Question 1

Fully engaged in the Twitter hype for GPT-4 and constantly up to date on the NeurIPS threads the CTO pitches a neural network architecture for predicting if the client will default on the loan. The CTO suggests the various columns of the table act as features to a deep neural network with 50 hidden layers. With your experience in the field you think that there could be a better model for such a sensitive task as this. Make an argument for why a decision tree is better for use in this particular situation (hint: you can mention runtime at inference, but there is an even stronger case to be made for a decision tree that we are looking for here! Especially consider that this ML model will be making highly sensitive decisions.)

A decision tree is much better for this situation for a number of reasons, but one of the most important is the ability for a decision tree to be able to rank the relevance of features. This is a crucial factor into how the company can decide to determine which variables in the format are worth considering the most. Another important factor is that decision trees are virtually immune to outliers, and given that this data is from the 1930s (almost a whole century worth of data), it can be generally assumed that there can be many outliers.

Question 2

The ethics review board at the company rejects the initial proposal from the CTO on the basis that algorithm could easily end up rejecting loan applicants based on race or sex. To fix this the CTO proposes that the sex and race columns be removed from the dataset. Will this completely prevent the machine learning model from discriminating based on race/sex? Why or why not?

While removing sex and race out of the equation would lead to less rejections and discrimination, I think it's a pretty far cry from saying that it would completely remove all discriminations based on race/sex. One reason is that the name variable is still included which is generally a huge indicator of someone's race anyways, so it doesn't change all that much. And going on a few more assumptions, the zip code can be an indicator of areas that are predominantly one race which the model might pick up on. Lastly, the pronouns variable is still included which determines the person's sex, so removing sex again doesn't change much.

Question 3

Algorithmic discrimination is a serious problem with the wide dissemination of machine learning models. (?) Unfortunately, these harms can go unnoticed due to the false assumption that math and algorithms are unbiased and objective. Machine learning models are only as good as the data used to train them. Research an instance where a machine learning model was used to make critical decisions, but was later found to be biased (excluding the example with bank loans). Summarize what the purpose of the model was, and how it ended up causing harm. Include a link to a news article or research paper discussing this issue.

https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai

In 2015, Google was caught under fire for misclassifying black people as "gorillas" in people's phones. The original purpose of Google's image labeling system in images was to make it easy for users when they wanted to find pictures in their camera roll of certain objects. For example, typing "car" in the camera roll would show up images that contained a car. But when a user would look at their pictures with black people in them, it would classify them as gorillas, sparking immense controversy. And when users would type "black people" in their camera roll, only photos of people in black and white would show up rather than photos of black people. This led to Google having to issue out an apology, and currently I'm not sure if this problem has improved since then.

Question 4

What is the reason for having a separate train and test set when constructing a machine learning model? Why not use all the possible data for training and testing?

The reason for having a separate training and test set for the ML model is mainly because we want to prepare the model for various situations and not result in "hardcoding" the answers, for lack of a better term. We want the model to be able to handle any kind of data it gets exposed to, so we prime it with training data and then run the test set to see how it performs. If we used all the possible data for training and testing, then we are essentially limiting its scope to what its trained on, and if we use another test set, it probably won't be able to perform as well as we want it to.