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

Intruduction

Analysis Aim

Find key factors influencing individuals' income levels.

Data Source

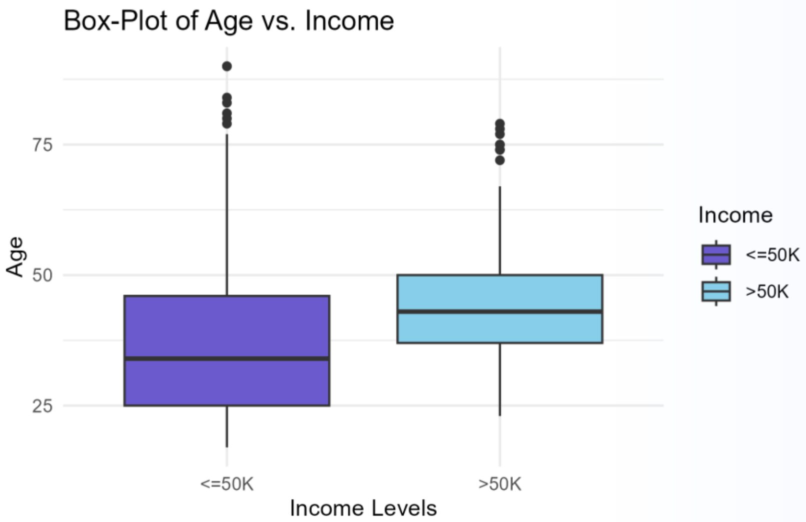
United States Census Bureau, 1994.

Analysis Approaches

Income are categorised into two levels: low-income ($\leq$ \$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

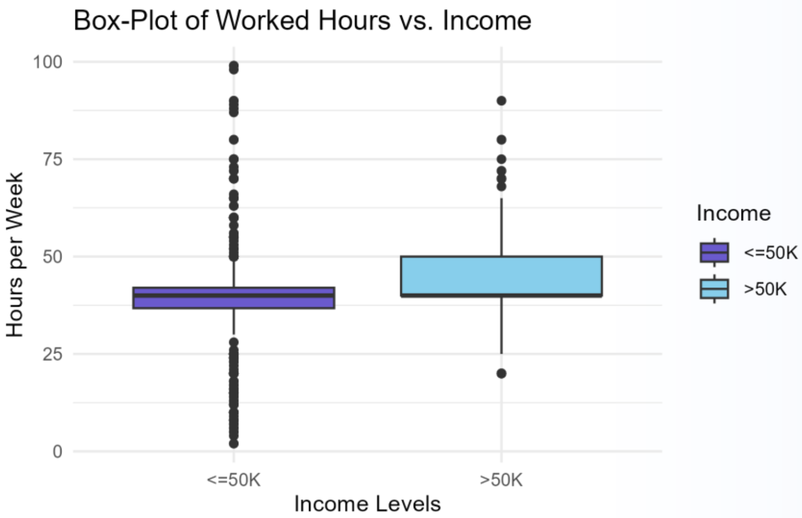


Numerical Variables

- **Age**

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

Exploratory Data Analysis



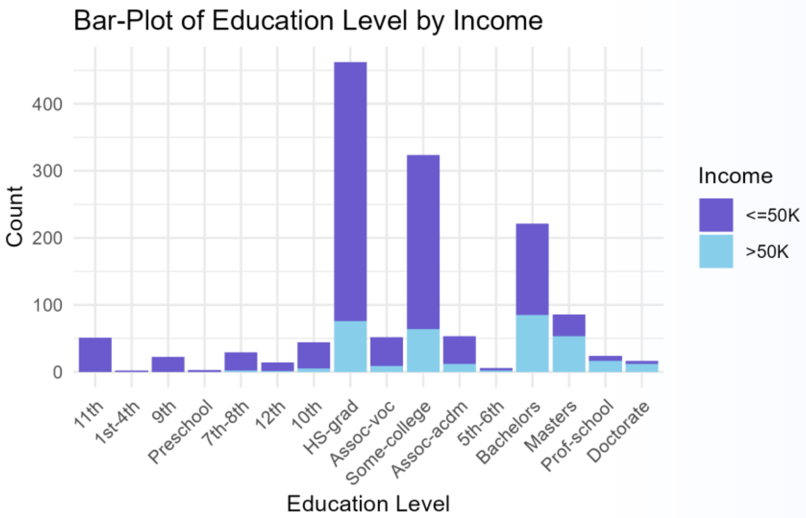
Numerical Variables

- **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



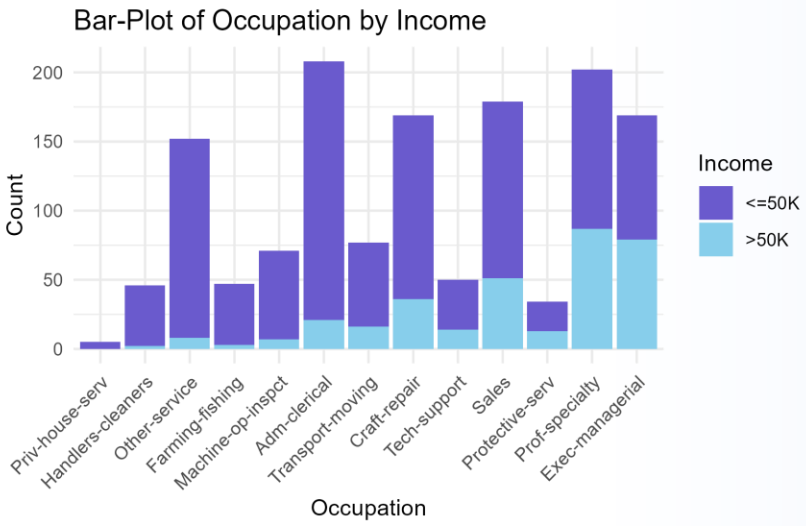
Categorical Variables

- **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

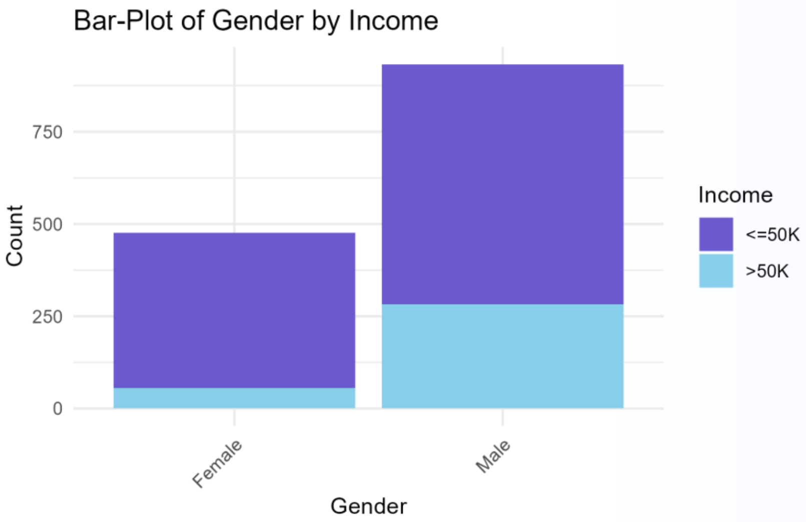


Categorical Variables

- **Occupation**

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

Exploratory Data Analysis

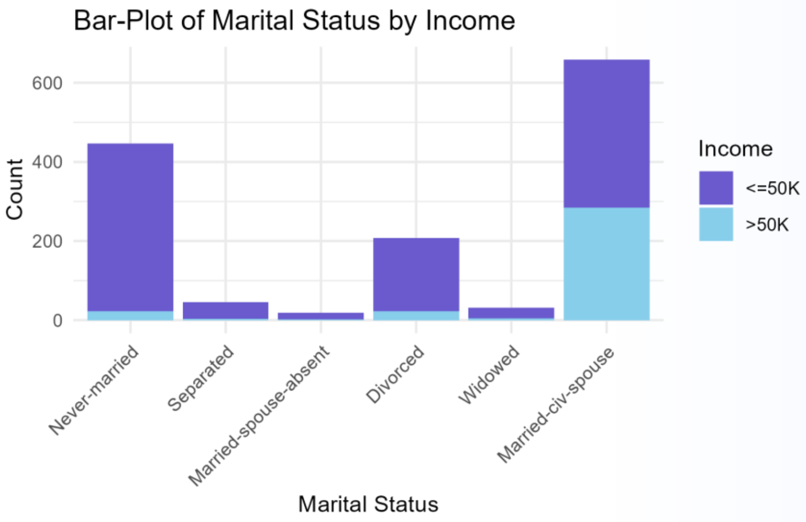


Categorical Variables

- **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

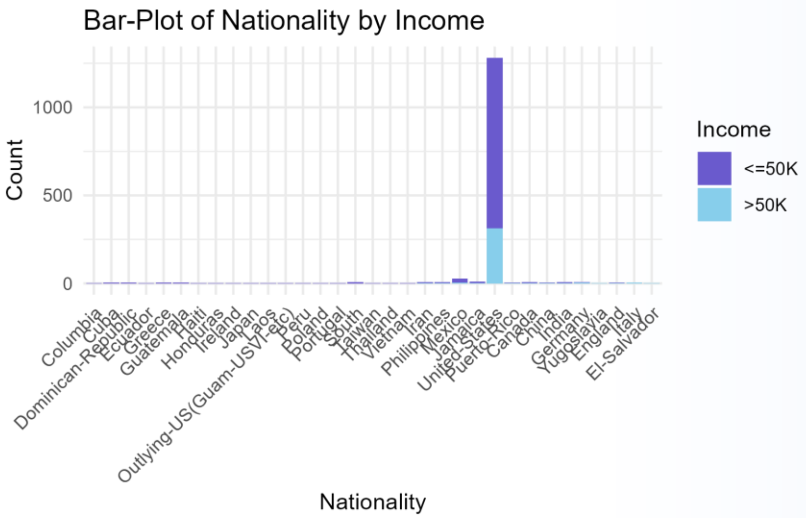


Categorical Variables

- **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**.

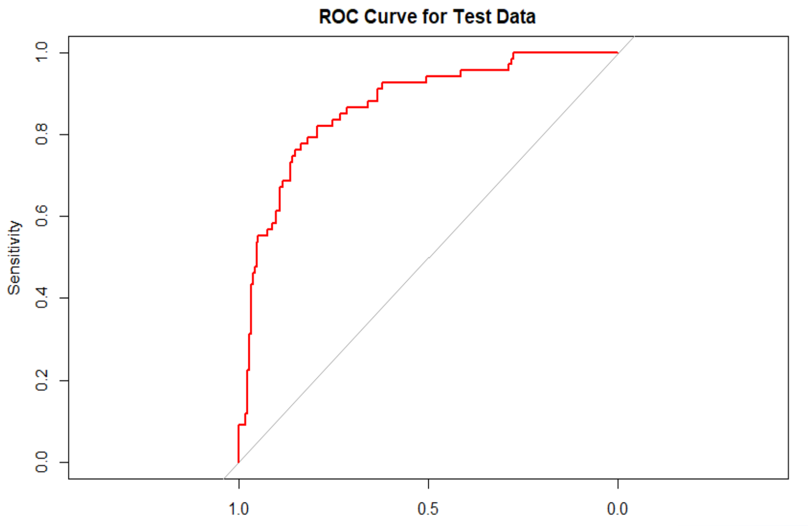
Generalized Linear Model (GLM)

Full Model
- AIC = 911.12
- Variables: Age, Education, Marital Status, Occupation, Sex, Hours Worked

Stepwise Elimination
- stepwise backward elimination
- dropped Nationality, Sex

Final Model
- 882.19
- Retained key predictors for better fit and interpretability

GLM: Performance Evaluation



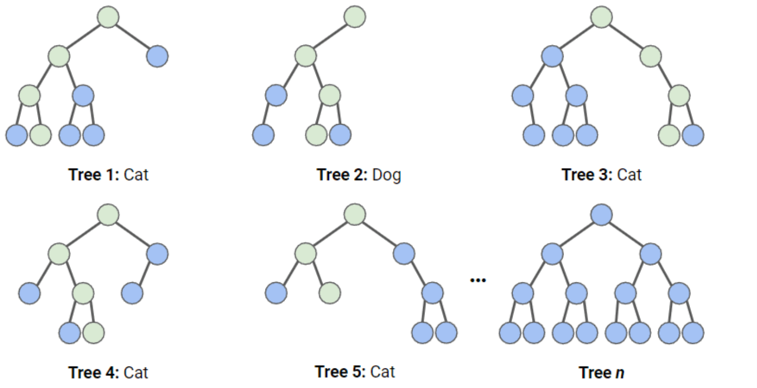
Accuracy: 85.4% (95% CI: 80.7% - 89.3%)

Sensitivity: 94.9% (strong for low-income)

Specificity: 55.2% (weaker for high-income)

AUC: 0.868 (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.

We analyze its accuracy using standard metrics, visualizations, and cross-validation to enhance reliability. The study also explores feature importance and the model's decision-making process.

Confusion Matrix and Performance Metrics

82.9%	91.2%	57.8%
Accuracy	Sensitivity	Specificity
Overall model prediction success rate	Correctly identifying incomes ≤$50K	Correctly identifying incomes >$50K

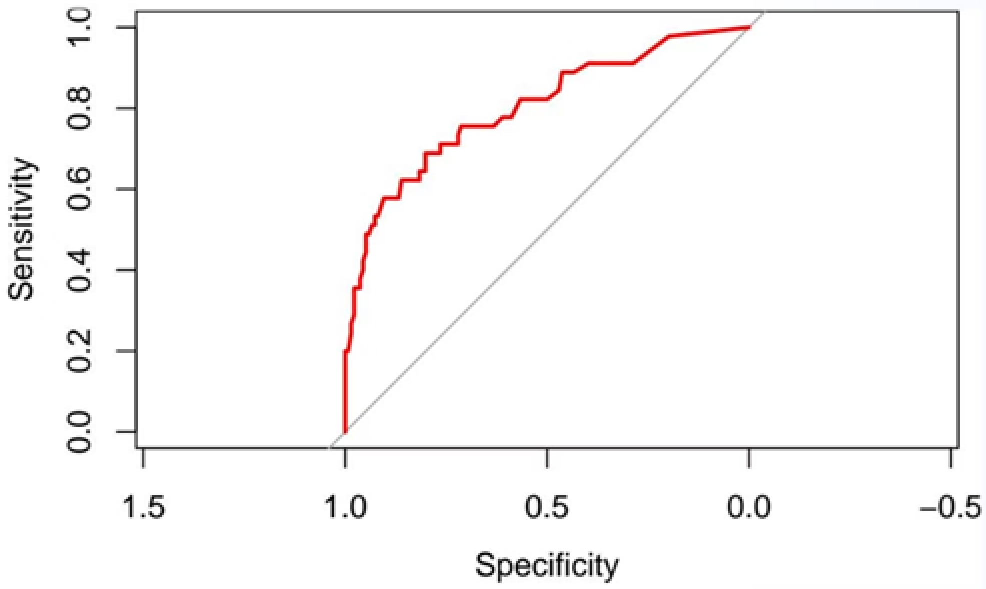
Our model shows strong overall performance. It excels at identifying lower incomes but struggles with predicting higher ones.

Confusion Matrix Analysis

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

True positives and negatives indicate correct predictions. False positives and negatives represent misclassifications.

ROC Curve Evaluation



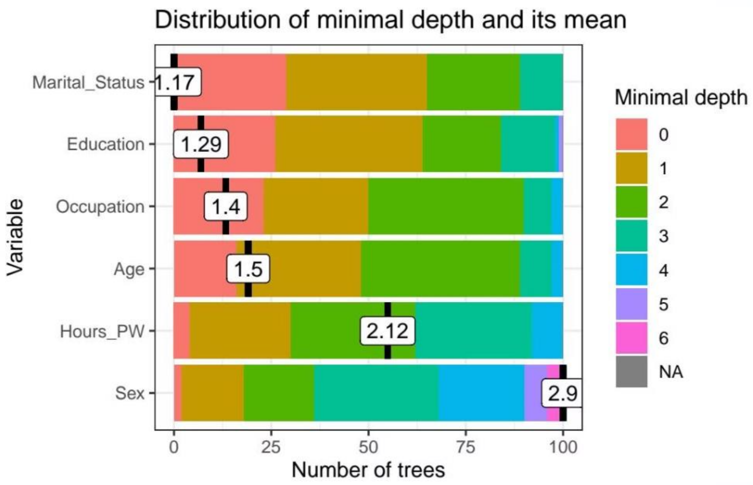
The ROC curve shows the trade-off between sensitivity and specificity.

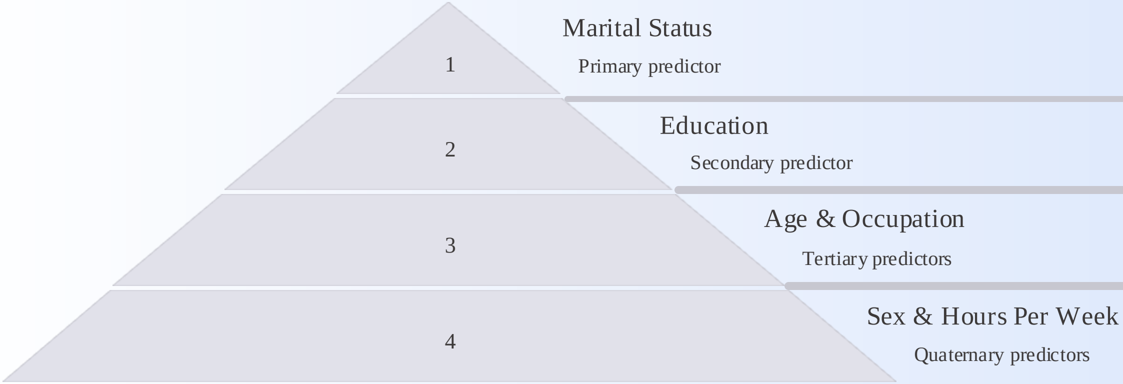
AUC Score: 0.802

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.

Variable Importance: Minimal Depth Distribution



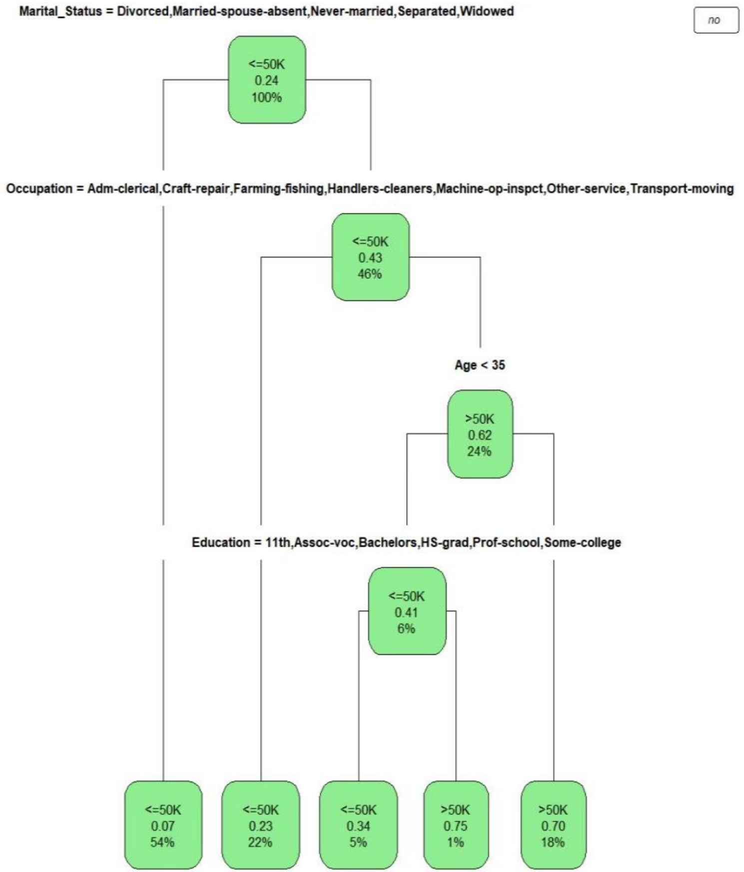


Lower minimal depth indicates greater variable importance. Marital status provides the most information to the model.

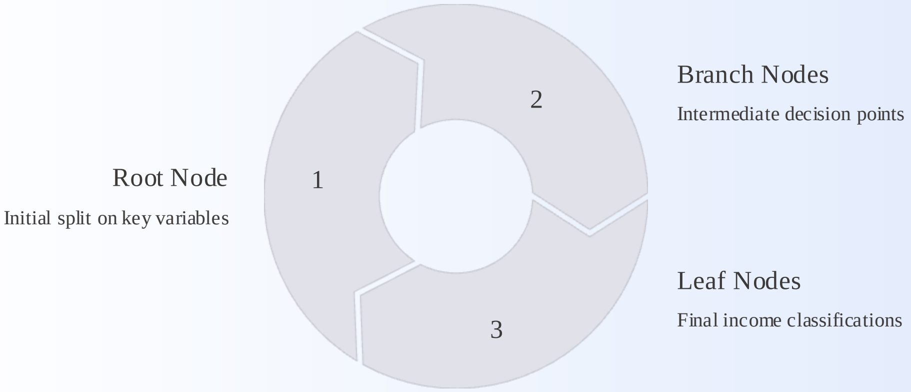
- Marital status is the most influential predictor in determining income levels.
- Education follows as a secondary predictor, indicating that higher education levels generally lead to higher income.
- Age & Occupation are tertiary predictors, suggesting that both experience and job type play a role.
- Sex & Hours Per Week have a lower impact but still contribute to the model.

Implication: Policies targeting marital status and education may have the most substantial impact on income inequality.

Decision Tree Visualization



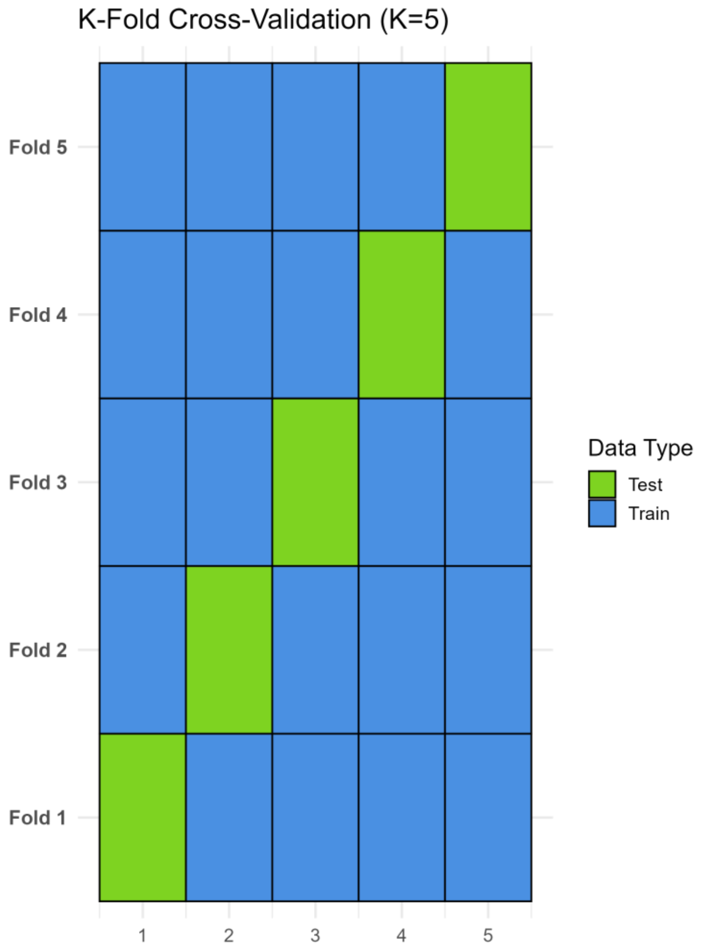
The tree shows how combinations of variables lead to different income predictions.



- Marital Status is the primary determinant—unmarried individuals are mostly classified as ≤$50K.
- Occupation significantly impacts income—certain jobs (e.g., administrative, farming, service-related) are more likely to fall into the ≤$50K category.
- Age plays a role—those over 35 have a higher chance of earning >$50K.
- Education influences income, but its effect depends on other factors like age and occupation.

Conclusion:Marital status, occupation, age, and education are key drivers of income.The tree structure shows how these variables interact to classify income levels.

K-Fold Cross-Validation Results (K=5)



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

Cross-validation is a technique used to improve model reliability by testing performance on different data subsets.

The dataset is split into 5 equal parts (K=5):

- The model is trained on 4 parts and tested on 1 part.
- The process is repeated **5 times**, each time with a different test set.
- The final model accuracy is obtained by **averaging the results**.

Key Results :

- Cross-validation accuracy: 81.4%.
- Sensitivity (TPR): 86.2% (Good at predicting low-income individuals)
- Specificity (TNR): 63.0% (Improved, but still weaker for high-income classification)

GLM: Performance Evaluation

Model	Accuracy	Sensitivity	Specificity	AUC
GLM	85.4%	94.9%	55.2%	0.868
RF	82.9%	86.7%	68.4%	0.802

GLM vs. RF: Accuracy (85.4% vs. 82.9%), AUC (0.868 vs. 0.802).

GLM: Higher accuracy, better for low-income prediction.

RF: Higher specificity (better for high-income detection).

Both models viable; GLM preferred for this data.

Conclusions

Prediction Performance

The random forest model achieves 81-83% accuracy in predicting income levels, based on 1994 U.S. Census data.

Critical Factors

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

Demographic Insights

Age, occupation, sex, and working hours influence income. However, the impact of working hours diminishes beyond a certain point.

Model Limitations

The model is better at identifying lower incomes (≤$50K) than higher incomes (>$50K).