Linear Classification

By: Perry Son, Waheed Anwar

Overview

Linear classification involves a qualitative target. The linear models work with a dataset of x & y values, which represents the predictors and targets. These models can cleanly separate classes; conduct computationally inexpensive operations; and generate intuitive probabilistic output. However, these models are prone to underfitting. They also cannot capture complex non-linear decision boundaries.

The data consists of information from the 1994 Census. The dataset can be found on Kaggle(https://www.kaggle.com/datasets/uciml/adult-census-income).

Loading Data

I printed the first 6 observations. Some of the features have values of "?".

```
set.seed(1111)
current_path = rstudioapi::getActiveDocumentContext()$path
setwd(dirname(current_path ))
adult_data <- read.csv("adult.csv")
head(adult_data)
```

```
##   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
##          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
##   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
```

There are 32561 observations.

```
nrow(adult_data)
```

```
## [1] 32561
```

I printed the column names of the data. Most of these features appear to be qualitative.

```
colnames(adult_data)
```

```
##  [1] "age"            "workclass"      "fnlwgt"         "education"
##  [5] "education.num"  "marital.status" "occupation"     "relationship"
##  [9] "race"           "sex"            "capital.gain"   "capital.loss"
## [13] "hours.per.week" "native.country" "income"
```

I checked for NaN's in the data's columns.

```
na_count <-sapply(adult_data, function(y) sum(length(which(is.na(y)))))
na_count
```

```
##            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
```

I also checked for columns with "?" values. There are some observations that contain these specific values.

```
question_count <-sapply(adult_data, function(y) sum(length(which(y=="?"))))
question_count
```

```
##            age      workclass         fnlwgt      education  education.num
##              0           1836              0              0              0
## marital.status     occupation   relationship           race            sex
##              0           1843              0              0              0
##   capital.gain   capital.loss hours.per.week native.country         income
##              0              0              0            583              0
```

I dropped observations that contained a "?" character.

```
adult_data <- adult_data[!(adult_data$occupation=="?"),]
adult_data <- adult_data[!(adult_data$workclass=="?"),]
adult_data <- adult_data[!(adult_data$native.country=="?"),]
```

The dataset is clearly skewed towards individuals who make less than 50k. Using data augmentation(ex: undersampling, SMOTE) is outside the scope of this assignment.

```
adult_data$income<-as.factor(adult_data$income)
levels(adult_data$income) <- c('<=50K', '>50K')
summary(adult_data$income)
```

```
## <=50K  >50K
## 22654  7508
```

With 80/20 ratio, I split the data into training and test sets.

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

Using R's Built-in Functions for Data Exploration

People's work schedule range from 1 to 99 hours per week.

```
range(train$hours.per.week)
```

```
## [1]  1 99
```

Average age is 38.

```
mean(train$age)
```

```
## [1] 38.43189
```

There is barely any correlation between hours-per-week and age.

```
cor(train$hours.per.week, train$age)
```

```
## [1] 0.1019156
```

Median age is 37 years old.

```
median(train$age)
```

```
## [1] 37
```

People's age range from 17 to 90 years old.

```
range(train$age)
```

```
## [1] 17 90
```

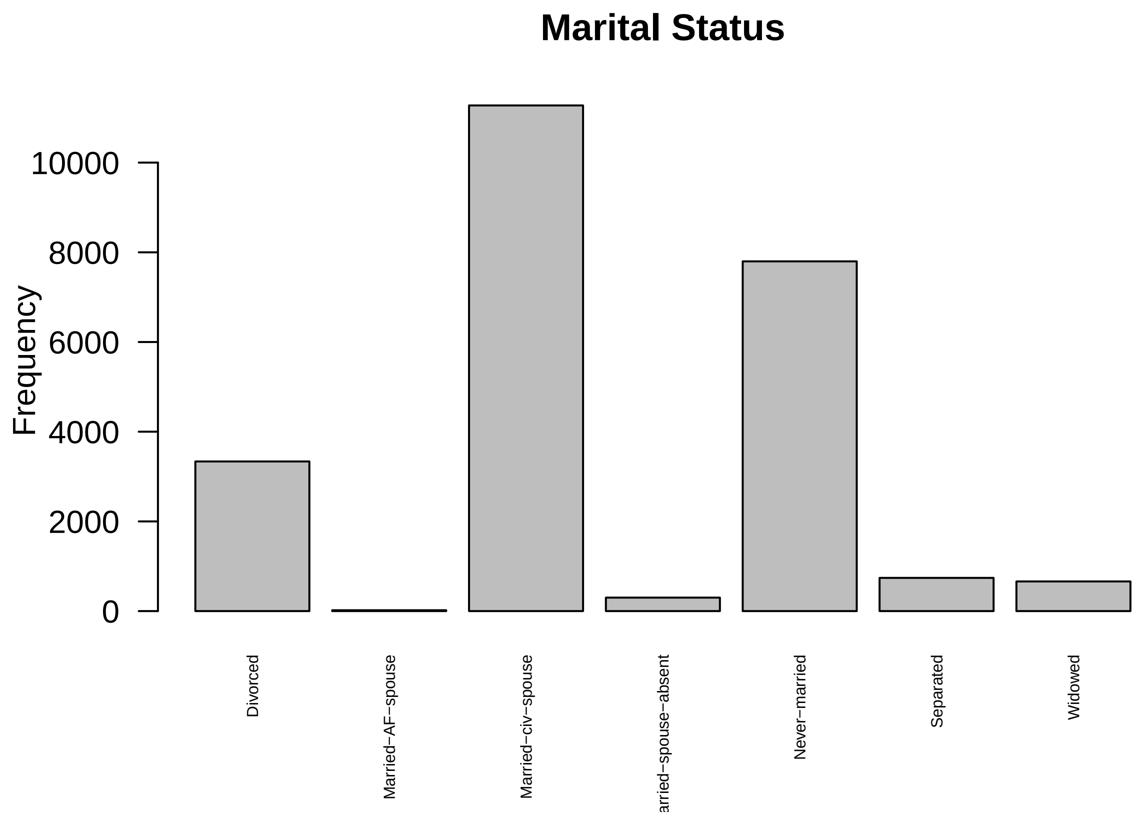
Graphs of Training Data

Not that many people worked for the military. There were similar amounts of executives and repairmen.

```r
barplot(table(train$occupation),  ylab="Frequency", main="Occupations", las = 2,
        cex.names = 0.65)
```

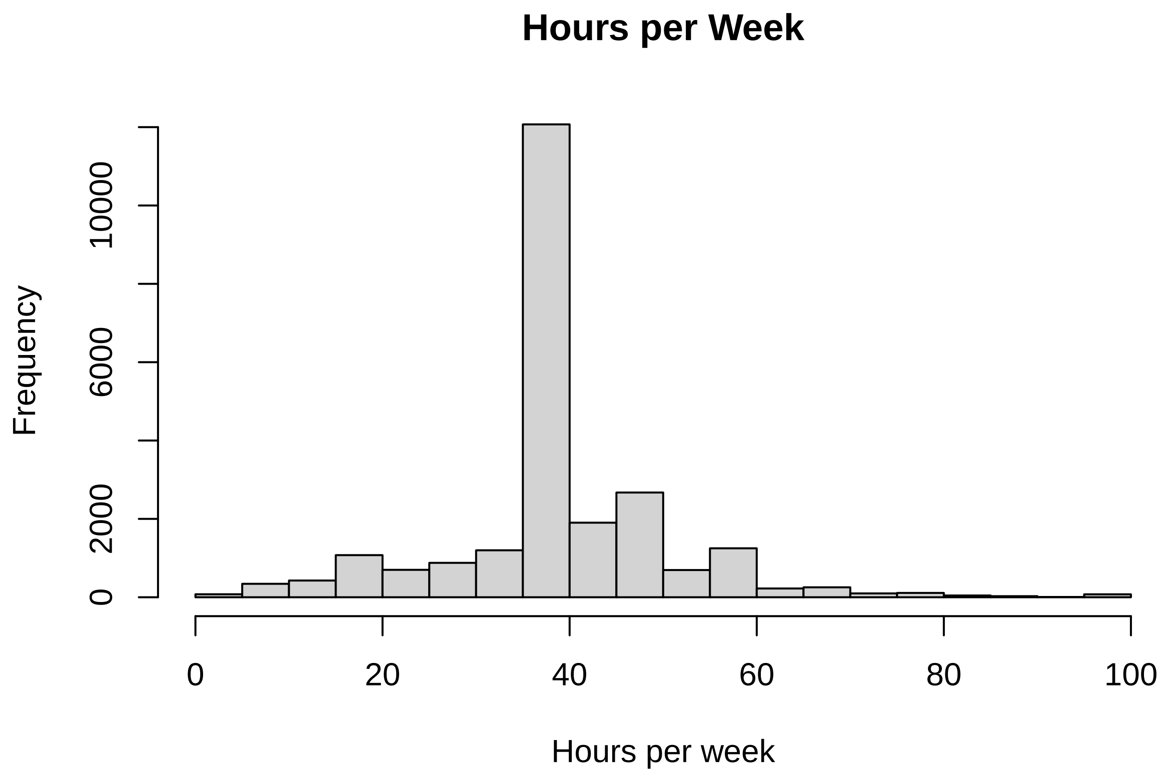
Many individuals were married to civilian spouses. However, the 2nd highest bar consisted of non-married individuals.

```r
barplot(table(train$marital.status),ylab="Frequency", main="Marital Status", las = 2,
        cex.names = 0.5)
```



People tended to work between 35 and 40 hours.

```
hist(train$hours.per.week, main = "Hours per Week", xlab="Hours per week")
```



Logistic Regression

Residual deviance is much lower than null deviance. AIC is 15810.

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

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

```r
summary(glm1)
```

```
##
## Call:
## glm(formula = income ~ ., family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.0642  -0.5151  -0.1900   0.0000   3.9356
##
## Coefficients: (1 not defined because of singularities)
##                                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                              -6.772e+00  8.272e-01  -8.187 2.68e-16
## age                                       2.469e-02  1.915e-03  12.894  < 2e-16
## workclassLocal-gov                       -6.970e-01  1.262e-01  -5.522 3.35e-08
## workclassPrivate                         -4.839e-01  1.048e-01  -4.619 3.85e-06
```

```
## workclassSelf-emp-inc                 -3.278e-01  1.381e-01  -2.374 0.017616
## workclassSelf-emp-not-inc             -9.745e-01  1.229e-01  -7.928 2.23e-15
## workclassState-gov                    -8.128e-01  1.403e-01  -5.794 6.87e-09
## workclassWithout-pay                  -1.311e+01  2.206e+02  -0.059 0.952602
## fnlwgt                                 7.525e-07  1.980e-07   3.800 0.000145
## education11th                          1.699e-01  2.338e-01   0.726 0.467554
## education12th                          4.373e-01  2.990e-01   1.462 0.143632
## education1st-4th                      -2.151e-01  5.082e-01  -0.423 0.672125
## education5th-6th                      -4.180e-01  4.114e-01  -1.016 0.309666
## education7th-8th                      -6.002e-01  2.749e-01  -2.183 0.029006
## education9th                          -3.130e-01  3.003e-01  -1.042 0.297354
## educationAssoc-acdm                    1.242e+00  2.006e-01   6.190 6.02e-10
## educationAssoc-voc                     1.280e+00  1.918e-01   6.671 2.55e-11
## educationBachelors                     1.867e+00  1.787e-01  10.443  < 2e-16
## educationDoctorate                     2.949e+00  2.488e-01  11.853  < 2e-16
## educationHS-grad                       7.561e-01  1.737e-01   4.352 1.35e-05
## educationMasters                       2.243e+00  1.913e-01  11.727  < 2e-16
## educationPreschool                    -1.874e+01  1.072e+02  -0.175 0.861162
## educationProf-school                   2.737e+00  2.297e-01  11.914  < 2e-16
## educationSome-college                  1.088e+00  1.763e-01   6.171 6.79e-10
## education.num                                 NA         NA      NA       NA
## marital.statusMarried-AF-spouse        2.833e+00  6.155e-01   4.603 4.17e-06
## marital.statusMarried-civ-spouse       2.244e+00  2.919e-01   7.687 1.51e-14
## marital.statusMarried-spouse-absent    7.193e-02  2.599e-01   0.277 0.781933
## marital.statusNever-married           -4.536e-01  1.004e-01  -4.520 6.18e-06
## marital.statusSeparated               -1.494e-01  1.886e-01  -0.792 0.428233
## marital.statusWidowed                  2.537e-01  1.757e-01   1.444 0.148776
## occupationArmed-Forces                -1.081e+00  1.537e+00  -0.703 0.481829
## occupationCraft-repair                 1.404e-01  9.090e-02   1.545 0.122415
## occupationExec-managerial              8.448e-01  8.782e-02   9.619  < 2e-16
## occupationFarming-fishing             -9.788e-01  1.588e-01  -6.163 7.13e-10
## occupationHandlers-cleaners           -6.067e-01  1.581e-01  -3.837 0.000125
## occupationMachine-op-inspct           -2.014e-01  1.162e-01  -1.733 0.083095
## occupationOther-service               -7.344e-01  1.319e-01  -5.566 2.61e-08
## occupationPriv-house-serv             -4.118e+00  1.691e+00  -2.435 0.014879
## occupationProf-specialty               5.804e-01  9.324e-02   6.224 4.84e-10
## occupationProtective-serv              7.049e-01  1.407e-01   5.010 5.43e-07
## occupationSales                        4.123e-01  9.396e-02   4.388 1.14e-05
## occupationTech-support                 7.278e-01  1.247e-01   5.835 5.38e-09
## occupationTransport-moving            -3.656e-02  1.132e-01  -0.323 0.746740
## relationshipNot-in-family              5.500e-01  2.881e-01   1.909 0.056258
## relationshipOther-relative            -2.675e-01  2.641e-01  -1.013 0.311155
## relationshipOwn-child                 -6.119e-01  2.843e-01  -2.152 0.031396
## relationshipUnmarried                  4.956e-01  3.063e-01   1.618 0.105682
## relationshipWife                       1.439e+00  1.191e-01  12.080  < 2e-16
## raceAsian-Pac-Islander                 7.763e-01  3.219e-01   2.412 0.015863
## raceBlack                              4.375e-01  2.703e-01   1.618 0.105556
## raceOther                              3.562e-01  4.162e-01   0.856 0.392076
## raceWhite                              5.952e-01  2.575e-01   2.311 0.020817
## sexMale                                8.365e-01  9.040e-02   9.253  < 2e-16
## capital.gain                           3.173e-04  1.203e-05  26.375  < 2e-16
## capital.loss                           6.281e-04  4.363e-05  14.397  < 2e-16
## hours.per.week                         3.070e-02  1.918e-03  16.005  < 2e-16
## native.countryCanada                  -8.670e-01  7.501e-01  -1.156 0.247734
```

```
## native.countryChina                        -1.578e+00  7.483e-01  -2.108 0.035032
## native.countryColumbia                     -3.065e+00  1.063e+00  -2.883 0.003945
## native.countryCuba                         -6.408e-01  7.620e-01  -0.841 0.400386
## native.countryDominican-Republic           -2.693e+00  1.256e+00  -2.145 0.031973
## native.countryEcuador                      -9.173e-01  9.978e-01  -0.919 0.357915
## native.countryEl-Salvador                  -1.999e+00  9.378e-01  -2.131 0.033053
## native.countryEngland                      -9.094e-01  7.586e-01  -1.199 0.230652
## native.countryFrance                       -6.360e-01  8.725e-01  -0.729 0.466041
## native.countryGermany                      -4.142e-01  7.303e-01  -0.567 0.570613
## native.countryGreece                       -1.729e+00  8.826e-01  -1.959 0.050096
## native.countryGuatemala                    -1.856e+00  1.267e+00  -1.465 0.142998
## native.countryHaiti                        -1.258e+00  1.025e+00  -1.227 0.219741
## native.countryHoland-Netherlands           -1.149e+01  8.827e+02  -0.013 0.989611
## native.countryHonduras                     -1.840e+00  2.999e+00  -0.613 0.539675
## native.countryHong                         -1.414e+00  9.822e-01  -1.439 0.150087
## native.countryHungary                      -1.006e+00  1.090e+00  -0.923 0.356026
## native.countryIndia                        -1.083e+00  7.264e-01  -1.491 0.136092
## native.countryIran                         -8.151e-01  8.192e-01  -0.995 0.319735
## native.countryIreland                      -8.112e-01  1.017e+00  -0.798 0.425112
## native.countryItaly                        -2.075e-01  7.691e-01  -0.270 0.787337
## native.countryJamaica                      -1.419e+00  8.657e-01  -1.639 0.101277
## native.countryJapan                        -3.700e-01  8.029e-01  -0.461 0.644971
## native.countryLaos                         -1.106e+00  1.108e+00  -0.999 0.318025
## native.countryMexico                       -1.453e+00  7.170e-01  -2.027 0.042694
## native.countryNicaragua                    -1.525e+00  1.069e+00  -1.427 0.153691
## native.countryOutlying-US(Guam-USVI-etc)  -1.342e+01  2.704e+02  -0.050 0.960432
## native.countryPeru                         -1.789e+00  1.089e+00  -1.642 0.100524
## native.countryPhilippines                  -5.883e-01  6.919e-01  -0.850 0.395147
## native.countryPoland                       -8.191e-01  8.056e-01  -1.017 0.309254
## native.countryPortugal                     -7.027e-01  9.326e-01  -0.753 0.451155
## native.countryPuerto-Rico                  -1.868e+00  8.379e-01  -2.230 0.025764
## native.countryScotland                     -9.492e-01  1.165e+00  -0.815 0.415108
## native.countrySouth                        -2.174e+00  7.851e-01  -2.769 0.005615
## native.countryTaiwan                       -1.035e+00  8.218e-01  -1.259 0.208019
## native.countryThailand                     -1.661e+00  1.053e+00  -1.578 0.114680
## native.countryTrinadad&Tobago              -1.124e+00  1.135e+00  -0.991 0.321793
## native.countryUnited-States                -7.880e-01  6.786e-01  -1.161 0.245549
## native.countryVietnam                      -2.358e+00  1.003e+00  -2.350 0.018755
## native.countryYugoslavia                   -8.852e-01  1.000e+00  -0.885 0.376118
##
## (Intercept)                                ***
## age                                        ***
## workclassLocal-gov                         ***
## workclassPrivate                           ***
## workclassSelf-emp-inc                      *
## workclassSelf-emp-not-inc                  ***
## workclassState-gov                         ***
## workclassWithout-pay
## fnlwgt                                     ***
## education11th
## education12th
## education1st-4th
## education5th-6th
## education7th-8th                           *
```

```
## education9th
## educationAssoc-acdm                     ***
## educationAssoc-voc                      ***
## educationBachelors                      ***
## educationDoctorate                      ***
## educationHS-grad                        ***
## educationMasters                        ***
## educationPreschool
## educationProf-school                    ***
## educationSome-college                   ***
## education.num
## marital.statusMarried-AF-spouse         ***
## marital.statusMarried-civ-spouse        ***
## marital.statusMarried-spouse-absent
## marital.statusNever-married             ***
## marital.statusSeparated
## marital.statusWidowed
## 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
## relationshipNot-in-family               .
## relationshipOther-relative
## relationshipOwn-child                   *
## relationshipUnmarried
## relationshipWife                        ***
## raceAsian-Pac-Islander                  *
## raceBlack
## raceOther
## raceWhite                               *
## sexMale                                 ***
## capital.gain                            ***
## capital.loss                            ***
## hours.per.week                          ***
## native.countryCanada
## native.countryChina                     *
## native.countryColumbia                  **
## native.countryCuba
## native.countryDominican-Republic        *
## native.countryEcuador
## native.countryEl-Salvador               *
## native.countryEngland
## native.countryFrance
## native.countryGermany
## native.countryGreece                    .
```

```
## native.countryGuatemala
## native.countryHaiti
## native.countryHoland-Netherlands
## native.countryHonduras
## native.countryHong
## native.countryHungary
## native.countryIndia
## native.countryIran
## native.countryIreland
## native.countryItaly
## native.countryJamaica
## native.countryJapan
## native.countryLaos
## native.countryMexico                     *
## native.countryNicaragua
## native.countryOutlying-US(Guam-USVI-etc)
## native.countryPeru
## native.countryPhilippines
## native.countryPoland
## native.countryPortugal
## native.countryPuerto-Rico                *
## native.countryScotland
## native.countrySouth                      **
## native.countryTaiwan
## native.countryThailand
## native.countryTrinadad&Tobago
## native.countryUnited-States
## native.countryVietnam                    *
## native.countryYugoslavia
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 27148  on 24128  degrees of freedom
## Residual deviance: 15618  on 24033  degrees of freedom
## AIC: 15810
##
## Number of Fisher Scoring iterations: 13
```

Naive Bayes

The priors for <=50K and >50K are 0.7498 and 0.25020. Mean age for making around 50K or less is ~36 years old. Mean age for making more than 50K is ~44 years old.

```
library(e1071)
```

```
## Warning: package 'e1071' was built under R version 4.1.3
```

```
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
## 0.7498031 0.2501969
##
## Conditional probabilities:
##        age
## Y           [,1]     [,2]
##   <=50K 36.61972 13.48521
##   >50K  43.86268 10.26529
##
##        workclass
## Y        Federal-gov    Local-gov      Private Self-emp-inc Self-emp-not-inc
##   <=50K 0.0255914216 0.0652774707 0.7674662834 0.0212801238     0.0787640946
##   >50K  0.0486996853 0.0818287229 0.6481696207 0.0805035614     0.0955772735
##        workclass
## Y          State-gov  Without-pay
##   <=50K 0.0410126023 0.0006080035
##   >50K  0.0452211363 0.0000000000
##
##        fnlwgt
## Y           [,1]     [,2]
##   <=50K 190431.2 106286.0
##   >50K  188349.1 102659.5
##
##        education
## Y               10th         11th         12th      1st-4th      5th-6th
##   <=50K 0.0338271059 0.0434446164 0.0161397303 0.0067433120 0.0121047977
##   >50K  0.0079509690 0.0086135498 0.0041411297 0.0009938711 0.0014908067
##        education
## Y            7th-8th          9th   Assoc-acdm    Assoc-voc    Bachelors
##   <=50K 0.0224408578 0.0199535706 0.0331085563 0.0421180632 0.1303338492
##   >50K  0.0044724201 0.0033129038 0.0329633924 0.0475401690 0.2850753686
##        education
## Y          Doctorate      HS-grad      Masters    Preschool  Prof-school
##   <=50K 0.0042560248 0.3596617289 0.0314503648 0.0020451028 0.0061905815
##   >50K  0.0366075865 0.2125227762 0.1230743747 0.0000000000 0.0548285572
##        education
## Y       Some-college
##   <=50K 0.2361817378
##   >50K  0.1764121252
##
##        education.num
## Y            [,1]     [,2]
##   <=50K  9.632545 2.424370
##   >50K  11.614544 2.368827
##
##        marital.status
## Y           Divorced Married-AF-spouse Married-civ-spouse Married-spouse-absent
```

```
##   <=50K 0.1650453239      0.0005527305       0.3378841477          0.0151448154
##   >50K  0.0579758158      0.0014908067       0.8550604605          0.0043067749
##        marital.status
## Y       Never-married    Separated      Widowed
##   <=50K  0.4101812956 0.0382489498 0.0329427371
##   >50K   0.0624482359 0.0079509690 0.0107669372
##
##        occupation
## Y       Adm-clerical Armed-Forces Craft-repair Exec-managerial Farming-fishing
##   <=50K 0.1419964625 0.0003869113 0.1384037143    0.0923612646    0.0390780455
##   >50K  0.0649329137 0.0001656452 0.1225774391    0.2603942355    0.0145767765
##        occupation
## Y       Handlers-cleaners Machine-op-inspct Other-service Priv-house-serv
##   <=50K      0.0570970595      0.0750055273  0.1356953350    0.0061353084
##   >50K       0.0119264535      0.0314725857  0.0182209707    0.0001656452
##        occupation
## Y       Prof-specialty Protective-serv        Sales Tech-support
##   <=50K   0.0984965731    0.0189033827 0.1129228388 0.0281892549
##   >50K    0.2385290707    0.0291535531 0.1293688918 0.0376014577
##        occupation
## Y       Transport-moving
##   <=50K     0.0553283219
##   >50K      0.0409143614
##
##        relationship
## Y           Husband Not-in-family Other-relative   Own-child   Unmarried
##   <=50K 0.300519567   0.304278134    0.037640946 0.194616405 0.133318594
##   >50K  0.758323671   0.106841146    0.005135001 0.008944840 0.028325327
##        relationship
## Y              Wife
##   <=50K 0.029626354
##   >50K  0.092430015
##
##        race
## Y       Amer-Indian-Eskimo Asian-Pac-Islander       Black       Other
##   <=50K        0.011165156        0.027968163 0.108556268 0.008733142
##   >50K         0.004638065        0.034785489 0.049030976 0.002815968
##        race
## Y             White
##   <=50K 0.843577272
##   >50K  0.908729501
##
##        sex
## Y          Female      Male
##   <=50K 0.3824895 0.6175105
##   >50K  0.1484181 0.8515819
##
##        capital.gain
## Y            [,1]       [,2]
##   <=50K  148.4982   949.5256
##   >50K  3993.6420 14595.5477
##
##        capital.loss
## Y            [,1]     [,2]
```

```
##   <=50K  52.97258 308.1574
##   >50K  190.60726 587.7037
##
##       hours.per.week
## Y           [,1]     [,2]
##   <=50K 39.30013 11.91462
##   >50K  45.80818 10.77619
##
##       native.country
## Y           Cambodia       Canada        China     Columbia         Cuba
##   <=50K 5.527305e-04 3.426929e-03 2.100376e-03 2.708379e-03 2.929472e-03
##   >50K  9.938711e-04 4.306775e-03 3.312904e-03 3.312904e-04 3.312904e-03
##       native.country
## Y       Dominican-Republic      Ecuador  El-Salvador      England       France
##   <=50K       2.708379e-03 1.105461e-03 4.034933e-03 2.487287e-03 7.738227e-04
##   >50K        1.656452e-04 6.625808e-04 8.282259e-04 3.809839e-03 1.656452e-03
##       native.country
## Y            Germany       Greece    Guatemala        Haiti Holand-Netherlands
##   <=50K 3.924386e-03 8.843688e-04 2.708379e-03 1.934557e-03       5.527305e-05
##   >50K  6.294517e-03 1.159516e-03 1.656452e-04 4.969356e-04       0.000000e+00
##       native.country
## Y           Honduras         Hong      Hungary        India         Iran
##   <=50K 4.421844e-04 5.527305e-04 4.421844e-04 2.542560e-03 1.050188e-03
##   >50K  1.656452e-04 8.282259e-04 3.312904e-04 5.466291e-03 2.650323e-03
##       native.country
## Y            Ireland        Italy      Jamaica        Japan         Laos
##   <=50K 8.843688e-04 1.934557e-03 3.040018e-03 1.381826e-03 4.974574e-04
##   >50K  4.969356e-04 3.312904e-03 9.938711e-04 3.478549e-03 3.312904e-04
##       native.country
## Y             Mexico    Nicaragua Outlying-US(Guam-USVI-etc)         Peru
##   <=50K 2.575724e-02 1.381826e-03               4.421844e-04 1.216007e-03
##   >50K  4.472420e-03 3.312904e-04               0.000000e+00 3.312904e-04
##       native.country
## Y        Philippines       Poland     Portugal  Puerto-Rico     Scotland
##   <=50K 5.858943e-03 1.879284e-03 1.271280e-03 4.256025e-03 3.316383e-04
##   >50K  8.447905e-03 1.656452e-03 6.625808e-04 9.938711e-04 3.312904e-04
##       native.country
## Y              South       Taiwan     Thailand Trinadad&Tobago United-States
##   <=50K 2.432014e-03 8.843688e-04 5.527305e-04    7.185496e-04  9.045987e-01
##   >50K  1.987742e-03 2.815968e-03 4.969356e-04    3.312904e-04  9.304290e-01
##       native.country
## Y            Vietnam   Yugoslavia
##   <=50K 2.763652e-03 5.527305e-04
##   >50K  4.969356e-04 6.625808e-04
```

Model Evaluation

The logistic regression model had an accuracy of 0.85. For future work, a weighted accuracy should be used to determine relative importance of false-positive and false-negative errors.

```
probs <- predict(glm1, newdata=test, type="response")
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
```

```
## prediction from a rank-deficient fit may be misleading
```

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

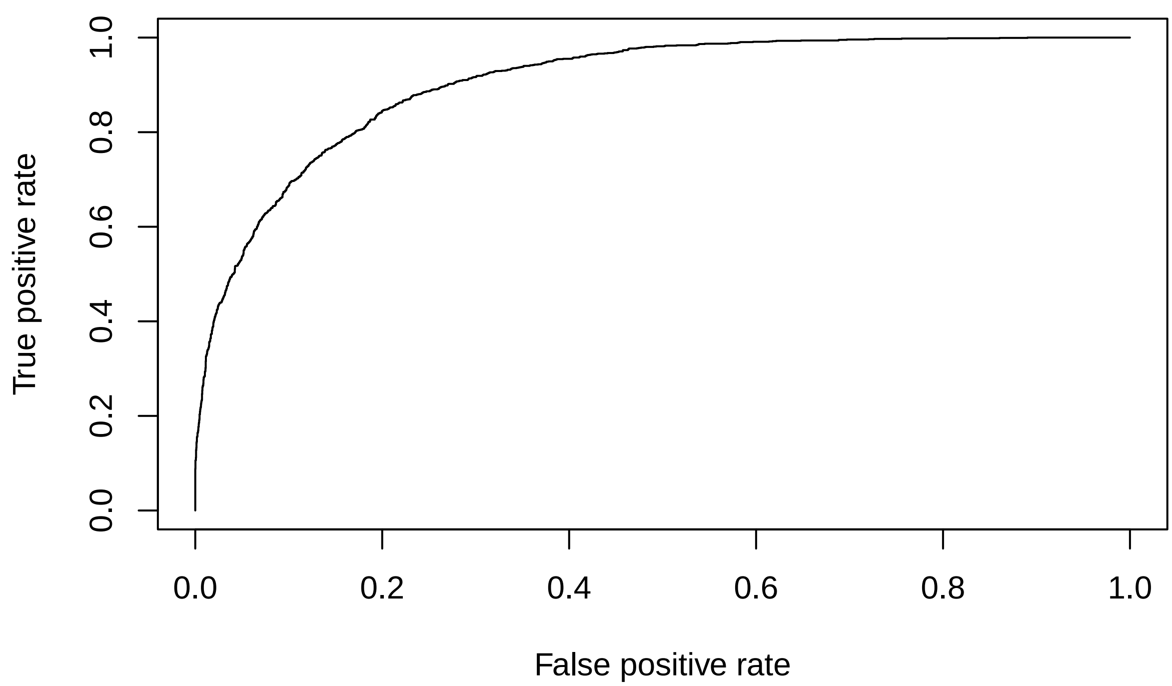
```
## [1] "glm1 accuracy =  0.853472567545168"
```

```r
pred <- ifelse(pred==2, ">50K", "<=50K")
table(pred, test$income)
```

```
##
## pred    <=50K >50K
##   <=50K  4246  568
##   >50K    316  903
```

The ROC curve shows the trade-off between predicting true and false positives.

```r
library(ROCR)
pr <- prediction(probs, test$income)
prf <- performance(pr, measure = "tpr", x.measure = "fpr")
plot(prf)
```



AUC measures the area under the curve. AUC is 0.90, which is very close to 1.0 (score for a perfect classifier).

```
auc <- performance(pr, measure = "auc")
auc <- auc@y.values[[1]]
auc
```

```
## [1] 0.9046322
```

The Naive Bayes model maintained an accuracy of 0.82.

```
p2 <- predict(nb1, test, type="class")
tab2 <- table(p2, test$income)
print(paste("nb1 accuracy = ", sum(diag(tab2)) / sum(tab2)))
```

```
## [1] "nb1 accuracy =  0.826454500248633"
```

```
tab2
```

```
##
## p2      <=50K >50K
##   <=50K  4248  733
##   >50K    314  738
```

Thoughts on Results

The Naive Bayes model has an accuracy of 0.82, while the logistic regression model has an accuracy of 0.85. Naive Bayes tends to perform better on smaller dataset. Given the large amount of observations, it makes sense for logistic regression to outperform Naive Bayes. This is measured in terms of accuracy.

Comparison between Naive Bayes and Logistic Regression

The strength of Naive Bayes is its assumption regarding independent predictors. Even in a false case, Naive Bayes can still be effective since it performs well on small datasets. Its summary is also very intuitive for human readers.However, a false assumption can negatively impact the performance. Meanwhile, logistic regression works well with binary classification due to qualitative targets. With a larger data set, the logistic regression will fare better than Naive Bayes. This is good for binary classification because it involves independent and dependent variables. The downside of logistic regression can be its high-bias and low-variance nature. It may have the problem of not fitting the data set accurately.

Metrics

Accuracy's benefit is its usage as a very simple metric for classification. This metric tells us the ratio of number of correct predictions over the total number of predictions. However, accuracy does not scale well with complex tasks. These tasks might require other suitable metrics. For example, we also used ROC and AUC. ROC is involved with the probability curve. AUC represents the area under the curve. For example, if the AUC lies in the range of [0.5, 1], then we can distinguish between these two thresholds. However, if the dataset is not implemented properly, this indicates the presence of falsified data or wide threshold differences. In that case, AUC might not be a suitable metric.