Nutritional Label for Cook County Mental Health Predictions
Responsible Data Science Project Draft

Team Members: Alene Rhea, Kelsey Markey

Background

This report will analyze the Cook County Mental Health Prediction project created by the team members during a 2019 term project for the NYU course, “Introduction to Data Science”. The project aimed to predict mental health-related legal outcomes from a set of judicial and case-based features available only at initiation. The goal of the ADS is to identify people who are likely suffering from mental illness as early as possible in the legal process, so as to provide them adequate support during their movement through the criminal justice system, while simultaneously minimizing the cost incurred by the county.

There is a natural trade-off between the dual goals of offering support to as many people as possible, and minimizing costs. In deployment, each additional positive prediction would represent one more person receiving support, as well as more money spent by the county to provide that support. Model hyperparameters were selected to optimize Area Under the Receiver-Operating Curve, with an eye to recall. The choice of AUC as a tuning metric speaks to the importance of the trade-off between cost and inclusion.

Input and Output

Data Source

The data comes from the December 2, 2019 updates to the Initiation, Disposition, and Sentencing datasets on the Cook County Open Data Portal (https://datacatalog.cookcountyil.gov). The project was moved to GitHub in January 2020, at which point the data cleaning and model tuning notebooks were adapted slightly to ensure further reproducibility (https://github.com/kelseymarkey/cook-county-mental-health-prediction). The Initiation dataset includes all cases at the start of the legal process and is used for model prediction. The Disposition and Sentencing datasets represent, respectively, the resolution and judgement imposed by the courts, and are used to construct the target variable for the training dataset (see Figure 1).

The scope of the project was limited to a single large county so that applicable laws would be uniform. Cook County was selected because it’s very populous, and has a well-kept open data portal with fairly well-documented metadata (Markey, Rhea et. al., 2019). The fact that the Cook County Criminal Justice System is at the forefront of the movement to use specialty treatment courts and programs to address mental illness means that the Sentencing and Dispositions datasets contain outcome information which could be used as proxies to predict mental illness (Braude & Alaimo, 2007). The choice of Cook County was further intended to maximize potential impact: the number of individuals with mental illness in the Cook County jail has been reported to be as high as 30%, exceeding the national average by nearly 10% (Behavioral Health Innovations, 2015).

Candidate Keys and IDs

In both the Initiation and Dispositions datasets, each row represents one charge against a participant in a case. Accordingly, the combination of 'CASE_PARTICIPANT_ID' and 'CHARGE_ID' form a candidate key in those datasets.

It seems as though a new row in Sentencing is generated whenever a case participant is sentenced or resentenced. However, because there is no unique identifier for a sentence, there is no simple or obvious candidate key in sentencing. Using a method inspired by the Apriori / candidate generation algorithm, we discovered that any candidate key must contain a minimum of 8 features (Section 2, Jupyter Notebook). (Note that only IDs and sentencing-specific features were considered for this task.) Two such 8-feature candidate keys exist, and 7 of their features are shared: CASE_PARTICIPANT_ID COMMITMENT_TERM, COMMITMENT_TYPE, CURRENT_SENTENCE, SENTENCE_DATE, SENTENCE_PHASE, and SENTENCE_TYPE. To these 7 features, we can add either CHARGE_ID or CHARGE_VERSION_ID to yield a candidate key. It's important to remember that the uniqueness of these candidate keys may not be guaranteed by the business rules which generate the data. It is entirely possible that the next data update to the Sentencing dataset could contain duplicate tuples over these 8 features.

There are 4 ID types in each dataset: CASE_ID, CASE_PARTICIPANT_ID, CHARGE_ID, and CHARGE_VERSION_ID. According to documentation available on the Cook County Open Data Portal (https://datacatalog.cookcountyil.gov), CASE_ID is an 'internal unique identifier for each case,' CASE_PARTICIPANT_ID is an 'internal unique identifier for each person associated with a case,' CHARGE_ID is an 'internal unique identifier for each charge filed,' and CHARGE_VERSION_ID is an 'internal unique identifier for each version of a charge associated with charges filed.'

In all three datasets, each CHARGE_VERSION_ID corresponds to only a single CHARGE_ID, but each CHARGE_ID can have several CHARGE_VERSION_IDS. This indicates that charges retain their original CHARGE_ID as they are updated.
Each CASE_ID can be associated with multiple CASE_PARTICIPANT_IDs, but each CASE_PARTICIPANT_IDs is associated with only a single CASE_ID. This means that there can be multiple defendants in a single case, and a person will receive a new CASE_PARTICIPANT_ID for each case they are associated with. There is no way to link one person's records across different cases.

Model Features

In order to replicate a use case where MHI is predicted at initiation, only the 27 attributes in the Initiation dataset were used to construct input features to pass to the model (See Table 4 in Appendix). After one-hot encoding and feature engineering, the data contains 5616 features. (Note that the three raw datasets use capital letters for attribute names, whereas the dataframe used as model input uses lowercase; this difference will be replicated throughout this paper to distinguish between pre-and-post processed features.)

Target Variable

The ADS uses a binary target variable called “mental health indicator” (MHI), which indicates whether or not there is a record of the individual incurring a mental health-related court outcome. An MHI of 1 indicates the presence of a mental health outcome in an individual’s court records, and an MHI of 0 indicates the absence of such an outcome.

The assignment of MHI is based on proxy features, such as sentence_type = “Inpatient Mental Health Services” or charge_disposition = “Finding Guilty but Mentally Ill.” It’s important to note that the use of these proxies is likely to produce a positive feedback loop which reinforces the courts’ conception of mental illness. These proxies were chosen via an exhaustive search of values in the Sentencing and Disposition dataset which identified 15 values across six columns (see Table 5 in Appendix).

MHI was engineered by merging the four proxy features from Sentencing with the two proxy features from Dispositions, using the common identifier CASE_PARTICIPANT_ID. The ADS aggregates to the level of case_participant_id, and assigns an MHI of 1 if *any* of the associated charges contain any of the 15 values of interest.

Case_participant_id is used to link the target variable to the features.

Output

The ADS’ classifier can be used to predict either the (future) binary MHI status of an individual, or a score meant to estimate the probability of a mental-health related outcome.

Data Profiling

Input features, both those used for the engineering of MHI (from Sentencing and Disposition) and for prediction (from Initiation), were examined for object type, number of missing values, cardinality (i.e. number of unique values), percent unique values, and value distribution; results can be seen in Tables 1, 2, and 3 (Section 1, Jupyter Notebook). The “Value Information” column in these tables reports: 1) standard statistical measures (minimum, maximum mean, median) for numeric variables, 2) the percent true and false for boolean variables (“PRIMARY_CHARGE”), 3) the top two categories or ids and their normalized value_counts categories for categorical and id variables.

Missing value heatmaps were also created to better understand the relationship between features with missing values (Figures 2-4; Section 8, Jupyter Notebook). These heatmaps display all features that have more than 0 missing values, and visualize the correlation matrix for missing values. They help to confirm some expected dependencies between missing features such as 1) the perfect positive correlation between missing DISPOSITION_CHARGED_SECTION and missing DISPOSITION_CHARGED_ACT (the legal section and legal act for a charge, in both Sentencing and Disposition, and as ACT and SECTION in Initiation), 2) the perfect correlation between missing ARRAGINMENT_DATE and missing LENGTH_OF_CASE_IN_DAYS (in Sentencing, suggesting that if there is no arraignment date, the case did not proceed), 3) perfect or near perfect correlation between missing COMMITMENT_TERM, COMMITMENT_TYPE, and COMMITMENT_UNIT (in

Sentencing and Disposition), 4) strong correlation between missing DISPOSTION_CHARGED_CLASS and DISPOSTION_CHARGED_AOIC (correlation = 0.8 in Sentencing and Disposition, AOIC is an ID for law of the charge which would be expected to correlate with the legal class for a charge), 5) COURT_NAME and COURT_FACILITY (in Sentencing, correlation = 0.6), and 6) EVENT and EVENT_DATE (in Initiation). In all three datasets RACE and GENDER are often missing together (correlation = 0.8), perhaps suggesting that demographic information is either entered completely or not at all. Additionally, AGE_AT_INCIDENT and INCIDENT_BEGIN_DATE are often missing together (correlation = 0.8-0.9 in all three datasets), perhaps relating to procedures in the documentation of an arrest.

Relationships between missing features inspired further exploration of functional dependencies and business rules, using a method described by Heiko Muller's "Data Profiling & Data Cleaning" lecture (Mueller, 2019). Each of the three original datasets was viewed in tabular form at the highest level possible (i.e., zoomed all the way out), and the team then looked for visual patterns.

In Sentencing, we also found that when CHARGE_DISPOSITION is "Nolle Prosecution," CHARGE_DISPOSITION_REASON is far more likely to be non-empty. "Nolle Prosecution" indicates that charges have been dropped (i.e. not pursued) by the prosecutor, so it makes sense that the disposition reason would be better documented in this case. We also found that DISPOSITION_CHARGED_CHAPTER is of highly variable length and format. When DISPOSITION_CHARGED_CHAPTER is short, DISPOSITION_CHARGED_SECTION is non-empty, and DISPOSITION_CHARGED_ACT is much more likely to be non-empty as well. This indicates that chapter, section, and act are often stored all together in DISPOSITION_CHARGED_CHAPTER. In Initiation, we find that CHAPTER, ACT, SECTION, and CLASS are far more uniform in length and format, indicating that the data input process at initiation is more standardized, and perhaps more centralized, than the process at sentencing.

High-level correlation heatmaps were created for each of the three input datasets, and include all non-categorical and non-datetime features (Figures 5-7). In these we see clearly the relationships between CASE_PARTICIPANT_ID, CASE_ID, CHARGE_ID and CHARGE_VERSION_ID discussed earlier. We also see some feature correlations similar to those in the missing value heatmaps, with strong correlations between ACT and CHAPTER in Initiation and DISPOSITION_CHARGED_ACT and DISPOSTION_CHARGED_CLASS in Sentencing. In all three datasets the features OFFENSE_CATEGORY and UPDATED_OFFENSE_CATEGORY have high correlations both with each other and with many other features, particularly ACT, CHAPTER, CLASS, and UNIT. This is in line with expectations of relationships between the legal category, act, chapter, class, and law enforcement unit of a charge. We also see correlations between UNIT (law enforcement unit), COURT_FACILITY, and COURT_NAME in Disposition, which might be related to Cook County specialized courts. In Sentencing and Disposition CHARGE_DISPOSTION_REASON is strongly correlated with many other features, such as COURT_FACILITY (and COURT_NAME), SENTENCE_TYPE, PRIMARY_CHARGE, OFFENSE_CATEGORY (and UPDATED_OFFENSE_CATEGORY), UNIT, COMMITMENT TYPE, and COMMITMENT_UNIT. This is also likely due to Cook County specialized courts, since the CHARGE_DISPOSTION_REASON feature holds "additional information about the result of the charge" and when it is not missing (99.66% of the time in

Sentencing and 73.31% in Disposition) it is populated with values such as “Drug Court Graduate”, “PG to Other Courts”, and “Mental Health Graduate”. Finally, we see in Disposition and Sentencing that OFFENSE_CATEGORY is often correlated with the ID features, potentially reflecting cases and participants with multiple charges or cases where multiple case participants are charged with the same charge.

Later in the project protected features will be removed from the model in order to study changes in performance, so initial investigations were done to understand possible proxies for protected classes (Section 4 of Jupyter Notebook). We first looked to see if there was a relationship between CHARGE_COUNT and demographic classes (RACE, AGE_AT_INCIDENT, and GENDER) by comparing the distribution of CHARGE_COUNT across the entire dataset to that of a sub-populations within the class. Since CHARGE_COUNT is a numeric feature, Kolmogorov–Smirnov tests were also performed to better understand differences in distributions. We did not find significant differences in CHARGE_COUNT across genders (Male, Female, and Unknown) or age (binned in 10 year increments from 17-96). CASE_PARTICIPANT_IDs with race = black (around 66% of the dataset) had a distribution most similar to that of the entire population (K-S value = 0.006), whereas those with RACE = Unknown (~ 1% of the dataset) had the highest K-S value of 0.147. There were some distribution differences in other less-represented sub-populations such as race = Biracial, but low sample group sizes prevented these results from being significant. Differences in OFFENSE_CATEGORY across racial sub-populations was studied in the same way, and the highest K-L value (0.833) was obtained for the Hispanic group, which also was more likely to have OFFENSE_CATEGORY = “aggravated dui” or “driving with suspended or revoked license” than “narcotics” or “unlawful use of weapon” like the entire population distribution (Figures 8-9).

It was also hypothesized that race could be reconstructed by location, either through INCIDENT_CITY, LAW_ENFORCEMENT_AGENCY, OR UNIT (the law enforcement unit associated with the arrest). The same process was repeated for these features and revealed clear differences in incident city, law enforcement agency, and unit across race, with certain values being predominant in certain racial sub-populations and not others. This was particularly clear again for the Hispanic population across incident cities, where we saw both a high K-L value (1.05) and distinct differences in distribution and common incident cities(Figure 10-11). This was logical considering the racial and ethnic differences across cities, neighborhoods, and the law enforcement agencies and units that serve them.

Diving further into the relationship between location and law enforcement, we studied the relationship between INCIDENT_CITY and LAW_ENFORCEMENT_AGENCY (Section 5 of Jupyter Notebook). It seemed at first that there might be a functional dependency between these two features, for example when INCIDENT_CITY = “bartlett” and LAW_ENFORCEMENT_AGENCY = “bartlett pd”, however further investigation showed that these relationships are not always one-to-one and that in some cases a specific LAW_ENFORCEMENT_AGENCY does not represent a single INCIDENT_CITY (i.e. LAW_ENFORCEMENT_AGENCY = “amtrak national railroad passenger corp”).

We have reproduced in this report the target variable distributions across subpopulations of race, gender, and age that were included in the original report (Figures 12-15; Section 6, Jupyter Notebook; figures in appendix). We plan to discuss these distributions, and compare them to available ground truth figures; see below in Validation section. MHI prevalence is also studied with respect to time and presence of narcotics (Figures 16-18).

Implementation and Validation

Data Pre-Processing

The ADS is concerned with analysis at the level of the individual person, so the data is aggregated over each CASE_PARTICIPANT_ID, effectively collapsing multiple charges to a single row representing all the charges against a given person in a given case. case_participant_id becomes the fundamental identifier, and primary key, of the resultant dataset.

Categorical features are one-hot encoded before aggregation. The aggregation function for each feature is determined by whether or not the feature ever varies between the charges of the same CASE_PARTICIPANT_ID: median is used for features that are constant across charges, and sum is used for those that vary across charges (see Table 6 in Appendix). The only exception to this rule is for CHARGE_COUNT, where the max is taken to indicate the total number of charges associated with the CASE_PARTICIPANT_ID.

The Initiation dataset was then filtered so that it only included rows corresponding to CASE_PARTICIPANT_IDs that were also present in Disposition and/or Sentencing. This ensured that each case had completed its movement through the legal process and could be assigned an MHI based on features in the later datasets. The existence of re-sentencing leads to a censoring problem wherein the newest cases have lower base rates, because they've not had as much time for a proxy feature to trigger a positive MHI. [See Figures 14-15 in Appendix.]

There was extensive data cleaning performed on the datasets in order to prepare the training set, some which violate best practice protocols. To start, all string variables were converted to lowercase, all numeric columns were converted to integers, and date features were converted to datetime assigning missing or unknown dates a filler value of “1900-01-01 00:00:00”. This filler value is problematic since new features were later created (i.e. season, incident_length, weekday) based on these values, which were likely assigned values which do not represent the true data. Additionally, all missing non-numeric inputs were replaced with “unknown” (a common filler already being used by Cook County) and age were converted to integers with null and outlying values (greater than 100) replaced by median age. The latter is a data leakage issue since missing values are imputed on the complete dataset, prior to splitting into training, test, and validation sets. Additionally, since age was an important factor in the resulting model, this ADS could be improved by building a model to predict and impute missing ages (after splitting). All gender values except “Male” and “Female” were converted to “Unknown” (this included null values, “Unknown”, “Male name, no gender given”, and “Unknown Gender”), despite the fact that each of these values could potentially encode different information (for example transgender individuals). Messy encoding of race was left untouched, and justified by the idea that different encodings may represent different perceptions of race, if not actual ethnicities. It's possible, however, that the different encodings are instead procedural vestiges which should have been cleaned.

Several new features are engineered: age_over_100, age_unknown, weekday [based on arrest date], season [based on arrest date], incident length, [incident end - incident begin], latitude, and longitude [from incident city]. The ADS does not thoughtfully handle the encoding of latitude and longitude from INCIDENT_CITY, setting latitude and longitude to 0 when INCIDENT_CITY = unknown. These coordinates represent a location in the Atlantic Ocean off the west African coast, and the function should be improved by instead interpolating these missing values to some central location within Cook County.

Before modeling, the six datetime columns and four IDs are removed from the dataset (with case_participant_id remaining as the index). Since the dataset was very class imbalanced, the negative class was downsampled in the training set. To do this, 100% of positive instances were sampled and the negative class was sampled without replacement until the positive class comprised 50% of the training set population. The validation and test sets were not downsampled, to replicate deployment.

Implementation Overview

After pre-processing, the ADS feeds features and labels into a gradient-boosting ensemble classifier to predict either the (future) binary MHI status of an individual, or a score meant to estimate the probability of a mental-health related outcome. The classifier hyperparameters were tuned to maximize AUC, with an eye toward recall.

If the ADS were deployed as an autonomous decision maker, users (e.g. Cook County representatives or NGO/non-profit workers) could use operational or budgetary constraints to specify probability thresholds at which to offer additional support and services.

Validation

To avoid data-leakage, the ADS uses RECEIVED_DATE to create a single time-based training/validation/test split. The ADS has not been tested for robustness. We plan to use bootstrapping and/or walk-forward testing to test model variance. We would also like to test the model on new data if there is a data release before the final report is produced.

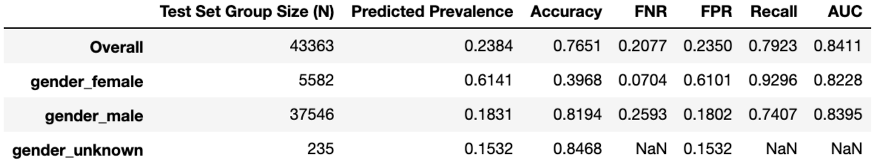
We plan to address the suitability of the target variable by comparing MHI base rates (i.e. prevalence) across sub-populations to epidemiological mental illness data. Baseline data sources for the epidemiological data have been located (https://www.nimh.nih.gov/health/statistics/mental-illness.shtml, https://www.nami.org/learn-more/mental-health-by-the-numbers), and will be compared to the base rates found in sub-populations. We will also look for more specific epidemiological data sources which deal directly with mental health within the criminal justice system and/or Cook County.

Outcomes

Fairness Metrics

We have calculated predicted prevalence, accuracy, false positive rate, false negative rate, recall, and AUC for each gender and racial subgroup present in the test set (Tables 7-8; Section 7, Jupyter Notebook). Intersectional statistics can be found in the appendix (Table 9).

Due to class imbalance, accuracy is not a particularly useful performance metric for this ADS; AUC can instead be used to measure the classifier's discrimination prowess. We include FNR and FPR in order to understand how errors are distributed across subgroups. Error rate analysis is crucial for this ADS, because a false positive represents a negligible monetary harm to the county, and perhaps some small potential for damage to the pride or reputation of the individual, whereas a false negative could have devastating effects on the well-being of an individual who is denied much-needed support. We have additionally included recall, although it is equivalent to 1-FNR, because the original ADS team deemed it essential to monitor the portion of true positives which are correctly identified.

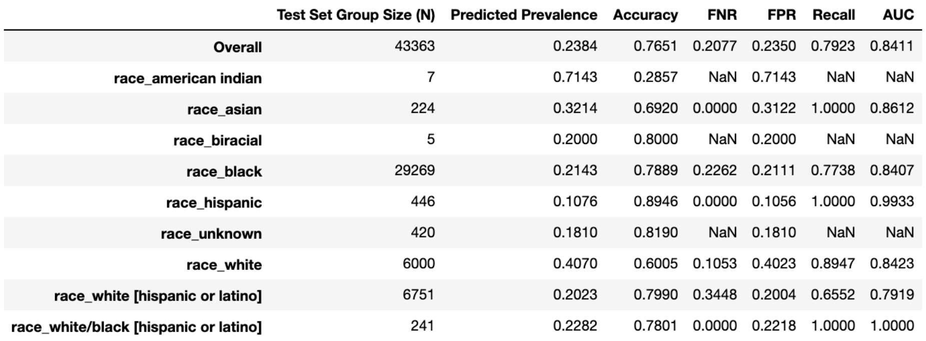


	Test Set Group Size (N)	Predicted Prevalence	Accuracy	FNR	FPR	Recall	AUC
Overall	43363	0.2384	0.7651	0.2077	0.2350	0.7923	0.8411
gender_female	5582	0.6141	0.3968	0.0704	0.6101	0.9296	0.8228
gender_male	37546	0.1831	0.8194	0.2593	0.1802	0.7407	0.8395
gender_unknown	235	0.1532	0.8468	NaN	0.1532	NaN	NaN

In the table displaying the fairness metrics across gender subpopulations, we can see that males are way overrepresented in this dataset, comprising over 86% of test-set instances. This is typical in a criminal justice setting, however this type of imbalance in training data can lead to poorer classifier performance for underrepresented groups. Indeed we find that the accuracy for females is far below that for males; more meaningfully, the AUC is lower as well.

To dig into what these differences mean, we can look at error rates. We find that for the female group, an astonishing 61% of negative instances are incorrectly identified as needing additional support, and 92% of positive instances are successfully caught. For the male group, 18% of negative instances are incorrectly identified, and only 84% of positive instances are successfully caught. Ultimately, it is clear that if this ADS were deployed, it would divert limited resources from men in need to provide unnecessary support to women without special mental health needs.

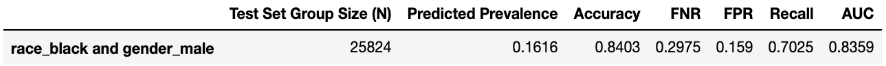
We find that the sample for gender_unknown is too small to calculate FNR, recall, or AUC. This is because no actual positives are present in the test set. This is equivalent to a test-set base rate of 0%. We find a very high overall accuracy for this group, which is achieved by assigning a negative label at an unusually high rate. We also find that the FPR is lower than it is for either the male or female group. This could prove problematic; if the target variable misses cases of mental illness, false positives could actually provide some benefit to individuals.



	Test Set Group Size (N)	Predicted Prevalence	Accuracy	FNR	FPR	Recall	AUC
Overall	43363	0.2384	0.7651	0.2077	0.2350	0.7923	0.8411
race_american indian	7	0.7143	0.2857	NaN	0.7143	NaN	NaN
race_asian	224	0.3214	0.6920	0.0000	0.3122	1.0000	0.8612
race_biracial	5	0.2000	0.8000	NaN	0.2000	NaN	NaN
race_black	29269	0.2143	0.7889	0.2262	0.2111	0.7738	0.8407
race_hispanic	446	0.1076	0.8946	0.0000	0.1056	1.0000	0.9933
race_unknown	420	0.1810	0.8190	NaN	0.1810	NaN	NaN
race_white	6000	0.4070	0.6005	0.1053	0.4023	0.8947	0.8423
race_white [hispanic or latino]	6751	0.2023	0.7990	0.3448	0.2004	0.6552	0.7919
race_white/black [hispanic or latino]	241	0.2282	0.7801	0.0000	0.2218	1.0000	1.0000

Accuracy and AUC vary widely between racial subgroups. By far the lowest accuracy is .29 for the American Indian subgroup, however there are only 7 individuals in that group, and no actual positives. The highest accuracy is 89% for the hispanic subgroup. The white/black [hispanic or Latino] subgroup achieves a perfect AUC of 1, meaning that there exists a threshold at which the classifier achieves an FPR of 0 and a TPR of 1. We know that this threshold must be lower than 0.5, because we are seeing a non-zero FPR at 0.5. (Note that this AUC should not be taken too seriously, as there are only 2 actual positives in the test set.) The lowest AUC we see is 0.79, surprisingly for the seemingly overlapping group of white [hispanic or Latino]. These two have fairly close accuracies, however, which speaks to the effect of class imbalance on accuracy.

Setting aside subgroups that do not have any positive instances in the test set, we find that the white subgroup has the highest FPR at 40%, indicating that this group will be offered un-needed specialized support at a disproportionate rate. The Asian, hispanic, and white/black [hispanic or Latino] all have perfect FNRs and recall, meaning that ADS correctly identified all positive instances of those groups in the test. Again we see a big difference with the white [hispanic or Latino] category, which has the overall highest FNR of 34%. The meaning of the difference between white [hispanic or Latino] and white/black [hispanic or Latino] is not immediately clear, though it seems a meaningful difference, if perhaps only as a vestige of methodology differences between departments.



	Test Set Group Size (N)	Predicted Prevalence	Accuracy	FNR	FPR	Recall	AUC
race_black and gender_male	25824	0.1616	0.8403	0.2975	0.159	0.7025	0.8359

It’s worth paying special attention to the black, male intersectional subgroup, because it makes up such a large portion of the dataset (60% in the test set). We find here an especially low predicted prevalence, FPR, and recall. This indicates that the ADS is biased against this very large group.

We plan to expand the above analysis to include age (binned into groups). We also plan to study the disparate impact present in the input data, and compare that to the disparate impact produced by the classifier, in terms of both race and gender. For this analysis, we will use race_white and gender_male as the respective “advantaged” categories, and group all other values into the “disadvantaged” category. Legal frameworks and the 80% DI threshold will be considered.

Interpretable Explanations

We used the LIME submodular picker to identify a set of 8 descriptive examples to explain the behavior of the ADS using 5 features each (Figure 19; Section 3, Jupyter Notebook). We find in several instances that these local explanations are not very good, meaning that the prediction made by the classifier is in fact not very well explained by the 5 features. We also do not have any positive instances represented in the set intended to provide a global explanation.

Several features stand out in these explanations, the importance of which may have otherwise gone unnoticed.

Chapter 625 shows up again and again, often in combination with Class x or Act 5; the absence of these values contributes to a positive classification. It is difficult to understand these features without a comprehensive understanding of Illinois penal code, but perfunctory research indicates that Chapter 625 may have to do with vehicles [https://codes.findlaw.com/il/chapter-625-vehicles/#!tid=NF5C556ECCF424FAC86A9D31D8C9591B9]. Given the unusually high proportion of aggravated DUI’s in the hispanic subgroup [ref to chart above], we hypothesize that there may be a connection between the importance of Chapter 625 and the low prevalence (both actual and predicted) of MHI in the group. We plan to explore this relationship further.

We find the charge_offense_title being something other than aggravated battery often contributes to a negative classification; it is possible that this has to do with the established correlation between mental illness and domestic violence (Tsirigotis & Luczak, 2017; Behavioral Health Innovations, 2015).

An updated_offense_category of narcotics often contributes to a negative prediction. The team made this same finding when attempting to construct an understandable decision tree for the original project, and noted that this seemed inconsistent with the fact that a large portion of Mental Health court programs involves drug support (Markey, Rhea et al., 2019). When building the ADS, the team also identified section 402, which relates to narcotics and their possession [https://www.cyberdriveillinois.com], as being important enough to receive its own feature.

Latitude being below 41.88 is seen several times to contribute to a negative MHI. We plan to conduct research into Chicago’s geography to determine whether this is a meaningful neighborhood division, and whether it might carry with it hidden racial information.

Gender is consistently one of the most heavily weighted features, with male always contributing to a negative classification, and female contributing to a positive one. In addition to gender, the sensitive features race and age_at_incident each show up in these examples. LIME is ultimately most useful for helping an individual understand a specific prediction. An individual is likely to be outraged to see these sensitive features as explanations, which points to the untrustworthiness of this ADS.

We would like to utilize the causal framework of QII to explain whether the ADS uses these sensitive features directly or indirectly, and to the impact of feature interactions. We plan to reach out to the authors of the QII paper to inquire as to existing implementations.

We plan to implement classic feature importance analysis, as well.

Additional methods for analyzing ADS performance include examining differences in model performance without explicitly using protected features RACE, GENDER, and AGE_AT_INCIDENT, as well as any features identified as strong proxies for those features.

Summary

- **Do you believe that the data was appropriate for this ADS?**

We have serious doubts as to whether this data was appropriate for the ADS.

First of all, a high degree of domain knowledge is necessary to understand the business rules governing the data, as well as some important features in it.

There is also a significant problem in that the ADS only learns cases of mental illnesses through the assignment of the target variable based on court-identified mental health outcomes. This effectively gives the court full control over what is considered a mental health disability and it is possible that there may exist bias towards certain types or presentations of mental illness. Our assessment of the appropriateness of this data will be heavily influenced by how strongly it diverges from ground-truth epidemiological data.

Finally, the data is specific to Cook County and thus the ADS cannot be exported to other jurisdictions. Furthermore, stakeholders within Cook County may have access to additional data which would be more useful than the data used here (e.g., medical records, previous criminal history, etc.).

- **Do you believe the implementation is robust, accurate, and fair? Discuss your choice of accuracy and fairness measures, and explain which stakeholders may find these measures appropriate.**

The team that produced the ADS clearly favors AUC as an accuracy measure. FNR and AUC would appeal to Cook County representatives who are keen to reduce costs. Subpopulations with high FNRs and low FPRs are likely to argue for the importance of these error rates.

Questions of robustness will be answered after attempts to perform further validation testing.

- **Would you be comfortable deploying this ADS in the public sector, or in the industry? Why so or why not?**

As mentioned in the original report (Markey, Rhea et al., 2019), the model is certainly not implementable because of its explicit use of protected classes (age, race, and gender) during prediction. Highly disparate error rates across sensitive subpopulations indicate that the ADS is amplifying bias; this will be confirmed via disparate impact calculations. As outlined in the Outcomes section, we plan to study changes in model performance without the explicit input of protected classes, which should guide additional discussions of implementability.

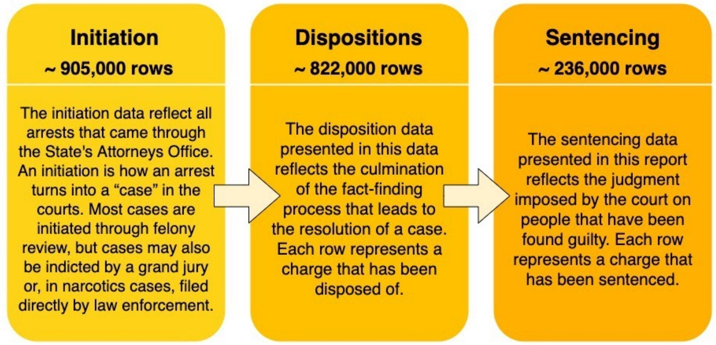
The original report also notes concept drift and censoring as potential pitfalls which would require careful monitoring during deployment:

> "As described above in the Training, Validation, and Test Sets section, concept drift is likely to be an issue in deployment. As the Cook County mental health court program continues to expand, we can expect an increase in base rate over time, which could eventually degrade model performance. Hence, there ought to be careful monitoring of the MHI base rate and of the legal and policy factors which may influence it. Periodic re-training may be necessary as models become out of date. Model custodians may also decide to exclude older data from training to mitigate concept drift -- this could be tested empirically, and would need to be considered in conjunction with the effects of censoring bias. Type-1 censoring bias is likely to have the opposite effect on our model, and mitigating it would require developing a heuristic cut-off point for the age of cases to be included in training (e.g., only training models on cases which have been in the system for 6 months of longer). Monitoring model sensitivity in deployment may prove challenging under censoring, but developing the aforementioned heuristic would give custodians a set time at which to evaluate an individual's MHI."
>
> (Markey, Rhea et al., 2019)

- **What improvements do you recommend to the data collection, processing, or analysis methodology?**

Data collection could certainly be better standardized, although that is not within the control of the ADS. Missing value imputation ought not take into account values in the test set. Walk-forward validation could be used to assess model variance. Fairness metrics need to be considered in evaluation.

Appendix: Figures



Initiation
~ 905,000 rows

The initiation data reflect all arrests that came through the State's Attorneys Office. An initiation is how an arrest turns into a "case" in the courts. Most cases are initiated through felony review, but cases may also be indicted by a grand jury or, in narcotics cases, filed directly by law enforcement.

Dispositions
~ 822,000 rows

The disposition data presented in this data reflects the culmination of the fact-finding process that leads to the resolution of a case. Each row represents a charge that has been disposed of.

Sentencing
~ 236,000 rows

The sentencing data presented in this report reflects the judgment imposed by the court on people that have been found guilty. Each row represents a charge that has been sentenced.

Figure 1: Descriptions of Cook County legal datasets used in this study

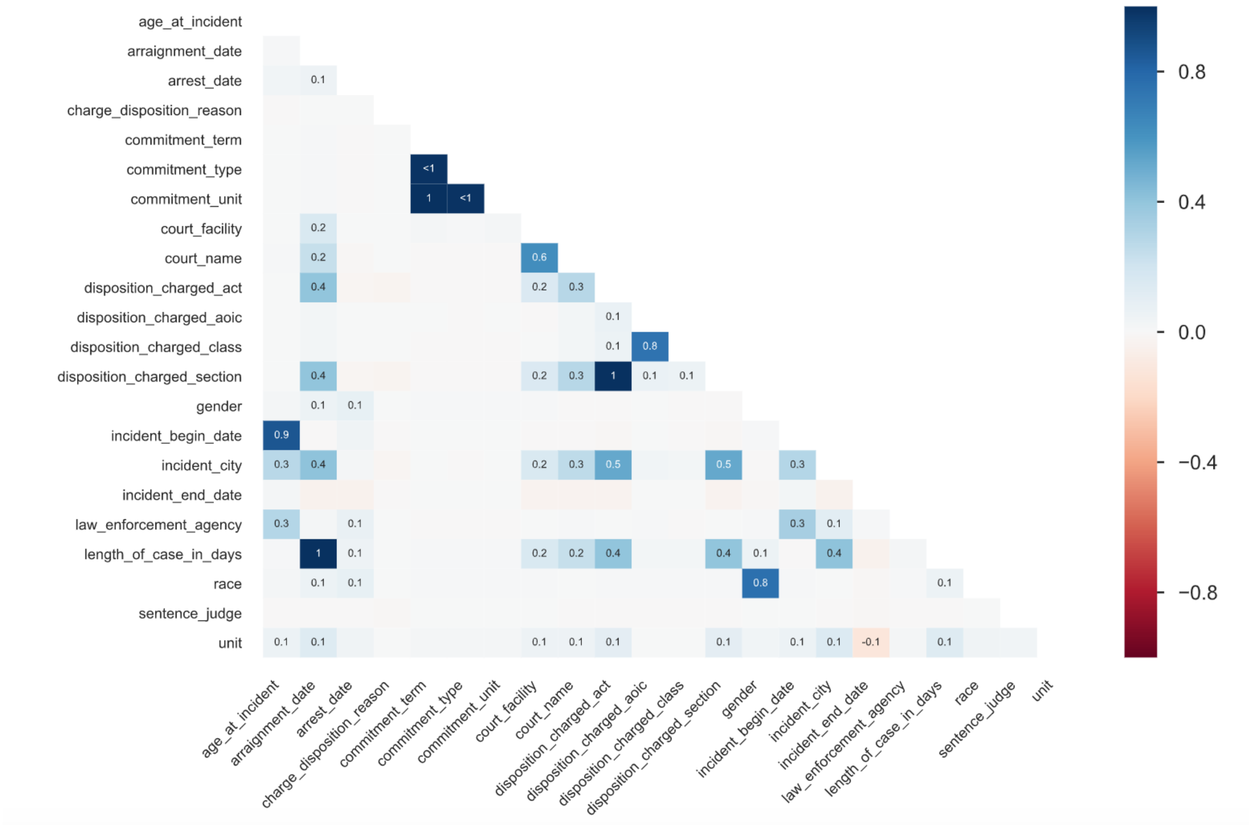


Figure 2: Sentencing Missing Value Heatmaps

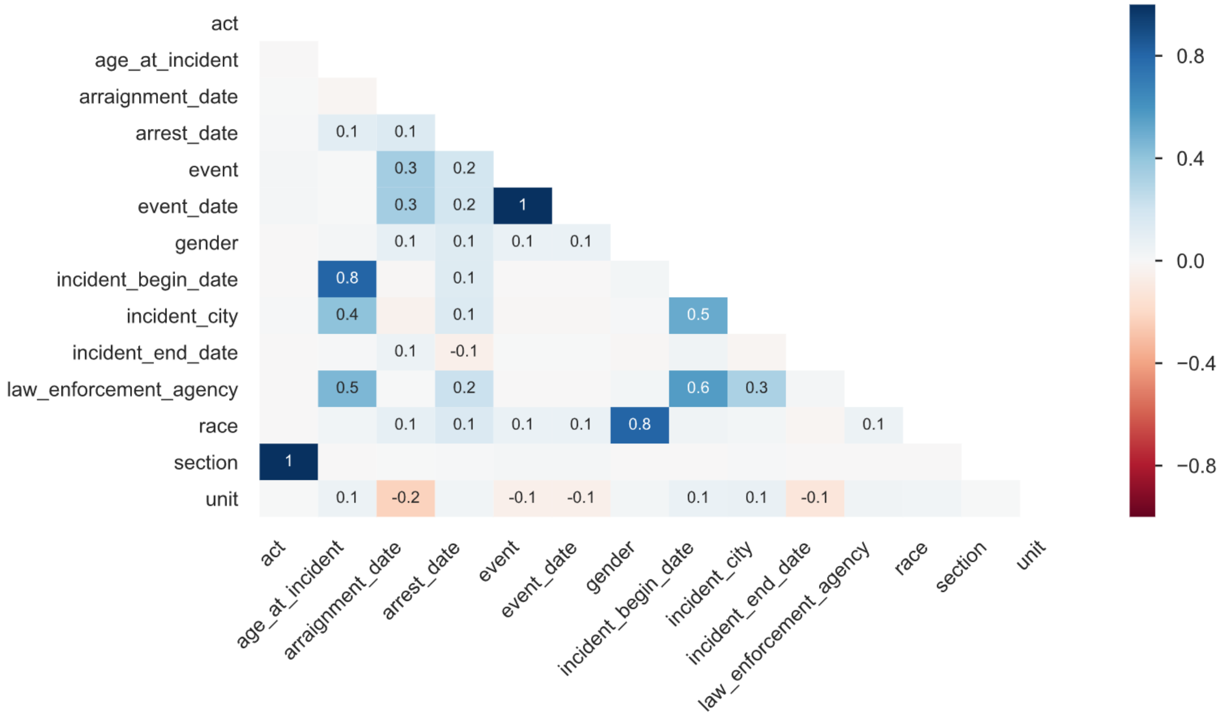


Figure 3: Initiation Missing Value Heatmaps

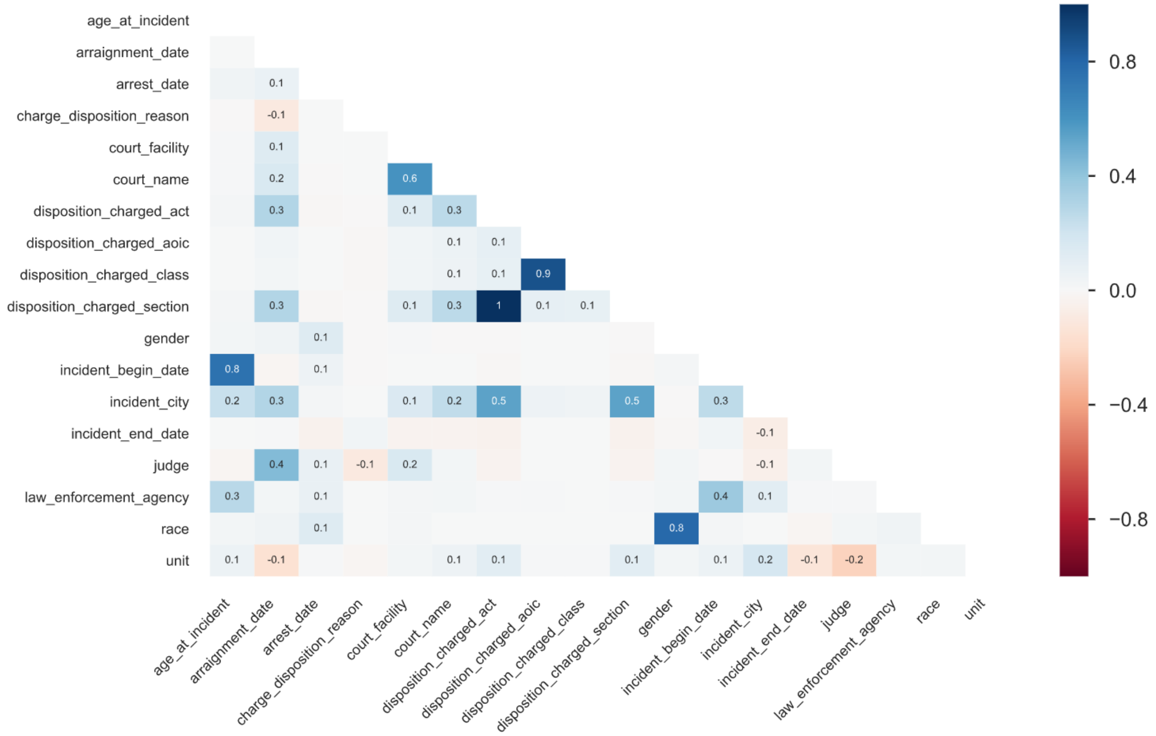


Figure 4: Disposition Missing Value Heatmaps

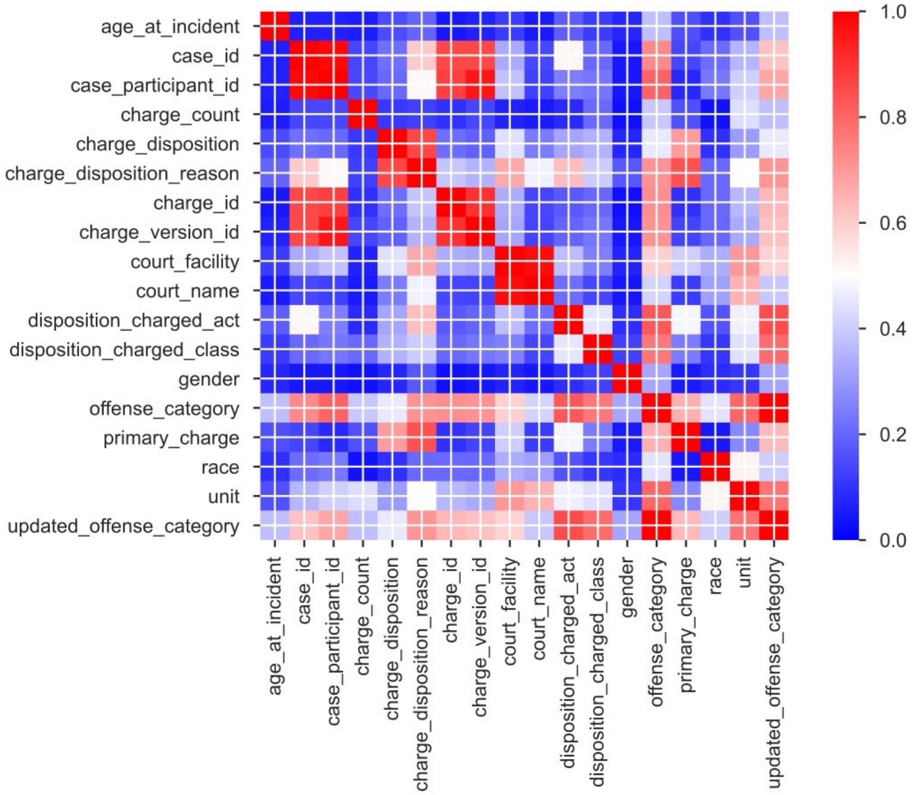


Figure 5: Disposition Correlation Heatmap

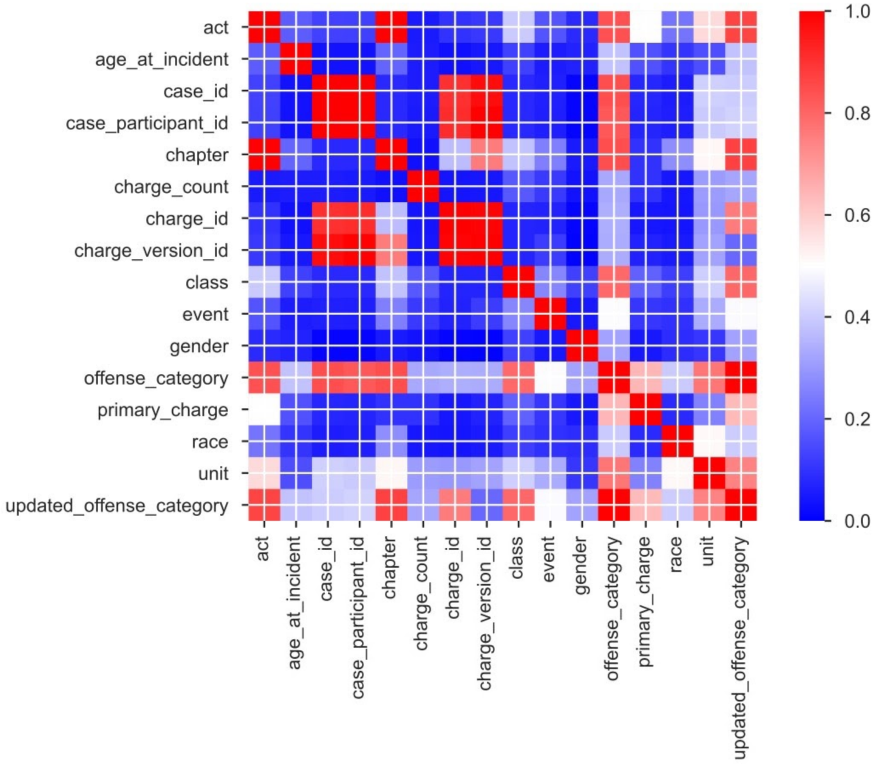


Figure 6: Initiation Correlation Heatmap

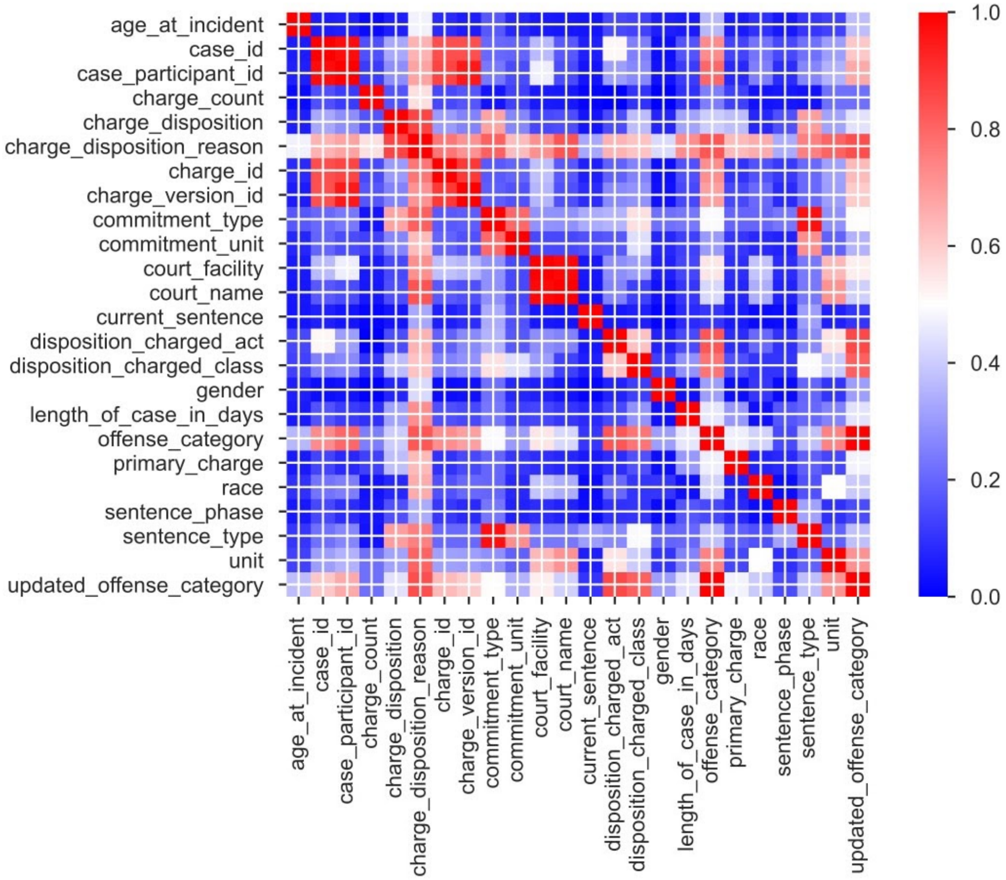


Figure 7: Sentencing Correlation Heatmap

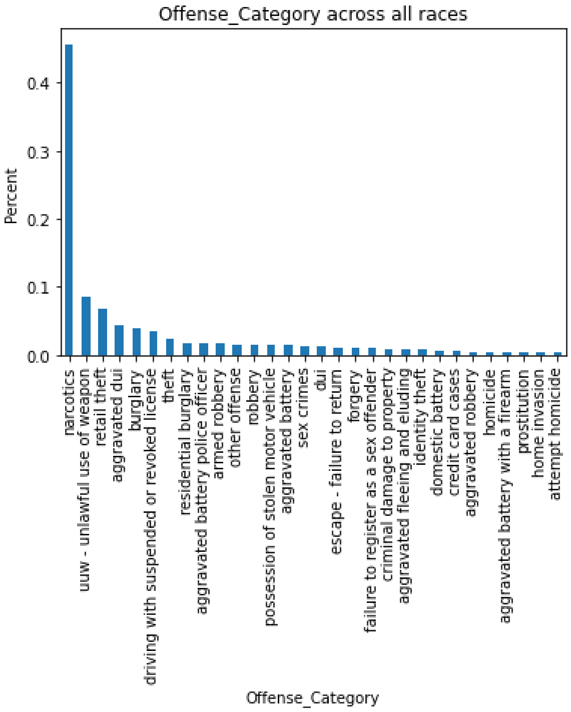


Figure 8: Offense_category across all races

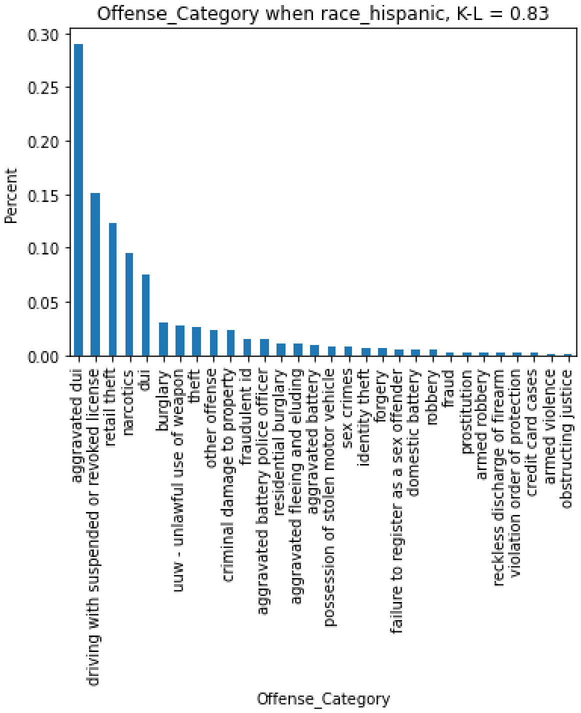


Figure 9: Offense_category for Race = Hispanic

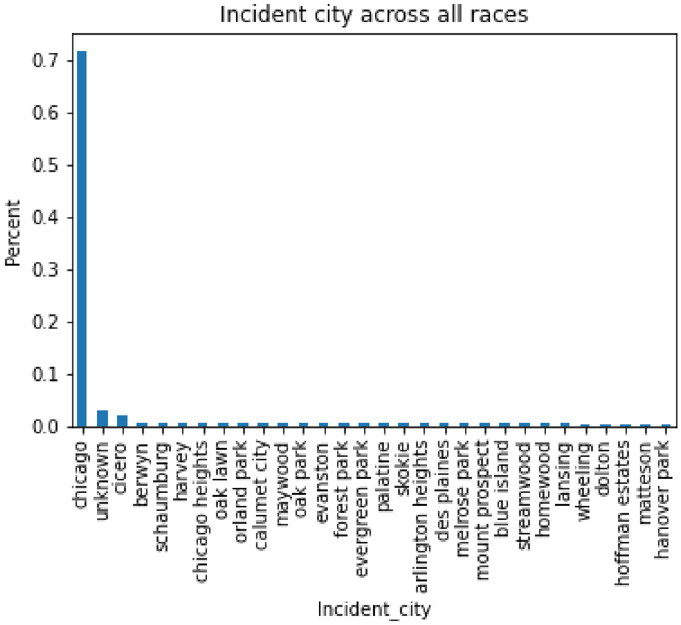


Figure 10: Incident_city across all races

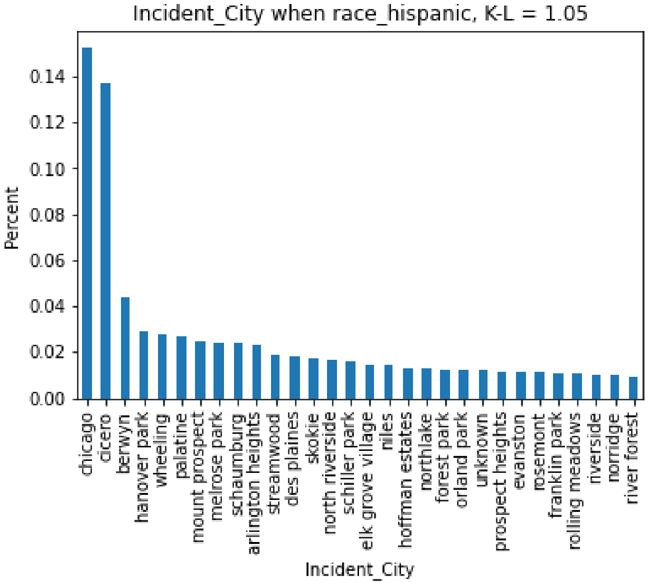
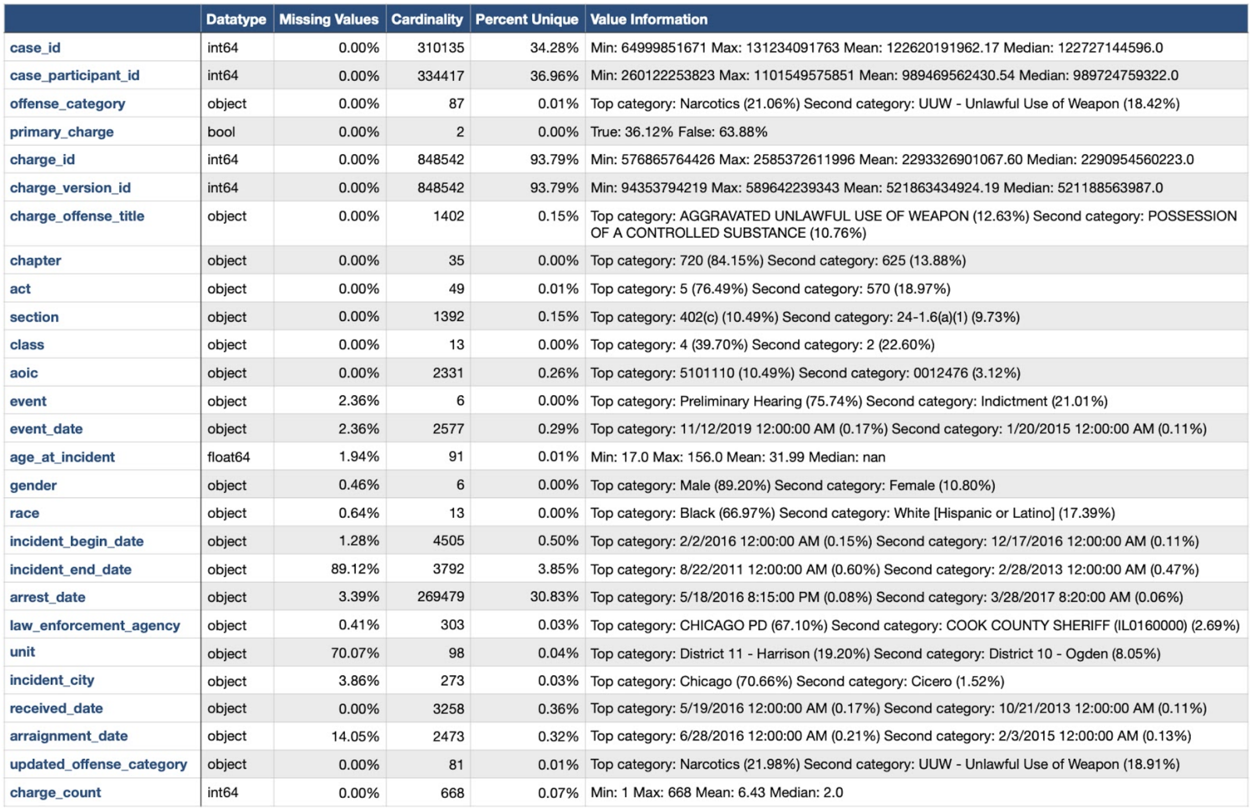
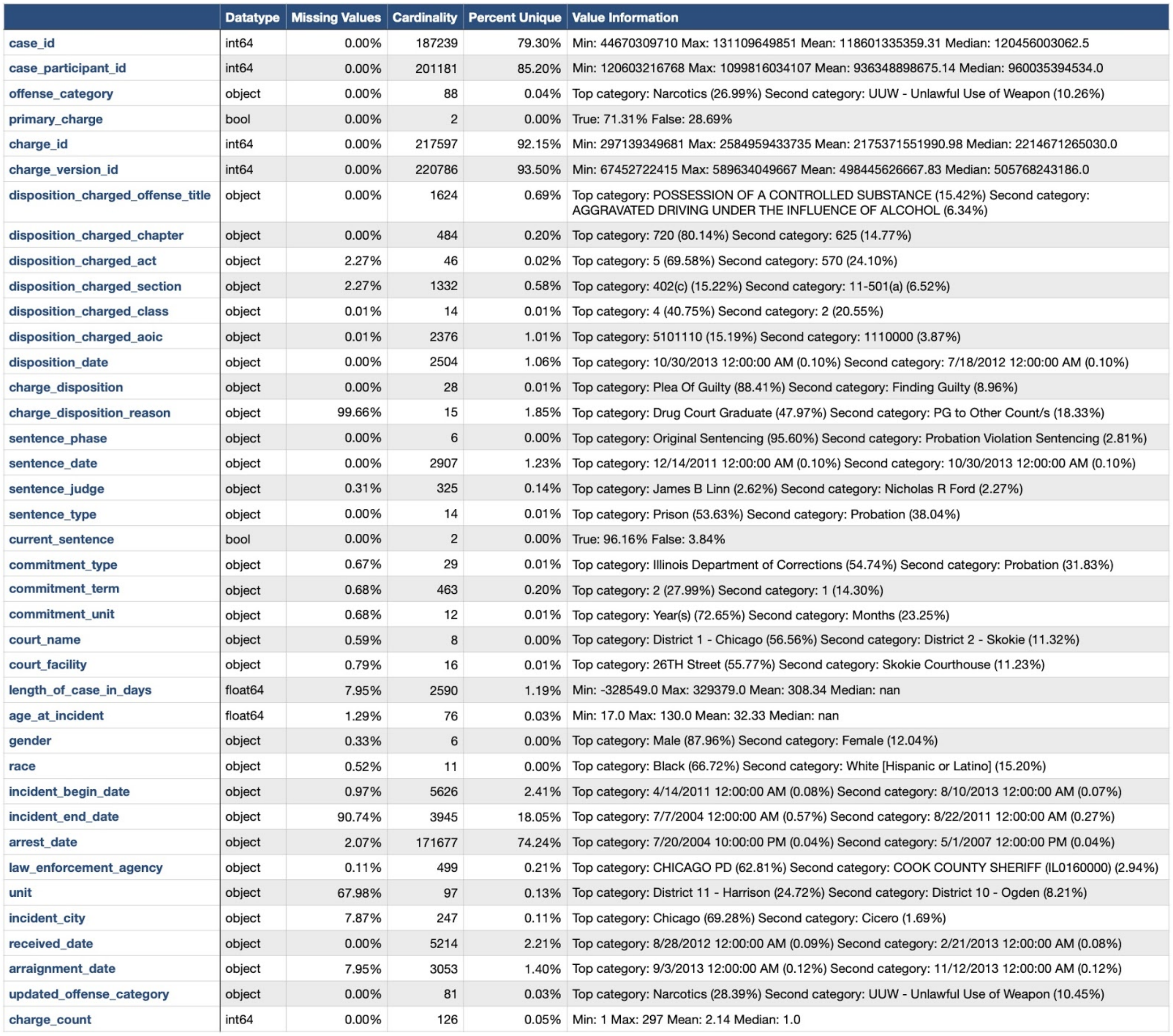


Figure 11: Incident_city when Race = Hispanic



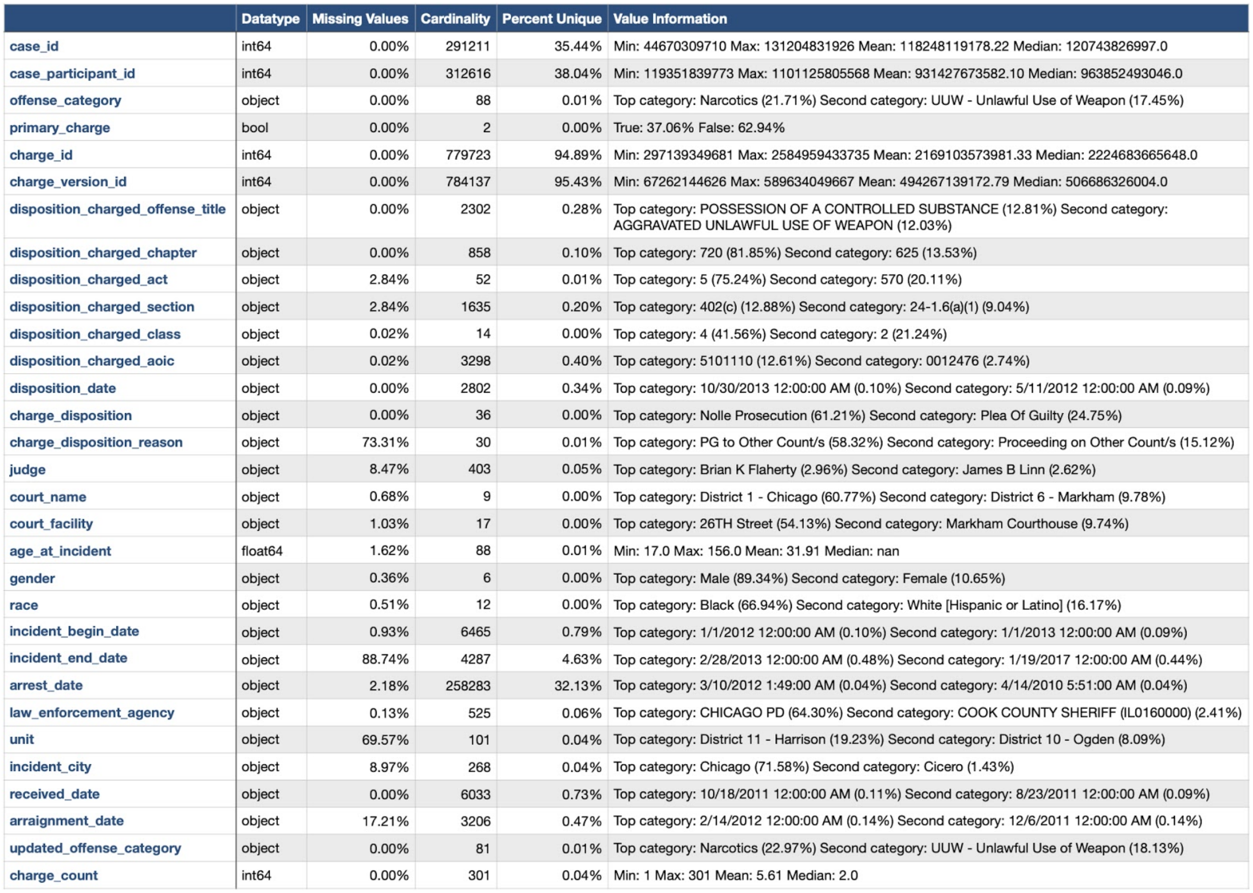
	Datatype	Missing Values	Cardinality	Percent Unique	Value Information
case_id	int64	0.00%	310135	34.28%	Min: 64999851671 Max: 131234091763 Mean: 122620191962.17 Median: 122727144596.0
case_participant_id	int64	0.00%	334417	36.96%	Min: 260122253823 Max: 1101549575851 Mean: 989469562430.54 Median: 989724759322.0
offense_category	object	0.00%	87	0.01%	Top category: Narcotics (21.06%) Second category: UUW - Unlawful Use of Weapon (18.42%)
primary_charge	bool	0.00%	2	0.00%	True: 36.12% False: 63.88%
charge_id	int64	0.00%	848542	93.79%	Min: 576865764426 Max: 2585372611996 Mean: 2293326901067.60 Median: 2290954560223.0
charge_version_id	int64	0.00%	848542	93.79%	Min: 94353794219 Max: 589642239343 Mean: 521863434924.19 Median: 521188563987.0
charge_offense_title	object	0.00%	1402	0.15%	Top category: AGGRAVATED UNLAWFUL USE OF WEAPON (12.63%) Second category: POSSESSION OF A CONTROLLED SUBSTANCE (10.76%)
chapter	object	0.00%	35	0.00%	Top category: 720 (84.15%) Second category: 625 (13.88%)
act	object	0.00%	49	0.01%	Top category: 5 (76.49%) Second category: 570 (18.97%)
section	object	0.00%	1392	0.15%	Top category: 402(c) (10.49%) Second category: 24-1.6(a)(1) (9.73%)
class	object	0.00%	13	0.00%	Top category: 4 (39.70%) Second category: 2 (22.60%)
aoic	object	0.00%	2331	0.26%	Top category: 5101110 (10.49%) Second category: 0012476 (3.12%)
event	object	2.36%	6	0.00%	Top category: Preliminary Hearing (75.74%) Second category: Indictment (21.01%)
event_date	object	2.36%	2577	0.29%	Top category: 11/12/2019 12:00:00 AM (0.17%) Second category: 1/20/2015 12:00:00 AM (0.11%)
age_at_incident	float64	1.94%	91	0.01%	Min: 17.0 Max: 156.0 Mean: 31.99 Median: nan
gender	object	0.46%	6	0.00%	Top category: Male (89.20%) Second category: Female (10.80%)
race	object	0.64%	13	0.00%	Top category: Black (66.97%) Second category: White [Hispanic or Latino] (17.39%)
incident_begin_date	object	1.28%	4505	0.50%	Top category: 2/2/2016 12:00:00 AM (0.15%) Second category: 12/17/2016 12:00:00 AM (0.11%)
incident_end_date	object	89.12%	3792	3.85%	Top category: 8/22/2011 12:00:00 AM (0.60%) Second category: 2/28/2013 12:00:00 AM (0.47%)
arrest_date	object	3.39%	269479	30.83%	Top category: 5/18/2016 8:15:00 PM (0.08%) Second category: 3/28/2017 8:20:00 AM (0.06%)
law_enforcement_agency	object	0.41%	303	0.03%	Top category: CHICAGO PD (67.10%) Second category: COOK COUNTY SHERIFF (IL0160000) (2.69%)
unit	object	70.07%	98	0.04%	Top category: District 11 - Harrison (19.20%) Second category: District 10 - Ogden (8.05%)
incident_city	object	3.86%	273	0.03%	Top category: Chicago (70.66%) Second category: Cicero (1.52%)
received_date	object	0.00%	3258	0.36%	Top category: 5/19/2016 12:00:00 AM (0.17%) Second category: 10/21/2013 12:00:00 AM (0.11%)
arraignment_date	object	14.05%	2473	0.32%	Top category: 6/28/2016 12:00:00 AM (0.21%) Second category: 2/3/2015 12:00:00 AM (0.13%)
updated_offense_category	object	0.00%	81	0.01%	Top category: Narcotics (21.98%) Second category: UUW - Unlawful Use of Weapon (18.91%)
charge_count	int64	0.00%	668	0.07%	Min: 1 Max: 668 Mean: 6.43 Median: 2.0

Table 1: Profiling of the Initiation dataset



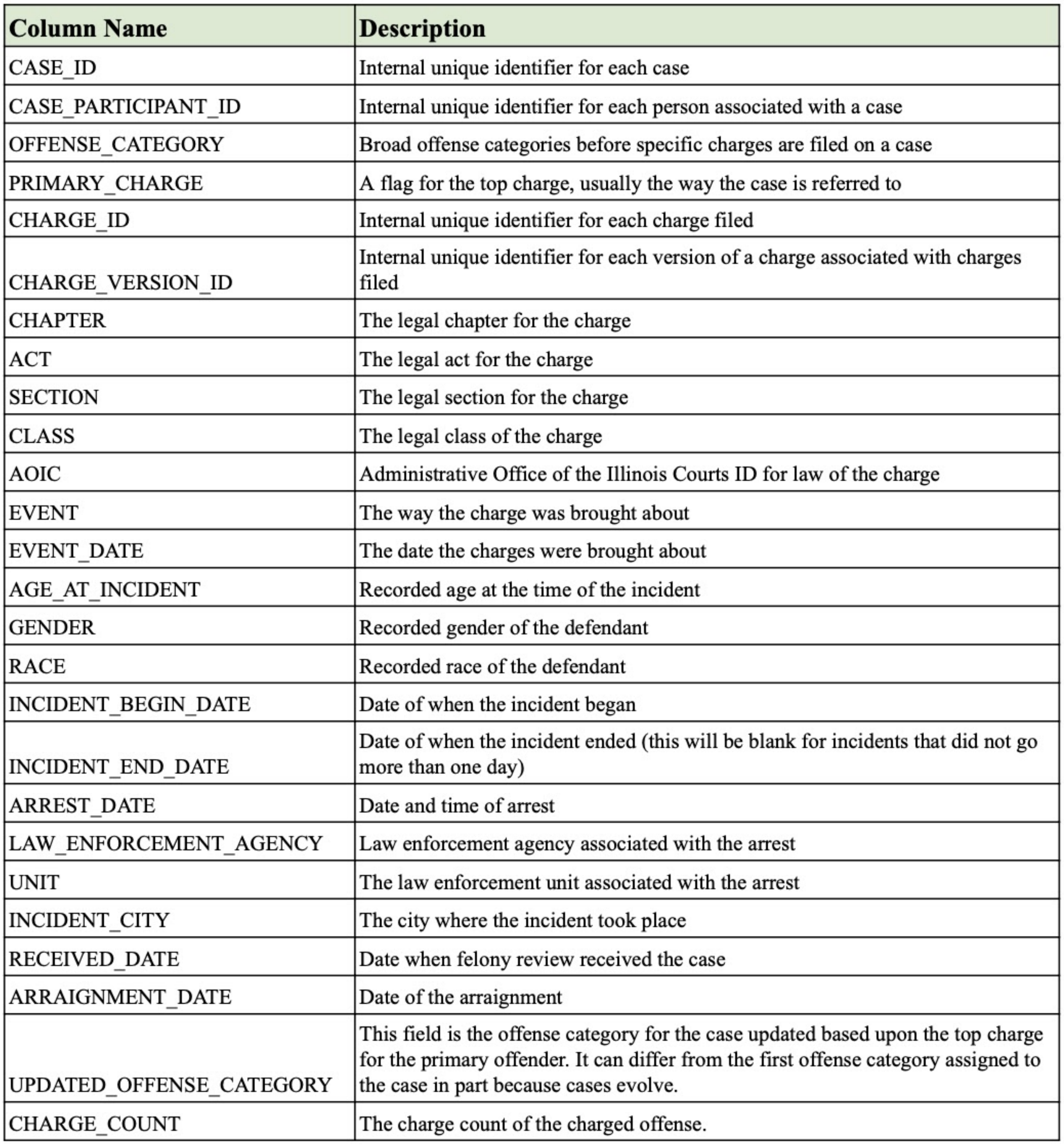
	Datatype	Missing Values	Cardinality	Percent Unique	Value Information
case_id	int64	0.00%	187239	79.30%	Min: 44670309710 Max: 131109649851 Mean: 118601335359.31 Median: 120456003062.5
case_participant_id	int64	0.00%	201181	85.20%	Min: 120603216768 Max: 1099816034107 Mean: 936348898675.14 Median: 960035394534.0
offense_category	object	0.00%	88	0.04%	Top category: Narcotics (26.99%) Second category: UUW - Unlawful Use of Weapon (10.26%)
primary_charge	bool	0.00%	2	0.00%	True: 71.31% False: 28.69%
charge_id	int64	0.00%	217597	92.15%	Min: 297139349681 Max: 2584959433735 Mean: 2175371551990.98 Median: 2214671265030.0
charge_version_id	int64	0.00%	220786	93.50%	Min: 67452722415 Max: 589634049667 Mean: 498445626667.83 Median: 505768243186.0
disposition_charged_offense_title	object	0.00%	1624	0.69%	Top category: POSSESSION OF A CONTROLLED SUBSTANCE (15.42%) Second category: AGGRAVATED DRIVING UNDER THE INFLUENCE OF ALCOHOL (6.34%)
disposition_charged_chapter	object	0.00%	484	0.20%	Top category: 720 (80.14%) Second category: 625 (14.77%)
disposition_charged_act	object	2.27%	46	0.02%	Top category: 5 (69.58%) Second category: 570 (24.10%)
disposition_charged_section	object	2.27%	1332	0.58%	Top category: 402(c) (15.22%) Second category: 11-501(a) (6.52%)
disposition_charged_class	object	0.01%	14	0.01%	Top category: 4 (40.75%) Second category: 2 (20.55%)
disposition_charged_aoic	object	0.01%	2376	1.01%	Top category: 5101110 (15.19%) Second category: 1110000 (3.87%)
disposition_date	object	0.00%	2504	1.06%	Top category: 10/30/2013 12:00:00 AM (0.10%) Second category: 7/18/2012 12:00:00 AM (0.10%)
charge_disposition	object	0.00%	28	0.01%	Top category: Plea Of Guilty (88.41%) Second category: Finding Guilty (8.96%)
charge_disposition_reason	object	99.66%	15	1.85%	Top category: Drug Court Graduate (47.97%) Second category: PG to Other Count/s (18.33%)
sentence_phase	object	0.00%	6	0.00%	Top category: Original Sentencing (95.60%) Second category: Probation Violation Sentencing (2.81%)
sentence_date	object	0.00%	2907	1.23%	Top category: 12/14/2011 12:00:00 AM (0.10%) Second category: 10/30/2013 12:00:00 AM (0.10%)
sentence_judge	object	0.31%	325	0.14%	Top category: James B Linn (2.62%) Second category: Nicholas R Ford (2.27%)
sentence_type	object	0.00%	14	0.01%	Top category: Prison (53.63%) Second category: Probation (38.04%)
current_sentence	bool	0.00%	2	0.00%	True: 96.16% False: 3.84%
commitment_type	object	0.67%	29	0.01%	Top category: Illinois Department of Corrections (54.74%) Second category: Probation (31.83%)
commitment_term	object	0.68%	463	0.20%	Top category: 2 (27.99%) Second category: 1 (14.30%)
commitment_unit	object	0.68%	12	0.01%	Top category: Year(s) (72.65%) Second category: Months (23.25%)
court_name	object	0.59%	8	0.00%	Top category: District 1 - Chicago (56.56%) Second category: District 2 - Skokie (11.32%)
court_facility	object	0.79%	16	0.01%	Top category: 26TH Street (55.77%) Second category: Skokie Courthouse (11.23%)
length_of_case_in_days	float64	7.95%	2590	1.19%	Min: -328549.0 Max: 329379.0 Mean: 308.34 Median: nan
age_at_incident	float64	1.29%	76	0.03%	Min: 17.0 Max: 130.0 Mean: 32.33 Median: nan
gender	object	0.33%	6	0.00%	Top category: Male (87.96%) Second category: Female (12.04%)
race	object	0.52%	11	0.00%	Top category: Black (66.72%) Second category: White [Hispanic or Latino] (15.20%)
incident_begin_date	object	0.97%	5626	2.41%	Top category: 4/14/2011 12:00:00 AM (0.08%) Second category: 8/10/2013 12:00:00 AM (0.07%)
incident_end_date	object	90.74%	3945	18.05%	Top category: 7/7/2004 12:00:00 AM (0.57%) Second category: 8/22/2011 12:00:00 AM (0.27%)
arrest_date	object	2.07%	171677	74.24%	Top category: 7/20/2004 10:00:00 PM (0.04%) Second category: 5/1/2007 12:00:00 PM (0.04%)
law_enforcement_agency	object	0.11%	499	0.21%	Top category: CHICAGO PD (62.81%) Second category: COOK COUNTY SHERIFF (IL0160000) (2.94%)
unit	object	67.98%	97	0.13%	Top category: District 11 - Harrison (24.72%) Second category: District 10 - Ogden (8.21%)
incident_city	object	7.87%	247	0.11%	Top category: Chicago (69.28%) Second category: Cicero (1.69%)
received_date	object	0.00%	5214	2.21%	Top category: 8/28/2012 12:00:00 AM (0.09%) Second category: 2/21/2013 12:00:00 AM (0.08%)
arraignment_date	object	7.95%	3053	1.40%	Top category: 9/3/2013 12:00:00 AM (0.12%) Second category: 11/12/2013 12:00:00 AM (0.12%)
updated_offense_category	object	0.00%	81	0.03%	Top category: Narcotics (28.39%) Second category: UUW - Unlawful Use of Weapon (10.45%)
charge_count	int64	0.00%	126	0.05%	Min: 1 Max: 297 Mean: 2.14 Median: 1.0

Table 2: Profiling of the Sentencing dataset



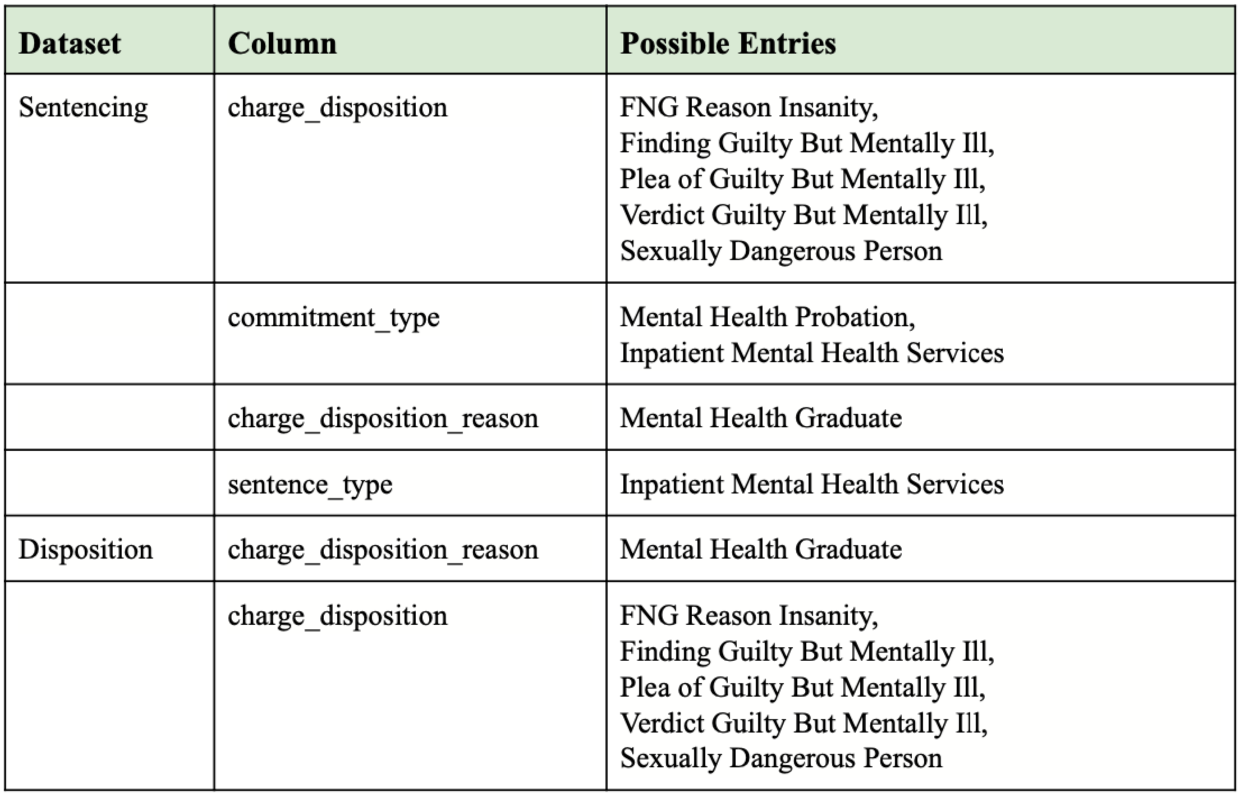
	Datatype	Missing Values	Cardinality	Percent Unique	Value Information
case_id	int64	0.00%	291211	35.44%	Min: 44670309710 Max: 131204831926 Mean: 118248119178.22 Median: 120743826997.0
case_participant_id	int64	0.00%	312616	38.04%	Min: 119351839773 Max: 1101125805568 Mean: 931427673582.10 Median: 963852493046.0
offense_category	object	0.00%	88	0.01%	Top category: Narcotics (21.71%) Second category: UUW - Unlawful Use of Weapon (17.45%)
primary_charge	bool	0.00%	2	0.00%	True: 37.06% False: 62.94%
charge_id	int64	0.00%	779723	94.89%	Min: 297139349681 Max: 2584959433735 Mean: 2169103573981.33 Median: 2224683665648.0
charge_version_id	int64	0.00%	784137	95.43%	Min: 67262144626 Max: 589634049667 Mean: 494267139172.79 Median: 506686326004.0
disposition_charged_offense_title	object	0.00%	2302	0.28%	Top category: POSSESSION OF A CONTROLLED SUBSTANCE (12.81%) Second category: AGGRAVATED UNLAWFUL USE OF WEAPON (12.03%)
disposition_charged_chapter	object	0.00%	858	0.10%	Top category: 720 (81.85%) Second category: 625 (13.53%)
disposition_charged_act	object	2.84%	52	0.01%	Top category: 5 (75.24%) Second category: 570 (20.11%)
disposition_charged_section	object	2.84%	1635	0.20%	Top category: 402(c) (12.88%) Second category: 24-1.6(a)(1) (9.04%)
disposition_charged_class	object	0.02%	14	0.00%	Top category: 4 (41.56%) Second category: 2 (21.24%)
disposition_charged_aoic	object	0.02%	3298	0.40%	Top category: 5101110 (12.61%) Second category: 0012476 (2.74%)
disposition_date	object	0.00%	2802	0.34%	Top category: 10/30/2013 12:00:00 AM (0.10%) Second category: 5/11/2012 12:00:00 AM (0.09%)
charge_disposition	object	0.00%	36	0.00%	Top category: Nolle Prosecution (61.21%) Second category: Plea Of Guilty (24.75%)
charge_disposition_reason	object	73.31%	30	0.01%	Top category: PG to Other Count/s (58.32%) Second category: Proceeding on Other Count/s (15.12%)
judge	object	8.47%	403	0.05%	Top category: Brian K Flaherty (2.96%) Second category: James B Linn (2.62%)
court_name	object	0.68%	9	0.00%	Top category: District 1 - Chicago (60.77%) Second category: District 6 - Markham (9.78%)
court_facility	object	1.03%	17	0.00%	Top category: 26TH Street (54.13%) Second category: Markham Courthouse (9.74%)
age_at_incident	float64	1.62%	88	0.01%	Min: 17.0 Max: 156.0 Mean: 31.91 Median: nan
gender	object	0.36%	6	0.00%	Top category: Male (89.34%) Second category: Female (10.65%)
race	object	0.51%	12	0.00%	Top category: Black (66.94%) Second category: White [Hispanic or Latino] (16.17%)
incident_begin_date	object	0.93%	6465	0.79%	Top category: 1/1/2012 12:00:00 AM (0.10%) Second category: 1/1/2013 12:00:00 AM (0.09%)
incident_end_date	object	88.74%	4287	4.63%	Top category: 2/28/2013 12:00:00 AM (0.48%) Second category: 1/19/2017 12:00:00 AM (0.44%)
arrest_date	object	2.18%	258283	32.13%	Top category: 3/10/2012 1:49:00 AM (0.04%) Second category: 4/14/2010 5:51:00 AM (0.04%)
law_enforcement_agency	object	0.13%	525	0.06%	Top category: CHICAGO PD (64.30%) Second category: COOK COUNTY SHERIFF (IL0160000) (2.41%)
unit	object	69.57%	101	0.04%	Top category: District 11 - Harrison (19.23%) Second category: District 10 - Ogden (8.09%)
incident_city	object	8.97%	268	0.04%	Top category: Chicago (71.58%) Second category: Cicero (1.43%)
received_date	object	0.00%	6033	0.73%	Top category: 10/18/2011 12:00:00 AM (0.11%) Second category: 8/23/2011 12:00:00 AM (0.09%)
arraignment_date	object	17.21%	3206	0.47%	Top category: 2/14/2012 12:00:00 AM (0.14%) Second category: 12/6/2011 12:00:00 AM (0.14%)
updated_offense_category	object	0.00%	81	0.01%	Top category: Narcotics (22.97%) Second category: UUW - Unlawful Use of Weapon (18.13%)
charge_count	int64	0.00%	301	0.04%	Min: 1 Max: 301 Mean: 5.61 Median: 2.0

Table 3: Profiling of the Disposition dataset



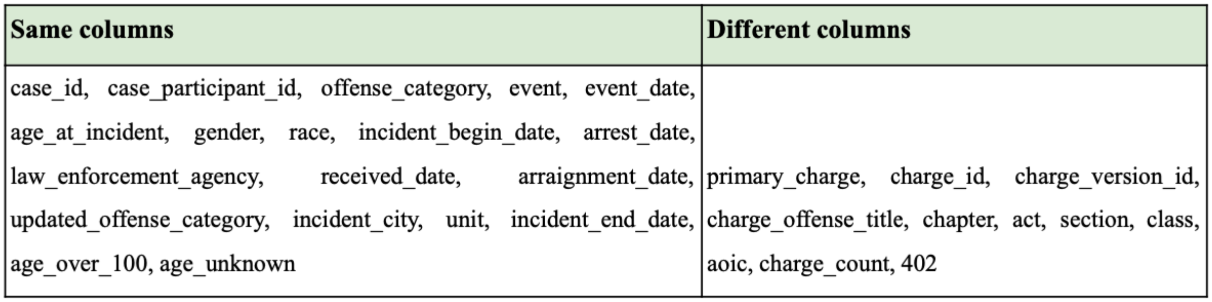
Column Name	Description
CASE_ID	Internal unique identifier for each case
CASE_PARTICIPANT_ID	Internal unique identifier for each person associated with a case
OFFENSE_CATEGORY	Broad offense categories before specific charges are filed on a case
PRIMARY_CHARGE	A flag for the top charge, usually the way the case is referred to
CHARGE_ID	Internal unique identifier for each charge filed
CHARGE_VERSION_ID	Internal unique identifier for each version of a charge associated with charges filed
CHAPTER	The legal chapter for the charge
ACT	The legal act for the charge
SECTION	The legal section for the charge
CLASS	The legal class of the charge
AOIC	Administrative Office of the Illinois Courts ID for law of the charge
EVENT	The way the charge was brought about
EVENT_DATE	The date the charges were brought about
AGE_AT_INCIDENT	Recorded age at the time of the incident
GENDER	Recorded gender of the defendant
RACE	Recorded race of the defendant
INCIDENT_BEGIN_DATE	Date of when the incident began
INCIDENT_END_DATE	Date of when the incident ended (this will be blank for incidents that did not go more than one day)
ARREST_DATE	Date and time of arrest
LAW_ENFORCEMENT_AGENCY	Law enforcement agency associated with the arrest
UNIT	The law enforcement unit associated with the arrest
INCIDENT_CITY	The city where the incident took place
RECEIVED_DATE	Date when felony review received the case
ARRAIGNMENT_DATE	Date of the arraignment
UPDATED_OFFENSE_CATEGORY	This field is the offense category for the case updated based upon the top charge for the primary offender. It can differ from the first offense category assigned to the case in part because cases evolve.
CHARGE_COUNT	The charge count of the charged offense.

Table 4: Cook County Data Portal’s description of attributes in Initiation dataset [citation: 1]



Dataset	Column	Possible Entries
Sentencing	charge_disposition	FNG Reason Insanity, Finding Guilty But Mentally Ill, Plea of Guilty But Mentally Ill, Verdict Guilty But Mentally Ill, Sexually Dangerous Person
	commitment_type	Mental Health Probation, Inpatient Mental Health Services
	charge_disposition_reason	Mental Health Graduate
	sentence_type	Inpatient Mental Health Services
Disposition	charge_disposition_reason	Mental Health Graduate
	charge_disposition	FNG Reason Insanity, Finding Guilty But Mentally Ill, Plea of Guilty But Mentally Ill, Verdict Guilty But Mentally Ill, Sexually Dangerous Person

Table 5: Columns used for the assignment of MHI



Same columns	Different columns
case_id, case_participant_id, offense_category, event, event_date, age_at_incident, gender, race, incident_begin_date, arrest_date, law_enforcement_agency, received_date, arraignment_date, updated_offense_category, incident_city, unit, incident_end_date, age_over_100, age_unknown	primary_charge, charge_id, charge_version_id, charge_offense_title, chapter, act, section, class, aoic, charge_count, 402

Table 6: Features that remained the same (left) and varied (right) during aggregation by case_participant_id

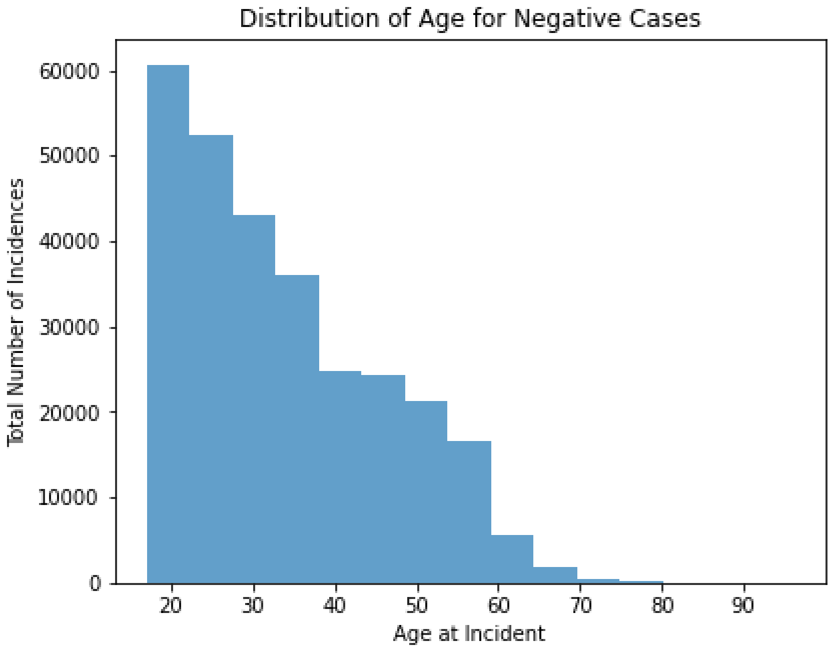


Figure 12: Distribution of age in negative class, over entire dataset.

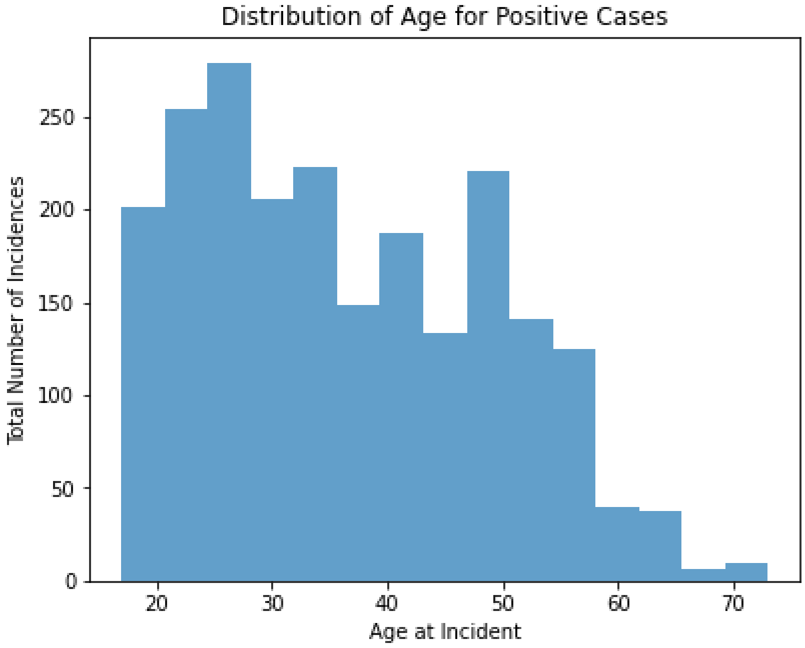


Figure 13: Distribution of age in positive class, over entire dataset.

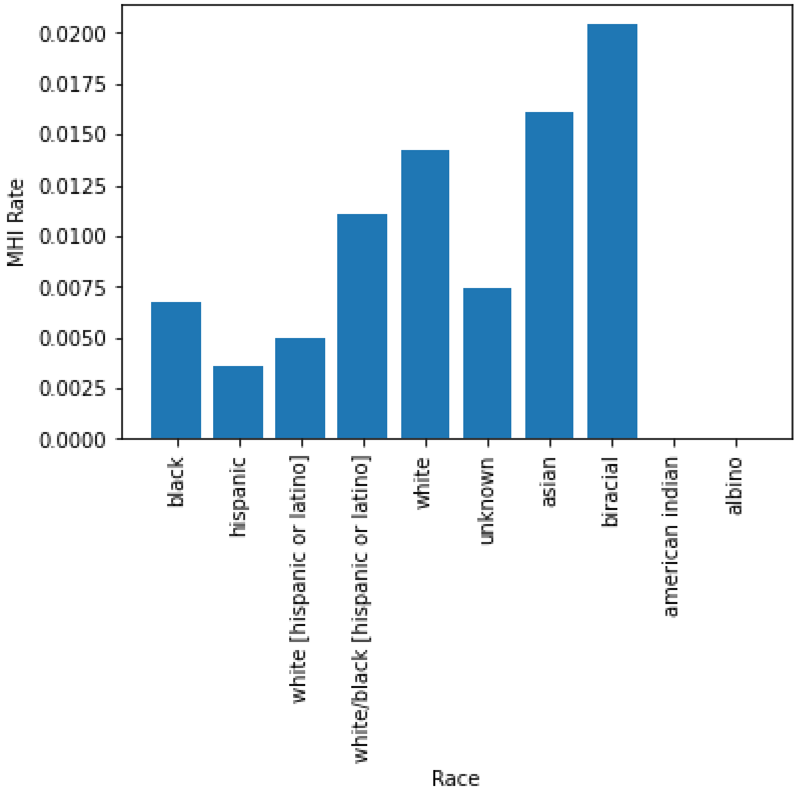


Figure 14: Base rate for racial subpopulations in entire dataset.

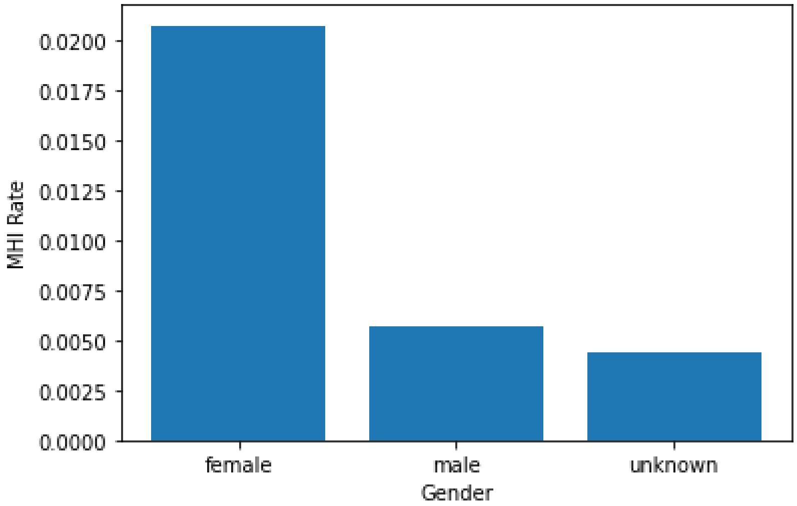


Figure 15: Base rate for gender subpopulations in entire dataset.

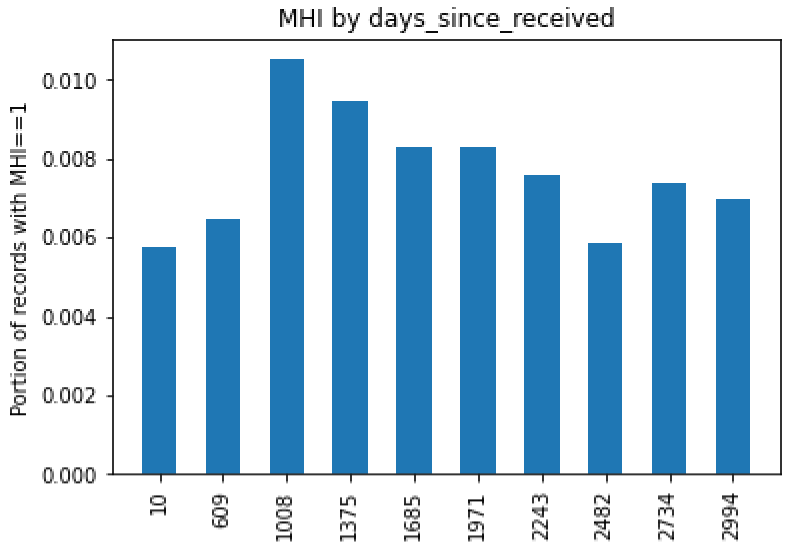


Figure 16: Base rates distribution over age of instance in entire dataset.

Training base rate: 0.007684097130940968
Validation base rate: 0.009155059496356425
Training+Validation base rate: 0.007943678020591706
Test base rate: 0.0059958951179577055

Convert to Figure 17: Base rates in test, validation, and training sets.

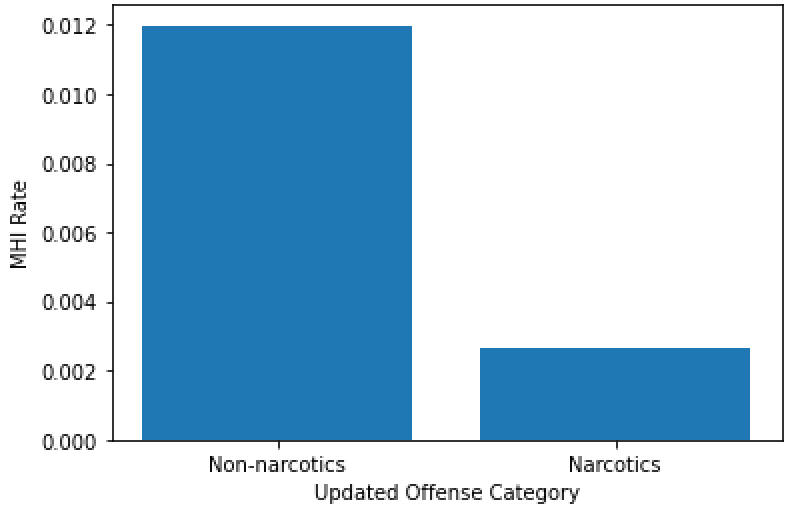
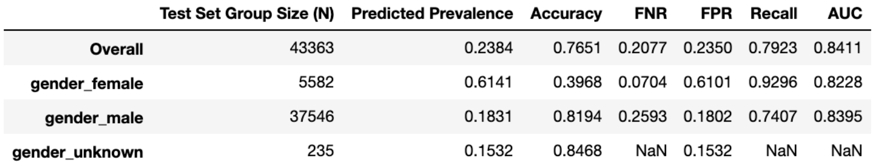
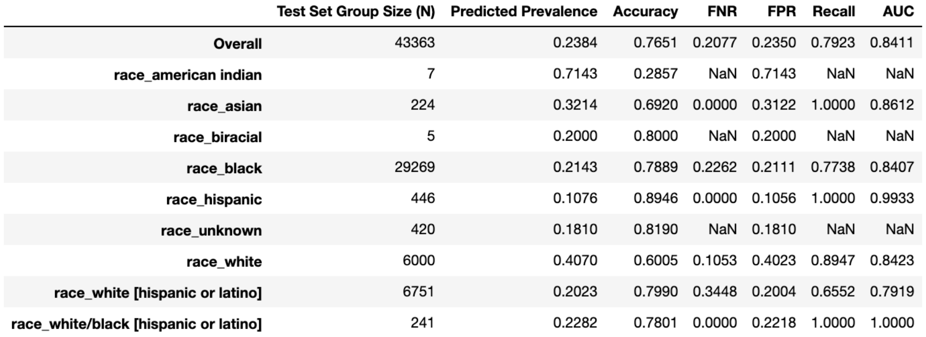


Figure 18: Base rate distribution in entire dataset, based on presence of narcotics in updated_offense_category



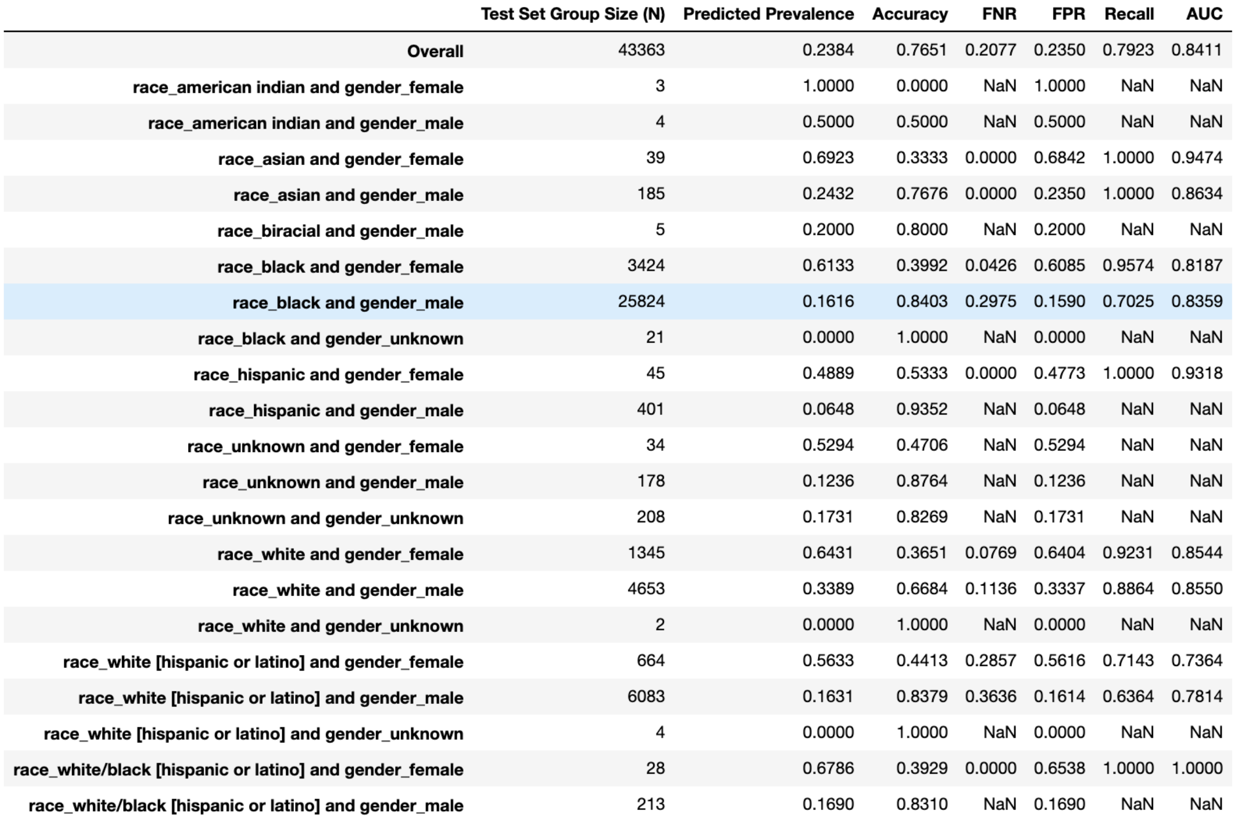
	Test Set Group Size (N)	Predicted Prevalence	Accuracy	FNR	FPR	Recall	AUC
Overall	43363	0.2384	0.7651	0.2077	0.2350	0.7923	0.8411
gender_female	5582	0.6141	0.3968	0.0704	0.6101	0.9296	0.8228
gender_male	37546	0.1831	0.8194	0.2593	0.1802	0.7407	0.8395
gender_unknown	235	0.1532	0.8468	NaN	0.1532	NaN	NaN

Table 7: Fairness metrics by gender subgroups



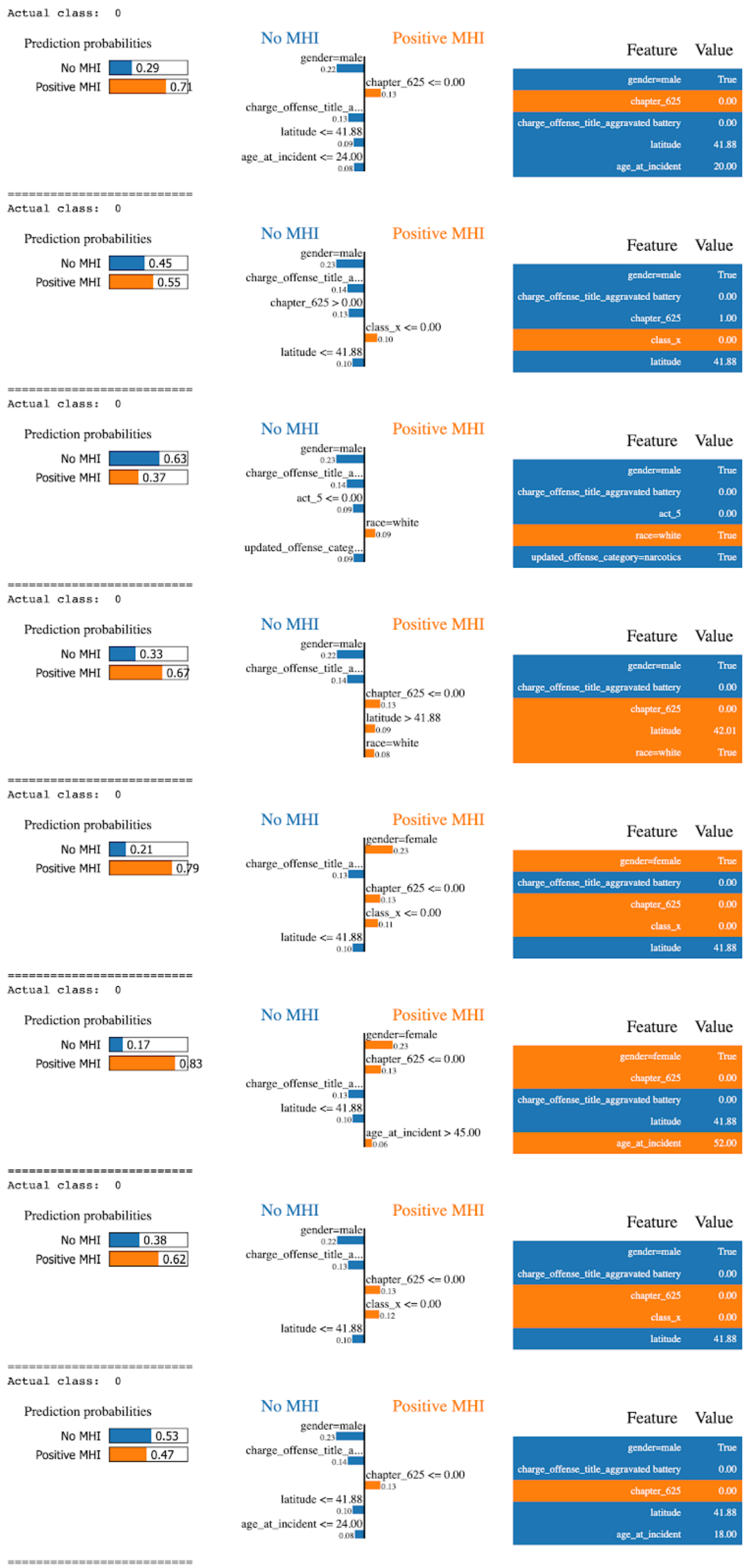
	Test Set Group Size (N)	Predicted Prevalence	Accuracy	FNR	FPR	Recall	AUC
Overall	43363	0.2384	0.7651	0.2077	0.2350	0.7923	0.8411
race_american indian	7	0.7143	0.2857	NaN	0.7143	NaN	NaN
race_asian	224	0.3214	0.6920	0.0000	0.3122	1.0000	0.8612
race_biracial	5	0.2000	0.8000	NaN	0.2000	NaN	NaN
race_black	29269	0.2143	0.7889	0.2262	0.2111	0.7738	0.8407
race_hispanic	446	0.1076	0.8946	0.0000	0.1056	1.0000	0.9933
race_unknown	420	0.1810	0.8190	NaN	0.1810	NaN	NaN
race_white	6000	0.4070	0.6005	0.1053	0.4023	0.8947	0.8423
race_white [hispanic or latino]	6751	0.2023	0.7990	0.3448	0.2004	0.6552	0.7919
race_white/black [hispanic or latino]	241	0.2282	0.7801	0.0000	0.2218	1.0000	1.0000

Table 8: Fairness metrics by racial subgroups



	Test Set Group Size (N)	Predicted Prevalence	Accuracy	FNR	FPR	Recall	AUC
Overall	43363	0.2384	0.7651	0.2077	0.2350	0.7923	0.8411
race_american indian and gender_female	3	1.0000	0.0000	NaN	1.0000	NaN	NaN
race_american indian and gender_male	4	0.5000	0.5000	NaN	0.5000	NaN	NaN
race_asian and gender_female	39	0.6923	0.3333	0.0000	0.6842	1.0000	0.9474
race_asian and gender_male	185	0.2432	0.7676	0.0000	0.2350	1.0000	0.8634
race_biracial and gender_male	5	0.2000	0.8000	NaN	0.2000	NaN	NaN
race_black and gender_female	3424	0.6133	0.3992	0.0426	0.6085	0.9574	0.8187
race_black and gender_male	25824	0.1616	0.8403	0.2975	0.1590	0.7025	0.8359
race_black and gender_unknown	21	0.0000	1.0000	NaN	0.0000	NaN	NaN
race_hispanic and gender_female	45	0.4889	0.5333	0.0000	0.4773	1.0000	0.9318
race_hispanic and gender_male	401	0.0648	0.9352	NaN	0.0648	NaN	NaN
race_unknown and gender_female	34	0.5294	0.4706	NaN	0.5294	NaN	NaN
race_unknown and gender_male	178	0.1236	0.8764	NaN	0.1236	NaN	NaN
race_unknown and gender_unknown	208	0.1731	0.8269	NaN	0.1731	NaN	NaN
race_white and gender_female	1345	0.6431	0.3651	0.0769	0.6404	0.9231	0.8544
race_white and gender_male	4653	0.3389	0.6684	0.1136	0.3337	0.8864	0.8550
race_white and gender_unknown	2	0.0000	1.0000	NaN	0.0000	NaN	NaN
race_white [hispanic or latino] and gender_female	664	0.5633	0.4413	0.2857	0.5616	0.7143	0.7364
race_white [hispanic or latino] and gender_male	6083	0.1631	0.8379	0.3636	0.1614	0.6364	0.7814
race_white [hispanic or latino] and gender_unknown	4	0.0000	1.0000	NaN	0.0000	NaN	NaN
race_white/black [hispanic or latino] and gender_female	28	0.6786	0.3929	0.0000	0.6538	1.0000	1.0000
race_white/black [hispanic or latino] and gender_male	213	0.1690	0.8310	NaN	0.1690	NaN	NaN

Table 9: Fairness metrics, intersectional racial/gender subgroups



```
Actual class:  0
```

Feature	Value
gender=male	True
chapter_625	0.00
charge_offense_title_aggravated battery	0.00
latitude	41.88
age_at_incident	20.00

```
=========================
Actual class:  0
```

Feature	Value
gender=male	True
charge_offense_title_aggravated battery	0.00
chapter_625	1.00
class_x	0.00
latitude	41.88

```
=========================
Actual class:  0
```

Feature	Value
gender=male	True
charge_offense_title_aggravated battery	0.00
act_5	0.00
race=white	True
updated_offense_category=narcotics	True

```
=========================
Actual class:  0
```

Feature	Value
gender=male	True
charge_offense_title_aggravated battery	0.00
chapter_625	0.00
latitude	42.01
race=white	True

```
=========================
Actual class:  0
```

Feature	Value
gender=female	True
charge_offense_title_aggravated battery	0.00
chapter_625	0.00
class_x	0.00
latitude	41.88

```
=========================
Actual class:  0
```

Feature	Value
gender=female	True
chapter_625	0.00
charge_offense_title_aggravated battery	0.00
latitude	41.88
age_at_incident	52.00

```
=========================
Actual class:  0
```

Feature	Value
gender=male	True
charge_offense_title_aggravated battery	0.00
chapter_625	0.00
class_x	0.00
latitude	41.88

```
=========================
Actual class:  0
```

Feature	Value
gender=male	True
charge_offense_title_aggravated battery	0.00
chapter_625	0.00
latitude	41.88
age_at_incident	18.00

```
=========================
```

Figure 19: Local explanations chosen by LIME’s submodular picker

Works Cited

1. Behavioral Health Innovations. Mental Health and Justice in Cook County Bond Courts An Examination of the Management of Persons with Mental Illness in Felony Bond Court. Report prepared for the Administrative Office of the Illinois Courts, July 2015.
2. Markey, Rhea, Hutchinson, and Teng. (2019). Predicting Mental Health-Related Dispositions and Sentences from Cook County Court Data. https://github.com/kelseymarkey/cook-county-mental-health-prediction/blob/master/FINAL%20PAPER.pdf
3. Braude, L., & Alaimo, C. (2007). A large court system tackles a huge problem: stakeholders in an Illinois county work toward better outcomes for mentally ill offenders. Behavioral healthcare, 27(3), 41-44. https://www.psychcongress.com/article/large-court-system-tackles-huge-problem
4. Mueller, Heiko. (2019). Data Profiling & Data Cleaning. https://dataresponsibly.github.io/courses/documents/spring20/Lecture3.pdf
5. Tsirigotis, K., & Łuczak, J. (2018). Resilience in women who experience domestic violence. Psychiatric quarterly, 89(1), 201-211. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5807488/.