Mental Health Prediction in Cook County Courts: Nutritional Label for an Automated Decision System

NYU Responsible Data Science, Spring 2020
Kelsey Markey (kcm312), Alene Rhea (akr435)

1 Background

This report analyzes 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 was moved to GitHub [1] in January 2020, at which point the data cleaning and model tuning notebooks were adapted slightly to ensure further reproducibility.

The project aimed to create an automated decision system (ADS) 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.

The ADS uses data from the December 2, 2019 updates to the Initiation, Disposition, and Sentencing datasets on the Cook County Open Data Portal.[2] Because the Cook County Criminal Justice System is at the forefront of the movement to use specialty treatment courts and programs to address mental illness,[3] the Sentencing and Dispositions datasets contain outcome information which could be used as proxies to predict mental illness. 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 (Figure 1). The Initiation dataset is used for model prediction and includes felony cases handled by the Criminal, Narcotics, and Special Prosecution Bureaus.[4]

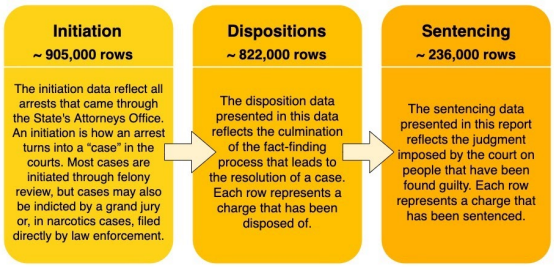


Figure 1: Descriptions of Cook County legal datasets used in this study.[5]

2 Input and Output

The business rules governing the datasets are difficult to decipher, and demand highly specific domain expertise. The authors directed questions to representatives of the Cook County Open Data Portal, the Cook County States Attorney's Office, the County Clerk, and the Clerk of the Circuit Courts. None of these inquiries received a response.

2.1 Candidate Keys and IDs

There are four 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, each of these is an "internal unique identifier": CASE_ID for each case, CASE_PARTICIPANT_ID for each person associated with a case, CHARGE_ID for each charge filed, and CHARGE_VERSION_ID for each version of a charge associated with charges filed.[2]

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 or sentence version, 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 eight features (Section 2, Jupyter Notebook). (Note that only IDs and sentencing-specific features were considered for this task.) Two such eight-feature candidate keys exist, and seven of their features are shared: CASE_PARTICIPANT_ID, COMMITMENT_TERM, COMMITMENT_TYPE, CURRENT_SENTENCE, SENTENCE_DATE, SENTENCE_TYPE, and SENT-ENCE_PHASE. To these seven 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 eight features.

2.2 Model Features

In order to replicate a use case where mental illness is predicted at initiation, only the 27 attributes in the Initiation dataset were used to construct input features to pass to the model (Figure 15). 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 features pre-and-post processing.)

2.3 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.

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

Figure 2: Columns used for the assignment of MHI.[5]

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." These proxies are identified from 15 values across six columns in the Sentencing and Disposition datasets (Figure 2). 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.

2.4 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.

2.5 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 (Figure 16, 17, and 18; Section 1, Jupyter Notebook). The "Value Information" column in these tables reports: 1) standard statistical measures for numeric variables, 2) the percent true and false for boolean variables, 3) the top two categories or ids and their normalized value_counts for categorical and id variables.

Missing value heatmaps were also created to visualize the correlation matrix for all features with missing values (Figure 19; Section 9, Jupyter Notebook). These help to confirm some expected dependencies between missing features such as the perfect positive correlation between missing EVENT and EVENT_DATE as well as missing ACT and SECTION (the legal act and legal section for a charge). Interestingly, RACE and GENDER are often missing together (correlation = 0.8), perhaps suggesting that demographic information is either entered completely or not at all. AGE_AT_INCIDENT and INCIDENT_BEGIN_DATE are also often missing together, 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 and Data Cleaning" lecture.[6] 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. We found that in Sentencing, 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.

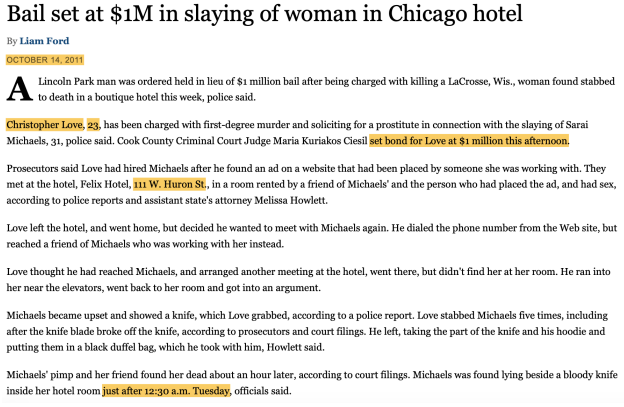
We also studied the relationship between LAW_ENFORCEMENT_AGENCY and INCIDENT_CITY (Section 5 of Jupyter Notebook). It seemed at first that there might be a functional dependency between these two features, for example when INCIDENT_CITY = "bartlett" and LAW_ENFORCEMENT_AGENCY = "bartlett pd", however 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 (e.g. LAW_ENFORCEMENT_AGENCY = "amtrak national railroad passenger corp").

High-level correlation heatmaps were created for each of the three input datasets, and include all non-categorical and non-datetime features (Figure 20). In these we see clearly the relationships between the ID variables discussed earlier. We also see feature correlations similar to those in the missing value heatmaps, such as strong correlations between ACT and CHAPTER in Initiation. 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.

CHARGE_DISPOSITION_REASON, a feature used in the construction of the target variable, is strongly correlated with many other attributes in both Sentencing and Dispositions, including COURT_FACILITY, SENTENCE_TYPE, OFFENSE_ CATEGORY, UNIT, and COMMITMENT_TYPE. This is likely related to Cook County specialized courts, since the CHARGE_DISPOSITION_REASON feature holds "additional information about the result of the charge," [7] 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."

2.6 Privacy

Defendant names are not present in the datasets and participants in a court case are instead identified with case_participant_id. Presumably, names were removed from the dataset for privacy, however we were able to quickly demonstrate that this dataset is vulnerable to linkage attacks. To do this, we selected in the Initiation dataset all cases with INCIDENT_BEGIN_DATE = "10/10/2011 12:00:00 AM" (randomly chosen) and CHARGE_OFFENSE_TITLE = "FIRST DEGREE MURDER" (chosen because we hypothesized that murder cases would be more public than other types of crimes). This query returned four rows from two different CASE_PARTICIPANT_IDs with two charges each (Figure 21).



Bail set at $1M in slaying of woman in Chicago hotel

By Liam Ford

OCTOBER 14, 2011

A Lincoln Park man was ordered held in lieu of $1 million bail after being charged with killing a LaCrosse, Wis., woman found stabbed to death in a boutique hotel this week, police said.

Christopher Love, 23, has been charged with first-degree murder and soliciting for a prostitute in connection with the slaying of Sarai Michaels, 31, police said. Cook County Criminal Court Judge Maria Kuriakos Ciesil set bond for Love at $1 million this afternoon.

Prosecutors said Love had hired Michaels after he found an ad on a website that had been placed by someone she was working with. They met at the hotel, Felix Hotel, 111 W. Huron St., in a room rented by a friend of Michaels' and the person who had placed the ad, and had sex, according to police reports and assistant state's attorney Melissa Howlett.

Love left the hotel, and went home, but decided he wanted to meet with Michaels again. He dialed the phone number from the Web site, but reached a friend of Michaels who was working with her instead.

Love thought he had reached Michaels, and arranged another meeting at the hotel, went there, but didn't find her at her room. He ran into her near the elevators, went back to her room and got into an argument.

Michaels became upset and showed a knife, which Love grabbed, according to a police report. Love stabbed Michaels five times, including after the knife blade broke off the knife, according to prosecutors and court filings. He left, taking the part of the knife and his hoodie and putting them in a black duffel bag, which he took with him, Howlett said.

Michaels' pimp and her friend found her dead about an hour later, according to court filings. Michaels was found lying beside a bloody knife inside her hotel room just after 12:30 a.m. Tuesday, officials said.

Figure 3: Chicago Tribune article linking accused's name to case_participant_id.[9]

Next, a simple search was done on Google for "October 10 2011 homicide Chicago" for which the first search result was a page "Tracking homicides in Chicago", which pulls information from the Cook County Medical Examiner's Office, Chicago Police Department, and the Chicago Breaking News Center.[8] On this page there were only three homicide events on 10/10/2011, the third of which was marked as occurring in community area "Near North Side", which aligns with the value in the UNIT feature ("DISTRICT 18- NEAR NORTH") of two of our rows. A link was also provided to "Read More" at a Chicago Tribune article which contained case information that matched the age, gender, and dates provided in the dataset (Figure 3).[9] The article also names the defendant, removing any privacy given by the use of CASE_PARTICIPANT_ID and allowing, with reasonable certainty, for the linkage of the accused's name.

Cook County's attempts to anonymize this data have evidently failed; their privacy goals and strategies ought to be re-evaluated.

3 Implementation and Validation

3.1 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 (Figure 23). 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 is then filtered so that it only includes rows corresponding to CASE_PARTICIPANT_IDs that are also present in Disposition and/or Sentencing. This ensures that each case in the training data is far enough along in the legal process to be assigned an MHI. 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 (Figure 24).

There was extensive data cleaning performed on the datasets in order to prepare the training set, some which violates best practice protocols. All string variables were converted to lowercase, all numeric columns were converted to integers, and date features were converted to datetime while 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. All missing non-numeric inputs were replaced with "unknown" (a common filler already being used by Cook County). Age_at_incident was converted to integers, with null and outlying values (greater than 100) replaced by the median age. The median age was computed using the full dataset; this constitutes data leakage. Missing values should always be imputed using training data only.

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 is missing. These coordinates represent a location in the Atlantic Ocean off the west African coast. The function should be improved by instead interpolating these missing values to some central location within Cook County.

The six datetime columns and four IDs are removed from the dataset (with case_ participant_id remaining as the index).

To deal with the significant class imbalance in the dataset (only 0.77% of the cases are positive), the negative class is downsampled in the training set. To do this, 100% of positive instances are sampled, and the negative class is sampled without replacement until it comprises 50% of the training set instances. The validation and test sets are not downsampled, to replicate deployment.

3.2 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. 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.

3.3 Validation

During development, the team used RECEIVED_DATE to create a single time-based training/validation/test split. The earliest 70% of cases are used for training, the next 15% of cases are used for validation, and the latest 15% of cases are used as a final test set. The splits had to be time aware to ensure that the ADS only predicted forward in time. As discussed above, re-sentencing leads to a lower base rates in the test set (Figure 25).

The ADS has not been tested for robustness. This is a major flaw in methodology; it is crucial to know how stable a model is, and how small variations in data might affect it.

The singular validation set was used to select a model class and tune hyperparameters to maximize AUC. Due to the high cost of false negatives, recall was evaluated at each model iteration; models which made only small improvements in AUC at the expense of recall were rejected. Recall was not selected as the optimization target because a trivial model which assigns a positive outcome to all cases would achieve perfect recall, yet would be useless as a classifier. Instead, the team used AUC to select a model which would perform well at all classification thresholds. This decision was bolstered by the fact that the team did not have access to information about the resource constraints of service providers.

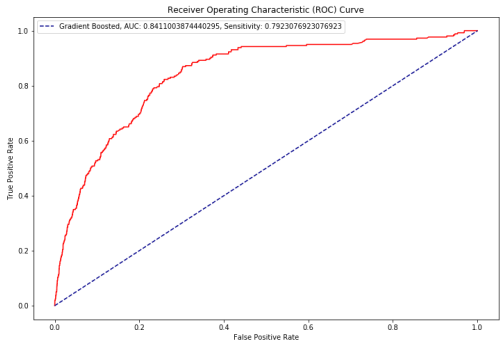


Figure 4: Receiver Operating Characteristic (ROC) curve for original gradient-boosting model's performance on test set.

The test set was used to evaluate the tuned model's performance on unseen data. The model achieved an AUC of 0.84 and a recall of 0.79. The ROC curve in Figure 4 shows the model's performance at different classification thresholds. The original project team offered the following suggestion for selecting a classification threshold: "the model achieves a recall of over 90% fairly quickly, and then flattens out. The false positive rate at that point is just over 40%. The researchers posit that this point represents the classification threshold best suited to the business case."[5]

The performance evaluation described above is only valid insofar as MHI is an accurate proxy for the true goal of identifying cases of serious mental illness. We can test the suitability of MHI as a target variable by comparing MHI base rates (i.e. prevalence) to epidemiological data.

We will first consider the following two sources of "ground truth": 1) the 2017 National Survey on Drug Use and Health (NSDUH) by the Substance Abuse and Mental Health Services Administration (SAMHSA),[10] and 2) the 2018 Healthy Chicago Survey, by the Chicago Department of Public Health (CDPH).[11][12] Note that public health data is not available for Cook County as a whole, as there are two separate Public Health Departments for Chicago and suburban Cook County.

Figure 5 compares MHI to CDPH's designation of "serious psychological distress" (SPD) and SAMHSA's "serious mental illness" (SMI) and "any mental illness" (AMI). We find that the prevalence of AMI (18.9%) is several times higher than that of SMI (4.5%). The prevalence of SPD in Chicago (6.6%) appears to be higher than that of SMI in the US, however the portion of people receiving services for SMI and SPD are similar. The overall MHI base rate of 0.77% is an order of magnitude lower than all of these measures.

Also included for comparison is the staggering estimate that the Cook County Sheriff gives for the prevalence of mental illness within Cook County Jail.[14] 30% is sometimes given as an estimate of AMI, and sometimes of SMI.[15] The researchers have been unable to find any study to back up the Sheriff's estimate. Given that SMI within jails is estimated nationally at 20-25%,[16] and that Cook County is known to be rife with mental illness following the closure of several area mental hospitals,[17] we will assume that 30% is an estimate of SMI.

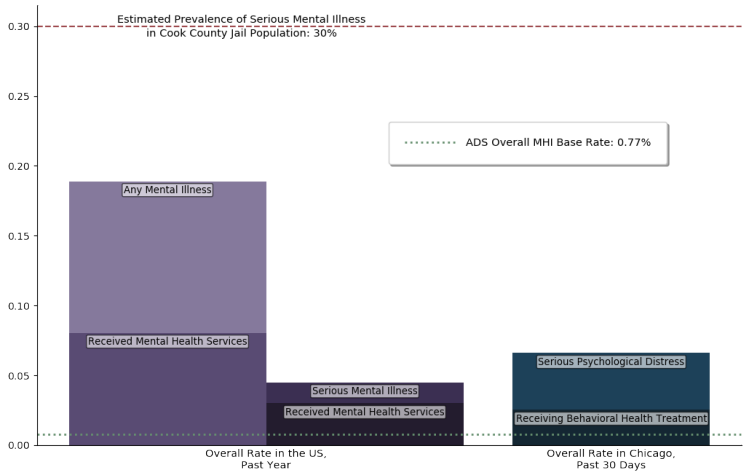


Figure 5: MHI base rate as compared to treatment and prevalence data from SAMHSA,[10] CDPH,[11–12] and the Cook County Sheriff.[14]

We can see in Figure 5 that this number dwarfs all other measures of mental illness, with the comparison to MHI being the starkest. A snapshot of the jail population at any given time is likely to have a higher proportion of mental illness than the set of all people who pass through the jail, because people with mental illness tend to stay in jail longer.[18] Thus, the true proportion of mentally ill people in the ADS' data is likely to less than 30%, although still significantly higher than the rates reported for the general population.

Figure 6 shows the prevalence of SMI, SPD, and MHI among sub-populations. We can see that the demographic trends within the ADS' data are not present in Chicago or the US as a whole.

SMI rates decrease dramatically with age, but we see an opposite trend in MHI. This is worrisome, because it indicates that MHI may be a poor proxy for SMI. However, we do know that the distribution of SMI can vary considerably for people within the criminal justice system. The researchers have been unable to locate any epidemiological studies which study the prevalence of mental illness within different incarcerated age groups.

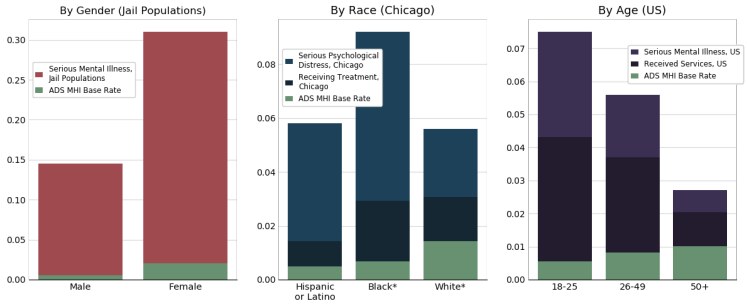


Figure 6: MHI base rate as compared to treatment and prevalence data from SAMHSA,[10] CDPH,[11–12] and the Treatment Advocacy Center[18]. (* indicates non-Hispanic.)

Looking at the racial breakdown of SPD in Chicago, we see interesting disparities with regards to who is receiving treatment. The rate of SPD is by far the highest for non-Hispanic black people, yet they receive treatment at about the same rate as non-Hispanic white people. Hispanic and Latino people experience SPD at about the same rate as the non-Hispanic white group, yet they receive treatment far less often. The distribution of MHI rates is much closer to the distribution of treated SPD than to actual SPD.

MHI is far higher in women than in men. The Chicago rates of SPD and its treatment are approximately equal between the genders (Figure 26), however a 2009 study analyzing inmate interviews found that 31% of jailed women and 14.5% of jailed men had symptoms of SMI. The authors of the 2009 study pointed out that these were almost certainly underestimates at the time, and that the true portion is growing over time.[18] An earlier study found that 75% of jailed women and 63% of jailed men had "mental problems."[15] Hence, although we have no firm source of ground truth for the portion of jailed men and women suffering from serious mental illness, it seems that the divergence in MHI's distribution across genders may in fact be a warranted artifact of a true disparity.

Overall, it's clear that MHI is vastly underestimating the true prevalence of serious mental illness in its target population. As we saw with race, MHI seems to be a better proxy for treated mental illness than for actual mental illness. This makes sense, because several of MHI's constituent outcomes actually require the case participant to have an active case with the State Department of Health.[19] This is problematic, because of the large gaps in treatment rates that we see in Chicago (Figure 26). Due to the lack of reliable sources of ground truth for rates of serious mental illness within the criminal justice system, it is difficult to assess the extent to which the use of MHI distorts the reality of the ADS' intended target.

3.4 Sensitive Feature Removal

The ADS team noted in their original report that the ADS' use of age, gender, and race was a major problem. They concluded that further work was necessary to develop a model which did not rely on these sensitive features or on other features that would be able to predict them.[5] We take up that work here. To identify proxy features for age, race, and gender, the pairwise mutual information was calculated between each sensitive feature and all other features in the Initiation dataset. We found that all three sensitive features have high mutual information with ID and date features (Figure 27). Since the ADS drops these features before modeling, we looked more closely at the mutual information with the remaining features (Figure 7).

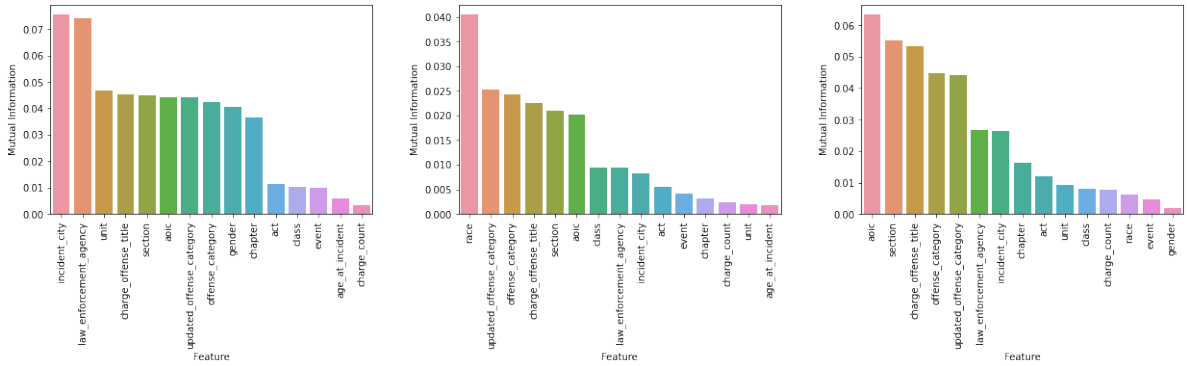


Figure 7: Pairwise mutual information between race, gender, and age and all other features used in the final model.

After looking at the results, we defined a mutual information threshold of 0.05 to identify proxies, because there was a clear cut-off at this level in all three cases. We found that INCIDENT_CITY and LAW_ENFORCEMENT_AGENCY were proxies for race under our definition. As described in "Data Profiling" above, these two features are often related to each other.

The distributions of INCIDENT_CITY and LAW_ENFORCEMENT_AGENCY were compared between the entire dataset and racial subpopulations (Figure 28), and revealed clear differences across race. This was particularly apparent for incident_city and the Hispanic population which had the highest Kullback–Leibler divergence value (1.05) and distinct differences in distribution and common incident cities (Figure 28). This speaks to the racial and ethnic differences across cities and neighborhoods, and confirms that location-based features cannot be treated as independent from race.

GENDER did not have a high mutual information with any features except for race. AOIC, SECTION, and CHARGE_OFFENSE_TITLE, which all describe the legal code of the charge, were identified as proxies

for age. These charge-related features seem as though they could be genuinely relevant to the task, but we decided to drop them along with the other proxies for this experiment.

We trained two new classifiers: one that dropped only the three sensitive features (race, gender, and age), and another that dropped the five proxy features described above in addition to the sensitive features. In both cases we removed all dummy features that resulted from the transformation of these features, as well as any features that were engineered from them (e.g. latitude and age_unknown).

	Recall	AUC	Accuracy	Number of positive cases predicted	Actual number of positive cases in test set	Predicted prevalence	Actual prevalence in test set
Including sensitive features and proxies	0.792	0.841	0.765	10337	260	0.238	0.006
Sensitive features removed	0.804	0.843	0.759	10628		0.245	
Sensitive features and proxies removed	0.812	0.828	0.742	11370		0.262	

Figure 8: ADS performance without sensitive features and proxies.

Surprisingly, when sensitive features were removed as explicit input, we found that AUC and recall improved (Figure 8). When proxies were dropped as well, recall improved further, but there was a slight drop in AUC. Whether this decline in performance is warranted is a matter of opinion, although the authors believe firmly that the location-based features should be removed.

4 Outcomes

4.1 Disparate Impact

Disparate impact (DI) is a useful metric for analyzing the fairness of outcome allocation in a set of predictions. By comparing DI in the predictions to DI in the actual data, we can also see the degree to which the ADS amplifies or reduces bias.

We calculated DI in terms of both race and gender in the original test set and as well as the full original dataset, and compared those figures to the DI in predictions yielded by the transformation of those datasets by three different classifiers: the original classifier used in the report, the classifier without sensitive features, and the classifier without both sensitive features and their proxies (Figure 9). When analyzing DI on race, race_white was used as the privileged class, and all others were treated as non-privileged. When analyzing DI on gender, the privileged class was gender_male, and gender_female and gender_unknown were grouped together under non-privileged. An MHI of 1 was always considered the positive outcome, because the ADS is designed for positive intervention.

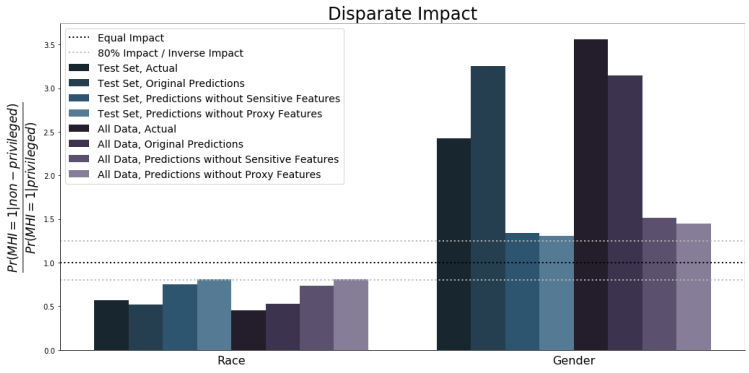


Figure 9: Disparate impact of predictions and true data, across classifiers.

We found that in the original dataset, DI was below 1 for race (0.57 in the test set and 0.46 in the entire dataset), and above 1 for gender (2.42 in the test set and 3.56 in the entire dataset). This means that Cook County courts have been about twice as likely to issue a mental-health related outcome for a white person as they they have for a non-white person, and about three times as likely to do so for a woman as compared to a man. The substantial difference in DI for gender between the test set and all of the data is likely due to statistical bias resulting from the time-based split (Figure 25). The distribution of MHI is changing as Cook County continues to expand its specialty court programs;[19] [20] we see here that outcomes in the more recent cases which comprise the test set are allocated more evenly.

It's important to note that disparate impact is not necessarily unfair. As demonstrated above in the validation section, statistical parity with regards to gender ought not be a goal in this setting, where the best available sources of ground truth indicate that the needs of the groups are indeed disparate. In the case of race, however, statistical parity seems a reasonable baseline goal. Whether ADS architects should aim for a DI above 1 to address the heightened SPD in Chicago's black community is an open question which demands further consideration.

For both gender and race, the original classifier pushes DI toward 1 (i.e., toward statistical parity) in the full dataset, and further from 1 in the test set. This result is a function of the sets' different starting points – both measures land in similar places for the classifier predictions (0.52 and 0.53 for race, and 3.25 and 3.15 for gender). This effect speaks to the potential pitfalls of deploying a high-stakes ADS under significant concept drift: even if a system appears to be mitigating bias during development, it could actually do the opposite to new data.

Removing sensitive features from the ADS inputs had a dramatic effect on DI, especially with respect to gender. This simple change cut DI on gender in half for both datasets. This indicates that the model was relying heavily on sensitive features to make its predictions. The researchers posit that the resultant DI is actually lower than desirable given the epidemiological evidence. If this version of the ADS were deployed, outcomes may need to be reassigned after processing, perhaps in the form of a special program to target women with classifier scores just shy of the classification threshold.

The DI for race improved substantially with the removal of sensitive features. After the removal of proxy features, DI actually reached 0.81 for both datasets. The 80% rule does not specifically apply here in a legal sense, but it is an important benchmark nonetheless. The fact that these simple pre-processing steps drove DI past this benchmark speaks to their importance. It's likely that lowering the MI threshold for race proxies below 0.05 (i.e., removing additional features) could drive DI up even further.

4.2 Additional Fairness Metrics

False positive rate, recall, AUC, and accuracy were calculated for each gender, racial, and age subgroup present in the test set (Figures 10, 11, and 13; Section 7, Jupyter Notebook). Intersectional race/gender statistics can be found in Figure 29.

Due to class imbalance, accuracy is not a particularly useful performance metric for this ADS; AUC is a more appropriate measure of the classifier's discrimination prowess. AUC represents the ability of the classifier to correctly identify positives without misidentifying negatives, across all possible classification thresholds. We also include accuracy for its familiarity; non-expert end-users are likely to look at it, so it is important to consider the portrait of the ADS painted by accuracy.

Error rate analysis is crucial for this ADS. It's especially important to understand how errors are distributed across subgroups, because false positives and false negatives affect the stakeholders very differently. A false positive represents a negligible monetary harm to the county or service provider, and perhaps some small potential for damage to the pride or reputation of the individual; a false negative, on the other hand, could have devastating effects on the well-being of an individual who is denied much-needed support.

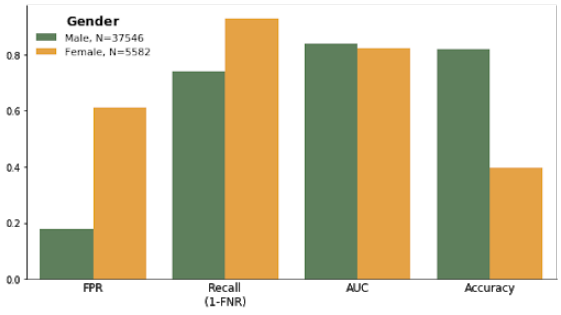


Figure 10: Fairness metrics across gender.

Rather than FNR, we have calculated recall, because it was used by the original ADS team as a key performance indicator. Recall is equal to 1-FNR, and it indicates the portion of positive cases which are correctly identified.

FPR gives the portion of negative cases which are misclassified. Due to the extreme class imbalance in the underlying data, FPR also gives an approximation of the predicted prevalence.

In Figure 10, we can see that males are very 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. This is likely to instill distrust in users.

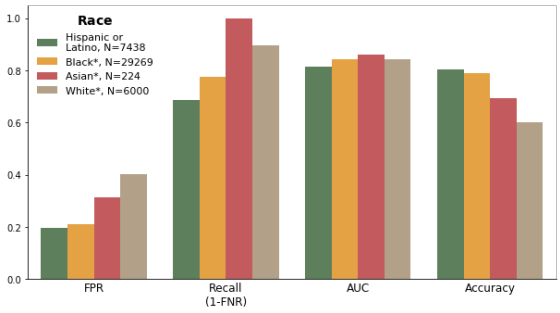


Figure 11: Fairness metrics across race.

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. 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.

Fairness metrics vary widely between racial sub-populations. By far the lowest accuracy is 0.29 for the American Indian subgroup, however there are only seven individuals in that group, and no actual positives. Setting aside groups that do not have any positive instances in the test set, we find that the white group has the highest FPR at 40%, indicating that this group will be offered unneeded specialized support at a disproportionate rate. The Asian group has perfect recall, meaning that ADS correctly identified all positive instances of those groups in the test. If we group all Hispanic sub-groups together, we find that this group has the lowest FPR and recall of all the racial subgroups. This indicates that the ADS disadvantages Hispanic and Latino people.

Race	Gender	Test Set Group Size (N)	FPR	Recall	AUC	Accuracy
All	All	43363	0.2350	0.7923	0.8411	0.7651
	Male	37546	0.1802	0.7407	0.8395	0.8194
Black*	All	29269	0.2111	0.7738	0.8407	0.7889
	Male	25824	0.1590	0.7025	0.8359	0.8403

Figure 12: Intersectional fairness metrics spotlighting black men.

Special attention is due to the black, male intersectional subgroup, because it makes up such a large portion of the dataset (60% in the test set), and because black men have historically borne the brunt of the US criminal justice system's harms.[21] We find an especially low FPR and recall for this beleaguered demographic, even as compared to the overall non-Hispanic black and the overall male groups (Figure 12).

The error rate trends for age are clear, as demonstrated by Figure 13. With the notable exception of the over-60 group, we see that as age increases, FPR and recall increase while accuracy decreases. This is most startling when looking at the youngest age groups: the ADS misses an unacceptable 40% of positive cases for 17-19 year-olds. As discussed in the Validation section above, the choice of MHI as a proxy for mental illness already disadvantages young people. We see here that the model is actually severely amplifying that bias.

Overall, we find that AUC tends to be fairly stable across sub-populations. Higher FPR is generally associated with lower FNR and lower accuracy, which makes sense in the context of class imbalance.

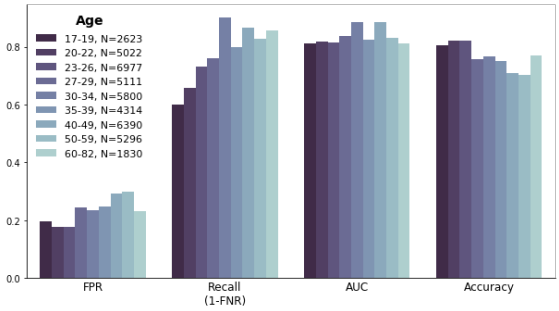


Figure 13: Fairness metrics by age.

4.3 Interpretable Explanations

Gradient-boosting classifiers are notorious for their lack of transparency. The team used the LIME submodular picker [22] to identify a set of eight descriptive examples to explain the behavior of the ADS using five features each (Figure 30; 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 five 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.[23] Given the unusually high proportion of aggravated DUI's in the Hispanic subgroup (Figure 31), 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 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.[24][25]

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.[5] When building the ADS, the team also identified section 402, which relates to narcotics and their possession,[26] as being important enough to receive its own feature

Latitude below 41.88 is seen several times to contribute to a negative MHI. Latitude was engineered from INCIDENT_CITY, which we have identified as a proxy for race. Chicago's South Side is 78% black, whereas the city as a whole is 31% black.[27] Assigning negative outcomes based on being south of a certain latitude line is therefore cause for concern, and is reminiscent of the redlining which befell the South Side in the 1930s.[28]

Figure 14 has been adapted to indicate the 41.88 latitude line. The bottom of census tract 839100 (highlighted in yellow) has been used as a proxy for the latitude line in question (extended by dashed lines), because its southern border lies at 41.881.[29] We see that Chicago's white residents are concentrated above 41.88, and that Cook County's black residents are concentrated below it.

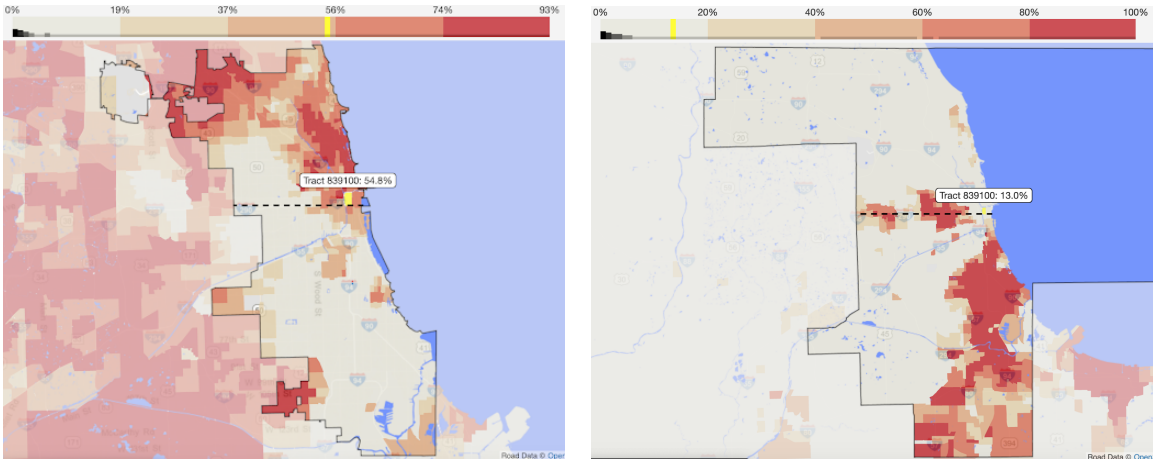


Figure 14: Percent of Chicago residents that are white (left) and percentage of Cook County residents that are black (right), above and below the 41.88 latitude line.[31]

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.

The local importance identified by LIME does not imply global feature importance. To investigate global importance, we conducted classical feature importance using scikit-learn's feature_importances_ (Figure 32). Globally, the most important feature was gender_female, which was over 30% more than the next highest feature importance score, updated_offense_category_narcotics. Both of these are in line with the features commonly occurring in LIME's local explanations, which suggests that the ADS' relies heavily on these two features for its predictions.

Neither LIME nor classic feature importance can provide conclusive answers to whether or how the model is using the features discussed above. We hoped to utilize the causal framework of QII to explain whether the ADS uses these sensitive features directly or indirectly, and the impact of feature interactions. However, after reaching out to the authors of the QII paper to inquire as to existing implementations, we learned that there is not an actively maintained implementation, and the existing implementation only supports a limited number of datasets.[32] As such, the only causal evidence available to us are the changes that occurred when sensitive features were removed. We believe this is sufficient to conclude that the original ADS was in some way directly using the sensitive features.

5 Summary

Overall, the researchers have serious doubts as to whether the data is appropriate for the ADS. We found that the target variable vastly underestimates the incidence of serious mental illness in the criminal justice system, and that its distribution may not be representative of the target population. It is likely that the courts are biased towards certain types or presentations of mental illness, and that their reliance on pre-existing cases in the Health Department is introducing structural bias from the healthcare system. By using court outcomes to define a mental health proxy, the ADS effectively gives the courts full control over who is eligible for mental health services. Using court-assigned outcomes will therefore produce a potentially dangerous positive feedback loop which reinforces the courts' conception of mental illness.

A high degree of domain knowledge is necessary to understand the business rules governing the data, as well as some important features in it. This compounds the lack of transparency in the model type, and is likely to erode public trust in the model's decisions.

There are several improvements which could be made to the ADS pipeline to improve it. Data collection could certainly be better standardized across police units, although that is not within the control of the ADS architects. Missing value imputation ought not take into account values in the test set. The sensitive features age, race, and gender should be dropped from the dataset; the costs and benefits of dropping proxies for these features should be carefully analyzed.

The original version of the ADS, which explicitly relies on sensitive features, results in problematic disparate impact on race, and amplifies bias against young people. Removing sensitive features and their proxies has proven to be a viable strategy for mitigating this unfairness while preserving AUC. The choice to remove more proxy features will depend on the evaluation of trade-offs in metrics which represent the interests of different stakeholders. High AUC will please service providers who are keen to reduce costs by maximizing impact. Improving recall will benefit people suffering from mental illness. Focusing on sub-population recall will be important for those groups where cases are missed at a disproportionately high rate (for example, teenagers and black men). Imposing statistical parity would benefit racial minorities (particularly black and Latino people), but may result in the under-serving of women.

Model variance remains untested. Sensitivity of the model to slight perturbations in the input data could result in unstable, unreliable predictions. This issue ought to be addressed with bootstrapping or with walk-forward cross-validation. Model performance could also be validated using future data releases from Cook County. Concept drift and censoring necessitate careful monitoring in deployment.

The appropriateness of the ADS must ultimately be considered within the specific context of the Cook County criminal justice ecosystem. Cook County Jail has required pre-bond mental health screenings since 2012, and is working to expand early screening programs to county courthouses.[33] The people performing these screenings possess professional expertise as well as a great deal more information than the ADS has

access to, including previous criminal record, housing history, and, crucially, in-person interviews. In this context, the ADS does not appear to offer any tangible benefit to stakeholders; we conclude that the project is not viable as an ADS. It is possible, however, that the project could be converted from a decision system to a tool for studying structural bias in court system outcomes.

6 References

Notes

[1]https://github.com/kelseymarkey/cook-county-mental-health-prediction

[2]https://datacatalog.cookcountyil.gov

[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]https://datacatalog.cookcountyil.gov/Courts/Initiation/7mck-ehwz

[5]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

[6]Mueller, Heiko. (2019). Data Profiling and Data Cleaning. https://dataresponsibly.github.io/courses/documents/ spring20/ Lecture3.pdf

[7]https://datacatalog.cookcountyil.gov/Courts/Dispositions/apwk-dzx8

[8]http://homicides.redeyechicago.com/date/2011/10/

[9] Ford, L. (2018, September 5) Bail set at $1M in slaying of woman in Chicago hotel. Retrieved from https://www.chicagotribune.com/news/ct-xpm-2011-10-14-chi-bail-set-at-1m-in-slaying-of-woman-in-chicago-hotel-20111014-story.html

[10]https://www.nimh.nih.gov/health/statistics/mental-illness.shtml

[11]https://www.chicagohealthatlas.org/indicators/serious-psychological-distress

[12] https://www.chicagohealthatlas.org/indicators/behavioral-health-treatment

[13]https://www.cookcountysheriff.org/departments/mental-health-policy-advocacy/

[14]https://www.cookcountysheriff.org/departments/mental-health-policy-advocacy/

[15]https://www.bjs.gov/content/pub/pdf/mhppji.pdf

[16]https://courts.illinois.gov/Administrative/Reports/MH_Justice_Cook_County_Bond_Court.pdf, page 9

[17]https://www.theatlantic.com/politics/archive/2015/06/americas-largest-mental-hospital-is-a-jail/395012/

[18]https://www.treatmentadvocacycenter.org/evidence-and-research/learn-more-about/3695

[19] http://www.cookcountycourt.org/ABOUTTHECOURT/CountyDepartment/CriminalDivision/SpecialtyTreatmentCourts/FelonyMentalHealthCourtProgram.aspx

[20]https://cookcountyhealth.org/press_releases/cook-county-health-expanding-access-to-mental-health-treatment-with-4m-grant/

[21]https://www.vsb.org/docs/valawyermagazine/dec00dunnaville.pdf

[22]Ribeiro, M. T., Singh, S., Guestrin, C. (2016, August). "Why should I trust you? Explaining the predictions of any classifier". In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).

[23]https://codes.findlaw.com/il/chapter-625-vehicles/!tid=NF5C556ECCF424FAC86A9D31D8C9591B9

[24]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/.

[25]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.

[26]https://www.cyberdriveillinois.com

[27]https://statisticalatlas.com/neighborhood/Illinois/Chicago/South-Chicago/Race-and-Ethnicity

[28]https://www.wbez.org/stories/new-redlining-maps-show-chicago-housing-discrimination/37c0dce7-0562-474a-8e1c-50948219ecbb

[29]https://geocoding.geo.census.gov/geocoder/geographies/coordinates?x=-87.63y=41.88benchmark=4vintage=4

[30]Adapted from https://statisticalatlas.com/place/Illinois/Chicago/Race-and-Ethnicitydata-map/tract and https://statisticalatlas.com/county/I County/Race-and-Ethnicitydata-map/tract

[31]Adapted from https://statisticalatlas.com/place/Illinois/Chicago/Race-and-Ethnicitydata-map/tract and https://statisticalatlas.com/county/I County/Race-and-Ethnicitydata-map/tract

[32]https://github.com/cmu-transparency/tool-qii

[33]https://www.cookcountysheriff.org/mental-health-template/

7 Appendix

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.

Figure 15: Cook County Data Portal's description of attributes in Initiation dataset. [5]

	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

Figure 16: Profiling results for Initiation.

	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

Figure 17: Profiling results for Disposition.

	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

Figure 18: Profiling results for Sentencing.

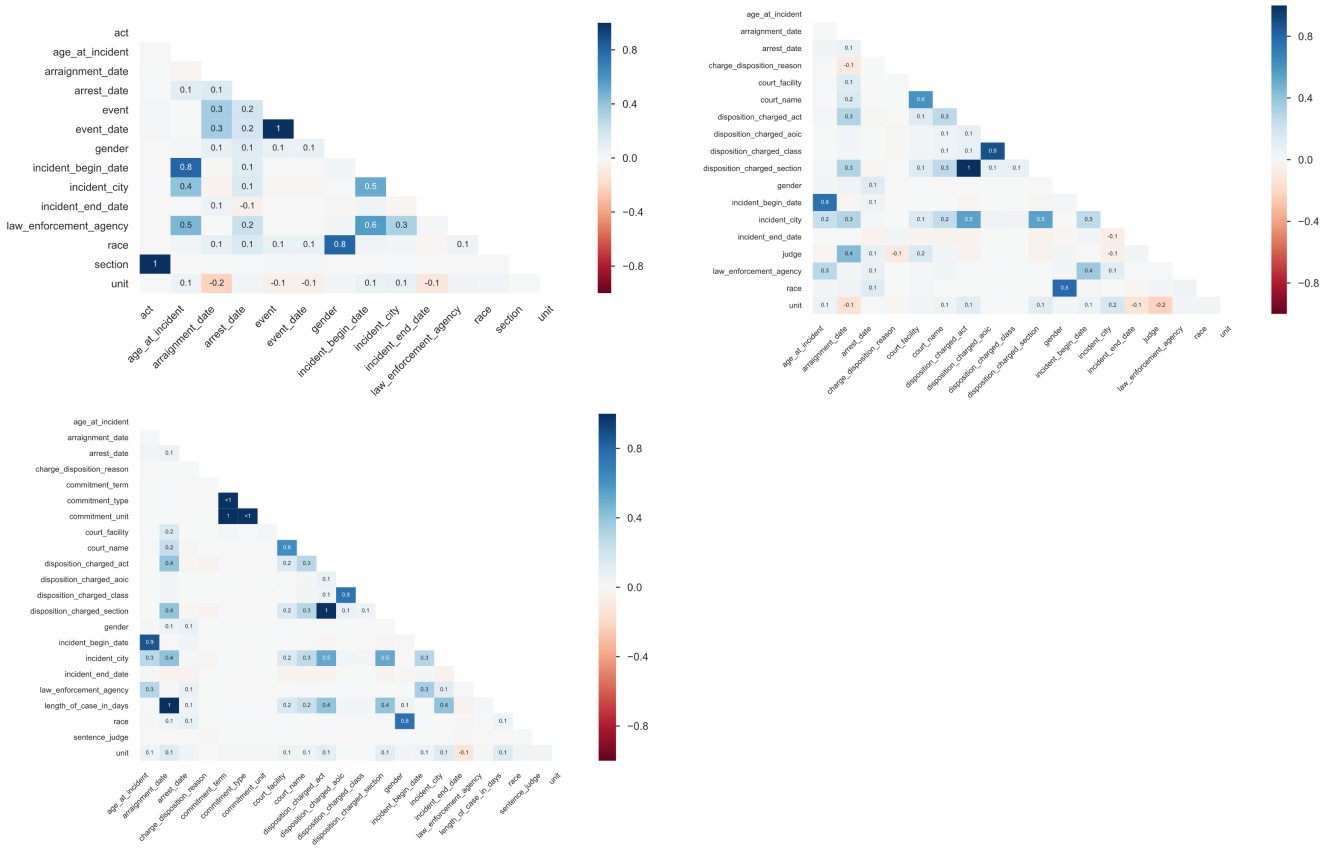


Figure 19: Missing value heatmaps for Initiation (top left), Disposition (top right), and Sentencing (bottom left)

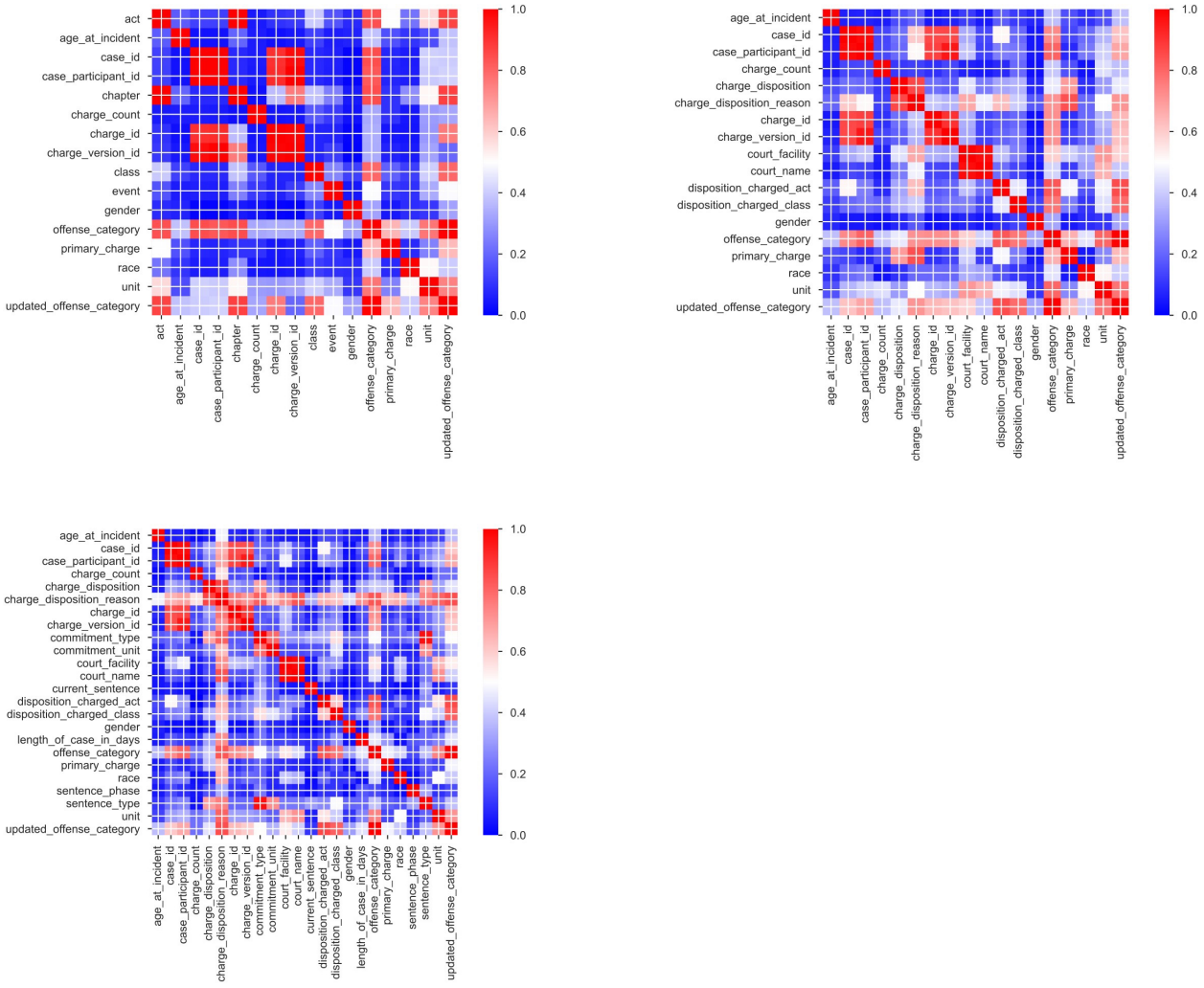


Figure 20: Correlation matrices for Initiation (top left), Disposition (top right), and Sentencing (bottom left)

CASE_ID	CASE_PARTICIPANT_ID	OFFENSE_CATEGORY	CHARGE_ID	CHARGE_OFFENSE_TITLE	EVENT_DATE	AGE_AT_INCIDENT	GENDER	RACE	INCIDENT_BEGIN_DATE	INCIDENT_END_DATE	ARREST_DATE	UNIT	INCIDENT_CITY	RECEIVED_DATE
115,778,940,777	902,927,365,437	Homicide	2,062,867,44(	FIRST DEGREE MURDER	11/1/2011 12:00:00 AM	23	Male	Black	10/10/2011 12:00:00 AM	10/11/2011 12:00:00 AM	10/11/2011 8:40:00 PM	District 18 - Near North	Chicago	10/13/2011 12:00:00 AM
115,778,940,777	902,927,365,437	Homicide	2,058,965,46(	FIRST DEGREE MURDER	11/1/2011 12:00:00 AM	23	Male	Black	10/10/2011 12:00:00 AM	10/11/2011 12:00:00 AM	10/11/2011 8:40:00 PM	District 18 - Near North	Chicago	10/13/2011 12:00:00 AM

Figure 21: Cases in the Initiation dataset where INCIDENT_BEGIN_DATE = 10/10/2011 and CHARGE_OFFENSE_TITLE = FIRST DEGREE MURDER.

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

Figure 22: Features that remained the same (left) and varied (right) during aggregation by case_participant_id.[5]

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

Figure 23: Features that remained the same (left) and varied (right) during aggregation by case_participant_id.[5]

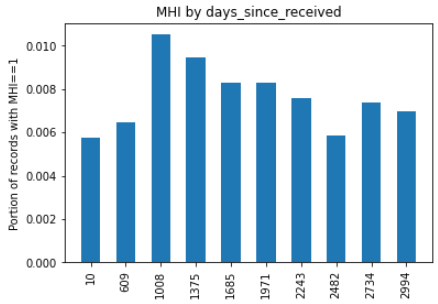


Figure 24: MHI base rates of cases that have existed in Cook County Courts for varying numbers of days.[5]

Dataset	Base rate
Training	0.0077
Validation	0.0092
Training + Validation	0.0079
Test	0.0060

Figure 25: Base rates in test, validation, and training sets.

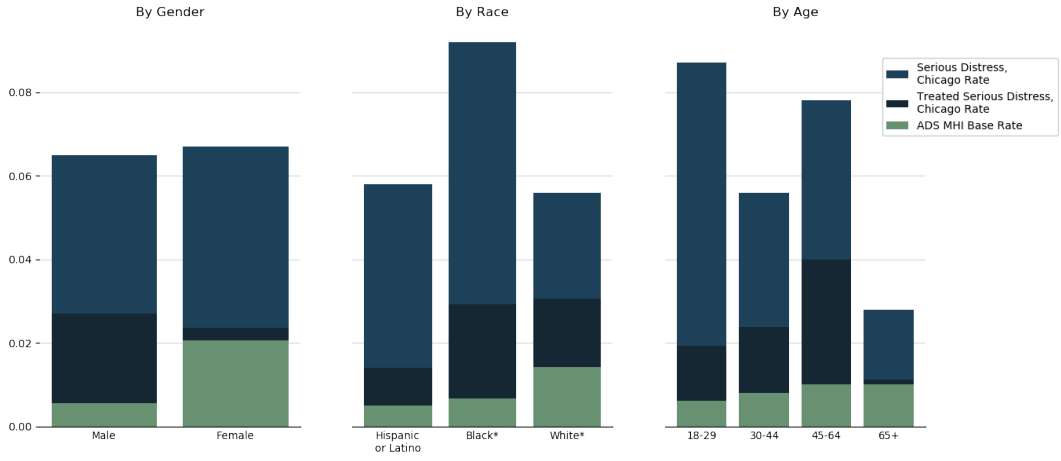


Figure 26: MHI as compared to Chicago-specific treatment and prevalence data from CDPH.[11]

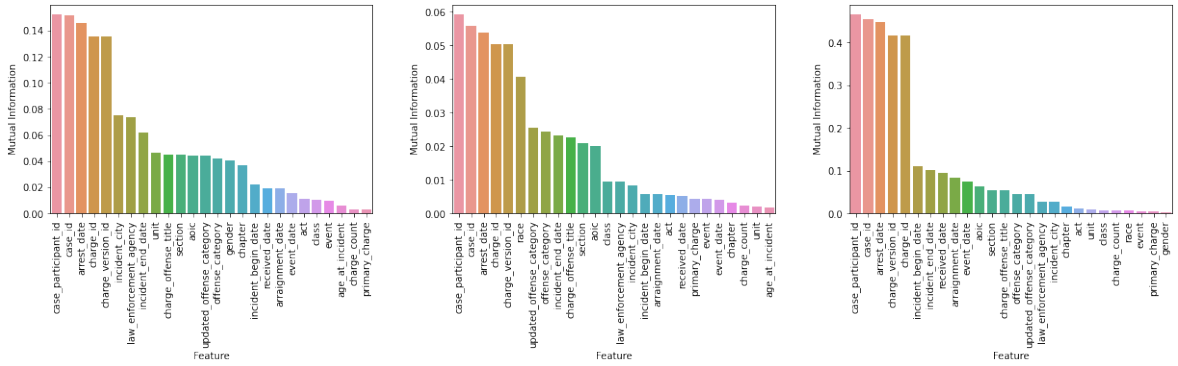


Figure 27: Pairwise mutual information between race, gender, and age and all other features in the Initiation dataset.

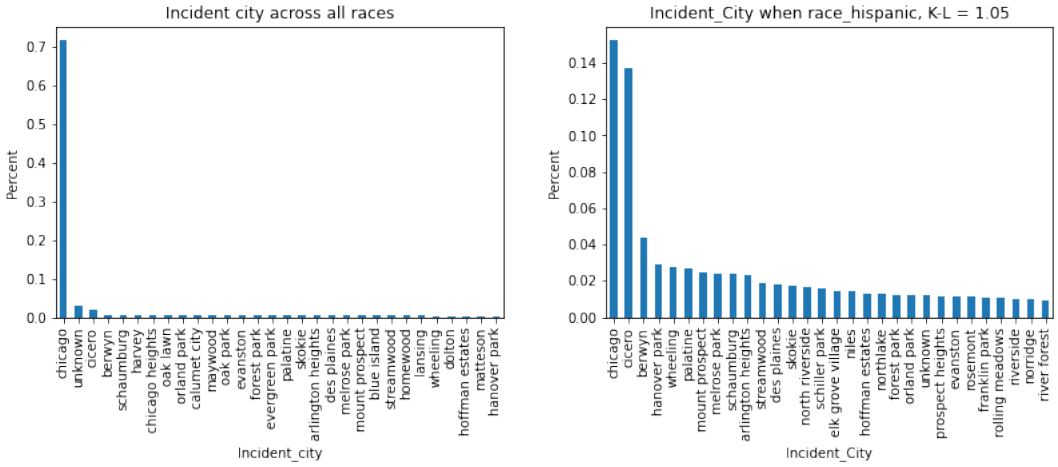


Figure 28: Distribution of incident_city across all races (left) and race = Hispanic (right).

		Test Set Group Size (N)	FPR	Recall	AUC	Accuracy
Race	**Gender**					
All	**All**	43363	0.2350	0.7923	0.8411	0.7651
	Female	5582	0.6101	0.9296	0.8228	0.3968
	Male	37546	0.1802	0.7407	0.8395	0.8194
	Unknown	235	0.1532	NaN	NaN	0.8468
Hispanic Or Latino	**All**	7438	0.1954	0.6875	0.8128	0.8041
	Female	737	0.5598	0.8000	0.8079	0.4450
	Male	6697	0.1558	0.6364	0.7837	0.8435
	Unknown	4	0.0000	NaN	NaN	1.0000
Black*	**All**	29269	0.2111	0.7738	0.8407	0.7889
	Female	3424	0.6085	0.9574	0.8187	0.3992
	Male	25824	0.1590	0.7025	0.8359	0.8403
	Unknown	21	0.0000	NaN	NaN	1.0000
White*	**All**	6000	0.4023	0.8947	0.8423	0.6005
	Female	1345	0.6404	0.9231	0.8544	0.3651
	Male	4653	0.3337	0.8864	0.8550	0.6684
	Unknown	2	0.0000	NaN	NaN	1.0000
Asian*	**All**	224	0.3122	1.0000	0.8612	0.6920
	Female	39	0.6842	1.0000	0.9474	0.3333
	Male	185	0.2350	1.0000	0.8634	0.7676
American Indian*	**All**	7	0.7143	NaN	NaN	0.2857
	Female	3	1.0000	NaN	NaN	0.0000
	Male	4	0.5000	NaN	NaN	0.5000
Two Or More*	**All**	5	0.2000	NaN	NaN	0.8000
	Male	5	0.2000	NaN	NaN	0.8000
Unknown	**All**	420	0.1810	NaN	NaN	0.8190
	Female	34	0.5294	NaN	NaN	0.4706
	Male	178	0.1236	NaN	NaN	0.8764
	Unknown	208	0.1731	NaN	NaN	0.8269

Figure 29: Fairness metrics, intersectional racial/gender subgroups. (* indicates non-Hispanic.)

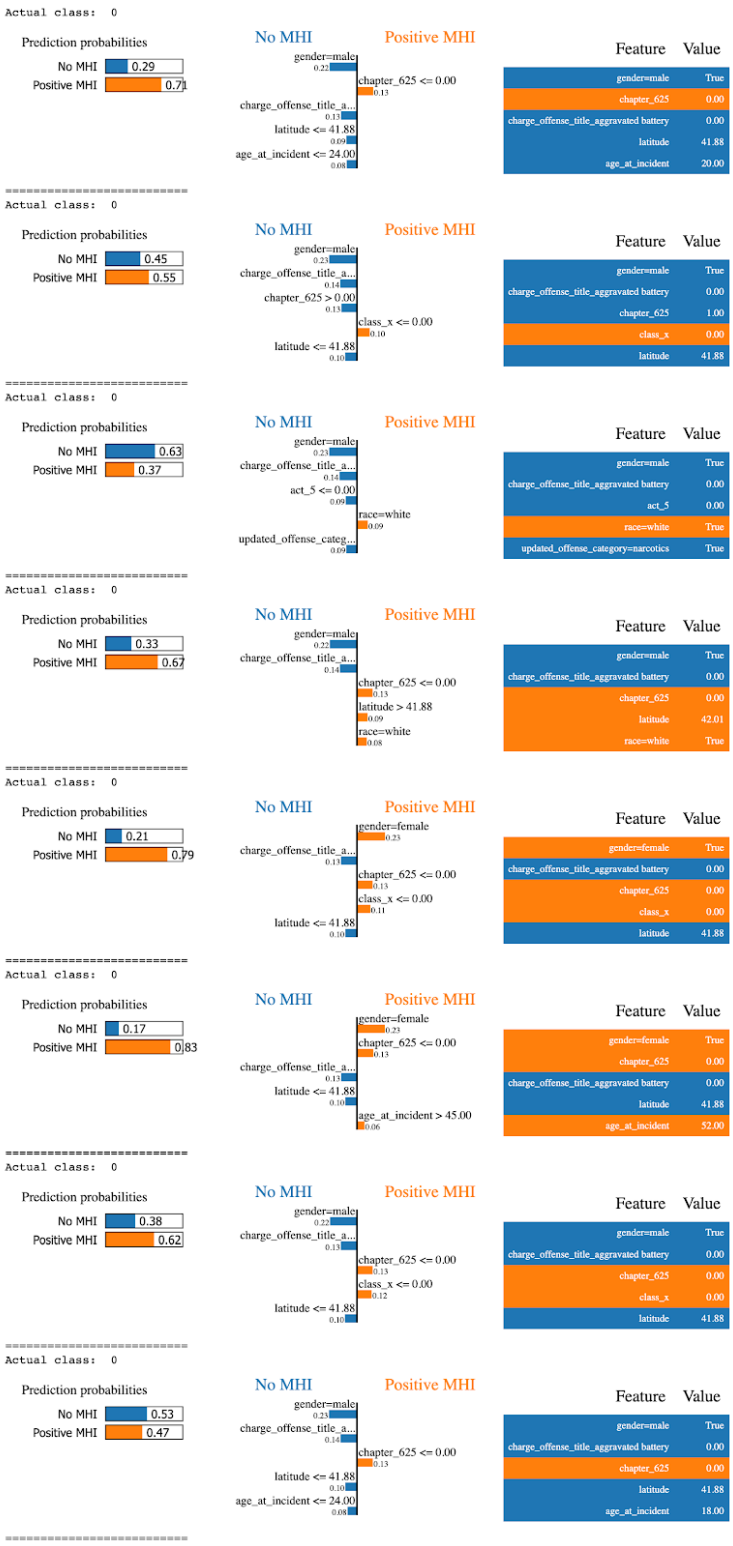


Figure 30: Local explanations chosen by LIME's submodular picker.

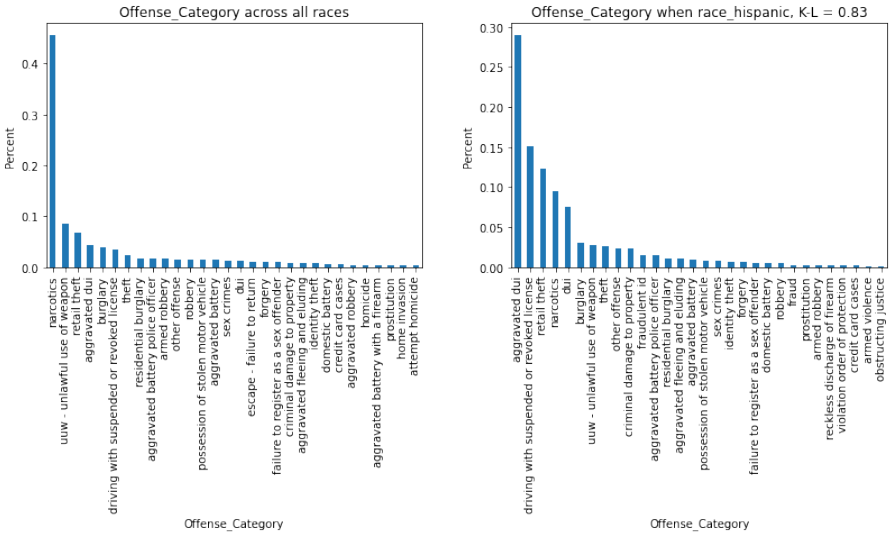


Figure 31: Distribution of offense_category across all races (left) and race = Hispanic (right).

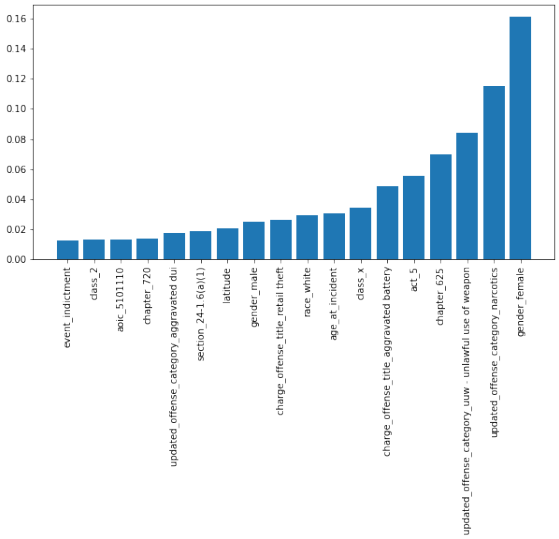


Figure 32: Classic feature importance for ADS' gradient boosting model.