Analysis of Factors Influencing Individual Income Levels

Group 30

```r
library(tidyverse)
library(ggplot2)
library(caret)
library(pROC)
library(GGally)
library(rpart)
library(pROC)
library(randomForest)
library(randomForestExplainer)
library(party)
```

1 Introduction

This analysis aims to identify key socioeconomic factors that influence whether an individual earns more than $50,000 per year, using data from the **1994 US Census**. The dataset includes demographic and employment-related variables such as **age, education level, marital status, occupation, sex, hours worked per week, and nationality**, with income categorized into two groups: **$50K and >$50K per year**.

To address this, a **Generalized Linear Model (GLM)** will be applied to evaluate the impact of these factors on income levels. The findings will provide insights into the most significant predictors of higher earnings, contributing to a deeper understanding of income distribution patterns. The results will be summarized and presented in a structured format.

2 Exploratory Data Analysis

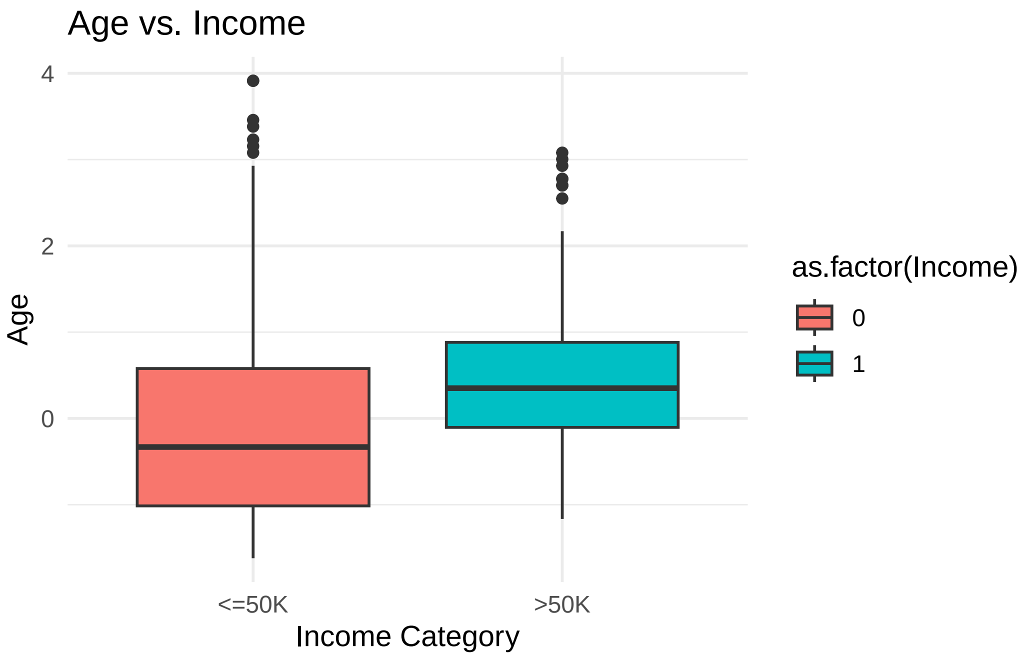
```r
df <- read.csv("dataset30.csv", stringsAsFactors = FALSE)
df <- df %>%
  mutate(across(everything(), ~str_remove_all(., ","))) %>%
  mutate(Hours_PW = as.numeric(Hours_PW),
         Age = as.numeric(Age),
         Income = ifelse(Income == ">50K", 1, 0))
df <- df %>% filter(Occupation != "?" & Nationality != "?")
df <- df %>% select(Age, Education, Marital_Status, Occupation, Sex, Hours_PW, Income)
df <- df %>% mutate(across(where(is.character), as.factor))
str(df)
```

```
'data.frame':   1409 obs. of  7 variables:
 $ Age           : num  79 44 51 30 52 33 50 51 41 35 ...
 $ Education     : Factor w/ 16 levels "10th","11th",..: 2 1 13 12 8 5 10 9 12 10 ...
 $ Marital_Status: Factor w/ 6 levels "Divorced","Married-civ-spouse",..: 2 1 2 2 2 2 1 2 4 2
 $ Occupation    : Factor w/ 13 levels "Adm-clerical",..: 3 1 9 6 11 4 3 12 6 11 ...
 $ Sex           : Factor w/ 2 levels "Female","Male": 2 1 1 2 2 2 2 2 2 1 ...
 $ Hours_PW      : num  7 42 50 40 48 40 45 40 40 30 ...
 $ Income        : num  0 0 1 0 1 0 1 1 0 1 ...
```

```r
df$Age <- scale(df$Age)
set.seed(1111)
trainIndex <- createDataPartition(df$Income, p = 0.8, list = FALSE)
train_data <- df[trainIndex, ]
test_data <- df[-trainIndex, ]
```

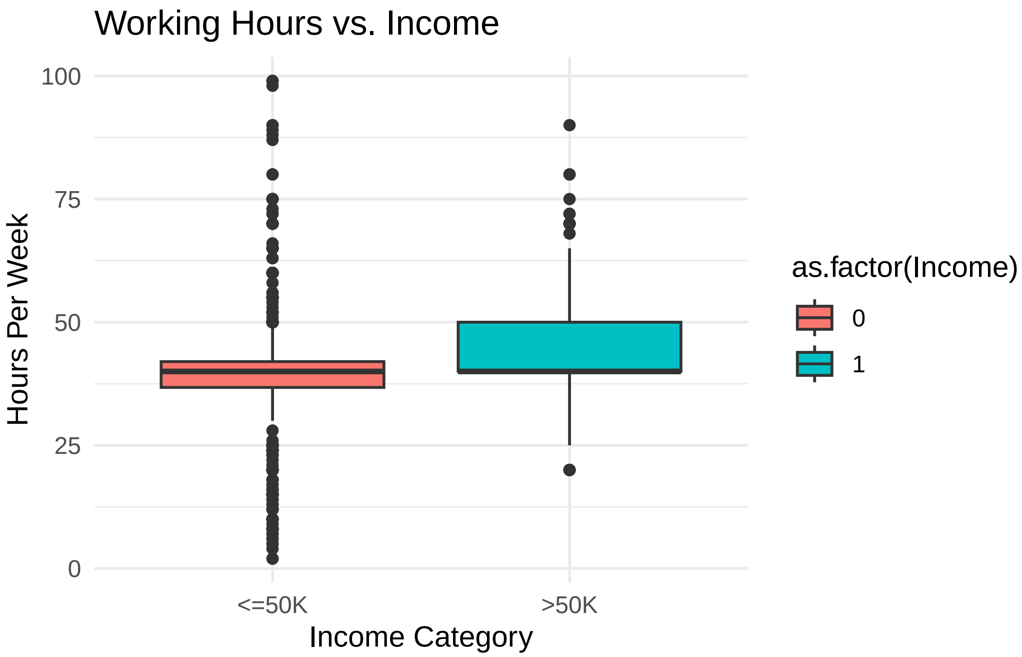
This data analysis first involved cleaning, transforming, selecting, and standardizing the raw data. Then, the dataset was split into **80% training set and 20% test set** to ensure data quality and enhance the model's generalization ability.

```r
# Age vs. Income
ggplot(df, aes(x = as.factor(Income), y = Age, fill = as.factor(Income))) +
  geom_boxplot() +
  scale_x_discrete(labels = c("<=50K", ">50K")) +
  ggtitle("Age vs. Income") +
  xlab("Income Category") +
  ylab("Age") +
  theme_minimal()
```



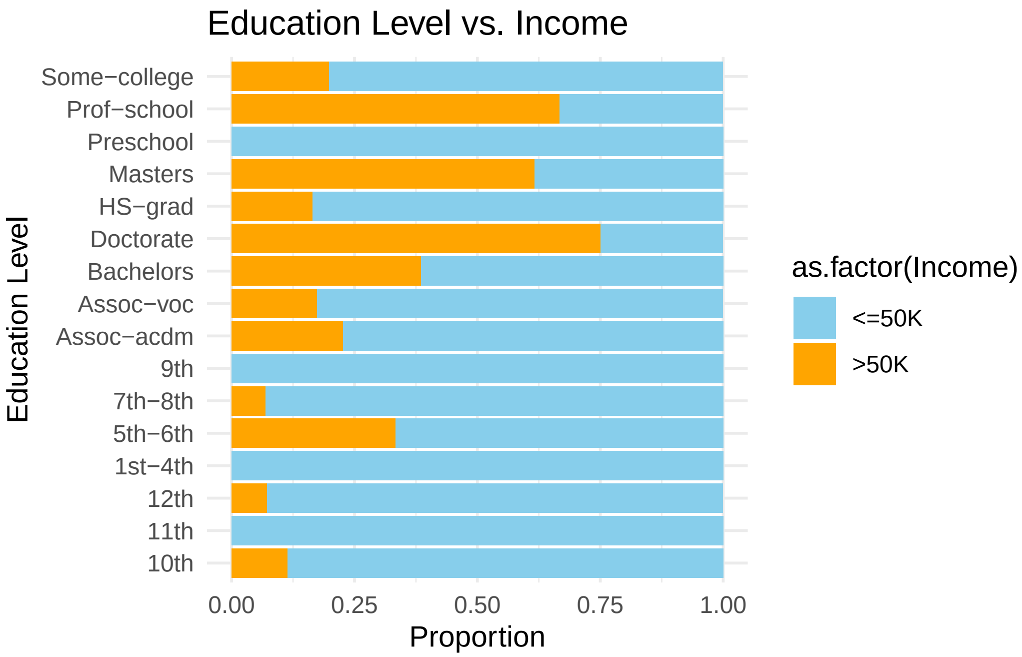
As can be seen from the figure, there is a certain positive correlation between age and income: the median age of the high-income group (income >50K) is higher than that of the low-income group (income 50K), indicating that older people are more likely to has the higher income.

```
#Working Hours vs. Income
ggplot(df, aes(x = as.factor(Income), y = Hours_PW, fill = as.factor(Income))) +
  geom_boxplot() +
  scale_x_discrete(labels = c("<=50K", ">50K")) +
  ggtitle("Working Hours vs. Income") +
  xlab("Income Category") +
  ylab("Hours Per Week") +
  theme_minimal()
```



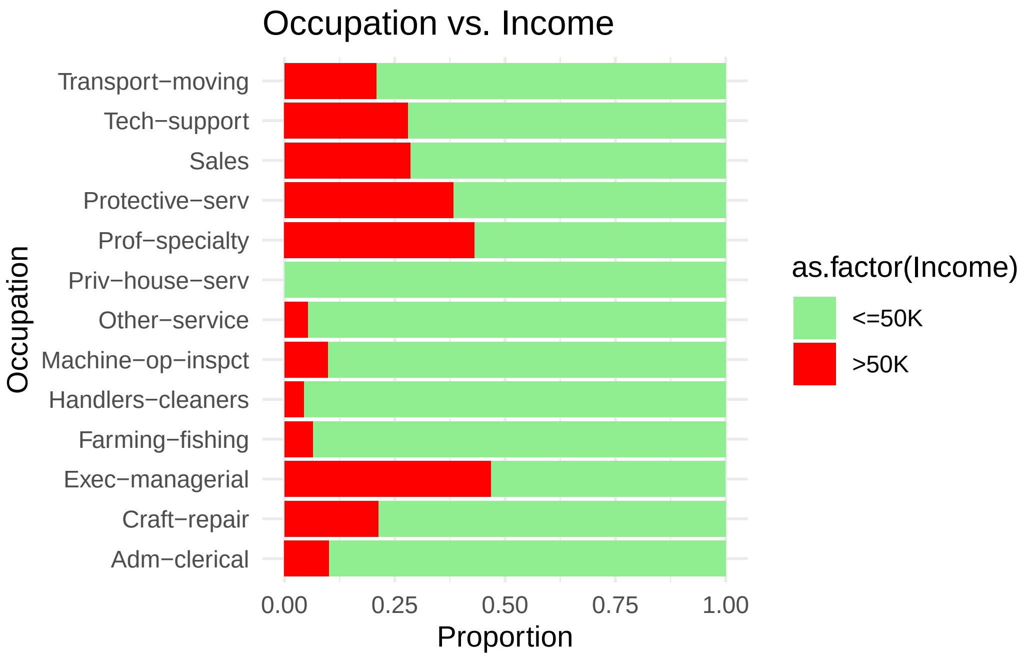
Working hours have a significant impact on income, higher earners tend to work longer hours, but hours alone do not fully determine income. The working hours of low-income groups are more concentrated, but there are more outliers: individuals with extremely low working hours (<25 hours) and extremely high working hours (>50 hours).

```r
#Education Level vs. Income (Bar Chart)
ggplot(df, aes(x = Education, fill = as.factor(Income))) +
  geom_bar(position = "fill") +
  coord_flip() +
  scale_fill_manual(values = c("skyblue", "orange"), labels = c("<=50K", ">50K")) +
  ggtitle("Education Level vs. Income") +
  xlab("Education Level") +
  ylab("Proportion") +
  theme_minimal()
```



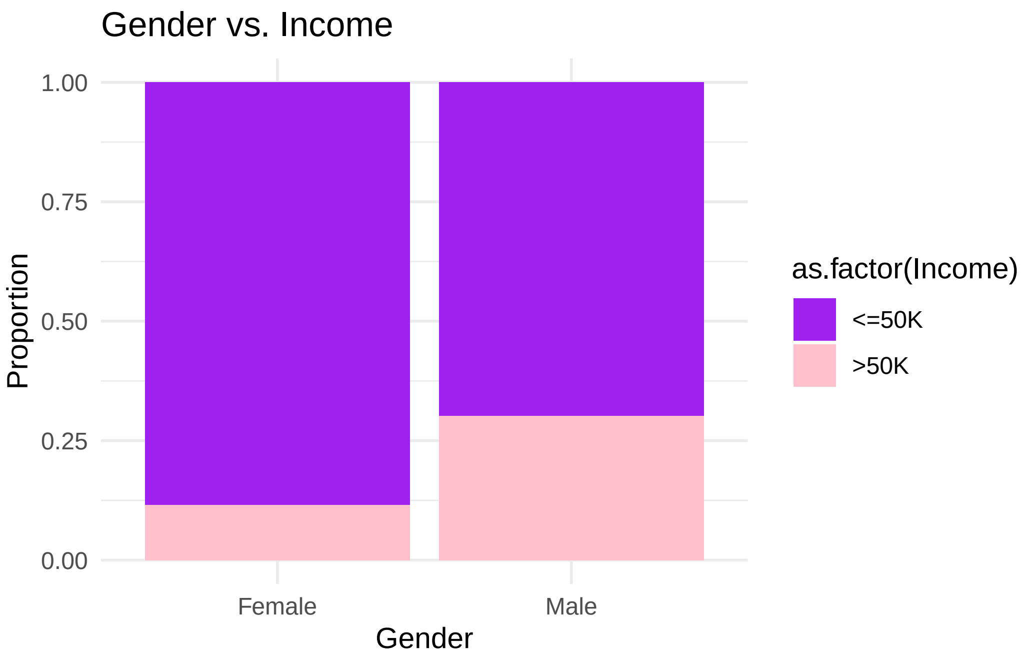
Education level is positively correlated to income, and higher degree is more likely to earn a high income. In particular, the proportion of high income of master's, doctor's and professional school graduates is significantly higher, while those with low education level are mainly concentrated in low-income groups. This suggests that education level plays an important role in income.

```r
#Occupation vs. Income (Bar Chart)
ggplot(df, aes(x = Occupation, fill = as.factor(Income))) +
  geom_bar(position = "fill") +
  coord_flip() +
  scale_fill_manual(values = c("lightgreen", "red"), labels = c("<=50K", ">50K")) +
  ggtitle("Occupation vs. Income") +
  xlab("Occupation") +
  ylab("Proportion") +
  theme_minimal()
```



The figure shows clear income differences between different occupational categories. A higher percentage of "Exec-managerial" and "Prof-specialty" occupations make more than 50,000, indicating that these occupations are more likely to generate higher income. By contrast, the proportion of "Priv-house-serv" and "Handlers-cleaners" is significantly lower.

```
#Gender vs. Income (Bar Chart)
ggplot(df, aes(x = Sex, fill = as.factor(Income))) +
  geom_bar(position = "fill") +
  scale_fill_manual(values = c("purple", "pink"), labels = c("<=50K", ">50K")) +
  ggtitle("Gender vs. Income") +
  xlab("Gender") +
  ylab("Proportion") +
  theme_minimal()
```



There are significant differences in income distribution between men and women. The proportion of men earning more than 50,000 is significantly higher than that of women, while the women income less than 50,000 is relatively higher.

3 Formal Data Analysis

3.1 Generalised Linear Model

```
full_model <- glm(Income ~ ., data = train_data, family = binomial)
stepwise_model <- step(full_model, direction = "backward")
```

```
Start:  AIC=883.77
Income ~ Age + Education + Marital_Status + Occupation + Sex + 
    Hours_PW


                 Df Deviance    AIC
- Sex             1   812.19 882.19
<none>                811.77 883.77
- Hours_PW        1   817.42 887.42
- Age             1   824.87 894.87
- Occupation     12   852.98 900.98
- Education      15   898.91 940.91
- Marital_Status  5   925.84 987.84


Step:  AIC=882.19
Income ~ Age + Education + Marital_Status + Occupation + Hours_PW

                 Df Deviance     AIC
<none>                812.19  882.19
- Hours_PW        1   818.59  886.59
- Age             1   826.05  894.05
- Occupation     12   853.51  899.51
- Education      15   899.58  939.58
- Marital_Status  5   946.75 1006.75
```

The **logistic regression** model is built to predict income level **(`Income`)** by considering various socioeconomic factors, such as **age, education, marital status, occupation, sex,** and **hours worked per week**. To further refine the model, **stepwise regression** is applied using **backward elimination**, which systematically removes variables that do **not significantly (sex)** contribute to the prediction based on the **Akaike Information Criterion (AIC)**. The initial model's AIC was **883.77**, while the optimized model's AIC was reduced to **882.19**, indicating that the model became more concise while maintaining a good fit. As a result, this approach not only simplifies the model but also helps to reduce overfitting and enhance interpretation by retaining only the most significant predictors.

```r
test_predictions <- predict(stepwise_model, newdata = test_data, type = "response")
test_predicted_classes <- ifelse(test_predictions > 0.5, 1, 0)
confusionMatrix(factor(test_predicted_classes), factor(test_data$Income))
```

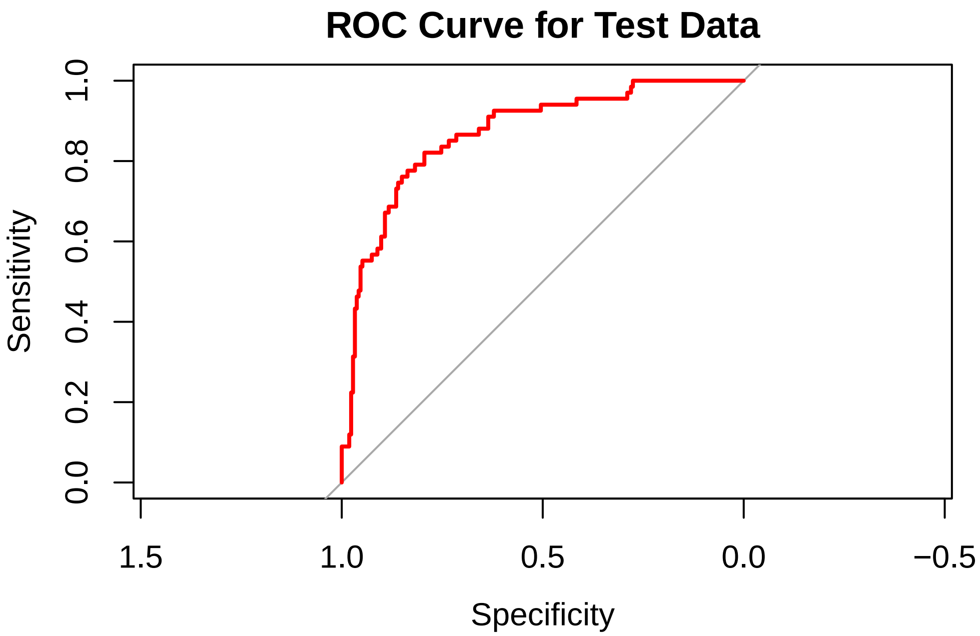
```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 203  30
         1  11  37

               Accuracy : 0.8541
                 95% CI : (0.8073, 0.8932)
    No Information Rate : 0.7616
    P-Value [Acc > NIR] : 8.846e-05
```

```
                  Kappa : 0.5549

 Mcnemar's Test P-Value : 0.004937

            Sensitivity : 0.9486
            Specificity : 0.5522
         Pos Pred Value : 0.8712
         Neg Pred Value : 0.7708
             Prevalence : 0.7616
         Detection Rate : 0.7224
   Detection Prevalence : 0.8292
      Balanced Accuracy : 0.7504

       'Positive' Class : 0
```

```
conf_matrix <- confusionMatrix(factor(test_predicted_classes), factor(test_data$Income))
accuracy <- conf_matrix$overall["Accuracy"]
print(accuracy)
```

```
 Accuracy
0.8540925
```

```
roc_curve_test <- roc(test_data$Income, test_predictions)
plot(roc_curve_test, col = "red", main = "ROC Curve for Test Data", xlim = c(1, 0), ylim = c
```



```
auc(roc_curve_test)
```

```
Area under the curve: 0.8683
```

The evaluation results of this logistic regression model on the test set indicate a high classification capability, with an accuracy of **85.41%** within the **95% confidence interval (80.73% - 89.32%)**, demonstrating stable predictive performance. The confusion matrix further shows that the model has a strong ability to identify low-income individuals, achieving a **sensitivity of 94.86%**, meaning that most low-income individuals are correctly classified. However, the **specificity is only 55.22%**, suggesting that the model has some difficulty in correctly identifying high-income individuals, with a considerable number being unclassified as low-income. Despite this, the model's **accuracy (ACC)** still reaches **85.4%**. In terms of overall classification performance, the **AUC value of 0.8683** indicates that the model performs well in distinguishing between high-income and low-income individuals, and the **ROC curve** demonstrates strong discriminatory power.

3.2 Random Forest

```
df = read.csv("dataset30.csv")
```

```r
#clear the data

df = df %>%
  mutate(across(everything(), ~str_remove_all(., ","))) %>%
  mutate(Hours_PW = as.numeric(Hours_PW),
         Age = as.numeric(Age))
df = df %>% filter(Occupation != "?" & Nationality != "?")
df = df %>% select(Age, Education, Marital_Status, Occupation, Sex, Hours_PW, Income)
df = df %>% mutate(across(where(is.character), as.factor))
x = df[,1:6]
y = df[,7]

x_pre = x[1:1228,]
x_te = x[1229:1409,]
y_pre = y[1:1228]
y_te = y[1229:1409]
pre = cbind(y_pre,x_pre)
rf_model = randomForest(y_pre~.,data = pre,ntree = 100,importance = T)
predictions <- predict(rf_model, newdata = x_te)
tb = table(predictions,y_te)
TP = tb[1,1]
FP = tb[1,2]
FN = tb[2,1]
TN = tb[2,2]
print(tb)
```

```
           y_te
predictions <=50K >50K
      <=50K   122   20
      >50K     14   25
```

After using the method of Random Forest , we do not need to do any judgement on the value of predictions ,the random forest automatic done the **decision tree** for the data. The table shows that most of the value full in the range TP and FP, which mean most of the data fit well.

```r
acc = (TP + TN)/(TP + FP + FN + TN)
TPR = TP/(TP + FN)
TNR = TN/(FP + TN)
print(acc)
```

```
[1] 0.8121547
```

```
print(TPR)
```

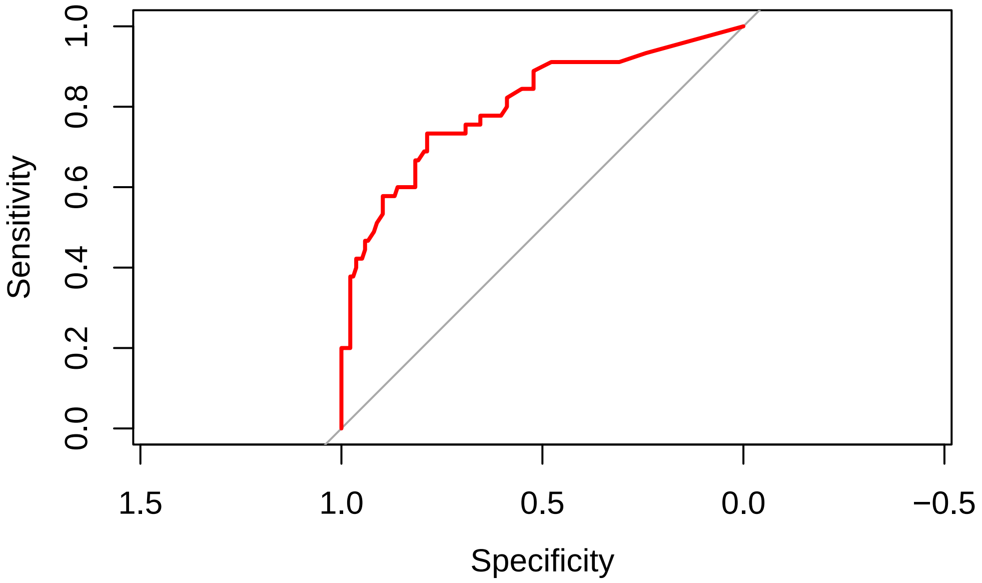
```
[1] 0.8970588
```

```
print(TNR)
```

```
[1] 0.5555556
```

The accuracy of random forest is up to **79.55%** which shows that it give lots of information to the predictors, and th achieving a **sensitivity of 87.506%** and **specificity of 55.55%** ,which mean it have a low ability in predict the negative size of value which is the income >50%.

```
pred_probs = predict(rf_model,newdata = x_te, type = "prob")[,2]
roc_c = roc(y_te,pred_probs)
plot(roc_c,col = "red")
```



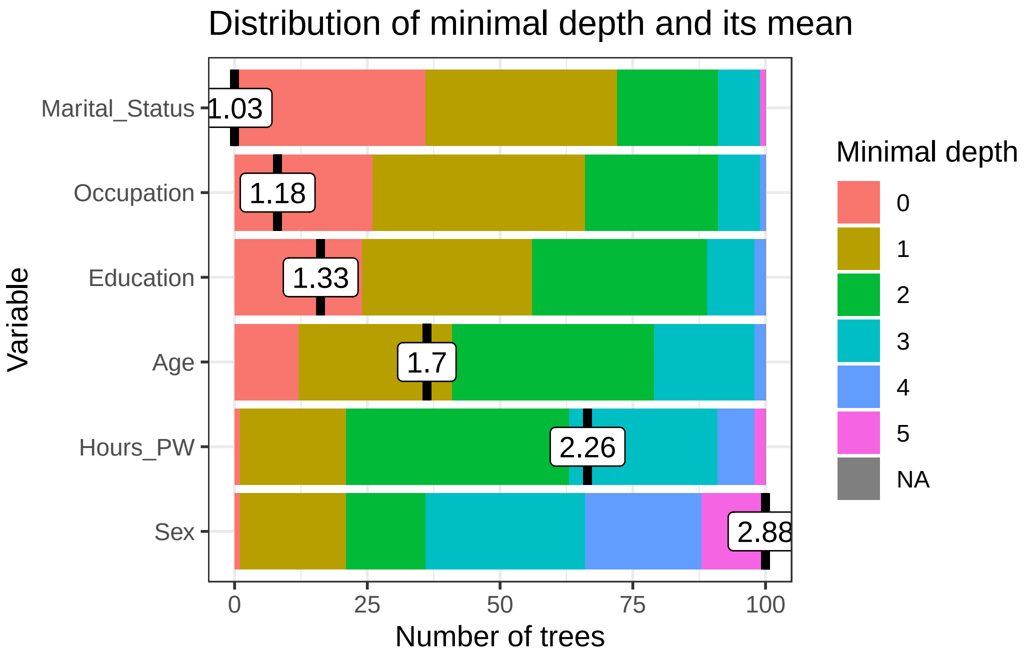
```
auc(roc_c)
```

```
Area under the curve: 0.8033
```

Using the **ROC and AUC** to detect the balance of the model ,from the plot of ROC we see that the curve are far away from the straight line ,which can also reflect by the number of AUC is **79.68**, that means the model show more information in the True Positive range than False Negative range ,so it give a good fit to the data.

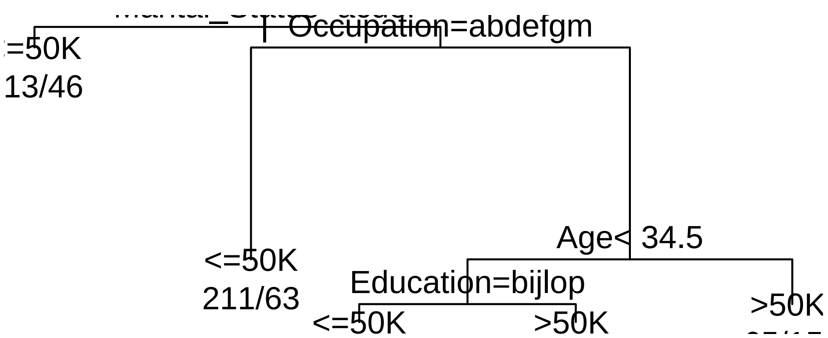
```
#plot the rf
#more minimal depth have high weight in predict
explain_forest(rf_model, data = pre)
```

```
plot_min_depth_distribution(rf_model)
```



This is the plot of **minimal depth** ,the plot give outlay the first time of the "tree branch" split by each variable ,the closer the split from the root the more influence it comes .So form the plot **Marital_Status** give the most information to the model ,then comes the **Education**. However that didn't mean that the other variable didn't appeared on the plot don't make any influence ,we still need to consider the times that it split for each variable.

```
rpart_tree <- rpart(y_pre ~ ., data=pre)
plot(rpart_tree,compress = TRUE)
text(rpart_tree, use.n = TRUE)
```



The decision tree plot is the plot that give the most obvious information to the model. We can see clearly from each branch that how the data are alienation to different slop. From this data we can see that there are no such much split in the plot maybe cause by the small amount of data or there are no strong influence with most of the variable.

3.2.1 The Resample for Model

```
K = 5
set.seed(1111)
folds = cut(1:1409, breaks=K, labels=FALSE)
sen = sep = acc =numeric(K)
for(k in 1:K){
  x.train = x[which(folds!=k),]
  x.text = x[which(folds==k),]
  y.train = y[which(folds!=k)]
  y.text = y[which(folds==k)]
  pre = cbind(y.train,x.train)
  rf_fit = randomForest(y.train~.,data = pre,ntree = 100,importance = T)
  predictions = predict(rf_fit,newdata =x.text)
  tb = confusionMatrix(y.text,predictions)
  tb.class = tb$byClass
  tb.overall = tb$overall
  sen[k] = tb.class[1]
  sep[k] = tb.class[2]
  acc[k] = tb.overall[1]
}
##the mean of 5 time predict-text outcome
sen_m = sum(sen)/5
sep_m = sum(sep)/5
acc_m =sum(as.numeric(acc))/5
print(sen_m)
```

```
[1] 0.8622353
```

```
print(sep_m)
```

```
[1] 0.6302956
```

```
print(acc_m)
```

```
[1] 0.8140506
```

the re-sample are usually used to do repeat experiment ,this is for helping the model to give a more precise ACC and any other data. The method shows here is the **K-fold** for k equal to 5 . which mean split the hole data into 5 part and set each part as test data . which will give 5 different outcome ,to make the outcome useful ,we take **mean** for these outcome (thus ACC: **81.4%**),this would be a more representative outcome then the ACC before.

3.3 Comparative Analysis between the GLM and Random Forest

From the data we get above we can have a compare to two model. The data show above:(LEFT FOR GLM AND RIGHT FOR RF) ACC:0.8541 - 0.8140 Sensitivity:0.9486 - 0.8751 Specificity:0.5522 - 0.5555 AUC:0.8683 - 0.7968 After looking at these data we can see that two data we can get the conclusion that two model are giving the near ACC and AUC ,and the GLM is higher, it can show that GLM are giving more information to the data, but maybe after changing the depth of Random Forest ,the fitting rate of it would be higher than the GLM. So we can say that both of the model can be use in this data fitting but GLM is a better one . tip: the RF model are higher in Specificity and lower in Sensitivity show that it have a better detection on the Negative side but less in Positive side.

4 Conclusions

This study utilized data from the **1994 U.S. Census** to analyze key socioeconomic factors influencing individual income levels. The results indicate that **education level, occupation, marital status, and age** are crucial determinants of income, with higher education and specialized professions significantly increasing earning potential. Additionally, **sex**, as men are more likely to earn higher incomes. Furthermore, **weekly working hours** have a certain impact on income, though with diminishing marginal returns.