PREDICTIVE MODELING & CRIME DETECTION

Yug Jyotirmay Singh | Sagrika Chandra | Kartikeya Gupta | Parth Ghumare

INTRODUCTION

We are aware of the fact that any community is never free of crime, which is an alarming aspect of the society. Hence, it is essential to reduce crime rates in order to foster prosperity and peace. Through our project, we want to identify and analyze crime patterns and prevent these before they take place. The ability to predict crimes that can occur in the future can help law enforcement agencies in strategically devising preventative measures and exercising caution.

Our potentiaL audience will be law enforcement bodies and state authorities.

This, in turn, would aid in dropping crime rates in a specific neighbourhood, and our project would be useful not only for law enforcement bodies but also for society at large. This would also help efficiently divert resources such as patrol cars and police personnel.

OUR DATASET

Combining information from the socio-economic data from the 1990 Law Enforcement Management and Admin Stats survey and crime data in the 1995 FBI UCR, the dataset contains socio-economic and crime-related information on 2,215 communities across the United States.

	state	communityname	fold	population	householdsize	racepctblack	racePctWhite	racePctAsian	racePctHisp	agePct12t21	...	PctForeignBorn	PctBornSameState	PctSameHouse85
0	1	Alabastercity	7	0.01	0.61	0.21	0.83	0.02	0.01	0.41	...	0.03	0.70	0.40
1	1	AlexanderCitycity	10	0.01	0.41	0.55	0.57	0.01	0.00	0.47	...	0.00	0.93	0.66
2	1	Annistoncity	3	0.03	0.34	0.86	0.30	0.04	0.01	0.41	...	0.04	0.77	0.59
3	1	Athenscity	8	0.01	0.38	0.35	0.71	0.04	0.01	0.39	...	0.03	0.78	0.56
4	1	Auburncity	1	0.04	0.37	0.32	0.70	0.21	0.02	1.00	...	0.12	0.49	0.12
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1988	56	Gillettecity	9	0.01	0.53	0.00	0.96	0.02	0.06	0.47	...	0.01	0.32	0.34
1989	56	GreenRivercity	9	0.00	0.67	0.01	0.91	0.03	0.21	0.56	...	0.06	0.35	0.55
1990	56	Laramiecity	3	0.03	0.40	0.02	0.90	0.14	0.12	0.89	...	0.10	0.37	0.24
1991	56	RockSpringscity	7	0.01	0.45	0.02	0.92	0.06	0.14	0.48	...	0.05	0.46	0.59
1992	56	Sheridancity	8	0.01	0.29	0.00	0.97	0.03	0.04	0.40	...	0.05	0.45	0.49

1993 rows × 104 columns

VARIABLES OF INTEREST

ViolentCrimesPerPop

total number of violent crimes per 100K population GOAL attribute (to be predicted).

Population

population for community

fold

fold number for non-random 10 fold cross validation, potentially useful for debugging, paired tests - not predictive (numeric)

State

US state (by number) - not counted as predictive above, but if considered, should be considered nominal

HouseholdSize

mean people per household (numeric-decimal)

PRE-PROCESSING DATA

Our feature variable of interest is ViolentCrimesPerPop. This variable is modeled against other features of the dataset to determine their effect on the number of violent Crimes.

The ViolentCrimesPerPop variable was transformed into a new feature variable named ‘highCrime’. The values that were above 0.1 were considered 1 otherwise 0.

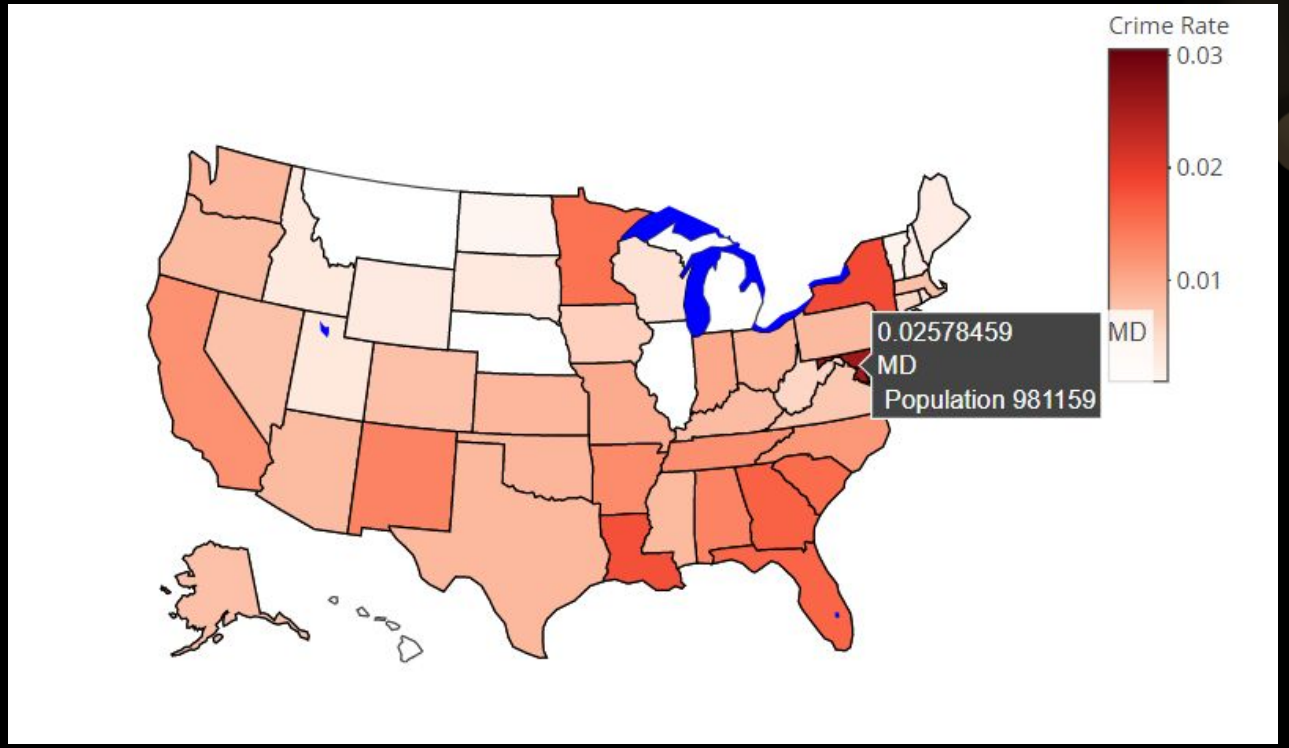
To reduce redundancy and avoid misleading results , the variables with a high proportion of NA values (greater than 75%) were removed.

We saw that the variables householdsize and PersPerOccupHous represent the same information, mean people per household. PersPerOccupHous was removed from the dataset.

Data Exploration

Chester City has an extremely high violent crime rate, while Atlanta, Miami, and Newark have very high populations in comparison to the others

Violent Crime rates in the US



District of Columbia leads in terms of violent crime rates in the United States of America.

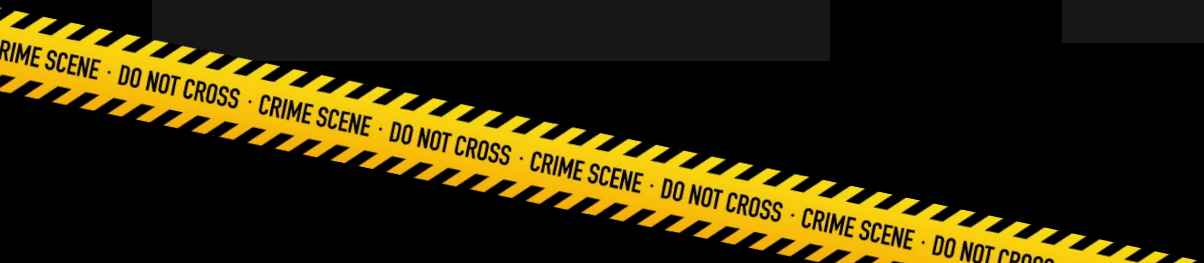
ANALYSIS

01 LINEAR REGRESSION

Linear regression is a statistical method that allows us to summarize and study relationships between continuous (quantitative) variables. It is a linear approach to modeling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables).

02 DECISION TREE

Decision tree learning uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). The advantage of the Decision Tree is the interpretability of the results.



RESULTS

	Model	Multiple R-sq	MSE
0	Linear Regression	0.6694	118166
1	Decision Tree	0.6268	133361

- **The linear regression model that we built is in itself a very good model with a multiple R2 of 66.94%.**
- **Decision tree is great for interpretation but significantly under-performs compared to linear regression.**

IMPLICATIONS

Violent crime rates in a community increase with:

- **Increase in the number of kids born to people who never married**
- **Decrease in the percent of kids of age 12-17 in two-parent households**
- **Increase in the percentage of male divorcees**
- **Increase in the percentage of people sharing the same room**
- **Decrease in the percentage of households with investment/rent income**