Project 2
Algorithmic Descision Making

Description In exploring algorithmic decision-making, especially in the context of criminal recidivism prediction, we want you to have first-hand experience in interrogating the model built by a machine learning algorithm. Your investigation includes not only being able to transparently see the code for the machine learning algorithm (which is what some lawmakers argue is a necessary safeguard in deploying algorithms for such decision-making), but also assessing the "fairness" of the model produced by the algorithm on a number of criteria while thinking about the limitations of the data given. Ultimately, we want you to determine what you would do to make the decision-making model produced by the algorithm more "fair," and justify this position in a short memo to a company executive.

The data and code

Information on the data and the code can be found in the file titled `Project2 - data explanation.pdf`. You should be sure to read all of that information carefully before starting with the task below.

The task

Say you are employed at a company named *JudgeSoft* that is producing an algorithm that will train a model to make a prediction as to whether an individual charged with a crime is likely to recidivate or not. The predictions of the model learned by the algorithm will be used by Broward County judges as one of many factors in determining whether to release an individual charged with a crime until the time of their trial (i.e., grant them bail) or keep them detained (i.e., deny bail) until the trial. Thus, whether the individual is likely to commit another crime (i.e., recidivate) will impact the judge's decision.

Having been just recently hired by JudgeSoft you find that a machine learning algorithm has already been developed for this project. The company Chief Technology Office (knowing that you are interested in ethics and technology after seeing CSC-395 on your transcript) asks you to investigate the algorithm, answer some questions about it, and make some recommendations before a final product is shipped. More specifically, she wants you to answer the questions below. You should write up the answer to all these questions in a single PDF file titled "Project2" and submit it on Gradescope by 10:30pm on the date it is due.

Questions to answer:

1. Without running the program (or, at least, not considering the results if you did run the program), review all the code files in the `AlgorithmicDecisionMaking` project to understand how the overall program works.

   (a) Note that the programmers of the algorithm had heard about "protected" characteristics, such as age, gender, and race, and that impacted their code. Do you believe that the algorithm itself (as coded) includes any biases? Briefly (in a paragraph or two) explain your answer. Include references to concepts from class as needed to justify your answer.

   (b) Again, prior to running the algorithm (or, at least, not considering the results if you did run the program), consider the case where the algorithm was coded to include the use of the input features age, gender, and race during the learning process. Would you believe that in this case the algorithm includes any biases? Briefly (in a paragraph or two) explain your answer. Include references to concepts from class as needed to justify your answer.

2. Next, we will run the code! The program can be run from the terminal by typing `python3 algorithmicdecisionmaking.py` once you have navigated to the location where the code is stored. This trains the original model using the training data, uses this learned model to compute statistical results, and prints some of the results.

Although there are six race-based input features, for this part of your investigation you can focus on just African-Americans and Caucasians as was done in the ProPublica investigation. Edit the code in `algorithmicdecisionmaking.py` using the `print_results` function (and any other statistics you believe are relevant that you want to add to the code) for African-Americans and Caucasians on the training data and the testing data. Report the statistics you computed. Comment on how the various results you obtained from the model compares with ProPublica's analysis of the differences between predictions for African-Americans and Caucasians in the COMPAS algorithm. Explain the similarities/differences you see.

3. Based on the results you obtained from question 2 above, do you believe that the model learned by the algorithm is "biased"? Explain why or why not. Justify your answer with respect to at least three of the notions of anti-classification, classification parity, calibration, and disparate impact. Use quantitative results as appropriate to make your argument.
4. Modify the code (`select_features_to_use` function) to include all the input features for age, gender, and race (along with the other input features already included in the model) and train a new model with these input features on the training data. Based on the results you obtained with the new model, do you believe that this new model is more or less "fair" than the old model (built in question 2) which did not use the protected characteristics? Explain your decision, including how you define the notion of "fair". Justify your answer with respect to at least three of the notions of anti-classification, classification parity, calibration, and disparate impact on the data. Use quantitative data as appropriate to make your argument. Note: you might want to consider protected characteristics beyond race, such as gender and age, in your answer.
5. Modify the code however you like, including (but, not limited to):

    - Choice of input features to include while training the model
    - Making manual adjustments to the weights. To do this, you can manually change the weights in the file `model.txt` for different subpopulations. For example, increasing the weight for a particular input feature makes individuals with that feature more likely to be classified as positive (prediction value 1) by the model, as their overall weighted sum will increase as a result. Then, to use these manual weights (instead of training the model with the training data), edit the `algorithmicdecisionmaking.py` file to read in that data (rather than train) with the following line of code: `pm = PerceptronModel("model.txt",features_to_use)`.
    - Whatever else you would like to do (e.g., training different models for different subpopulations, other means for setting different prediction thresholds based on various features, changing the number of epochs for training, interrogating the data set, researching the potential provenance of data from the criminal justice system, etc.)

    Your goal in making these modifications is to make the decision-making algorithm (the resulting learned model) as "fair" as possible in your assessment. Write an approximately 500 word (around 2 page) memo to the company CTO about your work. (Note that statistics and their labels/descriptions, as well as tables and diagrams you might want to include, do not count as part of the 500 words of text.) You can assume the CTO has a solid technical background—she knows how to code well, understands what machine learning is and how it works, and has done all the readings and attended all the classes in CSC-395, so she knows about different concepts of fairness, and understands machine learning and statistics. Your memo should explain the final model you have come up with in detail (e.g., choice of features, other adjustments you would make, etc.) and explain why you believe these choices have made the model as "fair" as possible (especially with respect to the various "fairness" criteria you are aware of), while also noting potential limitations. Try to be as clear as possible with regard to your definition of "fair". Use quantitative measurements and qualitative arguments to justify your claims.

    Note: this part of the assignment is not an assessment of how much you know about machine learning. For example, building a complicated model for which you have difficulty justifying its "fairness" will not impress the CTO to whom you are writing this memo. The goal here is to make an argument that shows your assessment of a model's "fairness" in a real-world setting, where you have control over what that model (and the algorithm that built it) is.

Learning Outcomes Completion of this assignment will contribute to your ability to fulfill the following learning outcomes:
5. Identify the impact of different technologies on various identities.
6. Identify strategies for creating more equitable and inclusive technical environments and software for diverse identities.
8. Engage in current ongoing discussions of algorithms, ethics, and society.
9. Understand advanced algorithms in the context of their application(s).

Acknowledgments This assignment was written and designed by Rob Reich, Jeremy Weinstein, and Mehran Sahami, all at the Stanford University for their CS182 course. Student Michael Dworsky also categorized the recidivism data and developed an early prototype for the assignment. They have given me permission to use their materials.
Any changes made have been minor with the aim of providing context, or adapting to the specifics of CSC-395 at Grinnell College.