Communities and crime

Prediction of violent crime in the USA

Marie Lontsie Zanmene Yanis Bosch

23rd May, 2024

Outline

1. The dataset
2. Preprocessing
3. Correlation analysis
4. Regression
5. Performance analysis
6. 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

The dataset

- As in most countries, violent crime is driven by socio-economic factors
- There seems to be a strong link between income inequality and crime
- Does our data confirm this?
- Which of these factors are of the highest importance?

Preprocessing

- Before studying these correlations we must make sure our data is clean
- The values are already normalised, we must thus turn our attention to missing values

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?

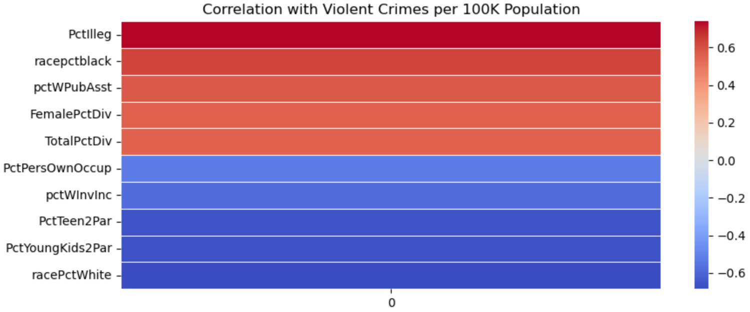
- Delete the column, but we would lose 1994 entries
- Use imputation, which should work well in this case
- Delete the row, and lose one out of 1994 rows = minimal loss of information

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

Correlation analysis

- Before applying a regression algorithm, it would be interesting to check which predictors are significant
- Thus we plot a graph with the correlation between the predictors and violent crime
- We exclude all predictors with a correlation that lies close to 0

Correlation analysis



Correlation analysis

INSERT CLOSER ANALYSIS OF SOME OF PREDICTORS WITH BEST CORRELATION (OR INVERSE CORRELATION)

Regression

- Given that our response variable is continuous, we have to perform regression to predict it
- Idea: Use random forest regression

Regression

What is random forest regression?

- Based on ensemble learning
  - = method where multiple ML algorithms are combined
- Utilises subsets of the data to create multiple trees (= bagging)
- The obtained results are averaged to create the final result

Regression

What are the advantages of random forest regression?

- Performs well with little to no hyperparameter tuning
- Rarely overfits
- Low sensitivity to noise
- Good at noticing general patterns in the data

Regression

What are the disadvantages of random forest regression?

- Bad at extrapolation
- Makes predictions only in the range of data contained in the training set

Regression

Why can we use random forest regression?

- Our data seems to be diverse enough to cover a realistic range of crime rates
- It seems unlikely that we might have to predict a crime rate that is much higher than in our training set
- We have quite a few predictors left, even after cleaning, thus overfitting could be an issue

Sources

- https://en.wikipedia.org/wiki/Listwise_deletion
- https://en.wikipedia.org/wiki/Imputation_(statistics)
- https://www.theanalysisfactor.com/mean-imputation/
- https://www.theanalysisfactor.com/when-listwise-deletion-works/
- https://cnvrg.io/random-forest-regression/
- https://minds.wisconsin.edu/bitstream/handle/1793/77496/Violent%