DATA 695 Capstone Project

From Data to Insights: Investigating Social Factors in Crime using Machine Learning

Imad Ahmad

Zakir Ullah

GitHub Link

Contents

Introduction

Crime is a complex phenomenon with far-reaching effects on individuals, communities, and society. Crime is defined as the intentional commission of an act that is usually deemed socially harmful or dangerous and specifically defined, prohibited, and punishable under criminal law [1]. Decisions to commit crimes are determined by a complex interplay of factors, including physical abnormalities, psychological disorders, social and economic circumstances, personal choices, and impulse. While some individuals might plan crimes for financial gain or excitement while others act out of anger or fear. Punishment attempts discourage criminal behavior, but its effectiveness varies - factors like police presence and economic conditions influence crime rates. Education plays an essential role; low educational levels have been shown to correlate with criminal behavior, particularly property crimes [2]. It is important to understand the social factors that influence criminal behavior to develop effective strategies for crime prevention.

The use of Machine learning techniques has become a popular approach to crime analysis and prediction. By employing large datasets with complex algorithms, machine learning models are capable of recognizing patterns in crimes that occur regularly while forecasting future crime events [3]. Machine learning methods have revolutionized crime analysis by processing extensive data, spotting patterns, and making predictions. These techniques find application in predictive policing, crime pattern recognition, suspect identification, link analysis, sentiment analysis, fraud detection, resource allocation, crime classification, and early intervention.

Motivation

Recent advances in data science, machine learning and data mining have opened new ways to analyze data and derive insights. This project explores the use of machine-learning techniques to investigate the social factors that influence crime and gain valuable insights. This project has increased our knowledge and understanding about machine learning. It also gave us a chance to identify and solve problems related to data science in crimes analysis. The project has enabled us to utilize different tools and techniques for performing exploratory data analysis and using machine learning algorithms, which are important for the development of machine learning models. Overall, this project was a valuable opportunity to gain practical experience in machine learning and contribute to creating a more inclusive and accessible society.

Related Work

Different Machine Learning techniques have been used for crime prediction. Shah et al. (2021) discussed in their paper a study that employed violent crime patterns with WEKA data mining software where linear regression algorithms were applied to predict violent crimes. This paper also discussed a research project

that used ensemble models, logistic regression, neural networks, K-nearest neighbor methods and Bayesian methods to predict crime patterns at urban locations [4].

Machine learning utilizes statistical models and algorithms to process data and make predictions, while deep learning uses multilayered neural networks to model complex relationships. Both approaches offer solutions for crime prediction challenges. Decision trees, random forests and support vector machines have all been employed by machine learning (ML) algorithms to accurately forecast crime patterns based on individual cities' crime data. Furthermore, these algorithms offer insights into crime trends and correlations with environmental and demographic factors. Deep learning techniques, specifically convolutional and recurrent neural networks, have proven their worth in crime prediction. Trained in crime data with spatial or temporal components to accurately forecast crime patterns. Furthermore, deep learning also shows its worth when used for computer vision analysis by monitoring surveillance footage for signs of criminal activities or being integrated with drones or aerial technologies for enhanced monitoring capabilities. Predictive policing has long employed machine learning algorithms extensively for locating crime hotspots as well as allocating resources efficiently to ensure effective law enforcement [5].

Education and crime have an intricate relationship, with empirical evidence pointing towards increased levels of education leading to less involvement in criminal activities. Yet some research indicates that certain crimes such as tax fraud or embezzlement could increase as educational levels increase. Areas with lower educational levels typically experience higher crime rates. Education can have an enormously consequential effect on various forms of crime. While white-collar offenses such as money laundering and insider trading tend to be more attainable for those with higher incomes due to financial resources available to them, while theft and vandalism tend to be committed more by individuals with less formal education who may disregard potential long-term repercussions when making their decisions. Education can also have an impact on time preferences and self-control. Furthermore, education plays an integral part in shaping individuals' perceptions and actions regarding criminality [6]

Methodology

We have used the systematic literature review (SLRs) method for this project. This article contributes to three areas. First, by outlining existing studies on machine learning (ML) and deep learning (DL) applications to crime detection in neighborhoods; outlining publicly available datasets for future research; and proposing ways to address research gaps related to neighborhood crime studies. Our goal is to enhance understanding, guide future studies, and shed insight into these technologies' roles in crime prediction.

Machine learning in crime analysis involves various steps, such as data collection, preprocessing, feature selection, model training and prediction. The process we followed for this study consists of the following steps.

Data Collection: The data for this study was collected from Open Government portal of Canada.

Data preprocessing: The data was processed, and duplicate records or missing values were removed.

Feature Selection: Key features that are most pertinent to crime prediction are selected during this step, reducing dimensionality of data, and improving efficiency and accuracy of models. Techniques employed for feature selection may include statistical methods, correlation analyses and domain knowledge.

Training Machine Learning Models: When training machine learning models such as decision trees, random forests, support vector machines or neural networks using labeled data sets, the models learn patterns and relationships between input features and the target variable (such as crime occurrence).

Model Evaluation: Trained models are assessed against various evaluation metrics such as accuracy, precision, recall and F1-score. Cross-validation techniques may also be utilized to gauge their performance and ensure their generalizability.

Prediction: Once trained and evaluated, models can be used to make predictions on unexplored data that helps identify high-risk areas or potential crime incidents. This could include hotspots that need further investigation as well as high-risk areas like hotspots.

Research Questions

The following questions were investigated in this study.

1. Which machine learning algorithm is the strongest at predicting crime rates?
2. Which variables tend to influence crime rates the most

Dataset

We have used the following datasets for performing data analysis and building machine learning models.

1. Year-over-year percentage change, key indicators for preliminary quarterly data, adult criminal court, and youth court
2. Postsecondary enrolments, by registration status, institution type, status of student in Canada and gender
3. Non-resident visitors entering Canada, by country of residence.
4. Census Profile

All these datasets have been taken from the Open Government portal of Canada. The portal hosts a wide range of government datasets from various departments and agencies. These datasets can be freely accessed, downloaded, and used by the public, researchers, businesses, and developers to analyze, create applications, and gain insights from the data. The datasets used in this study contains information licensed under the Open Government License – Canada. The variables of interests in these datasets were Province, Year, Crimes, Population, Registrants, Visitors, Births, Deaths, Immigrants.

Results

Data Cleaning

We began by cleaning and combining the datasets. The relevant code can be found in our accompanying jupyter notebook, here are the substantial changes we made:

- Removing all data outside of 2017-2021
- Combining Northwest Territories, Nunavut, and Yukon in 'Territories' as some datasets did this
- Normalizing the numerical values so that they reflect the population (by dividing them by the population)

The head of our final dataset can be seen below:

Province	Year	normalized Crimes	normalized Registrants	normalized Visitors	normalized Births	normalized Deaths	normalized Immigrants	normalized Emigrants	normalized Returning Emigrants	normalized Temporary Migration	normalized Net Interprovincial Migration	normalized Net non-permanent residents
AB	2017	0.236092	0.045562	0.520263	0.012443	0.006131	0.009088	0.001542	0.001174	0.000533	-0.000763	-0.00031
AB	2018	0.241617	0.046412	0.55258	0.012118	0.005964	0.009434	0.001522	0.001129	0.000533	-0.000471	0.000701
AB	2019	0.244361	0.046483	0.453584	0.011578	0.006186	0.008089	0.001068	0.000878	0.000476	-0.000542	0.000066
AB	2020	0.211207	0.046027	0.085557	0.011116	0.006846	0.005429	0.001128	0.000985	0.000354	-0.002139	-0.000937
BC	2017	0.169127	0.057886	1.767273	0.008939	0.007692	0.008457	0.002218	0.001223	0.001081	0.002822	0.005911

EDA

Upon obtaining our combined dataset, we set out to perform an initial exploratory data analysis to gain a keen sense of our data. We first compared the crime rates for each province. Keep in mind these are normalized crime rates which reflect the population. We obtained the graph below:

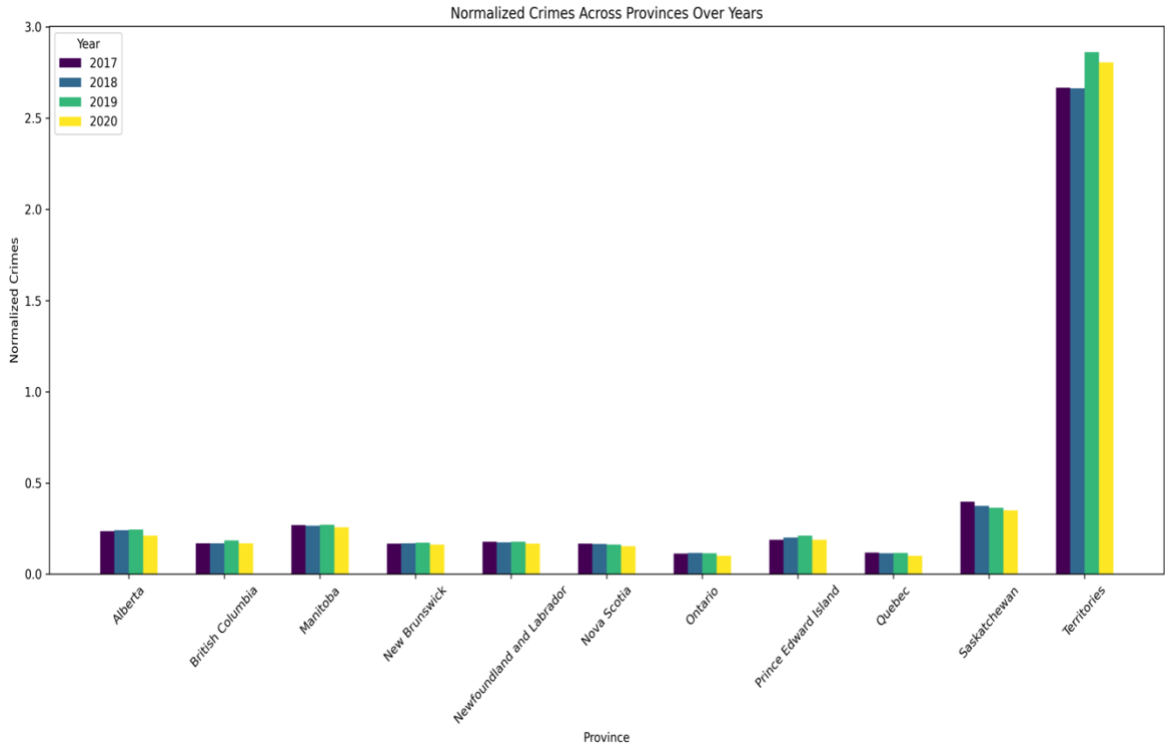

Figure 1: Crime rates by year for each Province.

We can see from the above graph that the territories have the highest 'per-capita' crime rate of any province. We expect this to be a strong contributing variable in our study. We then aimed to look at which variables are most strongly correlated with the crime rate. This can be seen in our heatmap below:

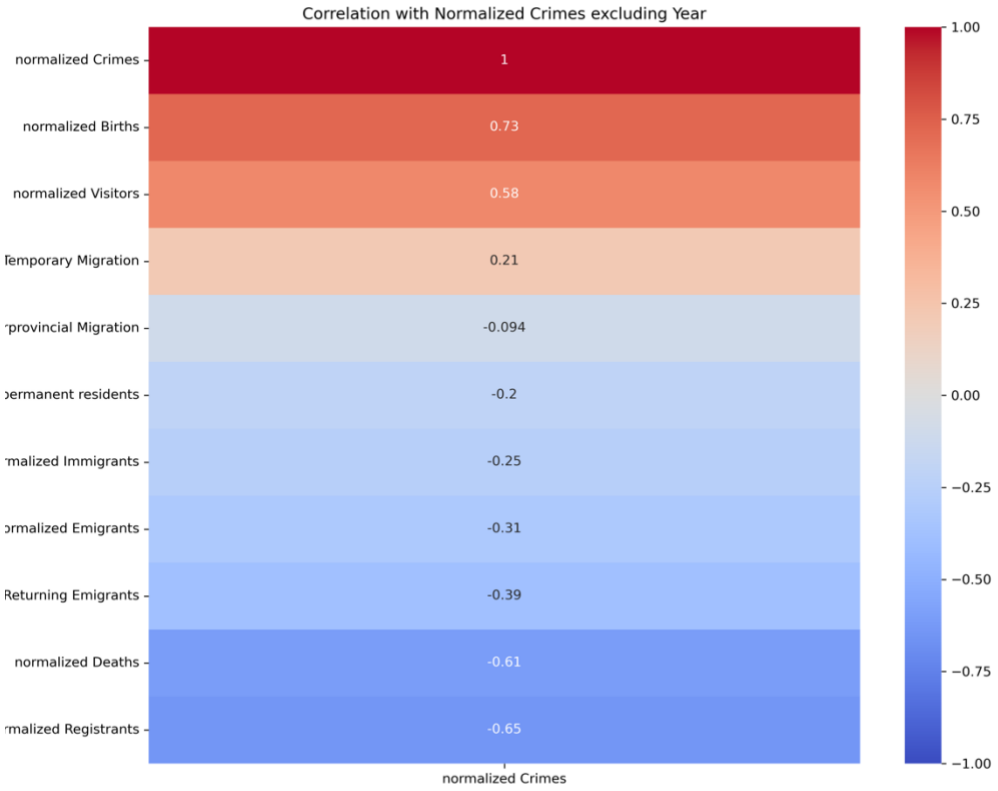

Figure 2: Correlations between crime rates and other variables.

We see that the two strongest correlations come from birth and death in opposite directions. I suspect these variables may exhibit strong multicollinearity, so we will check this later.

Feature Selection

The next step was figuring out which features would be relevant to use in our machine learning model. We encoded the province column and split the data into train and test. We then ranked each variable on feature importance in a decision tree. This algorithm measures feature importance by examining which variables decrease impurity in a tree the most. We gained the below table (shortened):

Feature	Importance
normalized Deaths	9.93E-01
normalized Births	4.33E-03
Year	1.03E-03
normalized Registrants	9.85E-04
normalized Net Interprovincial Migration	9.01E-04

normalized Emigrants	9.77E-05
normalized Net Temporary Migration	7.72E-05

From this, we suspected that the above variables would be significant. To ensure we were selecting the correct features, we then used recursive feature elimination, which can be seen from the plot below:

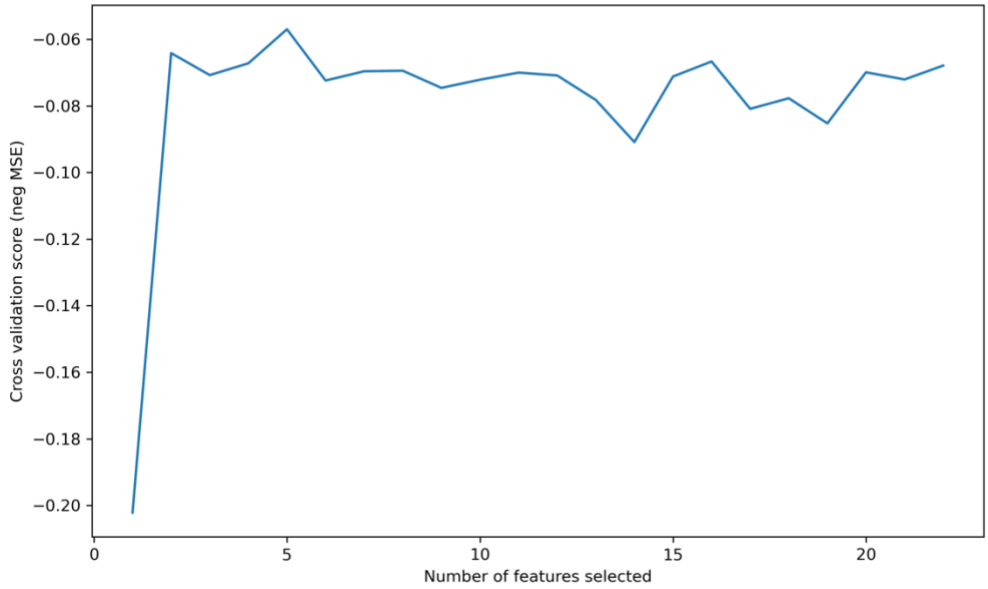

Figure 3: Recursive feature elimination.

The optimal number of features is 5, and the features deemed relevant were normalized (post-secondary) registrants, normalized Visitors, normalized Births, normalized Deaths, Province_Territories. As we previously suspected, it was significant to see if the province is a territory. Finally, we checked for multicollinearity, which can be seen from the heatmap below:

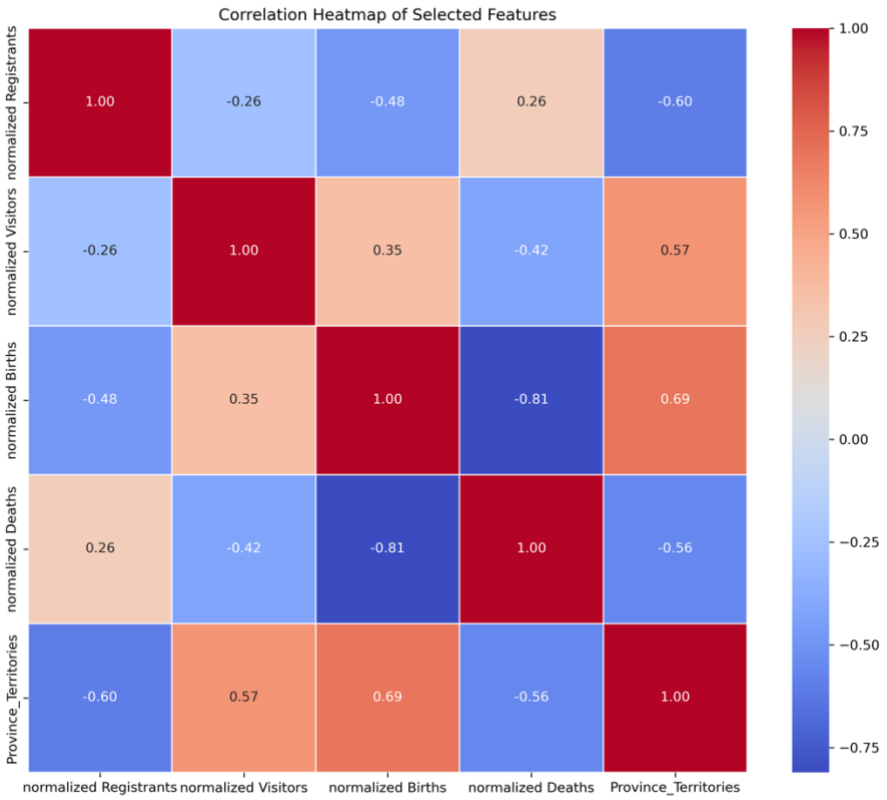

Figure 4: Correlation heatmap of features.

We can see from this heatmap that the two most correlated features were births and deaths. Because of this, we dropped the deaths column, as births showed higher feature importance in our previous analysis.

Regression Tree Model

Regression trees are a type of decision tree used for predicting continuous outcomes. They work by recursively splitting the data into subsets based on feature values, aiming to minimize the variance of the target variable within each subset. At each node, the tree determines the best feature to split on by considering how much each potential split would reduce the prediction error. Once the data is segmented, the prediction for a given leaf node is typically the average of the target values for the observations within that node. One crucial hyperparameter in regression trees is the tree depth, which dictates how many times the data can be split, or equivalently, the maximum number of decisions the tree can make. A deeper tree will capture more intricate patterns in the data, but it also runs the risk of overfitting to the training data, making its predictions less generalizable to unseen data. On the other hand, a shallow tree might be more interpretable and less prone to overfitting but may not capture essential patterns in the data. Balancing tree depth is crucial for achieving optimal model performance.

The first model we tried was a regression tree model. When we first ran this model, we got an R^2 value of 0.9998 and an MSE of 0.000147. This seemed suspiciously high, and so we decided to check the model for

overfitting.

We first examined the effect of tree depth. The tree may have been too deep and need pruning, so we saw how the MSE changed with depth, as can be seen below:

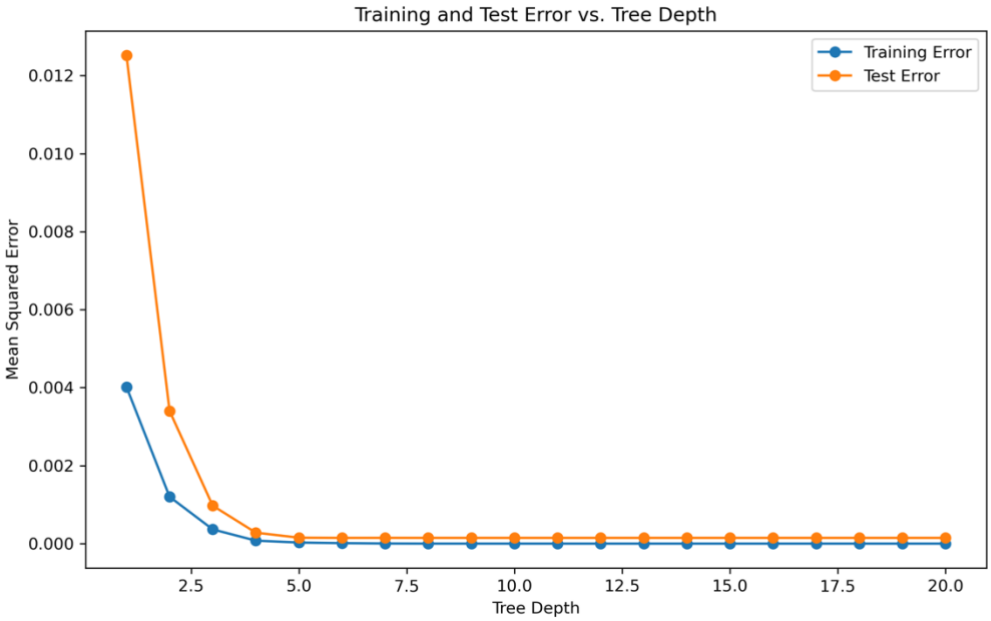

Figure 5: Mean squared error training and test error vs tree depth

We hit a plateau after a depth of around 5, with no increase after this and almost zero error. This leads us to believe overfitting has taken place.

Lasso Regression Model

We then decided to build a lasso regression model to see if we would get better results. Lasso regression, an abbreviation for Least Absolute Shrinkage and Selection Operator, is a type of linear regression that incorporates regularization to enhance prediction accuracy and interpretability. Unlike traditional linear regression, Lasso regression adds a penalty to the absolute values of the regression coefficients. The strength of this penalty is controlled by the hyperparameter, alpha. When alpha is set to zero, Lasso regression is equivalent to a standard linear regression. However, as the value of alpha increases, the imposed penalty becomes stronger, pushing some of the feature coefficients towards zero and effectively excluding them from the model. This property makes Lasso particularly useful for feature selection, especially in scenarios where there are many collinear features. By fine-tuning the alpha value, one can achieve a balance between model complexity (number of features used) and model accuracy.

This first we did was tune the alpha hyperparameter. We did this with a validation curve showing MSE for different alpha values. The curve can be seen below:

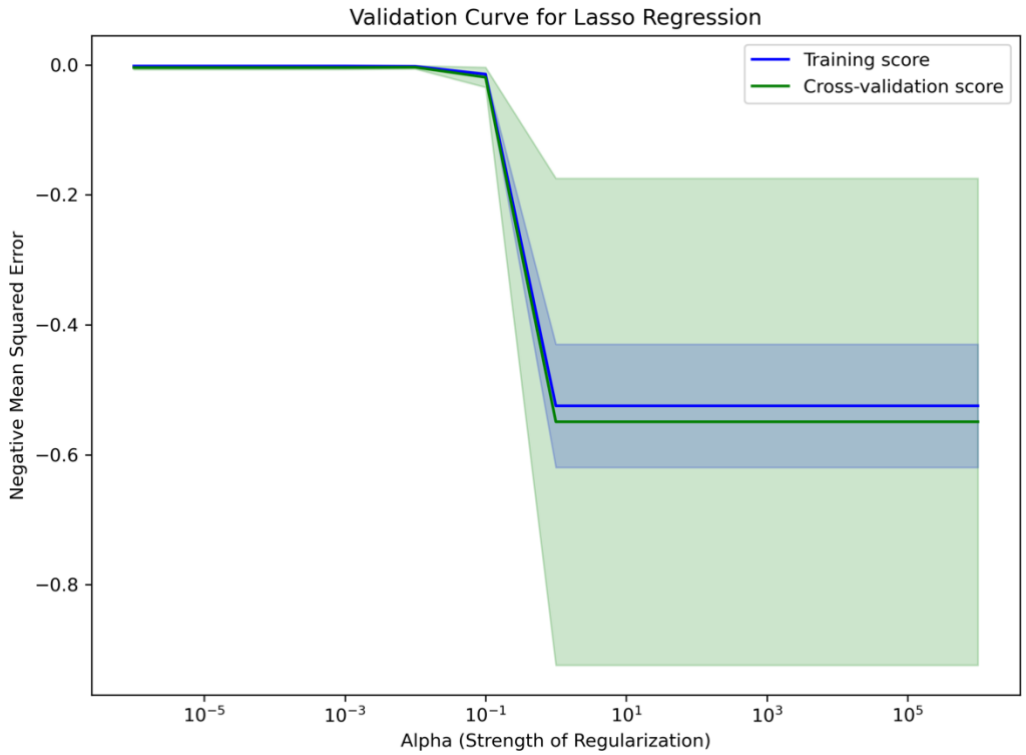

Figure 6: Mean square error vs alpha values

From this we can see that the optimal value of alpha is 0.01. We ran the model with these values and found a mean squared error of 0.0061 and an R^2 of 0.9899. This did seem high, and so we again explored the issue of overfitting. We compared the training and testing MSE and saw that the training MSE was over 164 times lower than the test MSE. This indicates that the model is overfitted to the training data, as it performs so much better on it.

Linear Regression Model

Having had issues with more complex models, we decided to explore a more simplified model in linear regression. We obtained a model with 99% accuracy. However, we saw similar performance in both the train and test dataset, and thus decided this model was optimal.

Overall, with clearly overfitted data, this dataset may have not been the right choice, as many of the variables showed slight collinearity. A future study should choose less related variables.

Conclusion and Future Work

Overall, though many of our models were overfitted due to our data, we still were able to obtain a clearer picture of crime rate cases and predictions. In terms of our guiding questions, it is unclear which ML algorithm worked the best. We achieved the highest R^2 with our regression tree model, however this appeared to be overfitted. Further testing would be required to properly address this question.

For our second guiding question, we saw that births and deaths were the most significant contributing

factors to crime rates. Births were positively correlated while deaths were negatively. This indicates that younger populations (those with higher birth than death rates) are more susceptible to higher crime rates.

Overall, we have seen that crime is a complex and multifaceted problem. Many other non-macro factors should be considered. A future study should include less correlated variables.

References

[1] "Crime | Definition, History, Examples, Types, Classification, & Facts | Britannica," Jul. 21, 2023. https://www.britannica.com/topic/crime-law (accessed Aug. 13, 2023).

[2] Causes of Crime - Explaining Crime, Physical Abnormalities, Psychological Disorders, Social And Economic Factors, Broken Windows, Income And Education. https://law.jrank.org/pages/12004/Causes-Crime.html. Accessed 13 Aug. 2023.

[3] Sloan, Cynthia Rudin, MIT. "Predictive Policing: Using Machine Learning to Detect Patterns of Crime." Wired, 22 Aug. 2013. www.wired.com, https://www.wired.com/insights/2013/08/predictive-policing-using-machine-learning-to-detect-patterns-of-crime/

[4] Shah, N., Bhagat, N., & Shah, M. (2021). Crime forecasting: A machine learning and computer vision approach to crime prediction and prevention. Visual Computing for Industry, Biomedicine, and Art, 4(1), 9. https://doi.org/10.1186/s42492-021-00075-z

[5] Mandalapu, V., Elluri, L., Vyas, P., & Roy, N. (2023). Crime prediction using machine learning and deep learning: A systematic review and future directions. IEEE Access, 11, 60153–60170. https://doi.org/10.1109/ACCESS.2023.3286344

[6] Education and crime—Criminal justice—Iresearchnet. (n.d.). Criminal Justice. Retrieved August 14, 2023, from https://criminal-justice.iresearchnet.com/correlates-of-crime/education-and-crime/

[7] S. E. Brown, F.-A. Esbensen, and G. Geis, Criminology: explaining crime and its context, 7th ed. New Providence, NJ: LexisNexis/Anderson Pub, 2010. P-12

[8] PA S. An Overview on MobileNet: An Efficient Mobile Vision CNN [Internet]. Medium. 2020. Available from: https://medium.com/@godeep48/an-overview-on-mobilenet-an-efficient-mobile-vision-cnn-f301141db94d

[9] Statistics Canada. Table 35-10-0177-01 Incident-based crime statistics, by detailed violations, Canada, provinces, territories, Census Metropolitan Areas and Canadian Forces Military Police https://doi.org/10.25318/3510017701-eng

[10] Statistics Canada. Table 37-10-0018-01 Postsecondary enrolments, by registration status, institution type, status of student in Canada and gender https://doi.org/10.25318/3710001801-eng

[11] Statistics Canada. Table 24-10-0050-01 Non-resident visitors entering Canada, by country of residence https://doi.org/10.25318/2410005001-eng

[12] Statistics Canada. Census Profile https://www12.statcan.gc.ca/census-recensement/2021/dp-pd/prof/index.cfm?Lang=E