CensusReport_AbhishekTalari

Abhishek Talari

2023-12-05

Introduction

The data set used for this project is from US Adult Census with a repository of 32,561 entries, provided by UCI Machine Learning Repository This data was extracted from the 1994 Census Bureau. The variables in this data set are - age, workclass, fnlwgt, education, education-num,marital-status, occupation, relationship, race, sex,capital-gain, capital-loss, hours-per-week, native-country, and income.

The goal of this project is to predict whether a person makes over $50K a year, using machine learning models. Three models are built and compared with respect to their accuracies.

Let us look at the structure of the data.

```
str(adult)

## 'data.frame':    32561 obs. of  15 variables:
##  $ age           : int  39 50 38 53 28 37 49 52 31 42 ...
##  $ workclass     : chr  " State-gov" " Self-emp-not-inc" " Private" "
Private" ...
##  $ fnlwgt        : int  77516 83311 215646 234721 338409 284582 160187
209642 45781 159449 ...
##  $ education     : chr  " Bachelors" " Bachelors" " HS-grad" " 11th" ...
##  $ education.num : int  13 13 9 7 13 14 5 9 14 13 ...
##  $ marital.status: chr  " Never-married" " Married-civ-spouse" " Divorced"
" Married-civ-spouse" ...
##  $ occupation    : chr  " Adm-clerical" " Exec-managerial" " Handlers-
cleaners" " Handlers-cleaners" ...
##  $ relationship  : chr  " Not-in-family" " Husband" " Not-in-family" "
Husband" ...
##  $ race          : chr  " White" " White" " White" " Black" ...
##  $ sex           : chr  " Male" " Male" " Male" " Male" ...
##  $ capital.gain  : int  2174 0 0 0 0 0 0 0 14084 5178 ...
##  $ capital.loss  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ hours.per.week: int  40 13 40 40 40 40 16 45 50 40 ...
##  $ native.country: chr  " United-States" " United-States" " United-States"
" United-States" ...
##  $ income        : chr  " <=50K" " <=50K" " <=50K" " <=50K" ...
```

Attributes

The Data

a. age: the age of an individual. Values are Integer bigger than 0.

b. workclass:: a general term to represent the employment status of an individual. Values are Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.

c. fnlwgt: final weight. this is the number of people the census believes the entry represents Values are continuous.

d. education:: the highest level of education achieved by an individual. Values are Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acad, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.

e. education-num:: the highest level of education achieved in numerical form.

f. marital-status: Married-civ-spouse, Divorced, etc.

g. occupation:: the general type of occupation of an individual. Values are tech-support, Craft-repair, Other-service, Sales, etc.

h. relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.

i. race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.

j. sex: Female, Male.

k. capital-gain:: capital gains for an individual.

l. capital-loss:: capital loss for an individual.

m. hours-per-week:: the hours an individual has reported to work per week

n. native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, etc.

O. income::>50K or <=50K.

Here in this project we are going to predict the income for an individual.

Let us first search for any ‘NA’ values present in the data set.

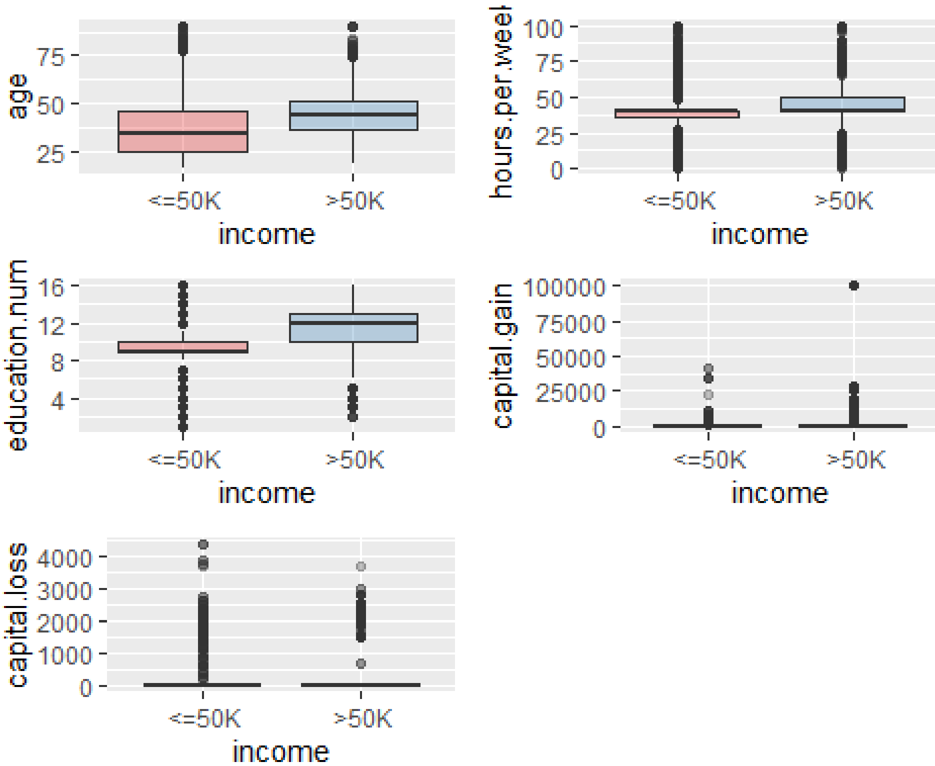
```
adult %>% anyNA()

## [1] FALSE
```

For simplicity of this analysis, i) the weighting factor and ii) relationship (Role in the family can be assessed from gender and marital status) are discarded. Thus, the following 2 variables are deleted - relationship and fnlwgt.

```
adult <- subset(adult,select = -c(fnlwgt,relationship))
```

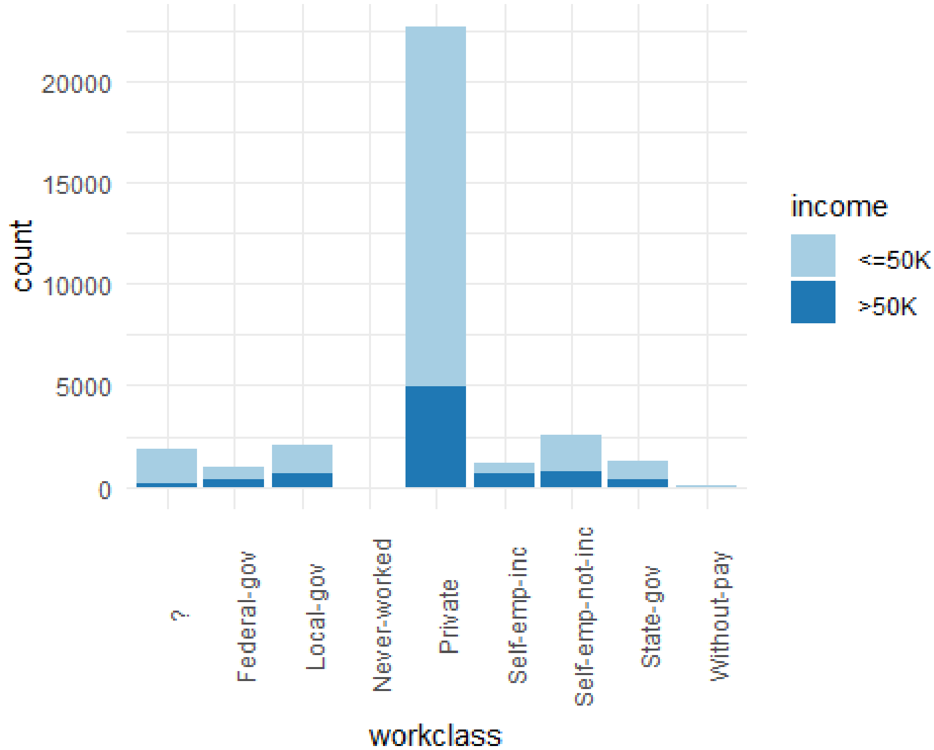
Explotory Analysis

a. To understand about which features would be most helpful for this analysis, let us plot a boxplot for all continuous variables.



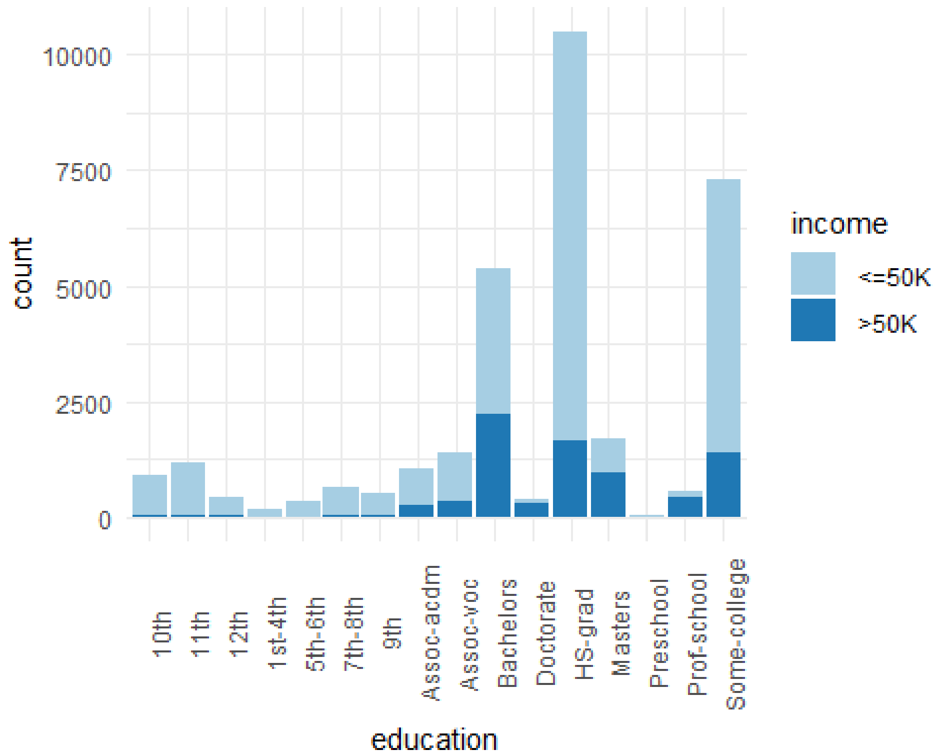
From the above box plots, we can see that all variables can affect the outcome.

b. Let us plot a bar plot for working classes and income.



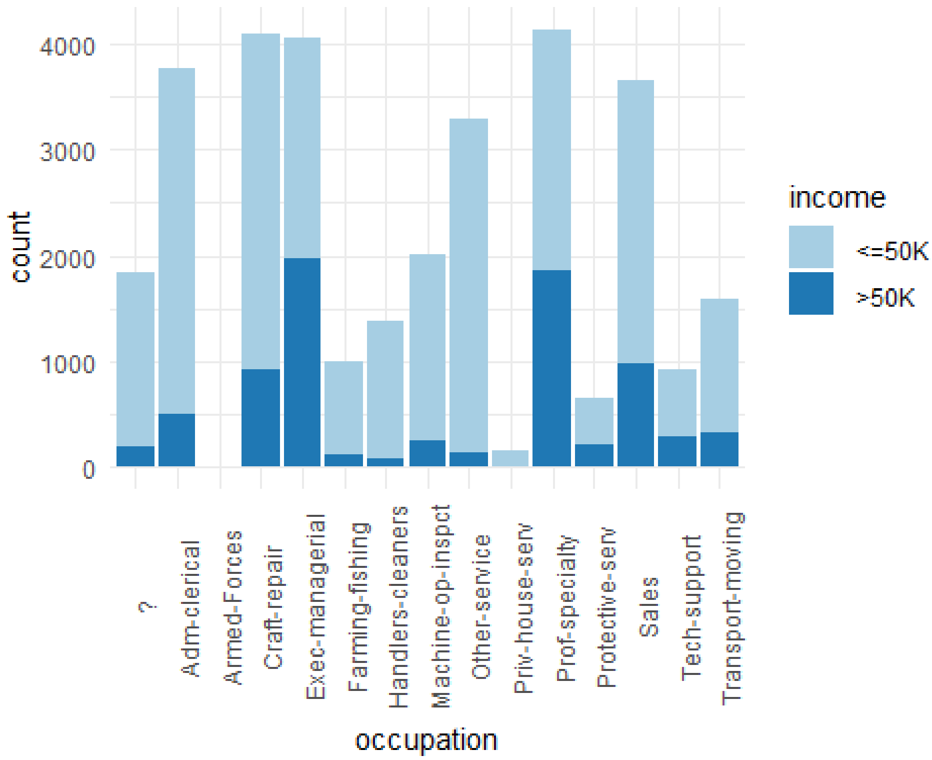
Majority of individuals work in private sector and all the working class people seem to have a good chance of earning more than $50K.

c.Plot for education variable comparison in relation to income



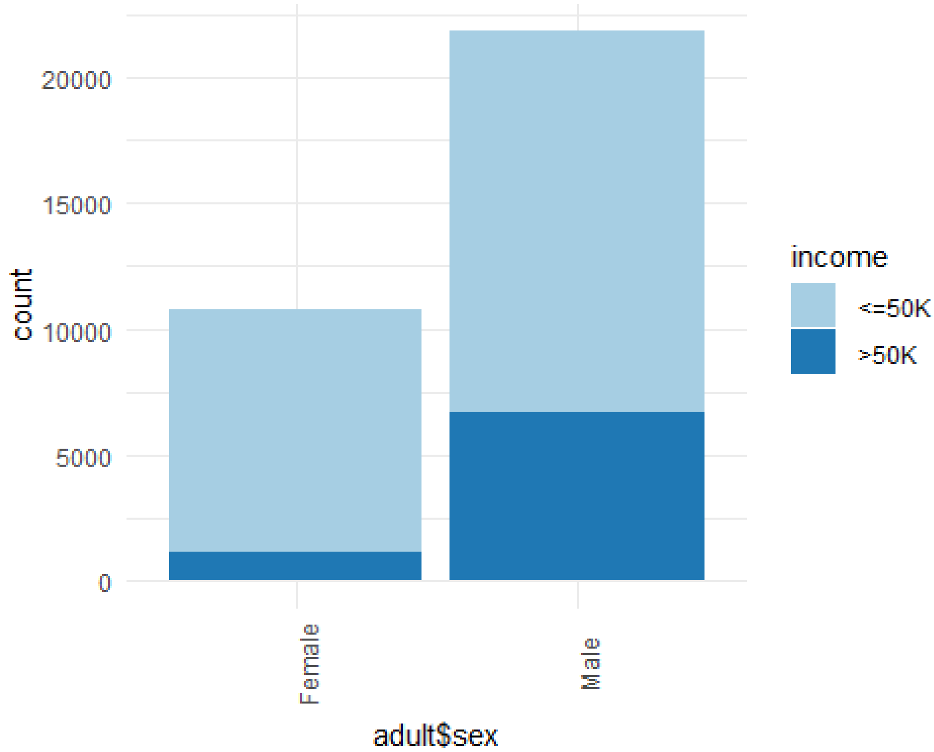
The variable education represents the latest education level for individuals. It appears that the individuals below 12th have very less chances of earning more than 50K.

d. Plot for occupation comparison in relation to income



From the plot we can see that people with exec-managerial and prof-specialty as occpation stand out at having a higher than 50K income.

e. Plot for sex comparison in relation to income



From the plot we can see that the percentage of males who make greater than 50K is much greater than the percentage of females who make greater than 50K.

Data Partition

Split the data into training and testing data 80:20 (standard approach of splitting)

```
trainIndex <- createDataPartition(adult$income, times=1,p =0.8, list=FALSE)

train <- adult[trainIndex,]
test <-  adult[-trainIndex,]
```

Machine Learning Techniques - Model Fitting

###Logistic Regression Model

Let us build a logistic regression model to predict the dependent variable "over 50k", using all of the other variables in the dataset as independent variables and using the training set to build the model.

```
train$income = factor(train$income)
censusglm <- glm( income ~ . , family = binomial , data = train )
summary(censusglm)
```

```
##
## Call:
## glm(formula = income ~ ., family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.1027  -0.5061  -0.2113  -0.0432   3.7652
##
## Coefficients: (2 not defined because of singularities)
##                                            Estimate Std. Error z value
## (Intercept)                              -7.908e+00  3.695e-01 -21.404
## age                                       2.516e-02  1.799e-03  13.987
## workclass Federal-gov                     9.967e-01  1.706e-01   5.842
## workclass Local-gov                       3.307e-01  1.561e-01   2.118
## workclass Never-worked                   -1.138e+01  5.480e+02  -0.021
## workclass Private                         4.796e-01  1.398e-01   3.429
## workclass Self-emp-inc                    6.272e-01  1.674e-01   3.746
## workclass Self-emp-not-inc                6.062e-02  1.532e-01   0.396
## workclass State-gov                       2.005e-01  1.690e-01   1.187
## workclass Without-pay                    -1.334e+01  3.481e+02  -0.038
## education 11th                            1.897e-01  2.249e-01   0.844
## education 12th                            4.584e-01  2.925e-01   1.567
## education 1st-4th                        -5.169e-01  5.423e-01  -0.953
## education 5th-6th                        -1.541e-01  3.488e-01  -0.442
## education 7th-8th                        -5.464e-01  2.558e-01  -2.136
## education 9th                            -8.891e-02  2.759e-01  -0.322
## education Assoc-acdm                      1.295e+00  1.908e-01   6.786
## education Assoc-voc                       1.250e+00  1.831e-01   6.826
## education Bachelors                       1.881e+00  1.693e-01  11.109
## education Doctorate                       2.960e+00  2.339e-01  12.653
## education HS-grad                         7.874e-01  1.647e-01   4.782
## education Masters                         2.237e+00  1.813e-01  12.333
## education Preschool                      -2.002e+01  1.658e+02  -0.121
## education Prof-school                     2.865e+00  2.188e-01  13.092
## education Some-college                    1.140e+00  1.671e-01   6.819
## education.num                                    NA         NA      NA
## marital.status Married-AF-spouse          2.256e+00  6.099e-01   3.699
## marital.status Married-civ-spouse         2.162e+00  7.366e-02  29.352
## marital.status Married-spouse-absent     -6.907e-02  2.621e-01  -0.264
## marital.status Never-married             -5.040e-01  9.107e-02  -5.534
## marital.status Separated                 -1.460e-01  1.766e-01  -0.827
## marital.status Widowed                   -6.919e-02  1.676e-01  -0.413
## occupation Adm-clerical                   2.091e-01  1.103e-01   1.896
## occupation Armed-Forces                  -8.982e-01  1.545e+00  -0.581
## occupation Craft-repair                   2.199e-01  9.609e-02   2.288
## occupation Exec-managerial                9.444e-01  9.811e-02   9.626
## occupation Farming-fishing               -8.740e-01  1.607e-01  -5.437
## occupation Handlers-cleaners             -5.021e-01  1.630e-01  -3.081
## occupation Machine-op-inspct             -2.724e-02  1.185e-01  -0.230
## occupation Other-service                 -6.145e-01  1.391e-01  -4.416
```

```
## occupation Priv-house-serv                   -4.240e+00  1.698e+00  -2.497
## occupation Prof-specialty                     6.861e-01  1.051e-01   6.525
## occupation Protective-serv                    7.480e-01  1.475e-01   5.071
## occupation Sales                              4.675e-01  1.014e-01   4.612
## occupation Tech-support                       8.437e-01  1.328e-01   6.353
## occupation Transport-moving                          NA         NA      NA
## race Asian-Pac-Islander                       4.917e-01  2.874e-01   1.711
## race Black                                    2.767e-01  2.442e-01   1.133
## race Other                                    2.127e-02  3.796e-01   0.056
## race White                                    4.553e-01  2.313e-01   1.968
## sex Male                                      8.003e-02  5.805e-02   1.379
## capital.gain                                  3.222e-04  1.154e-05  27.920
## capital.loss                                  6.267e-04  4.131e-05  15.171
## hours.per.week                                3.039e-02  1.802e-03  16.861
## native.country Cambodia                       1.719e+00  7.122e-01   2.414
## native.country Canada                         5.701e-01  3.165e-01   1.801
## native.country China                         -4.710e-01  4.290e-01  -1.098
## native.country Columbia                      -1.907e+00  8.418e-01  -2.265
## native.country Cuba                           3.528e-01  3.673e-01   0.961
## native.country Dominican-Republic            -1.293e+00  1.064e+00  -1.215
## native.country Ecuador                        3.692e-02  7.453e-01   0.050
## native.country El-Salvador                   -9.383e-01  6.978e-01  -1.345
## native.country England                        5.992e-01  3.651e-01   1.641
## native.country France                         8.704e-01  5.819e-01   1.496
## native.country Germany                        7.404e-01  3.174e-01   2.333
## native.country Greece                        -1.297e-01  6.516e-01  -0.199
## native.country Guatemala                     -1.970e-01  1.001e+00  -0.197
## native.country Haiti                          5.272e-01  6.389e-01   0.825
## native.country Holand-Netherlands            -1.260e+01  1.455e+03  -0.009
## native.country Honduras                      -1.132e+01  4.471e+02  -0.025
## native.country Hong                           1.191e-01  7.459e-01   0.160
## native.country Hungary                        6.417e-01  9.007e-01   0.712
## native.country India                         -7.389e-02  3.585e-01  -0.206
## native.country Iran                           3.343e-01  4.996e-01   0.669
## native.country Ireland                        7.872e-01  6.655e-01   1.183
## native.country Italy                          9.864e-01  3.698e-01   2.667
## native.country Jamaica                        3.310e-01  5.038e-01   0.657
## native.country Japan                          8.467e-01  4.494e-01   1.884
## native.country Laos                          -7.327e-02  9.114e-01  -0.080
## native.country Mexico                        -3.165e-01  2.887e-01  -1.096
## native.country Nicaragua                     -3.328e-01  8.013e-01  -0.415
## native.country Outlying-US(Guam-USVI-etc) -1.324e+01  4.692e+02  -0.028
## native.country Peru                           3.297e-01  9.800e-01   0.336
## native.country Philippines                    6.489e-01  3.051e-01   2.127
## native.country Poland                         3.917e-01  4.371e-01   0.896
## native.country Portugal                       5.983e-01  6.680e-01   0.896
## native.country Puerto-Rico                    3.640e-02  4.423e-01   0.082
## native.country Scotland                       3.085e-01  9.905e-01   0.311
## native.country South                         -5.724e-01  4.591e-01  -1.247
## native.country Taiwan                         2.960e-01  5.658e-01   0.523
```

```
## native.country Thailand                       3.477e-01  1.080e+00   0.322
## native.country Trinadad&Tobago               -3.844e-02  8.696e-01  -0.044
## native.country United-States                  3.919e-01  1.559e-01   2.514
## native.country Vietnam                       -5.967e-01  6.169e-01  -0.967
## native.country Yugoslavia                     7.886e-01  7.166e-01   1.100
##                                             Pr(>|z|)
## (Intercept)                                  < 2e-16 ***
## age                                          < 2e-16 ***
## workclass Federal-gov                       5.17e-09 ***
## workclass Local-gov                         0.034170 *
## workclass Never-worked                      0.983424
## workclass Private                           0.000605 ***
## workclass Self-emp-inc                      0.000179 ***
## workclass Self-emp-not-inc                  0.692331
## workclass State-gov                         0.235402
## workclass Without-pay                       0.969426
## education 11th                              0.398867
## education 12th                              0.117027
## education 1st-4th                           0.340492
## education 5th-6th                           0.658670
## education 7th-8th                           0.032698 *
## education 9th                               0.747278
## education Assoc-acdm                        1.16e-11 ***
## education Assoc-voc                         8.72e-12 ***
## education Bachelors                          < 2e-16 ***
## education Doctorate                          < 2e-16 ***
## education HS-grad                           1.74e-06 ***
## education Masters                            < 2e-16 ***
## education Preschool                         0.903890
## education Prof-school                        < 2e-16 ***
## education Some-college                      9.18e-12 ***
## education.num                                     NA
## marital.status Married-AF-spouse            0.000216 ***
## marital.status Married-civ-spouse            < 2e-16 ***
## marital.status Married-spouse-absent        0.792134
## marital.status Never-married                3.12e-08 ***
## marital.status Separated                    0.408300
## marital.status Widowed                      0.679703
## occupation Adm-clerical                     0.058013 .
## occupation Armed-Forces                     0.560904
## occupation Craft-repair                     0.022132 *
## occupation Exec-managerial                   < 2e-16 ***
## occupation Farming-fishing                  5.42e-08 ***
## occupation Handlers-cleaners                0.002062 **
## occupation Machine-op-inspct                0.818216
## occupation Other-service                    1.01e-05 ***
## occupation Priv-house-serv                  0.012513 *
## occupation Prof-specialty                   6.80e-11 ***
## occupation Protective-serv                  3.96e-07 ***
## occupation Sales                            3.98e-06 ***
```

```
## occupation Tech-support                     2.11e-10 ***
## occupation Transport-moving                       NA
## race Asian-Pac-Islander                     0.087089 .
## race Black                                  0.257266
## race Other                                  0.955316
## race White                                  0.049024 *
## sex Male                                    0.167986
## capital.gain                                 < 2e-16 ***
## capital.loss                                 < 2e-16 ***
## hours.per.week                               < 2e-16 ***
## native.country Cambodia                     0.015766 *
## native.country Canada                       0.071686 .
## native.country China                        0.272255
## native.country Columbia                     0.023523 *
## native.country Cuba                         0.336800
## native.country Dominican-Republic           0.224554
## native.country Ecuador                      0.960494
## native.country El-Salvador                  0.178702
## native.country England                      0.100775
## native.country France                       0.134695
## native.country Germany                      0.019668 *
## native.country Greece                       0.842229
## native.country Guatemala                    0.843963
## native.country Haiti                        0.409317
## native.country Holand-Netherlands           0.993090
## native.country Honduras                     0.979800
## native.country Hong                         0.873195
## native.country Hungary                      0.476185
## native.country India                        0.836707
## native.country Iran                         0.503425
## native.country Ireland                      0.236850
## native.country Italy                        0.007647 **
## native.country Jamaica                      0.511160
## native.country Japan                        0.059563 .
## native.country Laos                         0.935923
## native.country Mexico                       0.272948
## native.country Nicaragua                    0.677881
## native.country Outlying-US(Guam-USVI-etc) 0.977482
## native.country Peru                         0.736571
## native.country Philippines                  0.033455 *
## native.country Poland                       0.370252
## native.country Portugal                     0.370431
## native.country Puerto-Rico                  0.934414
## native.country Scotland                     0.755464
## native.country South                        0.212500
## native.country Taiwan                       0.600866
## native.country Thailand                     0.747614
## native.country Trinadad&Tobago              0.964738
## native.country United-States                0.011950 *
## native.country Vietnam                      0.333457
```

```
## native.country Yugoslavia                 0.271120
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 28759  on 26048  degrees of freedom
## Residual deviance: 16883  on 25956  degrees of freedom
## AIC: 17069
##
## Number of Fisher Scoring iterations: 14
```

Confusion matrix:

	<=50K	>50K
<=50K	4651	614
>50K	293	954

Computed accuracy :

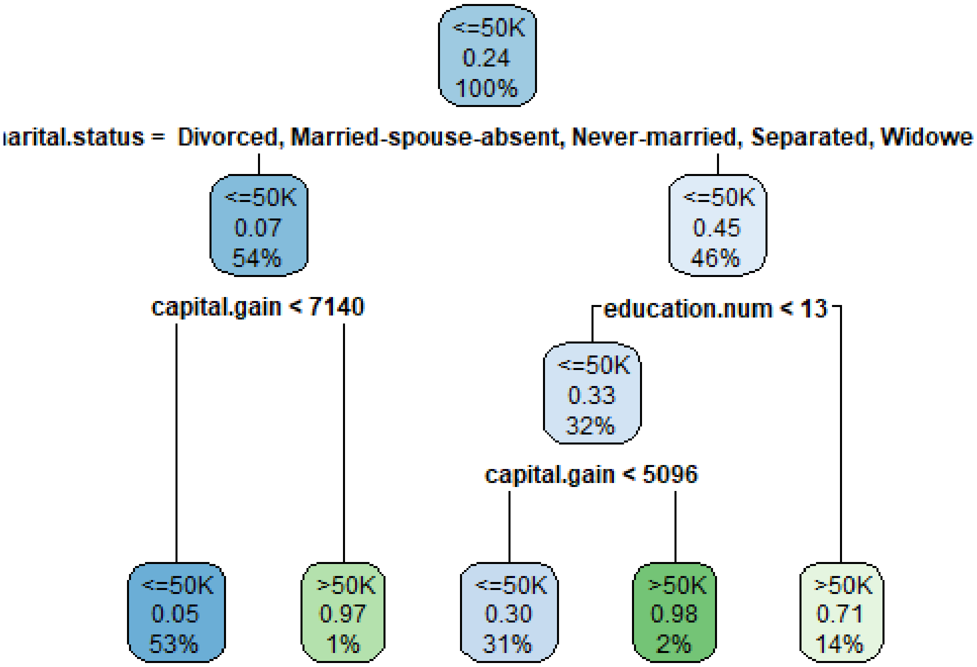
```
## [1] 0.8607187

## Warning: `data_frame()` was deprecated in tibble 1.1.0.
## i Please use `tibble()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

Model	Accuracy
Generalized Linear Model	0.8607187

Decision Tree Model

Decision tree can be used to visually and explicitly represent decisions and decision making for our data set.A decision tree describes data (but the resulting classification tree can be an input for decision making).

```
censustree <- rpart( income ~ . , method="class", data = train  )
# tree plot
rpart.plot(censustree)
```



From the above graph we can see that the Primary split is on marital.status and second node splits are based on capital.gain, education.

This can be verified by the variable importance as given below:

```
censustree$variable.importance

## marital.status   capital.gain  education.num            sex     occupation
##    1877.047594     818.866204     729.599308     609.026208     542.211651
##            age hours.per.week      workclass native.country   capital.loss
##     446.744533     239.976376     176.536146      17.390898      11.457533
##           race
##       3.273581
```

Confusion matrix:

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  <=50K  >50K
##      <=50K   4703   738
##      >50K     241   830
##
##                Accuracy : 0.8497
##                  95% CI : (0.8407, 0.8583)
##     No Information Rate : 0.7592
##     P-Value [Acc > NIR] : < 2.2e-16
##
```

```
##                   Kappa : 0.5389
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9513
##             Specificity : 0.5293
##          Pos Pred Value : 0.8644
##          Neg Pred Value : 0.7750
##              Prevalence : 0.7592
##          Detection Rate : 0.7222
##    Detection Prevalence : 0.8355
##       Balanced Accuracy : 0.7403
##
##        'Positive' Class :  <=50K
##
```

Computed accuracy:

Model	Accuracy
Decision Tree Model	0.8496622

Random Forest Model

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

```
train$income = factor(train$income)
censusforest <- randomForest(income ~ . ,data = train,importance = TRUE)
censusforest

##
## Call:
##  randomForest(formula = income ~ ., data = train, importance = TRUE)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 3
##
##         OOB estimate of  error rate: 13.74%
## Confusion matrix:
##         <=50K  >50K class.error
##  <=50K  18608  1168  0.05906149
##  >50K    2410  3863  0.38418619
```

Confusion matrix:

```
## Confusion Matrix and Statistics
##
##           Reference
```

```
## Prediction  <=50K  >50K
##      <=50K   4678   563
##      >50K     266  1005
##
##                Accuracy : 0.8727
##                  95% CI : (0.8644, 0.8807)
##     No Information Rate : 0.7592
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.6277
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9462
##             Specificity : 0.6409
##          Pos Pred Value : 0.8926
##          Neg Pred Value : 0.7907
##              Prevalence : 0.7592
##          Detection Rate : 0.7184
##    Detection Prevalence : 0.8048
##       Balanced Accuracy : 0.7936
##
##        'Positive' Class :  <=50K
##
```

Computed accuracy:

Model	Accuracy
Decision Tree Model	0.8496622
Random Forest Model	0.8726966

Results

Out of the 3 models that we trained, the accuracy for Random forest is the highest. Accuracy for Random forest is better than Decision tree, as random forest is an ensemble of many decision trees.

Conclusion

We started with the objective to create models that can predict if an individual can earn more than >50K. After data cleaning and identifying the independent variables, we built 3 models - Generalized linear model, Decision tree model and Random forest model, and trained the model.The results shows that Random forest model has better accuracy over the other two models used.

As part of further exploration, the ensemble of multiple models can be used to fine-tune the model further.

Refrence

[1] David R., et al. Modern Business Statistics. Cram101 Textbook Reviews, 2017 [2] Decision Tree Learning." Wikipedia, Wikimedia Foundation, 11 Apr. 2019, en.wikipedia.org/wiki/Decision_tree_learning. [3]"Random Forest." Wikipedia, Wikimedia Foundation, 9 Apr. 2019, en.wikipedia.org/wiki/Random_forest. [4] Lemon, Chet, et al. Predicting If Income Exceeds $50,000 per Year Based on 1994 US Census Data with Simple Classification Techniques. Predicting If Income Exceeds $50,000 per Year Based on 1994 US Census Data with Simple Classification Techniques.