Prediction Model for Income Classification

Under $50000 vs Over $50000

steve dubois

4/22/2019

ABSTRACT

The prominent inequality of wealth and income is a huge concern especially in the United States. The likelihood of diminishing poverty is one valid reason to reduce the world's surging level of economic inequality. The principle of universal moral equality ensures sustainable development and improves the economic stability of a nation. Governments in different countries have been trying their best to address this problem and provide an optimal solution. The aim here to show the usage of machine learning techniques in providing a solution to the income equality problem. The UCI Adult Dataset has been used for the purpose. Specifically, several machine learning classification models have been compared to predict whether a person's yearly income in the US falls in the income category of either greater than 50K dollars or less/equal to 50K dollars category based on a certain set of attributes. So, what_Y(>50, <=50) is predicted given (X1, X2, X3,...Xn), where Y is an income level, and X is a statistic feature of an individual.

LIBRRARIES USED - R PACKAGES

```
library(knitr)
library(ggvis)
library(ISLR)
library(e1071)
library(gmodels)
library(tidyverse)
library(tidyr)
library(dplyr)
library(readr)
library(ggplot2)
library(randomForest)
library(caret)
library(data.table)
library(gbm)
library(rpart)
library(rpart.plot)
library(plotly)
library(ggvis)
library(neuralnet)
library(MASS)
```

LOADING CENSUS DATA

```
train <- fread("http://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data")
test <- fread("http://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.test")
```

```
## Warning in fread("http://archive.ics.uci.edu/ml/machine-learning-databases/
```

```
## adult/adult.test"): Detected 1 column names but the data has 15 columns
## (i.e. invalid file). Added 14 extra default column names at the end.
```

Initializing headers

```
##
##  TRUE
## 32561
```

```
##
## Under_50K  More_50K
##        76        24
```

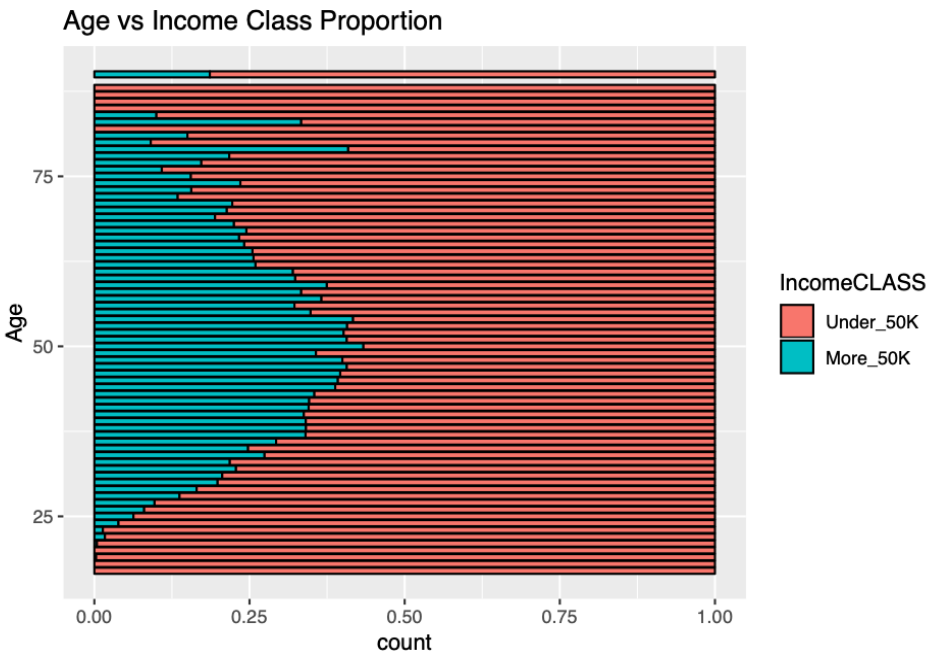
```
## [1] 0
```

```
## [1] 0
```

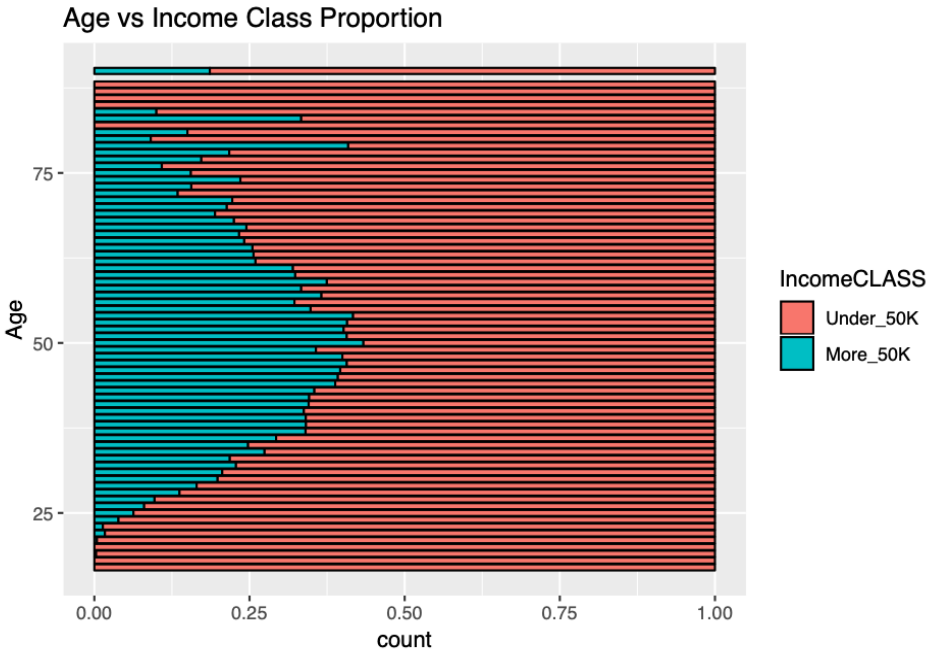
SETTING CATEGORICAL; FEATURES

EXPLORATORY DATA ANALYSIS USING GGPLOT

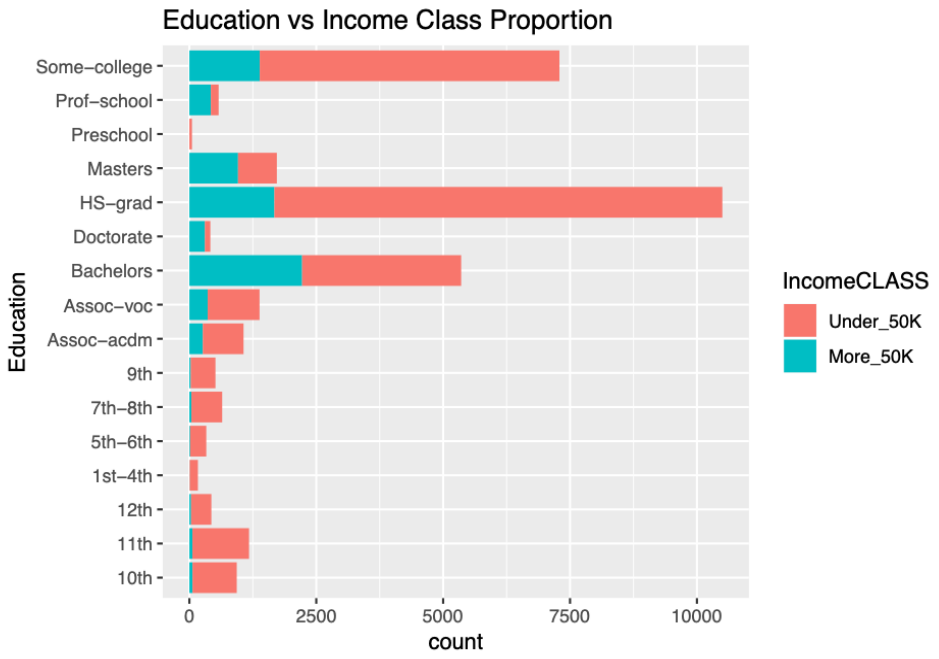
```
P <- ggplot(train,aes(x = Age, fill = IncomeCLASS)) + geom_bar(position = "fill", color = "black") + co
P1 <- P + labs(title = "Age vs Income Class Proportion")
P1
```



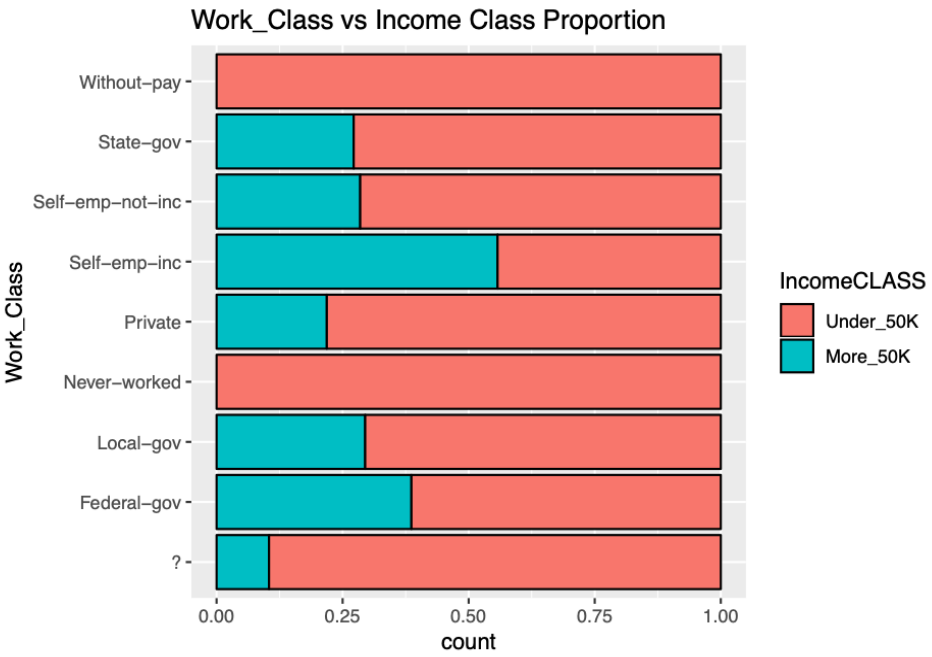
```
P2 <- ggplot(train,aes(x = Age, fill = IncomeCLASS)) + geom_bar(position = "fill", color = "black") + c
P3 <- P2 + labs(title = "Age vs Income Class Proportion")
P3
```




```
Q <- ggplot(train,aes(x = Education, fill = IncomeCLASS)) + geom_bar() + coord_flip()
Q1 <- Q + labs(title = "Education vs Income Class Proportion")
Q1
```

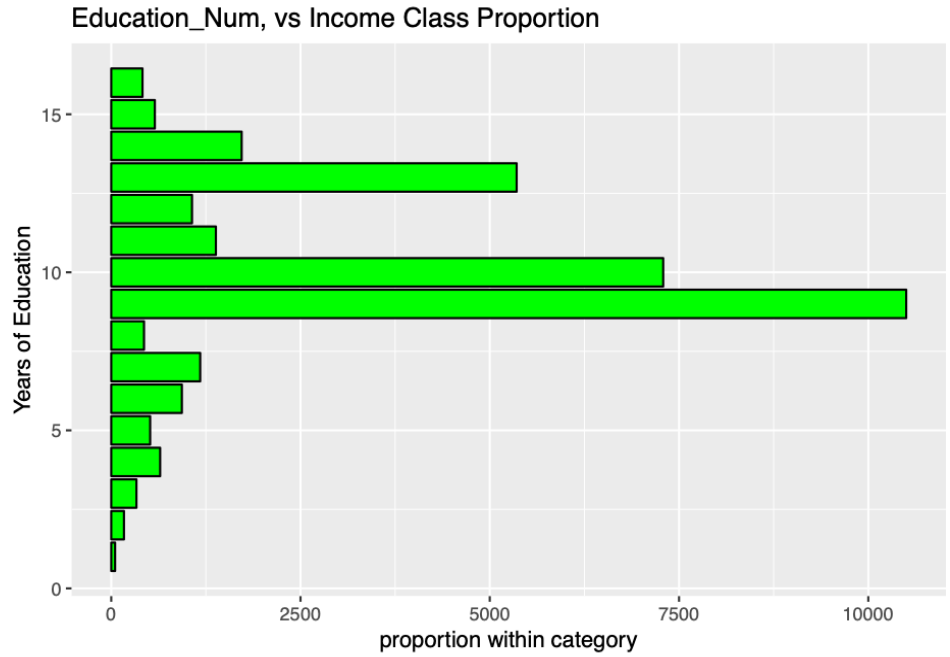



```
R <- train %>% ggplot(aes(x = Work_Class, fill = IncomeCLASS)) + geom_bar(position = "fill", color = "bl
R1 <- R + labs(title = "Work_Class vs Income Class Proportion")
R1
```

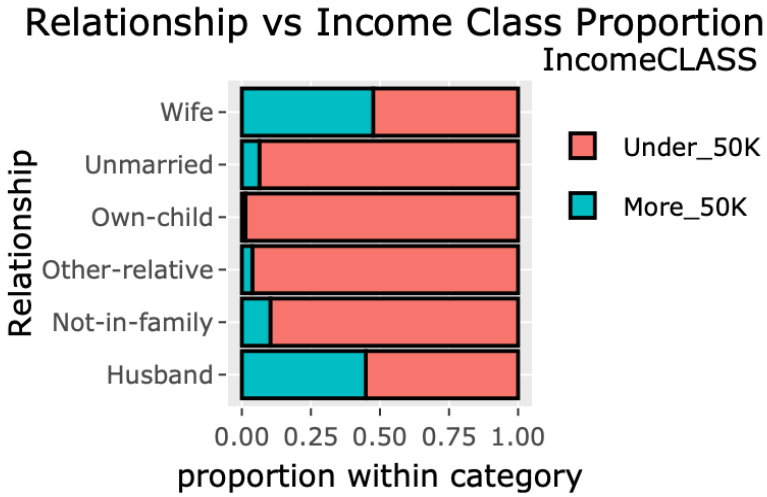


We find that the people employed in private companies have more people with income above 50k and Self Employed people having a higher proportion of peoplw with income greater than 50k.

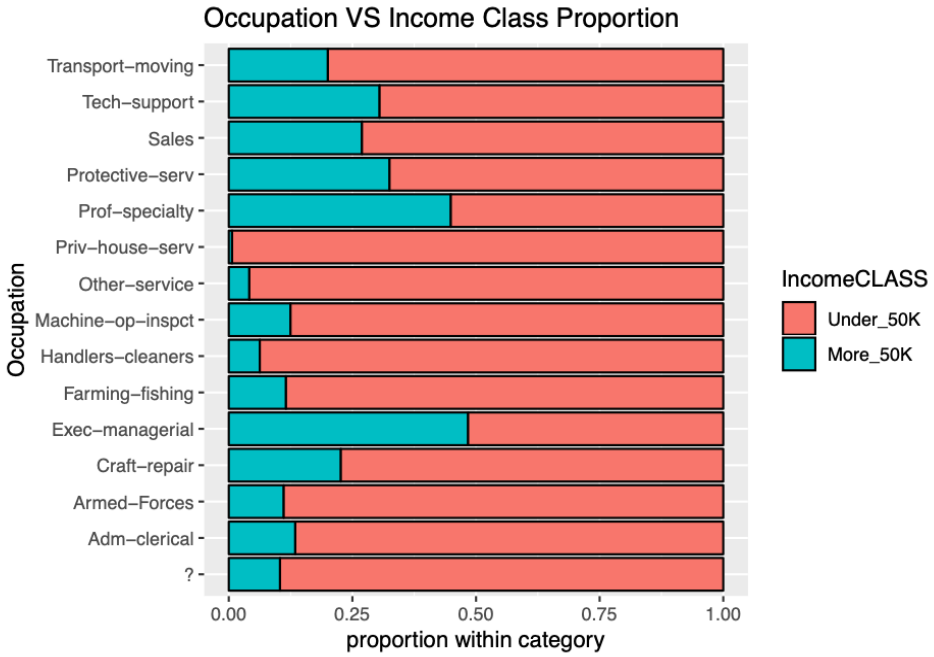
```
S <- ggplot(train,aes(x = Education_Num, fill = IncomeCLASS)) + ggtitle("Length of Eduction VS Income C
S1 <- S + labs(title = "Education_Num, vs Income Class Proportion")
S1
```




```
T <- ggplot(train,aes(x = Relationship, fill = IncomeCLASS)) + ggtitle("Relationship VS Income Class Pr
T1 <- T + labs(title = "Relationship vs Income Class Proportion")
ggplotly(T1)
```




```
jj <- ggplot(train,aes(x = Occupation, fill = IncomeCLASS)) + ggtitle("Occupation VS Income Class Propo
jj
```



PERFORMANCE METRICS for MODEL SELECTION:

Accuracy Statistic: Kappa Statistic:

NAIVE BAYES MODEL

```
#train Naive Bayes
model_Naive <- naiveBayes(IncomeCLASS ~ ., data = train)
pred_Nb <- predict(model_Naive, test)
confusionMatrix(pred_Nb, test$IncomeCLASS)
```

```
## Confusion Matrix and Statistics
##
##            Reference
## Prediction  Under_50K More_50K
##   Under_50K     11563     1968
##   More_50K        872     1878
##
##                Accuracy : 0.8256
##                  95% CI : (0.8196, 0.8314)
##     No Information Rate : 0.7638
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.4638
##  Mcnemar's Test P-Value : < 2.2e-16
##
```

```
##             Sensitivity : 0.9299
##             Specificity : 0.4883
##          Pos Pred Value : 0.8546
##          Neg Pred Value : 0.6829
##              Prevalence : 0.7638
##          Detection Rate : 0.7102
##    Detection Prevalence : 0.8311
##       Balanced Accuracy : 0.7091
##
##        'Positive' Class : Under_50K
##
```

```
CrossTable(pred_Nb, test$IncomeCLASS)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  16281
##
##
##              | test$IncomeCLASS
##      pred_Nb | Under_50K |  More_50K | Row Total |
## -------------|-----------|-----------|-----------|
##    Under_50K |     11563 |      1968 |     13531 |
##              |   146.006 |   472.069 |           |
##              |     0.855 |     0.145 |     0.831 |
##              |     0.930 |     0.512 |           |
##              |     0.710 |     0.121 |           |
## -------------|-----------|-----------|-----------|
##     More_50K |       872 |      1878 |      2750 |
##              |   718.400 |  2322.753 |           |
##              |     0.317 |     0.683 |     0.169 |
##              |     0.070 |     0.488 |           |
##              |     0.054 |     0.115 |           |
## -------------|-----------|-----------|-----------|
## Column Total |     12435 |      3846 |     16281 |
##              |     0.764 |     0.236 |           |
## -------------|-----------|-----------|-----------|
##
##
```

```
summary(pred_Nb)
```

```
## Under_50K  More_50K
##     13531      2750
```

```r
cm_Nb <- data.frame(confusionMatrix(pred_Nb, test$IncomeCLASS)[3])
kable(cm_Nb)
```

	overall
Accuracy	0.8255635
Kappa	0.4638227
AccuracyLower	0.8196457
AccuracyUpper	0.8313648
AccuracyNull	0.7637737
AccuracyPValue	0.0000000
McnemarPValue	0.0000000

TRAIN THE RPART DECISION TREE MODEL

```r
# rpart decision tree
set.seed(32323)
V <- 10
T <- 4
TrControl <- trainControl(method = "repeatedcv",
                          number = V,
                          repeats = T)


model_part <- caret::train(IncomeCLASS ~., data = train, method = "rpart",  control = rpart::rpart.cont
pred_rpart <- predict(model_part, test, type = "raw")

confusionMatrix(pred_rpart, test$IncomeCLASS)
```

```
## Confusion Matrix and Statistics
##
##            Reference
## Prediction  Under_50K More_50K
##   Under_50K     11803     1895
##   More_50K        632     1951
##
##                Accuracy : 0.8448
##                  95% CI : (0.8391, 0.8503)
##     No Information Rate : 0.7638
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5148
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9492
##             Specificity : 0.5073
##          Pos Pred Value : 0.8617
##          Neg Pred Value : 0.7553
##              Prevalence : 0.7638
##          Detection Rate : 0.7250
##    Detection Prevalence : 0.8413
##       Balanced Accuracy : 0.7282
```

```
##
##        'Positive' Class : Under_50K
##
```

```
CrossTable(pred_rpart, test$IncomeCLASS)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  16281
##
##
##              | test$IncomeCLASS
##   pred_rpart | Under_50K |  More_50K | Row Total |
## -------------|-----------|-----------|-----------|
##    Under_50K |     11803 |      1895 |     13698 |
##              |   171.840 |   555.598 |           |
##              |     0.862 |     0.138 |     0.841 |
##              |     0.949 |     0.493 |           |
##              |     0.725 |     0.116 |           |
## -------------|-----------|-----------|-----------|
##     More_50K |       632 |      1951 |      2583 |
##              |   911.290 |  2946.410 |           |
##              |     0.245 |     0.755 |     0.159 |
##              |     0.051 |     0.507 |           |
##              |     0.039 |     0.120 |           |
## -------------|-----------|-----------|-----------|
## Column Total |     12435 |      3846 |     16281 |
##              |     0.764 |     0.236 |           |
## -------------|-----------|-----------|-----------|
##
##
```

```
summary(pred_rpart)
```

```
## Under_50K  More_50K
##     13698      2583
```

```
model_part$finalModel
```

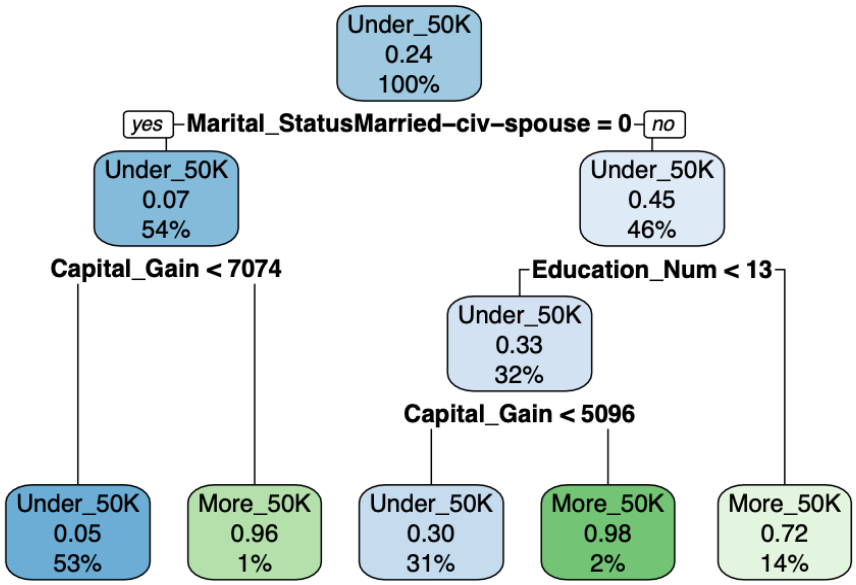
```
## n= 32561
##
## node), split, n, loss, yval, (yprob)
##       * denotes terminal node
##
##  1) root 32561 7841 Under_50K (0.75919044 0.24080956)
##    2) Marital_StatusMarried-civ-spouse< 0.5 17585 1149 Under_50K (0.93466022 0.06533978)
```

```
##      4) Capital_Gain< 7073.5 17274  849 Under_50K (0.95085099 0.04914901) *
##      5) Capital_Gain>=7073.5 311   11 More_50K (0.03536977 0.96463023) *
##    3) Marital_StatusMarried-civ-spouse>=0.5 14976 6692 Under_50K (0.55315171 0.44684829)
##      6) Education_Num< 12.5 10507 3478 Under_50K (0.66898258 0.33101742)
##       12) Capital_Gain< 5095.5 9979 2961 Under_50K (0.70327688 0.29672312) *
##       13) Capital_Gain>=5095.5 528   11 More_50K (0.02083333 0.97916667) *
##      7) Education_Num>=12.5 4469 1255 More_50K (0.28082345 0.71917655) *
```

```
cm_rpart <- data.frame(confusionMatrix(pred_rpart, test$IncomeCLASS)[3])
kable(cm_rpart)
```

	overall
Accuracy	0.8447884
Kappa	0.5148460
AccuracyLower	0.8391340
AccuracyUpper	0.8503193
AccuracyNull	0.7637737
AccuracyPValue	0.0000000
McnemarPValue	0.0000000

```
rpart.plot(model_part$finalModel)
```



TRAIN THE RANDOM FOREST MODEL

```
model.rf <- randomForest(IncomeCLASS~., data = train, ntree = 750, importance = TRUE)
pred.rf <- predict(model.rf, test)
summary(pred.rf)
```

```
plot(model.rf)
confusionMatrix(test$IncomeCLASS, pred.rf)
cm.rf <- data.frame(confusionMatrix(pred.rf, test$IncomeCLASS)[3])
kable(cm.rf)
```

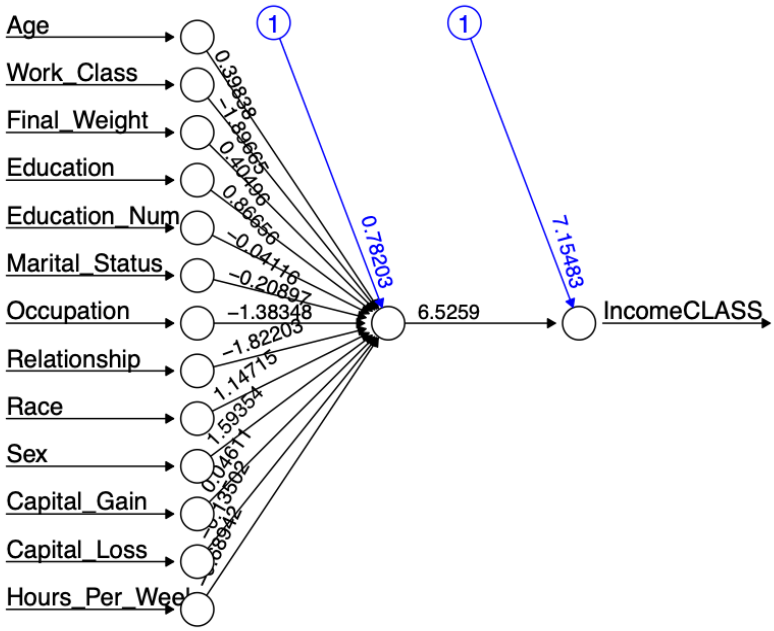
NEURAL NETWORKS

```
train_n <- train %>% sapply(as.numeric)
train_n <- as.data.frame(train_n)

model_neural <-  neuralnet(IncomeCLASS ~ ., data = train_n, hidden = 1, rep = 5, act.fct = "logistic", l
1
```

```
## [1] 1
```

```
plot(model_neural, rep = "best")
```



summarize results

```
summary(model_neural)
```

```
##                   Length Class      Mode
## call                   7 -none-     call
## response           32561 -none-     numeric
```

```
## covariate           423293 -none-     numeric
## model.list               2 -none-     list
## err.fct                  1 -none-     function
## act.fct                  1 -none-     function
## linear.output            1 -none-     logical
## data                    14 data.frame list
## exclude                  0 -none-     NULL
## net.result               5 -none-     list
## weights                  5 -none-     list
## generalized.weights      5 -none-     list
## startweights             5 -none-     list
## result.matrix           95 -none-     numeric
```

```
test_n <- as.matrix(sapply(test, as.numeric))
test_n <- as.data.frame(test_n)
model_pred <- compute(model_neural, test_n)
pr.nn <- model_pred$net.result
```

Accuracy (test set)

```
for (i in length(train))
original_values <- test_n[,14]
pr.nn_2 <- max.col(pr.nn)
outs <- mean(pr.nn_2 == original_values)
outs
```

```
## [1] 0.7637737
```

FINAL MODEL COMPARISON

RPART

```
kable(cm_rpart)
```

	overall
Accuracy	0.8447884
Kappa	0.5148460
AccuracyLower	0.8391340
AccuracyUpper	0.8503193
AccuracyNull	0.7637737
AccuracyPValue	0.0000000
McnemarPValue	0.0000000

RANDOM FOREST

```
####                    overall|
####|:--------------|---------:|
####|Accuracy       | 0.8643818|
####|Kappa          | 0.6007945|
```

```
####|AccuracyLower  | 0.8590268|
####|AccuracyUpper  | 0.8696063|
####|AccuracyNull   | 0.7637737|
####|AccuracyPValue | 0.0000000|
####|McnemarPValue  | 0.0000000|
```

NEURAL NETWORK

```
cor(max.col(pr.nn), test_n[,14])
```

```
## Warning in cor(max.col(pr.nn), test_n[, 14]): the standard deviation is
## zero
```

```
## [1] NA
```

```
Accuracy <- mean(pr.nn_2 == original_values)
Accuracy
```

```
## [1] 0.7637737
```

CONCLUSIONS:

The analysis confirmed (and quantified) what is considered common sense:

Age, education, occupation, and marital status (or relationship kind) are good for predicting income (above a certain threshold).

(1) if a person earns more than $50000 he is very likely to be a married man with large number of years of education;
(2) single parents, younger than 25 years, who studied less than 10 years, and were never-married make less than $50000.

Inferences

About 46% of the people are in a relationship called “Husband” or “Wife” which is then further classified based on Education Level where nearly 14% who earn above $50 K have the education of Bachelors, Prof-school, Masters and Doctorate.

The other education levels have income predominantly below $50 k with just 2% having salaries above $50k who also have capital gains greater than $5096

With respect to other relationships, only 1% have income above $50 k and with capital gains greater than $7074.

In the relationship of Education and Number of People Earning $>$ 50 k and separated by Work Class. We find that Bachelors graduates working in Private companies have a higher number of people earning above 50 k.

In the relationship of Average hours per week with respect to gender and separated by Work Class and we find that Males typically work more hours per week on Average across all work classes.

The third sheet shows that the relationship of marital status and income levels separated by Work Class. Majority of the people in Married with Civilian spouse have an income greater than 50 k and majorly in the private sector.

The fourth work sheet shows the impact of occupation, capital gain and capital loss on the income levels which has details of work class too. This is a comprehensive visualization across 4 different parameters. We find that Executives at Managerial Level have more people with income greater than 50 k and Professional Speciality has more capital gains.