Determining Factors that Affect Income in the United States

by
Kevin Rodriguez, Abraham Caceres, Avina Patel, Pankti Sheth

Overview

The data that is being analyzed is census information from 1994 about whether or not income is above $50,000. There are a total of 14 variables, eight of which are qualitative, while the other six quantitative.

Our goal is to determine whether a person is able to make $50,000 or more based on these variables. We will use logistic regression, decision trees, random forest method, and stepwise selection to compare the variables, as well as doing an exploratory analysis of the data.

Quantitative variables

Age in years

Final sampling weight

Education in years

Work hours per week

Capital gains

Capital loss

Qualitative variables

Income

Workclass

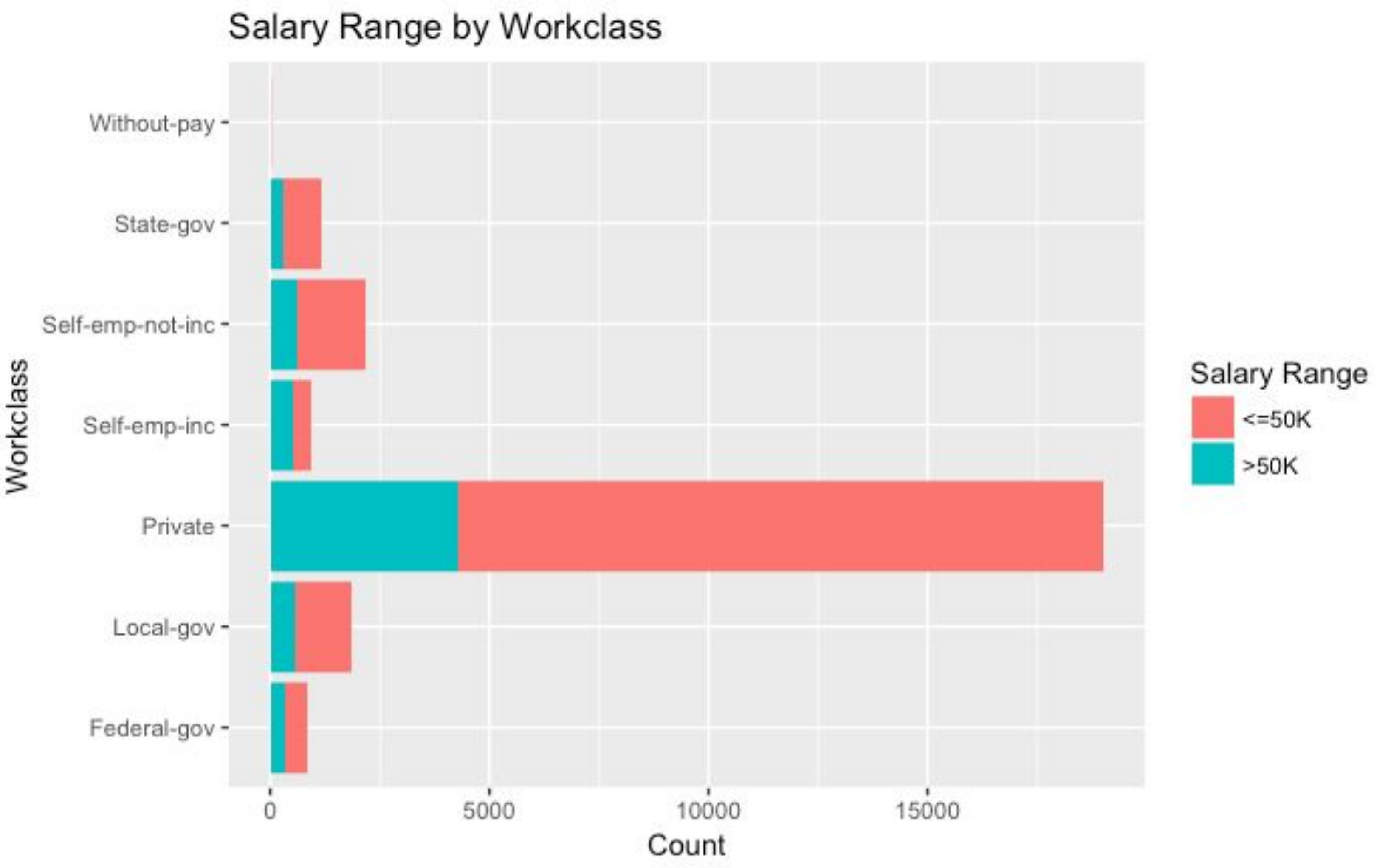
Education

Marital status

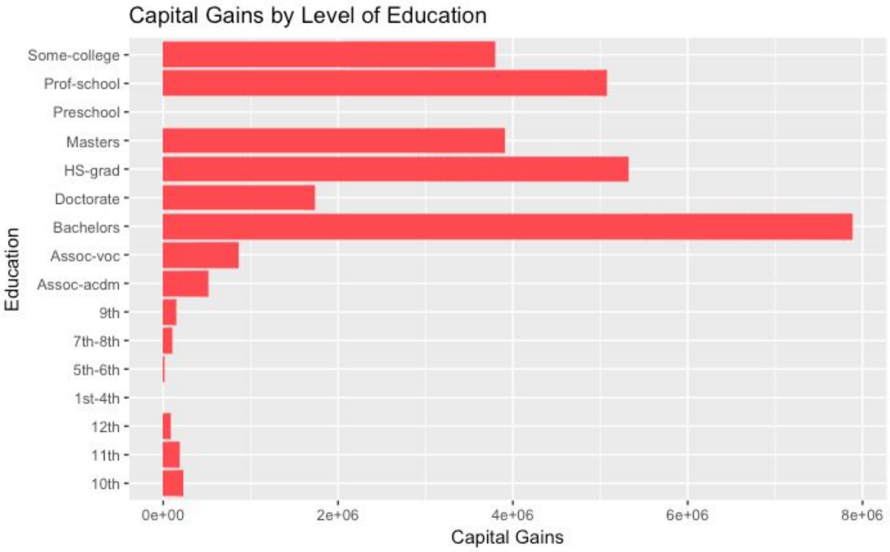
Gender

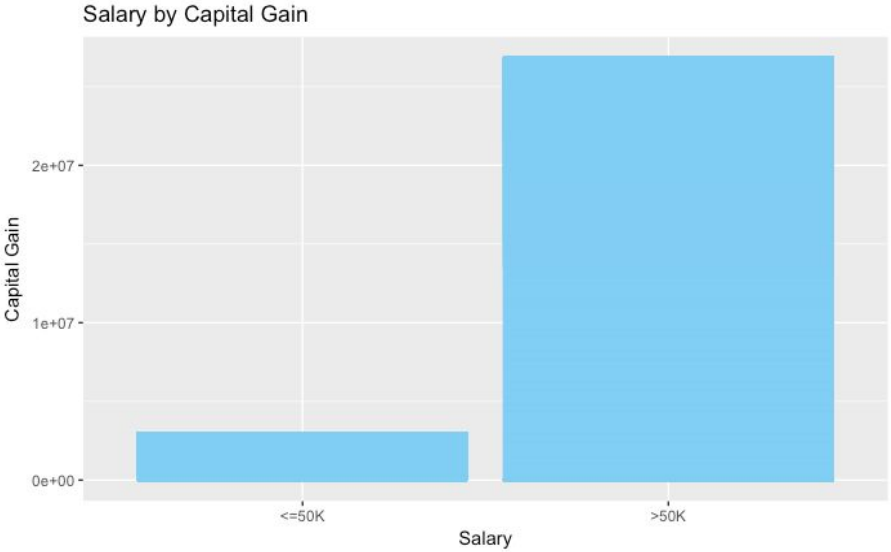
Race

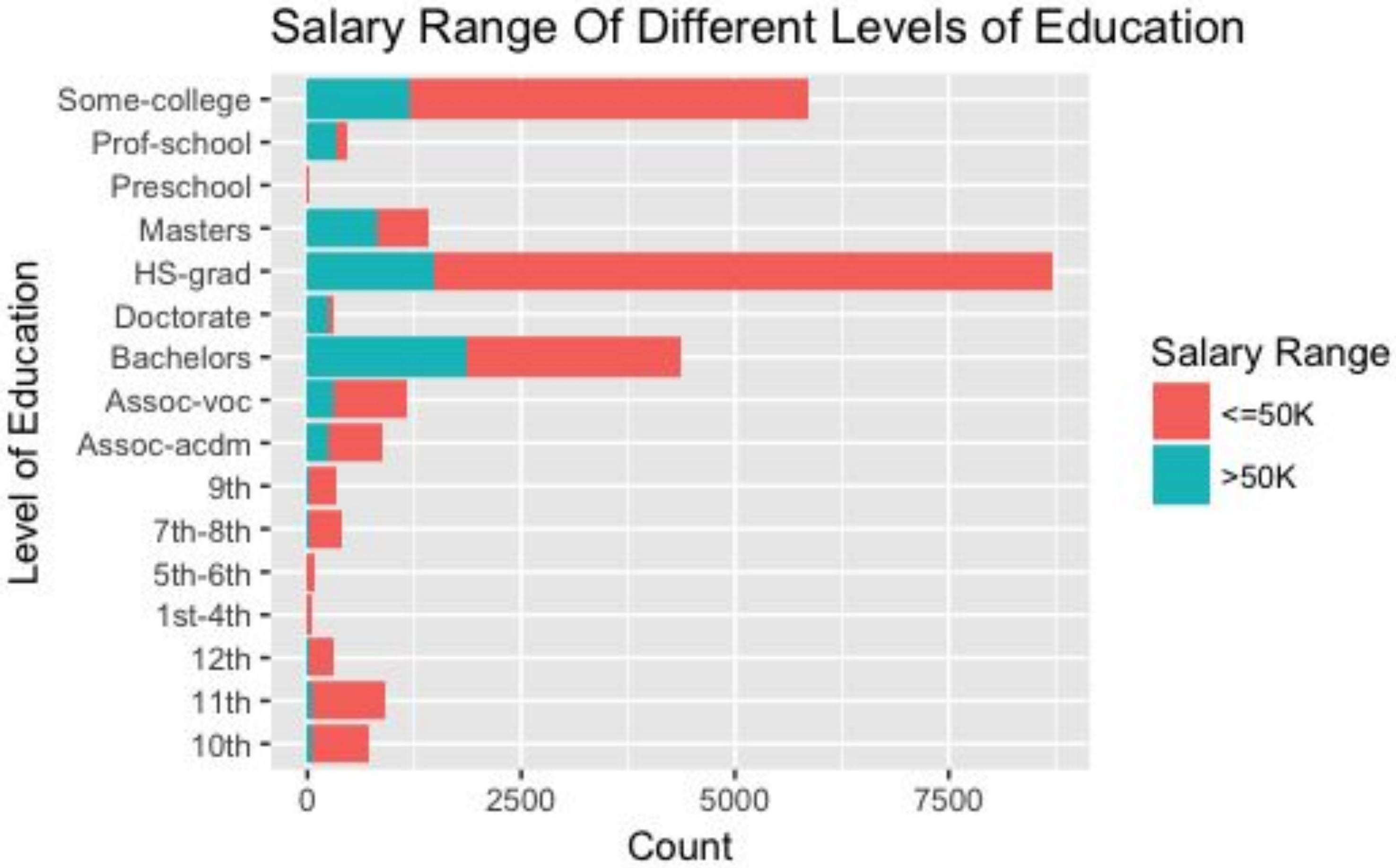
Occupation

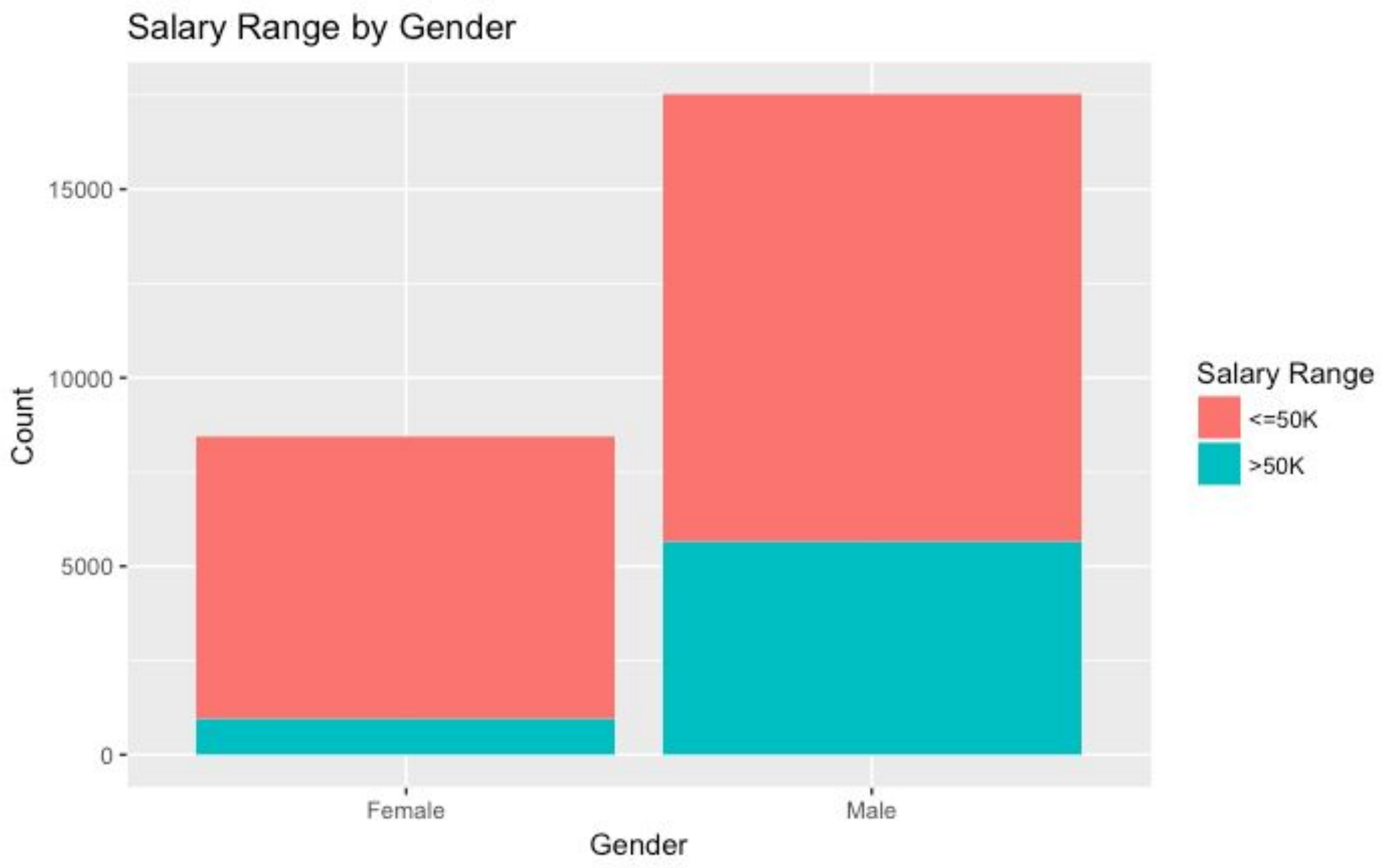


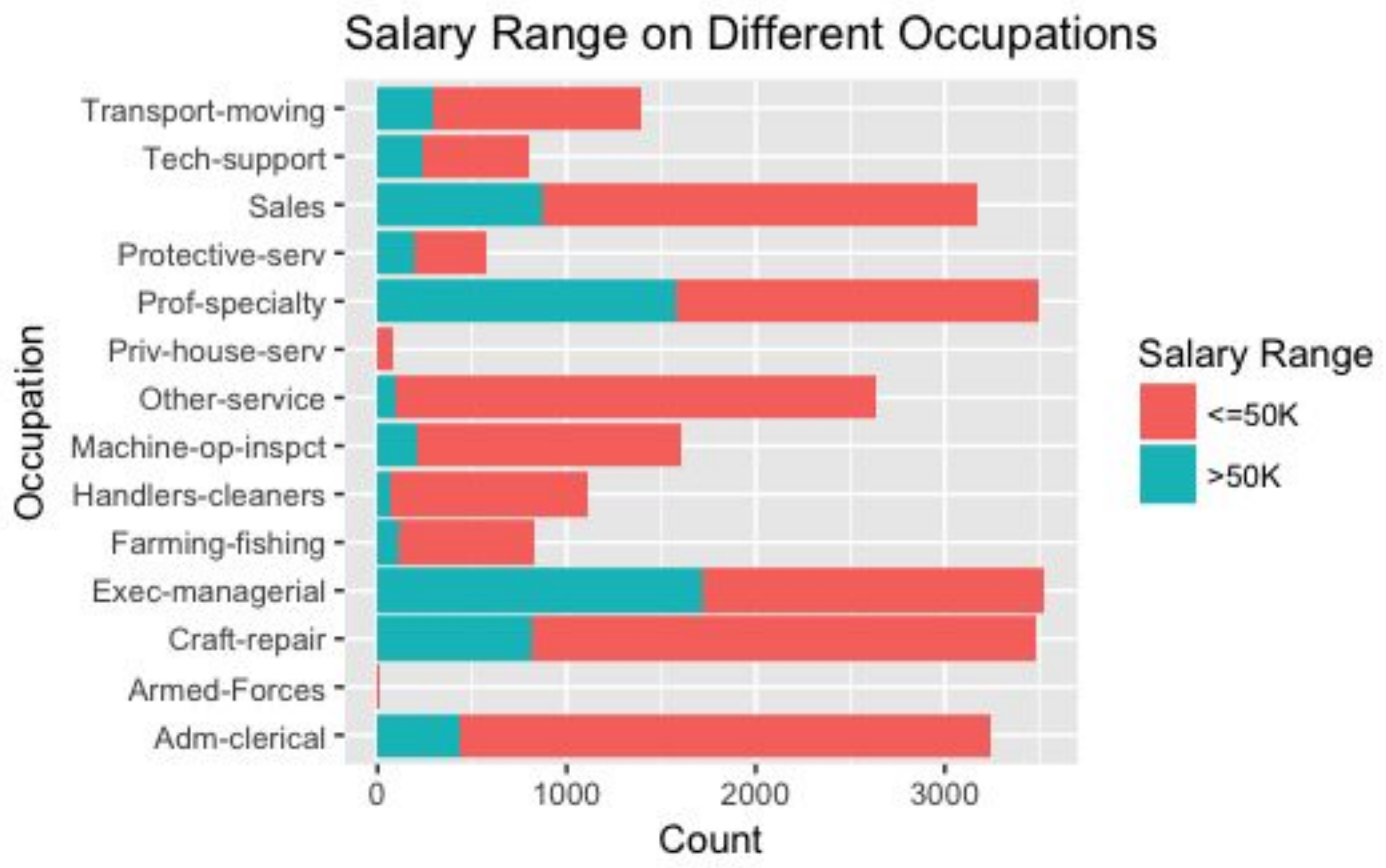
Salary Range by Workclass
Without-pay
State-gov
Self-emp-not-inc
Self-emp-inc
Private
Local-gov
Federal-gov
Workclass
0
5000
10000
15000
Count
Salary Range
<=50K
>50K


Capital Gains by Level of Education
Some-college
Prof-school
Preschool
Masters
HS-grad
Doctorate
Bachelors
Assoc-voc
Assoc-acdm
9th
7th-8th
5th-6th
1st-4th
12th
11th
10th
Education
0e+00
2e+06
4e+06
6e+06
8e+06
Capital Gains

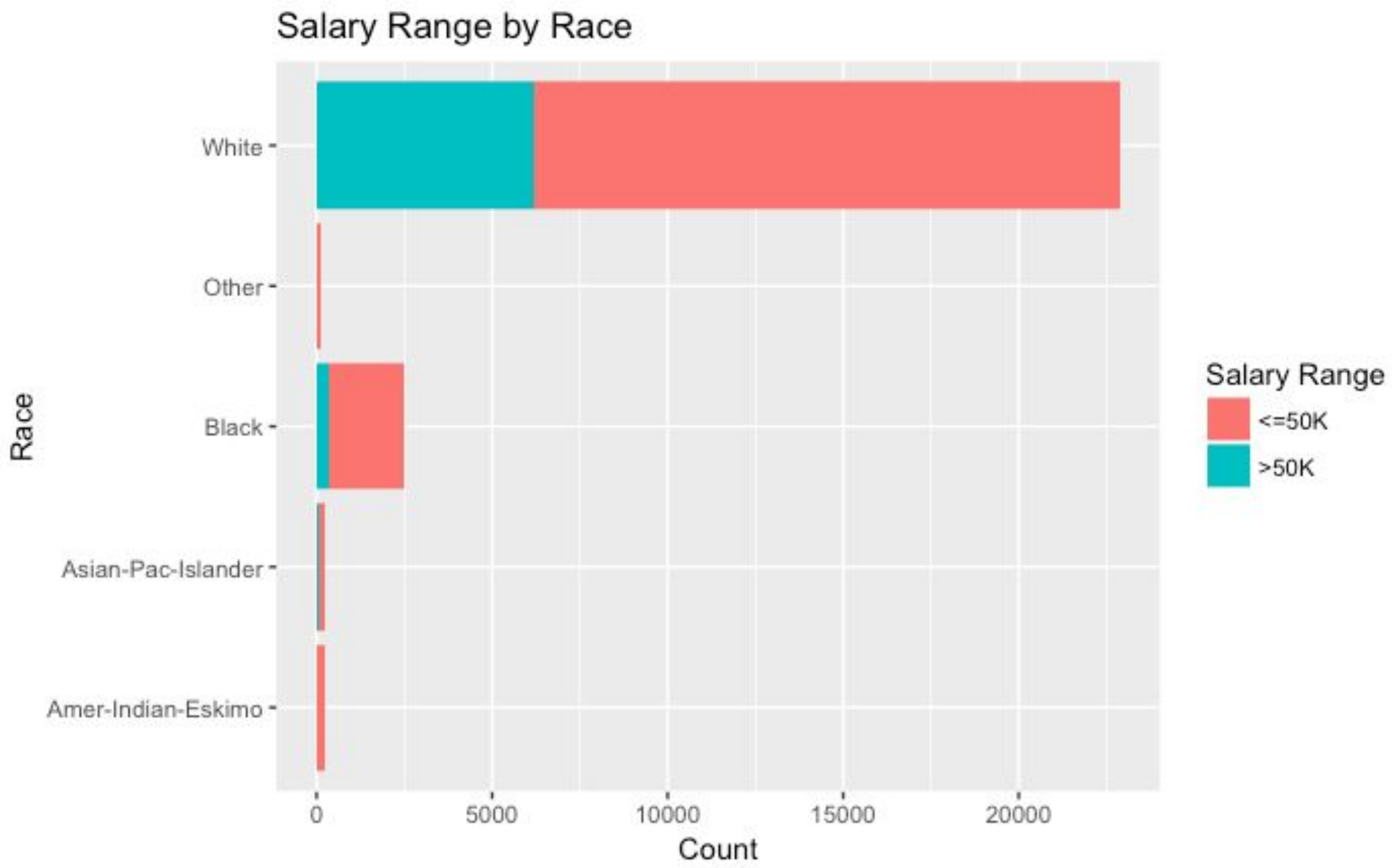

Salary by Capital Gain
2e+07
1e+07
0e+00
Capital Gain
<=50K
>50K
Salary


Salary Range Of Different Levels of Education
Some-college
Prof-school
Preschool
Masters
HS-grad
Doctorate
Bachelors
Assoc-voc
Assoc-acdm
9th
7th-8th
5th-6th
1st-4th
12th
11th
10th
Level of Education
0
2500
5000
7500
Count
Salary Range
<=50K
>50K

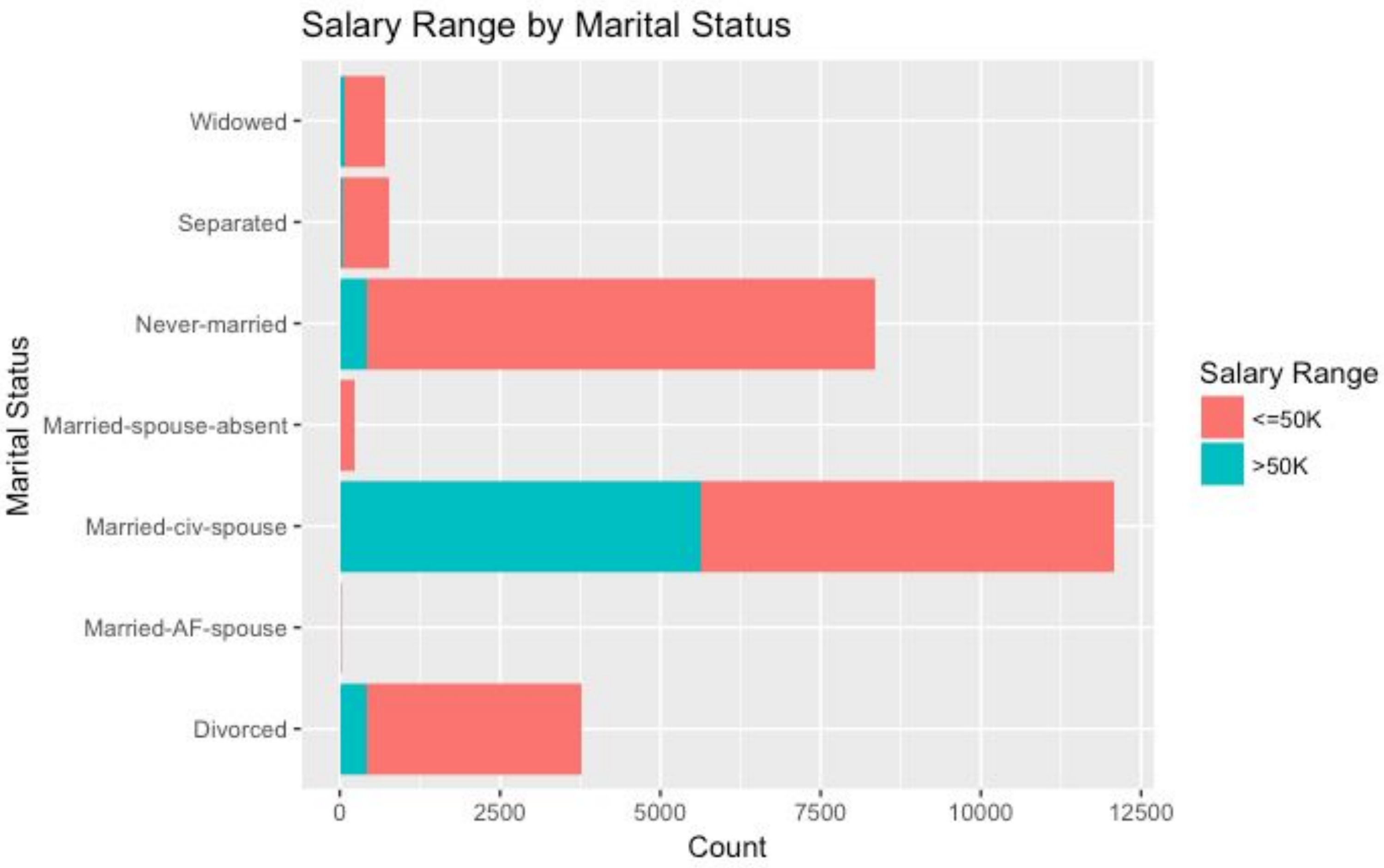

Salary Range by Gender
15000
10000
5000
0
Count
Female
Male
Gender
Salary Range
<=50K
>50K

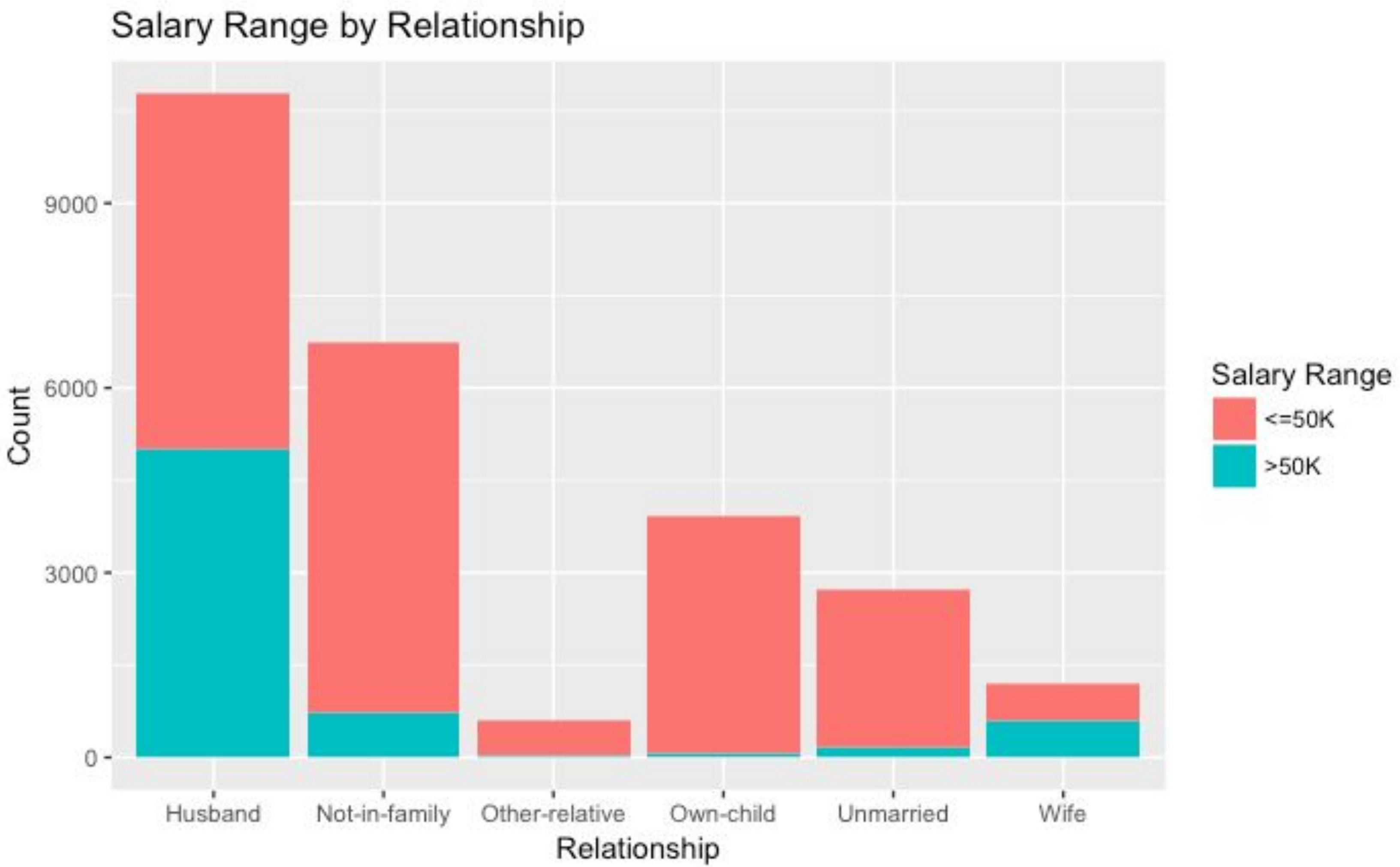


Salary Range on Different Occupations
Transport-moving
Tech-support
Sales
Protective-serv
Prof-specialty
Priv-house-serv
Other-service
Machine-op-inspct
Handlers-cleaners
Farming-fishing
Exec-managerial
Craft-repair
Armed-Forces
Adm-clerical
Occupation
0
1000
2000
3000
Count
Salary Range
<=50K
>50K

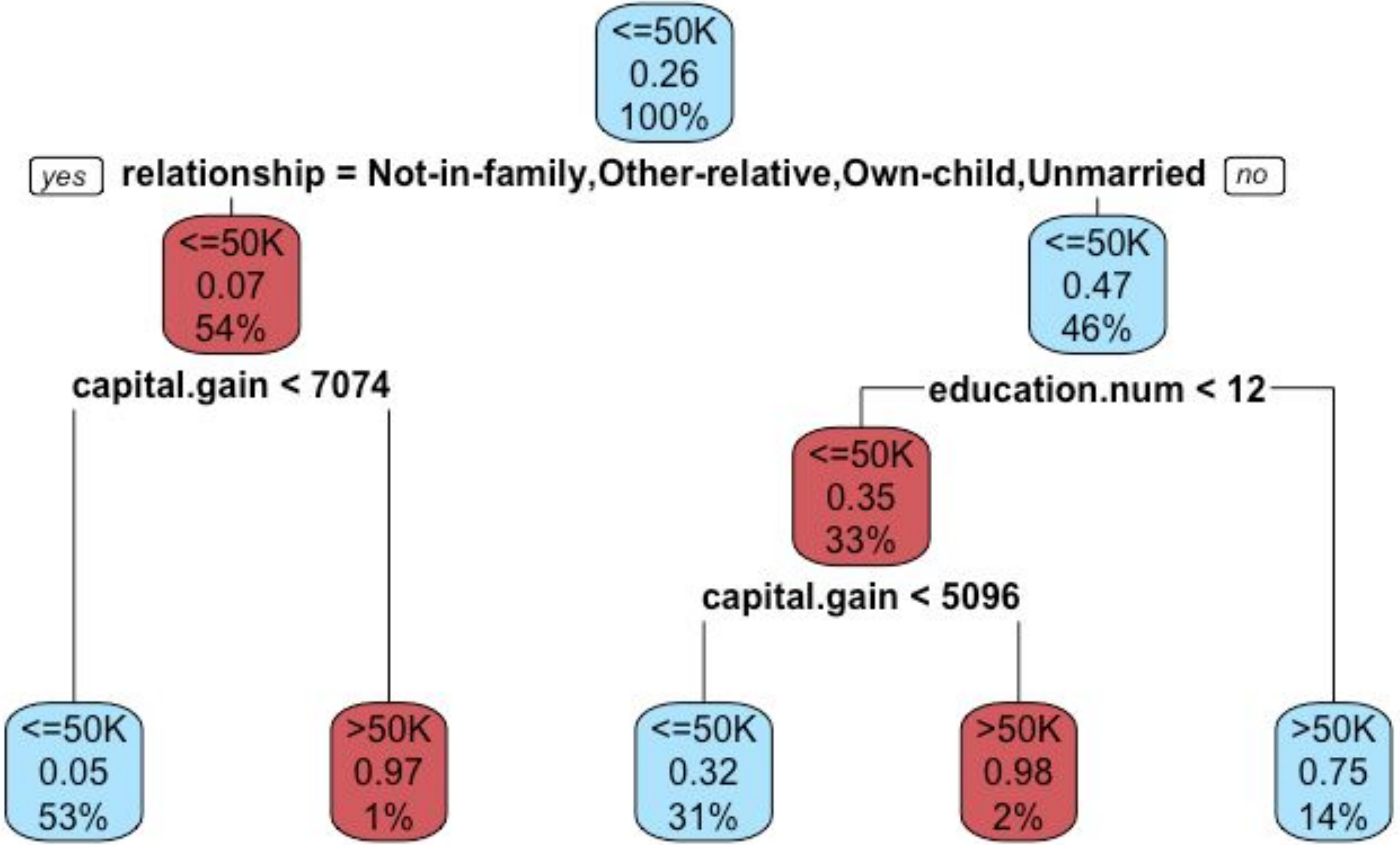


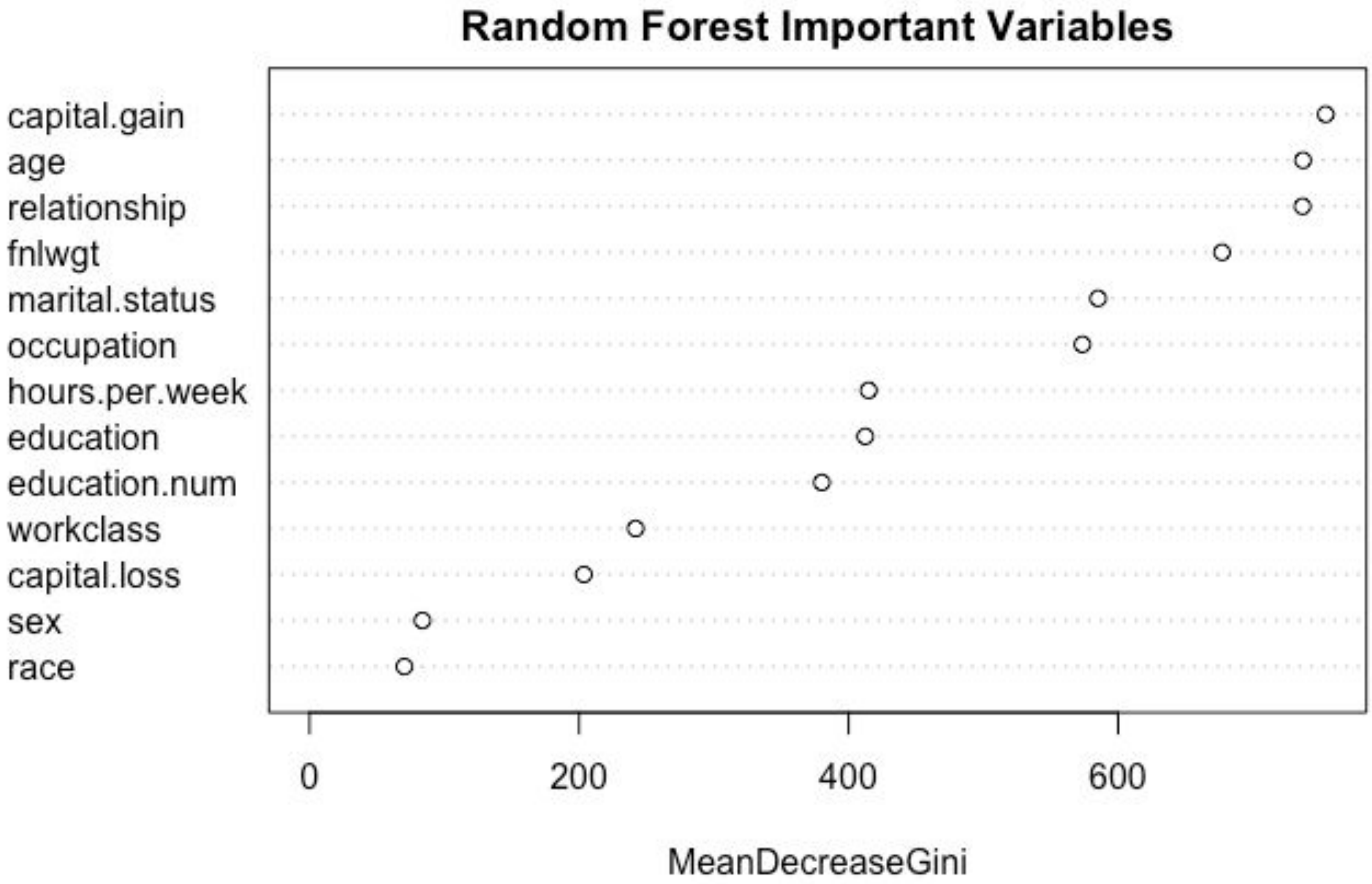
Salary Range by Race
White
Other
Black
Asian-Pac-Islander
Amer-Indian-Eskimo
Race
0
5000
10000
15000
20000
Count
Salary Range
<=50K
>50K


Salary Range by Marital Status
Widowed
Separated
Never-married
Married-spouse-absent
Married-civ-spouse
Married-AF-spouse
Divorced
Marital Status
0
2500
5000
7500
10000
12500
Count
Salary Range
<=50K
>50K



Salary Range by Relationship
9000
6000
3000
0
Count
Husband
Not-in-family
Other-relative
Own-child
Unmarried
Wife
Relationship
Salary Range
<=50K
>50K


<=50K
0.26
100%
yes
relationship = Not-in-family,Other-relative,Own-child,Unmarried
no
<=50K
0.07
54%
<=50K
0.47
46%
capital.gain < 7074
education.num < 12
<=50K
0.35
33%
capital.gain < 5096
<=50K
0.05
53%
>50K
0.97
1%
<=50K
0.32
31%
>50K
0.98
2%
>50K
0.75
14%


Random Forest Important Variables
capital.gain
age
relationship
fnlwgt
marital.status
occupation
hours.per.week
education
education.num
workclass
capital.loss
sex
race
0
200
400
600
MeanDecreaseGini

Logistic Regression

```
null.model <- glm(income ~ 1, family = binomial(link = "logit"),  data
= train )

full.model <- glm(income ~ age + workclass + fnlwgt + education +
marital.status +capital.gain + capital.loss+ hours.per.week +
occupation + relationship + race + sex,family = binomial(link =
"logit"),  data = train )

step(null.model, direction = "forward", scope = formula(full.model), k =
2, steps = 3 )
```

Final Model from the step function

```
glm(formula = income ~ relationship + education + capital.gain,
    family = binomial(link = "logit"), data = train)
```

Based on the step function the three most important variables based on the AIC values are the variables: relationship, education and capital.gain

```
Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -1.755e+00  1.884e-01  -9.318  < 2e-16 ***
relationshipNot-in-family  -2.374e+00  6.219e-02 -38.170  < 2e-16 ***
relationshipOther-relative -2.723e+00  2.547e-01 -10.689  < 2e-16 ***
relationshipOwn-child      -3.833e+00  1.592e-01 -24.071  < 2e-16 ***
relationshipUnmarried      -2.637e+00  1.094e-01 -24.101  < 2e-16 ***
relationshipWife            1.665e-01  8.189e-02   2.033    0.042 *
education11th              -5.428e-02  2.630e-01  -0.206    0.836
education12th               4.766e-01  3.382e-01   1.409    0.159
education1st-4th           -5.219e-01  1.059e+00  -0.493    0.622
education5th-6th           -8.850e-01  7.555e-01  -1.171    0.241
```

education7th-8th	-2.676e-01	2.930e-01	-0.913	0.361	
education9th	-4.629e-01	3.581e-01	-1.293	0.196	
educationAssoc-acdm	1.464e+00	2.191e-01	6.683	2.35e-11	***
educationAssoc-voc	1.469e+00	2.097e-01	7.004	2.49e-12	***
educationBachelors	2.348e+00	1.943e-01	12.083	< 2e-16	***
educationDoctorate	3.773e+00	2.740e-01	13.769	< 2e-16	***
educationHS-grad	8.344e-01	1.921e-01	4.345	1.40e-05	***
educationMasters	2.837e+00	2.058e-01	13.788	< 2e-16	***
educationPreschool	-8.045e+00	1.298e+02	-0.062	0.951	
educationProf-school	3.330e+00	2.470e-01	13.485	< 2e-16	***
educationSome-college	1.284e+00	1.938e-01	6.628	3.39e-11	***
capital.gain	3.011e-04	1.249e-05	24.100	< 2e-16	***

Goodness of Fit

Analysis of Deviance Table

Model 1: income ~ age + workclass + fnlwgt + education + marital.status +
capital.gain + capital.loss + hours.per.week + occupation +
relationship + race + sex

Model 2: income ~ relationship + education + capital.gain

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)	
1	18117	11986				
2	18151	13239	-34	-1253.5	< 2.2e-16	***

Based on the anova with the full model and the reduced model, the p-value is less than 2.2e-16 which is significantly less than .05. Therefore, our reduced model is a better fit compared to the full model.

Multicollinearity using Variance Inflation Factor

Multicollinearity corresponds to a situation where the data contain highly correlated predictor variables.

A variance inflation factor (*VIF*) quantifies how much the variance is inflated.

As a rule of thumb, a VIF value that exceeds 5 or 10 indicates a problematic amount of collinearity. In our example, there is no collinearity.

relationship	education	capital.gain
1.093182	1.085869	1.012079

Receiver Operating Characteristic (ROC) Curve and (AUROCC)

The ROC curve shows the relationship between sensitivity and specificity.

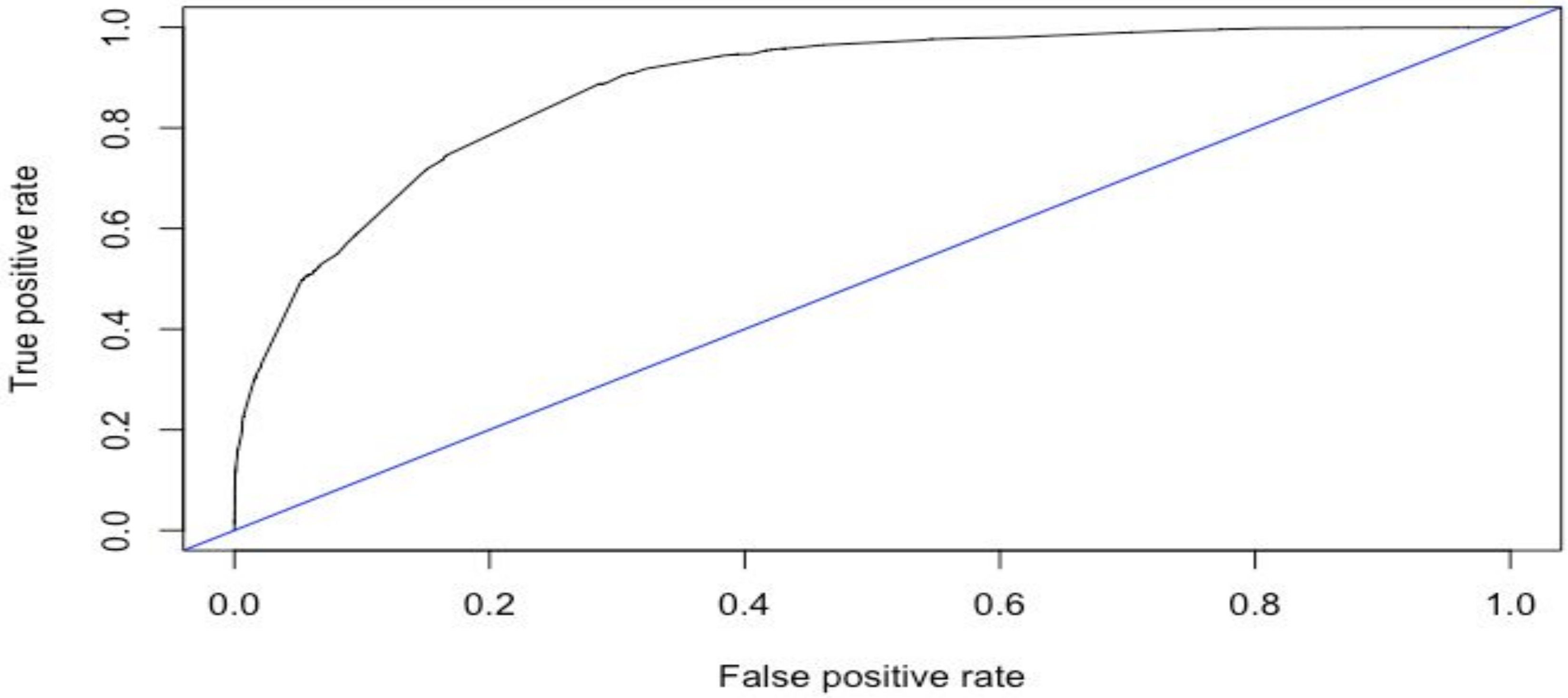
It can be also used to test accuracy; the closer the graph is to the top and left-hand borders, the more accurate the test.

Likewise, the closer the graph to the diagonal, the less accurate the test.

A perfect test has an area under the ROC curve (AUROCC) of 1. The diagonal line in a ROC curve represents perfect chance. In other words, a test that follows the diagonal has no better odds of detecting something than a random flip of a coin. The area under the diagonal is .5. Therefore, a useless test (one that has no better odds than chance alone) has a AUROCC of .5.

In our graph; the area under the ROC curve .88

Continuation



Conclusion

In conclusion, our three most significant variables using a logistic regression are relationship, education and capital gain which is similar to variables of the decision tree. When applying a logistic regression and using the ROC curve. Our final model predicts 88% of the observations from the testing dataset.

Using the decision tree, the accuracy we have 84% accuracy.

When performing a random forest, only two of our three variables are the same from the logistic regression and decision tree. The three variables are relationship, capital gain and age. However, the accuracy is around 86%