RAMAKRISHNA MISSION RESIDENTIAL COLLEGE (AUTONOMOUS),

NARENDRAPUR

KOLKATA – 700103

PREDICTION OF THE PROBABILITY OF INDIAN STUDENTS GETTING PLACED AFTER MBA, USING BINARY LOGISTIC REGRESSION

[BY SANKHADEEP MITRA]

Acknowledgement

I would like to express my sincere gratitude towards my professors of Department of Statistics, Ramakrishna Mission Residential College, Narendrapur, Dr Dilip Kumar Sahoo, Dr Parthasarathi Chakraborti, Sri Tulsidas Mukherjee, Sri Palas Pal & Sri Subhadeep Banerjee for encouraging me to explore different topics for my project work and guiding me to complete it. I am very thankful to our HOD Dr Dilip Kumar Sahoo for allowing me to conduct this project. This project was done under the supervision of Mr. Tulsidas Mukherjee.

I would also like to thank my classmates for their positive support and guidance. I feel thankful to the college staff for giving me such a big opportunity. I wish to enroll in more such events in the future.

- Author

Contents

Topic	Page No
1. Abstract	1
2. Strategy and Methodology	2
(a) Logistic Regression	3
(b) Univariate Analysis	8
(c) Multivariate Analysis	12
3. Calculations	17
4. Conclusion	21
5. Reference	22

Abstract

Every year, a large number of students join the quest of finding a placement in a renowned company after completing their respective degrees (usually Master's). In this modern era, nothing comes to you without competition. Thus, this project predicts the chances of Indian students to get placed after their Master's.

This data set consists of Placement data of Indian students in a campus. It includes secondary and higher secondary school percentage and specialization. It also includes degree specialization, type and Work experience and salary offers to the placed students. It was built with the purpose of helping students have an idea about the recruitment of companies and find what chance they have with their profiles. According to their scores the possibilities of chance of admit is calculated. Our model predicts the chances of Indian students from getting selected for a job in the campus placement conducted by different Indian universities based on various factors like percentage of marks in the 10^{th} and 12^{th} boards, and Bachelor's and Master's degree, stream of Bachelor's degree, specialisation in the Master's, working experience etc, using *logistic regression*.

Strategy and Methodology

The sample data looks like

Gender	ssc_p	ssc_b	hsc_p	hsc_b	hsc_s	degree_p	degree_t	Workex	etest_p	Specialisation	mba_p	Status	Salary
M	67	Others	91	Others	Commerce	58	Sci&Tech	No	55	Mkt&HR	58.8	Placed	270000
M	79.33	Central	78.33	Others	Science	77.48	Sci&Tech	Yes	86.5	Mkt&Fin	66.28	Placed	200000
M	65	Central	68	Central	Arts	64	Comm&Mgmt	No	75	Mkt&Fin	57.8	Placed	250000
M	56	Central	52	Central	Science	52	Sci&Tech	No	66	Mkt&HR	59.43	Not Placed	
M	85.8	Central	73.6	Central	Commerce	73.3	Comm&Mgmt	No	96.8	Mkt&Fin	55.5	Placed	425000
M	55	Others	49.8	Others	Science	67.25	Sci&Tech	Yes	55	Mkt&Fin	51.58	Not Placed	
F	46	Others	49.2	Others	Commerce	79	Comm&Mgmt	No	74.28	Mkt&Fin	53.29	Not Placed	
M	82	Central	64	Central	Science	66	Sci&Tech	Yes	67	Mkt&Fin	62.14	Placed	252000
M	73	Central	79	Central	Commerce	72	Comm&Mgmt	No	91.34	Mkt&Fin	61.29	Placed	231000
M	58	Central	70	Central	Commerce	61	Comm&Mgmt	No	54	Mkt&Fin	52.21	Not Placed	
M	58	Central	61	Central	Commerce	60	Comm&Mgmt	Yes	62	Mkt&HR	60.85	Placed	260000
M	69.6	Central	68.4	Central	Commerce	78.3	Comm&Mgmt	Yes	60	Mkt&Fin	63.7	Placed	250000
F	47	Central	55	Others	Science	65	Comm&Mgmt	No	62	Mkt&HR	65.04	Not Placed	

where

ssc_p = Secondary Education percentage- 10th Grade

ssc_b = Board of Education- Central/ Others

hsc_p= Higher Secondary Education percentage- 12th Grade

hsc_b= Board of Education- Central/ Others

hsc_s= Specialization in Higher Secondary Education

degree_p= Degree Percentage

degree_t= Under Graduation (Degree type)- Field of degree education

Workex= Work Experience

etest_p= Employability test percentage (conducted by college)

Specialisation= Post Graduation (MBA) Specialization

mba_p= MBA percentage

Status= Status of placement- Placed/Not placed

Salary= Salary offered by corporate to candidates

➢ LOGISTIC REGRESSION

In statistics, the logistic model (or logit model) is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick.

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression). Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable, where the two values are labelled "0" and "1". In the logistic model, the log-odds (the logarithm of the odds) for the value labelled "1" is a linear combination of one or more independent variables ("predictors"); the independent variables can each be a binary variable (two classes, coded by an

indicator variable) or a continuous variable (any real value). The corresponding probability of the value labelled "1" can vary between 0 (certainly the value "0") and 1 (certainly the value "1"), hence the labelling; the function that converts log-odds to probability is the logistic function, hence the name. The unit of measurement for the log-odds scale is called a logit, from logistic unit, hence the alternative names. Analogous models with a different sigmoid function instead of the logistic function can also be used, such as the probit model; the defining characteristic of the logistic model is that increasing one of the independent variables multiplicatively scales the odds of the given outcome at a constant rate, with each independent variable having its own parameter; for a binary dependent variable this generalizes the odds ratio.

In a binary logistic regression model, the dependent variable has two levels (categorical).

Let us try to understand logistic regression by considering a logistic model with given parameters, then seeing how the coefficients can be estimated from data. Consider a model with one predictor, x, and one binary (Bernoulli) response variable Y, which we denote $p = P(Y = 1)$. We assume a linear relationship between the predictor variables and the log-odds (also called logit) of the event that $Y = 1$. This linear relationship can be written in the following mathematical form (where l is the log-odds, b is the base of the logarithm, and β_i are parameters of the model):

$$l = log_b \frac{p}{1-p} = \beta_0 + \beta_1 x$$

We can recover the odds by exponentiating the log-odds:

$$\frac{p}{1-p} = b^{\beta_0 + \beta_1 x}$$

By simple algebraic manipulation, we get

$p = \frac{1}{1+b^{-(\beta_0+\beta_1 x)}} = S_b(\beta_0 + \beta_1 x)$,

where S_b is the sigmoid function with base b. The above formula shows that once β_i's are fixed, we can easily compute either the log-odds that $Y = 1$ for a given observation, or the probability that $Y = 1$ for a given observation. The main use-case of a logistic model is to be given an observation x and estimate the probability p that $Y = 1$. In most cases, the base b is taken to be e, or 2 or 10 for the ease of calculation.

Probability of passing an exam versus hours of study

To answer the following question:

A group of 20 students spends between 0 and 6 hours studying for an exam. How does the number of hours spent studying affect the probability of the student passing the exam?

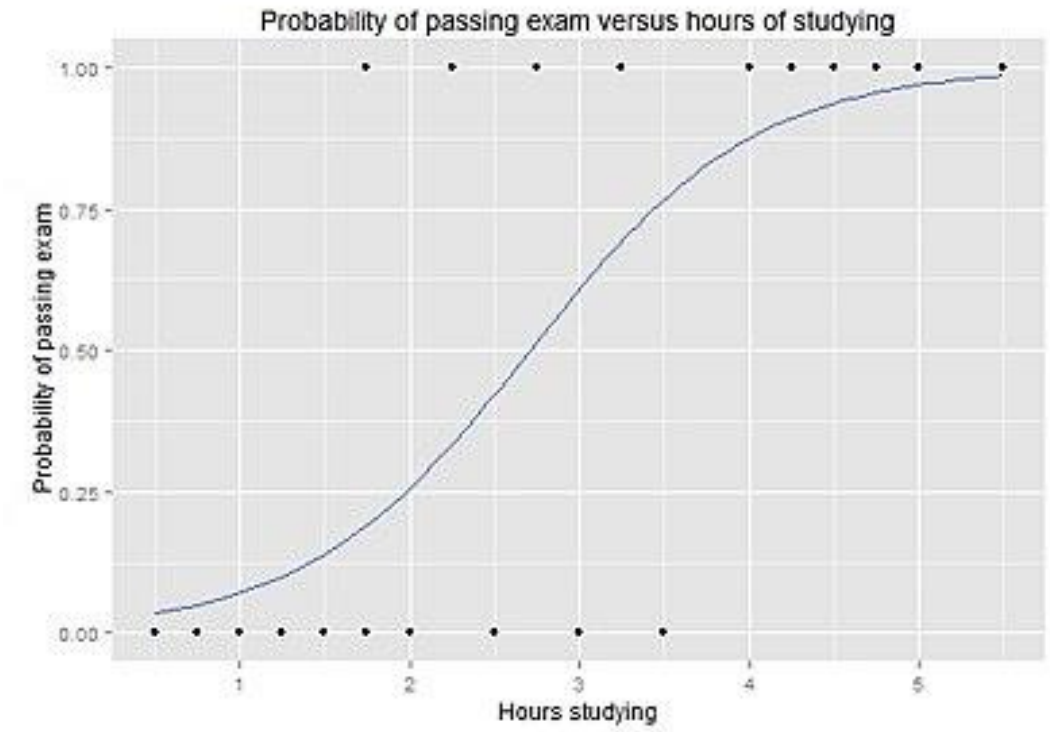
The reason for using logistic regression for this problem is that the values of the dependent variable, pass and fail, while represented by "1" and "0", are not cardinal numbers. If the problem was changed so

that pass/fail was replaced with the grade 0–100 (cardinal numbers), then simple regression analysis could be used.

The table shows the number of hours each student spent studying, and whether they passed (1) or failed (0).

Hours	Pass
0.5	0
0.75	0
1	0
1.25	0
1.5	0
1.75	0
2	1
2.25	0
2.5	1
2.75	0
3	1
3.25	0
3.5	1
3.75	0
4	1
4.25	1
4.5	1
4.75	1
5	1
5.25	1
5.5	1

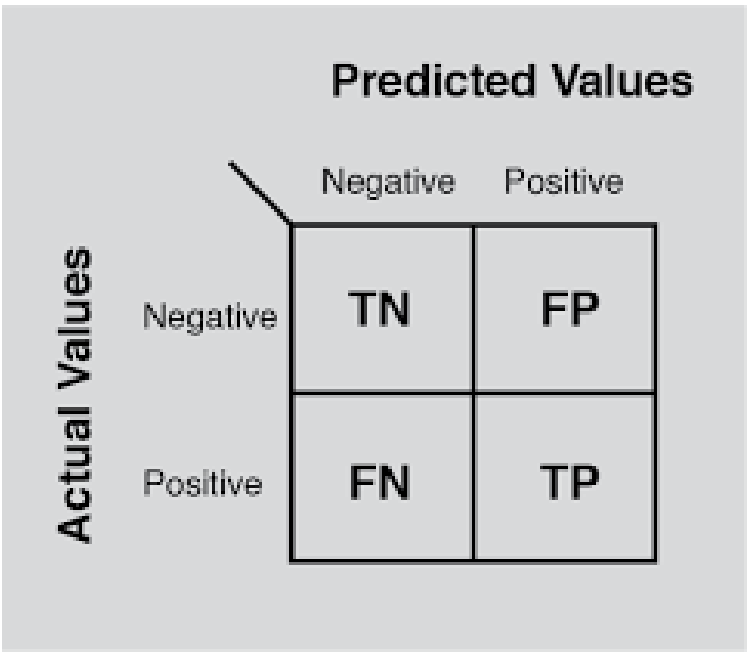
The graph shows the probability of passing the exam versus the number of hours studying, with the logistic regression curve fitted to the data.



Graph of a logistic regression curve showing probability of passing an exam versus hours studying

- Confusion Matrix

It is a tabular representation of Actual vs Predicted values. This helps us to find the accuracy of the model and avoid overfitting. This is how it looks like



We can calculate the accuracy of our model by

$$\frac{\text{True Positive + True Negatives}}{\text{True Positive + True Negatives + False Positives + False Negatives}}$$

➢ UNIVARIATE ANALYSIS

To study the dependence of the study variable (status) on the attribute type factors - 12th standard board (hsc_s), stream in the Bachelor's (degree_t), any previous work experience (workex), specialisation in MBA (specialisation), we construct two-way tables of

admit with each of these factors. Following are the two-way tables:

	hsc_s		
status	Arts	Commerce	Science
Not Placed	5	34	28
Placed	6	79	63

	degree_t		
status	Comm&Mgmt	Others	Sci&Tech
Not Placed	43	6	18
Placed	102	5	41

	workex	
status	No	Yes
Not Placed	57	10
Placed	84	64

status	specialisation	
	Mkt&Fin	Mkt&HR
Not Placed	25	42
Placed	95	53

By fitting a univariate logistic model

$$log\frac{p}{1-p} = \beta_0 + \beta_1 x$$

to the data, we get the following estimates of the coefficients in the model, for each of the 4 factors.

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.2877	0.7638	-0.377	0.706
hsc_sCommerce	1.2119	0.7995	1.516	0.130
hsc_sScience	0.9621	0.7995	1.203	0.229

Factor- 12th standard stream

From the table, we see that no stream is significant, at 5% level, in determining the chances of being placed.

```
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)        0.8979     0.2065   4.348 1.37e-05 ***
degree_tOthers    -1.4088     0.7589  -1.856   0.0634 .
degree_tSci&Tech  -0.2330     0.3561  -0.654   0.5129
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Factor- Stream in the Bachelor’s

From the table, we see that no stream is significant, at 5% level, in determining the chances of being placed.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.3709     0.1857   1.997  0.04584 *
workexYes     1.5544     0.4452   3.492  0.00048 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Factor- Work experience

From the table, we see that any previous work experience is significant, at 5% level, in determining the chances of being placed.

```
Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)             1.3218     0.2517   5.252  1.5e-07 ***
specialisationMkt&HR   -1.1211     0.3374  -3.323 0.000891 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Factor- MBA specialisation

From the table, we see that specialisation in MBA is significant, at 5% level, in determining the chances of being placed. In fact, the negative coefficient in the model indicates that there are higher chances in getting placed for those doing specialisation in the other course, Marketing & Finance.

➢ MULTIVARIATE ANALYSIS

Multivariate analysis (MVA) is based on the principles of multivariate statistics. Typically, MVA is used to address the situations where multiple measurements are made on each experimental unit and the relations among these measurements and their structures are important. Here we'll do logistic regression with multiple explanatory variables. If there are multiple explanatory variables, the above expression $\beta_0 + \beta_1 x$

can be revised to $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$. Then when this is used in the equation relating the log odds of a success to the values of the predictors, the linear regression will be a multiple regression with m explanators; the parameters $\beta_i, i = 0,1,2, \ldots, m$ are all estimated.

Similarly,

$$log_b \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$$

which reduces to

$$p = \frac{1}{1 + b^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m)}}$$

where usually $b = e$.

Step 1

By converting the response variable from numeric to factor we create the desire format of dependent. We also check whether there is any missing value or not in our complete data set.

Step 2

We partition our data set of 215 observations into 80-20% for training and testing data by simple random sampling without replacement. We have 175 observations of 14 variables in training data and 40 observations of 14 variables in testing data. Now we run

a logistic regression model on the training data set. As our response is a binary variable with values 0 & 1, we use binary logistic regression. The following table shows the result with all predictors (except gender):

```
Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          -14.03551    4.63944  -3.025 0.002484 **
ssc_p                  0.23036    0.05309   4.339 1.43e-05 ***
ssc_bOthers            0.59335    0.82968   0.715 0.474517
hsc_p                  0.11055    0.03955   2.795 0.005183 **
hsc_bOthers            0.19124    0.82260   0.232 0.816162
hsc_sCommerce         -1.16980    1.55664  -0.751 0.452359
hsc_sScience          -0.28741    1.71335  -0.168 0.866780
degree_p               0.13196    0.05238   2.519 0.011766 *
degree_tOthers        -2.38510    1.93855  -1.230 0.218566
degree_tSci&Tech      -2.22882    0.91628  -2.432 0.014997 *
workexYes              1.99387    0.77453   2.574 0.010045 *
etest_p               -0.01863    0.02465  -0.756 0.449783
specialisationMkt&HR  -0.31332    0.61385  -0.510 0.609760
mba_p                 -0.22213    0.06035  -3.680 0.000233 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From the above table we observe different statistics and their corresponding significance level for all the predictors. We observe that ssc_b, hsc_b, hsc_s, etest_p, specialisation are not significant at 5% level.

Step 3

Now we rerun the code by dropping the statistically insignificant variables. Here is the following result:

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -15.47858    4.12215  -3.755 0.000173 ***
ssc_p              0.22015    0.04466   4.930 8.24e-07 ***
hsc_p              0.09562    0.03555   2.690 0.007148 **
degree_p           0.12585    0.04973   2.531 0.011388 *
degree_tOthers    -1.71938    1.94811  -0.883 0.377458
degree_tSci&Tech  -1.62721    0.64843  -2.509 0.012091 *
workexYes          2.23116    0.74460   2.996 0.002732 **
mba_p             -0.19992    0.05332  -3.749 0.000177 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

This is our final model which we will use for prediction.

To check the performance of our fitted model we use the train data set for prediction. Here we have first few predicted values:

```
          1           2           3           4           6           7
0.868661911 0.998613079 0.861976336 0.005807882 0.536563118 0.204358012
```

Here is the following table from where we can predict them:

	sl_no	gender	ssc_p	ssc_b	hsc_p	hsc_b	hsc_s	degree_p	degree_t	workex	etest_p	specialisation	mba_p	status	salary
1	1	M	67.00	Others	91.00	Others	Commerce	58.00	Sci&Tech	No	55.00	Mkt&HR	58.80	Placed	270000
2	2	M	79.33	Central	78.33	Others	Science	77.48	Sci&Tech	Yes	86.50	Mkt&Fin	66.28	Placed	200000
3	3	M	65.00	Central	68.00	Central	Arts	64.00	Comm&Mgmt	No	75.00	Mkt&Fin	57.80	Placed	250000
4	4	M	56.00	Central	52.00	Central	Science	52.00	Sci&Tech	No	66.00	Mkt&HR	59.43	Not Placed	NA
6	6	M	55.00	Others	49.80	Others	Science	67.25	Sci&Tech	Yes	55.00	Mkt&Fin	51.58	Not Placed	NA
7	7	F	46.00	Others	49.20	Others	Commerce	79.00	Comm&Mgmt	No	74.28	Mkt&Fin	53.29	Not Placed	NA

Now look at the 1st, 2nd, 3rd observations. We can see that the chances of them to get placed are very high (>50%), so these students are predicted to be placed. On the other hand, the 4th, 6th and 7th students have a very low chance (<50%) of being placed.

Calculations

If $p =$ Probability of being placed, and

$1 - p =$ Probability of not being placed, then

Odds of being placed $= p/(1 - p)$.

Now,

$$Logit(Y) = log(odds) = log\left(\frac{p}{1-p}\right)$$
$$= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$$

Therefore, $log\ (p/1 - p) = Y$

By rearranging, we get

$p = \frac{1}{1+e^{-Y}}$

Now we rewrite the equation by using the estimates of the final model

$$log\left(\frac{p}{1-p}\right) = Y$$

$$= -15.47858 + 0.22015(ssc\_p) + 0.09562(hsc\_p) + 0.12585(degree\_p) - 0.19992(mba\_p) + -1.71938(I_{degreeOthers}) - 1.62721(I_{degreeSci\&Tech}) + 2.23116(I_{workex}),$$

where I_A is the indicator variable.

For the 1st student,

$ssc\_p = 67$

$hsc\_p = 91$

$degree\_p = 58$

$mba\_p = 58.8$

$I_{degreeOthers} = 0$

$I_{degreeSci\&Tech} = 1$

$I_{workex} = 0$

So, $Y = 1.889684$ and $p = 0.8687195$

As the probability is more than 0.5, we conclude that 1st student will get placed.

For the 2nd student,

$ssc\_p = 79.33$

$hsc\_p = 78.33$

$degree\_p = 77.48$

$mba\_p = 66.28$

$I_{degreeOthers} = 0$

$I_{degreeSci\&Tech} = 1$

$I_{workex} = 1$

So, $Y = 6.579945$ and $p = 0.998614$

As the probability is more than 0.5, we conclude that 2nd student will get placed, in fact, this student has a very high chance of getting placed.

For the 4th student,

$ssc\_p = 56$

$hsc\_p = 52$

$degree\_p = 52$

$mba\_p = 59.43$

$I_{degreeOthers} = 0$

$I_{degreeSci\&Tech} = 1$

$I_{workex} = 0$

So, $Y = -5.142196 \; and \; p = 0.005810879$

As the probability is less than 0.5, we conclude that 4th student will not get placed.

- Accuracy of Training Data

If the value of p is greater than 0.5, we predict that event as 1, i.e., the student will get placed otherwise 0, i.e., the student will not get placed.

Now we run the model for training data and get the following confusion matrix:

	Actual	
Predicted	Not Placed	Placed
0	45	9
1	11	110

Hence the accuracy of our model

= (45+110)/ (45+9+11+110)

= 0.8857143

- Accuracy of Testing Data

We run the model for testing data and get the following confusion matrix:

	Actual	
Predicted	Not Placed	Placed
0	10	3
1	1	26

Hence the accuracy of our model

= (10+26)/ (10+3+1+26)

= 0.9

Conclusion

After studying the data by fitting a Binary Logistic Regression model to the dataset,

- We are able to predict the probability of the Indian students getting placed, after their MBA, for the training and testing data with an accuracy of approximately 88%.
- We also conclude that among the predictors, percentage of marks in MBA, any previous working experience and percentage of marks in the 10^{th} and 12^{th} standard play a significant role.
- Also having a Science & technology in the Bachelor's and a work experience adds to the chances.
- The specialisation in MBA is found to be significant in the univariate case and the opposite in multivariate case. As a result, MBA specialisation does not make any significant difference in campus recruitment.

Reference

1. Sources of information
   https://en.wikipedia.org/wiki/Logistic_regression

2. Source of data
   Campus Recruitment | Kaggle

3. R code for the project

Logistic regression

Read datafile
```
mydata <- read.csv(file.choose(), header=T)
str(mydata)
mydata$gender <- as.factor(mydata$gender)
mydata$ssc_b <- as.factor(mydata$ssc_b)
mydata$hsc_b <- as.factor(mydata$hsc_b)
mydata$hsc_s <- as.factor(mydata$hsc_s)
mydata$degree_t <- as.factor(mydata$degree_t)
mydata$workex <- as.factor(mydata$workex)
mydata$specialisation <-
as.factor(mydata$specialisation)
mydata$status <- as.factor(mydata$status)
```

2-way table of factor variables
```
xtabs(~status + hsc_s, data=mydata)
xtabs(~status + degree_t, data=mydata)
```

```
xtabs(~status + workex, data=mydata)
xtabs(~status + specialisation, data=mydata)

# Partition data - train(80%) & test(20%)
set.seed(1234)
ind <- sample(2, nrow(mydata), replace=T, prob=c(0.8, 0.2))
train <- mydata[ind==1,]
test <- mydata[ind==2,]

# Logistic regression model
mymodel <- glm(status ~ ssc_p + ssc_b + hsc_p + hsc_b + hsc_s + degree_p + degree_t + workex + etest_p + specialisation + mba_p, data=train, family= 'binomial')
summary(mymodel)
mymodel1 <- glm(status ~ ssc_p + hsc_p + degree_p + degree_t + workex + mba_p, data=train, family= 'binomial')
summary(mymodel1)

# Univariate model
model1 <- glm(status ~ hsc_s, data= train, family= 'binomial')
summary(model1)
model2 <- glm(status ~ degree_t, data= train, family= 'binomial')
summary(model2)
model3 <- glm(status ~ workex, data= train, family= 'binomial')
```

```
summary(model3)
model4 <- glm(status ~ specialisation, data= train,
family= 'binomial')
summary(model4)

# Prediction
p1 <- predict(mymodel1, train, type='response')
head(p1)
head(train)

# Misclassification error- train data
pred1 <- ifelse(p1>0.5, 1, 0)
tab1 <- table(Predicted= pred1, Actual= train$status)
tab1
sum(diag(tab1))/sum(tab1)

# Misclassification error- test data
p2 <- predict(mymodel1, test, type='response')
pred2 <- ifelse(p2>0.5, 1, 0)
tab2 <- table(Predicted= pred2, Actual= test$status)
tab2
sum(diag(tab2))/sum(tab2)
```