Classification

Aditi Chaudhari

2022-09-25

Classification

Classification is a supervised learning task that attempts to identify what class an observation falls into. To be more precise, the linear models in classification create linear boundaries to separate regions for different classes an observation can fall into. An advantage to using classification algorithms is that they help classify observations when the target variable is qualitative. However, classification algorithms are disadvantaged in that they are not as useful when our target variable is quantitative. Linear regression would be more beneficial in the latter case.

Data Exploration

Let's delve into exploring the logistic regression model!

First, data from the adult.csv file is read into a data frame. The data was obtained from https://www.kaggle.com/datasets/uciml/adult-census-income?resource=download.

```
df <- read.csv("adult.csv")
```

Firstly, let's simply see what our data looks like using the head() function, which selects the first n rows of a data frame. The target variable will be income, so understanding how the data is stored in the income variable is key.

```
head(df, n=10)
```

```
##    age   workclass fnlwgt    education education.num marital.status
## 1   90           ?  77053      HS-grad             9        Widowed
## 2   82     Private 132870      HS-grad             9        Widowed
## 3   66           ? 186061 Some-college            10        Widowed
## 4   54     Private 140359      7th-8th             4       Divorced
## 5   41     Private 264663 Some-college            10      Separated
## 6   34     Private 216864      HS-grad             9       Divorced
## 7   38     Private 150601         10th             6      Separated
## 8   74   State-gov  88638    Doctorate            16  Never-married
## 9   68 Federal-gov 422013      HS-grad             9       Divorced
## 10  41     Private  70037 Some-college            10  Never-married
##           occupation   relationship  race    sex capital.gain capital.loss
## 1                  ?  Not-in-family White Female            0         4356
## 2    Exec-managerial  Not-in-family White Female            0         4356
## 3                  ?      Unmarried Black Female            0         4356
## 4  Machine-op-inspct      Unmarried White Female            0         3900
## 5     Prof-specialty      Own-child White Female            0         3900
## 6      Other-service      Unmarried White Female            0         3770
## 7       Adm-clerical      Unmarried White   Male            0         3770
## 8     Prof-specialty Other-relative White Female            0         3683
```

```
## 9     Prof-specialty  Not-in-family White Female            0         3683
## 10      Craft-repair      Unmarried White   Male            0         3004
##    hours.per.week native.country income
## 1              40  United-States  <=50K
## 2              18  United-States  <=50K
## 3              40  United-States  <=50K
## 4              40  United-States  <=50K
## 5              40  United-States  <=50K
## 6              45  United-States  <=50K
## 7              40  United-States  <=50K
## 8              20  United-States   >50K
## 9              40  United-States  <=50K
## 10             60              ?   >50K
```

Next, let's take a look at the structure of the data frame. An important point to note is that income is of type character, but we would want it to be a factor.

```r
str(df)
```

```
## 'data.frame':    32561 obs. of  15 variables:
##  $ age           : int  90 82 66 54 41 34 38 74 68 41 ...
##  $ workclass     : chr  "?" "Private" "?" "Private" ...
##  $ fnlwgt        : int  77053 132870 186061 140359 264663 216864 150601 88638 422013 70037 ...
##  $ education     : chr  "HS-grad" "HS-grad" "Some-college" "7th-8th" ...
##  $ education.num : int  9 9 10 4 10 9 6 16 9 10 ...
##  $ marital.status: chr  "Widowed" "Widowed" "Widowed" "Divorced" ...
##  $ occupation    : chr  "?" "Exec-managerial" "?" "Machine-op-inspct" ...
##  $ relationship  : chr  "Not-in-family" "Not-in-family" "Unmarried" "Unmarried" ...
##  $ race          : chr  "White" "White" "Black" "White" ...
##  $ sex           : chr  "Female" "Female" "Female" "Female" ...
##  $ capital.gain  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ capital.loss  : int  4356 4356 4356 3900 3900 3770 3770 3683 3683 3004 ...
##  $ hours.per.week: int  40 18 40 40 40 45 40 20 40 60 ...
##  $ native.country: chr  "United-States" "United-States" "United-States" "United-States" ...
##  $ income        : chr  "<=50K" "<=50K" "<=50K" "<=50K" ...
```

We can convert the income variable to be a factor using the as.factor() function.

```r
df$income <- as.factor(df$income)
```

Now, we can randomly divide the data into a training set containing 80% of the original data and a test set containing 20% of the original data.

```r
i <- sample(1:nrow(df), nrow(df) * 0.80, replace=FALSE)
train = df[i,]
test <- df[-i,]
```

Let's take a look at the structure of the training data frame to see how the data type of the income variable has changed. It is now a factor with 2 levels, one of which is "<=50k" and the other is ">50k".

```r
str(train)
```

```
## 'data.frame':    26048 obs. of  15 variables:
##  $ age           : int  55 20 54 24 39 17 34 20 50 46 ...
##  $ workclass     : chr  "Private" "Private" "Private" "Private" ...
##  $ fnlwgt        : int  234125 227491 117674 243190 238415 193769 172928 236523 238959 177114 ...
##  $ education     : chr  "HS-grad" "HS-grad" "Masters" "Assoc-acdm" ...
##  $ education.num : int  9 9 14 12 13 5 9 6 13 12 ...
```

```
##  $ marital.status: chr  "Married-civ-spouse" "Never-married" "Married-civ-spouse" "Separated" ...
##  $ occupation    : chr  "Other-service" "Sales" "Prof-specialty" "Craft-repair" ...
##  $ relationship  : chr  "Husband" "Not-in-family" "Husband" "Unmarried" ...
##  $ race          : chr  "White" "Asian-Pac-Islander" "White" "Asian-Pac-Islander" ...
##  $ sex           : chr  "Male" "Female" "Male" "Male" ...
##  $ capital.gain  : int  0 0 0 8614 0 0 0 0 99999 0 ...
##  $ capital.loss  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ hours.per.week: int  40 25 40 40 50 20 65 40 60 27 ...
##  $ native.country: chr  "United-States" "United-States" "United-States" "United-States" ...
##  $ income        : Factor w/ 2 levels "<=50K",">50K": 1 1 2 2 2 1 1 1 2 1 ...
```

Using the summary() function in R provides us with summary statistics for each column. It is important to note that there are more data points in "<=50K" level than there are in the ">50k" level for the income factor.

```
summary(train)
```

```
##       age         workclass            fnlwgt         education
##  Min.   :17.00   Length:26048       Min.   :  12285   Length:26048
##  1st Qu.:28.00   Class :character   1st Qu.: 117789   Class :character
##  Median :37.00   Mode  :character   Median : 178459   Mode  :character
##  Mean   :38.57                      Mean   : 189590
##  3rd Qu.:48.00                      3rd Qu.: 236861
##  Max.   :90.00                      Max.   :1484705
##  education.num   marital.status      occupation        relationship
##  Min.   : 1.00   Length:26048       Length:26048       Length:26048
##  1st Qu.: 9.00   Class :character   Class :character   Class :character
##  Median :10.00   Mode  :character   Mode  :character   Mode  :character
##  Mean   :10.08
##  3rd Qu.:12.00
##  Max.   :16.00
##      race               sex             capital.gain    capital.loss
##  Length:26048       Length:26048       Min.   :    0   Min.   :   0.00
##  Class :character   Class :character   1st Qu.:    0   1st Qu.:   0.00
##  Mode  :character   Mode  :character   Median :    0   Median :   0.00
##                                        Mean   : 1088   Mean   :  88.36
##                                        3rd Qu.:    0   3rd Qu.:   0.00
##                                        Max.   :99999   Max.   :4356.00
##  hours.per.week  native.country       income
##  Min.   : 1.00   Length:26048       <=50K:19751
##  1st Qu.:40.00   Class :character   >50K : 6297
##  Median :40.00   Mode  :character
##  Mean   :40.44
##  3rd Qu.:45.00
##  Max.   :99.00
```

Let's find the size of the training data set.

The nrow() function shows that there are 26,048 observations.

```
nrow(train)
```

```
## [1] 26048
```

The ncol() function shows that there are 15 variables

```
ncol(train)
```

```
## [1] 15
```

Using the colSums() function, we can see that there are no missing values in any of the columns. It is important to remove missing values prior to performing logistic regression.

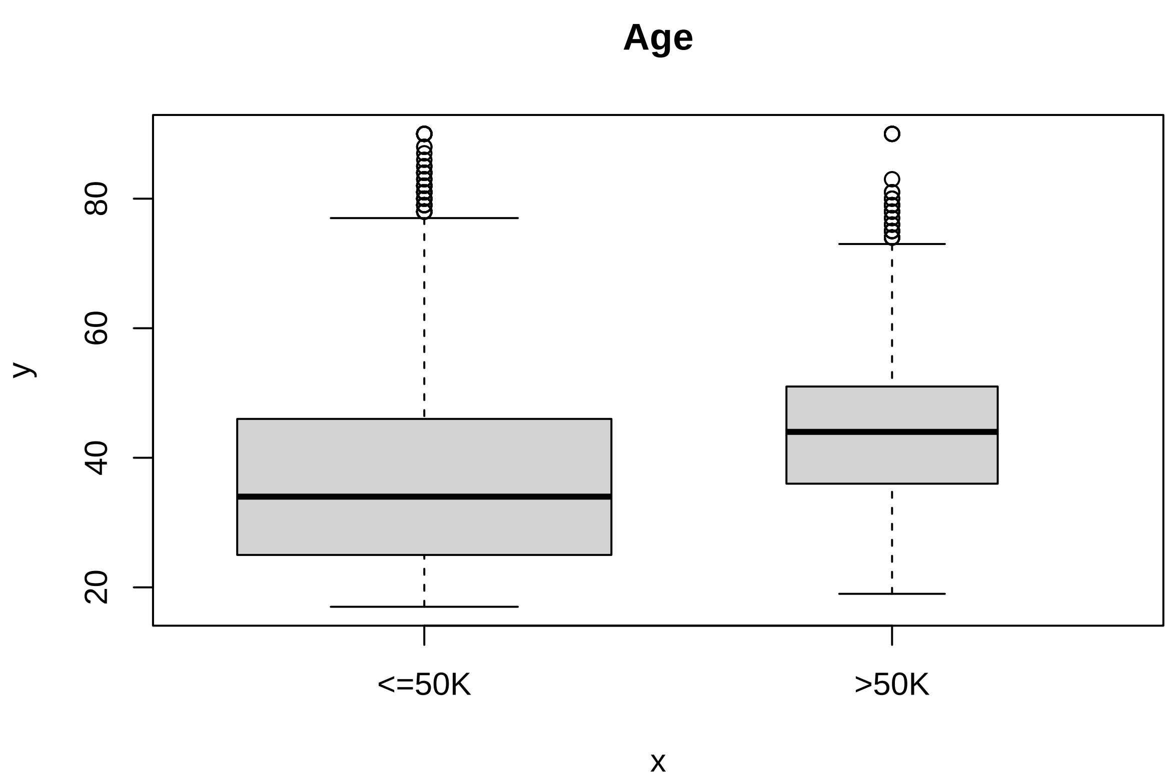
```
colSums(is.na(train))
```

```
##            age      workclass         fnlwgt      education  education.num
##              0              0              0              0              0
## marital.status     occupation   relationship           race            sex
##              0              0              0              0              0
##   capital.gain   capital.loss hours.per.week native.country         income
##              0              0              0              0              0
```

Data Visualization

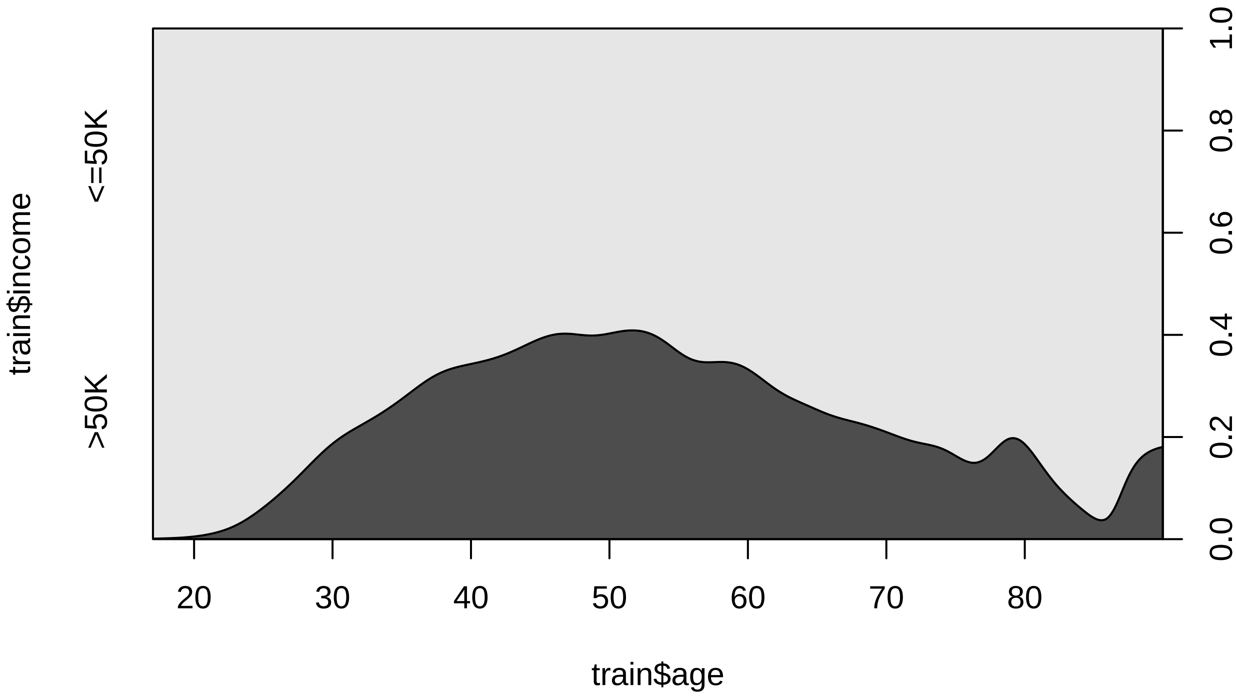
We can use a box-plot to visualize how age affects income. The graph below shows that <=50k is more common than >50k. More importantly, the box-plot shows that >50k observations are associated with those that are slightly older.

```
plot(train$income, train$age, data=train, main= "Age", varwidth=TRUE)
```



We can also use a conditional density plot to visualize how age affects income. The rectangle is the total probability space with the lighter grey indicating <=50k and the darker grey indication <50k.

```
cdplot(train$income~train$age)
```



Logistic Regression

Let's fit a logistic regression model to the data using the glm() function. A summary of the glm model that was created reveals 4 things: the glm() call, the residual distribution, the coefficients with statistical significance metrics, and metrics for the model. The deviance residual is a mathematical transformation of the loss function and details how a certain observation can contribute to the overall likelihood. It can be used to form RSS-like statistics. We can see statistical significance metrics at the bottom of the output. The null deviance measures the lack of fit of the model but only considers the intercept, whereas the residual deviance measures the lack of fit of the entire model. Since the residual deviance is lower than the null deviance, our model is a good fit. The AIC, which stands for the Akaike Information Criteria, is useful in comparing models and typically, the lower the AIC is, the better. The coefficient is 0.039647, which quantifies the difference in the log odds of a target variable.

```r
glm1 <- glm(income~age, data=train, family=binomial)
summary(glm1)
```

```
##
## Call:
## glm(formula = income ~ age, family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5427  -0.7523  -0.5972  -0.4875   2.0586
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.74500    0.04771  -57.54   <2e-16 ***
```

```
## age          0.03969    0.00108   36.75   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 28813  on 26047  degrees of freedom
## Residual deviance: 27400  on 26046  degrees of freedom
## AIC: 27404
##
## Number of Fisher Scoring iterations: 4
```

Naive Bayes Model

Naive Bayes is another classification algorithm. The prior for income, called A-priori, below is 0.759137 for <= 50k and 0.240863 for >50k. The likelihood data is shown in the output as conditional probabilities. Discrete data, such as sex, is broken down into <=50k and >50k for each attribute. For instance, if someone is making >50k, they are 15% likely to be female or 85% likely to be male according to the Naive Bayes model shown below. For continuous variables, such as age, we are given the mean and standard deviation for the two classes. The Naive Bayes model shown below reveals that the mean age for those making <=50k is around 36, while the mean age for those making >50k is around 44.

```r
library(e1071)
nb1 <- naiveBayes(income~.,data=train)
nb1
```

```
##
## Naive Bayes Classifier for Discrete Predictors
##
## Call:
## naiveBayes.default(x = X, y = Y, laplace = laplace)
##
## A-priori probabilities:
## Y
##    <=50K     >50K
## 0.758254 0.241746
##
## Conditional probabilities:
##        age
## Y           [,1]     [,2]
##   <=50K 36.75363 14.04133
##   >50K  44.25917 10.49579
##
##        workclass
## Y                   ?  Federal-gov    Local-gov Never-worked      Private
##   <=50K 0.0657181915 0.0230368083 0.0598957015 0.0002531517 0.7169257253
##   >50K  0.0239796729 0.0485945688 0.0776560267 0.0000000000 0.6322058123
##        workclass
## Y       Self-emp-inc Self-emp-not-inc    State-gov  Without-pay
##   <=50K 0.0195939446     0.0753379576 0.0386309554 0.0006075642
##   >50K  0.0784500556     0.0932189932 0.0458948706 0.0000000000
##
##        fnlwgt
## Y           [,1]     [,2]
##   <=50K 189965.6 106128.5
```

```
##   >50K  188412.3 102850.5
##
##       education
## Y              10th         11th         12th      1st-4th      5th-6th
##   <=50K 0.0364032201 0.0458710951 0.0158979292 0.0064806845 0.0128601083
##   >50K  0.0076226775 0.0073050659 0.0041289503 0.0009528347 0.0017468636
##       education
## Y           7th-8th          9th   Assoc-acdm    Assoc-voc    Bachelors
##   <=50K 0.0243531973 0.0205052909 0.0314920764 0.0411624728 0.1269809124
##   >50K  0.0047641734 0.0033349214 0.0347784659 0.0471653168 0.2850563761
##       education
## Y         Doctorate      HS-grad      Masters    Preschool  Prof-school
##   <=50K 0.0044554706 0.3577540378 0.0311376639 0.0017214318 0.0061769024
##   >50K  0.0384309989 0.2147054153 0.1210100048 0.0000000000 0.0544703827
##       education
## Y       Some-college
##   <=50K 0.2367475065
##   >50K  0.1745275528
##
##       education.num
## Y           [,1]     [,2]
##   <=50K  9.588477 2.438085
##   >50K  11.618390 2.373883
##
##       marital.status
## Y          Divorced Married-AF-spouse Married-civ-spouse Married-spouse-absent
##   <=50K 0.1613082882      0.0004556731       0.3342109260          0.0157966685
##   >50K  0.0592345561      0.0012704462       0.8534222646          0.0044465619
##       marital.status
## Y       Never-married    Separated      Widowed
##   <=50K  0.4133967900 0.0382765430 0.0365551111
##   >50K   0.0630458949 0.0079402890 0.0106399873
##
##       occupation
## Y                  ? Adm-clerical Armed-Forces Craft-repair Exec-managerial
##   <=50K 0.0659713432 0.1306769278 0.0002531517 0.1294617994    0.0849577237
##   >50K  0.0239796729 0.0671748452 0.0001588058 0.1167222487    0.2536128315
##       occupation
## Y       Farming-fishing Handlers-cleaners Machine-op-inspct Other-service
##   <=50K    0.0355931345      0.0527061921      0.0701230317  0.1260695661
##   >50K     0.0147689376      0.0098459584      0.0327139908  0.0173098301
##       occupation
## Y       Priv-house-serv Prof-specialty Protective-serv        Sales
##   <=50K    0.0064806845   0.0926535365    0.0184800770 0.1079439016
##   >50K     0.0001588058   0.2347149436    0.0271557885 0.1238685088
##       occupation
## Y       Tech-support Transport-moving
##   <=50K 0.0262265202     0.0524024100
##   >50K  0.0366841353     0.0411306972
##
##       relationship
## Y          Husband Not-in-family Other-relative   Own-child   Unmarried
##   <=50K 0.292845932   0.302921371    0.038884107 0.202875804 0.128651714
##   >50K  0.756074321   0.108146737    0.005240591 0.008575512 0.028743846
```

```
##        relationship
## Y              Wife
##   <=50K 0.033821072
##   >50K  0.093218993
##
##        race
## Y       Amer-Indian-Eskimo Asian-Pac-Islander       Black       Other
##   <=50K        0.010936155        0.030023796 0.109108400 0.009923548
##   >50K         0.004128950        0.034619660 0.047959346 0.003493727
##        race
## Y             White
##   <=50K 0.840008101
##   >50K  0.909798317
##
##        sex
## Y          Female      Male
##   <=50K 0.3886892 0.6113108
##   >50K  0.1483246 0.8516754
##
##        capital.gain
## Y            [,1]       [,2]
##   <=50K  150.8508   987.4192
##   >50K  4026.1078 14587.0989
##
##        capital.loss
## Y            [,1]     [,2]
##   <=50K  52.62554 308.7334
##   >50K  200.46149 603.4808
##
##        hours.per.week
## Y           [,1]     [,2]
##   <=50K 38.81879 12.35058
##   >50K  45.51771 11.00774
##
##        native.country
## Y                  ?     Cambodia       Canada        China     Columbia
##   <=50K 1.650549e-02 4.050428e-04 3.493494e-03 2.025214e-03 2.126475e-03
##   >50K  1.842147e-02 7.940289e-04 5.399397e-03 2.699698e-03 3.176116e-04
##        native.country
## Y               Cuba Dominican-Republic      Ecuador  El-Salvador      England
##   <=50K 2.885930e-03       2.987191e-03 1.063237e-03 3.847906e-03 2.480887e-03
##   >50K  3.176116e-03       3.176116e-04 4.764173e-04 1.111640e-03 4.128950e-03
##        native.country
## Y             France      Germany       Greece    Guatemala        Haiti
##   <=50K 7.088249e-04 3.594755e-03 6.075642e-04 2.531517e-03 1.772062e-03
##   >50K  1.270446e-03 5.240591e-03 9.528347e-04 3.176116e-04 6.352231e-04
##        native.country
## Y       Holand-Netherlands     Honduras         Hong      Hungary        India
##   <=50K       5.063035e-05 5.569338e-04 6.075642e-04 4.556731e-04 2.430257e-03
##   >50K        0.000000e+00 1.588058e-04 7.940289e-04 4.764173e-04 6.034620e-03
##        native.country
## Y               Iran      Ireland        Italy      Jamaica        Japan
##   <=50K 1.063237e-03 9.113463e-04 1.873323e-03 2.885930e-03 1.417650e-03
##   >50K  1.905669e-03 7.940289e-04 3.334921e-03 1.429252e-03 3.176116e-03
```

```
##        native.country
## Y                Laos       Mexico    Nicaragua Outlying-US(Guam-USVI-etc)
##   <=50K 6.075642e-04 2.440383e-02 1.215128e-03               5.569338e-04
##   >50K  3.176116e-04 4.446562e-03 3.176116e-04               0.000000e+00
##        native.country
## Y                Peru  Philippines        Poland     Portugal  Puerto-Rico
##   <=50K 1.063237e-03 5.214926e-03 1.974584e-03 1.367019e-03 4.202319e-03
##   >50K  3.176116e-04 7.463872e-03 1.111640e-03 3.176116e-04 1.429252e-03
##        native.country
## Y            Scotland        South        Taiwan     Thailand Trinadad&Tobago
##   <=50K 3.037821e-04 2.328996e-03 1.215128e-03 5.569338e-04    7.594552e-04
##   >50K  4.764173e-04 1.588058e-03 1.905669e-03 4.764173e-04    3.176116e-04
##        native.country
## Y       United-States      Vietnam   Yugoslavia
##   <=50K  8.921067e-01 2.531517e-03 3.037821e-04
##   >50K   9.147213e-01 6.352231e-04 7.940289e-04
```

Evaluating the Test Data

Evaluating the logistic regression model with the test data shows a 75% accuracy. The error rate is about 25%.

```r
probs <- predict(glm1, newdata=test, type="response")
pred <- ifelse(probs>0.5, 2, 1)
acc1 <- mean(pred==as.integer(test$income))
print(paste("glm1 accuracy = ", acc1))
```

```
## [1] "glm1 accuracy =  0.751573775525871"
```

```r
table(pred, as.integer(test$income))
```

```
##
## pred    1    2
##    1 4870 1519
##    2   99   25
```

A confusion matrix is created. 4829 is True Positive, in which the items are true and were classified as true. 1545 is False Positive, in which the items were false and classified as true. 117 is False Negative, in which the items were true and classified as false. Finally, 22 is True Negative, in which the items were false and classified as false.

The sensitivity, which is the true positive rate, is 97.6%. The specificity, which is the true negative rate, is approximately 1.4%.

Let's now evaluate the Naive Bayes model with the test data.

```r
p2 <- predict(nb1, test)
(tab2 <- table(p2, test$income))
```

```
##
## p2      <=50K >50K
##   <=50K  4633  795
##   >50K    336  749
```

```r
acc2= sum(diag(tab2)/sum(tab2))
print(acc2)
```

```
## [1] 0.8263473
```

The accuracy for Naive Bayes is about 83% and is slightly higher than the accuracy for logistic regression. The Naive Bayes model may have outperformed the logistic regression model due to the fact that Naive Bayes models tend to perform better with smaller data sets.

Strengths and Weaknesses of Logistic Regression and Naive Bayes

The strengths of the logistic regression model are that it separates classes relatively well if the classes are linearly separable, it is computationally inexpensive, and it provides a nice probabilistic output. The weakness of the logistic regression model is that it is prone to underfitting. The strengths of the Naive Bayes model are that it works well with smaller data sets, it is easy to implement and interpret, and it handles high dimensions well. The weaknesses for Naive Bayes are that other classifiers may outperform it for larger data sets, guesses are made for values in the test set that did not occur in the training set, and the predictors must be independent for good performance.

Benefits and Drawbacks of Classification Metrics

Classification can be evaluated using many metrics. In this notebook, we used accuracy, sensitivity, and specificity. Accuracy is the number of correct predictions divided by the total number of predictions. It is a good measure, but does not give information on the true positive rate and the true negative rate. Sensitivity gives information on the true positive rate, while specificity gives information on the true negative rate.