

Abstract

Income inequality in the United States is characterized by a combination of demographic, educational, and occupational factors. This paper investigates which variables are most influential in predicting whether or not the annual income of a person exceeds \$50,000 annually and evaluates the characteristics of individuals who exceed their predicted income classification. Using an income distribution dataset from Kaggle, statistical models such as logistic regression, random forest, decision tree, and XGBoost were implemented and the results of each model was compared. The analysis of the results examined model performance, variable coefficients, and false negative cases.

Results indicate that logistic regression with balanced class weights provide the most accurate analysis of the data. The model found that the factors that generally influenced income the most were factors that people could control such as hours per week they worked or the occupation that they work in. The analysis also found that false negative cases where the person was predicted to make less than \$50,000 but exceeded this threshold were generally attributed to capital gain and loss through investments and assets.

Overall, the findings demonstrate that income predictions are mostly influenced by variables that the subject can control rather than assigned variables that one has no control over such as race and gender.

Introduction

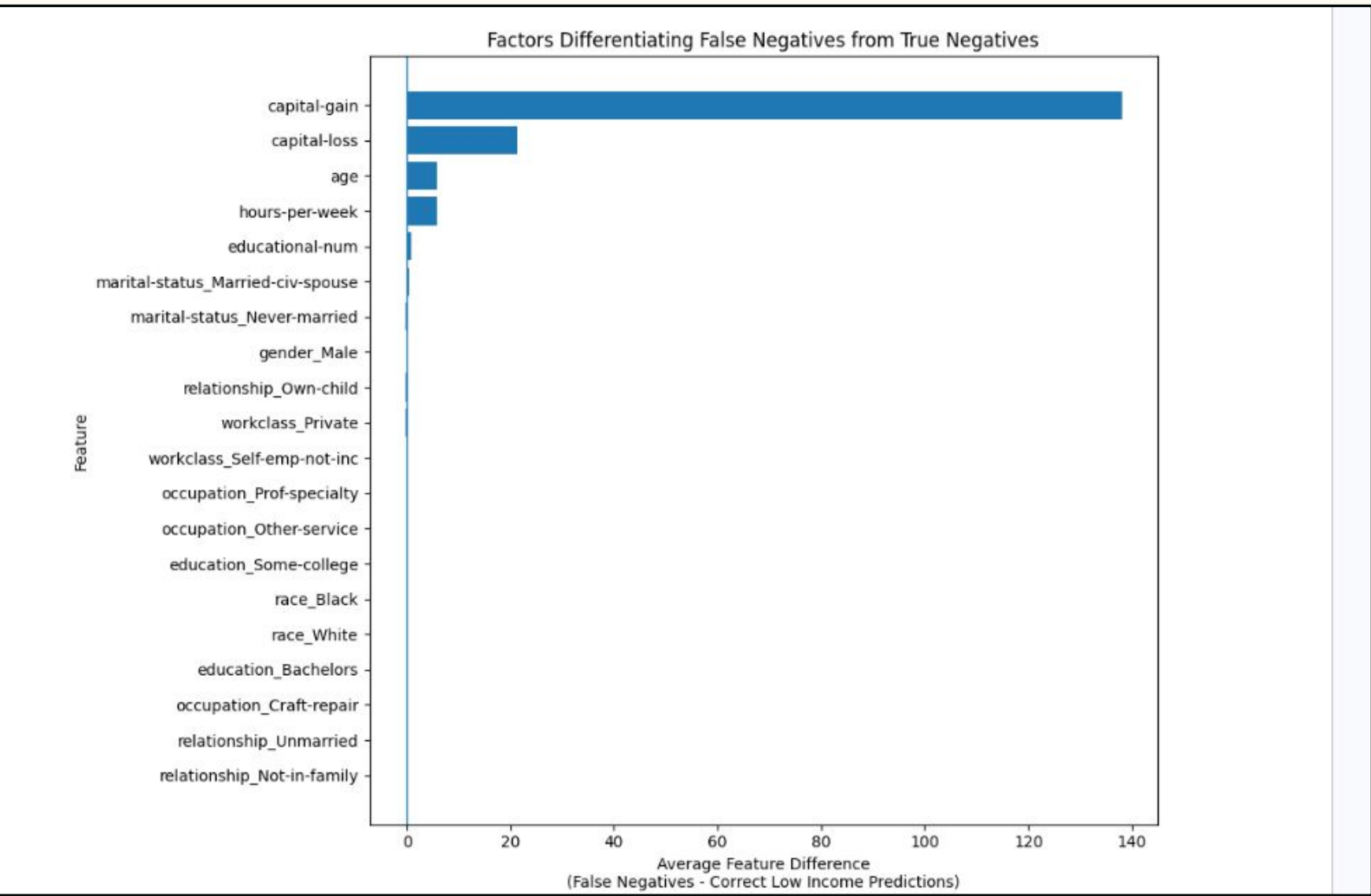
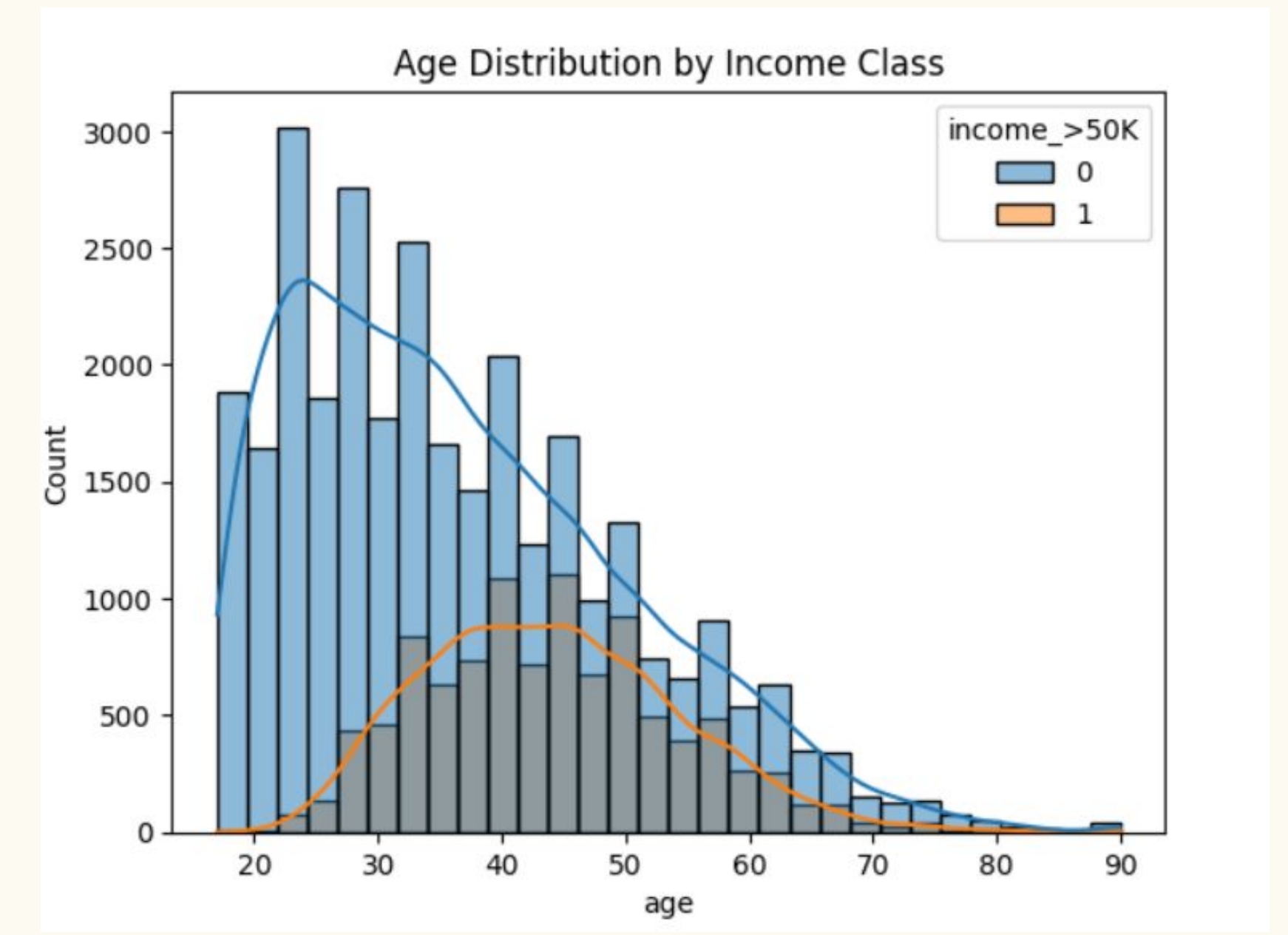
- Income Distribution is a major socioeconomic problem in the United States
- Data science and statistical analysis can provide various insights regarding income distribution
- This research paper examines demographic, educational, and occupational factors regarding income distribution
- It also identifies variables that influences individuals to exceed their projected income and differentiates false negatives from true negatives
- Research Question: Which factors are most influential in predicting whether an individual will earn more than \$50,000 annually and what factors distinguish those who outperform their predictions.

Methods

- I used datasets from Kaggle, UC Irvine ML data library, and census data from IPUMS USA. selected based on usability and implementation.
- Used demographic, occupational, and educational variables to predict a person’s income
- Applied logistic regression to the data and compared results to those from other statistical models such as Decision Tree, Random Forest, and XGBoost.
- Applied class balanced weighting for optimal model accuracy
- Used data visualization to compare the net effects of each variable on a person’s income (how much did that variable influence the prediction).

Results

- Most individuals were predicted to make under \$50,000 with a very high confidence; still 1200 false negatives occurred
- The strongest factors distinguishing false and true negatives were hours per week, age, capital gain/loss, and education
- The highest concentration of high earners was around age 50 and as age increased income increased on average to a certain extent
- Strongest predictors were related to education, marital status, and education.



Discussion

- Profits from investments and assets helped distinguish false negatives from true negatives
- False negatives were also caused by hours per week working a job, age, and education
- Demographic factors only cannot explain income distribution, income is more likely to be controlled by factors that the subject has control over
- Concentration of high earners in ages 30-50 highlight the importance of gaining experience and career development in order to earn more money.

Conclusion

- Logistic regression provided an effective framework for predicting whether a person will make more than 50,000 USD annually
- Logistic regression performed the best out of all statistical models for prediction accuracy and balanced analysis
- Income is ultimately dependent on multiple independent variables, and variables that the subject can control tend to be the most influential
- The findings establish that data science and statistical analysis can provide an accurate framework for performing analysis on income distribution

References

“Adult.” *UCI Machine Learning Repository*, <https://archive.ics.uci.edu/dataset/2/adult>. Accessed 1 August 2026.

Cambridge University. “The K-Statistics Approach to Income Distribution Analysis.” *Cambridge University Press*, 2025, <https://www.cambridge.org/core/elements/abs/statistics-approach-to-income-distribution-analysis/46C28A9C3B9F5AA8179CA3E5017E31A9#metrics>.

Labonte, Marc, et al. <https://www.congress.gov/crs-product/R44705>.

“Income Dataset.” *Kaggle*, <https://www.kaggle.com/datasets/mastmustu/income>. Accessed 1 August 2026.

IPUMS USA, <https://usa.ipums.org/usa/>. Accessed 1 August 2026.

Saez, Emmanuel, and Gabriel Zucman. “The Rise of Income and Wealth Inequality in America: Evidence from Distributional Macroeconomic Accounts.” *Journal of Economic Perspectives*. *American Economics Association*, <https://www.aeaweb.org/articles?id=10.1257/jep.34.4.3>.

