STAT 1500 Group Project:

UCI Adult Income Dataset

Research Question:

- How do education level and gender influence income levels, and to what extent do they contribute to income differences?

Workload Distribution

1. Introduction & Data Summary (Fawwaz) April 4-8

2. Data Wrangling & Preprocessing (Ian) March 21-27

3. Exploratory Data Analysis (EDA) & Visualization (Ian & Jack) March 21-27

4. Analysis & Model Development (Ike) March 28-April 3

5. Conclusion, Recommendations & Report Formatting (Fawwaz) April 4-8

Introduction

The UCI Adult Income dataset is a popular resource for studying how demographic and employment factors relate to income. In this project, our primary goal is to determine whether and how education level and gender (sex) influence an individual's likelihood of earning above $50k per year.

Data Summary

- The dataset contains approximately 48,842 rows, each representing an individual, and 15 features: Age, Work Class, Education, Marital Status, Occupation, Relationship, Race, Sex, Capital Gain/Loss, Hours per Week, Native Country, and Income bracket (>50k or = 50k).
- We found that three features—Work Class, Occupation, and Native Country—had missing values marked as "?", which ultimately led us to drop about 14% of the observations. Our final cleaned dataset retains around 30,000+ rows, ensuring a sufficient sample size.
- We also recoded "Income" as a binary variable (1 if >50K, else 0) and "Sex" as 1 for Male and 0 for Female. This makes it easier to perform the classification and statistical tests described below.

What Does Our Data Contain?

Feature	Type	Description	Qualitative or Quantitative
Age	Integer	Age of the individual	Quantitative
Work class	Categorical	Type of employment(self, private, etc)	Qualitative
Fnlwgt	Integer	Weight of individual	Quantitative
Education*	Categorical	Highest level of education	Qualitative
Education num*	Integer	Encoding of education level	Quantitative
Marital status	Categorical	Status(single, married, etc)	Qualitative
Occupation	Categorical	Current job	Qualitative
Relationship	Categorical	Status(husband, wife)	Qualitative
Race	Categorical	Rate of individual	Qualitative
Sex*	Binary	Male or Female	Quantitative
Capital gain	Integer	Gain income	Quantitative
Capital loss	Integer	Loss income	Quantitative

Hours per week	Integer	Working hours weekly	Quantitative
Native country	Categorial	Country of origin	Qualitative
Income **	Binary	>= 50k or <= 50k	Quantitative

Key: * = Variables used for testing, ** = Target Variable

Variables with missing data

a. Work Class = people did not log in their employment type

b. Occupation = people did not log in their current job role

c. Native Country = people did not log in their country of origin

Issues and Problems faced

1. Missing column/header names:

- When importing and reading the data, column headers were missing. To fix this, when opening the data, we set the header function to FALSE and explicitly renamed the R-given columns with their appropriate names.

2. Dealing with Missing Variables:

- In the Native Country column, 583 rows are missing values, totaling 2%(1.79) of the data. We decided to omit these rows completely because they would not significantly affect the target test.

- In the occupation column, there are 1843 rows with missing values. These values add up to ~6%(5.66) of the data, which is not significant enough to keep, so we have omitted the rows with missing data.

- The work class column has 1836 rows with missing data, which adds up to ~6(5.64)% of the data. This is not significant enough to keep, so we have also decided to omit the rows with missing values.

- Overall, the missing values add up to ~14% of the data, some of which are found in the same rows. This percentage of omitted data will not affect the results or calculations at all since it is too small.

3. Converting Qualitative Data to Quantitative Data:

- We've decided to convert both income and sex to binary data (1 or 0) because there were only 2 options to input for both categories. It would also make it easier to construct a frequency table, which we would use for graphing if the data were in binary form.

- Income: 0 represents <= 50K and 1 represent >50K
- Sex: 0 represents Female and 1 represents Male

4. Class Imbalances:

- In our cleaned dataset there was a significant difference as the number of Male entries is significantly larger than the number of Female entries. This will lead to inaccuracies in our calculations as there will be class bias and poor generalization sentiments.
- To handle this discrepancy, we will use proportion when comparing the 2 directly.
- Little to no class imbalances were detected in the education levels of our dataset, although our most significant entries appear to be on the 'HS-grad' level and 'Some College'

Patterns and Trends Noticed

1. Education level and Income(no gender):

- A clear positive trend is shown: As educational levels increase, the proportion of high-income earners (>50k) increases.
- At the highest education levels, more than half the individuals are high-income earners(>50k)
- At low education levels, almost all are low-income earners (<50k)

2. Income and Sex:

- Clear discrepancies between the parties in terms of income are shown.
- There is a greater proportion of high-income males compared to high-income females.
- There is a greater proportion of low-income females compared to low-income males.

3. Income, Sex, and Education:

- For Males and Females, the more advanced the education, the higher the proportion of people earning > 50k
- Males have a noticeably higher proportion of people earning >50k compared to women at all education levels
- Females, even at the highest education level, have a lower proportion of individuals earning >50k than males who are not as educated.

4. Educational Level by Sex:

- Males have a higher median education level than females
- Males have a wider IQR(Inter-Quartile Range), indicating males have more variation in their education levels compared to females
- Both genders show some outliers, with females showing more outliers on the lower education level

5. Income by Sex:

- Males show a significantly higher proportion of individuals earning >50k
- Females show a significantly higher proportion of individuals earning <50k

Interpretation of Findings

Why Logistics Regression?

- The logistic regression model uses Education_Num and sex as predictors for Income
- The exponentiated coefficients indicate how a unit increase in the education level or being a male affects the odds of earning more than $50k

Logistic Regression Findings

- If the coefficient for Education_Num is positive, higher education increases the probability of earning above $50k. If the coefficient for sex is positive, then males are more likely to earn above $50k.
- As seen in the model, the coefficients of Education_Num and Sex are positive, 55.78 and 37.01, respectively. This means that both variables increase the probability of earning more than $50k

Why Decision Tree Model?

- A classification tree is trained using Education_Num and sex to predict Income.

Decision Tree Findings

- The decision tree shows that the probability of earning more than $50k is mostly based on your education level, with the percentage being 75%. Although the model also illustrates that sex is a factor, the percentage increase in earnings of earning more than $50k by being a male is 18% and being female is 7%, which are both insignificant.

Linear Regression Findings

- The values show that after each additional year, the increase in income for education level and sex is estimated between ~$1400 and ~$5000.
- Our R-squared value is 0.1585, which illustrates that there is a 15.85% variance in income. This suggests that other factors that were not measured also play a large role.
- Residual spread shows a minimum of -16154 and a maximum of 32208, indicating that the model has underpredicted for some individuals and overpredicted for others.
- The linear regression shows that education and sex are strong predictors of income, but a random forest could

Why T-Test Model?

- Our T-test will compare the income distributions between males and females.
- T-Test is highly reliable as we have a large sample size of degrees of freedom = 26670
- If the p-value is below 0.05, it will indicate a significant difference in income based on gender

T-Test Result Findings

- The test statistic value(- 43.816) indicates that females have a lower mean income than males. An example of this is seen in the sample estimation, where females have a value of 0.114 and males have a value of 0.314. Dividing these two values will show that males earn 2.75 times higher income than females
- Our p-value(2.2e-16), is significantly below the significance level of 0.05. Therefore, we have to reject the null hypothesis as there is strong evidence that income differs by sex

Conclusion

Our analysis of the UCI Adult Income Dataset illustrates that both **education** (measured by Education_num) and **gender** (male vs. female) have statistically significant impacts on an individual's likelihood of earning more than $50k per year. The **logistic regression model** confirms that:

- Each additional year of education leads to higher odds of earning above $50k.
- Being male (in this dataset) further increases the likelihood of higher income, even when controlling the educational attainment.

Similarly, the **decision tree** and **random forest** models underscore the importance of education, showing clear splits at higher education levels; however, sex remains a strong predictor, reinforcing the presence of a notable gender gap.

To quantify the extent of this gap, the **two-sample t-test** reveals a highly significant difference in average income between males and females, meaning the mean proportion of high earners is substantially higher among men. Moreover, while the **linear regression** on a synthetic "Income_Continuous" variable provides additional evidence of the positive relationship between education and income, its relatively low R-squared value (~15%) implies that other factors (e.g., occupation, hours worked, or industry) also play a large role in determining earnings.

Overall, these findings highlight persistent inequalities along gender lines and the clear benefit of increased educational attainment. Nonetheless, the presence of other relevant variables—like occupation type, work hours, and age—warrants further exploration to get a fuller picture of income determinants.

Recommendations

1. Expand Feature Set
   Future models should incorporate additional predictors—such as occupation, marital status, and weekly hours worked—to more accurately capture the complexity of real-world wage determination.

2. Address Class Imbalance
   Since the dataset shows a larger proportion of males than females, advanced techniques (e.g., oversampling, undersampling, or class weights) may reduce potential model bias. Implementing these methods could help ensure more balanced predictions across genders.

3. Deeper Gender Gap Analysis
   Given that men in this dataset earn >$50K at a higher rate, future research might explore intersectional factors, such as race, marital status, and industry sector. This would help clarify whether the observed gap is consistent across subpopulations or influenced by specific demographic segments.

4. Improve Data Quality & Coverage
   Roughly 14% of the data was discarded due to missing entries in Work Class, Occupation, or Native Country. Encouraging more complete data collection, or using imputation strategies, would preserve a larger, more representative sample and potentially reveal trends missed in the reduced dataset.