Datasheet for Inmates with Violent Crimes

Siobhan B. Durcan, MSc Candidate at the University of Sussex

December 2021

Inmates with Violent Crimes is a data set derived from the US Survey of Prison Inmates 2016 (SPI) (combined US State and Federal Data, version 4) [1]. A datasheet for the SPI can be found here SPI Datasheet [2].

The information in this document concerns only the Inmates with Violent Crimes data set. The code to reproduce the Inmates with Violent Crimes data set can be found on the US Prisoners Github. The repository also contains a Google Colab notebook that imports the package and transforms the data set.

1 Datasheet

Motivation

For what purpose was the data set created?

The motivation for creating the data set was to provide a new criminal justice data set suitable for use in machine learning research. The original data set is not accessible to machine learning practitioners; in order to understand what data is contained in each of the 2104 variables, two documents, each of c. 1000 pages long, must be carefully read and cross referenced. Variables were selected for inclusion in the data set based on ability to predict if the prisoner would receive a harsh sentence. Further detail can be found in the paper Inmates with Violent Crimes stored in the Github repository.

Who created this dataset and on behalf of which entity?

The Inmates with Violent crimes data was created by Siobhan B. Durcan as part of an MSc at the University of Sussex, supervised by Novi Quadrianto of the Advanced Data Analytics Group at the University of Sussex. The original data was produced by the US Justice Bureau of Statistics; see the SPI 2016 Datasheet for more details.

Who funded the creation of the data set?

This project was completed as part of a taught Masters programme. No grants or funding were awarded for this particular project.

Any other comments? The funding for the original data came from US government sources- see the SPI 2016 Datasheet.

Composition

What do the instances that comprise the data set represent?

Each instance captures the answers of one prisoner interviewed as part of the SPI .

How many instances are there in total (of each type, if appropriate)?

There are 10248 instances in the Inmates with Violent Crimes data set.

Does the data set contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?

The Inmates with Violent Crimes subset is filtered to contain only instances where the prisoner has received a custodial sentence and where the controlling offense was a violent crime. In most cases where an inmate has been sentenced for multiple offenses, the inmate's controlling offense is the most serious offense or the offense for which they received the longest sentence. The distributions of key subgroups in the Inmates with Violent Crimes data set are consistent with that of the original SPI data set.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features?

Each instance consists of features describing the offenses the inmate has been sentenced for, demographic information, education, citizenship, children, living situation and income at time of arrest, criminal history, victim demographics and victim injuries. A full list of features can be found in the appendices of the paper on Github.

Is there a label or target associated with each instance? If so, please provide a description.

The label 'sentence above 25 years' has been created for each instance. If prisoners have a flat sentence (meaning it is for a specific length of time rather than a range) of 25 years or above, the label is set to harsh (1). If prisoners have a sentence of life (or variations of this such as life without parole), the label is set to 1 (harsh). In instances where the sentence is a range- i.e., 5-25 years, the label is set to harsh if either the minimum or maximum range is equal to or above 25 years.

Is any information missing from individual instances?

Some inmates provided a Don't Know or Refuse to Answer response to questions. If a one hot encoded and scaled version of the data is generated using the Colab notebook provided, these missing values are filled. The variable configuration document on Github provides the encoding for Missing and Don't Know/Refuse values.

Are relationships between individual instances made explicit ?

There are no relationships between individual instances.

Are there recommended data splits?

Assuming the threshold of 25 years is used for harsh/not harsh, there are no further recommendations regarding data splits. The US Prisoners package allows users to adjust the threshold at which a sentence is considered harsh. The Colab notebook provided on the Github will produce the version of the data set that is discussed in this Datasheet

Are there any errors, sources of noise, or redundancies in the data set?

Some variables in the original data were subject to a programming error. Only a few of these variables are used in the Inmates with Violent Crimes subset, and no issues seem to have been introduced as a result.

Is the data set self-contained, or does it link to or otherwise rely on external resources?

The original SPI data is needed in order to reproduce the data set using the Colab notebook provided.

Does the data set contain data that might be considered confidential ?

No. The public use version of the data set was used, and all variables that could identify participants were removed.

Does the data set contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

Yes. The data set contains information about violent crimes such as murder, rape, sexual abuse, bodily injuries suffered as a result of violent crimes, domestic abuse and crimes with child victims. The references to these topics are limited to discussing how variables containing this information were processed. Many of the variable names are codes (i.e., V1234) rather than semantic descriptions, but the Appendix contains a semantic description of variables, as does the variable configuration file in the code-base (which acts as a metadata dictionary).

Does the data set relate to people?

Yes

Does the data set identify any sub-populations (e.g., by age, gender)?

Yes, the data set contains a number of features that indicate sub populations.

Offender white (yes:4379, no: 5869) Offenders were able to select multiple race categories they identified with, and offender white was set to true if white was selected. The SPI also collected information about how others perceive offender race, but this was not considered when deriving the offender white variable.

Offender male (yes: 8341, no: 1907) Offender male was set to true if the offender's self-described sex identify was male, but did not consider sex assigned at birth.

Victim white (yes: 5704, no: 4544) In the case of offenders with multiple victims, victim white is set to true if the offender identified that all or more than half of the victims were white, but not if they identified less than half of the victims were white.

Victim male (yes: 5512, no: 4736) In the case of offenders with multiple victims, victim male is set to true if the offender identified that all or more than half of the victims were male, but not if they identified less than half of the victims as male.

The variables above have been created as binary variables for the purpose of machine learning bias algorithms. Victim male has a much higher distribution of the positive class (men), but the other variables are balanced.

The below variables were not derived for the purposes of bias assessment, but do represent sub populations in the data set. Upon inspection, it was found that these variables are not well balanced in the data

Born in US (yes: 9417, no: 831)

Special Educational Needs (yes: 2482, no: 7766)

Dyslexia/Discalculia (yes: 1502,

no: 8746)

ADD/ADHD (yes: 2304, no: 7944)

Heterosexual (yes: 9305, no: 942)

Age (18-24: 2969, 25-34: 2614, 35-44: 2097, 45-54: 1142, 55-64: 402)

Marital Status (Never Married: 5968, Divorced: 2137, Married: 1292, Separated: 436, Widowed: 404)

Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the data set?

It may be possible to match information in the data set to non-anonymised sources (such a newspaper articles) and identify either offenders or victims. However, there are no known instances were this is possible and doing so would directly contradict use the terms of use of the SPI.

Does the data set contain data that might be considered sensitive in any way?

The data set contains the sensitive features described above, as well as information about citizenship status, financial status and medical history. However, the SPI data that was used to create this data set was fully anonymised.

Any other comments? If the one-hot-encoded features are selected according to mutual information, offender sex is the fourteenth most informative metric in relation to a harsh sentence (with a threshold of 25 years). If ranking features according to chi2 metric, offender sex ranks as 30, widowed ranks 44 and divorced as 48. A Decision Tree classifier trained on this data determined that divorced inmates should receive a harsh sentence, whereas inmates with other marital statuses would not.

Collection Process

How was the data associated with each instance acquired?

The data was not collected from participants as part of this project. This project sourced the data from the Bureau of Justice Statistic's Departments website. The original data was collected directly from participants using computer assisted personal interviewing.

What mechanisms or procedures were used to collect the data?

The survey was downloaded as a tsv file from the ICSPR website and processed using Python. The code was checked for accuracy as it was developed by inspecting data outputs against expected results.

If the data set is a sample from a larger set, what was the sampling strategy?

The data set contains instances where the inmate had received a custodial sentence for a violent crime (or several crimes of which the controlling offense was a violent crime).

Who was involved in the data collection process (e.g., students, crowd workers, contractors) and how were they compensated (e.g., how much were crowd workers paid)?

The SPI documentation and/or datasheet should be consulted for details regarding how the original data was collected. The public-use data used in this project was sourced by an MSc student from the ICSPR website.

The original data was collected in 2016. The version of the data used in this project is version 4 on the ICSPR website, and was downloaded in August 2022.

Were any ethical review processes conducted (e.g., by an institutional review board)?

No ethical reviews were conducted as part of the further processing on this data.

Does the data set relate to people? If not, you may skip the remaining questions in this section.

Yes

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

The data was obtained via ICSPR website.

Were the individuals in question notified about the data collection?

The individuals were notified of data collection and consented to participation in the US Survey of Prison Inmates. New consent was not obtained for the further processing of this data as part of this project.

Did the individuals in question consent to the collection and use of their data?

The US Bureau of Justice Statistics have not published consent forms for the SPI. The survey documentation notes that prisoners were notified about data collection prior to the survey and that two minutes in each interview was used to re obtain consent.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?

N/A: new consent was not obtained for this project.

Has an analysis of the potential impact of the data-set and its use on data subjects (e.g., a data protection impact analysis) been conducted?

Yes, a review was conducted by the author. The Bureau of US Justice statistics anonymised survey responses and redacted variables that could be used for re-identification purposes. It was therefore determined that no further data protection impact analysis was required.

Any other comments?

No other comments.

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done?

Extensive preprocessing was required in order to transform the original SPI data into a format suitable for use in machine learning tasks.

The logic used to derive the label (sentence harsh or not harsh) and the sensitive attributes (offender white, offender male, victim white, victim male) is described above. Other derived variables include:

Victim below 12: for multiple victims, this is set to 1 (yes) if the youngest victim was below 12, regardless of the age of the oldest victim.

Victim injuries: victim injury information from the multiple victim section and single victim sections were collapsed into one set of variables. If the offense listed was rape or murder, the variables for victim raped and victim died were updated to true (as the survey permitted the offense to be rape without rape being listed as an injury).

It's important to note again that victim information was provided by the offender and not the victims.

Each prisoner provided up to five current offenses and corresponding offense types. This information is represented in the data set as offense counts. For the highest frequency offense codes, there is a column for each possible offense code, and the value on each row is the frequency of that code in the list of offenses provided by the participants. The maximum value for any offense code would be 5; this would occur if the prisoner had been sentenced for 5 counts of the same offense code. Locating the location of current and controlling offenses, and their corresponding types, was a labour intensive task. It's recommended to review the paper and the accompanying diagrams for more information.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?

The raw data is available on the ICSPR website. The code provided on Github, the paper and associated documentation can be used to identify the changes applied to the data in order to produce the Inmate with Violent Crimes data set.

Is the software used to preprocess/clean/label the instances available?

The package US Prisoners was created to produce the subset.The code uses standard Python libraries such as numpy and pandas. The package allows users to configure what variables from the survey to include, set how Missing and Don't Know responses are encoded for each included variable and select whether one hot encoding, ordinal encoding or scaling should be applied. The library encodes data in such a way that feature names are retained on the one hot encoded data set, which allows for easier inspection of models. Users can also choose to produce a dense version of the data set without feature preparation.The Github also contains a Colab notebook that loads in the US Prisoners package and processes the SPI data using the appropriate variable configuration and parameter settings to produce the Inmates with Violent Crimes subset.

Any other comments?

Ordinal encoding of offenses would require a judgement about what offenses were worse than others. However, offense seriousness is a multivariate issue that considers the details of the offense and number of victims. Given that violent crimes such as rape, murder and bodily harm are contained within the data set, it did not feel appropriate to make a judgement about how offense codes should be ranked.

Uses

Has the data set been used for any tasks already? If so, please provide a description.

As part of the same project that produced this data set, the Inmates with Violent Crimes subset was used to predict if an offender would receive a harsh or not harsh sentence with a Decision Tree, Multi Layer Perceptron and Linear Discriminant Analysis models. .

Is there a repository that links to any or all papers or systems that

use the data set? If so, please provide a link or other access point.

The Github contains the MSc paper submitted as part of this project. See the SPI Datasheet for papers that use the original data.

What (other) tasks could the data set be used for?

Assuming the code is configured to produce the Violent Crimes subset, the data set would contain other variables that could be predicted; such as if time already served will be applied to the prisoners sentence or if they must undergo a drug or alcohol treatment programme in addition to their custodial sentence. If the configuration was changed to include more variables from the original data set, other labels such as if an inmate is likely to take up a work programme or receive visits from children could be created.

Is there anything about the composition of the data set or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?

All of the information in the Survey was self-reported by offenders; as such, any conclusions drawn from the data or models trained on it may not reflect ground truth. The data set was assessed for bias using Statistical Parity (SP) and Disparate Impact (DI), and some disparity between outcomes based on protected attributes was identified. Varying how the label is determined or how features are derived may produce different fairness metric scores. Where the protected attribute is that of an offender, the favourable label is 0 (a less harsh sentence), whereas for a victim a favourable label is 1 (a more harsh sentence).

Offender white (SP: -0.04, DP: 0.93) Offender male (SP: 0.12, DP: 1.24) Victim white (SP: 0.07, DI: 1.01) Victim male (SP: -0.05, DI: 0.89).

Are there tasks for which the data set should not be used? If so, please provide a description.

The Inmates with Violent Crimes should not be used as a fixed population analysis. The SPI documentation provides extensive guidance on survey weighting and how to conduct analysis using the full data set. The data set should also not be used to make predictions about how likely offenders are to receive a custodial sentence as the population consists only of people who have received a sentence.

Models trained on the data used variables associated to arrest year as a signal for whether the inmate should receive a harsh sentence or not. This doesn't reflect the human reasoning that would have been applied at the time of sentencing; there is likely to be some correlation with certain years and shifts in criminal justice policy.

If working with the offense code columns (offense counts or controlling offense codes), note that it is not possible to distinguish between abortion, child abuse and gang crime as they are grouped under the same controlling offense in US law (code 180).

Any other comments? Another consideration that arose is that the victim information (sex, age, race, injuries suffered, relationship to offender) was reported by the offender. In this sense, victims have little agency in how this data was produced and so it would be advisable not to use this data to predict victim information; rather, the victim information should be considered as proxies that may or not not correspond with ground truth.

Distribution

Will the data set be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the data set was created? If so, please provide a description.

The Inmates with Violent Crimes data set will not be distributed itself, but a Colab notebook that reproduces the new data set with will be available. Users will need to source the original SPI data from the ICSPR website.

How will the data set will be distributed (e.g., tarball on website, API, GitHub) Does the data set have a digital object identifier (DOI)?

The US Prisoners package and the Colab notebook implementing it is available on Github. The variable configuration file and the parameters passed to the data processor are set to produce the Inmates with Violent Crimes data set as described in this datasheet. If more variables are included, if the sentence threshold is changed or other pre-processing hyperparmaeters adjusted, the data set will vary. No project artefacts have a DOI.

When will the data set be distributed?

The Colab notebook and the US Prisoners package will be available from August 2022.

Will the data set be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?

The code will be available under an MIT license.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances?

The author believes that the Terms of Use on the Survey of Prison Inmates would apply to the Inmates with Violent Crimes data set. The SPI terms require that permission be obtained to distribute the data. This permission has not been requested (due to the nature of this work as part of an MSc), hence the only code to produce the data set has been provided (rather than the data set itself). The Terms of Use Can be found here https://www.icpsr.umich.edu/web/ICPSR/studies/37692/terms

Do any export controls or other regulatory restrictions apply to the data set or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

The Terms of Use of the SPI do not reference export controls or regulatory restrictions.

Any other comments?

No.

Maintenance

Who will be supporting/hosting/maintaining the data set?

There are no commitments to maintaining the processed subset or associated code by the University of Sussex, although Siobhan B. Durcan (as an individual) will endeavour to do so.

How can the owner/curator/manager of the data set be contacted (e.g., email address)?

via the Github page.

Is there an erratum? If so, please provide a link or other access point.

There is currently no erratum. If one is created, it will be stored on Github.

Will the data set be updated (e.g., to correct labeling errors, add new instances, delete instances)?

There is no commitment to updating the data set. The repository can be followed on Github in order to be notified of any updates that are made.

If the data set relates to people, are there applicable limits on the retention of the data associated with the instances)?

It is assumed that any copies of the Inmates with Violent Crimes data created using the provided code can be retained for as long as the SPI data is available on the ICSPR website.

Will older versions of the data set continue to be supported/hosted/maintained?

There is no commitment to supporting the processed subset.

If others want to extend/augment/build on/contribute to the data set, is there a mechanism for them to do so?

The easiest way to enhance the data set would be by adding additional variables to the variable configuration. The Colab notebook shows how the variable configuration can be edited. Anyone who would like to change the base configuration can submit a pull request on Github, or fork the repository. In order to add derived new variables, a method on the dataset processor class can be added. The variable derivation methods follow a documented pattern of: copying the relevant columns, applying the transformation, dropping columns that were used to derive this variable and updating the variable configuration file with the new variable name and how it should be encoded or scaled.

Any other comments?

No.

References

[1]United States. Bureau of Justice Statistics, Survey of Prison Inmates, United States, 2016, Inter-university Consortium for Political and Social Research, 2021

[2] M. Zilka, B. Butcher and A. Weller, ”A Survey and Datasheet Repository of Publicly Available US Criminal Justice Datasets,” https://openreview.net/.