Income Prediction

Brett

April 11, 2019

Brett Waugh 11 April 2019 adult\_logistic.R Looks at data about people in the United States in 1996 and makes a prediction about income. Most of script is taken from <https://rpubs.com/H_Zhu/235617> with very few revisions.

Get the data from here: <https://archive.ics.uci.edu/ml/datasets/Adult>. Then read in the data.

After adding in the data, check it to make sure everything looks correct.

str(adult)

## 'data.frame': 32561 obs. of 15 variables:  
## $ age : int 39 50 38 53 28 37 49 52 31 42 ...  
## $ workclass : Factor w/ 9 levels "?","Federal-gov",..: 8 7 5 5 5 5 5 7 5 5 ...  
## $ fnlwgt : int 77516 83311 215646 234721 338409 284582 160187 209642 45781 159449 ...  
## $ education : Factor w/ 16 levels "10th","11th",..: 10 10 12 2 10 13 7 12 13 10 ...  
## $ education\_num : int 13 13 9 7 13 14 5 9 14 13 ...  
## $ marital\_status: Factor w/ 7 levels "Divorced","Married-AF-spouse",..: 5 3 1 3 3 3 4 3 5 3 ...  
## $ occupation : Factor w/ 15 levels "?","Adm-clerical",..: 2 5 7 7 11 5 9 5 11 5 ...  
## $ relationship : Factor w/ 6 levels "Husband","Not-in-family",..: 2 1 2 1 6 6 2 1 2 1 ...  
## $ race : Factor w/ 5 levels "Amer-Indian-Eskimo",..: 5 5 5 3 3 5 3 5 5 5 ...  
## $ sex : Factor w/ 2 levels "Female","Male": 2 2 2 2 1 1 1 2 1 2 ...  
## $ capital\_gain : int 2174 0 0 0 0 0 0 0 14084 5178 ...  
## $ capital\_loss : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ hours\_per\_week: int 40 13 40 40 40 40 16 45 50 40 ...  
## $ native\_country: Factor w/ 42 levels "?","Cambodia",..: 40 40 40 40 6 40 24 40 40 40 ...  
## $ income : Factor w/ 2 levels "<=50K",">50K": 1 1 1 1 1 1 1 2 2 2 ...

Next, we will remove some variables and create some graphs. Looking at the percentages for capital gains and capital losses, it is clear that very few people partake in this. Due to how few individuals partake in it, it is best to remove it from the study to not skew the results. Native country will also be removed.

# Three variables dropped.  
adult$education <- NULL  
adult$fnlwgt <- NULL  
adult$relationship <- NULL  
  
# Percentage of observatiosn with no capital gain or loss.  
sum(adult$capital\_gain == 0)/length(adult$capital\_gain)

## [1] 0.9167102

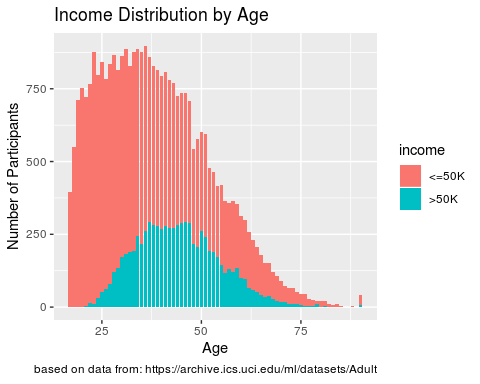
sum(adult$capital\_loss == 0)/length(adult$capital\_loss)

## [1] 0.9533491

# Variables appear to be unimportant for analysis and are dropped.   
adult$capital\_gain <- NULL  
adult$capital\_loss <- NULL  
adult$native\_country <- NULL

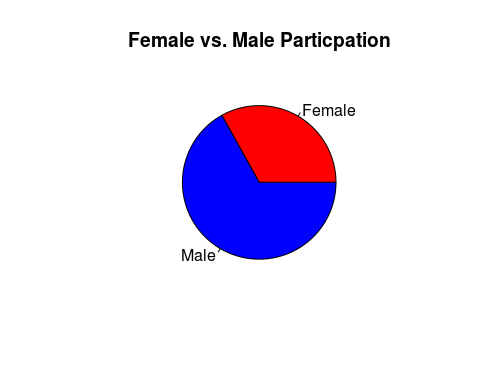
The first graph is a histogram of age by income group.

# Histogram of age by income group.   
ggplot(adult, aes(x=age, group=income, fill=income)) +   
 geom\_bar() +   
 labs(x = "Age", y = "Number of Participants", title = "Income Distribution by Age", caption = "based on data from: https://archive.ics.uci.edu/ml/datasets/Adult")

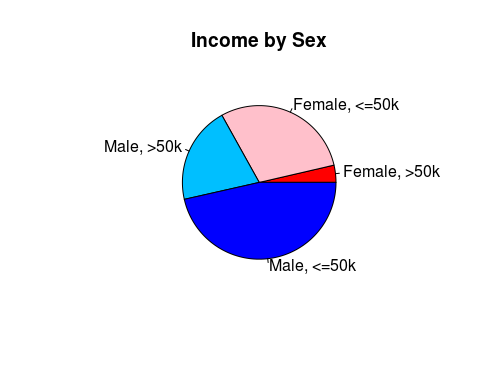
 As you can see in this graph, there are drastically more people making under 50k than above 50k. The center for those making below 50k is around 32, while the center for those making above 50k is around 34. This is a small difference in age, which may mean there are other variables in play here.

The next set of graphs deal with sex and income.

# Subset the data for the pie charts.   
numFe <- nrow(adult[adult$sex=="Female",])  
numMa <- nrow(adult[adult$sex=="Male",])  
numFeL <- nrow(adult[adult$sex=="Female" & adult$income=="<=50K",])  
numFeG <- nrow(adult[adult$sex=="Female" & adult$income==">50K",])  
numMaG <- nrow(adult[adult$sex=="Male" & adult$income==">50K",])  
numMaL <- nrow(adult[adult$sex=="Male" & adult$income=="<=50K",])  
  
# Pie chart of sex partcipation.   
slices <- c(numFe, numMa)  
lbls <- c("Female", "Male")  
pie(slices, labels = lbls, main="Female vs. Male Particpation", col=c("red", "blue"))

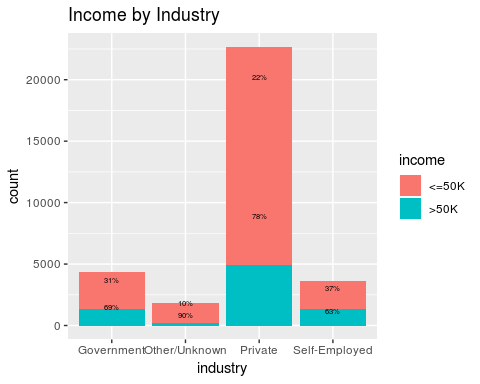


# Pie chart of sex's incomes.  
slices <- c(numFeG, numFeL, numMaG, numMaL)  
lbls <- c("Female, >50k", "Female, <=50k","Male, >50k", "Male, <=50k")  
pie(slices, labels = lbls, main="Income by Sex", col=c("red","pink","deepskyblue","blue"))

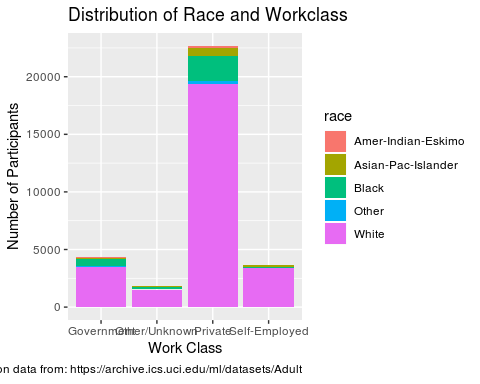
 These graphs show that there are significantly more female participants than male participants in the study. This could be representative of the overall population trend, but it may also skew the results of the study. There does appear to be a significant difference in income when sex is considered.

After preparing the data more, and combining several of the fields we can create a barplot for the data.

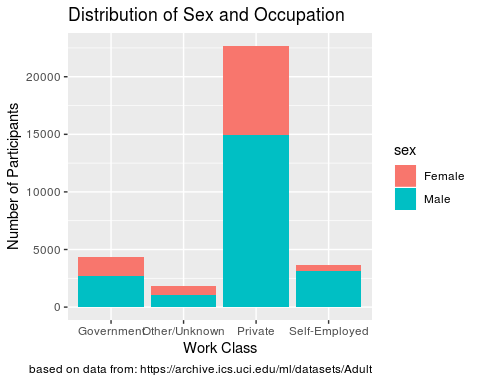
# Barplot of counts by industry with in group proportions.  
ggplot(df, aes(x = industry, y = count, fill = income)) +  
 geom\_bar(stat = "identity") +  
 geom\_text(aes(y = pos, label = label), size = 2) +   
 ggtitle('Income by Industry')



# Barplot of counts of industry by race.   
ggplot(adult, aes(x=workclass, fill=race, group=race)) +   
 geom\_bar() +  
 labs(x = "Work Class", y = "Number of Participants", title = "Distribution of Race and Workclass", caption = "based on data from: https://archive.ics.uci.edu/ml/datasets/Adult")



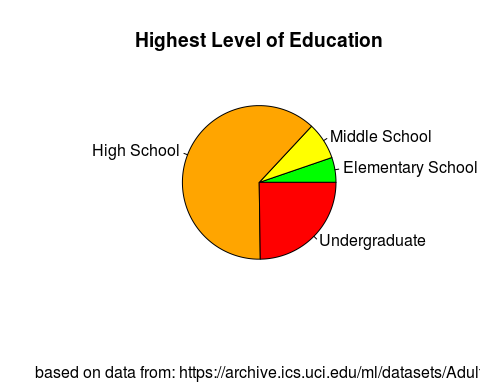
# Barplot of counts of industry by sex.   
ggplot(adult, aes(x=workclass, fill=sex, group=sex)) +   
 geom\_bar() +  
 labs(x = "Work Class", y = "Number of Participants", title = "Distribution of Sex and Occupation", caption = "based on data from: https://archive.ics.uci.edu/ml/datasets/Adult")

 From the “Income by Industry” barplot, we can see that the Private industry employees the most workers, followed by the government, then self-employed, then other/unknown. The Private industry also has the greatest number of people making above 50k per year, but Self-Employed industries have the greatest percentage of people making above 50k. The industry with the most number of people making 50k or less is Other/Unknown.

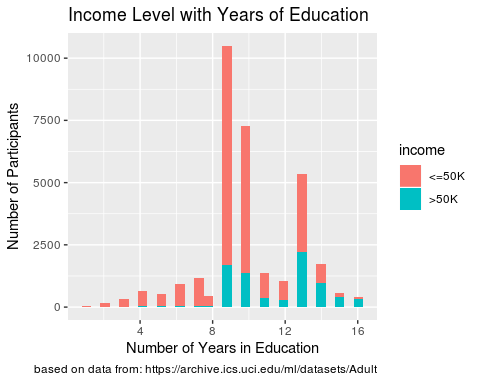
The “Distribution of Race and Workclass” shows that White has the most in all categories. People who are Black most commonly work in Private, with second place being the Government. The numbers for Ameri-Indian-Eskimo, Asian-Pac-Islander, and Other are too difficult to distiguish.

“Distribution of Sex and Occupation” shows that most women work in Private, followed by the Government. This also shows that there are very few women who are Self-Employed.

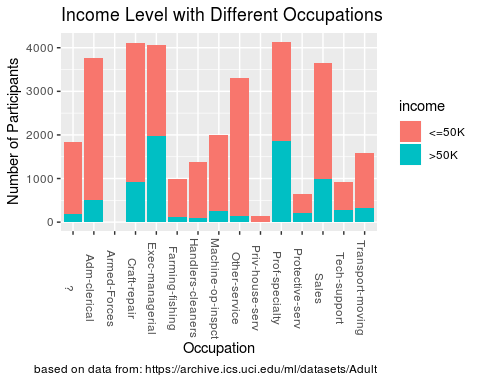
# Dividing the   
elementary <- nrow(adult[adult$education\_num<6,])  
middle <- nrow(adult[adult$education\_num<9 & adult$education\_num>5,])  
high <- nrow(adult[adult$education\_num<13 & adult$education\_num>8,])  
undergrad <- nrow(adult[adult$education\_num>12,])  
  
# Pie chart of years in school.   
slices <- c(elementary, middle, high, undergrad)  
lbls <- c("Elementary School", "Middle School", "High School", "Undergraduate")  
pie(slices, labels = lbls, main="Highest Level of Education", col=c("green","yellow","orange","red"), sub="based on data from: https://archive.ics.uci.edu/ml/datasets/Adult")



# Barplot of counts by years of education with in group proportions.  
ggplot(adult, aes(x = education\_num, group=income, fill = income)) +  
 geom\_histogram() +  
 labs(x = "Number of Years in Education", y = "Number of Participants", title = "Income Level with Years of Education", caption = "based on data from: https://archive.ics.uci.edu/ml/datasets/Adult")

 The piechart shows that most people in 1994 were finishing their education in High School. It appears that few only completed Middle and Elementary school. There is also a greater number of individuals continuing their education into college. Immediately, it is apparent that the greatest number of people making 50k or less have nine or ten years of education. There are few individuals in the study with less than nine years of education making above 50k. The number of years with the greatest percentage above 50k are sixteen years, with fifteen years coming in a close second. From this study, all of those with only one year of education make 50k or less.

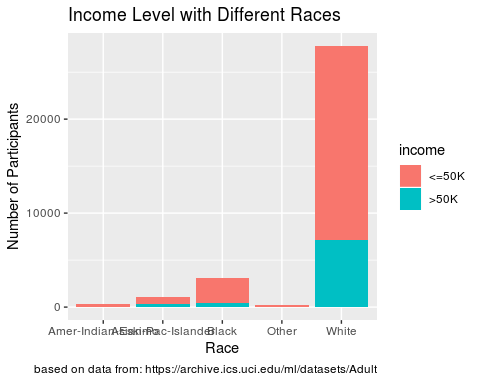
# Barplot of countrs by occupation with in group propoertions.   
ggplot(adult, aes(x = occupation, group=income, fill = income)) +  
 geom\_histogram(stat="count") +  
 labs(x = "Occupation", y = "Number of Participants", title = "Income Level with Different Occupations", caption = "based on data from: https://archive.ics.uci.edu/ml/datasets/Adult") +  
 theme(axis.text.x = element\_text(angle = -90))

 Income level can very greatly depending on which field someone goes into. From this dataset, the greatest number of people making 50k or less is from the Blue-Collar occupations. The greatest number of people making more than 50k are in White-Collar occupations. While the greatest number may be in White-Collar, Professional occupations holds the greatest percentage of people making above 50k. The occupations with the greatest percentage making 50k or less is in Other/Unknown and Service.

# Barplot of counts by marital status with in group proportions.  
ggplot(adult, aes(x = marital\_status, group= income, fill = income)) +  
 geom\_bar(stat = "count") +  
 labs(x = "Occupation", y = "Number of Participants", title = "Income Level with Different Occupations", caption = "based on data from: https://archive.ics.uci.edu/ml/datasets/Adult") +  
 theme(axis.text.x = element\_text(angle = -90))

 Marital status appears to play a significant role in determing how much someone will make. The greatest number of people making both types of income are Married. Married also has the greatest percentage of individuals making more than 50k. The greatest percentage of people making less than 50k is Single. However, Divorced, Seperated, Single, and Widowed also all have above 90% making 50k or less.

# Barplot of counts by race and income.  
ggplot(adult, aes(x = race, group = income, fill = income)) +  
 geom\_histogram(stat = "count") +  
 labs(x = "Race", y = "Number of Participants", title = "Income Level with Different Races", caption = "based on data from: https://archive.ics.uci.edu/ml/datasets/Adult")



Looking at income distribution by race can be a tricky situation as some people can be from multple races, or are sometimes unaware of ancestry. The greatest number of people making above 50k are White, but the greatest number of people making 50k or less are also White. Amer-Indian-Eskimo, Asian-Pac-Islander, Other, and Black all have high percentages of people making 50k or less.

At this point, we only have nine remaining features to work with.

summary(adult)

## age workclass education\_num   
## Min. :17.00 Government : 4351 Min. : 1.00   
## 1st Qu.:28.00 Other/Unknown: 1857 1st Qu.: 9.00   
## Median :37.00 Private :22696 Median :10.00   
## Mean :38.58 Self-Employed: 3657 Mean :10.08   
## 3rd Qu.:48.00 3rd Qu.:12.00   
## Max. :90.00 Max. :16.00   
##   
## marital\_status occupation   
## Divorced : 4443 Prof-specialty :4140   
## Married-AF-spouse : 23 Craft-repair :4099   
## Married-civ-spouse :14976 Exec-managerial:4066   
## Married-spouse-absent: 418 Adm-clerical :3770   
## Never-married :10683 Sales :3650   
## Separated : 1025 Other-service :3295   
## Widowed : 993 (Other) :9541   
## race sex hours\_per\_week income   
## Amer-Indian-Eskimo: 311 Female:10771 Min. : 1.00 <=50K:24720   
## Asian-Pac-Islander: 1039 Male :21790 1st Qu.:40.00 >50K : 7841   
## Black : 3124 Median :40.00   
## Other : 271 Mean :40.44   
## White :27816 3rd Qu.:45.00   
## Max. :99.00   
##

The next stage in the process is creating a model. For this particular dataset, logistic regression will be used.

summary(m1)

##   
## Call:  
## glm(formula = income ~ ., family = binomial("logit"), data = training\_set)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.7087 -0.5640 -0.2402 -0.0550 3.5226   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)  
## (Intercept) 4.249127 375.247824 0.011 0.99097  
## age 0.029340 0.001670 17.564 < 2e-16  
## workclassOther/Unknown -13.509252 375.247711 -0.036 0.97128  
## workclassPrivate 0.092273 0.055952 1.649 0.09912  
## workclassSelf-Employed -0.088158 0.072240 -1.220 0.22233  
## education\_num 0.297134 0.009617 30.898 < 2e-16  
## marital\_statusMarried-AF-spouse 2.623295 0.537076 4.884 1.04e-06  
## marital\_statusMarried-civ-spouse 2.097051 0.068304 30.702 < 2e-16  
## marital\_statusMarried-spouse-absent -0.091175 0.224534 -0.406 0.68469  
## marital\_statusNever-married -0.392235 0.083041 -4.723 2.32e-06  
## marital\_statusSeparated -0.102700 0.163946 -0.626 0.53104  
## marital\_statusWidowed -0.039551 0.155261 -0.255 0.79892  
## occupationAdm-clerical -12.933740 375.247724 -0.034 0.97250  
## occupationArmed-Forces -13.322993 375.249878 -0.036 0.97168  
## occupationCraft-repair -12.950148 375.247723 -0.035 0.97247  
## occupationExec-managerial -12.165141 375.247723 -0.032 0.97414  
## occupationFarming-fishing -14.097399 375.247719 -0.038 0.97003  
## occupationHandlers-cleaners -13.725472 375.247745 -0.037 0.97082  
## occupationMachine-op-inspct -13.343338 375.247728 -0.036 0.97163  
## occupationOther-service -13.866072 375.247734 -0.037 0.97052  
## occupationPriv-house-serv -26.677623 391.374021 -0.068 0.94566  
## occupationProf-specialty -12.403742 375.247724 -0.033 0.97363  
## occupationProtective-serv -12.604409 375.247738 -0.034 0.97320  
## occupationSales -12.710254 375.247724 -0.034 0.97298  
## occupationTech-support -12.341954 375.247733 -0.033 0.97376  
## occupationTransport-moving -13.113249 375.247725 -0.035 0.97212  
## raceAsian-Pac-Islander 0.154417 0.249791 0.618 0.53645  
## raceBlack 0.309849 0.238284 1.300 0.19349  
## raceOther -0.392717 0.373977 -1.050 0.29367  
## raceWhite 0.394851 0.227629 1.735 0.08281  
## sexMale 0.167759 0.053866 3.114 0.00184  
## hours\_per\_week 0.030660 0.001692 18.120 < 2e-16  
##   
## (Intercept)   
## age \*\*\*  
## workclassOther/Unknown   
## workclassPrivate .   
## workclassSelf-Employed   
## education\_num \*\*\*  
## marital\_statusMarried-AF-spouse \*\*\*  
## marital\_statusMarried-civ-spouse \*\*\*  
## marital\_statusMarried-spouse-absent   
## marital\_statusNever-married \*\*\*  
## marital\_statusSeparated   
## marital\_statusWidowed   
## occupationAdm-clerical   
## occupationArmed-Forces   
## occupationCraft-repair   
## occupationExec-managerial   
## occupationFarming-fishing   
## occupationHandlers-cleaners   
## occupationMachine-op-inspct   
## occupationOther-service   
## occupationPriv-house-serv   
## occupationProf-specialty   
## occupationProtective-serv   
## occupationSales   
## occupationTech-support   
## occupationTransport-moving   
## raceAsian-Pac-Islander   
## raceBlack   
## raceOther   
## raceWhite .   
## sexMale \*\*   
## hours\_per\_week \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 28685 on 26048 degrees of freedom  
## Residual deviance: 18635 on 26017 degrees of freedom  
## AIC: 18699  
##   
## Number of Fisher Scoring iterations: 14

confint(m1)

## 2.5 % 97.5 %  
## (Intercept) -24.49466683 NA  
## age 0.02607002 0.03261848  
## workclassOther/Unknown NA 15.41480660  
## workclassPrivate -0.01716809 0.20217877  
## workclassSelf-Employed -0.22973319 0.05346097  
## education\_num 0.27836987 0.31606860  
## marital\_statusMarried-AF-spouse 1.52658091 3.65981176  
## marital\_statusMarried-civ-spouse 1.96438361 2.23218165  
## marital\_statusMarried-spouse-absent -0.54940997 0.33336304  
## marital\_statusNever-married -0.55484099 -0.22923904  
## marital\_statusSeparated -0.43337234 0.21038904  
## marital\_statusWidowed -0.35097817 0.25844775  
## occupationAdm-clerical NA 15.97089043  
## occupationArmed-Forces -440.48097485 -507.75895506  
## occupationCraft-repair NA 15.95742475  
## occupationExec-managerial NA 16.74276105  
## occupationFarming-fishing NA 14.81381408  
## occupationHandlers-cleaners NA 15.16030954  
## occupationMachine-op-inspct NA 15.55742694  
## occupationOther-service NA 15.02713650  
## occupationPriv-house-serv NA 2.99945254  
## occupationProf-specialty NA 16.50296449  
## occupationProtective-serv NA 16.28428976  
## occupationSales NA 16.19642220  
## occupationTech-support NA 16.55258360  
## occupationTransport-moving NA 15.78941271  
## raceAsian-Pac-Islander -0.32319070 0.65824476  
## raceBlack -0.14371645 0.79272685  
## raceOther -1.14189263 0.32979462  
## raceWhite -0.03646172 0.85833868  
## sexMale 0.06223155 0.27340403  
## hours\_per\_week 0.02735151 0.03398486

We can see the relationships between some of the variables here. Notice some of the features with the low P Values: age, education\_num, martital\_StatusMarried, martitalStatusSingle, occupationProfessional, OccupationSales, occupationWhite-Collar, sexMale, and hours\_per\_week. These features seem to have a significant impact on determing the income of an individual.

# confusion matrix   
tb <- table(pred, testing\_set$income)  
tb

##   
## pred <=50K >50K  
## <=50K 4526 707  
## >50K 386 893

resultGiven <- (4526+893)/(4526+707+386+893)  
resultGiven

## [1] 0.832156

From the confusion matrix we can determine the accuracy of the model to be 0.832 or 83.2%. This is a great score for the model, because it does not suggest overfitting while maintaining a high degree of accuracy. Given information about an individual, the model should be able to accurately predict if the individual makes greater than 50k 82.7% of the time.

The second model uses a slightly different approach to the problem. This model is a “General Linear Model”. The logistic model is more appropirate for this dataset, so it will be interesting to see how well a linear model performs on the dataset.

# ---- Second model ---- #  
ControlParameters <-trainControl(method="repeatedcv",   
 number=10,  
 repeats=10)  
  
otherModel <-train(income ~ .,   
 data=training\_set,  
 method="glm",  
 trControl= ControlParameters  
)  
  
otherModel

## Generalized Linear Model   
##   
## 26049 samples  
## 8 predictors  
## 2 classes: '<=50K', '>50K'   
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold, repeated 10 times)   
## Summary of sample sizes: 23444, 23444, 23443, 23445, 23444, 23445, ...   
## Resampling results:  
##   
## Accuracy Kappa   
## 0.8303392 0.4965605

resultOtherModel <- 0.830  
resultOtherModel

## [1] 0.83

While this model took a different approach to the problem, it still yielded an accuracy of 0.830 or 83.0%. Another benefit to this model is that it uses 10-fold cross validation, making it understandable to lose an additional percentage from the first model.

Below are Support Vector Machine (SVM) models. I played around with the C value to see if there was a significant impact on the resutls.

# First SVM with C=1  
svmfit1 = svm(income ~ ., data = adult, kernal="radial", cost=1, na.action =na.omit, scale = TRUE)  
print(svmfit1)

##   
## Call:  
## svm(formula = income ~ ., data = adult, kernal = "radial", cost = 1,   
## na.action = na.omit, scale = TRUE)  
##   
##   
## Parameters:  
## SVM-Type: C-classification   
## SVM-Kernel: radial   
## cost: 1   
## gamma: 0.03125   
##   
## Number of Support Vectors: 12170

# Results of training set.   
pred\_train1 <- predict(svmfit1, training\_set)  
result1 <- mean(pred\_train1==training\_set$income)  
result1

## [1] 0.8351184

# Results of test set.  
pred\_test1 <- predict(svmfit1, testing\_set)  
mean(pred\_test1==testing\_set$income)

## [1] 0.8336916

The results from the training set are 0.835 or 83.5% accurate, while the results from the test set are 0.833 or 83.3% accurate. Because the values are similar we have good confidence that the model was trained correctly.

# Second SVM with C=10  
svmfit10 = svm(income ~ ., data = adult, kernal="radial", cost=10, na.action =na.omit, scale = TRUE)  
print(svmfit10)

##   
## Call:  
## svm(formula = income ~ ., data = adult, kernal = "radial", cost = 10,   
## na.action = na.omit, scale = TRUE)  
##   
##   
## Parameters:  
## SVM-Type: C-classification   
## SVM-Kernel: radial   
## cost: 10   
## gamma: 0.03125   
##   
## Number of Support Vectors: 11855

# Results of training set.   
pred\_train10 <- predict(svmfit10, training\_set)  
result10 <- mean(pred\_train10==training\_set$income)  
result10

## [1] 0.8347345

# Results of test set.  
pred\_test10 <- predict(svmfit10, testing\_set)  
mean(pred\_test10==testing\_set$income)

## [1] 0.8358415

The results from the training set are 0.835 or 83.5% accurate, while the results from the test set are 0.836 or 83.6% accurate. Because the values are similar we have good confidence that the model was trained correctly.

# Third SVM with C=25  
svmfit25 = svm(income ~ ., data = adult, kernal="radial", cost=25, na.action =na.omit, scale = TRUE)  
print(svmfit25)

##   
## Call:  
## svm(formula = income ~ ., data = adult, kernal = "radial", cost = 25,   
## na.action = na.omit, scale = TRUE)  
##   
##   
## Parameters:  
## SVM-Type: C-classification   
## SVM-Kernel: radial   
## cost: 25   
## gamma: 0.03125   
##   
## Number of Support Vectors: 11781

# Results of training set.   
pred\_train25 <- predict(svmfit25, training\_set)  
result25 <- mean(pred\_train25==training\_set$income)  
result25

## [1] 0.8364621

# Results of test set.  
pred\_test25 <- predict(svmfit25, testing\_set)  
mean(pred\_test25==testing\_set$income)

## [1] 0.8367629

The results from the training set are 0.836 or 83.6% accurate, while the results from the test set are 0.837 or 83.7% accurate. Because the values are similar we have good confidence that the model was trained correctly.

With the results from above, I used the C value of 25 to make the last SVM. To further support the results, I used cross validation of ten.

# Final SVM with C=  
svmfitFinal = svm(income ~ ., data = adult, kernal="radial", cost=25, na.action =na.omit, scale = TRUE, cross=10)  
print(svmfitFinal)

##   
## Call:  
## svm(formula = income ~ ., data = adult, kernal = "radial", cost = 25,   
## cross = 10, na.action = na.omit, scale = TRUE)  
##   
##   
## Parameters:  
## SVM-Type: C-classification   
## SVM-Kernel: radial   
## cost: 25   
## gamma: 0.03125   
##   
## Number of Support Vectors: 11781

# Results of training set.   
pred\_trainFinal <- predict(svmfitFinal, training\_set)  
resultFinal <- mean(pred\_trainFinal==training\_set$income)  
resultFinal

## [1] 0.8364621

# Results of test set.  
pred\_testFinal <- predict(svmfitFinal, testing\_set)  
resultFinal <- mean(pred\_testFinal==testing\_set$income)

Final results come to 0.836 or 83.6% for the testing set, and 0.837 or 83.7% for the test set. This leads us to conclude that the models have been trained correctly.

# Dataframe for model results.  
modelResults = data.frame(model=c("Given model", "Other model", "First SVM", "Second SVM", "Third SVM", "Final SVM"))  
modelResults$accuracy <- c(resultGiven, resultOtherModel, result1, result10, result25, resultFinal)   
  
# Plot for results from models.   
plot(modelResults$model, modelResults$accuracy, type="p", main="Results of Models", sub="based on data from: https://archive.ics.uci.edu/ml/datasets/Adult", xlab="Model", ylab="Accuracy")

