Adult Income Bias

Group 1
Christian Adcock | Daniel Guthrie | Donovan Manogue | Ethan Stanks

Lenders can use income prediction models to assess borrowers' default risk. While specific characteristics, such as race, sex, and age, are prohibited by law from use in assessing risk, other variables can serve as proxies, inadvertently encoding biases. By exploring and analyzing a 1994 Census dataset containing income information, we can identify how these structural biases are encoded and how they may lead to discrimination.

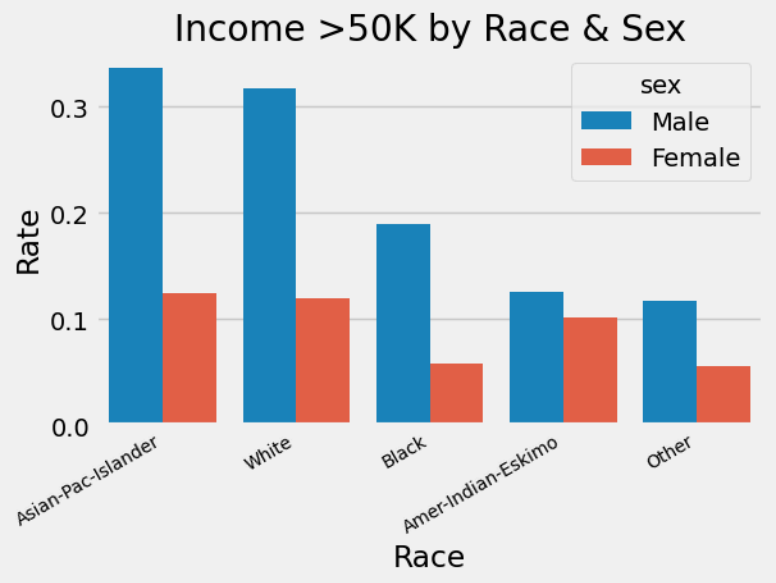
Encoding of Explicit Bias

The dataset includes several variables that directly encode protected demographic attributes, which pose a high risk of explicit bias in any predictive model trained on it. For example, the dataset contains race and sex, meaning a model does not need to infer this demographic information. It will learn to associate these demographic characteristics with certain income levels, directly leading to discrimination and furthering existing structural inequality.

In addition to protected characteristics, the dataset also shows imbalances in these variables. White males represent the largest share of observations (close to 60%), while white females make up another large portion (around 26%). In contrast, some race-sex subgroups appear extremely rarely, such as Amer-Indian Eskimo females, which make up well under 1% of the dataset. This imbalance means the dataset primarily reflects the labor outcomes of white people, particularly white men. As a result, any model trained on this data will be optimized for these groups and may perform poorly for smaller, marginalized groups.

The dataset also encodes bias through unequal distributions of the income label across demographic groups. In the dataset, about 30.6% of males earn more than $50,000, while only 10.9% of females do. Zooming in further, the grouped analysis showed that high-income

observations (those above $50,000) are heavily concentrated among white males. At the same time, many other race-sex combinations have very few high-income observations. The graph below shows the rate at which individuals of a certain sex in each race earn over $50,000.



As you can see, females in any race rarely make more than $50,000 in their jobs, as ~90% earn less. This makes race and sex powerful predictors in the training data, but it also means a model may treat membership in certain groups as evidence of low income. It might also treat being male as a strong positive indicator of making over $50,000. In effect, the dataset teaches the model that demographic identity itself is predictive of income, reinforcing inequality.

Additionally, the analysis of high-income rates across race-sex groups showed that the differences are not uniform. For example, Asian Pacific Islander males had a higher proportion of high-income outcomes than White males, while Black and Amer-Indian-Eskimo individuals had much lower proportions. This suggests that the dataset reflects real-world structural differences in employment distribution and socioeconomic opportunity. Even if the model accurately reflects real-world data and their distributions, it could still be unfair because it is learning patterns that reflect systemic inequality rather than individual merit.

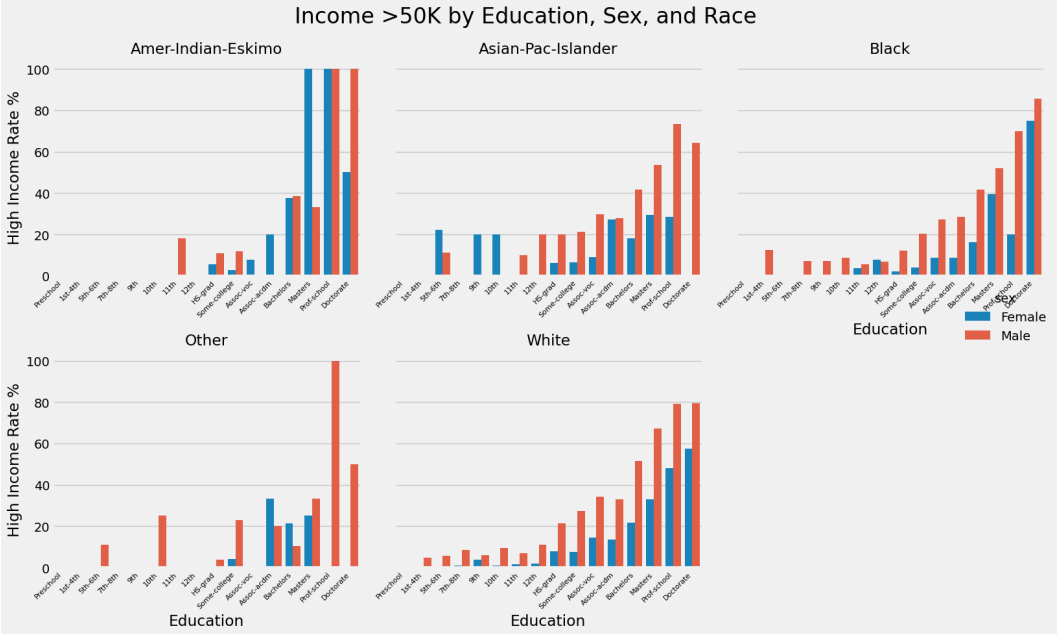
The dataset also contains far fewer low-income examples than high-income examples overall, with 24,720 entries below $50,000 and 7,841 above $50,000. This can make the model biased toward the majority class. When combined with race and sex disparities, the model can end up especially prone to predicting less than $50,000 for minority groups because many minority subgroups have both lower base rates of above $50,000 incomes and fewer total observations. This can produce a system that disproportionately denies high-income predictions to groups that already have fewer positive examples.

Another issue with this dataset is the distribution of null values across race and sex. The analysis of null entries showed that missing values are more common among White individuals and males. Because these null values need to be accounted for before use in a model, preprocessing choices (such as dropping rows with missing values) can shift the dataset's demographic distribution and potentially increase the dominance of certain groups. This can further strengthen biased patterns and make the final model less representative of minority populations.

Proxy Variables and their Possible Influence

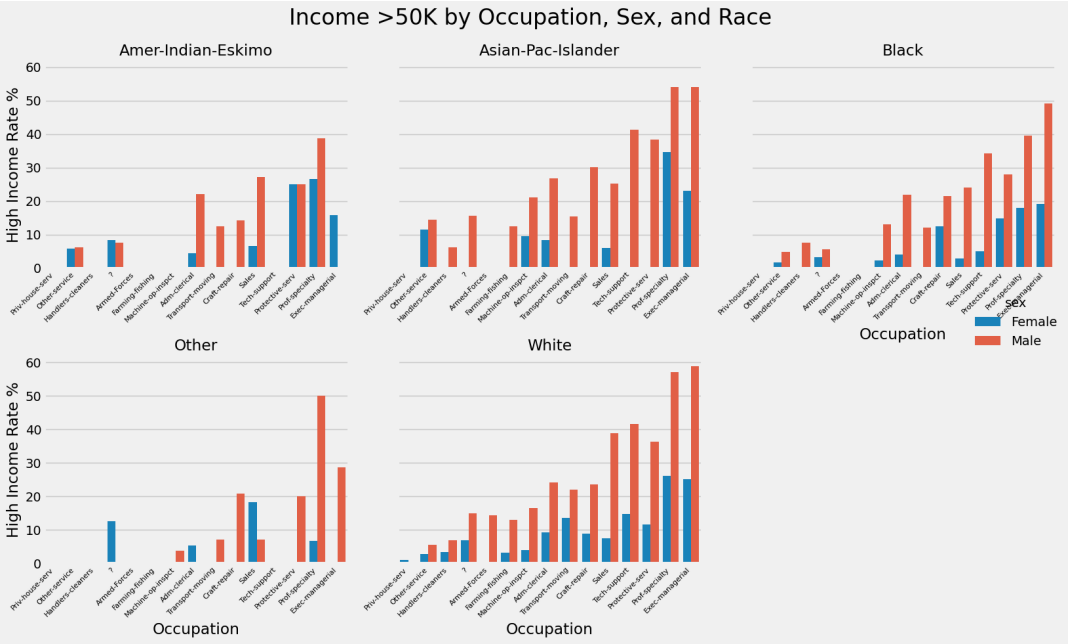
In addition to the dataset, which features examples of sampling bias encoded in its features, bias can also be found in proxy variables. A proxy variable represents information correlated with explicitly biased variables. In this dataset, a proxy variable does not include features like sex and race, but it will include features that can serve as proxies for those protected demographic attributes. Allowing proxy variables to remain in the dataset for analysis means that the sampling bias has not been removed; it has been renamed. In this section, we will explore proxy variables such as Education and Occupation.

Education level is often viewed as a neutral measure of job qualifications, but systemic differences can influence educational opportunities. This means that education level can serve as a proxy for demographic attributes such as race and sex. The graph below compares the High Income Rate % with Education level, split by sex and race. As seen across the graphs, there are disparities in the High Income Rate % between races and sexes. On average, White Males have a higher High Income Rate % than White Females and members of other demographics. The lack of representation across lower educational levels for American Indian, Alaska Native, and Asian Pacific Islander groups also plays into this bias. Rather than treating individuals equally based on their Education level, the dataset perpetuates the bias shown earlier. This suggests that education level can be a proxy for race and sex.



Occupation is expected to be a powerful predictor of Income. Despite this, it can also serve as a proxy for labor-market inequalities, such as differences in access to opportunity. The graph below compares the High Income Rate % with the Occupation split by sex and race.

Similar trends to the Education graphs are evident, with White Men being more represented in the highest levels than any other demographic.



While Education and Occupation may be strong predictors of Income, they can also serve as proxy variables, introducing unwanted bias. Dealing with proxy variables can be difficult because there needs to be a balance between fairness and model performance. Removing these variables will help to eliminate bias from the model, but model performance will suffer.

Conclusion

This analysis demonstrates how structural biases in data can lead to discrimination by exploring the encoding of explicit bias and proxy variables. Many examples of sampling bias were found in the dataset, which were discovered through exploratory data analysis in Python. We found that White Males were overrepresented in most features, including proxy variables. This indicates that removing protected demographic information will not completely remove bias from a predictive model. To create fair predictive models, our efforts as data scientists must account for how inequalities in the world around us shape the data we use.