Course Project

Cesar Espitia

Duubar Villalobos

CUNY SPS Data 609

December 9, 2018

Table of Contents

[Abstract 3](#_Toc531722725)

[Course Project 4](#_Toc531722726)

[Data Exploration 4](#_Toc531722727)

[Summary Statistics 4](#_Toc531722728)

[Data Preparation 6](#_Toc531722729)

[Model Building 6](#_Toc531722730)

[Conclusion 9](#_Toc531722731)

[Appendix A: Summary Full 10](#_Toc531722732)

Abstract

This assignment focused on NY City Public School demographic data and SAT Test scores. The dataset contains 8867 records that encompass the entire school system. The variables for the data are school demographic variables such as number of enrollments, school names, breakout by grade and percentages, but the SAT data is only available for one of the 5 years. The purpose for this project is to analyze the data, perform any data manipulation / clean-up and use two (2) methods learned in the class which in this case was a generalized linear model to predict SAT Scores and then differential equations to predict the demographics and then repredict the SAT scores using the improved demographic data. The final model provided an AIC = 3530.8.

Keywords: NY City schools, SAT, data621

Course Project

The following is the analysis and write-up based upon my interpretation of the data and predict the average SAT scores based upon demographic school data.

# Data Exploration

The purpose of this step is to get a ‘feel’ for the dataset. The following information describes the data from different angles including completeness, statistical summaries, visuals to determine the shape and effect of each variable and other items deemed pertinent.

## Summary Statistics

The first step is to look at the data to determine some items including completeness and the shape of each variable. The following are the results of summarizing the data in a table and the visualization of each variables density function (PDF).

Table 1

Parital Summary Statistics for NY City School Demographic Data

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Min | | 1st Q | | Median | | Mean | | 3rd Q | | Max | | NA |
| Total.Enrollment | 1 | 319 | | 483 | | 599.6 | | 710 | | 5534 | | 49 | |
| Grade9 | 0 | 0 | | 0 | | 52.43 | | 69 | | 1457 | | 49 | |
| Grade10 | 0 | 0 | | 0 | | 50.93 | | 67 | | 3692 | | 49 | |
| Grade11 | 0 | 0 | | 0 | | 39.15 | | 44 | | 1393 | | 49 | |
| Grade12 | 0 | 0 | | 0 | | 38 | | 39 | | 1380 | | 49 | |
| Female | 0 | 143 | | 238 | | 291.4 | | 357 | | 2356 | | 49 | |
| Female.1 | 0 | 46.3 | | 48.8 | | 48.46 | | 51.4 | | 100 | | 49 | |

Note: Source: NY Data portal

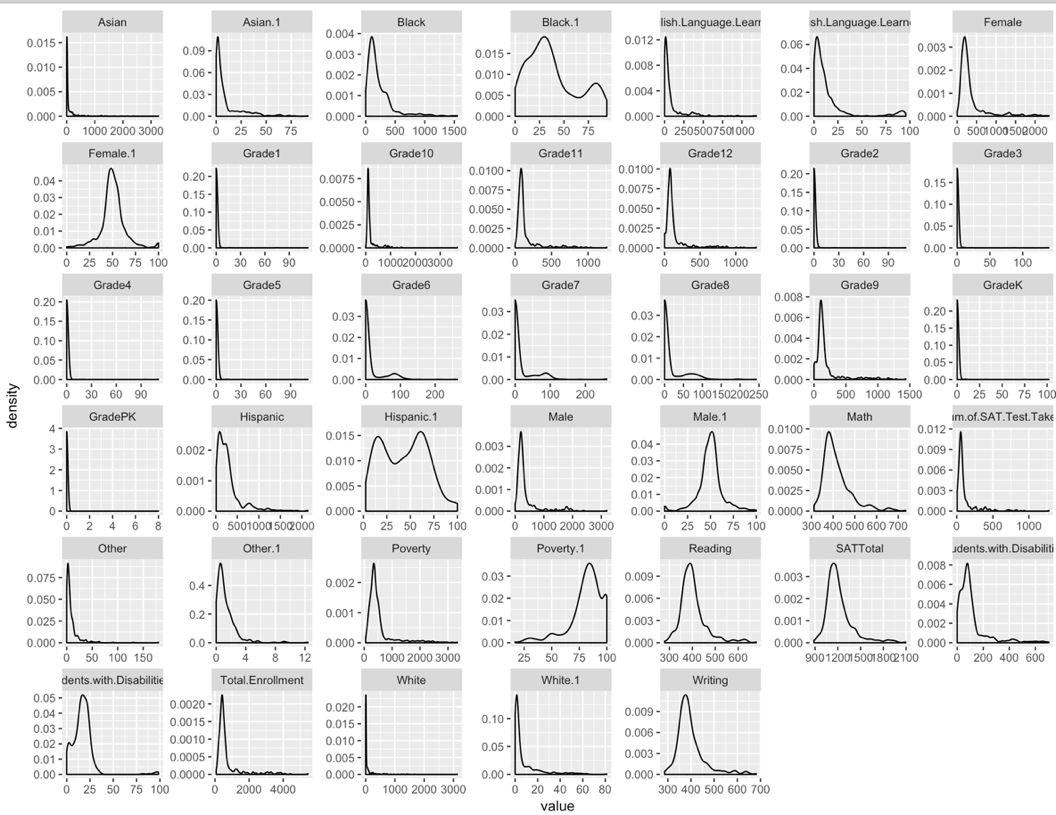
**

Figure 1. PDF for Each Dataframe Variable.

In looking at both, Table 1, Figure 1 and Appendix B (correlation matrix) together, we can note specific items that may skew our model building results. In this model, there was 10% of the data that was NA/null.

*PDF:*Figure 1 shows the PDF of each variable, this allows us to see if the data is normal or night. mean to remove the effects of severe skewness. All other variables were left as is because the shape didn’t warrant it.

*Correlation:* We look for correlated variables that we can make decisions on and determine which variable might be closely related to others either due to collinearity or other underlying factors that are visible at first glance in the dataset. The following variables were removed "Female","Male","Asian","Black","Hispanic","Other","White","Students.with.Disabilities","English.Language.Learners","Poverty","Grade4","Grade5","GradePK","GradeK","Grade1","Grade2","Grade3","Grade6","Grade7","Grade8". The demographic ones were counts that also had percentages. For the grades they did not add value to the model as they don’t affect the value of SAT scores.

# Data Preparation

The purpose of this step is to take the findings from the exploration and transform the data as needed. The following information describes the transformations done in order to prepare the data for model building and model selection.

For this analysis, 10% of the data had NAs and were imputed using the mean of the dataset. No variables had any transformations due to any sever skewness in the PDF graphs above. No new variables were created as there was nothing that was missing in the dataset. With this in mind, no secondary correlation check was done.

# Model Building

The purpose of this step is to take the modified dataset and begin exploring potential models that will be used on the final dataset provided. The following information describes the two (2) models built for this step and the relevant analysis to provide reasons for model selection in the next step.

**MODEL 1**

The first model takes in the data as manipulated in step two. In this first model, we have an AIC of 251.59. The data in Table 2, shows that the model has an accuracy of 90.1%.

Call:

glm(formula = SATTotal ~ . - Math - Writing - Reading, family = gaussian(),

data = train)

Deviance Residuals:

Min 1Q Median 3Q Max

-264.276 -53.580 0.231 48.382 307.101

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 24612.50060 85423.53951 0.288 0.77346

Total.Enrollment -0.01898 0.04443 -0.427 0.66952

Grade9 -0.21360 0.07170 -2.979 0.00314 \*\*

Grade10 0.01475 0.04905 0.301 0.76385

Grade11 -0.13444 0.12434 -1.081 0.28051

Grade12 0.01908 0.11330 0.168 0.86637

Female.1 -371.57014 857.15540 -0.433 0.66499

Male.1 -371.49222 857.16270 -0.433 0.66506

Asian.1 144.99326 74.91238 1.936 0.05392 .

Black.1 140.02171 74.91314 1.869 0.06264 .

Hispanic.1 141.41140 74.92149 1.887 0.06012 .

Other.1 145.69256 74.72295 1.950 0.05219 .

White.1 142.40984 74.92992 1.901 0.05837 .

Students.with.Disabilities.1 -0.11318 0.50653 -0.223 0.82335

English.Language.Learners.1 -2.83135 0.33174 -8.535 0.000000000000000869 \*\*\*

Poverty.1 -4.13095 0.57744 -7.154 0.000000000007240609 \*\*\*

Num.of.SAT.Test.Takers 0.60400 0.11378 5.308 0.000000223826564019 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 7114.704)

Null deviance: 8886824 on 299 degrees of freedom

Residual deviance: 2013461 on 283 degrees of freedom

**AIC: 3530.8**

Number of Fisher Scoring iterations: 2

|  |
| --- |
|  |
| Comparison of Predicted Values |

Figure 2. Histogram of Model 1 Prediction of SAT Total Scores.

No variables seem in terms of predictability and therefore no values will be removed for the second method.

**MODEL 2**

The second model only first predicts the demographic data more accurately using the method 1.2 in the text book prior to predicting SAT Test Scores. This model had an AIC of :

DUUBAR INSERT YOUR GLM RESULTS HERE.

|  |
| --- |
|  |
| Comparison of Predicted Values |

Figure 3. Histogram of Model 2 Prediction of SAT Total Scores.

In this model, the created variable ***new*** no longer has significance due to the fact that ***medv*** and ***tax*** are still in the model.

**TEST DATA**

The dataset had 129 entries and 21 columns and was modified to fit the final variables and scaling used in Model 2 from above.

# Conclusion

Three two (2) models were presented after exploring and manipulating the data as necessary. With using a multi-criteria approach for this exercise, it became clear that the Model 2 was selected and provided an AIC of X which was adequate for the data but doesn’t necessarily indicate the best model if it were solely based upon AIC (Model 1 would have been chosen) which is the equivalent of R-squared for binary regression models. If more time were available, the creation of other new variables that were not correlated could have been generated with better insight into the data set.

Appendix A: Summary Full

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Min | | 1st Q | | Median | | Mean | | 3rd Q | | Max | | NA |
| Total.Enrollment | 1 | 319 | | 483 | | 599.6 | | 710 | | 5534 | | 49 | |
| GradePK | 0 | 0 | | 0 | | 14.03 | | 30 | | 967 | | 49 | |
| GradeK | 0 | 0 | | 33 | | 47.73 | | 85 | | 406 | | 49 | |
| Grade1 | 0 | 0 | | 33 | | 49.31 | | 88 | | 383 | | 49 | |
| Grade2 | 0 | 0 | | 27 | | 47.75 | | 86 | | 349 | | 49 | |
| Grade3 | 0 | 0 | | 3.5 | | 46.1 | | 83 | | 373 | | 49 | |
| Grade4 | 0 | 0 | | 0 | | 44.19 | | 80 | | 380 | | 49 | |
| Grade5 | 0 | 0 | | 0 | | 43.34 | | 79 | | 398 | | 49 | |
| Grade6 | 0 | 0 | | 0 | | 42.75 | | 62 | | 812 | | 49 | |
| Grade7 | 0 | 0 | | 0 | | 42.06 | | 58 | | 819 | | 49 | |
| Grade8 | 0 | 0 | | 0 | | 41.79 | | 56 | | 839 | | 49 | |
| Grade9 | 0 | 0 | | 0 | | 52.43 | | 69 | | 1457 | | 49 | |
| Grade10 | 0 | 0 | | 0 | | 50.93 | | 67 | | 3692 | | 49 | |
| Grade11 | 0 | 0 | | 0 | | 39.15 | | 44 | | 1393 | | 49 | |
| Grade12 | 0 | 0 | | 0 | | 38 | | 39 | | 1380 | | 49 | |
| Female | 0 | 143 | | 238 | | 291.4 | | 357 | | 2356 | | 49 | |
| Female.1 | 0 | 46.3 | | 48.8 | | 48.46 | | 51.4 | | 100 | | 49 | |
| Male | 0 | 159 | | 249 | | 308.2 | | 370 | | 3250 | | 49 | |
| Male.1 | 0 | 48.6 | | 51.2 | | 51.54 | | 53.7 | | 100 | | 49 | |
| Asian | 0 | 4 | | 14 | | 92.78 | | 70 | | 3340 | | 49 | |
| Asian.1 | 0 | 1.1 | | 3.5 | | 10.72 | | 12 | | 94.7 | | 49 | |
| Black | 0 | 47 | | 115 | | 169.5 | | 218 | | 1552 | | 49 | |
| Black.1 | 0 | 9.8 | | 27.8 | | 34.48 | | 55.8 | | 98.8 | | 49 | |
| Hispanic | 1 | 79 | | 169 | | 241.1 | | 311.8 | | 2478 | | 49 | |
| Hispanic.1 | 0.4 | 18.9 | | 38.2 | | 41.45 | | 62.4 | | 100 | | 49 | |
| Other | 0 | 2 | | 5 | | 9.514 | | 11 | | 310 | | 49 | |
| Other.1 | 0 | 0.5 | | 1 | | 1.661 | | 2 | | 38.3 | | 49 | |
| White | 0 | 4 | | 12 | | 86.69 | | 72 | | 3230 | | 49 | |
| White.1 | 0 | 1 | | 2.6 | | 11.69 | | 13.5 | | 93.6 | | 49 | |
| Students.with.Disabilities | 0 | 56 | | 89 | | 110.5 | | 135 | | 842 | | 49 | |
| Students.with.Disabilities.1 | 0 | 13.7 | | 18 | | 20.79 | | 23.1 | | 100 | | 49 | |
| English.Language.Learners | 0 | 16 | | 40 | | 82.27 | | 99 | | 1233 | | 49 | |
| English.Language.Learners.1 | 0 | 4.1 | | 9 | | 13.36 | | 17.7 | | 100 | | 49 | |
| Poverty | 1 | 242 | | 383 | | 479.1 | | 584 | | 3842 | | 49 | |
| Poverty.1 | 3.3 | 75.3 | | 87.9 | | 81.93 | | 97.8 | | 100 | | 49 | |
| Num.of.SAT.Test.Takers | 6 | 41 | | 62 | | 110.3 | | 95 | | 1277 | | 8446 | |
| Math | 312 | 371 | | 395 | | 413.4 | | 437 | | 735 | | 8446 | |
| Writing | 286 | 360 | | 381 | | 394 | | 411 | | 682 | | 8446 | |
| Reading | 279 | 368 | | 391 | | 400.9 | | 416 | | 679 | | 8446 | |
| SATTotal | 887 | 1102 | | 1169 | | 1208 | | 1257 | | 2096 | | 8446 | |