Final Project

Sachin M

The UCI Adult Income Dataset comes from the 1994 Census Bureau database by Ronny Kohavi and Barry Becker. This dataset contains 15 columns in total: age, workclass, fnlwgt (final weight), education, education-num, marital-status, occupation, relationship, race, sex, capital-gain, capital-loss, hours-per-week, native-country, and income. With so many columns, maintaining high data quality is important. Training on poor-quality data could lead to unreliable or biased model predictions, reduced accuracy, and misleading conclusions. This report will analyze how feasible the Adult Income Dataset (AID) is for company analytics in 2025 with a focus on 2 data quality dimensions: Completeness to discuss potential missed data from bias related to sex and Timeliness to discuss the relevancy of data from 1994.

Garbage in, garbage out. Poor quality data will inevitably lead to poor insights and lost revenue, and so data quality dimensions are necessary to preserve and augment the quality of a dataset (Amir, 2025). The AID is assessed along two quality dimensions: Timeliness and Completeness. Timeliness in data quality refers to how promptly data is recorded and stored. Timeliness is important because there are some applications that rely on time-sensitive data to make critical business decisions. Take, for example, the calculation of expected incomes for 2025 using AID. Due to inflation rates, it might make an incorrect prediction and result in missed opportunities for a company. In other words, a company might want to calculate the average salary for an employee based on their job title. By not accounting for inflation, which has more than doubled since 1994, the calculation will most often be significantly lower than the real market offer, losing qualified candidates in the process (Federal Reserve Bank of Minneapolis, 2025). Perhaps a nonprofit wishes to analyze how many women pursued higher education through ‘education’ and ‘sex’ variables. Since the number of women pursuing degrees has been steadily increasing from <25% in 1994 to 47% in 2024, data from 1994 is therefore outdated and insufficient to make conclusions in 2025 (Hurst, K. 2024). Moreover, the original 1994 Current Population Survey only interviewed people living in households and excluded institutionalized populations and the homeless (U.S Census Bureau, 1994). This creates additional bias for homeless women and those in shelters or prisons attempting to pursue higher education, which directly infringes on ethics and fairness principles of data quality (Amir. D, 2025). Even when data was once accurate and complete, lack of timeliness can turn it into a misleading, unstable source for present day decisions. Completeness is the data quality dimension that ensures that required data is present. Completeness is an important dimension because operating on data that lacks values, makes faulty assumptions, or is simply insufficient for analysis welcomes many problems. Consider the ‘income’ variable in the dataset. This variable reports only whether the salary range is less than or equal to $50,000 or greater than $50,000, but the differences between individuals can be huge. Two self-employed people may make $1 million and $60,000, respectively. Thus, it is impossible for, say, a marketing company to accurately target middle- or high-class neighborhoods with income-specific items. The ‘income’ also does not take into consideration external assistance such as outstanding debts from student loans or anything of the like. Two people may report the same $50,000 gross income, but 42% of borrowers with at least $25,000 of loan debt are far more likely to struggle with payments, effectively reducing their disposable income below the dataset’s $50,0000 threshold (Federal Reserve, 2024). These two examples illustrate how completeness is a particularly strong dimension.

Some recommendations for each include not using the AID for a current-year income prediction model. The data is completely out of date and will not be helpful in analyzing trends in the present day due to numerous factors such as significantly more women pursuing higher education, the cost of living changing in each state, which would change incomes accordingly, and the absence of taking in external factors such as loans or assistance. Another recommendation would be to adjust the income threshold, $50,000, for inflation. Specifically, $1 in 1994 is equal to $2.17 at today's rate (Federal Reserve Bank of Minneapolis, 2025). This is balloons; the salary range is from $50,000 to $108,500, rendering any previous analyses difficult to use if they do not account for this. The third and final recommendation would be to explicitly treat unknown values in the dataset as distinct “Unknown” value types. In AID, many of the unknowns are present in the workclass category (see Figure 1). By letting them be their own value instead of dropping the entire row, statistical approaches can be applied to find errors in data collection pipelines by assessing columns that have unusual amounts of unknown values.

In summary, the 1994 UCI Adult Income Dataset is no longer suitable for 2025 business or analytical use. Its shortcomings include not accounting for inflation and the advancement of women’s participation in higher education while also not understanding external/contextual factors such as debt-to-income ratios and sampling bias. These will, without a doubt, cause biased predictions, unfair decisions, and lost opportunities. Companies must retire this dataset, adjust for inflation, and treat missing values explicitly if they want to produce trustworthy and actionable insights in 2025.

https://github.com/user1342234/COMP331-Final-Project