Determinants of High Income

A Generalized Linear Model Analysis of 1994 US Census Socioeconomic Factors

[Mohan Wang] [Huiyang Liao] [Qiwei Wang]

Group: 27

CONTENT

01

The aims of the analysis

Background

This study is based on the 1994 U.S. Census database (US Census 1994), which contains information on individuals' income levels and various socioeconomic factors. We use this data to analyze which factors influence an individual's income level, specifically whether their annual income exceeds $50,000.

Question of interest

The government aims to determine which socioeconomic factors are the best predictors of whether an individual earns more than $50,000 per year. To achieve this, we need to:

- Examine which variables (such as age, education, working hours, etc.) are significantly associated with income levels.

- Use a Generalized Linear Model (GLM) to analyze the impact of different variables on income.

- Summarize the findings through Exploratory Data Analysis (EDA) and modeling results, and present conclusions using visualizations.

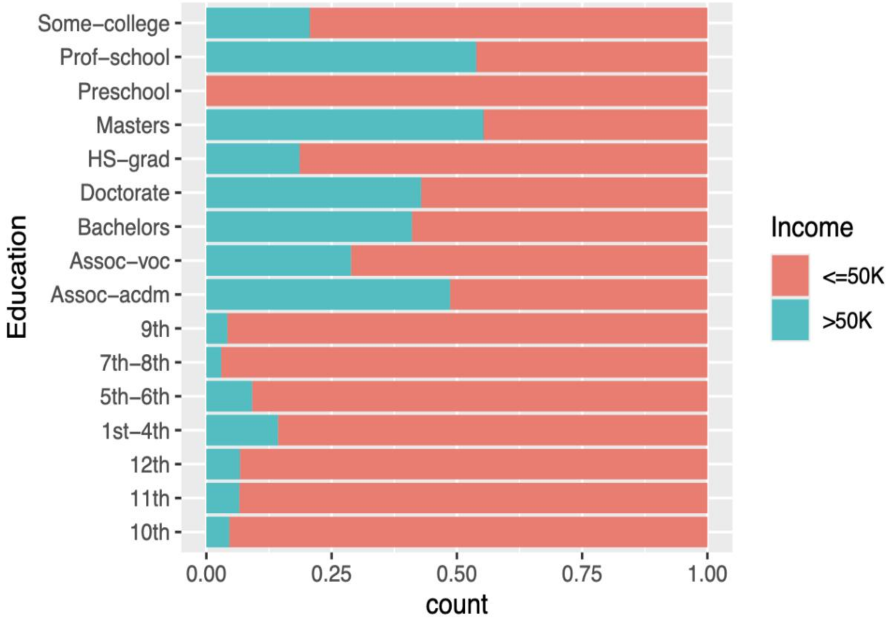
02

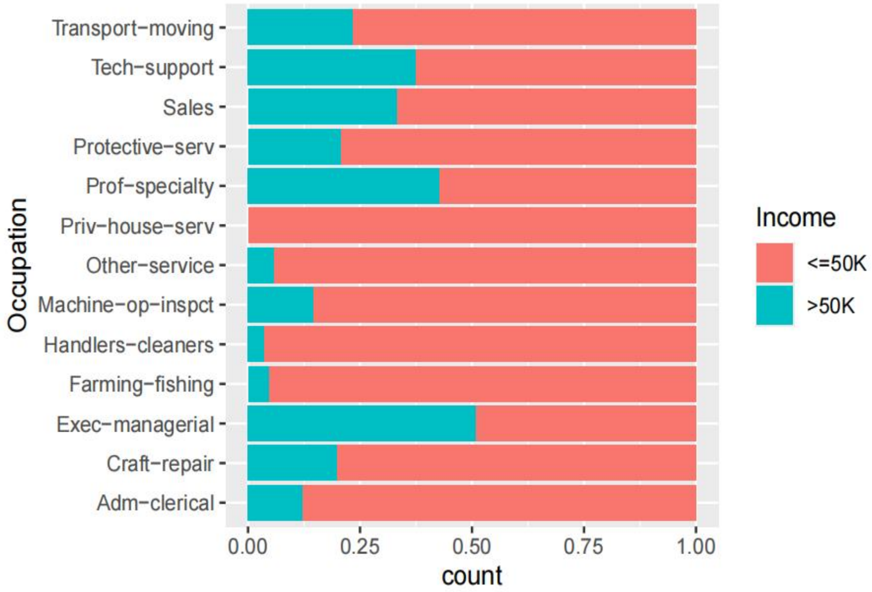
Exploratory data analysis

Data Cleaning

- Delete extra commas.
- Convert ‘?’ to ‘NA’ .
- Remove the NA value from the data.
- Determine the corresponding data type.

Original data analysis





By visualising all the variables first, we can see that there are some variables with perfectly linear relationships among the classes (for example, Preschool in Education and Priv-house_serv in Occupation).

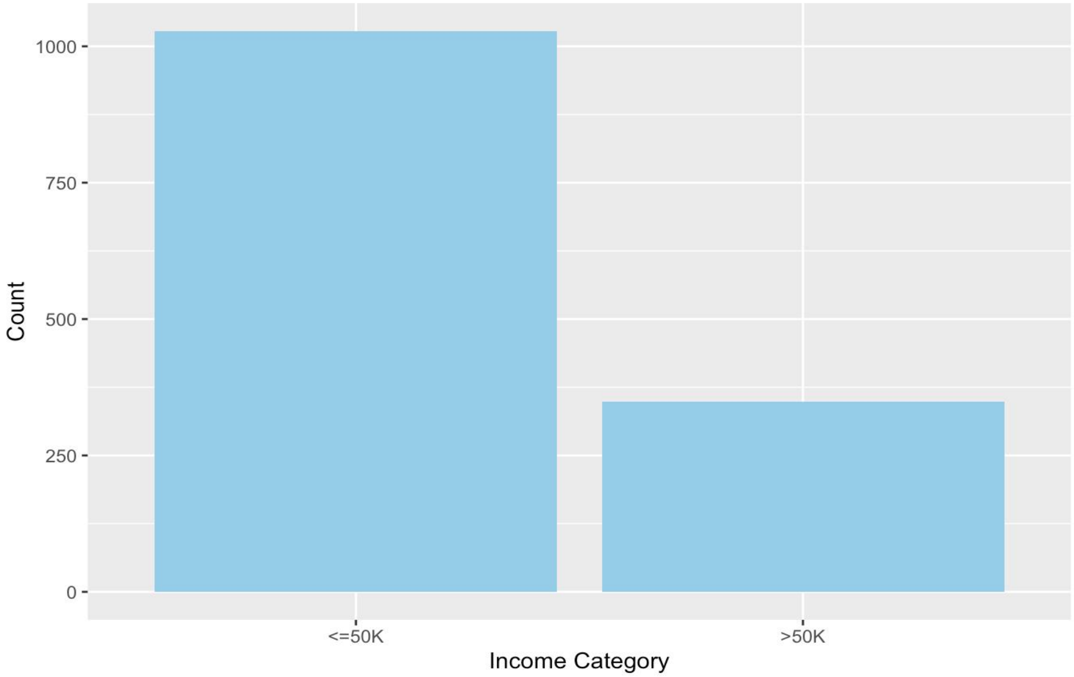
Advanced data process

```
$Age
 [1] 37 49 20 64 32 58 42 25 30 66 59 44 46 17 47 26 33 48
52 40 50 34
[23] 43 22 35 62 39 27 38 41 23 56 24 21 18 45 51 29 19 28
65 57 67 36
[45] 54 60 31 61 55 71 53 73 63 70 68 90 77 75 72 74 69 78
80 81

$Education
[1] "Bachelors"          "College Education"  "High School"
[4] "Basic Education"    "Advanced Education"

$Marital_Status
[1] "Married"
[2] "Divorced/Spouse Absent/Separated/Widowed"
[3] "Never-married"
```

```
$Occupation
[1] "White-Collar"    "Blue-Collar"     "Service"
[4] "Sales"           "Adm-clerical"    "Farming-fishing"

$Sex
[1] "Male"   "Female"

$Hours_PW
 [1] 80 45 55 42 40 50 14 65 35 58 30  9 15 12 70 24  8  6
52 20 16 10
[23] 60 36 48 38 18 43 39 90 72 32 49 56 84 44 54 25 37 99
46 47 53 28
[45] 17  5 75 68 22 21  3 13 27 11 34 33 98 85

$Nationality
[1] "United-States" "Others"

$Income
[1] 1 0
```

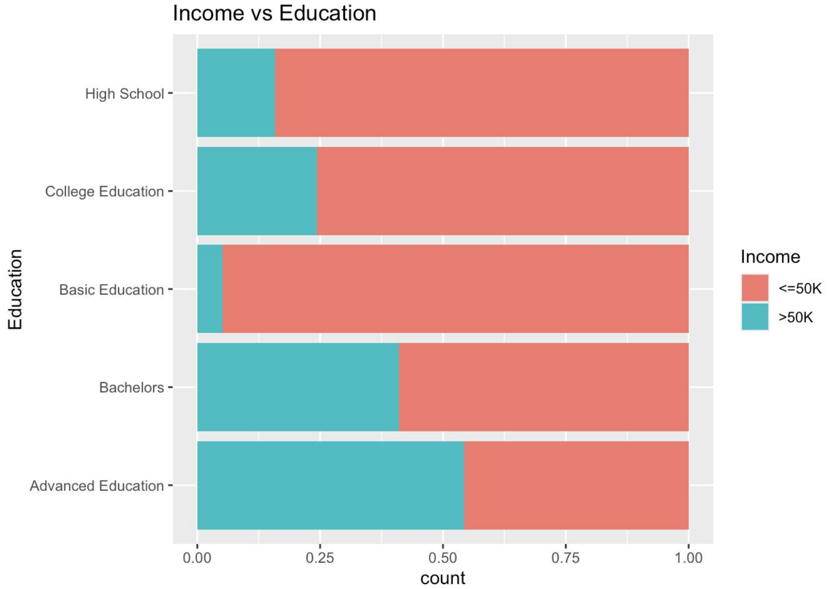
In order to avoid affecting the model fitting results later, we sorted and merged some date. For instance, we merged the original education levels into five categories: Basic Education, High School, College Education, Bachelors, and Advanced Education

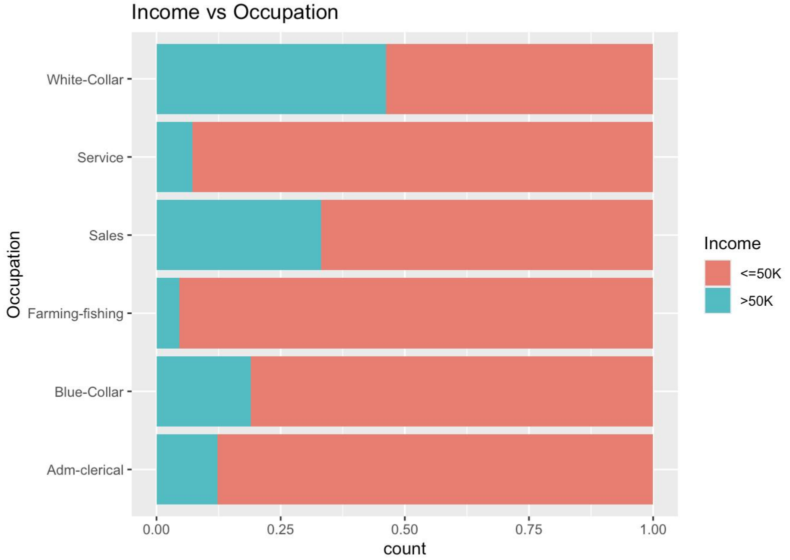
Visualise the income distribution



The chart shows that the majority of individuals have an income of $50K or less.This indicates a class **imbalance** in the target variable, which could influence the performance of classification models.

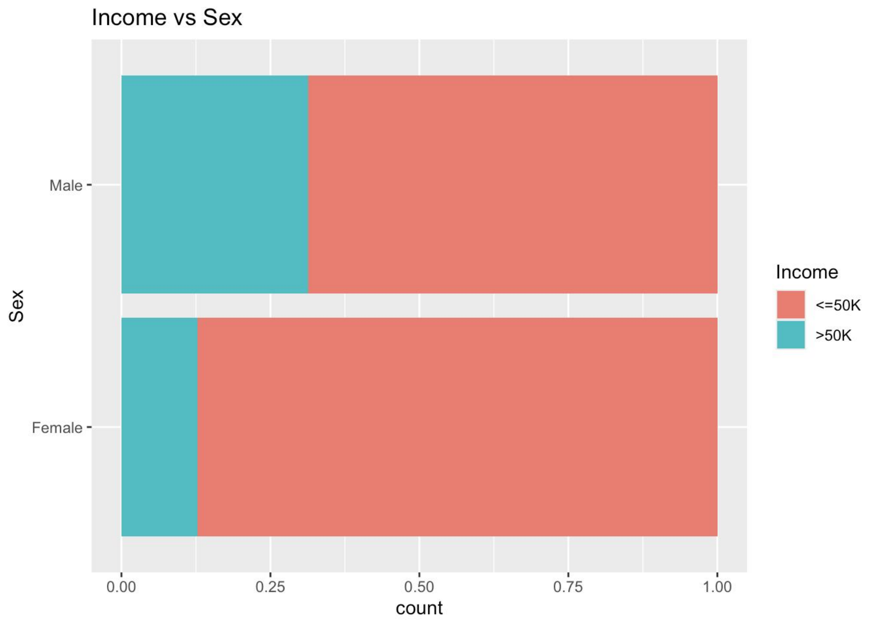
Exploring Categorical Variables' Relationship with Income

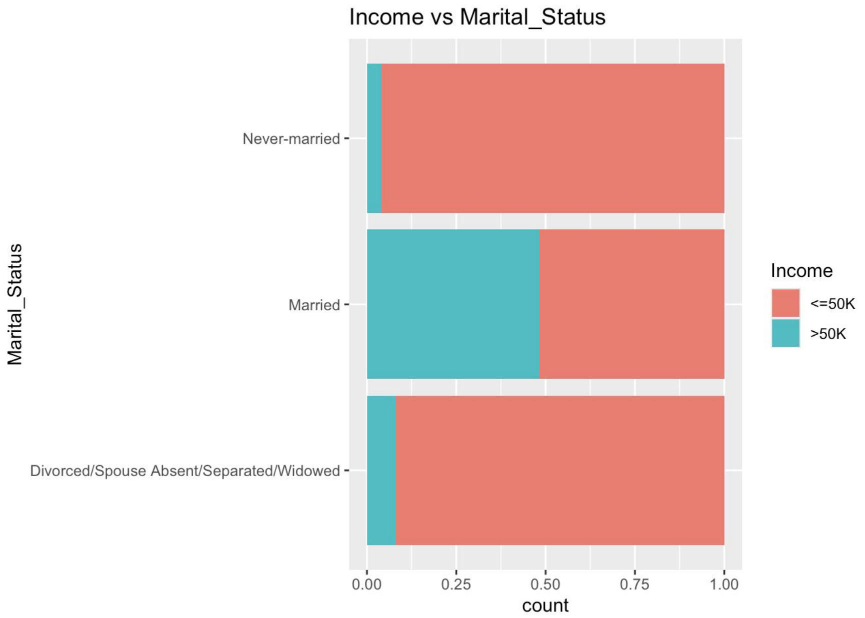




Individuals with higher education levels (especially bachelor's degrees and above) are more likely to earn over 50K; white-collar and sales jobs have a higher proportion of high-income earners than blue-collar and service jobs.

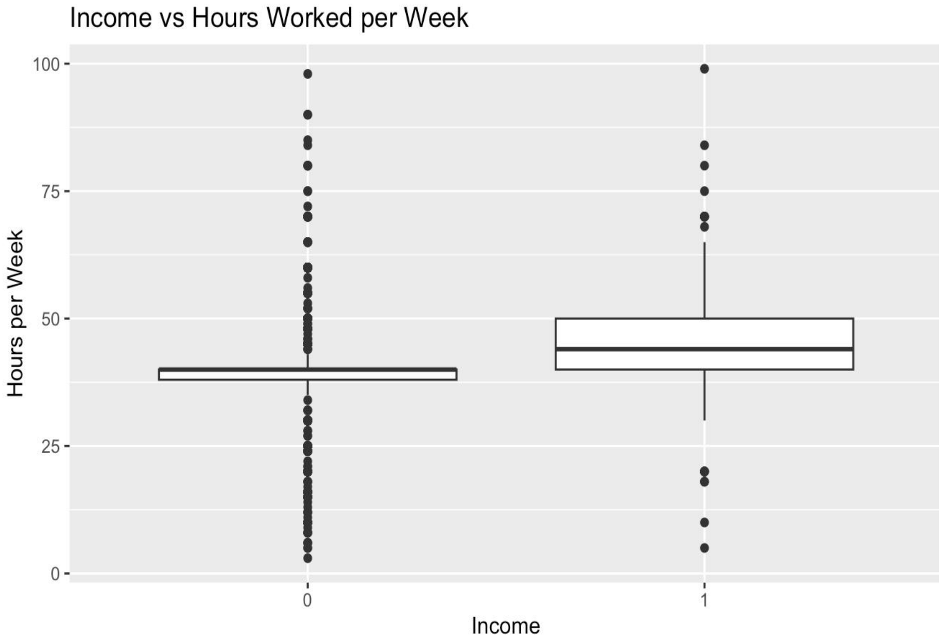
Exploring Categorical Variables' Relationship with Income

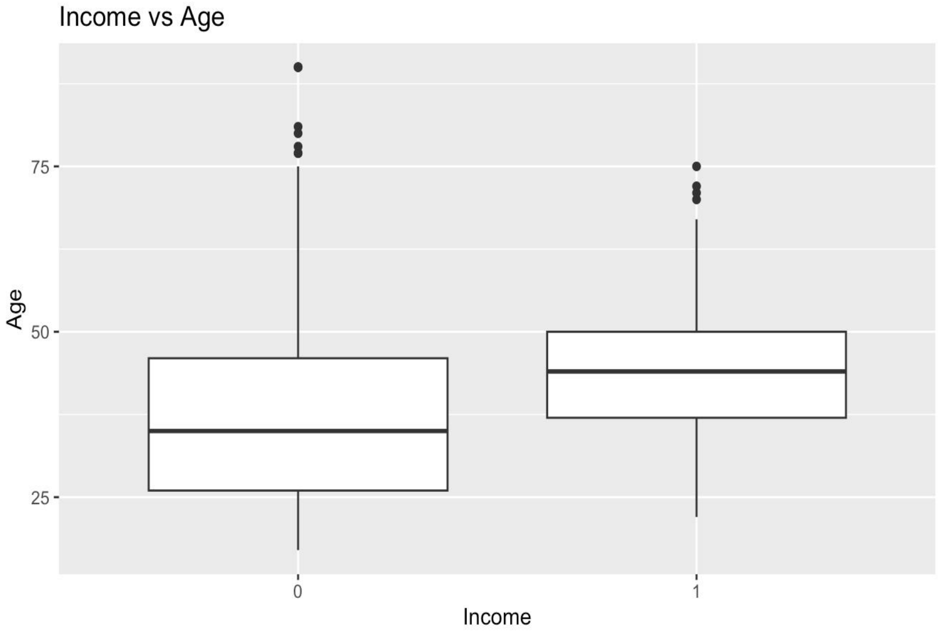




Males are more likely to have high incomes than females; and married individuals tend to earn more than those who are single, divorced, or separated, possibly due to dual-income households or other economic advantages.

Exploring numeric variables' Relationship with Income





Higher-income individuals (>50K) tend to work more hours per week, with fewer working very short hours, though both income groups have outliers with extreme work hours. They are also older on average, with a wider age distribution, and both groups include outliers at older ages.

03

Statistical modelling and results

Data preprocessing

We convert categorical variables (such as education, occupation) into factor types to ensure that the model correctly identifies categorical variables. The order of factors for the response variable Income is set to <=50k for the base group, which affects the direction of interpretation of subsequent OR values.

GLM and stepwise regression

- First time using GLM as modling. **All possible independent variables** are included in the model.

- Stepwise Regress is used for feature selection, **removing** the variables **'Nationality'** and **' Sex'**, detected with **AIC**.

```
Start:  AIC=998.68
Income ~ Age + Education + Marital_Status + Occupation + Sex +
    Hours_PW + Nationality

                 Df Deviance     AIC
- Nationality     1   966.80  996.80
- Sex             1   967.99  997.99
<none>                966.68  998.68
- Age             1   978.81 1008.81
- Education       4   996.49 1020.49
- Hours_PW        1  1000.02 1030.02
- Occupation      5  1036.31 1058.31
- Marital_Status  2  1199.76 1227.76

Step:  AIC=996.8
Income ~ Age + Education + Marital_Status + Occupation + Sex +
    Hours_PW

                 Df Deviance     AIC
- Sex             1   968.10  996.10
<none>                966.80  996.80
+ Nationality     1   966.68  998.68
- Age             1   978.86 1006.86
- Education       4   996.68 1018.68
- Hours_PW        1  1000.15 1028.15
- Occupation      5  1036.36 1056.36
- Marital_Status  2  1200.16 1226.16

Step:  AIC=996.1
Income ~ Age + Education + Marital_Status + Occupation + Hours_PW

                 Df Deviance     AIC
<none>                968.10  996.10
+ Sex             1   966.80  996.80
+ Nationality     1   967.99  997.99
- Age             1   979.70 1005.70
- Education       4   997.53 1017.53
- Hours_PW        1  1000.16 1026.16
- Occupation      5  1038.37 1056.37
- Marital_Status  2  1227.95 1251.95
```

Model checking

Four methods were used here to test the model fitting:

- GVIF test
- Hosmer-Lemeshow test
- Confusion matrix and accuracy
- ROC and AUC

GVIF test

	GVIF	Df	GVIF^(1/(2*Df))
Age	1.136553	1	1.066092
Education	1.518523	4	1.053605
Marital_Status	1.232088	2	1.053563
Occupation	1.652907	5	1.051538
Hours_PW	1.094464	1	1.046166

As the adjusted **GVIF values** are **all < 1.1**, indicating that there is **no significant multicollinearity** between all variables.

Hosmer-Lemeshow test

```
	Hosmer and Lemeshow goodness of fit (GOF) test

data:  data$Income, fitted(step_model)
X-squared = NA, df = 3, p-value = NA

[1] 0.1154056
```

We assessed model calibration using the **Hosmer-Lemeshow test** and **Brier Score**.
The Brier Score was **0.115**, which <0.25 suggesting that the predicted probabilities were fitted well.

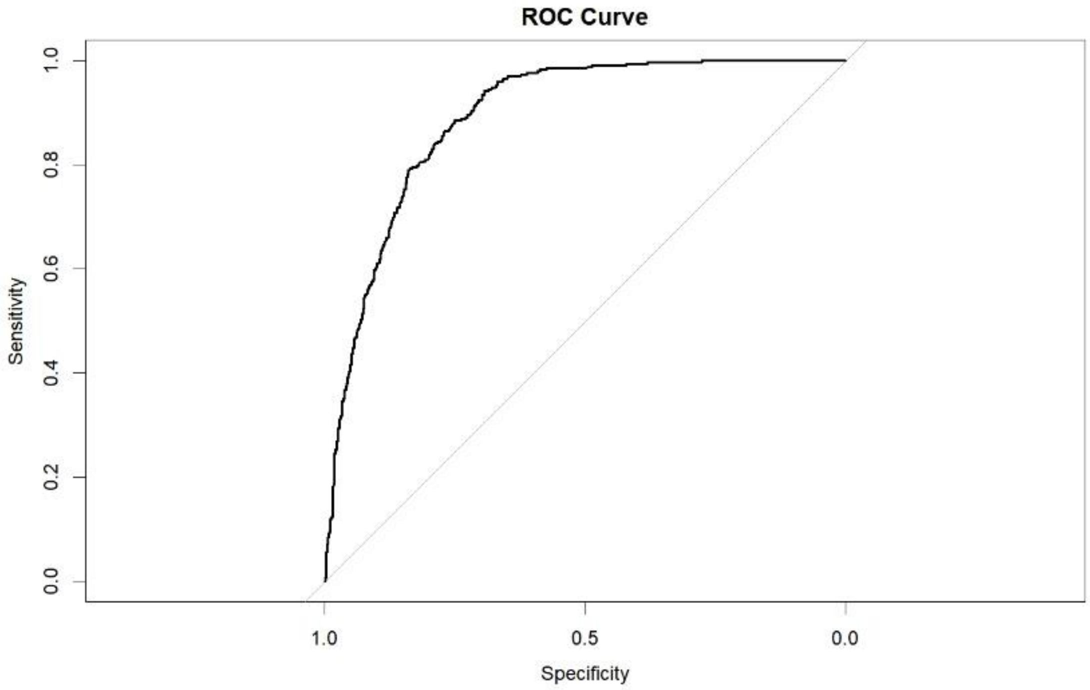
Confusion matrix and accuracy

```
> conf_matrix
          Actual
Predicted <=50K >50K
    <=50K   932  144
    >50K     96  204
> accuracy
[1] 0.8255814
```

The accuracy value of model prediction is **0.8255814**, indicating that about **82.56%** of prediction is correct.
However, considering the imbalance of the sample sizes between two different incomes, **ROC and AUC** could be conducted as further detections.

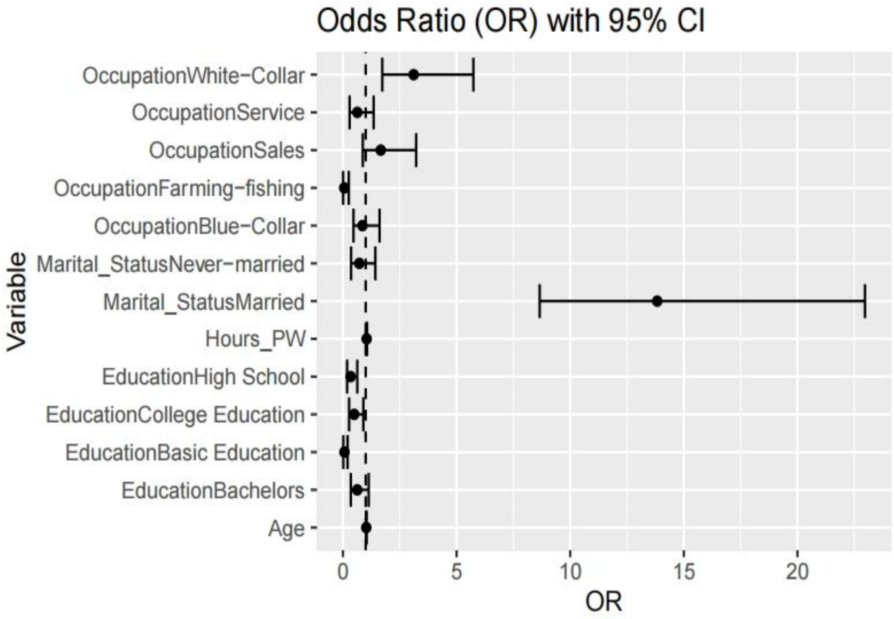
ROC and AUC

The ROC curve shows the curve bending sharply toward the **top-left corner**, which indicates high sensitivity and specificity. And the **AUC is 0.89**, indicating the model is effective at distinguishing between income groups.



Modeling results and analysis

```
Coefficients:
                             Estimate Std. Error
(Intercept)                 -5.106738   0.664669
Age                          0.025275   0.007447
EducationBachelors          -0.464104   0.301976
EducationBasic Education    -2.720209   0.619985
EducationCollege Education  -0.703813   0.305130
EducationHigh School        -1.077184   0.316037
Marital_StatusMarried        2.626701   0.248220
Marital_StatusNever-married -0.328516   0.350581
OccupationBlue-Collar       -0.155595   0.316453
OccupationFarming-fishing   -2.845193   0.857272
OccupationSales              0.508621   0.331698
OccupationService           -0.458251   0.388492
OccupationWhite-Collar       1.134498   0.304997
Hours_PW                     0.042542   0.007682
                            z value Pr(>|z|)
(Intercept)                  -7.683 1.55e-14 ***
Age                           3.394 0.000689 ***
EducationBachelors           -1.537 0.124320
EducationBasic Education     -4.388 1.15e-05 ***
EducationCollege Education   -2.307 0.021077 *
EducationHigh School         -3.408 0.000653 ***
Marital_StatusMarried        10.582  < 2e-16 ***
Marital_StatusNever-married  -0.937 0.348727
OccupationBlue-Collar        -0.492 0.622942
OccupationFarming-fishing    -3.319 0.000904 ***
OccupationSales               1.533 0.125180
OccupationService            -1.180 0.238173
OccupationWhite-Collar        3.720 0.000199 ***
Hours_PW                      5.538 3.06e-08 ***
```



04

Conclusions and extensions

Conclusions

Model performance:

- Model Fitting: The final GLM, eliminating the variables ‘Nationality’ and ‘Sex’, achieved an AIC of 996.1 and residual deviance of 968.1, indicating strong explanatory power.

- Interpretability: The adjusted values of GVIF of all variables are <1.1, indicating that there is no significant multicollinearity, and this model have good interpretability.

- Predictive Accuracy: The AUC of 0.893 demonstrates excellent discrimination between income groups.

Limitations & Future Directions

Non-significant factors: Sex and Nationality have no statistically significant impact on income (p=0.348), other relevant variables such as regional cost-of-living differences or industry growth trends can be found in further studies.

Temporal Bias: The dataset is from 1994. Using updated data for analysis could help improve the model.

Nonlinear Effects: Age and work hours may affect income in a nonlinear way. Future studies should explore threshold effects or diminishing returns for these variables.

Collinearity: The dataset may have perfect multicollinearity, causing instability in the model. Using PCA or feature selection could help reduce bias.

THANKS FOR LISTENING

Group 27