 **LOGISTIC REGRESSION**

**Submitted By: Team 02 Date Submitted:**

Amit Verma 100329477 11/27/2019

Sparsh Sharma 100329093 **Course:**

Simardeep Kaur 100329207 DANA-4820

Table of Contents

[Introduction 3](#_Toc25784726)

[Variables Description 3](#_Toc25784727)

[Variables Types 5](#_Toc25784728)

[Project Objective 5](#_Toc25784729)

[Two Sample t-Test (Quantitative Variable) 6](#_Toc25784730)

[Chi-Square Test 8](#_Toc25784731)

[Variable Selection 10](#_Toc25784732)

[Tree Algorithm 12](#_Toc25784733)

[Selecting final variables 13](#_Toc25784734)

[Likelihood Ratio Test 13](#_Toc25784735)

[Splitting data to train and test 15](#_Toc25784736)

[Multicollinearity 16](#_Toc25784737)

[Non-Interaction Model 17](#_Toc25784738)

[Interaction Model 18](#_Toc25784739)

[Hosmer-Lemeshow Test 19](#_Toc25784740)

[AIC for Interaction Model 19](#_Toc25784741)

[Likelihood Ratio Test 20](#_Toc25784742)

[Final Model 21](#_Toc25784743)

[Classification Table 21](#_Toc25784744)

[ROC Curve 22](#_Toc25784745)

[SAS Analysis 23](#_Toc25784746)

[Forward Selection 27](#_Toc25784747)

[Backward Elimination (Interaction Model) 36](#_Toc25784748)

[Final Model SAS 43](#_Toc25784749)

[Classification Table 44](#_Toc25784750)

[ROC Curve 44](#_Toc25784751)

[Conclusion 45](#_Toc25784752)

# Introduction

The US Adult income dataset was extracted by Barry Becker from the 1994 US Census Database.[[1]](#footnote-1) The data set consists of anonymous information such as occupation, age, native country, race, capital gain, capital loss, education, work class and more. This data was extracted from the census bureau database found at <https://archive.ics.uci.edu/ml/machine-learning-databases/adult/>

Number of observations: 32560

Number of variables: 14

The goal here is to predict the value column Income which has two possible values “>50K” and “<=50K”.

# Variables Description

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variables** | **Description** | **Quantitative/**  **Categorical** | **Scale of Measurement** | **Type of variable** | **No. of categories** |
| Age | Age of the individual | Quantitative | Interval/Ratio | Explanatory |  |
| Workclass | Working class an individual belongs to | Categorical | Nominal | Explanatory | 8 |
| fnlgwt | Represents final weight based on socio-economic characteristics of the  population | Quantitative | Interval/Ratio | Explanatory |  |
| Education | Highest education of the individual | Categorical | Ordinal | Explanatory | 16 |
| Education num | Each education level is assigned an education number. | Quantitative | Interval/Ratio | Explanatory |  |
| Marital Status | Individual’s marital status | Categorical | Nominal | Explanatory | 7 |
| Occupation | Profession of an individual | Categorical | Nominal | Explanatory | 13 |
| Relationship | Family relationship of an individual | Categorical | Nominal | Explanatory | 6 |
| Race | Ethnicity of an individual | Categorical | Nominal | Explanatory | 5 |
| Sex | Gender of an individual | Categorical | Nominal | Explanatory | 2 |
| Capital Gain | Total profit from an individual’s profession | Quantitative | Interval/Ratio | Explanatory |  |
| Capital Loss | Total loss from an individual’s profession | Quantitative | Interval/Ratio | Explanatory |  |
| Hours/Week | Total hours per week individual works in his/her profession | Quantitative | Interval/Ratio | Explanatory |  |
| Native country | Home country of the individual | Categorical | Nominal | Explanatory | 42 |
| Income | Net income earned by an individual per year | Quantitative | Interval/Ratio | Response (Dichotomous) |  |

## Variables Types

* **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, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc.), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holland-Netherlands.

# Project Objective

The objective of the project is to determine whether an individual earns more than 50K annually or not, considering various factors like age, his/her work class, hours per week the individual works, his/her native country, education etc. Logistic regression will be performed, and various tests will be implemented like 2-sample t-test, Chi-square test of independence, Hosmer-Lemeshow test, Likelihood Ratio test etc. to find out which explanatory variables are highly significant to the dichotomous response variable i.e. Income.

# Two Sample t-Test (Quantitative Variable)

The two-sample t test is performed on quantitative variables in order to check if the means are equal or not. If the means are unequal, then the variable is significant.

**Step 1:** Hypothesis:

Ho: Age is insignificant for prediction of income.

Ha: Age is significant for prediction of income.

**Step 2:** Perform variance test

**R Code**

#Age

var.test (adult\_data$Age[adult\_data$Income == " >50K"],adult\_data$Age[adult\_data$Income == " <=50K"])

**Output**

F test to compare two variances

data: adult\_data$Age[adult\_data$Income == " >50K"] and adult\_data$Age[adult\_data$Income == " <=50K"]

F = 0.58173, num df = 7507, denom df = 22653, p-value < 2.2e-16

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

0.5607432 0.6037101

sample estimates:

ratio of variances

0.5817293

The p-value is 2.2e-16 which is less than α(0.05). In conclusion, there is significant difference between the two variances.

**Step 3:** Perform t-test

**R-Code**

t.test(adult\_data$Age[adult\_data$Income == " >50K"],adult\_data$Age[adult\_data$Income == " <=50K"],var.equal = FALSE)

**Output**

Welch Two Sample t-test

data: adult\_data$Age[adult\_data$Income == " >50K"] and adult\_data$Age[adult\_data$Income == " <=50K"]

t = 49.505, df = 16701, p-value < 2.2e-16

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

7.05999 7.64211

sample estimates:

mean of x mean of y

43.95911 36.60806

The **p-value** of the test is 2.2e-16 which is less than α(0.05). We can then reject null hypothesis and conclude that age is significant variable in predicting the income.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Two Sample t-Test** | | | | |
|  | **Variance Test** | **Mean Test** | |  |
| **Variable** | **p-value** | **Mean of x** | **Mean of y** | **Hypothesis** |
| Age | 2.20E-16 | 43.95911 | 36.60806 | reject Ho |
| Capital Gain | 2.20E-16 | 3937.6798 | 148.8938 | reject Ho |
| Capital Loss | 2.20E-16 | 193.7507 | 53.4480 | reject Ho |
| Education Number | 2.20E-16 | 11.606420 | 9.629116 | reject Ho |
| Fnlgwt | 0.1132 | 188150 | 190338.6 | fail to reject Ho |
| Hours/Week | 2.20E-16 | 45.70658 | 36.60806 | reject Ho |

**Observation:**

After conducting two sample t test on all quantitative variables, we selected Age, Capital Gain, Capital Loss, Education Number, Hours/Week as they have p-value less than α(0.05).Also, the means of x and y are different. The variable fnlgwt is rejected as it has p-value greater than 0.05 making it insignificant.

# Chi-Square Test

The Chi-Square test is performed on categorical variables which tests the association of income with other categorical variables.

* **Marital Status**

**Step1:** Hypothesis

**Ho:**  Income is independent of Marital Status

**Ha:** Income is dependent on Marital Status

**Step 2:** Perform Chi-Square test

**R-Code:**

summary(adult\_data$Marital\_Status)

tbl\_MaritalStatus=table (adult\_data$Income, adult\_data$Marital\_Status)

tbl\_MaritalStatus

chisq.test(tbl\_MaritalStatus)

**Output:**

Single Married

<=50K 14638 8016

>50K 1068 6440

> chisq.test(tbl\_MaritalStatus)

Pearson's Chi-squared test with Yates' continuity correction

data: tbl\_MaritalStatus

X-squared = 5735.4, df = 1, p-value < 2.2e-16

Since p-value is less than α(0.05), we reject the null hypothesis. So, we conclude that Income is dependent on Marital Status.

* **Education Level**

**Step1:** Hypothesis

**Ho:**  Income is independent of Education Level

**Ha:** Income is dependent on Education Level

**Step 2:** Perform Chi-Square test

**R-Code:**

summary(adult\_data$education\_level) tbl\_educationlevel=table(adult\_data$Income,adult\_data$education\_level)

tbl\_educationlevel

chisq.test(tbl\_educationlevel)

**Output:**

Primary Intermediate Secondary Advanced

<=50K 3516 9938 8396 804

>50K 225 2217 3868 1198

> chisq.test(tbl\_educationlevel)

Pearson's Chi-squared test

data: tbl\_educationlevel

X-squared = 2598.5, df = 3, p-value < 2.2e-16

* **Native Region**

**Step1:** Hypothesis

**Ho:**  Income is independent of Native Region

**Ha:** Income is dependent on Native Region

**Step 2:** Perform Chi-Square test

**R-Code:**

summary(adult\_data$native\_region)

tbl\_nativeregion=table(adult\_data$Income,adult\_data$native\_region)

tbl\_nativeregion

chisq.test(tbl\_nativeregion)

**Output:**

Outlying-US United-States Asia Central\_America Europe

<=50K 270 20509 303 1228 344

>50K 110 6995 143 111 149

> chisq.test(tbl\_nativeregion)

Pearson's Chi-squared test

data: tbl\_nativeregion

X-squared = 224.81, df = 4, p-value < 2.2e-16

|  |  |  |  |
| --- | --- | --- | --- |
| **Chi-square Test** | | | |
| **Variable** | **p-value** | **X-squared** | **Accept/Reject** |
| Workclass | 2.20E-16 | 804.16 | Reject Ho |
| Occupation Status | 2.20E-16 | 3120.7 | Reject Ho |
| Marital Status | 2.20E-16 | 5735.4 | Reject Ho |
| Education Level | 2.20E-16 | 2598.5 | Reject Ho |
| Native Region | 2.20E-16 | 224.81 | Reject Ho |
| Relationship | 2.20E-16 | 6233.8 | Reject Ho |
| Race | 2.20E-16 | 304.24 | Reject Ho |
| Sex | 2.20E-16 | 1415.3 | Reject Ho |

**Observation:**

After performing Chi-Square test on all the categorical variables, we can conclude that all the variables are significant as their p-value is less than α(0.05).

# Variable Selection

We have used stepwise variable selection method to determine which variables are significant in our dataset.

**R-Code:**

adult<-data.frame(adult\_data$Age, adult\_data$Workclass,adult\_data$fnlgwt,adult\_data$education\_level,adult\_data$`Education num`,adult\_data$Marital\_Status,adult\_data$occupation\_status,adult\_data$Relationship,adult\_data$Race,adult\_data$Sex,adult\_data$`Capital Gain`,adult\_data$`Capital Loss`,adult\_data$`Hours/Week`,adult\_data$native\_region,adult\_data$Income)

adult\_stepwise<- glm (adult$adult\_data.Income ~., family = binomial, data = adult) %>%

stepAIC(trace = FALSE)

**Output:**

Call:

glm(formula = adult$adult\_data.Income ~ adult\_data.Age + adult\_data.Workclass +

adult\_data.fnlgwt + adult\_data.education\_level + adult\_data..Education.num. +

adult\_data.Marital\_Status + adult\_data.occupation\_status +

adult\_data.Relationship + adult\_data.Race + adult\_data.Sex +

adult\_data..Capital.Gain. + adult\_data..Capital.Loss. + adult\_data..Hours.Week. +

adult\_data.native\_region, family = binomial, data = adult)

Deviance Residuals:

Min 1Q Median 3Q Max

-5.2145 -0.5290 -0.2054 -0.0221 3.6491

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -8.196e+00 3.936e-01 -20.824 < 2e-16 \*\*\*

adult\_data.Age 2.813e-02 1.621e-03 17.355 < 2e-16 \*\*\*

adult\_data.Workclass Local-gov -7.259e-01 1.099e-01 -6.603 4.02e-11 \*\*\*

adult\_data.Workclass Private -4.748e-01 9.141e-02 -5.194 2.06e-07 \*\*\*

adult\_data.Workclass Self-emp-inc -2.449e-01 1.203e-01 -2.036 0.04178 \*

adult\_data.Workclass Self-emp-not-inc -1.036e+00 1.062e-01 -9.757 < 2e-16 \*\*\*

adult\_data.Workclass State-gov -8.330e-01 1.235e-01 -6.748 1.50e-11 \*\*\*

adult\_data.Workclass Without-pay -1.259e+01 1.202e+02 -0.105 0.91655

adult\_data.fnlgwt 7.936e-07 1.735e-07 4.574 4.79e-06 \*\*\*

adult\_data.education\_levelIntermediate 8.835e-02 9.645e-02 0.916 0.35965

adult\_data.education\_levelSecondary 2.405e-01 1.126e-01 2.135 0.03273 \*

adult\_data.education\_levelAdvanced 2.907e-01 1.488e-01 1.954 0.05075 .

adult\_data..Education.num. 2.557e-01 1.380e-02 18.539 < 2e-16 \*\*\*

adult\_data.Marital\_StatusMarried 9.432e-01 1.670e-01 5.648 1.63e-08 \*\*\*

adult\_data.occupation\_statusSecondaryOcc 2.814e-01 4.898e-02 5.746 9.15e-09 \*\*\*

adult\_data.occupation\_statusTertiaryOcc 8.184e-01 5.196e-02 15.752 < 2e-16 \*\*\*

adult\_data.occupation\_statusOthers -6.553e-01 1.074e-01 -6.099 1.07e-09 \*\*\*

adult\_data.Relationship Not-in-family -9.899e-01 1.694e-01 -5.843 5.13e-09 \*\*\*

adult\_data.Relationship Other-relative -1.356e+00 2.306e-01 -5.880 4.10e-09 \*\*\*

adult\_data.Relationship Own-child -2.147e+00 2.097e-01 -10.242 < 2e-16 \*\*\*

adult\_data.Relationship Unmarried -9.585e-01 1.856e-01 -5.166 2.40e-07 \*\*\*

adult\_data.Relationship Wife 1.276e+00 1.036e-01 12.326 < 2e-16 \*\*\*

adult\_data.Race Asian-Pac-Islander 7.036e-01 2.712e-01 2.595 0.00946 \*\*

adult\_data.Race Black 3.976e-01 2.385e-01 1.667 0.09555 .

adult\_data.Race Other 2.497e-02 3.684e-01 0.068 0.94596

adult\_data.Race White 5.929e-01 2.276e-01 2.605 0.00918 \*\*

adult\_data.Sex Male 8.461e-01 7.844e-02 10.787 < 2e-16 \*\*\*

adult\_data..Capital.Gain. 3.149e-04 1.050e-05 29.991 < 2e-16 \*\*\*

adult\_data..Capital.Loss. 6.485e-04 3.817e-05 16.989 < 2e-16 \*\*\*

adult\_data..Hours.Week. 2.893e-02 1.644e-03 17.594 < 2e-16 \*\*\*

adult\_data.native\_region United-States 1.180e-01 1.826e-01 0.646 0.51807

adult\_data.native\_regionAsia -3.101e-01 2.121e-01 -1.462 0.14372

adult\_data.native\_regionCentral\_America -4.921e-01 2.203e-01 -2.234 0.02549 \*

adult\_data.native\_regionEurope 2.272e-01 2.218e-01 1.024 0.30573

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 33851 on 30161 degrees of freedom

Residual deviance: 19857 on 30128 degrees of freedom

AIC: 19925

Number of Fisher Scoring iterations: 12

**Observation:**

From the above output, we can see that the variables adult\_data.native\_regionEurope, adult\_data.native\_regionAsia, adult\_data.native\_region United-States, adult\_data.Race Other,adult\_data.education\_levelAdvanced,adult\_data.education\_levelIntermediate, adult\_data.Workclass Without-pay, are insignificant as their p-values are greater than α(0.05).

# Tree Algorithm

Tree Algorithm is used to build classification and regression models in the form of a tree structure. Tree algorithm identifies the most significant variables by ranking them in decreasing order.

**R-Code:**

tree <- rpart(adult$adult\_data.Income ~ .,data = adult, method = "class")

imp <- varImp(tree)

rownames(imp)[order(imp$Overall, decreasing=TRUE)]

**Output:**

[1] "adult\_data..Capital.Gain." "adult\_data..Education.num." "adult\_data.Relationship" "adult\_data.Marital\_Status"

[5] "adult\_data.occupation\_status" "adult\_data.education\_level" "adult\_data..Capital.Loss." "adult\_data.Age"

[9] "adult\_data..Hours.Week." "adult\_data.Workclass" "adult\_data.fnlgwt" "adult\_data.Race"

[13] "adult\_data.Sex" "adult\_data.native\_region"

**Observation:**

The above output shows that the most important variable in our dataset is Capital Gain

followed by Education Number, Relationship, Marital Status and so on.

# Selecting final variables

We performed the 2-sample t test and Chi-square test but found out that all the variables (except fnlgwt) are significant. So, to eliminate the variables in our model, we conducted a Tree Algorithm which determined the rank of variables that were the most important in our dataset. Out of the 14 variables, we selected top 7 ranked by the algorithm.

1. Capital Gain
2. Education Number
3. Relationship
4. Marital Status
5. Occupation Status
6. Education Level
7. Capital Loss

**Rejected Variables**

1. **Education Number:** This variable is rejected due to its redundency, as it represents the column “education\_level” in numeric form.
2. **Relationship:** This variable represents almost same information as given in “Marital\_status”. So, we rejected this variable.
3. **Education Level:** Occupation status is described by the level of education. So, occupation status is retained, and education level is removed.

**Selected Variables**

1. Capital Gain
2. Marital Status
3. Occupation Status
4. Capital Loss

# Likelihood Ratio Test

* **Likelihood Ratio Test for significant variable**

**Step 1:** Hypothesis

Ho : Null model is appropriate [log(Income) = α]

Ha : Full model is appropriate [log(Income) = α + (Capital Gain)]

**R-Code:**

adultdata<-read.csv("H:/Documents/Downloads/adultdata1.csv",header=TRUE)

null\_model <- glm(Income ~1 , family=binomial(link = "logit"), data=adultdata)

full\_model <- glm(Income ~Capital.Gain , family=binomial(link = "logit"), data=adultdata)

anova(null\_model,full\_model,test="Chisq")

#Calculating G\*2

null\_model$deviance-full\_model$deviance

#Calculating df

null\_model$df.residual-full\_model$df.residual

**Output:**

Analysis of Deviance Table

Model 1: Income ~ 1

Model 2: Income ~ Capital.Gain

Resid. Df Resid. Dev Df Deviance Pr(>Chi)

1 30161 33851

2 30160 30699 1 3151.9 < 2.2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> null\_model$deviance-full\_model$deviance

[1] 3151.909

> #Calculating df

> null\_model$df.residual-full\_model$df.residual

[1] 1

**Step 2:** Calculate

= Null deviance –Residual deviance

= 3151.909

Step 3: Calculate degrees of freedom

df for intercept only model = 1

df for our model = 2

So, df = 2-1

= 1

Step 4:

From chi-square table, we found that p-value < α(0.05).

Step 5: Conclusion

Since p-value is less than α, we can reject Ho. Thus, we have sufficient statistical evidence

to conclude that Capital gain affects income of the adult.

**Likelihood Ratio Test for insignificant variable**

**Step 1:** Hypothesis:

Ho : Null model is appropriate [log(Income) = α]

Ha : Full model is appropriate [log(Income) = α + (fnlgwt)]

**R-Code:**

null\_model1 <- glm(Income ~1 , family=binomial(link = "logit"), data=adult\_data)

full\_model1 <- glm(Income ~fnlgwt, family=binomial(link = "logit"), data=adult\_data)

anova(null\_model1,full\_model1,test="Chisq")

**Output:**

Model 1: Income ~ 1

Model 2: Income ~ fnlgwt

Resid. Df Resid. Dev Df Deviance Pr(>Chi)

1 30161 33851

2 30160 33848 1 2.4326 0.1188

**Step 2:** Calculate

= Null deviance –Residual deviance

= 33851-33848

= 3

**Step 3**: Calculate degrees of freedom

df for intercept only model = 1

df for our model = 2

**Step 4:**

From chi-square table, we found that p-value > α(0.05).

**Step 5:** Conclusion

Since p-value is greater than α, we fail to reject Ho. Thus, we have sufficient statistical evidence to conclude that fnlgwt does not affect income of the adult.

# Splitting data to train and test

After getting significant variables, we split the dataset into train and test in a ratio of 70:30.

**R-Code:**

trainIndex = sample(1:nrow(adultdata), size = round(0.7\*nrow(adultdata)), replace=FALSE)

adult\_train = adultdata[trainIndex ,]

adult\_test = adultdata[-trainIndex ,]

nrow(adult\_train)

nrow(adult\_test)

**Output:**

>nrow(adult\_train)

[1] 21113

>nrow(adult\_test)

[1] 9049

# Multicollinearity

Multicollinearity takes place when an explanatory variable is highly correlated with one or more explanatory variables. We performed VIF test to check if the variables have multicollinearity problem or not.

* *For Categorical variables:* First we are performing vif test on categorical variables as follows:

**R-Code:**

vif(glm(Income~occupation\_status+Marital\_Status,family=binomial(link = "logit"), data=adult\_train ))

**Output:**

GVIF Df GVIF^(1/(2\*Df))

occupation\_status 1.043744 3 1.007161

Marital\_Status 1.043744 1 1.021638

**Observation:**

After performing VIF test on categorical variables, we can see that the GVIF value of occupation status and Marital Status is around 1 which is smaller than 3.16. So, we conclude that occupation status and marital Status are not correlated with each other.

* *For Quantitative variables:* Now, we will check multicollinearity for quantitative variables.

**R-Code:**

vif(glm(Income~Capital.Gain+Capital.Loss,family=binomial(link = "logit"), data=adult\_train ))

**Output:**

**Output:**

Capital.Gain Capital.Loss

1.005459 1.005459

**Observation:**

After performing VIF test on quantitative variables, the VIF values for variables Capital Gain and Capital Loss are around 1 which is smaller than 10 indicating that both these variables are not correlated with each other and have no multicollinearity problem.

# Non-Interaction Model

**R-Code:**

fit<-glm(Income~occupation\_status+Marital\_Status+Capital.Gain+Capital.Loss,family = binomial(link='logit'),data = adult\_train)

summary(fit)

**Output:**

Deviance Residuals:

Min 1Q Median 3Q Max

-5.0444 -0.7183 -0.2798 -0.1370 3.0574

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -2.124e+00 1.170e-01 -18.153 < 2e-16 \*\*\*

occupation\_statusPrimaryOcc 9.013e-01 1.228e-01 7.339 2.16e-13 \*\*\*

occupation\_statusSecondaryOcc 1.444e+00 1.216e-01 11.875 < 2e-16 \*\*\*

occupation\_statusTertiaryOcc 2.567e+00 1.207e-01 21.269 < 2e-16 \*\*\*

Marital\_StatusSingle -2.540e+00 4.988e-02 -50.917 < 2e-16 \*\*\*

Capital.Gain 3.376e-04 1.197e-05 28.197 < 2e-16 \*\*\*

Capital.Loss 6.701e-04 4.190e-05 15.994 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 23698 on 21112 degrees of freedom

Residual deviance: 15589 on 21106 degrees of freedom

AIC: 15603

Number of Fisher Scoring iterations: 7

**Observation**

All the variables taken in non-interaction model are significant as p-values are less that α as shown above in the output.

# Interaction Model

Interaction model has a presence of interaction terms which indicate that the effect of one variable is dependent on another variable.

Below is the R-code of interaction model having an interaction of Marital Status and Capital Gain.

**R-Code:**

fit1<-glm(Income~occupation\_status+Marital\_Status+Capital.Gain+Capital.Loss+Marital\_Status:Capital.Gain,

family = binomial,data = adult\_train)

summary(fit1)

**Output:**

Deviance Residuals:

Min 1Q Median 3Q Max

-4.8388 -0.7245 -0.2734 -0.1346 3.0688

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -2.094e+00 1.167e-01 -17.940 < 2e-16 \*\*\*

occupation\_statusPrimaryOcc 8.902e-01 1.223e-01 7.277 3.4e-13 \*\*\*

occupation\_statusSecondaryOcc 1.432e+00 1.211e-01 11.820 < 2e-16 \*\*\*

occupation\_statusTertiaryOcc 2.550e+00 1.203e-01 21.186 < 2e-16 \*\*\*

Marital\_StatusSingle -2.606e+00 5.247e-02 -49.661 < 2e-16 \*\*\*

Capital.Gain 2.912e-04 1.467e-05 19.853 < 2e-16 \*\*\*

Capital.Loss 6.700e-04 4.210e-05 15.915 < 2e-16 \*\*\*

Marital\_StatusSingle:Capital.Gain 1.152e-04 2.498e-05 4.612 4.0e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 23698 on 21112 degrees of freedom

Residual deviance: 15567 on 21105 degrees of freedom

AIC: 15583

Number of Fisher Scoring iterations: 7

**Observation:**

All the variables in the interaction model have p-value less than α making them significant.

# Hosmer-Lemeshow Test

Hosmer-Lemeshow test is used for ungrouped data to determine the goodness of fit of the model. If the test generates a low p-value, then we can say that the model is a poor fit. We performed the test to determine whether the model is a good fit or not.

**R-Code:**

hosmer\_fit <- glm(Income~occupation\_status+Marital\_Status+Capital.Gain+Capital.Loss, family = binomial(link='logit'), data = adult\_train )

logitgof(adult\_train$Income, fitted(hosmer\_fit))

**Output:**

Hosmer and Lemeshow test (binary model)

data: adult\_train$Income, fitted(hosmer\_fit)

X-squared = 43.334, df = 7, p-value = 2.875e-07

**Observation:**

Since, the p-value is very low, we can conclude that the model is not a good fit.

# AIC for Interaction Model

We performed AIC test for interaction model.

**R-Code:**

Model1<-glm(Income~occupation\_status+Marital\_Status+Capital.Gain+Capital.Loss+

Marital\_Status:Capital.Gain,family = binomial, data = adult\_train)

fitLM <- stepAIC(Model1)

**Output:**

Start: AIC=15583.24

Income ~ occupation\_status + Marital\_Status + Capital.Gain +

Capital.Loss + +Marital\_Status:Capital.Gain

Df Deviance AIC

<none> 15567 15583

- Marital\_Status:Capital.Gain 1 15589 15603

- Capital.Loss 1 15826 15840

- occupation\_status 3 17084 17094

**Observation:**

After performing AIC test on our interaction model, we observed that the variable Capital Gain is removed from the model as its AIC value was less than 15583.24. The rest of the variables stay in the model as their AIC value is more than 15583.24 as shown in the above output.

# Likelihood Ratio Test

We performed likelihood ratio test for interaction model having interaction term and compared with non-interaction model.

**Step 1:** Hypothesis:

**Ho:** Reduced model is appropriate [logit(Income)= α + (occupation\_status)+ (Marital\_Status)+ (Capital.Gain)+ (Capital.Loss)]

**Ha:** Full model is appropriate [logit(Income)= α + (occupation\_status)+ (Marital\_Status)+ (Capital.Gain)+(Capital.Loss)+(Marital\_Status\*Capital.Gain)]

**R-Code:**

# Non Interaction Model

model\_non\_interaction <- glm(Income ~occupation\_status+Marital\_Status+Capital.Gain+Capital.Loss , family=binomial(link = "logit"), data=adult\_train)

# Interaction Model

Model\_interaction <- glm(Income~occupation\_status+Marital\_Status+Capital.Gain+Capital.Loss+

Marital\_Status:Capital.Gain, family=binomial(link = "logit"), data=adult\_train)

anova(model\_non\_interaction,model\_interaction,test="Chisq")

**Output:**

Analysis of Deviance Table

Model 1: Income ~ occupation\_status + Marital\_Status + Capital.Gain +

Capital.Loss

Model 2: Income ~ occupation\_status + Marital\_Status + Capital.Gain +

Capital.Loss + +Marital\_Status:Capital.Gain

Resid. Df Resid. Dev Df Deviance Pr(>Chi)

1 21106 15589

2 21105 15567 1 22.095 2.595e-06 \*\*\*

**Step 2:** Calculate

= Residual deviance(Reduced model) –Residual deviance(Full model)

= 22

**Step 3:** Calculate degrees of freedom

df for reduced model = 5

df for full model = 6

So, df = 2-1 = 1

**Step 4:**

From chi-square table, we found that p-value < α(0.05).

**Step 5:** Conclusion

Since p-value is less than α, we can reject Ho. Thus, we have sufficient statistical evidence to conclude that interaction model is significant.

# Final Model

To select our final model, first we have performed Hosmer Lemeshow test to check the goodness of fit for the model. After that we performed likelihood ratio test to check whether interaction model is significant or not. As a result, we noticed that interaction model is significant. So, we performed AIC on our interaction model to get the significant variables which gave our final model as below:

Logit(Income)= (-2.094) +(0.8902)\*occupation\_statusPrimaryOcc+

(1.432)\*occupation\_statusSecondaryOcc+

(2.55)\* occupation\_statusTertiaryOcc+

(-2.606)\* Marital\_StatusSingle + (0.0002912)\* Capital.Gain+

(0.00067)\* Capital.Loss+

(0.000152)\* Marital\_StatusSingle\*Capital.Gain

# Classification Table

We have generated classification table for our final model to compare the predicted number of successes to the actual number of successes and predicted number of failures to the actual number of failures.

**R-Code:**

head(predict(fit1,adult\_test, type = "response"))

predicted\_r<-as.numeric(predict(fit1,adult\_test, type = "response")>prop)

#Table for r model

xtabs(~adult\_test$income\_test + predicted\_r)

**Output:**

predicted\_r

adult\_test$income\_test 0 1

0 5468 1380

1 504 1697

**Observation:**

Based on the output the overall misclassification rate for the model is

1884/9049=.208

For actual income <=50K

Misclassification rate 1380/6848=.201

For actual income >=50K

Misclassification rate 504/2201=.2289

**Observation:**

# ROC Curve

ROC Curve is used to determine the sensitivity and specificity of the predictions for all possible cut off πo values. We generated the ROC Curve for our final model as below.

**R-Code:**

rocplot<-roc(Income~fitted(Model\_interaction),data=adult\_train)

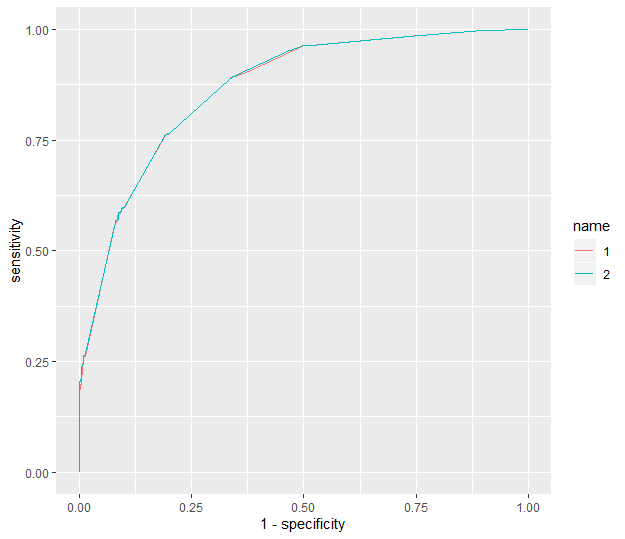
plot.roc(rocplot,legacy.axes = TRUE)

auc(rocplot)

**Output:**

auc(rocplot)

Area under the curve: 0.8669



**Observation:**

From the above output, we can see that the ROC Curve of our model is high and area under the curve is 0.8669 which is greater than 0.5. So, we conclude that the final model is the best fit having high curve and more area under the curve.

# SAS Analysis

In order to find a model which is a good fit, we have also done SAS analysis.

**SAS Code:**

Reading SAS File:

**Output:**

libname path 'H:\sastest';

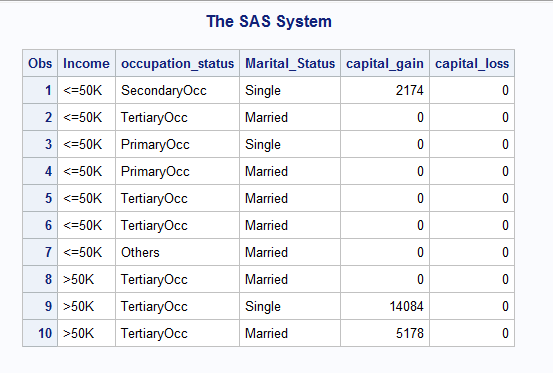
**proc** **import** datafile ='H:\sastest\adultdata\_final.xlsx'

out= path.adultdata

dbms=xlsx;

**proc** **print** data=path.adultdata(obs=**10**);

**run**;



**Logistic Model**

**SAS Code:**

**proc** **logistic** data = path.adultdata;

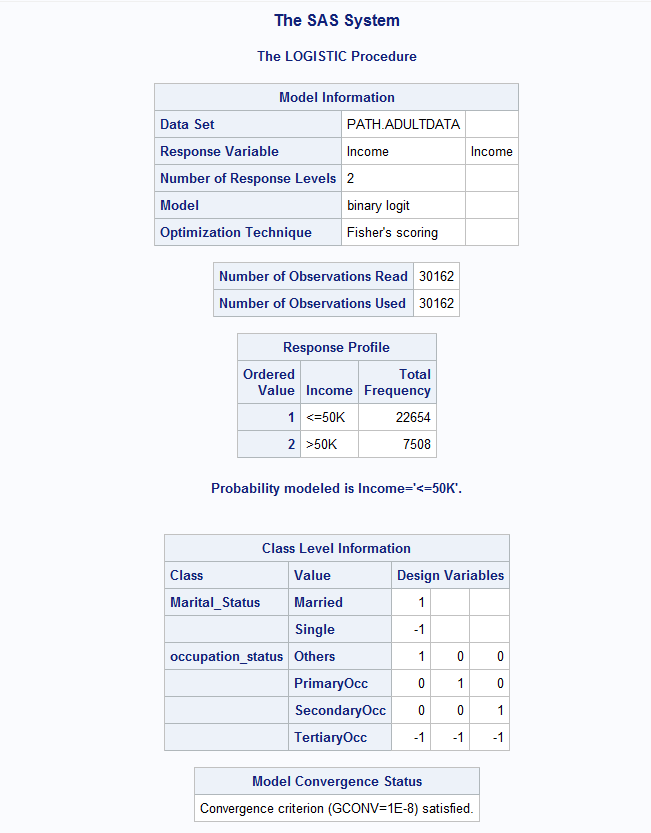
class Marital\_Status;

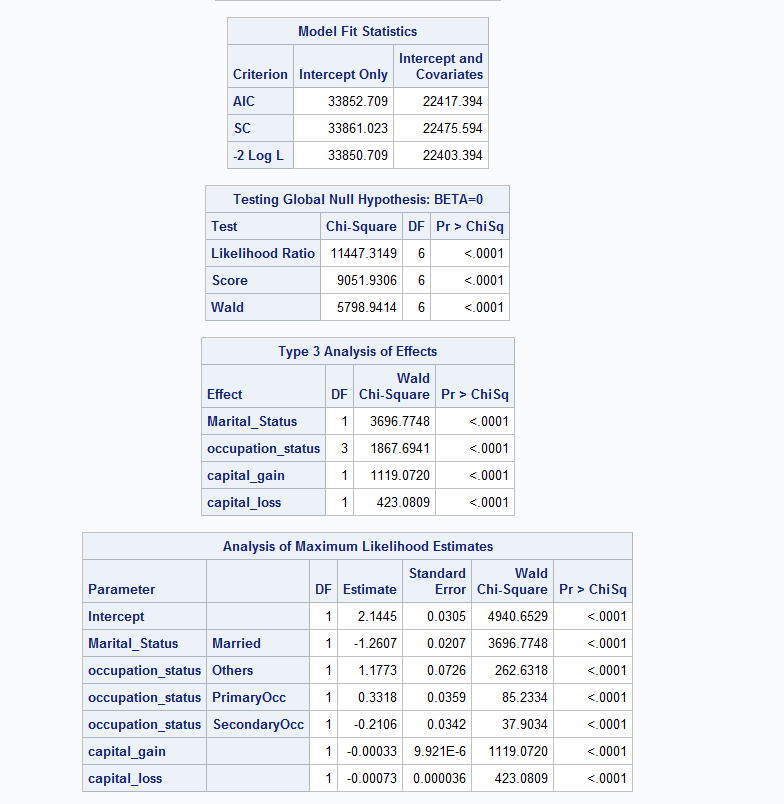
class occupation\_status;

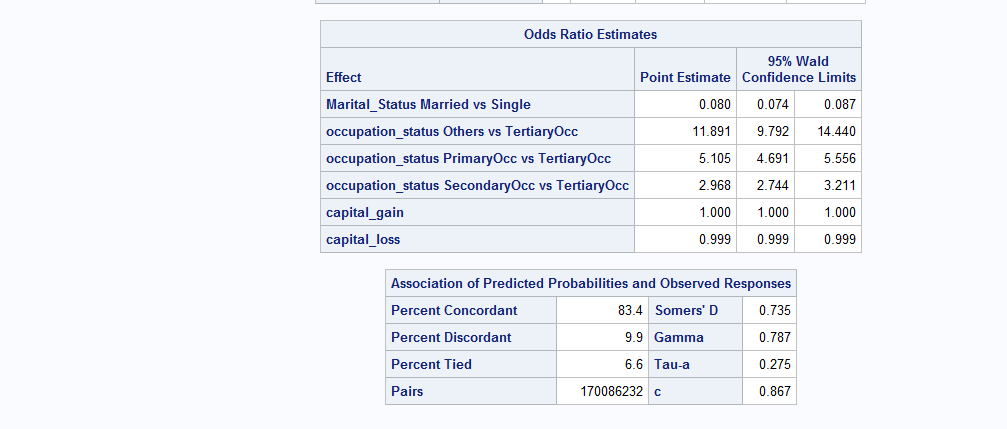
model Income = Marital\_Status occupation\_status capital\_gain capital\_loss;

**run**;

**Output:**







**Observation:**

After applying logistic regression on our model, we can see from the output that the p-value for all the variables is less than 0.05, so we can say that all the variables are significant.

## Forward Selection

Forward selection has been performed on non interaction model to select the significant variables taking α=0.05.

**SAS Code:**

**proc** **logistic** data = path.adultdata;

class Marital\_Status (param=ref ref=first);

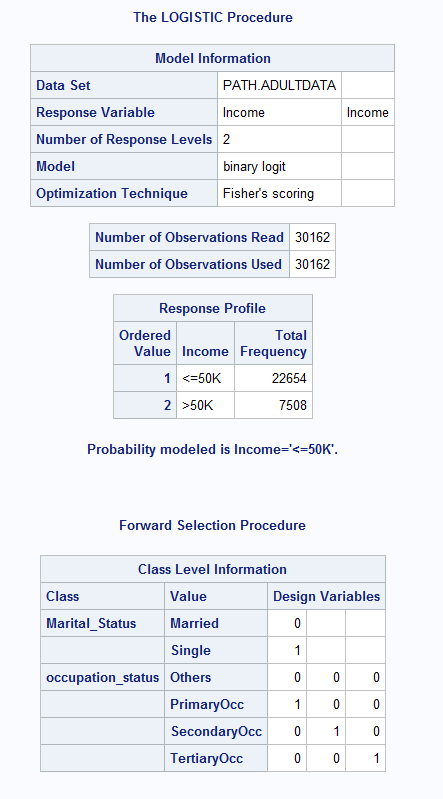
class occupation\_status (param=ref ref=first);

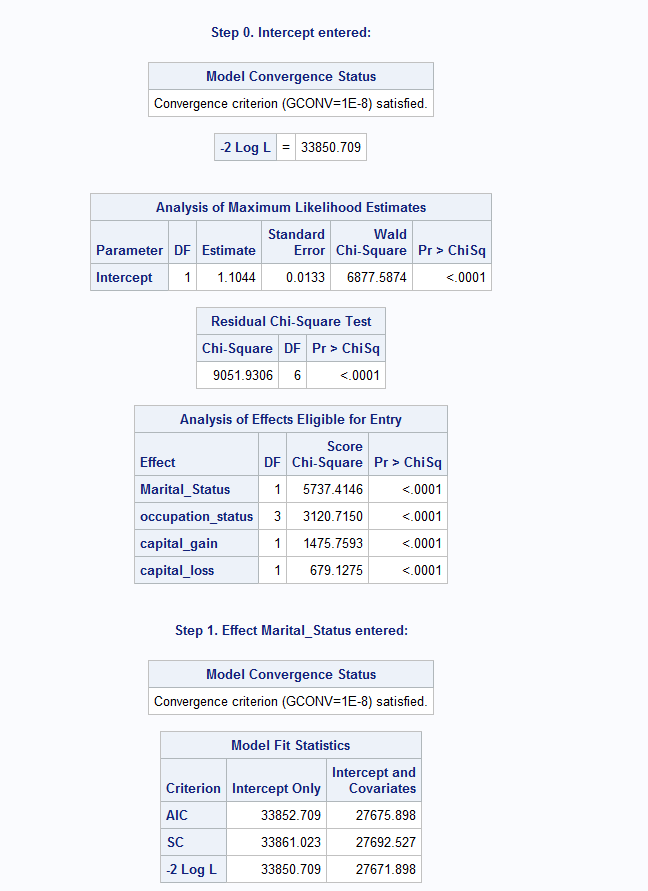
model Income = Marital\_Status occupation\_status capital\_gain capital\_loss

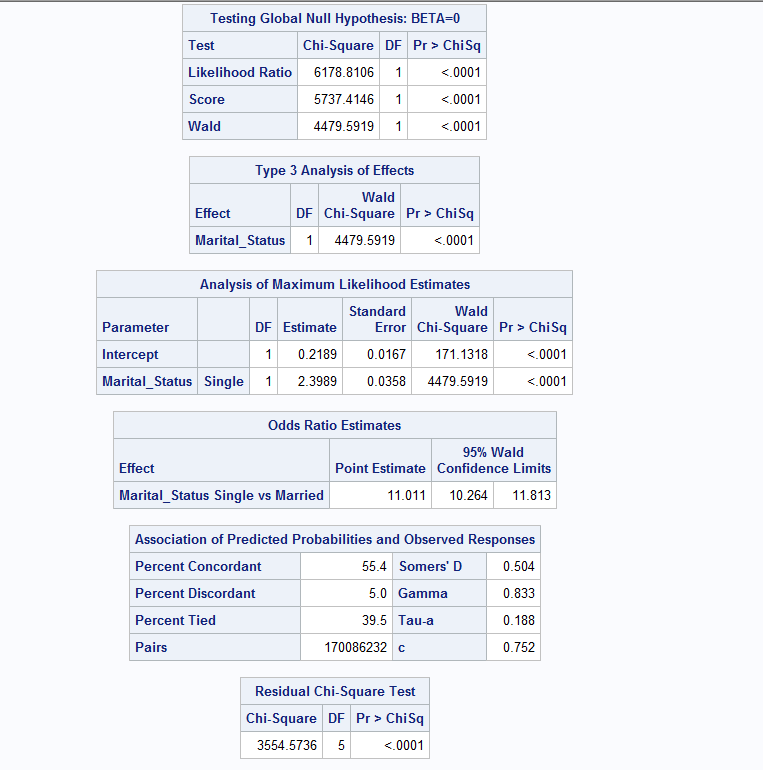
/selection=forward slentry=**0.05** hierarchy=single details;

**run**;

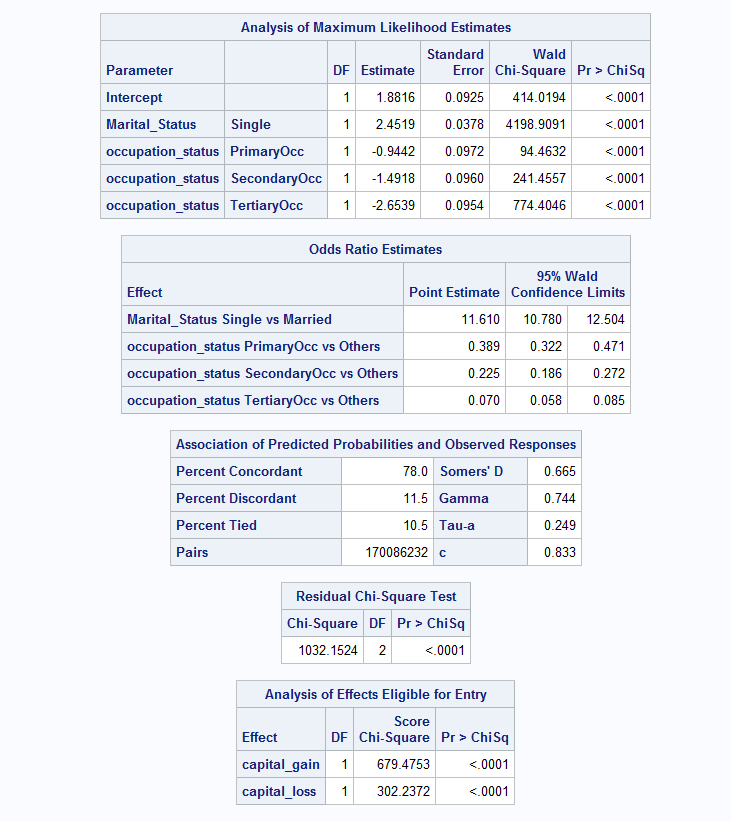
**Output:**

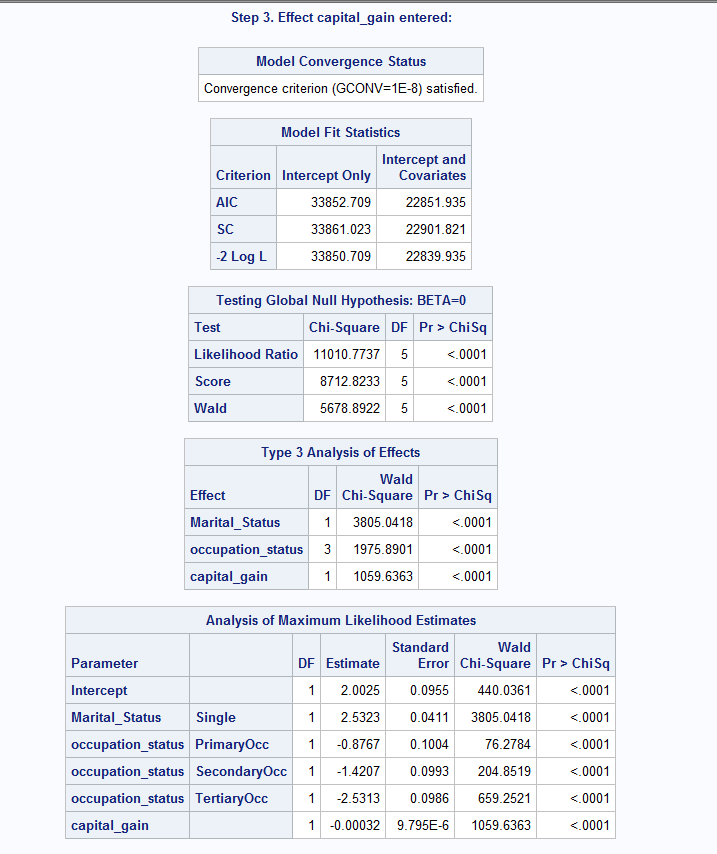


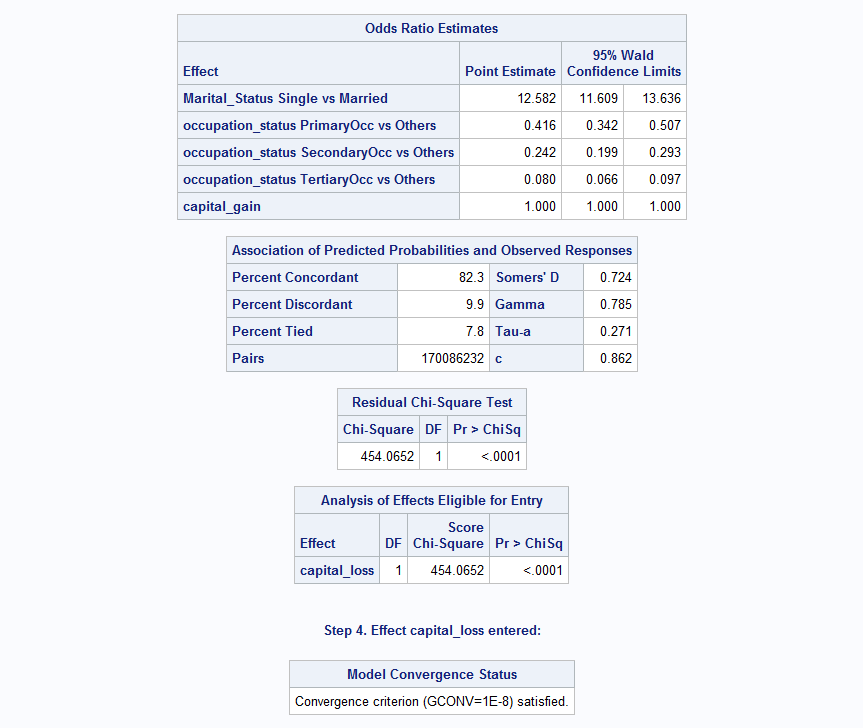




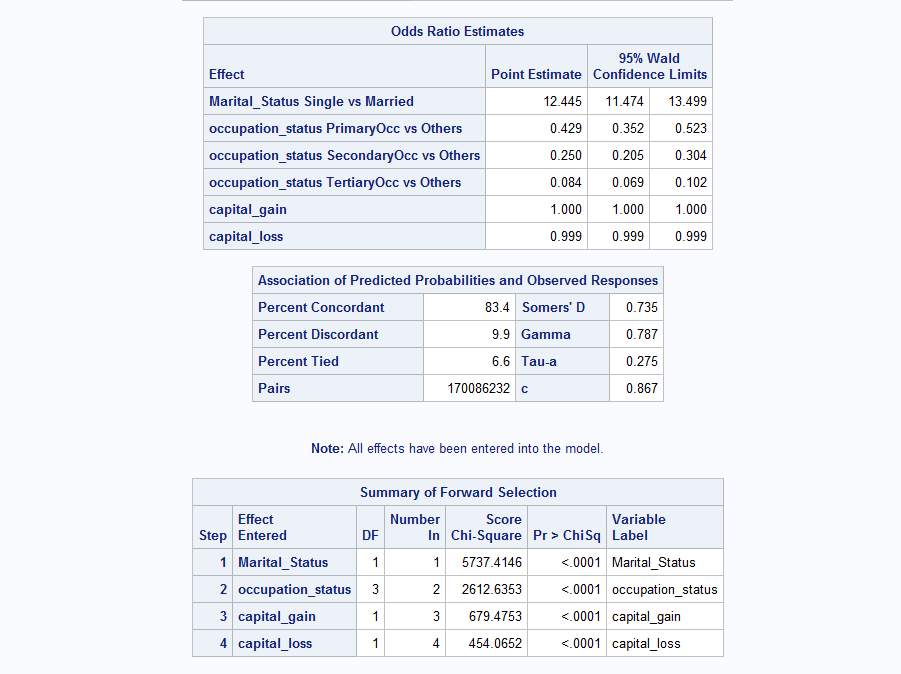












**Observation:**

After performing forward selection, all the variables have been entered into the model as their p-value is less than α(0.05), so we can conclude that all the variables are significant.

## Backward Elimination (Interaction Model)

We have performed backward elimination to our interaction model to check which variables are significant and can stay in the model. Below is the SAS code for our model:

**SAS Code**:

**proc** **logistic** data = path.adultdata;

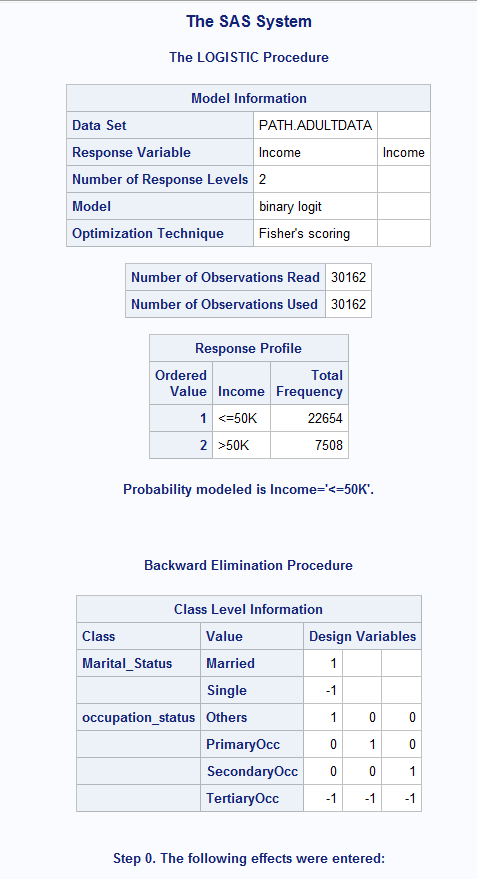
class Marital\_Status;

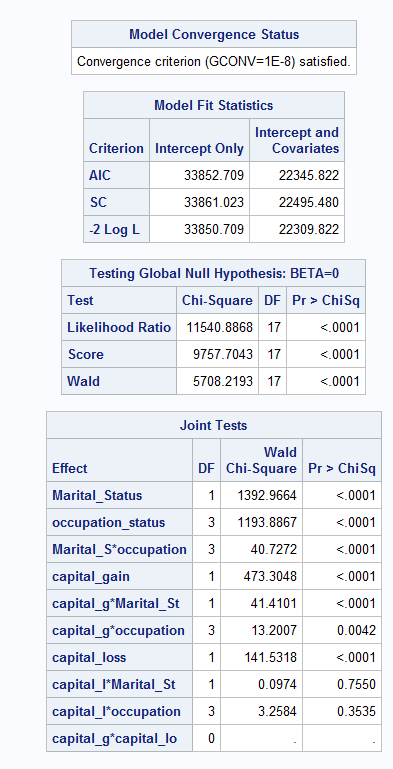
class occupation\_status;

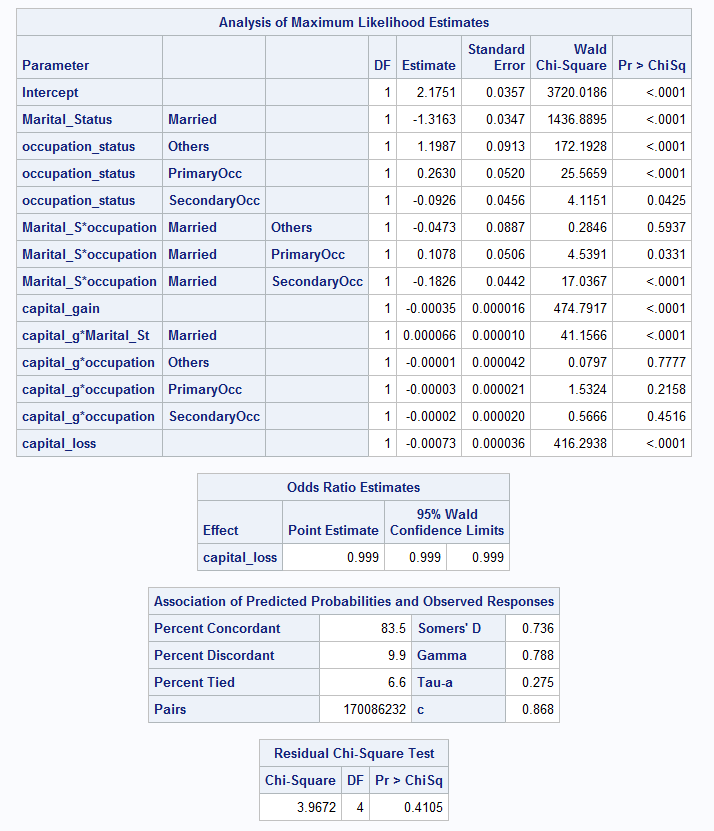
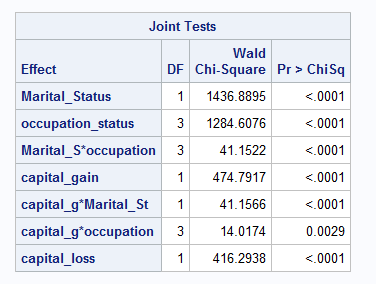
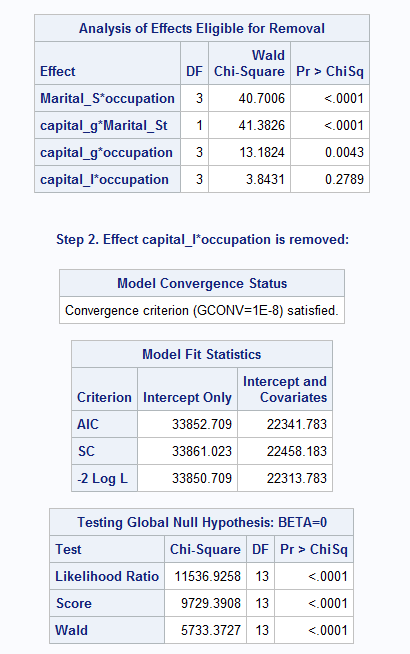
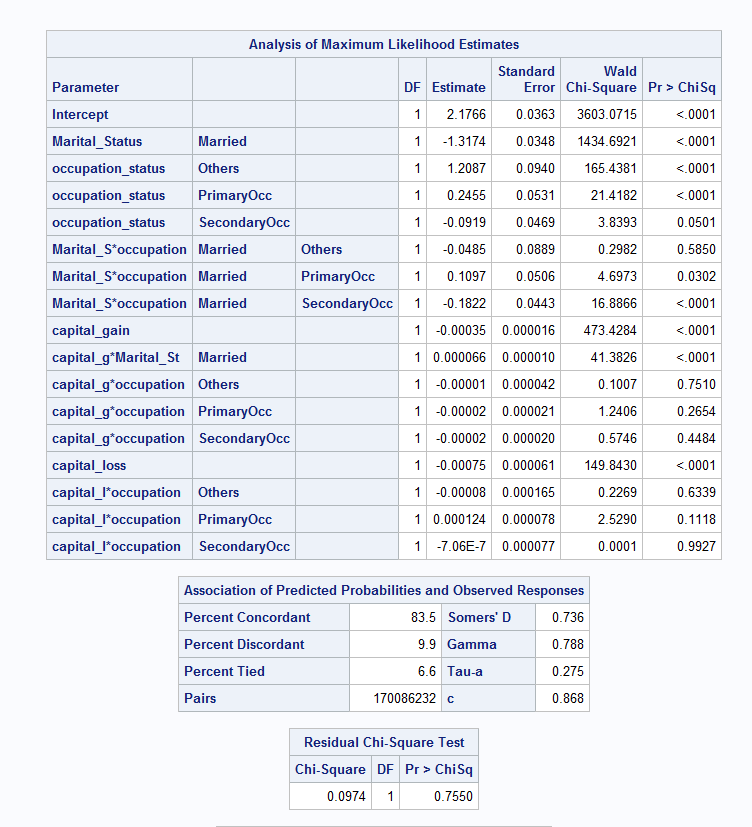
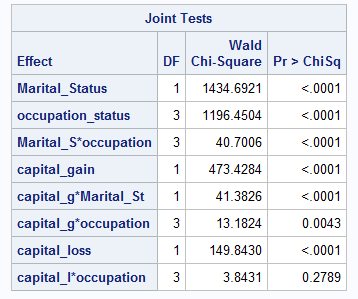
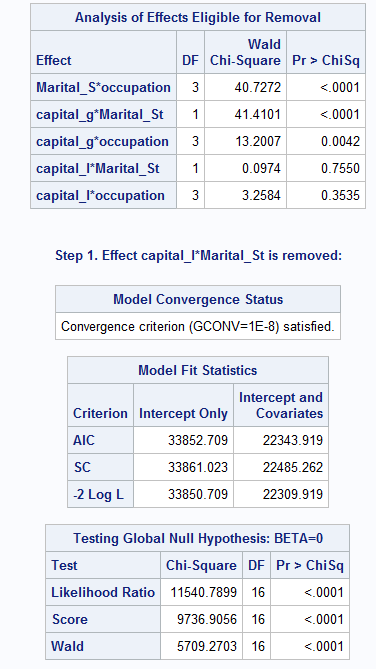
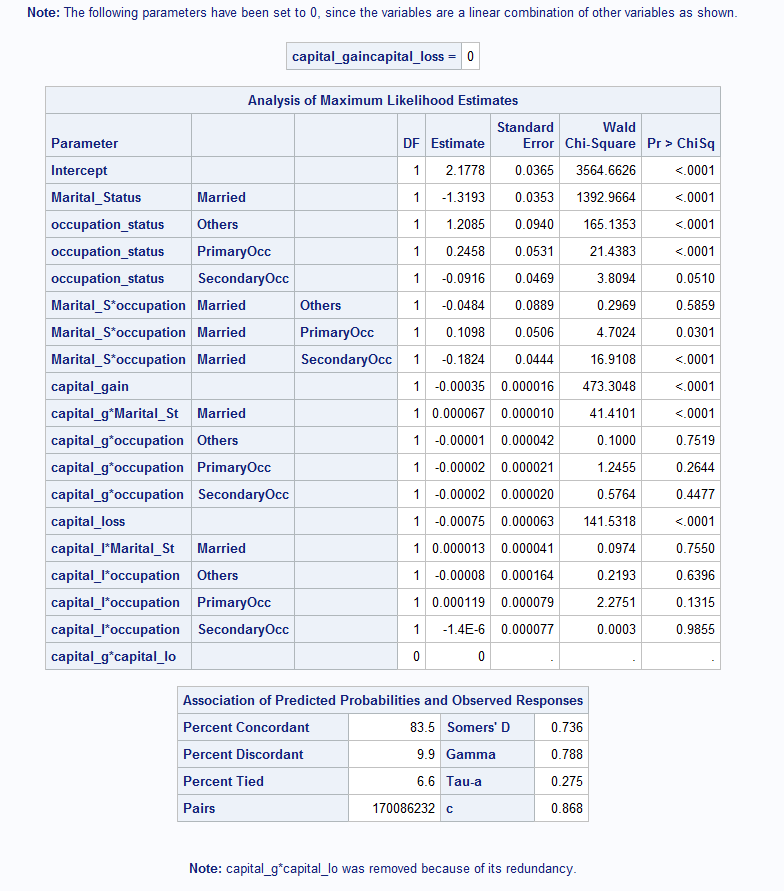
model Income = Marital\_Status|occupation\_status|capital\_gain| capital\_loss @**2** /selection=backward slstay=**0.05** hierarchy=single details;

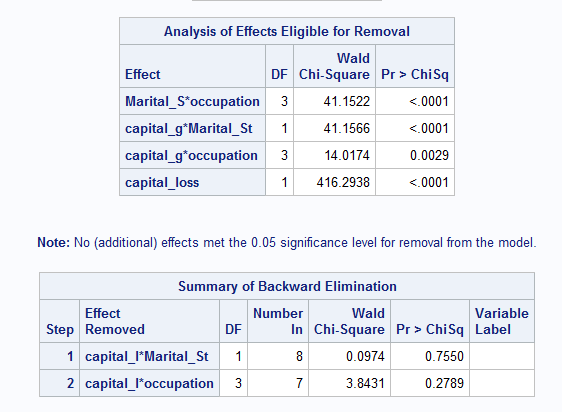
**run**;

**Output:**









**Observation:**

After performing backward elimination, in Step1 capital\_loss\*Marital\_status was eliminated as the p value was 0.7550 in the Joint Test table and in Step 2 capital\_loss\*occupation\_status was eliminated as the p-value was 0.2789 in the Joint Test table(both values being greater than α=0.05).

# Final Model SAS

After performing backward elimination to our model, we have arrived with the below mentioned final model.

Logit(Income) = 2.1751 + (-1.3163) \* Marital\_Status + 1.1987 \* occupation\_status(Others) +0.2630 \* occupation\_status(PrimaryOcc) + (-0.0926) \* occupation\_status(SecondaryOcc) + (-0.0473) \*Marital\_Status \* occupation\_status(Others) + 0.1078 \* Marital\_Status \* occupation\_status(PrimaryOcc) + (-0.1826) \* Marital\_Status \* occupation\_status(SecondaryOcc) + (-0.00035) \* capital\_gain + 0.000066 \* capital\_gain \* Marital\_Status + (-0.00001) \* capital\_gain \* occupation(Others) + (-0.00003) \* capital\_gain \* occupation (PrimaryOcc) + (-0.00002) \* capital\_gain \* occupation(SecondaryOcc) + (-0.00073) \* capital\_loss

# Classification Table

We have generated classification table for our final model to compare the predicted number of successes to the actual number of successes and predicted number of failures to the actual number of failures.

head(predict(SAS\_model,adult\_test, type = "response"))

predicted\_sas<-as.numeric(predict(SAS\_model,adult\_test, type = "response")>prop)

#Table for r model

xtabs(~adult\_test$income+predicted\_sas )

**Output:**

predicted\_sas

adult\_test$income 0 1

0 5458 1390

1 504 1697

**Observation:**

Based on the output the overall misclassification rate for the SAS model is

1894/9049=.209

For actual income <=50K

Misclassification rate 1390/6848=.202

For actual income >=50K

Misclassification rate 504/2201=.2289

# ROC Curve

ROC Curve is used to determine the sensitivity and specificity of the predictions for all possible cut off πo values. We generated the ROC Curve for our final model as below.

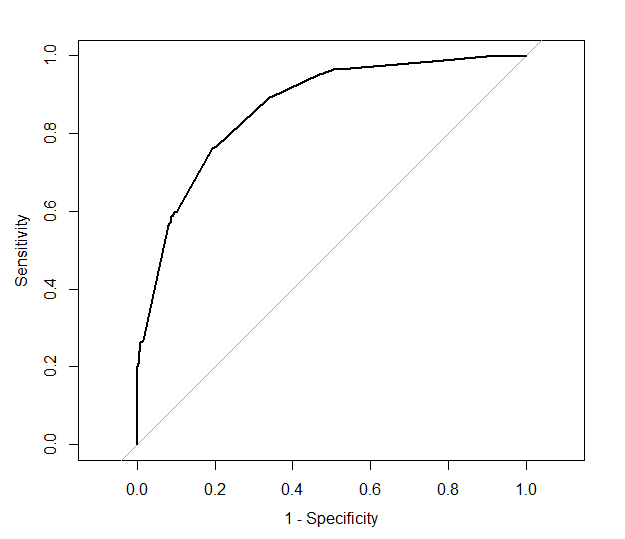
rocplot1<-roc(Income~fitted(SAS\_model),data=adult\_train)

plot.roc(rocplot1,legacy.axes = TRUE)

auc(rocplot1)

**Output:**

Area under the curve: 0.8674



**Observation:**

From the above output, we can see that the ROC Curve of our model is high and area under the curve is 0.8674 which is greater than 0.5. So, we conclude that the final model is the best fit having high curve and more area under the curve.

# Conclusion

After analyzing the classification tables and ROC Curve for the two models, we select the SAS model to be the best model among the two because area under the curve for SAS model (0.8674) is more than the area under the curve of R model (0.8669). Also, SAS model has higher curve than R model making it the better model. Also, the misclassification rate for R model is 0.208 and for SAS model is 0.209 which also concludes that SAS model is appropriate than R model.

1. Source: <https://archive.ics.uci.edu/ml/datasets/census+income> [↑](#footnote-ref-1)