Analysis of US Census Income Data and Bias in Machine Learning

Ronen Reouveni, Paul Mackanos III, and Kathleen McConnell

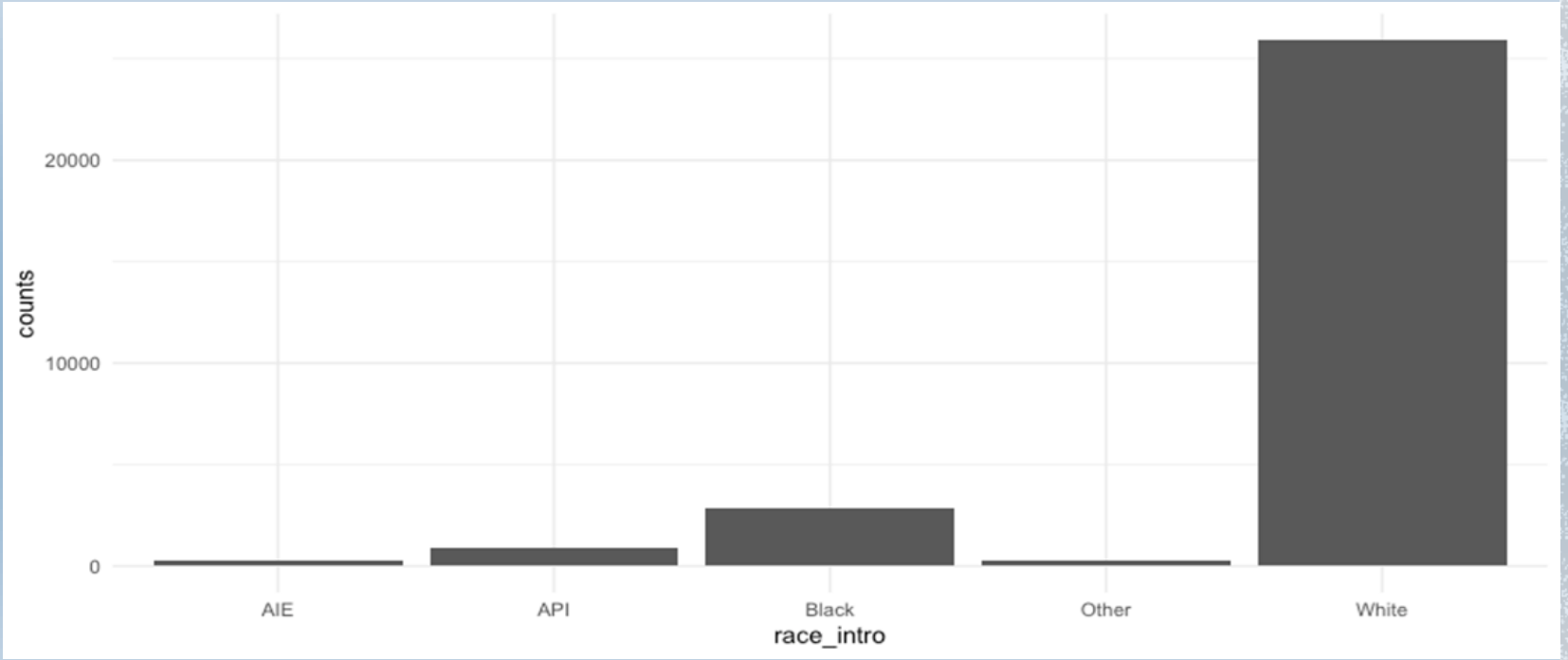
Data Attributes

Before any data munging, this data set consisted of approximately 32,561 observations of 15 variables.

- >$50,000 income per year
- <=$50,000 income per year
- Age
- Work class
- Fnlwgt (# of people census takers believe that observation represents)
- Education
- Marital status
- Occupation
- Relationship
- Race
- Sex
- Capital gain
- Capital loss
- Hours per week
- Native country

Race Data

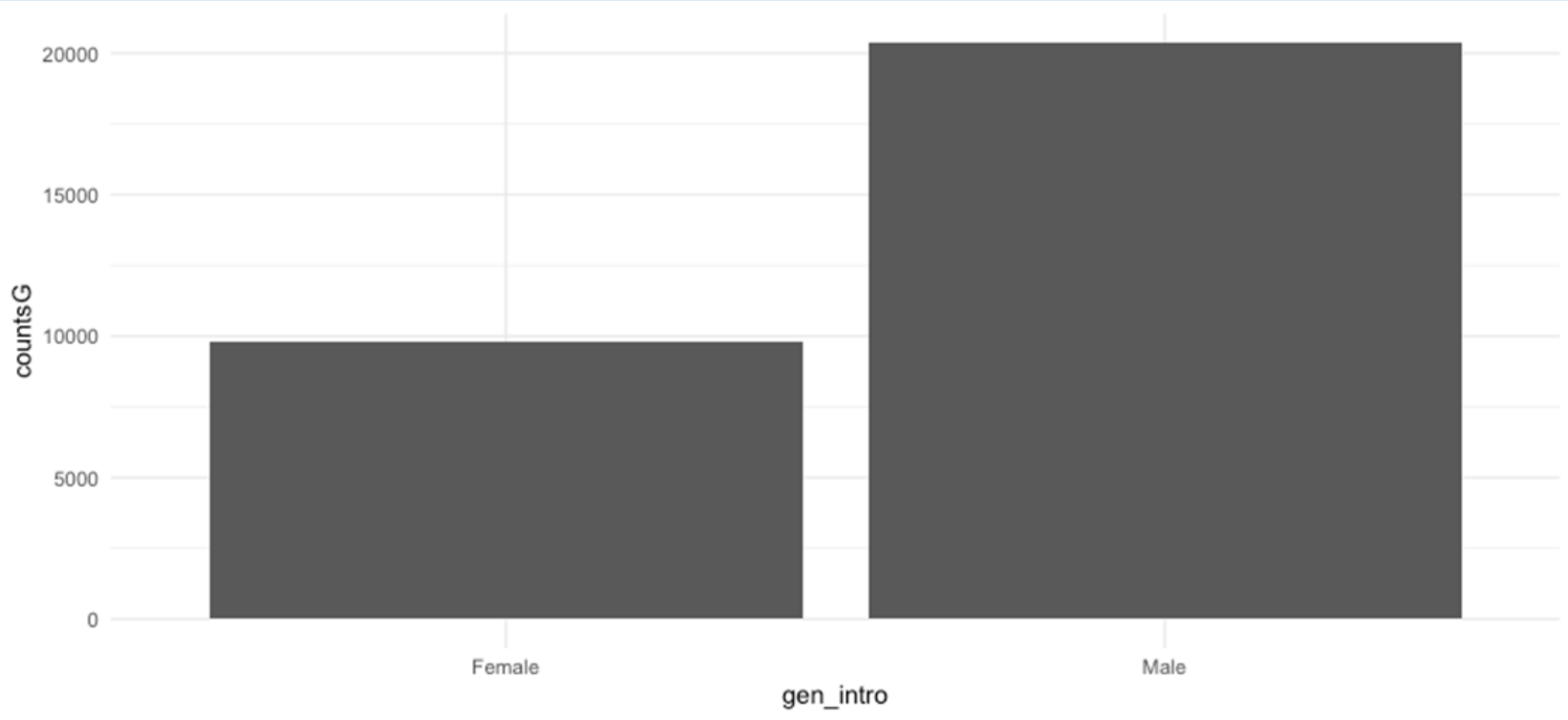
Table 1: Counts of each Race Contained in Sample Data



- The distribution of race in the source data is very skewed white, a deficiency of samples for other races means training and testing should be set close to 50/50 to get enough samples to test for accurate modeling.
- By training on half the data, the model is still stable, but we have enough testing data to examine the bias
  - Amer-Indian-Eskimo: 286
  - Asian-Pac-Islander: 895
  - Black: 2,817
  - Other: 231
  - White: 25,933

Gender Data

Table 2: Gender Count in Sample Data



- The data also skews heavily male. In theory we would think this distribution should be closer to 50/50.
  - Male: 20,380
  - Female: 9,782

The Questions

Can we find a bias in the frequency of type II errors given race and gender? A type II error is defined as miss classifying an observation as below 50k a year when they are above 50k a year.

What are the most important predictors according to mean Gini decrease?

What are some of the correlations between those inputs and income?

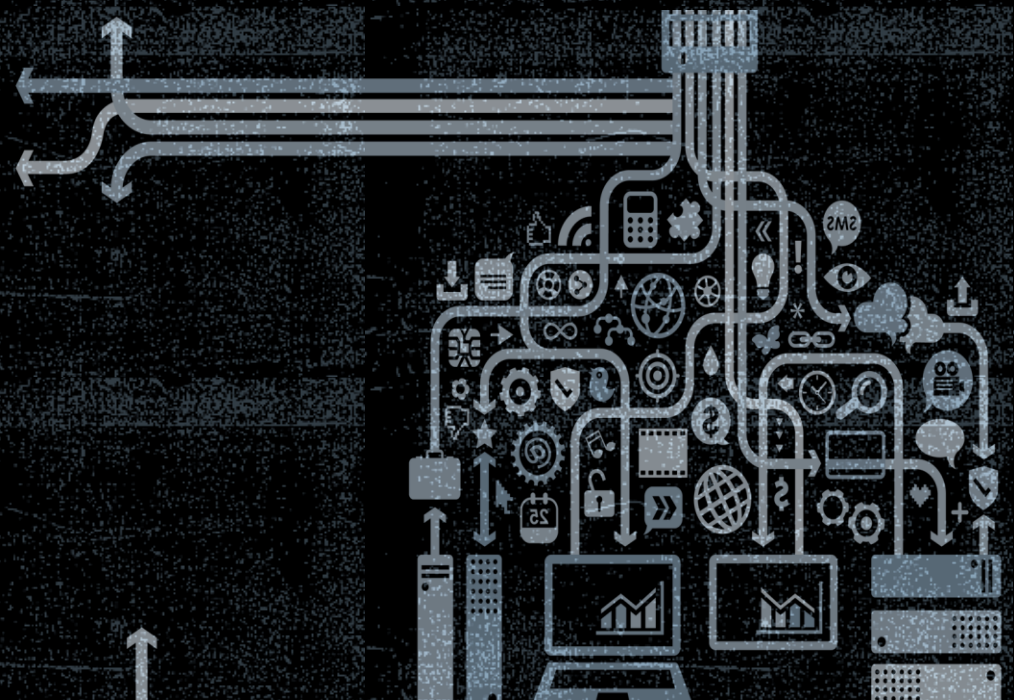
What is the simplest model we can implement with an error rate of at most 20%?

The Clean Up

- This data was extensively studied to determine the usable variables
- After this initial analysis, the data set was forwarded to the preprocessing phase, where all the errors in the data were removed to make it usable for further analysis
- Made column names consistent
- Summarized columns of data
- The number of rows in the data set (myCleanData) was 30,162 observations of 15 variable columns after munging
  - Reduced observations by 2,399

Modeling Techniques

- Support Vector Machines
- Random Forest
- Random Forest for Bias Determination
- Correlations
- Decision Trees



Code used to install packages and libraries:

```
> install.packages("randomForest")
> install.packages("pROC")
> install.packages("reshape2")
> install.packages("caret")
> install.packages("ISLR")
> install.packages("tree")
> install.packages("kernlab")
> library(caret)
> library(e1071)
> library(randomForest)
> library(pROC)
> library(ggplot2)
> library(reshape2)
> library(cluster)
> library(ISLR)
> library(tree)
> library(kernlab)
```

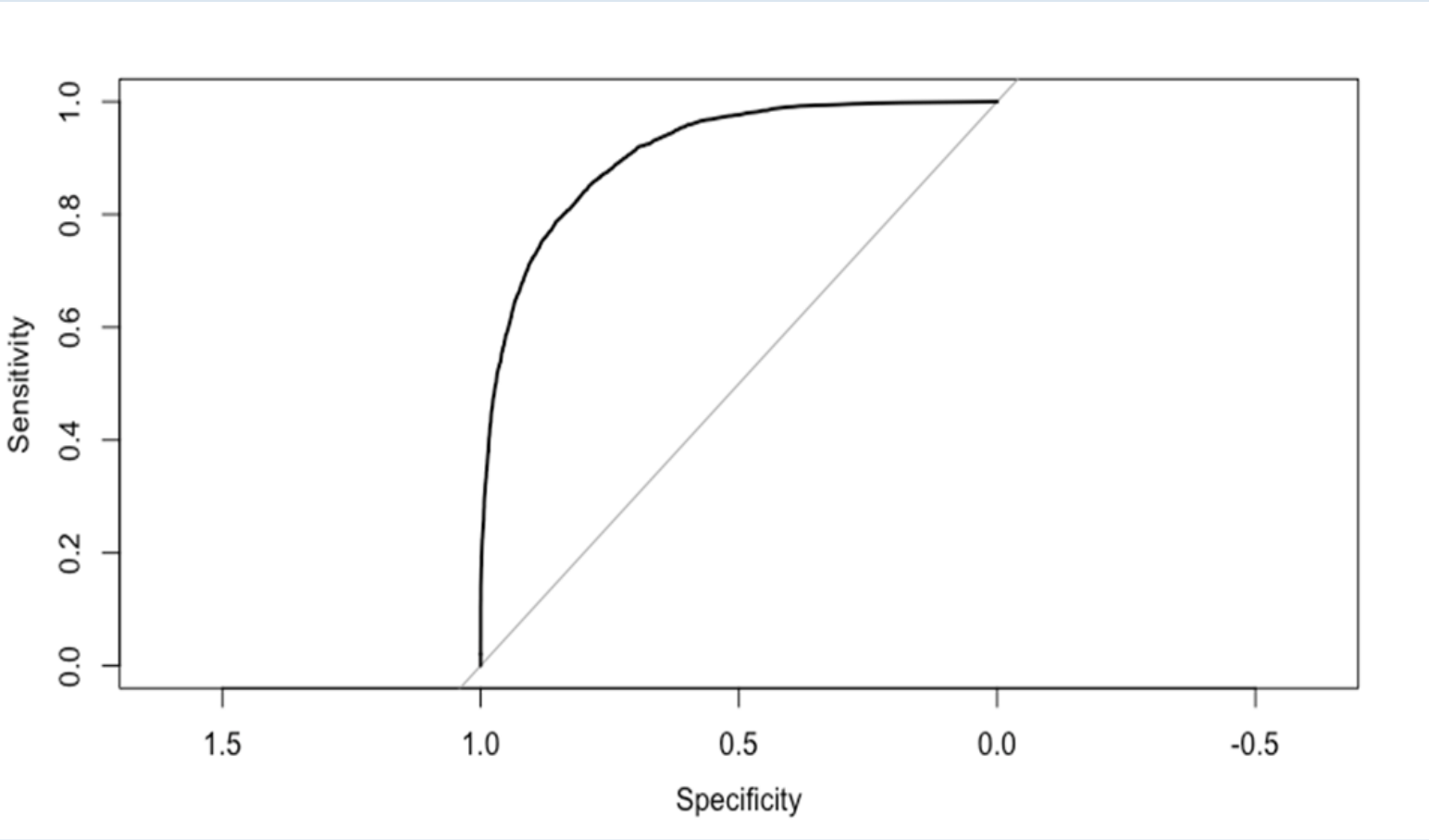
```
svm(formula = income ~ ., data = myCleanData[train_idx,
    ])
Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1
Number of Support Vectors:  5574
> svm_noT_pred <- predict(svm_noT, myCleanData[-train_idx,])
> svm_result <- table(svm_noT_pred,myCleanData$income[-
train_idx])
> svm_rate <- (svm_result[1,1] +
svm_result[2,2])/(svm_result[1,1] + svm_result[2,2] +
svm_result[1,2] + svm_result[2,1])
> svm_rate
[1] 0.8474239

> #svm class rate 84.74%
> support2 <- ksvm(income ~., data =
myCleanData[train_idx,], kernal="besseldot", kpar =
"automatic", C=3)
> support2
Support Vector Machine object of class "ksvm"
SV type: C-svc  (classification)
 parameter : cost C = 3
Gaussian Radial Basis kernel function.
 Hyperparameter : sigma =  0.0778132721216806
Number of Support Vectors: 5422
Objective Function Value : -13281.22
Training error : 0.125787
> supp_pred2 <- predict(support2, myCleanData[-train_idx,])
> supp_result2 <- table(supp_pred2,myCleanData$income[-
train_idx])
> supp_rate2 <- (supp_result2[1,1] +
supp_result2[2,2])/(supp_result2[1,1] + supp_result2[2,2] +
supp_result2[1,2] + supp_result2[2,1])
> supp_rate2
[1] 0.8545852
```

Support Vector Machines

- An excellent tool for classification, novelty detection, and regression
- These SVM models show that with 85% certainty, we can classify an observation as above or below $50,000 a year in income
  - This is important to keep in mind in establishing a baseline as to issues with bias.

Table 3 ROC Curve1



The table shows our model performing well above the 50% trend line. Next, we will see which inputs the Random Forest thought was most important.

Random Forest

- Performs better than the SVM, and we will, therefore, use it to analyze bias in the following sections
- It is crucial to keep in mind these strong classification prediction rates
  - In a social science setting like this one, 86% accuracy is strong evidence of relationships between the predictors and income

Mean Decrease in Gini

- It is the average of a variable's total decrease in node impurity, weighted by the proportion of samples reaching that node in each individual decision tree in the random forest
- Used as a measure of how important a variable is for estimating the value of the target variable across all the trees that make up the forest
- A higher number indicates higher variable importance

The most important take away from this is how little importance the Random Forest attributed to race and sex

rF2$importance

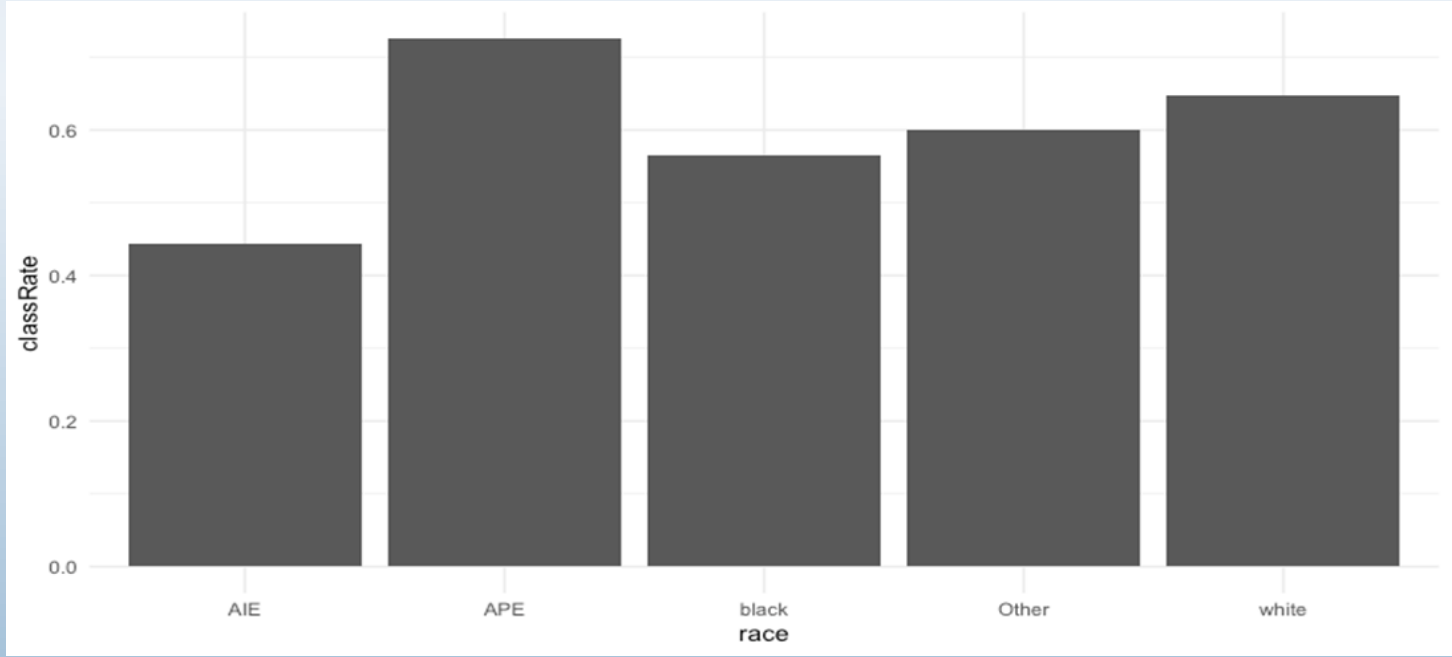
MeanDecreaseGini	
age	559.97793
workclass	193.26359
fnlwgt	528.56823
education	334.93850
educationNum	349.64303
maritalStatus	527.79199
occupation	467.10795
relationship	495.83711
race	65.19531
sex	70.30795
capitalGain	577.54788
capitalLoss	164.58231
hoursPerWeek	342.42957
nativeCountry	119.0994

How to Quantify Bias?

- Split Data into training and testing set
- Train a Random Forest on the training data
- Subset the testing data by race
- Use the Random Forest to predict values for the testing data set
- Calculate the percent of time the Random Forest predicts under $50,000 a year when the true value is above $50,000 a year
- Repeat steps for gender

Random Forest Bias Determination

Table 4: Race Classification Rate

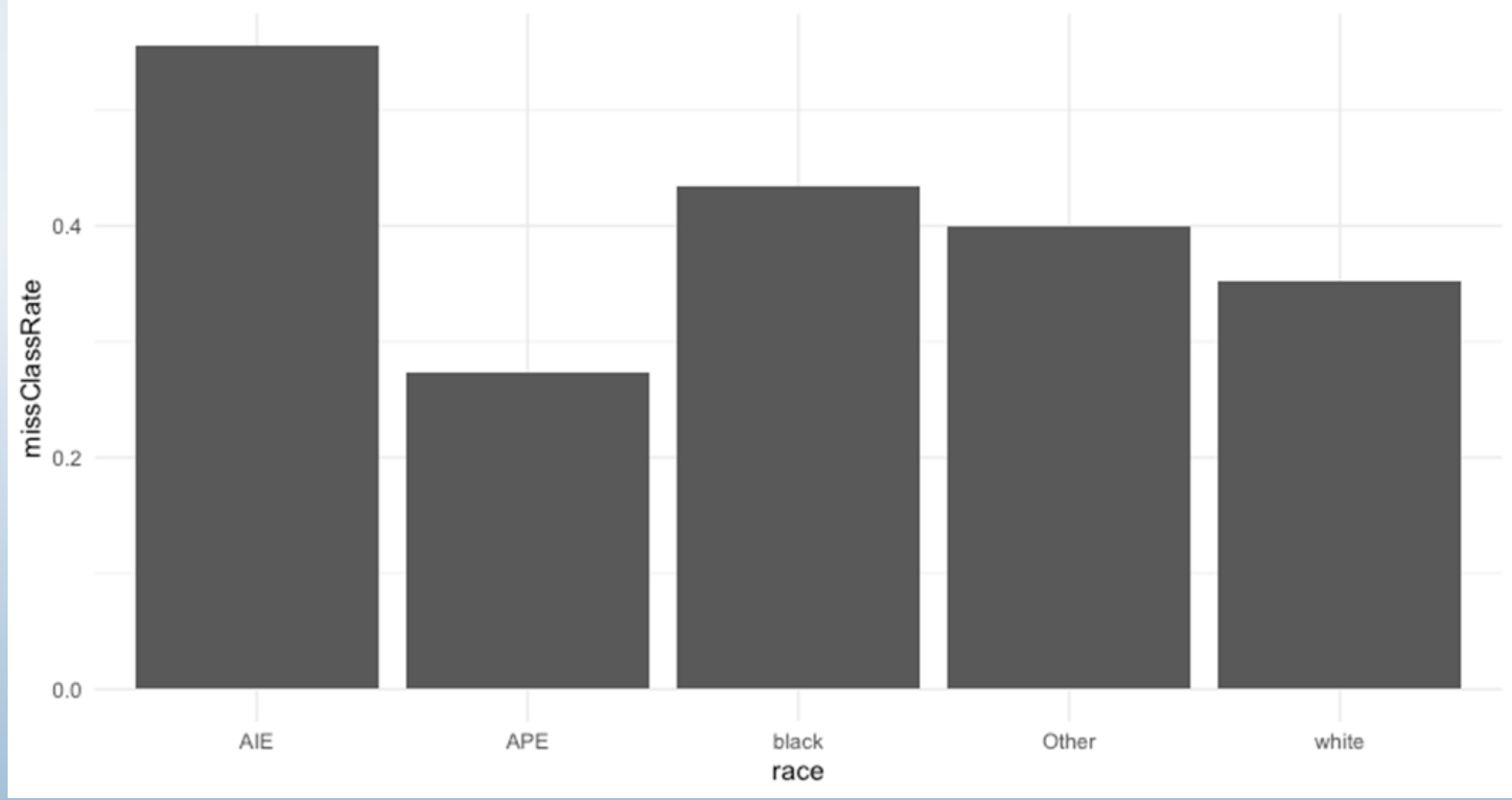


- Shows the probability of correct classification for type II
- All are below the overall classification rate, but there is a significant difference across races

Random Forest Bias Determination, cont.

- Shows the probability of misclassification of being under $50,000 a year when the true value is above $50,000 a year given race
  - Roughly 20% more likely to be misclassified given American Indian vs. white
  - Nearly 10% more likely to be misclassified in this setting, given a race of black vs. White
  - Asian Pacific Islander is the least likely to be miss classified.
- Shows significant evidence that we have a bias
- Most interestingly, we show these biases despite Gini decrease being almost irrelevant for race

Table 5: Race Misclassification Rate



Random Forest Bias Determination, cont.

Table 6: Gender Classification Rate

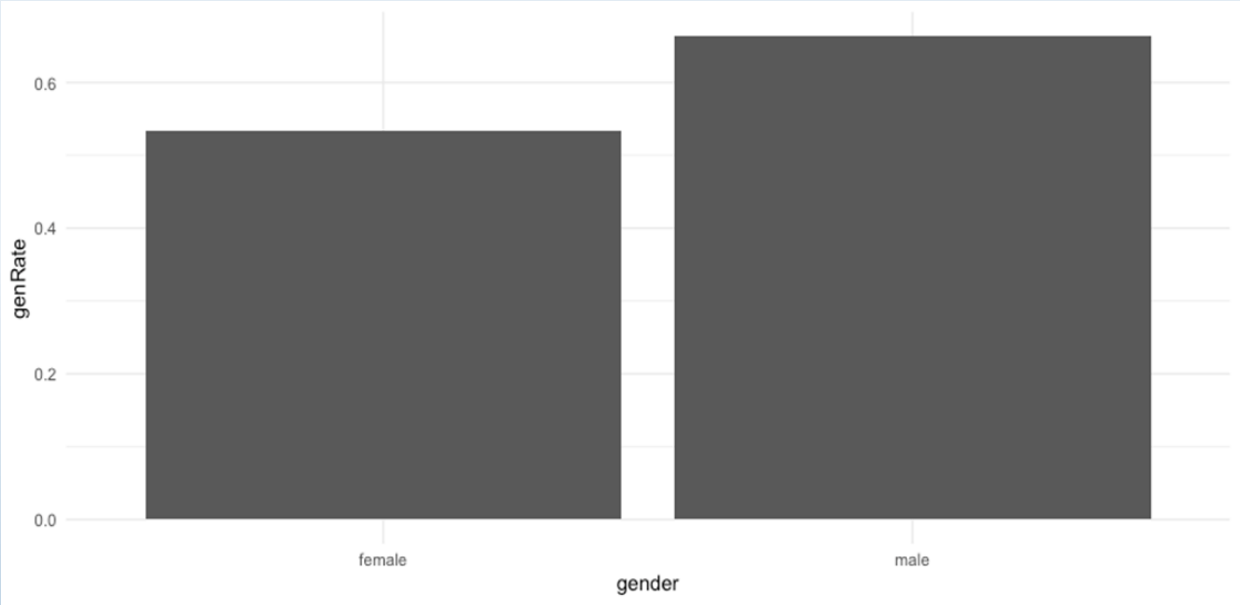
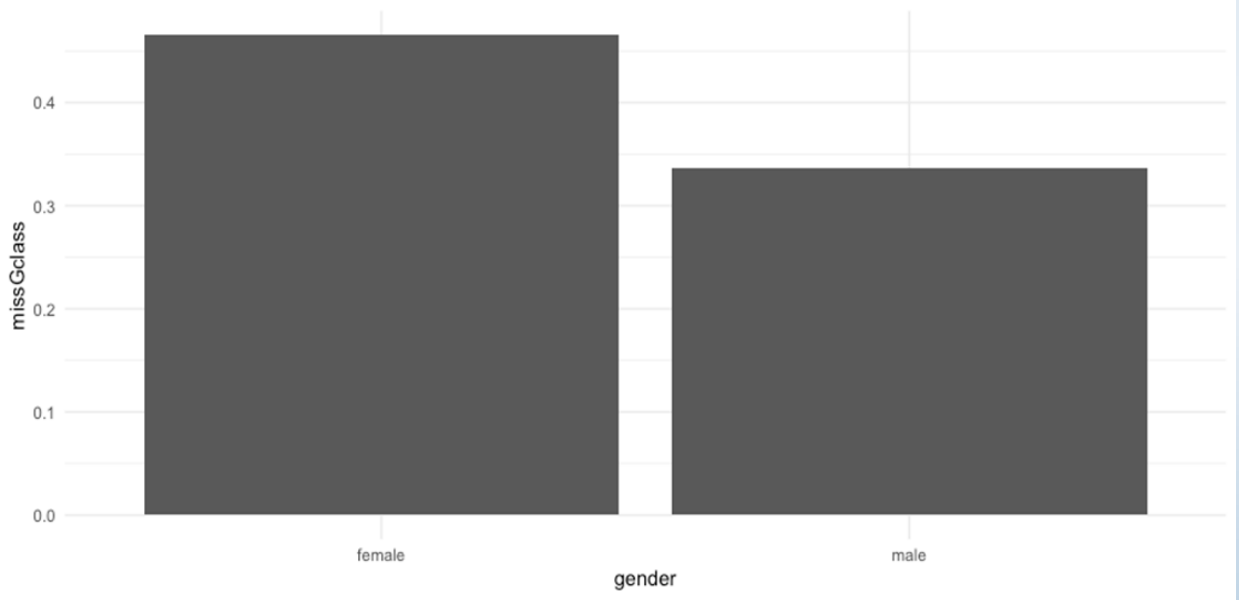
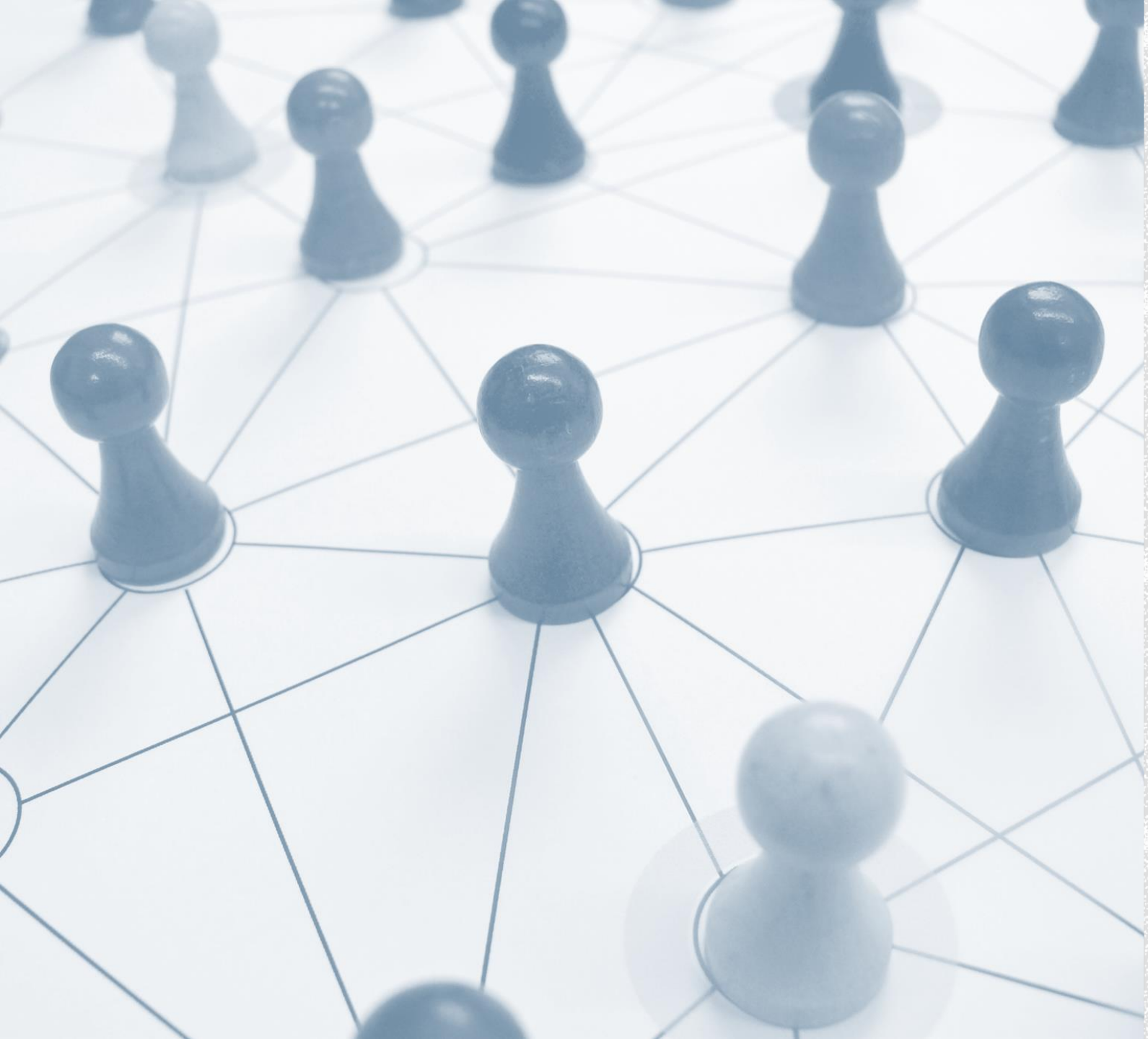


Table 7: Gender Misclassification Rate



- There is a significantly higher chance of being misclassified as under $50,000 than over $50,000 given being a female
- The difference between the genders Type II misclassification rate is about 10%
- This proves bias results based on race

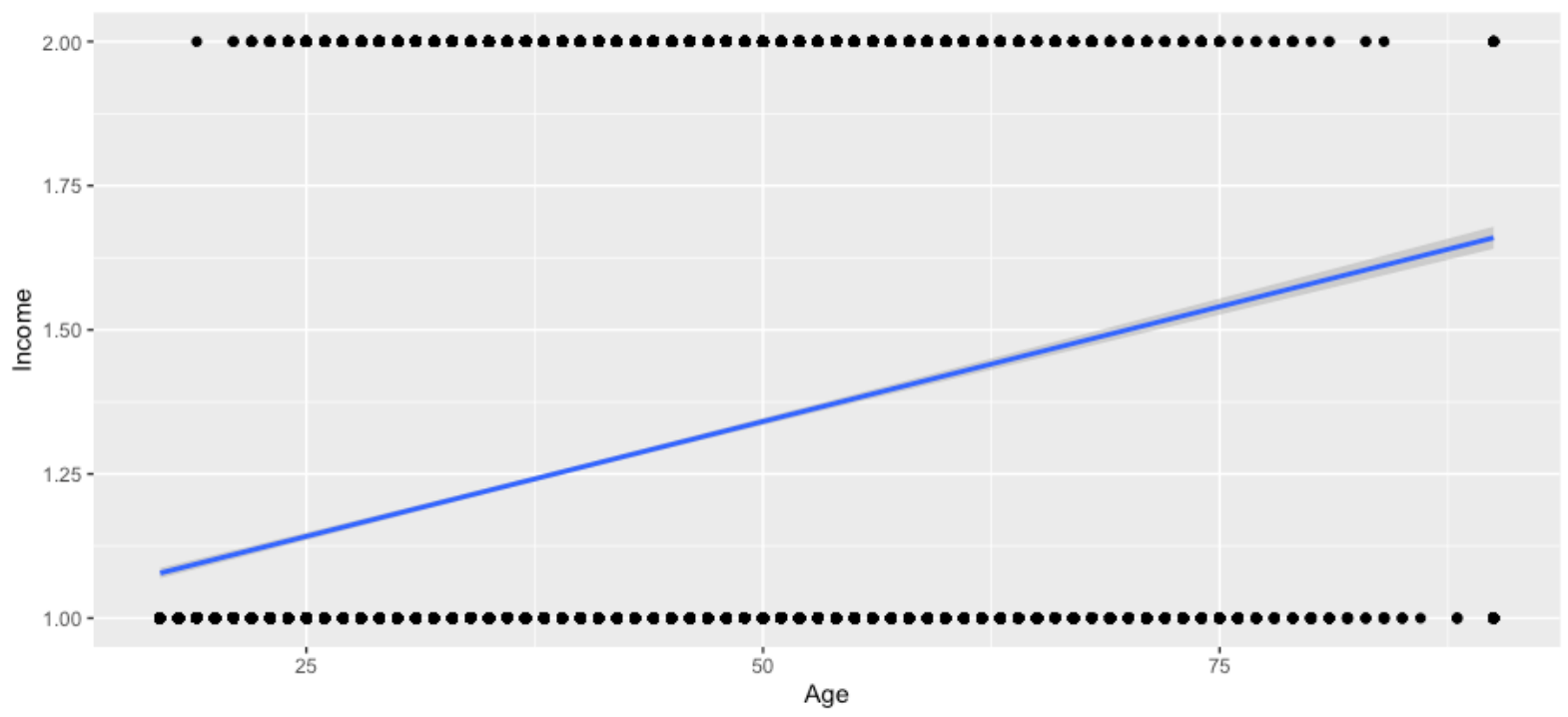


Correlations

- Created a data frame with income and the three variables with the highest Mean Decrease in Gini from our Random Forest outputs. We also added race.
- Interestingly, education was not ranked in the four top values in terms of importance.
- Pearson Product Moment Correlation
  - default correlation function in R.
  - Values for *r* between +1 and -1 (for example, *r* = 0.8 or -0.4) indicate that there is variation around the line of best fit.
- Biserial Correlation
  - used to measure the strength and direction of the association that exists between one continuous variable and one dichotomous variable
  - Since the Pearson's product-moment Correlation (cor and cor.test) produces the same output as running a Point Biserial Correlation (biserial.cor), we continued with the Pearson’s test.

Correlations, cont.

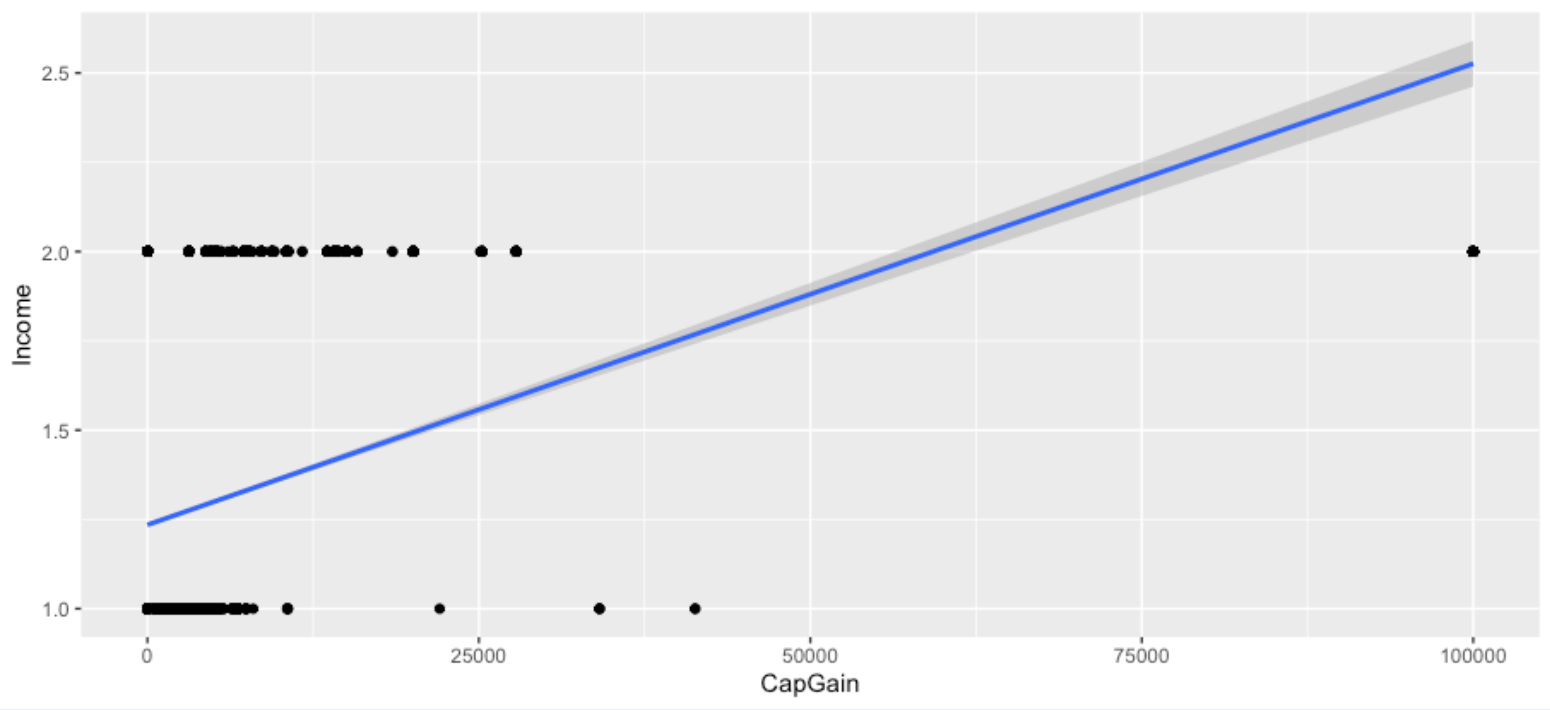
Table 8: Pearson's Product-moment Correlation Coefficient Trend Line (Income and Age)



- Intuitive
- One of the two top attributes
- Pearson's product-moment correlation 0.2419981

Correlations, cont.

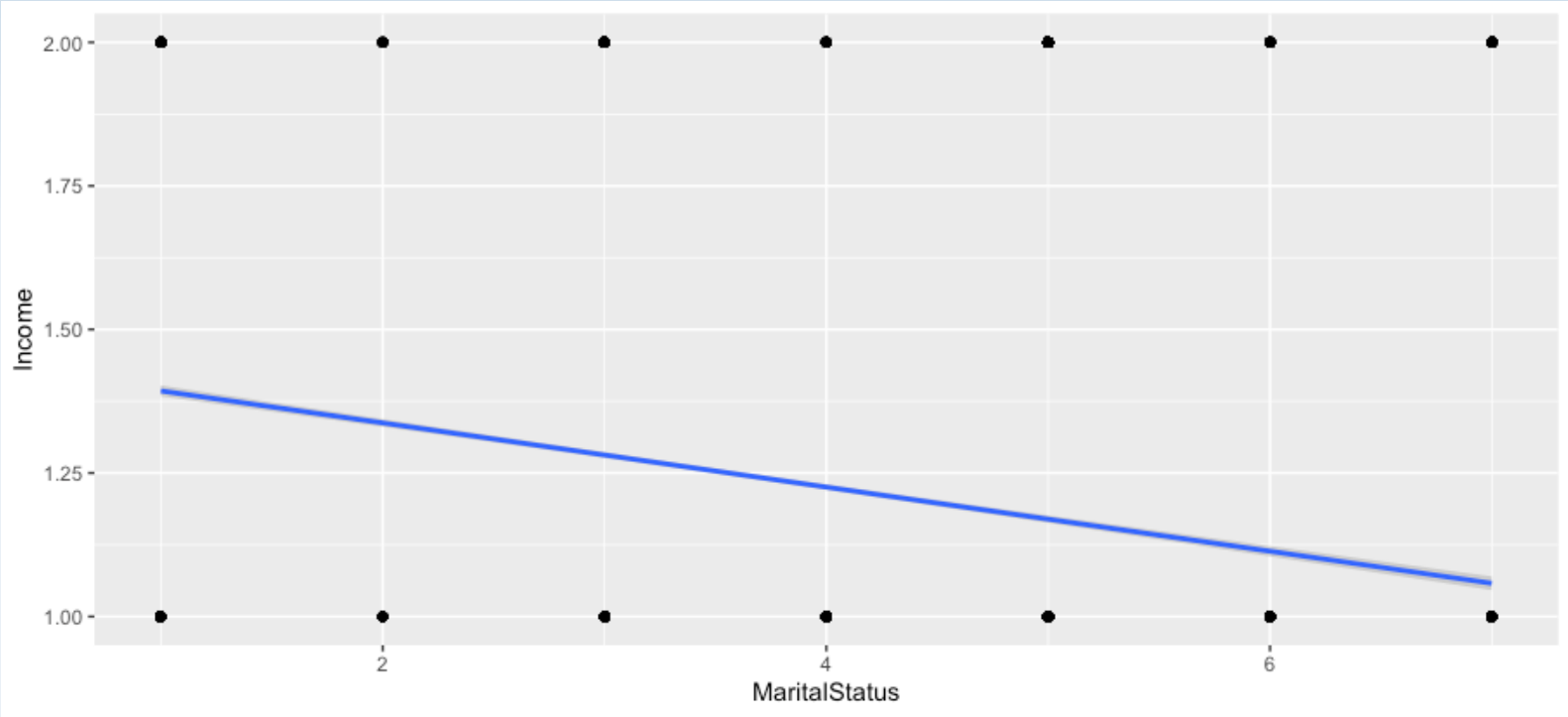
Table 9: Pearson's Product-moment Correlation Coefficient Trend Line (Income and Capital Gains)



- Intuitive
- One of two top attributes
- Pearson's product-moment correlation 0.2211962

Correlations, cont.

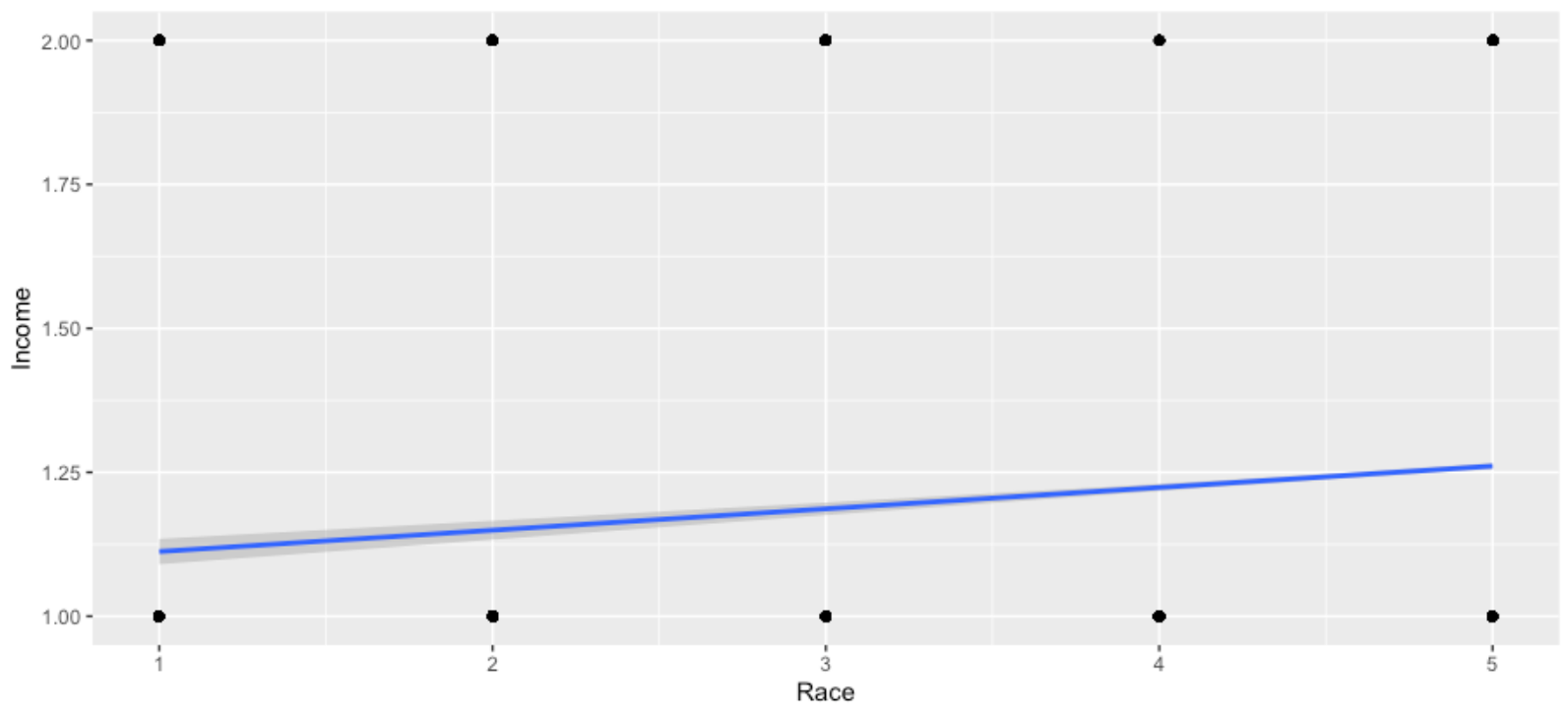
Table 10: Pearson's Product-moment Correlation Coefficient Trend Line (Income and Marital Status)



- 1= Divorced, 2= Married-AF-spouse, 3= Married-civ-spouse, 4= Married-spouse-absent, 5= Never-married, 6= Separated, 7= Widowed
- Non-intuitive
- Would have expected Married or Never-Married to be highest.
- Pearson's product-moment correlation -0.1935184

Correlations, cont.

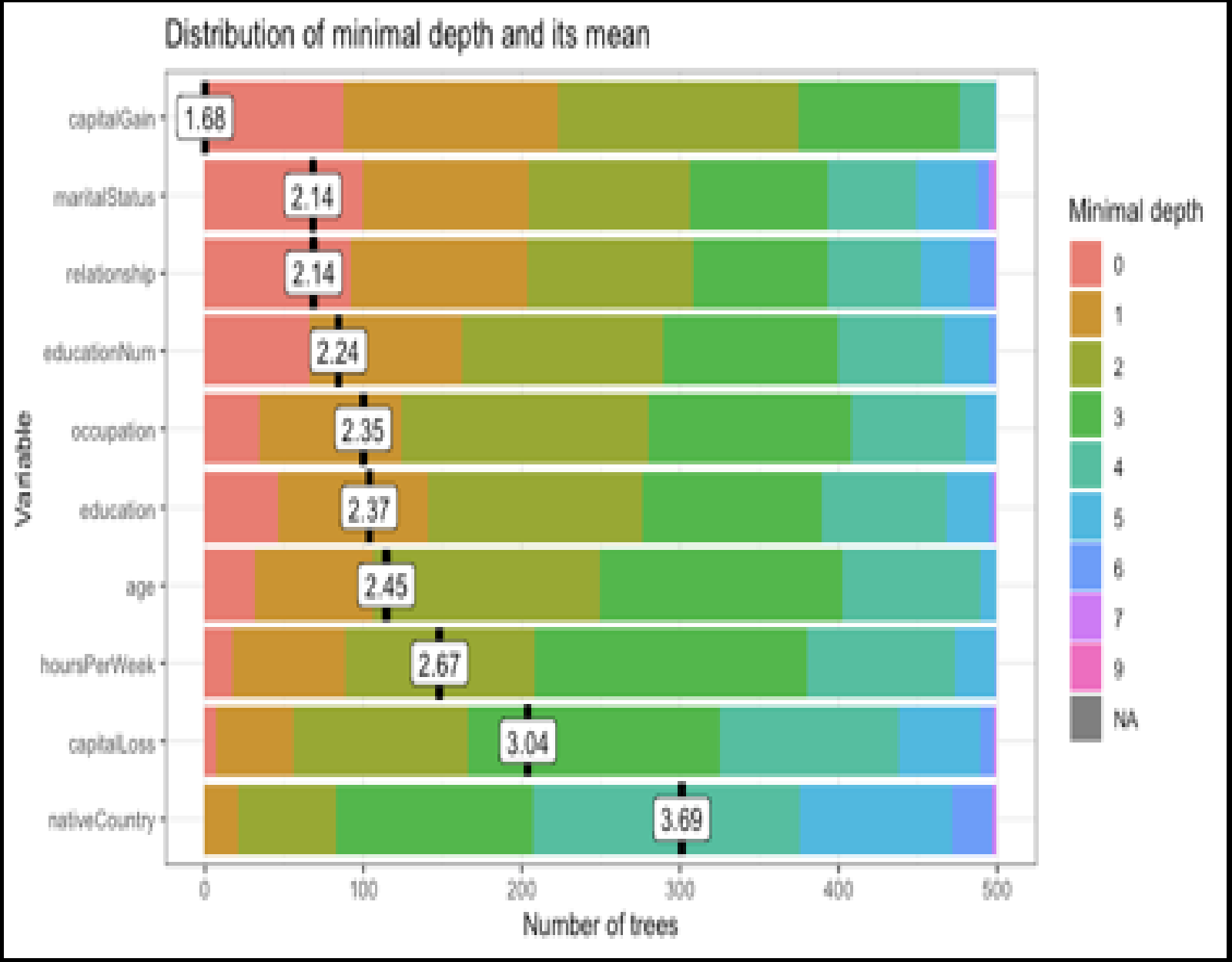
Table 11: Pearson's Product-moment Correlation Coefficient Trend Line (Income and Race)



- Income vs. Race
- Intuitive
- 1= Amer-Indian-Eskimo, 2= Asian-Pac-Islander, 3= Black, 4= Other, 5= White
- Very small Pearson's coefficient 0.0716576
- Not as significant as Age and Capital Gains

Random Forest Explainer

Table 12:Random Forest Explainer




```
#devtools::install_github("MI2DataLab/randomForestExplainer")

install.packages("randomForestExplainer")
```

The minimal depth is how deep into a tree the parameter is used.

Random Forest Explainer, cont.

Table 13: Importance Ranking Plots

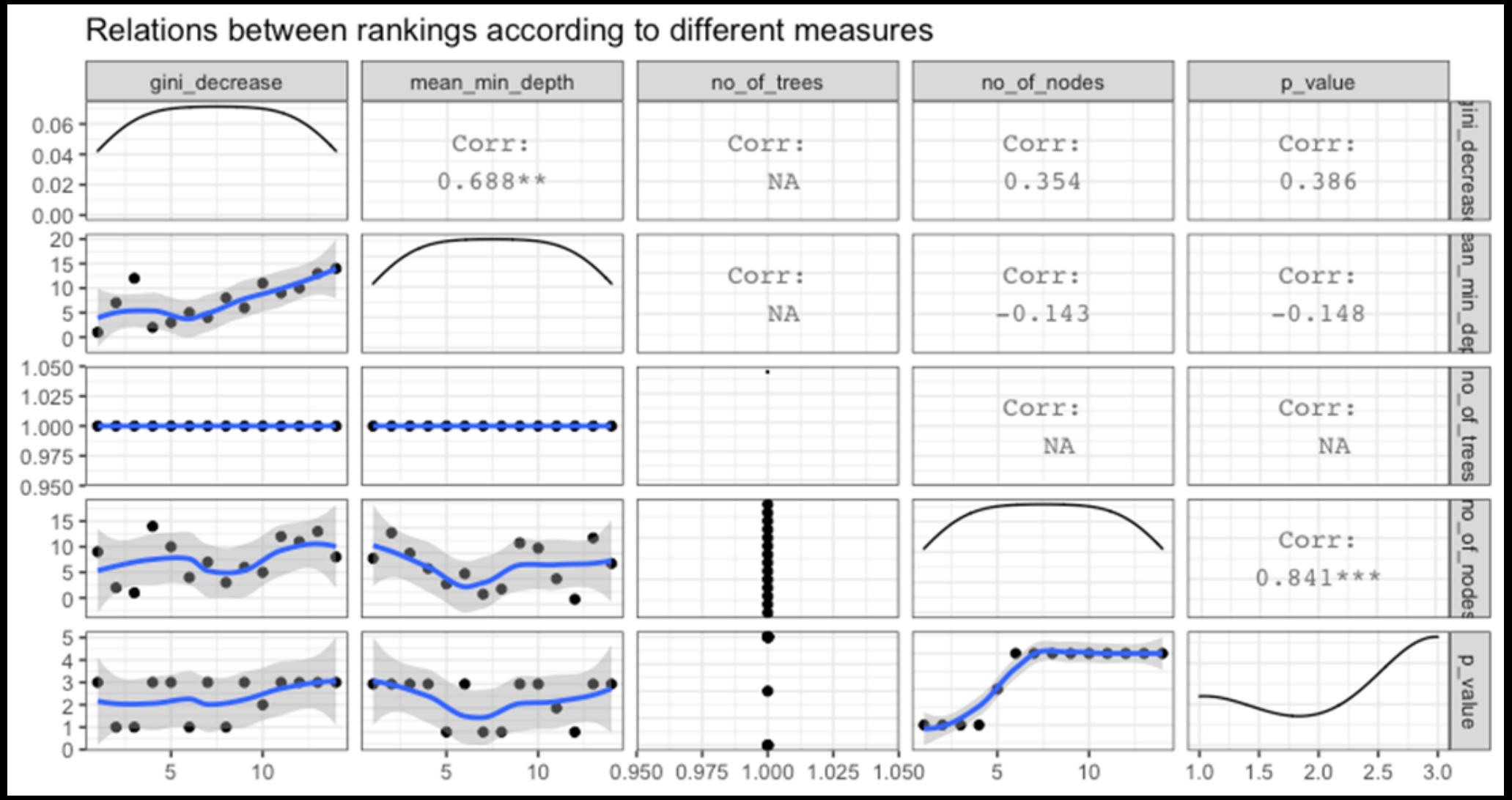
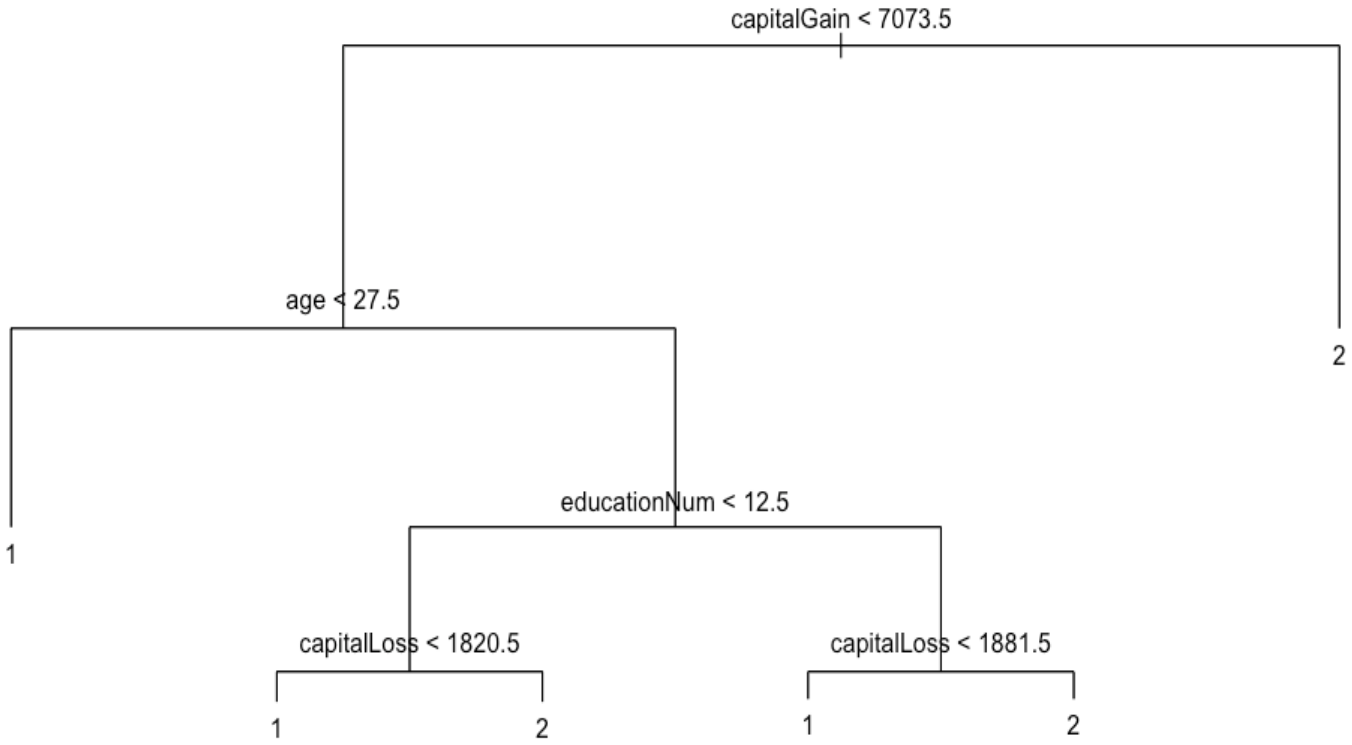
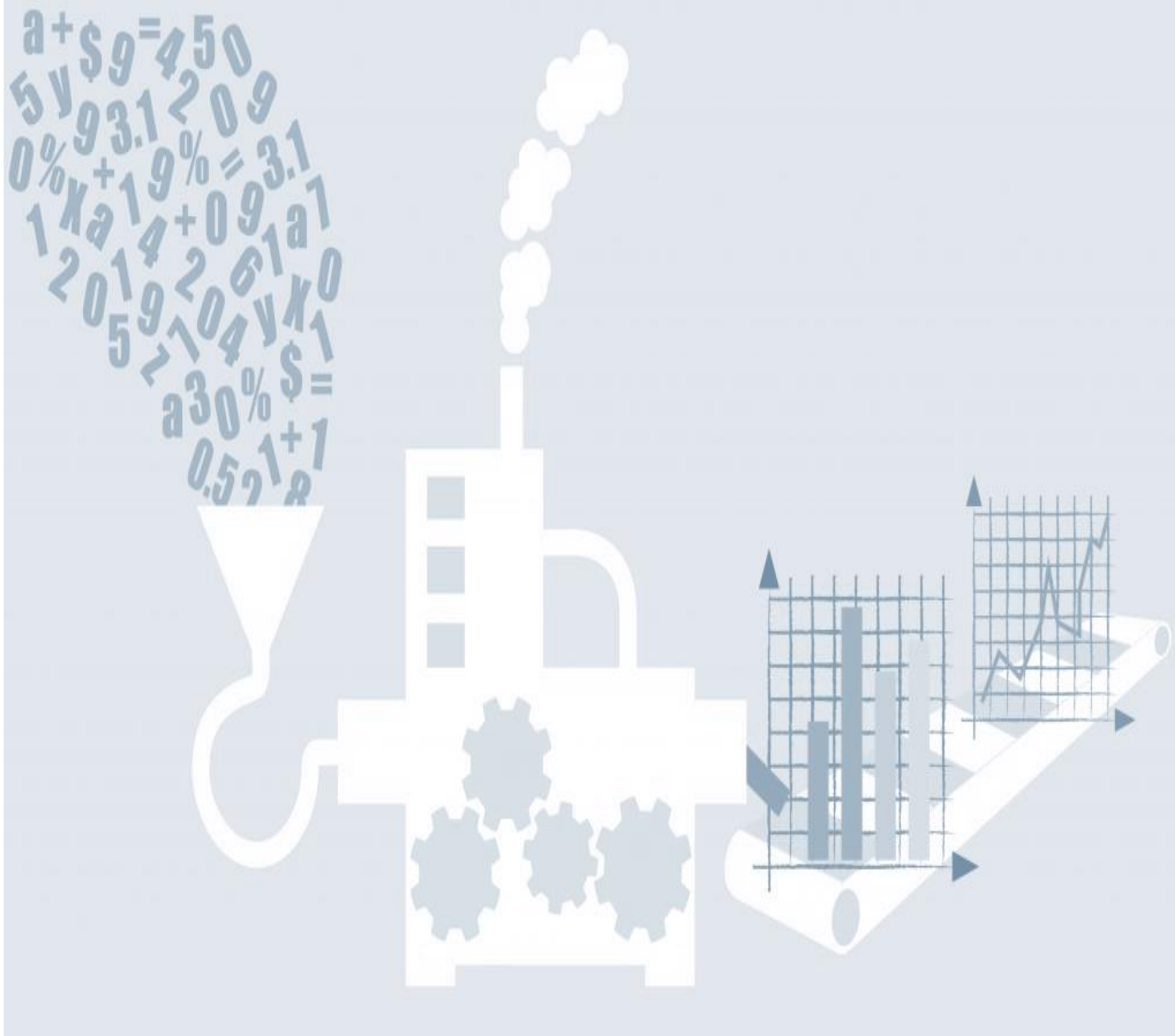


Table 15: Decision Tree with Integers



Decision Trees

- This decision tree has a error rate of 19% or a classification rate of roughly 81%
- Although our Random Forest performed at closer to 87% this decision tree is extremely intuitive and easy to visualize. It is also the foundation to how a Random Forest works



Finding the Answers

- The data tells us there is bias
- We could continue to work with this data to find more information
  - Ex. k fold cross validation
- Many attributes are provided however, we focused on race and gender
- Found both negative and positive correlations

How is this knowledge useful?

- This information would serve lending institutions by revealing areas with bias that have a negative effect on their business or a negative effect on their choice of loanee
  - A person being misclassified as meeting the approval criteria in error, increases the risk on return of the monies loaned out
  - A person being misclassified as not meeting the approval criteria would be a lost customer therefore, lost revenue
  - If a banking institution uses machine learning to approve or deny customers, then they are responsible for the disproportionate amount of people that will be wrongfully denied based on race or gender

Questions?