Which factors lead to high crime rates?

Fardous Sabnur

I) Introduction

In this paper, I will perform exploratory data analysis and implement multiple Machine Learning models to check if Education, Poverty, Youth Population and Foreign-Born population effects crime rates in NYC. I'm going to use the NYC Crime Stats data from Kaggle. The Education, Poverty, Youth Population and Foreign-Born population data are US Census data. It's difficult to predict crime and may be impossible but there are many factors that can cause one area to be more dangerous with higher crime rates than others. We can use Machine Learning can help find which factor can lead to higher crime rates, and where to invest in to reduce crime rates and increase the well-being of the people living in an area. Since Crime data was collected from a different source than all the other variable, the final result might not show the most accurate results. I handpicked the variables used in this data, there are many other factors like tourist spots, school districts that can cause an area to have higher crime rates.

I will implement different machine learning models to find which factors lead to high crime rates in an area and which don't have any effect on crime rates. Machine learning methods used for this process are Multiple Linear Regression, Ridge Regression, Tree-Based models – Decision Tree and Random Forest. To evaluate the performance of each model, Mean squared error and r-squared is used. By comparing the result of different models crime rate, the best model and best variable will be determined.

II) Description of Data

NYC Crime Stats: This dataset explores crime in a geospatial way and includes demographic of the perpetrator. There are a lot of different explanatory variables, I'm going to use the following variable for my analysis:

- "ofns_desc": Description of offenses, there are 87 unique offenses in this column
- "arrest_date": date of arrest
- "perp_sex" and "perp_race": Race and sex of perpetrator
- "age-group": Age of the perpetrator
- "latitude" and "longitude": geographic coordinates

Census Data: Initially I downloaded a few datasets from DATA2GO.nyc website and picked a few columns from those datasets that I thought would be important to predict crime. Then I created a new dataset and inserted all those important columns there manually. The data's for DATA2GO.nyc were collected from the Census website. The new dataset includes the following variables:

- "GEOID": Unique code to identify census tracts
- "Total Population": The total population of the census tract
- "Foreign-Born (# of people)": Anyone who is not a U.S. citizen at birth. This includes naturalized U.S. citizens, lawful permanent residents (immigrants), temporary migrants (such as foreign students), humanitarian migrants (such as refugees and asylees), and unauthorized migrants.
- "Youth Population (# of youth under 18)": Percentage of people under the age of 18

- "Completed High School or High School and Some College (# of adults 25+)": Percentage of people over the age of 25 who completed High School or more
- "Poverty (# of individuals in households with incomes below poverty)": Percentage of people in households with incomes below poverty.

III) Data Cleaning and Data wrangling

The following changes were made to clean and merge the two datasets mentioned above:

1) I dropped some columns and included the ones listed above.
2) I selected crimes for which the arrest dates were between 2014 to 2018 because the census data reported are also from 2014 to 2018
3) Using Geopandas library, converted all the given longitude and latitude to census tracts. That process gave unique GEOIDs for each tract and names of the tracts
4) There are 2168 census tracts in NYC but after the conversion I got 2339 census tracts. There were a lot of extra tracts that are not in NYC.
5) Then I visualized some of the data which would be shown in the next section.
6) Next, I grouped all the crimes by census tracts. My approach was to count how many crimes occurred in a census tract and created a new dataset labeled "newmerged_data" to only include GEOID and crime rates per GEOID.
7) Next, I merged the two datasets by GEOID
8) Since the NYC Crime Data had many extra Census Tracts, I removed all the tracts that were not included in the Census Data.
9) There were some tracts that were included in the census data but not in the NYC Crime Data. For those GEOIDs, I set the crime rate to 0. Which is probably not true but my

only other option was to drop the full row which would lead to other issues while mapping.

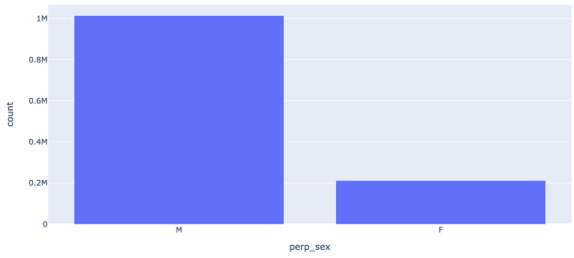
10) The "newmerged_data" had some null values, I set those nulls to 0's as well.

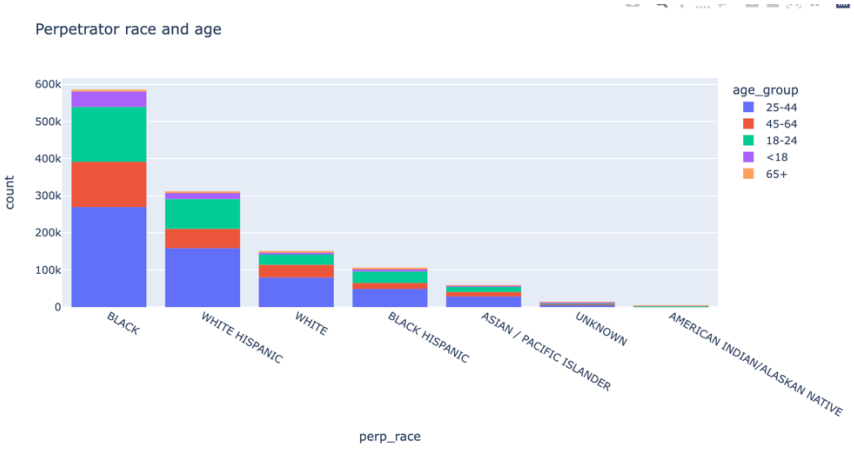
11) This is what the "newmerged_data" looks like now:

	GEOID	Crimes	GEO_LABEL	ForeignBorn	YouthPopulation	Education	Poverty	Total Population (#)
0	36005000100	159.0	Census Tract 1, Bronx County, New York	1057.0	171.0	2976.996310	0.0	7080.0
1	36005000200	148.0	Census Tract 2, Bronx County, New York	1551.0	960.0	2352.904918	1028.0	4542.0
2	36005000400	254.0	Census Tract 4, Bronx County, New York	1051.0	1127.0	3285.670285	549.0	5634.0
3	36005001600	261.0	Census Tract 16, Bronx County, New York	1822.0	1501.0	3348.367408	1264.0	5917.0
9	36005002300	916.0	Census Tract 23, Bronx County, New York	890.0	1102.0	2529.497451	2187.0	4600.0

IV) Data Visualization:

- The graph below shows the sex ratio of the perpetrator for the NYC crime data from 2014-2018. On average, male are five times more likely to commit crimes than female

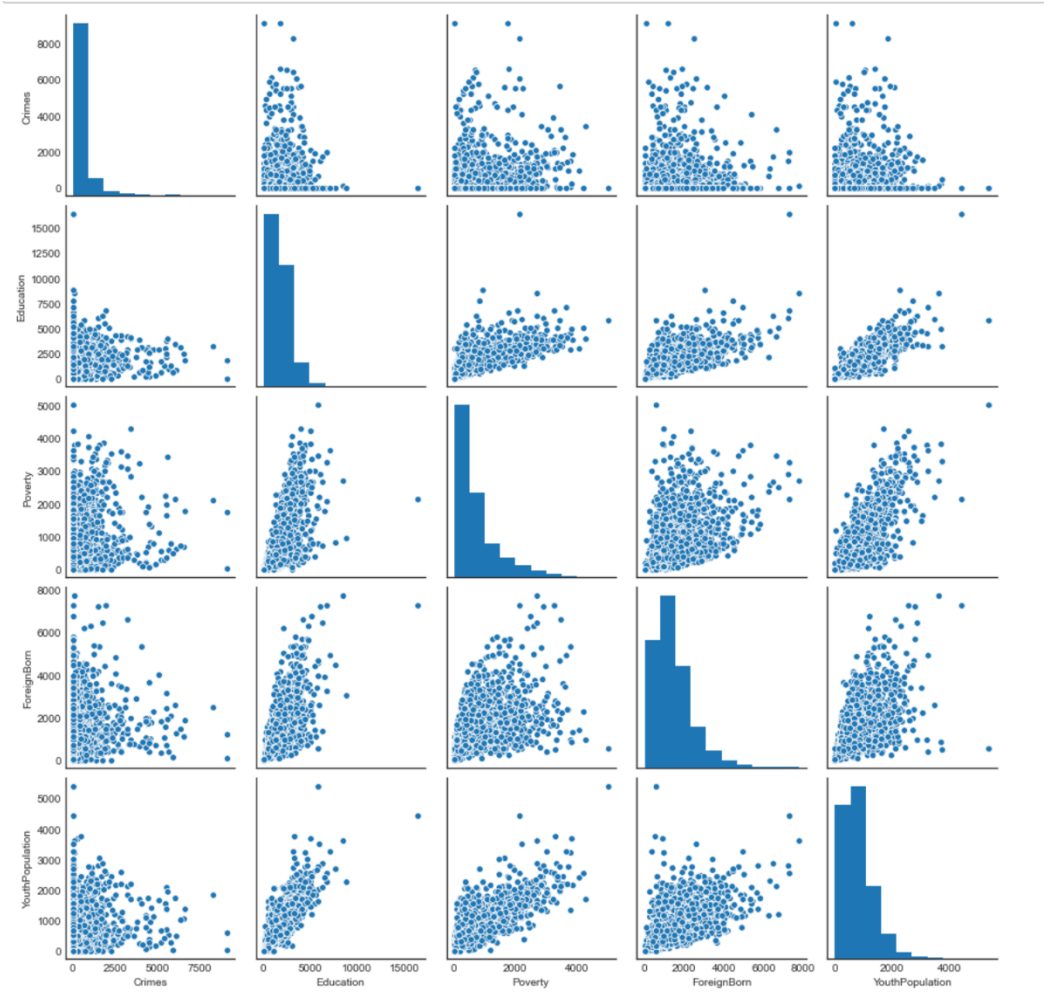


- 

The graph to the left shows the perpetrators race and age. People of age 25-44 have

the highest crime counts. That's why I picked the education variable for people age 25+ to see if completing high school degree or more has any effects on the crime rate for these age groups. My initial assumption was, if a neighborhood has many educated people, they'll have better life and commit less crimes. The graph also shows that people of Black and Hispanic ethnicities commit most crimes in NYC. Most Black and Hispanic people in NYC live in low-income communities. That's why I picked the Poverty data with the number of individuals in households with incomes below poverty level. To see if higher poverty rates lead to higher crime rates.

- The graph below shows the correlation between the NYC Crime Stat data's crime counts



and the Census Data variables. None of them show strong correlation with Crime, which could be due to the NYC Crime Data not having all the accurate crime information. But outside of crime, some of the other variable have strong correlations among them.

- The table below gives the numerical correlations for the graphs above for better interpretability. Poverty and Crimes have the highest correlation but it's still a weak correlation. ForeignBorn and Crime has the weakest correlation. This doesn't mean that Poverty is statistically significant factor causing crimes. We must run statistical and machine learning models to find the result.

	Crimes	Education	Poverty	ForeignBorn	YouthPopulation
Crimes	1.000000	0.117935	0.286118	0.098041	0.129188
Education	0.117935	1.000000	0.641690	0.670521	0.818343
Poverty	0.286118	0.641690	1.000000	0.484048	0.763665
ForeignBorn	0.098041	0.670521	0.484048	1.000000	0.584357
YouthPopulation	0.129188	0.818343	0.763665	0.584357	1.000000

V) Analysis

Multiple Linear Regression

I began my analysis by using a Multiple Linear Regression model to find which factors had the most impact on crime rates. This regression model can be used to find the effect on crime for multiple variables at once. After running the model two variables were statistically significant. For the rest of the model, I will perform all the models twice. Once on all four variables and then on two significant variables.

First, I used the statsmodels.formula.api,ols() ordinary least square function because this function shows how well a given model fits the data, and what variables "explain" or

affect the outcome. Then I split the data for X_train, y_train, X_test and y_test. I'll use this same split for all the models with all four variables. Then I fitted the function on the training set of four variables. Poverty and YouthPopulation are statistically significant because their p-values are less than 0.05. With higher poverty in an areas crime increases by .5467 and with higher youth population, crime goes down by 0,3538. On the other hand, Education and Foreign Born Population are not statistically significant.

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	380.7043	34.610	11.000	0.000	312.832	448.576
Education	0.0086	0.030	0.287	0.774	-0.050	0.067
Poverty	0.5467	0.039	14.168	0.000	0.471	0.622
ForeignBorn	0.0044	0.024	0.189	0.850	-0.042	0.051
YouthPopulation	-0.3538	0.066	-5.346	0.000	-0.484	-0.224

Which is a bit surprising to me, because I thought areas with higher education rates would have less crime rates. One reason to explain this could be that the perpetrators who commit crimes in the high education neighborhood probably do not live in that neighborhood. A lot of political party's state that immigrants are threat to native citizens and want to ban immigration. That's why I included the foreign-born population. It's not significant, meaning having high immigrant population doesn't affect the crime rate. The adjusted R-squared for this model is 0.101, which means that the variables used in this model for crime are not adding value to the model.

R-squared:	0.101
Adj. R-squared:	0.099
F-statistic:	60.75
Prob (F-statistic):	1.09e-48
Log-Likelihood:	-17615.
AIC:	3.524e+04
BIC:	3.527e+04

Then I split the X and y variables again into training and test data to X_train_sig, y_train_sig, X_test_sig and y_test_sig. I will use this same split on all the other models that only use the the two significant variables, Poverty and YouthPopulation. Then I performed the same function again, on these splits. The adjusted R squared increased by 0.01, and BIC decreased by a very little amount. This result is not very significant. I will continue to perform all the models on all four variables and then only on the two significant variables to see if the result changes significantly.

Next, I used the skl_lm.LinearRegression(): Scikit-learn Linear Regression function. This serves a different purpose for the data. This function follows the machine learning tradition where the main supported task is choosing the "best" model for prediction. When I implement the model on all four variables, the r-square is 0.125 and mean squared error is 643965.22. This r-square is not that high and the MSE is way too high. Overall, the variables are not great predicting factors for crime.

Then I used the same function on only Poverty and YouthPopulation. The r-square is 0.128 and the MSE is 642298.618. This r-square is also not that high but its higher than the previous model and the MSE is also way to high but it’s lower than previous model.

Ridge Regression

For the second model I chose Ridge Regression. This model involves shrinking the estimates coefficients towards zero. This shrinkage reduces the variance, some of the coefficients may shrink to exactly zero. Ridge regression applies a penalty to the magnitude of parameter values to reduce overfitting. Then I fit the model to the training data for all four variables. The r-square is negative, so the chosen model does not follow the trend of the data. Below are the coefficients assigned to each variable for this model: Poverty has the highest coefficient like previous model, YouthPopulation has the second highest coefficient.

```
Poverty            0.890104
Education          0.350065
ForeignBorn        0.288111
YouthPopulation    0.414111
```

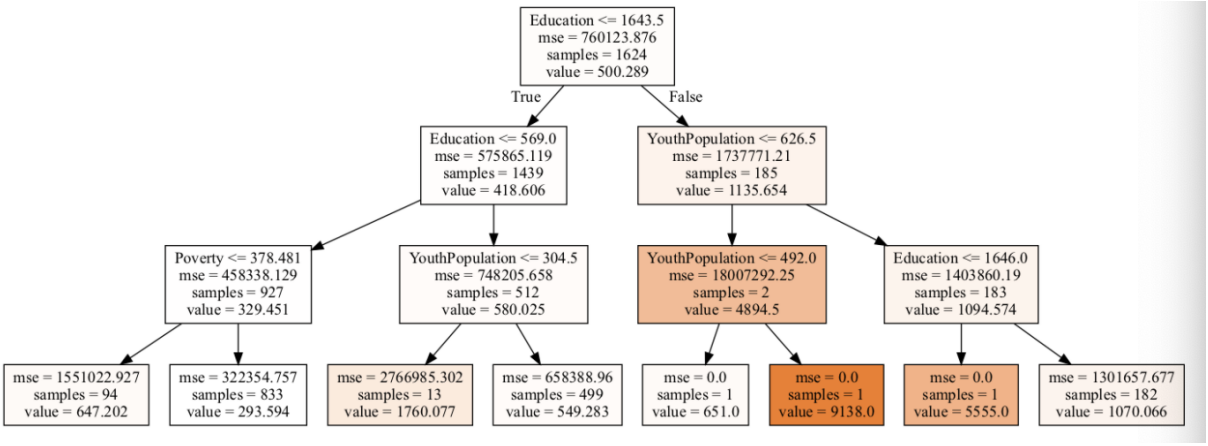
The mean squared error for this model is higher than the Multiple Linear Regression model. The MSE is 737411.82.

Now like previous model if we now only include the two variables with highest coefficients, YouthPopulation and Poverty. The MSE increases to 737569.74. The R^2 is still negative. This time removing Education and ForeignBorn does not lead to a lower MSE.

Regression Tree

I used the DecisionTreeRegressor(max_depth=3) function on all four variables for making the tree. This was the result. There are three possible cases where the tree can overfit the data because those are the times when MSE is equal to zero. The other times MSE is high like the

previous models.



For this model, MSE is 704976.33 and the r_squared is 0.042. The results are not extremely significant but it's better than the last model with ridge regression.

When I remove the ForeignBorn and Education variable the Mean squared error decreases to 670354.23. The R_squared increases to 0.0899. Removing the two variable leads to a better result.

Random Forest

For the model with all four variables, the MSE is 1179039.90 and the r_squared is 0.0258.

When I remove the ForeignBorn and Education variable the Mean squared error decreases to 1154883.721. The R_squared decrease to 0.020. It's unclear if removing the two-variable lead to a better result because MSE went down and R_squared went up.

VI) Conclusion

Accuracy rates for the Models:

Model Name	**R_squared**	**MSE**

Multiple Linear Regression (all variable)	0.125	643965.223
Ridge Regression (all variable)	-7.447	737411.819
Regression Trees (all variable)	0.0429	704976.33
Random Forest (all variable)	0.0258	1179039.9
Multiple Linear Regression (two variable)	0.128	642298.618
Ridge Regression (two variable)	-1.696	737569.748
Regression Trees (two variable)	0.0899	670354.236
Random Forest (two variable)	0.02	1154883.72

The Multiple Linear regression with two variables, Poverty and Youth Population have the best R-squared score and the lowest mean squared error. The R squared score must be close to one for the model to show that these variables are very significant in showing crime rate. Maybe it's the quality of the NYC Crime Stats data that doesn't have accurate information. Also, there are many other factors that lead to high crime rates. Crime data also include traffic violation crimes. In Manhattan, people are very rich and educated but the streets are narrow and it's easier to get into vehicle accidents. There are also many tourists. There are many factors that affect crime rates. The data must be dealt with differently to find better results.