An Analysis of Income Prediction Using Machine Learning

By Sheraz Ahmad

1. Introduction

Artificial intelligence systems are now used in decision making about hiring, lending and criminal justice, and many experts expect such systems to play a large role in decisions about who gets to go outside and when. Yet these systems tend to reflect and exacerbate societal biases, producing discriminatory results. This report looks at the bias running through predictive models about income, a critical algorithm in terms of how it affects people's economic potential and quality of life. More generally, it is well known that if AI systems are trained on historical data that is biased, they will maintain and, in some cases, even further bias (Barocas, S. and Selbst, A.D. 2016). A model that predicts income, for instance, can have a bias against women in that it systematically undervalues their earnings potential where pay gaps historically existed, with consequences for their ability to receive loans or advance in a career.

Studies have shown that biased inferences in the financial sector have caused minority groups to receive higher interest rates or to be rejected from a service (Bolukbasi, T., et al. 2016). In these, as in much of hiring, models have tended to disadvantage women and racial minorities by making it more likely that their applications sit in the queue for wealthier jobs (Dwork, C. et al. 2012). These biases are not merely technical glitches, rather, they amplify real-world inequalities, exacerbating the difficulties faced by marginalized groups in overcoming systemic barriers.

This report examines these matters through an income forecasting model developed with U.S. Census data. By examining performance among different gender groups, we identify the origins of biases and explore their consequences in the real world. The results add to the increasing evidence that fairness in AI necessitates more than merely precise predictions—it requires thoughtful examination of who gains from those predictions and who could be disadvantaged.

2. Model Development and Application of Fairness Criteria

Data Exploration

The analysis was carried out with the UCI Adult Income Dataset (https://archive.ics.uci.edu/dataset/2/adult), which includes demographic and job-related data from the 1994 U.S. Census. The dataset contains 48,842 records with attributes such as age, educational attainment, job title, hours worked per week, and gender. The target variable was binary, categorizing individuals as either making over $50,000 per year or not.

Exploratory Data Analysis

The first examination showed notable differences in gender within the dataset. Around 67% of the entries indicated male individuals, whereas just 33% indicated female individuals. This disparity reached income distribution, where merely 10% of women were regarded as high earners in contrast to 30% of men. The level of education proved to be a significant predictor of income, as those with advanced degrees like doctorates or master's degrees were considerably more inclined to belong to the high-income group. Weekly working hours were also linked to income, since individuals who worked over 40 hours per week were more often identified as high earners.

Age Distribution by Income Level

The histogram shows a distinct trend, indicating that the highest earning years occur between ages 35 and 50, followed by a noticeable drop in top earners after age 60. The stacked bars indicate that lower income (0) prevails among younger (<25) and older (>60) age groups, whereas middle-aged people display higher percentages of high earners (1).

Income Distribution by Education Level

Higher educational qualifications (Masters, Doctorates) indicate the greatest percentage of high-income earners (>50K), whereas individuals possessing only high school diplomas or less predominantly fall into the low-income bracket. The difference emphasizes that education is a significant indicator of income levels in the data.

Weekly Working Hours Distribution

The boxplot indicates that those who work more than 40 hours a week tend to be higher earners, although this correlation isn't definitive. Significantly, women (1) exhibit marginally reduced working hours compared to men (0) in both income categories, possibly indicating social work-participation trends.

Gender and Income Proportion

The bar chart clearly shows the disparity in income: approximately 30% of men are high earners, whereas only around 10% of women achieve this level. This gap remains evident even when factoring in education and work hours, indicating more profound structural biases in the data.

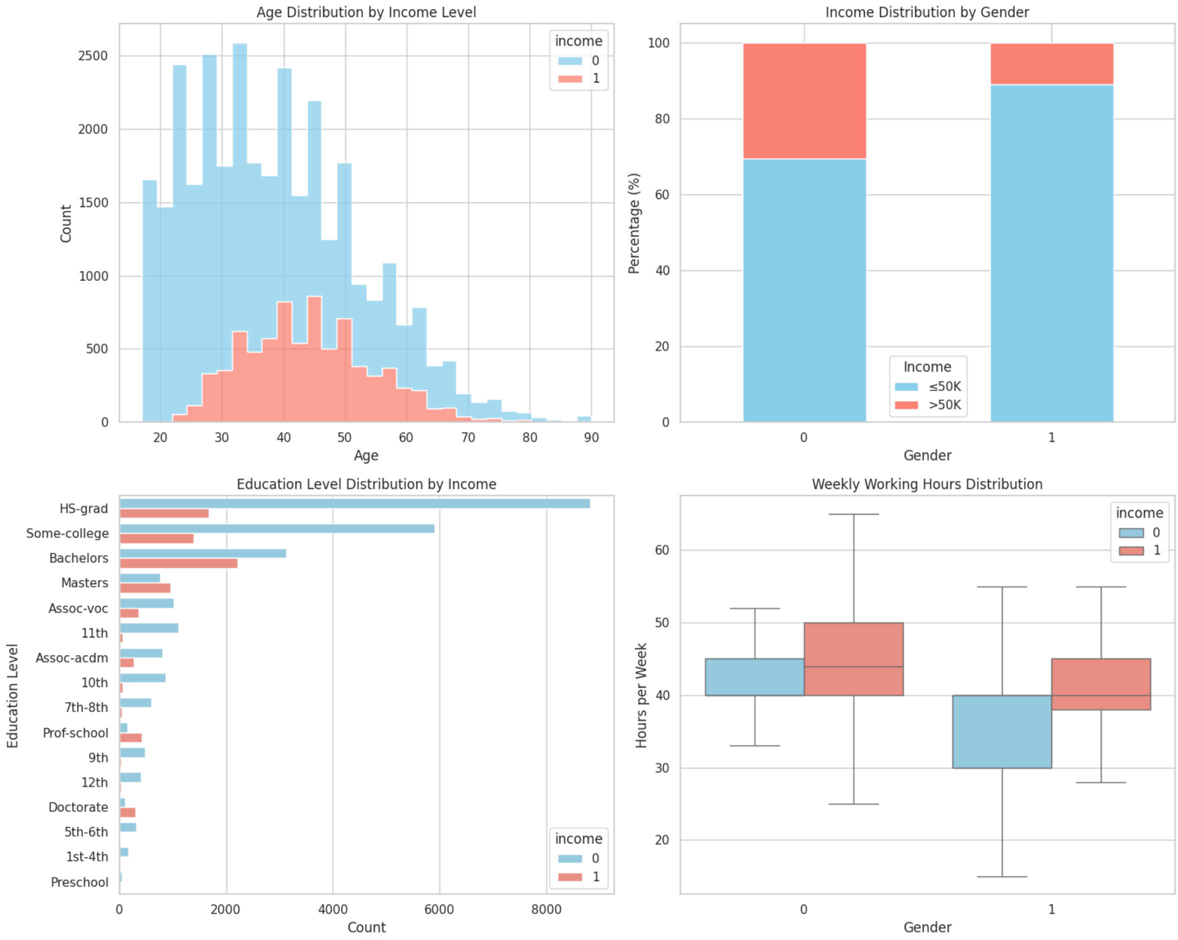


Figure 1. Exploratory Data Examination of Income Distribution Trends: (A) Age breakdown indicating peak earning periods (35-50), (B) Education level significantly related to income, (C) Working hours distribution by gender and income level, and (D) Notable gender imbalance in higher income categories (>$50K).

Feature Correlation Heatmap

The correlation heatmap shows important connections between numerical attributes and income levels. The highest positive correlation with income (r=0.34) is seen in education level (education-num), with age (r=0.23) and weekly working hours (r=0.23) following. Capital gains present a moderate correlation (r=0.22), whereas capital losses exhibit a weaker yet still positive association (r=0.15).

Significantly, education level shows stronger links to income than both work experience (age) and work intensity (hours per week), indicating that formal education is especially crucial in influencing earnings. The heatmap further shows negligible multicollinearity among predictor variables, as most correlations between features are below r=0.15, suggesting they offer independent information for the prediction model.

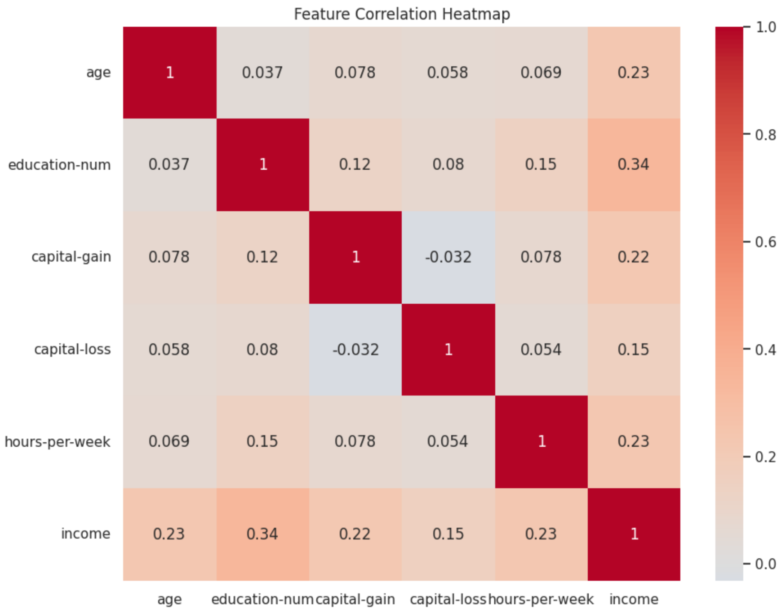


Figure 3. Feature correlation heatmap showing education-num (education level) as income's strongest predictor (r=0.34), that is followed by age (r=0.23) and working hours (r=0.23).

Preprocessing

Multiple preprocessing steps were carried out to ready the data for modeling. Values that were indicated with a "?" in the original dataset were eliminated to maintain data quality. Categorical variables such as work class, education level, and occupation were transformed using one-hot encoding.

The gender attribute was converted into a binary variable, assigning 0 to male and 1 to female. StandardScaler was used to standardize the ranges of numerical features such as age, working hours, and capital gains. Stratification was then used to maintain the original income distribution in both groups after the dataset was divided into training and testing sets using a 70-30 split.

Model Development Process

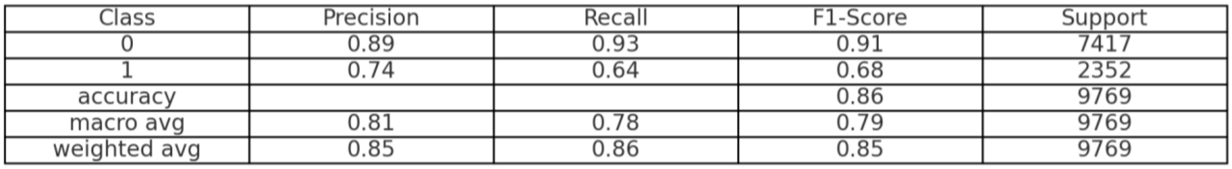
For this analysis, a Random Forest classifier was selected due to its ability to handle numerical and categorical data effectively and its resistance to overfitting. The model's interpretability through feature importance analysis was another important consideration in the selection process (Hardt, M. et al. 2016). To prevent direct gender-related bias in the model's results, the gender attribute was purposefully left out of the input variables during training. This approach made it possible to assess whether biases resulting from other related attributes would be present in the model.

Cross-validation techniques were used to adjust the model's hyperparameters, including the maximum depth and tree count. Care was taken to ensure that there was no overt gender bias in the training process, but it also allowed for the evaluation of potential indirect biases that could arise from other related attributes. The stratified sampling method in the train-test division ensured that income distributions remained consistent across both gender categories in the assessment data.

Performance Evaluation Metrics

The tested model reached an overall accuracy of 86% on the test dataset. A thorough analysis of the classification metrics uncovered significant trends in the model's performance. For the low-income bracket (≤$50K), the model showed excellent precision (0.89) and recall (0.93), reflecting its effectiveness in accurately identifying individuals within this category. Nonetheless, for the upper income bracket (>$50K), although precision stayed satisfactory at 0.74, recall decreased to 0.64, indicating that the model overlooked a considerable number of real high earners. Below is the table for the model evaluation metrics:

Table 1. Model performance metrics by income category: The classifier demonstrates best results for low-income forecasts (Class 0: Precision=0.89, Recall=0.93) but exhibits diminished effectiveness for high-income detection (Class 1: Precision=0.74, Recall=0.64), achieving an overall accuracy of 86%. The F1-score difference (0.91 vs 0.68) emphasizes the model's difficulty in consistently identifying high earners.



Class	Precision	Recall	F1-Score	Support
0	0.89	0.93	0.91	7417
1	0.74	0.64	0.68	2352
accuracy			0.86	9769
macro avg	0.81	0.78	0.79	9769
weighted avg	0.85	0.86	0.85	9769

The confusion matrix offered additional understanding of the model's performance. In total, the model accurately identified 6,878 individuals as low earners while incorrectly categorizing 539 as high earners. On the other hand, it accurately pinpointed 1,495 high earners but did not acknowledge 857 who ought to have been included in this group. When normalized, these statistics indicated that 92.73% of genuine low earners were accurately identified, whereas 36.44% of high earners were mistakenly categorized as low-income.

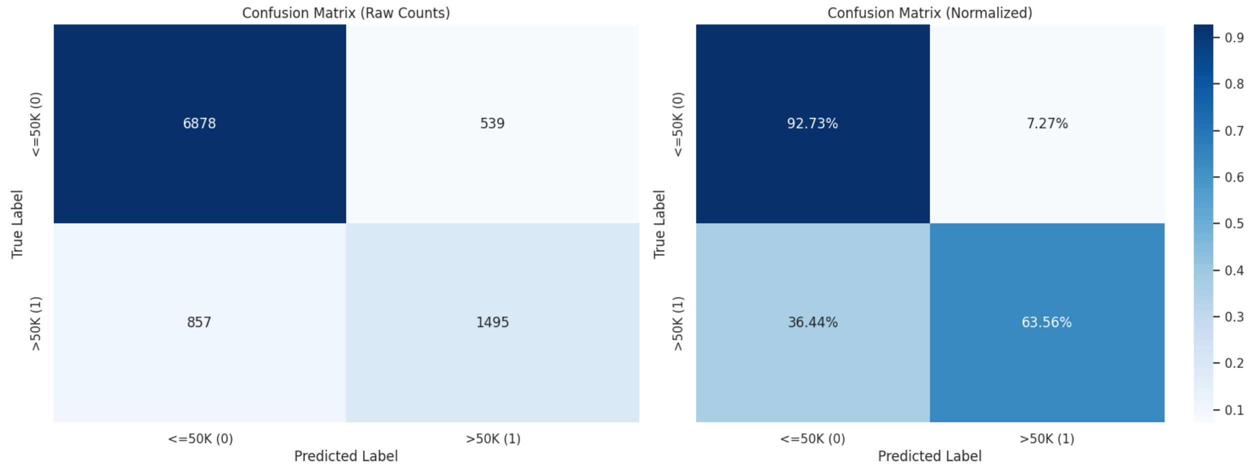


Figure 4. Confusion matrices for the income prediction model: (Left) Raw counts depicting classification effectiveness, (Right) Normalized percentages indicating that 36.4% of high earners (>$50K) were incorrectly classified as low income.

The assessment uncovers significant insights regarding the fairness of our model among different gender groups. Although the overall accuracy seems greater for females (92.9% compared to 82.2%), this surface-level figure conceals more profound inequalities. The confusion matrices indicate that males are given positive predictions (>$50K) almost three times more frequently than females (26.4% compared to 9.3%), which breaches demographic parity. The true positive rates are more aligned (64.2% compared to 59.8%), indicating a somewhat equal opportunity; however, this presents a worrying compromise - females encounter significantly lower overall positive prediction rates.

Notably, the increased accuracy among females mainly arises from accurately identifying low-income cases, which does not inherently imply equitable treatment. These findings emphasize that the three fairness criteria can present contradictory narratives: we could attain equal opportunity while entirely overlooking demographic parity, or we might observe favorable accuracy statistics that obscure systematic biases in the distribution of predictions (Mehrabi, N. et al. 2021). True fairness in practice necessitates balancing these conflicting metrics while taking into account the unique context of income forecasting and its effects on society.

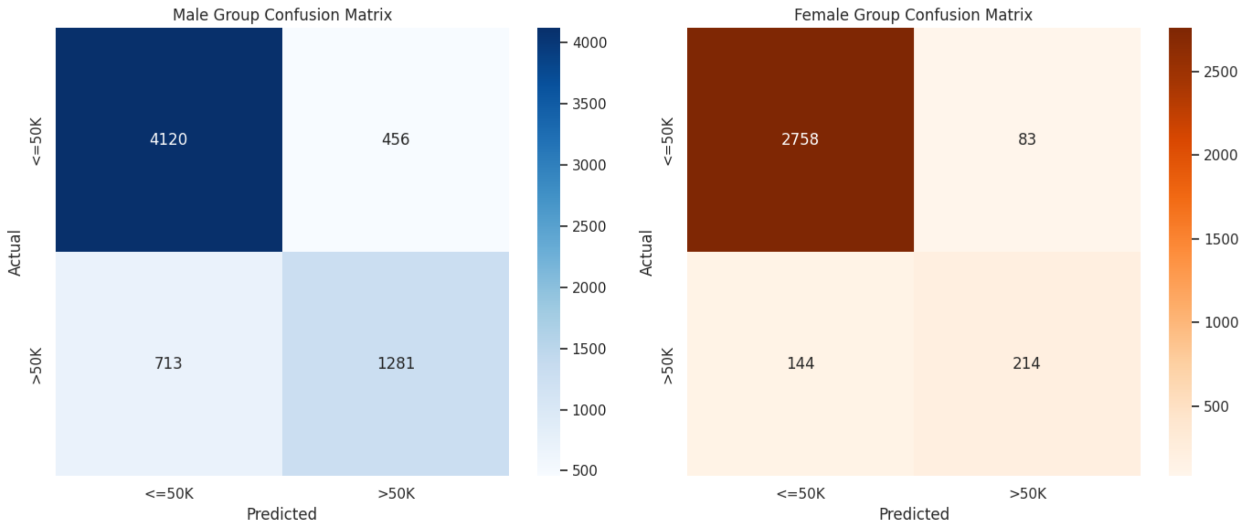


Figure 5. confusion matrices for the two groups.

3. Findings

Interpretation of Results and Ethical Implications

The performance of the model shows substantial biases that have real-world ethical implications. Although the equal opportunity standard is met (with only a 4.5% disparity in true positive rates across genders), the demographic parity findings are concerning, men are almost three times more likely to be forecasted as high earners compared to women (26.4% versus 9.3%). This indicates that the model consistently undervalues women's high-income potential, perpetuating negative stereotypes that women are less probable to occupy high-paying positions.

The difference in accuracy (10.7% higher for women) is deceptive—it arises from the model being excessively cautious in forecasting high incomes for women, often categorizing them as low earners. This fosters an illusion of equitable accuracy while concealing unequal opportunities.

Ethical Impact:

- **For job applicants:** If employed in recruitment or loan decisions, the model might unjustly disadvantage women by undervaluing their income potential.
- **For businesses:** Depending on these skewed forecasts may result in discriminatory recruitment methods, legal liabilities, and harm to reputation.
- **For society:** The model reinforces existing income inequalities by relying on biased data instead of addressing imbalances.

Metrics of fairness frequently clash—here, the model seems "fair" in terms of opportunity but does not achieve demographic parity, indicating that technical fairness does not necessarily equate to ethical fairness. To lessen damage, we need to exceed accuracy and take steps to actively amend biases in model forecasts.

Evidence of Model Bias

The examination uncovers distinct proof of gender bias within the income prediction model via various fairness metrics:

1. **Accuracy Disparity**

The model exhibits a 10.7 percentage point discrepancy in accuracy (Female: 92.9% compared to Male: 82.2%), breaching the equal accuracy standard. This indicates fundamentally varying performance between gender groups.

2. **Demographic Imbalance**

Almost three times more likely to get high income forecasts, male recipients (26.4%). This systematic bias in positive classification rates was also evidenced by a difference of 9.3% versus 9.3% for females.

3. **Error Pattern Analysis**

While equal opportunity was maintained (4.5% TPR disparity), these other biases were not.

   a. Increased false positive rates for males (excessive prediction of high income)
   b. Varying error distributions across groups

Limitations of Fairness Criteria

It is shown that current fairness frameworks impose important constraints.

1. **Metric-Specific Blind Spots**
   a. The distribution of error types is not considered by uniform accuracy.
   b. Valid discrepancies in outcomes are ignored by demographic parity.
   c. Equal opportunity focuses on exactly what is good and focuses on what is good for everyone.

2. **Implementation Challenges**
   a. In real world application, criteria are usually in contradictory (impossibility theorem).
   b. Did not take into consideration the intersectional impact.
   c. Provide no direction as to permissible difference limits.

The criteria of fairness themselves are limited. Equivalent accuracy can conceal compensating for mistakes in which types of errors cancel out statistically but nevertheless make something unjust. results. Demographic parity assumes that equal outcome rates are just by nature, ignoring actual demographic variations in income distribution. The limited emphasis of equal opportunity on The true positive rates ignore other important error categories such as false positives, which may harm. specific groups.

However, these restrictions make the point that no single metric fully captures fairness, and as such, a comprehensive assessment strategy. Intersectional biases are not considered by the standards too. It is that which may have different impact on subgroups when multiple protected characteristics are involved together. Instead, these results demonstrate that algorithmic bias is widespread and that one cannot satisfy the standard of accuracy well. It then assesses and addresses it with existing fairness models.

References

Barocas, S. and Selbst, A.D. (2016) *Big Data's Disparate Impact*. California Law Review, 104(3), pp. 671–732.

Bolukbasi, T. et al. (2016) *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings*. Advances in Neural Information Processing Systems, 29, pp. 4349–4357.

Chouldechova, A. (2017) *Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments*. Big Data, 5(2), pp. 153–163.

Dwork, C. et al. (2012) *Fairness Through Awareness*. Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, pp. 214–226.

Hardt, M. et al. (2016) *Equality of Opportunity in Supervised Learning*. Advances in Neural Information Processing Systems, 29, pp. 3315–3323.

Mehrabi, N. et al. (2021) *A Survey on Bias and Fairness in Machine Learning*. ACM Computing Surveys, 54(6), pp. 1–35.

Obermeyer, Z. et al. (2019) *Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations*. Science, 366(6464), pp. 447–453.