**Adult Census Income Analysis and Prediction Using Logistic Regression**

**Contents**

Introduction 1

Data set Description 2

Analysis 4

Model Building 13

Conclusion 19

References 20

**Introduction**

A **census** is the procedure of systematically acquiring and recording [information](https://en.wikipedia.org/wiki/Information) about the members of a given [population](https://en.wikipedia.org/wiki/Statistical_population). This term is used mostly in connection with [national population and housing censuses](https://en.wikipedia.org/wiki/Population_and_housing_censuses_by_country); other common censuses include traditional culture, business, income, agricultural, and traffic censuses. In this project we worked on the Census data set. It is also known as “Census Income” dataset. This data was extracted from the [1994 Census bureau database](http://www.census.gov/en.html) and our aim is to predict whether income exceeds $50K/year based on census. We have a summary of our analysis and exploration of the Adult Census Data to come up with meaningful, important and interesting attributes of the data. The question is inspected in two different approaches – traditional statistical modeling and machine learning techniques. Logistic regression is used as the statistical modeling tool as the outcome is binary. Two different machine learning techniques – Support vector machine, decision tree, are used to answer the same question.

**Data set description**

The data extracted from the [1994 Census bureau database](http://www.census.gov/en.html) consist of 32560 records and a binomial label indicating a salary of <50K or >50K USD. In the census data every record represents a person with 14 attributes, the last element of a record is one of the labels {“>=50K”,”<50K”}.

**Attribute Information:**

Listing of attributes:   
>50K, <=50K.   
age: continuous.   
workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.   
fnlwgt: continuous.   
education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.   
education-num: continuous.   
marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.   
occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.   
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: continuous.   
capital-loss: continuous.   
hours-per-week: continuous.   
native-country: United-States, Cambodia, England, China, Cuba,Jamaica, France, Germany, Mexico

**Load the data**

data=read.csv("C:/Users/Deepthi/Downloads/adult.csv")

dim(data)

## [1] 32560 15

Since the dataset doesn’t have any column names we will give columns names from the data description.

feauture\_name=c("Age","Workclass","fnlwgt","Education","Education-num","Marital-status","Occupation","Relationship","Race","Sex","Capital-gain","Capital-loss","Hours-per-week","Native-country","Income")  
colnames(data)=feauture\_name  
head(data)

## Age Workclass fnlwgt Education Education-num Marital-status  
## 1 50 Self-emp-not-inc 83311 Bachelors 13 Married-civ-spouse  
## 2 38 Private 215646 HS-grad 9 Divorced  
## 3 53 Private 234721 11th 7 Married-civ-spouse  
## 4 28 Private 338409 Bachelors 13 Married-civ-spouse  
## 5 37 Private 284582 Masters 14 Married-civ-spouse  
## 6 49 Private 160187 9th 5 Married-spouse-absent  
## Occupation Relationship Race Sex Capital-gain Capital-loss  
## 1 Exec-managerial Husband White Male 0 0  
## 2 Handlers-cleaners Not-in-family White Male 0 0  
## 3 Handlers-cleaners Husband Black Male 0 0  
## 4 Prof-specialty Wife Black Female 0 0  
## 5 Exec-managerial Wife White Female 0 0  
## 6 Other-service Not-in-family Black Female 0 0  
## Hours-per-week Native-country Income  
## 1 13 United-States <=50K  
## 2 40 United-States <=50K  
## 3 40 United-States <=50K  
## 4 40 Cuba <=50K  
## 5 40 United-States <=50K  
## 6 16 Jamaica <=50K

# A high level discription of the data

str(data)

## 'data.frame': 32560 obs. of 15 variables:  
## $ Age : int 50 38 53 28 37 49 52 31 42 37 ...  
## $ Workclass : Factor w/ 9 levels "?","Federal-gov",..: 7 5 5 5 5 5 7 5 5 5 ...  
## $ fnlwgt : int 83311 215646 234721 338409 284582 160187 209642 45781 159449 280464 ...  
## $ Education : Factor w/ 16 levels "10th","11th",..: 10 12 2 10 13 7 12 13 10 16 ...  
## $ Education-num : int 13 9 7 13 14 5 9 14 13 10 ...  
## $ Marital-status: Factor w/ 7 levels "Divorced","Married-AF-spouse",..: 3 1 3 3 3 4 3 5 3 3 ...  
## $ Occupation : Factor w/ 15 levels "?","Adm-clerical",..: 5 7 7 11 5 9 5 11 5 5 ...  
## $ Relationship : Factor w/ 6 levels "Husband","Not-in-family",..: 1 2 1 6 6 2 1 2 1 1 ...  
## $ Race : Factor w/ 5 levels "Amer-Indian-Eskimo",..: 5 5 3 3 5 3 5 5 5 3 ...  
## $ Sex : Factor w/ 2 levels "Female","Male": 2 2 2 1 1 1 2 1 2 2 ...  
## $ Capital-gain : int 0 0 0 0 0 0 0 14084 5178 0 ...  
## $ Capital-loss : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ Hours-per-week: int 13 40 40 40 40 16 45 50 40 80 ...  
## $ Native-country: Factor w/ 42 levels "?","Cambodia",..: 40 40 40 6 40 24 40 40 40 40 ...  
## $ Income : Factor w/ 2 levels "<=50K",">50K": 1 1 1 1 1 1 2 2 2 2 ...

**Inference: The structure of a dataset gives basic data structures present in the**

**data. The basic data structures used in r include vectors, lists, matrices, data**

**frames and factors. It is used to handle multiple values that means we can’t**

**work with data having single value.**

**Exploratory analysis of the data**

library(dplyr)

continuous <-select\_if(data, is.numeric)  
summary(continuous)

## Age fnlwgt Education-num Capital-gain   
## Min. :17.00 Min. : 12285 Min. : 1.00 Min. : 0   
## 1st Qu.:28.00 1st Qu.: 117832 1st Qu.: 9.00 1st Qu.: 0   
## Median :37.00 Median : 178363 Median :10.00 Median : 0   
## Mean :38.58 Mean : 189782 Mean :10.08 Mean : 1078   
## 3rd Qu.:48.00 3rd Qu.: 237055 3rd Qu.:12.00 3rd Qu.: 0   
## Max. :90.00 Max. :1484705 Max. :16.00 Max. :99999   
## Capital-loss Hours-per-week   
## Min. : 0.00 Min. : 1.00   
## 1st Qu.: 0.00 1st Qu.:40.00   
## Median : 0.00 Median :40.00   
## Mean : 87.31 Mean :40.44   
## 3rd Qu.: 0.00 3rd Qu.:45.00   
## Max. :4356.00 Max. :99.00

**Inference: The exploratory data analysis gives the basic statistical calculations like mean, median etc... for each variable present in the data. The above data gives the analysis of the numeric variables. The attribute age has a mean of 38.75 and hours per week shows a mean of 40.**

# Checking for missing values

colSums(is.na(data))

## 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

# Checking for empty values

colSums(data=='?')

## 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

datanew=data[!data$Workclass=="?"&!data$Occupation=="?"&!data$`Native-country`=="?", ]  
colSums(datanew=="?",)

## 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

**Inference: Null is a term used to represent a missing value .Here we are**

**checking if there are any null values in our data because if there is any null**

**values in the data it will affect the models and will reduce the accuracy.**

**Hence, we want to omit the null values if there are any. The above code shows**

**that there arenull values in the attributes work class, occupation and native country. So we will remove the empty values to increase the accuracy of model.**

datanew$fnlwgt=NULL  
head(datanew)

## Age Workclass Education Education-num Marital-status  
## 1 50 Self-emp-not-inc Bachelors 13 Married-civ-spouse  
## 2 38 Private HS-grad 9 Divorced  
## 3 53 Private 11th 7 Married-civ-spouse  
## 4 28 Private Bachelors 13 Married-civ-spouse  
## 5 37 Private Masters 14 Married-civ-spouse  
## 6 49 Private 9th 5 Married-spouse-absent  
## Occupation Relationship Race Sex Capital-gain Capital-loss  
## 1 Exec-managerial Husband White Male 0 0  
## 2 Handlers-cleaners Not-in-family White Male 0 0  
## 3 Handlers-cleaners Husband Black Male 0 0  
## 4 Prof-specialty Wife Black Female 0 0  
## 5 Exec-managerial Wife White Female 0 0  
## 6 Other-service Not-in-family Black Female 0 0  
## Hours-per-week Native-country Income  
## 1 13 United-States <=50K  
## 2 40 United-States <=50K  
## 3 40 United-States <=50K  
## 4 40 Cuba <=50K  
## 5 40 United-States <=50K  
## 6 16 Jamaica <=50K

**The ‘fnlwgt’ final weight estimate refers to population totals derived from CPS by creating “weighted tallies” of any specified socio-economic characteristics of the population. This variable is removed from the training data set due to it’s diminished impact on income level.**

# Checking the number of unique values in each column

sapply(datanew, function(x) length(unique(x)))

## Age Workclass Education Education-num Marital-status   
## 72 7 16 16 7   
## Occupation Relationship Race Sex Capital-gain   
## 14 6 5 2 118   
## Capital-loss Hours-per-week Native-country Income   
## 90 94 41 2

After removing the null values we are again taking the summary of the dataset which will be more accurate.

dim(datanew)

## [1] 30161 14

continuous <-select\_if(datanew, is.numeric)  
summary(continuous)

## Age Education-num Capital-gain Capital-loss   
## Min. :17.00 Min. : 1.00 Min. : 0 Min. : 0.00   
## 1st Qu.:28.00 1st Qu.: 9.00 1st Qu.: 0 1st Qu.: 0.00   
## Median :37.00 Median :10.00 Median : 0 Median : 0.00   
## Mean :38.44 Mean :10.12 Mean : 1092 Mean : 88.38   
## 3rd Qu.:47.00 3rd Qu.:13.00 3rd Qu.: 0 3rd Qu.: 0.00   
## Max. :90.00 Max. :16.00 Max. :99999 Max. :4356.00   
## Hours-per-week   
## Min. : 1.00   
## 1st Qu.:40.00   
## Median :40.00   
## Mean :40.93   
## 3rd Qu.:45.00   
## Max. :99.00

**Correlation**

cor(datanew$`Education-num`,datanew$`Hours-per-week`)

## [1] 0.1525282

cor(datanew$`Education-num`,datanew$Age)

## [1] 0.04352541

cor(datanew$`Education-num`,datanew$`Capital-gain`)

## [1] 0.1244132

cor(datanew$`Education-num`,datanew$`Capital-loss`)

## [1] 0.07965634

cor(datanew$`Education-num`,datanew$Age)

## [1] 0.04352541

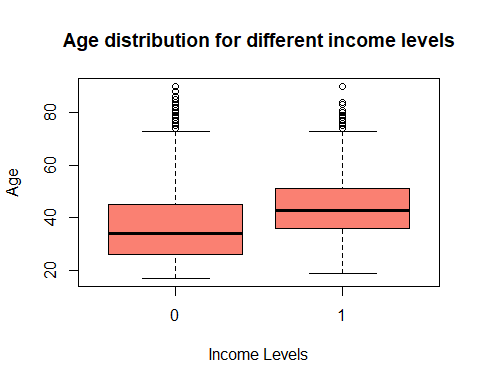
cor(datanew$`Education-num`,datanew$Age)

## [1] 0.04352541

**Inference: Correlation is a method of statistical evaluation used to study the strength of a relationship between two numerically measured variables. Its value varies from -1 to 1. If the value of correlation is -1 which shows a strong negative correlation, if the value is 1 it means there exist a strong negative correlation and the value of correlation = 0 means there is a weak relation exist between the variables . Here we are analyzing the correlation between education num and other numeric variables. The analysis shows that the all values are close to zero. Therefore there exists a weak relation between the variables.**

Boxplot

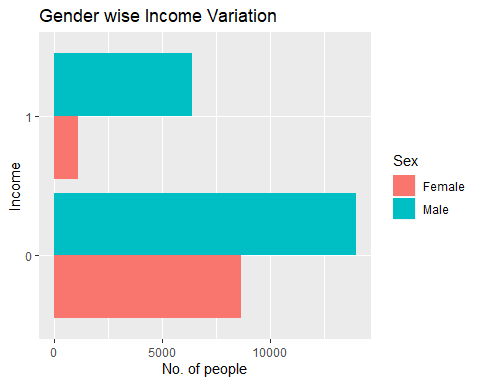
boxplot (Age ~Income, data = datanew,   
main ="Age distribution for different income levels",  
xlab ="Income Levels", ylab ="Age", col ="salmon")



**The boxplot of age distribution for different income levels shows outliers and overlapping.**

**Does Sex have any influence on income?**

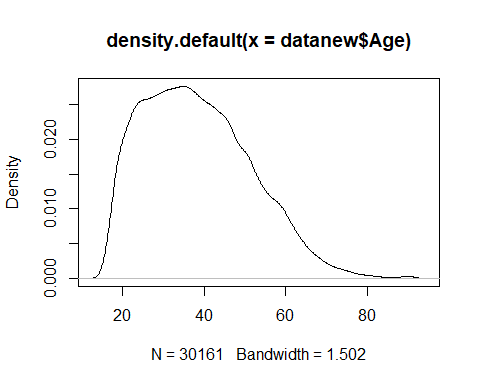
library(ggplot2)  
ggplot(datanew,aes(x=Income,fill=Sex)) +geom\_bar(position="dodge")+coord\_flip()+labs(title ="Gender wise Income Variation",y="No. of people",x="Income")



**Inference: The above graph shows gender wise income variation of a group of people. Blue color represents male and red color represents female. The figure shows a large variation in the number of males and females. And the number of people who receives an income >50k and <=50k shows a great variation. From the above graph we can conclude that more people are having income <=50k and the income of males are more than that of women. Therefore income varies according to gender and the numbers of women workers are far less than that of men workers.**

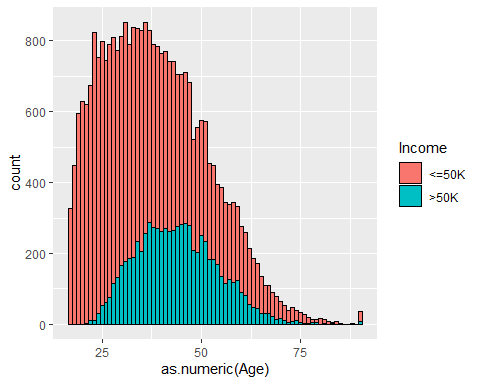
**Density plot**

d=density(datanew$Age)  
plot(d)



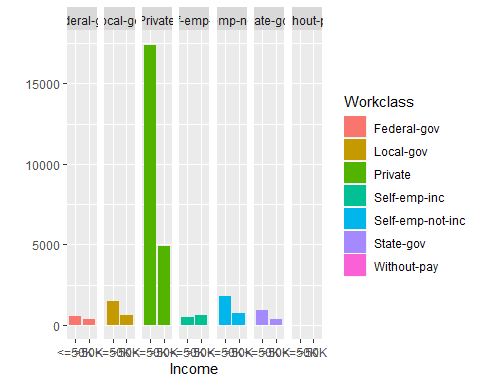
**Inference: A density plot is a representation of distribution of a numericvariable. It uses a kernel density estimate to show the probability density function of the variable. Density plots can be considered as plots of smoothed histograms. The visual representation of density plot will show whether the variable is normally distributed or not. The above graph shows that the density plot of attribute ‘age’ which shows a normal distribution since the shape of the graph is almost a bell shaped curve which is the representation of normal distribution.**

ggplot(datanew) +aes(x=as.numeric(Age), group=Income, fill=Income) +  
geom\_histogram(binwidth=1, color='black')

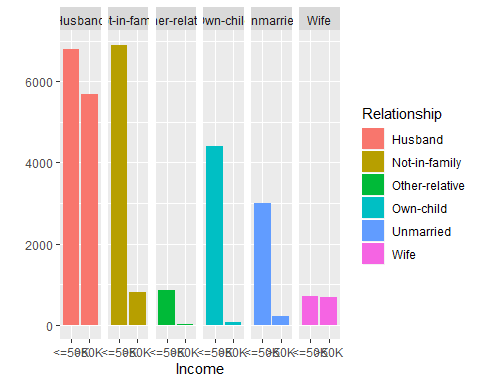


**Inference: We see that most people have an income of less than or equal to 50k. More number of middle age people earn greater than 50k.**

qplot (Income, data = datanew, fill =) +facet\_grid (. ~Workclass)



**Inference: Most people from private sector earn less than 50k. Most of the income generated comes from the private sector. People who work in the other sectors are very less in comparison to private sector.**



**Inference: Husbands and people not in family are high in number among people who earn an income. A fairly large number of husbands earn greater than 50k. There is**

**only a slight difference in number of husbands who earn less than or equal to 50k and greater than 50k. Number of wives who contribute to income are very less.**

Building training and test data

set.seed(100)  
Index <-sample(1:nrow(datanew), 0.7\*nrow(datanew))  
trainingData <-datanew[Index, ] # model training data  
  
testData <-datanew[-Index, ]

**Since the attribute ‘Income’ has factor we will convert those to binary using the code below.**

levels(datanew$Income)=c(0,1)

str(datanew$Income)

## Factor w/ 2 levels "0","1": 1 1 1 1 1 1 2 2 2 2 ...

The number of people who receive an income <50k and >=50k

table(datanew$Income)

##   
## 0 1   
## 22653 7508

**Logistic model**

**Logistic regression is a predictive modelling algorithm that is used when the Y variable is binary categorical. It is used to predict a class, i.e., a probability. Logistic regression can predict a binary outcome accurately. The output of the function is always between 0 and 1.**

logistic\_mod <-glm(Income ~., data=trainingData,family=binomial(link="logit")) # build the model

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

Income\_Pred <-predict(logistic\_mod, testData) # predict distance  
## prediction from a rank-deficient fit may be misleading

summary(logistic\_mod)

##   
## Call:  
## glm(formula = Income ~ ., family = binomial(link = "logit"),   
## data = trainingData)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -5.0962 -0.5143 -0.1890 -0.0097 3.5409   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error z value  
## (Intercept) -6.233e+00 8.694e-01 -7.169  
## Age 2.558e-02 2.051e-03 12.476  
## WorkclassLocal-gov -6.204e-01 1.342e-01 -4.624  
## WorkclassSelf-emp-not-inc -9.847e-01 1.305e-01 -7.545  
## Education1st-4th -3.093e-01 5.649e-01 -0.548  
## Education5th-6th -4.535e-01 4.360e-01 -1.040  
## Education7th-8th -4.981e-01 2.940e-01 -1.694  
## Education9th -1.006e-01 3.092e-01 -0.325  
## Pr(>|z|)   
## (Intercept) 7.57e-13 \*\*\*  
## Age < 2e-16 \*\*\*  
## WorkclassLocal-gov 3.76e-06 \*\*\*  
## WorkclassPrivate 3.65e-05 \*\*\*  
## WorkclassSelf-emp-inc 0.008781 \*\*   
## WorkclassSelf-emp-not-inc 4.53e-14 \*\*\*  
## WorkclassState-gov 1.39e-07 \*\*\*  
## Education7th-8th 0.090226 .   
## Education9th 0.744846   
## EducationAssoc-acdm 1.86e-10 \*\*\*  
## EducationAssoc-voc 3.76e-11 \*\*\*  
## EducationBachelors < 2e-16 \*\*\*  
## EducationDoctorate < 2e-16 \*\*\*  
## EducationHS-grad 2.53e-05 \*\*\*  
## EducationMasters < 2e-16 \*\*\*  
## EducationPreschool 0.950920   
## EducationProf-school < 2e-16 \*\*\*  
## EducationSome-college 4.43e-09 \*\*\*  
## `Education-num` NA   
## `Marital-status`Married-AF-spouse 3.13e-05 \*\*\*  
## `Marital-status`Married-civ-spouse 9.10e-07 \*\*\*  
## `Marital-status`Married-spouse-absent 0.635476   
## `Marital-status`Never-married 1.42e-05 \*\*\*  
## OccupationCraft-repair 0.955209   
## OccupationExec-managerial < 2e-16 \*\*\*  
## OccupationFarming-fishing 1.29e-08 \*\*\*  
## OccupationHandlers-cleaners 2.76e-06 \*\*\*  
## OccupationMachine-op-inspct 0.068611 .   
## OccupationOther-service 3.83e-09 \*\*\*  
## OccupationPriv-house-serv 0.091787 .   
## OccupationProf-specialty 1.61e-06 \*\*\*  
## OccupationProtective-serv 0.000449 \*\*\*  
## OccupationSales 0.011123 \*   
## OccupationTech-support 4.85e-06 \*\*\*  
## OccupationTransport-moving 0.395231   
## RelationshipNot-in-family 0.747602   
## RelationshipOther-relative 0.020982 \*   
## RelationshipOwn-child 0.002340 \*\*   
## RelationshipUnmarried 0.950779   
## RelationshipWife < 2e-16 \*\*\*  
## RaceAsian-Pac-Islander 0.009689 \*\*   
## RaceBlack 0.042489 \*   
## RaceOther 0.331025   
## RaceWhite 0.014047 \*   
## SexMale < 2e-16 \*\*\*  
## `Capital-gain` < 2e-16 \*\*\*  
## `Capital-loss` < 2e-16 \*\*\*  
## `Hours-per-week` < 2e-16 \*\*\*  
## `Native-country`Canada 0.478659   
## `Native-country`China 0.005781 \*\*   
## `Native-country`Columbia 0.011382 \*   
## `Native-country`Cuba 0.310778   
## `Native-country`Dominican-Republic 0.043067 \*   
## `Native-country`Ecuador 0.322146   
## `Native-country`El-Salvador 0.054665 .   
## `Native-country`Germany 0.408024   
## `Native-country`Greece 0.008621 \*\*   
## `Native-country`Guatemala 0.079936 .   
## `Native-country`Haiti 0.318761   
## `Native-country`Hungary 0.732834   
## `Native-country`India 0.004232 \*\*   
## `Native-country`Iran 0.125918   
## `Native-country`Laos 0.098206 .   
## `Native-country`Mexico 0.051498 .   
## `Native-country`Nicaragua 0.090445 .   
## `Native-country`Puerto-Rico 0.054224 .   
## `Native-country`Scotland 0.283230   
## `Native-country`South 0.001145 \*\*   
## `Native-country`Vietnam 0.005168 \*\*   
## `Native-country`Yugoslavia 0.722668   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 23571 on 21111 degrees of freedom  
## Residual deviance: 13544 on 21017 degrees of freedom  
## AIC: 13734  
##   
## Number of Fisher Scoring iterations: 14

predict <-predict(logistic\_mod, testData, type ='response')  
## prediction from a rank-deficient fit may be misleading

**Inference: The difference between Null deviance and Residual deviance tells us that the model is a good fit. Null deviance is the value when you only have intercept in your equation with no variables and Residual deviance is the value when you are taking all the variables into account. AIC is a counterpart of adjusted r square in multiple regressions. It’s an important indicator of model fit.The smaller the AIC is better.**

# confusion matrix  
table\_mat <-table(testData$Income, predict >0.5)  
table\_mat

##   
## FALSE TRUE  
## <=50K 6249 492  
## >50K 933 1375

accuracy\_Test <-sum(diag(table\_mat)) /sum(table\_mat)  
accuracy\_Test

## [1] 0.842524

**Inference:The confusion matrixis a better choice to evaluate the classification performance compared with the different metrics. The general idea is to count the number of times True instances are classified are False. The logistic model is showing 84.25 % accuracy, that is, that much percentage it is close to the correct value.**

**Decision tree**

**A decision tree is a diagram or chart that people use to determine a course of action or show a statistical probability. A decisiontree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility.An organization may deploy decisiontreesas a kind of decision support system.It’s a powerful and popular predictive machine learning technique that is used for both classification and regression. So it is also called classification and regression tree (CART).**

library(rpart)  
tree2 <-rpart(Income ~., data = trainingData, method ='class', cp =1e-3)  
tree2.pred.prob <-predict(tree2, newdata = testData, type ='prob')  
tree2.pred <-predict(tree2, newdata = testData, type ='class')  
# confusion matrix   
tb2 <-table(tree2.pred, testData$Income)  
tb2

##   
## tree2.pred <=50K >50K  
## <=50K 6397 1015  
## >50K 344 1293

#table\_mat <- table(testData$Income, predict > 0.5)  
#table\_mat  
accuracy\_Test2 <-sum(diag(tb2)) /sum(tb2)  
accuracy\_Test2

## [1] 0.8498177

**Inference: Accuracy is the value which shows how much the predicted value close to actual value. Decision tree has shown an accuracy of 84.98%.**

**Support Vector Machine**

**support-vector machines (SVM) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. It’s a discriminative classifier that constructs a hyperplane in a high dimensional space used for classification**.

library(kernlab)  
svm4 <-ksvm(Income ~., data = trainingData)  
svm4.pred.prob <-predict(svm4, newdata = testData, type ='decision')  
svm4.pred <-predict(svm4, newdata = testData, type ='response')  
# confusion matrix   
tb4 <-table(svm4.pred, testData$Income)  
tb4

##   
## svm4.pred <=50K >50K  
## <=50K 6311 1040  
## >50K 430 1268

accuracy\_Test4 <-sum(diag(tb4)) /sum(tb4)  
accuracy\_Test4

## [1] 0.8375511

**Inference: The prediction result of SVM has an accuracy of 83.75%, and a misclassification rate of 16.25%.**

**Conclusion**

We have a summary of our analysis and exploration of the Adult Census Data to come up with meaningful, important and interesting attributes of the data. Our classifiers are extrapolating patterns from the data, and this shows promise to be successfully can predict income based on Census information. The main techniques we used here are logistic regression, support vector machine and decision tree. All these methods give different amount of accuracy like,

Logistic 84.25%; Decision tree 84.98%; SVM 83.75%.

All these methods show slight variation in their accuracy levels. Since Decision tree is showing the maximum accuracy over here it is considered as the best analysis method for predicting the income using adult census dataset.

**References**

|  |  |
| --- | --- |
| [1] | S. S. S. Vidya Chockalingam, "Income Classification using Adult Census Data". |
| [2] | A. A. Antonov, "Classification and association rules for census income data," march 2014. |
| [3] | A. Agarwal, "medium," Towards Data Science. [Online]. |
| [4] | S. Jalalian, "Adult​ Census Income Analysis and Prediction," 28 August 2019. |
| [5] | H. Zhu, "Predicting Earning Potential using the Adult Dataset," 5 December 2016. |
| [6] | "https://www.guru99.com/r-generalized-linear-model.html#8," [Online]. |