Classifying Income of Census US Data With Machine Learning

Dhuha Alghanmi

3/26/2019

Introduction

A census is a procedure of systematically acquiring and recording information about the members of a given population. The term is used mostly in connection with national population and housing censuses; other common censuses include agriculture, business, and traffic censuses . The US Adult Census dataset is a repository of 48,842 entries provided by the UCI Machine Learning Repository This data was extracted from the 1994 Census Bureau by Ronny Kohavi and Barry Becker (Data Mining and Visualization, Silicon Graphics). The prediction task is to determine whether a person makes over $50K a year using simple machine learning model.

first let's look at the srtucture of the data

```r
str(adult)
```

```
## 'data.frame':    32561 obs. of  15 variables:
##  $ age           : int  90 82 66 54 41 34 38 74 68 41 ...
##  $ workclass     : Factor w/ 9 levels "?","Federal-gov",..: 1 5 1 5 5 5 5 8 2 5 ...
##  $ fnlwgt        : int  77053 132870 186061 140359 264663 216864 150601 88638 422013 70037 ...
##  $ education     : Factor w/ 16 levels "10th","11th",..: 12 12 16 6 16 12 1 11 12 16 ...
##  $ education.num : int  9 9 10 4 10 9 6 16 9 10 ...
##  $ marital.status: Factor w/ 7 levels "Divorced","Married-AF-spouse",..: 7 7 7 1 6 1 6 5 1 5 ...
##  $ occupation    : Factor w/ 15 levels "?","Adm-clerical",..: 1 5 1 8 11 9 2 11 11 4 ...
##  $ relationship  : Factor w/ 6 levels "Husband","Not-in-family",..: 2 2 5 5 4 5 5 3 2 5 ...
##  $ race          : Factor w/ 5 levels "Amer-Indian-Eskimo",..: 5 5 3 5 5 5 5 5 5 5 ...
##  $ sex           : Factor w/ 2 levels "Female","Male": 1 1 1 1 1 1 2 1 1 2 ...
##  $ 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: Factor w/ 42 levels "?","Cambodia",..: 40 40 40 40 40 40 40 40 40 1 ...
##  $ income        : Factor w/ 2 levels "<=50K",">50K": 1 1 1 1 1 1 1 2 1 2 ...
```

Attribute

The Data -age: the age of an individual. – Integer bigger than 0. -workclass:: a general term to represent the employment status of an individual. –Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked. -fnlwgt: final weight. this is the number of people the census believes the entry represents –continuous. -education:: the highest level of education achieved by an individual. –Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acad, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool. -education-num:: the highest level of education achieved in numerical form. -marital-status: Married-civ-spouse, Divorced, etc. -occupation:: the general type of occupation of an individual. –Tech-support, Craft-repair, Other-service, Sales, etc. -relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried. -race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black. -sex: Female, Male. -capital-gain:: capital gains for an individual. -capital-loss:: capital loss for an individual. -hours-per-week:: the hours an individual has reported to work per week -native-country:

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

checking to see if any NA values

```
adult %>% anyNA()
```

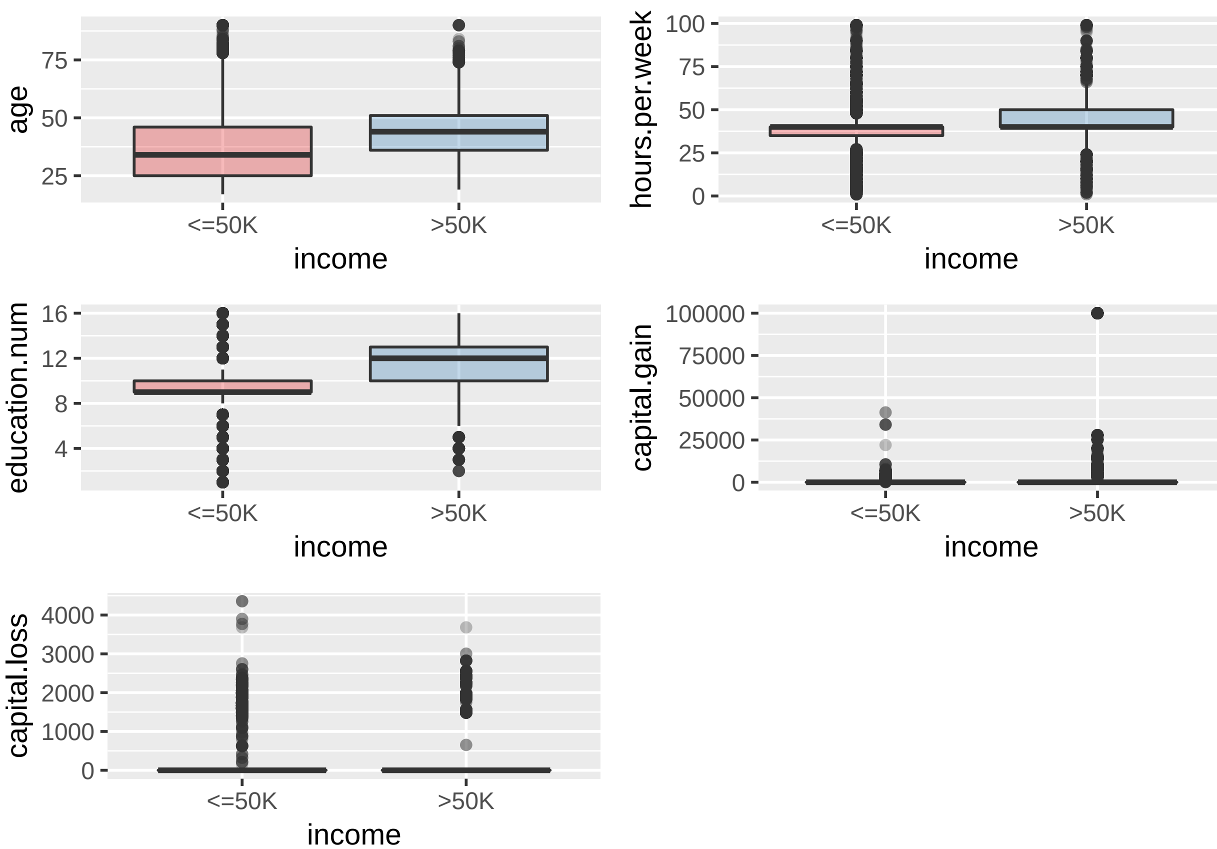
```
## [1] FALSE
```

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

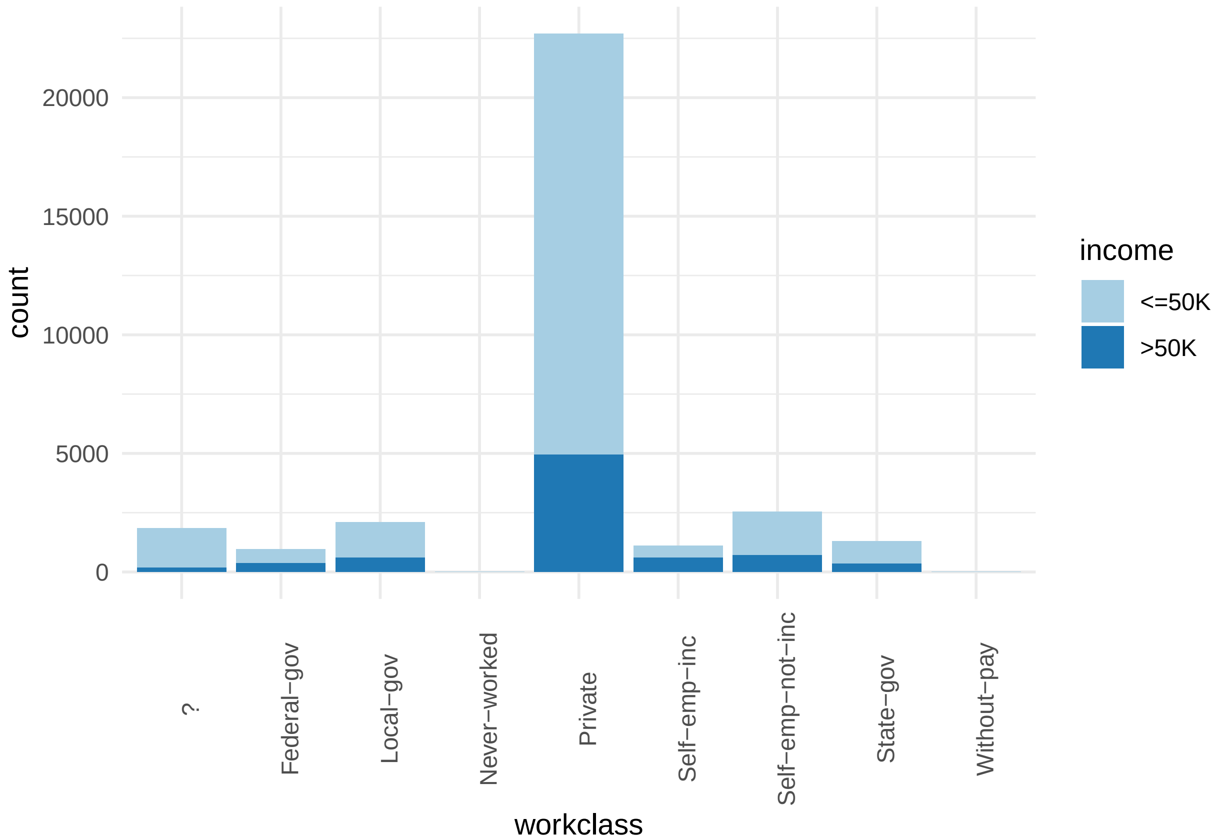
```
adult$fnlwgt <- NULL
adult$relationship <- NULL
```

Explotory Analysis

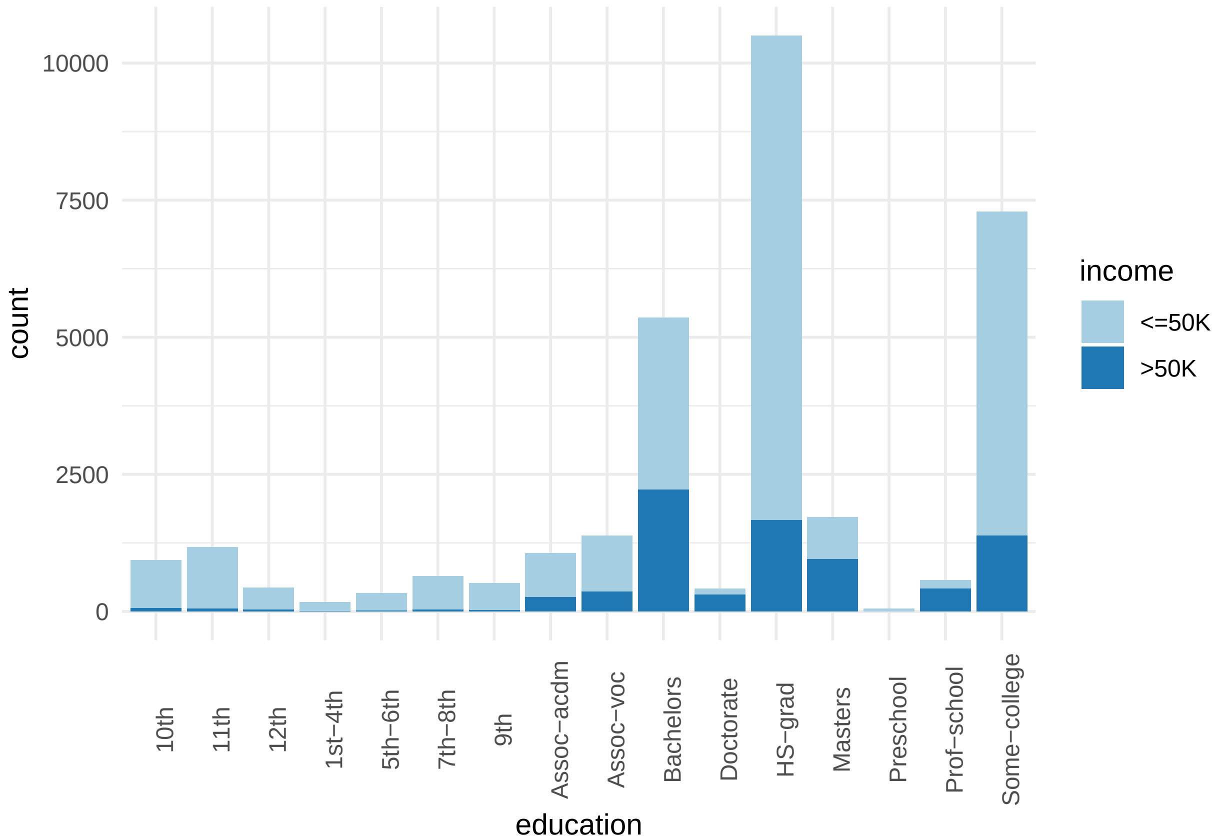
To gain insights about which features would be most helpful for this analysis I plotted a boxplot for all continuous variable



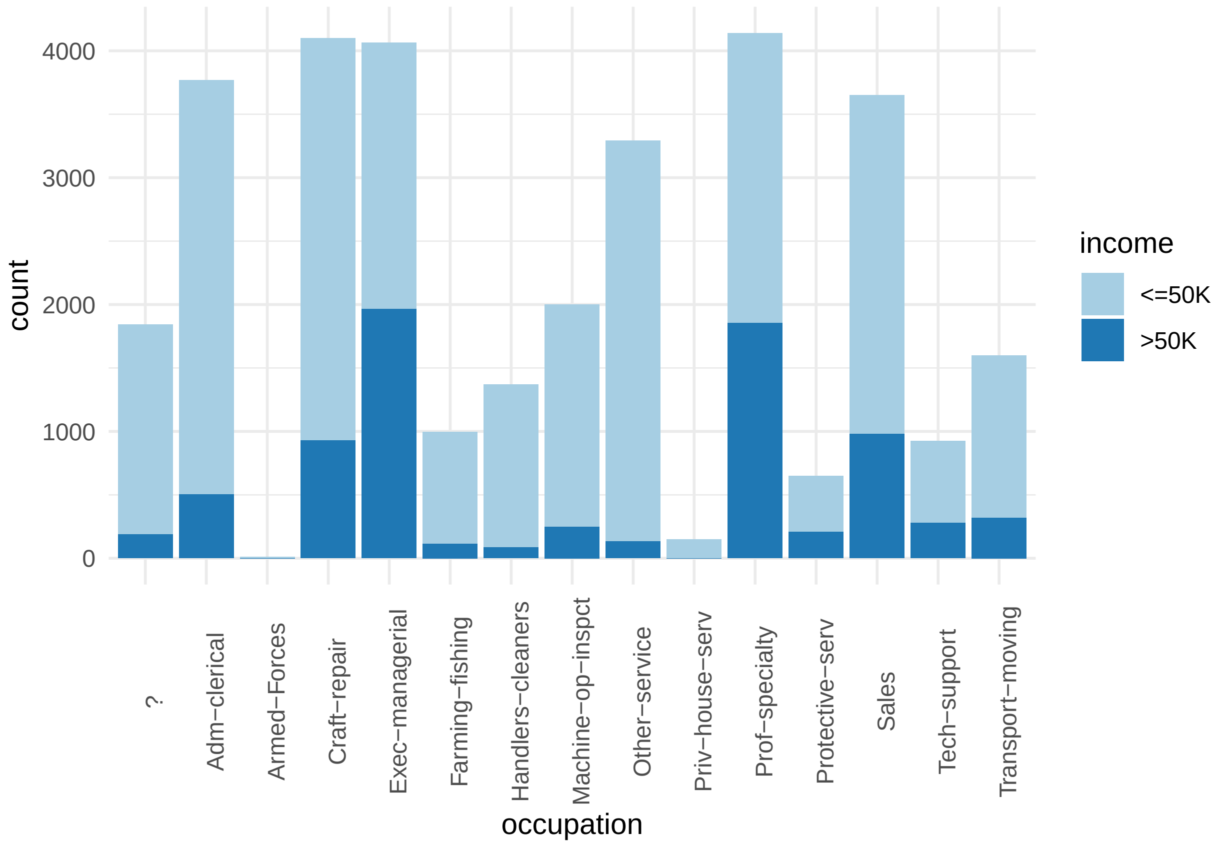
from this graph we can see that all variables can affect the outcome.



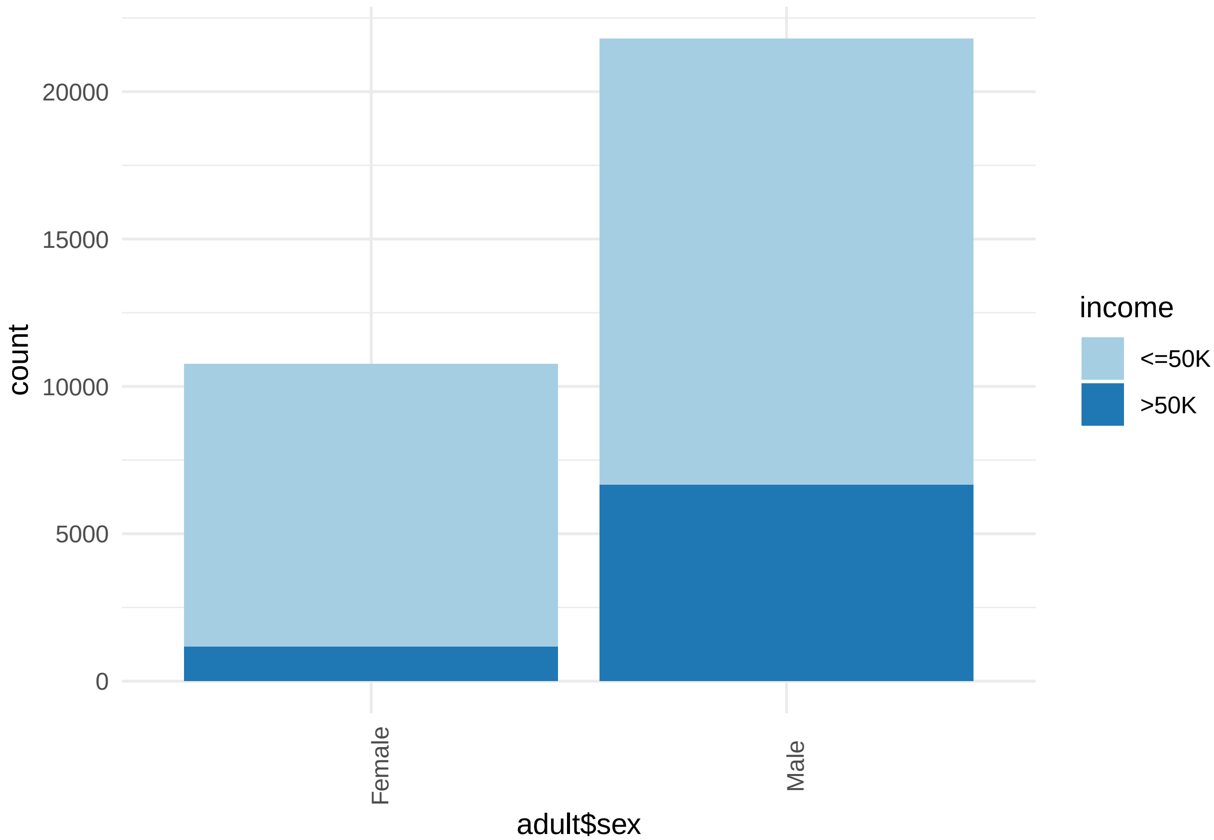
The majority of individuals in the data work in private sector and all workclass seem to have a good chance of earning more than $50K.



The vaiable education represnt the latest education level for indivisuals , which most of indivisuals are high-school graduate. Doctotate , Masters and Profissional-school seems to have the majority of Income higher than 50K income and but for the first grade to high school the chances are less of earning over 50K.



the dataset indivisuals occupation's does not seems uniform.as seen exec-managerial and prof-specialty stand out at having a higher than 50K income oppesite from Farming-fishing and Handlers-cleaners which stand in the lower than 50K income.



the individuals are mostly males also the percentage of males who make greater than $50,000 is much greater than the percentage of females that make the same amount.

Data Partition

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

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

splittd the data to 80% for training the models and 20% for testing.

Machine Learning Techniques - Model Fitting

###Logistic Regression Model

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

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

```
##
## Call:
## glm(formula = income ~ ., family = binomial, data = train)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.0770  -0.4974  -0.2056  -0.0411   3.7404
##
## Coefficients: (2 not defined because of singularities)
##                                       Estimate Std. Error z value
## (Intercept)                          -8.206e+00  3.838e-01 -21.381
## age                                   2.388e-02  1.816e-03  13.150
## workclassFederal-gov                  1.197e+00  1.731e-01   6.917
## workclassLocal-gov                    4.589e-01  1.580e-01   2.904
## workclassNever-worked                -1.122e+01  5.111e+02  -0.022
## workclassPrivate                      6.920e-01  1.410e-01   4.909
## workclassSelf-emp-inc                 8.913e-01  1.692e-01   5.267
## workclassSelf-emp-not-inc             2.226e-01  1.542e-01   1.443
## workclassState-gov                    3.932e-01  1.702e-01   2.310
## workclassWithout-pay                 -1.280e+01  3.654e+02  -0.035
## education11th                         4.665e-02  2.400e-01   0.194
## education12th                         6.429e-01  2.989e-01   2.151
## education1st-4th                     -7.569e-01  6.065e-01  -1.248
## education5th-6th                     -8.353e-02  3.548e-01  -0.235
## education7th-8th                     -2.941e-01  2.543e-01  -1.156
## education9th                         -1.746e-01  3.007e-01  -0.581
## educationAssoc-acdm                   1.358e+00  2.003e-01   6.780
## educationAssoc-voc                    1.410e+00  1.927e-01   7.318
## educationBachelors                    2.073e+00  1.792e-01  11.570
## educationDoctorate                    3.096e+00  2.405e-01  12.870
## educationHS-grad                      8.576e-01  1.747e-01   4.908
## educationMasters                      2.397e+00  1.904e-01  12.588
## educationPreschool                   -1.982e+01  1.640e+02  -0.121
## educationProf-school                  3.008e+00  2.267e-01  13.271
## educationSome-college                 1.196e+00  1.771e-01   6.754
## education.num                                NA         NA      NA
## marital.statusMarried-AF-spouse       2.729e+00  5.128e-01   5.322
## marital.statusMarried-civ-spouse      2.142e+00  7.430e-02  28.833
## marital.statusMarried-spouse-absent   9.330e-02  2.438e-01   0.383
## marital.statusNever-married          -5.723e-01  9.209e-02  -6.215
## marital.statusSeparated              -2.189e-01  1.810e-01  -1.209
## marital.statusWidowed                 8.181e-02  1.626e-01   0.503
## occupationAdm-clerical                8.569e-02  1.098e-01   0.780
## occupationArmed-Forces               -1.132e+00  1.550e+00  -0.731
## occupationCraft-repair                1.088e-01  9.437e-02   1.153
## occupationExec-managerial             8.176e-01  9.725e-02   8.408
## occupationFarming-fishing            -1.028e+00  1.594e-01  -6.451
## occupationHandlers-cleaners          -6.192e-01  1.606e-01  -3.856
## occupationMachine-op-inspct          -9.431e-02  1.164e-01  -0.810
## occupationOther-service              -7.374e-01  1.394e-01  -5.290
## occupationPriv-house-serv            -3.956e+00  2.087e+00  -1.895
## occupationProf-specialty              5.471e-01  1.043e-01   5.246
## occupationProtective-serv             6.141e-01  1.463e-01   4.198
## occupationSales                       3.157e-01  9.996e-02   3.159
## occupationTech-support                6.538e-01  1.333e-01   4.905
## occupationTransport-moving                   NA         NA      NA
## raceAsian-Pac-Islander                7.662e-01  3.000e-01   2.554
```

```
## raceBlack                                5.188e-01  2.569e-01   2.020
## raceOther                               -4.237e-02  4.015e-01  -0.106
## raceWhite                                6.778e-01  2.446e-01   2.771
## sexMale                                  1.315e-01  5.858e-02   2.244
## capital.gain                             3.198e-04  1.141e-05  28.040
## capital.loss                             6.603e-04  4.156e-05  15.889
## hours.per.week                           2.883e-02  1.796e-03  16.051
## native.countryCambodia                   1.048e+00  7.247e-01   1.446
## native.countryCanada                     6.070e-01  3.314e-01   1.831
## native.countryChina                     -5.684e-01  4.452e-01  -1.277
## native.countryColumbia                  -2.641e+00  1.130e+00  -2.338
## native.countryCuba                       4.381e-01  3.825e-01   1.145
## native.countryDominican-Republic        -1.433e+00  1.059e+00  -1.353
## native.countryEcuador                   -2.577e-01  8.505e-01  -0.303
## native.countryEl-Salvador               -1.504e-01  5.562e-01  -0.270
## native.countryEngland                    6.918e-01  3.591e-01   1.927
## native.countryFrance                     1.011e+00  5.642e-01   1.791
## native.countryGermany                    5.214e-01  3.137e-01   1.662
## native.countryGreece                    -1.262e+00  7.675e-01  -1.644
## native.countryGuatemala                 -5.309e-01  9.702e-01  -0.547
## native.countryHaiti                     -7.398e-01  9.156e-01  -0.808
## native.countryHoland-Netherlands        -1.263e+01  1.455e+03  -0.009
## native.countryHonduras                  -7.607e-01  2.321e+00  -0.328
## native.countryHong                       1.688e-01  7.747e-01   0.218
## native.countryHungary                    1.099e+00  1.162e+00   0.946
## native.countryIndia                     -3.694e-01  3.664e-01  -1.008
## native.countryIran                       4.551e-01  5.467e-01   0.832
## native.countryIreland                    4.848e-01  7.588e-01   0.639
## native.countryItaly                      1.066e+00  3.763e-01   2.833
## native.countryJamaica                    9.509e-02  5.283e-01   0.180
## native.countryJapan                      9.776e-01  4.758e-01   2.055
## native.countryLaos                       6.537e-03  8.986e-01   0.007
## native.countryMexico                    -3.360e-01  2.906e-01  -1.156
## native.countryNicaragua                 -1.069e+00  1.088e+00  -0.983
## native.countryOutlying-US(Guam-USVI-etc) -1.286e+01  3.924e+02  -0.033
## native.countryPeru                      -4.175e-01  8.952e-01  -0.466
## native.countryPhilippines                4.032e-01  3.170e-01   1.272
## native.countryPoland                    -4.021e-02  4.850e-01  -0.083
## native.countryPortugal                  -2.776e-03  7.364e-01  -0.004
## native.countryPuerto-Rico               -1.791e-01  4.704e-01  -0.381
## native.countryScotland                   1.372e-01  9.321e-01   0.147
## native.countrySouth                     -7.928e-01  5.051e-01  -1.570
## native.countryTaiwan                    -3.215e-01  5.263e-01  -0.611
## native.countryThailand                  -1.191e+00  1.185e+00  -1.005
## native.countryTrinadad&Tobago           -9.408e-02  8.631e-01  -0.109
## native.countryUnited-States              3.932e-01  1.538e-01   2.557
## native.countryVietnam                   -6.238e-01  6.354e-01  -0.982
## native.countryYugoslavia                 4.530e-01  7.136e-01   0.635
##                                         Pr(>|z|)
## (Intercept)                              < 2e-16 ***
## age                                      < 2e-16 ***
## workclassFederal-gov                    4.63e-12 ***
## workclassLocal-gov                      0.003688 **
## workclassNever-worked                   0.982488
```

```
## workclassPrivate                        9.16e-07 ***
## workclassSelf-emp-inc                   1.38e-07 ***
## workclassSelf-emp-not-inc               0.148986
## workclassState-gov                      0.020909 *
## workclassWithout-pay                    0.972062
## education11th                           0.845881
## education12th                           0.031481 *
## education1st-4th                        0.212018
## education5th-6th                        0.813901
## education7th-8th                        0.247593
## education9th                            0.561421
## educationAssoc-acdm                     1.20e-11 ***
## educationAssoc-voc                      2.52e-13 ***
## educationBachelors                       < 2e-16 ***
## educationDoctorate                       < 2e-16 ***
## educationHS-grad                        9.20e-07 ***
## educationMasters                         < 2e-16 ***
## educationPreschool                      0.903795
## educationProf-school                     < 2e-16 ***
## educationSome-college                   1.44e-11 ***
## education.num                                 NA
## marital.statusMarried-AF-spouse         1.02e-07 ***
## marital.statusMarried-civ-spouse         < 2e-16 ***
## marital.statusMarried-spouse-absent     0.701949
## marital.statusNever-married             5.14e-10 ***
## marital.statusSeparated                 0.226591
## marital.statusWidowed                   0.614859
## occupationAdm-clerical                  0.435181
## occupationArmed-Forces                  0.464936
## occupationCraft-repair                  0.248866
## occupationExec-managerial                < 2e-16 ***
## occupationFarming-fishing               1.11e-10 ***
## occupationHandlers-cleaners             0.000115 ***
## occupationMachine-op-inspct             0.417674
## occupationOther-service                 1.22e-07 ***
## occupationPriv-house-serv               0.058069 .
## occupationProf-specialty                1.56e-07 ***
## occupationProtective-serv               2.69e-05 ***
## occupationSales                         0.001584 **
## occupationTech-support                  9.36e-07 ***
## occupationTransport-moving                    NA
## raceAsian-Pac-Islander                  0.010637 *
## raceBlack                               0.043404 *
## raceOther                               0.915958
## raceWhite                               0.005589 **
## sexMale                                 0.024800 *
## capital.gain                             < 2e-16 ***
## capital.loss                             < 2e-16 ***
## hours.per.week                           < 2e-16 ***
## native.countryCambodia                  0.148270
## native.countryCanada                    0.067042 .
## native.countryChina                     0.201687
## native.countryColumbia                  0.019366 *
## native.countryCuba                      0.252141
```

```
## native.countryDominican-Republic        0.176158
## native.countryEcuador                    0.761900
## native.countryEl-Salvador                0.786907
## native.countryEngland                    0.054035 .
## native.countryFrance                     0.073304 .
## native.countryGermany                    0.096524 .
## native.countryGreece                     0.100130
## native.countryGuatemala                  0.584228
## native.countryHaiti                      0.419082
## native.countryHoland-Netherlands         0.993074
## native.countryHonduras                   0.743114
## native.countryHong                       0.827498
## native.countryHungary                    0.344222
## native.countryIndia                      0.313448
## native.countryIran                       0.405147
## native.countryIreland                    0.522857
## native.countryItaly                      0.004617 **
## native.countryJamaica                    0.857162
## native.countryJapan                      0.039901 *
## native.countryLaos                       0.994196
## native.countryMexico                     0.247734
## native.countryNicaragua                  0.325708
## native.countryOutlying-US(Guam-USVI-etc) 0.973849
## native.countryPeru                       0.640986
## native.countryPhilippines                0.203454
## native.countryPoland                     0.933917
## native.countryPortugal                   0.996992
## native.countryPuerto-Rico                0.703353
## native.countryScotland                   0.883006
## native.countrySouth                      0.116487
## native.countryTaiwan                     0.541265
## native.countryThailand                   0.314840
## native.countryTrinadad&Tobago            0.913207
## native.countryUnited-States              0.010570 *
## native.countryVietnam                    0.326266
## native.countryYugoslavia                 0.525573
## ---
## 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: 16684  on 25956  degrees of freedom
## AIC: 16870
##
## Number of Fisher Scoring iterations: 14
```

all variable seem to be Significant except for native country.

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type =
## ifelse(type == : prediction from a rank-deficient fit may be misleading
```

	<=50K	>50K
<=50K	4609	637

	<=50K	>50K
>50K	335	931

Computed accuracy :

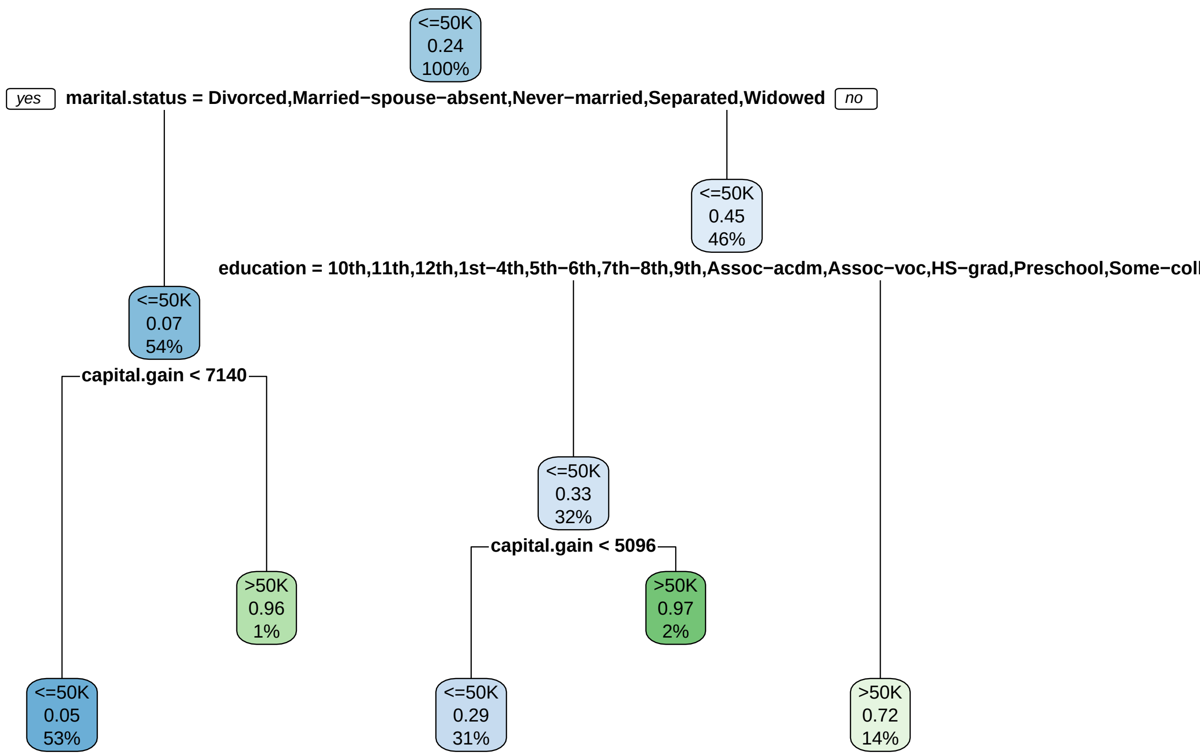
```
## [1] 0.8507371
```

Model	Accuracy
Generalized Linear Model	0.8507371

Decision Tree Model

A classification tree model is created using the rpart package. It is one of the predictive modeling approaches used in statistics, data mining and machine learning. Tree models where the target variable can take a discrete set of values are called classification trees.In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. In data mining, a decision tree describes data (but the resulting classification tree can be an input for decision making).

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



The Primary splits of the first node are marital.status,capital.gain and education,second node primary splits: capital.gain, education, occupation and hour per week as seen in the decision tree.

```r
censustree$variable.importance
```

```
## marital.status   capital.gain      education  education.num            sex
##     1884.43222      822.82590      790.19635      790.19635      608.56705
```

```
##      occupation            age hours.per.week      workclass native.country
##       583.21381      442.60854      256.65312      174.46114       19.16193
##    capital.loss
##        10.35384
```

Confusion matrix and auccarcy for the tree model:

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction <=50K >50K
##      <=50K  4675  784
##      >50K    269  784
##
##                Accuracy : 0.8383
##                  95% CI : (0.8291, 0.8472)
##     No Information Rate : 0.7592
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5019
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.9456
##             Specificity : 0.5000
##          Pos Pred Value : 0.8564
##          Neg Pred Value : 0.7445
##              Prevalence : 0.7592
##          Detection Rate : 0.7179
##    Detection Prevalence : 0.8383
##       Balanced Accuracy : 0.7228
##
##        'Positive' Class : <=50K
##
```

Model	Accuracy
Generalized Linear Model	0.8507371
Decision Tree Model	0.8382985

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.

```r
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: 13.89%
## Confusion matrix:
##       <=50K >50K class.error
## <=50K 18370 1406  0.07109628
## >50K   2213 4060  0.35278176
```

Confusion matrix and auccarcy for the random forest model:

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction <=50K >50K
##      <=50K  4596  541
##      >50K    348 1027
##
##                Accuracy : 0.8635
##                  95% CI : (0.8549, 0.8717)
##     No Information Rate : 0.7592
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.6102
##  Mcnemar's Test P-Value : 1.199e-10
##
##             Sensitivity : 0.9296
##             Specificity : 0.6550
##          Pos Pred Value : 0.8947
##          Neg Pred Value : 0.7469
##              Prevalence : 0.7592
##          Detection Rate : 0.7058
##    Detection Prevalence : 0.7889
##       Balanced Accuracy : 0.7923
##
##        'Positive' Class : <=50K
##
```

Results and Conclusion

First I visuialize and analysis of the data that was and performing machine learning algorithims GLM, Decision Tree(CARET)and Random Forest the least accuracy concuctet was 0.8407 for th Caret model and the highest 0.8639435 for the Random Forest Model which is expected beacause the random forest is an ensemble of Decicion Tree.

Model	Accuracy
Generalized Linear Model	0.8507371
Decision Tree Model	0.8382985
Random Forest Model	0.8634828

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.