Variable Weight in Determining Sentence Length for Federal Cocaine Trafficking Offenders

DS 3001 – December 14, 2024

Anna Girerd, Marisa Guajardo, Owen Himmel,
Clarissa Kusmana, Christine Lee

ABSTRACT

Federal judges are responsible both in charging offenders with a criminal offense, and mandating a sentence length. This evidently involves subjective decision-making, though judges attempt to serve unbiased. With the usage of machine learning, however, data scientists have the power to analyze whether these decisions are, in fact, unbiased, and whether certain trends and patterns can be elucidated. Using federal defendant charges and sentencing data from October 2018 to October 2019, sourced from the U.S. Sentencing Commission, we investigate which factors are most predictive of sentence length for federal cocaine trafficking offenders, and how plea deals impact these predictions. Through our usage of decision tree, random forest, and LASSO regression modeling within the Python programming language, we were able to discover that the variables of most importance in determining sentence length are acceptance of a plea deal, final offense level, drug quantity, and weapon enhancements. This illustrates less of a trend in bias, as offense-based characteristics carried more weight than individual characteristics. More research is needed, however, to validate the shortcomings in our modeling.

INTRODUCTION

Prior to the creation of the U.S. Sentencing Guidelines in 1984, federal judges had broad discretion in sentencing with little to no oversight. The Guidelines were created to address unwanted sentencing disparities that arose from the bias involved in judges' subjective decision-making. Under the current Guidelines, judges are prompted to uniformly sentence defendants with identical criminal histories and crimes, regardless of their demographic characteristics and the geographical location of the District Court. Despite this effort to make sentencing more equitable and transparent, disparities still persist, which raises questions about the factors influencing sentencing outcomes and why such disparities are present.

For this project, we are focusing on addressing the question "Do the factors outlined in the Federal Sentencing Guidelines, along with defendants' demographic factors, influence the sentence length of a defendant who is charged with cocaine trafficking?" with a focus on the impact of the defendant's demographic characteristics and plea/trial status. We believe this question is worth exploring because even though Sentencing Guidelines stipulate that judges uniformly sentence defendants with identical criminal histories and severity of crime, differences in demographic factors still result in sentencing disparities. We are especially interested in exploring whether defendants' demographics, the characteristics of the crime, and the acceptance/rejection of plea deals influence their sentence lengths.

In addition to predicting sentence length using the given variables, our models will also assess which variable has the most significant impact on a defendant's sentence length. This allows us to investigate whether extralegal factors that are discouraged by the U.S. Sentencing Guidelines from being taken into account when imposing a sentence (namely specific offender characteristics like race, age, gender, and level of education) are weighted equitably against other offense-specific variables.

For the analysis portion, we used supervised learning techniques, including linear regression, random forests, and LASSO regression, to identify and quantify the impact of various factors on sentencing length. Our predictors include demographic characteristics (e.g., age, gender, race, education level), offense-specific factors (e.g., drug quantity, presence of weapon enhancements), and trial outcomes (e.g., plea versus trial). We hypothesized that factors such as offense severity, represented by the Base Offense Level, and the presence of mitigating or aggravating circumstances would emerge as significant predictors of sentence length. In addition, we anticipated that plea deals would substantially reduce sentence length, even after controlling for other factors.

DATA AND BACKGROUND

Our data includes federal defendant charges and sentences over the fiscal year 2018-2019 in the United States. The data set is sourced from the U.S. Sentencing Commission, which collects data on federal sentencing trends annually from all federal District Courts across the country. For the purpose of generalizability, and to avoid applying the conclusions to defendants convicted of unrelated offenses, we have narrowed down the population of interest to only defendants who had been convicted of trafficking either powder or crack cocaine. We are not distinguishing between the two groups of cocaine (crack versus powder) because following several amendments, the Sentencing Guidelines indicate that similar quantities of crack and powder cocaine should be sentenced similarly. We are particularly interested in plea deals based on our EDA suggesting that subjects who accept a plea deal have shorter sentences. We'd like to dig into this to see if, for example, the plea deal subjects are from a more affluent background than those of non plea deal.

The dataset is very large, which could make it difficult to explore the data, but we hope to resolve this issue by narrowing down our focus to one type of crime. Additionally, the variable names are not intuitive and identifying them requires tedious investigation of the codebook. There are also some missing values, and the code book does not address clearly how to go about handling these. We will have to come up with a method of cleaning the dataset to filter out these missings. The data dictionary is almost useless because it is extremely lengthy and requires a lot of sifting through to obtain the context needed.

The Sentencing Guidelines ensure that judges uniformly sentence defendants with identical criminal histories and severity of crime. However, despite the application of the Sentencing Guidelines, some disparities are still present in the data. Although we require further EDA and analysis to pinpoint the source of the disparity, sentencing is generally considered to be

inherently unpredictable. Perhaps the capriciousness of sentencing decisions shows that the Guidelines are operating properly, as judges have the discretion to individualize sentences according to the circumstances of each defendant – such as the presence of mitigating or aggravating factors. On the other hand, this makes creating a predictive model difficult, as the data may not fully capture all nuances in the observations.

METHODS

We decided to focus on which factors are most predictive of sentence length for federal cocaine trafficking offenders, and how plea deals impact these predictions. We were interested in testing the predictive value of a wide range of variables, including identity-based variables such as age, gender, race, and education level, specific offense characteristics like severity of crime (Base Offense Level), weapon enhancement charge, drug quantity, type of drug, and trial outcomes like whether a plea deal was taken. In our data, each row represents an offender. Observations include many hundreds of data points including basic biographical info, location, crime, and sentencing.

To analyze the data, we chose to use supervised learning because there is a clear response variable, which is the sentence length. The dataset also provides labeled examples, where we have explanatory variables (age, gender, race, education level) and our response variable (sentence length). We're interested in understanding how different factors (including plea deals) influence the sentence length, which supervised learning models can help elucidate through feature importance and coefficient analysis. We plan to use a linear regression model to capture relationships between explanatory variables (such as age, weapon enhancement, education level, citizenship status) and sentence length. We will build a model with the desired variables and compare the weights of standardized coefficients to determine those with the greatest effect. Then, cross validating the model on unseen data we can determine importance by removing variables and evaluating change in accuracy. To determine explanatory variables of importance we will use decision trees in the form of a random forest. Based on results from the trees we can make informed decisions on which variables contribute most to sentence length prediction.

We believe that using regression modeling is beneficial, even if the model's overall predictive accuracy is limited, due its high interpretability power. Regression analysis quantifies the effects of both defendant characteristics and offense-specific factors on sentence lengths, allowing us to assess whether certain demographic groups experience systematic differences in sentencing outcomes. Judges have the discretion to individualize sentences to the circumstances of each defendant instead of blindly following calculations, since the Sentencing Guidelines empower courts to take into account aggravating/mitigating circumstances. This flexibility is likely to result in sentencing disparities and inconsistent trends, making the sentencing process unpredictable and will likely result in limited predictive power for the model.

We can be confident our approach works if the variables we find important while building and training the model show similar importance on test data during cross validation.

Building from this, a successful predictive algorithm would achieve high accuracy (high R^2 score for regression or mean absolute error below a certain threshold) for sentence lengths.

There are a few challenges that we anticipate. First, multicollinearity might be an issue, as some variables may be highly correlated, which could distort model interpretations. For instance, the Base Offense Level is directly calculated based on the weight of the drug involved, with larger quantities of cocaine resulting in a higher Level that reflects the severity of the offense. Using variance inflation factor (VIF) checks can help detect this. Furthermore, if multicollinearity is an issue we could use PCA to decorrelate variables. Another challenge could be missing data. Many subjective factors that influence sentencing outcomes, such as a judge's perception of the defendant's remorse or reasons for departures from the Guidelines, may not be captured in the dataset. If there are missing values, especially in key variables, this could reduce model accuracy. Techniques like imputation or focusing on subsets with complete data may help.

Additionally, there is a very large number of independent variables. Given the 100s of variables for each observation, determining “the most predictive” is a challenge. We identified variables of interest through EDA and research such as age, gender, race, education level, citizenship status, and weapon enhancement. We can establish a framework to identify the most effective predictors of sentence length among selected variables, such as x, y, and z, rather than evaluating all possible predictors.

Finally, interpretability may be difficult with this analysis. Strength and direction of variable effects may not be immediately clear from the random forest model. Linear regression coefficients should illustrate importance better. However, if PCA is needed or if there are many variables, the interpretation of coefficients may be a challenge. The combination of techniques to identify and quantify variables will hopefully combat this, then a thorough examination of what interactions between variables makes sense could help determine meaning from the numbers.

RESULTS AND ANALYSIS

For our project,we wanted to analyze the data to investigate: **Which factors are most predictive of sentence length for federal cocaine trafficking offenders, and how plea deals might impact these predictions?** These results and modeling approach could, in turn, be utilized by the Department of Justice or for identifying bias in how sentence length may be determined.

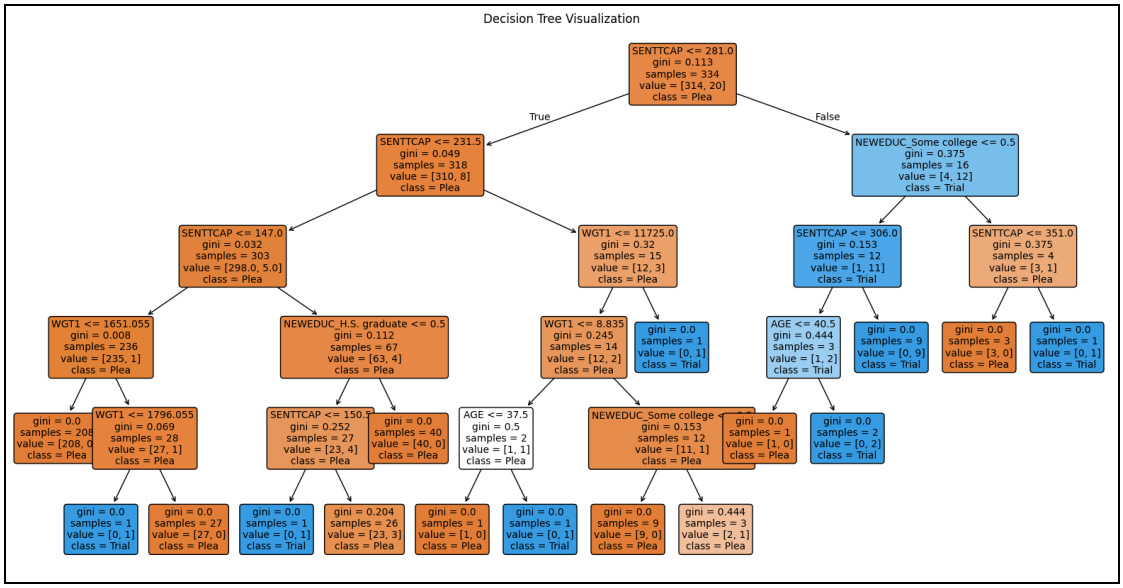

Figure 1: Variable Weights in Plea vs Trial Outcome

The dataset we used was very large and included hundreds of variables. For this reason, we focused on identifying the variables of most importance to our analysis. After cleaning the data, we created a decision tree classifier object with a maximum depth of 5 to prevent overfitting. This decision tree (Figure 1) predicts whether a case is classified as "Plea" or "Trial" based on key features, such as sentence length (SENTTCAP), drug weight (WGT1), and education level (NEWEDUC) of the defendant. From the beginning, SENTTCAP appears to be the most influential factor, with the first split occurring at SENTTCAP <= 281.0. Cases with sentence lengths below the 281-month threshold are predominantly classified as "Plea," while cases with longer sentences move to the right for further evaluation.

As the tree branches out, additional splits refine the classification. For instance, shorter sentences (where SENTTCAP <= 147.0) almost exclusively fall into the "Plea" category, with a high degree of confidence, indicated by the Gini index close to 0. Meanwhile, cases with longer sentences are more likely classified as "Trial," especially combined with factors such as higher education levels or higher drug quantity. Education level also plays a notable role, with defendants having "Some college" are less often classified as "Plea," while higher education levels increase the likelihood of getting a "Trial" classification. Similarly, drug quantity appears in several splits, which means it plays an important role in distinguishing between "Plea" and "Trial" cases in borderline scenarios.

Finally, the tree shows that shorter sentences and lower education levels strongly correlate with the "Plea" group. On the other hand, longer sentences and higher education levels are more correlated with "Trial" observations. The model demonstrates high confidence in many of its classifications, with several leaf nodes achieving perfect purity, where the Gini index is almost 0. However, there does exist an imbalance in class distribution, with "Plea" as the majority, and over 96% of defendants in this population within the dataset taking a plea deal and only 4% taking the case to trial. This imbalance is also present in the structure of the tree.

Specifically, at the root node, the value is [314, 20], which means 314 cases are classified as "Plea" and only 20 as "Trial."

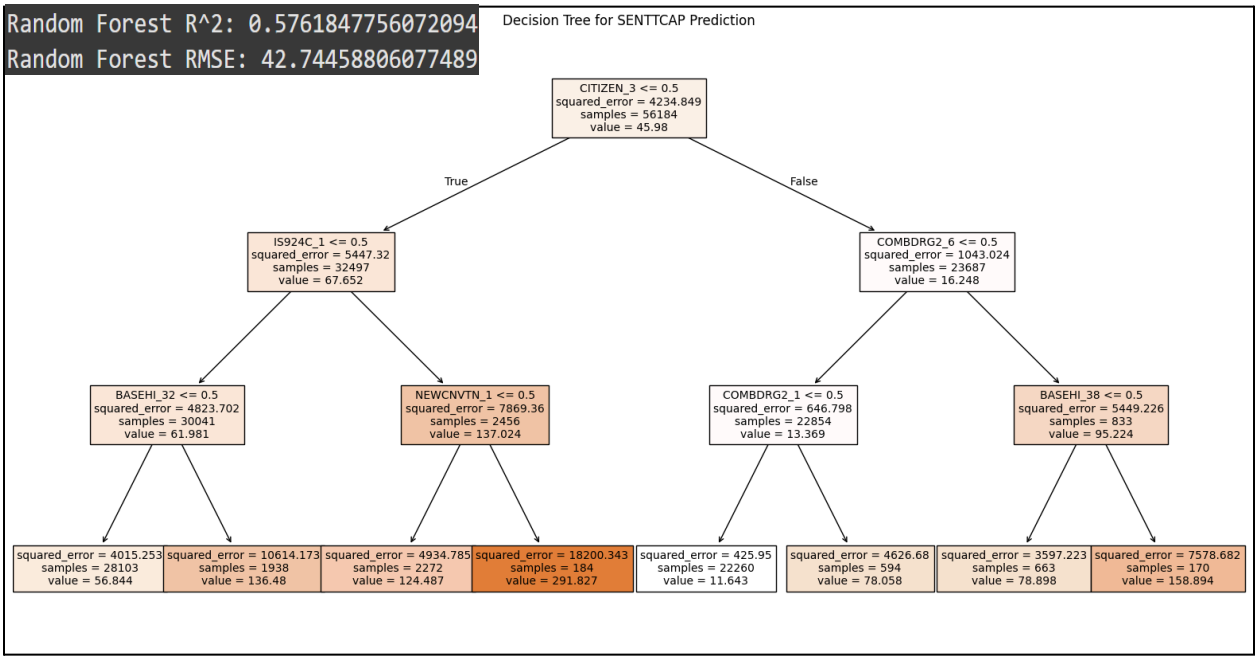

Figure 2: Predicting Sentence Length with Random Forest

Figure 2 shows a sample tree and accuracy statistics for a random forest model predicting sentence length. Variables of interest were citizenship, race, age, education, drug type, and weapon enhancement.The sample decision tree split at illegal alien status, and was followed by splits at weapon enhancement and drug type. Age, race, and education did not appear in the sample tree. Using an 80/20 train test split the model had little efficacy. The RMSE metric showed an average of a ~42 month difference between predicted sentences and real sentences. This suggests a strong omitted variable bias. Because of the complexities of criminal sentencing predictions would need to rely on either more inputs or more nuanced variable choice. Regarding ability for examining potential discrimination, a data set would need to be constructed with near identical crimes. Demographic data appears to not hold significant importance when compared to things like weapon enhancement or drug type. Given the inaccuracy of the model we decided to further refine our predictive strategy with the addition of Random Forest and LASSO Regression modeling as outlined below.

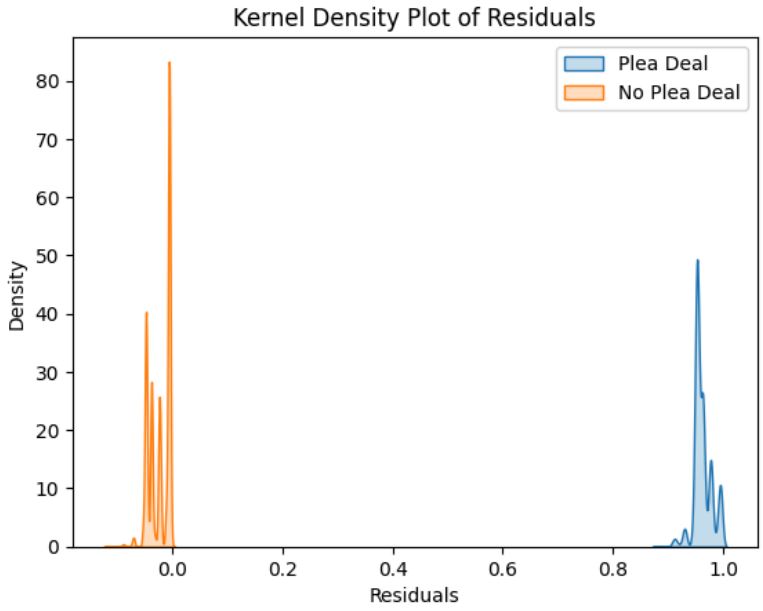

Figure 3: Nuisance Variable Influence on Plea vs Trial Outcome

Next, as demonstrated in Figure 3, we split the data into two features, nuisance variables (X_nuisance) and the target variable (y_target). We used this to train a Random Forest Regressor to predict the target variable using only the nuisance variables. The maximum depth parameter is set to 5 to control the complexity of the model. The model's predictions are used to calculate residuals, which represent the difference between the actual target values and the predicted values. We created a kernel density plot (Figure 3) to visually compare the distribution of residuals from the Random Forest Regressor for two groups: cases with a plea deal and cases without a plea deal. This helps assess whether the nuisance variables have a different impact on the prediction of plea deals for these two groups.

The sharp separation indicates that the two groups are fundamentally different, and nuisance variables alone cannot account for the distinction. This suggests a strong relationship between plea deals and other unmodeled factors and demonstrates that plea deals and no plea deals differ significantly, and the differences are not explained by nuisance variables. The model explains only 2% of the variation in plea deals using dummy-encoded nuisance variables with an R-squared of 0.02. The dataset is heavily imbalanced with 22,230 No Plea Deals versus 589 Plea Deals, which skews the model’s performance innately. We would need to address the class imbalance to improve predictions for plea deals.

The influence of whether a plea or trial is taken evidently is large on sentence length outcome. For this reason, we decided to perform a LASSO regression analysis to look at a larger

number of variables and their influence. The resulting visualization is quite large, so we've cropped the image to include the important variables.

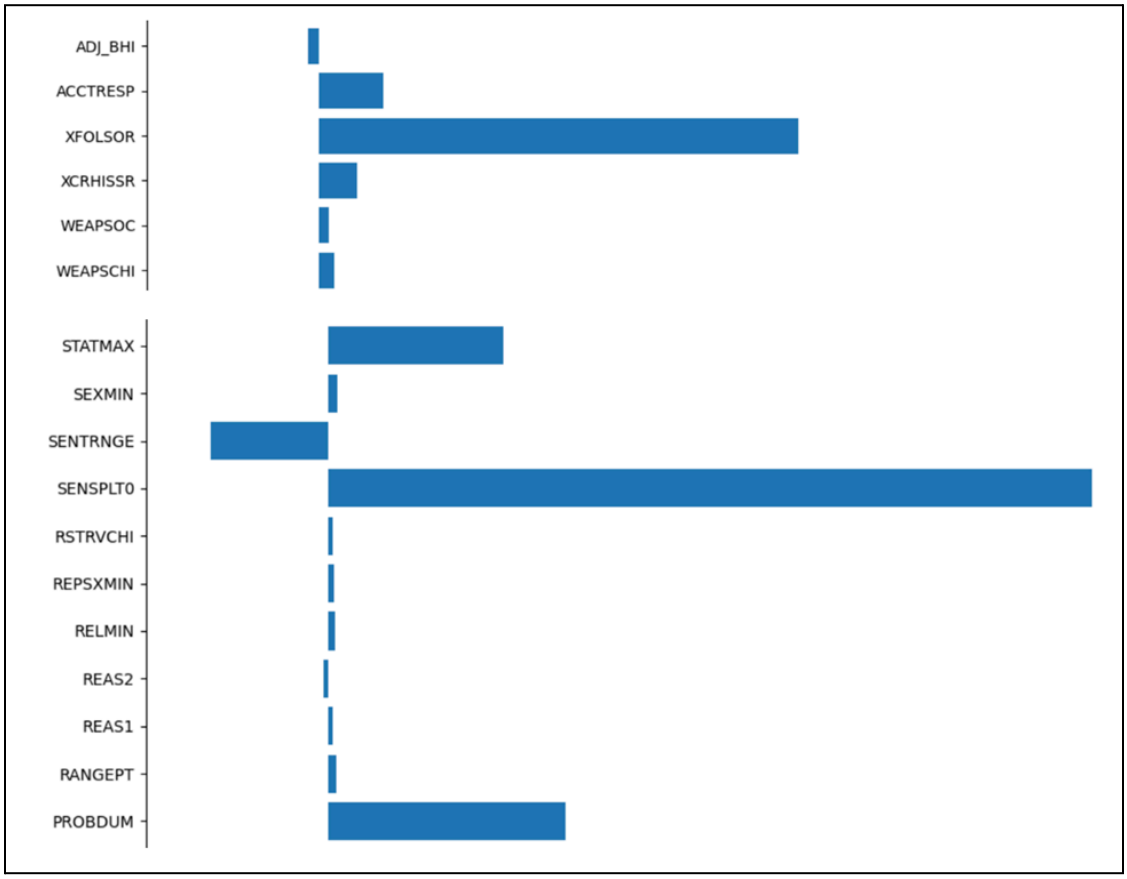

Figure 4: LASSO Variable Weight in Sentence Length

The Lasso reduced the feature set from 211 to 86, making the model simpler and more interpretable without significantly compromising performance. SENSPLT0 was a key feature (Figure 3), but this variable is total prison sentence in months with zeros, which is essentially the same as our response variable, so we disregarded this finding. The variable WEAPSOC (Figure 4) indicates whether there is a Specific Offense Characteristic (SOC) enhancement related to the use or presence of a weapon in a case. If a weapon was involved in the commission of a crime (e.g., carrying, brandishing, or using a weapon), a SOC enhancement could increase the severity of the sentence. While the Lasso indicates that the presence of a weapon enhancement is relevant in the prediction of sentencing outcomes, it surprisingly isn't as salient as other variables. XFOLSOR (Figure 4) seems to have a stronger influence. This is a numeric variable for final offense level as determined by the court on a scale of 1 to 43. PROBDUM is another variable

with a relatively stronger influence, and this variable indicates whether the defendant received probation.

We performed a regression analysis on the Lasso non-zero coefficients to see how well the model worked. The R^2 and adjusted R^2 of 0.94 suggest that the model is quite effective in predicting the sentence length (SENTCAP), but the MSE indicates there is still some room for improvement in prediction accuracy.

Through our analysis of this dataset, we have found some interesting takeaways about the impact of different variables on sentencing length for cocaine trafficking. First, demographic variables such as race, sex, and education levels did not have as big of an influence as we first predicted. There are a lot of variables of significance that are related to sentencing guidelines, such as MNTHDEPT, which indicates the difference in months between the guideline minimum and the sentence length. MAND1 indicates the status of any mandatory minimums at sentencing, which was a variable included in the Lasso, but had a smaller influence than other variables like PROBDUM, for example.

CONCLUSION

Our analysis and models produce several key findings. Firstly, plea deals were consistently associated with shorter sentences. Decision tree analysis revealed that sentence length (SENTTCAP) was a primary predictor in distinguishing between plea and trial outcomes, with shorter sentences overwhelmingly associated with plea deals. Secondly, contrary to our initial expectations, demographic variables such as race, gender, and education level had relatively minor impacts on sentence length compared to offense-specific factors. This suggests that the Sentencing Guidelines may mitigate, but not entirely eliminate, disparities in sentencing outcomes based on demographic characteristics. We also observed unbalanced data: over 96% of defendants in the dataset accepted plea deals, which makes the class proportion unbalanced and this influences model performance – while also showing that a vast majority of defendants chose to plead guilty rather than taking the case to trial in federal drug trafficking cases.

In addition, variables related to the specific nature of the offense – such as drug quantity, weapon enhancements, and Base Offense Level – are determined to be significant predictors of sentence length. The presence of a weapon enhancement (WEAPSOC) and the final offense level (XFOLSOR) were especially influential, which shows the importance of case-specific details in defining the initial sentencing range.

Finally, even though linear regression provided interpretable insights, it struggled to capture the complexity of sentencing decisions. Random forest models showed higher predictive accuracy, but sacrificed interpretability. LASSO regression struck a balance by reducing the feature set from 211 to 86 variables and highlighting key predictors without overfitting. Despite these advances, the models' predictive accuracy was limited by unobserved variables and the inherent unpredictability of judicial discretion.

Nevertheless, the outputs still give us insights about the data. The Guidelines offer judges suggestions of appropriate sentence ranges given the quantity of drug, type of drug,

mitigating/aggravating circumstances, etc. But judges do not have to follow the sentencing range recommended by the Guidelines and can impose a time outside of that range based on the aforementioned uncountable variables. This means factors included in our models and the Guidelines themselves are imperfect indicators of sentencing outcomes. The unpredictable nature of sentencing can be a double edged sword, as shown by the model evaluations, and this is further exacerbated by the unbalanced data. But the capriciousness of sentencing decisions show that the Guidelines are operating properly, as judges have the discretion to individualize sentences according to the circumstances of each defendant.

Fitting punishment to crime is no straightforward thing. US Sentencing Guidelines give a foundation for consistency in sentence length both for and between individuals, however, these guidelines have come under criticism for inconsistency when applied to defendants of different races, particularly drug cases. Identifying or verifying these patterns in bias or discrimination through machine learning models is tricky. Searching for a certain trend seems unlikely to result in trustworthy results. Nevertheless, investigations using thoughtful methods and variable choices could be valuable in an objective evaluation of United States sentencing. Future work will rely upon a more intimate knowledge of the sentencing guidelines themselves and careful consideration of previously omitted variables.

REFERENCES

We would like to acknowledge Professor Terence Johnson and his teachings in DS 3001 Foundations in Machine Learning at the University of Virginia in the assistance and completion of this research project.

United States Sentencing Commission. “Monitoring of Federal Criminal Sentence Series.” 29 Mar. 2022. https://www.icpsr.umich.edu/web/ICPSR/studies/37990.