Intermediate Project Report: Adult Income Analysis

Matthew Sawyer and Destanie Pitt

March 19th 2025

1 Introduction & Objective

This project analyzes the Adult Income dataset to determine whether an individual earns more than $50,000 per year based on the various features such as age, education, occupation, and working hours. The main objectives of this project are to:

- Identify factors effecting income level
- Evaluate classification modeling techniques (an individual makes more or less than $50,000)

2 Data and Exploratory Data Analysis (EDA)

The dataset used in this study is the UCI Adult dataset which consists of demographic and employment related attributes. It includes features such as age, work class,education level, marital status, occupation, relationship, race, sex, capital gain, capital loss, hours per week, and native country. Additionally, the classification variable indicates whether an individual's income exceeds $50,000 a year or not.

- **Data Cleaning:** The dataset contained missing values represented by "?". As part of the data cleaning process, these placeholders were replaced with "NA" and removed to ensure consistency in the analysis.
- **Data Transformation:** Ensuring categorical variables were properly encoded using one-hot encoding.
- **Visual Analysis:** Creating scatter plots, box blots, confusion matrices, and correlation matrices to identify trends and potential relationship between demographic and employment related attributes and income.

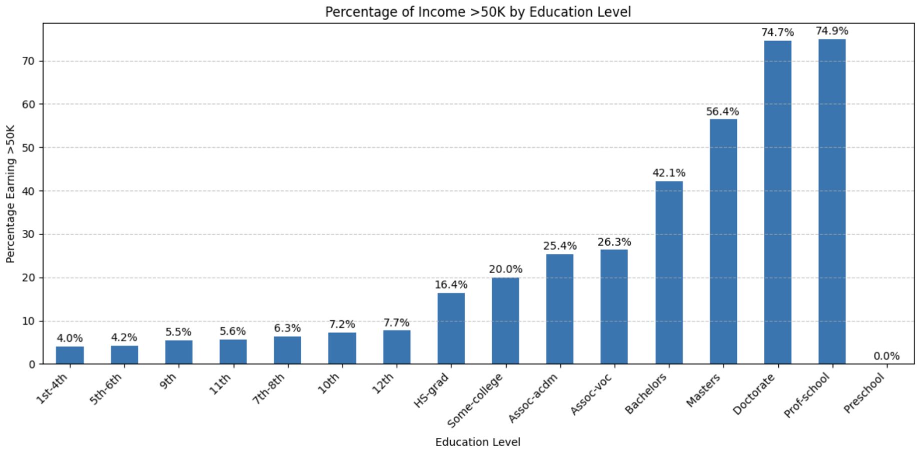


Figure 1: Bar chart illustrating the percentage of individuals earning more than $50k by education level.

Figure 1 is a bar chart that examines the relationship between education level and the percentage of individuals earning more than $50K per year. By displaying the income distribution across different education levels, the visualization highlights how earnings increase with higher education. The data reveals a clear upward trend, where individuals with higher education levels are more likely to earn above $50K. Notably:

- Graduate degrees show the highest earning, with over 74% of individuals surpassing the $50K threshold.
- Individuals with less than a high school diploma have significantly lower chances of earning more than $50K These results suggest that higher education strongly correlates with increased income, reinforcing the idea that advanced degrees provide greater financial opportunities.

3 Modeling Approaches

To address the research questions, several basic modeling techniques were applied.

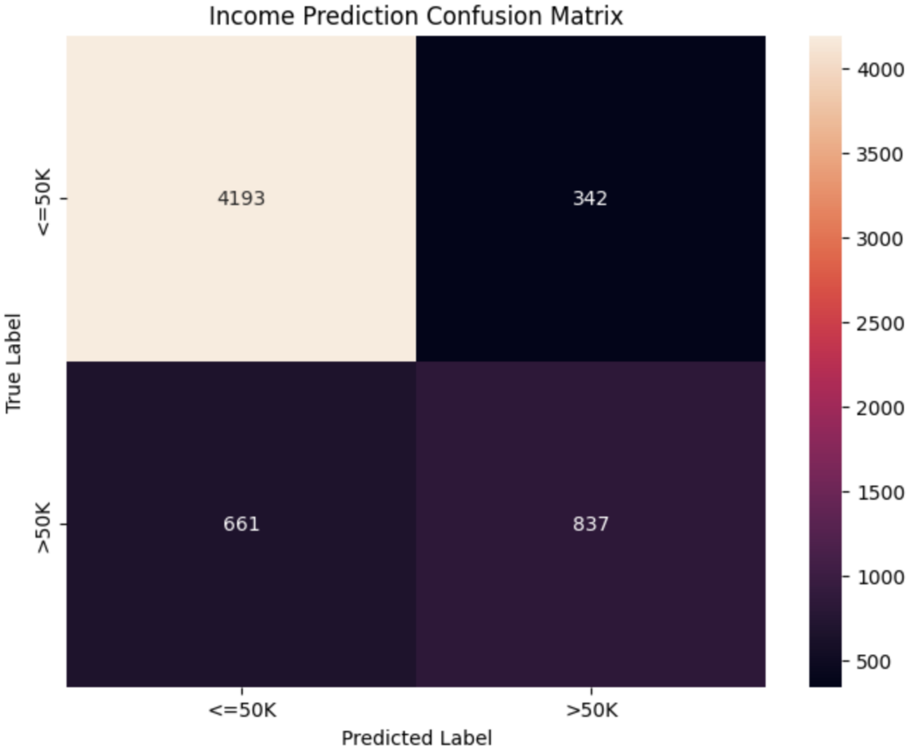
3.1 3.1 Regression Analysis

Regression analysis was ultimately deemed unsuitable for the income prediction due to the following:

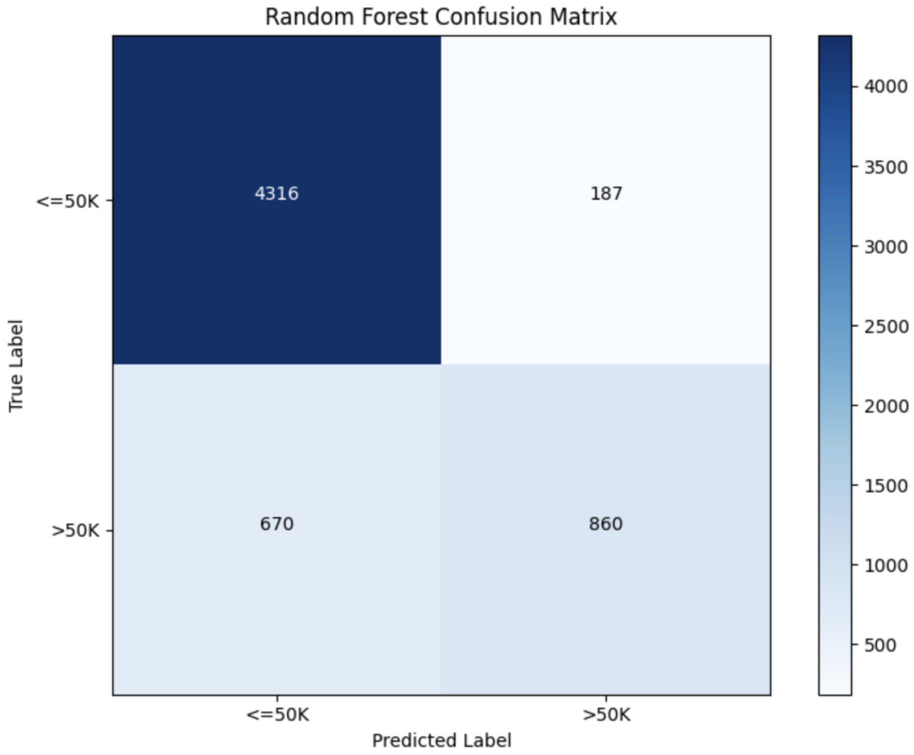
- **Binary Target Variable:** The target variable, income, is categorical ($<= 50K$ or $> 50K$), meaning it does not follow a continuous numerical scale. Since regression models assume a continuous dependent variable, it made them inappropriate for this problem. Given the limitations of regression analysis for this binary classification problem, logistic regression and other classification based approaches were explored instead.

3.2 3.2 Classification Analysis

- **Logistic Regression:** This method was used to model the probability of income being $> 50K, <= 50K$. As show by the following figure,



This confusion matrix represents the performance of the income prediction model. The model correctly classified 837 individuals as earning more than $50K, and 4,193 individuals as earning $<= 50K$. The model incorrectly classified 342 individuals as earning $> 50K$ and 661 individuals as earning $<= 50K$. The model appears to perform better at predicting lower income individuals with fewer false positives. However, there 661 false negatives, meaning it underestimates higher earners. The model is reasonably effective, improvements could focus on reducing false negatives to improve its ability to correctly classify individuals earning more than $50k.



- **Random Forest:** Random forest was the second model we evaluated on the data. It appears to more accurately classify individuals according to income, especially those who earn greater than \$50,000. This model had a classification precision of 0.87 for those earning $<= 50,000$, and a classification precision of 0.82 for those earning $> 50,000$.

The following table shows the test error rates for our chosen methods:

Method	Test Error Rate
Logistic Regression	0.17
Random Forest	0.15