Project Report

On

Candidate Joining Predictive Model using Data Mining Techniques

Submitted in requirement of partial fulfillment of
Post Graduate Diploma in Management (Executive)
Specialization: Management Information Systems

By
Ankur Shrivastava
(Student PGDMe)
Roll No: 1625600055

INSTITUTE OF MANAGEMENT TECHNOLOGY
CENTRE FOR DISTANCE LEARNING,
GHAZIABAD

ACKNOWLEDGMENTS

It was a great opportunity for me to work with Malwa Institute of Science and Technology, Indore. I am extremely grateful to those who have shared their expertise and knowledge with me and without whom the completion of this project would have been virtually impossible.

I would like to thank my project guide **Mr. Akshat Shrivastava, Group Head HR, Malwa Institute of Science and Technology, Indore**, who has been a constant source of inspiration for me during the completion of this project. He gave me invaluable inputs during my endeavor to complete this project.

MALWA GROUP OF INSTITUTIONS

MALWA INSTITUTE OF SCIENCE & TECHNOLOGY, INDORE
Icon Of The Future...

MALWA INSTITUTE OF SCIENCE & MANAGEMENT, INDORE
Icon Of The Future...

CERTIFICATE

This is to certify that MR. ANKUR SHRIVASTAVA a student of IMT-CDL Ghaziabad has completed ~~Minor~~/Major project work on CANDIDATE JOINING PREDICTIVE MODEL USING DATA MINING TECHNIQUES under my guidance and supervision.

I certify that this is an original work and has not been copied from any source.

Signature of guide _____

Name of the Project guide AKSHAT SHRIVASTAVA

Malwa Institute of Science & Technology INDORE (Head HR)

CAMPUS :
VILLAGE LIMBODAGARI, BEHIND AUROBINDO HOSPITAL, POST PALIA VIA HATOD, INDORE DIST. 453 111
TEL.: +91 731 6 777 777, 6 777 788 FAX: +91 731 6 777 799
E-mail: info@mistindore.com Website: www.mistindore.com
CITY OFFICE :
FF-33, SCHEME NO. 54, VIJAY NAGAR, INDORE (M.P.)
TEL. : +91 731 4089333 FAX : +91 731 4074333

INSTITUTE OF MANAGEMENT TECHNOLOGY
Centre for Distance Learning
(Recognized by Distance Education Council)

View Synopsis

Enrollment No: 1625600055

Name: ANKUR SHRIVASTAVA

Area of Specialisation: Systems

Title of the Project: **Candidate Joining Predictive Model using Data Mining Techniques**

Statement of the problem: Malwa Institute of Science and Technology, Indore is currently facing the problem of low conversion rate of candidates joining the institute post extension of offers of employment. Sometimes, the institute is kept waiting for days by a candidate before declining resulting in disruption of the normal academic cycle of the institute. This results in high overhead and lost productivity for the institute. if the person-hours thus wasted on non-joining candidates could be reduced, they could be used for more meaningful and productive activities.

TOPIC- Evaluation & Comments from Faculty
CORRECT

Objectives of the study: The study aims to analyze various factors during the full recruitment cycle right from ad posting to joining of candidates from the data collected during this process and then build a predictive model of whether a candidate is likely to join the institute or not depending on the various factors identified during the study. This will help the institute to narrow down its focus on candidates most likely to join.

OBJECTIVE- Evaluation & Comments from Faculty
CORRECT

Methodology: Secondary Data

METHODOLOGY- Evaluation & Comments from Faculty
CORRECT

Explanation of the Method: Sample population will the entire dataset of candidates who have gone through the recruitment process available with the institute.

Company Name: Malwa Institute of Science and Technology, Indore

Company Profile: Malwa Institute of Science & Technology (www.mistindore.com) is affiliated to Rajiv Gandhi Technical University and approved by AICTE. The institute has dynamic vision to crystallize in its myriad forms which in its fulfilment develops into a multi disciplinary technological centre of excellence for providing technical education and research opportunities to students in addition to developing the complete personality of an individual so as to instil high levels of technical, managerial skills with discipline.

Questionnaire (10 to 15 questions)

Number of respondents:

Area of study: Indore

Method you will use to classify data:

Method you will use to present data:

QUESTIONNAIRE- Evaluation & Comments from Faculty
CORRECT

References: (1) Data Mining and Business Intelligence by M. Sudheep Elayidom (2) Data Mining Concepts and Techniques by Jiawai Han, Micheline Kamber & Jian Pei. (3) Data Mining Techniques by Gordon S. Linoff and Michael J.A. Berry

Chapterization Scheme:

1. Introduction to the Study
2. Company profile
3. Theoretical Perspective
4. Study Objective
5. Data Preparation
6. Data Analysis and Interpretation
7. Model Building

8. Test and Evauation
9. Conclusions and Recommendations
10. Bibliography

CHAPTERIZATION- Evaluation & Comments from Faculty
CORRECT

Profile of Project Guide

Name:	Akshat Shrivastava
Age:	47
Educational Qualification:	B.A. (Econ) Osmania University, Executive MBA IIM Calcutta, PGCHRM XLRI Jamshedpur
Years of Experience:	24
Current organisation:	Malwa Institute of Science & Technology, Indore
Current designation:	Group Head HR
Brief profile:	
Address:	
House No:.	118, Sharma Enclave
Street:	Girdhar Nagar
City:	Indore
State:	Madhya Pradesh
Phone Number (Residence):	0731-2497206
Phone Number (Office):	0731-6777777
Mobile:	9179221139
Email:	akshat_s@hotmail.com

PROJECT GUIDE- Evaluation & Comments from Faculty
CORRECT

FINAL COMMENTS FROM FACULTY
Approved

FACULTY DETAILS
Faculty Name: N.M. Mishra
Email: nmmishra@imtcdl.ac.in

CLOSE

TABLE OF CONTENTS

CHAPTER **PAGE NUMBER**

CHAPTER - 1

Introduction

INTRODUCTION

People are the most important resources of any organization and the quality of people hired by an organization has the potential to make or break the organization. Wisely investing in people defines the ability of an organization to perform, compete, innovate and succeed. Therefore the organizations must hire the right skills and mindset and who are a good cultural fit. (Jean Paul Isson & Jesse S. Harriott, 2016) [1].

Traditional talent acquisition approaches that are solely based upon resumes and interviews are no longer sufficient in proactively identifying the best applicants. This methodology also fails to identify which candidates will be successful at the job once hired (Jean Paul Isson & Jesse S. Harriott, 2016) [1].

Then there is also the problem of recruitment Turnaround Time.

Depending on the popularity of the open job requisition, hiring managers can quickly become inundated with myriad resumes and job applicant profiles that they are required to sift through in order to uncover the best candidate matches. To succeed, employers must act quickly, precisely, and cost-effectively. This requires the use of advanced analytics tools and solutions when looking to fill these types of roles (Jean Paul Isson & Jesse S. Harriott, 2016) [1].

Hiring the wrong person for a job can undermine an organization's success and interfere with its ability to compete and lead in the market. According to Ryan Holmes, CEO of Hootsuite, "One subpar employee can throw an entire department into disarray. Team members end up investing their own time into training someone who has no future with the company. (Jean Paul Isson & Jesse S. Harriott, 2016) [1].

A similar problem is faced by Malwa Institute of Science and Technology (MIST) located at Indore. MIST is currently facing the problem of low conversion rate of candidates joining the institute post shortlisting of resumes. Sometimes, the institute is kept waiting for days by a candidate before declining resulting in disruption of the normal academic cycle of the institute. This results in high overhead and lost productivity for the institute. If the person-hours thus wasted on non-joining candidates could be reduced, they could be used for more meaningful and productive activities.

This study attempts to analyze the various factors that could be responsible for the eventual onboarding of the candidates from the data available from the institute. The study will also attempt to apply two popular classification algorithms to the data and see whether they can perform any better than the baseline model. The predictive model thus built will help in predicting whether a candidate is likely to join the institute or not depending on the various factors identified during the study. This will help the institute to narrow down its focus on candidates most likely to join.

CHAPTER – 2

ORGANIZATION PROFILE

Malwa Institute of Science & Technology (www.mistindore.com) is a higher education institute located at Limbodagari, Sanwer Road, Indore (M.P.) It is run by Patel Education & Welfare Society (PEWS). The society has an aim of providing high quality technical education in Madhya Pradesh. The Society and its educational institutions are the dream of Late Shri Hargovindh Patel Shivhare, one of the famous Industrialists of his time. His son, Shri R.S. Shivhare the chairman of PEWS is an eminent industrialist, Entrepreneur, Philanthropist and a Visionary. He firmly believes that modern education has to have its roots in strong values. Shri R.S. Shivhare is also the Vice President of Indus Global Education & Welfare society as well as the Vice Chairman of Governing Body of Malwa Institute of Technology, Malwa Institute Of Pharmacy and Chairman of Malwa Institute of Science & Technology (MIST). He has great vision for Education endeavors. Within a short span of six years, he has taken the institute to a high velocity growth trajectory. These six years have catapulted Malwa Institute of Science and Technology (MIST) a name to reckon with in the academic landscape of Madhya Pradesh.

Vision

MIST has dynamic vision to emerge as a center of excellence and innovation proactively catalyzing growth of corporate & management world by imparting futuristic technical education at par with international standards.

Mission

MIST commits itself to create an empowered society comprising of professionals who possess right mind set, knowledge, skills, ethical values and heightened sense of moral and social responsibilities, personally succeed in a way that leads to collective global success.

CHAPTER – 3

THEORETICAL PERSPECTIVE

“Arguably the most practical tool and greatest potential for organizational management is the emergence of predictive analytics.”

(Jac Fitz-Enz and John R. Mattox II, 2014)[2]

“Analytics present a tremendous opportunity to help organizations understand what they don’t yet know... By identifying trends and patterns, HR professionals and management teams can make better strategic decisions about the workforce challenges that they may soon face.”

(Huselid, 2014).

Recruiting is the process of developing a pool of qualified applicants who are interested in working for the organization and from which the organization might reasonably select the best individual or individuals to hire for employment. Recruitment entails substantial costs. Organizations can reduce this cost by ensuring high retention of existing employees. Also, careful planning of recruitment goes a long way in reducing the cost of recruitment. (Debra Nelson, James Quick, Preetam Khandelwal, Angelo DeNisi, Ricky Griffin, Anita Sarkar, 2014) [3]

Towards the goal of recruitment, the organization wants to **optimize the size of the pool of qualified applicants**, that is, have enough candidates to be able to make choices but not so many that processing would become overwhelming. The recruitment process should generate a pool of applicants that is both qualified and **interested in working for the organization**. (Debra Nelson, James Quick, Preetam Khandelwal, Angelo DeNisi, Ricky Griffin, Anita Sarkar, 2014) [3]

This study aims to down the cost of recruitment by narrowing down the list of candidates who are likely to join by analyzing past recruitment data for various factors that could be related to a candidate joining or not joining an institute.

The study also aims to apply two of the popular classification methods: Logistic Regression and Decision Trees.

Logistic Regression is a form of regression analysis which is used for prediction of discrete variables using a mix of continuous and discrete predictors. Instead of building a predictive model for "Y (Response)" directly, the approach models Log Odds (Y); hence the name Logistic regression. The dependent variables are categorical and the independent variables are continuous or categorical.

A decision tree is a flowchart-like tree structure, where each internal node (non-leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label. The topmost node in a tree is the root node. Decision tree induction is the learning of decision trees from class-labeled training tuples. (Jiawei Han, Micheline Kamber & Jian Pei, 2012)[4]. There are various algorithms for decision tree induction like ID3, C4.5 and CART (Classification and Regression Trees). This study will utilize the CART algorithm.

CHAPTER – 4

STUDY OBJECTIVE & METHODOLOGY

PRIMARY OBJECTIVES

- To explore and analyze the various factors related to a candidate eventually joining the organization from the pool of shortlisted candidates.

SECONDARY OBJECTIVES

- To apply classification algorithms like Logistic Regression and Decision Trees and compare their performance at prediction of candidate joining.

METHODOLOGY

- This study will analyze the secondary data supplied by MIST which is generated during the recruitment process from the time of shortlisting to eventual joining of candidates.

- The study will involve the following steps:

1. **Data Preparation**- This step will involve collecting and cleaning data and making it ready for analysis by 'R' statistical software. This step will mainly be carried out in Microsoft Excel. This step will also involve the creation of a data dictionary which will explain the various variables in the dataset.
2. **Data Analysis and Interpretation-** This step will involve loading and transforming data in 'R' , visualizing data , running appropriate statistical tests and drawing inferences about the relationship between variables. This step will also involve making the data ready for modeling.

3. **Model Building**- This step will involve applying the classification algorithms on the data prepared thus far and interpreting the results.
4. **Test and Evaluation**- This step will involve testing the model prepared in the previous step on a portion of the dataset not used for model building and checking the accuracy of the model.
5. **Conclusions and Recommendations**- This step will involve making final conclusion, explaining any anomalies or shortcomings in the model. Also, any recommendations will be provided regarding usage of the model.

- **Software Tools and Utilities-** This study will utilize Microsoft Excel, R and R Studio softwares. Microsoft Excel will be utilized primarily for cleaning purposes and R and R Studio for all the remaining steps.
- **Methods / Tools of Analysis-** Tools used for analysis are:

- Chi-square test
- Karl Pearson's coefficient of correlation
- Graph
- Percentage

1. CHI-SQUARE TEST

There may be situation in which it is not possible to make any rigid assumption about distribution of the population from which samples being drawn. This limitation has led to the development of a group of alternative techniques known as non-parametric tests. Chi-square describes the magnitude of the discrepancy between theory and observation.

$$\chi^2 = \sum_{i=1} \left[\frac{(O_i - E_i)^2}{E_i}\right] \text{ with n-1 degrees of freedom}$$

2. KARL PEARSON'S COEFFICIENT OF CORRELATION

Correlation analysis helps us in determining the degree of relationship between 2 or more variables. The value of the coefficient of correlation as obtained by the below formula shall always lie between +1 and -1. When r = +1, it means there is perfect positive correlation between the variables. When r = -1, there is perfect negative correlation between the variables and when r = 0, there is no relationship between the two variables.

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \cdot \sum y^2}}$$

$$x = (X - \overline{X}) \quad ; \quad y = (Y - \overline{Y})$$

3. PERCENTAGE ANALYSIS

Percentage analysis shows the entire population in terms of percentages.

$$\text{Percentage} = \frac{\text{No. of respondents}}{\text{Total respondents}} * 100$$

4. GRAPHS

Graphical method was used in order to represent the factor in various graphical methods like pie-chart, bar diagram and cylinder.

5. ANOVA

The one-way Analysis of Variance (ANOVA) is used with one categorical independent variable and one continuous variable. The independent variable can consist of any number of groups (levels).

Note- All these tools are available as functions in R.

CHAPTER – 5

DATA PREPARATION

The data used in this study is secondary in nature, that is, data was not collected for the purpose of this study but what was collected over their periodic recruitment drives.

Below is the snapshot of the 'raw' data received from MIST*.

MALWA GROUP OF INSTITUTIONS

List of selected candidates for CS-IT Engineering Depart

S.No.	Qualification & Expereience	Presently working at	Current Salary	Expected Salary
1	B.E.- 69%,M.Tech.- R.A. **Work Exp-7.10 yrs**	SanghVi Innovative Academy	25,600	Same
2	B.E.- 70%, M.E.-76% Ph.D Pursuing **Work Exp 2 years**	Last working with T..I.T bhopal till November 2013	Last Drawn Salary 25 K	Expected 25 K + Accomodation or Rs.35 K
3	M.E.- 67.2%, M.C.A.- 63% **Work Exp-11.6 yrs**	Till June 2014 - Oriental University	Last Drawn Salary 45K	Same
4	B.E.63% M.Tech.-66% **Work Exp-1.6 years**	Till Aug - 2009 GSBA - Roorkee	Rs26,000	As per norms
5	B.E. / M.Tech.-72% **Work Exp-6.11 yrs**	BM College of Technology	Rs.24,000	Rs.22,000
6	B.E.-75.75%, M.E.-74.93%, GATE Qualified **Work Exp-4.7 yrs**	Till Jan-14, Prestige Instt of Engg & research	Rs.20,500	Same
7	B.E.- 65%, M.E.-72.43% **Work Exp-6.5 yrs**	At LNCT as Asst Professor	Rs 25,000	Rs. 25,000
8	B.E.-69.56%, M.E.-7.71 **Work Exp-1 year**	Last Working with Prestige Instt of engg. & Science	Rs.18,600	Rs.20,000 to Rs.22,000
9	B.E.-69.94%, M.E.-70.9% **Work Exp-9 years**	Till Sept-14 with Vindhya Instt of Technology	Rs.24,984	Min Rs, 21,600
10	B.E,-66.07%, M.E.-72% **Work Exp 6.9 yrs**	Last Prof Ram Meghe Instt. Of Technology & Research, Amravati	Rs.45420	Rs.21,600

11	B.E.-74.49%, M.Tech.-82.5 **Work Exp -4.8 yrs**	SanghVi Institute of Management and Science	Rs19220	Rs.25,000
12	B.E.- 62.3%, M.Tech.-7.84- **Fresher**	Fresher	NA	As per norms
13	B.E.- 62%, M.E.- 65.5% **Work Exp-6 yrs**	Last O.P. Jindal Institute of Technology , Raigarh, CG	Rs 16,000	Rs21,600
14	B.E.-76.69%, M.E.-7.81 **, Fresher**	Fresher	NA	Rs.16,000 to Rs.18,000
15	B.E.-66.91%, M.Tech-75.80 Ph.D Pursuing **Work Exp-3.7 years**	Last Working Orienetal University Left for PhD.	Rs. 20,000	Rs. 17,000 to Rs.20,000

Recommendation	**Offer Made**	**joined**
V. Good Additional Ability CRT.	yes	yes
ME from Greenvich University London, Brings in Fresh outlook. Best Technical Feedback	yes	yes
Not eligible Cannot Be taken as not a BE	no too expensive	no
Lots of GAP in Career	no	no
Not very clear on Objective	no	no
	no	no
Confused and not knowing why he wants to switch job	no	no
Could be considered if We go for only male candiodates	no	no
	no	no

Qualified & Experienced still less expectation	yes	no
	no	no
Promising Fresher with Less Expectation hence Recommended	yes	yes
	no	no
	no	no
	no	no

*Note: Name column has been removed to protect identities.

From the data, it is clear that there are several columns that have been merged together which need to be separated out. Qualification and Experience, for instance need to be separated into separate columns as each of them can represent a separate factor in the joining of candidates. Along with this, there is other information that needs to be pulled out in separate columns. Whether or not a candidate is pursuing Phd, is he/she GATE qualified or whether she is from same city as the institute. All this information needs to be pulled into separate columns. All this is done in excel itself and a new sheet which is more friendly for analysis is prepared.

A snapshot of the cleaned sheet is shown below.

gender	hqual	hpercent	gate	pPhd	exp	working	org	same_city
M	BE	69	no	no	7.1	yes	SanghVi Innovative Academy	yes
M	ME	76	no	yes	2	no	TIT, Bhopal	no
F	ME	67.2	no	no	11.6	no	Oriental University	yes
F	MTech	66	no	no	1.6	no	GSBA, Roorkee	no
F	MTech	72	no	no	7	yes	BM College of Technology	yes
M	ME	75	yes	no	4.7	no	Prestige Instt	yes
M	ME	72.5	no	no	6.5	yes	LNCT	yes

np	dept	csal	esal	offered	remarks	osal	joined
0	CSIT	25600	25600	yes	V. Good Additional Ability CRT.	25600	yes
0	CSIT	25000	35000	yes	ME from Greenvich University London, Brings in Fresh outlook. Best Technical Feedback	35000	yes

0	CSIT	45000	45000	no	Not eligible Cannot Be taken as not a BE	NA	no
0	CSIT	26000	26000	no	Lots of GAP in Career	NA	no
1	CSIT	24000	22000	no	Not very clear on Objective	NA	no
0	CSIT	20500	20500	no		NA	no
1	CSIT	25000	25000	no	Confused and not knowing why he wants to switch job	NA	no

Data Dictionary- At this point, a data dictionary describing the various variables in the cleaned sheet needs to be defined.

Below are the various variables used in the sheet are as follows:

1. gender: Gender of the candidate whether Male or Female. It can take values 'M' or 'F'.

2.hqual: Highest qualification of the candidate. It can take various values like BE, Btech, ME, Mtech, Mcom etc.

3.hpercent: Percentage obtained in the highest qualification.

4.gate: Whether a candidate has qualified GATE exam or not.

5.pPhd: Whether a candidate is pursuing Phd. or not.

6.exp: Number of years of experience of the candidate.

7.working: Whether a candidate is currently working or not.

8.org: Current organization of the candidate. None if she is not working.

9.same_city: Whether a candidate is from the same city as the Institute is located in.

10.np: Notice period of the candidate in months. Usually 0 if the candidate is not

currently working.

11.dept: The department for which the candidate has applied like Mechanical, Civil etc.

12.csal: Current salary of the candidate in Rs/month. Last drawn salary in case of experienced candidates not currently working. 0 in case of freshers.

13.esal: Candidate's expectation of the salary from the job if selected in Rs/month.

14.offered: Whether the candidate was offered the job or not.

15.remarks: Remarks from the interview panel on the candidate.

16.osal: Salary offered to the selected candidates in Rs/month.

17.joined: Whether the candidate finally joined or not.

CHAPTER – 6

DATA ANALYSIS AND INTERPRETATION

After tidying the data, the file containing the recruitment data is loaded into R environment.

```
#Load the recruitment dataset.
recruit <- read_excel("Recruit1.xlsx", 1)

#First look at data in R
str(recruit)

Classes ‘tbl_df’, ‘tbl’ and 'data.frame':     49 obs. of  18 variables:
 $ name     : chr  "Sunny Bagga" "Siddhartha Saha" "Archana Sharma" "Alka Sharma"
...
 $ gender   : chr  "M" "M" "F" "F" ...
 $ hqual    : chr  "BE" "ME" "ME" "MTech" ...
 $ hpercent : num  69 76 67.2 66 72 75 72.5 77.1 71 72 ...
 $ gate     : chr  "no" "no" "no" "no" ...
 $ pPhd     : chr  "no" "yes" "no" "no" ...
 $ exp      : num  7.1 2 11.6 1.6 7 4.7 6.5 1 9 7 ...
 $ working  : chr  "yes" "no" "no" "no" ...
 $ org      : chr  "SanghVi Innovative Academy" "TIT, Bhopal" "Oriental University"
"GSBA,  Roorkee" ...
 $ same_city: chr  "yes" "no" "yes" "no" ...
 $ np       : num  0 0 0 0 1 0 1 0 0 0 ...
 $ dept     : chr  "CSIT" "CSIT" "CSIT" "CSIT" ...
 $ csal     : chr  "25600" "25000" "45000" "26000" ...
 $ esal     : chr  "25600" "35000" "45000" "26000" ...
 $ offered  : chr  "yes" "yes" "no" "no" ...
 $ remarks  : chr  "V. Good  Additional Ability CRT." "ME from Greenvich University
London, Brings in Fresh outlook. Best Technical Feedback" "Not eligible Cannot Be
taken as not a BE" "Lots of GAP in Career" ...
 $ osal     : chr  "25600" "35000" "NA" "NA" ...
 $ joined   : chr  "yes" "yes" "no" "no" ...


summary(recruit)
```

```
   name              gender             hqual              hpercent
 Length:49          Length:49          Length:49          Min.   :55.00
 Class :character   Class :character   Class :character   1st Qu.:68.75
 Mode  :character   Mode  :character   Mode  :character   Median :72.00
                                                          Mean   :72.20
                                                          3rd Qu.:76.35
                                                          Max.   :83.00
                                                          NA's   :2
    gate               pPhd               exp           working
 Length:49          Length:49          Min.   : 0.000   Length:49
 Class :character   Class :character   1st Qu.: 1.000   Class :character
 Mode  :character   Mode  :character   Median : 2.000   Mode  :character
                                       Mean   : 4.284
                                       3rd Qu.: 6.500
                                       Max.   :31.000

    org             same_city             np            dept
 Length:49          Length:49          Min.   :0.000   Length:49
 Class :character   Class :character   1st Qu.:0.000   Class :character
 Mode  :character   Mode  :character   Median :0.000   Mode  :character
                                       Mean   :0.375
                                       3rd Qu.:1.000
                                       Max.   :1.000
                                       NA's   :1
    csal               esal             offered            remarks
 Length:49          Length:49          Length:49          Length:49
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character
```

Since there are some text values that have been read in as “NA”, those will have to be replaced with R’s built-in NA values.

```
# replace text "NA" with R's NA
```

```
for(x in 1:nrow(recruit)){
  for(y in 1:ncol(recruit)){
    if(!is.na(recruit[x,y])){
      if(recruit[x,y] == "NA"){
        recruit[x,y] <- NA
      }
    }
  }
}
```

The categorical variables read in as text will have to be converted to factor variables to aid in analysis.

```
#Converting categorical variables red in as text to factor variables
recruit$gender <- as.factor(recruit$gender)
recruit$hqual <- as.factor(recruit$hqual)
recruit$gate <- as.factor(recruit$gate)
recruit$pPhd <- as.factor(recruit$pPhd)
recruit$working <- as.factor(recruit$working)
recruit$org <- as.factor(recruit$org)
recruit$same_city <- as.factor(recruit$same_city)
recruit$np <- as.factor(recruit$np)
recruit$dept <- as.factor(recruit$dept)
recruit$offered <- as.factor(recruit$offered)
recruit$joined <- as.factor(recruit$joined)
recruit$csal <- as.numeric(recruit$csal)
recruit$esal <- as.numeric(recruit$esal)
recruit$osal <- as.numeric(recruit$osal)
```

Removing the name column.

```
#Removing the name column as that won't be needed for analysis
recruit <- subset(recruit, select = -c(name))
```

Relooking at data

```
#looking at the data again
str(recruit)
```

```
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':    49 obs. of  17 variables:
 $ gender   : Factor w/ 2 levels "F","M": 2 2 1 1 1 2 2 2 2 1 ...
 $ hqual    : Factor w/ 6 levels "BCom","BE","MCom",..: 2 4 4 5 5 4 4 4 4 4 ...
 $ hpercent : num  69 76 67.2 66 72 75 72.5 77.1 71 72 ...
 $ gate     : Factor w/ 2 levels "no","yes": 1 1 1 1 1 2 1 1 1 1 ...
 $ pPhd     : Factor w/ 2 levels "no","yes": 1 2 1 1 1 1 1 1 1 1 ...
 $ exp      : num  7.1 2 11.6 1.6 7 4.7 6.5 1 9 7 ...
 $ working  : Factor w/ 2 levels "no","yes": 2 1 1 1 2 1 2 1 1 1 ...
 $ org      : Factor w/ 28 levels "Accounts Manager",..: 20 24 15 6 3 16 11 16 27 17
...
 $ same_city: Factor w/ 2 levels "no","yes": 2 1 2 1 2 2 2 2 2 1 ...
 $ np       : Factor w/ 2 levels "0","1": 1 1 1 1 2 1 2 1 1 1 ...
 $ dept     : Factor w/ 5 levels "Accounting","Civil",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ csal     : num  25600 25000 45000 26000 24000 ...
 $ esal     : num  25600 35000 45000 26000 22000 20500 25000 20000 21600 21600 ...
 $ offered  : Factor w/ 2 levels "no","yes": 2 2 1 1 1 1 1 1 1 2 ...
 $ remarks  : chr  "V. Good  Additional Ability CRT." "ME from Greenvich University
London, Brings in Fresh outlook. Best Technical Feedback" "Not eligible Cannot Be
taken as not a BE" "Lots of GAP in Career" ...
 $ osal     : num  25600 35000 NA NA NA NA NA NA NA 21600 ...
 $ joined   : Factor w/ 2 levels "no","yes": 2 2 1 1 1 1 1 1 1 1 ...
```

```
summary(recruit)
```

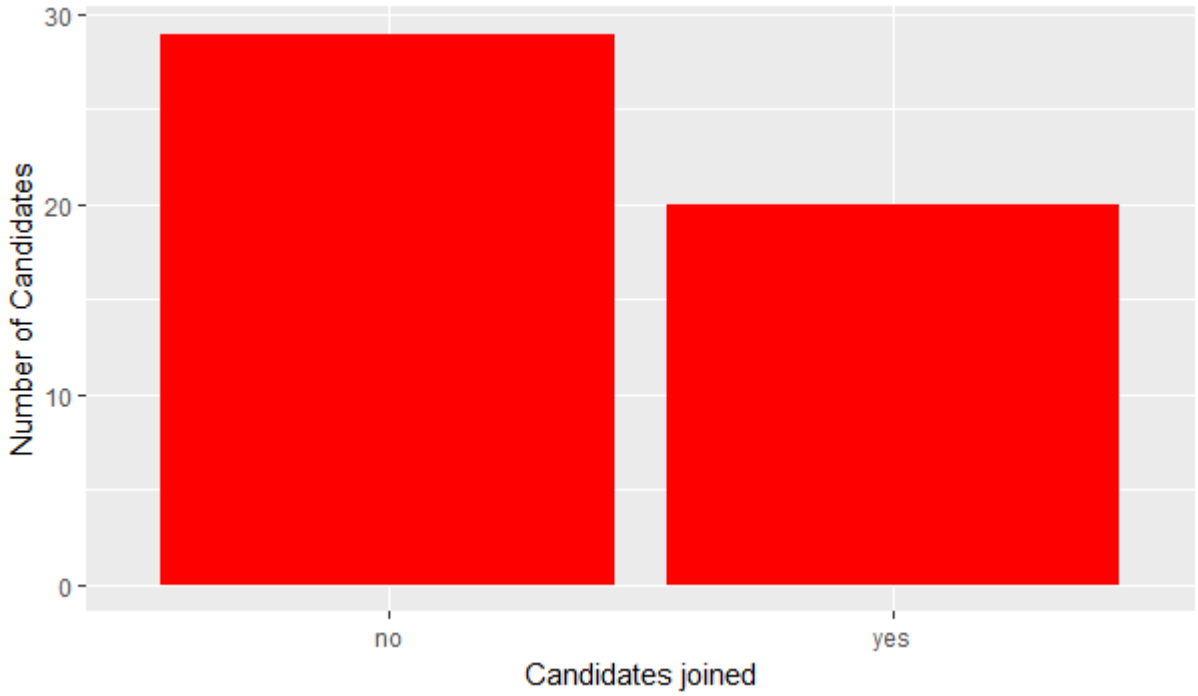
```
gender   hqual       hpercent      gate     pPhd         exp          working
 F:13   BCom : 1   Min.   :55.00   no :43   no :44   Min.   : 0.000   no :27
 M:36   BE   : 6   1st Qu.:68.75   yes: 6   yes: 5   1st Qu.: 1.000   yes:22
        MCom : 2   Median :72.00                     Median : 2.000
        ME   :25   Mean   :72.20                     Mean   : 4.284
        MTech:14   3rd Qu.:76.35                     3rd Qu.: 6.500
```

```
       Phd  : 1   Max.   :83.00                    Max.   :31.000
                  NA's   :2
                                      org     same_city    np              dept
 none                                   :10   no : 5    0   :30   Accounting : 4
 Oriental University                    : 4   yes:44    1   :18   Civil      : 9
 Prestige Instt                         : 4             NA's: 1   CSIT       :19
 Indore Institute of Sc. And Technology: 3                       Electronics: 2
 LNCT Indore                            : 2                      Mechanical :15
 (Other)                                :24
 NA's                                   : 2
     csal            esal       offered    remarks               osal
 Min.   : 7500   Min.   :10000   no :22   Length:49          Min.   :10000
 1st Qu.:18000   1st Qu.:20000   yes:27   Class :character   1st Qu.:15600
 Median :24000   Median :22500            Mode  :character   Median :21600
 Mean   :25594   Mean   :24688                               Mean   :22259
 3rd Qu.:30000   3rd Qu.:26500                               3rd Qu.:25000
 Max.   :70000   Max.   :50000                               Max.   :50000
 NA's   :12      NA's   :1                                   NA's   :20
 joined
 no :29
 yes:20
```

Data Visualization:

Here the data will be visualized to explore the relationships between dependent and independent variables.

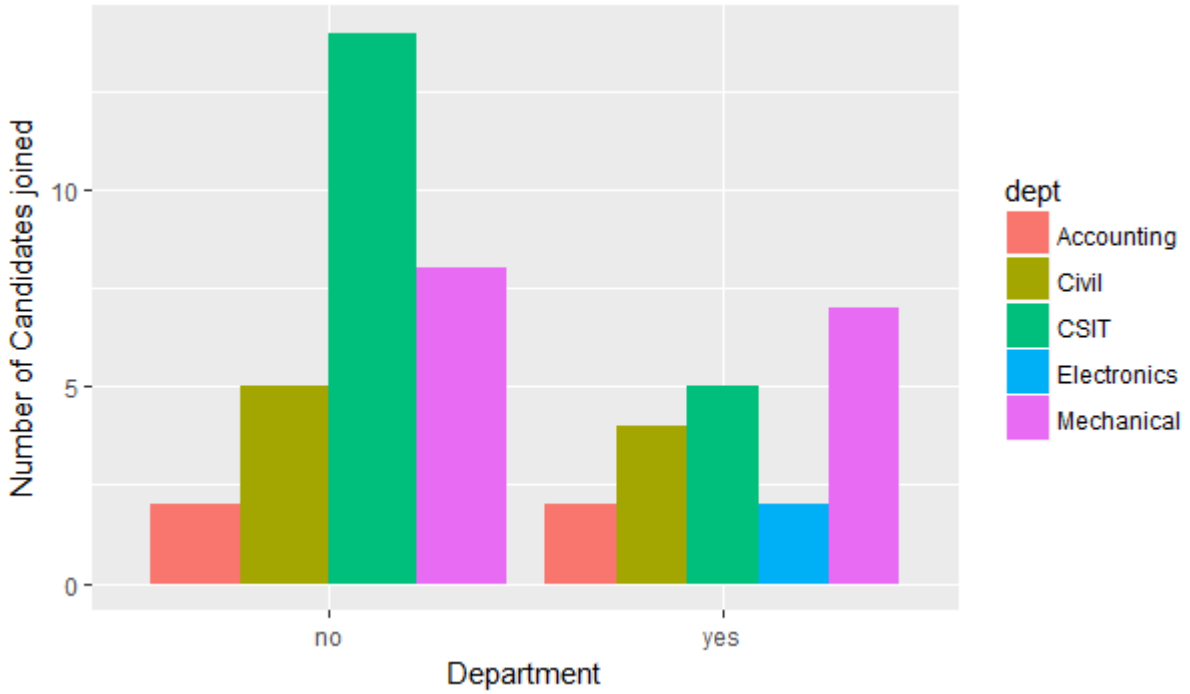
```
library(ggplot2)
#Plot of actual numbers who joined
ggplot(recruit, aes(x = joined)) + geom_bar(fill = "red") + xlab("Candidates
joined") + ylab("Number of Candidates")
```



Percentage of candidates joining = 20/49 = 40.82%

Next, the number of candidates joining by each department is plotted and see if there is any variation in it.

```
#Plot of actual numbers who joined by each department
ggplot(recruit, aes( x = joined, fill = dept)) + geom_bar(position = "dodge") +
xlab("Department") + ylab("Number of Candidates joined")
```



To determine whether the candidates joining different departments are different or the result observed is due to chance, a chi-square test is run.

```
#Running chi squre tests
chisq.test(recruit$joined, recruit$dept)
```

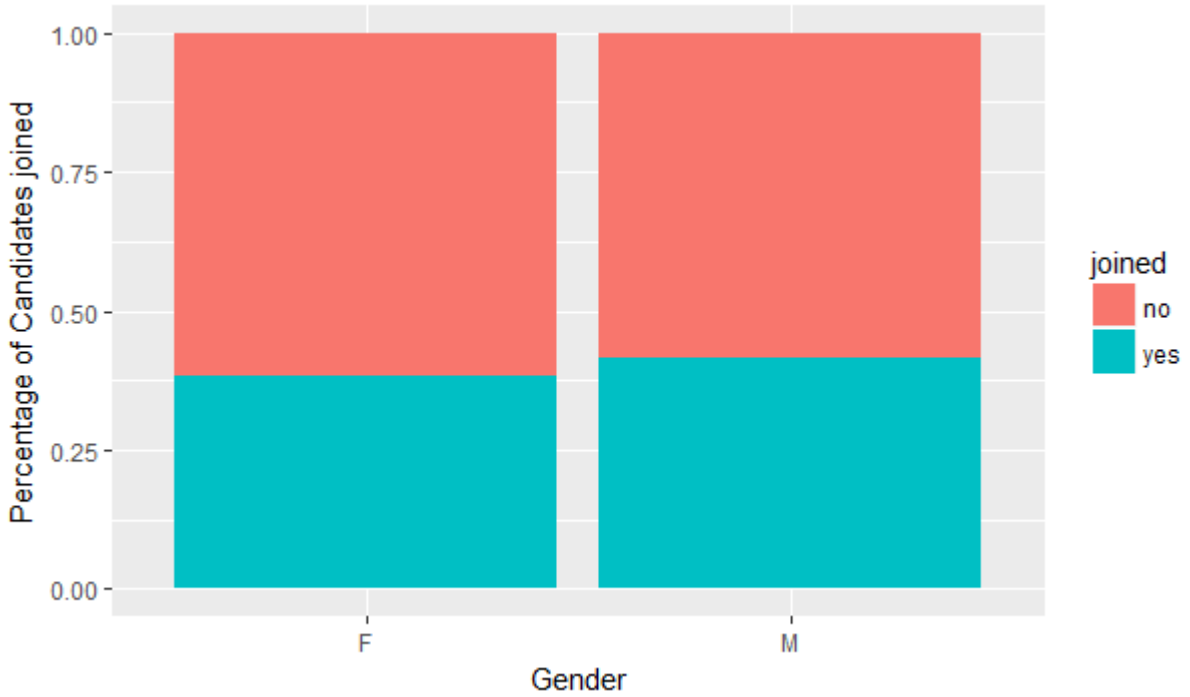
```
Pearson's Chi-squared test

data:  recruit$joined and recruit$dept
X-squared = 4.955, df = 4, p-value = 0.2919
```

With observed p-value of 0.2919, we can be sure that the observed differences are due to random chance at 95% confidence level.

Similarly, we explore the relationships between other independent variables and the dependent variable.

```
#Percentage of people joined by each department
ggplot(recruit, aes( x = dept, fill = joined)) + geom_bar(position = "fill") +
xlab("Department") + ylab("Percentage of Candidates joined")
```


```
#Percentage of people joined by gender
ggplot(recruit, aes( x = gender, fill = joined)) + geom_bar(position = "fill") +
xlab("Gender") + ylab("Percentage of Candidates joined")
```



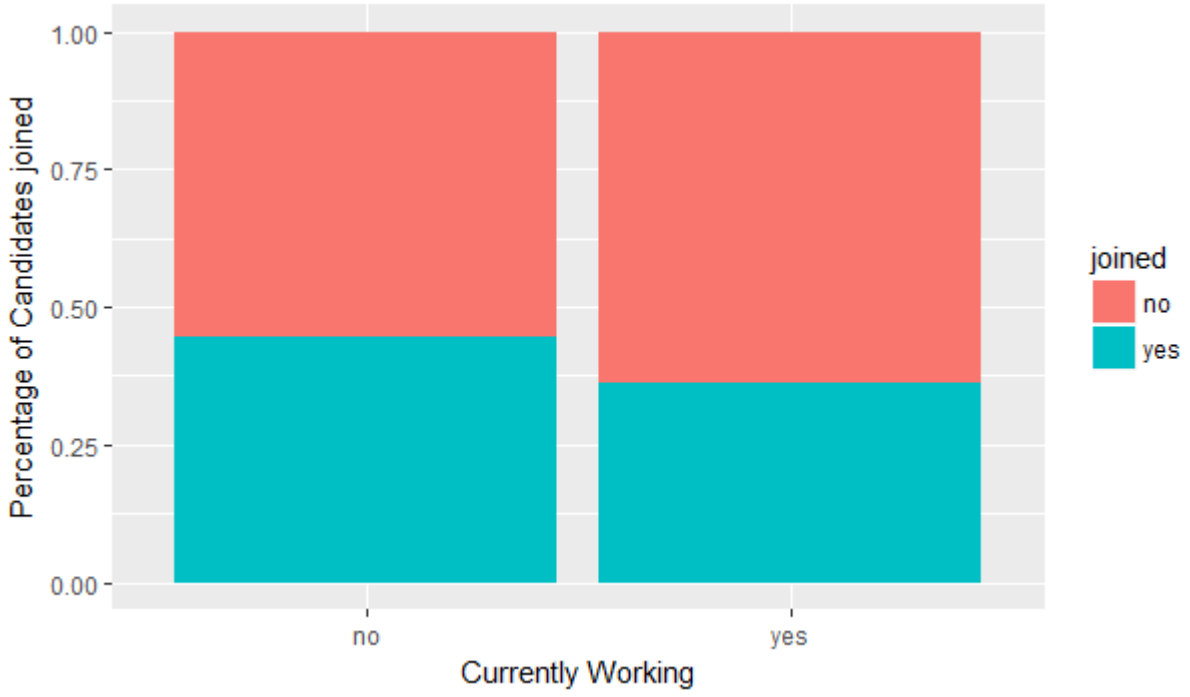
Visually, there seems to be no difference between the percentage of candidates joining by gender.

```
#Running chi squre tests
chisq.test(recruit$joined, recruit$gender)
```

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

data:  recruit$joined and recruit$gender
X-squared = 1.481e-31, df = 1, p-value = 1
```

Running chi-square test confirms what is seen visually. With a p-value of 1, the minute observed difference is entirely due to random chance.

```
#Percentage of people joined by working or not
ggplot(recruit, aes( x = working, fill = joined)) + geom_bar(position = "fill") +
xlab("Currently Working") + ylab("Percentage of Candidates joined")
```




```
#Running chi squre tests
chisq.test(recruit$joined, recruit$gender)
```

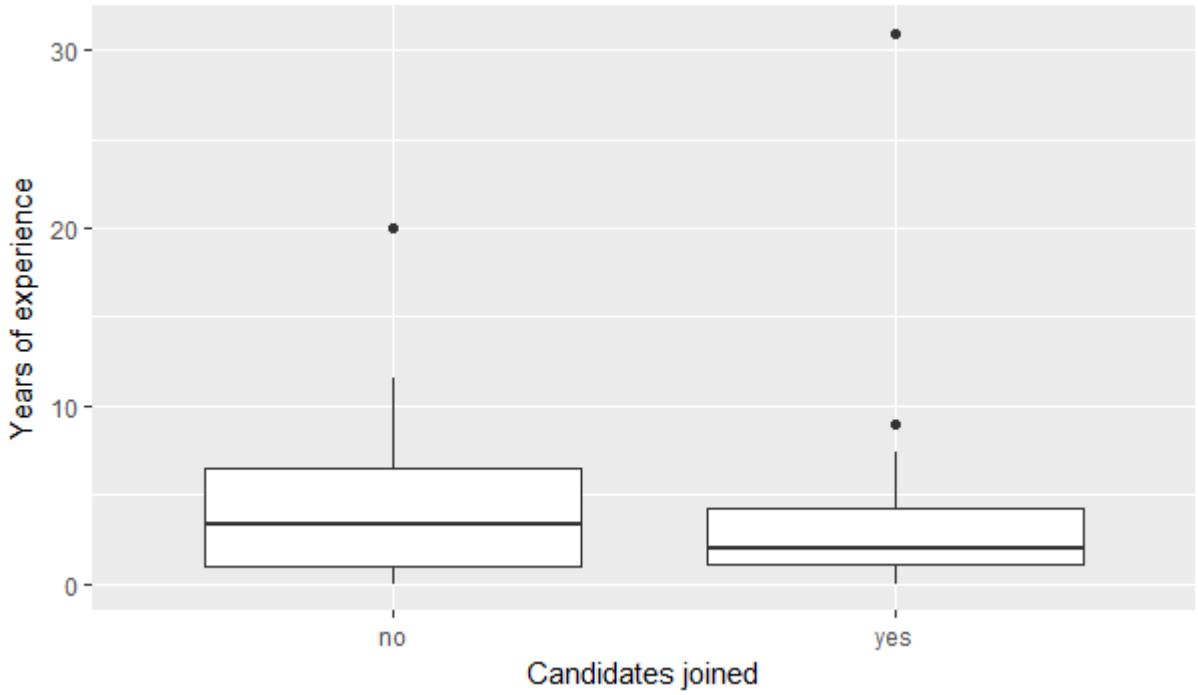
```
        Pearson's Chi-squared test with Yates' continuity correction
```

```
data:  recruit$joined and recruit$working
X-squared = 0.078545, df = 1, p-value = 0.7793
```

With p-value of 0.7793, the observed differences in the number of candidates joining by their current work status is not statistically significant at 95% confidence level.

#Number of people joined by years of experience

ggplot(recruit, aes(x = joined, y = exp)) + geom_boxplot() () + xlab("Candidates joined") + ylab("Years of experience")

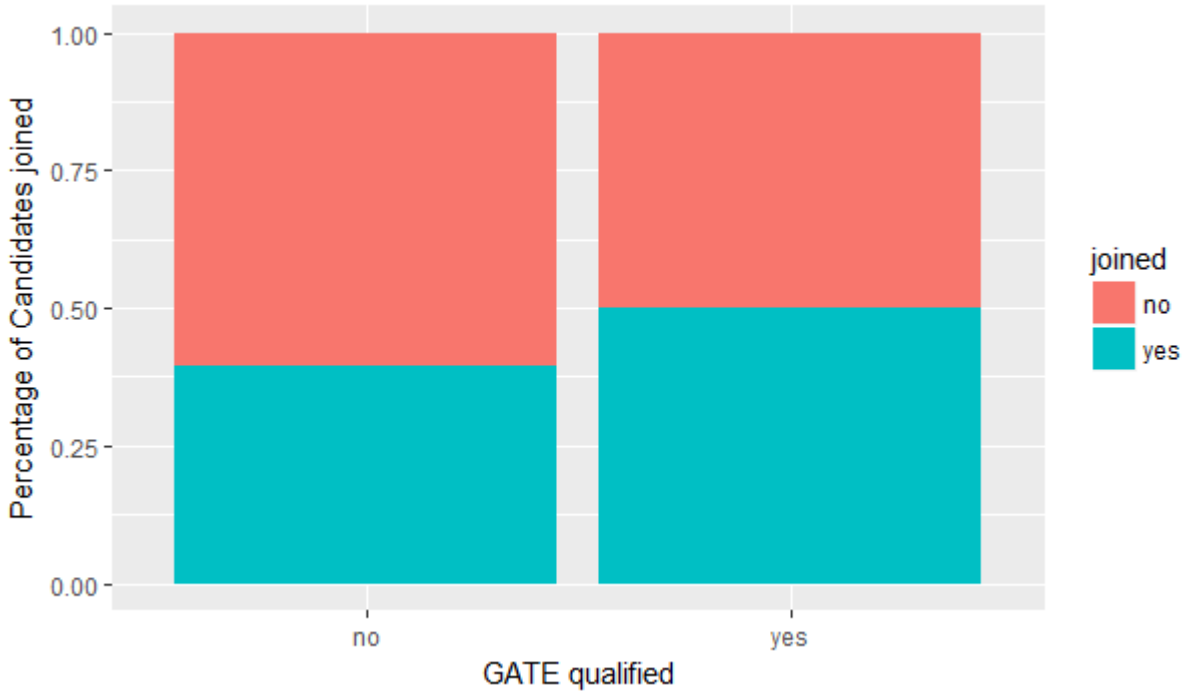



```
#Running one-way anova
exp.aov <- aov(recruit$exp ~ recruit$joined)
summary(exp.aov)

                Df Sum Sq Mean Sq F value Pr(>F)
recruit$joined  1    1.7   1.691   0.054  0.818
Residuals      47 1478.5  31.458
```

Since the p-value is high at 0.818, the observed differences in means of years of experiences of candidates joining is entirely due to chance at 95% confidence level.

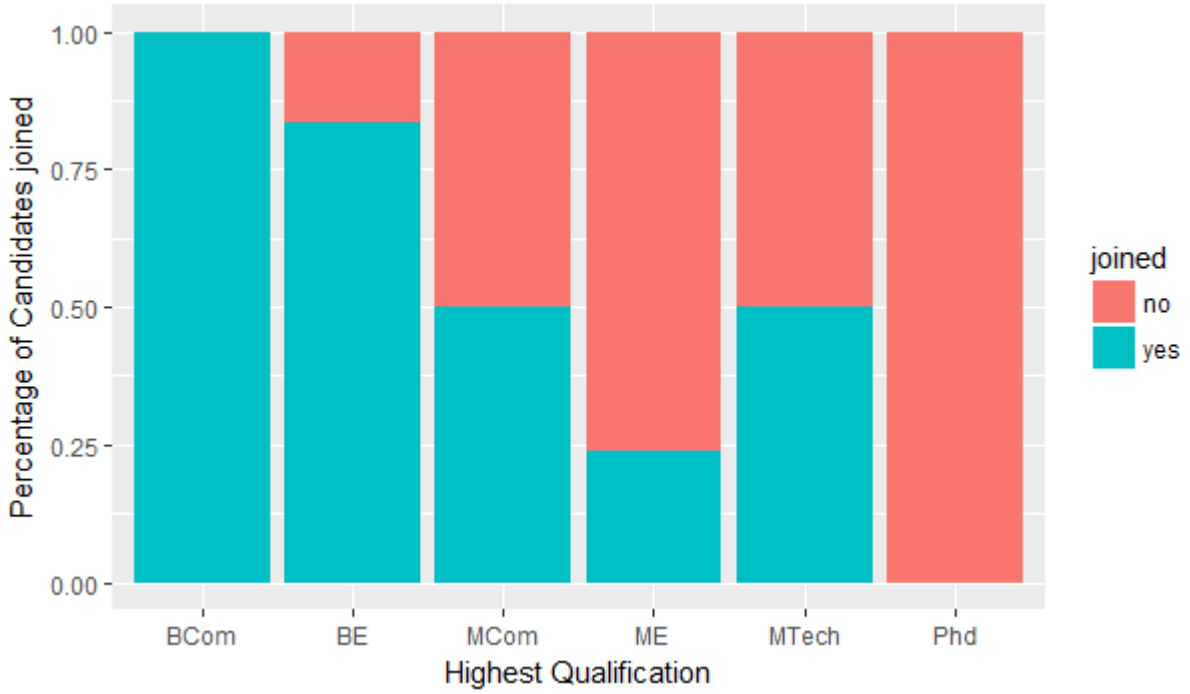
```
#Percentage of people joined by GATE qualified
ggplot(recruit, aes( x = gate, fill = joined)) + geom_bar(position = "fill") +
xlab("GATE qualified") + ylab("Percentage of Candidates joined")

#Running chi squre test
chisq.test(recruit$joined, recruit$gate)

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

data:  recruit$joined and recruit$gate
X-squared = 0.0020466, df = 1, p-value = 0.9639
```

Since p-value is high at 0.9639, the observed differences in the number of candidates joining because of GATE qualified is entirely due to random chance at 95% confidence level.




```
#Percentage of people joined by Highest qualification
```

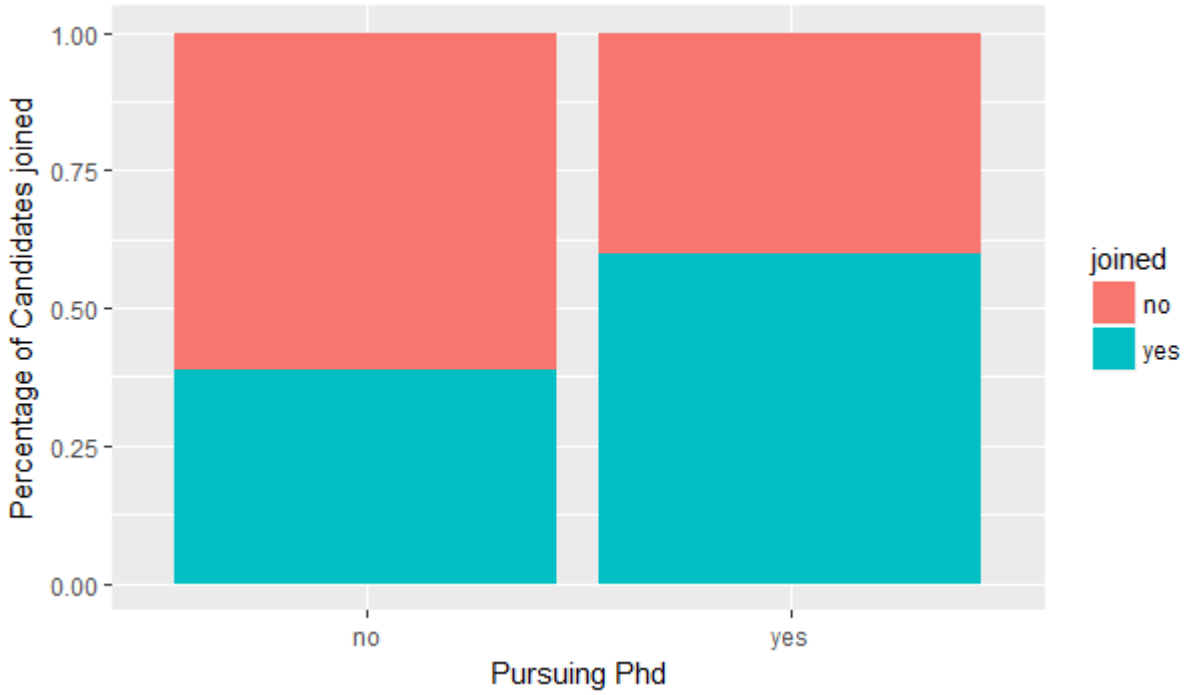
```
ggplot(recruit, aes( x = hqual, fill = joined)) + geom_bar(position = "fill") +
xlab("Highest Qualification") + ylab("Percentage of Candidates joined")
```




```
#Running chi squre test
chisq.test(recruit$joined, recruit$hqual)

        Pearson's Chi-squared test

data:  recruit$joined and recruit$hqual
X-squared = 10.115, df = 5, p-value = 0.07205
```

Since p-value is 0.07205, the observed differences in the number of candidates joining because of Highest qualification is entirely due to random chance at 95% confidence level but is significant at 90% confidence level.

```
#Percentage of people joined by pursuing Phd
ggplot(recruit, aes( x = pPhd, fill = joined)) + geom_bar(position = "fill") +
xlab("Pursuing Phd") + ylab("Percentage of Candidates joined")
```

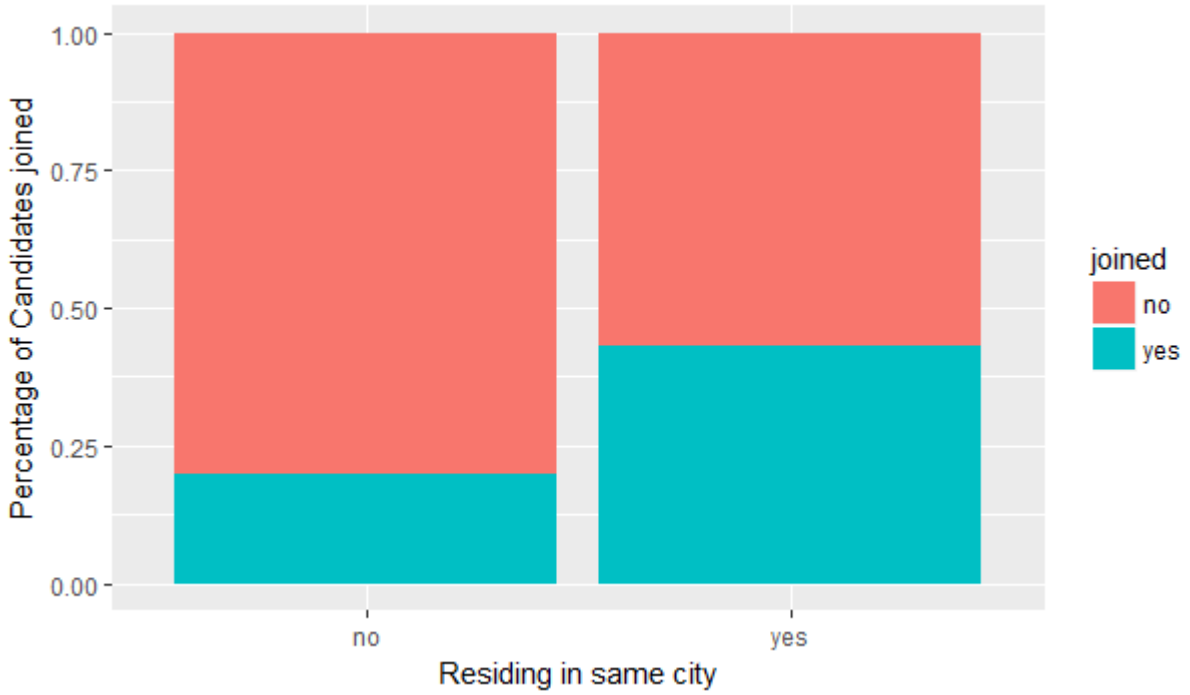



```
#Running chi squre test
chisq.test(recruit$joined, recruit$pPhd)
```

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

data:  recruit$joined and recruit$pPhd
X-squared = 0.19441, df = 1, p-value = 0.6593
```

Since p-value is 0.6593, the observed differences in the number of candidates joining because of pursuing Phd is entirely due to random chance at 95% confidence level.

```
#Percentage of people joined by same_city
ggplot(recruit, aes( x = same_city, fill = joined)) + geom_bar(position = "fill") +
xlab("Residing in same city") + ylab("Percentage of Candidates joined")
```

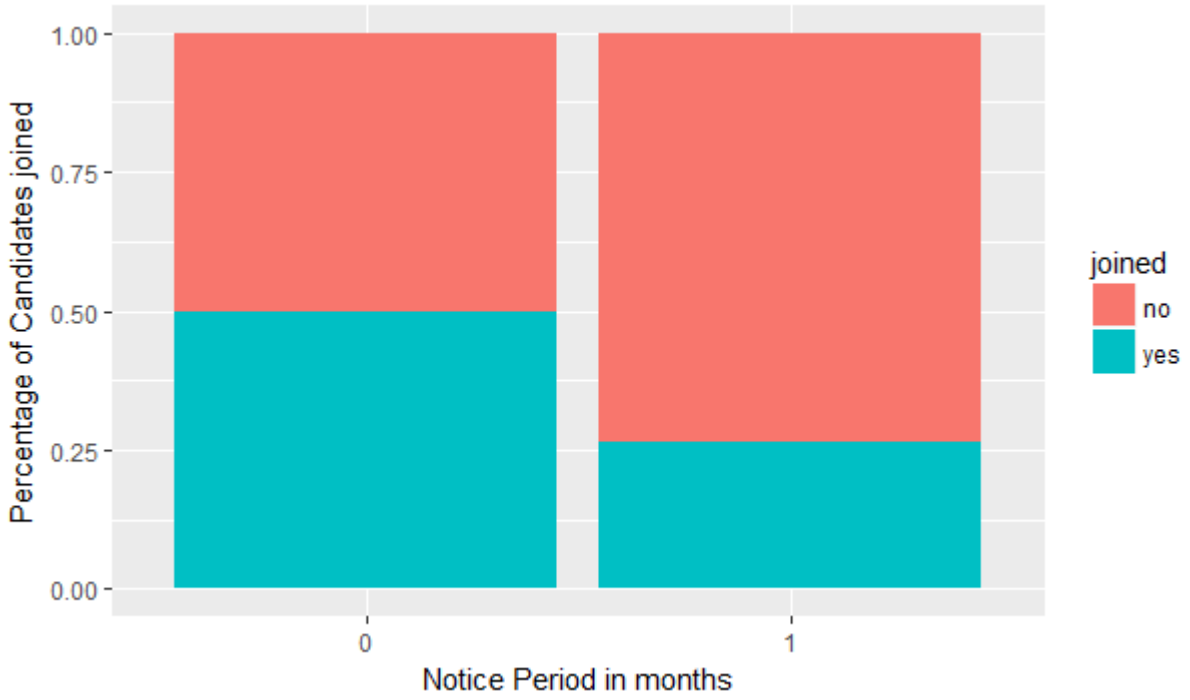



```
#Running chi squre test
chisq.test(recruit$joined, recruit$same_city)
```

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

data:  recruit$joined and recruit$same_city
X-squared = 0.26967, df = 1, p-value = 0.6036
```

Since p-value is 0.6593, the observed differences in the number of candidates joining because of same_city is entirely due to random chance at 95% confidence level.

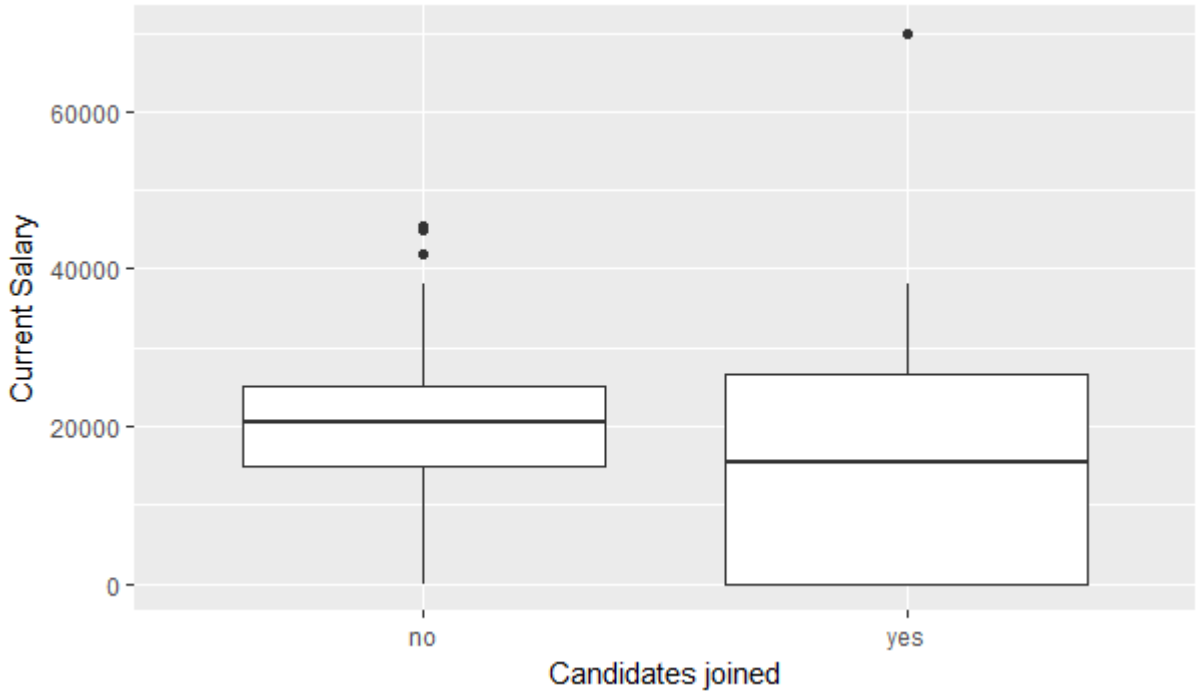
```
#Percentage of people joined by notice period
ggplot(recruit, aes( x = np, fill = joined)) + geom_bar(position = "fill") +
xlab("Notice Period in months") + ylab("Percentage of Candidates joined")
```




```
#Running chi squre test
chisq.test(recruit$joined, recruit$np)

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

data:  recruit$joined and recruit$np
X-squared = 1.8097, df = 1, p-value = 0.1785
```

Since p-value is 0.1785, the observed differences in the number of candidates joining because of notice period is entirely due to random chance at 95% confidence level.

```
#Plot of people joining by current salary
ggplot(recruit, aes( x = joined, y = csal)) + geom_boxplot() + xlab("Candidates
joined") + ylab("Current Salary")
```

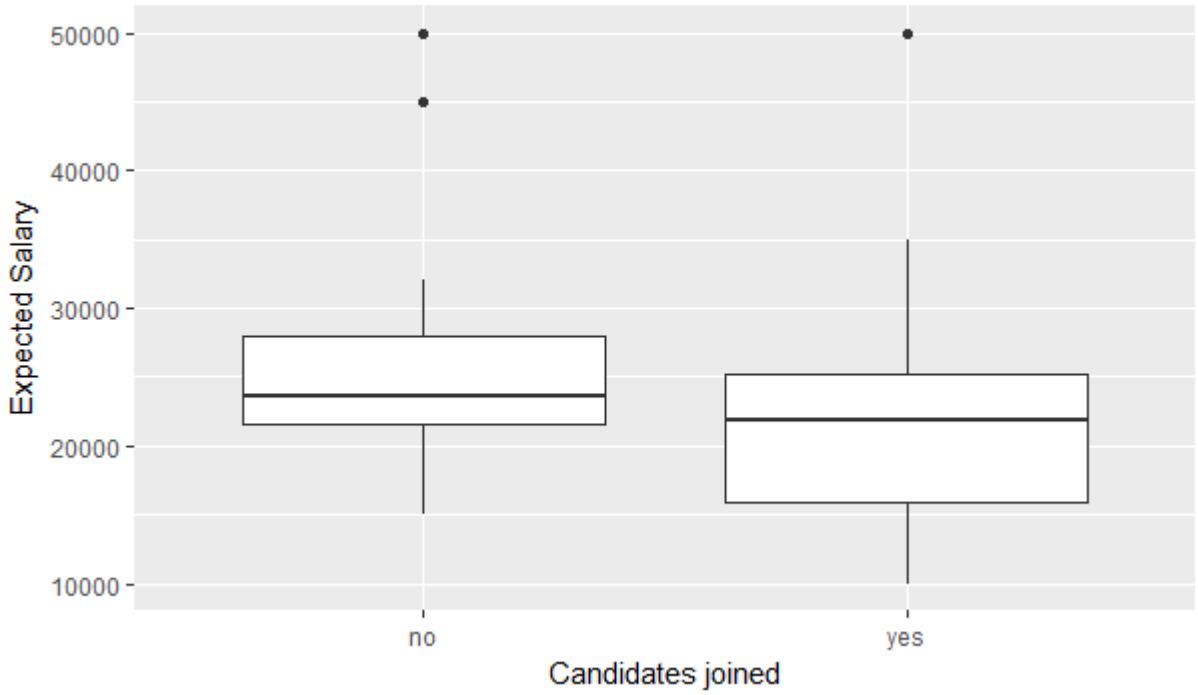



```
#Running one-way anova
csal.aov <- aov(recruit$csal ~ recruit$joined)
summary(csal.aov)
```

```
               Df    Sum Sq   Mean Sq F value Pr(>F)
recruit$joined  1 4.255e+07  42554288   0.188  0.667
Residuals      47 1.066e+10 226723082
```

Since p-value is 0.667, the observed differences in the means of current salary because of candidates joining is entirely due to random chance at 95% confidence level.

```
#Plot of people joining by expected salary
ggplot(recruit, aes( x = joined, y = esal)) + geom_boxplot() + xlab("Candidates
joined") + ylab("Expected Salary")
```




```
#Running one-way anova
esal.aov <- aov(recruit$csal ~ recruit$joined)
summary(esal.aov)
```

```
               Df    Sum Sq   Mean Sq F value Pr(>F)
recruit$joined  1 4.255e+07  42554288   0.188  0.667
Residuals      47 1.066e+10 226723082
```

Since p-value is 0.667, the observed differences in the means of current salary because of candidates joining is entirely due to random chance at 95% confidence level.

Correlation between Current salary and Expected Salary

```
#Correlation between current salary and expected salary
cor(recruit$csal, recruit$esal, use = "pairwise.complete.obs")
```

```
[1] 0.7098949
```

Correlation between current salary and Expected salary is high which is expected.

Having explored the relationships between the dependent variable and the independent variables, we will move to the modeling section.

CHAPTER – 7

MODEL BUILDING

Baseline Accuracy

Baseline accuracy of a model is simply the percentage of the most frequently occurring value of the dependent variable. So, the baseline accuracy of the given dataset will be

```
table(recruit$joined)
no yes
29  20

#baseline accuracy
29/49
[1] 0.5918367
```

Hence, baseline accuracy is 59.18%

Next step is to remove unnecessary variables not required for analysis. These would be organization, remarks, Offered and Salary offered.

```
#Remove unnecessary variables
recruit1 <- subset(recruit, select = -c(org, offered, remarks, osal))
```

The whole of data will not be used for training the model. Instead, the dataset will be split into a training set and testing set. 70% of the data will be used for training and the rest for testing.

```
library(caTools)
# Randomly split data
set.seed(2017)
split = sample.split(recruit1$joined, SplitRatio = 0.7)
split
```

After creating the split vector, the training and testing sets need to be created.

```
# Create training and testing sets
train = subset(recruit1, split == TRUE)
```

```
test = subset(recruit1, split == FALSE)
```

Now run the command for generating logistic regression model.

After checking various combinations of independent variables, the following combination of Independent variables seems to give the best result. The testing and evaluation of the model will be explained in a later chapter.

```
# Logistic Regression Model
logmodel = glm(joined ~  hqual + gate + pPhd +exp +working + same_city + np + dept +
csal , data=train, family=binomial)
summary(logmodel)

Call:
glm(formula = joined ~ hqual + gate + pPhd + exp + working +
    same_city + np + dept + csal, family = binomial, data = train)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.72960  -0.36558  -0.00951   0.20620   1.86562

Coefficients: (1 not defined because of singularities)
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      1.305e+01  7.537e+00   1.732   0.0833 .
hqualMCom       -2.069e+01  1.075e+04  -0.002   0.9985
hqualME         -2.984e+00  2.968e+00  -1.005   0.3147
hqualMTech       1.598e+00  2.968e+00   0.538   0.5904
hqualPhd        -2.785e+01  1.075e+04  -0.003   0.9979
gateyes         -4.152e+00  3.902e+00  -1.064   0.2873
pPhdyes          2.052e+00  2.294e+00   0.894   0.3711
exp              7.505e-01  5.415e-01   1.386   0.1658
workingyes       2.504e+01  5.687e+03   0.004   0.9965
same_cityyes    -4.825e+00  3.061e+00  -1.576   0.1149
```

```
np1             -2.875e+01  5.687e+03  -0.005   0.9960
deptCivil       -7.533e+00  5.158e+00  -1.460   0.1442
deptCSIT        -6.381e+00  3.671e+00  -1.739   0.0821 .
deptElectronics  1.831e+01  7.593e+03   0.002   0.9981
deptMechanical          NA         NA      NA       NA
csal            -3.514e-04  1.961e-04  -1.792   0.0732 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 46.070  on 33  degrees of freedom
Residual deviance: 18.766  on 19  degrees of freedom
AIC: 48.766

Number of Fisher Scoring iterations: 18
```

From the model, we can make out that variables deptCSIT representing department = CSIT is significant at 90% level of confidence. Also, the variable csal representing Current Salary is significant at 90% confidence level.

Now we will build a Decision Tree model using the same training and testing sets.

```
#Building Decision Trees model
library(rpart)
library(rpart.plot)
treemodel = rpart(joined ~ ., data=train, minbucket = 2)
summary(treemodel)
Call:
rpart(formula = joined ~ ., data = train, minbucket = 2)
  n= 34

          CP nsplit rel error   xerror      xstd
```

```
1 0.21428571      0 1.0000000 1.000000 0.2049800
2 0.10714286      2 0.5714286 1.357143 0.2068021
3 0.03571429      4 0.3571429 1.428571 0.2049800
4 0.01000000      6 0.2857143 1.428571 0.2049800

Variable importance
       np   working      csal      dept  hpercent       exp      esal     hqual
       15        14        13        13        12        11         8         7
     pPhd    gender same_city
        3         2         2

Node number 1: 34 observations,    complexity param=0.2142857
  predicted class=no   expected loss=0.4117647  P(node) =1
    class counts:    20    14
   probabilities: 0.588 0.412
  left son=2 (11 obs) right son=3 (23 obs)
  Primary splits:
      np       splits as  RL,        improve=3.348059, (0 missing)
      csal     < 14982 to the right, improve=2.409982, (0 missing)
      hpercent < 67.6  to the left,  improve=2.121212, (1 missing)
      hqual    splits as  -RLLRL,    improve=1.902167, (0 missing)
      exp      < 0.5   to the right, improve=1.613445, (0 missing)
  Surrogate splits:
      working splits as  RL,        agree=0.941, adj=0.818, (0 split)
      dept    splits as  LRRRL,     agree=0.765, adj=0.273, (0 split)
      hqual   splits as  -RLRRL,    agree=0.735, adj=0.182, (0 split)
      exp     < 4.4   to the right, agree=0.735, adj=0.182, (0 split)
      esal    < 24500 to the right, agree=0.735, adj=0.182, (0 split)

Node number 2: 11 observations
  predicted class=no   expected loss=0.09090909  P(node) =0.3235294
    class counts:    10     1
   probabilities: 0.909 0.091

Node number 3: 23 observations,    complexity param=0.2142857
  predicted class=yes  expected loss=0.4347826  P(node) =0.6764706
```

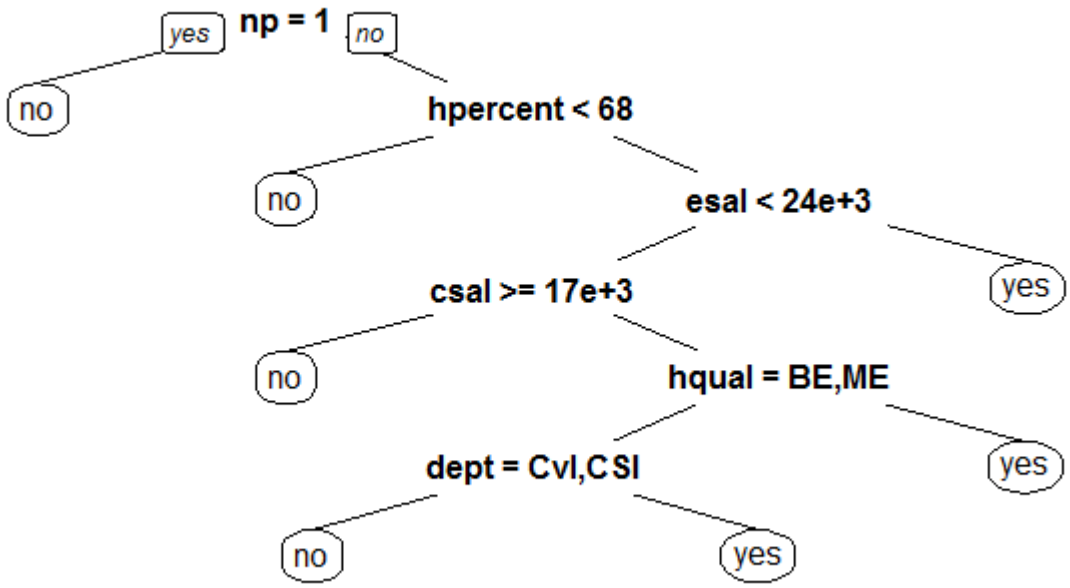
```
    class counts:    10    13
   probabilities: 0.435 0.565
  left son=6 (3 obs) right son=7 (20 obs)
  Primary splits:
      hpercent < 67.6  to the left,  improve=2.2043480, (0 missing)
      dept     splits as  -RLRR,     improve=1.7134390, (0 missing)
      csal     < 3750  to the right, improve=0.8376812, (0 missing)
      esal     < 20800 to the right, improve=0.8376812, (0 missing)
      working  splits as  LR,        improve=0.8281573, (0 missing)

Node number 6: 3 observations
  predicted class=no   expected loss=0  P(node) =0.08823529
    class counts:     3     0
   probabilities: 1.000 0.000

Node number 7: 20 observations,    complexity param=0.1071429
  predicted class=yes  expected loss=0.35  P(node) =0.5882353
    class counts:     7    13
   probabilities: 0.350 0.650
  left son=14 (16 obs) right son=15 (4 obs)
  Primary splits:
      esal     < 24500 to the left,  improve=1.2250000, (0 missing)
      hqual    splits as  -R-LR-,    improve=0.9000000, (0 missing)
      dept     splits as  -RLRR,     improve=0.6000000, (0 missing)
      csal     < 24992 to the left,  improve=0.5761905, (0 missing)
      hpercent < 69.1  to the right, improve=0.5444444, (0 missing)
  Surrogate splits:
      csal < 24992 to the left,  agree=0.90, adj=0.50, (0 split)
      pPhd splits as  LR,        agree=0.85, adj=0.25, (0 split)
      exp  < 8     to the left,  agree=0.85, adj=0.25, (0 split)

Node number 14: 16 observations,    complexity param=0.1071429
  predicted class=yes  expected loss=0.4375  P(node) =0.4705882
    class counts:     7     9
   probabilities: 0.438 0.562
  left son=28 (5 obs) right son=29 (11 obs)
```

```
  Primary splits:
      csal     < 17482 to the right, improve=1.911364, (0 missing)
      hqual    splits as  -R-LR-,    improve=1.906746, (0 missing)
      exp      < 5.5   to the right, improve=1.446429, (0 missing)
      esal     < 20800 to the right, improve=1.125000, (0 missing)
      hpercent < 69.1  to the right, improve=0.875000, (0 missing)
  Surrogate splits:
      exp       < 3.35  to the right, agree=0.938, adj=0.8, (0 split)
      pPhd      splits as  RL,        agree=0.750, adj=0.2, (0 split)
      working   splits as  RL,        agree=0.750, adj=0.2, (0 split)
      same_city splits as  LR,        agree=0.750, adj=0.2, (0 split)
      dept      splits as  -RRRL,     agree=0.750, adj=0.2, (0 split)

Node number 15: 4 observations
  predicted class=yes  expected loss=0  P(node) =0.1176471
    class counts:     0     4
   probabilities: 0.000 1.000

Node number 28: 5 observations
  predicted class=no   expected loss=0.2  P(node) =0.1470588
    class counts:     4     1
   probabilities: 0.800 0.200

Node number 29: 11 observations,    complexity param=0.03571429
  predicted class=yes  expected loss=0.2727273  P(node) =0.3235294
    class counts:     3     8
   probabilities: 0.273 0.727
  left son=58 (7 obs) right son=59 (4 obs)
  Primary splits:
      hqual    splits as  -L-LR-,    improve=0.9350649, (0 missing)
      dept     splits as  -LLRR,     improve=0.3636364, (0 missing)
      csal     < 13200 to the left,  improve=0.3636364, (0 missing)
      esal     < 20800 to the right, improve=0.2969697, (0 missing)
      hpercent < 72.3  to the right, improve=0.0969697, (0 missing)
  Surrogate splits:
      gender   splits as  RL,        agree=0.818, adj=0.5, (0 split)
```

```
      hpercent < 78    to the left,  agree=0.818, adj=0.5, (0 split)
      dept     splits as  -LRLL,     agree=0.818, adj=0.5, (0 split)
      csal     < 13200 to the left,  agree=0.818, adj=0.5, (0 split)

Node number 58: 7 observations,    complexity param=0.03571429
  predicted class=yes  expected loss=0.4285714  P(node) =0.2058824
    class counts:     3     4
   probabilities: 0.429 0.571
  left son=116 (5 obs) right son=117 (2 obs)
  Primary splits:
      dept     splits as  -LLRR,     improve=1.0285710, (0 missing)
      hpercent < 72.25 to the right, improve=0.5952381, (0 missing)
      hqual    splits as  -R-L--,    improve=0.0952381, (0 missing)
      gate     splits as  LR,        improve=0.0952381, (0 missing)
      esal     < 18800 to the right, improve=0.0952381, (0 missing)

Node number 59: 4 observations
  predicted class=yes  expected loss=0  P(node) =0.1176471
    class counts:     0     4
   probabilities: 0.000 1.000

Node number 116: 5 observations
  predicted class=no   expected loss=0.4  P(node) =0.1470588
    class counts:     3     2
   probabilities: 0.600 0.400

Node number 117: 2 observations
  predicted class=yes  expected loss=0  P(node) =0.05882353
    class counts:     0     2
   probabilities: 0.000 1.000
```

Now, plotting the Decision Tree

```
#Plot the Decision tree.
prp(treemodel)
```



The plot suggests various decision alternatives by which to arrive at a possible scenario of whether a candidate is likely to join or not. One such decision route is notice period is 0 month , highest qualification percentage is greater than or equal to 68, Expected salary is less than 24K, Current Salary is less than or equal to 17K, highest qualification is BE or ME and the department is not Civil or CSIT, then the candidate is likely to join.

As can be seen from the plot, there are other decision routes also.

After building these models, we will look into testing and evaluating both the models.

Chapter – 8

Test and Evaluation

To assess the models, predicted outcomes on the testing set are compared against the actual outcomes using a **confusion matrix** or a **classification matrix**.

	Predicted class = 0	Predicted class = 1
Actual class = 0	True Negatives (TN)	False Positives (FP)
Actual class = 1	False Negatives (FN)	True Positives (TP)

N = number of observations

Overall Accuracy = (TN + TP)/N

Sensitivity = TP/(TP + FN)

Specificity = TN/(YN + FP)

We will build confusion matrix for the logistic regression model.

```
#Since, the number of levels are different in training and testing datasets
logmodel$xlevels[["hqual"]] <- union(logmodel$xlevels[["hqual"]],
levels(test$hqual))
#Prediction using the predict function
predictlog = predict(logmodel, newdata = test, type = "response")
table(test$joined, predictlog > 0.5)
```

```
      FALSE TRUE
  no      3    6
  yes     0    6
```

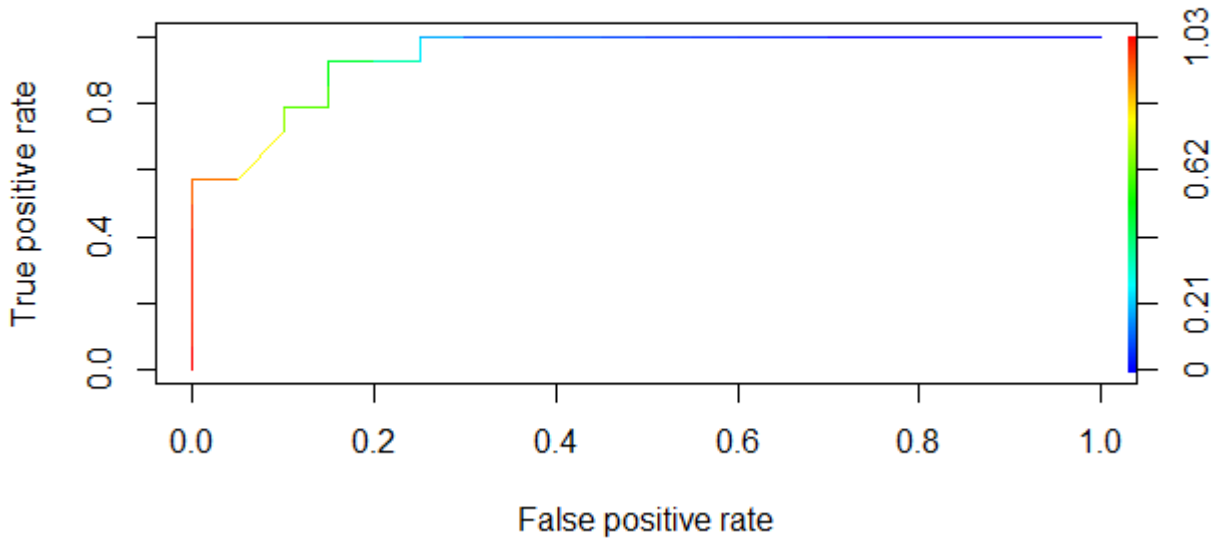
Thus overall accuracy of the model = (3+6)/15 = 60%.

This is only marginally better than the baseline accuracy of 59%.

Next the ROC (Receiver Operating Characteristics) Curve is plotted.

ROC curve is the plot of TPR (True Positive Rate) on the vertical axis against FPR (False Positive Rate) on the horizontal axis. More the area under the ROC curve or higher the ROC curve, better is the model.

```
predicttrainlog = predict(logmodel, type = "response")
#Load ROCR package for plotting ROC curve
library(ROCR)

# Prediction function
ROCRtrainlogpred = prediction(predicttrainlog, train$joined)

# Performance function
ROCRtrainlogperf = performance(ROCRtrainlogpred, "tpr", "fpr")

# Plot ROC curve
plot(ROCRtrainlogperf, colorize=TRUE)
```




```
performance(ROCRtrainlogpred, "auc")@y.values
0.9428571
```

Area under the ROC curve gives a high value of 94.28% but this could also mean that the model is overfitting the training data. If a model performs very well on the training data but performs much worse on the test data, it is said to be overfitting the data.

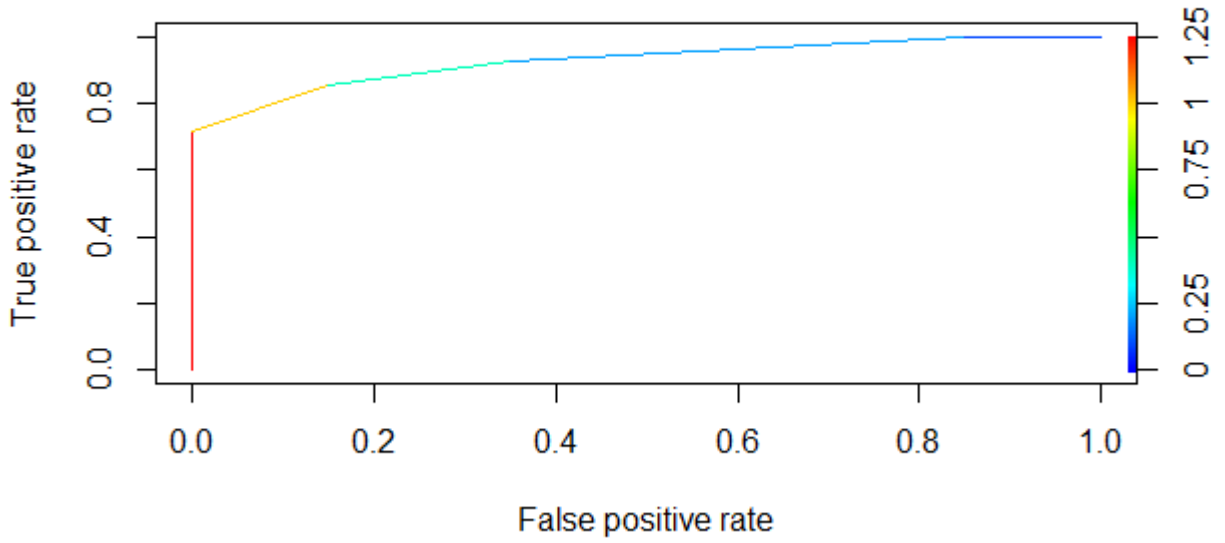
We can similarly calculate the performance metrics for the Decsin Tree model as well.

Creating Confusion matrix:

```
predictCART = predict(treemodel, newdata = test, type = "class")
table(test$joined, predictCART)
  predictCART
      no yes
  no   8   1
  yes  4   2
(8+2)/(15)
[1] 0.6666667
```

The accuracy of the CART model is 66.7% which is better than both the baseline and the logistic regression model.

```
predicttrainCART = predict(treemodel)

# Prediction function
ROCRtrainCARTpred = prediction(predicttrainCART[,2], train$joined)

# Performance function
ROCRtrainCARTperf = performance(ROCRtrainCARTpred, "tpr", "fpr")

# Plot ROC curve
plot(ROCRtrainCARTperf, colorize=TRUE)
```




```
performance(ROCRtrainCARTpred, "auc")@y.values
[1] 0.9285714
```

Area under the ROC curve for CART model gives a high value of 92.86% which means the model can differentiate very well between true positives and false positives on the training data. Thus, the CART model is better model for the given dataset as its accuracy is higher and the tendency to overfit is also lesser.

CHAPTER – 9

CONCLUSIONS AND RECOMMENDATIONS

As can be seen from the metrics, both the models perform very well on the training dataset but perform much worse on the test dataset. One reason could be the fairly limited number of 49 observations available in the dataset. There should be at least 100 observations for a model with 10 predictors, as a rule of thumb. These models could be deployed with limited accuracy in the field.

It is recommended that the institute collect much more data regarding recruitment and selection and at a more granular level, so that these algorithms can improvised and their predictive power improved. It is also recommended that institute keep candidate resumes and selection questionnaire and interview ratings so that the scope of the analytics models can be increased further.

CHAPTER – 10

BIBLIOGRAPHY

BIBLOGRAPHY

1. Jean Paul Isson & Jesse S. Harriott, People Analytics in the Era of Big Data, Wiley, 2016
2. Jac Fitz-Enz and John R. Mattox II, Predictive Analytics for the Human Resources, Wiley, 2014
3. Debra L. Nelson, James Campbell Quick, Preetam Khandelwal, Angelo S. DeNisi, Ricky W. Griffin, Anita Sarkar, Human Behavior and People Processes, Cengage Learning, 2014
4. Jiawei Han, Micheline Kamber and Jian Pei, Data Mining Concepts and Techniques, Morgan Kaufmann, 2012.
5. M. Sudheep Elayidom, Data Mining and Business Intelligence, Cengage Learning, 2015
6. Gordon S. Linoff and Michael J. A. Berry, Data Mining Techniques for Marketing, Sales and Customer Relationship Management, Wiley, 2013
7. David R. Anderson, Dennis J. Sweeney, Thomas A. Williams, S. Christian Albright, Christopher J. Zappe, Wayne L. Winston and Cliff T. Ragsdale, Data Analysis and Business Decision Making, Cengage Learning, 2014

APPENDIX

Source Code

Recruit1.R

```
# *------------------------------------------------------------------
# | PROGRAM NAME: Analysis of MIST recruitment data
# | DATE: 25/05/2017
# | CREATED BY:  Ankur Shrivastava
# | PROJECT FILE: Recruit1.R
# *------------------------------------------------------------------
# | PURPOSE: The purpose of this file is to analyze the various
# |          factors responsible for a candidate joining MIST
# *------------------------------------------------------------------
# | COMMENTS:
# |
# |  1:
# |*-----------------------------------------------------------------
# | DATA USED:  MIST recruitment data
# |
# |
# |*-----------------------------------------------------------------

install.packages("devtools", dependencies = TRUE)
install.packages("Rcpp", dependencies = TRUE)
install.packages("readxl", dependencies = TRUE)
install.packages("caTools", dependencies = TRUE)
install.packages("rpart", dependencies = TRUE)
install.packages("rpart.plot", dependencies = TRUE)
install.packages("ROCR", dependencies = TRUE)

library(devtools)
library(Rcpp)
library(readxl)
```

```
#Load the recruitment dataset.
recruit <- read_excel("Recruit1.xlsx", 1)

#First look at data in R
str(recruit)
summary(recruit)

# replace text "NA" with R's NA
for(x in 1:nrow(recruit)){
  for(y in 1:ncol(recruit)){
    if(!is.na(recruit[x,y])){
      if(recruit[x,y] == "NA"){
        recruit[x,y] <- NA
      }
    }
  }
}

#Converting categorical variables read in as text to factor variables
recruit$gender <- as.factor(recruit$gender)
recruit$hqual <- as.factor(recruit$hqual)
recruit$gate <- as.factor(recruit$gate)
recruit$pPhd <- as.factor(recruit$pPhd)
recruit$working <- as.factor(recruit$working)
recruit$org <- as.factor(recruit$org)
recruit$same_city <- as.factor(recruit$same_city)
recruit$np <- as.factor(recruit$np)
recruit$dept <- as.factor(recruit$dept)
recruit$offered <- as.factor(recruit$offered)
recruit$joined <- as.factor(recruit$joined)
recruit$csal <- as.numeric(recruit$csal)
recruit$esal <- as.numeric(recruit$esal)
recruit$osal <- as.numeric(recruit$osal)

#Removing the name column as that won't be needed for analysis
recruit <- subset(recruit, select = -c(name))
```

```
#looking at the data again
str(recruit)
summary(recruit)

library(ggplot2)
#Plot of actual numbers who joined
ggplot(recruit, aes(x = joined)) + geom_bar(fill = "red") + xlab("Candidates
joined") + ylab("Number of Candidates")

#Plot of actual numbers who joined by each department
ggplot(recruit, aes( x = joined, fill = dept)) + geom_bar(position = "dodge") +
xlab("Department") + ylab("Number of Candidates joined")

#Running chi squre tests
chisq.test(recruit$joined, recruit$dept)

#Percentage of people joined by each department
ggplot(recruit, aes( x = dept, fill = joined)) + geom_bar(position = "fill") +
xlab("Department") + ylab("Percentage of Candidates joined")

#Percentage of people joined by gender
ggplot(recruit, aes( x = gender, fill = joined)) + geom_bar(position = "fill") +
xlab("Gender") + ylab("Percentage of Candidates joined")

#Running chi squre tests
chisq.test(recruit$joined, recruit$gender)

#Percentage of people joined by working or not
ggplot(recruit, aes( x = working, fill = joined)) + geom_bar(position = "fill") +
xlab("Currently Working") + ylab("Percentage of Candidates joined")

#Running chi squre test
chisq.test(recruit$joined, recruit$working)

#Number of people joined by years of experience
```

```
ggplot(recruit, aes( x = joined, y = exp)) + geom_boxplot() + xlab("Candidates
joined") + ylab("Years of experience")

#Running one-way anova
exp.aov <- aov(recruit$exp ~ recruit$joined)
summary(exp.aov)

#Percentage of people joined by GATE qualified
ggplot(recruit, aes( x = gate, fill = joined)) + geom_bar(position = "fill") +
xlab("GATE qualified") + ylab("Percentage of Candidates joined")

#Running chi squre test
chisq.test(recruit$joined, recruit$gate)

#Percentage of people joined by Highest qualification
ggplot(recruit, aes( x = hqual, fill = joined)) + geom_bar(position = "fill") +
xlab("Highest Qualification") + ylab("Percentage of Candidates joined")

#Running chi squre test
chisq.test(recruit$joined, recruit$hqual)

#Percentage of people joined by pursuing Phd
ggplot(recruit, aes( x = pPhd, fill = joined)) + geom_bar(position = "fill") +
xlab("Pursuing Phd") + ylab("Percentage of Candidates joined")

#Running chi squre test
chisq.test(recruit$joined, recruit$pPhd)

#Percentage of people joined by same_city
ggplot(recruit, aes( x = same_city, fill = joined)) + geom_bar(position = "fill") +
xlab("Residing in same city") + ylab("Percentage of Candidates joined")

#Running chi squre test
chisq.test(recruit$joined, recruit$same_city)

#Percentage of people joined by notice period
```

```
ggplot(recruit, aes( x = np, fill = joined)) + geom_bar(position = "fill") +
xlab("Notice Period in months") + ylab("Percentage of Candidates joined")

#Running chi squre test
chisq.test(recruit$joined, recruit$np)

#Plot of people joining by current salary
ggplot(recruit, aes( x = joined, y = csal)) + geom_boxplot() + xlab("Candidates
joined") + ylab("Current Salary")

#Running one-way anova
csal.aov <- aov(recruit$csal ~ recruit$joined)
summary(csal.aov)

#Plot of people joining by expected salary
ggplot(recruit, aes( x = joined, y = esal)) + geom_boxplot() + xlab("Candidates
joined") + ylab("Expected Salary")

#Running one-way anova
esal.aov <- aov(recruit$csal ~ recruit$joined)
summary(esal.aov)

#Correlation between current salary and expected salary
cor(recruit$csal, recruit$esal, use = "pairwise.complete.obs")

#----------------------------Modeling Section-------------------------------------
------

table(recruit$joined)
#baseline accuracy
29/49

#Remove unnecessary variables
recruit1 <- subset(recruit, select = -c(org, offered, remarks, osal))

library(caTools)
```

```
# Randomly split data
set.seed(2017)
split = sample.split(recruit1$joined, SplitRatio = 0.7)
split

# Create training and testing sets
train = subset(recruit1, split == TRUE)
test = subset(recruit1, split == FALSE)

# Logistic Regression Model
logmodel = glm(joined ~  hqual + gate + pPhd +exp +working + same_city + np + dept +
csal , data=train, family=binomial)
summary(logmodel)

#Building Decision Trees model
library(rpart)
library(rpart.plot)
treemodel = rpart(joined ~ ., data=train, minbucket = 2)
summary(treemodel)
#Plot the Decision tree.
prp(treemodel)

#----------Test and evaluation--------------------------------------------------

#Since, the number of levels are different in training and testing datasets
logmodel$xlevels[["hqual"]] <- union(logmodel$xlevels[["hqual"]],
levels(test$hqual))
#Prediction using the predict function
predictlog = predict(logmodel, newdata = test, type = "response")
table(test$joined, predictlog > 0.5)
(3+6)/15

predicttrainlog = predict(logmodel, type = "response")
#Load ROCR package for plotting ROC curve
library(ROCR)
```

```
# Prediction function
ROCRtrainlogpred = prediction(predicttrainlog, train$joined)

# Performance function
ROCRtrainlogperf = performance(ROCRtrainlogpred, "tpr", "fpr")

# Plot ROC curve
plot(ROCRtrainlogperf, colorize=TRUE)

performance(ROCRtrainlogpred, "auc")@y.values



#----------------Similarly for CART model---------------------------

predictCART = predict(treemodel, newdata = test, type = "class")
table(test$joined, predictCART)
(8+2)/(15)

predicttrainCART = predict(treemodel)

# Prediction function
ROCRtrainCARTpred = prediction(predicttrainCART[,2], train$joined)

# Performance function
ROCRtrainCARTperf = performance(ROCRtrainCARTpred, "tpr", "fpr")

# Plot ROC curve
plot(ROCRtrainCARTperf, colorize=TRUE)

performance(ROCRtrainCARTpred, "auc")@y.values
#-------------------End of File-------------------------------------------
```