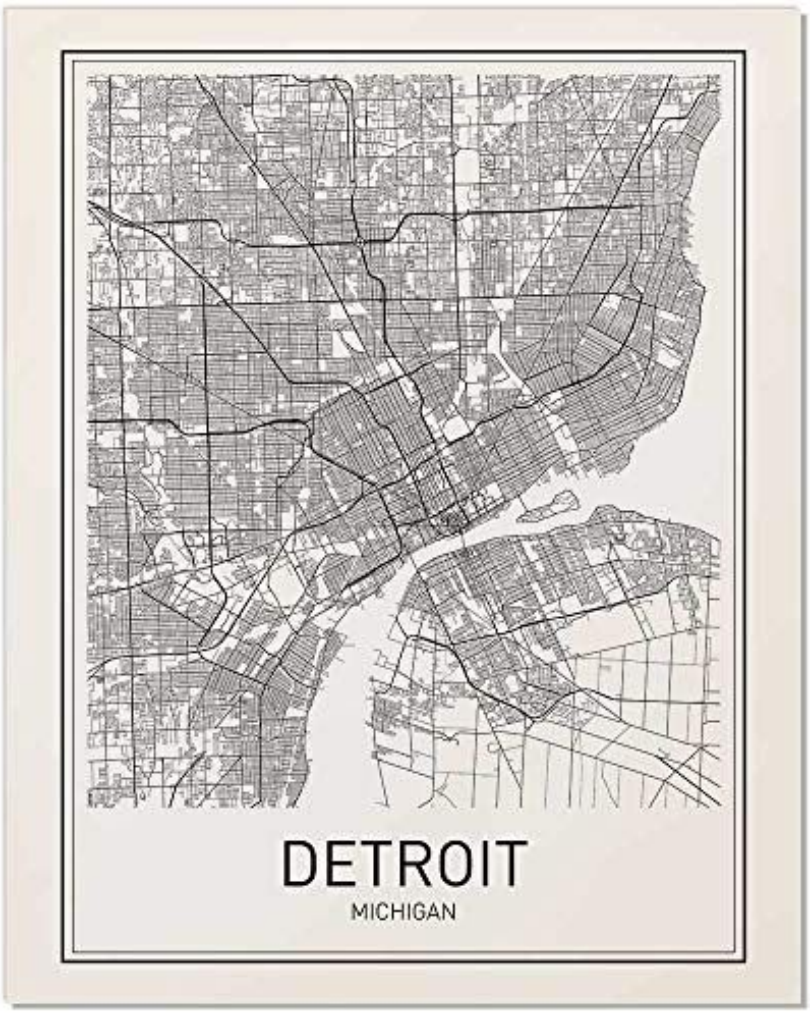
Detroit College Enrollment

Final Project for STATS 415

Clare Affinito, Justin Ritenburgh, Daniel Thompson



Introduction

When deciding what questions we wanted to investigate, we wanted to explore something relevant to us and our peers. Since our University is located in Ann Arbor, which is right outside Detroit, we decided it would be interesting to use data from Detroit. When exploring Detroit public data, we discovered a data set regarding college enrollment in different areas of Detroit. We shaped our questions around this data set because exploring what impacts college enrollment in Detroit seemed important, especially for college students in a neighboring city.

When choosing data sets from Data Driven Detroit, a database of publicly-available data related to Detroit and Michigan, we looked for a few things. First, we knew that we wanted data that we thought could be relevant to young people and education. Next, we knew that we wanted all of our data to be from the same time period, 2016-2017, to most accurately represent the relationships that we uncovered; we decided not to use several interesting data sets because they were a few years younger or older than the rest of what we wanted to use. Finally, we looked for data sets that had columns in common so that we could combine them into one master data frame that related all of our data sets based on the Zip Code and GEOID where it was collected.

Our goal was to be able to predict the percentage of college enrollment on the block group level based on predictors revolving Elevated Blood Lead Levels, which are most likely caused by lead in water, overall state of health of the Detroit community, and attitude towards healthcare in the Detroit area. Using techniques like Linear Regression, Lasso Methods, and KNN, we were able to attempt to predict college enrollment. In this process, we encountered many difficulties, but we were able to personally discover a lot through data exploration and use of these methods. Our observations are grouped by block group, not other factors such as socioeconomic standing or investment in education.

Data Sets

We combined a collection of data sets from Data Driven Detroit in order to create one large dataset that would be useful to answer the questions we had. Those datasets were (as stated by DDD):

- **College Enrollment:** This dataset contains college enrollment information for students who graduated in 2017. It defines enrollment as those who enrolled in a college or university (two-year or four-year institution) within 12 months of highschool graduation.

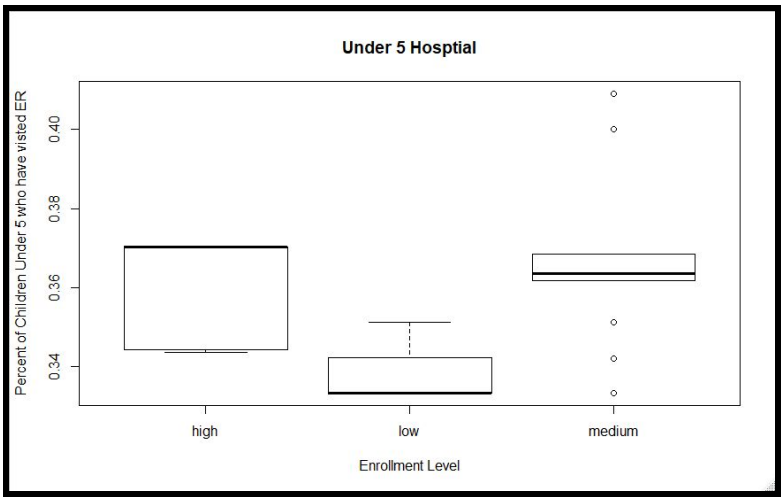
- **Lead Blood Levels:** This dataset contains lead blood levels for the state of Michigan in 2017. **Elevated Blood levels are blood lead levels above 4.5 micrograms of lead per deciliter of blood.** Null values were used when no blood lead testing was completed or if there was a need to protect the privacy of the individual.
- **Medicaid:** This dataset contains Medicaid data of patients under 18 in Michigan in 2017.

All of the data sets share a common column containing block group, a census defined geographic area within zip codes. The three data sets, as found on Data Driven Detroit, contain observations from the same 8205 block groups within the state of Michigan, although not in the same order. This makes them ideal candidates for analysis across each other. Because the scope of this project is just the city of Detroit, the data set was reduced to the 487 block groups contained by Detroit's 30 zip codes.

The major challenge concerning the data is that every predictor contains some missing data, and the amount of missing data varies from predictor to predictor. This left multiple opportunities for addressing the issue. Mean imputation was considered, although was not used for a variety of reasons. One being that in some cases, over 90% of observations had missing data. This would cause an overwhelming majority of all values to be the mean of a relatively small sample of the data. More concerningly, we cannot faithfully operate under the assumption that the observations with missing data came from the same distribution as the data without missing values. The source of the data stated that some values were omitted from the datasets over concerns for privacy and some variables were absent because of a lack of medical tests such as drawing blood. There certainly could be demographic differences between residents of Detroit who had submitted blood for testing and those who did not, so it would be unwise to assume that the mean of the data for Detroit residents whose data was available would be the same as those whose data was not recorded. Ultimately, by calling the R function na.omit() on the combined data set containing College Enrollment, Lead Blood Levels, and Medicaid. This ultimately reduced the overall sample to 95 observations.

One more unique issue that arose while making visualizations was the fact that many ratios occurred frequently in the data set. This is because there are 487 block groups within Detroit (population ~ 675,00) alone, so the average block group will have less than 1,400 residents. For statistics such as the percentage of children under 5 years old who have been admitted to the emergency room, the percentages can be drawn from samples of already small populations, causing common fractions to occur repeatedly. This was magnified by the fact that the range of block values for this statistic was no less than 0.3 and no greater than 0.42 for any

studied block group. For instance 0.33333 and 0.40000 were common values for the percentage of children under 5 who had been to the ER in a block group, likely because there may have been only 9, 10, 12, or 20 children in that age category in a block group and equivalent fractions such as 3/9 and 4/12 or 4/10 and 8/20 occurred frequently. This caused problems for scatter plot and boxplot visualizations, because those visualizations would make the plots appear to be displaying far less data than was actually being used in numerical analysis. The jitter() function in the ggplot2 package for R was a convenient solution to this, in order to make the visualizations still accurately represent the size of the data.



Due to some values occurring at a disproportionately high rate, box plots became very difficult to interpret, as the median and upper or lower quartiles were all indistinguishably similar or the same.

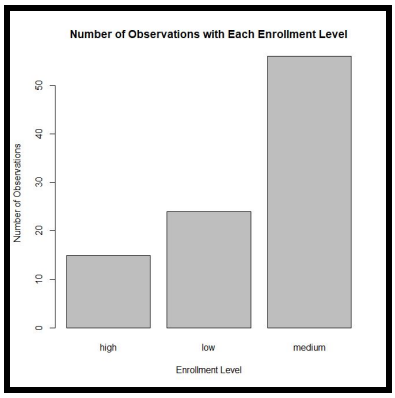
Observations

*Descriptions as stated by Detroit Driven Data

We originally had 4,822 observations that we brought down to 95 observations for privacy and to counteract problems with data collection. These observations had 29 variables that we chose to explore in order to gain the fullest possible picture of our data set. All of which are listed below:

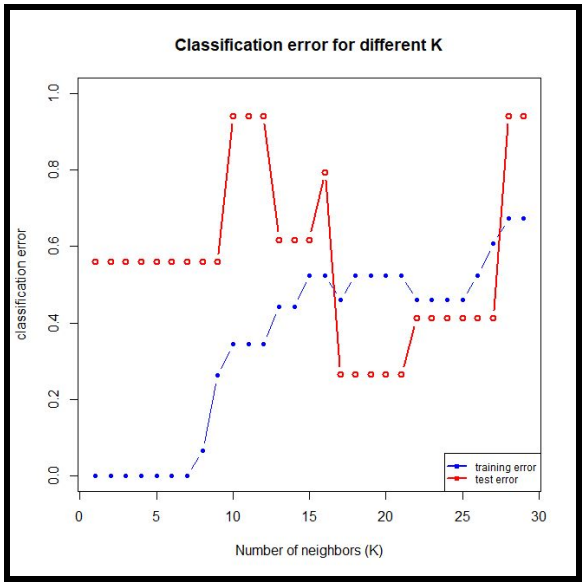
- **Total Grad:** Number of public high school graduates who received a diploma during the high school graduation year.
- **Total Enrollment:** Number of public high school graduates who enrolled in college.
- **Percent Enrollment**: Percent of public high school graduates who enrolled in college.
- **Under 18 Total:** Total number of hospital or ER visits through Medicaid for children (under 18)
- **Under 5 Total:** Total number of hospital or ER visits through Medicaid for children under 5
- **Under 5 ER:** Total number of ER visits through Medicaid for children under 5
- **Under 5 Hospital:** Total number of hospital visits through Medicaid for children under 5

- **Over 5 Total:** Total number of hospital or ER visits through Medicaid for children over 5
- **Over 5 ER:** Total number of ER visits through Medicaid for children over 5
- **Total ER Visits:** Total number of ER visits through Medicaid for children (under 18)
- **Total Hospital Visits:** Total number of hospital visits through Medicaid for children under 18
- **Asthma Count:** Total number of hospital or ER visits for children through Medicaid related to Asthma
- **Diabetes 1 Count:** Total number of hospital or ER visits for children through Medicaid related to Type 1 Diabetes
- **Diabetes 2 Count:** Total number of hospital or ER visits for children through Medicaid related to Type 2 Diabetes
- **Diabetes Other Count:** Total number of hospital or ER visits for children through Medicaid related to other types of Diabetes (excluding Type 1 and 2)
- **Percentage Under 5 years old ER:** Percent of total children under 5 who visited the ER through Medicaid
- **Percentage Over 5 years old ER:** Percent of total children over 5 who visited the ER through Medicaid
- **Percentage Asthma:** Percent of total children (under 18) who visited the ER through Medicaid related to Asthma
- **Percentage Diabetes:** Percent of total children (under 18) who visited the ER through Medicaid related to Diabetes
- **CNT Tested:** Number of individuals who were tested.
- **Elevated Blood Lead Levels:** Number of individuals with an elevated blood lead level, defined as > 4.5 µg/dL.
- **Under 6 years old CNT Tested:** Number of individuals, under 6 years of age, who were tested.
- **Under 6 years old Elevated Blood Lead Levels:** Number of individuals, under 6 years of age, with an elevated blood lead level, defined as > 4.5 µg/dL.
- **Under 18 years old CNT Tested:** Number of individuals, under 18 years of age, who were tested.
- **Under 18 years old Elevated Blood Lead Levels:** Number of individuals, under 18 years of age, with an elevated blood lead level, defined as > 4.5 µg/dL.
- **Percentage Elevated Blood Levels:** Percent of individuals with an elevated blood lead level (defined as $>$ 4.5 µg/dL).
- **Percentage under 6 years old Elevated Blood Lead Levels:** Percent of individuals, under 6 years of age, with an elevated blood lead level (defined as > 4.5 µg/dL).
- **Percentage under 18 years old Elevated Blood Lead Levels:** Percent of individuals, under 18 years of age, with an elevated blood lead level (defined as > 4.5 µg/dL).

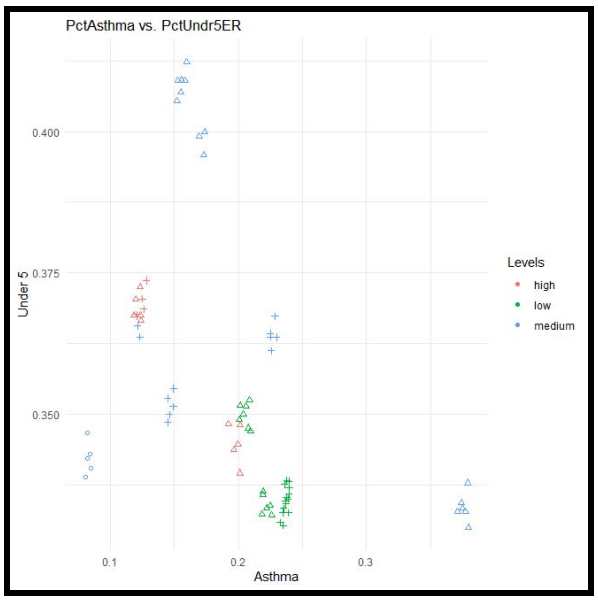


For our classification methods, we turned the percentage of students enrolled in higher education into a factor with three levels: low, medium, and high, corresponding to block groups with less than 35%, between 35% and 65%, and greater than 65% enrollment, respectively.

KNN for Classification



In order to predict which level any given block group would fall into, we separated the data into a training set making up 65% of the data and a test set making up 35% of the data. We then performed KNN many times to find the K that gave us the best result for our test data. K = 17 performed the best, giving us a classification error of 0.265 on our test data. We also tried other bounds for our levels and sizes for our training set, but we ran into issues with the test set being too small and not having each level represented, leading to similar results as if we had naively predicted every observation to belong to the most common level.



Displayed is a plot of the percentage of children who went to the ER for asthma-related symptoms plotted against the percentage of children under 5 who had visited the ER. The colors represent the actual levels of enrollment and the shapes correspond to whether KNN correctly predicted the level for that observation, with triangles representing correct predictions.

Lasso for Variable Selection and Linear Regression

(Intercept)	-1.0887639932
TotGrad	.
TotEnrl	0.0011858756
Under18_Tot	.
Under5_Tot	.
Under5_ER	.
Under5_Hosp	.
FivePlus_Tot	0.0075622087
FivePlus_ER	.
Tot_ER_Visits	.
Tot_Hosp_Visits	.
AsthmaCt	.
Diabetes1Ct	.
Diabetes2Ct	.
DiabetesOtherCt	.
PctUndr5ER	4.4374333324
PctOvr5ER	-1.0547094733
PctAsthma	.
PctDiabetes	.
CntTested	0.0006098732
EBLL	-0.0339019154
Under6CntTested	.
Under6EBLL	-0.0042742874
Under18CntTested	0.0037414795
Under18EBLL	.
PctEBLL	.
PctUnder6EBLL	0.4299921351
PctUnder18EBLL	.

With so many variables, we decided that we should use one of our selection techniques to choose only predictors that are not scaled to zero when a penalty is applied for the number of predictors. When we applied the Lasso technique, we reduced our number of predictors by more than half, to just nine. We then performed both a Lasso Regression and a Simple Linear Regression on the data using the predictors selected by Lasso.

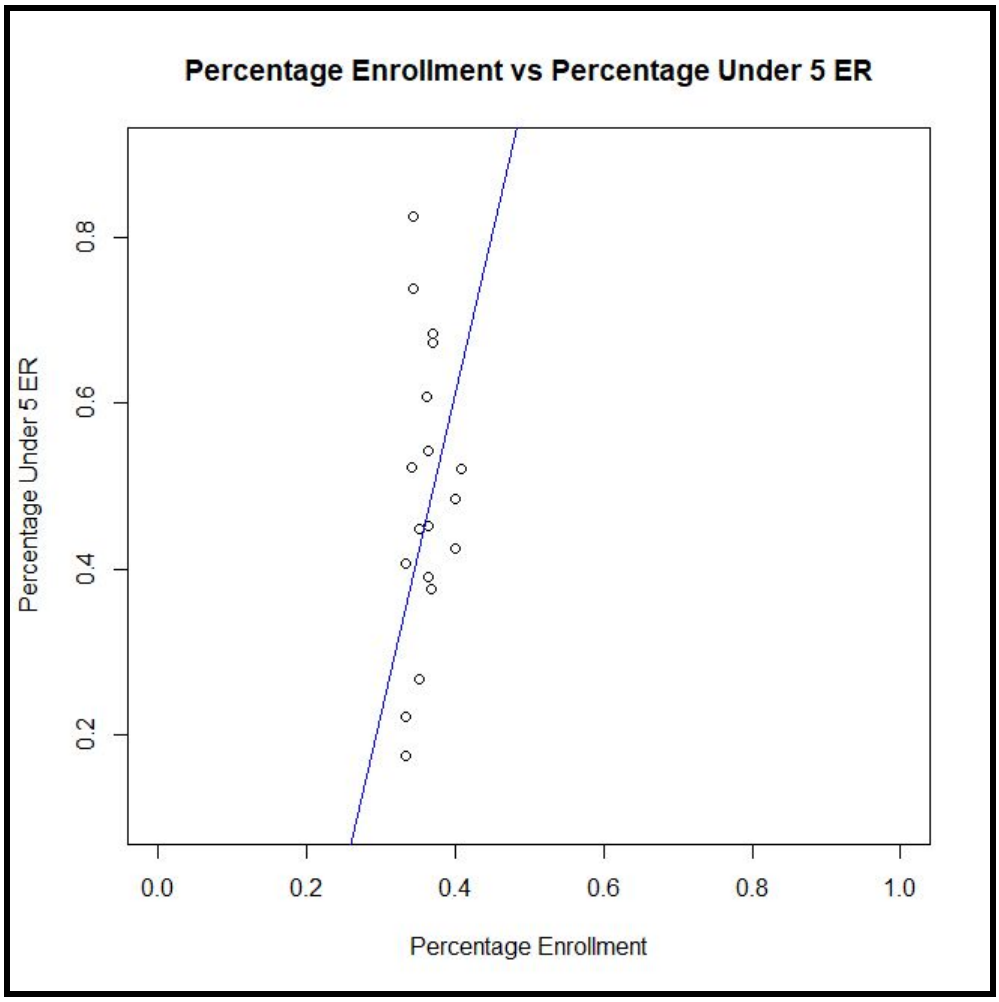
Method	Lasso	Linear Regression
Test MSE	0.038	0.096

Lasso Regression performed much better than Linear Regression with predictors chosen by Lasso, so shrinking the variables improved the accuracy of the prediction.

Simple Linear Regression on a Single Predictor

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.9453     0.3412  -2.771 0.007468 **
PctUndr5ER    3.8855     0.9637   4.032 0.000161 ***
```

We wanted a model that was more interpretable than the ones generated by KNN or by the Lasso methods, so we decided to choose from among the Lasso-chosen predictors a single one that we thought we could interpret if we built a model using it alone. Regressing the percentage enrollment in higher education on the percentage of children under 5 who had visited the ER in a given area surprisingly yielded a better result than either KNN or the Lasso methods, giving us an error of 0.029. Our interpretation of the model states that for every percent increase in the proportion of children under 5 who had visited the ER, the percentage enrollment increased by about 4%.



Limitations

Level determination:

Since there are no predefined ranges defining what is “Low”, “Medium”, and “High” enrollment, we had to determine a set of ranges that could be used to create levels for some of our analysis. Through testing different ranges for the Percentage of Enrollment, we were able to create proper levels for analysis, despite no predefined standard existing.

Data Collection:

Since we used Detroit Public Data, we were not able to control how the data was collected. When the data was collected originally, there were a lot of data points we could not use for either privacy reasons of the individual or improper collection. In order to make sure our data set followed basic statistical practice, we ultimately decided to remove those data points violating basic laws of data collection. We debated this at length and considered multiple different solutions, but removing those points from our data set was the best decision for our analysis.

Number of Variables:

We had a high number of variables, which may have made our analysis more accurate, but less interpretable.

Small Data Set:

Because we had to remove a lot of observations from our data set, we became more limited in making grand conclusions about college enrollment in Detroit.

Concluding Remarks

In attempting to predict college enrollment based on different predictors regarding Blood Lead Levels and Medicaid outcomes, we used a multitude of methods. After completing all of these methods, we discovered that our models lacked a level of interpretability that we desired. In order to make our explorations more interpretable, we performed linear regression on one interpretable predictor, ER visits for children under 5. In doing so, we achieved a strong level of predictability while improving interpretability. While we were faced with many limitations, we were able to create a solid model that was not only accurate, but easy to interpret.

Sources

Data Driven Detroit D3 Website
https://portal.datadrivendetroit.org/

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5629454/