Factors Influencing Whether an Individual Earns More Than $50K Per Year

Group 28: Chu Chu, Junhui Bai, Chaoyu Han, Yufei Luan, Zening Wang

Aims of Analysis

- **Understand income determinants** using 1994 U.S. Census data, focusing on whether individuals earn more than $50,000 per year.

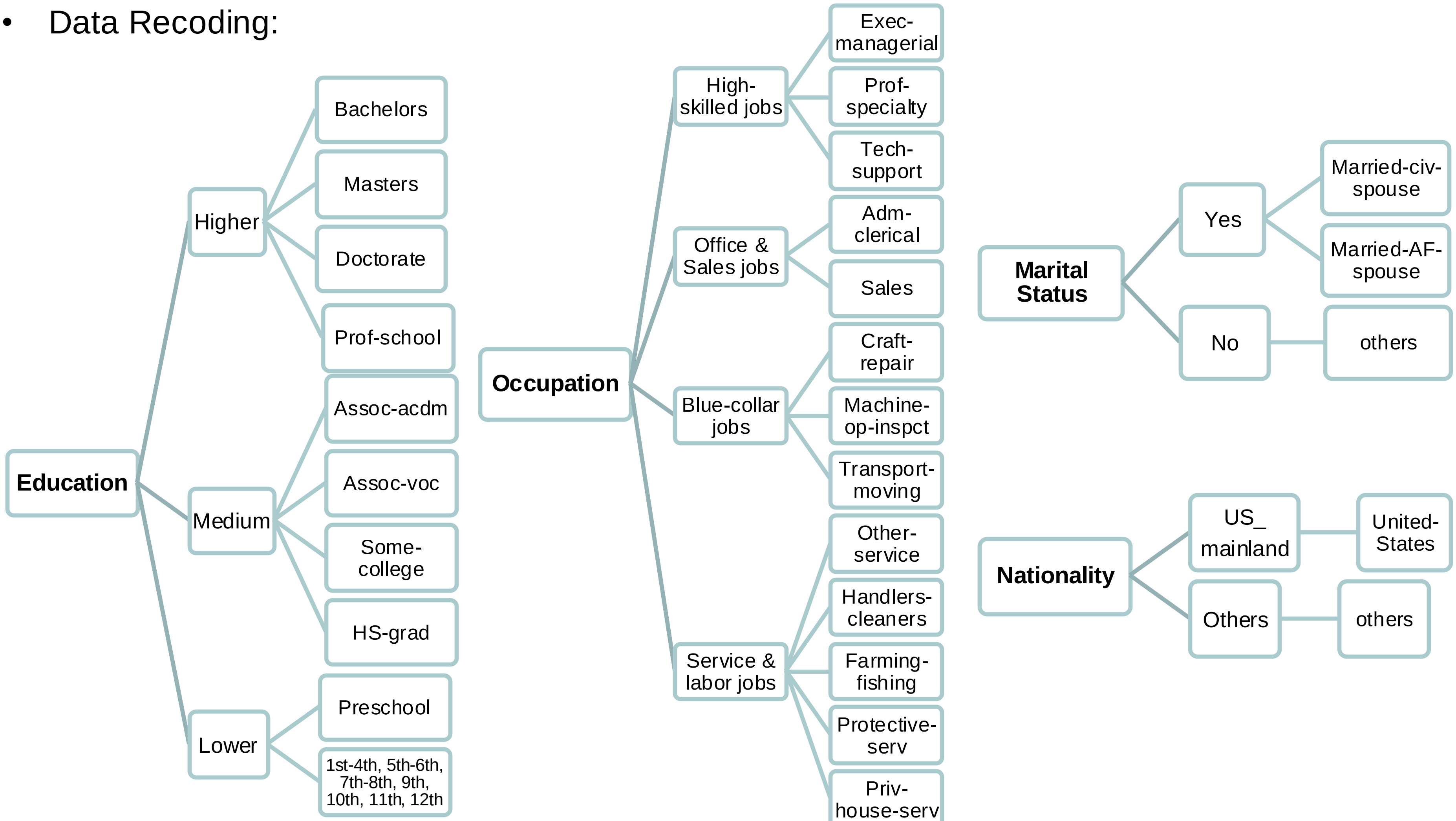
- **Compare models:** Generalized Linear Model (GLM) vs. machine learning models (e.g., Random Forest) in terms of predictive performance and interpretability.

- **Key research objectives:**

  1. Feature Identification – Find significant features that affect income levels.
  2. Income Classification – Predict if an individual earns >$50K (binary classification).
  3. Practical Implications – Interpret results to inform social & economic policy.

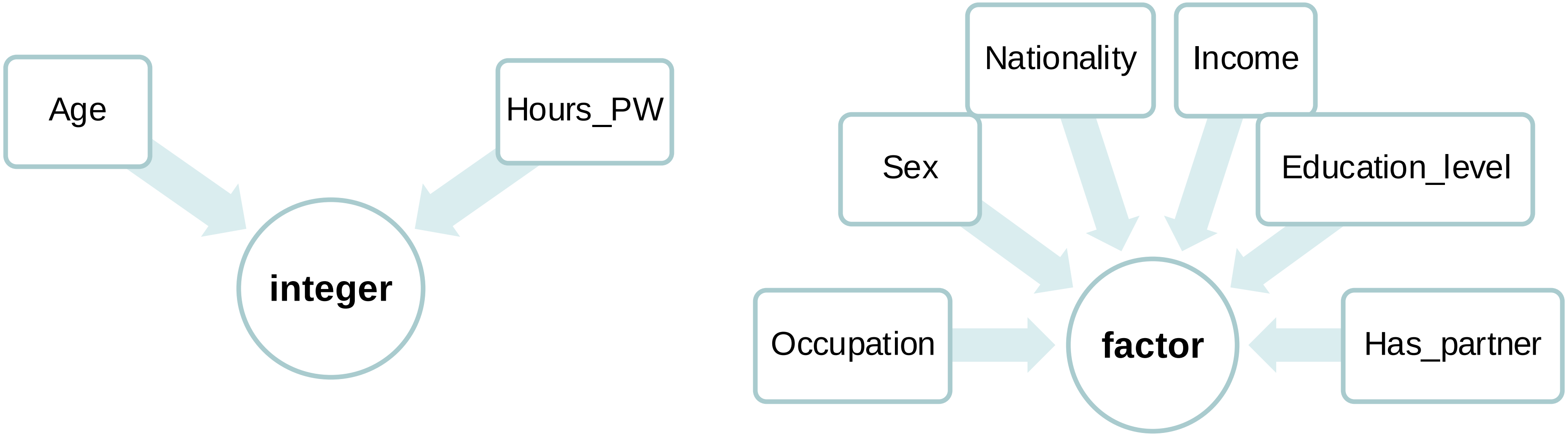
Exploratory Data Analysis

- Missing values were identified by treating '?,' as NA and subsequently removed using na.omit( ).
- Data Recoding:



Exploratory Data Analysis

- Data Type Checking and Conversion.

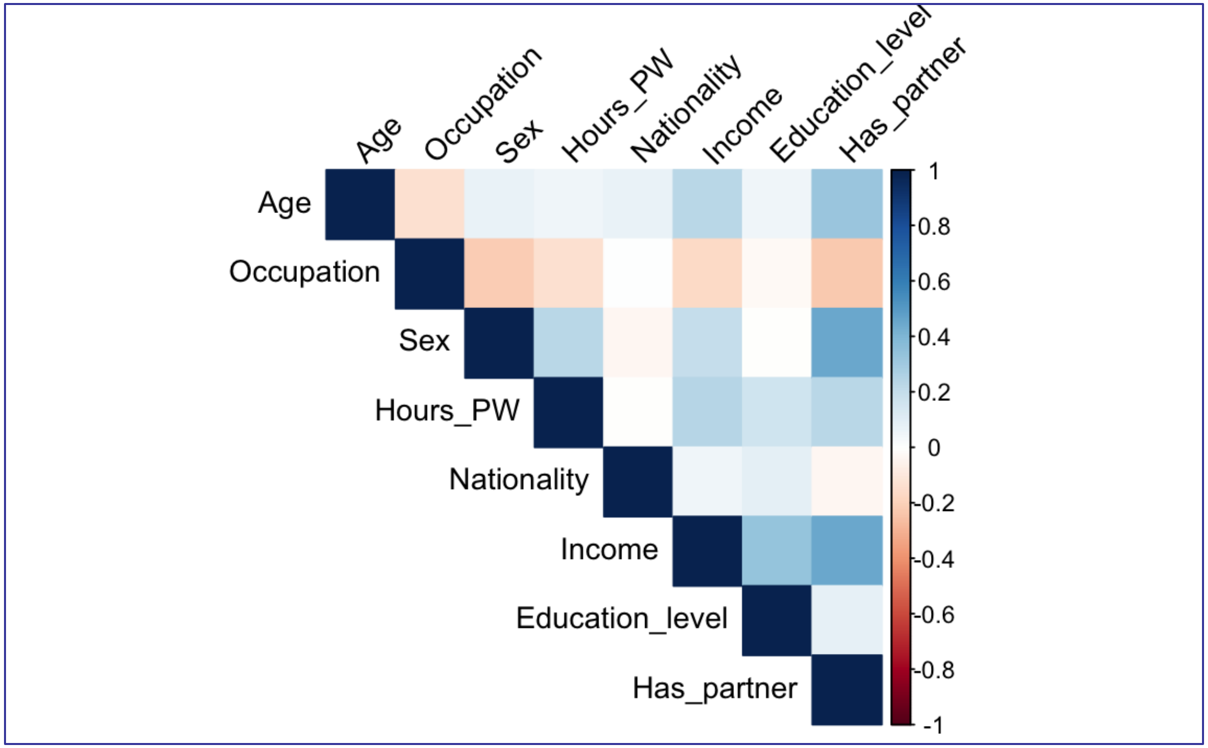


- Summary of Statistics by Variable.

Statistic	Age	Hours_PW
Min.	17.00	3.00
1st Qu.	28.00	40.00
Median	37.00	40.00
Mean	38.53	41.26
3rd Qu.	48.00	46.00
Max.	90.00	99.00

Exploratory Data Analysis

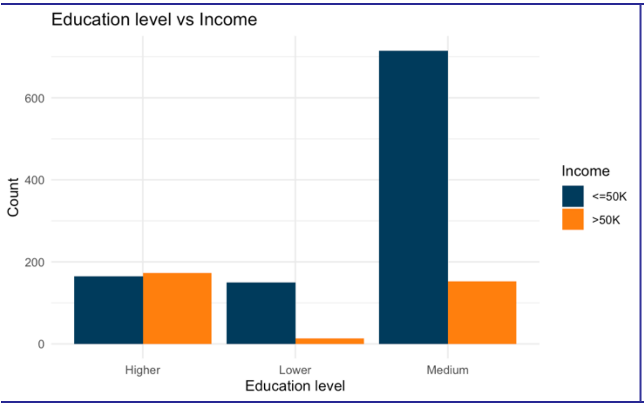
- Check for Multi-Collinearity issues.

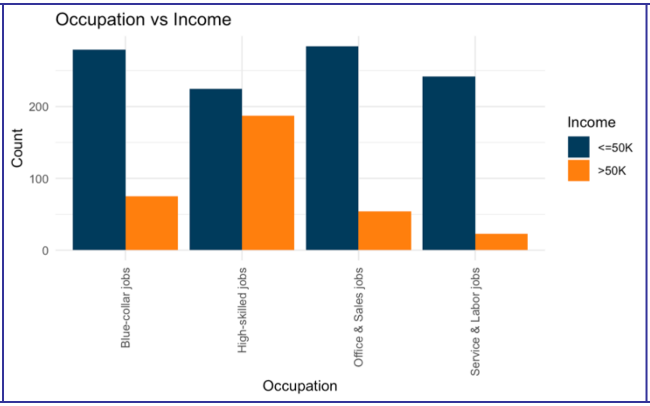


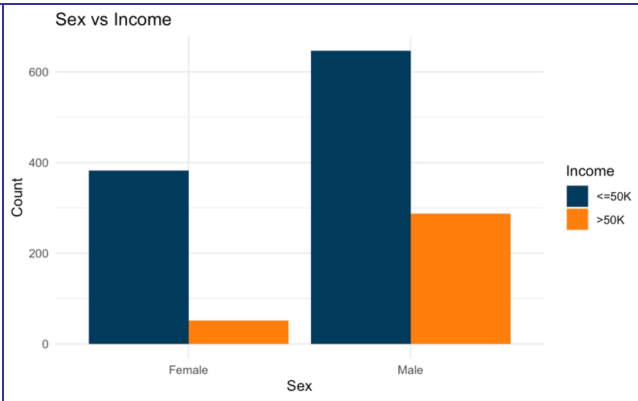
There is no serious multicollinearity problem.

Exploratory Data Analysis

- Data Visualization -- Categorical Variable.



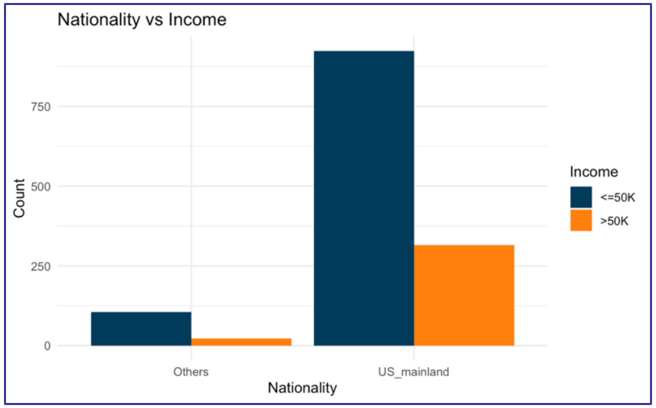


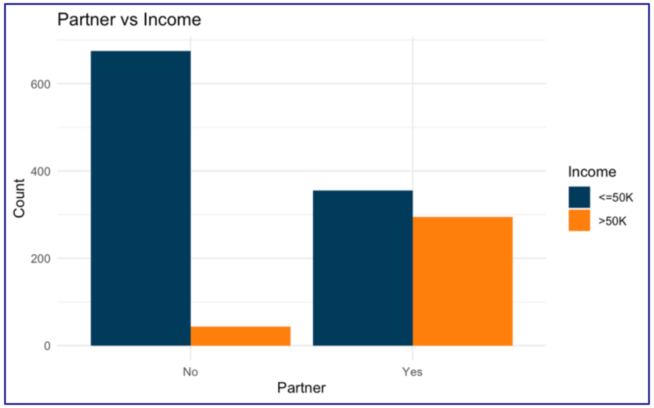


Education Level vs Income

Occupation vs Income

Sex vs Income





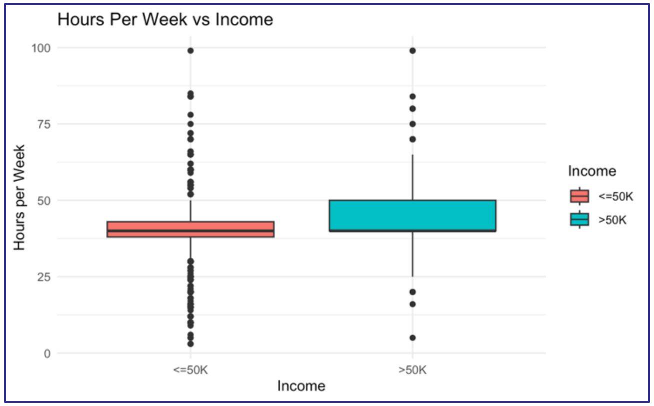
Nationality vs Income

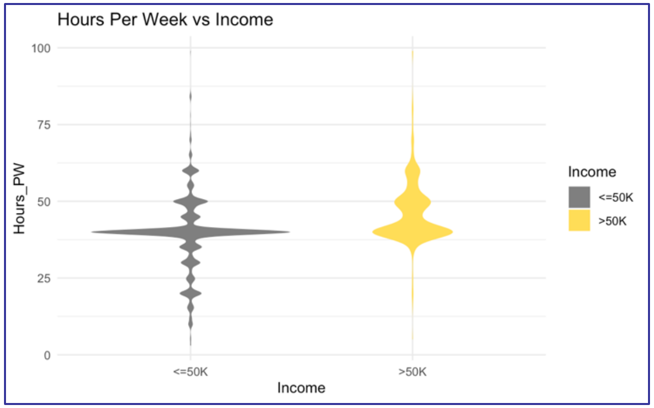
Partner vs Income

Exploratory Data Analysis

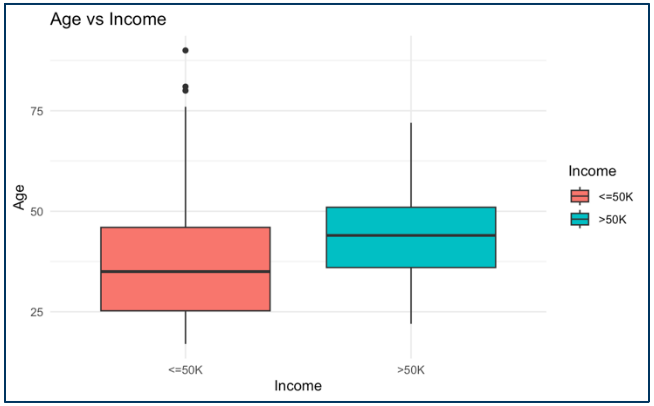
- Data Visualization -- Numerical Variable.

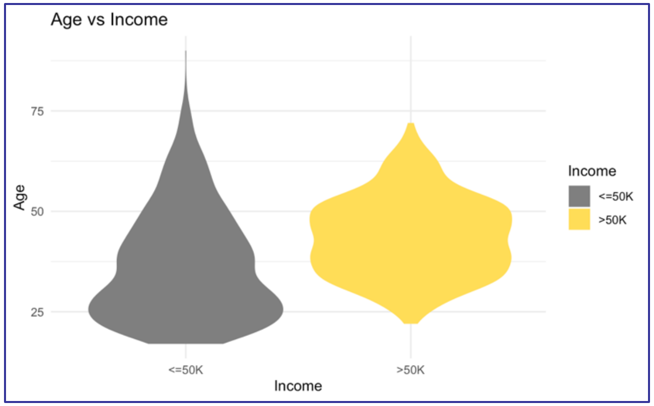
Hours Per Week vs Income





Age vs Income





Exploratory Data Analysis

- Check Sample Balance.
- Data Splitting: Training Set (80%) and Test Set (20%).

Dataset Income Category	Count	Proportion
≤50K	1030	75.24%
>50K	339	24.76%

Training data Income Category	Count	Proportion
≤50K	824	75.18%
>50K	272	24.82%

Test data Income Category	Count	Proportion
≤50K	206	75.46%
>50K	67	24.54%

Statistical Modelling & Results

- GLM

- Model Construction

Build Main Effects Model

- Created dummy variables for categorical predictors
- Built GLM with all main effect variables
- Stepwise AIC selection → Resulted in a simpler and interpretable base model

Add Interaction Terms

- Added :
  Education Level × Occupation
  Hours Worked × Occupation
- Chi-square test ($p = 0.0019$) ➔ Model fit improved

Finalize Optimized Model

- Removed insignificant interactions ($p > 0.05$)
- Kept : Hours Worked × Service & Labor Jobs
- Checked multicollinearity (VIF < 1.4)
  → Stable & strong final model

Statistical Modelling & Results

- GLM

- Base Model Summary

Variable	Coefficient	Effect Direction	p-value
Age	+0.0262	Positive	<0.001***
Hours_PW	+0.0384	Positive	<0.001***
Has_partner_Yes	+2.5359	Strong Positive	<0.001***
Education_level_Lower	-2.6628	Negative	<0.001***
Education_level_Medium	-1.3376	Negative	<0.001***
Nationality_US_mainland	+0.8734	Positive	0.012*
Occupation_High_skilled_jobs	+0.7780	Positive	0.0016***
Occupation_Service_Labor_jobs	-0.6674	Negative	0.0375*
Occupation_Office_Sales_jobs	-0.3869	Not Significant	0.1564

- Most variables are statistically significant ($p < 0.05$), but **OccupationOffice_Sales_jobs** is not.
- Model performance (AIC = 793.01; Residual Deviance = 773.01) indicates room for improvement.

→ Optimization needed to improve model fit and eliminate non-significant predictors.

Statistical Modelling & Results

- GLM

- Final Model Summary

Variable	Coefficient	Effect Direction	p-value
Age	+0.0265	Positive	<0.001***
Hours_PW	+0.0429	Positive	<0.001***
Has_partner_Yes	+2.5551	Strong Positive	<0.001***
Education_level_Lower	-2.5287	Negative	<0.001***
Education_level_Medium	-1.2623	Negative	<0.001***
Nationality_US_mainland	+0.8843	Positive	0.011*
Occupation_High_skilled_jobs	+0.9145	Positive	<0.001***
Hours_PW × Occupation_Service_Labor_jobs	+0.0186	Negative Interaction	0.005**

Coefficients indicate the effect on the log-odds of earning >$50K.

All shown variables are statistically significant ($p < 0.05$).

Statistical Modelling & Results

- GLM

• Final Model Equation

$$
\begin{aligned}
\log\left(\frac{P(\text{Income}=1)}{1-P(\text{Income}=1)}\right) = &-5.786 + 0.027 \cdot \text{Age} + 0.043 \cdot \text{Hours\_PW} \\
&+ 0.914 \cdot \text{Occupation}_{\text{High\_skilled\_jobs}} \\
&+ 0.884 \cdot \text{Nationality}_{\text{US\_mainland}} \\
&- 2.529 \cdot \text{Education}_{\text{Lower}} \\
&- 1.262 \cdot \text{Education}_{\text{Medium}} \\
&+ 2.555 \cdot \text{Has\_partner}_{\text{Yes}} \\
&- 0.019 \cdot \left(\text{Hours\_PW} \times \text{Occupation}_{\text{Service\_Labor\_jobs}}\right)
\end{aligned}
$$

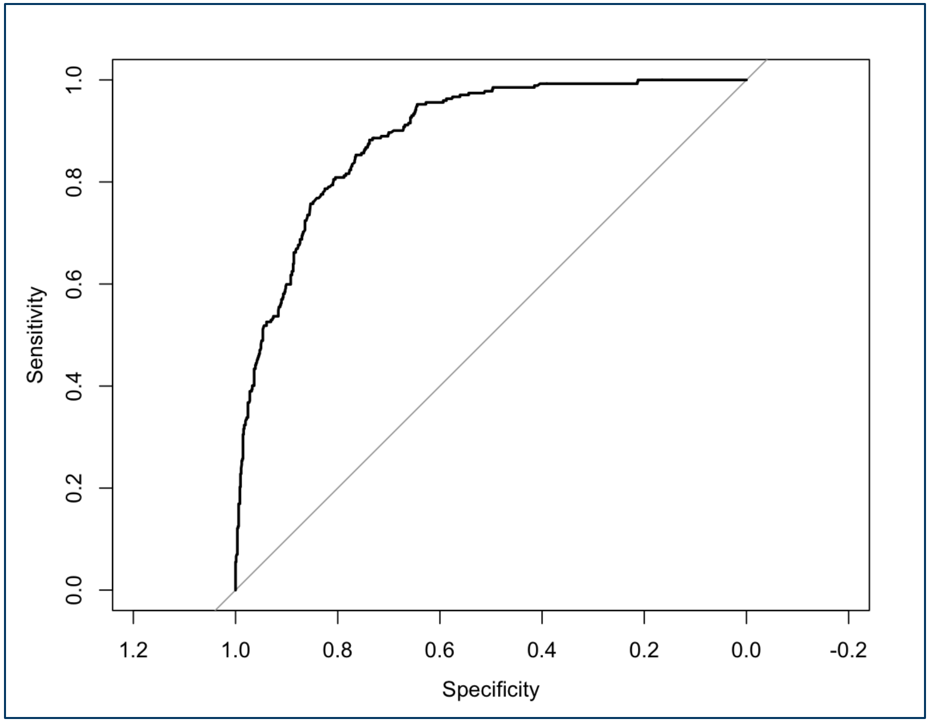
This final model includes the most influential main effects and one significant interaction term.

All coefficients are statistically significant ($p < 0.05$), and the model shows strong predictive performance.

Statistical Modelling & Results

- GLM

• Model Performance



- **AUC = 0.8892**

  → Indicates strong classification performance

- **AIC = 787.58**

  → Improved from base model (793.01), indicating better model fit

- **Residual Deviance = 769.58**

  → Suggests improved goodness of fit

- **All VIF values < 1.4**

  → Confirms absence of multicollinearity

Statistical Modelling & Results

- GLM

- Model Summary

  - **Final model includes 8 key predictors**, all statistically significant ($p < 0.05$).
  - **Has Partner** and **Education Level** show strongest effects.
  - Significant interaction:

    → **Longer working hours** have *reduced impact* in **labor-intensive jobs**.
  - Model is **interpretable, stable**, and shows **strong classification performance** .
  - Results highlight key drivers of income inequality.

Statistical Modelling & Results

- Random Forest

- Model Construction

Build Random Forest Model

- Trained with `ntree=500` and `mtry=√p`, where *p* is the number of predictors
- Enabled variable importance and proximity calculations

Extract Model Info

- Printed model summary
- Extracted one decision tree (`getTree()`)
- Collected Out-of-Bag (OOB) error estimates

Visualize & Evaluate

- Plotted OOB Error Curve
- Created Variable Importance Plot
- Prepared model for prediction on test set

Statistical Modelling & Results

- Random Forest

- Model Performance

- Key Metrics (Test Set Evaluation):

Accuracy: 80.59%

AUC (Area Under Curve): 0.8632

OOB Error Estimate: 17.06%

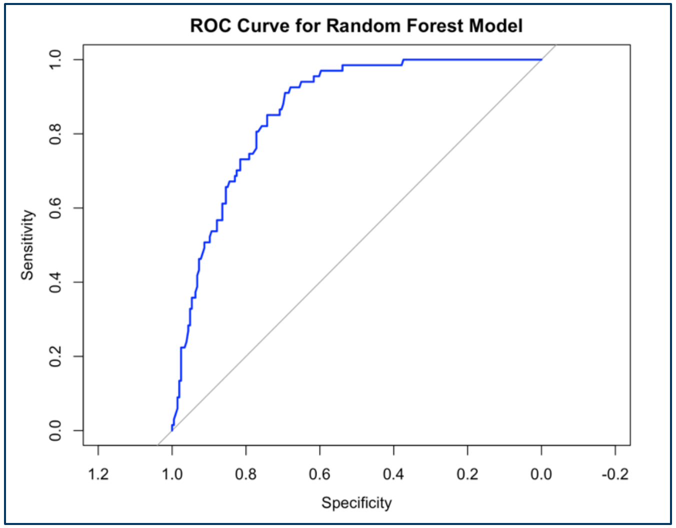
- **Class-specific Performance:**

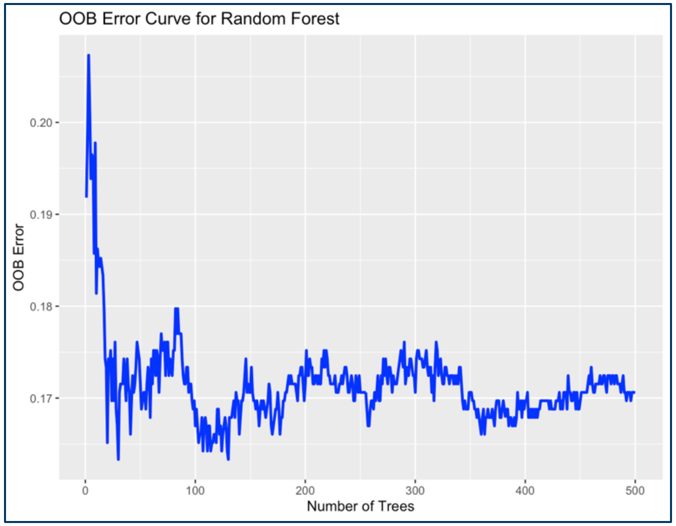
Sensitivity (≤$50K): 89.32%

Specificity (>$50K): 53.73%

Balanced Accuracy: 71.53%

Class Error (>50K): 40.89%

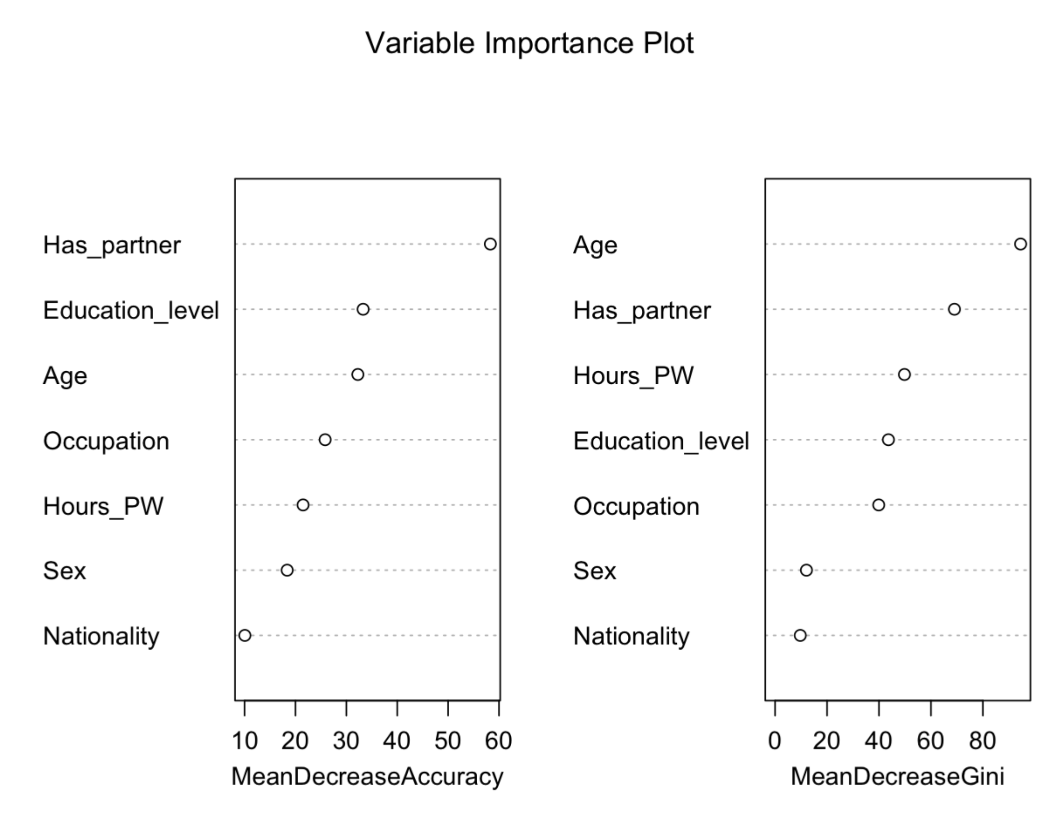




Statistical Modelling & Results

- Random Forest

• Variable Importance



Variable	MDA
Has_partner	58.31
Education_level	33.29
Age	32.26
Occupation	25.83
Hours_PW	21.49

(MDA: Mean Decrease Accuracy)

The **top five** predictors showed the **highest contributions** to classification accuracy based on Mean Decrease Accuracy.

Statistical Modelling & Results

- Model Comparison

Metric	GLM Model	Random Forest Model
AUC	0.8892	0.8632
Interpretability	Strong	Low (Black Box)
Interaction Modeling	Explicitly Modeled	Implicitly Included
Model Suitability	Interpretation & Inference	High-Dimensional, Nonlinear
Class Balance Sensitivity	Adjustable via Weighting	Relatively High

- Both models show strong AUC performance (GLM: 0.8892 > RF: 0.8632).
- GLM outperforms in interpretability and policy relevance.
- GLM explicitly models interactions and shows clearer effect directions.
- RF is better suited for complex, high-dimensional data but lacks transparency.

Overall, GLM is more appropriate for explaining income inequality in this study.

University of Glasgow

Conclusions

What determines whether an individual earns over $50K/year?

- **Education Level**
- **Working Hours (Hours_PW)**
- **Marital Status (Has Partner)**
- **Age**
- **Occupation Type**
- **Nationality**

Among these, education and marital status showed the strongest effects.

Significant interaction: Longer working hours yield less income benefit in labor-intensive jobs.

Policy Implications

- **Expand Access to Education** → Education is key to income mobility.
- **Improve Labor Policies** → Especially in physical jobs where long hours bring limited returns.
- **Support Social Stability** → Marital status shows strong financial relevance.
- **Address Regional Inequality** → Income advantages differ by nationality.

Future Work

1. **Address Class Imbalance**

   → Apply resampling techniques (e.g. SMOTE, undersampling) to improve prediction for high-income group.

2. **Improve Minority Class Prediction**

   → Explore model adjustments that reduce bias towards the majority class (≤$50K).

3. **Extend Data Coverage**

   → Consider gathering more balanced or diverse datasets for future analysis.

University of Glasgow

Thank You !