Neighborhood Disadvantage and High School Dropout

Lerong Wang

June 2018

Abstract

This paper studies neighborhood disadvantages and high school dropout and uses neighborhood disadvantages to predict high school dropout rates in NYC neighborhoods. An exploratory spatial analysis is conducted first, and then by running three models, including Ordinary Least Squares, Random Forest, and Decision Tree, the paper compares the predictive performance of the three models. The result shows that overall, Random Forest has the best predictive performance for the feature set. Also, the three models all suggest that Hispanic population, which is a demographic control variable, is the most significant feature for predicting high school dropout rate.

Keywords: Neighborhood disadvantage, high school dropout, predictive models, spatial analysis

1 Introduction

Since the establishment of public school in the United States, obtaining a high school education has been viewed as the great equalizer to upward social mobility and opportunity. Educators and administrators in the US public school system have been dedicated to lowering high school dropout rates to ensure more equitable opportunities and outcomes for all students (Datiri,

2013). Dropping out of high school has been associated with negative psychological adjustment and negative life outcomes. Thus, studying high school dropout has become an increasingly important social psychology, sociology, and public policy question.

Many studies suggest that individual-level factors are associated with high school dropout. Yet social scientists have long considered the social contexts in which individuals develop to influence life outcomes. In particular, strong theoretical claims have been made that neighborhood conditions influence socioeconomic trajectories above and beyond individual factors (Massey and Denton 1993; Rendn, 2014). In this paper, I will focus on the neighborhood context that may affect high school dropout, and more specifically, my research question is: Can we use neighborhood disadvantage to predict high school dropout?

2 Literature Review

Researchers in different fields have investigated the factors affecting young peoples educational attainment for a long time. More recently, the focus has shifted to the influence of external factors on educational attainment such as how the characteristics of young peoples neighbors and neighborhood affect his or her schooling (Vartanian Gleason, 1999). Some previous articles

related to this topic have focused on how living in a poverty neighborhood may affect young peoples schooling experience. They have concluded that growing up in a relatively disordered and poor neighborhood is a potential risk factor for high school dropout. Rumberger (2013) has discussed in the article poverty and high school dropouts that some neighborhoods, particularly those with high concentrations of African Americans, are communities of concentrated disadvantages with high levels of joblessness, family instability, poor health, substance abuse, poverty, crime, and these neighborhoods influence child and adolescent development negatively, and may increase the likelihood of dropping out of school. Harding (2003) showed that when two groups of children are at the same age, but experience different neighborhoods during adolescence, those in high poverty neighborhoods are more likely to drop out of high school than those in low poverty neighborhoods. Measurement of neighborhood characteristics is an ongoing empirical challenge, and thus it is not surprising that poverty becomes the most prevalent feature for measuring neighborhood effects because data about some socio-economic characteristics can be easily obtained from Census data (Donnelly, 2017).

A more recent study conducted by Donnelly (2017) covered a more comprehensive range of neighborhood characteristics and comprised an index of neighborhood disadvantages. Donnelly analyzed the effect of neighborhood disadvantage on high school dropout as well as a multilevel analysis on the extent to which school, neighborhood, and peer group mediate the impact of

disadvantaged neighborhood. The neighborhood characteristics used in this study include the unemployment rate, proportion of persons with income below the poverty line, proportion of households headed by females with own children ages 18 years and younger, proportion persons age 25 and older with no high school diploma or equivalency and so on. Socio-demographic control variables were also added to the multivariate models that the study built, since numerous studies showed that race and ethnic segregation is associated with lower educational attainment in the long run (Goldsmith, 2009). For the associations between neighborhood disadvantage and school dropout, the results from Donnellys analysis showed that youth who reside in the most disadvantaged neighborhoods are more than twice as likely to drop out as those who live in the most affluent neighborhoods.

Prior research on how neighborhood conditions affect high school dropout has been mixed, and a variety of data sources, methods and measures have been used. Donnelly (2017) used data from the National Longitudinal Study of Adolescent Health, and all analyses were conducted at individual-level. The outcome variable, school dropout, was treated as a dichotomous variable in the study. Vartanian and Gleason (1999) used the Panel Study of Income Dynamics merged with information on neighborhood characteristics from the U.S. Census and built a logistic regression model to measure the effects of neighborhood conditions and family background on the likelihood of high school dropout. Zaff and Malone (2016) took a different perspec-

tive, and they analyzed why the rate of youth leaving high school in some neighborhoods has improved. They used the change in the rate of youth leaving school in neighborhood as their outcome variable. By integrating three different datasets: the Geolytics Inc. Neighborhood Change Database, the Business Master Files, and the Common Core of Data, they examined whether community environment can be implicated in the reduction in the rate of youth who leave school within neighborhoods in some metropolitan areas throughout the United States. Also, several social experiments have also been used to analyze neighborhood effects on high school dropout. The first such social experiment was the Gautreaux Program administered by the nonprofit Leadership Council for Metropolitan Open Communities in Chicago in 1981. As part of public assistance, some low-income black families were assigned to a neighborhood in a different city, which were mostly poor and black, or a neighborhood in suburban, which was less poor and predominately white. The longitudinal research found that by the time children became young adults, only 5 percent of the suburban movers dropped out of high school, but 20 percent of those who moved to mostly poor and black neighborhood have dropped out (Rosenbaum, 1995).

My study will mirror that of Donnelly (2017), but I will make some adjustments and my contributions. First, my outcome variable will be the high school dropout rate in the neighborhood instead of the dichotomous variable in Donnellys study. Second, I plan to mirror some of the neighborhood char-

acteristics that Donnelly used to construct her neighborhood disadvantage index and her method to add socio-demographic control variables because I think by far, her research considered the most comprehensive neighborhood characteristics and possible confounding variables. I plan to do a model comparison using different statistical models such as linear regression and random forest. Finally, I hope to add a spatial analysis such as local spatial autocorrelation before doing the statistical analysis to detect possible outliers and clusters.

3 Data

3.1 Data Source

- Education data and socio-demographic characteristics for New York City Census Tracts Data come from American Community Survey 2008-2012, US Census Bureau. I downloaded a compiled version from Geoda data and lab website. Neighborhood characteristics are contained in this data and high school dropout rate is computed using the variables contained in this data.

- NYC Neighborhood Tabulation Area Shapefile Data. Downloaded from `https://www1.nyc.gov/site/planning/data-maps/open-data/dwn-nynta.`

page. The shapefile provided a boundary file for neighborhood area so that we can create a base map for the NYC Neighborhood Tabulation Area.

3.2 Data Processing

A list of raw variables from the data that will be used for this project is provided below. Each observation in the original data corresponds to one census tract in New York City. There are 2166 observations in total. I performed spatial aggregation in GeoDa and combined the individual observations within the same neighborhood tabulation area (have same ntacode) and aggregated the sum of each variable (male16to19, maledrop, female16to19, popunemplp, popinlabou, poptot, households, withpubass, poor and so on). Median household income is aggregated by computing the average median household income of different census tracts within the same neighborhood tabulation area. The descriptions of the variables are shown in Table 1.

high school dropout rate is calculated using

$$\frac{\text{maledrop} + \text{femaledrop}}{\text{male16to19} + \text{female16to19}}$$

unemployment rate is calculated using

$$\frac{\text{popunemplo}}{\text{popinlabou}}$$

poverty rate is calculated using

$$\frac{\text{poor}}{\text{poptot}}$$

proportion of households with public assistance income is calculated using

$$\frac{\text{withpubass}}{\text{households}}$$

The proportions of female population, male population, and people of different races are calculated separately using

$$\frac{\text{variable}}{\text{poptot}}$$

.

The summary statistics table of the variables that will be used to build the model is shown in Table 2, and Figure 1 shows the high school dropout rates distribution of each neighborhood tabulation area after performing spatial aggregation.

Variable Name	Desciptions
ntaname	New York City NTA (Neighborhood Tabulation Area) name
ntacode	Code associatd with the NTA
male16to19	Male Civilian Population 16 To 19 Years
maledrop	Male Not high school graduate, not enrolled (dropped out) - 16-19 age band
femal16to19	Female Civilian Population 16 To 19 Years
femaledrop	Female Not high school graduate, not enrolled (dropped out) - 16-19 age band
popunemplo	Unemployed total population
popinlabou	Total population in labour force
poptot	Total population
households	Total Households
withpubass	Households With Public Assistance Income
medianinco	Median household income (In 2012 Inflation Adjusted Dollars)
poor	Doing poorly as regard Ratio Of Income In 2012 To Poverty Level (Under 1.00)
female	Total Population Female
male	Total Population Male
pacific	Total Population Pacific Islander
mixed	Total Population Mixed race
hispanic	Total Population Hispanic
european	Total Population White
asian	Total Population Asian American
american	Total Population American Indian
african	Total Population African American

Table 1: List of raw variables

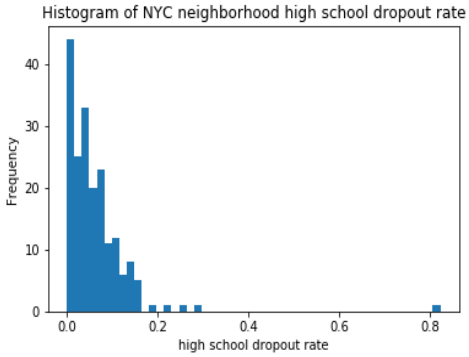


Figure 1: high school dropout rate distribution

Variable	N	Mean	St. Dev.	Min	Max
high school droupout rate	192	0.060419	0.074030	0	0.822222
unemployment rate	192	0.103509	0.037271	0.048299	0.224234
Households With Public Assistance Income	192	0.043396	0.032536	0	0.138621
poverty rate	192	0.188334	0.109335	0	0.555593
Median household income	192	58107	23671	18690	156604
male	192	0.476013	0.039998	0.346436	0.905554
female	192	0.523987	0.039998	0.094446	0.653564
pacific	192	0.000405	0.000862	0	0.004973
mixed	192	0.028857	0.018769	0.003587	0.162148
Hispanic	192	0.279489	0.212956	0.010923	0.858873
European	192	0.455828	0.28050	0.017508	0.976538
Asian	192	0.126666	0.142224	0	0.668978
American	192	0.003581	0.003224	0	0.017076
African	192	0.243516	0.268429	0.001474	0.938427

Table 2: Summary statistics table

4 Method

The first part of the analysis is an exploratory analysis of the spatial distribution of high school dropout rate in NYC neighborhoods. Local spatial autocorrelation was performed in Geoda. The Local Moran statistic was suggested in Anselin (1995) as a way to identify local clusters and spatial outliers. Queen contiguity spatial weight was generated and used to perform the spatial autocorrelation.

As I mentioned in the literature review section, a study conducted by Donnelly (2017) covered a relatively comprehensive range of neighborhood characteristics and comprised an index of neighborhood disadvantages. Donnelly did analysis on the effect of neighborhood disadvantage on high school dropout as well as a multilevel analysis on the extent to which school, neighborhood, and peer group mediate the effect of disadvantaged neighborhood. The neighborhood characteristics used in this study include unemployment rate, proportion of persons with income below poverty line, proportion of persons age 25 and older with no high school diploma or equivalency and so on. Socio-demographic control variables were also added to the multivariate models that the study built. My theoretical model is largely based on Donnelly's study.

My basic model is:

$$\begin{aligned} \text{high}_\text{s}\text{chool}_\text{d}\text{ropout}_\text{r}\text{ate} = \beta_0 &+ \beta_1 \text{unemploymentrate} \\ &+ \beta_2 \text{povertyrate} \\ &+ \beta_3 \text{medianhouseholdincome} \\ &+ \beta_4 \text{publicassistance} \\ &+ \beta_5 \text{demographiccontrol} \end{aligned}$$

I split the data into a training set (70%) and a test set (30%). I use the training set to fit the model and use the test set to evaluate the model. I use three different methods to build the model and predict high school dropout rate:

- Ordinary Least Squares Regression
- Random Forest
- Decision Tree

5 Result

5.1 Spatial distribution

Figure 2 is the natural break map of high school dropout rates in NYC neighborhoods. As we can see from the map, high school dropout rates are relatively higher in the borough of Bronx, since there are some aggregated

areas with high school dropout rates > 0.288 or high school dropout rates in between 0.124 and 0.288.

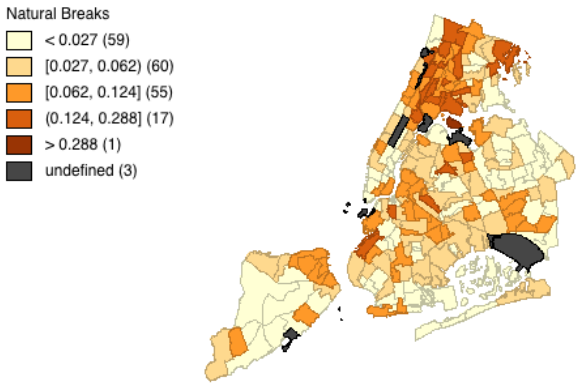


Figure 2: Natural Break Map of high school dropout rate

Figure 3 shows the LISA cluster map and the result of uni-variate local spatial autocorrelation. The high-high clusters and low-low clusters in the map have positive spatial autocorrelation, which means these areas are similar in high school dropout rate to their neighbors. The high-high clusters are in Bronx area, which corresponds to what we visualized from the natural break map. I further did some descriptive analysis to compare whether these areas have similar neighborhood characteristics that lead to similar high school dropout rates. Figure 4 shows the high school dropout rates of neighborhoods that are in the high-high clusters. The majority of the neighborhoods have dropout rate between 0.1 and 0.2. Then I further explore their neighborhood characteristics. Figure 5 shows the unemployment rate and poverty rate of the neighborhoods in high-high clusters, but it's hard to discover patterns

just from visualizing single neighborhood characteristics statistics. However, I found that over a third of the Hispanic residents of the South Bronx and roughly around a third of the Black residents of the area have incomes below the poverty, low median household income; The unemployment rate is around twice the New York state averages. More importantly, twice the national average of residents over 25 years of age report to the U.S. Census Bureau that they have not graduated from high school (Holzman, 2015). This fact further suggests that neighborhood disadvantages may be an indicator of high school dropout rate.

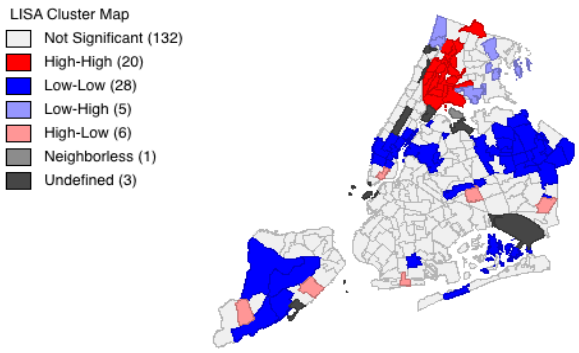


Figure 3: LISA Cluster Map

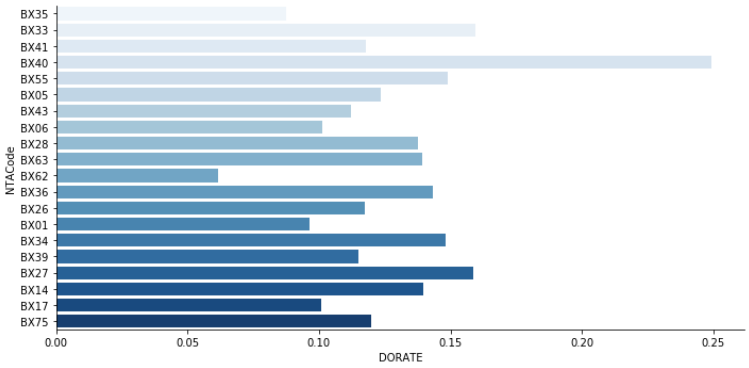
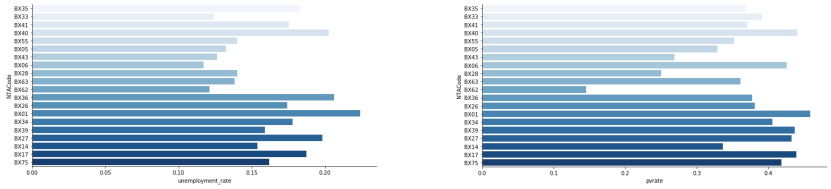


Figure 4: dropout rate of neighborhoods in the high-high clusters



(a) unemployment rate of the neighborhoods in high-high clusters

(b) poverty rate of the neighborhoods in high-high clusters

Figure 5: feature comparison

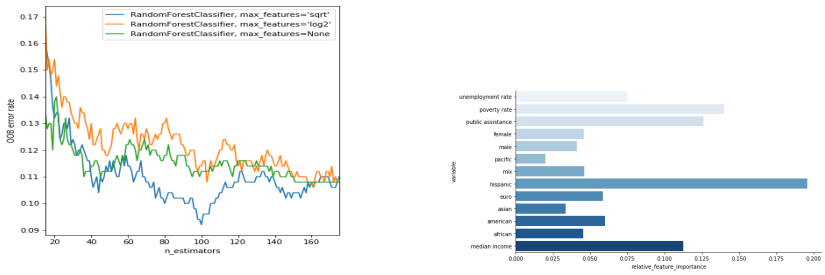
5.2 OLS

Then I move from spatial analysis to the three models I built for predicting high school dropout rate. The result of the OLS regression is shown in the appendix. The adjusted R-squared is 0.572, which means 57.2 % of the variance can be explained of the model. The only variable that is significant at the level of 0.05 is Hispanic population. Although most predictors

fail individual t-tests, the overall regression is significant, i.e. the predictors are jointly informative, according to the F-test. As poverty rate and public assistance increase, high school dropout rate in the neighborhood increases.

5.3 Random Forest Model

For the random forest model, I compared the OOB error rates of different max features and number of estimators. I used nestimators = 100, maxfeatures = 'sqrt'. The relative feature importance is shown in the right panel of figure 6. Hispanic population is the most relevant feature in this model.



(a) OOB error rate estimates (b) relativ feature importance

Figure 6: Random Forest Model

5.4 Decision Tree

After tuning the parameters in the decision tree to determine the optimal level of tree complexity, I decide to use minsamplesleaf=8, maxdepth=2 as the parameters for my decision tree. The result I got is shown in figure 7.

It appears that Hispanic population and poverty rate are more useful for partitioning the tree.

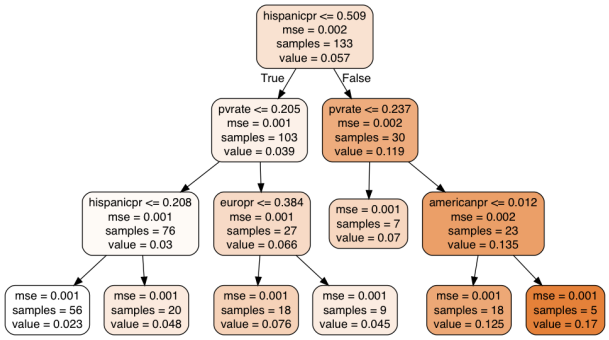


Figure 7: Decision tree

I use the test set to test the performance of the three models, and the comparison of the mean squared error is shown in table 3.

Model	Test MSE
OLS	15.025
Random Forest	14.170
Decision Tree	14.523

Table 3: Model comparison

6 Conclusion

By comparing the test MSE, Random Forest model has the best predictive performance among all the three models. Hispanic population, poverty rate and percentage of households with public assistance income are the top 3 important features in the Random Forest model. Among all the three models, Hispanic population, which a the demographic control variable, is the most significant variable for predicting high school dropout rate.

One limitation of the study is that the neighborhood characteristics considered in this research are limited. More neighborhood indicators such as crime rate are necessary. Moreover, there are several ways to divide a city into small regions such as neighborhood tabulation area, zip code tabulation area and school district. More theoretical frameworks are needed to determine which method makes most sense. Future work should also consider using different evaluation metrics to evaluate and compare the performance of the models.

References

[1] Anselin, L. (1995). *"Local indicators of spatial association LISA".* Geographical Analysis, 27, 93-115.

[2] Datiri, Dorothy Hines. *High School Dropout*, Salem Press Encyclopedia, 2013

[3] David J. Harding, *"Counterfactual Models of Neighborhood Effects: The Effect of Neighborhood Poverty on Dropping Out and Teenage Pregnancy,"* American Journal of Sociology 109, no. 3 (November 2003): 676-719.

[4] Donnelly, Louis. *Neighborhood disadvantage and school dropout.* Dissertation Abstracts International Section A: Humanities and Social Sciences, Vol 77(7-A)(E), 2017.

[5] Holzman, Michael. *The Dismal Bronx.* Dropout Nation, 19 June 2015, dropoutnation.net/2015/06/19/the-dysmal-bronx/.

[6] Massey, D. S., Denton, N. A. (1993). *American apartheid: Segregation and the making of the underclass.* Cambridge: Harvard University Press.

[7] Rendn, Maria G. *Drop Out and Disconnected Young Adults: Examining the Impact of Neighborhood and School Contexts.* The Urban Review, vol. 46, no. 2, 2013, pp. 169196., doi:10.1007/s11256-013-0251-8.

[8] Rosenbaum, J. E. (1995). *Changing the geography of opportunity by expanding residential choice: Lessons from the Gautreaux program.* Housing Policy Debate, 6(1), 231-269. doi: 10.1080/10511482.1995.9521186

[9] Rumberger, Russell W. *Poverty and High School Dropouts.* American Psychological Association, American Psychological Association, `www.apa.org/pi/ses/resources/indicator/2013/05/poverty-dropouts.aspx`.

[10] Vartanian, Thomas P., and Philip M. Gleason. *Do Neighborhood Conditions Affect High School Dropout and College Graduation Rates?* The Journal of Socio-Economics, vol. 28, no. 1, 1999, pp. 2141., doi:10.1016/s1053-5357(99)00011-6.

[11] Zaff, Jonathan F. Malone, Thomas, *"Who's Minding the Neighborhood? The Role of Adult Capacity in Keeping Young People on a Path to Graduation.",* America's Promise Alliance, Center for Promise, 2016

Appendices

OLS result

Dep. Variable:	y	**R-squared:**	0.611
Model:	OLS	**Adj. R-squared:**	0.572
Method:	Least Squares	**F-statistic:**	15.71
Date:	Mon, 28 May 2018	**Prob (F-statistic):**	1.84e-19
Time:	00:09:56	**Log-Likelihood:**	-338.14
No. Observations:	133	**AIC:**	702.3
Df Residuals:	120	**BIC:**	739.9
Df Model:	12		

	coef	std err	t	P> |t|	[0.025	0.975]
constant	-0.0101	4.537	-0.002	0.998	-8.993	8.973
urate	-2.0131	15.157	-0.133	0.895	-32.023	27.997
pvrate	2.0633	6.929	0.298	0.766	-11.655	15.782
assistance	37.3056	22.102	1.688	0.094	-6.455	81.066
femalepr	-12.7820	7.076	-1.806	0.073	-26.792	1.228
malepr	12.7719	7.523	1.698	0.092	-2.123	27.666
pacificpr	182.9742	327.901	0.558	0.578	-466.248	832.196
mixpr	1.2515	18.379	0.068	0.946	-35.138	37.641
hispanicpr	9.9028	4.509	2.196	0.030	0.975	18.830
europr	2.4872	5.609	0.443	0.658	-8.618	13.592
asianpr	-1.1014	6.473	-0.170	0.865	-13.918	11.715
americanpr	166.1614	99.969	1.662	0.099	-31.770	364.093
africanpr	3.9238	5.790	0.678	0.499	-7.539	15.387
medianinco	-1.724e-05	2.26e-05	-0.761	0.448	-6.21e-05	2.76e-05

Omnibus:	29.800	**Durbin-Watson:**	1.983
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	46.285
Skew:	1.101	**Prob(JB):**	8.90e-11
Kurtosis:	4.871	**Cond. No.**	3.20e+20