IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁂ Vol. 11, Issue 3, March 2022

DOI: 10.17148/IJARCCE.2022.11342

Prediction of dependency of crime rate on level of migrant population using Machine Learning

Soumili Mondal[1], Utsab Ghosh[2], Sulekha Das[3], Avijit Kumar Chaudhuri[4], Moumita Chakraborty[5]

[1,2]Student, CSE department, Techno Engineering College Banipur

[3,4,5]Assistant Professor, CSE department, Techno Engineering College Banipur

Abstract: Common perception is that America as a global leader in technology, employment opportunities, and living standards. But the shadow is darker just beneath the light. United States is also the global leader when it comes to incarcerations. One in every three US adults have criminal record.

Statistical numbers clearly suggest a racial imbalance in terms of arrests and victim counts. While colored people only make up 37% of US population, they account for 67% of prison population. Among various factors behind this uncomfortable truth, major ones are poverty, lack of education, unemployment, improper family planning etc. Apart from the mentioned reason the relationship between race and crime has been the subject of controversy in the developed nations like United States. Correlation of crime with racial disparity is prominent to the extent to have its influence on social movements and even legislation.

As for the purpose of prediction of expected criminal activity in a particular locality along with total population, proportions of colored people also make significant contribution.

The proposed model is tested on the "Communities and Crime Data Set" from the UCI Machine Learning Repository[3]

1. INTRODUCTION

Regression is a statistical method used in finance, and other disciplines that attempts to determine the relationship between one dependent variable (usually denoted by Y) and a series of other variables (independent variables).There are various types of regression analysis available in statistics like simple linear regression, multiple linear regressions, logistic regression, ordinal regression, etc. Linear and logistic regression methods are most common. "Linear regression attempts to draw a line that comes closest to the data by finding the slope and intercept that define the line and minimize regression errors"[6]. If two or more such variables have a linear relationship with the dependent variable or target variable, the regression is called a multiple linear regression.

The main aim in this paper is to understand the relationships between the selected controlled variables and crimes and also to predict violent crime using multiple linear regression model. Data that has been used in this research were taken from the 1990 US Census, law enforcement data from the 1990 US LEMAS survey, and crime data from the 1995 FBI UCR.

"The connection between immigration and crime is one of the most debatable topics in contemporary society. These discussions are not new, as debates on the issue date back more than 100 years." [1]

"There are important reasons to believe that immigrants should be involved in crime to a greater degree than native-born. For example, immigrants might suffer from acculturation and assimilation problems unlike the natives, and immigrants tend to settle in disorganized neighbourhoods characterized by structural characteristics often associated with crime, such as widespread poverty, ethnic heterogeneity, and a preponderance of young males."[1]

"There is no simple link between immigration and crime. According to major published research and studies larger immigrant concentrations in an area have least association with violent crime and, slight effects on property crime. But this is also true of disadvantaged native groups. There is also a case for ensuring that immigrants can legally obtain work in the receiving country, since the evidence shows that such legalization programs tend to reduce criminal activity among the targeted group." [2]

We have made a crime prediction model with the help of multiple linear regression in Python programming language. In linear regression we can use simple equations without the help of library function. We have taken the help of sklearn library in Python to set up the model because the number of independent variables are greater and simple linear equations were unable to do the prediction efficiently, since it is not possible to find a simple one dimensional link between immigration, emigration and crime rates.

DOI: 10.17148/IJARCCE.2022.11342

2. LITERATURE REVIEW

Author Ashwin Bahulkar has implemented a network algorithm [4] to set a relation between immigrant people and crime, in this paper it is important for us to know the community membership of high degree nodes" . [4]
There are many research that have been conducted using multiple regression. FECHETE, Flavia; NEDELCU, Anisor (2014) [7] analysed the economic performance of an organization using multiple regression since "the objectives of the organization can be measured as effectiveness or as efficiency." [7]
Rainer Bohme (2005) [8] conducted a research that proposes "MLR models as a method for quantitative evaluation of the accuracy in stegnalysis with respect to various moderating factors, such as parameter choice of the detector and properties of the carrier object."
RG Newton and DJ Spurrell [9] developed a MLR model for data analysis by using limited number of regression equations.
David J Lowe, Margaret W Emsley, Anthony Harding (2006) [10] used MLR techniques to predict construction costs of buildings based on collected data set in the UK.
Tiberiu Catalina, Vlad Iordache and Bogdan Caracaleanu (2013) [11] used MLR model to predict heating energy demand which is used during designing a new building, based on building's heat consumption.
MLR have been used by AT Bourgoyne, FS Young [12] to gain optimal drilling models and pore pressure detection. Old research in this topic were based upon limited data and were not as accurate as this newly developed model.
A research conducted in 2016 by S Alireza Eslamian, S Samuel Li, Fariborz Haghighat [13] used MLR for predicting and understanding urban water use.
Lilian Pun, Pengxiang Zhao, Xintao Liu (2019) [14] used MLR approach for estimating traffic flow by integrating five topological measures and road length.
Bok-Hee Jung, SoonGohn Kim (2014) [15] constructed a study model to evaluate the factors so as to understand festival satisfaction using MLR.

We have conducted a research to predict violent crime in various states in America, based on the 1990 US LEAMAS survey data. We have chosen eight factors from all the given variables in the data, that might influence the number (or rates) of violent crimes. The chosen categories are—

"percentage of immigrants who immigrated within last 3 years",
"percentage of immigrants who immigrated within last 5 years",
"percentage of immigrants who immigrated within last 8 years",
"percentage of immigrants who immigrated within last 10 years",
"percent of population who have immigrated within the last 3 years",
"percent of population who have immigrated within the last 5 years",
"percent of population who have immigrated within the last 8 years",
"percent of population who have immigrated within the last 10 years"

Immigration and its relationship with crime has been a debate in Western countries. However, one aspect of the debate is that there are accusations that higher levels of immigration might lead to higher amount of crime. Based on studies of many countries, we have the evidence that indicates that there is no simple link between crime and immigration. If the status of immigrants is legalized then it will have desired effects on crime rates, i.e. lesser crimes. Importantly, the evidence points to the substantial differences in the impact on property crime, depending on the labour market opportunities of immigrant groups. If immigrants don't get enough job opportunities or poor labour opportunities then they are more likely to get involved with property related crimes. Also if the rate of crime in a place is lower than the amount of immigrants who might want to go there should be higher. If a place has high crime rates then a big percent of the population might want to go to another place.

3. METHODOLOGY

3.1. Data

Data Set: Multivariate
Attributes: Real
Associated Tasks: Regression
Number of Instances: 1994
Number of Attributes: 128
Area: Social

Date Donated: 2009-07-13
Source: Creator: Michael Redmond (Redmond 'at' lasalle.edu);
-- Donor: Michael Redmond (Redmond 'at' lasalle.edu); Computer Science; La Salle University; Philadelphia, PA, 19141, USA -- Date: July 2009

Detailed information on the data set:
“Many variables are included so that algorithms that select or learn weights for attributes could be tested. However, unrelated attributes were not included; attributes were picked only if there were any direct or indirect connection to crime (N=122), plus the attribute to be predicted (Per Capita Violent Crimes).
Data is described based on original values. All numeric data was normalized into the decimal range 0.00-1.00. Attributes retain their distribution and skew. E.g. an attribute described as 'mean people per household' is actually the normalized (0-1) version of that value. The normalization preserves rough ratios of values within an attribute. However, the normalization does not preserve relationships between values between attributes.
Many communities are missing LEMAS data.”

Table.1. Fields we have chosen for our prediction-

Data	Attribute	Attribute information	Mean	Standard deviation
X1	'PctImmigRecent'	“percentage of _immigrants_ who immigrated within last 3 years”	0.32	0.22
X2	'PctImmigRec5'	“percentage of _immigrants_ who immigrated within last 5 years”	0.36	0.21
X3	'PctImmigRec8'	“percentage of _immigrants_ who immigrated within last 8 years”	0.40	0.20
X4	'PctImmigRec10'	“percentage of _immigrants_ who immigrated within last 10 years”	0.43	0.19
X5	'PctRecentImmig'	“percent of _population_ who have immigrated within the last 3 years”	0.18	0.24
X6	'PctRecImmig5'	“percent of _population_ who have immigrated within the last 5 years”	0.18	0.24
X7	'PctRecImmig8'	“ percent of _population_ who have immigrated within the last 8 years”	0.18	0.24
X8	'PctRecImmig10'	“percent of _population_ who have immigrated within the last 10 years”	0.18	0.23
Y	ViolentCrimesPerPop	“total number of violent crimes per 100K population (numeric - decimal) GOAL attribute (to be predicted)”	0.24	0.23

X vs Y scatter plot:

Chart.1. (Crime vs 'PctImmigRecent')

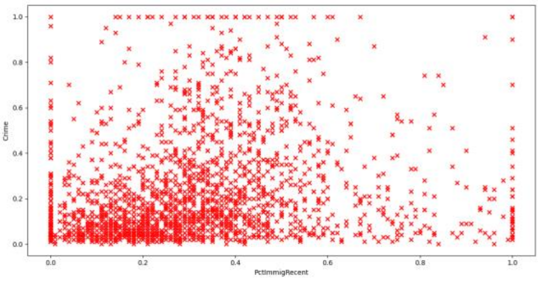
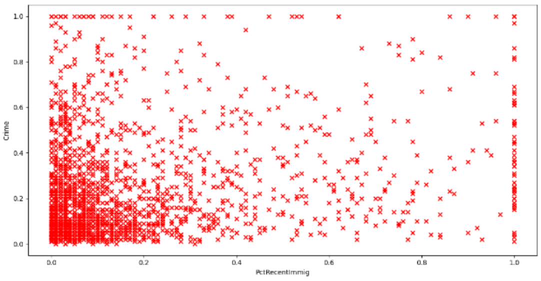


Chart.2. (Crime vs 'PctRecentImmig')



3.2. Research method

Multiple linear regression (MLR)or simply multiple regression is a method of statistics in regression that used to analyze the relationship between single response variable (dependent variable) with two or more explanatory/ controlled variables (independent variables). It's the "extension of ordinary least- squares" regression in essence.
Formula for calculation of multiple regression-
$Y_i = A + B_1X_{i1}+ B_2X_{i2}+ B_3X_{i3}+ \dots + B_pX_{ip}+ e$; for i=n number of observations,
Y_i= dependent variable,
A = y-intercept/ constant term,
e = residuals/error term,
B_p= slope coefficient for each independent variable,
X_i = independent variable.

MLR model is based on certain assumptions-

- Linear relationship exist between the dependent and independent variables.
- The independent variables are not very much related with each other.
- Y_i observations are selected randomly and independently from the dataset.
- Residuals should be normally distributed.

If we plot the MLR model then we get a straight line (linear relationship) that best approximates all the individual data points.
The least-squares estimates— B_1, B_2, B_3 …—are usually computed by statistical software. Since many variables are included in the regression model and for each variable we have multiple data. It's unlikely to be able to do a multiple regression by hand as MLR models are complex and the amount of data to analyze is huge. We can use specialized statistical software features/functions present within programs such as Microsoft Excel.
Still the model may not be perfectly accurate as each data point may differ from the outcome predicted by the model. The residual value is the difference between the original results and the predicted outcome and is included in the model for these small variations.
This method is used for this research because there was more number of independent variables, as opposed to simple linear regression. In this research, response variable is ViolentCrimesPerPop(Y) while "percentage of immigrants who immigrated within last 3 years"(X_1), "percentage of immigrants who immigrated within last 5 years"(X_2), "percentage of immigrants who immigrated within last 8 years"(X_3),"percentage of immigrants who immigrated within last 10 years"(X_4),"percent of population who have immigrated within the last 3 years"(X_5),"percent of population who have immigrated within the last 5 years"(X_6), "percent of population who have immigrated within the last 8 years"(X_7)and "percent of population who have immigrated within the last 10 years" (X_8) were controlled variables. There were eight general factors to build predicting violent crime rate.

Selecting suitable methods of multiple linear regression and interpreting the output:

We have used the sklearn library to set up the multiple linear regression model with the available data. Then we have compared the calculated value with the original value and calculated the accuracy value. We have also done ten-fold cross validation.

Cross Validation:

Cross-validation is a method in statistics which is used to estimate the skill of machine learning models. Cross-validation is mainly used in applied machine learning to estimate the skill of a machine learning model on certain data. It is a popular method because it is simple and it generally results in a less optimistic or less biased estimate of the model skill than other methods, like a simple train/test split. The data is first shuffled randomly and split into k groups. For each unique group we have unique test and training set and the data is evaluated for accuracy values. After we get our desired values we discard it and repeat till k number of times.

Confusion Matrix:

A confusion matrix is a table that is usually used to describe the performance of a classification model (also known as "classifier") on a set of test data for which the true values are known. The confusion matrix itself is simple. Author Robert Susmaga describes " The table assumes that there were A + B positive objects, but only A of them have been correctly recognized as positive, while B of them have been incorrectly recognized as negative. At the same time, out of C + D originally negative objects, only D have been recognized correctly as negative, while C incorrectly recognized as positive. The most popular measures based on the above table are: the recall, calculated as A/(A + B), the precision, calculated as A/(A + C), the F-measure, calculated as 2A/(2A + B + C) and the accuracy, which is calculated as (A + C)/(A + B + C + D)"[5].

Numbers of objects	**Classified as positive**	**Classified as negative**
Positive	**A**	**B**
Negative	**C**	**D**

We get the output in terms of:

"True positive(TP), True negative(TN), False positive(FP), False negative(FN)."

"TP indicates the number of positive examples classified accurately.

TN stands for True Negative which shows the number of negative examples classified accurately. FP is the number of actual negative examples classified as positive. FN is the number of actual positive examples classified as negative."

The accuracy, sensitivity, specificity, precision, recall, and f1-score of a model (through a confusion matrix) is calculated using the given formulas.

$$\text{Accuracy}= \frac{(TP+TN)}{(TP+TN+FP+FN)}$$

$$\text{Sensitivity}= \frac{TP}{(TP+TN)}$$

$$\text{Specificity}= \frac{TN}{(TN+FP)}$$

$$\text{Precision}= \frac{TP}{(TP+FP)}$$

$$\text{Recall}= \frac{TP}{(TP+FN)}$$

$$\text{F1\_Score} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}$$

Equations and values for our problem:

The multiple linear regression model can be represented with the following equation:

$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 + b_6X_6 + b_7X_7$

Where,

Y = Response variable: ViolentCrimesPerPop

X_1 =percentage of immigrants who immigrated within last 3 years,

X_2 =percentage of immigrants who immigrated within last 5 years,

X_3 = percentage of immigrants who immigrated within last 8 years,

X_4 = percentage of immigrants who immigrated within last 10 years,

X_5 =percent of population who have immigrated within the last 3 years,

X_6 =percent of population who have immigrated within the last 5 years,

X_7 =percent of population who have immigrated within the last 8 years,

X_8 = percent of population who have immigrated within the last 10 years were controlled variables,

a= constant (intercept)

Therefore the formulas are-

$$a=\frac{\sum y*\sum x^2-\sum x*\sum (x*y)}{n\sum x^2-(\sum x)^2}$$

$$b_i=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

Algorithm for the model:

➢ Step 1- Full data set has been read and split into 2 parts initially- test data and training data

➢ Step 2- Linear regression model from sklearn library is used to fit the x values and y from our data

➢ Step 3- Slope coefficients and the intercept have been calculated with the fitted data

➢ Step 4- Predict function has been used on the test data to calculate predicted value

➢ Step 5- Then predicted value and the original value have been compared to find out the number of the most accurate calculated values (those with deviations <=0.2)

➢ Step 6- After this we have done ten-fold cross validation (10% test data) on our data set, 10 times. The 10% data is selected randomly each time the loop runs (i.e. 10 times)

➢ Step 7- Accuracy and confusion matrix have been calculated for each loop in the same way as earlier.

4. RESULTS AND DISCUSSIONS:

Accuracy- 73.93483709273183 % over 399 counts of test data and 1595 training data.
Confusion matrix-

	True(T)	False(F)
Positive (P)	295	16
Negative (N)	78	10

Then in 10 fold cross validation 200 data is selected at random for 10 times. Then accuracy is calculated including the confusion matrix values.
Sample output of 10 fold cross validation-
(TP-True positive, TN- True Negative, FP- False positive, FN- False negative

As we can see from the below table this model gives us accuracy in the range of 70-80% approximately.

Table for calculated values of 10 fold cross validation

Test cases	Accuracy in %	TP	TN	FP	FN	Sensitivity	Specificity	Precision	F1 score
TC 1	76.5	153	33	8	6	0.82	0.80	0.95	0.96
TC 2	73.0	146	35	13	6	0.80	0.72	0.91	0.93
TC 3	73.0	146	37	11	6	0.79	0.77	0.92	0.94
TC 4	76.0	152	34	8	6	0.81	0.80	0.95	0.96
TC 5	73.5	147	38	9	6	0.79	0.80	0.94	0.95
TC 6	80.0	160	28	6	6	0.85	0.82	0.96	0.93
TC 7	68.5	137	49	8	6	0.74	0.86	0.94	0.95
TC 8	78.5	157	32	8	3	0.83	0.8	0.95	0.96
TC 9	77.0	154	28	14	4	0.84	0.67	0.92	0.94
TC 10	75.0	150	29	15	6	0.83	0.66	0.90	0.93

CONCLUSIONS

This paper uses multiple linear regressions (MLR) to understand the relationship between crime and immigration and accurately predict crime values. There is no simple relationship between crime and immigration. Larger immigrant concentrations in any area have no association with violent crimes and, pretty weak effects on property related crime. However, if immigrant groups face poor labor market opportunities then they might be more likely to commit property crime. It is important that the status of immigrants is legalized then it will have desired effects on crime rates, i.e. lesser crimes.

REFERENCES

1. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.198.6457&rep=rep1&type=pdf
2. https://wol.iza.org/articles/crime-and-immigration/long
3. https://archive.ics.uci.edu/ml/datasets/communities+and+crime
4. Community Detection with Edge Augmentation in Criminal Networks Ashwin Bahulkar∗ , Boleslaw K. Szymanski∗ , N. Orkun Baycik‡ and Thomas C. Sharkey† ∗Computer Science
5. Confusion Matrix Visualization Robert Susmaga Poznan University of Technology, Piotrowo 3a, 60-965 Poznan, Poland
6. https://sonalsart.com/what-is-the-key-difference-between-stepwise-and-hierarchical-multiple-regression/
7. https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=paper+using+multiple+regression&btnG=#d=gs_qabs&u=%23p%3DG_EKyXEy0n4J
8. https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=paper+using+multiple+regression&btnG=#d=gs_qabs&u=%23p%3DXmATw-xgylYJ
9. https://scholar.google.com/scholar?start=10&q=paper+using+multiple+regression&hl=en&as_sdt=0,5#d=gs_qabs&u=%23p%3DilB7ZXw8Q8QJ
10. https://scholar.google.com/scholar?start=10&q=paper+using+multiple+regression&hl=en&as_sdt=0,5#d=gs_qabs&u=%23p%3DrdbctIqvNjkJ
11. https://scholar.google.com/scholar?start=10&q=paper+using+multiple+regression&hl=en&as_sdt=0,5#d=gs_qabs&u=%23p%3DTNVYmEJGg1IJ
12. https://scholar.google.com/scholar?start=20&q=paper+using+multiple+regression&hl=en&as_sdt=0,5#d=gs_qabs&u=%23p%3DJ8EB57QdBCMJ
13. https://scholar.google.com/scholar?start=20&q=paper+using+multiple+regression&hl=en&as_sdt=0,5#d=gs_qabs&u=%23p%3D1OU-dh8lQXAJ
14. https://scholar.google.com/scholar?start=30&q=paper+using+multiple+regression&hl=en&as_sdt=0,5#d=gs_qabs&u=%23p%3DtSHDRNZF_kIJ
15. https://scholar.google.com/scholar?start=40&q=paper+using+multiple+regression&hl=en&as_sdt=0,5#d=gs_qabs&u=%23p%3DU65SIDWF6VgJ