CYO: Adult Income Predictions

Jesse Reid

9/13/2019

Overview

For this project, I chose to take a look at the Adult Census Income dataset from Kaggle (https://www.kaggle.com/uciml/adult-census-income). I used different models which incorporated several different variables to predict whether a persons incomes where above or below 50,000 annually.

Analysis

The models I used to predict incomes include the below:

1. Logistic Regression
2. Random Forest
3. Boosted Random Forest using Bernoulli Distributions

Data Cleaning and Creating Functions

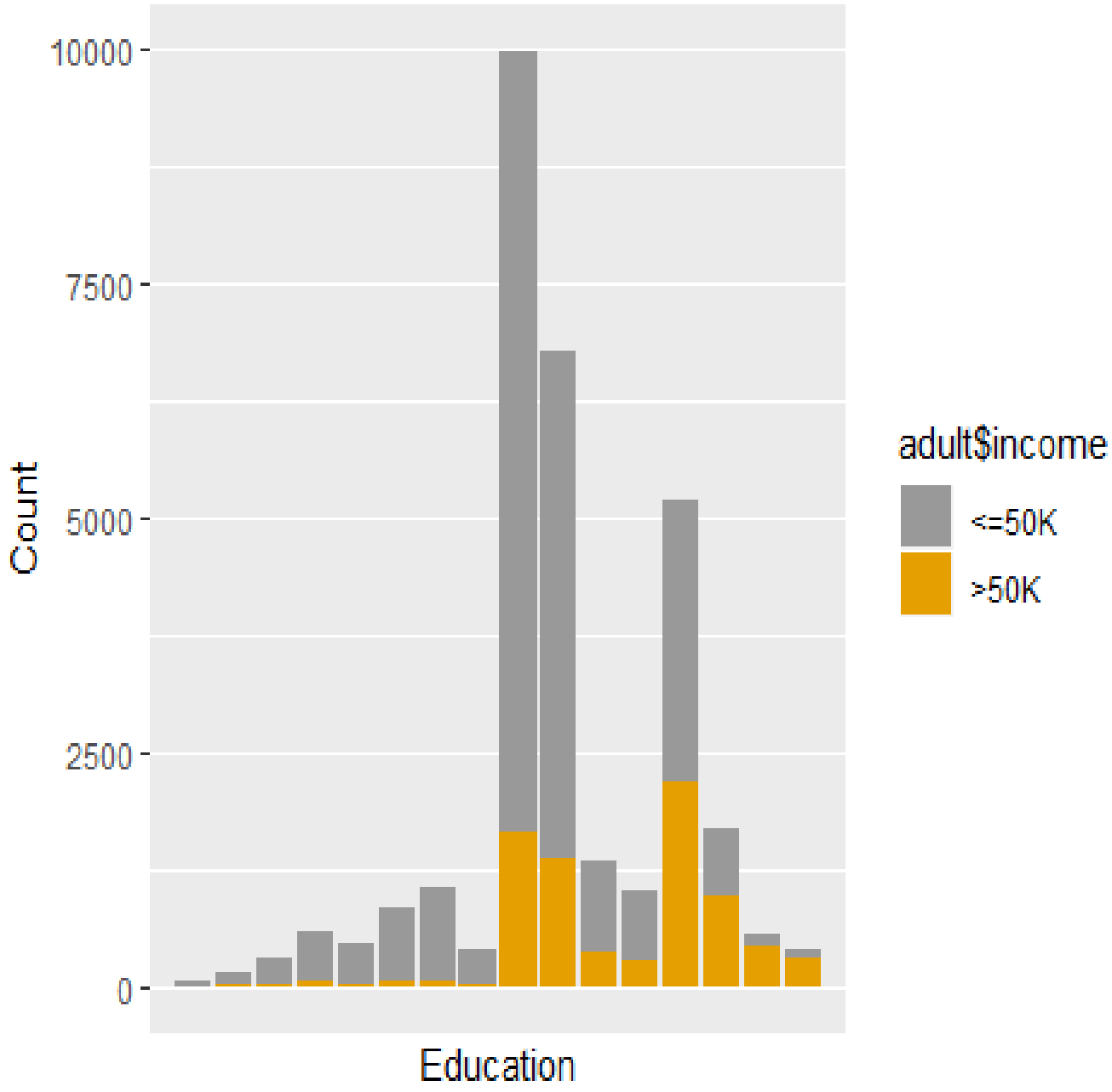
Before I used the data in my models, it was neccessary to do some cleaning. First, I combined Local and State government jobs into a category called SL-gov as well as combined all self-employed jobs into a category called self-emp. I then updated the Marital Status column to display either 'Married', 'Not Married', or 'Never Married'. Next, I grouped the countries into regions and displayed them in the native.country column. Lastly, I turned any missing data to 'NA' and then used na.omit to remove the data I did not need. I also created several fuctions to help facilitate my analysis.

Data Exploration

I decided to visualize some of the data before running my models.

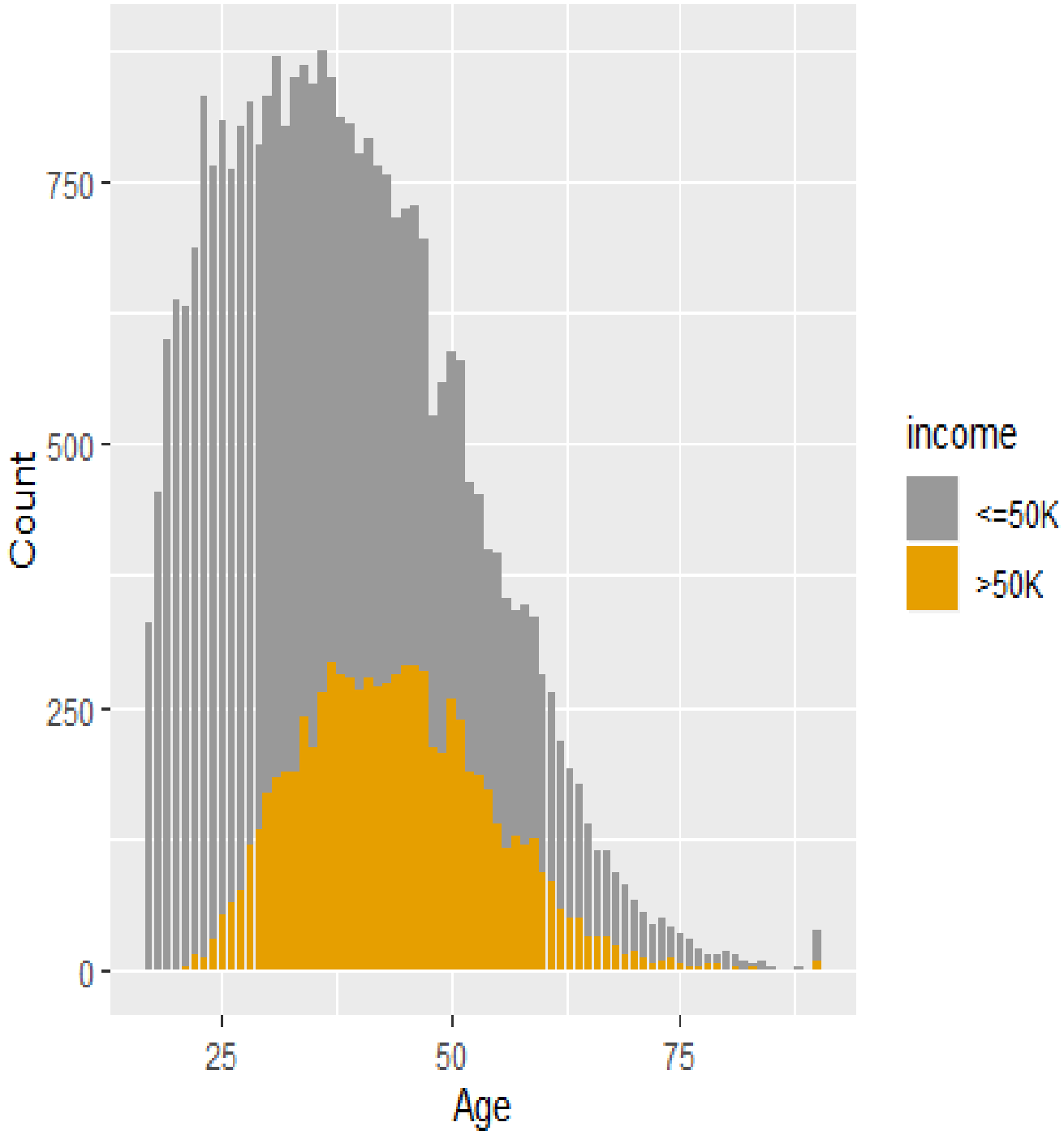
Education and Income

I first looked at the distribution of salaries above or below 50,000 based on education level. There were more individuals with salaries above 50,000 as their education level increased.



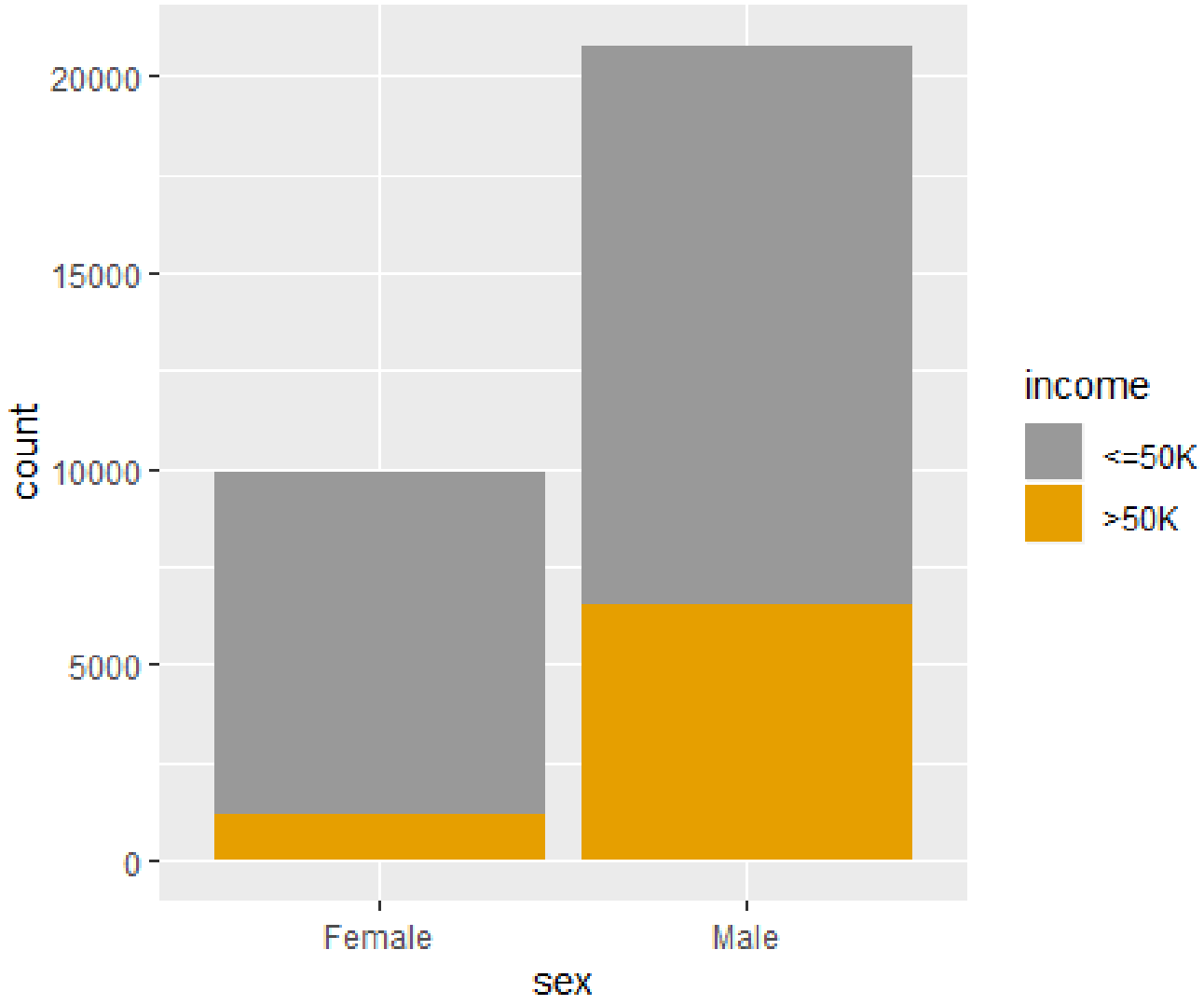
Age and Income

I also looked at the distribution of salaries by age. This plot is slighlty skewed to the right for salaries both above and below 50,000. Also, it appears that there are a larger number of individuals with salaries below 50,000 in this dataset.



Sex and Income

Next, I looked at salary distribution by sex. It is clear that there are more men in the dataset and proportionately, more males have salaries above 50,000.



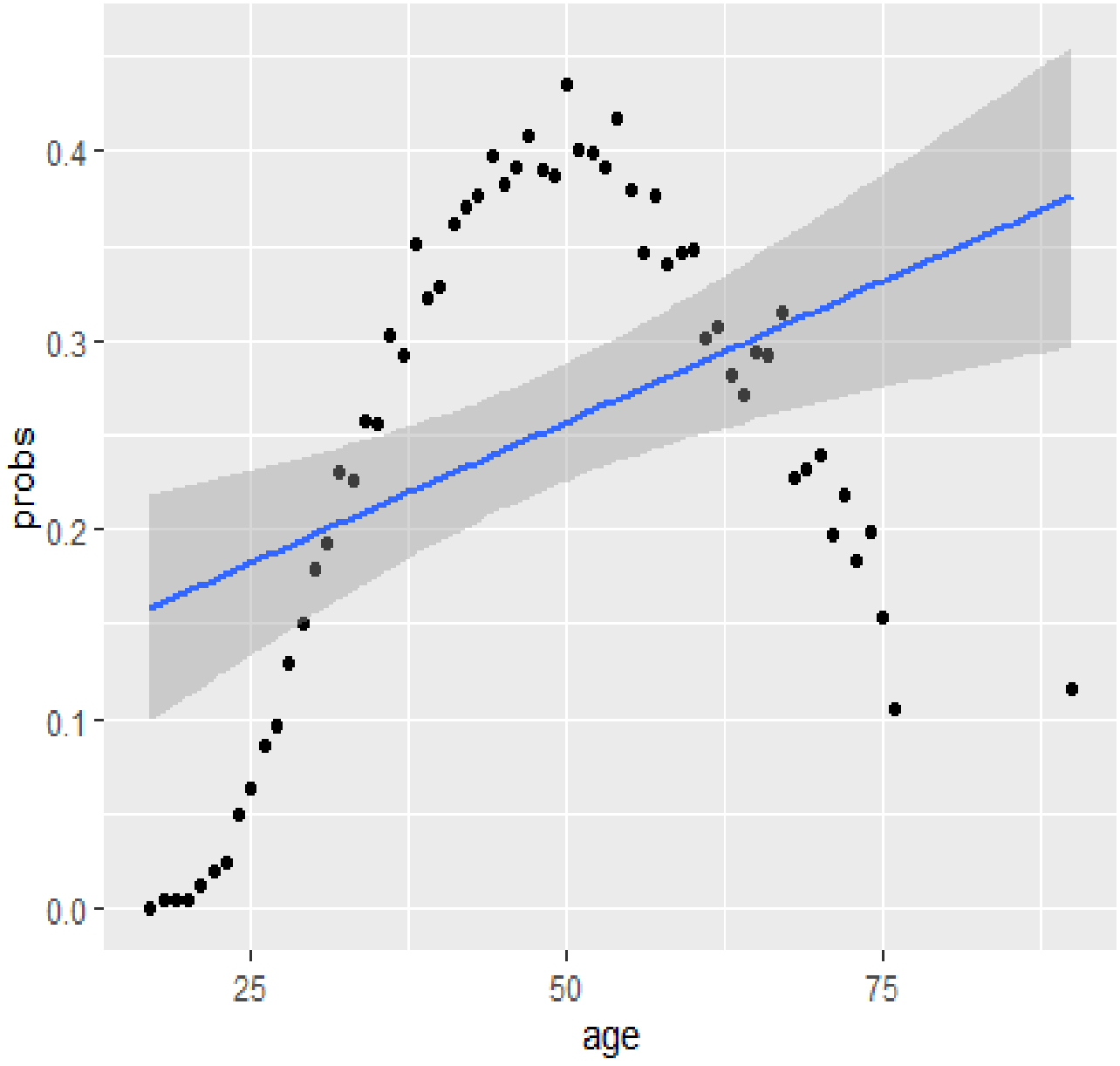
Model Testing and Results

Model 1: Logistic Regression

The first model I used to predict income involved Logistic Regression. I used variables such as age, education, and sex to predict salaries above or below 50,000.

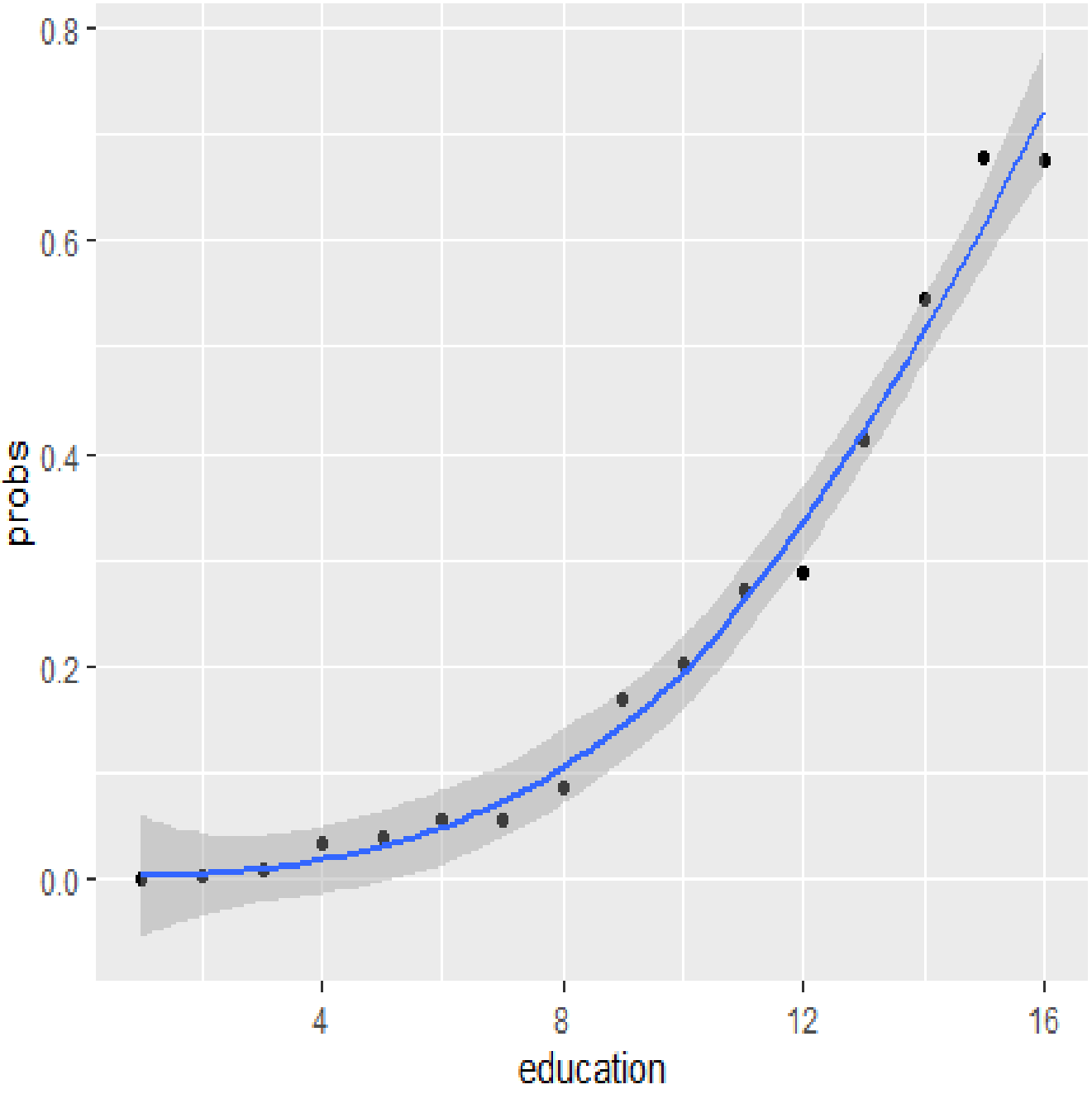
Logistic Regression: Age

Individuals between the ages of 35 and 55 have the highest chance of earning more than 50,000 annually.



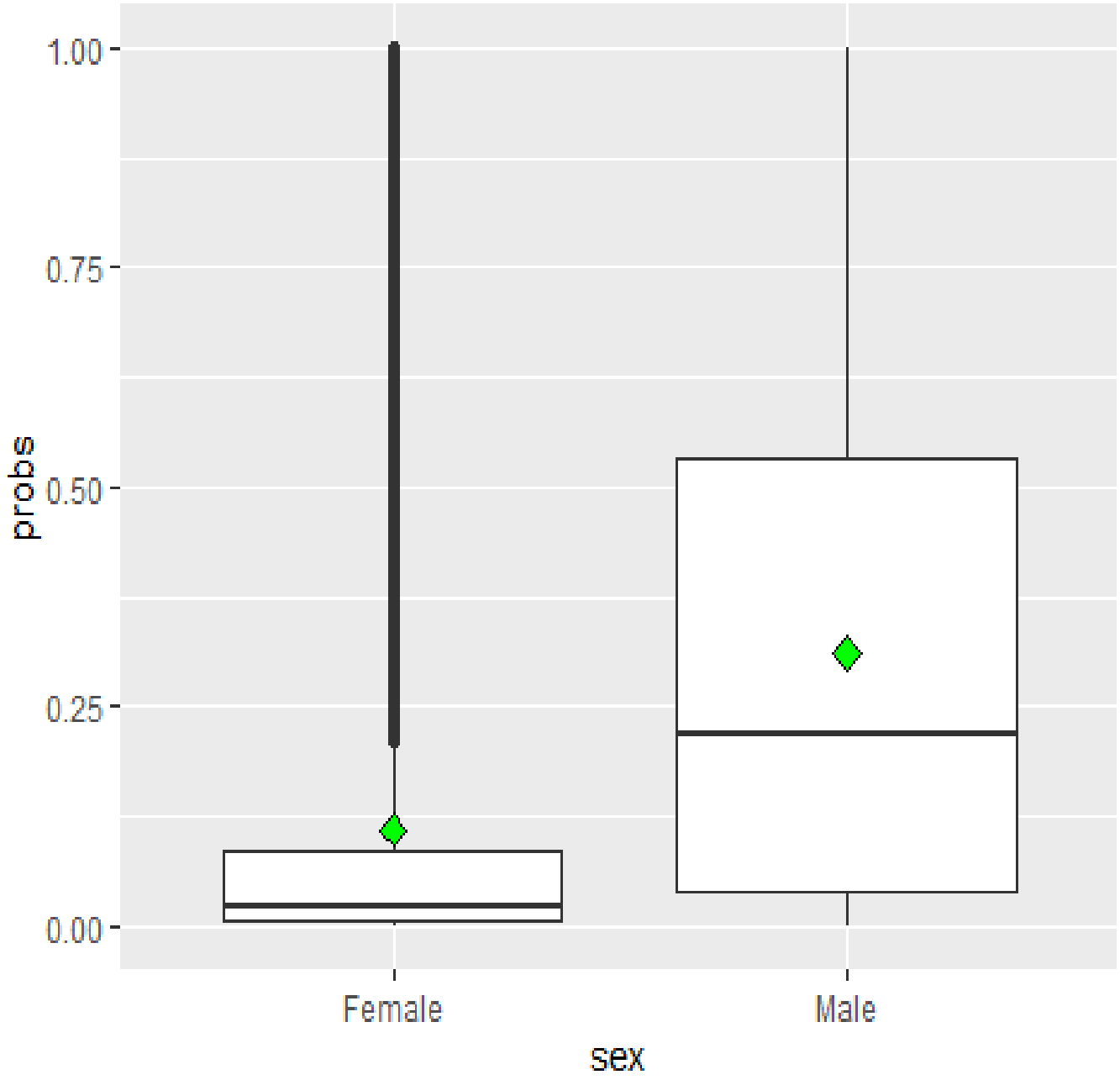
Logistic Regression: Education

Not surprisingly, the probability of having an annual salary above 50,000 increases the more years of education an individual has.

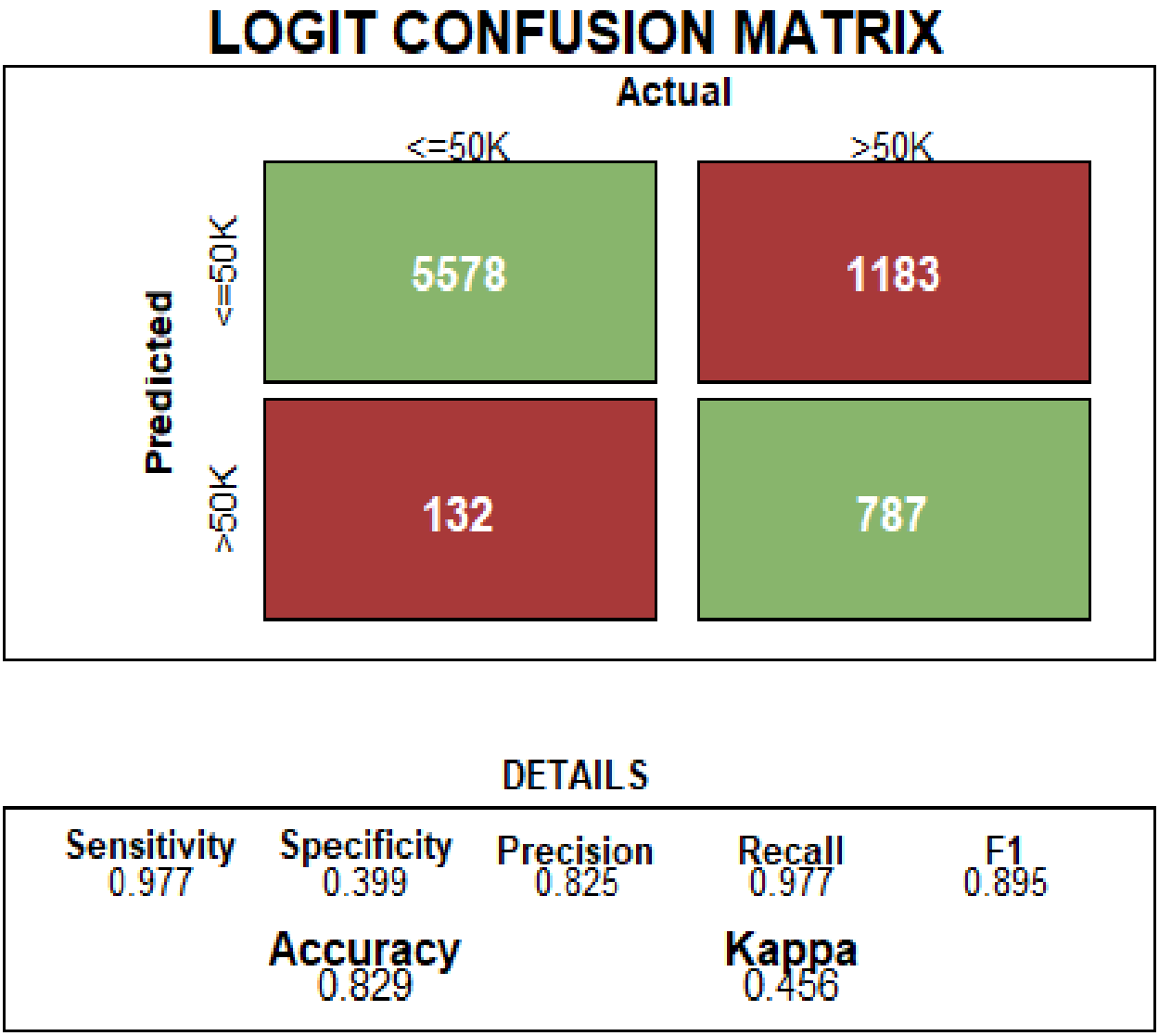


Logistic Regression: Sex

From the plot males have a higher probability of making more than 50,000 annually.

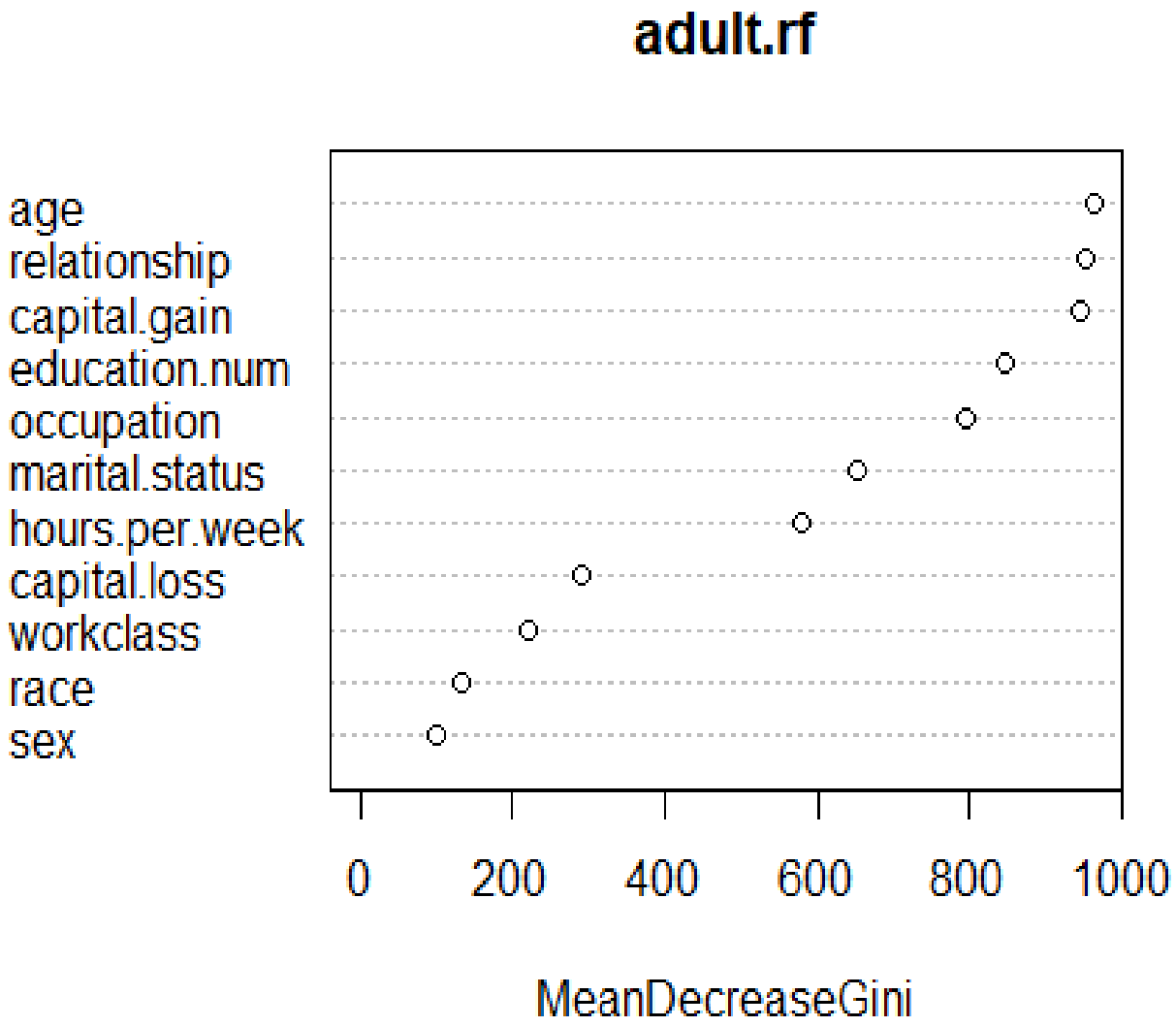


Confusion Matrix: Logistic Regression



Variables and Importance

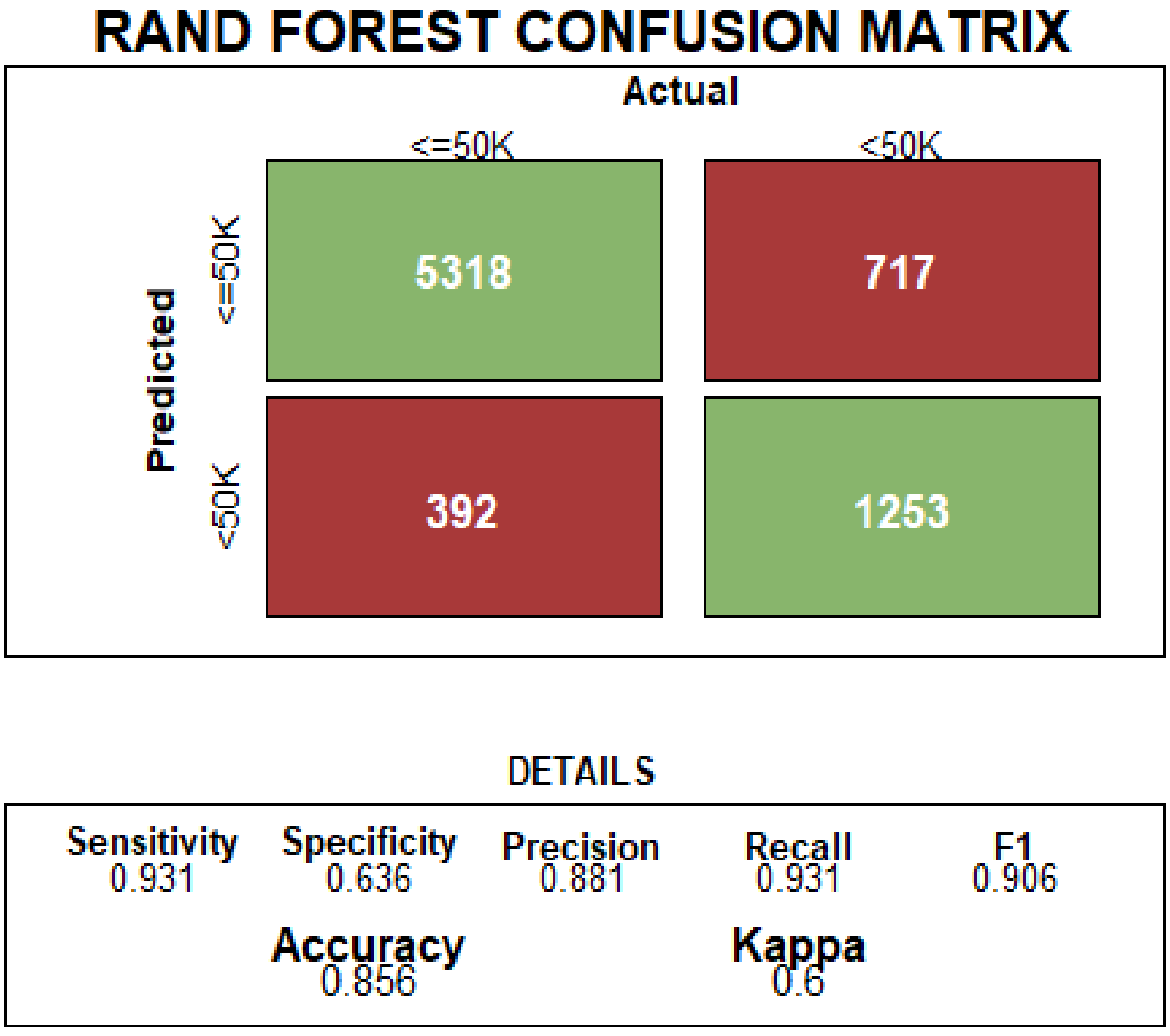
I looked at the different variables importance in predicting salaries. Age, realtionship, education and occupation were among the most important variables while class, race and sex where among the least important. This is displayed in the plot below.



Model 2: Random Forest

The Random Forests algorithm is one of the best among classification algorithms - able to classify large amounts of data with accuracy. Random Forests are an ensemble learning method (also thought of as a form of nearest neighbor predictor) for classification and regression that construct a number of decision trees at training time and outputting the class that is the mode of the classes output by individual trees. Random Forests are a wonderful tool for making predictions considering they do not overfit because of the law of large numbers. Introducing the right kind of randomness makes them accurate classifiers and regressors.

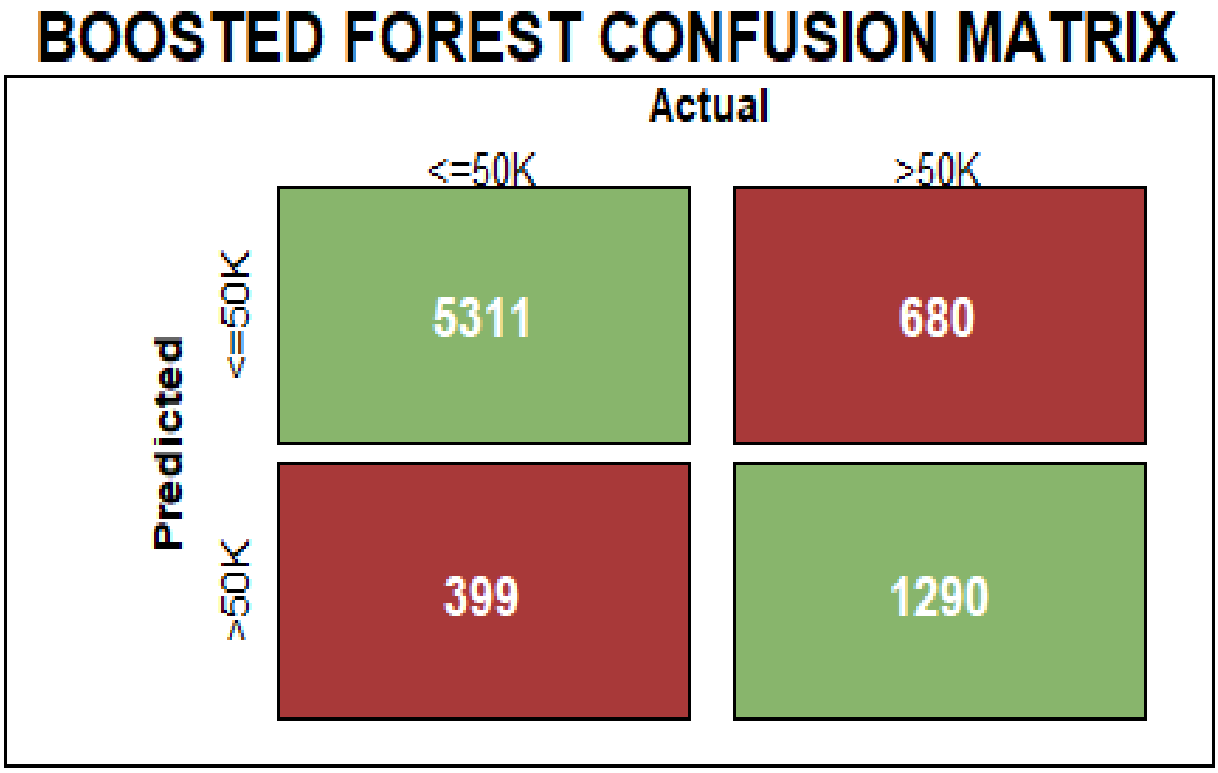
Below is the confusion matrix for the random forest model. The accuracy is greater than that of the logistic regression model, with lower sensitivity and higher specificity.

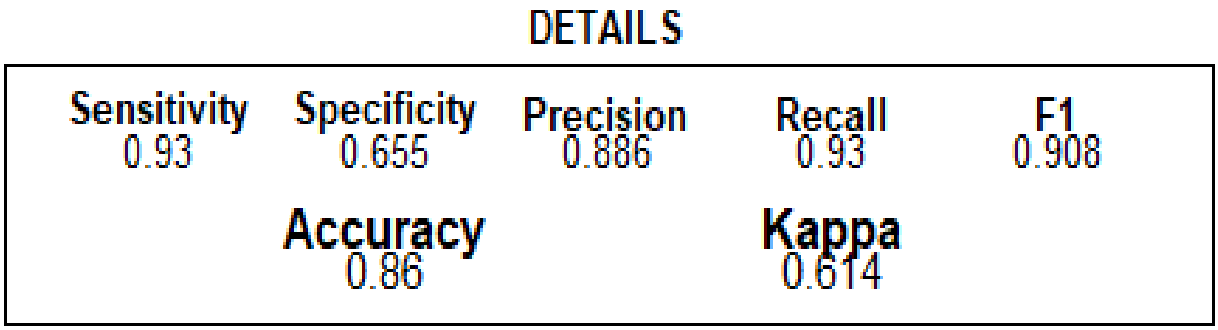


Model 3: Boosted Random Forest (BRF) with Bernoulli distributions

Random forests(RF) using Bernoulli distributions are much less data-dependent, when compared to original RF. BRF models help to close the gap between theoretical consistency and empirical soundness of RF classification.

The confusion matrix for the BRF model is displayed below. The accuracy, precision, and F1 score all slightly increased showing an improvement upon the normal RF model.





Conclusion

I used three different models on the Adult Income data set to predict individual's income (above or below 50,000/year). The logistic regression model showed that males between the ages of 35-55 who have higher levels of education are the most likely to have annual salaries above 50,000. The random forest models(both normal and boosted with Bernoulli distributions) improved upon logistic regression, showing the ability to predict incomes with a high level of accuracy.