Dear Dr. Penelope Dyson,

I hope this message finds you well. My name is Carmen Santana, and I am the data scientist from Awkward Problem Solutions™ assigned to assess the fairness of your project in recidivism prediction. I have begun my preliminary analysis of the data, and I have a few questions that need clarification to proceed effectively:

Initial Findings on Data Bias:
In my initial exploratory data analysis, I observed patterns suggesting a potential biases in how individuals are assessed for risk. For example, African-American individuals tend to receive higher risk scores, indicated by a Pearson correlation coefficient of 0.29, this is only happening for individuals labeled as African-americans. This correlation, while weak, suggests a slight trend where the likelihood of being labeled high risk increases with the proportion of African-American individuals. Conversely, Caucasian individuals often receive lower risk scores, with a correlation coefficient of 0.18, something that doesn't happen with the other races. These correlations do not imply causation but do suggest a pattern that warrants further investigation to ensure fairness, like performing an statistical analysis to find if this discrepancy is significative.

Prediction Goals:
For the prediction model, what is the priority of the outcome:
Do we focus on accurately predicting who will reoffend, knowing that we might miss some individuals?
Or should we aim to flag as many potential reoffenders as possible, even if this means some innocent people might be incorrectly labeled as likely to reoffend? This approach might increase bias towards certain racial groups. I saw that your dataset is unbalance (half of the individuals in your dataset are labeled as African-american), so I think we should start with looking for precision (the first approach), and check how the model behaves and make further calibrations and optimizations..

Data Columns Clarification:
I need a better understanding of the data fields provided:
Could you provide a brief description of each column in the dataset?
Specifically, which columns contain the predictions from your current model? I believe they might be labeled as type_of_assessment, decile_score, and score_text for general risk, and similarly labeled with a 'v_' prefix for violence risk.

For the columns labeled is_recid , is_violent_recid, and two_year_recid, do these represent the actual outcomes (whether someone reoffended), or are they predictions from your model? If they are predictions, do you have the real data to know if the individual really reoffended?
What does dob stand for? Is it 'date of birth'?
How is race data collected? Is it self-identified by individuals or determined by others?, are we going to get only the current categories or is it possible that we could get other categories like mixed races?
There are discrepancies in offense and arrest dates. When offense date (c_offence_date) is missing there is usually a value for arrest date (c_arrest_date:)). However some rows are missing both, does that mean that they were never arrested but they commited crime? Or does it mean that they didn't commit a crime? If they have c_offense_date but not c_arrest_date does it mean that they were never arrested or is it an error from the database and I should take as arrest date the offense date in those cases?
What do c_case_number, r_case_number, and vr_case_number represent, particularly when c_case_number is missing? In the data, there are some columns with NAN values (no values) in c_case_number, but with values in 'r_case_number' and 'vr_case_number', what does that mean? How can a convict recidivate if they don't have a first crime (c_case_number)? Or am I understanding wrongly what a c_case_number mean?
What do the values in c_charge_degree and vr_charge_degree represent?
At what age is someone considered a juvenile in this dataset?
Is there a cap on the number of felonies (juv_fel_count) and misdemeanors (juv_misd_count) a juvenile can be recorded as having committed?
Could you clarify what counts toward priors_count? Is it only previous jail time, or does it include detentions and accusations as well?
What does compas_screening_date represent?
Does the type_of_assessment (Risk of Recidivism) include assessments for violent crimes as well?
What does each decile score represent in decile_score and how do they relate to score_text? The same question goes for v_decile_score and v_score_text.
Does v_type_of_assessment specifically refer to violent crimes?
Do columns starting with vr_ relate to repeat offenses specifically for violent crimes?

Technical Requirements:
You mentioned setting up an API for this project. Could you specify what type of requests the API should handle? An example query would be extremely helpful.

Privacy Concerns:
I noticed a column labeled names containing actual names of individuals. Typically, personal data like this should be anonymized to protect privacy. Could we confirm if these need to be anonymized?

Data accuracy:
The age in age column seems incorrect, by nine years. It looks like the last time that it was updated was in 2016.

Thank you for your assistance, and I look forward to your responses so we can proceed effectively. As well, I am open for any suggestion and questions.

Sincerely,

Carmen Santana
Data Scientist, Awkward Problem Solutions™