**Using machine learning methods to predict rates of crime in Bristol and assessing their effectiveness**

**Introduction**

According to Bristol City Council the total number of recorded crimes in Bristol in 2020/21 was 46,821 which represents 101 crimes per 1,000 of the population and in a survey accompanying the document, 15.7% of Bristol residents said that fear of crime was affecting their day to day lives (Bristol City Council, 2021). In their latest figures Avon and Somerset police report that they are spending £372.2 million annually on police (Avon and Somerset Police, 2020).

Map

Description automatically generated

**Figure 1**-Map depicting the geographical locations of the Bristol Wards.

Earlier in 2021 each Bristol resident was required to fill in a census. The aim of this research is to see if it is possible to predict future crime rates in the 34 Bristol Wards, depicted in **Figure 1**, based on the population demographics gathered from the census. Previous research has shown that focusing police efforts on high-activity crime places, known as hot spotting, can lead to crime and disorder reductions without a subsequent movement of crime to nearby areas (Braga, et al., 2012).

Datasets containing historic data were used from Open Data Bristol (Open Data Bristol, 2021). To choose the datasets to merge from Open Data Bristol for predicting crime rates, a literature review was carried out to determine what researchers have determined to be the main indicators of crime in an area.

Due to factors such as inflammatory media coverage, in Western nations large amounts of the public believe that racial minorities contribute to higher levels of crime than the native population (Wortley, 2009). This doesn’t appear to be supported by studies and it is more likely to be social and economic factors that contribute to higher crime levels (Short, 1997). The percentage of people who identify as from a Black and Minority Ethnic group (BAME) and the index of Multiple Deprivation Score were included to assess these two features.

Elsewhere, it has been noted that in literary depictions- such as Captain Hook and Shakespeare’s Richard III – physical deformities have been used to associate the character with being the bad guy even though there has been little evidence to suggest people who are disabled are more likely to commit crimes (Dahl, 1993). Percentage of people who identify as their day-to-day activity being limited for the Ward have been added as the closest link to this just in case though.

An additional report suggests that decent housing helps to prevent crime and create stable neighbourhoods (Friedman, 2010). This was harder to link to the datasets provided but this has been represented by percentage of households that are overcrowded.

Finally, a random forest regressor was used to predict homicides in Brazilian cities using urban indicators and found that unemployment and illiteracy were the most important variables (Alves, et al., 2018). The percentage of 16-17 year olds not in education or employment (NEET) was loosely categorised as this group.

**Exploratory Data Analysis**

Graphical user interface, application, Word

Description automatically generated

**Table 1**- Table depicting the summary statistics of the dataset.

Python and the Pandas library were used to merge the six datasets and produce the summary statistics table seen in table 1. The first column is the target variable and the other five the input variables. As there are only thirty-four wards, there are only thirty four records to train the machine learning models on which is lower than ideal. As a minimum it is suggested that there are ten times the number of degrees of freedom which in this case would be fifty (Gonfalonieri, 2019). The low number of records is going to result in it being difficult to get a good bias-variance tradeoff – lack of data means models being high in bias; more complex models will very quickly begin to overfit and be high in variance.

The mean rate of crimes in all wards for this data is 113 per 100 people. The mean for the input variables varies from 4.98 to 25.39 with four of five percentages which suggests that standardising the data isn’t needed. There are no missing values which need to be considered.

Chart, bar chart

Description automatically generated

**Figure 2-** A bar chart depicting the crime rate in each Bristol Ward.

Figure 2 suggests that most wards have a crime rate between 80 and 110 per 1,000 people. There are two significantly higher peaks in the Central and Hotwells and Harbourside Wards. Looking at the Ward map in figure 1, these are in the centre of Bristol and the crimes may be being committed by people not living in the Ward areas- they contain large numbers of shops (shoplifting) and bars/nightclubs (incidents related to alcohol). They also contain the routes for protest marches such as for Extinction Rebellion and for Kill the Bill. As such they are likely to be known hot spots already targeted by police and the crime rate is less likely to be linked to the demographics of people living there. It might be more beneficial to train the models excluding these Wards, but it was decided that it would be best not to make assumptions and with so few records they have been left in.

Chart, scatter chart

Description automatically generated

**Figure 3-**A pairplot depicting the relationships between variables.

Graphical user interface, text, application

Description automatically generated

**Table 2-**A correlation matrix showing the correlation between variables.

The correlation matrix produced for table 2 suggests that there is a strong positive correlation between the crime rate and the percentage of 16-17 year old NEETs in the Ward’s area. There are weak positive correlations between crime rate and how deprived the Ward is, the percentage of households that are overcrowded and the percentage of people who live in the Ward that are BAME. There is no positive correlation between crime rates and the percentage of people in a Ward whose day-to-day activities are limited.

Figure 2 also highlights some strong correlations between the input variables. Ward’s containing overcrowded households have a very positive correlation with Ward’s with higher percentages of people identifying as BAME. There are also strong positive correlations between Ward’s that are classified as deprived and the percentage of 16-17 year old NEETS, people whose day-to-day activities are limited and households that are overcrowded. As such there is likely to be some redundant information and it would probably be best to remove features to reduce the number of dimensions and the variance to improve the accuracy when building the models. At this stage the percentage of people whose day-to-day activities is limited was removed as it had the weakest correlation with the crime rate. The percentage of overcrowded households was also removed because of the very strong correlation with Wards with large numbers of people identifying as BAME and this featuring having a lower correlation with the crime rate.

**Model selection and deployment**

To predict the crime rate in a Ward from the input variables selected, six supervised regression model were tested using Python and the Scikit-learn library. The metric root mean squared error (RMSE) was used to evaluate which model is the most accurate.

For all the models k-Fold cross validation was used to reduce the bias of the models. With such a small dataset, if just randomly splitting it into testing and training data once then any outliers could disproportionately affect the model’s performance. In this investigation the data was split into eight (k=8). Seven groups were used as the training data and one as the testing data. The model was trained and tested and the RMSE recorded. The data was then resampled with a different group used as the testing data. This was conducted on each model eight times and the mean RMSE calculated.

For models 2-6 nested cross-validation was used to optimise the hyperparameters of the models and reduce the root mean squared error further as the models deployed using the Sci kit learn library have default hyperparameters which aren’t always the best for the dataset that is being used. In this instance, k-fold cross-validation for model hyperparameter optimisation is nested inside k-fold cross-validation procedure for model selection (Brownlee, 2020).

1. **Multiple Linear Regression:-** To begin with, a simple Multiple Linear Regression model was created with the rate of crime assumed to be linearly dependant on the three input variables. To fit the model, a coefficient is added to each input variable which gives a weight to each one.
2. **Ridge Regression:-** As there are only three input variables, the linear model shouldn’t be too complicated and high in variance. Even so, a ridge regression model was created to minimise the model complexity by reducing the coefficients on each input variable. The degree to which it constrains the model is set using the alpha hyperparameter.
3. **Lasso Regression:-** Lasso, similar to Ridge, is another regularisation method that minimises the model complexity by reducing the coefficients of each input variable. Unlike Ridge, it can minimise the coefficients to zero and remove some input features completely (though it doesn’t in this instance). Again, the alpha hyperparameter can be tuned to constrain the model.
4. **Decision Tree Regressor:** A decision tree arrives at an estimate of the rate of crime by asking a series of questions to the data, each time narrowing down values by asking a series of true or false questions. By constantly splitting through questions, it is possible to get to the correct value each time with the training data, but this leads to high variance and poor performance on the test data. Hyperparameters are tuned to prevent this, for example by limiting the depth of the decision tree (Drakos, 2019).
5. **Random Forest Regressor:** This is a variant of the Decision Tree Regressor which combines multiple decision trees into one model. An average of all the decision trees is produced to make a prediction. It is a bagging technique which limits the number of features that can be split at each node and draws a random sample from the dataset when generating its splits (Chakure, 2019).
6. **K-Nearest Neighbour Regressor:** This regressor predicts the crime rate using a memory-based model. For a new data point it compares this to the closest data points from the training set and takes the average value. The number of neighbours to compare the new data point to is a hyperparameter that is tuned (Singh, 2018).

**Model results and evaluation**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | RMSE | Standard deviation | Coefficient for % of 16-17 year olds in Ward who are NEET | Coefficient for the Index of Multiple Deprivation | Coefficient for % of people in Ward who identify as BAME |
| Multiple Linear Regression | 53.85 | 25.97 | 12.99 | 1.61 | 2.05 |
| Ridge Regression | 53.90 | 26.24 | 12.64 | 1.53 | 2.04 |
| Lasso Regression | 53.84 | 25.98 | 12.88 | 1.56 | 2.04 |

**Table 3-** A table showing the metrics for the linear models

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | RMSE | Standard deviation | Feature importance for % of 16-17 year olds in Ward who are NEET | Feature importance for the Index of Multiple Deprivation | Feature importance for % of people in Ward who identify as BAME |
| Decision Tree Regression | 53.08 | 33.37 | 0.76 | 0.19 | 0.05 |
| Random Forest Regression | 51.15 | 29.83 | 0.76 | 0.19 | 0.05 |
| K-Nearest Neighbour Regression | 61.20 | 42.67 | N/A | N/A | N/A |

**Table 4-** A table showing the metrics for the non- linear models

The results in tables 3 and 4 show that the Random Forest Regression model performed the best with a RMSE of 51.15 and the K-Nearest Neighbour Regression model performed the worst with a RMSE of 61.20.

The three linear models had a RMSE only slightly higher than the Random Forest Regression model. All three linear models produced similar errors and coefficients, possibly because of the limited amount of data and low number of variables which meant that adding the regularisation techniques made little difference. The coefficient for percentage of sixteen to seventeen year olds NEET’s in the Ward is six times higher than the others meaning that this variable provides the biggest indicator for what the crime rate is going to be. These models run quickly and are easy to interpret.

The Decision Tree Regression Model and Random Forest Regression Model also acknowledge that the feature with greatest importance for predicting the crime rate is the percentage of sixteen to seventeen year olds NEETs in the Ward. What is different, though, is that where for the linear models the other two variables have a similar coefficient, these two models place a much greater emphasis on the Index of Multiple Deprivation comparative to the percentage of people in the ward who identify as BAME. Random Forest Regression, while producing the lowest RMSE, is the slowest to run.

**Conclusion**

Using a Random Forest Regression model, it is possible to predict the crime rate in a Bristol Ward with an error of 51.15 crimes per 1,000 people using the three variables chosen. At just under half the mean crime rate in a Ward (112 per 1,000) it is advised that this model is used to just get an idea of which Wards will have higher crime rates and then allocate police resources to investigate further.

Care should also be taken if using this model, as using an algorithm to predict crime rates with percentage of people who are BAME as an input variable could lead to controversy if released. As demonstrated, the feature which impacted the most on crime rates in a Ward was the percentage of sixteen to 17 year old NEETs and policies with education and employment for young people may be the most beneficial for the crime rate.

Finally, this model was trained on a very small set of data. A follow up investigation could take data from each year or from other cities if available to help improve the model.

Word count:2172

# References

Alves, L., Ribeiro, H. & Rodrigues, F., 2018. Crime prediction through urban metrics and statistical learning. *Physica A: Statistical Mechanics and its Applications,* Volume 505, pp. 435-443.

Avon and Somerset Police, 2020. *Your Policing Pounds.* [Online]   
Available at: https://www.bristol.gov.uk/documents/20182/391623/Avon+and+Somerset+Police+council+tax+document+2019-20.pdf/47de3399-31c2-0bcb-a670-a8f67f7d78c6  
[Accessed 17 08 2021].

Braga, A., Papachristos, A. & Hureau, D., 2012. Hot spots policing effects on crime. *Campbell Systematic Reviews,* 8(1), pp. 1-96.

Bristol City Council, 2021. *JSNA Health and Wellbeing Profile-Crime,* Bristol: s.n.

Brownlee, J., 2020. *Nested Cross-Validation for Machine Learning with Python.* [Online]   
Available at: https://machinelearningmastery.com/nested-cross-validation-for-machine-learning-with-python/  
[Accessed 17 08 2021].

Chakure, A., 2019. *Random Forest Regression.* [Online]   
Available at: https://medium.com/swlh/random-forest-and-its-implementation-71824ced454f  
[Accessed 17 8 2021].

Dahl, M., 1993. The role of the media in promoting images of disability: disability as metaphor, the evil crip. *Canadian Journal of Communication,* 18(1), pp. 75-80.

Drakos, G., 2019. *Decision Tree Regressor explained in depth.* [Online]   
Available at: https://gdcoder.com/decision-tree-regressor-explained-in-depth/  
[Accessed 17 8 2021].

Friedman, D., 2010. *Social impact of poor housing,* s.l.: s.n.

Gonfalonieri, A., 2019. *5 Ways to Deal with the Lack of Data in Machine Learning.* [Online]   
Available at: https://www.kdnuggets.com/2019/06/5-ways-lack-data-machine-learning.html  
[Accessed 17 08 2021].

Open Data Bristol, 2021. *Open Data Bristol.* [Online]   
Available at: https://opendata.bristol.gov.uk/pages/homepage/  
[Accessed 17 08 2021].

Short, J., 1997. *Poverty, Ethnicity, and Violent Crime.* 1st ed. New York: Routledge.

Singh, A., 2018. *A Practical Introduction to K-Nearest Neighbors Algorithm for Regression (with Python code).* [Online]   
Available at: https://www.analyticsvidhya.com/blog/2018/08/k-nearest-neighbor-introduction-regression-python/  
[Accessed 17 8 2021].

Wortley, S., 2009. Introduction. The Immigration-Crime Connection: Competing Theoretical Perspectives. *Journal of International Migration and Integration,* Volume 10.