Using machine learning to understand whether sentencing outcomes are predictable based on case and defendant data alone

Zachary Nathan[a], Emma Wolfe[a], and Sharanya Majumder[a]

[a]Dartmouth College

This manuscript was compiled on March 15, 2022

Racial disparities are pervasive in the criminal justice system, with concerning disproportionalities regarding the rate of incarceration, arrest, conviction for Black people compared to White people. Much of the work involving predictive analytics and artificial intelligence aimed at targeting the disparities in the criminal justice system have made the problems they attempt to solve worse. The purpose of our study was to determine whether it was possible to use case and defendant data from Cook County, Illinois to predict sentencing outcomes and to identify the features most important in making this prediction. Our key questions for this project are:

- **To what extend are sentencing outcomes predictable based on case/defendant data alone?**
- **Can case/defendant data predict whether an individual is sentenced to incarceration or probation and how long the sentence is?**
- **Which features of the case/defendant data are the most predictive of sentencing outcomes?**
- **Do these features indicate racial or socioeconomic disparities?**

Machine Learning | Sentencing | Disparities | Incarceration | Justice

Introduction

Last year, in the wake of the murder of George Floyd, protests erupted across the country concerning police brutality against people of color, specifically against Black people. However, injustices against people of color extend far past the policing system. Racial injustices have been a part of our criminal legal system since the founding of the United States. The passage of "Black Codes" and Jim Crow laws after the Civil War made it so economic and labor systems in Southern states relied heavily on the arrest and imprisonment of Black people, making it so slave-like environments were never outlawed. In modern day America, immense racial disparities still pervade the criminal justice system, with certain policies making it so people of color are disproportionately targeted, arrested, and incarcerated: African-American adults are 5.9 times as likely to be incarcerated than white people, African-Americans are more likely than white Americans to be arrested, convicted, and to experience lengthy prison sentences, and when black and white men commit the same crime, Black men on average receive a 20 percent longer sentence (1)

Racial biases are being further amplified by the use of big data and predictive algorithms by police departments across the United States, known as predictive policing. Predictive policing is said to lead to the "discovery of new or previously unknown trends" regarding crime (2). However, numerous studies have suggested that predictive policing tools perpetuate systemic racism, as they rely heavily on socioeconomic variables like education and zip code, acting as proxies for race. These predictive algorithms rely heavily on arrest rates, and as mentioned earlier, Black people are twice as likely to be arrested than White people. Arguments in favor of predictive policing state that automated techniques remove sources of human biases from the criminal justice system, but evidence suggests that human biases have simply been integrated into these predictive models because the models are trained using biased police data (3). The use of predictive policing algorithms has only increased in the past few years, making Black and Latino defendants look riskier than White defendants and further exacerbating racial biases.

The goal of our research project was to determine whether machine learning models can predict sentencing outcomes with only the case and defendant data, specifically whether a trial ends with incarceration. We were also interested in determining the important features into making such predictions regarding sentencing outcome, giving insight into the causes of sentencing disparities and whether they're rooted in racial or socioeconomic differences.

Related Work.

There has been limited empirical work predicting the factors most important for incarceration, but there have been governmental and independent groups that have conducted investigations into disparities regarding incarceration and pretrial detention status. A 2018 longitudinal study by the People's Policy project followed a nationally representative sample of Americans between grades 7 and 12 in 1994 and 1995 (4). Motivated by Black people being disproportionately represented in US prisons, Lewis was interested in looking into what factors explain the racial incarceration gap and whether its better explained by racial bias in sentencing or by a "racist economic system that produces a disproportionate population of impoverished African-Americans" who are victims of a criminal justice system that targets the poor. The People's Policy project concludes that the incarceration gap is due to both causes, but that only class has a statistically significant effect. However, this doesn't mean that race doesn't have an effect on the incarceration gap. Rather, the People's Policy project team states that they observe a "culmination of the race effect", as "if there is a small racial bias each step of the way, this study wouldn't find race to be statistically significant at any given step, but when added together, as it is in the last question, we could see a significant effect" (4).Our project builds upon this study by creating a predictive model for whether a trial ends with incarceration and which factors are most important in making such a prediction over a variety of felony classes instead of generalizing across all types of felonies to see if race has varying importance based on the type of felony.

An additional survey titled "Survey of Inmates in Local Jails 2002" conducted by the Department of Justice looked at national data on the race and ethnicity of people awaiting trial and measured pretrial detention status. The Prison Policy Initiative looked into this data and found that in large urban areas, "Black defendants accused of felonies are 25 percent more likely than white defendants to be held pretrial" (5). Young Black men are also 50 percent more likely to be detained pretrial than White defendants, even in states that have implemented pretrial reforms. The Prison Policy Initiative comes to similar conclusions as the People's Policy project does, stating that national studies of felony cases in large counties conclude that "the direct impact of race on pretrial decisions is weak, but that racial bias acts cumulatively to affect outcomes" (6). Our research project expands on the objectives of this study by the Prison Policy Initiative by looking at varying trial outcomes instead of pretrial data to observe how different can affect the trial outcomes for defendants.

1. Data

For our project, we used the Cook County Open Data to access two data sets, specifically looking into the Sentencing Data and the Initiation Dataset. Both the Sentencing and Initiation Data sets were originally released on February 13th, 2018, but were updated on January 3rd, 2022, which are the versions we used. Furthermore, we filtered the Sentencing data to only include charges where the "PRIMARY_CHARGE_FLAG" variable is set to true, sub-setting to only the most severe charge against the accused. Beyond these two data sets, we also looked at the Police Stations Data set to use the precise locations of police stations for each police district in the city of Chicago.

A. Sentencing Data.

The majority of the data we used was from the Sentencing Data set from the Cook County State's Attorney Office, which is focused on felony charges after the defendants have plead guilty to those charges. Originally, each row in the Sentencing data set is a single charge, making that the original unit of analysis. Because of this, individual defendants can have multiple sentences if they appeal the original sentencing decision, or if they are charged for multiple crimes for a single incident. As mentioned earlier, by filtering to only where "PRIMARY_CHARGE_FLAG" is true, the unit of analysis we used is at the case participant level. This data set includes information regarding the defendant's race, age, and gender. It also contains information concerning the incident and arrest dates, the Law Enforcement Agency involved and the Law Enforcement Unit within the Chicago Police Department associated with the arrest. The Sentencing Data set contains charges starting from October 10th, 2003 to September 9th, 2021. Since we didn't believe there to be any factors concerning year that would create abnormalities, we decided not to subset to any specific date range.

B. Initiation Data.

In addition to the data from the Sentencing Data set, we also incorporated a few columns of interest from the Initiation Data set. "Initiation" refers to how an arrest becomes a "case" in the court system. After an arrest, defendants are taken to bond course, where Assistant State's Attorneys make recommendations regarding whether the defendant presents a risk to public safety and the court sets the bond type and amount. The initiation data set includes the initial bond type, the current status of the bond type, the initial bail amount set at bond court, and the most recent bail amount set. For our purposes, we were only interested in the current bond type and the current bond amount.This data set also includes the legal class for each charge, and, like the Sentencing data, composed primarily of felony charges. For sentencing purposes, felonies are classified as follows in decreasing order of severity : First degree murder (Class M felonies), Class X felonies, Class 1 felonies, Class 2 felonies, Class 3 felonies, and class 4 felonies. The original unit of analysis of the Initiation data set is also at the charge level, with each row indicating a charge filed against one defendant in one case.

Like we did to the Sentencing data set, we filtered the Initiation data set to only include primary charges, so the unit of analysis we used is at the case participant level. We used exact matching to merge the "CASE_PARTICIPANT_ID", "PRIMARY_CHARGE_FLAG", "CLASS", BOND_TYPE_CURRENT", and "BOND_AMOUNT_CURRENT" to the Sentencing data set using a left join. We exact matched with the Sentencing data set using the "CASE_PARTICIPANT_ID" column, which is in both the Sentencing data set and the Initiation data set. Therefore, the individuals that are in our resulting data set include This data set contains charges starting from October 10th, 2011 up to September 9th, 2021. Because the Sentencing data set includes data from 2003 on, we do have missing values for felony class and bond amount and type for data from 2003 to 2011, which is a source of missingness regarding key variables in the Initiation data set.

C. Police Stations Data.

Lastly, we used the Police Stations data set from the Chicago Police Department. In particular, we used the "DISTRICT" and "DISTRICT_NAME" variables to merge using the law enforcement unit information in the Sentencing data, as well as the "LATITUDE" and "LONGITUDE" variables, so that we could incorporate latitude and longitude in the our analysis. The units were not changed between the original data and our calculations, remaining in degrees.

D. Limitations and Missing Data.

For many of the target columns, there is not a large amount of missing data inhibiting the analysis. (see Table 1) Notably, however, about 68% of the observations in the data do not have a recorded law enforcement unit, and, because this is the column used to merge the Police Station data, these observations without a law enforcement unit also do not have information on their police station, or the location of the police station. However, this problem cannot be addressed with the given data because the only column in Sentencing that includes information about the specific arresting police station is "LAW_ENFORCEMENT_UNIT".

Besides the missing law enforcement unit data, another limitation of the data is the types of columns that we have access to. Because the data has been de-identified, personal data is not included, and we have no way of accessing that information, which is important for the defendants' privacy, but does limit the scope of our research. A major example of this is past criminal history, as that certainly influences the way that judges sentence defendants, but is not available in the data. We also don't have information that may be influencing the judge but is not officially a component of sentencing, such as the financial status of the

Table 1. Missingness of Target Variables in the Combined Data

Column Name	Total Missing Entries	Proportion Missing
LAW_ENFORCEMENT_UNIT	128537	0.680400
BOND_AMOUNT_CURRENT	55211	0.292255
BOND_TYPE_CURRENT	52767	0.279318
CLASS	15250	0.080725
ARREST_DATE	4106	0.021735
AGE_AT_INCIDENT	2399	0.012699
INCIDENT_BEGIN_DATE	2003	0.010603
COMMITMENT_TERM	1462	0.007739
COMMITMENT_TYPE	1435	0.007596
RACE	1065	0.003531
GENDER	677	0.003531
SENTENCE_COURT_NAME	667	0.003531
LAW_ENFORCEMENT_AGENCY	168	0.000889
UPDATED_OFFENSE_CATEGORY	0	0.000000
CASE_PARTICIPANT_ID	0	0.000000

Each of the target columns, and their total missing entries as well as the proportion of observations that are missing data in that column.

defendant, any medical diagnoses they may have, or education and employment records. We will therefore be unable to see if machine learning finds these features (or any others) influential, because they are not in the data to begin with.

2. Methods

A. Data Cleaning and Merging.

A.1. Sentencing Data.

In order to prepare our data sets for the machine learning models, we took many steps to clean our data sets.

As mentioned before, the majority of our data was from the Sentencing data set. First, we created binary indicators for the relevant categorical data using one-hot encoding. We created these indicators for race, gender, law enforcement agency, commitment type, the sentence court district, and the updated offense category.

Law enforcement agency text data that served as an indicator for what law enforcement agency made the arrest was used to create "is_chicagopd_derived", which is a binary indicator for whether the law enforcement agency was Chicago PD. Chicago PD is by far the most common law enforcement agency, making up the agency for 118686 out of 188746 charges, which motivated our decision to create a binary indicator for it.

Data containing the race of the defendant as determined by law enforcement or which was self-reported, was used to create four binary indicators for the race of the defendant, "is_black_derived", "is_white_derived", "is_hispanic_derived", and "is_othereth_derived".

Text data stating the gender of the defendant as determined by law enforcement or which was self reported was used to make a single binary indicator. "is_male_derived", equating to 1 if the defendant was male and 0 if they were female. Data containing the age of the defendant in years was made into "age_derived", which was kept in years.

The full date that the defendant was arrested in year/month/date format was used to create "arrest_year_derived", which consisted of only the year of arrest, and to create "arrest_dayofyear_derived", converting these date objects to date-time objects. The full date that the incident of the crime occurred in year/month/date format was used to create "incident_year_derived" and "incident_dayofyear_derived" in the same way that the arrest date-time objects were created. Both of our "dayofyear" indicators are later encoded using sine and cosine as suggested by this blog post, since the machine learning models need to understand that "dayofyear" is a cyclical variable (London 2017).

Data referring to the length of the defendant's sentence and the units of the sentence were used to create "commitment_term_clean", which recorded the length of all sentences in days. The quantitative variable that consists of the type of commitment a defendant was sentenced to, including incarceration in various correction departments, boot camps, probation programs, and release programs, was used to create "is_incarceration_derived" and "is_probation_derived". "is_incarceration_derived" is set to 1 if the defendant was sentenced to a Department of Corrections or a boot camp. Boot camps were classified as incarceration because defendants were not free to leave for the duration of their sentence. "is_probation_derived" is set to 1 if the defendant was sentenced to a probation program, meaning that they were given provisional freedom based on the compliance with conditions set by the court. Both of these indicators took on a value of 0 if the defendant was sentenced to a release program.

Data containing the name of the district court where the defendant was sentenced was used to create six indicator columns, called "is_dist[n]_derived", where [n] took on a value between 1 and 6 to indicate which out of the six district courts the defendant was sentenced in.

Data with the defendant's last recorded offense category was used to create indicator variables by grouping these offenses into wider categories. We chose 10 wider categories, titled "narcotics", "weapon", "vehicular", "theft", "battery", "escape", "fraud", "sex", "homicide", and "personal". This method of consolidating offenses into broader categories is imperfect, as there aren't perfect fits for every type of offense, but these indicators were made to avoid having very small categories in cases where it was avoidable. In many cases, offenses are categorized into multiple groups, with the indicators set appropriately. For example, "Armed Robbery" offenses are categorized as both "weapon" and "theft", to indicate that the case involved theft as well as weapons.

A.2. Initiation Data.

After creating indicators for the Sentencing data set, we needed to clean the Initiation data set. First, we subset the full data set to only our columns of interest, which were "CASE_PARTICIPANT_ID", "PRIMARY_CHARGE_FLAG", "CLASS", "BOND_TYPE_CURRENT", and "BOND_AMOUNT_CURRENT". Like we did with the Sentencing data set, we filtered to only include primary charges.

Using the "CASE_PARTICIPANT_ID" column, we were able to exact match to our Sentencing data set. "CASE_PARTICIPANT_ID" is the same across data sets as long as they were published in the same year, which was the case for the version of the Sentencing data set and Initiation data set we used, allowing us to exact match. We chose to use a left join on the Sentencing data set because the majority of our data of interest was in the Sentencing data. Furthermore, although the Initiation data set is larger, it contains more than just guilty pleas, which were not within the scope of this research, leading us to merge the two data frames using a left join on the Sentencing data set.

With the Initiation data set, we created a few indicator variables for the felony class type, bond type, and bond amount. We created six indicators regarding felony class, as felony class can take on one of six values as mentioned previously: M, X, 1, 2, 3, and 4. Misdemeanors, or cases that began as felonies but were lowered to a misdemeanor charge, were identified by an indicator for whether a case was reclassified as a misdemeanor.

We also created indicators for bond type. There are four types of bonds granted to defendants: Individual bond, where the defendant pays nothing and is released on the promise that they will return to court for each scheduled court date and comply with all conditions of bail, Deposit bond, where the defendant must pay 10% of the bail amount set by a judge, Cash bond, where the defendant has to pay the full-face value of the bail amount ordered by a judge in order to secure release from custody, and No Bond, where there is no bond issued and the defendant remains in custody. We created an indicator variable "has_bond_derived" that takes on a value of 1 if the defendant was granted some type of bond and 0 if the defendant wasn't granted any type of bond. Lastly, we also created an indicator for the bond amount, "bond_amount_derived", where defendants granted cash bond had their bond amount listed in full and the defendants granted deposit bond had their bond amount listed as 10% of their original bond amount.

A.3. Police Stations Data.

We extracted four columns from the Police Stations data for our use: "DISTRICT", "DISTRICT NAME", "LATITUDE", and "LONGITUDE". To combine the data with our main data, we extracted the district number from the Sentencing column "LAW_ENFORCEMENT_UNIT", creating a new column called "DISTRICT" with just the district number. Then, we could perform a left merge with the combined Sentencing and Initiation data on the left and the Police Stations data on the right, so that all observations with a recorded law enforcement unit would also have the desired Police Stations data.

Notably, the "LATITUDE" and "LONGITUDE" columns were already in a cleaned form, so it was not necessary to change anything about the data after merging.

A.4. Imputation of Missing Data.

Most of the missing categorical data was imputed with 0, to indicate that the category was not present. For example, a defendant without a race recorded would be coded as 0 for "is_black_derived", "is_white_derived", "is_hisp_derived", and "is_othereth_derived". This was done for "ischicagopd_derived", all of the race indicator variables, "is_male_derived", "arrest_dayofyear_derived", "incident_dayofyear_derived", "senlength_derived", "is_incarceration_derived", "is_probation_derived", all of the is_district variables, all of the offense categories, all of the felony classes, "has_bond_derived", and "bond_amount_derived". The variables that were imputed with 0 that were not categorical had small enough missing percentages that it would not impact the model later on.

Some of the other, numerical data was imputed with the mean of that column, so that the missing data would not influence the model unduly. This was specifically done with "age_derived" (where outliers were also replaced with the mean), "arrest_year_derived", "ps_latitude_clean", and "ps_longitude_clean".

B. Data Analysis with Machine Learning.

B.1. Feature Selection and Training Data Creation.

To begin the machine learning, we first identified the feature and label columns. "senlength_derived", "is_incarceration_derived" and "is_probation_derived" were our target label columns. We then used the train test split function to define our training and testing data, with a testing data size of 20%. Our random seed was set to 1234 for the entire process, to ensure that our results could be replicated.

Additionally, we included a flag for the bond variables ("has_bond_derived" and "bond_amount_derived"), so that we could easily run our analysis either with or without them included as features. This was done because bonds are potentially confounding variables, which may or may not be fair to include as model inputs. In other words, we suspect that the underlying factors that determine sentencing outcomes have similar impacts on bond outcomes. Later on, we will describe how we controlled for the different felony classes, for the same reason.

To find the features that were most and least important to the data, we trained an extra trees classifier as recommended by scikit-learn's user guide on the subject. Our hyperparameters were 100 for "n_estimators", 5 for "max_depth", -1 for "n_jobs" (to enable parallelization), and the fixed random seed. We used a binary classifier model with the "is_incarceration_derived" label. Then, we ranked the most and least important features as identified by the model. Looking at the bottom of the list of features, we dropped some of the least influential. Among these features were "is_misdeameanor_derived", as well as the four variables indicating incident and arrest day of year- "cos_incident_dayofyear_derived", "sin_arrest_dayofyear_derived", "cos_arrest_dayofyear_derived", and "sin_incident_dayofyear_derived".

Additionally, while "is_black_derived" was highly influential, the other race indicators were not as much, and "is_hisp_derived" and "is_othereth_derived" only made up a very small portion of the data, so it was determined that only "is_black_derived" was needed, and the other binary race indicators could be discarded. Although "is_white_derived" was a relatively important feature, it is the inverse of "is_black_derived" in over 96% of cases, so we decided that one race indicator was sufficient to avoid data duplication. To validate the results of the classifier, we ran the .score() function on the testing data to see the model's accuracy.

B.2. Binary Classification.

After dropping the features with the lowest importance in the Extra Trees Classifier, we then performed a Gradient Boosting Classifier on the data, asking it to predict only the binary value of "is_incarcerated_derived". The hyperparameters for this model were "n_estimators" = 500, a "max_depth" of 5, and our constant random seed. After fitting the data, we again sorted the features by importance, to again identify what factors in the data were being used to make the predictions it was making, and used several of scikit-learn's metrics to measure the model's predictive ability, including the accuracy, precision, and recall scores, as well as the confusion matrix.

In addition to the Gradient Boosting Classifier, we also tried fitting a neural network to the data, again testing the model's ability to predict the binary values of "is_incarceration_derived". The hyperparameters for this model were "hidden_layer_sizes" of 50, an "alpha" of 0.00001, a "max_iter" of 500, and the constant random seed. For consistency, we measured this model using the same set of metrics.

B.3. Multiclass Classification.

We then wanted to see if we could generate a model that could predict incarceration as well as probation and release programs. To do this, we created a new target label that was a combination of the mutually-exclusive indicators "is_incarceration_derived" and "is_probation_derived", which will then have values of 0, 1, and 2. This was done using the formula:

$$new\_indicator = is\_probation + 2 * is\_incarcerated$$

Then, the train test split was recalculated with the new label set, now including the combination indicator, called "outcome_multiclass", still with 20% of the total data reserved for the testing section.

After creating the training and testing data sections, we trained multiclass gradient boosting classifier, with the same hyperparameters as that of the original binary class gradient boosting model. (n_estimators of 500, max_depth of 5, fixed random seed) We then tested the validity of the model using the same metrics (accuracy, precision, recall, confusion matrix) on the test data, and also found the most important features of the new model.

Using the same methods as with the binary classification, we also fit the multiclass data to the neural network classifier, with the same hyperparameters used as the binary neural network (hidden_layer_sizes of 50, alpha of 0.00001, max_iter of 500, and the fixed random seed.) After fitting the model with the training data, we again measured the model's effectiveness using the same metrics.

B.4. Subsetting by Felony Class.

We then subsetted the data into the different felony classes, to see what the model's effectiveness was when controlling for felony class by holding it constant in the training and testing data. For each felony class except "M" class, we removed all of the class indicators after subsetting, and then redefined the training and testing data with train test split, still assigning 20% of the data to the testing section. "M" class was excluded from this analysis because it only had 623 cases and was therefore too small to generate meaningful data. Then, the data in each section was fitted to both the binary and multiclass gradient boosting classifiers, with hyperparameters 300 for n_estimators, a max_depth of 5, and the fixed random seed. Each model was measured using the same metrics as above, and the features were ranked by importance for each individual felony class subset.

B.5. Sentence Length Regression.

To attempt to create a model that predicts sentence lengths, we first need to transform the sentence length label (senlength_derived) into a smooth distribution. Otherwise, the model would predict very short sentence lengths in all cases, as the raw distribution is heavily skewed in that direction. We tried multiple transformation methods– particularly a logarithmic transformation, a normalized logarithmic transformation, and the sigmoid of the normalized logarithmic transformation. The sigmoid of the normalized logarithmic transformation had the best normalizing affect, so this was the transformation used for the regression model. See Figure 1 for a visualization of these transformation steps.

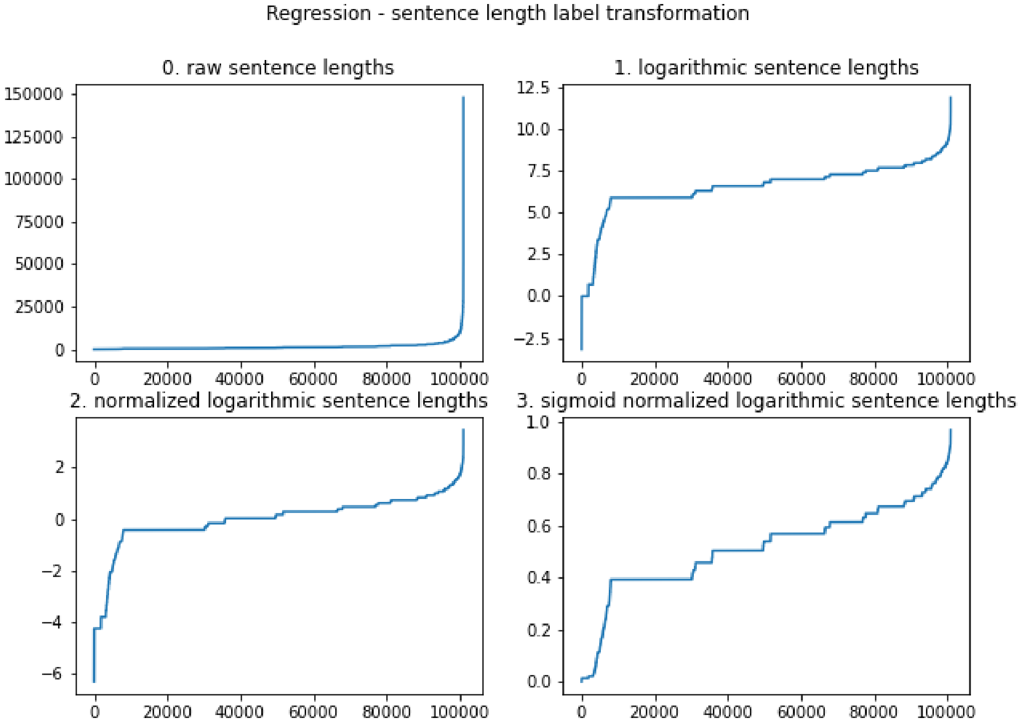

Fig. 1. Steps in transforming the sentence length label for the regression models. The goal was to transform raw data into a smooth bounded distribution without loss of information. Y-axes show the values after each transformation, while X-axes show the row number after sorting in ascending order of sentence length.

After normalizing, the train test split process was the same as all other models in our methods, with a 20% testing size. To create the regression model, we first trained a gradient boosting regressor, with hyperparameters of 500 for n_estimators, a max_depth of 3, and the fixed random seed. We validated the model with the score() method, and found the feature importances. We repeated the process with a neural network regressor as well, with hyperparameters of 50 for hidden_layer_sizes, an alpha of 0.0001, max_iter of 1000, and the fixed random state. We then validated the model with the score() method, allowing the comparison between the neural network and gradient boosting regressors.

3. Results

A. Binary Classification.

After cleaning and merging our data sets and after dropping features with the lowest importance, we were able to train a gradient boosting classifier to predict the binary value of whether or not an individual was incarcerated. The binary gradient boosting classifier with bond features had an accuracy of 71.3 percent, and without bond features had an accuracy of 68.9 percent. Tables 3 and 4 show the accuracy, precision, and recall scores for all models, with and without bond features respectively.

We also identified the most important features used to make this prediction, according to the interpretable gradient boosting model. Table 9 refers to the top five features used to make predictions regarding whether or not an individual is incarcerated. The most important features (excluding bond features) were whether or not the defendant was male, whether or not the defendant was black, their age, whether or not the defendant was charged with a Class 4 felony, and whether or not the defendant was charged with a Class X felony. Including bond features, the most important features became the bond amount, whether or not the defendant was black, whether the defendant had bond, whether or not the defendant was male, and the defendant's age. Examining the confusion matrix for the binary gradient boosting classifier, Table 5 demonstrates that the class accuracy for correctly predicting incarceration is .68 and the class accuracy for correctly predicting not incarceration is .74 when including bond features. As Table 7 show, when not including bond features, the class accuracy for correctly predicting incarceration is .72 and the class accuracy for correctly predicting not incarceration is .65. Since these accuracy rates aren't as high as we would prefer them to be, sentencing outcomes seem to be only somewhat predictable based on case/defendant data when using binary classification, predicting for whether or not an individual is incarcerated.

In all cases, the binary neural network classifier performed slightly worse than the binary gradient boosting classifier. Due to this performance gap, as well as the fact that neural networks are non-interpretable black box models, further analysis of the binary neural network classifier would be redundant. The model's scores are included only for the sake of comparison.

B. Multiclass Classification.

In addition to predicting a binary value of whether or not a defendant is incarcerated, we were also able to predict whether or not a defendant was sentenced to incarceration, probation, or neither (release). With bond features included, this classifier had an accuracy of 68.6 percent, and without bond features, it had an accuracy of 66.5 percent. Tables 3 and 4 show the accuracy, precision, and recall scores for all models, with and without bond features respectively.

As shown in Table 9, the features most important to the model's ability to predict incarceration, probation, and release were almost identical to the features most important in predicting whether or not a defendant was incarcerated. From most to least important and excluding bond features, these features were whether or not the defendant is male, whether or not the defendant is black, the age of the defendant, whether or not the defendant was charged with a class 4 felony, and whether or not the crime involved the use of a weapon. Including bond features, the most important features for the multiclass gradient boosting model were the bond amount, the age of the defendant, whether or not the defendant has bond, whether or not the defendant is black, and whether or not the defendant is male.

Tables 6 and 8 are confusion matrices that demonstrate the ability of the multiclass gradient boosted classifier to classify release, probation, and incarceration correctly. The model almost always failed to classify release, likely because release made up a very small portion of our data. However, the model had higher accuracy for incarceration and probation, meaning that the model has learned differences between probation and incarceration cases. Adding bond features didn't change the accuracy of class prediction by much for the model's ability to identify release and incarceration, but it did increase the class accuracy of probation by around 3 percent. Since multiclass classification almost never correctly predicts release correctly, we were more interested in the results of the binary confusion matrices, as by predicting whether or not an individual is incarcerated, this classifier effectively groups together probation and release as not incarceration and subsequently has higher class accuracy for not incarceration than either of the multiclass classifiers did for either probation or release.

The accuracy of both binary and multiclass classification are quite similar, both hovering in the mid-to-high 60s percentage range, regardless of whether bond features are included or not. The similar accuracy across models suggests that there are inherent limitations in the predictability of sentencing outcomes for this data set. That being noted, "is_black_derived" being the most important feature for both binary and multiclass classification when bond features aren't included does have frightening implications, demonstrating that for the whole data set, both models have been trained to place the most importance on whether or not the defendant is black when predicting the sentencing outcomes for an individual, indicating racial disparities.

C. Subset by Felony Class.

Beyond creating models predicting incarceration across the entire data set, we were interested in seeing how effective the model was once felony class was controlled for. Felony data was fitted to the multiclass gradient boosted classifier, enabling us to see the most important features used by the model for each felony class including bond features, as is shown in Table 10. Felony classes are shown left to right from most severe to least severe. The accuracy of the multiclass classifier on the felony subsets including bond features

Table 2. Data Overview

Data Set	Rows	Columns	Start Date	End Date
Sentencing	188914	41	October 10th, 2003	September 9th, 2021
Initiation	1041304	5 (directly selected target columns)	October 10th, 2011	September 9th, 2021
Police Station	23	4 (directly selected target columns)	N/A	N/A
Our Data (combined_all)	188914	31 (all derived columns)	October 10th, 2003	September 9th, 2021

An overview of the size and scope of the data used.

is shown in Table 4. At first look, the accuracy score for the multiclass classifier for Felony Class X looks very high, with an accuracy score of .86, but this is misleading as Class X felonies are easier to predict as the accuracy score is usually approximately equal to the proportion of cases ending in incarceration, and since Class X felonies are more severe, this proportion is higher. The accuracy scores for the rest of the felony classes are on par with the rest of our accuracy scores. Referring to Table 11, for felony classes X, 1, 3, and 4, age is the most important feature in predicting sentencing outcomes. For Class 2, the use of a weapon in the crime is the most important feature in predicting sentencing outcomes. Other binary indicators for offense categories, like "is_narcotics_"derived, "is_escape_derived", or "is_theft_derived" appear a few times as one of the most important features across felony classes that include offenses relating to the binary indicator. For example, "is_narcotics_derived" appears as one of the most important features in predicting sentencing outcomes for Class 1 felonies, which are classified as very serious crimes and which includes the possession of heroin, cocaine, or opiods.

Race of the defendant only appears as one of the most important features for Class 2 and Class 4 felonies. For Class 4 felonies, it is the second most important feature in determining sentencing outcomes. Theft is often categorized as a Class 4 felony and the binary indicator for whether the crime involves theft is less important than the indicator for race. Figures 2 and 3 visualize the most important demographic features for each felony class, for binary and multiclass classification models, respectively. Race becomes much more important as a feature for class 4 felonies relative to the other subsets. This suggests that racial biases play a greater role in sentencing outcomes for less severe crimes.

D. Regression Models.

Lastly, beyond understanding if case/defendant data can predict whether an individual is sentenced to incarceration or probation, we were also interested in whether sentence length is predictable using this data. The sigmoid of the normalized logarithmic transformation had the best normalizing effect, so our regression models were based off of this transformation. We trained both a gradient boosting regressor and a neural network regressor, with R^2 coefficients of 0.469 and 0.415 respectively with bond features included, and 0.453 and 0.424 respectively without. In other words, sentence lengths are not at all predictable by regression models using case and defendant data.

4. Discussion and Conclusion

Our predictive modeling of sentencing outcomes certainly faced some limitations. As suggested by the similarities between the metrics of the different types of predictive models, it is likely that these limitations are inherent in the data itself. Another limitation of ouranalysis occurred early in our data cleaning, since our two data frames did not consist of data from the same date range, creating missingness of key outcomes that ended up being highly important features for our models when predicting sentencing outcomes, such as age, bond amount, and bond type. Additionally, in our multiclass classification, the ability of the model to accurately predict release was extremely low, primarily because release was a small section of our data. With limitations from the data sets themselves, missingess that was created as a result of different date ranges, and difficulty in predicting whether or not a defendant is released, the accuracy of our models should demonstrate that there are difficulties with predicting sentencing outcomes only using case/defendant data. Furthermore, our data do not support the idea that case/defendant data alone can be used to predict sentencing lengths.

Table 3. Model Results (Including Bond Features)

Full Data Analysis			
Model Type	Accuracy	Precision	Recall
Binary Gradient Boosting Classifier	0.7133	0.7130	0.7133
Binary Neural Network Classifier	0.7016	0.7024	0.7016
Multiclass Gradient Boosting Classifier	0.6861	0.6765	0.6861
Multiclass Neural Network Classifier	0.6794	0.6748	0.6794
Felony Class Subsets (Binary)			
Felony Class	Accuracy	Precision	Recall
Class X (n = 11628)	0.8641	0.8424	0.8641
Class 1 (n = 23258)	0.7132	0.7080	0.7132
Class 2 (n = 42369)	0.7232	0.7223	0.7232
Class 3 (n = 28596)	0.7037	0.7023	0.7037
Class 4 (n = 66970)	0.6807	0.6782	0.6807
Felony Class Subsets (Multiclass)			
Felony Class	Accuracy	Precision	Recall
Class X (n = 11628)	0.8594	0.8347	0.8594
Class 1 (n = 23258)	0.7012	0.6926	0.7012
Class 2 (n = 42369)	0.6949	0.6853	0.6949
Class 3 (n = 28596)	0.6535	0.6422	0.6535
Class 4 (n = 66970)	0.6427	0.6357	0.6427
Regression Models			
Regression Model	R^2 coefficient		
Gradient Boosting Regressor	0.4690		
Neural Network Regressor	0.4152		

Test scores of all machine learning models. Precision and accuracy scores were calculated using the weighted average method.

Table 4. Model Results (Not Including Bond Features)

Full Data Analysis			
Model Type	Accuracy	Precision	Recall
Binary Gradient Boosting Classifier	0.6894	0.6890	0.6894
Binary Neural Network Classifier	0.6834	0.6850	0.6834
Multiclass Gradient Boosting Classifier	0.6648	0.6549	0.6648
Multiclass Neural Network Classifier	0.6598	0.6537	0.6598
Felony Class Subsets (Binary)			
Felony Class	Accuracy	Precision	Recall
Class X (n = 11628)	0.8568	0.8298	0.8568
Class 1 (n = 23258)	0.6814	0.6729	0.6814
Class 2 (n = 42369)	0.6932	0.6901	0.6932
Class 3 (n = 28596)	0.6563	0.6539	0.6563
Class 4 (n = 66970)	0.6593	0.6569	0.6593
Felony Class Subsets (Multiclass)			
Felony Class	Accuracy	Precision	Recall
Class X (n = 11628)	0.8517	0.8207	0.8517
Class 1 (n = 23258)	0.6713	0.6560	0.6713
Class 2 (n = 42369)	0.6696	0.6573	0.6696
Class 3 (n = 28596)	0.6180	0.6075	0.6180
Class 4 (n = 66970)	0.6208	0.6101	0.6208
Regression Models			
Regression Model	R^2 coefficient		
Gradient Boosting Regressor	0.4526		
Neural Network Regressor	0.4239		

Test scores of all machine learning models. Precision and accuracy scores were calculated using the weighted average method.

Table 5. Binary Gradient Boosting Confusion Matrix (Including Bond Features)

Actual Class	Predicted Class		Class Accuracy
	Not Incarceration	Incarceration	
Not Incarceration	11879	5575	0.6806
Incarceration	5256	15073	0.7415

The confusion matrix for the binary gradient boosting classifier, including bond features.

Table 6. Multiclass Gradient Boosting Confusion Matrix (Including Bond Features)

Actual Class	Predicted Class			Class Accuracy
	Release	Probation	Incarceration	
Release	75	949	319	0.0558
Probation	69	10409	5633	0.6461
Incarceration	43	4847	15439	0.7595

The confusion matrix for the multiclass gradient boosting classifier, including bond features.

Table 7. Binary Gradient Boosting Confusion Matrix (Not Including Bond Features)

Actual Class	Predicted Class		Class Accuracy
	Not Incarceration	Incarceration	
Not Incarceration	11399	6055	0.6531
Incarceration	5680	14649	0.7206

The confusion matrix for the binary gradient boosting classifier, not including bond features.

Table 8. Multiclass Gradient Boosting Confusion Matrix (Not Including Bond Features)

Actual Class	Predicted Class			Class Accuracy
	Release	Probation	Incarceration	
Release	72	883	388	0.0536
Probation	74	9929	6108	0.6163
Incarceration	42	5170	15117	0.7436

The confusion matrix for the multiclass gradient boosting classifier, not including bond features.

Table 9. Most Important Features for Interpretable Full-Data Models

Not Including Bond Features		**Including Bond Features**	
Binary Gradient Boosting	**Multiclass Gradient Boosting**	**Binary Gradient Boosting**	**Multiclass Gradient Boosting**
is_male_derived, 0.154	is_male_derived, 0.142	bond_amount_derived, 0.271	bond_amount_derived, 0.253
is_black_derived, 0.134	is_black_derived, 0.123	is_black_derived, 0.095	age_derived, 0.095
age_derived, 0.115	age_derived, 0.116	has_bond_derived, 0.094	has_bond_derived, 0.087
is_fclass4_derived, 0.091	is_fclass4_derived, 0.082	is_male_derived, 0.092	is_black_derived, 0.087
is_fclassX_derived, 0.0723	is_weapon_derived, 0.071	age_derived, 0.090	is_male_derived, 0.087

Top five features in each interpretable full-data model, with feature importance scores.

Table 10. Most Important Features for Felony Class Subsets (Including Bond Features)

Class X	Class 1	Class 2	Class 3	Class 4
is_narcotics_derived, 0.179	bond_amount_derived, 0.233	is_weapon_derived, 0.211	bond_amount_derived, 0.178	bond_amount_derived, 0.268
age_derived, 0.176	age_derived, 0.2125	is_bond_amount_derived, 0.198	is_male_derived, 0.144	age_derived, 0.134
bond_amount_derived, 0.160	is_male_derived, 0.088	age_derived, 0.148	age_derived, 0.137	is_black_derived, 0.117
incident_year_derived, 0.078	has_bond_derived, 0.073	is_male_derived, 0.077	has_bond_derived, 0.086	is_theft_derived, 0.073
is_male_derived, 0.063	arrest_year_derived, 0.063	has_bond_derived, 0.063	incident_year_derived, 0.055	is_male_derived, 0.056

When subsetted by felony class, the top five most important features identified by the multiclass gradient booster model.

Table 11. Most Important Features for Felony Class Subsets (Not Including Bond Features)

Class X	Class 1	Class 2	Class 3	Class 4
age_derived, 0.235	age_derived, 0.281	is_weapon_derived, 0.262	age_derived, 0.203	age_derived, 0.184
is_narcotics_derived, 0.209	is_male_derived, 0.135	age_derived, 0.198	is_male_derived, 0.180	is_black_derived, 0.178
incident_year_derived, 0.114	arrest_year_derived, 0.113	is_male_derived, 0.118	incident_year_derived, 0.075	is_theft_derived, 0.120
arrest_year_derived, 0.084	incident_year_derived, 0.82	is_black_derived, 0.072	is_escape_derived, 0.073	is_male_derived, 0.076
is_male_derived, 0.080	is_narcotics_derived, 0.068	arrest_year_derived, 0.061	arrest_year_derived, 0.069	arrest_year_derived, 0.076

When subsetted by felony class, the top five most important features identified by the multiclass gradient booster model.

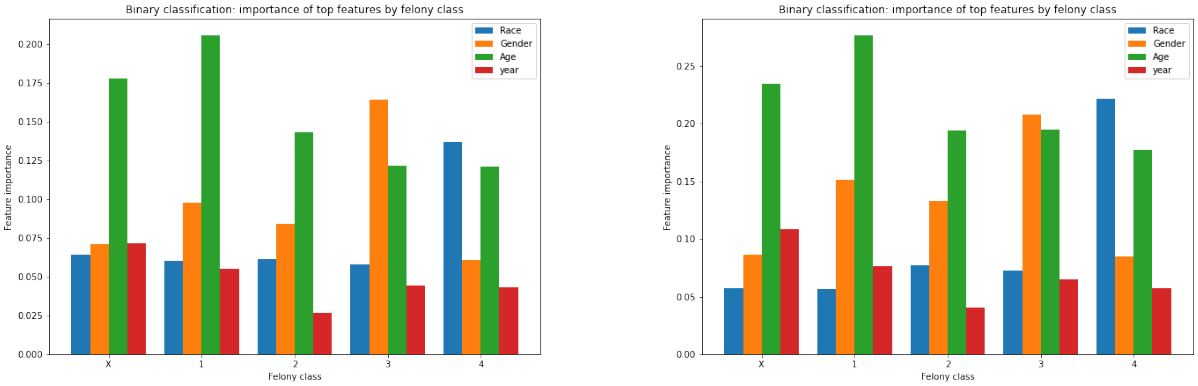

Fig. 2. Important demographic features for each felony class subset according to the binary gradient boosting classifier. Left is including bond features, right is not including bond features. Note that Race becomes a much more important feature for class 4 felonies.

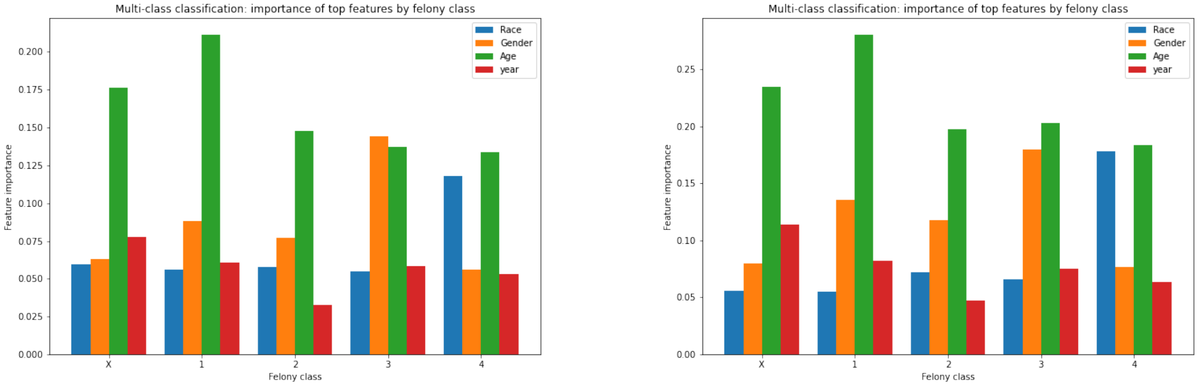

Fig. 3. Important demographic features for each felony class subset according to the multiclass gradient boosting classifier. Left is including bond features, right is not including bond features. Note that Race becomes a much more important feature for class 4 felonies.

Though we do examine difficulties with using case/defendant data to predict sentencing outcomes, this study has many further applications to expand upon initial results. One way in which future work can improve upon our results is by using Police Stations Data from Cook County and intersecting it with Chicago census tracts to add further demographic features. Adding census tract data has the potential to increase the accuracy of our predictive models, but there are also limitations here, as missing values for "LAW_ENFORCEMENT_UNIT", which is the Law Enforcement Unit within Chicago Police Department associated with the arrest, is our main source of missing data. Furthermore, future studies interested in examining the predictability of sentencing outcomes using case/defendant data could limit the data set to only include males, as "is_male_derived" appeared as the most important feature for both binary and multiclass classification excluding bond features, but this may have been so because most of the defendants in our data set were male. Lastly, another way future work can expand upon the results of our study is by creating an indicator for "is_black_derived" that goes beyond race of the defendant. In a study published by Ellis Monk, he finds that "skin tone is significantly associated with the probability of having been arrested and/or incarcerated, net of relevant controls," with chance for incarceration among African Americans increasing in a directly proportional manner with the darkness of their skin, ranging from very light to very dark.[7] Though we are unaware of a data set that examines skin tone for our population of interest, looking into whether skin tone among African American defendants is an important factor in predicting sentencing outcomes could have interesting implications regarding the presence colorism within racial disparities in the criminal justice system.

Though adjustments can be made to our models to more accurately predict sentencing outcomes, we believe that our results shed light on the features that are most predictive of incarceration and other sentencing outcomes. These features do indicate racial disparities, creating concern regarding the disproportionate harm the criminal justice system puts on Black people.

ACKNOWLEDGMENTS. Thank you to Professor Johnson and the rest of the QSS 20 teaching team for providing guidance and feedback throughout the term!

1. USS Commission, Demographic differences in sentencing (year?).
2. B Pearsall, Predictive policing: The future of law enforcement? *NIJ J.* (2010).
3. WD Heaven, Predictive policing algorithms are racist. they need to be dismantled. (https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-crimina (year?) Accessed: 2022-03-11.
4. Z Jilani, Class is a more potent predictor of incarceration than race. but racism drives it (https://theintercept.com/2018/02/05/mass-incarceration-class-predictor-race/) (2018) Accessed: 2022-03-11.
5. I Landon, Encoding cyclical continuous features - 24-hour time (https://ianlondon.github.io/blog/encoding-cyclical-features-24hour-time/) (2019) Accessed: 2022-03-11.
6. O Mitchell, The relationship between race, ethnicity, and sentencing outcomes: A meta-analysis of sentencing research. (2004).
7. E Monk, The color of punishment: African americans, skin tone, and the criminal justice system. *Ethn. Racial Stud.* **42** (2018).