Project 1

Ailene Torres

2022-05-08

Introduction

Schools are essential in developing children for their life beyond graduation. Parents understandably want to send their children to the best schools. The New York City Department of Education reviews schools annually in order to determine if schools are meeting set targets about student achievement and such. These reports are important for the schools to determine what factors impact their ability to meet student achievement targets. By understanding these factors, these schools can make necessary changes to improve the school's performance. In this analysis, I will attempt to provide some insight into the student achievement targets using data from the NYC Open Data Portal. I have used two specific data sets for my analysis. The first data set provides information on every DOE high school's Regents Exam results. The second data set provides information on the quality report that the DOE releases every year about each school. This data set includes information about which targets each school has met and the student/faculty demographics of each school. Using these data sets, I'll attempt to fit a model that predicts and explains meeting student achievement targets.

Setup and Feature Engineering

Libraries Here are the libraries I will use in this analysis:

```
tinytex::install_tinytex()
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(ggpubr)
library(RColorBrewer)
library(forcats)
library(caret)
```

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

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

Reading in the data I first read in the data and explore the features. Seeing how the features are labeled, we rename them into lower case, underline-separated words in order to have the same column name pattern in each data set.

In Excel, I added a `year` feature to each QR data set to indicate which year of data it contains.

```
regents <- read_csv("regents_scores_revert.csv")
qr_2017_2018 <- read_csv("2017_2018_QR_revert.csv")
qr_2016_2017 <- read_csv("2016_2017_QR_Results.csv")
qr_2015_2016 <- read_csv("2015_2016_QR_Results.csv")
qr_2014_2015 <- read_csv("2014_2015_QR_Results.csv")

## cleaning the column names for aesthetics
colnames(regents) <- tolower(colnames(regents)) %>%
gsub(" ", "_", .)

colnames(qr_2014_2015) <- tolower(colnames(qr_2014_2015)) %>%
gsub(" ", "_", .)

colnames(qr_2015_2016) <- tolower(colnames(qr_2015_2016)) %>%
gsub(" ", "_", .)

colnames(qr_2016_2017) <- tolower(colnames(qr_2016_2017)) %>%
gsub(" ", "_", .)

colnames(qr_2017_2018) <- tolower(colnames(qr_2017_2018)) %>%
gsub(" ", "_", .)

qr_2014_2015 <- qr_2014_2015[,-c(18:23)]
```

```
colnames(qr_2014_2015)[24] <- "percent_in_temp_housing"

qr_2015_2016 <- qr_2015_2016[,-c(18:23,30)]

qr_2016_2017 <- qr_2016_2017[,-c(18:28,35)]

qr_2017_2018 <- qr_2017_2018[,-c(24)]

lst <- list(qr_2014_2015, qr_2015_2016, qr_2016_2017, qr_2017_2018)
quality_report <- Reduce(function(x,y) merge(x,y,all=TRUE), lst)

unique(quality_report$student_achievement_rating)
```

```
## [1] "Approaching Target" "Meeting Target"     "Exceeding Target"
## [4] "N/A"                NA                   "Not Meeting Target"
```

Exploratory Data Analysis We will first gain some insights on the two data sets to see what we're working with.

```
str(regents)
```

```
## spec_tbl_df [33,031 x 15] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ school_dbn                 : chr [1:33031] "01M034" "01M034" "01M034" "01M034" ...
##  $ school_name                : chr [1:33031] "P.S. 034 Franklin D. Roosevelt" "P.S. 034 Franklin D.
##  $ school_level               : chr [1:33031] "K-8" "K-8" "K-8" "K-8" ...
##  $ regents_exam               : chr [1:33031] "Living Environment" "Living Environment" "Common Core
##  $ year                       : num [1:33031] 2015 2016 2017 2018 2018 ...
##  $ total_tested               : num [1:33031] 16 9 4 2 2 3 3 9 3 15 ...
##  $ mean_score                 : num [1:33031] 77.9 74 0 0 0 0 0 67.4 0 72.6 ...
##  $ number_scoring_below_65    : num [1:33031] 1 1 0 0 0 0 0 3 0 2 ...
##  $ percent_scoring_below_65   : num [1:33031] 6.3 11.1 0 0 0 0 0 33.3 0 13.3 ...
##  $ number_scoring_65_or_above : num [1:33031] 15 8 0 0 0 0 0 6 0 13 ...
##  $ percent_scoring_65_or_above: num [1:33031] 93.8 88.9 0 0 0 0 0 66.7 0 86.7 ...
##  $ number_scoring_80_or_above : num [1:33031] 7 2 0 0 0 0 0 0 0 5 ...
##  $ percent_scoring_80_or_above: num [1:33031] 43.8 22.2 0 0 0 0 0 0 0 33.3 ...
##  $ number_scoring_cr          : num [1:33031] 0 0 0 0 0 0 0 0 0 0 ...
##  $ percent_scoring_cr         : num [1:33031] 0 0 0 0 0 0 0 0 0 0 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   ‘School DBN‘ = col_character(),
##   ..   ‘School Name‘ = col_character(),
##   ..   ‘School Level‘ = col_character(),
##   ..   ‘Regents Exam‘ = col_character(),
##   ..   Year = col_double(),
##   ..   ‘Total Tested‘ = col_double(),
##   ..   ‘Mean Score‘ = col_double(),
##   ..   ‘Number Scoring Below 65‘ = col_double(),
##   ..   ‘Percent Scoring Below 65‘ = col_double(),
##   ..   ‘Number Scoring 65 or Above‘ = col_double(),
##   ..   ‘Percent Scoring 65 or Above‘ = col_double(),
##   ..   ‘Number Scoring 80 or Above‘ = col_double(),
##   ..   ‘Percent Scoring 80 or Above‘ = col_double(),
##   ..   ‘Number Scoring CR‘ = col_double(),
##   ..   ‘Percent Scoring CR‘ = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

We see that the regents data set is composed of 33,031 rows and 15 columns. The data looks to spread out amongst multiple years and grade levels.

```
str(quality_report)
```

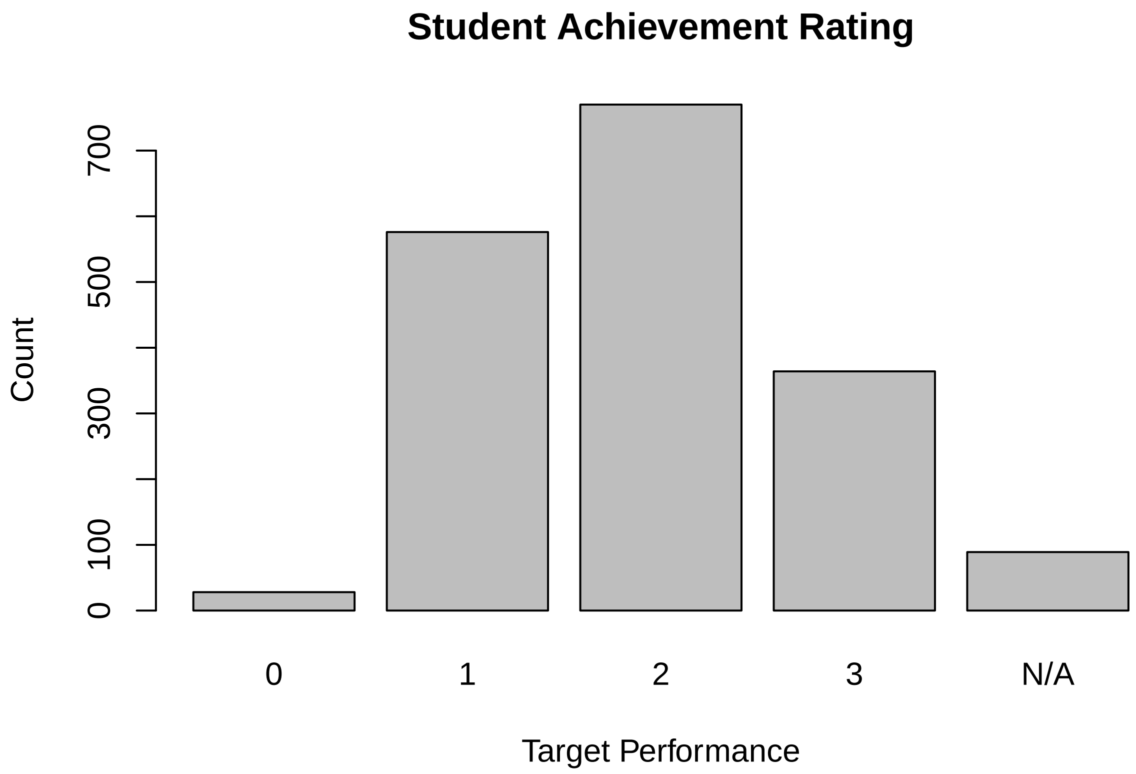
```
## 'data.frame':    1959 obs. of  35 variables:
##  $ dbn                                                   : chr  "01M292" "01M292" "01M292" "01M292"
##  $ school_name                                           : chr  "Henry Street School for Internationa
##  $ school_type                                           : chr  "High School" "High School" "High Sc
##  $ enrollment                                            : num  160 255 140 171 304 343 392 465 385
##  $ rigorous_instruction_rating                           : chr  "Approaching Target" "Not Meeting Ta
##  $ collaborative_teachers_rating                         : chr  "Meeting Target" "Approaching Target
##  $ supportive_environment_rating                         : chr  "Approaching Target" "Approaching Ta
##  $ effective_school_leadership_rating                    : chr  "Meeting Target" "Approaching Target
##  $ strong_family-community_ties_rating                   : chr  "Meeting Target" "Approaching Target
##  $ trust_rating                                          : chr  "Meeting Target" "Meeting Target" "E
##  $ student_achievement_rating                            : chr  "Approaching Target" "Meeting Target
##  $ rigorous_instruction_-_percent_positive               : num  0.8 0.72 0.81 0.82 0.85 0.91 0.82 0.
##  $ collaborative_teachers_-_percent_positive             : num  0.76 0.77 0.92 0.87 0.95 0.91 0.88 0
##  $ supportive_environment_-_percent_positive             : num  0.73 0.77 0.8 0.78 0.87 0.76 0.76 0.
##  $ effective_school_leadership_-_percent_positive        : num  0.89 0.72 0.92 0.91 0.93 0.95 0.89 0
##  $ strong_family-community_ties_-_percent_positive       : num  0.83 0.76 0.82 0.85 0.79 0.85 0.86 0
##  $ trust_-_percent_positive                              : num  0.91 0.87 0.94 0.92 0.95 0.92 0.92 0
##  $ average_grade_8_english_proficiency                   : num  2.31 2.18 2.47 2.75 2.27 2.6 2.85 2.
##  $ average_grade_8_math_proficiency                      : num  2.09 2.06 2.12 2.28 2.37 2.54 2.66 2
##  $ percent_english_language_learners                     : num  0.15 0.127 0.143 0.135 0.194 0.137 0
##  $ percent_students_with_disabilities                    : num  0.319 0.298 0.271 0.24 0.22 0.198 0.
##  $ percent_self-contained                                : num  0.013 0.015 0.036 0.012 0.003 0.009
##  $ economic_need_index                                   : num  0.881 0.832 0.832 0.898 0.812 0.771
##  $ percent_in_temp_housing                               : num  0.225 0.19 0.2 0.205 0.263 0.198 0.1
##  $ percent_hra_eligible                                  : num  0.638 0.663 0.621 0.813 0.599 0.548
##  $ percent_asian                                         : num  0.131 0.132 0.15 0.117 0.299 0.28 0.
##  $ percent_black                                         : num  0.225 0.244 0.243 0.246 0.25 0.274 0
##  $ percent_hispanic                                      : num  0.6 0.566 0.55 0.567 0.411 0.414 0.4
##  $ percent_white                                         : num  0.038 0.039 0.05 0.047 0.033 0.029 0
##  $ years_of_principal_experience_at_this_school          : num  0.9 3 1.9 2.9 5.5 6.5 7.5 8.5 16.8 1
##  $ percent_of_teachers_with_3_or_more_years_of_experience: num  0.591 0.667 0.5 0.684 0.696 0.577 0.
##  $ student_attendance_rate                               : num  0.811 0.766 0.867 0.886 0.88 0.908 0
##  $ percent_of_students_chronically_absent                : num  0.524 0.568 0.448 0.364 0.347 0.256
##  $ teacher_attendance_rate                               : num  0.972 0.971 0.973 0.965 0.971 0.966
##  $ year                                                  : num  2016 2015 2017 2018 2015 ...
```

We can see the quality_report data set is composed of 416 rows and 29 columns.

Student Achievement Targets Targets are a great way for schools to see how they are performing when it comes to student achievement. Student achievement is an overall rating for student test results, graduation rates, attendance, etc. Using the `student_achievement_rating` feature, we can see how schools are performing. The NA values will be removed before the model is created.

```
## Converting "Not Meeting Target" into 0, "Approaching Target" into 1, "Meeting Target" into 2, and "E
quality_report$student_achievement_rating[quality_report$student_achievement_rating == 'Not Meeting Targ
quality_report$student_achievement_rating[quality_report$student_achievement_rating == 'Approaching Targ
quality_report$student_achievement_rating[quality_report$student_achievement_rating == 'Meeting Target']
quality_report$student_achievement_rating[quality_report$student_achievement_rating == 'Exceeding Targe

barplot(table(quality_report$student_achievement_rating), main="Student Achievement Rating",
        xlab = "Target Performance", ylab = "Count")
```



```
## Renaming values back to their original state.
quality_report$student_achievement_rating[quality_report$student_achievement_rating == 0] <- 'Not Meeti
quality_report$student_achievement_rating[quality_report$student_achievement_rating == 1] <- 'Approachi
quality_report$student_achievement_rating[quality_report$student_achievement_rating == 2] <- 'Meeting Ta
quality_report$student_achievement_rating[quality_report$student_achievement_rating == 3] <- 'Exceeding
```

Filtering data Since we're working with high school data, we need to filter the high school Regents exam scores. We do this because students normally take Regents exams in high school. Both data sets will also be merged into one data set called school.

```
##filtering out scores from 2018 -- High schools
scores <- regents %>%
  filter((year == 2018 | year == 2017 | year == 2016 | year == 2015) & school_level == "High school")

## combining both datasets together
```

```
colnames(scores)[1] <- "dbn"
school <- merge(quality_report, scores, on = c("dbn", "school_name"))

school[school == "N/A"] <- NA
```

Student Testing Participation Not all students take the Regents exam at the same time. They have multiple opportunities to do so within a school year, but only a student's highest score is reported. The number of students who take Regents exams may also impact a school's student achievement rating. So, we engineer a new feature which lists for each Regents exam offered, the share of students who take it. This feature will later be merged back to the original data.

```
## calculating the number of students taking a regents exam relative to the school total enrollment
total_students <- school %>%
  group_by(school_name, year) %>%
  summarise(tested_sum = sum(total_tested, na.rm = T),
            count = n()) %>%
  mutate(participation_share = tested_sum / count) %>%
  select(school_name, year, participation_share)

school <- merge(school, total_students,
                on = c("school_name", "year"))

scores <- merge(scores, total_students,
                on = c("school_name", "year"))
```

Our output variable We will now define our ouput variable. For this analysis, we will define student achievement rating using the feature `student_achievement_rating`. Any rating listed as `"Meeting Target"` and `"Exceeding Target"` will be counted as passing, and anything else will not. Since we're only focusing on the `student_achievement_rating` feature, we will also ignore the other ratings and percentages from the data set. Redundant columns will also be filtered out.

```
## re-coding rating into passing (ceased == 1, other = 0)\
school <- na.omit(school)
school$passing <- ifelse((school$student_achievement_rating == "Meeting Target") | (school$student_achi
  factor()

mean(school$passing == 1)
```

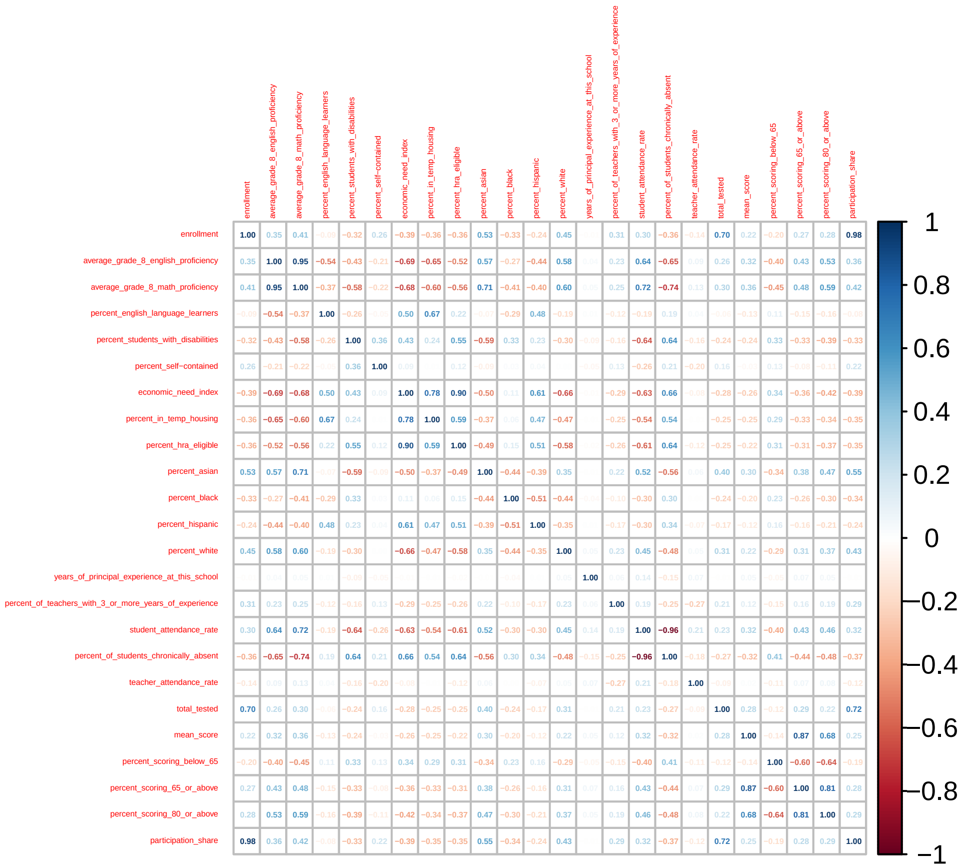
```
## [1] 0.6403073
```

```
## removing irrelevant columns as well as redundant columns such as
## number_ columns (we already have columns with the % values)
school <- school[,-c(6:11,13:18,40,42,44,46,47)]

## converting character to factor variables
char_vars <- c("school_name", "year", "dbn", "school_type", "school_level", "regents_exam", "student_ac
for(i in seq_along(char_vars)){
  school[,char_vars[i]] <- as.factor(school[,char_vars[i]])
}
```

Correlation between numeric variables

```
numeric_cols <- school[, sapply(school, is.numeric)]
pairs_matrix <- cor(numeric_cols, use = "complete.obs")
#, tl.cex = 0.40, number.cex = 0.15
corrplot(pairs_matrix, method = "number", tl.cex = 0.25, number.cex = 0.25)
```



	enrollment	average_grade_8_english_proficiency	average_grade_8_math_proficiency	percent_english_language_learners	percent_students_with_disabilities	percent_self−contained	economic_need_index	percent_in_temp_housing	percent_hra_eligible	percent_asian	percent_black	percent_hispanic	percent_white	years_of_principal_experience_at_this_school	percent_of_teachers_with_3_or_more_years_of_experience	student_attendance_rate	percent_of_students_chronically_absent	teacher_attendance_rate	total_tested	mean_score	percent_scoring_below_65	percent_scoring_65_or_above	percent_scoring_80_or_above	participation_share
enrollment	1.00	0.35	0.41	−0.09	−0.32	0.26	−0.39	−0.36	−0.36	0.53	−0.33	−0.24	0.45	−0.01	0.31	0.30	−0.36	−0.14	0.70	0.22	−0.20	0.27	0.28	0.98
average_grade_8_english_proficiency	0.35	1.00	0.95	−0.54	−0.43	−0.21	−0.69	−0.65	−0.52	0.57	−0.27	−0.44	0.58	0.04	0.23	0.64	−0.65	0.09	0.26	0.32	−0.40	0.43	0.53	0.36
average_grade_8_math_proficiency	0.41	0.95	1.00	−0.37	−0.58	−0.22	−0.68	−0.60	−0.56	0.71	−0.41	−0.40	0.60	0.05	0.25	0.72	−0.74	0.13	0.30	0.36	−0.45	0.48	0.59	0.42
percent_english_language_learners	−0.09	−0.54	−0.37	1.00	−0.26	−0.05	0.50	0.67	0.22	−0.07	−0.29	0.48	−0.19	0.01	−0.12	−0.19	0.19	0.04	−0.06	−0.13	0.11	−0.15	−0.16	−0.08
percent_students_with_disabilities	−0.32	−0.43	−0.58	−0.26	1.00	0.36	0.43	0.24	0.55	−0.59	0.33	0.23	−0.30	−0.09	−0.16	−0.64	0.64	−0.16	−0.24	−0.24	0.33	−0.33	−0.39	−0.33
percent_self−contained	0.26	−0.21	−0.22	−0.05	0.36	1.00	0.09		0.12	−0.09	0.03	0.04		−0.05	0.13	−0.26	0.21	−0.20	0.16	−0.03	0.13	−0.08	−0.11	0.22
economic_need_index	−0.39	−0.69	−0.68	0.50	0.43	0.09	1.00	0.78	0.90	−0.50	0.11	0.61	−0.66	−0.01	−0.29	−0.63	0.66	−0.08	−0.28	−0.26	0.34	−0.36	−0.42	−0.39
percent_in_temp_housing	−0.36	−0.65	−0.60	0.67	0.24		0.78	1.00	0.59	−0.37	0.06	0.47	−0.47		−0.25	−0.54	0.54		−0.25	−0.25	0.29	−0.33	−0.34	−0.35
percent_hra_eligible	−0.36	−0.52	−0.56	0.22	0.55	0.12	0.90	0.59	1.00	−0.49	0.15	0.51	−0.58	−0.02	−0.26	−0.61	0.64	−0.12	−0.25	−0.22	0.31	−0.31	−0.37	−0.35
percent_asian	0.53	0.57	0.71	−0.07	−0.59	−0.09	−0.50	−0.37	−0.49	1.00	−0.44	−0.39	0.35		0.22	0.52	−0.56	0.06	0.40	0.30	−0.34	0.38	0.47	0.55
percent_black	−0.33	−0.27	−0.41	−0.29	0.33	0.03	0.11	0.06	0.15	−0.44	1.00	−0.51	−0.44	−0.04	−0.10	−0.30	0.30		−0.24	−0.20	0.23	−0.26	−0.30	−0.34
percent_hispanic	−0.24	−0.44	−0.40	0.48	0.23	0.04	0.61	0.47	0.51	−0.39	−0.51	1.00	−0.35		−0.17	−0.30	0.34	−0.07	−0.17	−0.12	0.16	−0.16	−0.21	−0.24
percent_white	0.45	0.58	0.60	−0.19	−0.30		−0.66	−0.47	−0.58	0.35	−0.44	−0.35	1.00	0.05	0.23	0.45	−0.48	0.05	0.31	0.22	−0.29	0.31	0.37	0.43
years_of_principal_experience_at_this_school	−0.01	0.04	0.05	0.01	−0.09	−0.05	−0.01		−0.02		−0.04		0.05	1.00	0.06	0.14	−0.15	0.07		0.05	−0.05	0.07	0.05	
percent_of_teachers_with_3_or_more_years_of_experience	0.31	0.23	0.25	−0.12	−0.16	0.13	−0.29	−0.25	−0.26	0.22	−0.10	−0.17	0.23	0.06	1.00	0.19	−0.25	−0.27	0.21	0.12	−0.15	0.16	0.19	0.29
student_attendance_rate	0.30	0.64	0.72	−0.19	−0.64	−0.26	−0.63	−0.54	−0.61	0.52	−0.30	−0.30	0.45	0.14	0.19	1.00	−0.96	0.21	0.23	0.32	−0.40	0.43	0.46	0.32
percent_of_students_chronically_absent	−0.36	−0.65	−0.74	0.19	0.64	0.21	0.66	0.54	0.64	−0.56	0.30	0.34	−0.48	−0.15	−0.25	−0.96	1.00	−0.18	−0.27	−0.32	0.41	−0.44	−0.48	−0.37
teacher_attendance_rate	−0.14	0.09	0.13	0.04	−0.16	−0.20	−0.08		−0.12	0.06		−0.07	0.05	0.07	−0.27	0.21	−0.18	1.00	−0.09	0.02	−0.11	0.07	0.08	−0.12
total_tested	0.70	0.26	0.30	−0.06	−0.24	0.16	−0.28	−0.25	−0.25	0.40	−0.24	−0.17	0.31		0.21	0.23	−0.27	−0.09	1.00	0.28	−0.12	0.29	0.22	0.72
mean_score	0.22	0.32	0.36	−0.13	−0.24	−0.03	−0.26	−0.25	−0.22	0.30	−0.20	−0.12	0.22	0.05	0.12	0.32	−0.32		0.28	1.00	−0.14	0.87	0.68	0.25
percent_scoring_below_65	−0.20	−0.40	−0.45	0.11	0.33	0.13	0.34	0.29	0.31	−0.34	0.23	0.16	−0.29	−0.05	−0.15	−0.40	0.41	−0.11	−0.12	−0.14	1.00	−0.60	−0.64	−0.19
percent_scoring_65_or_above	0.27	0.43	0.48	−0.15	−0.33	−0.08	−0.36	−0.33	−0.31	0.38	−0.26	−0.16	0.31	0.07	0.16	0.43	−0.44	0.07	0.29	0.87	−0.60	1.00	0.81	0.28
percent_scoring_80_or_above	0.28	0.53	0.59	−0.16	−0.39	−0.11	−0.42	−0.34	−0.37	0.47	−0.30	−0.21	0.37	0.05	0.19	0.46	−0.48	0.08	0.22	0.68	−0.64	0.81	1.00	0.29
participation_share	0.98	0.36	0.42	−0.08	−0.33	0.22	−0.39	−0.35	−0.35	0.55	−0.34	−0.24	0.43		0.29	0.32	−0.37	−0.12	0.72	0.25	−0.19	0.28	0.29	1.00

We can see that a few of the enrollment/participation pairs have a positive correlation like `enrollment` and `participation_share`. We also see some positive correlation in some of the demographic features such as the `economic_need_index` and `percent_hra_eligible`. There are also some noticeable negative correlations. For example, the `economic_need_index` feature and the average math and english proficiency features have negative correlations.

Data Splitting

To train the models, we employ an 80-20 split of train and test, but we also introduce a new subset to only include exam scores from 2015-2018 high schools.

```
set.seed(1234)
# Create Training and Testing Data
school18 <- school %>%
  select(-dbn, -student_achievement_rating, -school_type, -school_level, -school_name)

split <- sample(1:nrow(school18), 0.8*nrow(school18), replace = F)
```

```
train18 <- school18[split,]
test18 <- school18[-split,]

str(train18)
```

```
## 'data.frame':    9995 obs. of  27 variables:
##  $ year                                              : Factor w/ 4 levels "2015","2016",..: 1 4 3
##  $ enrollment                                        : num  699 1174 2215 437 2225 ...
##  $ average_grade_8_english_proficiency               : num  2.33 2.76 2.57 2.52 2.65 2.23 2.23 2
##  $ average_grade_8_math_proficiency                  : num  2.18 2.5 2.27 2.37 2.68 2 2.07 2.28 4
##  $ percent_english_language_learners                 : num  0.067 0.124 0.196 0.057 0.192 0.197 0
##  $ percent_students_with_disabilities                : num  0.203 0.114 0.161 0.13 0.133 0.333 0
##  $ percent_self-contained                            : num  0.043 0.023 0.051 0.005 0.042 0 0.034
##  $ economic_need_index                               : num  0.751 0.67 0.618 0.503 0.702 0.809 0
##  $ percent_in_temp_housing                           : num  0.089 0.089 0.141 0.087 0.126 0.178 0
##  $ percent_hra_eligible                              : num  0.597 0.549 0.332 0.343 0.453 0.587 0
##  $ percent_asian                                     : num  0.039 0.297 0.342 0.087 0.333 0.06 0
##  $ percent_black                                     : num  0.279 0.441 0.223 0.705 0.313 0.283 0
##  $ percent_hispanic                                  : num  0.619 0.168 0.352 0.124 0.208 0.622 0
##  $ percent_white                                     : num  0.044 0.038 0.033 0.025 0.129 0.013 0
##  $ years_of_principal_experience_at_this_school      : num  4 6 5 6.8 1 2.3 4 3.9 9.9 6.4 ...
##  $ percent_of_teachers_with_3_or_more_years_of_experience: num  0.732 0.936 0.602 0.455 0.644 0.733 0
##  $ student_attendance_rate                           : num  0.792 0.875 0.857 0.846 0.867 0.843 0
##  $ percent_of_students_chronically_absent            : num  0.583 0.355 0.387 0.396 0.23 0.404 0
##  $ teacher_attendance_rate                           : num  0.972 0.96 0.964 0.966 0.97 0.965 0.9
##  $ regents_exam                                      : Factor w/ 18 levels "Algebra2/Trigonometry
##  $ total_tested                                      : num  243 93 3 171 37 104 9 154 1 146 ...
##  $ mean_score                                        : num  61 71.8 0 64.2 89 55.3 55.4 64.1 0 65
##  $ percent_scoring_below_65                          : num  51.9 28 0 44.4 2.7 59.6 77.8 42.9 0 2
##  $ percent_scoring_65_or_above                       : num  48.1 72 0 55.6 97.3 40.4 22.2 57.1 0
##  $ percent_scoring_80_or_above                       : num  14.8 36.6 0 14 89.2 6.7 0 3.2 0 26.7
##  $ participation_share                               : num  99 224.7 385.8 78.4 353.4 ...
##  $ passing                                           : Factor w/ 2 levels "0","1": 1 1 2 1 2 2 2
```

Classifier: Logistic Regression

In this model, we use maximum-likelihood estimation to fit a line in the form f(x) = a + b1x + $b2$z + . . . , where P(Y = 1) = logit-1(f(x)).

```
log_fit <- glm(passing ~ ., data = train18, family = binomial(link = "logit"))
summary(log_fit)
```

```
## 
## Call:
## glm(formula = passing ~ ., family = binomial(link = "logit"), 
##     data = train18)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -2.4896  -0.7987   0.2893   0.7211   2.8633  
## 
```

```
## Coefficients:
##                                                           Estimate Std. Error
## (Intercept)                                            -5.189e+01  3.995e+00
## year2016                                               -6.697e-01  9.919e-02
## year2017                                               -9.137e-01  1.384e-01
## year2018                                               -1.539e+00  1.627e-01
## enrollment                                             -1.273e-03  1.456e-04
## average_grade_8_english_proficiency                    -2.122e-01  4.145e-01
## average_grade_8_math_proficiency                        9.901e-01  3.496e-01
## percent_english_language_learners                       1.589e+00  4.784e-01
## percent_students_with_disabilities                      5.088e+00  8.253e-01
## ‘percent_self-contained‘                               -1.195e+01  1.115e+00
## economic_need_index                                    -4.550e+00  7.259e-01
## percent_in_temp_housing                                 7.983e-01  7.651e-01
## percent_hra_eligible                                    4.376e+00  6.377e-01
## percent_asian                                           3.886e+00  2.217e+00
## percent_black                                           2.268e+00  2.061e+00
## percent_hispanic                                        2.741e+00  2.075e+00
## percent_white                                           3.999e-02  2.058e+00
## years_of_principal_experience_at_this_school            3.775e-02  6.778e-03
## percent_of_teachers_with_3_or_more_years_of_experience -3.581e-01  1.985e-01
## student_attendance_rate                                 2.137e+01  2.022e+00
## percent_of_students_chronically_absent                 -1.379e+00  7.008e-01
## teacher_attendance_rate                                 3.297e+01  2.997e+00
## regents_examChinese                                    -1.054e+00  2.955e-01
## regents_examCommon Core Algebra                        -2.005e-01  1.713e-01
## regents_examCommon Core Algebra2                       -3.208e-01  1.731e-01
## regents_examCommon Core English                        -6.045e-01  1.672e-01
## regents_examCommon Core Geometry                       -4.345e-02  1.615e-01
## regents_examEnglish                                    -4.125e-01  1.856e-01
## regents_examFrench                                     -6.765e-01  2.378e-01
## regents_examGeometry                                   -2.402e-01  1.956e-01
## regents_examGlobal History and Geography               -2.644e-01  1.659e-01
## regents_examIntegrated Algebra                         -4.561e-01  1.882e-01
## regents_examItalian                                    -1.116e+00  3.261e-01
## regents_examLiving Environment                         -3.623e-01  1.681e-01
## regents_examPhysical Settings/Chemistry                -1.884e-01  1.684e-01
## regents_examPhysical Settings/Earth Science            -1.380e-01  1.607e-01
## regents_examPhysical Settings/Physics                  -3.960e-01  1.844e-01
## regents_examSpanish                                    -7.992e-01  1.964e-01
## regents_examU.S. History and Government                -5.850e-01  1.683e-01
## total_tested                                           -2.794e-04  2.247e-04
## mean_score                                              7.956e-03  1.170e-02
## percent_scoring_below_65                               -1.525e-02  5.601e-03
## percent_scoring_65_or_above                            -1.593e-03  8.847e-03
## percent_scoring_80_or_above                             1.061e-03  3.224e-03
## participation_share                                     7.027e-03  9.855e-04
##                                                        z value Pr(>|z|)    
## (Intercept)                                            -12.990  < 2e-16 ***
## year2016                                                -6.752 1.46e-11 ***
## year2017                                                -6.604 4.01e-11 ***
## year2018                                                -9.460  < 2e-16 ***
## enrollment                                              -8.745  < 2e-16 ***
## average_grade_8_english_proficiency                     -0.512 0.608678    
## average_grade_8_math_proficiency                         2.832 0.004623 **
## percent_english_language_learners                        3.322 0.000894 ***
## percent_students_with_disabilities                       6.166 7.02e-10 ***
## ‘percent_self-contained‘                               -10.715  < 2e-16 ***
## economic_need_index                                     -6.268 3.66e-10 ***
## percent_in_temp_housing                                  1.043 0.296797
## percent_hra_eligible                                     6.862 6.80e-12 ***
## percent_asian                                            1.753 0.079684 .
## percent_black                                            1.100 0.271187
## percent_hispanic                                         1.321 0.186403
## percent_white                                            0.019 0.984497
## years_of_principal_experience_at_this_school             5.569 2.56e-08 ***
## percent_of_teachers_with_3_or_more_years_of_experience  -1.804 0.071261 .
## student_attendance_rate                                 10.569  < 2e-16 ***
## percent_of_students_chronically_absent                  -1.968 0.049124 *
## teacher_attendance_rate                                 10.999  < 2e-16 ***
## regents_examChinese                                     -3.568 0.000360 ***
## regents_examCommon Core Algebra                         -1.171 0.241717
## regents_examCommon Core Algebra2                        -1.853 0.063889 .
## regents_examCommon Core English                         -3.615 0.000300 ***
## regents_examCommon Core Geometry                        -0.269 0.787939
## regents_examEnglish                                     -2.223 0.026212 *
## regents_examFrench                                      -2.845 0.004443 **
## regents_examGeometry                                    -1.228 0.219379
## regents_examGlobal History and Geography                -1.594 0.111045
## regents_examIntegrated Algebra                          -2.423 0.015388 *
## regents_examItalian                                     -3.423 0.000619 ***
## regents_examLiving Environment                          -2.155 0.031166 *
## regents_examPhysical Settings/Chemistry                 -1.119 0.263339
## regents_examPhysical Settings/Earth Science             -0.859 0.390308
## regents_examPhysical Settings/Physics                   -2.147 0.031782 *
## regents_examSpanish                                     -4.069 4.71e-05 ***
## regents_examU.S. History and Government                 -3.477 0.000507 ***
## total_tested                                            -1.244 0.213631
## mean_score                                               0.680 0.496674
## percent_scoring_below_65                                -2.723 0.006469 **
## percent_scoring_65_or_above                             -0.180 0.857136
## percent_scoring_80_or_above                              0.329 0.741975
## participation_share                                      7.131 1.00e-12 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 13055.8  on 9994  degrees of freedom
## Residual deviance:  9175.9  on 9950  degrees of freedom
## AIC: 9265.9
##
## Number of Fisher Scoring iterations: 6
```

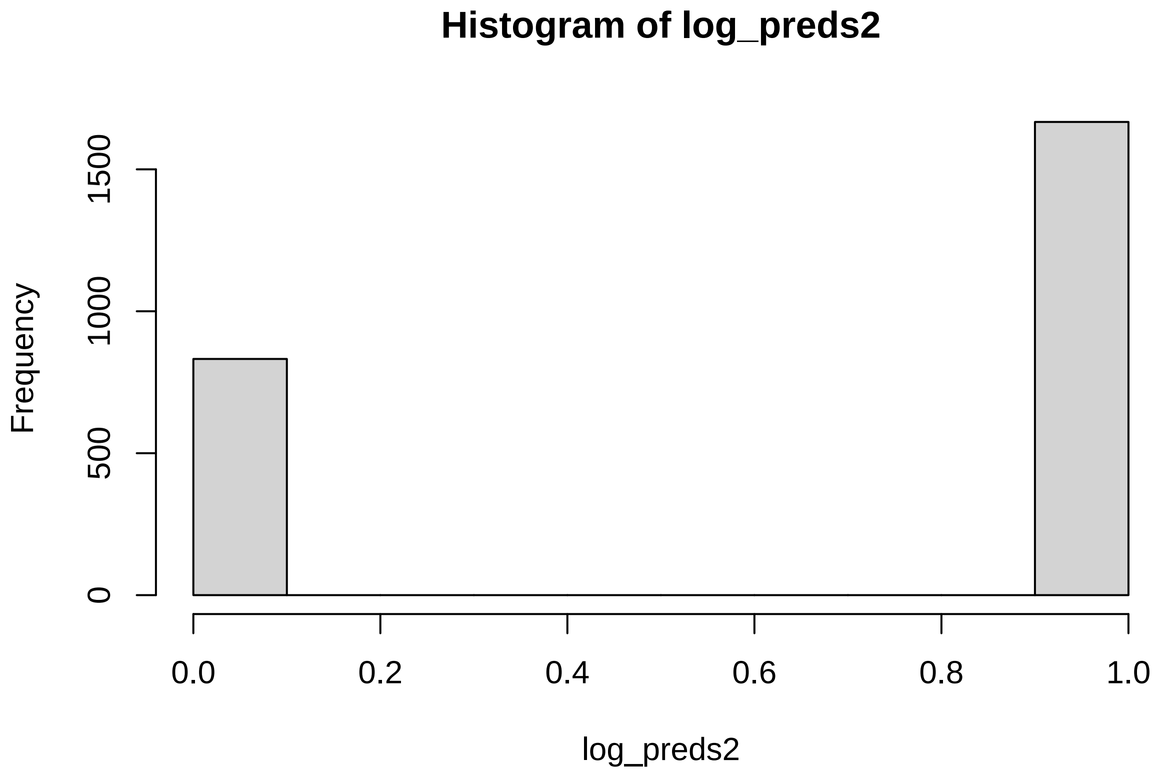
```r
log_preds <- predict(log_fit, newdata = test18, type = "response")

alpha <- 0.5
log_preds2 <- ifelse(log_preds > alpha, 1, 0)
```

```
confusionMatrix(factor(log_preds2), test18$passing)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0  571  261
##          1  330 1337
##
##                Accuracy : 0.7635
##                  95% CI : (0.7463, 0.78)
##     No Information Rate : 0.6395
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.4784
##
##  Mcnemar's Test P-Value : 0.005156
##
##             Sensitivity : 0.6337
##             Specificity : 0.8367
##          Pos Pred Value : 0.6863
##          Neg Pred Value : 0.8020
##              Prevalence : 0.3605
##          Detection Rate : 0.2285
##    Detection Prevalence : 0.3329
##       Balanced Accuracy : 0.7352
##
##        'Positive' Class : 0
##
```

```
hist(log_preds2)
```



With an accuracy of ~76%, the model performs slightly better than predicting at random according to the base rate of passing or a non-informative model.

There are two main ways we can tune our model. The first is feature selection– determining which features are most important– and the other is tuning the classification hyper-parameter. So instead of classifying an observation as 1 if its predicted probability is >= 50%, we could make our model more confident by lowering the hyperparameter from 50% to 40% (or some other number). For now, I will focus on feature selection.

```
summary(log_fit)
```

```
##
## Call:
## glm(formula = passing ~ ., family = binomial(link = "logit"),
##     data = train18)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4896  -0.7987   0.2893   0.7211   2.8633
##
## Coefficients:
##                                                           Estimate Std. Error
## (Intercept)                                             -5.189e+01  3.995e+00
## year2016                                                -6.697e-01  9.919e-02
## year2017                                                -9.137e-01  1.384e-01
## year2018                                                -1.539e+00  1.627e-01
## enrollment                                              -1.273e-03  1.456e-04
## average_grade_8_english_proficiency                    -2.122e-01  4.145e-01
## average_grade_8_math_proficiency                        9.901e-01  3.496e-01
## percent_english_language_learners                       1.589e+00  4.784e-01
## percent_students_with_disabilities                      5.088e+00  8.253e-01
## ‘percent_self-contained‘                               -1.195e+01  1.115e+00
## economic_need_index                                    -4.550e+00  7.259e-01
## percent_in_temp_housing                                 7.983e-01  7.651e-01
## percent_hra_eligible                                    4.376e+00  6.377e-01
## percent_asian                                           3.886e+00  2.217e+00
## percent_black                                           2.268e+00  2.061e+00
## percent_hispanic                                        2.741e+00  2.075e+00
## percent_white                                           3.999e-02  2.058e+00
## years_of_principal_experience_at_this_school            3.775e-02  6.778e-03
## percent_of_teachers_with_3_or_more_years_of_experience -3.581e-01  1.985e-01
## student_attendance_rate                                 2.137e+01  2.022e+00
## percent_of_students_chronically_absent                 -1.379e+00  7.008e-01
## teacher_attendance_rate                                 3.297e+01  2.997e+00
## regents_examChinese                                    -1.054e+00  2.955e-01
## regents_examCommon Core Algebra                        -2.005e-01  1.713e-01
## regents_examCommon Core Algebra2                       -3.208e-01  1.731e-01
## regents_examCommon Core English                        -6.045e-01  1.672e-01
## regents_examCommon Core Geometry                       -4.345e-02  1.615e-01
## regents_examEnglish                                    -4.125e-01  1.856e-01
## regents_examFrench                                     -6.765e-01  2.378e-01
## regents_examGeometry                                   -2.402e-01  1.956e-01
## regents_examGlobal History and Geography               -2.644e-01  1.659e-01
## regents_examIntegrated Algebra                         -4.561e-01  1.882e-01
## regents_examItalian                                    -1.116e+00  3.261e-01
## regents_examLiving Environment                         -3.623e-01  1.681e-01
## regents_examPhysical Settings/Chemistry                -1.884e-01  1.684e-01
## regents_examPhysical Settings/Earth Science            -1.380e-01  1.607e-01
## regents_examPhysical Settings/Physics                  -3.960e-01  1.844e-01
## regents_examSpanish                                    -7.992e-01  1.964e-01
## regents_examU.S. History and Government                -5.850e-01  1.683e-01
## total_tested                                           -2.794e-04  2.247e-04
## mean_score                                              7.956e-03  1.170e-02
## percent_scoring_below_65                               -1.525e-02  5.601e-03
## percent_scoring_65_or_above                            -1.593e-03  8.847e-03
## percent_scoring_80_or_above                             1.061e-03  3.224e-03
## participation_share                                     7.027e-03  9.855e-04
##                                                        z value Pr(>|z|)
## (Intercept)                                            -12.990  < 2e-16 ***
## year2016                                                -6.752 1.46e-11 ***
## year2017                                                -6.604 4.01e-11 ***
## year2018                                                -9.460  < 2e-16 ***
## enrollment                                              -8.745  < 2e-16 ***
## average_grade_8_english_proficiency                     -0.512 0.608678
## average_grade_8_math_proficiency                         2.832 0.004623 **
## percent_english_language_learners                        3.322 0.000894 ***
## percent_students_with_disabilities                       6.166 7.02e-10 ***
## ‘percent_self-contained‘                               -10.715  < 2e-16 ***
## economic_need_index                                     -6.268 3.66e-10 ***
## percent_in_temp_housing                                  1.043 0.296797
## percent_hra_eligible                                     6.862 6.80e-12 ***
## percent_asian                                           1.753 0.079684 .  
## percent_black                                           1.100 0.271187    
## percent_hispanic                                        1.321 0.186403    
## percent_white                                           0.019 0.984497    
## years_of_principal_experience_at_this_school            5.569 2.56e-08 ***
## percent_of_teachers_with_3_or_more_years_of_experience -1.804 0.071261 .  
## student_attendance_rate                                10.569  < 2e-16 ***
## percent_of_students_chronically_absent                 -1.968 0.049124 *  
## teacher_attendance_rate                                10.999  < 2e-16 ***
## regents_examChinese                                    -3.568 0.000360 ***
## regents_examCommon Core Algebra                        -1.171 0.241717    
## regents_examCommon Core Algebra2                       -1.853 0.063889 .  
## regents_examCommon Core English                        -3.615 0.000300 ***
## regents_examCommon Core Geometry                       -0.269 0.787939    
## regents_examEnglish                                    -2.223 0.026212 *  
## regents_examFrench                                     -2.845 0.004443 ** 
## regents_examGeometry                                   -1.228 0.219379    
## regents_examGlobal History and Geography               -1.594 0.111045    
## regents_examIntegrated Algebra                         -2.423 0.015388 *  
## regents_examItalian                                    -3.423 0.000619 ***
## regents_examLiving Environment                         -2.155 0.031166 *  
## regents_examPhysical Settings/Chemistry                -1.119 0.263339    
## regents_examPhysical Settings/Earth Science            -0.859 0.390308    
## regents_examPhysical Settings/Physics                  -2.147 0.031782 *  
## regents_examSpanish                                    -4.069 4.71e-05 ***
## regents_examU.S. History and Government                -3.477 0.000507 ***
## total_tested                                           -1.244 0.213631    
## mean_score                                              0.680 0.496674    
## percent_scoring_below_65                               -2.723 0.006469 ** 
## percent_scoring_65_or_above                            -0.180 0.857136    
## percent_scoring_80_or_above                             0.329 0.741975    
## participation_share                                     7.131 1.00e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 13055.8  on 9994  degrees of freedom
## Residual deviance:  9175.9  on 9950  degrees of freedom
## AIC: 9265.9
## 
## Number of Fisher Scoring iterations: 6
```

From the summary, we can see that a few features that are not significant. This includes some of the student demographics and Regents exam percentages. We hope that our model will improve by removing some of these features.

```
set.seed(1234)
# Create new Training and Testing Data
school18_new <- school18 %>%
  select(-percent_white, -total_tested, -mean_score, -percent_scoring_65_or_above,
         -percent_scoring_80_or_above)

split <- sample(1:nrow(school18_new), 0.8*nrow(school18_new), replace = F)
```

```
train18_new <- school18_new[split,]
test18_new <- school18_new[-split,]

str(train18_new)
```

```
## 'data.frame':    9995 obs. of  22 variables:
##  $ year                                                  : Factor w/ 4 levels "2015","2016",..: 1 4 3
##  $ enrollment                                            : num  699 1174 2215 437 2225 ...
##  $ average_grade_8_english_proficiency                   : num  2.33 2.76 2.57 2.52 2.65 2.23 2.23 2
##  $ average_grade_8_math_proficiency                      : num  2.18 2.5 2.27 2.37 2.68 2 2.07 2.28 4
##  $ percent_english_language_learners                     : num  0.067 0.124 0.196 0.057 0.192 0.197 0
##  $ percent_students_with_disabilities                    : num  0.203 0.114 0.161 0.13 0.133 0.333 0
##  $ percent_self-contained                                : num  0.043 0.023 0.051 0.005 0.042 0 0.034
##  $ economic_need_index                                   : num  0.751 0.67 0.618 0.503 0.702 0.809 0
##  $ percent_in_temp_housing                               : num  0.089 0.089 0.141 0.087 0.126 0.178 0
##  $ percent_hra_eligible                                  : num  0.597 0.549 0.332 0.343 0.453 0.587 0
##  $ percent_asian                                         : num  0.039 0.297 0.342 0.087 0.333 0.06 0
##  $ percent_black                                         : num  0.279 0.441 0.223 0.705 0.313 0.283 0
##  $ percent_hispanic                                      : num  0.619 0.168 0.352 0.124 0.208 0.622 0
##  $ years_of_principal_experience_at_this_school          : num  4 6 5 6.8 1 2.3 4 3.9 9.9 6.4 ...
##  $ percent_of_teachers_with_3_or_more_years_of_experience: num  0.732 0.936 0.602 0.455 0.644 0.733 0
##  $ student_attendance_rate                               : num  0.792 0.875 0.857 0.846 0.867 0.843 0
##  $ percent_of_students_chronically_absent                : num  0.583 0.355 0.387 0.396 0.23 0.404 0
##  $ teacher_attendance_rate                               : num  0.972 0.96 0.964 0.966 0.97 0.965 0.9
##  $ regents_exam                                          : Factor w/ 18 levels "Algebra2/Trigonometry
##  $ percent_scoring_below_65                              : num  51.9 28 0 44.4 2.7 59.6 77.8 42.9 0 2
##  $ participation_share                                   : num  99 224.7 385.8 78.4 353.4 ...
##  $ passing                                               : Factor w/ 2 levels "0","1": 1 1 2 1 2 2 2
```

After removing these features from the data set, we run the logistic regression model again.

```
log_fit_new <- glm(passing ~ ., data = train18_new, family = binomial(link = "logit"))
summary(log_fit_new)
```

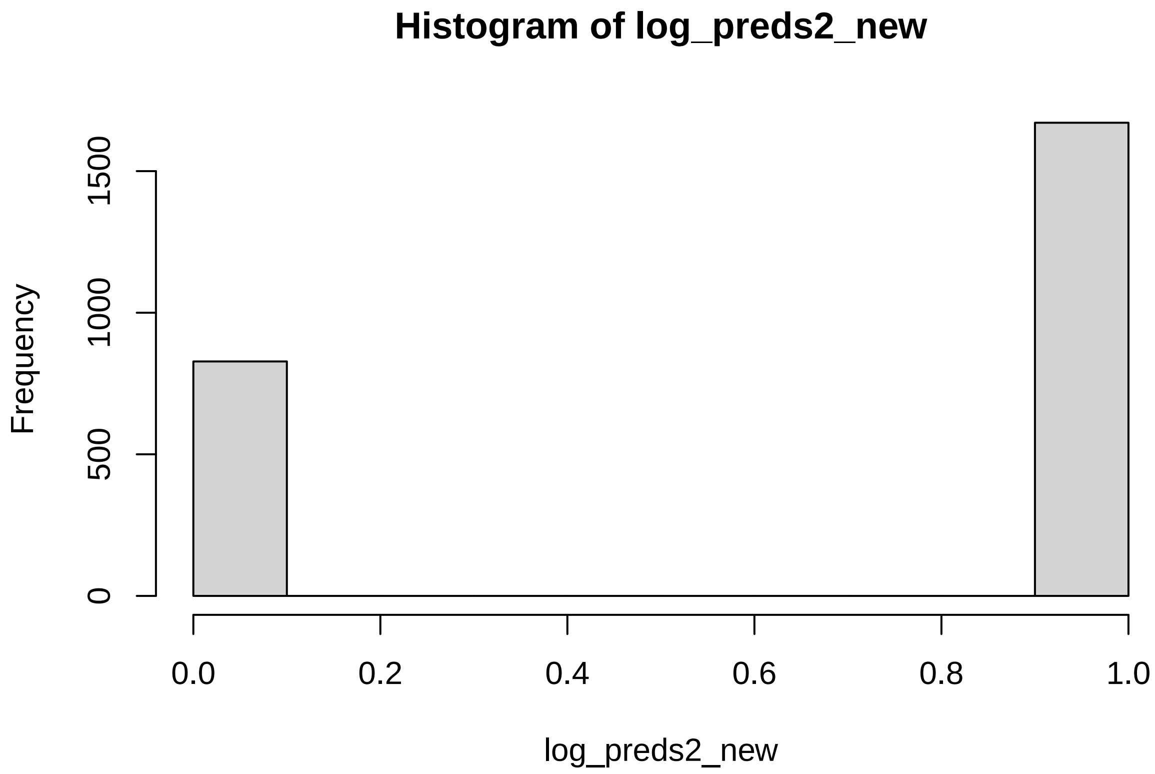
```
##
## Call:
## glm(formula = passing ~ ., family = binomial(link = "logit"),
##     data = train18_new)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4739  -0.8069   0.2897   0.7197   2.8250
##
## Coefficients:
##                                                          Estimate Std. Error
## (Intercept)                                            -5.195e+01  3.540e+00
## year2016                                               -7.042e-01  9.858e-02
## year2017                                               -9.459e-01  1.376e-01
## year2018                                               -1.568e+00  1.603e-01
## enrollment                                             -1.284e-03  1.460e-04
## average_grade_8_english_proficiency                    -1.998e-01  4.087e-01
## average_grade_8_math_proficiency                        1.075e+00  3.411e-01
## percent_english_language_learners                      1.488e+00  4.740e-01
## percent_students_with_disabilities                     5.041e+00  8.112e-01
## ‘percent_self-contained‘                              -1.183e+01  1.113e+00
## economic_need_index                                   -4.413e+00  7.196e-01
## percent_in_temp_housing                                7.971e-01  7.613e-01
## percent_hra_eligible                                   4.275e+00  6.350e-01
## percent_asian                                          3.920e+00  4.774e-01
## percent_black                                          2.182e+00  3.423e-01
## percent_hispanic                                       2.708e+00  3.613e-01
## years_of_principal_experience_at_this_school           3.815e-02  6.759e-03
## percent_of_teachers_with_3_or_more_years_of_experience -3.244e-01  1.966e-01
## student_attendance_rate                                2.148e+01  2.018e+00
## percent_of_students_chronically_absent                -1.463e+00  6.995e-01
## teacher_attendance_rate                                3.295e+01  2.995e+00
## regents_examChinese                                   -8.505e-01  2.766e-01
## regents_examCommon Core Algebra                       -1.145e-01  1.513e-01
## regents_examCommon Core Algebra2                      -2.318e-01  1.643e-01
## regents_examCommon Core English                       -4.135e-01  1.556e-01
## regents_examCommon Core Geometry                       1.257e-02  1.537e-01
## regents_examEnglish                                   -2.900e-01  1.795e-01
## regents_examFrench                                    -5.513e-01  2.309e-01
## regents_examGeometry                                  -1.990e-01  1.908e-01
## regents_examGlobal History and Geography              -1.705e-01  1.526e-01
## regents_examIntegrated Algebra                        -3.414e-01  1.780e-01
## regents_examItalian                                   -8.730e-01  3.143e-01
## regents_examLiving Environment                        -2.419e-01  1.529e-01
## regents_examPhysical Settings/Chemistry               -1.082e-01  1.564e-01
## regents_examPhysical Settings/Earth Science           -3.798e-02  1.544e-01
## regents_examPhysical Settings/Physics                 -2.993e-01  1.811e-01
## regents_examSpanish                                   -4.841e-01  1.734e-01
## regents_examU.S. History and Government               -3.975e-01  1.555e-01
## percent_scoring_below_65                              -1.310e-02  1.213e-03
## participation_share                                    6.878e-03  9.616e-04
##                                                       z value Pr(>|z|)
## (Intercept)                                           -14.675  < 2e-16 ***
## year2016                                               -7.143 9.13e-13 ***
## year2017                                               -6.873 6.31e-12 ***
## year2018                                               -9.780  < 2e-16 ***
## enrollment                                             -8.798  < 2e-16 ***
## average_grade_8_english_proficiency                    -0.489  0.62491
## average_grade_8_math_proficiency                        3.153  0.00162 **
## percent_english_language_learners                       3.140  0.00169 **
## percent_students_with_disabilities                      6.214 5.16e-10 ***
## ‘percent_self-contained‘                              -10.623  < 2e-16 ***
## economic_need_index                                    -6.132 8.65e-10 ***
## percent_in_temp_housing                                 1.047  0.29505
## percent_hra_eligible                                    6.732 1.67e-11 ***
## percent_asian                                           8.211  < 2e-16 ***
## percent_black                                           6.372 1.86e-10 ***
## percent_hispanic                                        7.495 6.65e-14 ***
## years_of_principal_experience_at_this_school            5.645 1.65e-08 ***
## percent_of_teachers_with_3_or_more_years_of_experience  -1.650  0.09892 .
## student_attendance_rate                                10.641  < 2e-16 ***
## percent_of_students_chronically_absent                 -2.092  0.03645 *
## teacher_attendance_rate                                 11.000  < 2e-16 ***
## regents_examChinese                                     -3.075  0.00211 **
## regents_examCommon Core Algebra                         -0.757  0.44921
## regents_examCommon Core Algebra2                        -1.410  0.15844
## regents_examCommon Core English                         -2.658  0.00787 **
## regents_examCommon Core Geometry                         0.082  0.93481
## regents_examEnglish                                     -1.615  0.10624
## regents_examFrench                                      -2.388  0.01695 *
## regents_examGeometry                                    -1.043  0.29679
## regents_examGlobal History and Geography                -1.117  0.26386
## regents_examIntegrated Algebra                          -1.918  0.05514 .
## regents_examItalian                                     -2.777  0.00548 **
## regents_examLiving Environment                          -1.582  0.11361
## regents_examPhysical Settings/Chemistry                 -0.692  0.48901
## regents_examPhysical Settings/Earth Science             -0.246  0.80573
## regents_examPhysical Settings/Physics                   -1.653  0.09835 .
## regents_examSpanish                                     -2.793  0.00523 **
## regents_examU.S. History and Government                 -2.557  0.01055 *
## percent_scoring_below_65                               -10.805  < 2e-16 ***
## participation_share                                      7.153 8.52e-13 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 13055.8  on 9994  degrees of freedom
## Residual deviance:  9196.5  on 9955  degrees of freedom
## AIC: 9276.5
##
## Number of Fisher Scoring iterations: 6
```

```
log_preds <- predict(log_fit_new, newdata = test18_new, type = "response")

alpha <- 0.5
log_preds2_new <- ifelse(log_preds > alpha, 1, 0)
confusionMatrix(factor(log_preds2_new), test18_new$passing)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0  566  262
##          1  335 1336
##
##                Accuracy : 0.7611
##                  95% CI : (0.7439, 0.7777)
##     No Information Rate : 0.6395
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.4726
##
##  Mcnemar’s Test P-Value : 0.003211
##
##             Sensitivity : 0.6282
##             Specificity : 0.8360
##          Pos Pred Value : 0.6836
##          Neg Pred Value : 0.7995
##              Prevalence : 0.3605
##          Detection Rate : 0.2265
##    Detection Prevalence : 0.3313
##       Balanced Accuracy : 0.7321
##
##        'Positive' Class : 0
##
```

```
hist(log_preds2_new)
```



We can remove more features since we didn't observe a change in accuracy.

```
set.seed(1234)
# Create new Training and Testing Data
school_new <- school18_new %>%
  select(-average_grade_8_english_proficiency, -percent_in_temp_housing)

split <- sample(1:nrow(school_new), 0.8*nrow(school_new), replace = F)

train_new <- school_new[split,]
test_new <- school_new[-split,]

str(train_new)
```

```
## ’data.frame’:    9995 obs. of  20 variables:
##  $ year                                                  : Factor w/ 4 levels "2015","2016",..: 1 4
##  $ enrollment                                            : num  699 1174 2215 437 2225 ...
##  $ average_grade_8_math_proficiency                      : num  2.18 2.5 2.27 2.37 2.68 2 2.07 2.28
##  $ percent_english_language_learners                     : num  0.067 0.124 0.196 0.057 0.192 0.197
##  $ percent_students_with_disabilities                    : num  0.203 0.114 0.161 0.13 0.133 0.333 0
##  $ percent_self-contained                                : num  0.043 0.023 0.051 0.005 0.042 0 0.03
##  $ economic_need_index                                   : num  0.751 0.67 0.618 0.503 0.702 0.809 0
##  $ percent_hra_eligible                                  : num  0.597 0.549 0.332 0.343 0.453 0.587
##  $ percent_asian                                         : num  0.039 0.297 0.342 0.087 0.333 0.06 0
##  $ percent_black                                         : num  0.279 0.441 0.223 0.705 0.313 0.283
##  $ percent_hispanic                                      : num  0.619 0.168 0.352 0.124 0.208 0.622
##  $ years_of_principal_experience_at_this_school          : num  4 6 5 6.8 1 2.3 4 3.9 9.9 6.4 ...
##  $ percent_of_teachers_with_3_or_more_years_of_experience: num  0.732 0.936 0.602 0.455 0.644 0.733
##  $ student_attendance_rate                               : num  0.792 0.875 0.857 0.846 0.867 0.843
##  $ percent_of_students_chronically_absent                : num  0.583 0.355 0.387 0.396 0.23 0.404 0
##  $ teacher_attendance_rate                               : num  0.972 0.96 0.964 0.966 0.97 0.965 0.
##  $ regents_exam                                          : Factor w/ 18 levels "Algebra2/Trigonometr
##  $ percent_scoring_below_65                              : num  51.9 28 0 44.4 2.7 59.6 77.8 42.9 0
##  $ participation_share                                   : num  99 224.7 385.8 78.4 353.4 ...
##  $ passing                                               : Factor w/ 2 levels "0","1": 1 1 2 1 2 2 2
```

Run the model again.

```
log_fit_new <- glm(passing ~ ., data = train_new, family = binomial(link = "logit"))
summary(log_fit_new)
```

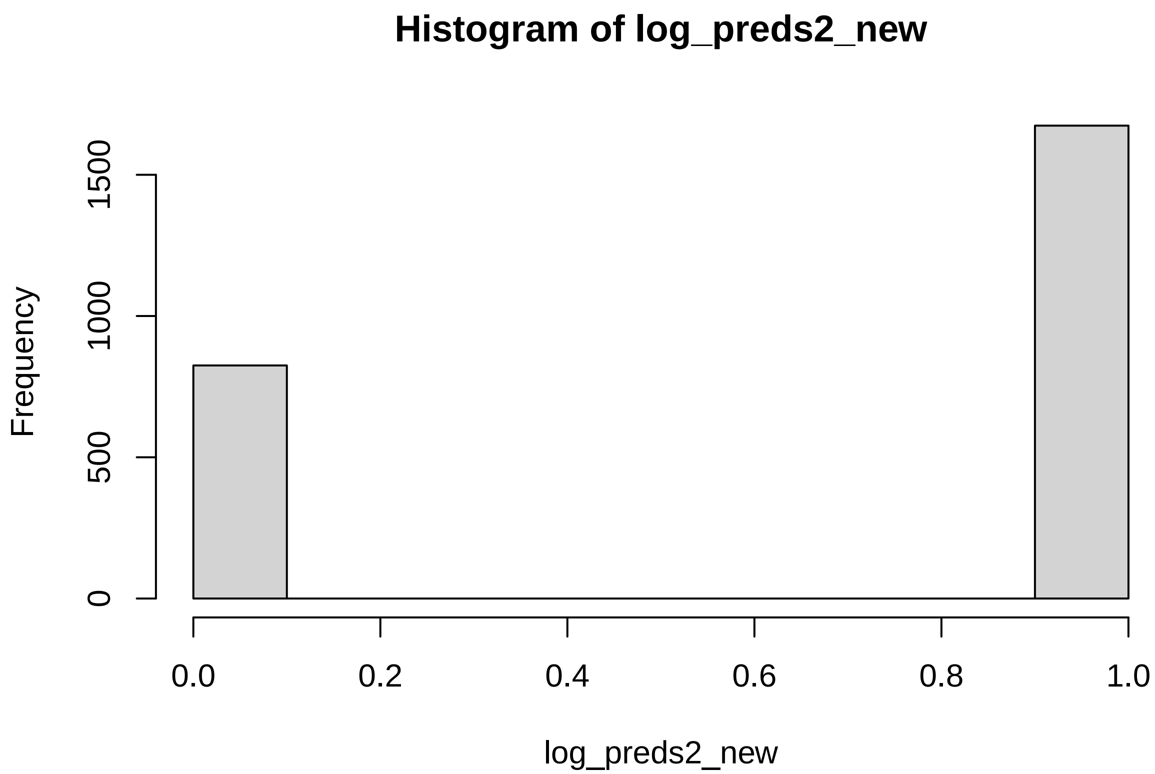
```
##
## Call:
## glm(formula = passing ~ ., family = binomial(link = "logit"),
##     data = train_new)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4762  -0.8060   0.2912   0.7213   2.8300
##
## Coefficients:
##                                                           Estimate Std. Error
## (Intercept)                                            -5.226e+01  3.493e+00
## year2016                                               -7.279e-01  8.273e-02
## year2017                                               -9.836e-01  9.057e-02
## year2018                                               -1.620e+00  1.204e-01
## enrollment                                             -1.284e-03  1.457e-04
## average_grade_8_math_proficiency                        9.339e-01  1.864e-01
## percent_english_language_learners                       1.751e+00  3.928e-01
## percent_students_with_disabilities                      5.236e+00  7.894e-01
## ‘percent_self-contained‘                               -1.194e+01  1.108e+00
## economic_need_index                                    -4.143e+00  6.799e-01
## percent_hra_eligible                                    4.223e+00  6.302e-01
## percent_asian                                           3.939e+00  4.578e-01
## percent_black                                           2.206e+00  3.410e-01
## percent_hispanic                                        2.699e+00  3.600e-01
## years_of_principal_experience_at_this_school            3.829e-02  6.757e-03
## percent_of_teachers_with_3_or_more_years_of_experience -3.403e-01  1.960e-01
## student_attendance_rate                                 2.125e+01  2.009e+00
## percent_of_students_chronically_absent                 -1.533e+00  6.972e-01
## teacher_attendance_rate                                 3.319e+01  2.985e+00
## regents_examChinese                                    -8.533e-01  2.763e-01
## regents_examCommon Core Algebra                        -1.135e-01  1.513e-01
## regents_examCommon Core Algebra2                       -2.325e-01  1.643e-01
## regents_examCommon Core English                        -4.120e-01  1.556e-01
## regents_examCommon Core Geometry                        1.190e-02  1.536e-01
## regents_examEnglish                                    -2.893e-01  1.795e-01
## regents_examFrench                                     -5.551e-01  2.307e-01
## regents_examGeometry                                   -1.993e-01  1.907e-01
## regents_examGlobal History and Geography               -1.696e-01  1.526e-01
## regents_examIntegrated Algebra                         -3.406e-01  1.780e-01
## regents_examItalian                                    -8.715e-01  3.141e-01
## regents_examLiving Environment                         -2.416e-01  1.529e-01
## regents_examPhysical Settings/Chemistry                -1.088e-01  1.564e-01
## regents_examPhysical Settings/Earth Science            -3.892e-02  1.544e-01
## regents_examPhysical Settings/Physics                  -2.994e-01  1.810e-01
## regents_examSpanish                                    -4.826e-01  1.733e-01
## regents_examU.S. History and Government                -3.969e-01  1.554e-01
## percent_scoring_below_65                               -1.307e-02  1.212e-03
## participation_share                                     6.892e-03  9.602e-04
##                                                        z value Pr(>|z|)
## (Intercept)                                            -14.960  < 2e-16 ***
## year2016                                                -8.798  < 2e-16 ***
## year2017                                               -10.860  < 2e-16 ***
## year2018                                               -13.457  < 2e-16 ***
## enrollment                                              -8.814  < 2e-16 ***
## average_grade_8_math_proficiency                         5.009 5.46e-07 ***
## percent_english_language_learners                        4.458 8.26e-06 ***
## percent_students_with_disabilities                       6.633 3.29e-11 ***
## ‘percent_self-contained‘                               -10.773  < 2e-16 ***
## economic_need_index                                     -6.094 1.10e-09 ***
## percent_hra_eligible                                     6.701 2.07e-11 ***
## percent_asian                                            8.605  < 2e-16 ***
## percent_black                                            6.470 9.80e-11 ***
## percent_hispanic                                         7.496 6.56e-14 ***
## years_of_principal_experience_at_this_school             5.666 1.46e-08 ***
## percent_of_teachers_with_3_or_more_years_of_experience  -1.736  0.08256 .  
## student_attendance_rate                                 10.574  < 2e-16 ***
## percent_of_students_chronically_absent                  -2.199  0.02786 *  
## teacher_attendance_rate                                 11.119  < 2e-16 ***
## regents_examChinese                                     -3.089  0.00201 ** 
## regents_examCommon Core Algebra                         -0.750  0.45304    
## regents_examCommon Core Algebra2                        -1.415  0.15702    
## regents_examCommon Core English                         -2.649  0.00808 ** 
## regents_examCommon Core Geometry                         0.077  0.93827    
## regents_examEnglish                                     -1.611  0.10707    
## regents_examFrench                                      -2.406  0.01614 *  
## regents_examGeometry                                    -1.045  0.29593    
## regents_examGlobal History and Geography                -1.111  0.26640    
## regents_examIntegrated Algebra                          -1.913  0.05572 .  
## regents_examItalian                                     -2.774  0.00553 ** 
## regents_examLiving Environment                          -1.580  0.11400
## regents_examPhysical Settings/Chemistry                 -0.695  0.48686
## regents_examPhysical Settings/Earth Science             -0.252  0.80094
## regents_examPhysical Settings/Physics                   -1.654  0.09814 .
## regents_examSpanish                                     -2.785  0.00535 **
## regents_examU.S. History and Government                 -2.554  0.01065 *
## percent_scoring_below_65                               -10.787  < 2e-16 ***
## participation_share                                      7.177 7.13e-13 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 13055.8  on 9994  degrees of freedom
## Residual deviance:  9197.9  on 9957  degrees of freedom
## AIC: 9273.9
##
## Number of Fisher Scoring iterations: 6
```

```
log_preds <- predict(log_fit_new, newdata = test_new, type = "response")

alpha <- 0.5
log_preds2_new <- ifelse(log_preds > alpha, 1, 0)
confusionMatrix(factor(log_preds2_new), test_new$passing)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0  567  258
##          1  334 1340
##
##                Accuracy : 0.7631
##                  95% CI : (0.7459, 0.7797)
##     No Information Rate : 0.6395
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.4766
##
##  Mcnemar’s Test P-Value : 0.002053
##
##             Sensitivity : 0.6293
##             Specificity : 0.8385
##          Pos Pred Value : 0.6873
##          Neg Pred Value : 0.8005
##              Prevalence : 0.3605
##          Detection Rate : 0.2269
##    Detection Prevalence : 0.3301
##       Balanced Accuracy : 0.7339
##
##        ’Positive’ Class : 0
##
```

```
hist(log_preds2_new)
```



Based on the two iterations of feature selection, we can now assume that the logistic regression model can achieve ~76% accuracy on our data sets.

Conclusion

Even though we removed a few features, the model didn't improve from its previous iteration. This could mean different things. It can mean that we need to remove more features or that we need a more powerful classifier model. The reason I chose logistic regression is because it's a classic classification model. It's easy to train and easy to tune. With more time, I would have liked to explore other classification models. A lot of iterations went into figuring out which dependent and independent variables worked best with logistic regression. What we can conclude from this model is that there are student demographic and test result features that influence a school's student achievement rating. However, there is no one specific feature that has an immense impact on student achievement ratings.

Critique

Group: Edgardo Zelaya and Julia Ulziisaikhan

The motivation for tackling their project was to see if the sentiment makeup of an article can be used to predict whether or not it is political. They used readily available New York Times article and comment data from a Kaggle competition. The data spanned from January to May 2017 and January to April 2018. I believe the mining portion of their project was 'mining' either positive, negative, or neutral sentiment from each comment, and then calculating the proportions of comments which had negative sentiment on a given article, and so on. They also engineered a polarization variable from these proportions, in order to capture quantitatively, which articles tended to have a high proportion of negative and positive sentiment and low proportion of neutral sentiment, so that the article attracted ‘polarized’ sentiment. Something I would have done differently would be to include more topics of article categorization. I think that the classification of articles into political and non-political bins can be too subjective, because there may be topics like abortion or reproductive health that may be considered political, but may have not been counted.