Using Machine Learning to Identify Top Antecedents Affecting Crime in US Communities

Kamil Samara[1]

[1] University of Wisconsin-Parkside, Kenosha, USA

Abstract. One of the main concerns for countries has been always crime activities. In recent years, with the development of data collection and analysis techniques, a massive number of data-related studies have been performed to analyze crime data. Studying indirect features is important yet challenging task. In this work we are using machine learning (ML) techniques to try to identify the top variables affecting crime rates in different US communities. The data used in this work was collected from the Bureau of the Census and Bureau of Justice Statistics. Out of the 125 variables collected in this data we will try to identify the top factors that correlate with higher crime rates either in a positive or a negative way. The analysis in this paper was done using the Lasso Regression technique provided in the Python library Scikit-learn

Keywords: Machine Learning, Lasso Regression, Crime.

1 Introduction

Crime as a socioeconomic complication has shown multifaceted associations with socioeconomic, and environmental aspects. Trying to recognize patterns and connections between crime and these factors is vital to understand the root causes of criminal activities. By detecting the source causes, legislators can instrument solutions for those source causes, eventually avoiding most crime sources [1].

In the age of information technology, crime statistics are recorded in data bases for studying and analysis. Manual analysis in impractical due to the vast size of stored data. The suitable solution here is to use data science and machine learning techniques to analysis the data. Using the descriptive and predictive powers of those solutions officials will be able to minimize crime.

The descriptive and predictive powers of machine learning techniques can give longstanding crime prevention solutions. This predictive analysis could be done on two levels. First, predicting when and where the crimes will happen. But this type of prediction is hard to implement because predictions are highly sensitive to complex disseminations of crimes in time and space. Second, focusing predictions on identifying the correlations of crimes with socioeconomic, and environmental aspects [2].

Lately machine learning methods have grew in popularity. Among the most popular approaches is Bayesian model, random forest, K-Nearest Neighbors (KNN), neural network, and support vector machine (SVM) [3].

As a step toward crime prediction using machine learning techniques, the proposed work in this paper uses the Lasso Regression technique to predict the top socioeconomic, and environmental factors related to crime rates in US cities. The study was preformed using data collected form the Bureau of the Census and Bureau of Justice Statistics.

The remaining of the paper is organized as follows: Section 2 is related work, Section 3 presents the work done is this study and Section 4 concludes the work.

2 Related Work

Crime is a global problem, which motivated my researchers to apply machine learning techniques to preform predictive analytics in effort to detect crime factors. The performed studies range in complexity depending on the volume of the datasets used in the study and the number of variables collected.

A common crime prediction analysis is the focus on temporal data. A common reason behind this emphasis is crime data sets contain data collected over many years. An example of such analysis is the work of Linning in [4]. Linning have studied the variation of crime throughout the year to predict a pattern over seasons. The main observation was crime peaks in the hot summer seasons as compared to cold winter seasons.

In a similar study [5], authors have examined the crime data of two major US cities and compared the statistical investigation of the crimes in these cities. The main goal of the study was to use agent-based crime environment simulation to identify crime hotspots.

In [6], Nguyen and his team used data from the Portland Police Bureau (PPB) augmented with census data from other public sources to predict crime category in the city of Portland using Support Vector Machine (SVM), Random Forest, Gradient Boosting Machines, and Neural Networks.

A unique approach to classify crime from crime reports into many categories using textual analysis and classification was done in [7]. The authors used five classification methods and concluded that Support Vector Machines (SVM) performed better than the other methods.

The researches in [8], used data extracted from the neighborhoodscout.com from the University of California-Irvine, in the state of Mississippi to predict crime patterns using additive linear regression.

Graph based techniques were used in [9] to mine datasets for correlation. The objective was to identify top and bottom correlative crime patterns. In the authors final remarks, they conclude that it successfully discovers both positive and negative correlative relations among crime events and spatial factors and is computationally effective.

3 Proposed Work

The scope of this work is to analyze crime data in effort to recognize top factors affecting crime plausibility. The features considered in this work are socio-economic factors like race, number of people in living in the same house, and mean house income.

Python was the programming language of choice in this work. To perform the regression part, the Lasso model in the Scikit-learn library was used. Scikit-learn is a free software machine learning library for the Python programming language.

3.1 Dataset

The dataset used in this paper is the "Communities and Crime Unnormalized Data Set" available at the University of California Irvine (UCI) Machine Learning Repository. This data set main focus is communities in the United States and was combined from the following sources: 1995 US FBI Uniform Crime Report, 1990 United States Census, 1990 United States LEMAS (Law Enforcement Management and Administrative) Statistics Survey. In July 2009, this data set was presented to the UCI Machine Learning Repository [10].

The data set includes 2215 total examples and 125 features for different communities across all states. Features contain data blended from a diverse source of crime-related information, extending from the number of vacant households to city density and percent of people foreign born, to average household income. Also included are measures of crimes considered violent, which are murder, rape, robbery, and assault. Only features that had plausible connection to crime were included. So unrelated features were not included [10].

3.2 Lasso Regression

Lasso (Least Absolute Shrinkage and Selection Operator) regression is part of the linear regression family that utilizes reduction. Reduction is where data values are contracted near a central value, like the average. The lasso technique promotes models with less parameters. The lasso regression is more suitable for models expressing high levels of multicollinearity [11].

Lasso regression performs L1 regularization. As shown in equation 1, L1 regularization works by enforcing a penalty equivalent to the absolute value of the magnitude of coefficients. L1 regularization encourages models with few coefficients. This can be achieved by reducing some coefficients to become zero and get removed from the model. L1 regularization helps produce simpler models since larger penalties result in coefficient values almost zero. On the other hand, Ridge regression (e.g., L2 regularization) doesn't result in removal of coefficients or sparse models. This makes the L1 regularization far easier to interpret than the L2 regularization [11].

$$RSS_{LASSO}(w,b) = \sum_{\{i=1\}}^{N}(y_i - (w.x_i + b))^2 + \alpha \sum_{\{i=1\}}^{p} |w_j| \quad (1)$$

Where:
Yi: target value
w. xi + b: predicted value
α: controls amount of L1 regularization (default = 1.0).

3.2 Feature Normalization

Before applying the Lasso regression on the dataset, a MinMax scaling of the features was done. It is crucial in several machine learning techniques that all features are on the same scale (e.g. faster convergence in learning, more uniform or 'fair' influence for all weights) [12].

For each feature Xi: compute the min value XiMIN and the max value XiMAX achieved across all instances in the training set. For each feature: transform a given feature xi value to a scaled version x' using e

$$x_i^{'} = (x_i - x_i^{MIN}) \,/\, (x_i^{MAX} - x_i^{MIN}) \tag{2}$$

3.3 Alpha Value Selection

The parameter α controls amount of L1 regularization done in the Lasso linear regression. The default value of alpha is 1.00. To use the Lasso regression efficiently the appropriate value of alpha must be selected.

To decide the appropriate value of alpha, a range of alpha values were compared using the r-squared value. R-squared (R2) is a statistical measurement which captures the proportion of the change of a variable that's explained by another variable or variables in a regression fashion [12].

The alpha values used in the comparison were: [0.5, 1, 2, 3, 5, 10, 20, 50]. The results of the comparison are shown in table 1. As we can see in table 1, the highest r-squared value was achieved at an alpha value of 2.

Table 1. Effect of alpha regularization.

Alpha Value	Features kept	R-squared value
0.5	35	0.58
1	25	0.60
2	20	0.63
3	17	0.62
5	12	0.61
10	6	0.58
20	2	0.50
50	1	0.30

3.4 Results

For alpha = 2.0, 20 out of 125 features have non-zero weight. Top features (sorted by abs. magnitude are shown in table 2. Higher weights indicate higher importance and impact of the feature on crime rate. Positive weights for features mean a positive correlation between the feature value and crime rate. On the other hand, negative weights for features mean a negative correlation between the feature value and crime rate.

Table 2. Features with non-zero weight (sorted by absolute magnitude).

Feature	Weight	Description
PctKidsBornNeverMar	1488.365	kids born to never married parents
PctKids2Par	-1188.740	kids in family with two parents
HousVacant	459.538	unoccupied households
PctPersDenseHous	339.045	persons in compact housing
NumInShelters	264.932	people in homeless shelters
MalePctDivorce	259.329`	divorced males
PctWorkMom	-231.423	moms of kids under 18 in labor force
pctWInvInc	-169.676	households with investment
agePct12t29	-168.183	in age range 12-29
PctVacantBoarded	122.692	vacant housing that is boarded up
pctUrban	119.694	people living in urban areas
MedOwnCostPctIncNoMtg	104.571	median owners' cost
MedYrHousBuilt,	91.412	median year housing units built
RentQrange	86.356	renting a house
OwnOccHiQuart	73.144	owning a house
PctEmplManu	-57.530	people 16 and over who are employed in manufacturing
PctBornSameState	-49.394	people born in the same state as currently living
PctForeignBorn	23.449	people foreign born
PctLargHouseFam	20.144	family households that are large (6 or more)
PctSameCity85	5.198	people living in the same city

Although there are many interesting features to discuss from the results shown in table 2, we will focus our interest on the top two features. The top antecedent from the list of features was Kids Born to Never Married with a positive weight of 1488.365. The second antecedent was Kids in Family Housing with Two Parents with a negative weight of -1188.740. These two top antecedents indicate the importance of a stable family (two parents are present) for raising kids whom less likely to commit crime in the future.

4 Conclusion

In this study, we employed machine learning through the use of Lasso linear regression in effort to predict socio-economic antecedents that affect crime rate in US cities. The regression model was implemented on a data set that was sourced from the 1995 US FBI Uniform Crime Report, 1990 United States Census, 1990 United States LEMAS (Law Enforcement Management and Administrative) Statistics Survey.

The regression results pointed out that the topmost influential factors affecting crime in US cities are related to stable families. Kids born into families with two parents are less likely to commit crime.

These findings should help policy makers to create strategies and dedicate fundings to help minimize crime.

References

1. Melossi, D.: Controlling Crime, Controlling Society: Thinking about Crime in Europe and America. 1st edn. Polity, 2008.
2. H.M.M.I.S.B Herath and D.M.R Dinalankara, "Envestigator:AI-based Crime Analysis and prediction platform" Proceedings of Peradeniya University International Research sessions,vol. 23,no. 508,pp. 525 2021.
3. K. C. Baumgartner, S. Ferrari, and C. G. Salfati, ``Bayesian networkmodeling of offender behavior for criminal profiling," in Proc. 44th IEEE Conf. Decis. Control, Eur. Control Conf. (CDC-ECC), Dec. 2005, pp. 2702-2709.
4. S. J. Linning, M. A. Andresen, and P. J. Brantingham, "Crime seasonality: Ex-amining the temporal fluctuations of property crime in cities with varying cli-mates," International journal of offender therapy and comparative criminology, Vol. 61, no. 16, pp. 1866–1891, 2017.
5. T. Almanie, R. Mirza, and E. Lor, "Crime prediction based on crime types and using spatial and temporal criminal hotspots," arXiv preprint arXiv:1508.02050, 2015.
6. T. T. Nguyen, A. Hatua, and A. H. Sung, "Building a learning machine classifier with inadequate data for crime prediction," Journal of Advances in Information Technology Vol, vol. 8, no. 2, 2017.
7. D. Ghosh, S. Chun, B. Shafiq, and N. R. Adam, "Big data-based smart city plat-form: Real-time crime analysis," in Proceedings of the 17th International DigitalGovernment Research Conference on Digital Government Research. ACM, 2016,pp. 58–66.
8. McClendon, Lawrence, and Natarajan Meghanathan. "Using machine learning algorithms to analyze crime data." Machine Learning and Applications: An International Journal (MLAIJ) 2.1 (2015): 1-12.
9. P. Phillips and I. Lee, "Mining top-k and bottom-k correlative crime patterns through graph representations," 2009 IEEE International Conference on Intelligence and Security Informatics, 2009, pp. 25-30, doi: 10.1109/ISI.2009.5137266.
10. A. Asuncion and D.J. Newman, “UCI Machine Learning Repository,” School of Information and Computer Science, University of California, Irvine, CA, 2007, URL: archive.ics.uci.edu/ml/datasets/Communities+and+Crime.
11. Kumar, D. (2021, December 26). A Complete understanding of LASSO Regression. The Great Learning. https://www.mygreatlearning.com/blog/understanding-of-lasso-

regression/#:~:text=Lasso%20regression%20is%20a%20regularization,i.e.%20models%20with%20fewer%20parameters).
12. Kuhn, M., Johnson, K.: Feature Engineering and Selection: A Practical Approach for Predictive Models. 1st edn. CRC Press, New York (2019).