Modelling neighbourhood crime-rate using sociodemographic factors in Toronto

Introduction

Crime, especially violent crimes, in cities have far-reaching negative impacts, including increased fear and anxiety among residents, a decrease in sense of safety, and reduced economic opportunities due to businesses avoiding high-crime areas[1,2,3]. In the past few years, the rate of crime in Canada has increased overall[4]. In Toronto specifically, the rate of homicides have nearly doubled from 2012 (1.38 per 100,000) to 2022 (2.06 per 100,000)[5]. As such, understanding the factors that give rise to crime within the city is of utmost importance.

The "law of crime concentration" states that the majority of criminal activities within a city tend to be localized in particular, limited areas such as city blocks, streets or neighborhoods[6]. In other words, crimes are clustered spatially[7]. While this has long-been known, the challenge to understand why certain areas are "hotspots" remains challenging. At its core crime is a complex, multifaceted issue. However, the key driving force to crime is inequality and disparity[8], and is thus intimately linked with various socio-demographic factors[9]. Many studies acknowledge these facts[5, 10,]. Historically, linear models have struggled to predict crime rates across geographical time and space[11]. Overall, evidence from several papers have demonstrated that crime typically occurs more highly across areas of unemployment, low-income[12,13], and high income disparities[13,14]. However, the strengths of these relationships vary significantly between studies, cities and countries. Additionally, while family dynamics and education have also had impacts on a person's future criminal activities[15,16], less consensus has been reached about a neighbourhood's level of educational attainment on a crime occurrence at a neighborhood level[17]. To deal with the issue of non-linearity in generalized linear models, several recent studies predicting crime rates at the city-level have employed the use of novel machine-learning models[18,19,20,21]. In particular, Alves et al. (2018) use of the Random Forest method managed to explain up to 97% of the variance in homicide rates between Brazilian cities. However, results from studies of crime rate at the neighborhood level have been limited.

Thus, the main question we seek to address is how much of the variance in the crime rate in Toronto neighbourhoods can be explained solely by neighborhood sociodemographic characteristics. Secondly, we wish to identify the key neighbourhood-level demographic features that contribute to crime hotspots within certain neighbourhoods. To accomplish this, we will investigate the use of multiple models, including simple Decision Tree Regressors, Random Forest Models, and Linear Regression. By studying this relationship through machine-learning approaches, we can begin to identify the influence and impact of unique and complex features of neighborhoods that give rise to crime in Toronto. Through this process, we can learn to understand the root causes and adopt comprehensive strategies that address not only law enforcement but also social and economic disparities to effectively stifle the growing crime rates and enhance community well-being.

Methods:

The Data

In order to find correlations between neighbourhood characteristics and the occurrences of crime, information was compiled from two datasets, the Neighbourhood Profiles dataset, and the Neighbourhood Crime Rates dataset. Both of these datasets are publicly available on The City of Toronto's Open Data Portal. The Neighbourhood Profiles dataset includes sociodemographic characteristics of the 158 neighbourhoods of Toronto collected from the 2016 census and includes features such as age, family-income, marriage status, education-level, distributions. The second dataset, Neighbourhood Crime Rates, contains labels and includes all city-reported and crimes for each of the 158 Toronto neighbourhoods from 2014 to 2022. In 2016, reported crime categories include assault (41.41%), auto thefts (7.35%), breaking and enterings (14.23%), homicide (0.17%), robberies (8.16%), and shootings (0.90%). No missing data was present in either datasets.

Feature and Outcome Engineering:

To explain neighbourhood crime rate, features selected from the neighbourhood profiles included the total number of individuals, average total household income, the gini-coefficient of total household income, unemployment rate, highest educational attainment of each census respondent, and the proportions of people from different ethnic and racial backgrounds per neighbourhood. For each neighbourhood the respondent's highest level of educational attainment were separated into five different categories: those without a diploma, with a highschool diploma or equivalent, with a non-bachelors post-secondary degree, with a bachelor's degree, and those with a graduate degree.

Additionally, while multiple race/ethnic categories were available as a response, we decided to recategorise results into seven broad categories including those of Black, Arab, Latin American, East & South East Asian, West Asian, Multi-racial/ethnic, and Non-racial minority backgrounds. Additionally, because neighbourhoods typically varied by population size, features such as highest degree of educational attainment and race were calculated as an intra-neighbourhood proportion to be inputted into the model instead. Total annual crime per-neighbourhood was calculated as a sum of all reported categories of crime within the year. The population for the last Census of the study period (2016) was used to calculate the total amount of reported crimes per 100,000 in 2016 (**APPENDIX - Dataset**).

Data Exploration and Model Building:

To explore the data for a bivariate correlation was constructed to observe the correlation between features and outcomes (**APPENDIX - Correlation Matrix**). In addition, in preparation for Linear Regression, Decision Tree and Random Forest analyses, the data was split into we used the train_test_split function from scikit-learn to make an approximately 80:20 split for the training (n = 126) and testing dataset (n=32). From this, a multivariate-Linear Regression model was constructed using the training dataset and validated it against the test dataset. Features selected in this model contained all those described in the previous section. Model performance results were evaluated based on scikit-learn's built in Pearson correlation (R^2) score function calculator and assessment of fit was performed by observing the distribution of residuals on the test dataset. Additionally optimal hyper parameters for Decision Tree and Random Forest Regressor were found using the sci-kit-learn's gridsearch. Models were then evaluated using pearson correlation scores and informative splits were identified using SHapley Additive explanation values (SHAP).

Results:

Multivariate Linear Regression Model

Fitting the data features into a multivariate linear regression model revealed that annual crime rates per 100,000 were larger in smaller neighbourhoods population-wise, with lower average total household income, lower income-inequality, and a larger proportion of individuals who self-identify as Black, Arab, or as a part of multiple visual minorities (**Table 1**). Conversely, crime seems to occur less often in neighbourhoods with a higher proportion of East Asian, West Asian, or non-visible minorities within their populations. Additionally, after controlling for these factors, conflicting results about neighbourhood level educational attainment seems to demonstrate that crime occurrence is least abundant in neighbourhoods who have a larger population of people whose highest degree of educational attainment was a post-secondary diploma or certificate. Additionally, the influence of the unemployment rate in Toronto neighbourhoods is limited as well. Running our model on the test dataset, revealed a coefficient of determination of 0.331, suggesting that 33.1% of the total variation in crime rates in Toronto could be explained by the linear relationship between crime and our selected features.

Table 1

Feature	Coefficient	Interpretation
Gini index for total household income	-3.44e+02	Every 0.1 increase in the gini-coefficient is associated with a decrease annual crime by 34 per 100,000
Average household size	-7.37e+02	Every 1 number increase in the average household size is associated with a decrease annual crime by 737 per 100,000
Average total household income	-4.16e-03	Every 1,000 dollar increase in the Average total Household Income per neighbourhood is associated with a decrease in annual crime by 4.15 per 100,000
Number of individuals in the neighbourhood	-5.32e-03	Every 1,000 increase in the population size of the neighbourhood is associated with a decrease annual crime by 5.32 per 100,000
Proportion of self-identifying Black individuals	7.23e+02	Every 1% increase in the proportion of Black individuals in the neighbourhood is associated with an increase annual crime by 7.23 per 100,000
Proportion of self-identifying Arab individuals	1.22e+04	Every 1% increase in the proportion of Arab individuals in the

		neighbourhood is associated with an increase annual crime by 122 per 100,000
Proportion of self-identifying Latin-American individuals	4.74e+03	Every 1% increase in the proportion of Latin American individuals in the neighbourhood is associated with an increase in annual crime by 47.4 per 100,000
Proportion of self-identifying East and South-East Asian individuals	-7.02e+02	Every 1% increase in the proportion of East and South-East Asian individuals in the neighbourhood is associated with a decrease in annual crime by 7.02 per 100,000
Proportion of self-identifying West Asian individuals	-1.44e+04	Every 1% increase in the proportion of West Asian individuals in the neighbourhood is associated with a decrease in annual crime by 144 per 100,000
Proportion of individuals self identifying as part of multiple visible minorities	2.95e+04	Every 1% increase in the proportion of individuals from multiple visible minorities in the neighbourhood is associated with an increase in annual crime by 295 per 100,000
Proportion of non-visible minority individuals	-1.29e+03	Every 1% increase in the proportion of individuals from not from any visible minority in the neighbourhood is associated with a decrease in annual crime by 12.9 per 100,000
Unemployment rate	-3.95e+01	Every 1% increase in unemployment rate decreases annual crime by 0.395 per 100,00
Proportion of individuals with no diploma	2.32e+4	Every 1% increase in the proportion of individuals without a diploma in the neighbourhood is associated with an increase in annual crime by 232 per 100,000
Proportion of individuals with highschool diploma or equivalent	1.23e+5	Every 1% increase in the proportion of individuals whose highest degree of educational attainment was a highschool diploma (or equivalent) in the neighbourhood is associated with an increase in annual crime by 1230 per 100,000
Proportion of individuals with a post secondary diploma or certificate below bachelor level	-6.68e+4	Every 1% increase in the proportion of individuals whose highest degree of educational attainment was a post-secondary diploma or certificate (below bachelor's level) in a neighbourhood is associated with a decrease in annual crime by 668 per 100,000
Proportion of individuals with a bachelor's degree	4.32e+4	Every 1% increase in the proportion of individuals whose highest degree of educational attainment was a bachelor's degree in the neighbourhood is associated with an increase in annual crime by 432 per 100,000
Proportion of individuals with graduate-level degree higher than a bachelor's.	9.32e+4	Every 1% increase in the proportion of individuals whose highest degree of educational attainment was a graduate degree in the neighbourhood is associated with an increase in annual crime by 932 per 100,000

Table 1. Multivariate Linear Regression results. Intercept for the model was set at 0 per 100,000. Total squared error was 339281.1592377367, R2 Score on the training data set was 0.478 and the R2 Score on testing dataset was 0.331.

Comparing predicted crime rates to actual crime rates per 100,000 in Toronto neighbourhoods (**APPENDIX - Geographical Depictions**) similarities in geographical distributions of crime is apparent. However, the Multivariate Linear Regression Model over-predicted crime across neighbourhoods near the Toronto harbourfront. Analysing the model's residual plots, the residuals appear to violate homoscedasticity and deviate slightly from the normality assumption, with the model underpredicting at the extreme ends of the distribution (**APPENDIX - MLR Residuals**).

Decision Tree Regressor

Constructing the Decision Tree Regressor proved challenging with a great amount of model instability. The final optimal hyperparameters found included {max depth: 10, max_features: sqrt, min_samples_leaf: 15, splitter: 'best'}. Ultimately, the decision tree regressor that best explained the variance in both the train and test dataset was found to have an R^2 performance of 0.225 and 0.350 respectively(**Figure 1.**)

Figure 1.

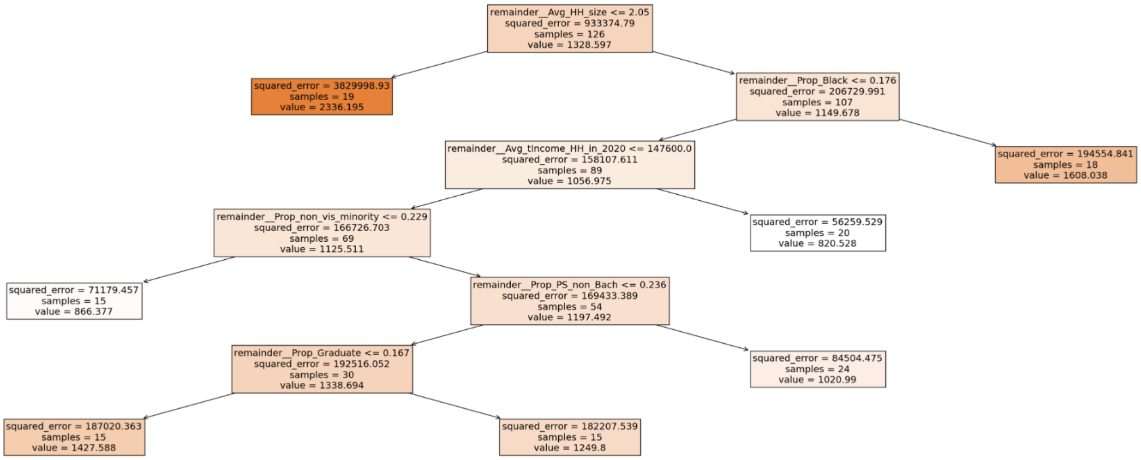

remainder__Avg_HH_size <= 2.05
squared_error = 933374.79
samples = 126
value = 1328.597
squared_error = 3829998.93
samples = 19
value = 2336.195
remainder__Prop_Black <= 0.176
squared_error = 206729.991
samples = 107
value = 1149.678
remainder__Avg_tincome_HH_in_2020 <= 147600.0
squared_error = 158107.611
samples = 89
value = 1056.975
squared_error = 194554.841
samples = 18
value = 1608.038
remainder__Prop_non_vis_minority <= 0.229
squared_error = 166726.703
samples = 69
value = 1125.511
squared_error = 56259.529
samples = 20
value = 820.528
squared_error = 71179.457
samples = 15
value = 866.377
remainder__Prop_PS_non_Bach <= 0.236
squared_error = 169433.389
samples = 54
value = 1197.492
remainder__Prop_Graduate <= 0.167
squared_error = 192516.052
samples = 30
value = 1338.694
squared_error = 84504.475
samples = 24
value = 1020.99
squared_error = 187020.363
samples = 15
value = 1427.588
squared_error = 182207.539
samples = 15
value = 1249.8

Figure 1. Decision Tree Regressor model results. DT__max_depth': 10, 'DT__max_features': 'sqrt', 'DT__min_samples_leaf': 15, 'DT__splitter': 'best'. The Pearson correlation (R^2) score on the training dataset and test dataset was 0.225 and 0.350 respectively.

Informative splits included in this decision tree regressor appeared to be centred around average household size being greater than 2.05, the proportion of black individuals in a neighbourhood being greater than 0.176, and the average household income being greater than 147,600 a year. According to the tree, crime appears to occur most commonly in neighbourhoods with a smaller average household size than 2.05 and those whose average household size is greater than 2.05, but have a greater proportion of black individuals living in it.

Random Forrest Regressor

A great amount of model instability was also present when constructing the Random Forest Regressor. However, the final optimised hyperparameters of the model included a total of 10 different estimator trees with a max depth of 5, with a minimum sample number of 5 per leaf, and whose total number of potential splits considered at each node was log2 of the total amount of features available. This Random Forest Regressor was able to perform at an R^2 of 0.461 and 0.212 respectively the training and testing datasets. Additionally, the SHAP values for this multi-tree regressor approach indicated that factors such as smaller household size, larger proportion of multi-ethnic/racial minorities, and higher unemployment rates within neighbourhoods were consistently increased within-neighbourhood level crime (**APPENDIX - SHAP values Random Forest**).

Discussions:

Table 2.

	Multivariate Linear Regression	Decision Tree Regressor	Random Forest Regressor
Train	0.478	0.225	0.461
Test	0.331	0.350	0.212

Table 2. Model comparison between multivariate linear regression, decision tree, and random forest approaches. Values demonstrated are performances based on pearson correlation values

Surprisingly, overall the multivariate linear regression model performed best on explaining neighbourhood level crime rate and was able to explain approximately 47.8% and 33.1% of the variation in crime rates based within the training and test dataset based on their own neighbourhood-level sociodemographic characteristics (**Table 2**). However, there were some issues with this model as some degree of non-linearity in modelled relationships was present. Running Decision Tree and Random Forest approaches, we had hoped to incorporate more complex interactions and non-linearity into our approach, however they failed to perform better than the Linear model. Explanations for this may include the presence of collinearity between features, causing most of the variation in subsequent splits to already have been explained by a previous split. Additionally, due to the number of features incorporated (n =17), and the limited amounts of neighbourhoods available to take observations from (n = 158), most of the complex interactions that we attempted to model for may not have had enough data points to appropriately capture the different ways sociodemographic characteristics may interact with each other in Torontonian neighbourhoods.

Additionally, we also sought to investigate which demographics of individuals were living in high-crime rate neighbourhoods. Reviewing our three models, small average house-hold size seems to remain a consistently important predictor of crime rates. This is likely a reflection of crime being clustered in areas high in dense urbanised-person apartments[22]. Additionally, a greater proportion of Black and multi-visible ethnic minority individuals within the neighbourhoods also remained a moderate predictor of neighbourhood crime across all approaches as well. Thus highlighting similar racial disparities associated with living in higher crime neighbourhoods[23]. Furthermore average total household income within neighbourhoods and the gini-index for neighbourhood-level income disparity seemed to have less of an impact in all approaches. Additionally, there appears to be a complex interaction between neighbourhood-level educational attainment as, after incorporating other SES features, higher degrees of educational attainment do not directly correlate with decreased crime rates. This is likely because educational attainment is closely correlated to the identified factors.
To better understand the current landscape in crime prediction or explanation and to gain insight on which critical features are potentially missing in our analysis, we decided to observe the prevalence of diversity, equity, and equality related terms within methodology sections of 52 different papers (**APPENDIX - References 52**). From the methodology sections, both context-specific and non-context-specific stop words relating to models, study designs, and populations were removed, thus leaving behind subjects of their investigations. The word cloud generated displayed a high prevalence of studies surrounding crime in city and urban settings, however, with little on the subject of rural crime (**APPENDIX - Word Cloud**). Additionally, many studies looked at socioeconomic and demographic variations relating to gender, violence, household, family structure, age, school/education, and housing. Additionally, many crime prediction studies also appeared to have incorporated elements of spatial and temporal analysis and have included features such as urban environment[11,18,]. On the subject of crime studies, many studies focused on violent crimes and assaults with some focusing on shootings and inter-partner violence.

As a result, it has become evident that the sole use of neighbourhood-level sociodemographic characteristics is unable to explain a majority of the variance of total aggregated neighbourhood-level crime. However, through our approaches, we have highlighted overall disparities between the crime rates people of Black, Arab, multi-ethnic/racial minorities experience within their neighbourhoods. Furthermore, our results also indicate the prevalence of crime being located in neighbourhoods with smaller average household sizes. Thus, enabling us to help direct actionable policies and intervention in the short run that could help these individuals deal with crime within their neighbourhoods. Future studies investigating neighbourhood crime rate occurrence in Toronto should look for datasets that include additional features such as urban environment, and look to use approaches such as Recurrent Neural Networks and Long-Short Term Memory models to incorporate data on crime rates temporally. Additionally, because of the heterogeneity of the type of crimes that occur within neighbourhoods, future studies should look to use sociodemographic features to predict the occurrence rate of specific types of crimes such as homicides, assaults, and thefts.

References:

1. Baranyi, G., Di Marco, M. H., Russ, T. C., Dibben, C., & Pearce, J. (2021). The impact of neighbourhood crime on mental health: A systematic review and meta-analysis. Social science & medicine (1982), 282, 114106. https://doi.org/10.1016/j.socscimed.2021.114106
2. Putrik, P., van Amelsvoort, L., Mujakovic, S., Kunst, A. E., van Oers, H., Kant, I., Jansen, M. W., & De Vries, N. K. (2019). Assessing the role of criminality in neighbourhood safety feelings and self-reported health: results from a cross-sectional study in a Dutch municipality. BMC public health, 19(1), 920. https://doi.org/10.1186/s12889-019-7197-z

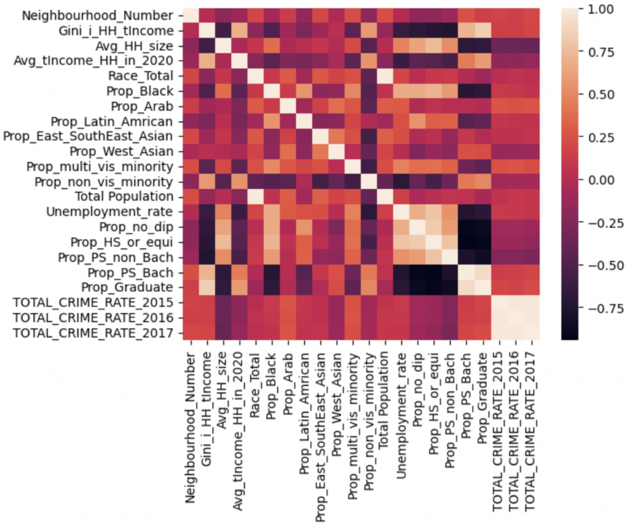
3. Lens, Michael & Meltzer, Rachel. (2016). Is crime bad for business? Crime and commercial property values in new york city. Journal of Regional Science. 56. n/a-n/a. 10.1111/jors.12254.
4. Statistics Canada. Table 35-10-0071-01 Number and rate of homicide victims, by Census Metropolitan Areas. https://doi.org/10.25318/3510007101-eng
5. Mohammadi, A., Bergquist, R., Fathi, G. et al. Homicide rates are spatially associated with built environment and socio-economic factors: a study in the neighbourhoods of Toronto, Canada. BMC Public Health 22, 1482 (2022). https://doi.org/10.1186/s12889-022-13807-4
6. Braga, A. A., Andresen, M. A., & Lawton, B. (2017). The law of crime concentration at places: Editors' introduction. *CrimRxiv*. https://doi.org/10.21428/cb6ab371.e61c7170
7. Johnson, S.D. (2010) A Brief History of the Analysis of Crime Concentration. European Journal of Applied Mathematics, 21, 349-370. https://doi.org/10.1017/S0956792510000082
8. De Courson, B., Nettle, D. Why do inequality and deprivation produce high crime and low trust?. Sci Rep 11, 1937 (2021). https://doi.org/10.1038/s41598-020-80897-8
9. Chong, V. E., Lee, W. S., & Victorino, G. P. (2015). Neighborhood socioeconomic status is associated with violent reinjury. *The Journal of surgical research*, *199*(1), 177–182. https://doi.org/10.1016/j.jss.2015.03.086
10. Alves, Luiz G.A. & Ribeiro, Haroldo V. & Rodrigues, Francisco A., 2018. "Crime prediction through urban metrics and statistical learning," Physica A: Statistical Mechanics and its Applications, Elsevier, vol. 505(C), pages 435-443. Handle: RePEc:eee:phsmap:v:505:y:2018:i:c:p:435-443 DOI: 10.1016/j.physa.2018.03.084
11. Luca, M., Campedelli, G. M., Centellegher, S., Tizzoni, M., & Lepri, B. (2023). Crime, inequality and public health: a survey of emerging trends in urban data science. *Frontiers in big data*, *6*, 1124526. https://doi.org/10.3389/fdata.2023.1124526
12. De Courson, B., & Nettle, D. (2021). Why do inequality and deprivation produce high crime and low trust?. *Scientific reports*, *11*(1), 1937. https://doi.org/10.1038/s41598-020-80897-8
13. Mundia, L., Matzin, R., Mahalle, S., Hamid, M. H., & Osman, R. S. (2016). Contributions of sociodemographic factors to criminal behavior. Psychology research and behavior management, 9, 147–156. https://doi.org/10.2147/PRBM.S95270
14. Browning, M., Fonberg, J., & Schellenberg, G. (2022). Neighbourhood characteristics of lower-income families in census metropolitan areas. Statistics Canada. https://doi.org/10.25318/36280001202200400002-eng
15. Anser, M.K., Yousaf, Z., Nassani, A.A. et al. Dynamic linkages between poverty, inequality, crime, and social expenditures in a panel of 16 countries: two-step GMM estimates. Economic Structures 9, 43 (2020). https://doi.org/10.1186/s40008-020-00220-6
16. Metz, N., & Burdina, M. (2018). Neighbourhood income inequality and property crime. Urban Studies, 55(1), 133-150. https://doi-org.myaccess.library.utoronto.ca/10.1177/0042098016643914
17. Nieuwenhuis, J., & Hooimeijer, P. (2016). The association between neighbourhoods and educational achievement, a systematic review and meta-analysis. Journal of housing and the built environment : HBE, 31(2), 321–347.https://doi.org/10.1007/s10901-015-9460-7
18. Kang H-W, Kang H-B (2017) Prediction of crime occurrence from multi-modal data using deep learning. PLoS ONE 12(4): e0176244. https://doi.org/10.1371/journal.pone.0176244
19. Bosick, S., & Fomby, P. (2018). Family Instability in Childhood and Criminal Offending during the Transition into Adulthood. The American behavioral scientist, 62(11), 1483–1504. https://doi.org/10.1177/0002764218787000
20. Jingyi He, Hao Zheng, Prediction of crime rate in urban neighborhoods based on machine learning, Engineering Applications of Artificial Intelligence, Volume 106, 2021, 104460, ISSN 0952-1976, https://doi.org/10.1016/j.engappai.2021.104460.
21. S. Kim, S. Lee. Nonlinear relationships and interaction effects of an urban environment on crime incidence: Application of urban big data and an interpretable machine learning method Sustainable Cities and Society, 91 (2023), Article 104419, https://doi.org/10.1016/j.scs.2023.104419
22. Widya Putra, D., Salim, W. A., Indradjati, P. N., &amp; Prilandita, N. (2023). Understanding the position of urban spatial configuration on the feeling of insecurity from crime in public spaces. Frontiers in Built Environment, 9. https://doi.org/10.3389/fbuil.2023.1114968
23. Ulmer, J. T., Harris, C. T., & Steffensmeier, D. (2012). Racial and Ethnic Disparities in Structural Disadvantage and Crime: White, Black, and Hispanic Comparisons. Social science quarterly, 93(3), 799–819. https://doi.org/10.1111/j.1540-6237.2012.00868.x

Appendix:

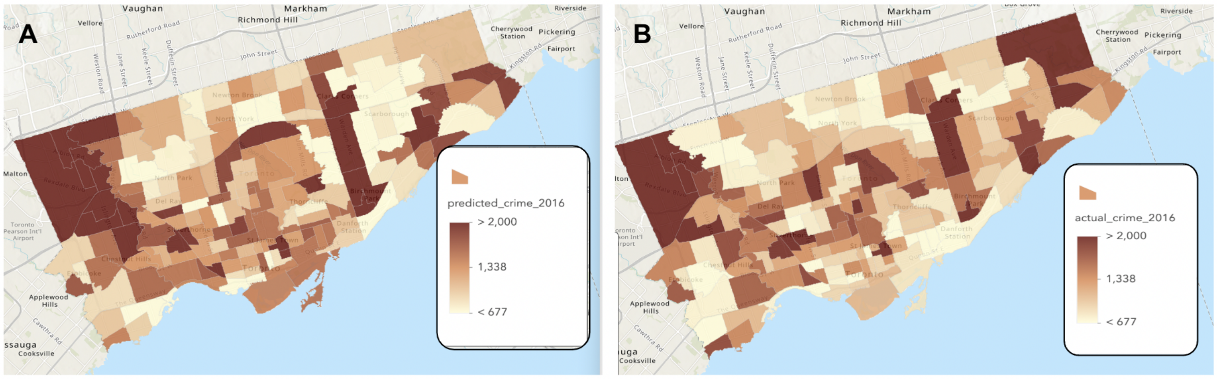
Dataset

https://docs.google.com/spreadsheets/d/1o1YIv3AwVg_Yhfb3G_96wo2crDurF_AR2yQZJ969az8/edit?usp=sharing

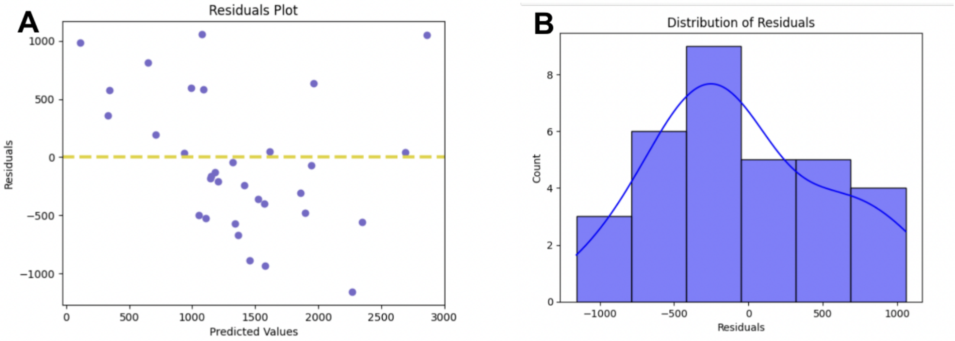
Correlation Matrix



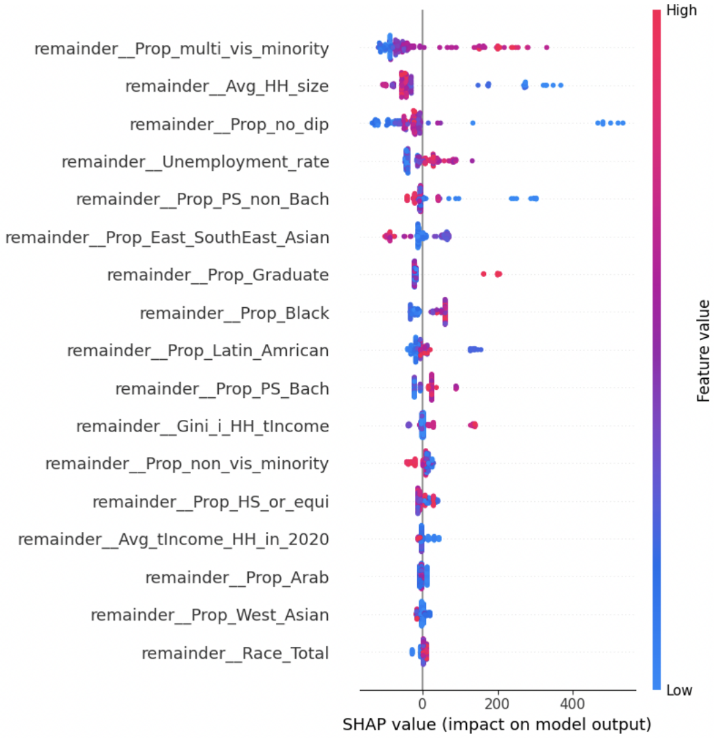
Geographical Depictions



MLR Residuals



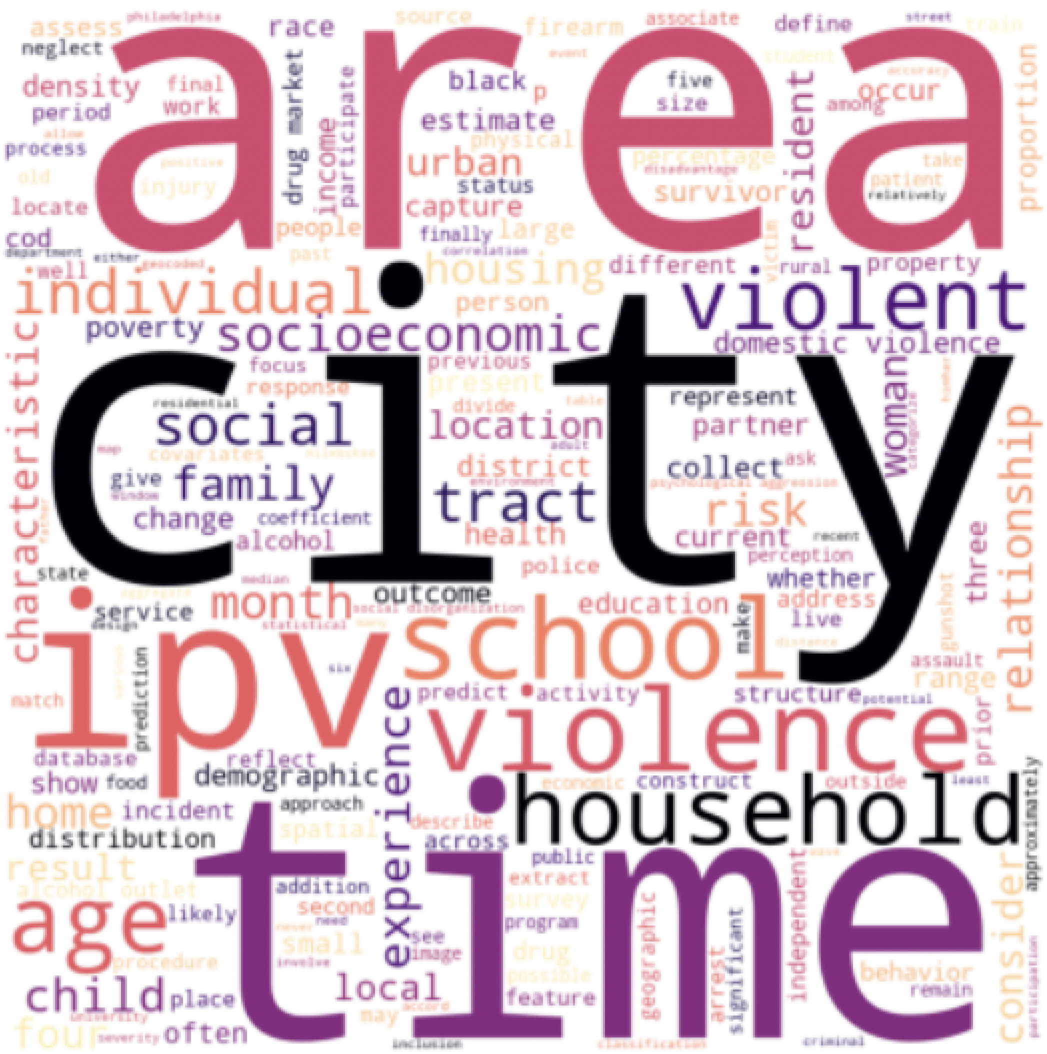
SHAP values Random Forest



References 52

https://docs.google.com/spreadsheets/d/1fyT1FKMa_wXoZuCPoxNtYh1fo1eOaKzPulqnJvguaNQ/edit?usp=sharing

Wordcloud


area
city
time
ipv
school
violence
household
violent
individual
socioeconomic
social
tract
family
location
age
child
experience
characteristic
relationship
resident
housing
urban
month
risk
woman
domestic violence
demographic
local
home
result
four
consider
assess
neglect
density
period
process
locate
cod
well
race
final
work
injury
drug market
income
participate
black
estimate
capture
status
large
people
finally
physical
firearm
associate
define
five
size
occur
among
percentage
survivor
patient
take
proportion
different
rural
property
person
poverty
focus
response
previous
present
represent
partner
collect
district
give
change
coefficient
alcohol
health
current
perception
police
whether
state
service
outcome
education
address
live
three
gunshot
assault
range
prior
make
predict
activity
structure
potential
prediction
database
show
food
reflect
incident
approach
spatial
describe
across
distribution
alcohol outlet
addition
second
likely
see
image
small
procedure
drug
possible
public
extract
survey
program
geographic
arrest
significant
independent
behavior
remain
approximately
economic
construct
outside
least
place
often
may
feature
inclusion
criminal
classification
severity
source
street
philadelphia
student
train
event
accuracy
relatively
disadvantage
past
correlation
victim
department
either
geocoded
residential
divide
table
adult
map
covariates
ask
environment
psychological aggression
recent
median
social disorganization
statistical
design
six
match
need
involve
university
accord
participation
allow
positive
old
p