University of Glasgow | School of Mathematics and Statistics

STATS5085 Data Analysis Skills (Level M) 2024-2025

Analysis of Factors Influencing Individual Income Levels

Group 30

Anurag Choudhary, Ziyu Dong, Keyang Liang,
Zhuohang Qin, Jingzhi Wang, Manyi Yang

Introduction

Analysis Aim

Find key factors influencing individuals' income levels.

Data Source

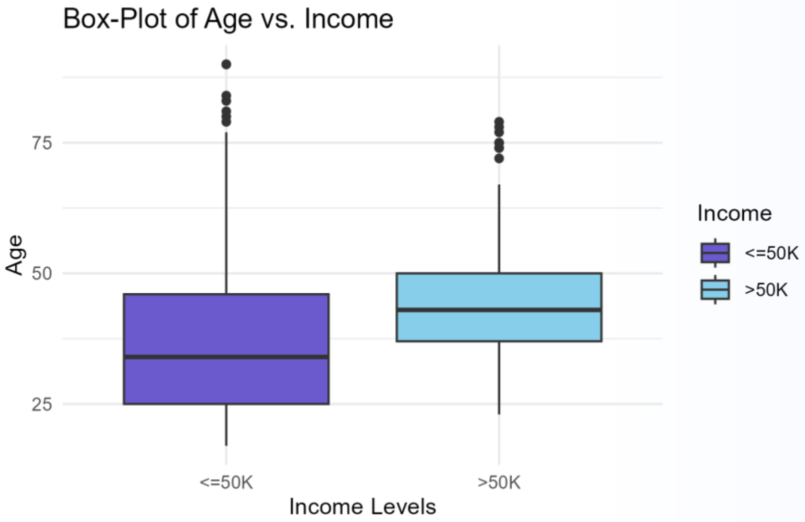
United States Census Bureau, 1994.

Analysis Approaches

Income are categorised into two levels: low-income (≤ \$50k per year), and high-income (> \$50k per year).

GLM and RF are applied and compared to find the relationship between income level and explanatory variables.

Exploratory Data Analysis

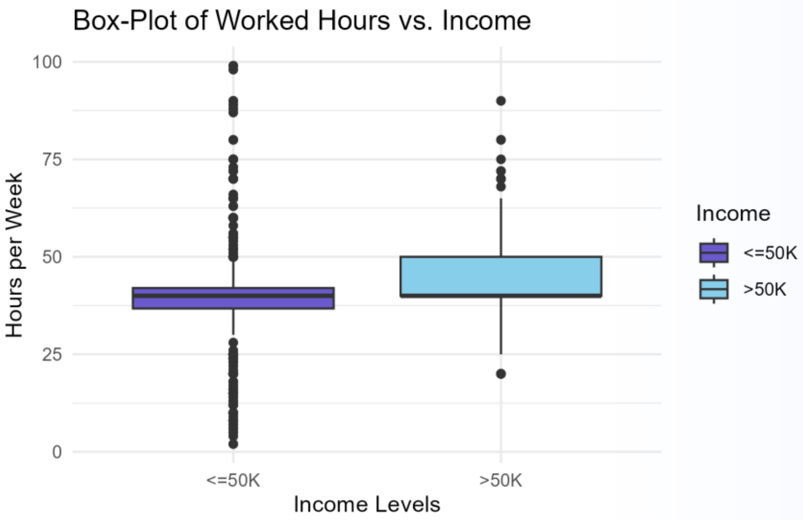


Numerical Variables

- **Age**

Age of high-income group tends to be greater than low-income group.

Exploratory Data Analysis



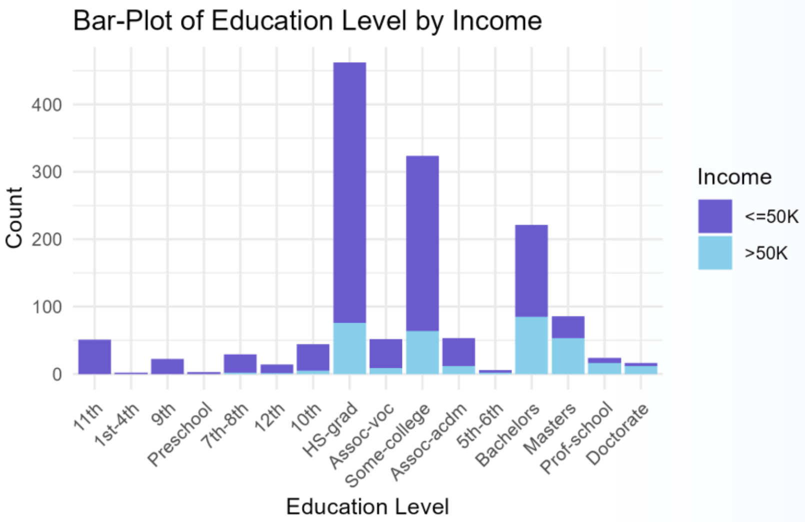
Numerical Variables

- **Working Hours per Week**

The middle 50% of high-income group tend to have more working hours than low-income group.

Meanwhile, low-income group has greater range of working hours and more outliers.

Exploratory Data Analysis



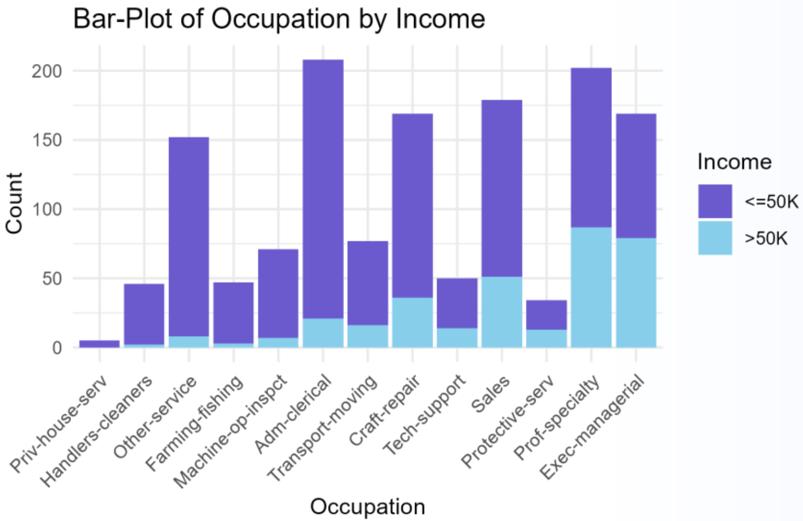
Categorical Variables

- Education Level

Doctorate, **professional school**, and **masters** have the highest proportion of high-income, which is more than 50%.

Education levels lower than **12th** have very low proportion of high-income.

Exploratory Data Analysis

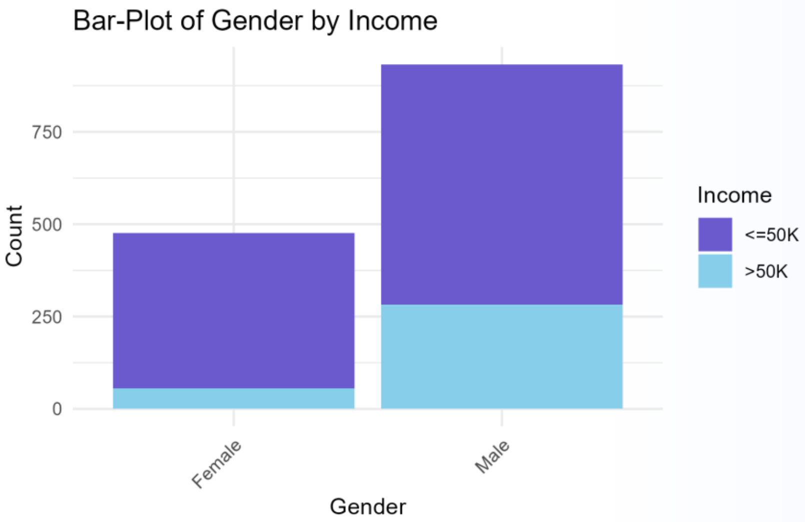


Categorical Variables

- Occupation

Executive managerial and **professional specialty** is nearly 50%, while **house serving** and **handlers cleaners** is almost 0.

Exploratory Data Analysis

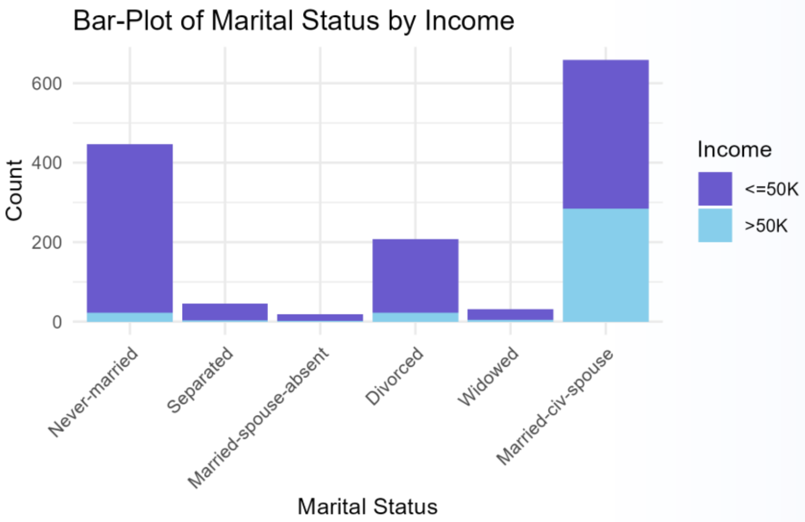


Categorical Variables

- Gender

Male has a higher proportion of high-income people, although the sample size between male and female is unbalanced, which may due to data collecting issues.

Exploratory Data Analysis

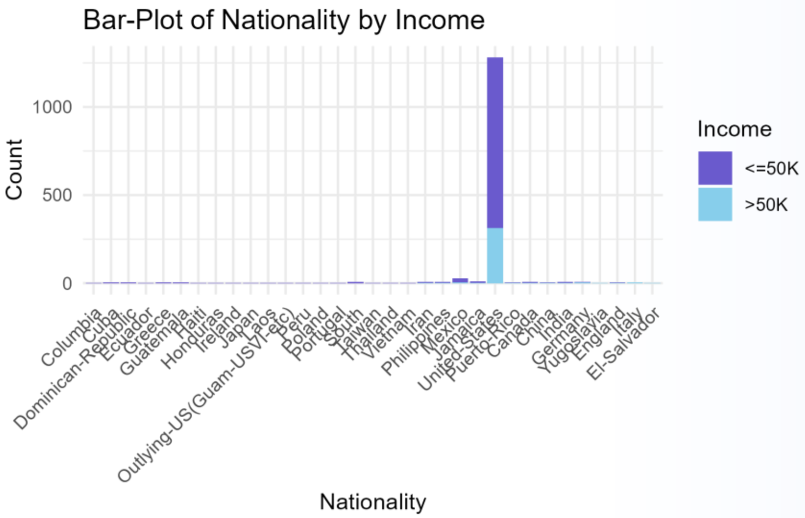


Categorical Variables

- Marital Status

The proportion of high-income people is the highest in the group **Married Civil Spouse**.

Exploratory Data Analysis



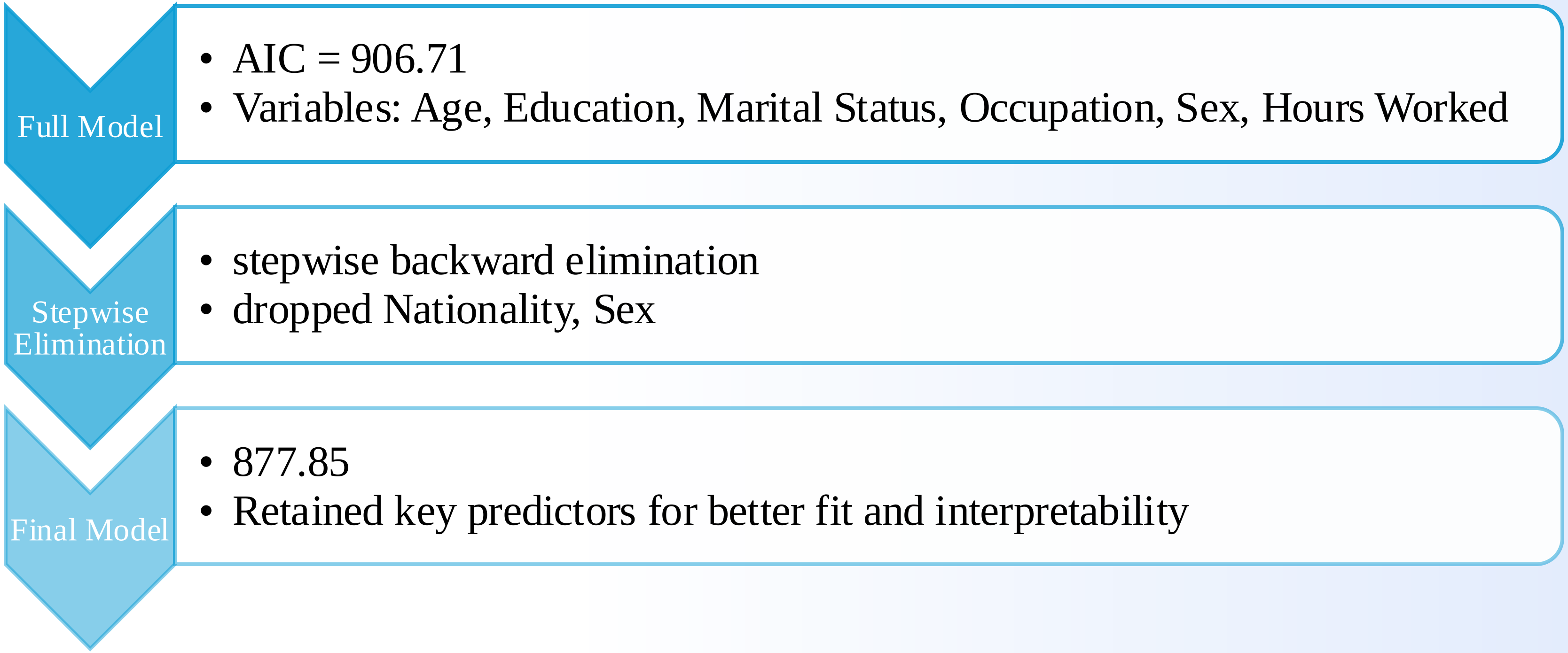
Nationality	United States	Mexico	Jamaica
Proportion in Observations	90.92%	1.92%	0.64%

Categorical Variables

- **Nationality**

The huge difference in sample size between groups indicates that, this variable may be a **bad choice for modelling**.

Generalized Linear Model (GLM)



Full Model

- AIC = 906.71
- Variables: Age, Education, Marital Status, Occupation, Sex, Hours Worked

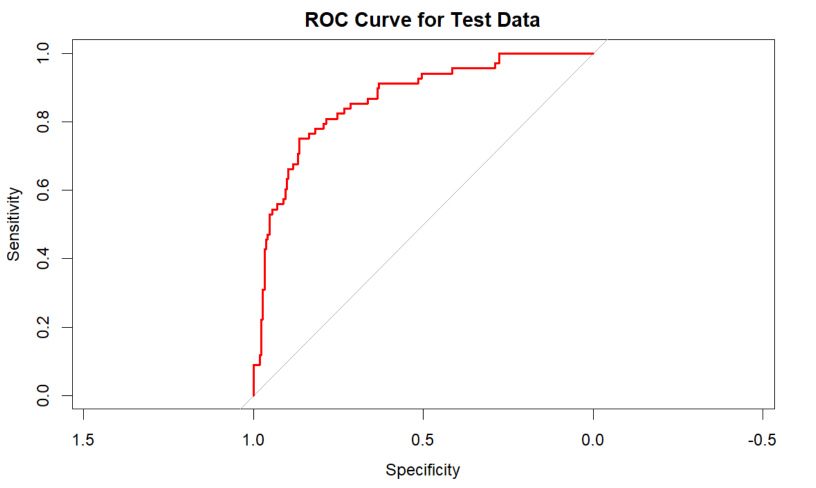
Stepwise Elimination

- stepwise backward elimination
- dropped Nationality, Sex

Final Model

- 877.85
- Retained key predictors for better fit and interpretability

GLM: Performance Evaluation



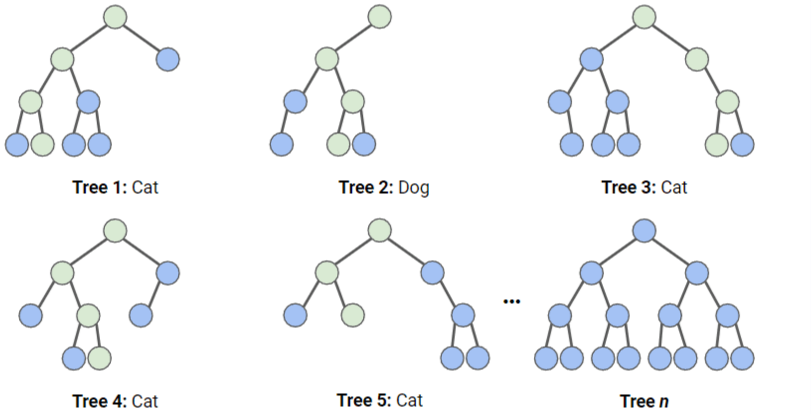
Accuracy 84.8% (95% CI: 80.0% - 88.7%)

Sensitivity 94.4% (strong for low-income)

Specificity 54.4% (weaker for high-income)

AUC 0.864 (strong discriminatory power)

Random Forest Analysis of Income Prediction



This presentation covers the implementation and performance of a **Random Forest** classifier for predicting whether an individual’s income exceeds $50,000 annually.

The accuracy is analyzed using standard metrics, visualizations, and cross-validation to enhance reliability.

The study also explores feature importance and the model’s decision-making process.

Random Forest: Performance

82.9%

Accuracy

91.2%

Sensitivity

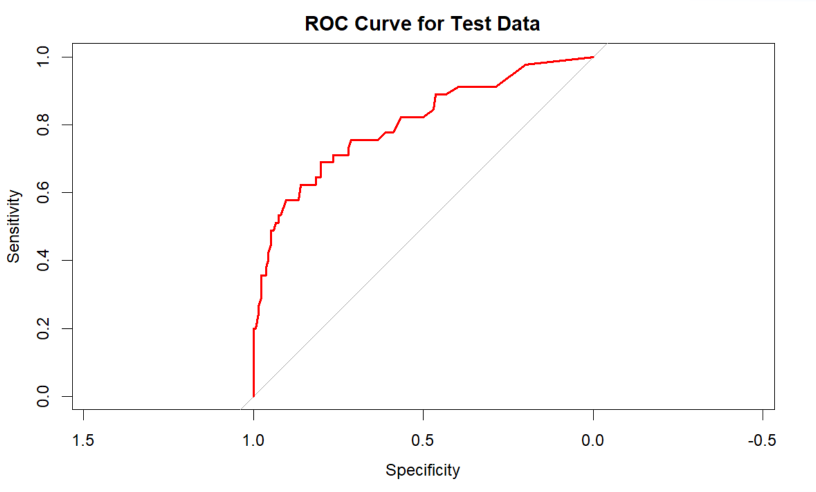
57.8%

Specificity

Confusion Matrix

Predictions vs Actual	Actual ≤$50k	Actual >$50k
Predicted ≤$50k	124 (TP)	19 (FP)
Predicted >$50k	12 (FN)	26 (TN)

Random Forest: ROC Curve Evaluation

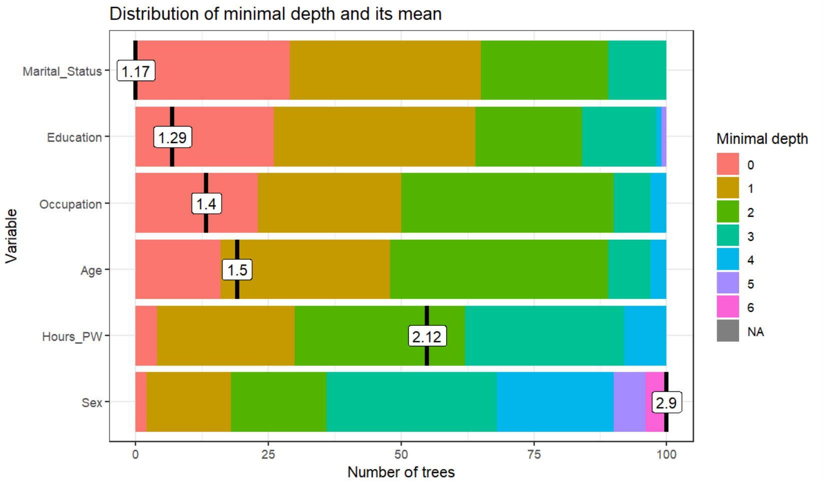


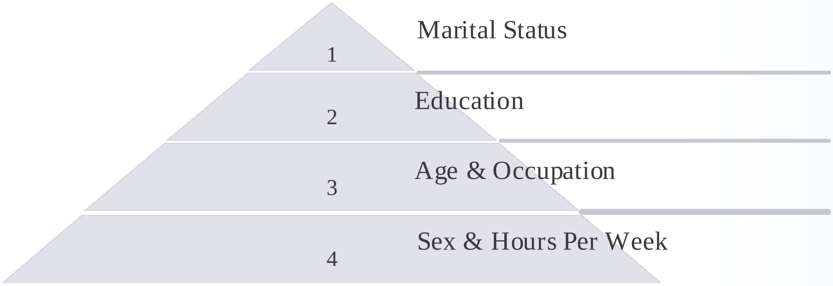
AUC Score 0.802

The Area Under Curve (AUC) score of 0.802 indicates strong discriminative ability.

AUC values range from 0.5 (random classification) to 1.0 (perfect classification), meaning our model performs significantly better than random guessing.

Random Forest: Minimal Depth Distribution





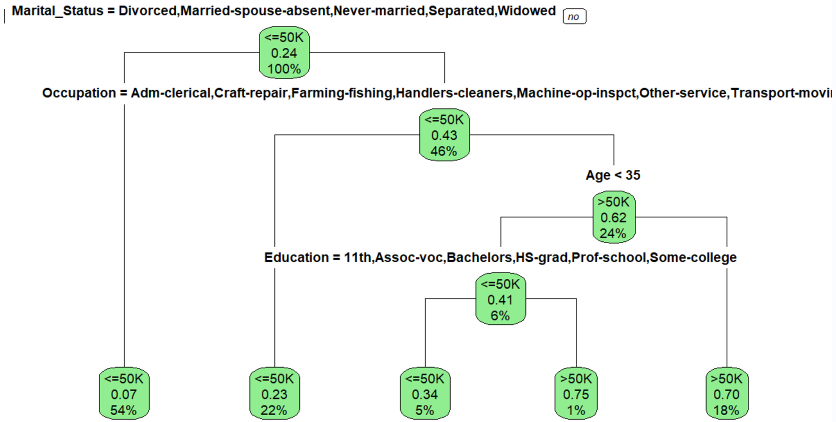
Marital status is the most influential predictor in predicting income levels.

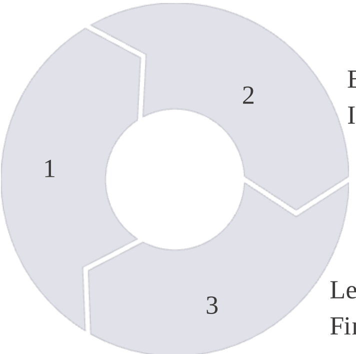
Education follows as a secondary predictor, indicating that higher education levels generally lead to higher income.

Age & **Occupation** are tertiary predictors, suggesting that both experience and job type play a role.

Sex & **Hours Per Week** have a lower impact but still contribute to the model.

Random Forest: Decision Tree Visualization


Marital_Status = Divorced,Married-spouse-absent,Never-married,Separated,Widowed
no
<=50K 0.24 100%
Occupation = Adm-clerical,Craft-repair,Farming-fishing,Handlers-cleaners,Machine-op-inspct,Other-service,Transport-movi
<=50K 0.43 46%
Age < 35
>50K 0.62 24%
Education = 11th,Assoc-voc,Bachelors,HS-grad,Prof-school,Some-college
<=50K 0.41 6%
<=50K 0.07 54%
<=50K 0.23 22%
<=50K 0.34 5%
>50K 0.75 1%
>50K 0.70 18%



Root Node
Initial split on key variables

Branch Nodes
Intermediate decision points

Leaf Nodes
Final income classifications

Marital Status is the primary determinant, where unmarried individuals are mostly classified as ≤$50k.

Occupation significantly impacts income, where certain jobs are more likely to fall into the ≤$50k category.

Age plays a role, where those over 35 have a higher chance of earning >$50k.

Education influences income, but its effect depends on other factors like age and occupation.

Random Forest: K-Fold Cross-Validation

1 Split Data

Divide dataset into 5 equal parts

2 Train Model

Train on 4 parts, test on 1

3 Rotate

Repeat 5 times with different test sets

4 Average Results

Calculate mean performance metrics

Key Results

Accuracy: 81.4%

Sensitivity (TPR): 86.2% (Good at predicting low-income individuals)

Specificity (TNR): 63.0% (Improved, but still weak for high-income classification)

Model Performance Comparison

	GLM	RF
Accuracy	84.8%	82.9%
Sensitivity	94.4%	91.2%
Specificity	54.4%	57.8%
AUC	0.864	0.802

GLM has higher accuracy, better for low-income prediction.

RF has higher specificity (better for high-income detection).

Both models are viable, but **GLM** performs better for this dataset.

Conclusion

Prediction Performance

The Generalised Linear Model and Random Forest achieves 85% and 83% accuracy respectively, based on 1994 U.S. Census data.

Critical Factors

Marital status, education level, occupation, and age are the strongest predictors of income.

Model Limitations

The model is better at identifying low-income (≤$50k) than high-income (>$50k).