Pre Analysis Planning
DS 3001

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

Research Question:
Which factors are most predictive of sentence length for federal cocaine trafficking offenders, and how do plea deals impact these predictions?

- Testing predictive value of the following variables:
  - Offender-based characteristics: age, gender, race, education level, citizenship status
  - Specific offense characteristics: severity of crime (Base Offense Level), weapon enhancement charge, drug quantity, type of cocaine
  - Trial outcomes: whether a plea deal is taken
- Predicted variable: sentence length in months

1. **What is an observation in your study?**

   In the data, each row represents an offender. Observations include many hundreds of data points including basic biographical info, location, crime, and sentencing.

2. **Are you doing supervised or unsupervised learning? Classification or regression?**

   We are doing 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'll use regression analysis, as we are using our model to predict a numeric variable.

   In addition, 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 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.

3. **What models or algorithms do you plan to use in your analysis? How?**

   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.

4. **How will you know if your approach "works"? What does success mean?**

   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.

5. **What are weaknesses that you anticipate being an issue? How will you deal with them if they come up? If your approach fails, what might you learn from this unfortunate outcome?**

   Multicollinearity: 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.

   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.

Sheer 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.

Interpretability: 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.