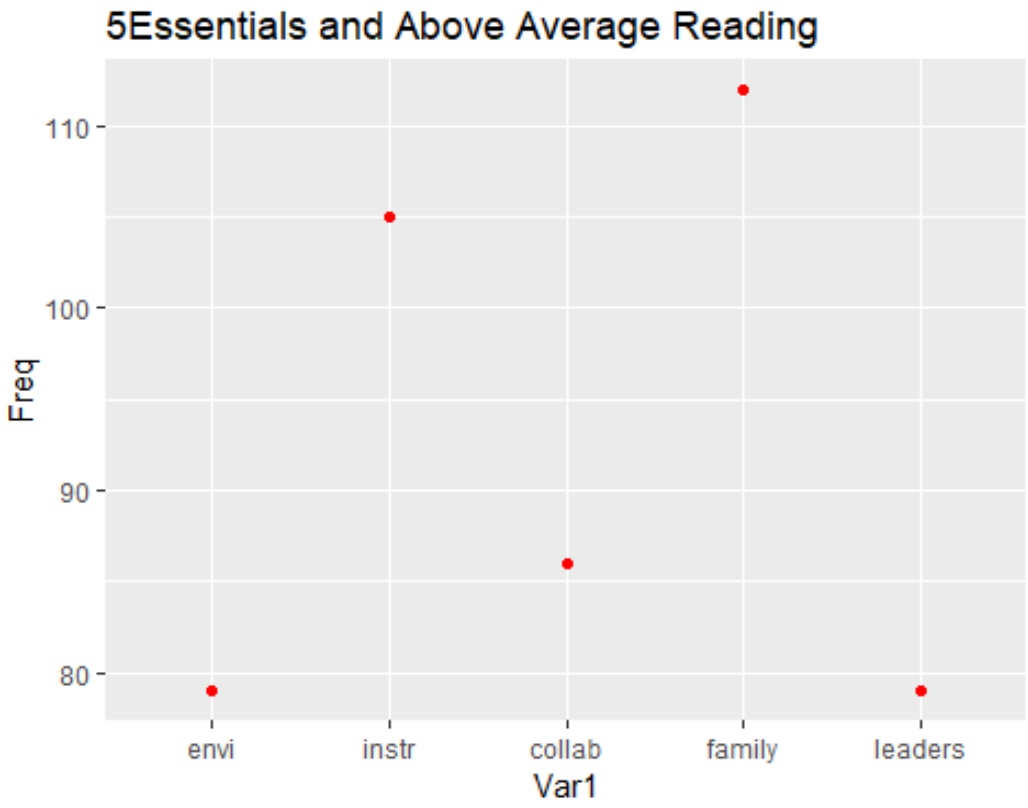

Figure 9: A Scatter Plot Showing the Frequency a 5Essential was Scored Strong/Very Strong for Schools Scoring for schools scoring above the 50th Percentile in Reading

Data Modelling

Main Features

The main features of my project's machine-learning problem were, as discussed above the 2016 reading percentiles for grade 3, as well as the 2016 math percentiles for grade 3. As there were two outcome variables, multiple models had to be trained.

Machine Learning Techniques

The decision was made to use both random forest and gradient boosting to create predictive models for grade 3 reading and grade 3 math. Random Forest trains an ensemble of trees, and the technique is known for reducing over-fitting because of an added element of randomness (at splits). Gradient boosting typically trains trees sequentially, using the errors from previous trees to inform the way that future trees are trained. Random Forest and Gradient Boosting can both be used on classification and regression problems.

Building the Models

The data was split into a training set (80% of the data) and a test set (20%). The training set is what is used to train the model, which is then tested against the test set.

```
#creating the training and test sets
# Number of rows for the training set (80% of the dataset)
n_train <- round(0.80 * n)

# Create a vector of indices which is an 80% random sample
set.seed(123)
train_indices <- sample(1:n, n_train)


# Subset the schools data frame to training indices only
schoolsmath_train <- schools14_16linmath[train_indices, ]


# Exclude the training indices to create the test set
schoolsmath_test <- schools14_16linmath[-train_indices, ]
```

Gradient Boosting Regression Model(s)

I began by creating the gradient boosting models (gbm) using the gbm package. I started with training 10,000 trees. I built gbm models for both math and reading. I'll show code from the reading model here, and code from the math model in the section on Random Forest. Important note: the findings for math and reading were exceedingly similar, for both gbm and random forest.

```
#create a gbm model using 10,000 trees
set.seed(1)

schoolsreadgb_model <- gbm(formula = r_2016 ~ .,

                           distribution = "gaussian",

                           data = schoolsread_train,

                           n.trees = 10000)
```

I printed the model

```
summary(schoolsreadgb_model)
```

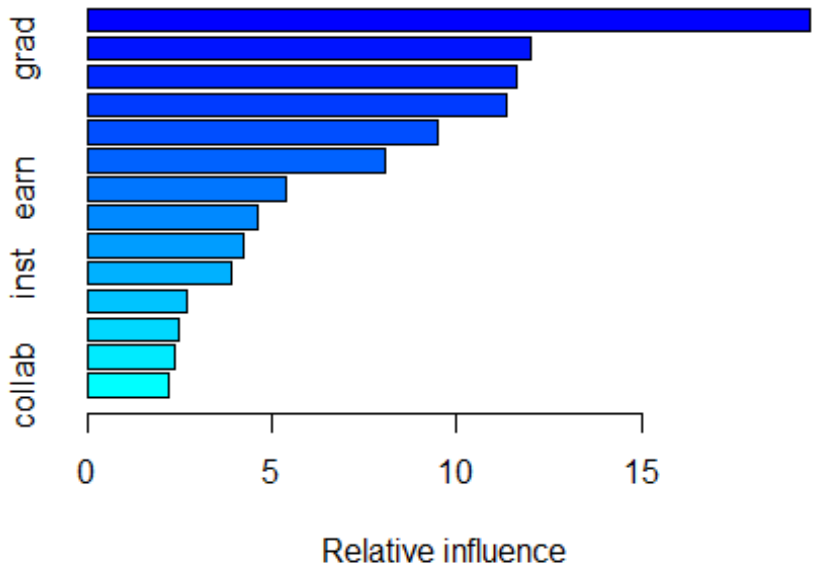

Figure 10: Bar graph showing the relative influence of variables on the gradient boosting model

```
##                  var   rel.inf
## r_2015.     r_2015. 19.564880
## grad           grad 12.012089
## m_2014       m_2014 11.627713
## r_2014       r_2014 11.356039
## m_2015       m_2015  9.455650
## income       income  8.066805
## earn           earn  5.403355
## no_stamps no_stamps  4.607552
## envi           envi  4.234968
## inst           inst  3.894889
## fam             fam  2.713096
## lead           lead  2.461236
## safe           safe  2.391178
## collab       collab  2.210551
```

I then used the gbm model to generate predictions on the test set:

```
pred <- predict.gbm(object = schoolsreadgb_model,
                    newdata = schoolsread_test,
                    n.trees = 10000,
                    type = "response")
```

I then evaluated the RMSE (root mean squared error - the lower, the better).

```
print(rmse10000_gbm)
[1] 22.02576
```

I wrote code to calculate the optimal trees, using both OOB (out of bag error) and CV (cross validation error).

I calculated the optimal trees using OOB error first:

```
# Optimal ntree estimate based on OOB
ntree_opt_oob <- gbm.perf(object = schoolsreadgb_model,
                          method = "OOB",
                          oobag.curve = TRUE)
```

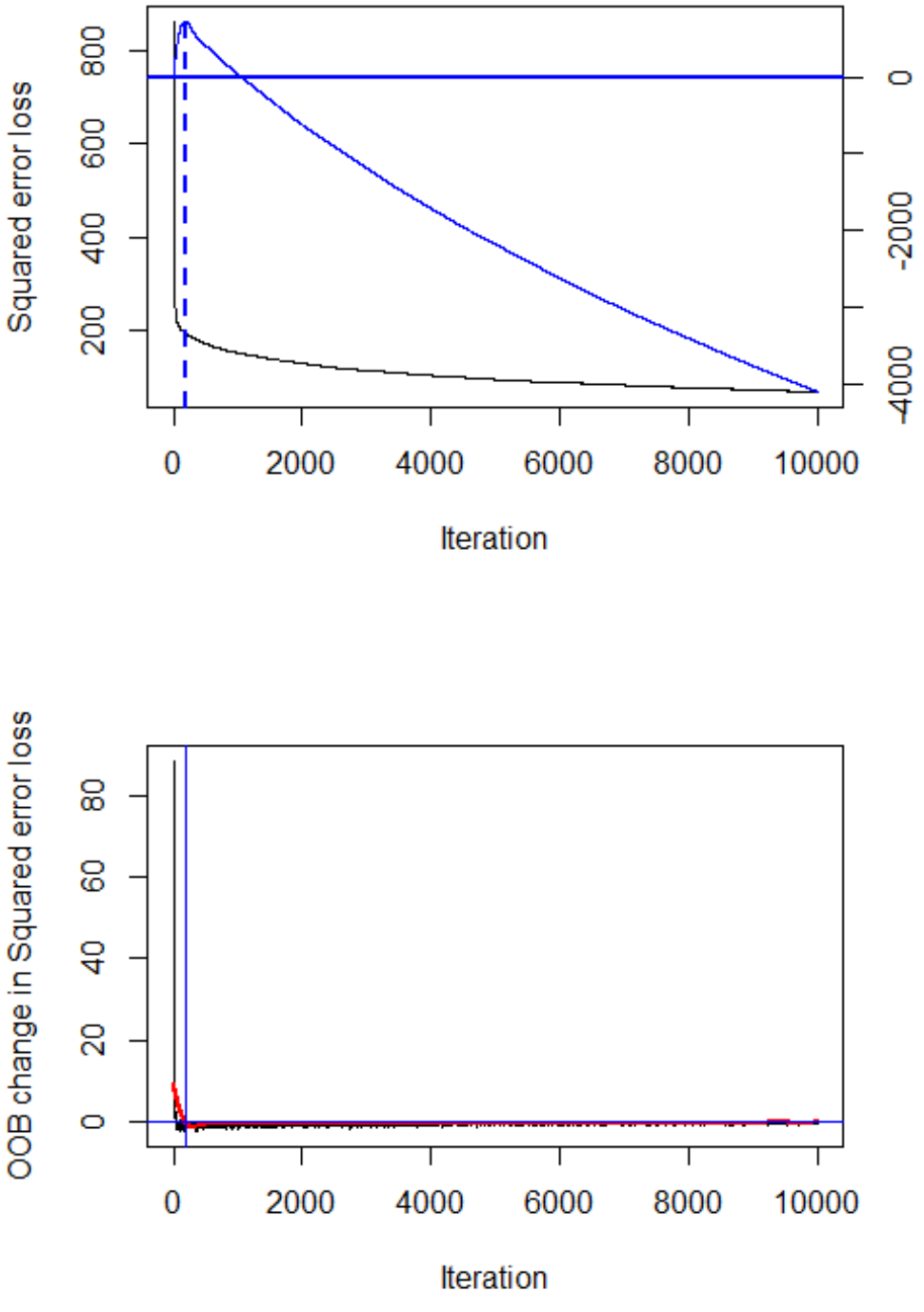

Figure 11: Plot of OOB Error for the Gradient Boosting Model

Then I looked at the optimal amount of trees based on CV error:

```
# Train a CV GBM model
set.seed(1)

schoolsreadgb_model_cv <- gbm(formula = r_2016 ~ .,
                              distribution = "gaussian",
                              data = schoolsread_train,
                              n.trees = 10000,
                              cv.folds = 2)


# Optimal ntree estimate based on CV

ntree_opt_cv <- gbm.perf(object = schoolsreadgb_model_cv,
                         method = "cv")
```

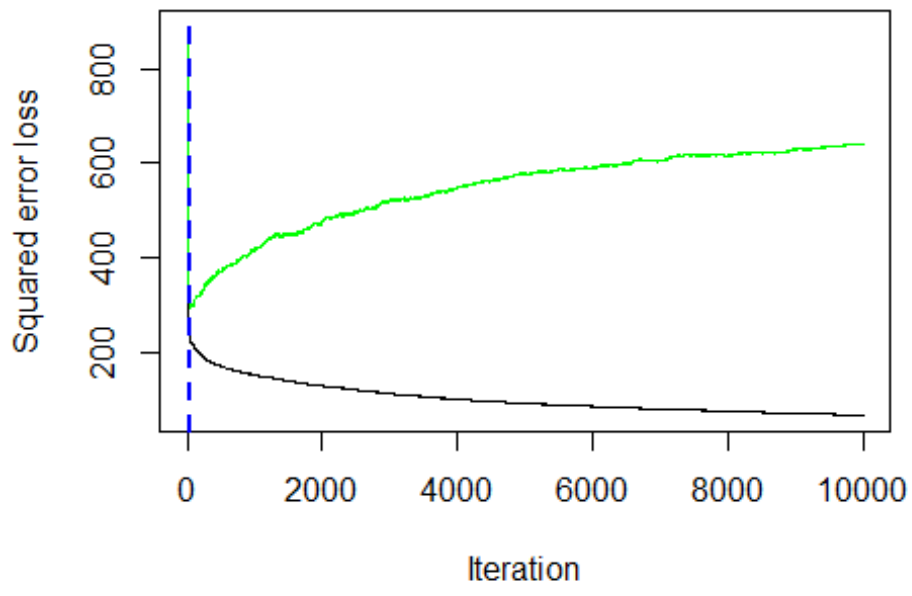

Figure 12: Plot Showing CV Error for the Gradient Boosting Model

I compared the two estimates of optimal n.trees:

```
# Compare the estimates
print(paste0("Optimal n.trees (OOB Estimate): ", ntree_opt_oob))

## [1] "Optimal n.trees (OOB Estimate): 180"

print(paste0("Optimal n.trees (CV Estimate): ", ntree_opt_cv))

## [1] "Optimal n.trees (CV Estimate): 51"
```

I then generated predictions using the optimized ntrees suggested by OOB and CV error:

```
# Use oob
preds1_gbm <- predict.gbm(object = schoolsreadgb_model,

                          newdata = schoolsread_test,

                          n.trees = ntree_opt_oob, type = "response")

#use cv
preds2_gbm <- predict.gbm(object = schoolsreadgb_model,

                          newdata = schoolsread_test,

                          n.trees = ntree_opt_cv, type = "response")
```

Finally, I calculated the RMSE for these 2 models:

```
print(paste0("Optimal n.trees (OOB Estimate): ", ntree_opt_oob))
[1] "Test set RMSE (OOB): 17.7129477263524"

print(paste0("Optimal n.trees (CV Estimate): ", ntree_opt_cv))
[1] "Test set RMSE (CV): 17.3691513680797"
```

These same steps were repeated to build another set of gradient boosting models - however with these models, all percentile variables from the data set were removed except the outcome variable, as I was trying to see what level of accuracy a model purely based on demographic information and the 5Essentials would have. The code

to generate these models is identical to what was provided above, but the RMSEs of these models was substantially higher:

The RMSE of the 10,000 trained trees gbm model for this model was:27.63879

The RMSE of the OOB model was: 27.8236254977981

The RMSE of the CV model was: 28.5695823504538

In short, for both math and reading 2016, with the percentile variables included, **the gradient boosting model that was strongest was the one built on n.trees having been optimized by CV error**. However, for both math and reading gbm models where percentile values had been removed, the gbm model trained on 10,000 trees was the strongest, with the lowest RMSE.

Random Forest

I began by building the Random Froest models, using the randomForest package. I built models for both math and reading. I'll show code from the math model here, as the sample code from the gradient boosting section featured the reading model.

First I trained a model, and then plotted that model:

```
# Train a Random Forest
#set seed for reproducibiilty
set.seed(1)

schoolsmathrf_model <- randomForest(formula = X2016.Math ~ .,
                                    data = schoolsmath_train)

plot(schoolsmathrf_model, main = "Random Forest Model for Schools,
Math, 2016")
```

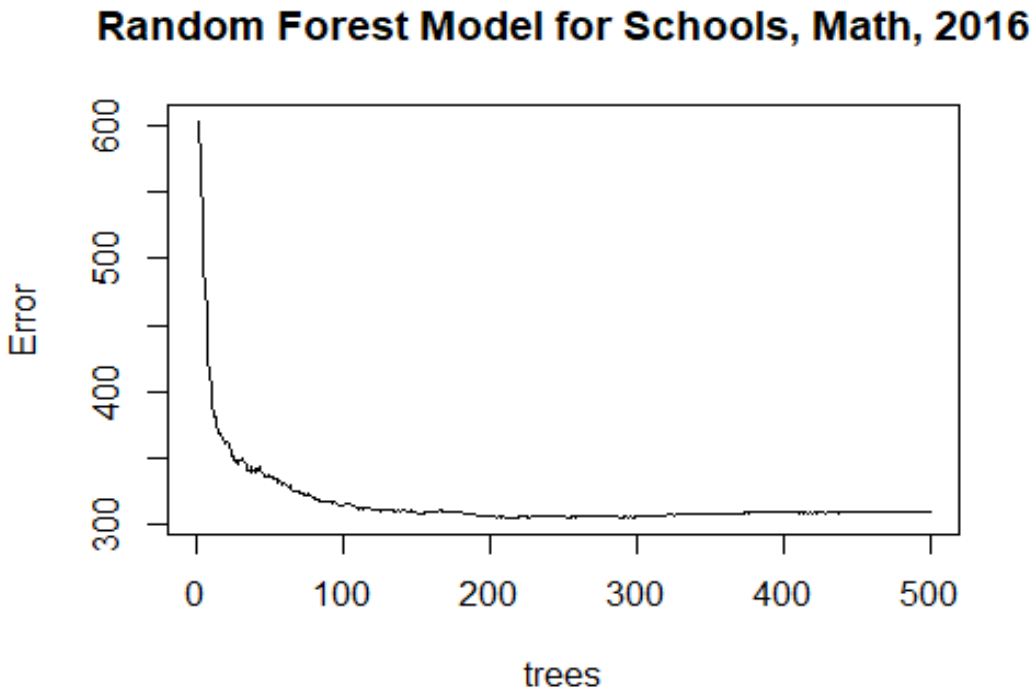

Figure 13: Plot of Random Forest Model

I printed the model:

```
# Print the model output
print(schoolsmathrf_model)

##
## Call:
##  randomForest(formula = X2016.Math ~ ., data = schoolsmath_train)
##                Type of random forest: regression
##                      Number of trees: 500
## No. of variables tried at each split: 4
##
##           Mean of squared residuals: 309.672
##                     % Var explained: 66.21
```

One can see that the level of explained variance is around 66.21, which while not exceedingly exceptional, is at least above 50%.

I generated predictions:

```
# Generate predicted classes using the model object

schoolsmathrf_prediction <- predict(object = schoolsmathrf_model,

                                    newdata = schoolsmath_test,

                                    type = "response")
```

I then grabbed the model RMSE

```
#calculating the RMSE
mathrf_rmse <- rmse(schoolsmath_test$X2016.Math,
schoolsmathrf_prediction)

print(mathrf_rmse)

## [1] 17.16823
```

For the next step, I tuned the RF model using mtry:

```
# Execute the tuning process

set.seed(1)

res_math <- tuneRF(x = subset(schoolsmath_train, select = -X2016.Math),

              y = schoolsmath_train$X2016.Math,

              ntreeTry = 500)
```

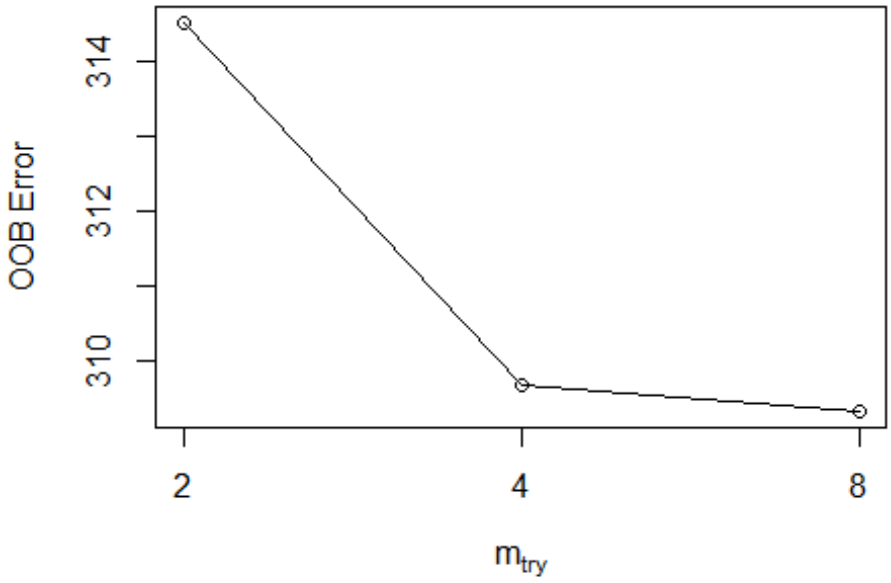

Figure 14: Plot Determining Optimum Mtry based on OOB Error for the Random Forest Model

```
# Look at results

print(res_math)

##   mtry OOBError
## 2    2 314.5058
## 4    4 309.6720
## 8    8 309.3321

# Find the mtry value that minimizes OOB Error

mtry_opt_m <- res_math[,"mtry"][which.min(res_math[,"OOBError"])]
```

I then created a model using the optimized mtry, generated predictions (following the steps outline above to build the first random forest) and then grabbed the RMSE for that model, which was: 17.1467

As with the gradient boosting models, these steps outlined above were repeated to build another set of Random Forest models - however with these models, all

percentile variables from the data set were removed except the outcome variable, trying to see what level of accuracy a model purely based on demographic information and the 5Essentials would have. The code to generate these models is identical to what was provided above, but the RMSEs of these models was substantially higher:

The RMSEs for these models were:

For the intial random forest model (without any percentiles included in the data, save the outcome variable), the RMSE was : 27.42462

For the random forest model with optimized mtry, the RMSE was: 26.94643

The Random Forest model with percentile variables included and a tuned mtry produced the smallest RMSE and is therefore the strongest model.

Recommendations and Final Thoughts

When working on this project to see if grade 3 academic outcomes in reading and math could be predicted by using a combination of the 5Essentials as well as demographic data pulled from the American Community Survey, I could sense rather early on that the chances of creating a strong predictive model, based purely on the 5Essentials and information provided in the data set, were low. However, I still wanted to create these models (after adding in some additional demographic information), because weak models can still provide insights and recommendations. Below, find some insights and recommendations based on this project's results.

Insight: Keep the focus on the schools

I created two sets of regression models for each machine learning technique - one that included the percentile variables from other years, and ones that did not. Rather expectedly, the models built with the percentile variables were stronger than the models built without, as the corrplot had showed that these variables had the strongest correlation with the outcome variables. This might seem banal and self-evident - of course a school's percentile attainments in 2014 and 2015 should help one to more accurately predict its percentile rank for 2016. But the stregnth of previous years percentiles for predicting percentiles in the future, *re-confirms* that the schools themselves are important.

When I first started looking at this project, I wondered if one thing was perhaps missing from the data set: the students. It seemed possible that academic outcomes might be influenced (and undoubtedly they are, to some extent) by the relative intelligence of the individual students, their desire to do well in school, their hard work. I was fairly certain that schools were vital, for allowing students the opportunity to develop and showcase their intelligence. But I wanted to check.

So the fact that the schools' performance in 2014 and 2015 math and reading were what was most predictive of 2016 percentiles, suggests that perhaps more than individual students and their characteristics, it is the institutions themselves that help determine how students do in math and reading.

While it was safe to assume from the start, the greater strength of models including percentiles from earlier years, underlines that schools themselves are important for academic outcomes, and that the focus need to remain on creating and maintaining strong schools.

Recommendation 1: Create long-term education policy and stick with it!

As stated earlier in the report, one of the main frustrations that I encountered with this project was that the metrics for evaluating academic performance altered in the early 2010s. As a result, I could only compare 2014, 2015 and 2016. However, to really do full justice to the 5Essentials paradigm and to actually assess its claims,

one would need to have comparable metrics for a 10-year period. While politics often leads to alterations in policy as administrations change, there needs to be a push for a long-term policy, for those creating education policy to understand that only in using long-term metrics can the long-term performance of the schools actually be assessed and known. To create data-driven, evidence-based decisions, education policy needs to adopt and stick to a long-term perspective.

Recommendation 2: Family Focus

The data analysis showed that while Ambitious Instruction was the Essential most frequently cited as strong or very strong throughout the data set, Family Involvement was the Essential most frequently cited a strong or very strong for schools scoring over the 50th percentile (above average) in grade 3 reading and math. What this suggests is that schools should particulary focus on cultivating family involvement in the life of the schools, if they want to create and maintain strong academic outcomes. One of the ways they might think about doing this is by making it mandatory (it is currently optional) for parents to fill out the survey upon which the progress report cards are based. Parents can offer a unique perspective that teachers might not have, about the school and student experience. And this might help to incorporate some of the experiences of students from kindergarten to grade 5, who are currently not eligible to participate in the survey, but whose parents could speak with them to get their perspective on school safety, health, and the supportiveness of the school environment. Making it mandatory for parents to fill out the survey would be a way of getting parents to sit down and engage with ideas about their children's education, perhaps a first step toward greater involvement, while adding further evidence to the data set.

Recommendation 3: More data!

While the Chicago Public Schools Progress Report Cards provide a number of statistics for each school, such as student attendance and teacher attendance, there are other statistics about these schools that should be included, so that one can get a better picture of a school environment. What is the average class size? What is the average ratio of teachers to students at a school? What are the rates of teacher turn-over for a particular year? What is the average rate of students failing grades? These are all statistics that the schools should be compiling and if added to the Progress Reports, would provide researchers with further relevant information for assessing the school.

Links to Project Code

Original Data Sets

Cleaned Up Data Set

Data Wrangling Script

Exploratory Data Analysis Scripts

Machine Learning Scripts