Abstract

Data science finds application in almost all fields relating to academics and society in general. The aim of this report is to use integrity-driven data analysis to obtain truthful, reliable and interpretable results about the factors that affect crime rate in US communities. Linear Regression, Lasso Regression, Random Forests and a K-Nearest Neighbour classifier are applied to the 'Communities and Crime Data Set' created by UCI's Machine Learning Repository. It was found that given the appropriate model certain community factors, such as the population proportion of certain races and median income levels, were reasonably accurate predictors of violent crime rates. Results varied significantly state by state, however, Southern states were typically associated with higher crime rates.

Introduction

The Journal of Law and Economics published an article estimating the annual cost of crime in the United States to be \$4.71-\$5.76 trillion dollars (Anderson, 2021). Various recent economic and political events, such as the introduction of the 'Violent Crime Control and Law Enforcement Act' and the legalisation of abortion, have greatly influenced crime rates in the US (of Justice & of America, 1995). The fluid and volatile nature of American society today has lead to a number of hypotheses relating to the causes of crime rates, for example the increase of immigration to the US (Wadsworth, 2010). In this report, an attempt is made to clarify **whether there exists concrete, statistical relationships between these factors (immigration, racial composition, income etc.) and crime.** The dataset used in this report is that of the 'Communities and Crime Data Set' created by UCI Machine Learning Repository. The hope is that one can extrapolate findings from this dataset that are representative of crime in the US in general. 'The data combines socio-economic data from the 1990 US Census, law enforcement data from the 1990 US LEMAS survey, and crime data from the 1995 FBI UCR' to create a dataset with 128 attributes (Repository, 2009). The methods of analysis chosen in this report, are hugely affected by the fact that all of the data has been normalized between 0 and 1. Therefore, **one must be careful to interpret results correctly when comparing attributes**.

It is inevitable that human errors will come into play when analysing data. We can draw ideas from psychology, such as priming (Tulving & Schacter, 1990) and hindsight bias (Fischhoff, 1975), to understand the harmful effects of these errors when performing data analysis. **Preregistration** is a common tool used to mitigate against such effects whereby research questions are laid out in advance of analysing the data itself (Nosek, Ebersole, DeHaven, & Mellor, 2018). This helps to distinguish between discovery and justification, ultimately leading to more reliable and interpretable results. In the context of this report, exploratory analysis is performed in combination with tailored analysis aimed at answering certain predetermined questions in the hope that this will provide an unbiased, yet well-rounded, picture of crime in the US. This is a standard approach in cases where a comprehensive list of predetermined questions is not feasible (Gelman & Loken, 2014). The predetermined questions we seek to answer are as follows:

- Is there a significant relationship between the **ethnic/racial composition** in a community and the level of violent crime? The hypothesis is motivated by similar studies, based on unrelated data, suggesting that there indeed exists such a relationship (Walker, Spohn, & DeLone, 2016) (Peterson & Krivo, 2005).
- Do **income and education** have notable effects on crime rates? Extensive research supports the existence of these effects (Patterson, 1991).
- Given certain features of a community, can crime rate be accurately predicted?

Background

There exists extensive research on factors affecting crime rates, particularly in the US. Several books comprehensively explore this topic through up-to-date literature, many in the context of race (Gabbidon & Greene, 2018). The broad conclusion from these texts is that the influence of variables contributing to crime is highly discriminatory, depending significantly on some communal factors and not others. (Ousey & Kubrin, 2009), for example, demonstrated an **inverse relationship between immigration rates and violent crime rates** based on US census data 1980 -2000 for large US cities. On the other hand, (Kirk, 2021) investigated the **proportional effects of incarceration policy on violent crime** in US neighbourhoods. As is the case for these two studies, the common method of analysis in this area is the use of standard regression models to investigate the relationship between a single societal factor and crime. In contrast, for this report **additional regression and classification machine learning algorithms** are implemented, examining the influence of a **broader range of attributes, such as income, race and education, on violent crime**. Existing research on the UCI dataset itself has focused primarily on classification tasks, using Naive Bayes and Decision Stump algorithms, for example, with the paramount goal of improving model classification accuracy (Iqbal, Murad, Mustapha, Panahy, & Khanahmadliravi, 2013), (McClendon & Meghanathan, 2015). The purpose of this report is to try and understand and measure the relationship of the community factors themselves with crime rather than to optimise prediction accuracy. The emphasis placed on **integrity-based statistics, through preregistration and multiverse analysis**, is also somewhat distinctive. Additionally, a unique subset of attributes is thoughtfully chosen, deemed to provide the potential for interesting statistical study. It is important also to mention the foundational article associated with this dataset, written by its creator (Redmond & Baveja, 2002). The paper presents an entirely new AI software 'that helps police departments develop a strategic viewpoint toward decision-making'. The building of similar software, as well as the use of black box-style analysis, is beyond the scope and intention of this report.

Data and Methods

Data and Chosen Attributes

To gain a more informed perspective on the target of this dataset, it is important to examine the definitions[1] associated with violent crime. The following refer to the FBI Uniform Crime Reports (UCR) data source previously mentioned (FBI, 2009).

- Murder: 'The wilful (non-negligent) killing of one human being by another' . Things not included by the UCR include accidental deaths, suicide and justifiable homicides.
- Rape: 'The carnal knowledge of a female forcibly and against her will'.
- Robbery: 'The taking or attempting to take anything of value from the care, custody, or control of a person or persons by force or threat of force or violence and/or by putting the victim in fear'.
- Assault: 'An unlawful attack by one person upon another for the purpose of inflicting severe or aggravated bodily injury'.

It is reasonable to assume that the US Census and FBI UCR are reliable sources. The LEMAS survey, however, leaves more room for inaccurate and biased data. By manually exploring the dataset and reviewing its documentation (Repository, 2009), it was found that **approximately**

[1] Definitions of violent crime acts can differ from state to state, however, data with significantly different measures were omitted prior to this report. There exists some more granular distinctions, deemed superfluous for this analysis, available at (FBI, 2009).

84% of the policing data (LEMAS) was missing. Additionally, the **county and community code columns both contained roughly 60% missing values**. There exists considerable theory regarding approaches for dealing with missing data, from simple solutions of replacement with the attribute mean, to more complex maximum likelihood approaches (Little & Rubin, 2019), (Allison, 2001). However, given the extent of both relevant and interesting attributes with no missing values, there exists potential for insightful analysis without the need for these processing steps. Due to this abundance, LEMAS variables were dropped completely[2]. The remaining dataset contained 1994 observations of 1 target variable and 104 attributes, each describing community aspects ranging from housing density to divorce rates. A small subset of these attributes, were chosen as seen in Table 1 below. The non-predictive state variable was included for exploratory purposes but was not used as an input to any models.

Features	Description
ViolentCrimesPerPop	Numeric - decimal Total number of violent crimes (murder, rape, robbery and assault) per 100K population.
state	Numeric - US state.
population	Numeric - decimal Population for community
racepctblack	Numeric - decimal percentage of population that is African American.
racePctWhite	Numeric - decimal percentage of population that is caucasian.
NumImmig	Numeric - decimal total number of people known to be foreign born.
medIncome	Numeric - decimal Median household income.
PctPopUnderPov	Numeric - decimal percentage of people under the poverty level.
PctBSorMore	Numeric - decimal Percentage of people 25 and over with a bachelor's degree or higher education.
PctUnemployed	Numeric - decimal Percentage of people 16 and over, in the labor force, and unemployed.

Table 1: Table of nine features and one target variable relating to violent crimes. This subset was chosen to reduce the complexity of analysis, containing features of societal relevance and exhibiting potentially interesting relationships.

Exploratory Data Analysis

Before conducting statistical tests on the chosen dataset, one must acknowledge the impact data (pre) processing has on results. The effect of how the data is transformed, categorised and engineered must be taken into account. The **conclusions in this report are sensitive to the inevitable arbitrariness of the data's construction**. One method to alleviate this effect is through the use of multiverse analysis (Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016). With this in mind, the results of many statistical analyses, on differently-preprocessed datasets, will be combined with the aim of **presenting a broader, more representative picture of reality**. Furthermore, the potential for randomly-favourable/cherry-picked results will be reduced by running, and averaging over, multiple iterations on analyses of a random nature. This helps to avoid a prevalent phenomenon in contemporary statistical analysis call p-hacking (Gelman & Loken, 2014)[3].

Basic exploratory data analysis (EDA), yields the following insights from the chosen dataset:

[2]This also avoids potential (non) response bias likely associated with this data source.

[3]The motivation for including these ideas is to show awareness of them. To perform full preregistration of a large set of possible data structures is infeasible given the scope of this report. However, a representative sample will be chosen and their results compared.

- 1994 observations, 9 features, 1 target variable, 46 states represented.
- No NA or other missing values.
- All variables are floating numbers except 'state' which is of type integer.
- Target variable (violent crime) statistics: mean[4] ≈ 0.2380, standard deviation ≈ 0.2330, min $= 0.0000$ and max $= 1.0000$.
- The normalized data is relatively sparse with 826 zero values for proportion of white people, for example. Race attributes are still balanced as 0 values are valid given the normalisation.

Due to the normalisation, and hence the larger concentration of smaller target variable values, a log-transformation is tested on the crime variable to potentially bring about the advantageous properties of normality, Figure 1. The non-transformed data is markedly right-skewed indicating a majority of safe communities relative to a few particularly dangerous ones.

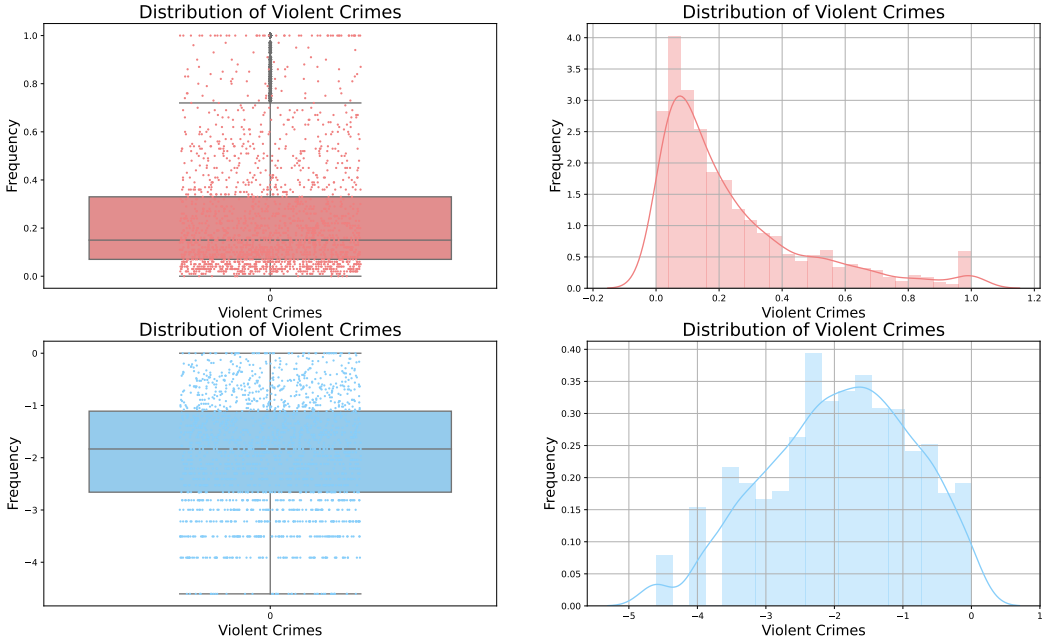

Figure 1: The distribution of violent crimes is given by the boxplots and histograms above. Jitter is added to the datapoints for visual effect. A density curve is appended over the histogram bins as a clearer suggestion of the real underlying distribution. The blue graphs indicate the log-transformed distribution. The effect of the log-transformation is clear with a shift upwards of the mean and quartile values as well as a distribution more resembling that of a normal distribution. This may make it more suitable for the statistical models to follow.

Figure 2 shows four correlation values greater than 0.7. In general, these relationships are expected, namely, one would expect there to be a higher poverty rate for lower median income communities and few black people in a community with more white people[5]. Some of these relationships can be explained by the definitions of the attributes, for example, a community with a larger population is likely to have a higher number of immigrants. Variables with **high correlations will have to be dealt with cautiously when working with models** such as linear regression models. The 'ViolentCrimesPerPop' column is of particular interest, showing a **strong relationship between**

[4] All numerical results are given to four significant figures.

[5] It is important to remember that we are still talking about normalised values rather than the absolute values themselves i.e. the interpretation of the relationships has somewhat changed.

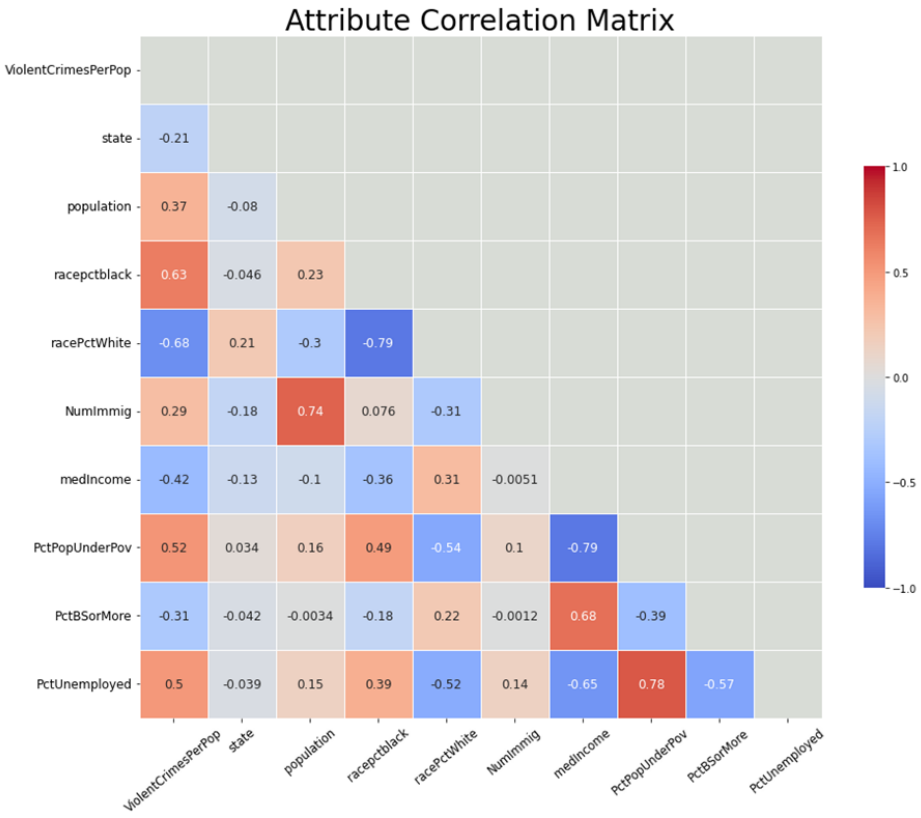

Figure 2: Heatmap showing correlations between attributes within reduced dataset. Darker red and blue values indicate higher positive and negative correlations respectively.

crime and black/white population proportion, as well as other variables such as poverty and unemployment.

Figure 3 illustrates eleven **states with the highest normalised violent crime rates** per 100K population along with the number of datapoints in the dataset relating to that state in brackets[6]. The top state (Washington) was removed as it contained only a single datapoint with a crime rate of 1 which resulted in a misleading graph. Each of the 46 states are not equally represented in the dataset so one must be careful when dealing with this imbalance. There is **potential for bias here, however, in this case more datapoints do not necessarily imply a higher rate of crime**, therefore, the between-state comparison can still yields valid insight. **The Southern states fill a disproportionate amount of the top spots**.

Figure 3: Ranking of crime rates per state. Size of state names are indicative of their rates of violent crime.

[6]State names were transcribed from the community names.

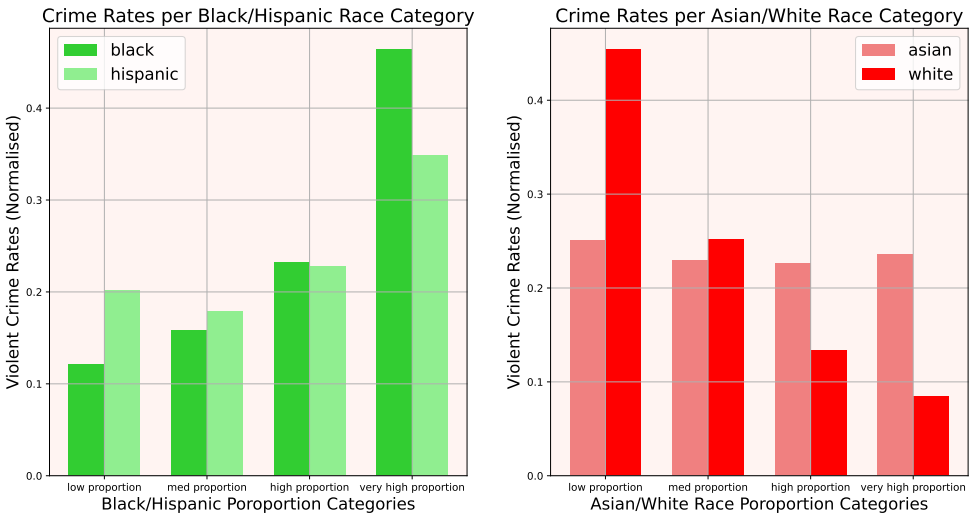

Figure 4: Violent crime rate per race category. Lighter coloured bars show weaker trend between crimes and race. Communities with a larger proportion of Hispanic and black people, in particular, tend to have higher rates of crime. The inverse is true for white people. A less distinct relationship is evident for the Asian race. Note these results are sensitive to the arbitrary category construction.

Results

- Multiverse Analysis: To avoid the processing-dependent results as discussed in the previous section, a multiverse approach is taken here. Various data cleaning and transformation steps are performed to form separate datasets using (log) **transformations, feature selection and outlier removal**. The model outputs, affected by each of these data processing decisions, are contrasted with one another to give a more representative view of the relationship between crime and community factors.

- p-hacking: Due to the randomness of the train/test split, results are averaged over 1000 splits. Relevant standard deviations are also provided in brackets.

One particular relationship of interest as mentioned in the introduction is that of **community ethnic/racial composition and crime rate.** To investigate this, two common models, linear regression and random forest, will be implemented on a modified version of the reduced dataset[7]. The thirteen attributes describe the community proportion and income of each race (white, black, Indian, Asian and Hispanic) as well as its number of immigrants. The importance of each feature in explaining violent crimes is measured and compared across the models. Some assumptions about the data must be made when interpreting model results[8]. Normal Linear Regression: Standard linear regression, including an intercept term, will be performed here. There must be no multicollinearity amongst the predictors. Errors are assumed to be normal and residuals are assumed to be homoscedastic. This model assumes a linear relationship between the predictors and the target variable. Random Forest: There are few assumptions regarding these types of models, however, they perform better when the data is independent and identically distributed. Again, variables with lower correlations are preferred. Cross-validation will be used to tune the important hyperparameters of this model: number of trees, number of features to consider when looking for the best split and minimum number of samples required to be at a leaf node. To adhere to these assumptions, 'racePctWhite' and 'NumImmig' are removed as predictors due to their strong correlation with

[7]This includes some variables not described in the introduction but which are similarly defined and contain no missing values. These variables form a good representative sample of the ethnic/racial variables in the dataset.

[8]As the data is numeric only, there is no need to worry about non-numeric/categorical inputs.

other variables. The results in Table 2 and Figure 5 give an overview of the relationship between ethnic/racial composition and crime rates.

Model	RMSE	MAE
Linear Regression Model (no transformations)	0.1432 (± 0.0075)	0.0972 (± 0.0046)
Linear Regression Model ('racePctblack')	0.1872 (± 0.0094)	0.1351 (± 0.0083)
Linear Regression Model (log transformation all variables)	0.1443 (± 0.0081)	0.0999 (± 0.0068)
Linear Regression Model (0s removed)	0.1322 (± 0.0094)	0.0938 (± 0.0062)
Random Forest (no transformations)	0.1265 (± 0.0075)	0.0829 (± 0.0044)
Random Forest (log transformation all variables)	0.1516 (± 0.0101)	0.1028 (± 0.0070)
Random Forest (0s removed)	0.1073 (± 0.0087)	0.0792 (± 0.0059)

Table 2: Mean (± standard deviation) Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) of (non) transformed NLMs and random forest models based on a 80/20 train-test split of the data. This exemplifies the multiverse analysis. The random forest model with 0s removed has the best violent crime prediction performance given the ethnic/racial predictors based on RMSE and MAE.

From Figure 5, there is evident agreement between the two models, especially with regards to the **strong relationship between crime and black/Hispanic proportion relative to other attributes**. Population[9] and black per capital income also seem to be good predictors of violent crime in a community. Asian-related attributes have little bearing on crime rates it seems. Model comparison yields inconclusive results when it comes to the effect of immigration. Low RMSE and MAE values, relative to the normalised scale, indicate small crime rate prediction errors on test data. Coupling this with reasonably high R^2 model values (≈ 0.64), **suggest a significant relationship between ethnicity/race and crime rates**. We fail to reject the preregistered hypothesis as per the introduction. In assessing model reliability, one must keep in mind that the results are only valid if the relevant assumptions are upheld. Given the relatively high correlations between the variables, one must be cautious in interpreting the results. Additionally, one must consider model bias. The models with 0 values removed are likely to be less biased as they alleviate the influence of extreme values more prominent with some attributes and not others. In saying this, the **built-in normalisation helps to alleviate potential bias associated with different scales.**

For the following analysis, the **original modified dataset with 9 features is used**. Due to the apparent sparsity of the data, a **Lasso regression model** is implemented. This model's regularisation term tends to remove poorer predictors and reduces overfitting. Tuning this hyperparameter involved 10-fold cross validation which provided a good balance between tuning accuracy and computational expense. As with the previous models, Figure 6 shows the Lasso regression coefficients averaged across multiple different datasets and train-test splits in accordance with the multiverse approach. Tying back to the preregistered questions, it seems that **income plays a significant role in predicting crime, whereas, the education variable is deemed irrelevant by regularisation**. This is somewhat of a different outcome to the one hypothesised in the introduction. **This highlights the importance of correctly analysing data over trying to match a preconceived idea.**

Using this reasonably accurate model, one can predict crime rate for an arbitrary community. For example, given a community with a relative low proportion of white people (0.2), large number

[9]The non-racial attribute population is also included as a comparative measure for models to follow. Results did not differ significantly in its absence.

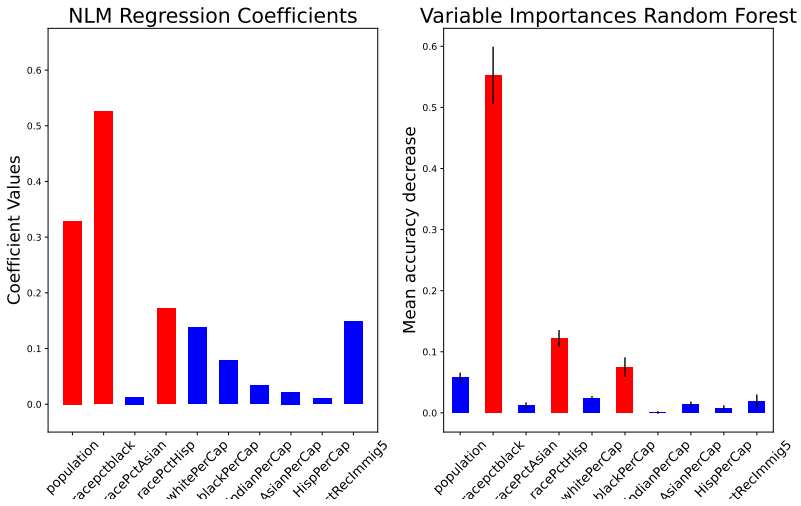

Figure 5: Mean NLM absolute-valued regression coefficients (left) vs random forest variable importance, including standard error bars, (right) for each feature. Models with 0s removed were used in both cases. Permutation importance is used to calculate variable importance. The three most important features are highlighted in red. **Determining variable importance through coefficient values is possible due to normalisation**.

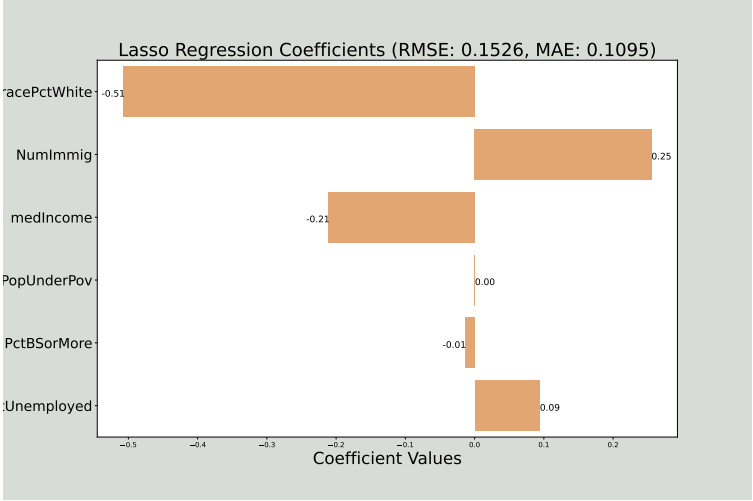

Figure 6: Mean Lasso regression coefficients with model RMSE and MAE. Low 'PctPopUnderPov' and 'PctBSorMore' variable values in particular confirms efficacy of Lasso regression sparsity. Three highest coefficient bars are shaded.

of immigrants (0.8), low median income (0.2) and an average amount of all other variables, the model predicts a crime rate of 0.7310. Linking back to the final pre-registered question, the RMSE and MAE values suggest that crime can indeed be accurately predicted given certain features of a community. Again, scaling reduces bias for the Lasso regression results.

Precise prediction values are not always necessary, therefore, it is often useful to categorise a target variable. As results are highly dependent on the choice of categorisation, different violent crime rate categories are implemented and compared. A **K-Nearest Neighbors (KNN) algorithm** is utilised in order to accurately classify communities based on the given nine features. Cross-validation is used to choose the optimal k value from a range of 3 - 20. Results are averaged over 1000 train-test splits. In Figure 7, crime rate was divided by its quartiles, however, when using even splits or fewer categories, the KNN algorithm performs significantly better with average accuracy scores of 0.7388 and 0.8897 respectively. This again supports the need for a multiverse analysis approach.

The results seem to show that, with the appropriate categorisation, the explanatory variables in this dataset are strong predictors of violent crime rates. However, we must be wary of these results as there exists some **bias introduced by the imbalance of some categories.** For example, the improved results of the evenly-spaced categories may be partially attributed to the fact that the vast majority of crime rates lie in the smallest partition. The algorithm is likely assigning the majority of unseen data point to this category regardless of community variables. Therefore, we could conclude that the **chosen community attributes are in fact not great predictors of categorised crime rates.**

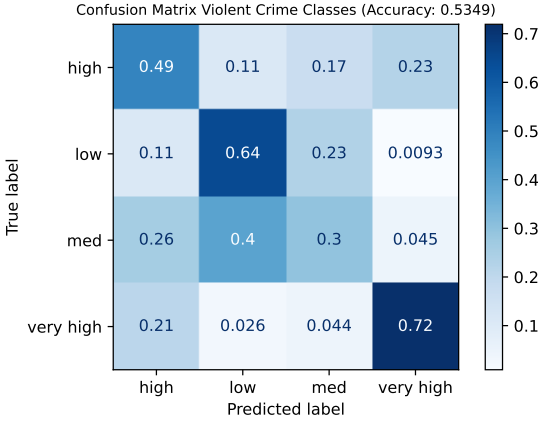

Figure 7: Confusion matrix showing relatively poor performance of KNN algorithm for quartile-based categorisation. A k value of 5 - 7 was optimal. Algorithm is best at classifying edge categories 'low' and 'very high'. Different categorisation choices lead to 'improved' results.

In order to determine whether the above results changed depending on state, one must compare on an even scale with a sufficient amount of data. Therefore, only the states with the highest number of related datapoints (> 100) were chosen: California, New Jersey, Texas, Massachusetts and Pennsylvania. This discrimination on the linear regression model yielded some interesting results. **The importance of certain community factors varied wildly by state.** For example, the proportion of black population was by far the most significant predictor in California, whereas the number of immigrants had a large effect on crime rate for Pennsylvania. The chosen attributes were relatively strong predictors of violent crime rates in New Jersey (MAE = 0.0680) but relatively poor predictors for Texas (MAE = 0.1261).

Conclusion

Given more state datapoints, I would have liked to construct a geographical map of crime rates throughout the US, perhaps, using the Folium Python library on longitude/latitudes coordinates transcribed from the state numbers. This would have enabled more insightful visualisation of cluster analysis. The given dataset, with its many normalised attributes, may have also been suitable for use within neural networks. A transition to the unnormalised version of the UCI dataset would have opened up more possibilities for exploring relationships between attributes other than crime rates. The improved interpretability would have been advantageous as this plays a large role in data science, especially when communicating results to a non-technical audience. The breadth of models included in this report came at a cost of statistical complexity. This was in line with the aforementioned intentions of the analysis, however, in other contexts I would have liked to improve the sophistication of the models. This could include the appreciation of factor variables in the regression analysis or more rigorous hyperparameter tuning for the random forest model.

All results in this report refer to the normalised interpretations of the familiar attribute definitions. As one might expect, common community factors, such as racial composition, size and income, have been shown to be reasonable predictors of (non-categorised) violent crime rates. More specifically, the proportion of white, black and Hispanic community populants, and their respective incomes, are responsible for explaining the majority of crime rates out of the chosen subset of important variables. The inaccuracy of the models described in this report can be attributed to the exclusion of possibly better predictors, the random nature of crime in US communities and, in part, to some acknowledged model biases. Southern states seem to contain the majority of violent communities. The drastic variation in results observed between states supports the choice to opt for a multiverse

approach throughout. The temptation to choose outcomes which randomly-favoured a preconceived idea was reduced through averaging over many iterations. Furthermore, the preregistration of some relevant hypotheses imposed needed structure on the analysis. This diagnosis is a true testament to the validity of such integrity-based approaches to statistical analysis.

References

Allison, P. D. (2001). *Missing data*. Sage publications.

Anderson, D. A. (2021). The aggregate cost of crime in the united states. *The Journal of Law and Economics*, *64*(4), 857–885.

FBI. (2009). *https://www2.fbi.gov/ucr/cius2009/offenses/violent_crime/.*

Fischhoff, B. (1975). Hindsight is not equal to foresight: the effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human perception and performance*, *1*(3), 288.

Gabbidon, S. L., & Greene, H. T. (2018). *Race and crime*. Sage Publications.

Gelman, A., & Loken, E. (2014). The statistical crisis in science data-dependent analysis—a "garden of forking paths"—explains why many statistically significant comparisons don't hold up. *American scientist*, *102*(6), 460.

Iqbal, R., Murad, M. A. A., Mustapha, A., Panahy, P. H. S., & Khanahmadliravi, N. (2013). An experimental study of classification algorithms for crime prediction. *Indian Journal of Science and Technology*, *6*(3), 4219–4225.

Kirk, E. M. (2021). Community consequences of mass incarceration: sparking neighborhood social problems and violent crime. *Journal of Crime and Justice*, 1–17.

Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data* (Vol. 793). John Wiley & Sons.

McClendon, L., & Meghanathan, N. (2015). Using machine learning algorithms to analyze crime data. *Machine Learning and Applications: An International Journal (MLAIJ)*, *2*(1), 1–12.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606.

of Justice, U. D., & of America, U. S. (1995). Violent crime control and law enforcement act of 1994: Briefing book.

Ousey, G. C., & Kubrin, C. E. (2009). Exploring the connection between immigration and violent crime rates in us cities, 1980–2000. *Social problems*, *56*(3), 447–473.

Patterson, E. B. (1991). Poverty, income inequality, and community crime rates. *Criminology*, *29*(4), 755–776.

Peterson, R. D., & Krivo, L. J. (2005). Macrostructural analyses of race, ethnicity, and violent crime: Recent lessons and new directions for research. *Annu. Rev. Sociol.*, *31*, 331–356.

Redmond, M., & Baveja, A. (2002). A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research*, *141*(3), 660–678.

Repository, U. M. L. (2009). *Communities and crime data set.*

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702–712.

Tulving, E., & Schacter, D. L. (1990). Priming and human memory systems. *Science*, *247*(4940), 301–306.

Wadsworth, T. (2010). Is immigration responsible for the crime drop? an assessment of the influence of immigration on changes in violent crime between 1990 and 2000. *Social Science Quarterly*, *91*(2), 531–553.

Walker, S., Spohn, C., & DeLone, M. (2016). *The color of justice: Race, ethnicity, and crime in america*. Cengage Learning.