**Who earned more than 50K in 1994?**

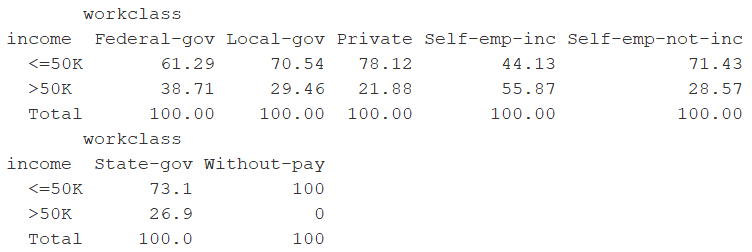
**Question of Interest**

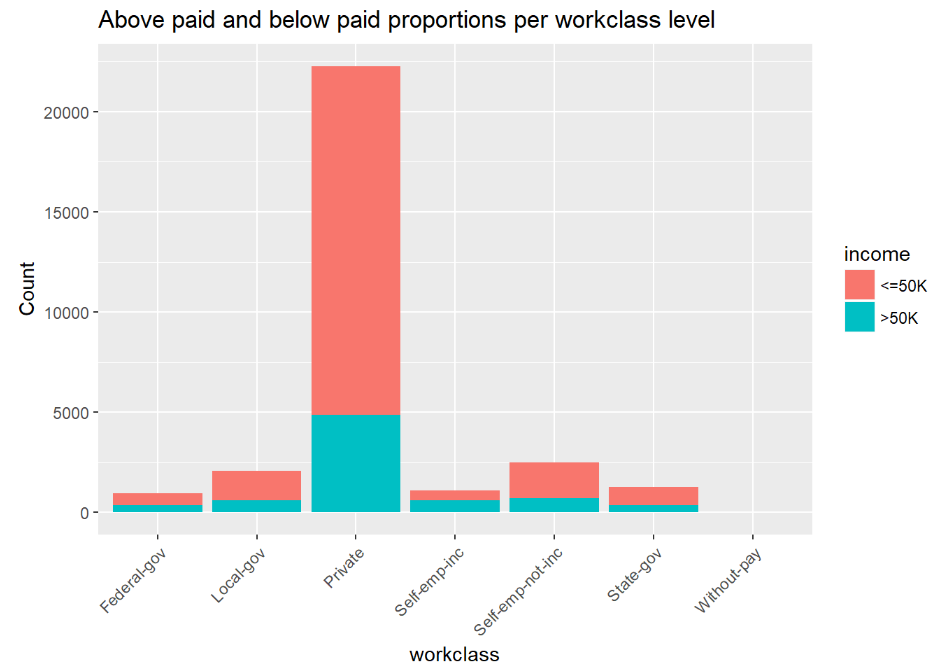
What factors or their weighted combinations can predict whether a person made over $50K a year in 1994 based on the given data.

**Exploratory Data Analysis (EDA)**

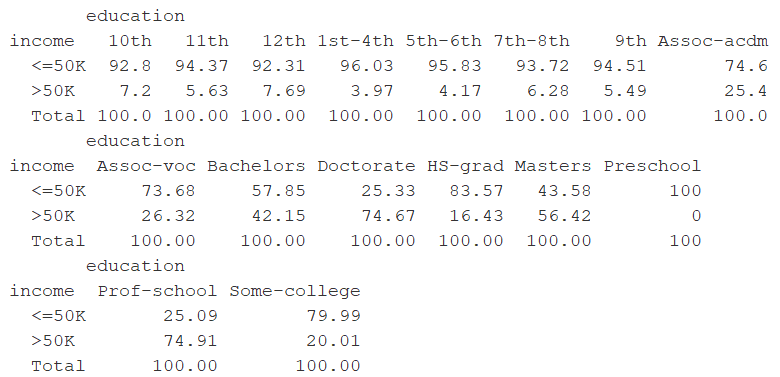
To better understand the factors that were most influential in resultant income of greater than $50K in 1994, we first explored the data to investigate the distributions of individual variables with respect to income category and in total. We further analyzed the relationship of those individual variables with the income category and the correlation between the variables themselves. The training and test sets are cleaned up first and the six new variables were created. Details of the EDA to assess the impact of explanatory variables on the likelihood of the high income are briefly summarized below.

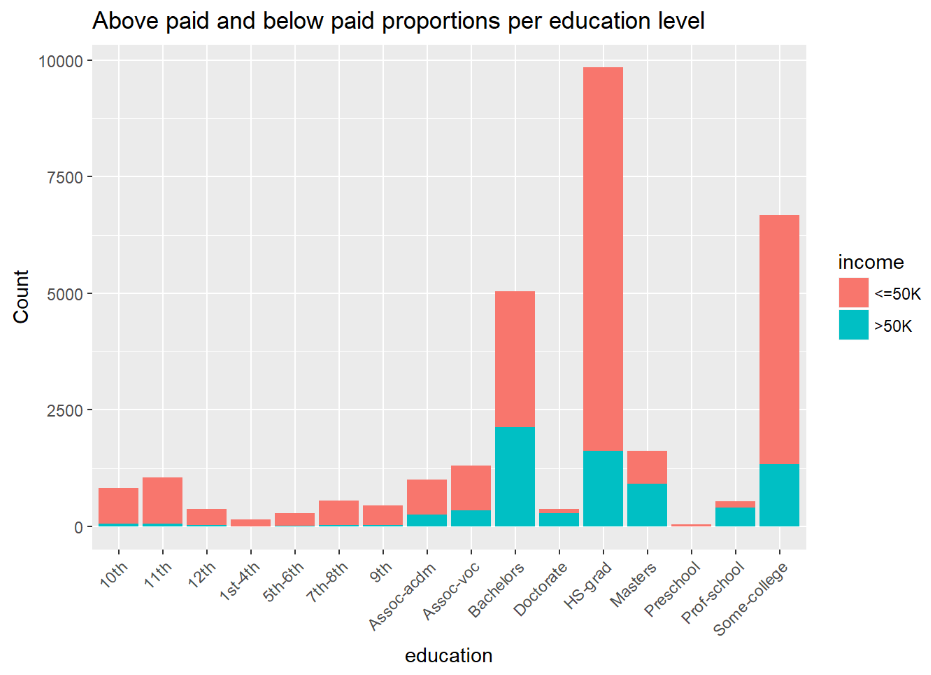
**Workclass:** The dataset comprises of 74% of the people who worked in Private sector. Incorporated self-employed workers had highest percentage (more than half, 56%) that earned more than 50K.



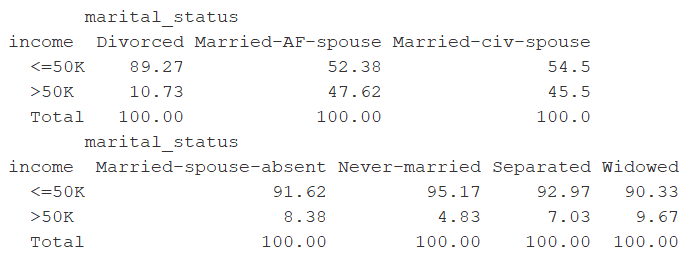


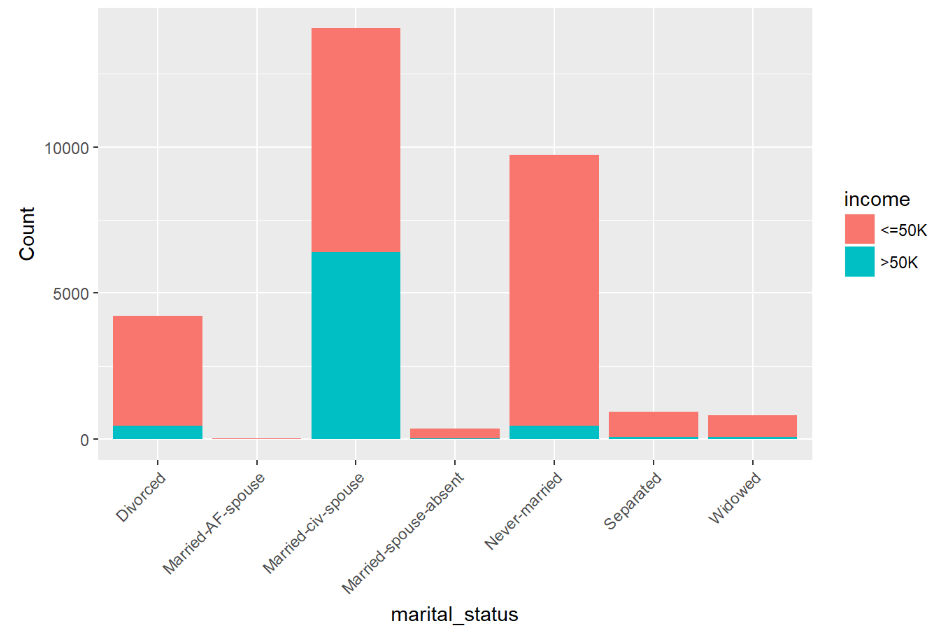
#### Education: Data is dominated by HS-grad (). As education level increases, percentage of people earning >50K increases. Higher education degree levels like Bachelor to PhD significantly contribute to higher pay grades.



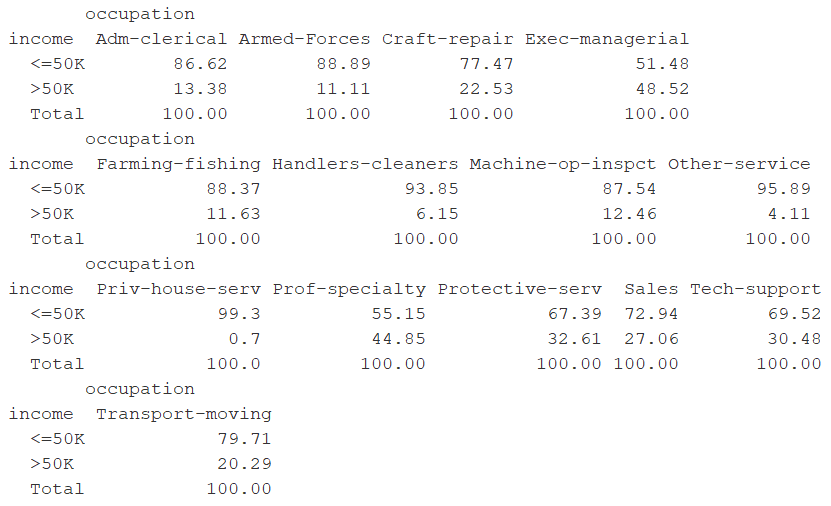


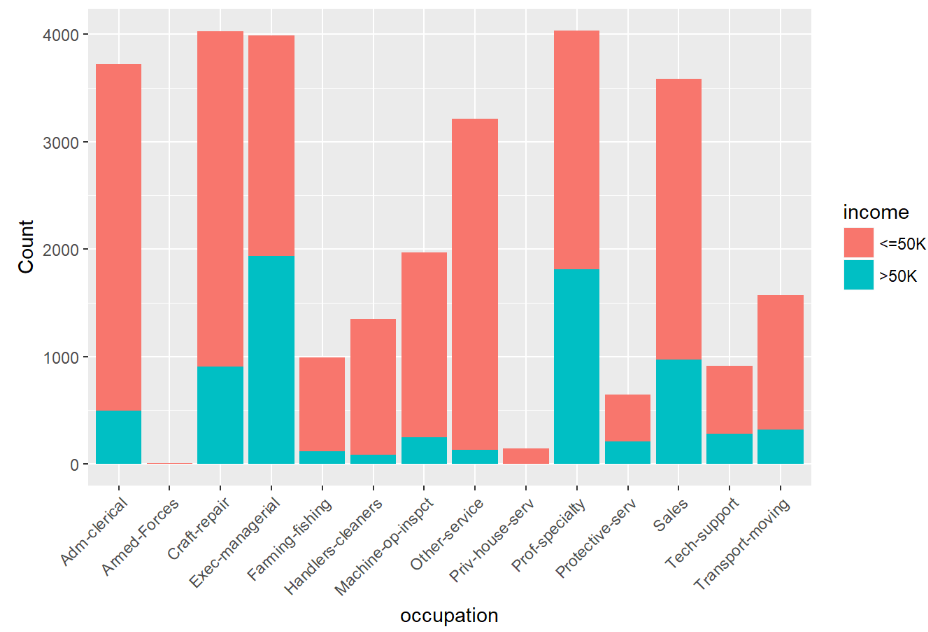
### **Marital status:** High percentage of >50K income earners are married civilians and living with spouse with good marital relationship.



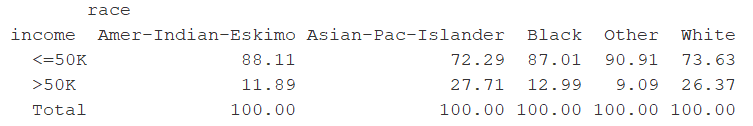


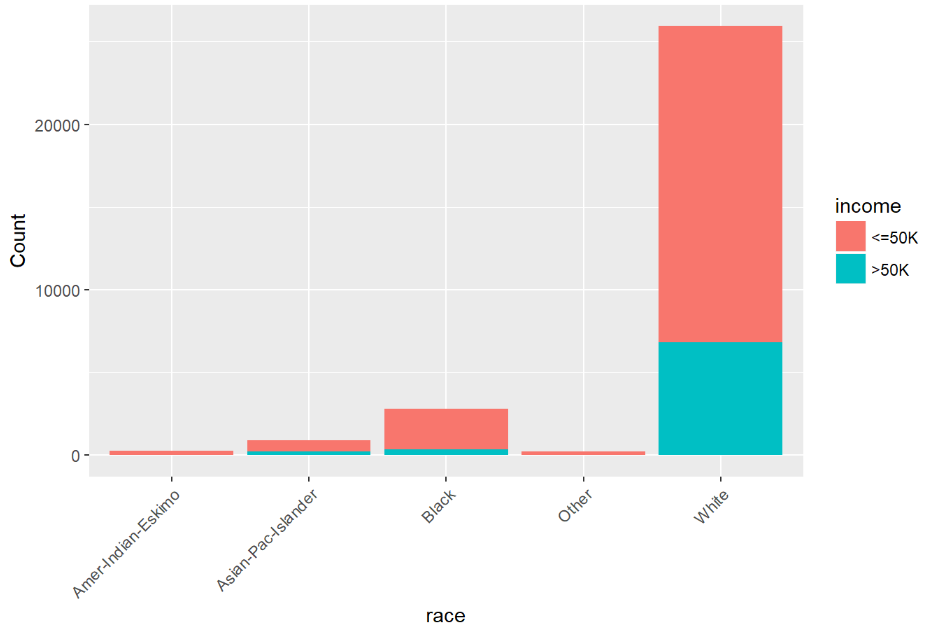
### **Occupation:** High percentage of >50K earner worked in executive managerial and professional specialty occupations**.**



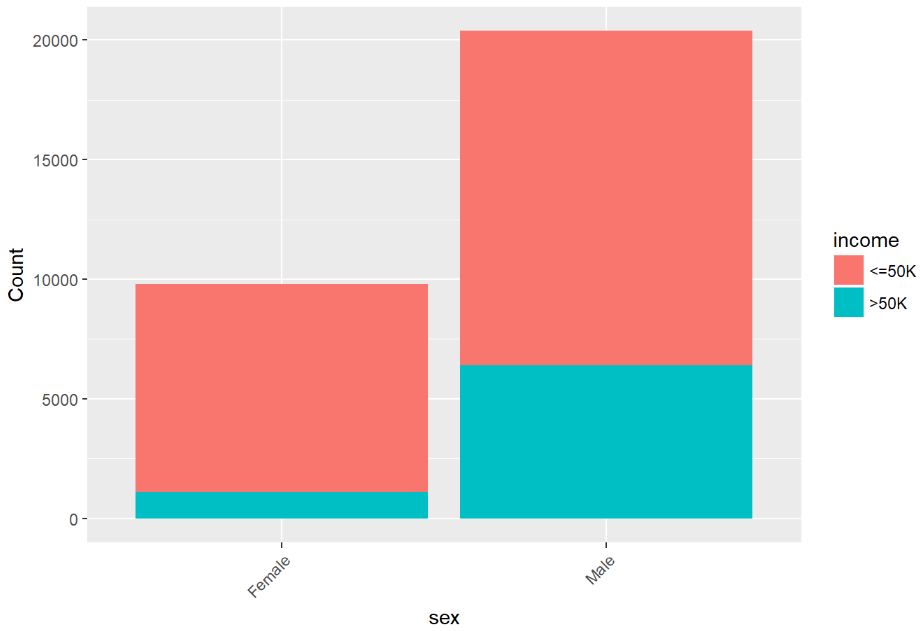
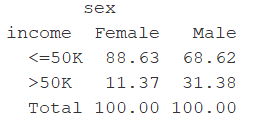


### **Race:** **Although** high percentage of high earners were Whites percentage of high earners was similar for Whites and Asian-Pac-Islanders.

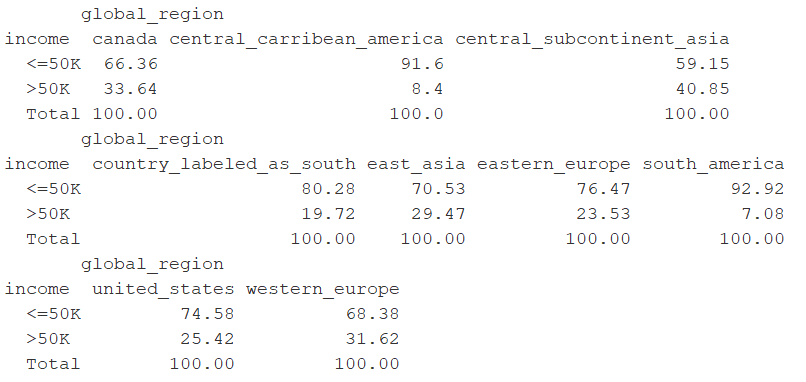


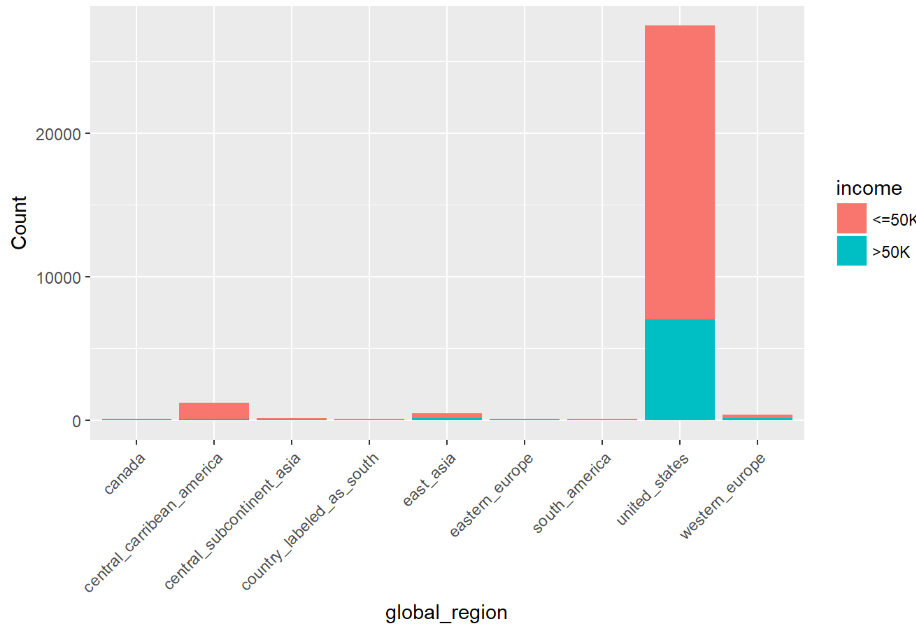


#### Sex: High paid workers were mostly Male.



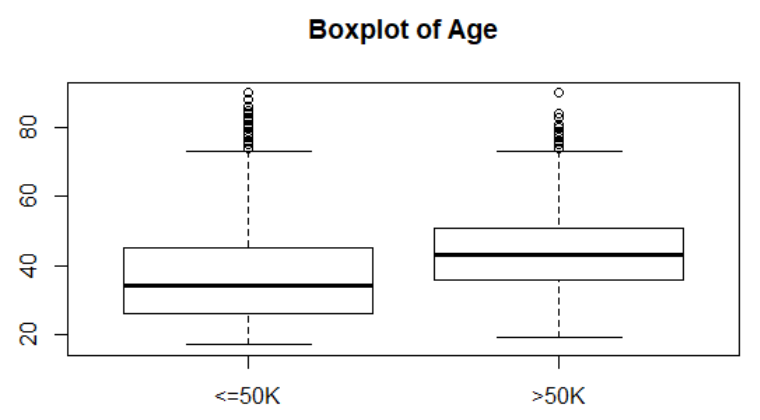
#### Global\_region: Although high percentage of high earners were from US percentage of high earners was higher for people from central\_subcontinent\_asia region.



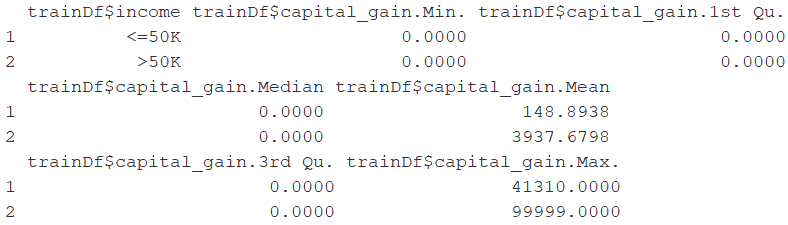


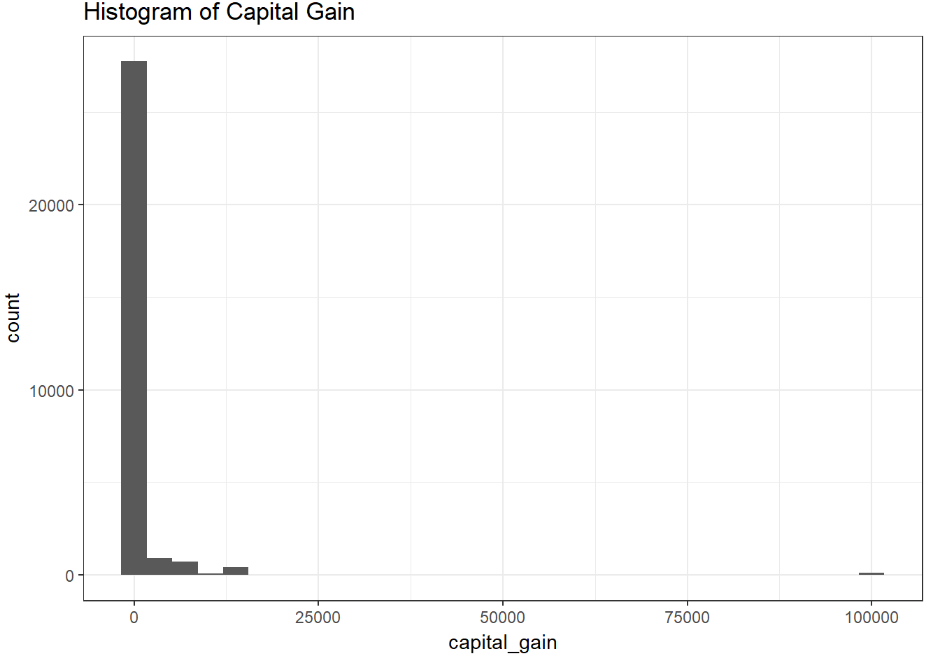
#### Summary stats for Continuous variables:

**Age:** Boxplot of Age shows people who earned >50k were between age of 36 to 51 with 43 as the median. Obviously, people between 17 to 24 were not earning high.

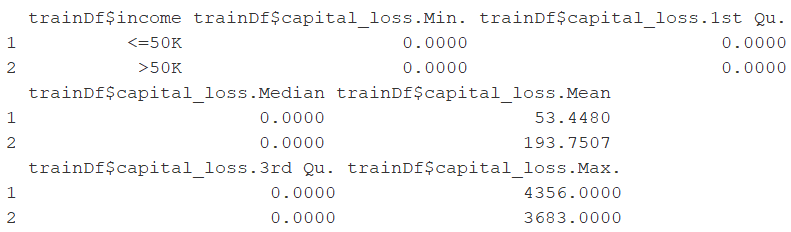


**Capital Gain:**





**Capital loss:**



**Predictive Model**

This dataset consists of majority of qualitative variables. Some of the qualitative variables are highly skewed. e.g. 86% of the data coming from “White” in “race” variable and 74 % of the data from “Private” sector “work class” variable.

We have different scales for different variables. To avoid few variables overweighing others in Principal Component Analysis (PCA) simply because of absolute scale we normalized all the variables. One of the limitation of PCA in this dataset is that the dependent variable levels are heavily skewed e.g. 75.1% observations from low income level and rest 24.9% from high income levels. So, we chose to use logistic regression to create a predictive model.

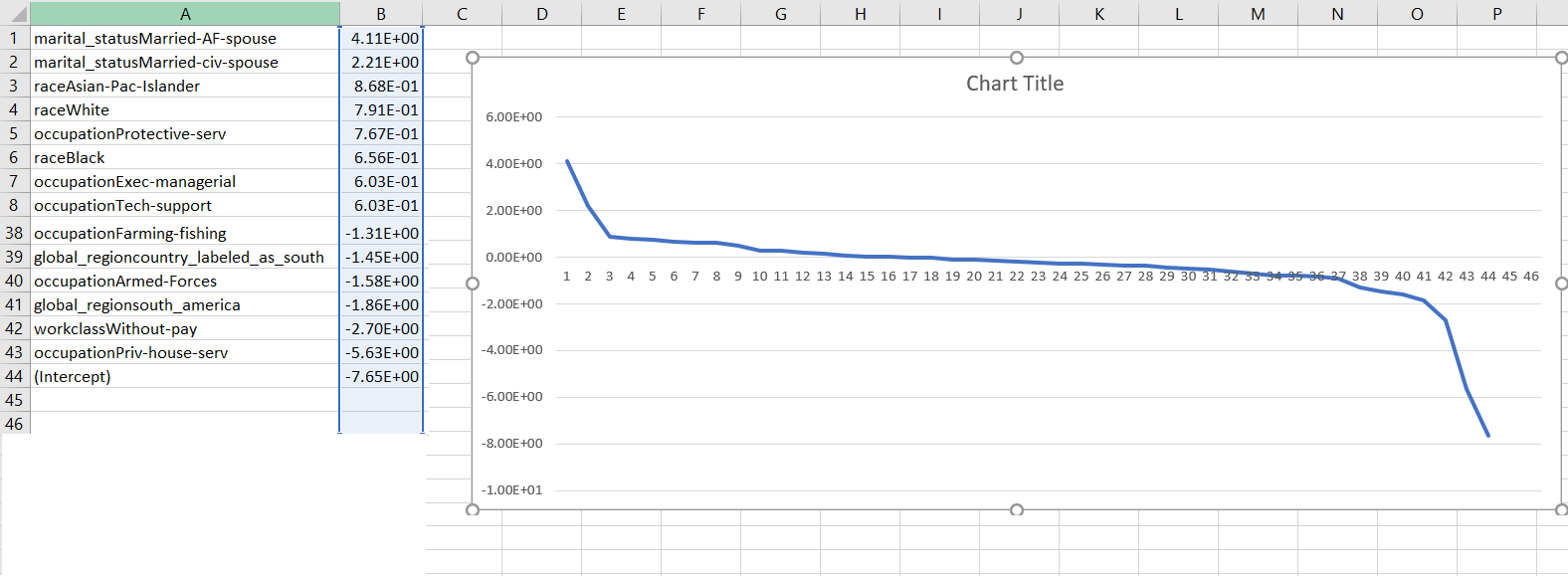
**Logistic regression:**

Since the response variable is binary, we are going to use logistic regression to predict who can earn >50k on test data set provided with this Adult census data. Based on EDA, we chose age, workclass, education\_num, marital\_status, occupation, race, sex, capital\_gain, capital\_loss, hours\_per\_week, global\_region as predictive explanatory variables in our logistic regression model. All the variables used in the model are statistically significant. Briefly writing the logistic regression equation below:

Ln(p/1-p) = y = -7.648 + (0.0341\*age) + (0.283\*education\_num) + ….

Both age (indirectly work experience) and education level increase the log odds of earning higher income in 1994.

The following image shows the coefficient values of the predictor variables and their plot(Copy pasted from R output to Excel sheet to order them and plot for the ease of identification and to get better idea of the factors significance using data visualization ) in significance order which are most important predictors of earning higher income.



### **Conclusion:** In short, higher income community comprised of people who are married civilians and living with spouse with good marital relationship (“marital\_status” catergory), people of “race” Asian-Pac-Islander or White and people who are working (“occupation” category) in Protective services or Exec-managerial or Tech-support. Lower income community comprised of people who are without pay (“workclass”), people who are in the “occupation” of Priv-house-serv, people from south\_america “region”.

**Goodness-of-fit test:** This logistic regression model is statistically significant. We also verified whether the first six predictions by our model matches with actual data.

Residual Deviance: 11579.03 and AIC: 11667.03

library(nnet)

library(ROCR)

mymodel<- multinom(income~., data = trainDf3)

#just to get ready-made Z-values and p-values of coeffients

#Residual Deviance: 11579.03 and AIC: 11667.03 are exactly same for both models using differnt R packages

model2<-glm(income~., data = trainDf3, family = "binomial")

summary(model2)

#Goodness of fit test: This model is statistically significant

with(model2, pchisq(null.deviance-deviance, df.null-df.residual, lower.tail = F ))

[1] 0

**Misclassification rate** is 0.179 and 0.19 on train and test datasets respectively.

|  |  |
| --- | --- |
|  |  |

##### **Accuracy** is 82.26% and 84.54% for train and test datasets respectively.

**ROC curve and AUC**

|  |  |
| --- | --- |
| **Train** | **Test** |

**Appendix: Code for all analyses**

**Appendix A: LogisticRegression.R**

The below commented code provides the steps taken in order to create, analyze, and select our models.

library(tigerstats)

library(sqldf)

library(ggplot2)

###Import data

trainDf <- read.csv("C:\\Manisha\\Stat II\\Project2\\train.csv", header = TRUE)

testDf <- read.csv("C:\\Manisha\\Stat II\\Project2\\test.csv", header = TRUE)

summary(trainDf)

##### EDA of Variable of Interest ####

# For building model, We decided not to use "education", "relationship" variables from the summary of the data.

# Instead of "native\_country", we are using "global\_region"

#Percentages of above and below paid Per level of categorical variables and plot for Catergorical variables

#Percentages of above and below paid Per workclass category

colPerc(xtabs(~income+workclass, trainDf))#Variable of Interest

workclass\_inc <-sqldf("select workclass, income, count(workclass) as Count from trainDf group by workclass, income")

# Plot of above paid and below paid counts per workclass level

ggplot(workclass\_inc, aes(x=workclass, y=Count, fill=income)) +

geom\_bar(stat="identity") +

theme(axis.text.x = element\_text(angle = 45, hjust = 1)) +

ggtitle('Above paid and below paid proportions per workclass level')

#Percentages of above and below paid Per education category

colPerc(xtabs(~income+education, trainDf))#Variable of Interest

education\_inc <-sqldf("select education, income, count(education) as Count from trainDf group by education, income")

# Plot of above paid and below paid counts per education level

ggplot(education\_inc, aes(x=education, y=Count, fill=income)) +

geom\_bar(stat="identity") +

theme(axis.text.x = element\_text(angle = 45, hjust = 1)) +

ggtitle('Above paid and below paid proportions per education level')

#Percentages of above and below paid Per marital\_status category

colPerc(xtabs(~income+marital\_status, trainDf))#Variable of Interest

maritalStatus\_inc <-sqldf("select marital\_status, income, count(marital\_status) as Count from trainDf group by marital\_status, income")

# Plot of above paid and below paid counts per education level

ggplot(maritalStatus\_inc, aes(x=marital\_status, y=Count, fill=income)) +

geom\_bar(stat="identity") +

theme(axis.text.x = element\_text(angle = 45, hjust = 1)) +

ggtitle('Above paid and below paid proportions per marital\_status level')

#Percentages of above and below paid Per occupation category

colPerc(xtabs(~income+occupation, trainDf))#Variable of Interest

occupation\_inc <-sqldf("select occupation, income, count(occupation) as Count from trainDf group by occupation, income")

# Plot of above paid and below paid counts per education level

ggplot(occupation\_inc, aes(x=occupation, y=Count, fill=income)) +

geom\_bar(stat="identity") +

theme(axis.text.x = element\_text(angle = 45, hjust = 1)) +

ggtitle('Above paid and below paid proportions per occupation level')

#Percentages of above and below paid Per race category

colPerc(xtabs(~income+race, trainDf))#Variable of Interest

race\_inc <-sqldf("select race, income, count(race) as Count from trainDf group by race, income")

# Plot of above paid and below paid counts per race level

ggplot(race\_inc, aes(x=race, y=Count, fill=income)) +

geom\_bar(stat="identity") +

theme(axis.text.x = element\_text(angle = 45, hjust = 1)) +

ggtitle('Above paid and below paid proportions per race level')

#Percentages of above and below paid Per sex category

colPerc(xtabs(~income+sex, trainDf))#Variable of Interest

sex\_inc <-sqldf("select sex, income, count(sex) as Count from trainDf group by sex, income")

# Plot of above paid and below paid counts per sex level

ggplot(sex\_inc, aes(x=sex, y=Count, fill=income)) +

geom\_bar(stat="identity") +

theme(axis.text.x = element\_text(angle = 45, hjust = 1)) +

ggtitle('Above paid and below paid proportions per sex level')

#Percentages of above and below paid Per global\_region category

colPerc(xtabs(~income+ global\_region, trainDf))#Variable of Interest

region\_inc <-sqldf("select global\_region, income, count(global\_region) as Count from trainDf group by global\_region, income")

# Plot of above paid and below paid counts per race level

ggplot(region\_inc, aes(x=global\_region, y=Count, fill=income)) +

geom\_bar(stat="identity") +

theme(axis.text.x = element\_text(angle = 45, hjust = 1)) +

ggtitle('Above paid and below paid proportions per global\_region level')

#Percentages of above and below paid Per hours\_per\_week\_category

colPerc(xtabs(~income+ hours\_per\_week\_category, trainDf))#Variable of Interest

hours\_inc <-sqldf("select hours\_per\_week\_category, income, count(hours\_per\_week\_category) as Count from trainDf group by hours\_per\_week\_category, income")

# Plot of above paid and below paid counts per hours\_per\_week\_category

ggplot(hours\_inc, aes(x=hours\_per\_week\_category, y=Count, fill=income)) +

geom\_bar(stat="identity") +

theme(axis.text.x = element\_text(angle = 45, hjust = 1)) +

ggtitle('Above paid and below paid proportions per Hours Per Week Category level')

# Finally summary of the dependent variable

ftable(trainDf$income)

#colPerc(summary(trainDf$income))

#### Summary stats by above and below paid groups for Continuous variables

aggregate(trainDf$capital\_gain~trainDf$income,data=trainDf,summary)

qplot(capital\_gain, data=trainDf, geom="histogram")+theme\_bw()+ggtitle('Histogram of Capital Gain')

aggregate(trainDf$capital\_loss~trainDf$income,data=trainDf,summary)

qplot(capital\_loss, data=trainDf, geom="histogram")+theme\_bw()+ggtitle('Histogram of Capital Loss')

#####Age: Boxplot of Age shows people who earn >50k are between age of 36 to 51 with 43 as the median

boxplot(age~income,trainDf,main="Boxplot of Age")

aggregate(trainDf$age~trainDf$income,data=trainDf,summary)

#Examine the correlation between the continous predictors

## The scatterplots and correlation matrix shows no significant correlation between any of the continuous variables

#pairs(trainDf[,c("age","education\_num","hours\_per\_week", "capital\_gain", "capital\_loss")])

my\_cor<-cor(trainDf[,c("age","education\_num","hours\_per\_week", "capital\_gain", "capital\_loss")])

my\_cor

############################################################

########After EDA code#####################################

#############################################################

trainTest <- rbind(trainDf, testDf)

# when I ran logistic regression code on a modified train dataset made up of equal obs for high income and

#low income(as per Dr. Turner's suggestion to get a good prediction model using this type of

#stratified sampling method), the sample workclass and occupation variables had atleast one level

#with Zero observations. Thus I am sampling

ftable(addmargins(table(trainTest$income,trainTest$occupation)))

ftable(addmargins(table(trainTest$income,trainTest$workclass)))

noPayobs<-trainTest[which(trainTest$workclass=="Without-pay"),]

#I decided to not include "relationship" factor variable since it seems redundant and for the sake of

#simplicity of the model.From the summary of this dataset, it seems that marital\_status and sex variables

# are capturing the information provided by "relationship" variable.

#To build the predictive model to predict who earn more than 50 K per annum,

#we decided to include age, workclass, education\_num, marital\_status, occupation,

#race, sex, capital\_gain, capital\_loss, hours\_per\_week, global\_region

#after reviewing plots and the summary stats for all the variables.

trainDf2 <-subset(trainDf, select=c("age", "workclass", "education\_num", "marital\_status", "occupation", "race", "sex", "capital\_gain", "capital\_loss", "hours\_per\_week", "global\_region", "income"))

testDf2 <-subset(testDf, select=c("age", "workclass", "education\_num", "marital\_status", "occupation", "race", "sex", "capital\_gain", "capital\_loss", "hours\_per\_week", "global\_region", "income"))

noPayobs <-subset(noPayobs, select=c("age", "workclass", "education\_num", "marital\_status", "occupation", "race", "sex", "capital\_gain", "capital\_loss", "hours\_per\_week", "global\_region", "income"))

#summary(trainDf2)

above50k <- trainDf2[trainDf2$income==">50K",]

below50k <- trainDf2[trainDf2$income=="<=50K",]

set.seed(111)

sampleBelow50 <- below50k[sample(nrow(below50k), size = 7508), ]

trainDf3 <-rbind(sampleBelow50, above50k, noPayobs)

#making sure that each category level of each categorical variable is represented in the sampled train data

xtabs(~income+workclass, data = trainDf3)

xtabs(~income+marital\_status, data = trainDf3)

xtabs(~income+occupation, data = trainDf3)

xtabs(~income+race, data = trainDf3)

xtabs(~income+sex, data = trainDf3)

xtabs(~income+global\_region, data = trainDf3)

#########################################################################

###############Logistic regression######################################

########################################################################

#install.packages("nnet")

library(nnet)

#install.packages("ROCR")

library(ROCR)

trainDf3$income<-ifelse(trainDf3$income=='>50K',1,0)

testDf2$income<-ifelse(testDf2$income=='>50K',1,0)

mymodel<- multinom(income~., data = trainDf3)

#just to get ready-made Z-values and p-values of coeffients

#Residual Deviance: 11579.03 and AIC: 11667.03 are exactly same for both models using differnt R packages

model2<-glm(income~., data = trainDf3, family = "binomial")

summary(model2)

#Goodness of fit test: This model is statistically significant

with(model2, pchisq(null.deviance-deviance, df.null-df.residual, lower.tail = F ))

predProb <- predict(mymodel, trainDf3, type = "prob")

#hist(predProb)

###### Confusion Matrix and Misclassification Rate of Train

confMatrix <- predict(mymodel, trainDf3)

tab <- table(confMatrix, trainDf3$income)

tab ### this prints confsion matrix

#### Misclassification Rate

1-sum(diag(tab))/sum(tab)

#see whether predProb(prediction Probalbilties) matches with actual data and it does for atleast first 6 obs

head(predProb)

head(trainDf3)

predProb <- prediction(predProb, trainDf3$income)

perfEval <- performance(predProb, "acc")

plot(perfEval)

maxYval <- which.max(slot(perfEval, "y.values")[[1]])

maxYval

acc <- slot(perfEval, "y.values")[[1]][maxYval]

acc ## thus the accuracy of the logistic regression on train data is 82.26%

###ROC

roc <- performance(predProb, "tpr", "fpr")

plot(roc)

abline(a= 0, b=1)

########### AUC(Area Under Curve) on Train ###################

auc <- performance(predProb, "auc")

auc<- unlist(slot(auc, "y.values"))

auc # area under the curve for train is .9062

###############################################################

###################ROC on test data

###########################################################

myTstmodel<- multinom(income~., data = trainDf3)

predProbTst <- predict(myTstmodel, testDf2, type = "prob")

#hist(predProbTst)

#see whether predProb(prediction Probalbilties) matches with actual data and it does for atleast first 6 obs

head(predProbTst)

head(testDf2)

###### Confusion Matrix and Misclassification Rate of test

confMatrixTst <- predict(myTstmodel, testDf2)

tabTst <- table(confMatrixTst, testDf2$income)

tabTst ### this prints confsion matrix

#### Misclassification Rate

1-sum(diag(tabTst))/sum(tabTst)

predProbTst <- prediction(predProbTst, testDf2$income)

perfEvalTst <- performance(predProbTst, "acc")

plot(perfEvalTst)

maxYvalTst <- which.max(slot(perfEvalTst, "y.values")[[1]])

maxYvalTst

accTst <- slot(perfEvalTst, "y.values")[[1]][maxYvalTst]

accTst ## thus the accuracy of the logistic regression on test data is 84.54%

# ROC test

rocTst <- performance(predProbTst, "tpr", "fpr")

plot(rocTst)

#abline(a= 0, b=1)

########### AUC(Area Under Curve) on Test data ###################

aucTst <- performance(predProbTst, "auc")

aucTst<- unlist(slot(aucTst, "y.values"))

aucTst # area under the curve for train is 0.90135

legend(.4,.4, aucTst, title= "AUC")