New York Pretrial Release Prediction Problem Set Handout

Please use the template notebook provided to complete the assignment. Where this handout provides a general guideline of what we expect for each question, the notebook outlines more granular steps. You should only submit the notebook which contains your code as well as answers to discussion questions in markdown cells.

I. Introduction

On January 1, 2020, a new bail law restricting the circumstances under which pretrial detention and cash bail could be imposed went into effect in New York state. State legislators sought to lower the population of people who were jailed because they could not afford bail before their trial. While well-intentioned, the legislation received criticism from people across the political spectrum. A common concern was that crime would increase as a result of these changes, despite a lack of data to support this claim. Subsequent reforms, which went into effect on July 1, 2020, broadened the situations under which cash bail could be set.

Arraignments are hearings where defendants are presented with and respond to the charges against them. This is also where presiding judges decide whether to detain them and set bail. In this problem set, we will be predicting and analyzing judge predictions of defendant flight risk at arraignment hearings. We will focus on the **`ROR_at_Arraign`** variable – which indicates whether or not the judge released the defendant on their own recognizance – in the provided dataset to be a representation of this. (Releasing someone on their own recognizance means they are released without having to post bail, which indicates that the judge believes you are not a flight risk and will likely return to court for your scheduled appearances). We are aiming to predict **`ROR_at_Arraign`** for this problem set because it is a more common outcome than being remanded to jail.

To do so, we will be using pretrial release data. Note that the data file we will use extend from January 2020 through the June of 2023; they are updated biannually, and now separate New York City from upstate New York. We will use the New York City data, which has been put in the Data Science GitHub repo for big datasets used in classes. This dataset is similar to the dataset used in Kleinberg et. al's paper, "Human Decisions and Machine Predictions."

By the end of this assignment, you will gain experience in the following:

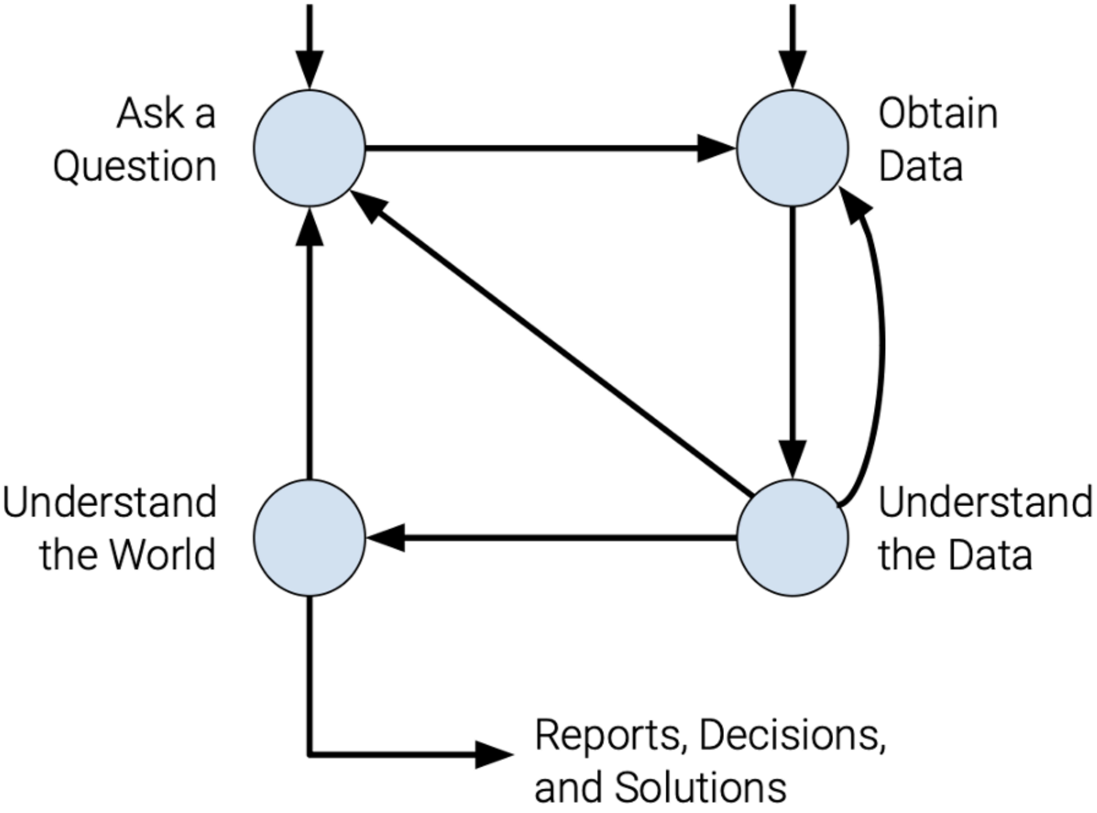
- Following the roadmap of the Data Science Lifecycle while reflecting on ethical considerations.
- Cleaning messy publicly available data and documenting your process clearly.
- Applying and rationalizing different predictive modeling techniques for a given dataset.

- Evaluating the relative performance of different models and justifying the selection of a particular model.
- Comparing models against an external benchmark and contextualizing results.

The Data Science Lifecycle

The Data Science Lifecycle gives us a way to understand the workflow when building something like a prediction model using data. The lifecycle is not linear, in that at any point you may have to go back and reassess previous steps. This is an important part of building anything using data from the real world. In addition to using data to make a model that leads to "reports, decisions, and solutions," it is equally as pivotal to infuse ethical considerations into each step of the lifecycle. This includes questions about bias, discrimination, fairness, transparency, and more. In some cases, it means asking questions about *if* your model should actually be made instead of *how* it should.

In building our judge decision model in this problem set, we will use the lifecycle to guide us.



II. Ask a Question

Our question: Can we use a model to predict what decision a judge will make at arraignment? More specifically, can a model predict whether a judge will release a defendant on their own recognizance?

This is a different question than Kleinberg et al., who are focused on calculating the supposed “risk” of an individual defendant using statistical methods. Kleinberg et al. ask, broadly: **Can machine learning improve human decision making?** in "Human Decisions and Machine Predictions," as they use pretrial release data to compare an algorithm's prediction of a released defendant's behavior (i.e., fleeing, committing another crime) to a judge's — effectively replicating the calculation that a judge is required to make when deciding where a defendant will await trial.

As technology advances to the point where incredibly large datasets can be processed, applications of machine learning have sprawled across industries and systems – including the United States criminal justice system. Take the COMPAS algorithm, which used a machine learning model to compute "risk scores" for defendants, or predictive policing models – all of which have been used to augment human decision making, in turn having a direct impact on people's lives. Kleinberg et al. ultimately conclude that machine learning can improve judge decision making, i.e. what they decide to do with the defendant at the time of their arraignment.

(a) What stakeholders are implicated by the question we are asking/ the problem we are trying to solve? (ex: judges, defendants, law enforcement, etc.). Compare this to the stakeholders that are implicated by the question we are asking. *Consider how Kleinberg et al. are framing the use of data science itself – as a supplement or replacement to human decision making – versus how we are framing it.*

(b) Though Kleinberg et al. are NOT suggesting this, think about a world where a machine learning or predictive model replaced the labor of human judges. Do you think this would be a good idea? Who would be impacted by this automation?

III. Obtain Data

The data-generating process— that is, why and how data was collected— has implications for how we approach the data and what we do with it. Along with the bail policy changes, the first iteration of the law was accompanied by a mandate to collect the pretrial release data (which we are using for this problem set). So that we do not have to use the State of New York’s server, the .csv file for the data is on the Data Science Modules data repository, where its url is

https://github.com/ds-modules/data/raw/main/LS123_Pretrial_NYC_2020jan_2023jun.csv

(a) How should the context in which this data was produced impact the way we understand it - from data cleaning to modeling?

(b) Take a look at the supplemental information for this dataset, including the data documentation and data dictionary (which will be useful to you throughout the rest of the problem set!). Is it transparent? Why or why not?

IV. Understand the Data

Data Cleaning and Exploratory Data Analysis

(a) **Leakage.** Leakage is a term used in machine learning for when a model makes decisions using information that it would not actually have, resulting in unnaturally high performance. What would leakage look like in the context of this problem and dataset?
Hint: think about the entire criminal sentencing process (and what happens before and after arraignment).

> (i) Inspect the features in the dataset and their associated descriptions in the documentation. List the ones that must be removed to prevent data leakage.

(b) **Missing values.**

Read this article on the importance of missing values in machine learning tasks: "Missing the missing values: The ugly duckling of fairness in machine learning." Then, reflect:

> (i) Which features have the greatest proportion of missing values? Why do you think that is (think about the data collection process, the meaning of and representation of each feature, etc.)?
>
> (ii) How will our decision about how we deal with missing values (i.e. dropping them altogether, replacing them with 0s, etc.) ultimately impact our model – not just in terms of accuracy, but in terms of fairness?

(c) **Exploring the features.**

In working with a dataset about the criminal justice system, we might be interested in exploring the demographics of people represented. In order to model, we are also interested in exploring the relative frequencies of values for each of the categorical variables. Creating visualizations and finding summary statistics is a great way to gain a better understanding of this.

> (i) How many people from each racial category are represented in the dataset? How about each gender category?
>
> (ii) How many people from each decision category are represented in this dataset (that is, how many records of each outcome)? How does the breakdown of race and gender differ for each outcome?

Adapting Our Dataset to Our Prediction Task

(a) **Time frame.** What should the timeframe of the dataset be if we're interested in model predictions under the revised statute that reinstituted opportunities for cash bail? How do you address this in your data cleaning?

(b) **Categorical variables.** In order to use categorical variables in a modeling task, they must be transformed using the process of one-hot encoding. Identify these features and perform one-hot encoding.

> (i) Notice that `Age_at_Crime` and `Age_at_Arrest` are two of the only numerical variables in our dataset. Investigate these two variables further by creating a histogram of both `Age_at_Crime` and `Age_at_Arrest`. Then, reflect on and deal with missing values for these two variables.

Feature Selection

Feature selection, by which we reduce our dataset down to what are deemed the most important features in order to embark on our prediction task, is a pivotal part of any machine learning task. Given what we understand about the criminal justice system, we could select the features that we deem most important to train our model on. We could also use feature selection techniques such as the correlation coefficient, which measures the correlation between each feature in our dataset to the variable we are trying to predict.

(a) The New York Criminal Justice Agency published a study where they used a similar dataset to determine which features had the most power for predicting whether a defendant would fail to appear. This study was ultimately used to create a "Release Assessment" tool intended to help judges make decisions at the time of arraignment. Judges can plug in a defendant's information into the eight factors that the study deemed were most important:

1. Years since the last bench warrant?
2. Two or more bench warrants in the last five years? (Y/N)
3. Misdemeanor convictions in the last year? (Y/N)
4. Number of misdemeanor convictions in the last three years?
5. Number of felony convictions in the last ten years?
6. Number of pending cases?
7. Years living at the last two addresses?
8. Reachable by phone? (Y/N)

Under this assessment system, each defendant begins with 25 points, which are deducted according to where they fall within each of these eight features. The score is then used to give a judge a recommendation on whether to release them on their recognizance or not.

(Note that in this problem set, we are not building a model that predicts whether or not a defendant will fail to appear in court or not, but rather to predict judge decisions themselves).

(i) What do you think are the potential pitfalls of using a release assessment algorithm in the judge's decision-making process at arraignment?

(ii) No feature in our dataset that indicates if a judge relied on the assessment tool to make their decision. What is the problem with this?

(iii) What metrics of success did the study's authors use to assess their predictions?

(iv) Built into this tool is the assumption that defendants under similar circumstances will behave the same way/ have the same outcome. What is the problem with this – and how does it fit into the larger discussion about human decision-making versus decision-making through ML models?

(b) Create a correlation matrix to find which features are most highly correlated with being released on the defendants' own recognizance at the time of arraignment. What features are most highly correlated? Why do you think that is?

(c) Are any "protected classes" (i.e. race, gender, etc.) included in the most highly correlated features? Should they be included in our model? Why or why not?

V. Understand the World → Reports, Decisions, and Solutions

Modeling, Prediction, and Performance Metrics

(a) Partition the dataset into a training, test, and validation set.

(b) Choose three different prediction techniques that are appropriate to the problem at hand. Describe the basic mechanism of each, benefits and drawbacks, and any other criteria you factored into your decision. Why is each a reasonable choice in this context.
Hint: take a look at the different techniques and justifications in the classification labs.

(c) Create models and report their accuracy as well as a false-positive/ false-negative rates through an AUC-ROC curve and a confusion matrix (note that this is one way of measuring the performance of our model – it gives us a sense of the metrics of recall, precision, specificity, and accuracy).

(d) Choose one of the three models you created and report the top 10 features with the most predictive value. Use pandas `coef_` or `feature_importances_`, depending on which model you pick. Then, provide a brief discussion on your interpretation of the most important features.

Interpretability

Create a dataframe that can be used to analyze your model's performance, especially relative to judges. This should contain (1) all the features the model was trained on, (2) the judge's arraignment release decision, (3) the probability score of the data point computed by the model, and (4) the model's ultimate prediction.

Evaluate the Model Further

Choose one model out of the three you created above. We want a way to compare this model to actual judge decisions, specifically focusing on different demographic groups. Because our data is a product of the criminal justice system – which disproportionately arrests and incarcerates Black men – it is within our interest to examine how our model performs, and if there are any differences, between racial groups.

(a) Compute the average judge decision, model probability score, and model prediction for 1) all cases, 2) cases with black defendants, 3) cases with non-black defendants, 4) cases with male defendants, 5) cases with non-male defendants, 6) cases with defendants at or above the median age, and 7) cases with defendants below the median age. What do you observe about race, gender, and age differences for these scores? Include visualizations to help your audience understand the group differences.

(Remember, the average in this context means the proportion of 1s, or in other words, the proportion of defendants of a given demographic group that a judge released on their own recognizance.)

Final Reflection Questions

1. What have you learned about the Data Science Lifecycle throughout this problem set? Are there any ethical considerations that remain unsolved?
2. This paper by Barabas et al. suggests that models like the one you made in this problem set – which aim to make predictions about judge decision making instead of about defendants themselves – is a form of studying the power structure as a whole. Reflect.
3. Imagine that you are a data scientist in an agency hired to assess judge decision-making in New York, given the pretrial release dataset. What insights have you gleaned from the process of making a model that predicts a judge's decision to release someone on their own recognizance or not? Is there anything you would have done differently (i.e. with cleaning the data, the choice of model, etc.) and why?
4. As Kleinberg et al. noted, criminal court judges in New York City are able to avail themselves of a risk assessment algorithm (which is designed to predict risk of non-appearance, which is the "official" reason for a judge to assign bail). The risk assessment algorithm is from an NGO, the New York City Criminal Justice Agency. The NGO does an intake interview and creates a simple risk assessment that is then reported to the judge, who may or may not take it into consideration in making the pretrial release decision. How do the predictive features you found align with those in the Criminal Justice Agency's algorithm? What additional data does the CJA have that you do not, and why do think those additional items are good predictors of appearance post-arraignment?