A Dataset for Studying Cash Bail Decisions in Pennsylvania

Joshua Williams
Computer Science Department
Carnegie Mellon University
jnwillia@cs.cmu.edu

J. Zico Kolter
Computer Science Department
Carnegie Mellon University
Bosch Center for AI
zkolter@cs.cmu.edu

This datasheet was created using the examples from [1]

1 Motivation

For what purpose was the dataset created? This dataset was created for the purpose of providing a nuanced analysis on how willing judges are to set cash bail for members of certain groups (eg. age, gender, race, etc) as opposed to others based on the associated societal 'cost' of defendants from this group not appearing for their court date. The analysis consisted of cases from January 2016 to December 2019, however, the full dataset consists of cases outside of this scope. The full, original analysis that gathered and used this dataset can be found at the following link:

2 Composition

What do the instances that comprise the dataset represent? This dataset consists of instances of court cases taken from the Unified Judicial System of Pennsylvania. These records are publicly available on the Unified Judicial System of Pennsylvania's website.

How many instances are there in total (of each type, if appropriate)? The dataset contains 116942 instances. As this dataset was used to analyze bail decisions it may be worth noting that there are 52087 instances in which a magistrate set cash bail and 4717 instances in which a defendant failed to appear for a court date.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? This dataset consists of a large sample of court cases from Pennsylvania, primarily from January 2016 to December 2019. This sample consists of cases from the Court of Common Pleas in Philadelphia County and Allegheny County, Pennsylvania. The Court of Common Pleas is generally used for more serious crimes, which in turn correlates with defendants being less likely to fail to appear than the population of all defendants (likely due to the increased use of cash bail and harsher sentences).

Cases from Allegheny County may also be taken from the presiding magistrate's Magisterial District Court, which sees a broader range of cases than the Court of Common Pleas, however, such cases may be transferred to the Court of Common Pleas under a different docket identifier. In our cleaning process, we exclude such duplicate cases and use the court records from the Magisterial District Court.

What data does each instance consist of? Each row within this dataset comprises data extracted from individual court cases. Data for the court cases are extracted from dockets obtained from the Unified Judicial System of Pennsylvania. Each row consists of the following features. It should be noted that __ represents an the offense grade; M = Misdemeanor, F = Felony, S = Summary Offense, Unlabeled = Unlabeled Offense grade, 1 = First Degree, 2 = Second Degree, 3 = Third Degree. (Eg. F1 = First Degree Felony):

1. **Date**: The month and year that the arrest took place (if not present, the docket filing date)
2. **Magistrate**: A numerical identifier for the presiding magistrate that initially assigned bail
3. **County**: The County where the court proceedings took place
4. **Lead Offense**: The most serious offense labelled by the docket (may be left empty)

5. **Lead Description**: A description of the lead offense's statute from PA Code - Title 18 (may be left empty)
6. **Number __ Charges** The number of charges for the specified offense grade within this case
7. **Bail Type**: The type of bail assigned to a defendant
8. **Bail Amount**: The amount required to be released on bond (non-zero for Monetary and Unsecured)
9. **Bail Status**: The status of a defendants bail (eg. whether bail was posted or revoked)
10. **Fail To Appear**: Whether a defendant did not appear for a court date
11. **Race**: A defendant's race
12. **Sex**: A defendant's sex
13. **Age**: A 5-year age range that contains a defendant's age at the time of the case
14. **Attorney**: The type of attorney for a defendant (ie. private, public, or waived the right)
15. **Prior __**: The number of prior convictions for the given offense grade at the time of the current case
16. **Zip Code Median Income**: The median income for a defendant's residential zip code
17. **Zip Code High School Graduates**: The high school graduation rate for a defendant's residential zip code
18. **Zip Code College Graduates**: The college graduation for a defendant's residential zip code
19. **Zip Code Employment**: The employment for a defendant's residential zip code
20. **Zip Code Poverty Rate**: The poverty rate for a defendant's residential zip code
21. **Zip Code Median Age**: The median age for a defendant's residential zip code

Is there a label or target associated with each instances Depending on the purpose, type of bail, bail amount, or failure to appear indicators may be used as labels or predictors. The analysis that this dataset was prepared from predicted probabilities of cash bail assignment based on other features and the probability of a defendant failing to appear for their court date and not failing to post bail.

Is any information missing from individual instances Information may be missing from individual instances; the underlying court records that were aggregated to form this dataset are prone to mistakes in their transcription (eg. entering 1905 instead of 2005 for a person's arrest date). Similarly, some information may not be present within the dockets, which in turn means that it will not be present within our dataset (eg. Not listing information about a defendant's attorney). Finally, some defendant addresses may not be searchable within census records (eg. homelessness).

In total, of the 116942 instances within the data, 88850 instances do not have any missing data.

Are relationships between individual instances made explicit? Should a defendant commit multiple offenses over several years, the court records for each offense will be included as separate instances. Thus, multiple instances within the dataset may be from the same defendant, however, these instances are still from different, unrelated court cases. Such instances should not be identifiable from the data itself.

Are there recommended data splits (eg., training, development/validation, testing)? We have no recommendations for splitting the data.

Are there any errors, sources of noise, or redundancies in the dataset? The underlying dockets in most cases do not explicitly state whether a defendant failed to appear or not. Even in such cases bail is reinstated, which implies that a defendant failed to appear for the court date for a justifiable reason (eg. Unable to get off of work, unable to find child care, illness, etc). We label a defendant as having failed to appear for their court date by considering cases in which a magistrate revokes bail, does not reinstate it, and a bench warrant was put out for their arrest or the docket explicitly states that the defendant failed to appear.

Such cases make up 4.03% of instances within the dataset, which is comparable to the average rate of missed court appearance for the Common Pleas Court in Philadelphia 4.30% as reported by the Philadelphia Inquirer [2]

Does this dataset contain data that might be considered confidential? All data gathered here is taken from the publicly available court records as gathered by the Unified Judicial System of Pennsylvania.

Does this dataset contain data that, if viewed directly might be offensive, insulting, threatening, or otherwise cause anxiety? No, this dataset solely consists of information extracted from court records. However, as these are records from the US criminal justice system, such data and related statistics may be upsetting.

Does the dataset identify subpopulations (eg. by age, gender)? This dataset labels instances of a defendant's age (within a range of 5 years from their true age), race, sex, and census data for the defendant's associated zipcode. All information was available as part of publicly available records.

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 dataset? Due to the sensitive nature of this data, we, in conversation with members of the ACLU of Pennsylvania, have anonymized this dataset by removing the docket numbers and other court id's, replacing magistrate names with numeric identifiers, grouping dates only by month and year, and grouping ages into bins of 5 years.

Based on the acceptable inputs to the Unified Judicial System of Pennsylvania's records from the Court of Common Pleas and Magisterial District Courts, we believe that a particular instance will be unable to identify any particular defendant unless one were to first retrieve all court cases from the county for a particular month and year. In doing so, one has all necessary data to identify an individual without necessitating any information from this dataset. As an added precaution, we shuffled all data within each month/year, in order to remove a correlation between the order of cases within this dataset and the date recorded in the docket.

3 Collection Process

How was the data associated with each instance acquired? We scraped cases over the period of January 2016 to December 2020 from the Unified Judicial System of Pennsylvania. Each record consists of a pdf that describes a defendant's prior court history and a more detailed json file for each case from the PAeDocket API. The data comprising each instance was then extracted from a combination of both sources.

We then use the defendant's participant address from the json record in tandem with the US Census 2018 American Community Survey 5-Year data to determine the census-based confounders.

What mechanisms or procedures were used to collect the data (e.g.,hardware apparatus or sensor, manual human curation, software program, software API)? This data was collected through custom python scripts.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? Signed inn 2018, Pennsylvania has implemented the first-of-its-kind, "Clean Slate" law, which has sealed millions of criminal records in an effort to allow Pennsylvania residents who have not been found guilty of serious crimes (eg. violence, sexual assault, homicide, etc) to more easily continue with their lives. This law seals records so that they do not show up in employer background searches nor in Pennsylvania's online docket database.

As a result of this law, this dataset is based around more serious cases from the Court of Common Pleas. Thus, this dataset comprises a subset of all court cases that over emphasizes serious court cases.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g.,how much were crowdworkers paid)? The data was collected solely by the authors of this datasheet.

Over what timeframe was the data collected? This dataset was collected from October 2019 to August 2020.

Were any ethical review processes conducted (e.g., by an institutional review board)? This work consists solely of publicly available data. As such, it did not go through an institutional review.

Does the dataset relate to people? This dataset consists of individual court cases, as such each instance relates to the defendant that was charged in each case.

Did the individuals in question consent to the collection and use of their data? No, the data was publicly available through the Pennsylvania Judiciary. For privacy reasons, the site does not provide the means for contacting individual defendants.

4 Preprocessing/Cleaning/Labeling

Was any preprocessing/cleaning/labeling of the data done Within this dataset, we removed all defendants below the age of 18 and bucketed the remaining ages into a range of 5 years (eg. 18-22, 23-27, ...).

Additionally, for the feature: Zip Code Median Income, the census data represents all incomes greater than $250000 per year, as "$250,000+". All instances of median income within this dataset that are equal to $250000 may be higher.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data(e.g.,to support unanticipated future uses)? All of the original court summaries and docket json files are saved locally. Additionally, in case of unanticipated issues, we have a saved, but not released, a copy of this dataset, that includes docket identifiers and magistrate names.

Is the software used to preprocess/clean/label the instances available? The custom scripts to retrieve the data are not being released, however, the raw data is publicly available. Individuals are open to retrieving this data from its original source.

5 Uses

Has the dataset been used for any tasks already? This dataset was originally gathered and used in an analysis of magistrate cash bail decisions over the years 2016 to 2019.

Is there a repository that links to any or all papers or systems thatuse the dataset? We do not plan to create such a repository. At the time of release the only use of this dataset is the aforementioned cash bail analysis.

What (other) tasks could the dataset be used for? We do not anticipate any uses of this dataset outside of other cash bail/failure to appear analyses. We give more detail below on cases for which this data should not be used.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? Due to the passing of the "Clean Slate" law in Pennsylvania, some individuals may request their information removed from the Unified Judicial System of Pennsylvania. As this data is anonymized and we are not in contact with the defendants herein, we are not checking if cases have been removed in order to remove them from the dataset. In other words, over time, some cases may be present in the released data here, while not being available online. However, as the majority of cases are based on serious offenses from the Court of Common Pleas, this is unlikely to be a significant number.

Are there tasks for which the dataset should not be used? While a user may be interested in understanding the amount that bail is set at, there are unknown, exogeneous variables that make this task likely unidentifiable with respect to the magistrates that set them.

Moreover, without magistrate names, this dataset should not be used to make decisions on a specific magistrate's propensity for setting cash bail. Only the aggregate trends of magistrate decisions or county-level treatment of defendants who belong to one group or another.

It may also be possible to learn associations between zip-code data and defendant treatment, however, these are only proxies for true defendant information (eg. nothing precludes an individual with an income of $100000 from living in a zip code with median income $20000). We would be wary of such a use with the data included here.

6 Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? Yes, the dataset is publicly available.

How will the dataset will be distributed (e.g., tarball on website, API,GitHub)? The dataset can be downloaded from the following github repo, maintained by the first author: `https://github.com/jnwilliams/padockets`

When will the dataset be distributed? The data was released in January 2021.

Will the dataset be distributed under a copyright or other intellectualproperty (IP) license, and/or under applicable terms of use (ToU)? The data included consists of publicly available instances from the Unified Judicial System of Pennsylvania. There is no copyright or other ToU.

If one would like to retrieve data from the original source, be aware of their terms of use and be reasonable in rate-limiting web crawlers.

Have any third parties imposed IP-based or other restrictions on thedata associated with the instances? There are no fees or restrictions

7 Maintenance

Who will be supporting/hosting/maintaining the dataset? The first author of the related analysis and this datasheet will be maintaining the dataset through the associated github repo.

Is there an erratum? All changes to the dataset will be included as separate commits on github. Verify changes through the commit history.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? No. We have no means for contacting defendants included here and do our best to ensure that their information is protected within this dataset.

If others want to extend/augment/build on/contribute to the dataset,is there a mechanism for them to do so? There is no mechanism within this dataset, but others are free to use the original source and create a new version that fits their needs.

References

[1] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., and Crawford, K. Datasheets for datasets. *arXiv preprint arXiv:1803.09010* (2018).

[2] Melamed, S. Philly da larry krasner stopped seeking bail for low-level crimes. here's what happened next., 2019. Available at Philadelphia Inquirer; https://www.inquirer.com/news/philly-district-attorney-larry-krasner-money-bail-criminal-justice-reform-incarceration-20190219.html.