Salary Analysis

Rory Quinlan

```r
library(tidyverse)
library(cluster)
library(dplyr)
library(ggplot2)
library(rpart)
library(rpart.plot)
library(glmnet)

salary_US0 = read_csv("salary.csv", show_col_types = FALSE)

salary_US1 <- salary_US0 %>% filter(`native-country`=="United-States")
```

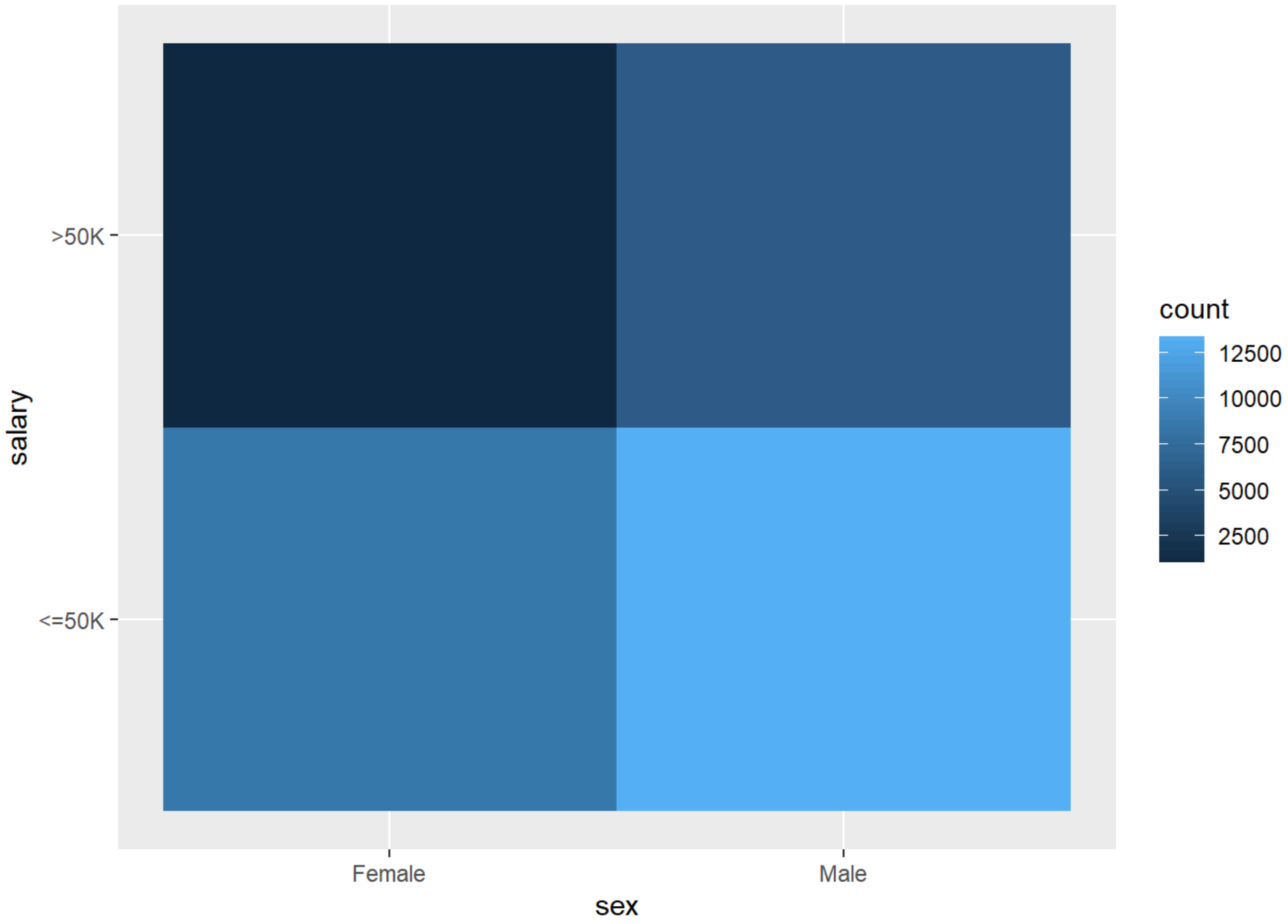
Data Exploration

```r
#summary each variable
salary_US2 <- as.data.frame(unclass(salary_US1), stringsAsFactors=TRUE)
summary(salary_US2)
```

```
##       age                   workclass         fnlwgt               education
##  Min.   :17.00   Private         :20135   Min.   :  12285   HS-grad     :9702
##  1st Qu.:28.00   Self-emp-not-inc: 2313   1st Qu.: 115895   Some-college:6740
##  Median :37.00   Local-gov       : 1956   Median : 176730   Bachelors   :4766
##  Mean   :38.66   ?               : 1659   Mean   : 187069   Masters     :1527
##  3rd Qu.:48.00   State-gov       : 1210   3rd Qu.: 234139   Assoc-voc   :1289
##  Max.   :90.00   Self-emp-inc    :  991   Max.   :1484705   11th        :1067
##                  (Other)         :  906                     (Other)     :4079
##  education.num                 marital.status            occupation
##  Min.   : 1.00   Divorced             : 4162   Exec-managerial:3735
##  1st Qu.: 9.00   Married-AF-spouse    :   23   Prof-specialty :3693
##  Median :10.00   Married-civ-spouse   :13368   Craft-repair   :3685
##  Mean   :10.17   Married-spouse-absent:  253   Adm-clerical   :3449
##  3rd Qu.:12.00   Never-married        : 9579   Sales          :3364
##  Max.   :16.00   Separated            :  883   Other-service  :2777
##                  Widowed              :  902   (Other)        :8467
##         relationship                   race          sex
##  Husband       :11861   Amer-Indian-Eskimo:  296   Female: 9682
##  Not-in-family : 7528   Asian-Pac-Islander:  292   Male  :19488
##  Other-relative:  696   Black             : 2832
##  Own-child     : 4691   Other             :  129
##  Unmarried     : 3033   White             :25621
##  Wife          : 1361
##
##   capital.gain    capital.loss     hours.per.week        native.country
##  Min.   :    0   Min.   :   0.00   Min.   : 1.00   United-States:29170
##  1st Qu.:    0   1st Qu.:   0.00   1st Qu.:40.00
##  Median :    0   Median :   0.00   Median :40.00
##  Mean   : 1089   Mean   :  88.51   Mean   :40.45
##  3rd Qu.:    0   3rd Qu.:   0.00   3rd Qu.:45.00
##  Max.   :99999   Max.   :4356.00   Max.   :99.00
##
##    salary
##  <=50K:21999
##  >50K : 7171
##
##
##
##
##
```

```
salary_US3 <- salary_US2 %>% select(-c(native.country,fnlwgt,education,relationship))
```

```
ggplot(data=salary_US3,mapping=aes(x=sex,y=salary))+geom_bin2d()
```



###

Logistics Regression

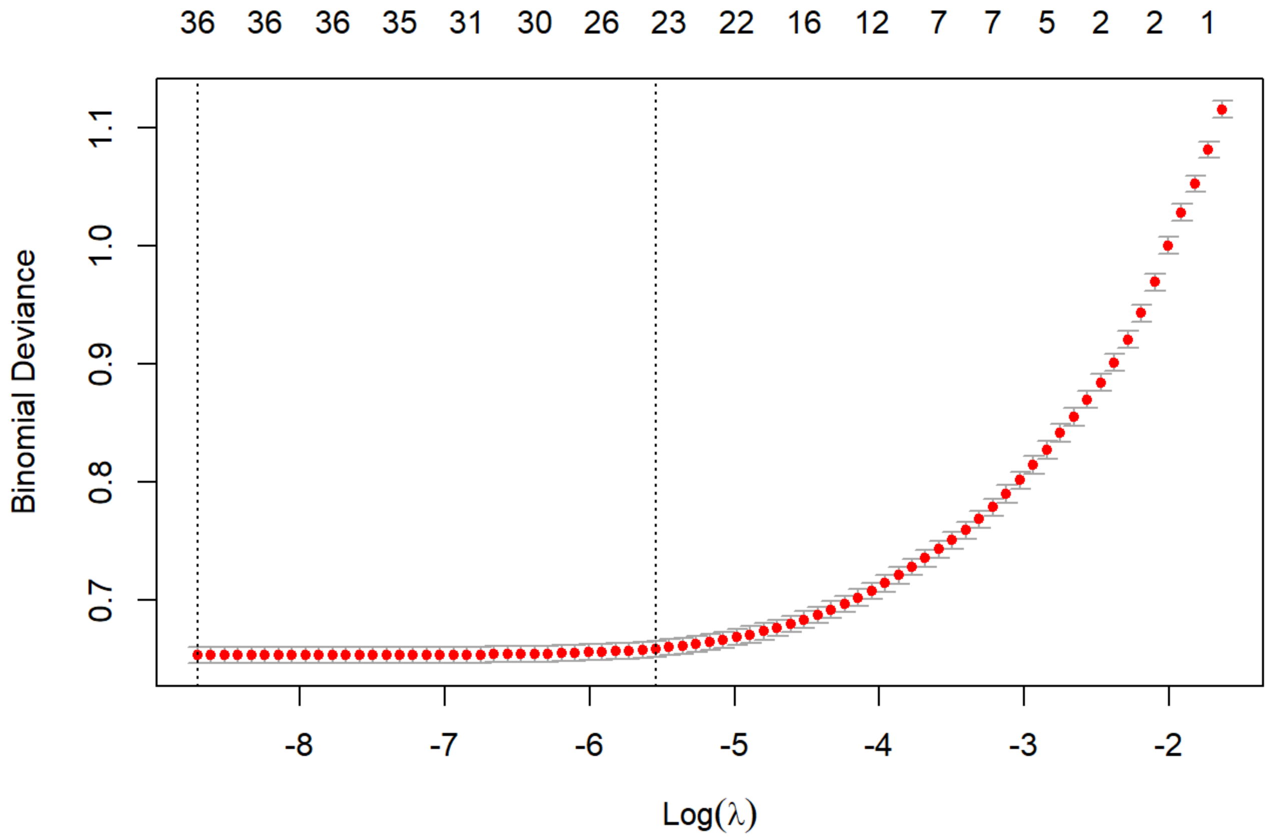
```
# Create for for regression
form_full <- as.formula("salary~age+workclass+education.num+
 marital.status+occupation+race+sex+capital.gain+
 capital.loss+hours.per.week")
set.seed(99)

# Split data for training and testing
train6 <- salary_US3 %>% sample_frac(size = 0.8)
test6 <- salary_US3 %>% setdiff(train6)

# Find best lambda with 5 fold cross validation
predictors <- model.matrix(form_full, data = train6)
fit1 <- cv.glmnet(predictors, train6$salary, family = "binomial")
fit1$lambda.1se
```

```
## [1] 0.003916005
```

```
# Plot fit
plot(fit1)
```



```
# Fit model with predictors, data, and binomial model

fit2 <- glmnet(predictors, train6$salary, family = "binomial", lambda = 0.004)
fit2
```

```
##
## Call:  glmnet(x = predictors, y = train6$salary, family = "binomial",      lambda = 0.004)
##
##   Df  %Dev Lambda
## 1 23 41.16  0.004
```

```
# Create function to return misclass rate
logistic.misclassrate <- function(dataset, y, fit, form){
 misclass_lr <- dataset %>%
 mutate(pred.logistic = predict(fit, newx = model.matrix(form, data = dataset),
 type = "class")) %>%
 mutate(misclassify = ifelse(y != pred.logistic, 1,0)) %>%
 summarize(misclass.rate = mean(misclassify))
return(misclass_lr$misclass.rate)
}


logistic.misclassrate(test6,test6$salary,fit2,form_full)
```

```
## [1] 0.1677222
```

Lambda min model

```
# Find lambda min
fit1$lambda.min
```

```
## [1] 0.0001656173
```

```
# Fit logistic regression with lambda
fit3 <- glmnet(predictors, train6$salary, family = "binomial", lambda = 0.0001)
logistic.misclassrate(test6,test6$salary,fit3,form_full)
```

```
## [1] 0.1635728
```

- The lambda min model has a lower misclassification rate and is the better model
- Can we improve this by trying another possible model?

Decision Tree

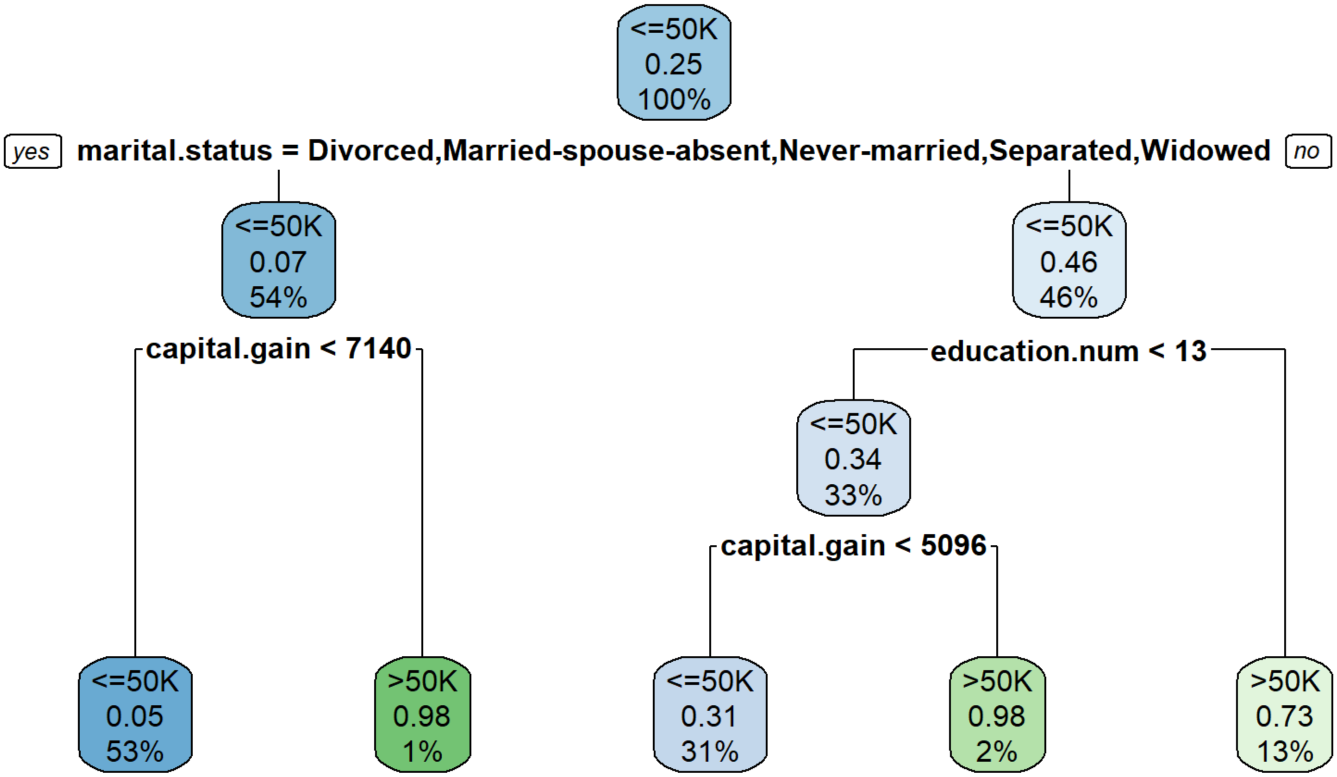
```
set.seed(99)

# Split the data
train1 <- salary_US3 %>% sample_frac(size = 0.8)
test1 <- salary_US3 %>% setdiff(train1)
library(glmnet)
```

```
# Select form for tree
form<- as.formula(
"salary ~sex+age+workclass+education.num+
 marital.status+occupation+race+sex+capital.gain+
 capital.loss+hours.per.week")

# Select form and data for model
mod_lr2 <- glm(form, data=train1,family=binomial)
```

```
# Fit and plot model
mod_tree <- rpart(form,data=train1)
rpart.plot(mod_tree)
```



```
prop.table(table(salary_US3$salary))
```

```
##
##     <=50K      >50K
## 0.7541652 0.2458348
```

```
confusMatrix <- function (data, y, mod)
 { confMatrix <- data %>%
 mutate(pred = predict(mod, newdata = data, type ="class"),y=y) %>%
 select (y, pred) %>% table() }
misclass <- function(confusion) {
misclass <- 1- sum(diag(confusion))/sum(confusion)
return(misclass)}
cMat <- confusMatrix(salary_US3, salary_US3$salary, mod_tree)
cMat
```

```
##        pred
## y       <=50K  >50K
##   <=50K 20931  1068
##   >50K   3542  3629
```

```
Rates<-c("Misclass", "True Positive", "True Negative")
Values<-c( misclass(cMat),cMat[1,1]/sum(cMat[,1]), cMat[2,2]/sum(cMat[,2]))

cbind(Rates,Values)
```

```
##      Rates           Values             
## [1,] "Misclass"      "0.158039081247857"
## [2,] "True Positive" "0.855269072038573"
## [3,] "True Negative" "0.77262082180115" 
```