Introduction

The goal of this project is to understand what factors are most strongly related to individual income, and whether income can be accurately predicted using demographic and employment information. Specifically, we study whether a person earns more than $50,000 per year based on census-style features such as age, education, work hours, occupation, and marital status.

This analysis has two main components. First, we treat income prediction as a supervised learning problem and evaluate how well different models perform. Second, we explore the structure of the data using unsupervised methods to see whether natural groupings exist even without using income labels. Together, these approaches allow us to study both prediction performance and data structure.

If successful, this analysis could be useful for understanding which features matter most for income-related outcomes and for building interpretable models that support decision-making in policy, economics, or business settings.

The DataSet

The dataset used in this project is a cleaned version of the U.S. Adult Census Income dataset. It contains demographic and employment information for over 48,000 individuals. The target variable is a binary indicator of whether annual income exceeds $50,000.

The available variables include age, education level, years of education (education-num), work hours per week, occupation, marital status, work class, relationship status, race, sex, capital gain and loss, and country of origin. These variables represent a mix of numerical and categorical data.

Before modeling, several preprocessing steps were required. Rows containing missing values or unknown categories were removed. The categorical variables were converted using one-hot encoding. Some variables were simplified, such as replacing education with education-num and converting sex into a binary variable. After preprocessing, the final dataset contained 45,222 observations and 43 features.

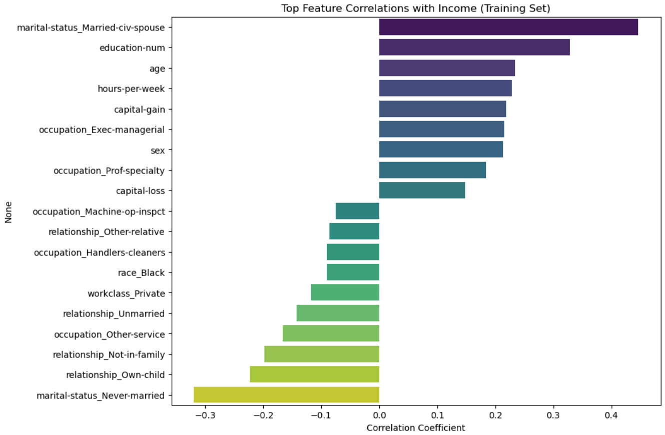
One concern about data quality is class imbalance. Only about 25% of individuals earn more than $50K, which affects metrics such as accuracy and recall. This imbalance motivated the use of ROC-AUC and F1-score in addition to accuracy.

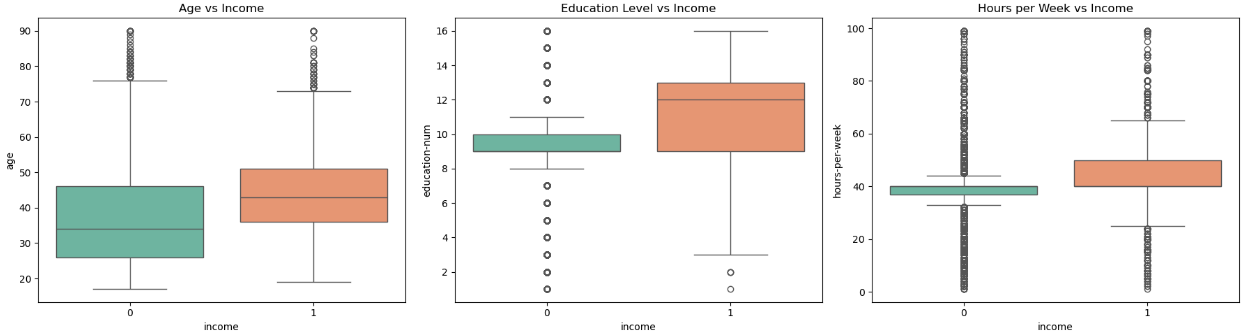
Results

Exploratory Data Analysis

We first explored the training set to understand which features are most correlated with income.

The correlation analysis shows that marital status (Married-civ-spouse) has the strongest positive correlation with high income (≈ 0.45). Education level, age, hours per week, and capital gain also show moderate positive correlations. In contrast, being never married or having an “own child” relationship shows a negative correlation with income.





Boxplots further support these findings. Individuals earning more than $50K tend to be older, have higher education levels, and work longer hours compared to the lower-income group.

Primary Question: Income Prediction

We trained two supervised models: Logistic Regression and Random Forest.

Logistic Regression achieved an accuracy of 0.85 and a ROC-AUC of 0.91. The model performs well overall but has lower recall for the high-income class, which is expected given class imbalance.

Random Forest slightly outperformed Logistic Regression, with an accuracy of 0.8535 and a ROC-AUC of 0.9057. Increasing the number of trees improved performance, with diminishing returns after about 50–100 estimators.

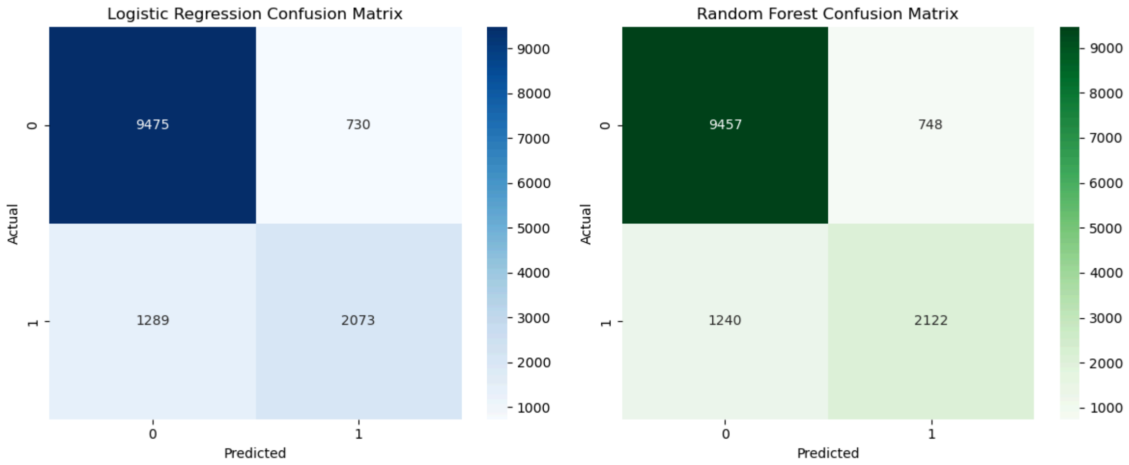


Figure 3: Confusion matrices for Logistic Regression and Random Forest

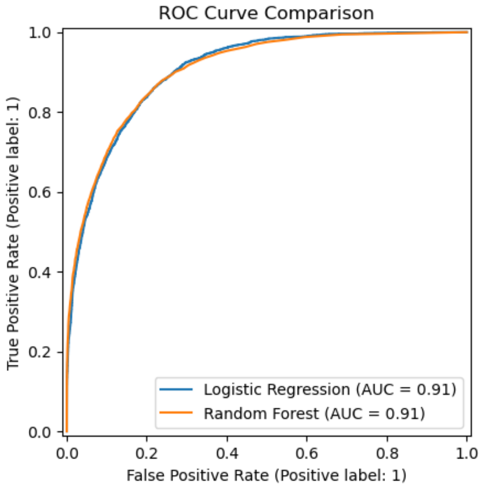


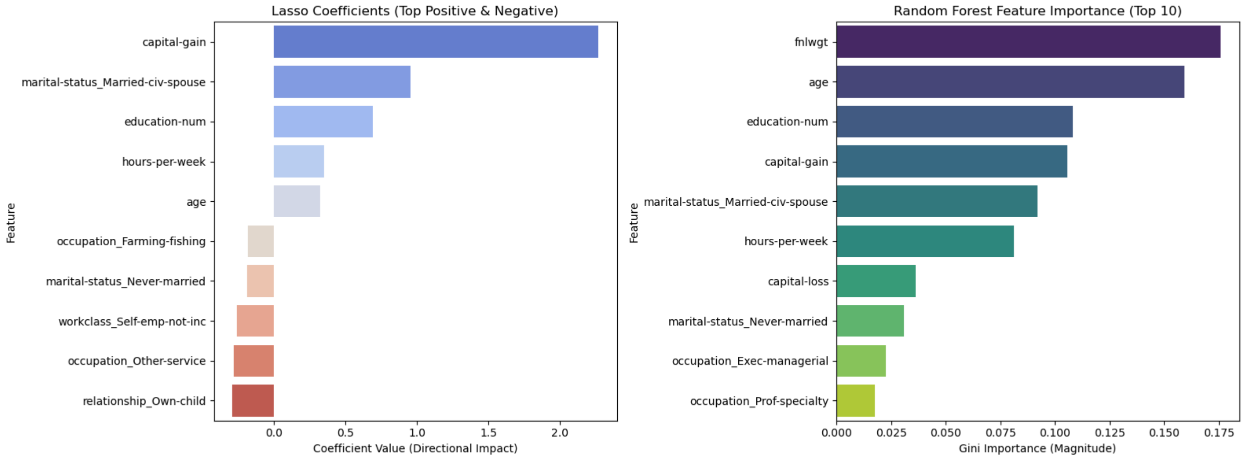
Figure 4: ROC curve comparison for Logistic Regression and Random Forest

The ROC curves for both models are very similar, suggesting that while Random Forest captures more non-linear patterns, both models are strong predictors. In summary, the predictive performance is reliable, but predictions for high-income individuals are more uncertain than for low-income individuals.

Secondary Question 1: Feature Importance and Interpretation

To better understand which features drive income predictions, we used Lasso regression and Random Forest feature importance.

Lasso identifies capital gain as the strongest positive predictor, followed by being married, higher education, longer work hours, and older age. Features such as having children, working in low-paying service occupations, or being never married reduce the probability of high income.



Random Forest highlights a different pattern. The feature fnlwgt appears as the most important variable, even though its correlation with income is near zero. This suggests that Random Forest may overemphasize high-cardinality features. Other important features include age, education, capital gain, and marital status.

Comparing these results suggests that Lasso provides a more interpretable and stable view of feature importance, while Random Forest is more prone to capturing noise.

Secondary Question 2: Unsupervised Structure

We used PCA and K-Means clustering to explore the structure of the data without using income labels.

The PCA results show that the first two principal components explain only 12.95% of the variance. The 2D projection shows heavy overlap between income groups, indicating that the data is high-dimensional and not linearly separable.

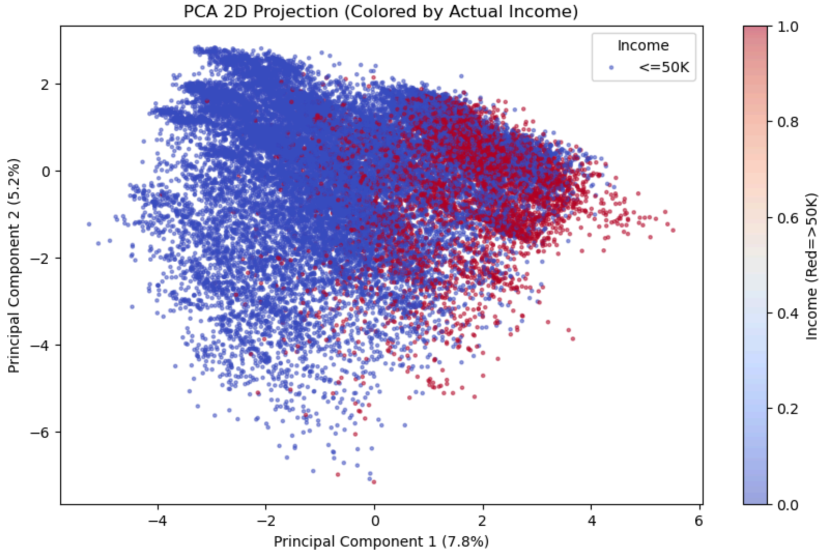


Figure 7: PCA 2D projection colored by income

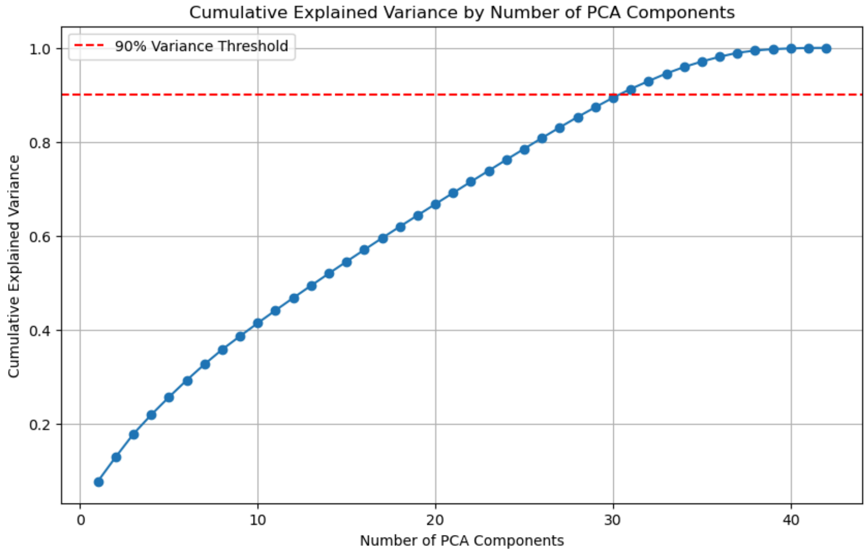


Figure 8: Cumulative explained variance vs. number of PCA components

K-Means clustering (k = 3) revealed meaningful structure. One cluster contains about 45% high-income individuals, while the other two clusters contain fewer than 8%. This high-income cluster consists of individuals who are older, more educated, and work longer hours. This shows that even without labels, the data contains a natural “high-achiever” group.

Discussion

This project shows that income is strongly associated with education, age, marital status, work hours, and capital gains. Supervised models can predict income with high accuracy and strong ROC-AUC, but class imbalance limits performance on high-income individuals.

From an interpretability perspective, Lasso provides clearer and more trustworthy feature importance than Random Forest. Unsupervised methods reveal that meaningful structure exists in the data, even without income labels.

Overall, the results are consistent, interpretable, and useful for guiding decisions or further analysis. A simple takeaway is that education and sustained work effort are central drivers of income, and combining multiple modeling approaches provides a more complete understanding than relying on a single method.