U.S. Adult Census: Income Prediction with Logistic Regressiona

- Benjamin Gerochi - Izzy Zhou - Michael Tham - Yui Mikuriya

Table of contents

Prepared for DSCI 310 by Group 8: - Benjamin Gerochi - Izzy Zhou - Michael Tham - Yui Mikuriya

1 (1) Summary

This report investigates income prediction using the UCI Adult Dataset(Kohavi and Becker 1996), which compiles demographic and income data from the 1994 U.S. Census. The primary objective is to predict whether an individual earns over $50,000 annually using factors such as age, education level, and hours worked per week. By employing a logistic regression model, the analysis effectively predicted income levels on test cases while assessing model performance using metrics like the ROC curve (AUC: Area Under the Curve), sensitivity, specificity, and accuracy. The findings underscore that while the model achieves robust overall accuracy, there are challenges with false positives that warrant further refinement.

The insights derived from this study not only validate the role of education and work intensity in income determination but also suggest avenues for future research, such as integrating geographic and intersectional demographic variables to capture the complexities of income disparities. Overall, the analysis offers a comprehensive approach to understanding income inequality and provides actionable information for policy makers and individuals aiming to navigate economic opportunities.

2 (2) Introduction

2.1 Dataset Overview

The dataset selected for this project is the UCI Adult Dataset(Kohavi and Becker 1996), available through the UCI Machine Learning Repository(Dua and Graff 2017). It contains demographic and income data collected by the **U.S. Census Bureau** and is widely used for predicting whether an individual's income exceeds **$50,000 per year** based on various demographic factors.

2.2 Dataset Details:

- **Dataset Name**: UCI Adult Dataset(Kohavi and Becker 1996)
- **Source**: 1994 U.S. Census database, compiled by Ronny Kohavi and Barry Becker

- **Total Observations**: 32,561
- **Total Variables**: 15

2.3 Variables and Their Types

Table 1: Variable Index and Descriptions

Variable Index	Variable Name	Type	Description
0	age	continuous	Age of the individual
1	workclass	categorical	Employment sector (e.g., Private, Self-emp-not-inc, State-gov)
2	fnlwgt	continuous	Final weight, representing the number of people the observation represents in the population
3	education	categorical	Highest level of education attained
4	education-num	continuous	Numerical representation of education level
5	marital-status	categorical	Marital status (e.g., Never-married, Married-civ-spouse)
6	occupation	categorical	Type of occupation (e.g., Adm-clerical, Exec-managerial)
7	relationship	categorical	Relationship of the individual to the household (e.g., Husband, Not-in-family)

Variable Index	Variable Name	Type	Description
8	race	categorical	Race of the individual (e.g., White, Black)
9	sex	categorical	Gender (Male/Female)
10	capital-gain	continuous	Capital gains earned
11	capital-loss	continuous	Capital losses incurred
12	hours-per-week	continuous	Average hours worked per week
13	native-country	categorical	Country of origin
14	income	categorical	Income level (<=50K, >50K)

This dataset includes both **categorical** and **numerical** variables, making it suitable for analyzing relationships between **demographic attributes** and **income levels**. Further **exploration and preprocessing** may involve handling **missing values** and **encoding categorical features**.

2.4 Research Question

How accurately can key demographic factors predict whether an individual's annual income exceeds $50,000?

This study aims to use demographic variables to predict income levels without pre-assuming key predictors. Our team initially analyzed different aspects of the dataset before deciding to focus on demographic influences on income such as age, education, and hours worked.

2.5 Literature Context

Prior research supports the importance of demographic factors in income prediction. Jo (Jo 2023) analyzed the **Adult dataset** and identified **capital gain, education, relationship status, and occupation** as key predictors. Similarly, Azzollini et al. (Azzollini, Breen, and Nolan 2023) found that demographic differences explained **40% of income inequality** across OECD countries, reinforcing the relevance of our analysis.

2.6 Objective

To develop and evaluate a predictive model that estimates the probability of an individual earning more than $50,000 annually based on their demographic characteristics:

- **Prediction:** Build a robust model to forecast whether an individual's annual income will exceed $50,000.
- **Model Evaluation:** Assess model performance to ensure that the model provides reliable predictions.

3 (3) Methods & Results

3.1 Loading the Required Libraries and Dataset

We will start by importing the necessary R libraries for data analysis and preprocessing.

We then load the dataset into R by referencing the downloaded file path.

Table 2: Raw Adult Income Dataset

39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K
49	Private	160187	9th	5	Married-spouse-absent	Other-service	Not-in-family	Black	Female	0	0	16	Jamaica	<=50K

3.2 Data Wrangling

We will begin by cleaning the Table 2. First, we remove missing values and convert the income column into a factor variable to ensure R treats it as a categorical variable. This transformation is crucial for statistical modeling and visualization, especially when income is used as a binary outcome in logistic regression. We also create new column names to streamline readability and analysis.

Table 3: Cleaned Adult Income Dataset

age	workclass	fnlwgt	education	education_num	marital_status	occupation	relationship	race	sex	capital_gain	capital_loss	hours_per_week	native_country	income
39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K

After removing missing values from the Table 3, we randomly sample 10% of the data (which contains a total of 32561 observations), bringing our sample size to 3256 data points. The sample is then split into training and testing sets (80-20 split) for prediction analysis.

Table 4: Training Set of Adult Dataset

age	workclass	fnlwgt	education	education_num	marital_status	occupation	relationship	race	sex	capital_gain	capital_loss	hours_per_week	native_country	income
51	Federal-gov	106257	Assoc-acdm	12	Married-civ-spouse	Tech-support	Husband	Black	Male	0	0	40	United-States	<=50K

Table 4: Training Set of Adult Dataset

age	workclass	fnlwgt	education	education_num	marital_status	occupation	relationship	race	sex	capital_gain	capital_loss	hours_per_week	native_country	income
50	State-gov	24185	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	38	United-States	>50K
34	Private	34088	HS-grad	9	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	60	Philippines	>50K
22	Private	31650	HS-grad	9	Never-married	Craft-repair	Not-in-family	White	Male	0	0	50	United-States	<=50K
40	Private	137142	HS-grad	9	Married-civ-spouse	Sales	Husband	White	Male	0	0	60	United-States	<=50K
43	Private	18880	Masters	14	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	45	United-States	>50K

Above, we can see that the Table 3 has been successfully split, with our Table 4 containing 2604 rows, representing about 80% of our sample size: 3256.

3.3 Exploratory Data Analysis and Visualization

3.3.1 Pairwise Plot:

To focus on the most relevant variables, we will exclude columns that do not directly contribute to addressing our research question. Hence, we have retained demographic predictors such as age, education level, and hours worked per week. These predictors were chosen based on prior literature Smith-Edgell (2024), theoretical considerations, and empirical evidence from exploratory analyses, which indicate that they have a significant influence on income levels.

Using our Table 4 with the irrelevant columns dropped, we create pairwise plots to examine relationships between continuous variables (`age`, `hours_per_week`, `education_num`) and the response variable, as well as associations among the input variables.

The Figure 1 shows that `age` is right-skewed, `hours_per_week` peaks around 40, and `education_num` has a bimodal distribution. Weak correlations (< 0.6) suggest minimal multicollinearity.

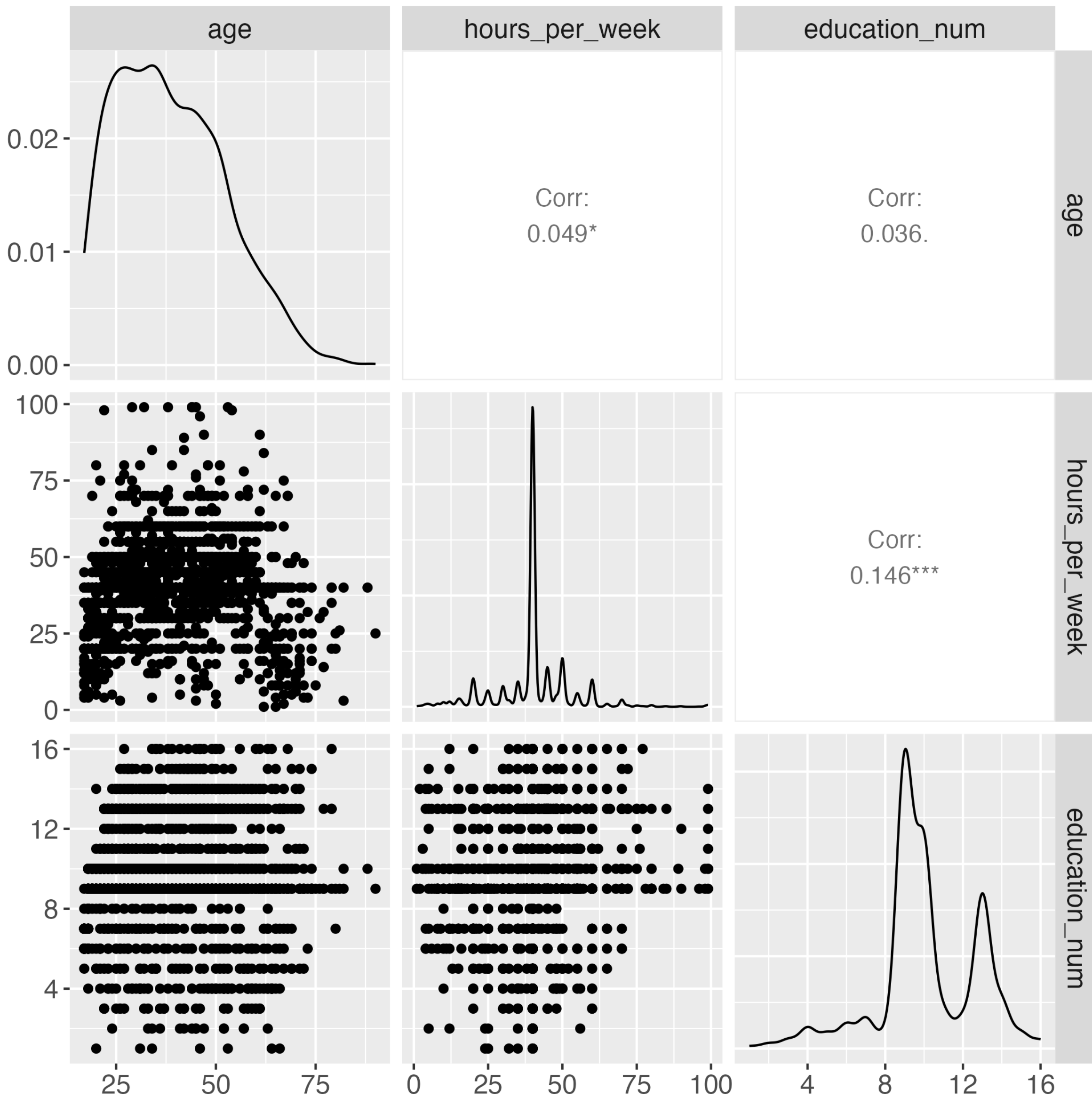

Figure 1: Pairwise Plot of Response and Predictors

The following code generates summary tables for continuous variables, with the code computing key summary statistics: mean, standard deviation, median, variance, maximum, and minimum.

Table 5: Summary Statistics Table of Relevant Predictors

name	mean	sd	median	variance	max	min
age	38.72504	13.560569	37	183.889023	90	17
education_num	10.04224	2.592533	10	6.721227	16	1
hours_per_week	40.10676	12.432628	40	154.570235	99	1

The Table 5 show that the average age is 39 years (SD = 13.56) with a range of 17 to 90. The average education level (education_num) is 10 years (SD = 2.59), reflecting high school or some college education. For hours_per_week, the average is 40.11 hours (SD = 12.43), with a maximum of 99 hours, indicating some individuals work significantly long hours.

3.4 Proposed Method: Logistic Regression for Prediction and ROC Curve for Model Evaluation

Why is Logistic Regression Appropriate?

Logistic regression is suitable for modeling binary outcomes like income categories (<=50K and >50K). It estimates the probability of an individual falling into a specific category based on predictors, then classifies the predictions based on a threshold.

3.4.1 Assumptions:

1. Independence of observations.
2. No high correlation among predictors.
3. A large enough sample size for reliable estimates.

3.4.2 Limitations:

1. Potential underfitting if too little predictors are included.

3.5 Fit the Logistic Regression Model

In the following code, we fit the logistic regression model to the Table 4 using the relevant predictors.

Table 6: Summary of the Logistic Regression Model

term	estimate	std.error	statistic	p.value
(Intercept)	-8.1289099	0.3776223	-21.526561	0
age	0.0471249	0.0040550	11.621338	0
education_num	0.3237071	0.0224509	14.418428	0
hours_per_week	0.0396302	0.0042943	9.228463	0

We can observe from the Table 6 that all predictors were deemed significant (based on the p-values). Furthermore, education number seemed to have the highest coefficient, demonstrating the greatest impact on model predictions.

3.6 Visualizing the ROC Curve

To evaluate the model, we will use the ROC curve to visualize the trade-off between sensitivity and specificity across classification thresholds. The AUC (Area Under the Curve) will be calculated to quantify model performance, with values closer to 1 indicating strong discrimination and values near 0.5 suggesting random guessing.

The Figure 2 shows us that the AUC (Area Under the Curve) values obtained for the model 0.7928102 is significantly above 0.5, indicating that the model performs much better than random guessing. The high AUC value suggests that the model has strong discriminatory power, effectively distinguishing between individuals earning `<=50K` and `>50K` based on the selected predictors.

3.7 Test the Model on the Testing Dataset

Now, we perform the classification analysis and apply the model to the testing dataset and visualize the results of the analysis in a confusion matrix.

3.8 Classification Results and Model Metrics

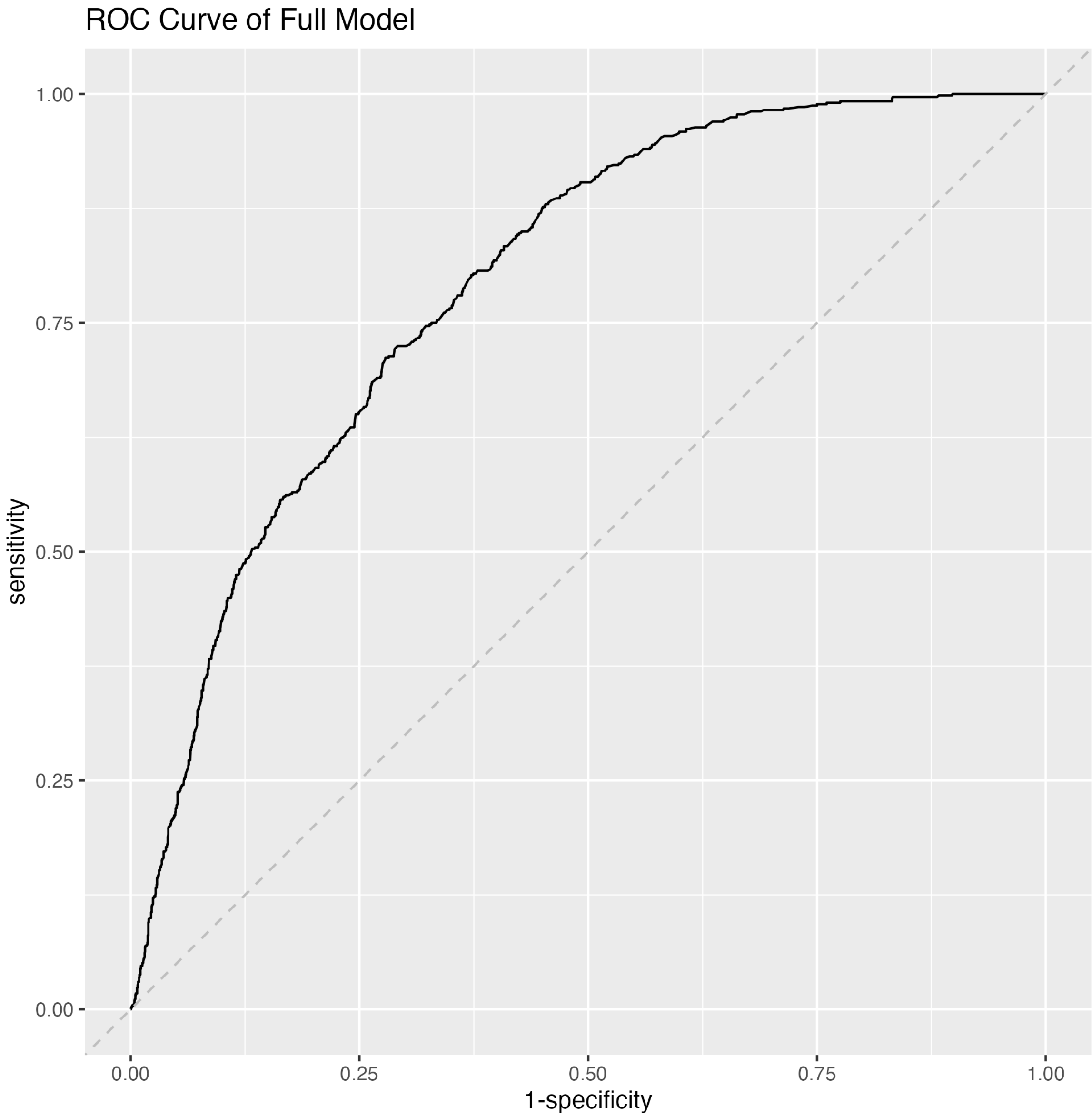

Figure 2: ROC Curve of the Logistic Regression Model

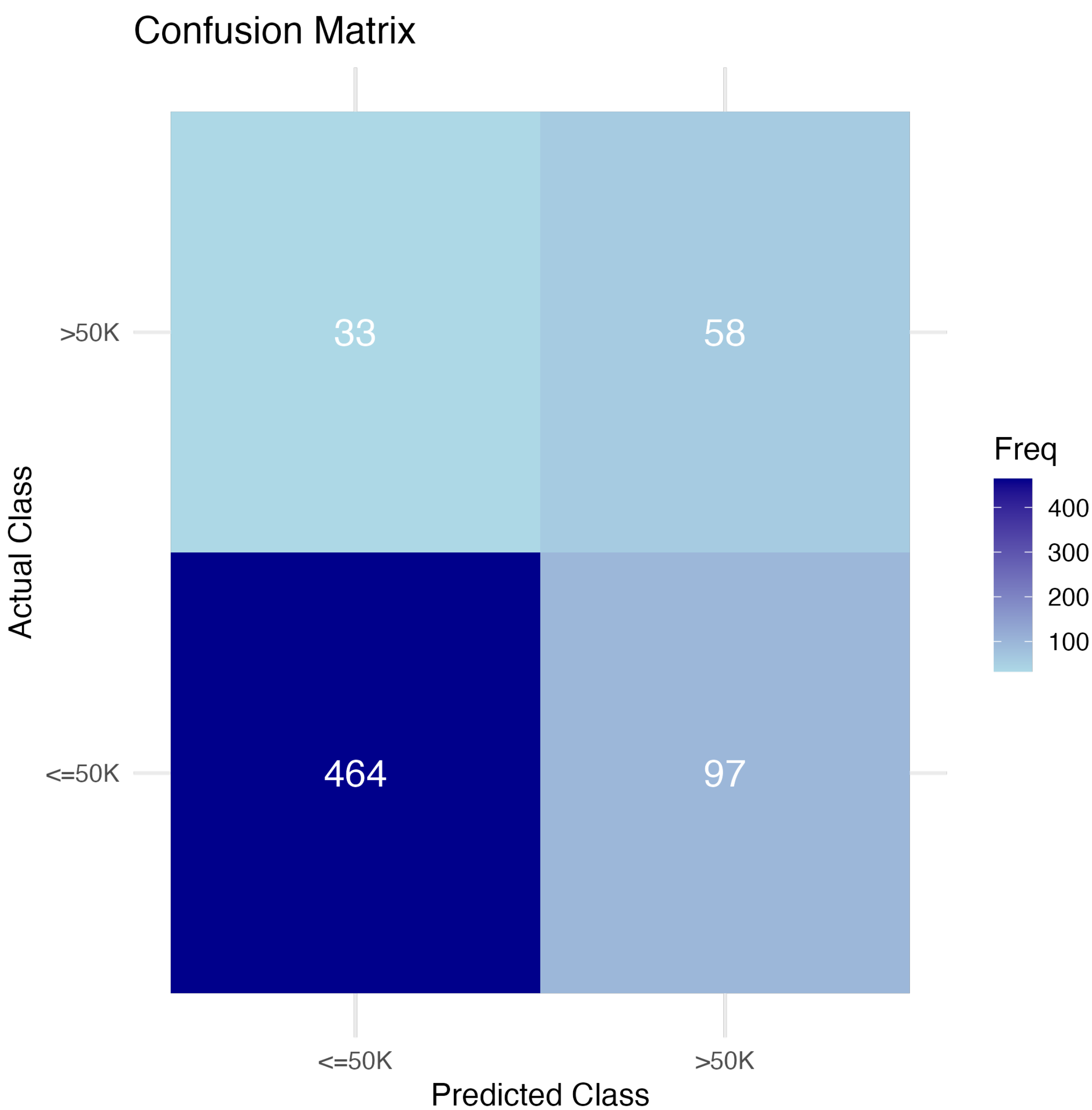

Figure 3: Confusion Matrix of Full Model on Testing Set

Table 7: Classification Results and Model Metrics

Metric	Value
Sensitivity	0.3741935
Specificity	0.9336016
Precision	0.6373626
Accuracy	0.8006135
Cohen's Kappa	0.3587629

From our Table 7, we observe the following metrics: 1. **Sensitivity (SN): 37.4193548%** - The model correctly identifies 37.4193548% of higher-income individuals. 2. **Specificity (SP): 93.360161%** - 93.360161% of lower-income individuals are correctly classified. 3. **Precision (PR): 63.7362637%** - 63.7362637% of predicted >50K individuals actually earn >50K, indicating many false positives. 4. **Accuracy (ACC): 80.0613497%** - 80.0613497% of overall predictions are correct. 5. **Cohen's Kappa ( ): 35.8762918%** - Moderate agreement, better than random chance but room for improvement.

3.9 Interpretation

- Strong specificity, but moderate sensitivity and low precision suggest improvements in identifying high-income individuals.
- High accuracy reflects solid overall performance but overlooks class imbalance.
- Moderate Cohen's Kappa indicates the need for refinement to improve consistency.

4 (4) Discussion

4.1 Summary of Findings and Implications

- The logistic regression model showed strong predictive power (AUC =0.7928102), demonstrating that the model can effectively distinguish income levels better than a baseline.
- These findings can inform policies aimed at reducing income inequality. Education and hours worked were key predictors, emphasizing the need for skill development and work-life balance.
- Understanding the factors behind income disparities can help individuals make more informed career decisions and pursue opportunities for skill enhancement.

4.2 Expectations and Results

- The model's AUC (0.7928102) is strong, reflecting the importance of predictors like age, education, and hours worked. Overall, the results are consistent with expectations from the research study:
  - **Age** correlates with experience, leading to higher salaries.
  - **Education** increases income, with those holding a degree earning significantly more.
  - **Hours Worked** reflects labor input, where more hours can translate to higher pay.

4.3 Future Research

- **Geographic Influence on Income:** Including geographic variables may reveal regional disparities in income linked to education and job opportunities.
- **Intersectionality of Demographics:** Exploring how race, gender, and marital status interact could improve the model's accuracy in predicting income.
- **Health and Disability Status:** Accounting for health conditions or disability could provide additional insight into income disparities by limiting education or work opportunities.

(5) References

Azzollini, L., R. Breen, and B. Nolan. 2023. "Demographic Behaviour and Earnings Inequality Across OECD Countries." *Journal of Economic Inequality* 21: 441–61. https://doi.org/10.1007/s10888-022-09559-1.

Dua, Dheeru, and Casey Graff. 2017. "UCI Machine Learning Repository." University of California, Irvine, School of Information and Computer Sciences. https://archive.ics.uci.edu/ml.

Jo, K. 2023. "Income Prediction Using Machine Learning Techniques." University of California, Los Angeles.

Kohavi, R., and B. Becker. 1996. "UCI Machine Learning Repository: Adult Data Set." UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/adult.

Smith-Edgell, A. 2024. "Proof Point: Financial Returns After a Post-Secondary Education Have Diminished." RBC Thought Leadership. https://thoughtleadership.rbc.com/proof-point-financial-returns-after-a-post-secondary-education-have-diminished/#:~:text=Incomes%20are%20positively%20correlated%20with%20higher%20education&text=Respondents%20with%20a%20bachelor's%20degree,median%20income%20in%20the%20sample.