Communities and crime

Prediction of violent crime in the USA

Marie Lontsie Zanmene Yanis Bosch

22nd May, 2024

Outline

1. The dataset
2. Preprocessing
3. Regression
4. Performance analysis
5. Conclusion

The dataset

- Data sources:
  - Socio-economic data from the 1990 US Census
  - Law enforcement data from the 1990 US LEMAS survey
  - Crime data from the 1995 FBI UCR
- Creator: Michael Redmond, La Salle University, Philadelphia
- Date: 13th July, 2009

The dataset

- Size: 1994 rows, 128 columns
- Example attributes: police officers per 100K population, median rent,...
- Goal: Prediction of violent crime in the USA

Preprocessing

Column Name	Missing values	Column Name	Missing values
PolicReqPerOffic	1675(84%)	PolicAveOTWorked	1675(84%)
PolicPerPop	1675(84%)	RacialMatchCommPol	1675(84%)
PctPolicWhite	1675(84%)	PctPolicBlack	1675(84%)
PctPolicHisp	1675(84%)	PctPolicAsian	1675(84%)
PctPolicMinor	1675(84%)	OfficAssgnDrugUnits	1675(84%)
NumKindsDrugsSeiz	1675(84%)	LemasSwFTFieldPerPop	1675(84%)
LemasTotReqPerPop	1675(84%)	LemasSwFTFieldOps	1675(84%)
LemasSwFTPerPop	1675(84%)	PolicCars	1675(84%)
PolicOperBudg	1675(84%)	LemasPctPolicOnPatr	1675(84%)
LemasGangUnitDeploy	1675(84%)	LemasSwornFT	1675(84%)
PolicBudgPerPop	1675(84%)	LemasTotalReq	1675(84%)
OtherPerCap	1(0.05%)		

Table 1: Total number of rows: 1994

Preprocessing

Listwise deletion:

- = Method for handling missing data
- Delete columns or rows that have any missing data at all
- Very simple method to deal with missing data
- Loss of information, and thus loss in the quality of the prediction
- Good method so long as we retain sufficient power after deletion

Preprocessing

Imputation:

- = Method for handling missing data
- Replace missing values with substituted data
- Ex: Median, Average,...
- Less loss of information
- May introduce bias in the correlation
- Leads to lower standard errors, which may lead to Type 1 errors

Preprocessing

Why can we use listwise deletion on the columns with 84% of missing data?

- Most of the entries are missing, thus we don't lose too much data
- We have very little data left to base our imputation on, which would make it a bad choice

Preprocessing

How do we handle the one missing entry in the OtherPerCap column?

- Delete the column, but we would lose 1994 entries
- Use imputation, which should work well in this case
- Delete the row, and lose one out of 1994 rows = minimal loss of information

We deleted the row containing the missing value to keep our code as simple as we can

- https://en.wikipedia.org/wiki/Listwise_deletion
- https://en.wikipedia.org/wiki/Imputation_(statistics)