Logistic Regression and Resampling using k-fold validation

Gosuddin Siddiqi

February 16, 2017

```r
#load the datasetset
dataset <- read.csv("census data.csv")
```

Creating a new column

```r
#initialize with 0
dataset$income.g50 <- 0


#based on value decide 0 or 1
dataset$income.g50[dataset$income == " >50K"] <- 1
```

Exploring Relationship

```r
mod <- glm(income.g50 ~ education + age + sex + race,
data=dataset[,!colnames(dataset)%in%"income"], family="binomial")
```

a. What are the odds ratios for high earnings (remember the output of summary() gives log odds ratios) for having a masters degree? Or a 1st - 4th grade education? Are these statistically significant? What about multiple comparisons?

```r
summary(mod)
```

```
##
## Call:
## glm(formula = income.g50 ~ education + age + sex + race, family = "binomial",
##     data = dataset[, !colnames(dataset) %in% "income"])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4997  -0.6802  -0.4460  -0.1114   2.8328
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -6.865618   0.504710 -13.603  < 2e-16 ***
## education 11th            0.296538   0.327880   0.904  0.36578
## education 12th            0.724955   0.386826   1.874  0.06092 .
## education 1st-4th        -0.180178   0.656420  -0.274  0.78371
## education 5th-6th        -1.065786   0.638832  -1.668  0.09525 .
## education 7th-8th        -0.055193   0.344602  -0.160  0.87275
## education 9th            -0.472650   0.422347  -1.119  0.26310
## education Assoc-acdm      1.793180   0.276081   6.495 8.30e-11 ***
## education Assoc-voc       1.806001   0.265792   6.795 1.08e-11 ***
## education Bachelors       2.498991   0.246065  10.156  < 2e-16 ***
## education Doctorate       3.465742   0.316322  10.956  < 2e-16 ***
## education HS-grad         1.099579   0.244388   4.499 6.82e-06 ***
## education Masters         2.910088   0.256902  11.328  < 2e-16 ***
## education Preschool     -10.727247 130.041047  -0.082  0.93426
```

```
## education Prof-school     3.834590   0.308031  12.449  < 2e-16 ***
## education Some-college    1.590668   0.246311   6.458 1.06e-10 ***
## age                       0.043369   0.002061  21.043  < 2e-16 ***
## sex Male                  1.291684   0.066444  19.440  < 2e-16 ***
## race Asian-Pac-Islander   1.009867   0.457049   2.210  0.02714 *
## race Black                1.119303   0.442857   2.527  0.01149 *
## race Other                0.213828   0.654001   0.327  0.74370
## race White                1.392564   0.431938   3.224  0.00126 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 11151.2  on 9999  degrees of freedom
## Residual deviance:  8900.2  on 9978  degrees of freedom
## AIC: 8944.2
##
## Number of Fisher Scoring iterations: 12
```

The odds ratio for higher income is exp(2.91) i.e. 18.3567986 when a person has a master's degree. The p values is <0.05 thus it is statistically significant.

The odds ratio for higher income is exp(-0.1801) i.e. 0.8344354 when a person has a master's degree. The p-value is >0.05 thus it is not statistically significant.

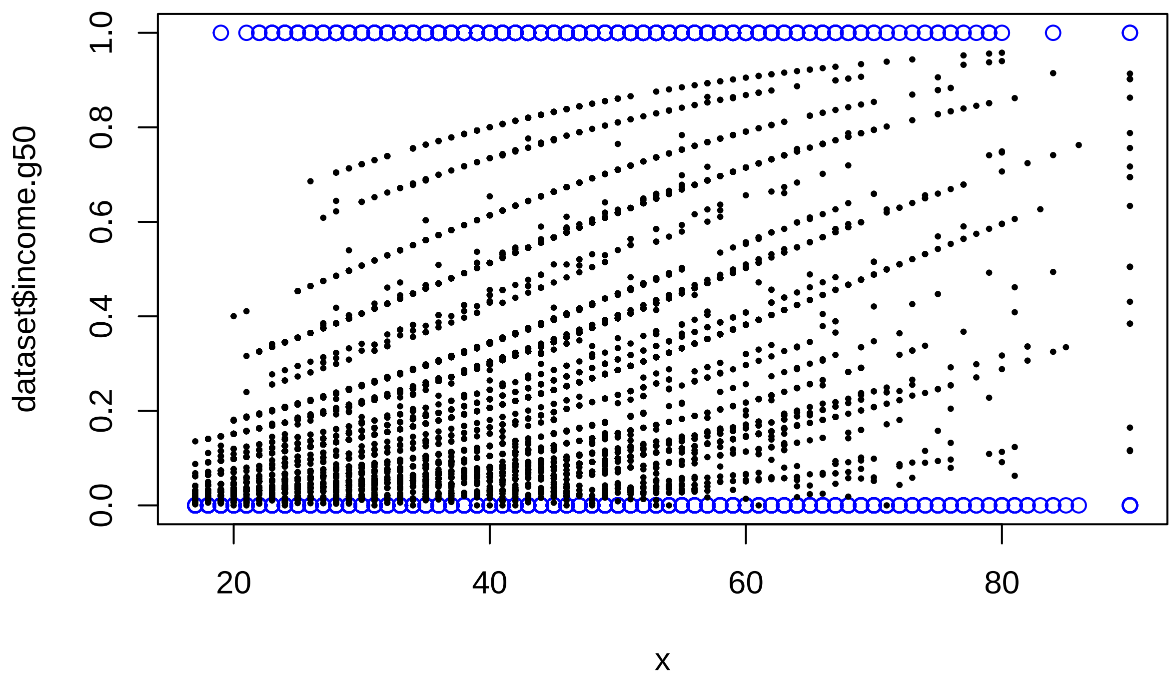
There are predictors which has p-value in the order of -16 and for them I can be confident that with multiple comparisons they would be significant.

b. What are the effects of age and sex? Again, are they statistically significant? Are they practically significant? Are they fair? Age: odd ratio is 1.0442511. Here it is fair because income increases with experience.

Sex: Odd ratio is 3.6364212 Being male increases the chances on earning higher incomes. Here it is unfair because there is discrimination based on sex.

3. Exploring Relationships II: Plot age by the outcome and the observed predicted probabilities. Why are the predicted probabilities so variable?

```r
x <- dataset$age
plot(x, dataset$income.g50, col="blue")
fits <- fitted(mod)
points(x, fits, pch=19, cex=0.3)
```



Since we have increased the number of features we observe variablity in the outcome

4. Explore some cutoffs for the probabilities: Tabulate the outcome with a cutoff of 0.25, 0.5, and 0.75. Which has the lowest percent error?

```
tab <- table(dataset$income.g50, fits>=0.25)
(tab[1,2]+tab[2,1])/sum(tab)
```

```
## [1] 0.2662
```

```
tab <- table(dataset$income.g50, fits>=0.5)
(tab[1,2]+tab[2,1])/sum(tab)
```

```
## [1] 0.2061
```

```
tab <- table(dataset$income.g50, fits>=0.75)
(tab[1,2]+tab[2,1])/sum(tab)
```

```
## [1] 0.2307
```

The code with cutoff as 0.5 has the lowest percent error of 20.61%

5. Examine this model.

a. Plot the ROC curve and calculate the AUC for this model.

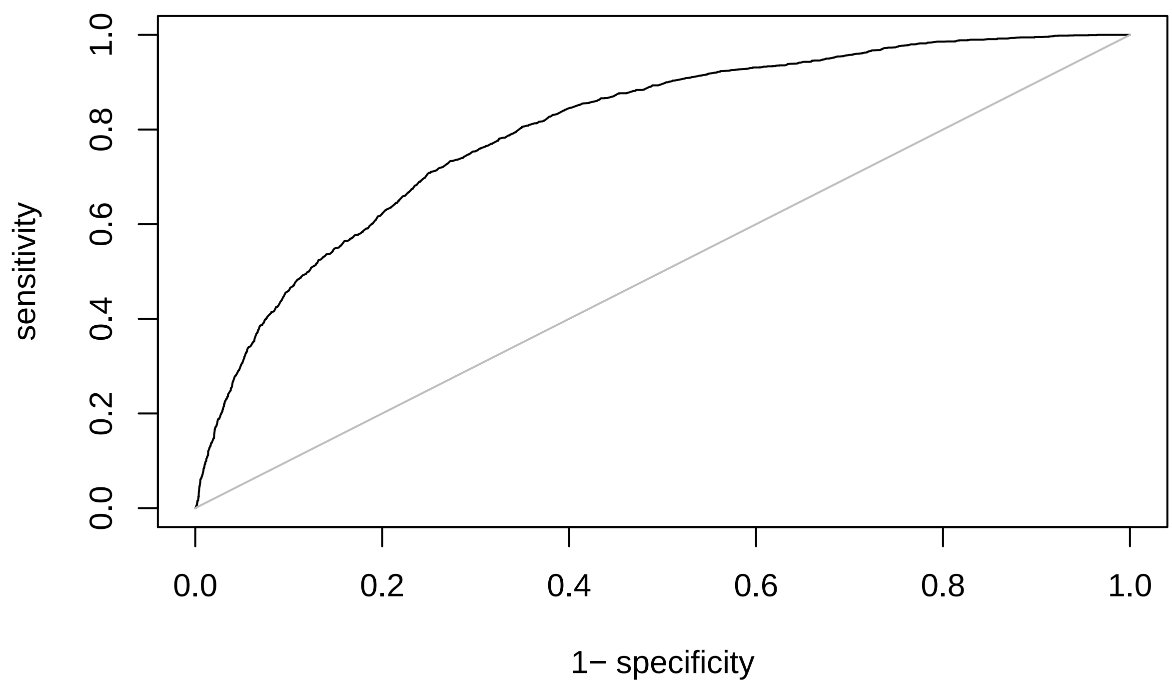
```
library(AUC)
```

```
## Warning: package 'AUC' was built under R version 3.2.5
```

```
## AUC 0.3.0
```

```
## Type AUCNews() to see the change log and ?AUC to get an overview.
```

```
y <- factor(dataset$income.g50)
rr <- roc(fits, y)
plot(rr)
```




```
auc(rr)
```

```
## [1] 0.8021133
```

b. How well does it fit?

The area under curve is around 80%. Implies a decent fit

6. Let's formulate another model.

a. Fit a model with all covariates (except "income"!). Do you see the same patterns for level of schooling?

```
mod <- glm(income.g50~.,
data=dataset[,!colnames(dataset)%in%c("income")], family="binomial")
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
mod
```

```
##
## Call:  glm(formula = income.g50 ~ ., family = "binomial", data = dataset[,
##     !colnames(dataset) %in% c("income")])
##
## Coefficients:
##                                (Intercept)
```

```
##                                -9.931e+00
##                                       age
##                                 2.801e-02
##                    work.class Federal-gov
##                                 1.071e+00
##                      work.class Local-gov
##                                 1.906e-01
##                   work.class Never-worked
##                                -1.359e+01
##                        work.class Private
##                                 3.800e-01
##                   work.class Self-emp-inc
##                                 8.081e-01
##               work.class Self-emp-not-inc
##                                 2.786e-01
##                      work.class State-gov
##                                 2.782e-01
##                    work.class Without-pay
##                                -1.459e+01
##                              final.weight
##                                 1.082e-06
##                            education 11th
##                                 5.089e-01
##                            education 12th
##                                 6.785e-01
##                         education 1st-4th
##                                 2.756e-01
##                         education 5th-6th
##                                -7.479e-01
##                         education 7th-8th
##                                 1.145e-01
##                             education 9th
##                                -2.324e-01
##                      education Assoc-acdm
##                                 1.466e+00
##                       education Assoc-voc
##                                 1.473e+00
##                       education Bachelors
##                                 2.092e+00
##                       education Doctorate
##                                 2.892e+00
##                         education HS-grad
##                                 9.690e-01
##                         education Masters
##                                 2.429e+00
##                       education Preschool
##                                -1.309e+01
##                     education Prof-school
##                                 3.346e+00
##                    education Some-college
##                                 1.412e+00
##                              years.school
##                                        NA
##            marital.status Married-AF-spouse
```

```
##                                 2.779e+00
##        marital.status Married-civ-spouse
##                                 2.571e+00
##     marital.status Married-spouse-absent
##                                 1.422e-01
##             marital.status Never-married
##                                -5.452e-01
##                 marital.status Separated
##                                -2.253e-01
##                   marital.status Widowed
##                                 2.295e-01
##                  occupation Adm-clerical
##                                 1.308e-01
##                  occupation Armed-Forces
##                                 5.030e-01
##                  occupation Craft-repair
##                                 3.629e-01
##               occupation Exec-managerial
##                                 9.648e-01
##               occupation Farming-fishing
##                                -8.811e-01
##             occupation Handlers-cleaners
##                                -5.332e-01
##             occupation Machine-op-inspct
##                                -4.981e-02
##                 occupation Other-service
##                                -7.228e-01
##               occupation Priv-house-serv
##                                -1.482e+01
##                occupation Prof-specialty
##                                 8.118e-01
##               occupation Protective-serv
##                                 7.941e-01
##                        occupation Sales
##                                 5.718e-01
##                  occupation Tech-support
##                                 8.180e-01
##              occupation Transport-moving
##                                        NA
##               relationship Not-in-family
##                                 1.066e+00
##              relationship Other-relative
##                                 9.728e-02
##                   relationship Own-child
##                                -1.158e-01
##                   relationship Unmarried
##                                 8.553e-01
##                        relationship Wife
##                                 1.414e+00
##                  race Asian-Pac-Islander
##                                 1.297e+00
##                               race Black
##                                 1.582e+00
##                               race Other
```

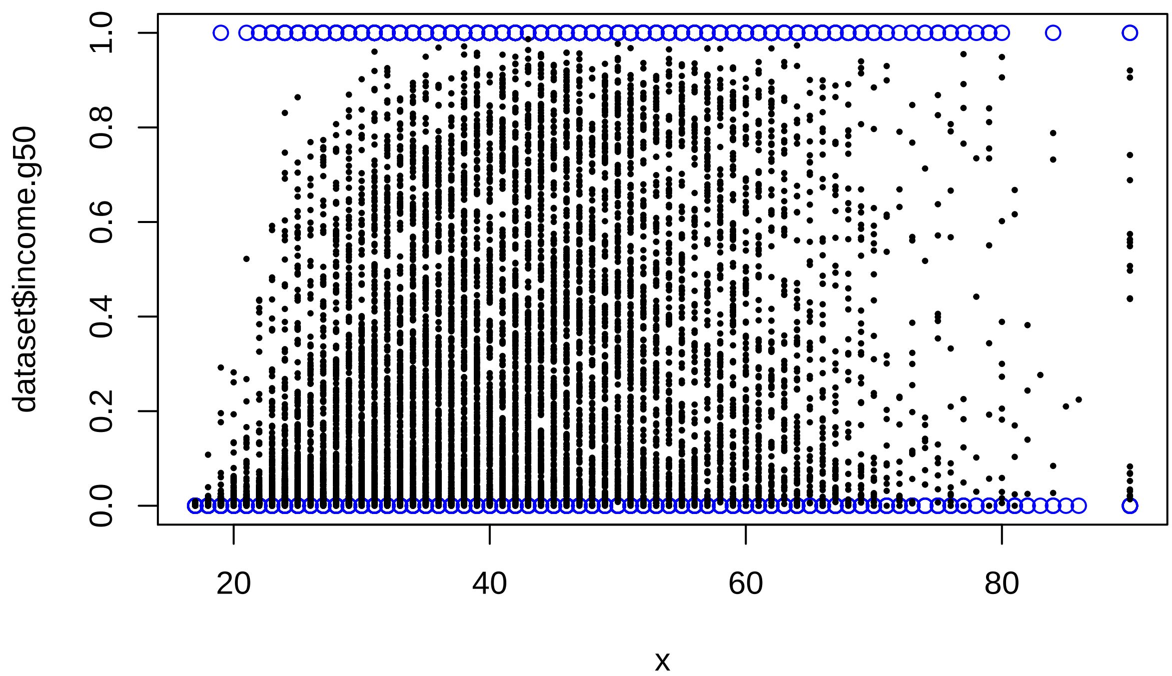
```
##                                   6.241e-01
##                                  race White
##                                   1.477e+00
##                                    sex Male
##                                   8.170e-01
##                              hours.per.week
##                                   2.709e-02
##                    native.country Cambodia
##                                  -1.570e+01
##                      native.country Canada
##                                  -5.204e-02
##                       native.country China
##                                  -6.827e-01
##                    native.country Columbia
##                                  -1.466e+01
##                        native.country Cuba
##                                   6.734e-02
##          native.country Dominican-Republic
##                                   4.164e-01
##                     native.country Ecuador
##                                  -1.097e+00
##                 native.country El-Salvador
##                                  -1.070e+00
##                     native.country England
##                                  -8.452e-02
##                      native.country France
##                                  -1.111e-01
##                     native.country Germany
##                                  -8.578e-02
##                      native.country Greece
##                                  -2.564e-01
##                   native.country Guatemala
##                                  -1.296e+01
##                       native.country Haiti
##                                   1.956e-01
##          native.country Holand-Netherlands
##                                  -1.142e+01
##                    native.country Honduras
##                                  -1.226e+01
##                        native.country Hong
##                                  -9.191e-01
##                     native.country Hungary
##                                  -6.953e-01
##                       native.country India
##                                  -7.849e-01
##                        native.country Iran
##                                  -7.106e-01
##                     native.country Ireland
##                                   2.469e+00
##                       native.country Italy
##                                   3.133e-01
##                     native.country Jamaica
##                                  -1.495e+00
##                       native.country Japan
```

```
##                                  5.607e-01
##                         native.country Laos
##                                 -1.290e+01
##                       native.country Mexico
##                                 -5.061e-01
##                    native.country Nicaragua
##                                 -1.277e+01
## native.country Outlying-US(Guam-USVI-etc)
##                                 -1.372e+01
##                         native.country Peru
##                                 -6.137e-01
##                  native.country Philippines
##                                  4.936e-01
##                       native.country Poland
##                                 -6.022e-01
##                     native.country Portugal
##                                 -5.827e-01
##                  native.country Puerto-Rico
##                                 -2.545e-01
##                     native.country Scotland
##                                  1.799e+00
##                        native.country South
##                                  1.198e+00
##                       native.country Taiwan
##                                  8.449e-01
##                     native.country Thailand
##                                 -1.556e+01
##              native.country Trinadad&Tobago
##                                 -1.489e+01
##                native.country United-States
##                                  9.913e-02
##                      native.country Vietnam
##                                 -1.438e+01
##                   native.country Yugoslavia
##                                  9.149e-01
##
## Degrees of Freedom: 9999 Total (i.e. Null);  9903 Residual
## Null Deviance:       11150
## Residual Deviance: 6998  AIC: 7192
```

Not for all schooling follows the same pattern

b. Plot the age by the outcome and the observed predicted probabilities. Do the predicted probabilities have the same pattern as the other model? Why or why not?

```r
x <- dataset$age
plot(x, dataset$income.g50, col="blue")
fits <- fitted(mod)
points(x, fits, pch=19, cex=0.3)
```



The pattern is not the same. This is because we are now considering more features in our model.

c. Calculate the percent error as before for cutoffs 0.25, 0.5, 0.75. Which cutoff has the lowest percent error? Does this model perform better than the other model?

```
tab <- table(dataset$income.g50, fits>=0.25)
(tab[1,2]+tab[2,1])/sum(tab)
```

```
## [1] 0.2071
```

```
tab <- table(dataset$income.g50, fits>=0.5)
(tab[1,2]+tab[2,1])/sum(tab)
```

```
## [1] 0.1659
```

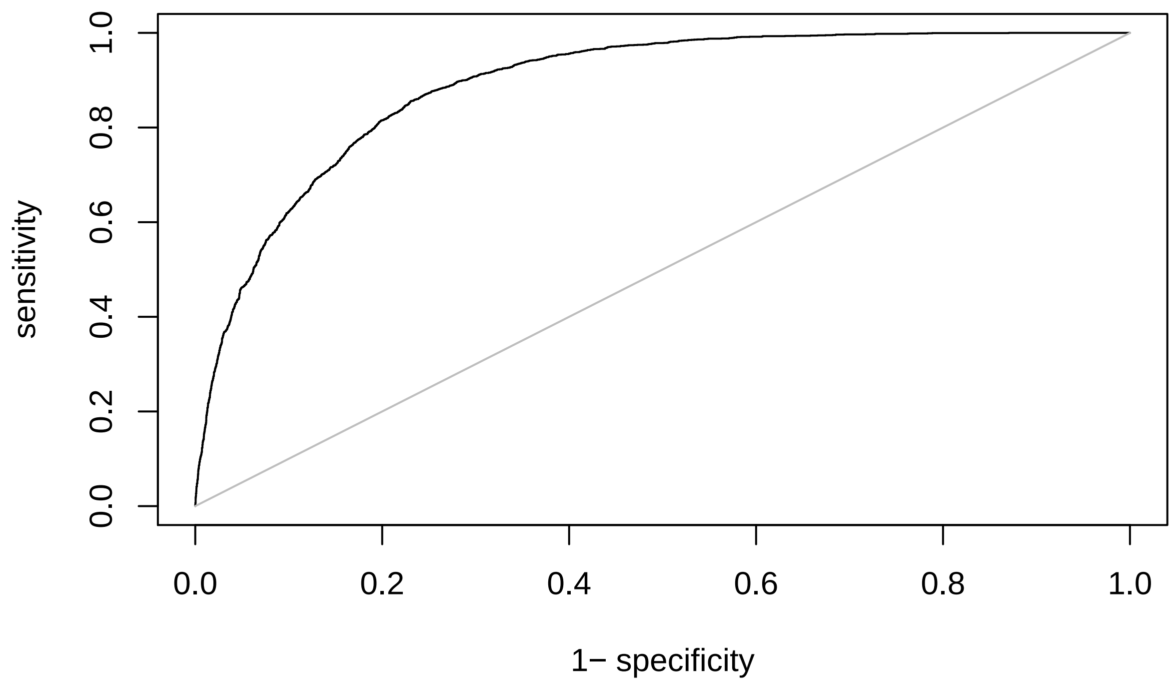
```
tab <- table(dataset$income.g50, fits>=0.75)
(tab[1,2]+tab[2,1])/sum(tab)
```

```
## [1] 0.1943
```

Yes this model outperforms the previous and has lower error rates, least at .50 cutoff

d. Plot the ROC and calculate the AUC. Again, does this model out perform the other model?

```
y <- factor(dataset$income.g50)
rr <- roc(fits, y)
plot(rr)
```




```
auc(rr)
```

```
## [1] 0.8893198
```

Yes this model outperform the other.

Extra credit (5 points): Run a k-fold validation on both models and decide which you would prefer to use for predicting high income.

```
get.cutoff <- function(fits, labs){
  youden <- sensitivity(fits, labs)$measure + specificity(fits, labs)$measure-1
  roc.ix <- which.max(youden)
  sens <- sensitivity(fits, labs)
  sens$cutoffs[roc.ix]
}
```

Considering all features.

```
set.seed(123)

k <- 10 # number of folds

acc <- NULL

#k-fold validation for the first model.
for(i in 1:k)
{
   # 95-5 split
   smp_size <- floor(0.95 * nrow(dataset))
```

```
    index <- sample(seq_len(nrow(dataset)),size=smp_size)

    #Splitting the data
    train <- dataset[index, ]
    test <- dataset[-index, ]

    # Fitting
    model <- glm(income.g50~ education + age + sex + race,family='binomial',data=train)

    # Predict results
    results_prob <- predict(model,subset(test,select=c(1:ncol(dataset)-1)),type='response')

    # If prob > 0.5 (Cutoff) then 1, else 0
    results <- ifelse(results_prob > 0.5,1,0)

    #Accuracy
    answers <- test$income.g50
    error <- mean(answers != results)

    acc[i] <- 1-error
}

mean(acc)
```

```
## [1] 0.7894
```

Considering the only education,age, sex and race

```
k <- 10 # k-fold

acc2 <- NULL


for(i in 1:k)
{
  #k-fold validation for the first model.
    smp_size2 <- floor(0.95 * nrow(dataset))
    index2 <- sample(seq_len(nrow(dataset)),size=smp_size2)

    #Splitting the data
    train <- dataset[index2, ]
    test <- dataset[-index2, ]

    # Fitting
    model2 <- glm(income.g50~.,family='binomial',data=train[,!colnames(dataset)%in%"income"])

    # Predict results
    results_prob <- predict(model2,subset(test,select=c(1:ncol(dataset)-1)),type='response')

    # If prob > 0.5 (Cutoff) then 1, else 0
    results <- ifelse(results_prob > 0.5,1,0)

    #Accuracy
    answers <- test$income.g50
    error <- mean(answers != results)
```

```
  acc2[i] <- 1-error
}

mean(acc2)
```

```
## [1] 0.8342
```

Since second model i.e. considering all variables gives better accuracy, I would prefer second model to predict higher incomes.