HR ANALYTICS

TABLE OF CONTENTS

INDEX

LITERATURE

The goal of human resources analytics is to provide an organization with insights for optimum utilization and managing employees so that business goals can be reached quickly and efficiently. The challenge of human resources analytics is to identify what data should be captured and how to use the data to model and predict capabilities, so the organization gets an optimal return on investment on its human capital.

Retaining key employees is a major stake for any organization. But are there reliable ways to figure out if and why the best and most experienced employees are leaving prematurely? Most firms these days are already integrating the benefits of using analytics to introduce special efforts in regaining employees as well as hiring decisions. Lot of factors play key role in identifying significant predictors offering insights and meaning that can be interpreted using a statistical model language like R.

In this project, we have used HR Analytics dataset from Kaggle that is fictitious in nature seemingly because no company will share its personally identifiable record.

BACKGROUND

Data set

This data set represents 14,999 employees and is composed of both currently employed and people who have already left the company with 30 variables defining the best possible way to answer the below questions and insights.

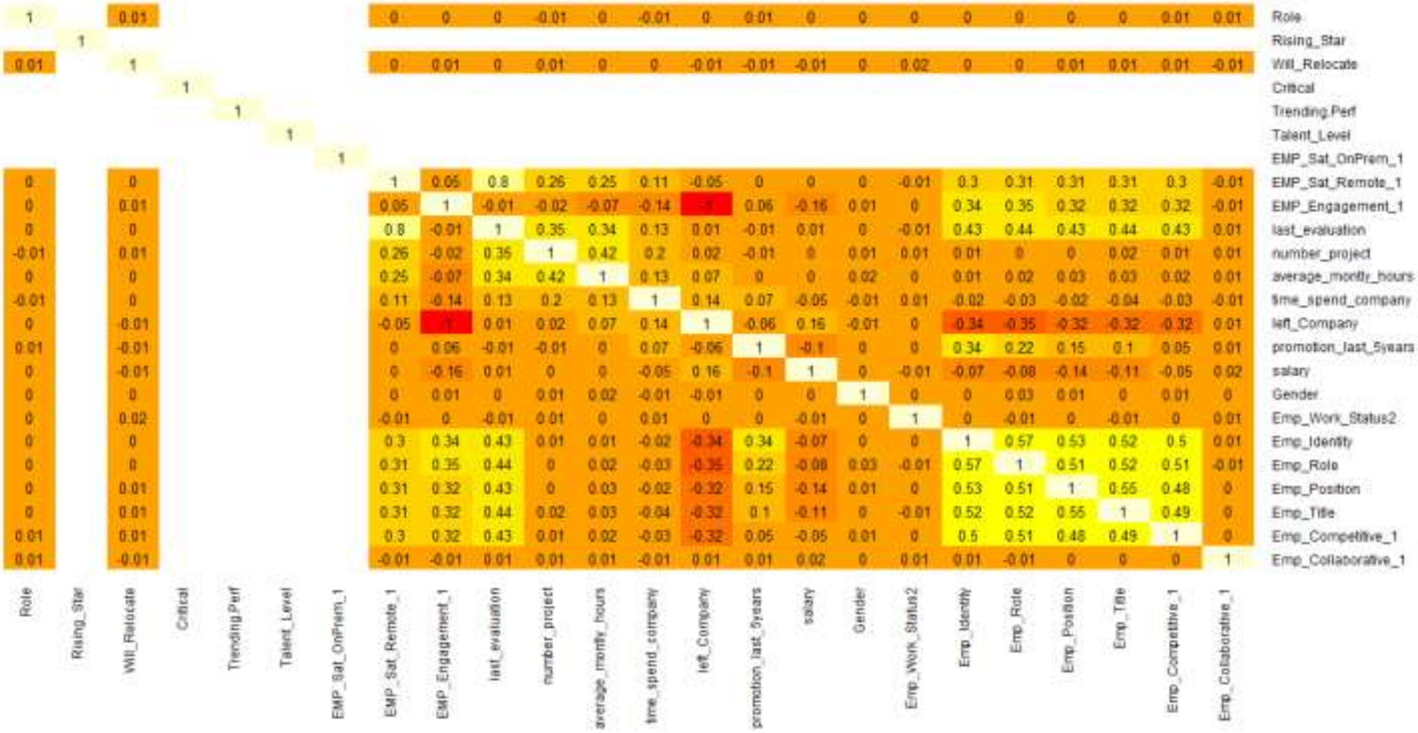
Initially after loading the dataset, we saw 25+ variables that had no significance for any of our analysis model and hence we decided to discard them. It is always recommended to run some basic checks and see if there are missing values or any unusual patterns amongst other things (in most data sets Kaggle gives you clean data). Right from the very first correlation that we ran, we were clear about incorporating few changes to the dataset. We compared the Kaggle dataset with the IBM HR analytics dataset and included a field called Employee_satisfaction from the latter and merged it with the existing file to create a new variable with the same name, representing an average of five other parameters from the file.

<u>Correlation matrix before and after making changes to the dataset.</u>

Before running the correlation, it was imperative to convert all the category variable values to factors and from factors to Numeric.

```
# Convert Category values to Factors

hr.df$Role <- factor(hr.df$Role, levels = c("Director","Level 1",
                                            "Level 2-4","Manager","Senior Director",
                                            "Senior Manager","VP"),
                                            labels = c(3,7,6,5,2,4,1))

hr.df$salary <- factor(hr.df$salary, levels = c("high", "low", "medium"),
                       labels = c(1, 3, 2))

hr.df$Gender <- factor(hr.df$Gender, levels = c("F", "M"),
                       labels = c(0, 1))

#Convert Factors into Numeric
hr.df$salary = as.numeric(paste(hr.df$salary))
hr.df$Gender = as.numeric(paste(hr.df$Gender))
hr.df$Role = as.numeric(paste(hr.df$Role))

#Remove not needed Categorical Variable for Heat Map
hrform.df <- hr.df[,c(-1,-2,-3,-4,-11)]

heatmap.2(cor(hrform.df), Rowv = FALSE, Colv = FALSE, dendrogram = "none",
          cellnote = round(cor(hrform.df),2), notecol = "black",
          key = FALSE, trace = 'none', margins = c(10,10))
```

Correlation run before making changes to the dataset

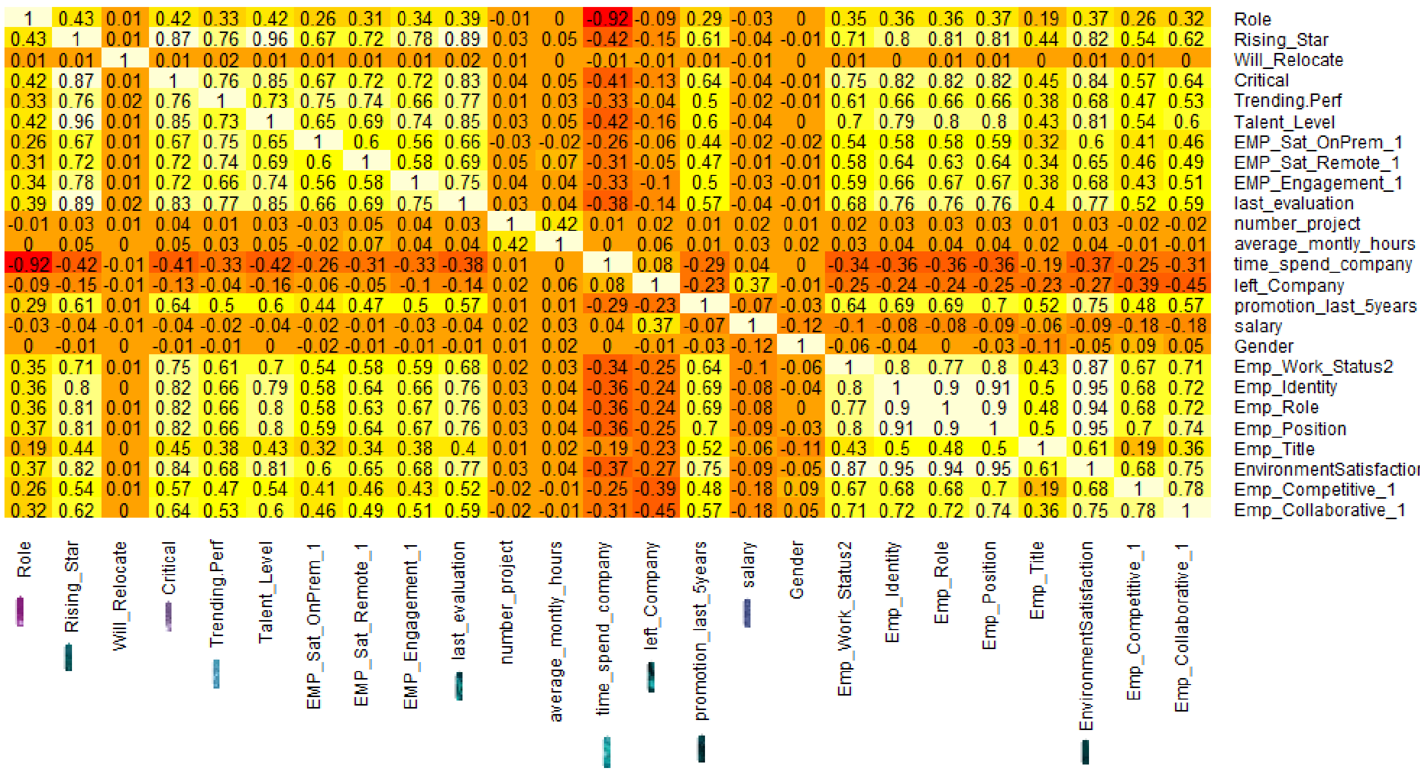


Correlation run after incorporating changes to the dataset

To improve the correlation significance between various predictors, we made changes against many variable records. (*Rising_Star, Role, Left_Company, promotion_last_5years, Critical, time_spend_company, Salary, Emp_Satisfaction*)

We included a new variable Emp_Satisfaction from IBM dataset and merged it with our file to create a new field storing averages of (Emp_Position + Emp_Work_Status2 + Emp_Identity + Emp_Title and Emp_role) all storing a value of scale between 1-10. 1 being the lowest and 10 highest.

Significant correlation exists amongst majority of the variables



	Role	Rising_Star	Will_Relocate	Critical	Trending.Perf	Talent_Level	EMP_Sat_OnPrem_1	EMP_Sat_Remote_1	EMP_Engagement_1	last_evaluation	number_project	average_montly_hours	time_spend_company	left_Company	promotion_last_5years	salary	Gender	Emp_Work_Status2	Emp_Identity	Emp_Role	Emp_Position	Emp_Title	EnvironmentSatisfaction	Emp_Competitive_1	Emp_Collaborative_1
Role	1	0.43	0.01	0.42	0.33	0.42	0.26	0.31	0.34	0.39	-0.01	0	-0.92	-0.09	0.29	-0.03	0	0.35	0.36	0.36	0.37	0.19	0.37	0.26	0.32
Rising_Star	0.43	1	0.01	0.87	0.76	0.96	0.67	0.72	0.78	0.89	0.03	0.05	-0.42	-0.15	0.61	-0.04	-0.01	0.71	0.8	0.81	0.81	0.44	0.82	0.54	0.62
Will_Relocate	0.01	0.01	1	0.01	0.02	0.01	0.01	0.01	0.01	0.02	0.01	0	-0.01	-0.01	0.01	-0.01	0	0.01	0	0.01	0.01	0	0.01	0.01	0
Critical	0.42	0.87	0.01	1	0.76	0.85	0.67	0.72	0.72	0.83	0.04	0.05	-0.41	-0.13	0.64	-0.04	-0.01	0.75	0.82	0.82	0.82	0.45	0.84	0.57	0.64
Trending.Perf	0.33	0.76	0.02	0.76	1	0.73	0.75	0.74	0.66	0.77	0.01	0.03	-0.33	-0.04	0.5	-0.02	-0.01	0.61	0.66	0.66	0.66	0.38	0.68	0.47	0.53
Talent_Level	0.42	0.96	0.01	0.85	0.73	1	0.65	0.69	0.74	0.85	0.03	0.05	-0.42	-0.16	0.6	-0.04	0	0.7	0.79	0.8	0.8	0.43	0.81	0.54	0.6
EMP_Sat_OnPrem_1	0.26	0.67	0.01	0.67	0.75	0.65	1	0.6	0.56	0.66	-0.03	-0.02	-0.26	-0.06	0.44	-0.02	-0.02	0.54	0.58	0.58	0.59	0.32	0.6	0.41	0.46
EMP_Sat_Remote_1	0.31	0.72	0.01	0.72	0.74	0.69	0.6	1	0.58	0.69	0.05	0.07	-0.31	-0.05	0.47	-0.01	-0.01	0.58	0.64	0.63	0.64	0.34	0.65	0.46	0.49
EMP_Engagement_1	0.34	0.78	0.01	0.72	0.66	0.74	0.56	0.58	1	0.75	0.04	0.04	-0.33	-0.1	0.5	-0.03	-0.01	0.59	0.66	0.67	0.67	0.38	0.68	0.43	0.51
last_evaluation	0.39	0.89	0.02	0.83	0.77	0.85	0.66	0.69	0.75	1	0.03	0.04	-0.38	-0.14	0.57	-0.04	-0.01	0.68	0.76	0.76	0.76	0.4	0.77	0.52	0.59
number_project	-0.01	0.03	0.01	0.04	0.01	0.03	-0.03	0.05	0.04	0.03	1	0.42	0.01	0.02	0.01	0.02	0.01	0.02	0.03	0.03	0.03	0.01	0.03	-0.02	-0.02
average_montly_hours	0	0.05	0	0.05	0.03	0.05	-0.02	0.07	0.04	0.04	0.42	1	0	0.06	0.01	0.03	0.02	0.03	0.04	0.04	0.04	0.02	0.04	-0.01	-0.01
time_spend_company	-0.92	-0.42	-0.01	-0.41	-0.33	-0.42	-0.26	-0.31	-0.33	-0.38	0.01	0	1	0.08	-0.29	0.04	0	-0.34	-0.36	-0.36	-0.36	-0.19	-0.37	-0.25	-0.31
left_Company	-0.09	-0.15	-0.01	-0.13	-0.04	-0.16	-0.06	-0.05	-0.1	-0.14	0.02	0.06	0.08	1	-0.23	0.37	-0.01	-0.25	-0.24	-0.24	-0.25	-0.23	-0.27	-0.39	-0.45
promotion_last_5years	0.29	0.61	0.01	0.64	0.5	0.6	0.44	0.47	0.5	0.57	0.01	0.01	-0.29	-0.23	1	-0.07	-0.03	0.64	0.69	0.69	0.7	0.52	0.75	0.48	0.57
salary	-0.03	-0.04	-0.01	-0.04	-0.02	-0.04	-0.02	-0.01	-0.03	-0.04	0.02	0.03	0.04	0.37	-0.07	1	-0.12	-0.1	-0.08	-0.08	-0.09	-0.06	-0.09	-0.18	-0.18
Gender	0	-0.01	0	-0.01	-0.01	0	-0.02	-0.01	-0.01	-0.01	0.01	0.02	0	-0.01	-0.03	-0.12	1	-0.06	-0.04	0	-0.03	-0.11	-0.05	0.09	0.05
Emp_Work_Status2	0.35	0.71	0.01	0.75	0.61	0.7	0.54	0.58	0.59	0.68	0.02	0.03	-0.34	-0.25	0.64	-0.1	-0.06	1	0.8	0.77	0.8	0.43	0.87	0.67	0.71
Emp_Identity	0.36	0.8	0	0.82	0.66	0.79	0.58	0.64	0.66	0.76	0.03	0.04	-0.36	-0.24	0.69	-0.08	-0.04	0.8	1	0.9	0.91	0.5	0.95	0.68	0.72
Emp_Role	0.36	0.81	0.01	0.82	0.66	0.8	0.58	0.63	0.67	0.76	0.03	0.04	-0.36	-0.24	0.69	-0.08	0	0.77	0.9	1	0.9	0.48	0.94	0.68	0.72
Emp_Position	0.37	0.81	0.01	0.82	0.66	0.8	0.59	0.64	0.67	0.76	0.03	0.04	-0.36	-0.25	0.7	-0.09	-0.03	0.8	0.91	0.9	1	0.5	0.95	0.7	0.74
Emp_Title	0.19	0.44	0	0.45	0.38	0.43	0.32	0.34	0.38	0.4	0.01	0.02	-0.19	-0.23	0.52	-0.06	-0.11	0.43	0.5	0.48	0.5	1	0.61	0.19	0.36
EnvironmentSatisfaction	0.37	0.82	0.01	0.84	0.68	0.81	0.6	0.65	0.68	0.77	0.03	0.04	-0.37	-0.27	0.75	-0.09	-0.05	0.87	0.95	0.94	0.95	0.61	1	0.68	0.75
Emp_Competitive_1	0.26	0.54	0.01	0.57	0.47	0.54	0.41	0.46	0.43	0.52	-0.02	-0.01	-0.25	-0.39	0.48	-0.18	0.09	0.67	0.68	0.68	0.7	0.19	0.68	1	0.78
Emp_Collaborative_1	0.32	0.62	0	0.64	0.53	0.6	0.46	0.49	0.51	0.59	-0.02	-0.01	-0.31	-0.45	0.57	-0.18	0.05	0.71	0.72	0.72	0.74	0.36	0.75	0.78	1

OBJECTIVES

The main objectives that we had set out before working on the dataset were :

- Identify the primary reasons for employees leaving both low and high performance
- Why do good employees leave?
- Will the employee leave the company?
- What is the likelihood of Employee getting a promotion?
- How much time will the employee spend in company?
- How satisfied are the employees in company?

DATA EXPLORATION

Read the HR Dataset

```
hr.df <- read.csv("HR.csv", header = TRUE)
```

Dataset Details

```
dim(hr.df)
```

```
## [1] 14999    30
```

Describe Dataset

```
summary(hr.df)
##        ID             Name                    Department          GEO
##  Min.   :    1   AARON   :     1   Finance         :1983   UK      :1772
##  1st Qu.: 3750   ABAD    :     1   Human Resources:1785   France  :1699
##  Median : 7500   ABADIE  :     1   IT              :3485   Korea   :1685
##  Mean   : 7500   ABARCA  :     1   Operations      :2500   Japan   :1669
##  3rd Qu.:11250   ABATE   :     1   Sales           :2500   China   :1667
##  Max.   :14999   (Other):14993   Support         : 247   Colombia:1659
##                  NA's    :     1   Warehouse       :2499   (Other) :4848
##               Role        Rising_Star    Will_Relocate       Critical
##  Director        : 660   Min.   :1.000   Min.   :0.0000   Min.   :0.000
##  Level 1         :3270   1st Qu.:2.000   1st Qu.:0.0000   1st Qu.:0.000
##  Level 2-4       :6889   Median :4.000   Median :0.0000   Median :1.000
##  Manager         :2420   Mean   :3.511   Mean   :0.4998   Mean   :0.682
##  Senior Director: 330   3rd Qu.:5.000   3rd Qu.:1.0000   3rd Qu.:1.000
##  Senior Manager :1326   Max.   :5.000   Max.   :1.0000   Max.   :1.000
##  VP              : 104
##  Trending.Perf     Talent_Level     Percent_Remote   EMP_Sat_OnPrem_1
##  Min.   : 1.000   Min.   : 1.000   Min.   :0.4000   Min.   : 0.000
##  1st Qu.: 6.000   1st Qu.: 5.000   1st Qu.:0.4000   1st Qu.: 5.000
##  Median : 8.000   Median : 7.000   Median :0.8000   Median : 7.000
##  Mean   : 7.171   Mean   : 6.451   Mean   :0.6173   Mean   : 6.615
##  3rd Qu.: 9.000   3rd Qu.: 8.000   3rd Qu.:0.8000   3rd Qu.: 8.000
##  Max.   :10.000   Max.   :10.000   Max.   :1.0000   Max.   :10.000
##
##  EMP_Sat_Remote_1 EMP_Engagement_1 last_evaluation  number_project
##  Min.   : 1.000   Min.   :1.000    Min.   : 3.000   Min.   :2.000
##  1st Qu.: 6.000   1st Qu.:2.000    1st Qu.: 5.000   1st Qu.:3.000
##  Median : 8.000   Median :3.000    Median : 7.000   Median :4.000
##  Mean   : 7.273   Mean   :2.997    Mean   : 7.017   Mean   :3.803
##  3rd Qu.: 9.000   3rd Qu.:4.000    3rd Qu.: 9.000   3rd Qu.:5.000
##  Max.   :10.000   Max.   :5.000    Max.   :10.000   Max.   :7.000
##
##  average_montly_hours time_spend_company  left_Company
##  Min.   : 40          Min.   : 1.000     Min.   :0.0000
##  1st Qu.:156          1st Qu.: 7.000     1st Qu.:0.0000
##  Median :200          Median : 9.000     Median :0.0000
```

```
##  Mean   :201          Mean   : 9.616    Mean   :0.3062
##  3rd Qu.:245          3rd Qu.:12.000    3rd Qu.:1.0000
##  Max.   :310          Max.   :22.000    Max.   :1.0000
##
##  promotion_last_5years    salary     Gender   Emp_Work_Status2
##  Min.   :0.0000          high  :1668  F:7596  Min.   : 1.00
##  1st Qu.:0.0000          low   :6857  M:7403  1st Qu.: 4.00
##  Median :0.0000          medium:6474          Median : 7.00
##  Mean   :0.4744                               Mean   : 6.41
##  3rd Qu.:1.0000                               3rd Qu.: 9.00
##  Max.   :1.0000                               Max.   :10.00
##
##   Emp_Identity       Emp_Role        Emp_Position      Emp_Title
##  Min.   : 1.000   Min.   : 1.000   Min.   : 1.000   Min.   : 1.000
##  1st Qu.: 2.000   1st Qu.: 2.000   1st Qu.: 2.000   1st Qu.: 2.000
##  Median : 7.000   Median : 7.000   Median : 7.000   Median : 3.000
##  Mean   : 6.143   Mean   : 6.143   Mean   : 6.067   Mean   : 3.287
##  3rd Qu.: 9.000   3rd Qu.: 9.000   3rd Qu.: 9.000   3rd Qu.: 5.000
##  Max.   :10.000   Max.   :10.000   Max.   :10.000   Max.   :10.000
##
##  Emp_Satisfaction Emp_Competitive_1 Emp_Collaborative_1
##  Min.   : 1.000   Min.   : 1.000    Min.   : 1.000
##  1st Qu.: 3.000   1st Qu.: 2.000    1st Qu.: 3.000
##  Median : 7.000   Median : 6.000    Median : 7.000
##  Mean   : 5.608   Mean   : 4.998    Mean   : 5.938
##  3rd Qu.: 8.000   3rd Qu.: 8.000    3rd Qu.: 9.000
##  Max.   :10.000   Max.   :10.000    Max.   :10.000
```

Meta Data

Attribute	Description
ID	Employee ID
Name	Employee Name
Department	Department
GEO	Geographical location
Role	Current Role or title of employee
Rising Star	Indicates the level of promise or promote-ability the employee has. Scale(1-5)
Will_Relocate	Is the employee willing to relocate? 0- No, 1- Yes
Critical	Is the employee critical to the organization? 0- No, 1- Yes
Trending Perf	How is the employee trending in performance this year? Scale (1-10)
Talent_Level	This field represents a subjective level of management's view of the employee. Scale (1-10)
Percent_Remote	The percentage of the employee's work that is done remotely.
EMP_Sat_OnPrem_1	One indicator from a survey that was sent to employees. On prem (On premise) means that the employee maintains a high percentage of work on the corporation’s physical work locations. Scale (1-10)
EMP_Sat_Remote_1	One indicator from a survey that was sent to employees. Remote (distance employee) means that the employee does a high percentage of work away from the corporation’s physical work locations. Scale (1-10)
EMP_Engagement_1	One indicator from a survey that was sent to employees. Engagement represents the employee's feeling about how they feel about being engaged in company activities. Scale(1-5)

last_evaluation	The score on the last employee evaluation.Scale (1-10)
number_project	The number of projects the employee works on throughout the year.
average_montly_hours	The average number of hours the employee works monthly.
time_spend_company	Years of service
left_Company	Did the employee leave the company? 0- No, 1- Yes
promotion_last_5years	Did the employee get promoted in last 5 years? 0- No, 1- Yes
salary	Relative pay grade (low, medium, high) by role.
Gender	M or F
Emp_Work_Status2	One indicator from a survey that was sent to employees. Status represents how strongly employee feels about their status level in the organization. Scale (1-10)
Emp_Identity	How the employee identifies themselves with the company. Scale (1-10)
Emp_Role	How the employee identifies themselves with the importance of their role in the company. Scale (1-10)
Emp_Position	How the employee identifies themselves with the importance of their position in the company. Scale (1-10)
Emp_Title	How the employee feels about their title.Scale (1-10)
Emp_Satisfaction	Average value of the above 5 variables. Scale out of 1-10
Emp_Competitive_1	One indicator from a survey that was sent to employees. How employee feels about the competitive nature of work in the organization. Scale (1-10)
Emp_Collaborative_1	One indicator from a survey that was sent to employees. How employee feels about the collaborative nature of work in the organization.Scale (1-10)

The complete R code file #Uploaded separately on E-Learning

HR Analytics_R script.txt

Promotion on basis of time spend in a company

```
timespend_prom <-xtabs(~promotion_last_5years+time_spend_company,data=hr.df)
timespend_prom
```

```
##                       time_spend_company
## promotion_last_5years    1    2    3    4    5    6    7    8    9   10
##                     0    8    5  145  253  316  272  731 1072  970  610
##                     1    6    7  195  377  425  402 1046 1540 1313  863
##                       time_spend_company
## promotion_last_5years   11   12   13   14   15   16   17   18   19   20
##                     0  350  451  504  529  420  308  341  243  172  120
##                     1  215  278   75  100   81   64   62   20   21   18
##                       time_spend_company
## promotion_last_5years   21   22
##                     0   61    2
##                     1    6    2
```

Employees who have been in the company for 7-9 years have been awarded the most number of promotions in the last 5 years and as the number of years spent at the company increases, the number of promotions decreases.

Department wise salary

```
dept_sal <-xtabs(~Department+salary,data=hr.df)
dept_sal
```

```
##                  salary
## Department         high  low medium
##   Finance           295 1162   1043
##   Human Resources   280 1126   1094
##   IT                277 1176   1047
##   Operations        284 1180   1036
##   Sales             269 1147   1084
##   Warehouse         255 1188   1056
```

The finance department has the highest number of high-wage workers whereas the warehouse department has the highest number of low-wage workers.

Promotion in last 5 years vs salary

```
Prom_sal <-xtabs(~promotion_last_5years+salary,data=hr.df)
Prom_sal
```

```
##                       salary
## promotion_last_5years high  low medium
##                     0  715 3884   3284
##                     1  945 3095   3076
```

Employees getting the maximum promotions in the last 5 years have had a low to medium increase in their salary, with very few of them promoted with a high wage

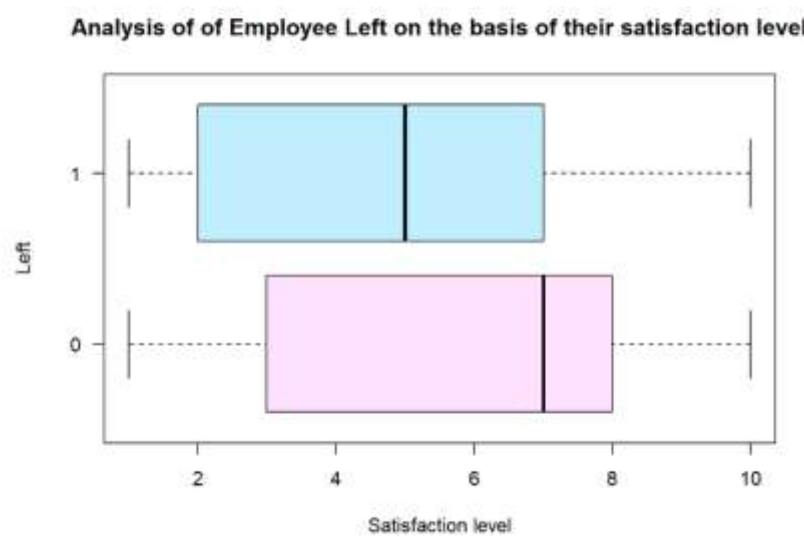
Box Plot describing relationship between Salary and Emp_Satisfaction

```
boxplot(Emp_Satisfaction ~salary,data=hr.df, horizontal=TRUE,
          ylab="Salary Level", xlab="Satisfaction level", las=1,
          main="Analysis of Salary of Employee on the basis of their satisfaction level",
          col=c("azure2","gold","darksalmon")
          )
```

Employees in the higher wage category have more satisfaction levels than lower wage level employees.

Box Plot describing relationship between Left_company and Emp_Satisfaction

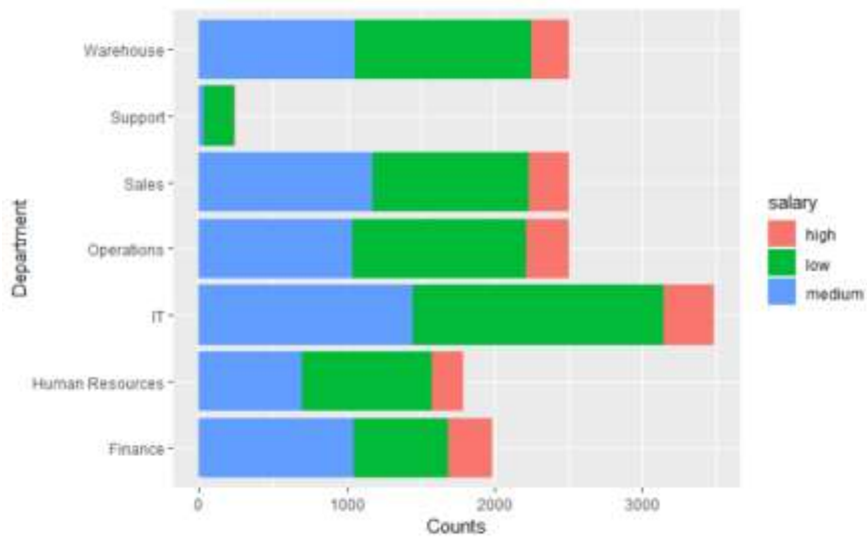
```
boxplot(Emp_Satisfaction ~left_Company, data=hr.df, horizontal=TRUE,
          ylab="Left", xlab="Satisfaction level", las=1,
          main="Analysis of of Employee Left on the basis of their satisfaction level",
          col=c("thistle1","lightblue1")
          )
```



As it can be seen, employees with lower satisfaction levels tend to leave the company.

Barplot to ascertain the salaries of employees by their department using GGPLOT

```
ggplot(aes(x = Department),data = hr.df ) +
  geom_bar(aes(fill = salary))  +
  xlab('Department') +
  ylab('Counts') +
  coord_flip()
```

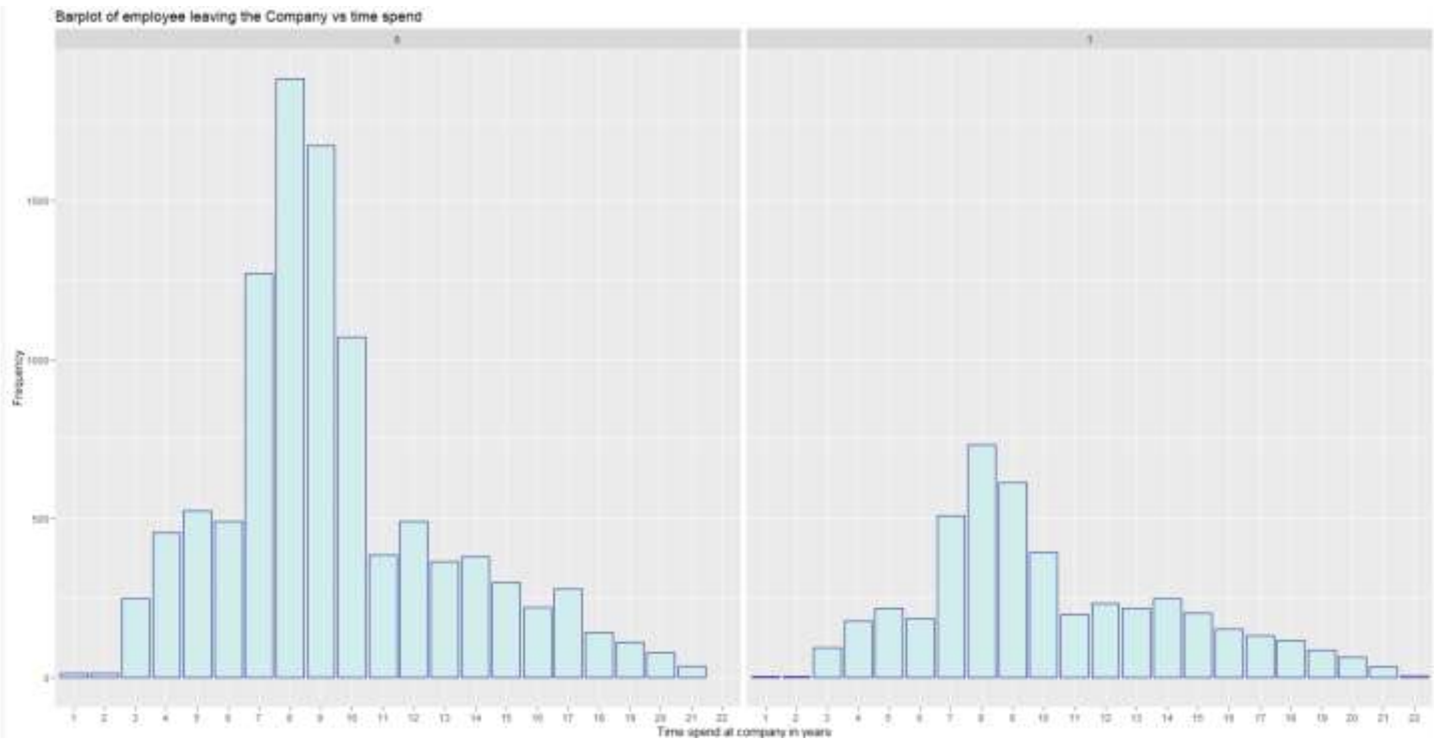


Interpretation

- IT department, having the maximum employees working in shows considerable variability in term of salary distribution.
- Sales, Operation and Warehouse departments have a similar trend in terms of salary distribution.
- Support dept, having the least count of employees working in have majority of the employees in the low salary bracket giving us more insights about potentially being the crowd about to leave the company or not performing well.

Barplot of employees leaving/not-leaving the company vs time spend using GGPLOT

```
ggplot(aes(x = factor(hr.df$time_spend_company)),data = hr.df) +
  geom_bar(fill = 'lightcyan2',color='navy') +
  xlab("Time spend at company in years") +
  ylab("Frequency")+
  labs(title = "Barplot of employee leaving the Company vs time spend")  +
  facet_wrap(~left_Company)
```



Interpretation

- From the second plot above that represents the employees having left the company, it is evident that employees tend to leave a company after spending 7-10 years with average being 8 years
- Very less number of employees leave the company within the first 2 years of joining
- There are employees who after spending 11-15 years leave the company, something we will figure out in the next chart

- From the first plot, we see majority of current employees have spent 7-10 years in the company with tough fight between employees having spent 8 years. This bracket might have intense competition in terms of promotion and salary as there are more employees
- Very few employees are in the 20-22 years category that says they belong to the higher bands within the company
- Company might have reduced its recruiting in the past 2 years as shown above with less number of employees having spent 2 years

Table showing department wise promotion

```
hr.df$promotion_last_5years<-factor(hr.df$promotion_last_5years,labels=c('False',"True"))

#Sreading out the data
promotiondf<-hr.df %>% group_by(Department, promotion_last_5years) %>%
  summarise(Count = n())

promotiondf<-promotiondf %>% spread(promotion_last_5years,Count)

#Changing column names
names(promotiondf)<-c("Department","Got No promotion","Promotion")
promotiondf
```

Department <fctr>	Got No promotion <int>	Promotion <int>
Finance	1095	888
Human Resources	988	797
IT	1797	1688
Operations	1282	1218
Sales	1307	1193
Support	107	140
Warehouse	1307	1192

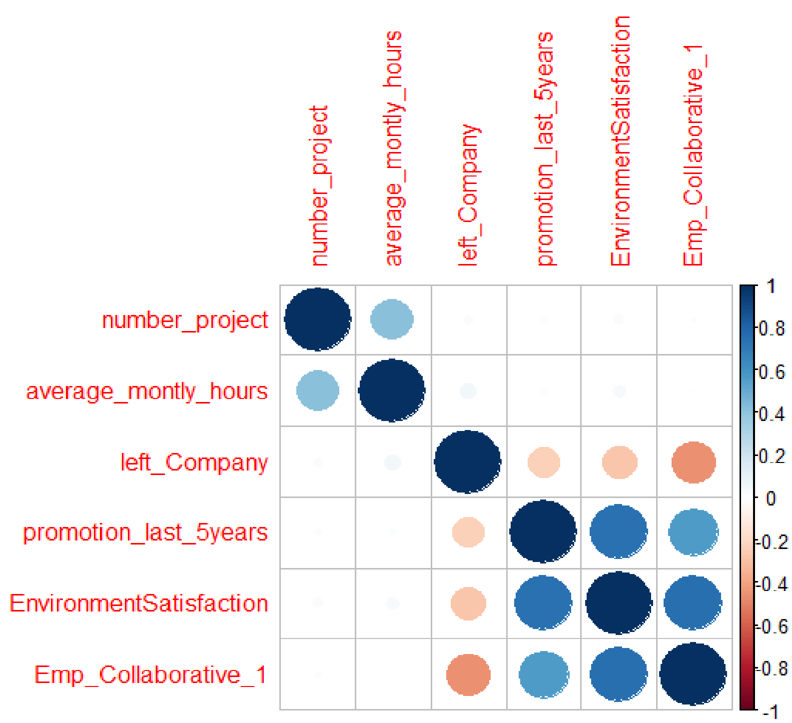
Correlation showing the important factors on which employee satisfaction depends on :

```
HR_correlation1 <- hr.df %>% dplyr::select(number_project,average_montly_hours,time_spend_company,left_Company,promotion_la
st_5years,Emp_Satisfaction)
M <- cor(HR_correlation1)
corrplot(M, method="circle")
```

Interpretation

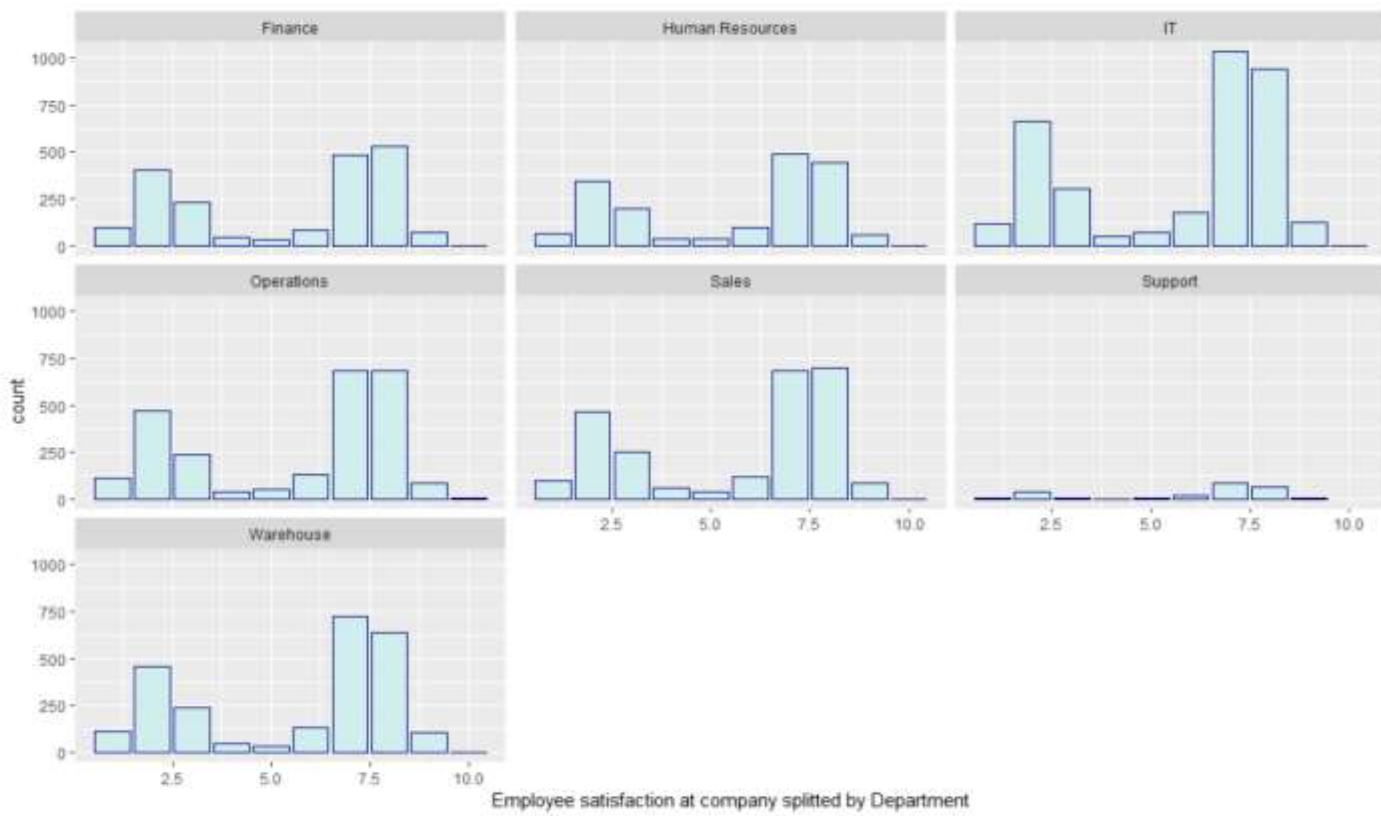
Employee_Satisfaction has a very positive correlation with promotion_received in last 5 years which directly gives us more insights for such employees to stay longer in a company.

Also , the satisfaction levels depend on Emp_Collaborative_1 which describes how collaborative an employee thinks his coworkers are. If an employee has a good relationship with their coworkers , then their satisfaction levels are also high.



Barplot showing department wise Employee_Satisfaction

```
ggplot(aes(x = Emp_Satisfaction),data = hr.df) +
  geom_bar(fill = 'lightcyan2',color='navy') +
  xlab("Employee satisfaction at company splitted by Department") +
  facet_wrap(~Department)
```



Interpretation

- IT department has got the most number of employees falling in both the categories(Satisfied and not satisfied) giving us takeaway that a high number of employees aren't happy with their work.
- We see a bimodal barplot for across departments telling us that employees are either not satisfied; with average between 2-4 and employees satisfied with average being 7-8.
- Very less employees are highly satisfied across the departments.

WHY DO GOOD EMPLOYEES LEAVE?

```
#people that left
leavers = subset(hr.df,hr.df[,19] == 1)

#filter out people with a good last evaluation. Taking rating 7 as the threshold
leaving_performers <- subset(leavers,leavers[,15] > 7)

#Analyzing reasons for such employees to have left the company
```

Are the number of projects employees assigned to the reason?

```
#Was number of projects, they were assigned to the reason?
table(leaving_performers$left_Company,leaving_performers$number_project)
```

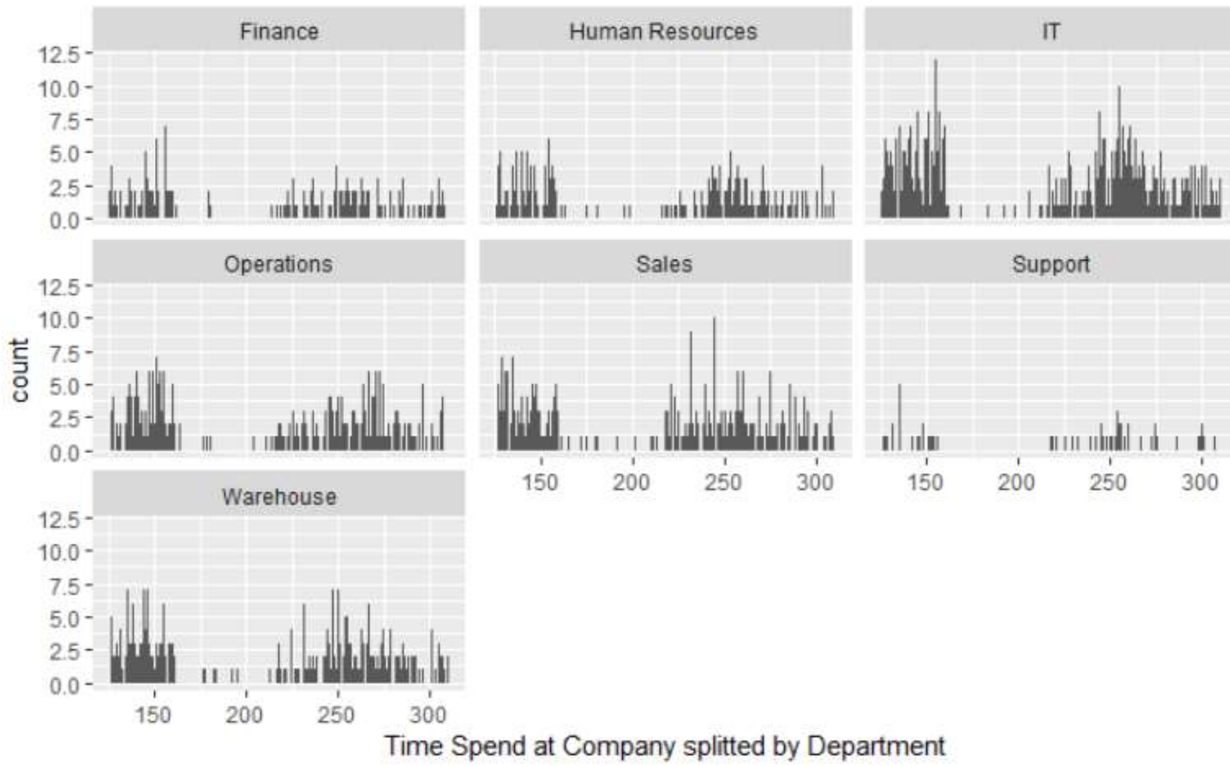
```
                 Good_Emp_leavers
No_of_projects
               2 646
               3  33
               4 239
               5 325
               6 350
               7 145
```

Interpretation

- The data shows that employees have left more when they were assigned to less number of projects.
- Probably , they felt that they were being under-utilized in the company and left the company.

Or the average monthly hours they work for across projects?

```
#or was it the average monthly hours they worked, the reason?
ggplot(aes(x = average_montly_hours),data = leaving_performers) +
  geom_bar() +
  xlab("Time Spend at Company splitted by Department") +
  facet_wrap(~Department)
```

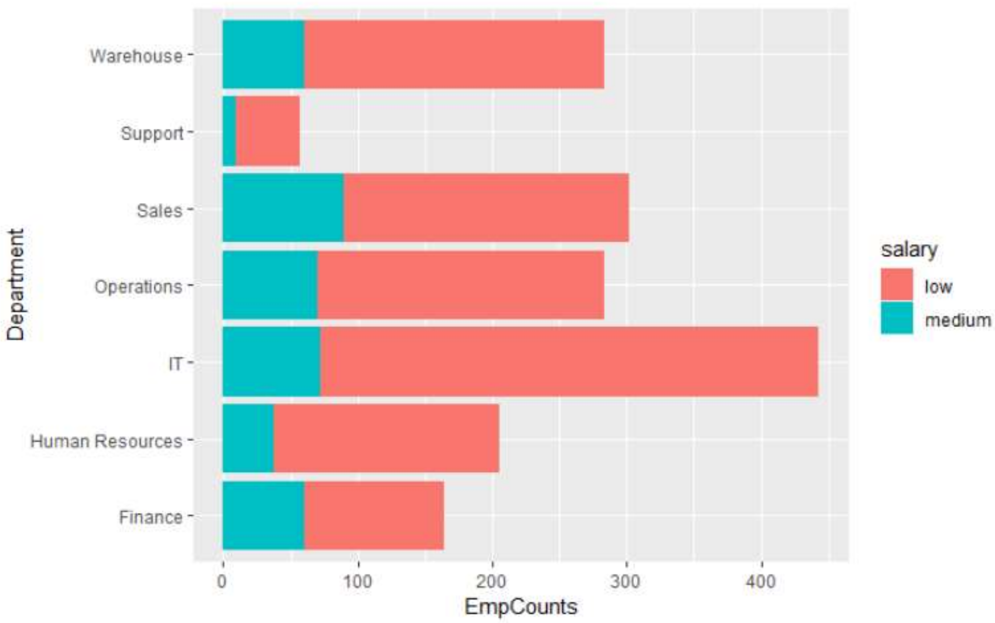


Interpretation

- Average monthly hours are the highest for multiple departments as shown above.
- In terms of the number of employees, IT department has the maximum count of employees working for more than 250 hours, suggesting a certain kind of load they have working across multiple projects as we have seen in the previous chart.

Probably salary could reveal more?

```
#or may be it was Salary
ggplot(aes(x = Department),data = leaving_performers ) +
  geom_bar(aes(fill = salary))  +
  xlab('Department') +
  ylab('EmpCounts') +
  coord_flip()

Sal_leavers <- xtabs(~Department+salary, data = leaving_performers)
Sal_leavers
```



Interpretation

- Salary gives us a final picture in concluding that last evaluation or a promotion gives no major boost in terms of financial satisfaction for any employee, also clearly seen from the table and chart above.
- Not a single employee having left got a high salary package despite having an excellent performance review.

```
                 salary
Department        high low medium
  Finance            0 104     60
  Human Resources    0 168     37
  IT                 0 371     72
  Operations         0 214     70
  Sales              0 213     89
  Support            0  48      9
  Warehouse          0 223     60
```

Conclusion is that these employees are highly valuable assets that should not have been lost.

MODEL ANALYSIS

After running descriptive diagnostics on the data, we move on to predictive analytics. In this section we aim to answer the questions that will help the management to mitigate the attrition rate of employees. This analysis is important in the sense that it assists HR personnel to analyze the factors that drive employees out of the organization and to take proactive actions in retaining employees.

Principal Component Analysis:

The central idea of using principal component analysis (PCA) in our project is to reduce the dimensionality of the HR Analytics data set, which consists of many interrelated variables, while retaining as much as possible of the variation present in the data set. This is achieved by normalizing the data and transforming to a new set of variables, the principal components (PCs), which are uncorrelated.

```
pcs.cor <- prcomp(na.omit(HR.df, scale. = T))
summary(pcs.cor)
pcs.cor$rot
```

```
> summary(pcs.cor)
Importance of components:
                           PC1     PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9
Standard deviation     49.9600 8.56978 3.54386 2.73303 2.15958 1.64728 1.50228 1.41017 1.33074
Proportion of Variance  0.9558 0.02812 0.00481 0.00286 0.00179 0.00104 0.00086 0.00076 0.00068
Cumulative Proportion   0.9558 0.98396 0.98877 0.99163 0.99342 0.99446 0.99532 0.99608 0.99676
                          PC10    PC11    PC12    PC13    PC14    PC15    PC16    PC17   PC18
Standard deviation     1.20984 1.11238 1.08261 0.98214 0.95673 0.87722 0.74884 0.66711 0.4995
Proportion of Variance 0.00056 0.00047 0.00045 0.00037 0.00035 0.00029 0.00021 0.00017 0.0001
Cumulative Proportion  0.99732 0.99780 0.99825 0.99861 0.99897 0.99926 0.99947 0.99965 0.9997
                          PC19    PC20    PC21    PC22    PC23    PC24    PC25
Standard deviation     0.48399 0.35423 0.33569 0.28350 0.25001 0.18461 0.16517
Proportion of Variance 0.00009 0.00005 0.00004 0.00003 0.00002 0.00001 0.00001
Cumulative Proportion  0.99983 0.99988 0.99992 0.99995 0.99998 0.99999 1.00000
```

```
> pcs.cor$rot
                                 PC1          PC2           PC3           PC4          PC5           PC6
Rising_Star              1.281663e-03 -0.1436166164 -0.0163924736  0.1374750334 -0.020230321  0.1122931887
Will_Relocate           -2.670791e-05 -0.0006896328 -0.0006867222  0.0007351206 -0.001795376  0.0002485325
Critical                 4.854457e-04 -0.0482510857 -0.0015756848  0.0330211723 -0.004324663  0.0132741492
Trending.Perf            1.475495e-03 -0.2320307400 -0.0012389284  0.3911912501 -0.368646452 -0.2104406472
Talent_Level             2.499956e-03 -0.2685663449 -0.0294871941  0.2514904251 -0.019072225  0.2720712354
Percent_Remote           1.232006e-03 -0.0022532212  0.0043643362 -0.0054443891  0.014242131  0.0068877737
EMP_Sat_OnPrem_1        -9.503701e-04 -0.1861249034  0.0268573681  0.3554778462 -0.354393837 -0.2452525480
EMP_Sat_Remote_1         2.823000e-03 -0.1674811630  0.0023045139  0.2284557575 -0.195000131  0.0029758845
EMP_Engagement_1         9.273460e-04 -0.1090923634 -0.0054981219  0.1224089120 -0.016962904  0.0506553903
last_evaluation          1.863458e-03 -0.2115431390 -0.0119159007  0.2313454207 -0.097266754  0.1268316660
number_project           1.030784e-02 -0.0007926080  0.0046678936  0.0078119530  0.025022709  0.0485537903
average_montly_hours     9.999196e-01  0.0056691208 -0.0006267445 -0.0024753687 -0.002492241 -0.0023617273
time_spend_company      -2.212766e-04  0.2176919760  0.9606812716  0.1501938227 -0.010813383  0.0273641193
left_Company             5.476736e-04  0.0146041665 -0.0125233371  0.0640595844 -0.001697372  0.0594831970
promotion_last_5years    1.277121e-04 -0.0415602283  0.0089850588 -0.0017138153  0.047207622 -0.0292508200
salary                   3.648783e-04  0.0078741859 -0.0052837247  0.0473507591  0.013394506  0.0422424041
Gender                   1.721712e-04  0.0006805467  0.0001475460 -0.0175787470 -0.034804905  0.0202709067
Emp_Work_Status2         1.662061e-03 -0.2898324009  0.0770892938 -0.1184747071  0.080122186 -0.1381294394
Emp_Identity             2.498228e-03 -0.3414849756  0.0958809479 -0.0255315047  0.246755092  0.1963246019
Emp_Role                 2.795385e-03 -0.3439698793  0.0942384185 -0.0245963085  0.232295860  0.2405679864
Emp_Position             2.587152e-03 -0.3489323486  0.0981217740 -0.0449969272  0.210174309  0.1461445952
Emp_Title                9.122571e-04 -0.1149280027  0.0300278599  0.1263305558  0.462562692 -0.6609429279
EnvironmentSatisfaction  2.023301e-03 -0.2881581776  0.0802438140 -0.0197215571  0.254369391 -0.0427382532
Emp_Competitive_1       -5.827584e-04 -0.2625324023  0.1510457943 -0.5316460580 -0.437149467  0.1104132416
Emp_Collaborative_1     -4.747312e-04 -0.2765032741  0.1046086253 -0.4135748072 -0.223946338 -0.4388964355
                                  PC7          PC8          PC9          PC10          PC11          PC12          PC13
Rising_Star             -0.0912333143  0.190306797 -0.095508002  1.418468e-01 -1.347391e-02  0.0609528729 -0.0032454546
Will_Relocate            0.0032395063 -0.001270128  0.003298887  4.569262e-03 -3.472264e-03 -0.0084829040 -0.0162238156
Critical                 0.0001428806  0.021465713 -0.002216414 -1.087765e-03 -3.474979e-05  0.0052207656  0.0010580641
Trending.Perf            0.0637916820 -0.020728179  0.321971722 -1.689805e-01  1.197792e-01 -0.6063656801  0.0047138230
Talent_Level            -0.2024572219  0.399210909 -0.220295756  3.410287e-01 -3.051593e-02  0.1828956110 -0.0142405173
Percent_Remote          -0.0056497056  0.007488686  0.038956166 -7.467692e-03 -3.039476e-02  0.0045509865 -0.0016052417
EMP_Sat_OnPrem_1         0.0419057152 -0.566210606 -0.453200087  8.185747e-02 -1.564234e-01  0.2959491892 -0.0094071992
EMP_Sat_Remote_1         0.0494128901  0.168760024  0.580534086 -2.640471e-01  2.873126e-02  0.6473693703 -0.0210224116
EMP_Engagement_1        -0.0594549098  0.157779018 -0.066373916  6.317095e-02 -1.988246e-02 -0.1359758770  0.0132995264
last_evaluation         -0.0914011318  0.283695507 -0.103516709  1.527784e-01  1.809193e-03 -0.2148467129  0.0307233011
number_project           0.0267662155  0.075195275  0.100499656 -1.309746e-01 -9.729063e-01 -0.1117078786 -0.0090778196
average_montly_hours    -0.0005505517 -0.002256520 -0.002843961  2.022031e-03  9.532110e-03  0.0007345420  0.0002245364
time_spend_company      -0.0115033491  0.074002393 -0.018620713  7.446212e-03  9.683437e-03 -0.0002601424 -0.0007139156
left_Company             0.0503629025 -0.031043307  0.021999616 -3.874040e-02  1.941344e-02 -0.0175794209  0.0120707781
promotion_last_5years    0.0014749130 -0.006677781  0.001035093  5.147681e-03 -8.617565e-04  0.0007640406 -0.0124975984
salary                   0.0227083530 -0.011886272  0.005986232 -3.699496e-02  1.091491e-02  0.0030897549  0.0166685292
Gender                  -0.0417521048  0.005537388 -0.003619927  5.268775e-05 -4.613058e-03 -0.0063421808 -0.0366950087
Emp_Work_Status2         0.8552133528  0.240628823 -0.152835954  6.684154e-02  1.212124e-02  0.0251723710 -0.1322801441
Emp_Identity            -0.0048451339 -0.204681471  0.043060384 -1.519780e-01  8.515084e-04 -0.0054075158  0.3399579742
Emp_Role                -0.1872067547 -0.229301442  0.004572779 -2.186725e-01  4.467462e-02 -0.0677931976 -0.7505762829
Emp_Position            -0.0500324408 -0.167338594  0.007597324 -1.163098e-01  3.519154e-02 -0.0317862628  0.5438550346
Emp_Title               -0.2103710136  0.003931919  0.240533596  4.093696e-01 -5.873708e-02  0.0341348591 -0.0435207886
EnvironmentSatisfaction  0.0782851002 -0.082122395  0.030196334 -8.702311e-03  4.107210e-03 -0.0105346656 -0.0115732968
Emp_Competitive_1       -0.0507275300 -0.184581666  0.310918115  5.260886e-01 -6.131052e-02 -0.0265742056 -0.0287416214
Emp_Collaborative_1     -0.3208140131  0.339838282 -0.296476866 -4.215975e-01  7.405363e-03  0.0512145198  0.0277576873
```

	PC14	PC15	PC16	PC17	PC18	PC19
Rising_Star	0.0064310777	-0.085706344	0.0081810486	-8.058570e-03	7.626720e-03	-0.0111966367
Will_Relocate	-0.0060816154	0.008833385	-0.0021507529	1.758535e-02	-9.993234e-01	-0.0175617460
Critical	-0.0009790745	0.008800712	-0.0025370580	-1.054601e-02	-1.196405e-03	-0.0240577239
Trending.Perf	0.0032584504	-0.300011790	-0.0423867433	2.406938e-02	3.602620e-03	0.0100244745
Talent_Level	-0.0026340191	-0.511096811	-0.0716652328	-4.580641e-04	-8.791277e-03	0.0183009979
Percent_Remote	0.0031815337	0.007597113	0.0022096434	-1.152349e-02	-2.639651e-03	-0.0059226386
EMP_Sat_OnPrem_1	0.0027023594	0.067745440	0.0209894027	1.759541e-03	-7.622675e-04	-0.0019517686
EMP_Sat_Remote_1	-0.0266964297	0.136871659	0.0395456003	2.555990e-02	-2.571410e-03	0.0032660035
EMP_Engagement_1	0.0192853136	0.237113150	0.9225990993	-7.915215e-03	-4.632523e-04	0.0213245021
last_evaluation	0.0447230717	0.745638828	-0.3696928354	-4.828989e-04	7.919617e-03	-0.0018692130
number_project	-0.0315457054	-0.032773383	-0.0263878227	-5.126458e-03	4.170625e-03	0.0032008556
average_montly_hours	0.0005535178	0.000228783	0.0003059456	2.703291e-04	-7.375131e-05	0.0003086971
time_spend_company	0.0003543576	-0.009568014	-0.0022709612	3.095407e-03	-8.010706e-04	-0.0005139261
left_Company	-0.0022452621	-0.007906117	0.0226468381	-2.267796e-01	9.081362e-03	-0.1647722282
promotion_last_5years	-0.0102146635	0.001510134	-0.0091703091	3.280772e-04	3.142037e-03	-0.0147004964
salary	0.0106459040	-0.006563071	-0.0108123051	-9.569930e-01	-1.824679e-02	-0.1051813269
Gender	-0.0150153195	-0.004614086	0.0173582354	1.407288e-01	1.864810e-02	-0.9774058135
Emp_Work_Status2	-0.0298547882	-0.016926284	0.0017299312	4.753691e-03	4.365682e-03	-0.0307328754
Emp_Identity	0.7434781552	-0.036845407	-0.0048163792	3.801290e-02	-1.010434e-02	-0.0225134977
Emp_Role	-0.1243839104	0.026296322	-0.0054354786	-1.211179e-03	1.128468e-02	0.0338174796
Emp_Position	-0.6506226627	0.007115696	-0.0037537459	1.719507e-02	-5.162028e-03	-0.0089736498
Emp_Title	0.0147386731	0.027846341	-0.0110481138	-4.536728e-02	1.956095e-03	-0.0356354498
EnvironmentSatisfaction	-0.0091101967	0.002358229	-0.0094995744	-2.321927e-05	1.162511e-03	-0.0094027492
Emp_Competitive_1	0.0153864178	0.024300633	0.0297157523	-6.034739e-02	4.070529e-03	0.0154403703
Emp_Collaborative_1	0.0488666891	-0.025267364	-0.0153744518	-5.927028e-02	-6.149473e-03	0.0031933498

	PC20	PC21	PC22	PC23	PC24	PC25
Rising_Star	0.0821246435	-0.0407854691	9.128701e-01	0.0097572808	1.398057e-01	0.0203169844
Will_Relocate	0.0108063569	-0.0048297936	7.821654e-03	0.0002828713	3.153480e-03	0.0028765870
Critical	0.0612814187	-0.0783928025	1.239248e-01	0.0501108734	-9.707898e-01	-0.1578293063
Trending.Perf	-0.0367541117	0.0033560056	8.909557e-03	-0.0018763151	1.485609e-02	-0.0011178085
Talent_Level	-0.0162260563	0.0141992300	-3.490414e-01	-0.0128936358	-1.474011e-02	-0.0116207830
Percent_Remote	-0.0131999100	0.0540311800	4.737684e-03	-0.0115128083	1.552861e-01	-0.9846197447
EMP_Sat_OnPrem_1	-0.0056597916	0.0054924132	-1.431864e-02	-0.0006405728	1.112428e-02	-0.0225080239
EMP_Sat_Remote_1	-0.0199553549	0.0001649381	-2.485230e-02	-0.0025674746	1.489210e-02	0.0277271526
EMP_Engagement_1	-0.0263016864	-0.0018218227	-7.287498e-02	-0.0055343793	1.039552e-03	0.0021166937
last_evaluation	0.0012226162	0.0066524756	-1.042676e-01	-0.0058211951	1.894929e-02	0.0029308215
number_project	0.0077292139	-0.0020385762	5.050422e-04	0.0033185115	-3.081073e-03	0.0344911299
average_montly_hours	-0.0003459031	-0.0001593214	2.091157e-05	-0.0000505129	-6.842805e-05	0.0007486444
time_spend_company	0.0020079655	-0.0034252916	5.198551e-03	0.0008277286	-4.416523e-03	0.0014961411
left_Company	0.9316993062	-0.1533259819	-9.974106e-02	0.0137523584	6.829062e-02	-0.0072832497
promotion_last_5years	-0.1690317097	-0.9415897557	-4.194364e-02	0.2633297366	8.159309e-02	-0.0391663144
salary	-0.2517383656	0.0425139193	1.513871e-02	-0.0021579182	-5.715334e-03	0.0168790747
Gender	-0.1317156390	0.0348194330	-5.395209e-03	-0.0099710265	8.605846e-03	0.0093073109
Emp_Work_Status2	-0.0037820579	0.0730369984	-3.080539e-03	0.1652305229	1.680691e-02	-0.0031614538
Emp_Identity	-0.0087989053	0.0624439194	-4.107275e-03	0.1763818219	1.347026e-02	0.0124269388
Emp_Role	0.0089907628	0.0795569136	-4.709165e-03	0.1753872338	1.268904e-02	0.0108304055
Emp_Position	-0.0058752217	0.0731227144	-1.564830e-03	0.1717339069	1.572997e-02	0.0059924310
Emp_Title	0.0560465212	0.0881930528	-1.018267e-02	0.1649636486	9.323549e-04	0.0146918905
EnvironmentSatisfaction	-0.0202040386	-0.2166717447	4.832883e-03	-0.8840159284	-1.869172e-02	0.0003513233
Emp_Competitive_1	0.0329677714	-0.0136412399	1.674579e-03	0.0006555995	-4.640870e-03	0.0066579882
Emp_Collaborative_1	0.0648664325	-0.0011321170	-1.178336e-02	-0.0031376582	7.294277e-03	-0.0062529209

After running PCA, we find that the first PC retained almost 95.4% of the variation present in all the original variables. Also, in PC1, average_montly_hours is the most significant variable .

Strengths: PCA is a versatile technique that works well in practice. It's fast and simple to implement, which means you can easily test algorithms with and without PCA to compare performance. In addition, PCA offers several variations and extensions (i.e. kernel PCA, sparse PCA, etc.) to tackle specific roadblocks.

Weaknesses: The new principal components are not interpretable, which may be a deal-breaker in some settings. In addition, you must still manually set or tune a threshold for cumulative explained variance

Running various models to answer the below questions-

A) Will the employee leave the company? Which employee?

1.Running Logistic Regression

Logistic regression extends the idea of linear regression to situation where outcome variable is categorical. It is widely used, especially where a structured model is used to explain or predict.

We make a model using logistic regression to predict if the employee will leave the company. We run the algorithm after excluding the “Name”, “Department” and “Geographical location”.

```
#Dataset for Logistic Regression
hr.logit <- hr.df[,5:30]
```

The model is trained on test data that comprises 60% of the total data and validated on the rest.

```
set.seed(13)
#Partitioning data into training (60%) and validation(40%) for logistic regression
train.index <- createDataPartition(hr.logit$left_Company , p = 0.6, list = FALSE)
train.df <-hr.logit[train.index,]
valid.df <- hr.logit[-train.index,]
```

```
#Logistic Regression for Leaving the company
lc<- glm(left_Company ~ ., data = train.df, family = "binomial")
options(scipen=999)
summary(lc)
```

Output:

```
Call:
glm(formula = left_Company ~ ., family = "binomial", data = train.df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7342  -0.5892  -0.2612   0.5167   3.7555
```

Deviance residuals is the measure of how far the line of regression is from the actual point. A perfect fit of the given point equates to 0 as the log (1) is zero. However, this never occurs.

```
Coefficients:
                          Estimate Std. Error z value          Pr(>|z|)    
(Intercept)             -2.3982829  0.5005114  -4.792    0.000001654029320 ***
RoleLevel 1              0.1207421  0.2936491   0.411             0.680942    
RoleLevel 2-4           -0.1674366  0.2361813  -0.709             0.478366    
RoleManager             -0.3027137  0.1820273  -1.663             0.096310 .  
RoleSenior Director      0.0094415  0.2176361   0.043             0.965397    
RoleSenior Manager      -0.3145143  0.1697704  -1.853             0.063942 .  
RoleVP                  -0.2369644  0.3343716  -0.709             0.478519    
Rising_Star              0.2010943  0.1027328   1.957             0.050295 .  
Will_Relocate           -0.0991628  0.0606357  -1.635             0.101968    
Critical                 1.1231274  0.1486025   7.558    0.000000000000041 ***
Trending.Perf            0.2668123  0.0229458  11.628 < 0.0000000000000002 ***
Talent_Level            -0.2822166  0.0451914  -6.245    0.000000000424017 ***
Percent_Remote          -0.3275806  0.1934508  -1.693             0.090388 .  
EMP_Sat_OnPrem_1         0.0037499  0.0203492   0.184             0.853797    
EMP_Sat_Remote_1         0.0884219  0.0246097   3.593             0.000327 ***
EMP_Engagement_1         0.1234475  0.0420940   2.933             0.003361 ** 
last_evaluation         -0.1803273  0.0339934  -5.305    0.000000112817001 ***
number_project          -0.0312286  0.0278120  -1.123             0.261503    
average_montly_hours     0.0031399  0.0006985   4.495    0.000006957069492 ***
time_spend_company       0.0050750  0.0211959   0.239             0.810769    
promotion_last_5years1 -0.3894301  0.0944767  -4.122    0.000037564886538 ***
salarylow                3.8743370  0.2522385  15.360 < 0.0000000000000002 ***
salarymedium             2.4251708  0.2529390   9.588 < 0.0000000000000002 ***
GenderM                  0.3320626  0.0633033   5.246    0.000000155791115 ***
Emp_Work_Status2         0.0468876  0.0283098   1.656             0.097674 .  
Emp_Identity             0.0804798  0.0366811   2.194             0.028233 *  
Emp_Role                -0.0067620  0.0355531  -0.190             0.849157    
Emp_Position             0.0942739  0.0365584   2.579             0.009917 ** 
Emp_Title               -0.3753306  0.0307164 -12.219 < 0.0000000000000002 ***
Emp_Satisfaction         0.0753069  0.1089411   0.691             0.489400    
Emp_Competitive_1       -0.2043303  0.0191879 -10.649 < 0.0000000000000002 ***
Emp_Collaborative_1     -0.4885122  0.0205072 -23.821 < 0.0000000000000002 ***
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 11085.5  on 8999  degrees of freedom
Residual deviance:  6887.3  on 8968  degrees of freedom
AIC: 6951.3

Number of Fisher Scoring iterations: 6
```

Interpretation:

Three stars indicate an extremely low P value (approximately 0), it signifies that probability of a dependent variable occurring in a certain way in accordance with the corresponding dependent variable is very low. This suggest that there is relationship between two variables in a way that independent variable largely effects the outcome of the dependent variable.

The predictors with two and three stars can be deemed important for predicting if the employee will leave the company.

Let’s go ahead and try to interpret how the coefficient estimate of “Critical” can be interpreted. The dependent variable here is “Left_Company” with “0” as still in the company and “1” as left the company. The independent variable “Critical” has “0” as not critical to the organization and “1” as critical to the organization. “0” comes first numerically for both the variables, the sequence is important in deciding the sign of coefficient estimates. The positive estimate 1.21 of “Critical” indicates that when the critical value is “0” it proves as a driving factor for the employee to leave the company resulting in “1” of the variable “Left_Company” and when the critical value is “1” it motivates the employee to stay resulting in “0” for variable “Left_Company”.

Based on the above summary and P-values of coefficient estimates it can be concluded that following predictors are important in deciding whether the employee will or will not leave the company. “Critical”, “Trending.perf”, “Talent Level”, “EMP_Sat_Remote_1”, “EMP_Engagement_1”, “last_evaluation”, “average_montly_hours”, “promotion_last_5years1”, “salarylow”, “salarymedium”, “GenderM”, “Emp_Position”, “Emp_Title”, “Emp_Competitive_1” and “Emp_Collaborative_1”

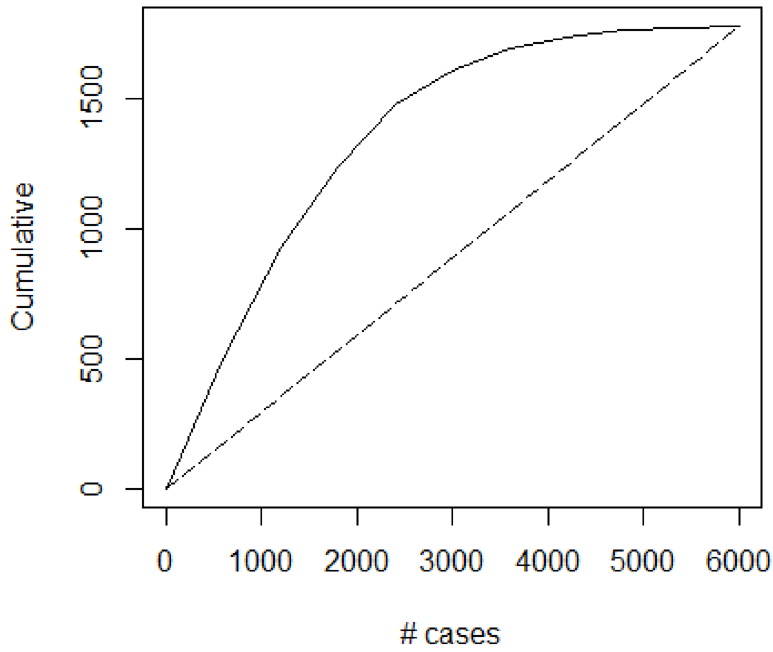
```
#calculate e to the power coefficients
exp(coef(lc))
```

```
         (Intercept)          RoleLevel 1        RoleLevel 2-4          RoleManager
          0.09087386           1.12833383           0.84583021           0.73881058
   RoleSenior Director   RoleSenior Manager             RoleVP          Rising_Star
          1.00948626           0.73014343           0.78901934           1.22274001
        Will_Relocate             Critical        Trending.Perf         Talent_Level
          0.90559529           3.07445428           1.30579529           0.75411031
       Percent_Remote      EMP_Sat_OnPrem_1     EMP_Sat_Remote_1     EMP_Engagement_1
          0.72066517           1.00375690           1.09244891           1.13139059
      last_evaluation       number_project average_montly_hours   time_spend_company
          0.83499685           0.96925401           1.00314484           1.00508790
promotion_last_5years1            salarylow         salarymedium              GenderM
          0.67744287          48.15076605          11.30415983           1.39384004
      Emp_Work_Status2         Emp_Identity             Emp_Role         Emp_Position
          1.04800421           1.08380696           0.99326082           1.09886073
            Emp_Title     Emp_Satisfaction    Emp_Competitive_1  Emp_Collaborative_1
          0.68706211           1.07821503           0.81519304           0.61353855
```

From the above values it is evident that Low salary has the highest impact on employees leaving the company followed by medium salary and criticalness.

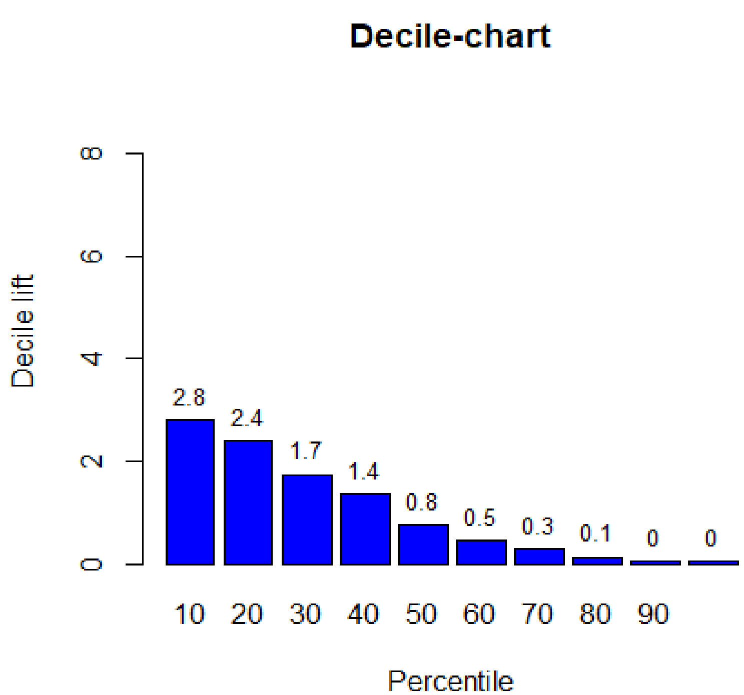
```
### Evaluate Performance of the Logit Model
### Predict propensities

pred <- predict(lc, valid.df[, -15], type = "response")

#Gains
gain <- gains(valid.df$left_Company , pred, groups = 10)
gain

#Lift
plot(c(0,gain$cume.pct.of.total*sum(pred))~c(0,gain$cume.obs),
     xlab = "# cases", ylab = "Cumulative", main = "", type = "l")
     lines(c(0,sum(pred))~c(0, dim(valid.df)[1]), lty = 5)
```

Lift Chart



As seen from the above lift chart, it is evident that the model curve has more area under it compared to the naïve rule represented by the straig ht line.

```r
#decile chart and values
heights <- gain$mean.resp/mean(valid.df$left_Company)
midpoints <- barplot(heights, names.arg = gain$depth,  ylim = c(0,9), col = "blue",
                     xlab = "Percentile", ylab = "Decile lift",
                     main = "Decile-chart")
text(midpoints, heights+0.5, labels=round(heights, 1), cex = 0.8)
```



- Decile chart follows an ideal structure representing maximum variation cover ed in initial deciles.
- First 5 deciles cover 90% of the variation.
- This can be considered as good model where the deciles are decreasing in order from start to end.
- Looking at the first decile, we can say that this model performs 2.8 time better than the one with Naïve rule.

```
#Confusion Matrix
#confusionMatrix(data = pred.scale, reference = valid.df$left_Company)
confusiontable <- table(Predicted = as.numeric(pred.scale) , Actual =as.numeric(valid.df$left_Company))
confusiontable
```

```
          Actual
Predicted    0    1
        0 3772  692
        1  388 1147
```

```
#Accuracy of Logistic Regression on predicting if the employee will leave the company
mean(pred.scale==valid.df$left_Company)*100
```

```
[1] 81.997
```

Strengths: Outputs have a nice probabilistic interpretation, and the algorithm can be regularized to avoid over fitting. Logistic models can be updated easily with new data using stochastic gradient descent.

Weaknesses: Logistic regression tends to underperform when there are multiple or non-linear decision boundaries. They are not flexible enough to naturally capture more complex relationships. Logistic regression attempts to predict outcomes based on a set of independent variables but if we include the wrong independent variable, the model will have little to no predictive value. Logistic regression also works well for predicting categorical outcomes but cannot predict continuous outcomes.

2.Running Linear Discriminant Analysis for the same question and comparing which model is better

Discriminant Analysis is a classical statistic technique used for classification. It also has business data applications and can be used for profiling. Linear discriminant analysis is used to find a linear combination of the predictors that gives maximum separation between the centers of the data. It is also used to minimize the variation within each group of data.

We use the lda() function to perform linear discriminant analysis in R . It finds directions that maximize the separation between the classes , then uses these directions to predict the classifications . These linear directions are linear combinations of predictor variables.

Assumptions :

- Predictors are normally distributed
- Different classes have class-specific means and same variance/covariance structure

```r
#-----Running Linear Discrimant Analysis for the above question------------

#### linear discriminant Regression
lda1 <- lda(left_Company~., data = valid.df, family="binomial")

# predict values
predict1 = predict(lda1, newdata=valid.df[,-c(14)])
names(predict1)

# model Accuracy
table(predict=predict1$class, actual=valid.df$left_Company)
mean(predict1$class == valid.df$left_Company)
```

```
> lda1
Call:
lda(left_Company ~ ., data = valid.df, family = "binomial")

Prior probabilities of groups:
        0         1 
0.6994888 0.3005112 

Group means:
  Rising_Star Will_Relocate  Critical Trending.Perf Talent_Level Percent_Remote EMP_Sat_OnPrem_1 EMP_Sat_Remote_1 EMP_Engagement_1
0    3.682237     0.5090562 0.7283127      7.283445     6.807118      0.6211630         6.784874         7.356530         3.131872
1    3.157544     0.4829882 0.5776627      6.994083     5.729290      0.6126479         6.265533         7.107988         2.795858
  last_evaluation number_project average_montly_hours time_spend_company promotion_last_5years   salary    Gender Emp_Work_Status2
0        7.258977       3.784874             198.1846           9.494757             0.5544963 2.174770 0.4944391         6.941532
1        6.523669       3.936391             205.6938          10.409763             0.3047337 2.724112 0.5029586         5.321006
  Emp_Identity Emp_Role Emp_Position Emp_Title EnvironmentSatisfaction Emp_Competitive_1 Emp_Collaborative_1
0     6.690181 6.703209     6.646965  3.597077                6.119479          5.806800            6.854147
1     4.948964 4.948225     4.826923  2.666420                4.537722          3.298077            3.992604

Coefficients of linear discriminants:
                                 LD1
Rising_Star              0.120969808
Will_Relocate           -0.087993238
Critical                 0.737109421
Trending.Perf            0.195081619
Talent_Level            -0.174570024
Percent_Remote          -0.202886876
EMP_Sat_OnPrem_1        -0.045633966
EMP_Sat_Remote_1         0.072866269
EMP_Engagement_1         0.081307523
last_evaluation         -0.086080890
number_project           0.025375781
average_montly_hours     0.001533263
time_spend_company      -0.003163039
promotion_last_5years   -0.129871169
salary                   0.845830116
Gender                   0.330569548
Emp_Work_Status2         0.025023489
Emp_Identity             0.037243520
Emp_Role                 0.023989720
Emp_Position             0.057227646
Emp_Title               -0.164081989
EnvironmentSatisfaction -0.042140892
Emp_Competitive_1       -0.126546177
Emp_Collaborative_1     -0.312814481
```

The first thing we can see are the Prior probabilities of groups. These probabilities are the ones that already exist in our training data. I.e. 69.94% of your training data corresponds to credit risk evaluated as 0 and 30.06% of your training data corresponds to credit risk evaluated as 1. (I assume that 0 means "risky credits" and 1 means "Non risky Cr edits").

The second thing that you can see are the Group means, which are the average of each predictor within each class. These values could suggest that the variable Left_Company might have a slightly greater influence on risky credits (69.9 4) than on non-risky credits (30.06).

```
> names(predict1)
[1] "class"     "posterior" "x"
```

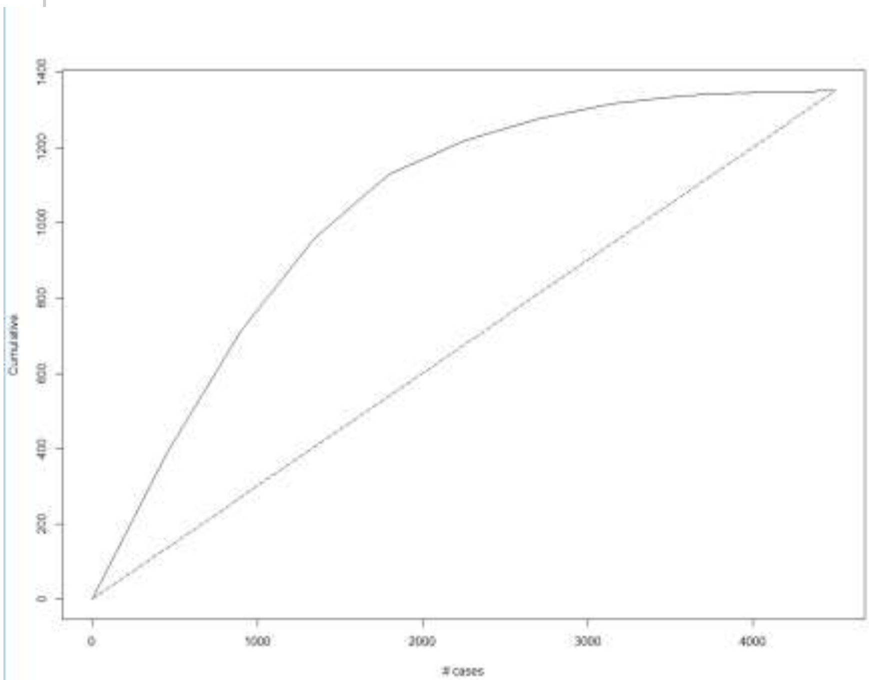
If we call names(predict1) it uses a leave-one-out cross-validation and returns a named list with components:

- class: the Maximum a Posteriori Probability (MAP) classification (a factor)
- posterior: posterior probabilities for the classes. Posterior has two columns with says the probability of that element being "0" and next says the probability of that element being "1"

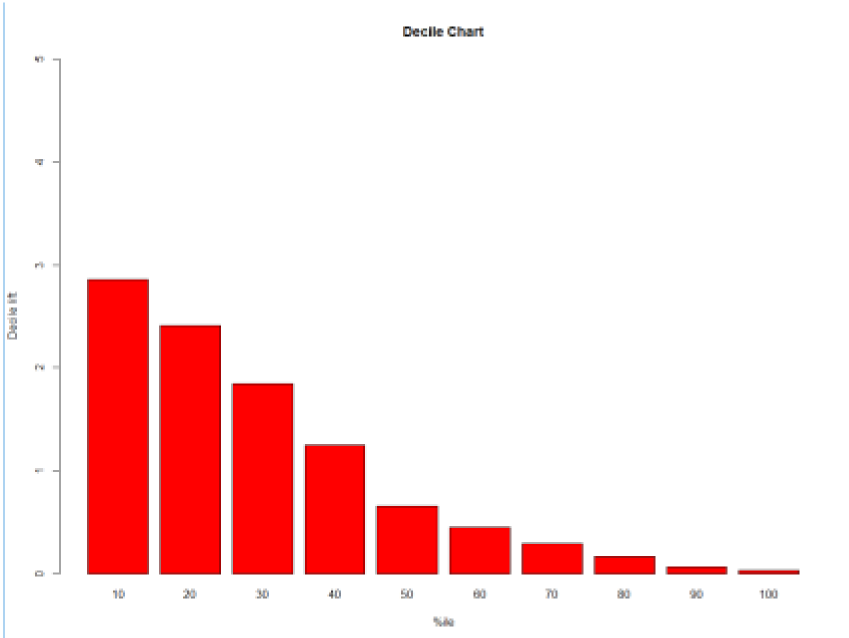
There is also a predict method implemented for lda objects. It returns the classification and the posterior probabilities of the new data based on the Linear Discriminant model. Below, I use half of the dataset to train the model and the other half is used for predictions.

```
> names(predict1)
[1] "class"     "posterior" "x"
> # model Accuracy
> table(predict=predict1$class, actual=valid.df$left_Company)
       actual
predict    0    1
      0 2852  483
      1  295  869
> mean(predict1$class == valid.df$left_Company)
[1] 0.8270727
```

```
#gains lift and decile
gain1 <- gains(valid.df$left_Company, predict1$posterior[,2], groups = 10)

### Plot Lift Chart
plot(c(0,gain1$cume.pct.of.total*sum(valid.df$left_Company))~c(0,gain1$cume.obs),
     xlab = "# cases", ylab = "Cumulative", main = "", type = "l")
lines(c(0,sum(valid.df$left_Company))~c(0, dim(valid.df)[1]), lty = 5)

### Plot decile-wise chart
Decile <- gain1$mean.resp/mean(valid.df$left_Company)
Decile1 <- barplot(Decile, names.arg = gain1$depth,  ylim = c(0,5), col = "red",
                   xlab = "%ile", ylab = "Decile lift",
                   main = "Decile Chart")
```



As seen from the lift chart, it is evident that the model curve has more area under it compared to the naïve rule represented by the straight line.



- Decile chart follows an ideal structure representing maximum variation covered in initial deciles.
- First 5 deciles cover 90% of the variation.
- This can be considered as good model where the deciles are decreasing in order from start to end.
- Looking at the first decile, we can say that this model performs 2.9 time better than the one with Naïve rule

Interpretation :

From both the Algorithms we can say that Linear Discriminant Analysis is much suited and most appropriate regression method thought its accuracy levels are almost near

LDA – 82.70

LR – 81.99

Advantages of Linear Discriminant analysis over Logistic regression (LR)

- **LR**: Based on Maximum likelihood estimation. **LDA**: Based on Least squares estimation; equivalent to linear regression with binary predictand (coefficients are proportional and R-square = 1-Wilk's lambda).
- **LR**: Estimates probability (of group membership) immediately (the predictand is itself taken as probability, observed one) and conditionally. **LDA**: estimates probability mediately (the predictand is viewed as binned continuous variable, the discriminant) via classificatory device (such as naive Bayes) which uses both conditional and marginal information.
- **LR**: Not so exigent to the level of the scale and the form of the distribution in predictors. **LDA**: Predictors desirably interval level with multivariate normal distribution.
- **LR**: No requirements about the within-group covariance matrices of the predictors. **LDA**: The within-group covariance matrices should be identical in population.
- **LR**: Not so sensitive to outliers. **LDA**: Quite sensitive to outliers.
- **LR**: Younger method. **LDA**: Older method.
- **LR**: Usually preferred, because less exigent / more robust. **LDA**: With all its requirements met, often classifies better than BLR (asymptotic relative efficiency 3/2 time higher then).

Strengths : Outputs have a nice probabilistic interpretation, and the algorithm can be regularized to avoid overfitting Regularization is a technique for penalizing large coefficients in order to avoid overfitting, and the strength of the penalty should be tuned.

Weakness : Logistic regression tends to underperform when there are multiple or non-linear decision boundaries. They are not flexible enough to naturally capture more complex relationships.

Which Employee will leave the company (Let us pick up the glm model from above and continue)

```
#Creating a data frame to structure the prediction output in a table
predAboutToLeave <- data.frame(pred)

#Add a column to the predAboutToLeave dataframe containing the performance
predAboutToLeave$performance = valid.df$Trending.Perf
predAboutToLeave

#Find out which valuable employee has the most proability of leaving the company
predAboutToLeave$Valuable_emp <- predAboutToLeave$performance * predAboutToLeave$pred

#Sorting
orderpred <- predAboutToLeave[order(predAboutToLeave$Valuable_emp,decreasing = TRUE),]

#Displaying the top 20 records
orderpred <- head(orderpred, n=20)
orderpred
```

	pred <dbl>	performance <int>	Valuable_emp <dbl>
13635	0.9852981	10	9.852981
11755	0.9833011	10	9.833011
11482	0.9829926	10	9.829926
11418	0.9800850	10	9.800850
11203	0.9793362	10	9.793362
2844	0.9788722	10	9.788722
14446	0.9779275	10	9.779275
14658	0.9778931	10	9.778931
5086	0.9772691	10	9.772691
9031	0.9761691	10	9.761691

1-10 of 20 rows Previous 1 2 Next

Interpretation:

We got the list of first 20 employees that the company should retain.

After grouping them per department we could email the different managers to tell them which valuable employees might leave soon.

Managers can ignore if the employees have already left and can be cautious in case they have not.

B) What is the likelihood of employees getting a promotion?

1.Running Linear Regression

Linear regression is the most basic and commonly used predictive analysis. Regression estimates are used to describe data and to explain the relationship between one dependent variable and one or more independent variables. At the center of the regression analysis is the task of fitting a single line through a scatter plot. It c onsists of 3 stages: 1) analyzing the correlation and directionality of the data, 2) estimating the model, i.e., fitting t he line, and 3) evaluating the validity and usefulness of the model.

We run the linear regression algorithm on non-categorical variables keeping "Rising_Star" as the dependent variable. The model is trained on test data that comprises 60% of the total data and validated on the rest.

```
#Partitioning data into training (60%) and validation(40%) for linear regression on "Rising_Star"
train.lm.rs.index <- createDataPartition(hrform.df$Rising_Star , p= 0.6, list = FALSE)
train.linear.rs <-hrform.df[train.lm.rs.index,]
valid.linear.rs <- hrform.df[-train.lm.rs.index,]

# Linear Regression for Rising Star
hr.rise <- lm(Rising_Star ~ ., data = train.linear.rs)
summary(hr.rise)
```

Interpretation

The significant coefficients (P Value two and three stars) for Rising_Star are:

Critical: Positive coefficient signifies that if the employee is critical ("1") the likely hood of promotion ("Rising_Star) a lso increases in number (1 through 5). For every one-unit change in Critical value, the independent variable is affected to change +0.239

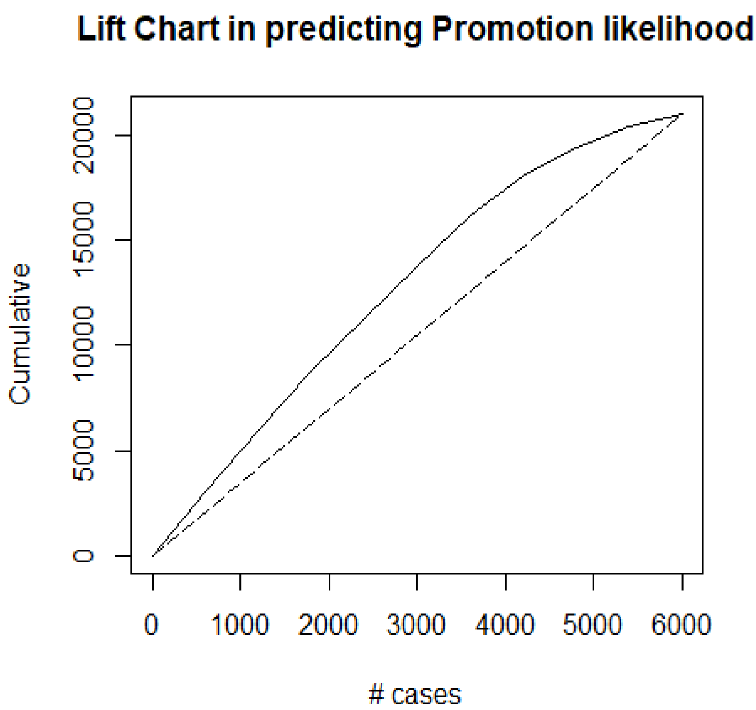
Trending.perf: For every unit change in Trending.perf, there is negative 0.0082 effect on Rising_Star.
Talent Leve: For every unit change in Trending.perf, there is positive 0.3473 effect on Rising_Star.

Similarly, variables EMP_SAT_OnPRem_1, EMP_SAT_Remote1, EMP_Engagement_1, last_Evaluation, number_projec ts and Emp_Collaborative_1 significantly determine the output of Rising_Star.

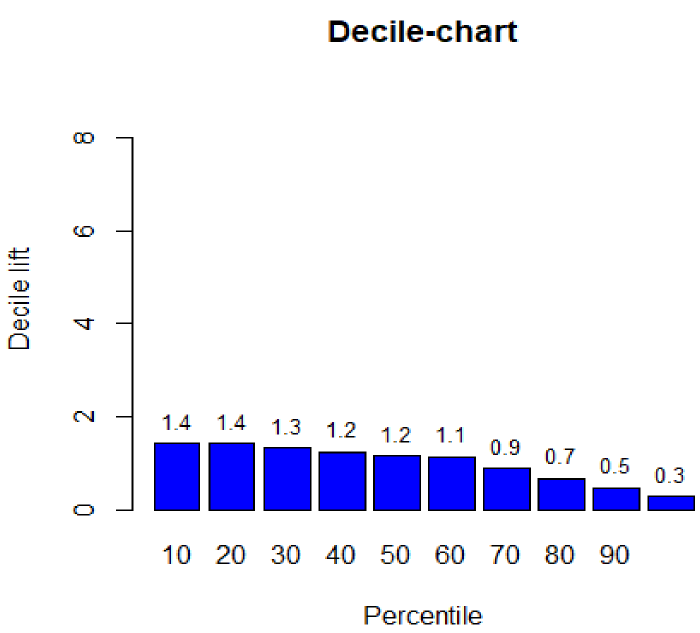
Adjusted R square value of 0.9528 can be considered as an excellent number exhibiting that approximately 95% of the variation in Rising_Star variable is captured by the input variables.

```
Coefficients:
                         Estimate  Std. Error t value             Pr(>|t|)    
(Intercept)           -0.160227851 0.061429812  -2.608             0.009114 ** 
Role                   0.012429008 0.006509563   1.909             0.056249 .  
Will_Relocate         -0.007068016 0.006330444  -1.117             0.264233    
Critical               0.239053981 0.016852727  14.185 < 0.0000000000000002 ***
Trending.Perf         -0.008287284 0.002465108  -3.362             0.000778 ***
Talent_Level           0.347302786 0.002816608 123.305 < 0.0000000000000002 ***
EMP_Sat_OnPrem_1       0.016147193 0.002132542   7.572   0.0000000000000404 ***
EMP_Sat_Remote_1       0.020028255 0.002591444   7.729   0.0000000000000120 ***
EMP_Engagement_1       0.072012753 0.004049837  17.782 < 0.0000000000000002 ***
last_evaluation        0.104104542 0.003335761  31.209 < 0.0000000000000002 ***
number_project         0.000043642 0.002831846   0.015             0.987704    
average_montly_hours   0.000008867 0.000070247   0.126             0.899558    
time_spend_company    -0.001456099 0.002124007  -0.686             0.493019    
left_Company           0.016929032 0.008713745   1.943             0.052072 .  
promotion_last_5years1 0.003809789 0.009687133   0.393             0.694119    
salary                -0.001509074 0.005149479  -0.293             0.769488    
Gender                 0.000446806 0.006553943   0.068             0.945649    
Emp_Work_Status2      -0.004629775 0.002995547  -1.546             0.122248    
Emp_Identity           0.004083190 0.003668345   1.113             0.265701    
Emp_Role               0.000410334 0.003556289   0.115             0.908144    
Emp_Position          -0.000382038 0.003662990  -0.104             0.916936    
Emp_Title              0.007179571 0.003050805   2.353             0.018627 *  
Emp_Satisfaction       0.004933944 0.011343134   0.435             0.663593    
Emp_Competitive_1     -0.004487523 0.001956697  -2.293             0.021847 *  
Emp_Collaborative_1    0.007733397 0.002087620   3.704             0.000213 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2998 on 8975 degrees of freedom
Multiple R-squared:  0.9529,	Adjusted R-squared:  0.9528 
F-statistic:  7574 on 24 and 8975 DF,  p-value: < 0.00000000000000022
```

```r
pred.linear.rs <- predict(hr.rise, valid.linear.rs)

#Gains
gain.linear.rs <- gains(valid.linear.rs$Rising_Star , pred.linear.rs, groups = 10)
gain.linear.rs

#Lift
plot(c(0,gain.linear.rs$cume.pct.of.total*sum(pred.linear.rs))~c(0,gain.linear.rs$cume.obs),
     xlab = "# cases", ylab = "Cumulative", main = "", type = "l")
lines(c(0,sum(pred.linear.rs))~c(0, dim(valid.linear.rs)[1]), lty = 5)

#decile chart and values
heights <- gain.linear.rs$mean.resp/mean(valid.linear.rs$Rising_Star)
midpoints <- barplot(heights, names.arg = gain.linear.rs$depth,  ylim = c(0,9), col = "blue",
                     xlab = "Percentile", ylab = "Decile lift",
                     main = "Decile-chart")
text(midpoints, heights+0.5, labels=round(heights, 1), cex = 0.8)

pred.linear.rs.round <- round(pred.linear.rs,0)

#Confusion Matrix
confusiontable.linear.rs <- table(Predicted = pred.linear.rs.round , Actual = valid.linear.rs$Rising_Star
confusiontable.linear.rs

#Accuracy
mean(pred.linear.rs.round==valid.linear.rs$Rising_Star)
```



- As seen from the above lift chart, it is evident that the model curve has comparatively more area (covers more variation) under it compared to the naïve rule represented by the straight line.



- Decile chart follows an ideal structure representing maximum variation covered in initial deciles.
- This can be considered as good model where the deciles are decreasing in order from start to end.
- Looking at the first decile, we can say that this model performs 1.4 time better than the one with Naïve rule.

Confusion Matrix

	Actual				
Predicted	1	2	3	4	5
1	820	39	0	0	0
2	1	739	0	0	0
3	0	3	726	0	0
4	0	0	7	1752	424
5	0	0	0	118	1370

Accuracy in predicting Validation data set is 90%

```
[1] 0.9013169
```

Strengths: Linear regression is straightforward to understand, explain and can be regularized to avoid over fitting. In addition, linear models can be updated easily with new data.

Weaknesses: Linear regression performs poorly when there are non-linear relationships. They are not naturally flexible enough to capture more complex patterns and adding the right interaction terms or polynomials can be tricky and time -consuming.

2.Running Knn model for the same question and comparing which model is better

KNN is used to classify or predict a new record based on similar records in the training data. It is a non-parametric method, it is data driven. There are no parameters to estimate as in linear regression. It is based on distance between records.

Rising star indicates the level of promise or promote-ability the employee has. Scale(1-5) 5 being the highest and 1 lowest

```
hr.logit <- hr.df[,5:30]

### Partitioning data
set.seed(123456789)
train.index <- sample(row.names(hr.logit), 0.6*dim(hr.logit)[1])
valid.index <- setdiff(row.names(hr.logit), train.index)
train.df <- hr.logit[train.index, ]
valid.df <- hr.logit[valid.index, ]

train.norm.df <- train.df
valid.norm.df <- valid.df
hr.norm.df <- hr.logit
### Normalize data using preProcess() from CARET
norm.values <- preProcess(train.df[, c(1,3:26)], method=c("center", "scale"))
train.norm.df[, c(1,3:26)] <- predict(norm.values, train.df[, c(1,3:26)])
valid.norm.df[, c(1,3:26)] <- predict(norm.values, valid.df[, c(1,3:26)])
### Run K-NN
nn <- knn(train = train.norm.df[, c(1,3:26)], test = valid.norm.df[, c(1,3:26)],
          cl = train.norm.df[, 2], k = 5)

### Nearest-neighbor Index
row.names(train.df)[attr(nn, "nn.index")]
```

output: character(0)

```
### Showing the accuracy by using confusion matrix
table(nn, valid.norm.df$Rising_Star)
CrossTable(x=nn,y=valid.norm.df$Rising_Star,prop.chisq=F)
```

output:

```
nn     1    2    3    4    5
  1  677  179    0    0    0
  2  125  634    7    0    0
  3    0    4  661   26    5
  4    0    0   65 1361  515
  5    0    0    2  481 1258
```

```
   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  6000


             | valid.norm.df$Rising_Star
          nn |         1 |         2 |         3 |         4 |         5 | Row Total |
-------------|-----------|-----------|-----------|-----------|-----------|-----------|
           1 |       677 |       179 |         0 |         0 |         0 |       856 |
             |     0.791 |     0.209 |     0.000 |     0.000 |     0.000 |     0.143 |
             |     0.844 |     0.219 |     0.000 |     0.000 |     0.000 |           |
             |     0.113 |     0.030 |     0.000 |     0.000 |     0.000 |           |
-------------|-----------|-----------|-----------|-----------|-----------|-----------|
           2 |       125 |       634 |         7 |         0 |         0 |       766 |
             |     0.163 |     0.828 |     0.009 |     0.000 |     0.000 |     0.128 |
             |     0.156 |     0.776 |     0.010 |     0.000 |     0.000 |           |
             |     0.021 |     0.106 |     0.001 |     0.000 |     0.000 |           |
-------------|-----------|-----------|-----------|-----------|-----------|-----------|
           3 |         0 |         4 |       661 |        26 |         5 |       696 |
             |     0.000 |     0.006 |     0.950 |     0.037 |     0.007 |     0.116 |
             |     0.000 |     0.005 |     0.899 |     0.014 |     0.003 |           |
             |     0.000 |     0.001 |     0.110 |     0.004 |     0.001 |           |
-------------|-----------|-----------|-----------|-----------|-----------|-----------|
           4 |         0 |         0 |        65 |      1361 |       515 |      1941 |
             |     0.000 |     0.000 |     0.033 |     0.701 |     0.265 |     0.324 |
             |     0.000 |     0.000 |     0.088 |     0.729 |     0.290 |           |
             |     0.000 |     0.000 |     0.011 |     0.227 |     0.086 |           |
-------------|-----------|-----------|-----------|-----------|-----------|-----------|
           5 |         0 |         0 |         2 |       481 |      1258 |      1741 |
             |     0.000 |     0.000 |     0.001 |     0.276 |     0.723 |     0.290 |
             |     0.000 |     0.000 |     0.003 |     0.257 |     0.708 |           |
             |     0.000 |     0.000 |     0.000 |     0.080 |     0.210 |           |
-------------|-----------|-----------|-----------|-----------|-----------|-----------|
Column Total |       802 |       817 |       735 |      1868 |      1778 |      6000 |
             |     0.134 |     0.136 |     0.122 |     0.311 |     0.296 |           |
-------------|-----------|-----------|-----------|-----------|-----------|-----------|
```

Interpretation

accuracy = 0.78

Using the 5 nearest records to predict the employee's rising star value, the accuracy is 0.78. The prediction behaves better when employee has a low rising star score especially when their rising star is 3. It's easy to find that 2 is more likely to be confused with 1 and 4 with 5. On the ground of that we could divide the employees into 3 parts which represent high, medium and low rising star or expectation in other words.

Strengths : Easy to use and understand, robust to noisy training data. Effective if the training data is large

Weakness : Determining K is the most crucial task despite its popularity. Computation cost is quite high because we nee d to compute distance of each query instance to all training samples. Some indexing (e.g. K-D tree) may reduce this com putational cost

C) How much time will the employee spend in company?

Running Linear Regression

We run the linear regression algorithm on non-categorical variables keeping “time_spend_company” as the dependent variable. The model is trained on test data that comprises 60% of the total data and validated on the rest.

```
#Linear Regression for time spend in company
set.seed(123)
#Partitioning data into training (60%) and validation(40%) for linear regression
train.lm.ts.index <- createDataPartition(hrform.df$time_spend_company , p= 0.6, list = FALSE)
train.linear.ts <-hrform.df[train.lm.ts.index,]
valid.linear.ts <- hrform.df[-train.lm.ts.index,]

hr_time.lm <- lm(time_spend_company ~ ., data = train.linear.ts )
summary(hr_time.lm)
```

```
Coefficients:
                         Estimate Std. Error  t value              Pr(>|t|)    
(Intercept)            25.2459961  0.1463631  172.489 < 0.0000000000000002 ***
Role                   -2.7287643  0.0142642 -191.301 < 0.0000000000000002 ***
Rising_Star            -0.0829792  0.0525187   -1.580              0.11414    
Will_Relocate           0.0073851  0.0316397    0.233              0.81545    
Critical               -0.0630477  0.0848120   -0.743              0.45727    
Trending.Perf          -0.0091227  0.0123881   -0.736              0.46150    
Talent_Level            0.0097108  0.0233011    0.417              0.67687    
EMP_Sat_OnPrem_1        0.0071092  0.0106756    0.666              0.50548    
EMP_Sat_Remote_1        0.0189551  0.0130244    1.455              0.14561    
EMP_Engagement_1       -0.0108493  0.0205069   -0.529              0.59678    
last_evaluation        -0.0045745  0.0176099   -0.260              0.79505    
number_project          0.0078684  0.0141169    0.557              0.57729    
average_montly_hours    0.0002659  0.0003491    0.762              0.44633    
left_Company            0.0105658  0.0437195    0.242              0.80904    
promotion_last_5years1 -0.1034780  0.0488336   -2.119              0.03412 *  
salary                  0.0710599  0.0257036    2.765              0.00571 ** 
Gender                  0.0072757  0.0327452    0.222              0.82417    
Emp_Work_Status2       -0.0102524  0.0149908   -0.684              0.49405    
Emp_Identity            0.0065488  0.0183455    0.357              0.72112    
Emp_Role               -0.0034782  0.0178678   -0.195              0.84566    
Emp_Position            0.0118968  0.0183459    0.648              0.51670    
Emp_Title              -0.0064747  0.0151110   -0.428              0.66832    
Emp_Satisfaction       -0.0524456  0.0569057   -0.922              0.35675    
Emp_Competitive_1       0.0104714  0.0096487    1.085              0.27783    
Emp_Collaborative_1    -0.0068588  0.0103888   -0.660              0.50914    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.499 on 8976 degrees of freedom
Multiple R-squared:  0.8463,    Adjusted R-squared:  0.8459 
F-statistic:  2059 on 24 and 8976 DF,  p-value: < 0.00000000000000022
```

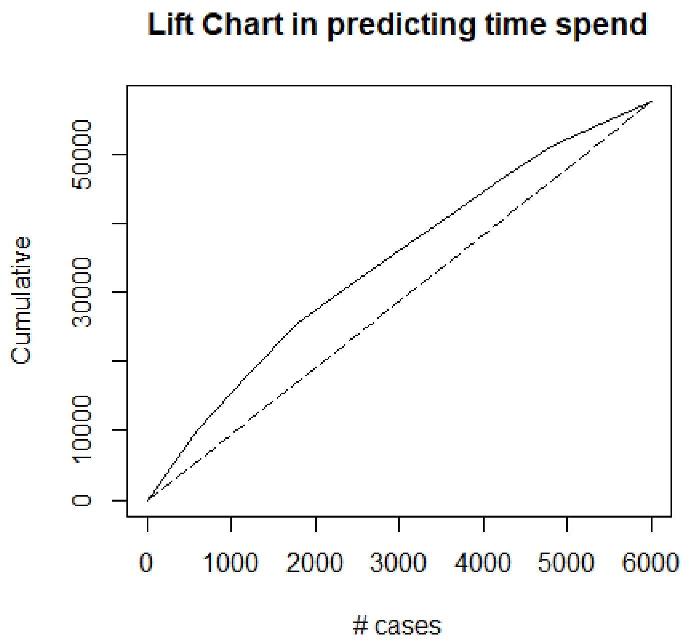
Interpretation

The significant coefficients (P Value one, two and three stars) for time_spend_company are Role, promotion_last_5years1 and salary.

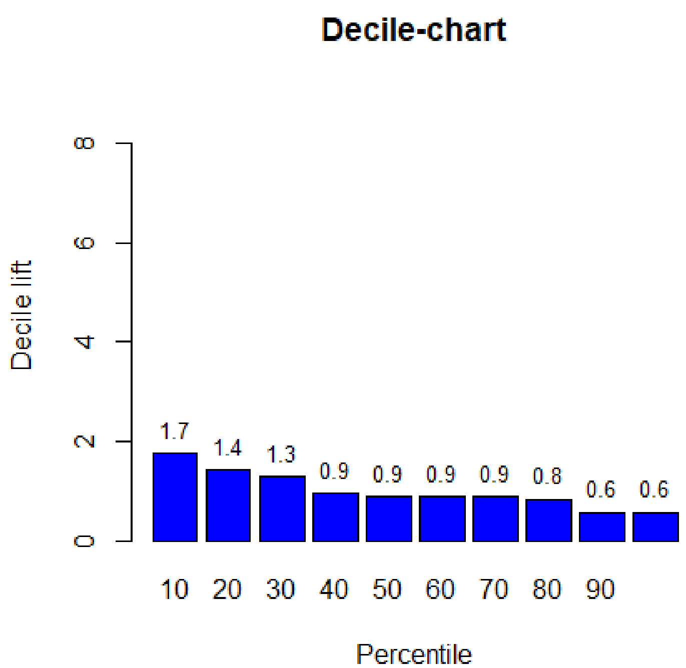
Adjusted R square value of **0.8459** can be considered as a good number exhibiting that approximately **85%** of the variation in time_spend_company variable is captured by the input variables.

Running lift and Decile chart

```
pred.linear.ts <- predict(hr_time.lm, valid.linear.ts)

#gains
gain.linear.ts <- gains(valid.linear.ts$time_spend_company , pred.linear.ts, groups = 10)
gain.linear.ts

#Lift
plot(c(0,gain.linear.ts$cume.pct.of.total*sum(pred.linear.ts))~c(0,gain.linear.ts$cume.obs),
     xlab = "# cases", ylab = "Cumulative", main = "", type = "l")
lines(c(0,sum(pred.linear.ts))~c(0, dim(valid.linear.ts)[1]), lty = 5)

#decile chart and values
heights <- gain.linear.ts$mean.resp/mean(valid.linear.ts$time_spend_company)
midpoints <- barplot(heights, names.arg = gain.linear.ts$depth,  ylim = c(0,9), col = "blue",
                     xlab = "Percentile", ylab = "Decile lift",
                     main = "Decile-chart")
text(midpoints, heights+0.5, labels=round(heights, 1), cex = 0.8)

pred.linear.ts.round <- round(pred.linear.ts,0)

#Accuracy
mean(pred.linear.ts.round==valid.linear.ts$time_spend_company)
```



As seen from the above lift chart, it is evident th As seen from the above lift chart, it is evident that the model curve has comparatively more ar ea(covers more variation) under it compared to t



Decile chart follows an ideal structure representing maximum variation covered in initial deciles.
This can be considered as good model where the deciles are decreasing in order from start to end.
Looking at the first decile, we can say that this model performs 1.7 time better than the one with Naïve rule.

D) How satisfied are the employees in company?

Running Linear Regression

We run the linear regression algorithm on non-categorical variables keeping "Emp_Satisfaction" as the dependent variable.
The model is trained on test data that comprises 60% of the total data and validated on the rest.

```
#Linear Regression for Employee Satisfaction
set.seed(123)
#Partitioning data into training (60%) and validation(40%) for linear regression
train.lm.es.index <- createDataPartition(hrform.df$EnvironmentSatisfaction , p= 0.6, list = FALSE)
train.linear.es <-hrform.df[train.lm.es.index,]
valid.linear.es <- hrform.df[-train.lm.es.index,]

hr_emp_sat.lm <- lm(EnvironmentSatisfaction ~ ., data = train.linear.es )
summary(hr_emp_sat.lm)

pred.linear.es <- predict(hr_emp_sat.lm, valid.linear.es)
```

```
Coefficients:
                         Estimate  Std. Error t value             Pr(>|t|)    
(Intercept)            0.09036542  0.05673420   1.593               0.1112    
Role                  -0.00017799  0.00597595  -0.030               0.9762    
Rising_Star            0.01147654  0.00982246   1.168               0.2427    
Will_Relocate         -0.00449842  0.00588222  -0.765               0.4444    
Critical              -0.02308707  0.01564745  -1.475               0.1401    
Trending.Perf          0.00055414  0.00230877   0.240               0.8103    
Talent_Level          -0.00920716  0.00431139  -2.136               0.0327 *  
EMP_Sat_OnPrem_1       0.00021497  0.00197892   0.109               0.9135    
EMP_Sat_Remote_1      -0.00049504  0.00243172  -0.204               0.8387    
EMP_Engagement_1      -0.00404135  0.00383356  -1.054               0.2918    
last_evaluation       -0.00344075  0.00329004  -1.046               0.2957    
number_project         0.00277301  0.00261346   1.061               0.2887    
average_montly_hours  -0.00004298  0.00006463  -0.665               0.5060    
time_spend_company    -0.00084766  0.00196004  -0.432               0.6654    
left_Company           0.00513397  0.00805445   0.637               0.5239    
promotion_last_5years1 0.14826061  0.00890727  16.645 <0.0000000000000002 ***
salary                -0.00304740  0.00479289  -0.636               0.5249    
Gender                -0.00684964  0.00610254  -1.122               0.2617    
Emp_Work_Status2       0.18940205  0.00192096  98.598 <0.0000000000000002 ***
Emp_Identity           0.20309943  0.00264245  76.860 <0.0000000000000002 ***
Emp_Role               0.20132969  0.00253920  79.289 <0.0000000000000002 ***
Emp_Position           0.19970034  0.00268879  74.271 <0.0000000000000002 ***
Emp_Title              0.18989458  0.00197312  96.241 <0.0000000000000002 ***
Emp_Competitive_1     -0.00088995  0.00180521  -0.493               0.6220    
Emp_Collaborative_1    0.00035620  0.00194406   0.183               0.8546    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2786 on 8975 degrees of freedom
Multiple R-squared:  0.9882,	Adjusted R-squared:  0.9882 
F-statistic: 3.131e+04 on 24 and 8975 DF,  p-value: < 0.00000000000000022
```

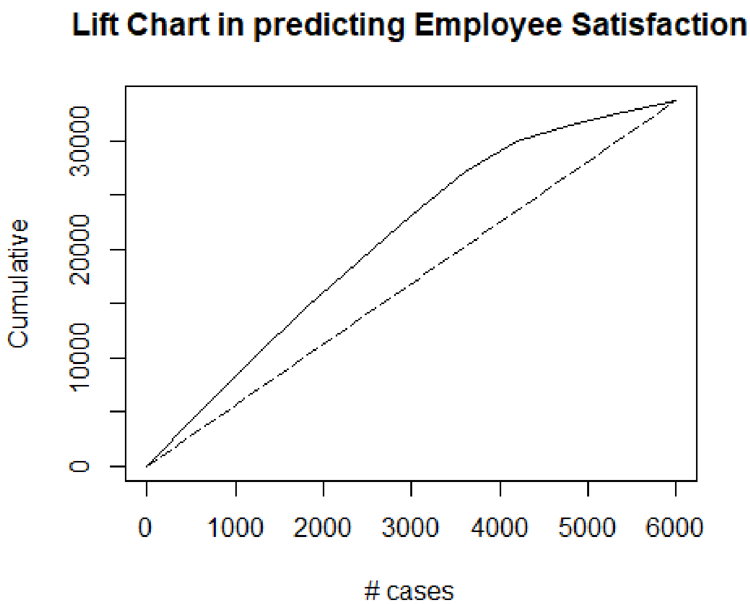
Interpretation:

The significant coefficients (P Value one, two and three stars) for Emp_Satisfaction are Talent_Level, promotion_last_5years1, Emp_work_Status2, Emp_Identity, Emp_Role, Emp_Position and Emp_Title.

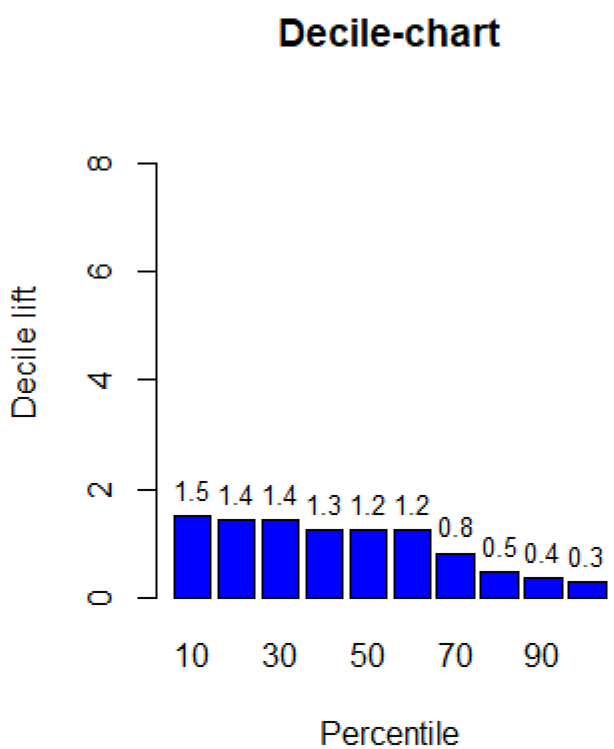
Adjusted R square value of **0.9882** can be considered as an excellent number exhibiting that approximately **99%** of the variation in Emp_Satisfaction variable is captured by the input variables.

Running Life and Decile chart

```
#gains
gain.linear.es <- gains(valid.linear.es$EnvironmentSatisfaction , pred.linear.es, groups = 10)
gain.linear.es

#Lift
plot(c(0,gain.linear.es$cume.pct.of.total*sum(pred.linear.es))~c(0,gain.linear.es$cume.obs),
     xlab = "# cases", ylab = "Cumulative", main = "", type = "l")
lines(c(0,sum(pred.linear.es))~c(0, dim(valid.linear.es)[1]), lty = 5)

#decile chart and values
heights <- gain.linear.es$mean.resp/mean(valid.linear.es$EnvironmentSatisfaction)
midpoints <- barplot(heights, names.arg = gain.linear.es$depth,  ylim = c(0,9), col = "blue",
                     xlab = "Percentile", ylab = "Decile lift",
                     main = "Decile-chart")
text(midpoints, heights+0.5, labels=round(heights, 1), cex = 0.8)

pred.linear.es.round <- round(pred.linear.es,0)

#Accuracy
mean(pred.linear.es.round==valid.linear.es$EnvironmentSatisfaction)
```



- As seen from the above lift chart, it is evident that the model curve has comparatively more area (covers more variation) under it compared to the naïve rule represented by the straight line.



- Decile chart follows an ideal structure representing maximum variation covered in initial deciles.
- This can be considered as good model where the deciles are decreasing in order from start to end.
- Looking at the first decile, we can say that this model performs 1.5 time better than the one with Naïve rule.

Accuracy in predicting the Employee satisfaction in Validation data set is 99%

```
mean(hrform.df$EnvironmentSatisfaction)
```

```
[1] 0.9943324
```

E) Running K-mean Clustering to find out which set of employees are more likely to exit

```r
###Some variables are factors so we need to transfer them into numeric
for (i in 1:26) {
  hr.logit[,i] <- as.numeric(hr.logit[,i])
}
###Normalize the data
hr.logit.norm <- sapply(hr.logit, scale)
###Function to calculate the AIC
kmeansAIC = function(km){

  m = ncol(km$centers)
  n = length(km$cluster)
  k = nrow(km$centers)
  D = km$tot.withinss
  return(D + 2*m*k)
}
###Finding the optimal k with lowest AIC
set.seed(123)
for (k in 1:30) {
  km <- kmeans(hr.logit.norm, k)
  print(k)
  print(kmeansAIC(km))
}
```

Output:

```
[1] 1
[1] 390000
[1] 2
[1] 226203.2
[1] 3
[1] 213838.9
[1] 4
[1] 198790.5
[1] 5
[1] 185537.8
[1] 6
[1] 188492.8
[1] 7
[1] 183974.1
[1] 8
[1] 169316.4
[1] 9
[1] 164959.3
[1] 10
[1] 158721.6
[1] 11
[1] 158954.4
[1] 12
[1] 153724.6
[1] 13
[1] 149703.7
[1] 14
[1] 146424.1
[1] 15
[1] 145616.1
```

```
[1] 16
[1] 144186.8
[1] 17
[1] 142037.3
[1] 18
[1] 140598.2
[1] 19
[1] 136633.8
[1] 20
[1] 135867.5
[1] 21
[1] 134374
[1] 22
[1] 135720
[1] 23
[1] 131888.1
[1] 24
[1] 132382.1
[1] 25
[1] 129481
[1] 26
[1] 129932.5
[1] 27
[1] 128610.8
[1] 28
[1] 125586.5
[1] 29
[1] 125885.8
[1] 30
[1] 125322
```

Interpretation:

AIC gets smaller when k increases and 20 is the optimal k since the AIC decreases much slower after it. That means the employees would be divided into 20 clusters.

```
###k-Means Clustering
km <- kmeans(hr.logit.norm, 20)
## Cluster size
km$size
## Cluster centroids
km$centers
```

output:

```
> km$size
 [1] 1257  435  739  566  376 1062  226 1147 1177  506 2242  399  447  591  570  426  281  734  717 1101
> ## Cluster centroids
> km$centers
```

```
   average_montly_hours time_spend_company left_Company
onmentSatisfaction
1           -0.09970569        -0.16519588  -0.64880885
       -1.39404141
2           -0.28313991         1.36179291  -0.11573547
        0.61289553
3            0.05782417        -0.44133889  -0.66434219
        0.74659200
4           -0.49985149        -0.08106666   0.67338234
       -1.32869745
5            0.30878960         1.97847905   0.20691584
       -1.20264835
6            0.02150059         1.29879711  -0.66434219
       -1.08285425
7            1.09895484        -0.40240769   1.50514801
       -0.02957039
8           -0.16663702        -0.45203728  -0.66434219
        0.80840082
9            0.06312677        -0.40468024  -0.66065572
        0.83806305
10           0.04548243        -0.27004126  -0.53142876
        1.14075057
11          -0.02197009        -0.41720015  -0.66337453
        0.84550112
12           0.66330773        -0.18277438   0.79829657
       -1.14793393
13          -0.73063299        -0.10588280   1.00524311
        0.34245562
14          -1.10568336         0.01086154   1.50514801
        0.43352897
15           0.53673528         1.31026027   1.50134189
       -1.32513589
16           1.11282969        -0.38644063   1.45931371
        0.43595798
17          -0.18490357         1.30279353  -0.06985555
       -1.38917950
18          -0.74850843        -0.10370324   1.49332518
       -1.46954630
19           1.11686707        -0.40681765   1.50514801
        0.58595138
20          -0.01779324        -0.40404068  -0.66434219
```

Strengths: K-Means is the most popular clustering algorithm because it's fast, simple, and flexible.

Weaknesses: We specify the number of clusters, which is never easy to do. In addition, if the true underlying clusters in your data are not globular, then K-Means will produce poor clusters.

The employees can be divided into 20 cluster and according to the centroids of each clusters, it seems that employees in cluster 7, 13, 14, 15, 16, 18, 19 are more likely to leave the company. By using km$cluster, we could identify every employee’s group number and predict if they are more likely to leave the company.

RESULTS

Why do good Employees Leave?

Number_of_projects, Average_monthly_hours and Salary play crucial role in determining why employees with good performance evaluation leave. These employees are highly valuable assets that should not have been lost.

Will the employee leave the company? Which employee?

We ran 2 models to conclude and predict which variables play crucial role in deciding whether the employee will or will not leave the company. Linear Diriment Analysis(LDA) model gave us a slightly better accuracy compared to Logistic Regression(GLM) model.

How dream would it be for any HR department of a company to predict which employee will leave next.

After adding Trending_Performance variable from the original dataset to the predicted output of GLM model, we got the list of the most vulnerable employees that might leave the company despite being the best in their business and performance. After grouping them per department we could email the different managers to tell them which valuable employees might leave soon and to discuss on one-to-one basis their dis-satisfaction.

What is the likelihood of the employees getting promotion?

Here we ran Linear Regression(LM) and Knn on non-categorical variables keeping “Rising_Star” as the dependent variable and came to the conclusion that if the employee is critical (“1”) the likelihood of promotion (“Rising_Star) also increases in number (1 through 5). For every one-unit change in Critical value, the independent variable is affected to change +0.239. The accuracy of LM was 90% whereas for Knn, it was 78%

How satisfied are the employees in company?

We ran the linear regression algorithm on non-categorical variables keeping “Emp_Satisfaction” as the dependent variable. The most significant parameters for Emp_Satisfaction are Talent_Level, promotion_last_5years1, Emp_work_Status2, Emp_Identity, Emp_Role, Emp_Position and Emp_Title. An adjusted R-Square value of 99% reveals that all the above parameters play the most impactful roles in deciding the level of emp_satisfaction also something in which the company could invest more in.

How much time will the employee spend in company?

We ran the linear regression algorithm on non-categorical variables keeping “time_spend_company” as the dependent variable. We concluded that the significant coefficients for time_spend_company are Role, promotion_last_5years1 and salary. Managers of different teams could possibly be alerted and advised about these factors to prevent critical and valuable employees from leaving.

Running K-mean Clustering to find out which set of employees are more likely to exit?

The employees were divided into 20 clusters(20 was the optimal K after running the model depending upon the AIC) and according to the centroids of each clusters, it seems that employees in cluster 7, 13, 14, 15, 16, 18, 19 are more likely to leave the company. By using km$cluster, we could identify every employee’s group number and predict if they are more likely to leave the company. Such potential groups could be evaluated again by higher management to ascertain the possibility of their exit

REFERENCES & CITATIONS

Book

Garrett Grolemund and Hadley Wickham. Jan 2017. *R for Data Science: Visualize, Model, Transform, Tidy and Import* O'Reilly

Website

Elite Data Science. 2016-2018. https://elitedatascience.com/machine-learning-algorithms

The Classroom. 2018. Disadvantages of Logistic Regression. https://www.theclassroom.com/disadvantages-logistic-regression-8574447.html

Ragul Ram J. 2017. HR Analytics- exploration and modelling with R. https://www.kaggle.com/ragulram/hr-analytics-exploration-and-modelling-with-r

Melanie Frazier. 2012. R color Cheatsheet. https://www.nceas.ucsb.edu/~frazier/RSpatialGuides/colorPaletteCheatsheet.pdf

Mikmart. 2018. Allow flipping panel.grid theme elements with coord_flip. https://github.com/tidyverse/ggplot2/issues/2908