Title: HR – Employee Attrition Solution
Group 117

First name	Last Name	Email	Monday or Tuesday class
Pragya	Shukla	Pshukla5@hawk.iit.edu	Monday

Table of Contents

1. INTRODUCTION

What is Attrition :

Attrition in human resources refers to the gradual loss of employees over time. HR professionals often assume a leadership role in designing company compensation programs, work culture and motivation systems that help the organization retain top employees. Attrition can be a manifest itself when employees voluntarily leave their jobs. This can happen for a variety of reasons: employees may move or retire, take another job, be ill-suited to the position they were hired to fill, or want employment that offers a more equitable work-life balance. Others may experience a lack of the freedom or autonomy they require to perform at expected levels. Human resources professionals inadvertently encourage attrition when they condone or ignore maltreatment of employees by management.
A major problem in high employee attrition is its cost to an organization. Job postings, hiring processes, paperwork and new hire training are some of the common expenses of losing employees and replacing them. Additionally, regular employee turnover prohibits your organization from increasing its collective knowledge base and experience over time. This is especially concerning if your business is customer facing, as customers often prefer to interact with familiar people. Errors and issues are more likely if you constantly have new workers.

> Employees who are engaged are more likely to stay with their organization, reducing overall turnover and the costs associated with it. They feel a stronger bond to their organization's mission and purpose, making them more effective brand ambassadors. They build stronger relationships with customers, helping their company increase sales and profitability.

How to tackle it :

By the help of analysis, we can built a model which can predict the probability of an employee to quit the job in near future. Also we can identify whether few attributes could be a main reason for an employee to leave the current job. This could help the HR and management to take preventive steps in orders to retain their best employees and losing them to their competitors.

2. DATA SETS

Target department – H/R

This dataset is created by IBM for analysis purpose and it contains attributes that may be considered effecting employee attrition rate.

It is far less expensive to 'keep' good employees once you have them, then the cost of attracting and training new ones.

HR truly needs to start thinking outside of its traditional thinking and methodologies to powerfully address the HR challenges and issues in the future.

Data set –
www.kaggle.com/HRAnalyticRhttps://www.kaggle.com/HRAnalyticR

Size : 49,654
Attributes : 18

More Information on attributes

Employee Personal Information

Birthdate_key – {Date}

Age – {Numerical date}

City_Name – {categorical data}

Gender_shot – {categorical data (M, F)}

Gender_Full – {categorical data (male, female)}

Company Related Information

Eployee_ID – {numerical}

Originhireddate_key – {Date}

Length of service – {numerical}

Job-Title – {categorical}

Termreason_desc – {categorical}

Status – {categorical (yes, no)}

Recorddate_key – {Date}

Terminationdate_key – {Date}

Department_name – {numerical}

Store_name – {numerical}

Status_year – {Date}

Business_unit – {categorical (headoffice,store)}

Termtype_desc – {categorical}

3. Problem to be Solved

What proportion of staff is leaving?

Which area is it occurring the most?

Does age , Gender or length of service affect it?

Can HR department of the company predict the attrition rate?

If yes which model would be best for this purpose?

- We will try to find the attributes which affect the attrition rate most
- Classify age attribute into group to identify which group shows the most attrition.
- Design a predictive model to predict employee attrition with higher accuracy and reliability
- Perform Hypothesis Testing to predict –
  - ❑ If male and Female employees have same attrition rate.
  - ❑ If employees of any length of service show same attrition rate.

4. Data Processing

Summary of Data

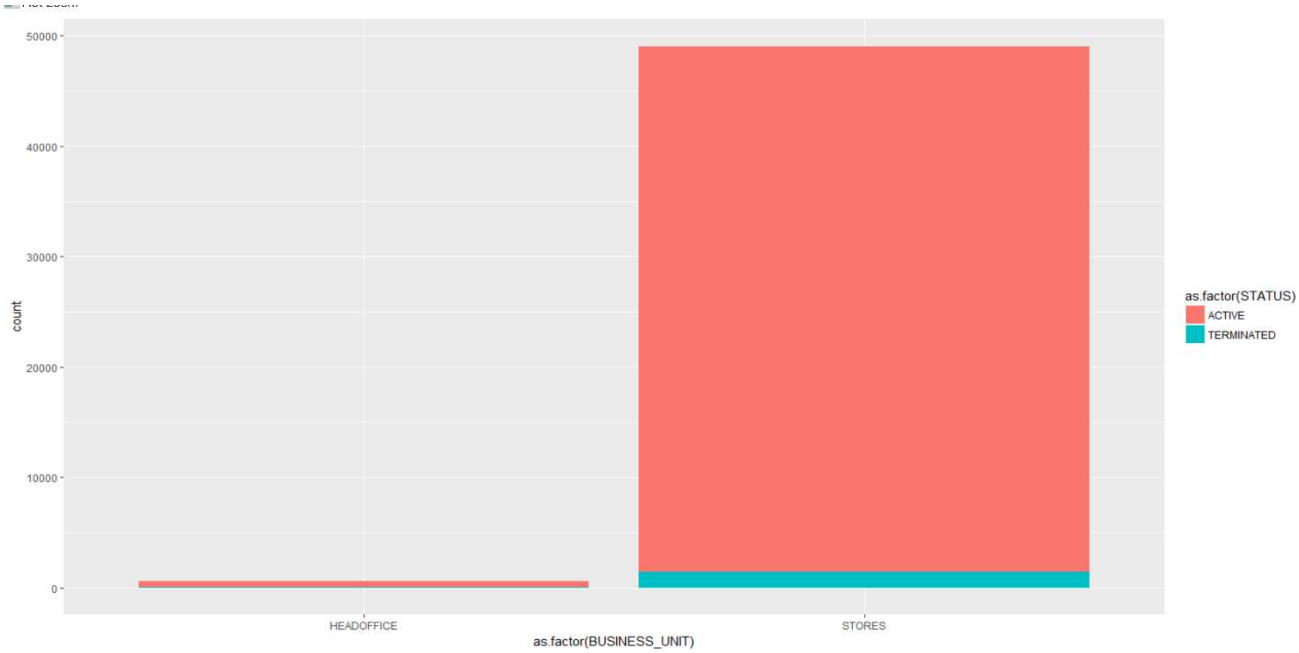
```
Console C:/Pragya/study/Data Analytics/project/R Churn Example/
> summary(EmployeeAttrition)
   EmployeeID          recorddate_key     birthdate_key      orighiredate_key
 Min.   :1318   12/31/2013 0:00: 5215   3/23/1973:   40   10/16/2005:   50
 1st Qu.:3360   12/31/2012 0:00: 5101   4/27/1956:   40   12/4/2004 :   50
 Median :5031   12/31/2011 0:00: 4972   8/4/1954 :   40   2/22/1995 :   50
 Mean   :4859   12/31/2014 0:00: 4962   1/12/1977:   30   2/26/2006 :   50
 3rd Qu.:6335   12/31/2010 0:00: 4840   1/19/1957:   30   8/9/1992  :   50
 Max.   :8336   12/31/2015 0:00: 4799   1/24/1957:   30   9/25/2006 :   50
                (Other)         :19764   (Other)  :49443   (Other)   :49353
 terminationdate_key      age          length_of_service          city_name
 1/1/1900   :42450   Min.   :19.00   Min.   : 0.00     Vancouver       :11211
 12/30/2014: 1079    1st Qu.:31.00   1st Qu.: 5.00     Victoria        : 4885
 12/30/2015:  674    Median :42.00   Median :10.00     Nanaimo         : 3876
 12/30/2010:   25    Mean   :42.08   Mean   :10.43     New Westminster: 3211
 11/11/2012:   21    3rd Qu.:53.00   3rd Qu.:15.00     Kelowna         : 2513
 2/13/2015 :   20    Max.   :65.00   Max.   :26.00     Burnaby         : 2067
 (Other)   : 5384                                     (Other)         :21890
         department_name           job_title       store_name   gender_short gender_full
 Meats            :10269   Meat Cutter  :9984   Min.   : 1.0   F:25898      Female:25898
 Dairy            : 8599   Dairy Person :8590   1st Qu.:16.0   M:23755      Male  :23755
 Produce          : 8515   Produce Clerk:8237   Median :28.0
 Bakery           : 8381   Baker        :8096   Mean   :27.3
 Customer Service: 7122    Cashier      :6816   3rd Qu.:42.0
 Processed Foods : 5911    Shelf Stocker:5622   Max.   :46.0
 (Other)          :  856   (Other)      :2308
       termreason_desc          termtype_desc     STATUS_YEAR          STATUS
 Layoff         : 1705   Involuntary   : 1705   Min.   :2006   ACTIVE    :48168
 Not Applicable:41853    Not Applicable:41853   1st Qu.:2008   TERMINATED: 1485
 Resignaton     : 2111   Voluntary     : 6095   Median :2011
 Retirement     : 3984                          Mean   :2011
                                                3rd Qu.:2013
                                                Max.   :2015

   BUSINESS_UNIT
 HEADOFFICE:  585
 STORES    :49068
```

Rate of Attrition over 10 years

```
> statuscount <- as.data.frame.matrix(EmployeeAttrition %>%
+                                       group_by(STATUS_YEAR) %>%
+                                       select(STATUS) %>%
+                                       table())
Adding missing grouping variables: `STATUS_YEAR`
> statuscount$Total <- statuscount$ACTIVE + statuscount$TERMINATED
> statuscount$PercentTerminated <- statuscount$TERMINATED/(statuscount$Total)*100
> statuscount
     ACTIVE TERMINATED Total PercentTerminated
2006   4445        134  4579          2.926403
2007   4521        162  4683          3.459321
2008   4603        164  4767          3.440319
2009   4710        142  4852          2.926628
2010   4840        123  4963          2.478340
2011   4972        110  5082          2.164502
2012   5101        130  5231          2.485184
2013   5215        105  5320          1.973684
2014   4962        253  5215          4.851390
2015   4799        162  4961          3.265471
> mean(statuscount$PercentTerminated)
[1] 2.997124
```

 # To determine which business Unit has more attrition rate :

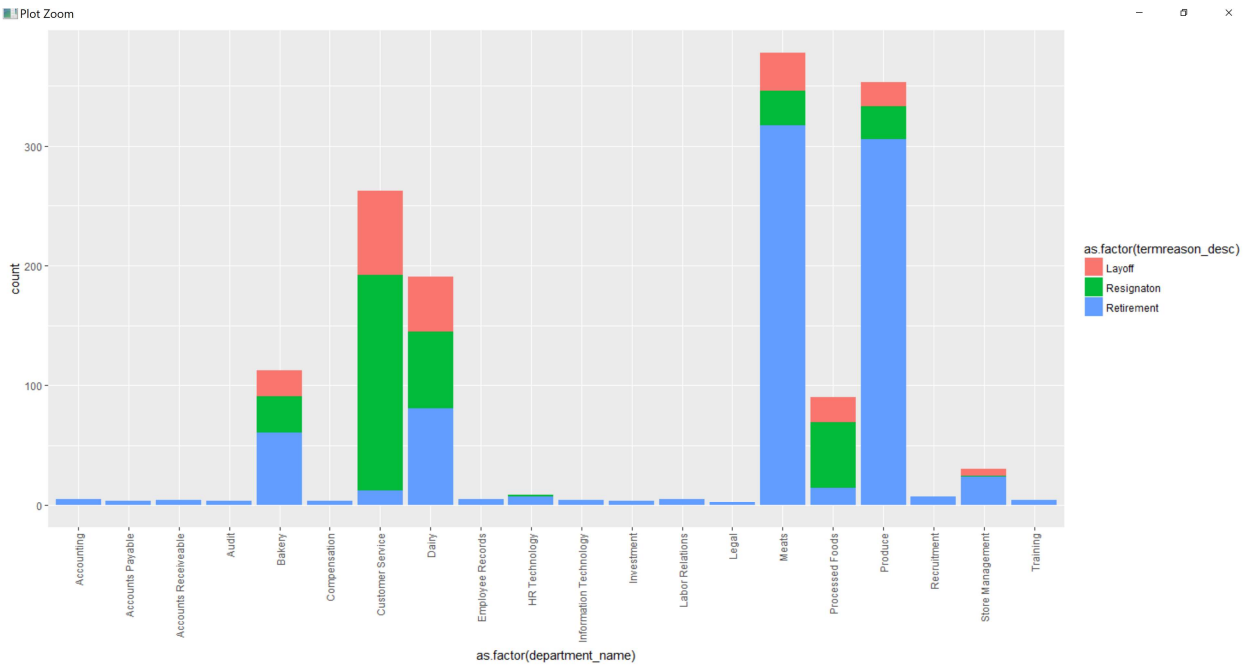
```
> ggplot() + geom_bar(aes(y = ..count..,x =as.factor(BUSINESS_UNIT),
+                         fill = as.factor(STATUS)),data=EmployeeAttrition,
+                     position = position_stack())
> 
```



Looking at the above graph we can say that store unit has more churn rate in comparison to head office

 # Attrition Rate by termination reason and department :

```
> # determinig the churn rate with reason in each department
> # filtering data on the basis of STATUS = TERMINATED
> ChurnData<- as.data.frame(EmployeeAttrition %>%
+                                 filter(STATUS=="TERMINATED"))
Warning message:
package 'bindrcpp' was built under R version 3.3.3
> ggplot() + geom_bar(aes(y = ..count..,x =as.factor(department_name),
+                         fill = as.factor(termreason_desc)),data=ChurnData,
+                     position = position_stack())+
+   theme(axis.text.x=element_text(angle=90,hjust=1,vjust=0.5))
> 
```



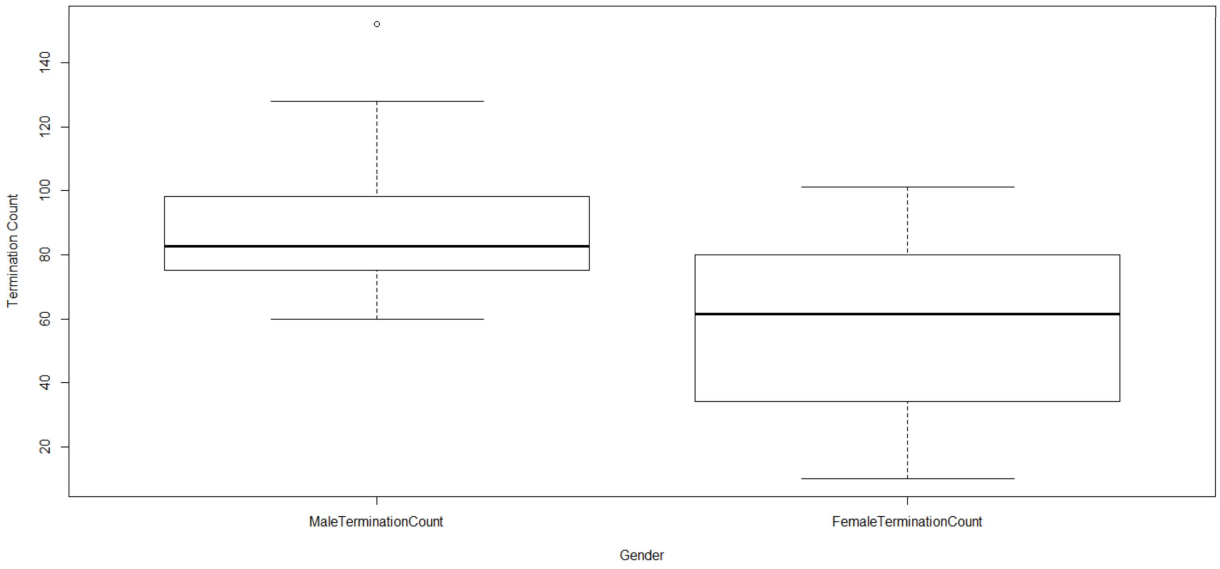
Seeing the above screen shot we can say that Customer Service area has the maximum resignation and meats and Product departments shows to have the maximum retirements.

Hypothesis Testing 1 :

Null Hypothesis : µo = male and female employees have same attrition rate

Alternate Hypothesis µ1 = male and female employees do not have same attrition rate

```
> hypotest1 <- EmployeeAttrition[,c(13,16,17)]
> hypotest1 <- as.data.frame(hypotest1 %>%
+                            filter(hypotest1$STATUS=="TERMINATED"))
> hypotest1total <- as.data.frame.matrix(hypotest1 %>%
+                                        group_by(STATUS_YEAR) %>%
+                                        select(gender_full) %>%
+                                        table())
Adding missing grouping variables: `STATUS_YEAR`
> MaleTerminationCount <- c(hypotest1total$Male)
> FemaleTerminationCount <- c(hypotest1total$Female)
> MaleTerminationCount
 [1]  74  81  80  64  59  35  32  10 101  34
> FemaleTerminationCount
 [1]  60  81  84  78  64  75  98  95 152 128
> boxplot(hypotest1total,
+         names = c("MaleTerminationCount","FemaleTerminationCount"),
+         xlab="Gender", ylab="Termination Count")
> Terminationdiff <- MaleTerminationCount - FemaleTerminationCount
```



Seeing the boxplot we can say that the average count of termination for male employees is far more than female employees. For male employees 50% of termination count is from 70 to 90 but for female employees it is much lower between 30 and 70. The variance for female employees is higher than that of the male employees. The male plot is positive skewed where as female plot is positive skewed.

Lets perform the z test to go to a definite result –

```
> z.test(Terminationdiff, NULL, alternative = "two.sided",
+        mu=0, sigma.x = sd(Terminationdiff), sigma.y = NULL, conf.level = 0.95)

        One-sample z-Test

data:  Terminationdiff
z = -2.8543, p-value = 0.004313
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -58.18999 -10.81001
sample estimates:
mean of x
    -34.5
```

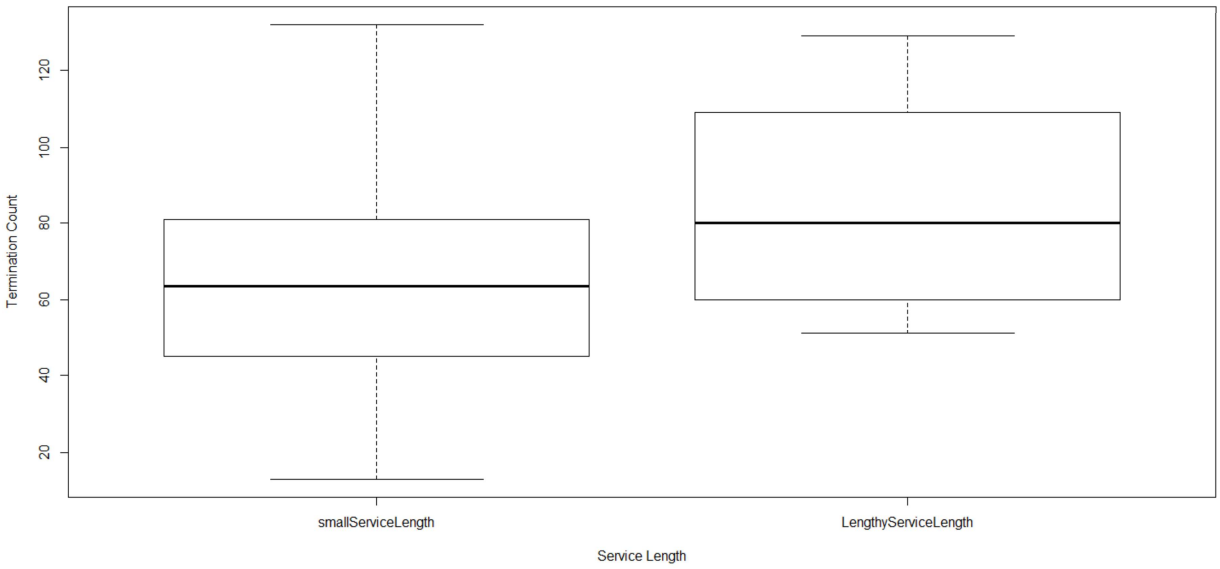
Seeing the p value above which is less than .05 we can say with 95% confidence that we can reject null hypothesis and accept alternate hypothesis that attrition rate for male and female is not same.

Hypothesis Testing 2 :

Null Hypothesis : µo = Employee with length of service less than 10 years and more than 10 years have the same attrition rate

Alternate Hypothesis µ1 = Employee with length of service less than 10 years and more than 10 years don't have the same attrition rate

```
> hypotest2 <- EmployeeAttrition[,c(7,16,17)]
> # crating different age groups
> lengthofservicegrp <- cut(hypotest2$length_of_service, breaks = c(0,11,30),
+                           labels = c("smallservicelen","lengthyservicelen")
+                   , right= FALSE)
> hypotest2 <- data.frame(hypotest2, lengthofservicegrp)
> hypotest2 <- hypotest2[1:49653, -1]
> hypotest2 <- as.data.frame(hypotest2 %>%
+                              filter(hypotest2$STATUS=="TERMINATED"))
> hypotest2total <- as.data.frame.matrix(hypotest2 %>%
+                                          group_by(STATUS_YEAR) %>%
+                                          select(lengthofservicegrp) %>%
+                                          table())
Adding missing grouping variables: `STATUS_YEAR`
> LengthyserTermination <- c(hypotest2total$lengthyservicelen)
> SmallserTermination <- c(hypotest2total$smallservicelen)
> boxplot(hypotest2total,
+         names = c("smallServiceLength","LengthyServiceLength"),
+         xlab="Service Length", ylab="Termination Count")
>
```



Seeing the boxplot we can say that the average count of termination for lengthy service length(more than 10 years) is more than that of small service length(less than 10 years). For small service length 50% termination count is between 40 to 80. And for lengthy service length it is 50 to 110. Variance for small length service is more than lengthy service length. The boxplot for small service length is normally distributed and for lengthy service length is slightly positive skewed.

Lets perform the z test to go to a definite result –

```
> Termination2diff <- SmallserTermination - LengthyserTermination
> z.test(Termination2diff, NULL, alternative = "two.sided",
+        mu=0, sigma.x = sd(Termination2diff), sigma.y = NULL, conf.level = 0.95)

        One-sample z-Test

data:  Termination2diff
z = -1.475, p-value = 0.1402
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -48.672612   6.872612
sample estimates:
mean of x
    -20.9
```

As p value is greater than .05 so with 95% confidence interval we have to accept null hypothesis that attrition rate for any length of service is same.

Anova Model :

To determine which age group has the highest attrition rate –

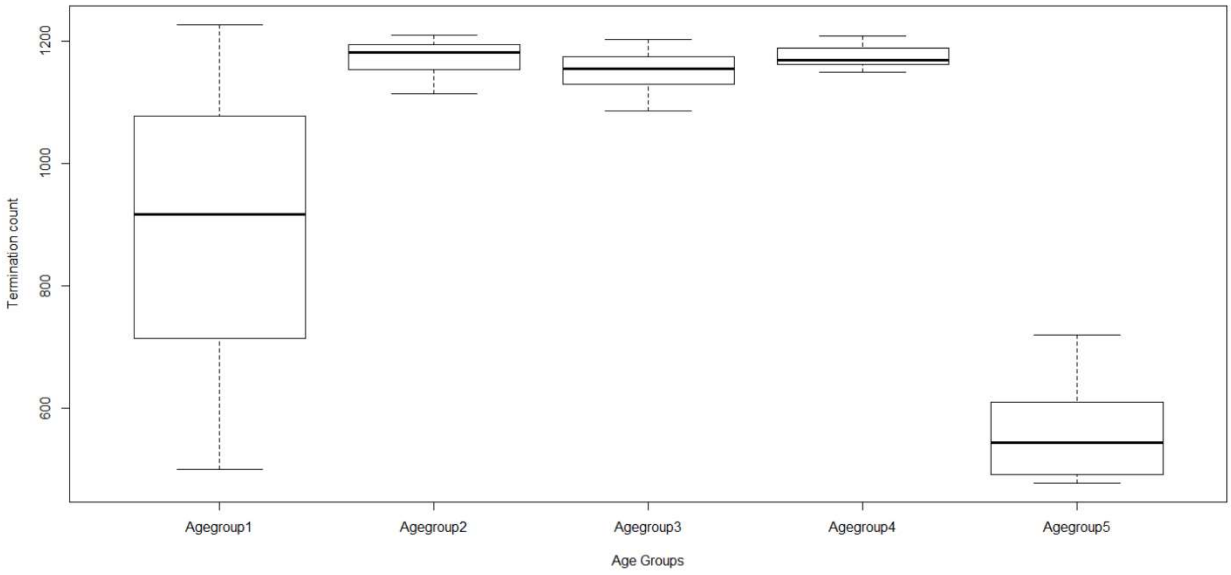
```
> # aniva testing start
> anovadata <- EmployeeAttrition[,c(6,16,17)]
> # crating different age groups
> agegroup <- cut(anovadata$age, breaks = c(18,29,39,49,59,69), labels = c("Agegroup1",
+                                                                           "Agegroup2", "Ageg
roup3",
+                                                                           "Agegroup4", "Ageg
roup5")
+                  , right= FALSE)
> anovadata <- data.frame(anovadata, agegroup)
> anovadata <- anovadata[1:49653, -1]
> View(anovadata)
> avdata <- anovadata
> avdata<- as.data.frame(avdata %>%
+                          filter(STATUS=="TERMINATED"))
> View(avdata)

> agegroupTotal1 <- as.data.frame.matrix(avdata %>%
+                                 group_by(agegroup) %>%
+                                 select(STATUS_YEAR) %>%
+                                 table())
Adding missing grouping variables: `agegroup`
> agegroupTotal1
          2006 2007 2008 2009 2010 2011 2012 2013 2014 2015
Agegroup1    3    5    8    7   11   42   39   19   91   10
Agegroup2    5    6    8    4   12   15   25   26   41   44
Agegroup3    2    8    7    5    4    3    8    4   33    8
Agegroup4    2    6    3    2    2    7    2    0   23   15
Agegroup5  122  137  138  124   94   43   56   56   65   85
```

```
> agesplit <- c("agegroup1", "agegroup2", "agegroup3", "agegroup4", "agegroup5",
+               "agegroup1", "agegroup2", "agegroup3", "agegroup4", "agegroup5",
+               "agegroup1", "agegroup2", "agegroup3", "agegroup4", "agegroup5",
+               "agegroup1", "agegroup2", "agegroup3", "agegroup4", "agegroup5",
+               "agegroup1", "agegroup2", "agegroup3", "agegroup4", "agegroup5",
+               "agegroup1", "agegroup2", "agegroup3", "agegroup4", "agegroup5",
+               "agegroup1", "agegroup2", "agegroup3", "agegroup4", "agegroup5",
+               "agegroup1", "agegroup2", "agegroup3", "agegroup4", "agegroup5",
+               "agegroup1", "agegroup2", "agegroup3", "agegroup4", "agegroup5",
+               "agegroup1", "agegroup2", "agegroup3", "agegroup4", "agegroup5"
+               )
> newanovadata <- as.data.table(agesplit)
```

```
> TotalTerminate <- c(agegroupTotal1$`2006`,agegroupTotal1$`2007`,agegroupTotal1$`2008`,
+                     agegroupTotal1$`2009`, agegroupTotal1$`2010`,agegroupTotal1$`2011`,
+                     agegroupTotal1$`2012`, agegroupTotal1$`2013`, agegroupTotal1$`2014`,
+                     agegroupTotal1$`2015`)
> newanovadata$TotalTerminate <-TotalTerminate
>
```

```
> boxplot(agegroupTotal, name = c("Agegroup1", "Agegroup2", "Agegroup3", "Agegroup4",
+                                 "Agegroup5"), xlab="Age Groups", ylab= "Termination count")
>
```

Box plot for each group –



```
> agegroupTotal <- as.data.frame.matrix(avdata %>%
+                                group_by(avdata$STATUS_YEAR) %>%
+                                  table())
Error in `[.default`(x, , i) : incorrect number of dimensions
> agegroupTotal
     Agegroup1 Agegroup2 Agegroup3 Agegroup4 Agegroup5
2006       501      1181      1203      1149       545
2007       614      1183      1192      1149       545
2008       714      1194      1175      1163       521
2009       830      1203      1154      1172       493
2010       957      1209      1158      1161       478
2011      1066      1193      1152      1189       482
2012      1160      1179      1156      1181       555
2013      1227      1153      1129      1201       610
2014      1077      1151      1116      1208       663
2015       876      1114      1086      1165       720
> boxplot(agegroupTotal, name = c("Agegroup1", "Agegroup2", "Agegroup3", "Agegroup4",
+                                 "Agegroup5"), xlab="Age Groups", ylab= "Termination count")
>
```

Seeing the above box plots we can say that for Age group 1 the termination count was normally distributed. For group2, group 3 and group 4 mean attrition count is high while mean attrition count for age group 5 is lower. Age group 1 has the largest variance.

Anova Model –

```
> anova1 = aov(TotalTerminate~agesplit, data = newanovadata)
> anova1
Call:
   aov(formula = TotalTerminate ~ agesplit, data = newanovadata)

Terms:
                agesplit Residuals
Sum of Squares   50574.4   21788.1
Deg. of Freedom        4        45

Residual standard error: 22.00409
> anova2= lm(TotalTerminate~agesplit)
> summary(anova2)

Call:
lm(formula = TotalTerminate ~ agesplit)

Residuals:
   Min     1Q Median     3Q    Max
-49.00 -12.03  -4.20   7.15  67.50

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         23.500      6.958   3.377  0.00152 **
agesplitagegroup2   -4.900      9.841  -0.498  0.62095
agesplitagegroup3  -15.300      9.841  -1.555  0.12700
agesplitagegroup4  -17.300      9.841  -1.758  0.08554 .
agesplitagegroup5   68.500      9.841   6.961 1.16e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22 on 45 degrees of freedom
Multiple R-squared:  0.6989,    Adjusted R-squared:  0.6721
F-statistic: 26.11 on 4 and 45 DF,  p-value: 3.121e-11
```

Seeing the above screen shot we can make out that all the groups don’t have same attrition rate. Keeping the Age group 1 as base we can see that age group 5 has different attrition rate than that of age group 2.

5. Methods and Process

As we know that we are trying to predict a categorical target which is whether an employee will leave the company or not. There are many model that can be used but today I will be using three models-

- Logistic Regression
- Naïve Bayes Classification
- Random Forest

Partitioning :

- Train data : Data for years 2006 – 2014
- Test data : Data for year 2015

```
> statuscount <- as.data.frame.matrix(train.data %>%
+                                       group_by(STATUS_YEAR) %>%
+                                       select(STATUS) %>%
+                                       table())
Adding missing grouping variables: `STATUS_YEAR`
> statuscount$Total <- statuscount$ACTIVE + statuscount$TERMINATED
> statuscount$PercentTerminated <- statuscount$TERMINATED/(statuscount$Total)*100
> statuscount
     ACTIVE TERMINATED Total PercentTerminated
2006   4445        134  4579          2.926403
2007   4521        162  4683          3.459321
2008   4603        164  4767          3.440319
2009   4710        142  4852          2.926628
2010   4840        123  4963          2.478340
2011   4972        110  5082          2.164502
2012   5101        130  5231          2.485184
2013   5215        105  5320          1.973684
2014   4962        253  5215          4.851390
> mean(statuscount$PercentTerminated)
[1] 2.967308
> statuscount <- as.data.frame.matrix(test.date %>%
+                                       group_by(STATUS_YEAR) %>%
+                                       select(STATUS) %>%
+                                       table())
Adding missing grouping variables: `STATUS_YEAR`
> statuscount$Total <- statuscount$ACTIVE + statuscount$TERMINATED
> statuscount$PercentTerminated <- statuscount$TERMINATED/(statuscount$Total)*100
> statuscount
     ACTIVE TERMINATED Total PercentTerminated
2015   4799        162  4961          3.265471
```

Seeing above screenshot we can see that if we split data as train data including first 9 years and test data as last 1 year data then the proportion of terminated data is almost in similar proportion.

```
> names(EmployeeAttrition)
 [1] "EmployeeID"          "recorddate_key"      "birthdate_key"
 [4] "orighiredate_key"    "terminationdate_key" "age"
 [7] "length_of_service"   "city_name"           "department_name"
[10] "job_title"           "store_name"          "gender_short"
[13] "gender_full"         "termreason_desc"     "termtype_desc"
[16] "STATUS_YEAR"         "STATUS"              "BUSINESS_UNIT"
> View(mydata)
Error in View : object 'mydata' not found
> mydata <- EmployeeAttrition[c(6,7,13,16,17,18)]
```

```
> crv$seed <- 40
> set.seed(crv$seed)
> train.data = subset(mydata,STATUS_YEAR<=2014)
> test.date = subset(mydata,STATUS_YEAR== 2015)
>
```

Logistic Regression Model :

First attempt –

```
" from the main menu.> names(EmployeeAttrition)
 [1] "EmployeeID"           "recorddate_key"       "birthdate_key"
 [4] "orighiredate_key"     "terminationdate_key" "age"
 [7] "length_of_service"    "city_name"            "department_name"
[10] "job_title"            "store_name"           "gender_short"
[13] "gender_full"          "termreason_desc"      "termtype_desc"
[16] "STATUS_YEAR"          "STATUS"               "BUSINESS_UNIT"
> mydata <- EmployeeAttrition[c(6,7,8,8,10,13,16,17,18)]
> library(rattle)
> crv$seed <- 40
> set.seed(crv$seed)
> train.data = subset(mydata,STATUS_YEAR<=2014)
> test.date = subset(mydata,STATUS_YEAR== 2015)
>
> #Logistic Regression model -
> Lmodel = glm(STATUS~., data = train.data, family = binomial)
> summary(Lmodel)

Call:
glm(formula = STATUS ~ ., family = binomial, data = train.data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1686  -0.1852  -0.1119  -0.0675   4.0705

Coefficients: (21 not defined because of singularities)
                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                       -1.030e+03  3.516e+02  -2.929 0.003400 **
age                                2.882e-01  6.765e-03  42.598  < 2e-16 ***
length_of_service                 -5.311e-01  1.432e-02 -37.094  < 2e-16 ***
city_nameAldergrove               -2.430e-01  4.456e-01  -0.545 0.585585
city_nameBella Bella               9.084e-01  5.824e-01   1.560 0.118854
city_nameBlue River                2.220e+00  1.120e+00   1.983 0.047374 *
city_nameBurnaby                  -7.115e-01  3.341e-01  -2.129 0.033213 *
city_nameChilliwack                5.156e-02  3.371e-01   0.153 0.878433
city_nameCortes Island             2.115e+00  5.515e-01   3.835 0.000126 ***
city_nameCranbrook                -8.592e-03  3.436e-01  -0.025 0.980053
city_nameDawson Creek              1.990e+00  3.859e-01   5.155 2.53e-07 ***
city_nameDease Lake                2.013e+00  8.186e-01   2.459 0.013949 *
city_nameFort Nelson               2.143e+00  3.307e-01   6.479 9.22e-11 ***
city_nameFort St John              6.478e-01  3.530e-01   1.835 0.066510 .
city_nameGrand Forks               2.268e+00  3.481e-01   6.517 7.17e-11 ***
city_nameHaney                     7.516e-01  4.022e-01   1.869 0.061684 .
city_nameKamloops                  3.612e-01  3.088e-01   1.170 0.242079
city_nameKelowna                   3.052e-01  3.092e-01   0.987 0.323690
city_nameLangley                   1.603e-01  3.474e-01   0.461 0.644592
```

department_nameInvestment	1.444e+01	3.496e+02	0.041	0.967046	
department_nameLabor Relations	1.476e+01	3.496e+02	0.042	0.966330	
department_nameLegal	1.523e+01	3.496e+02	0.044	0.965250	
department_nameMeats	1.249e+01	3.496e+02	0.036	0.971496	
department_nameProcessed Foods	1.090e+01	3.496e+02	0.031	0.975128	
department_nameProduce	1.199e+01	3.496e+02	0.034	0.972638	
department_nameRecruitment	1.480e+01	3.496e+02	0.042	0.966228	
department_nameStore Management	1.291e+01	3.496e+02	0.037	0.970538	
department_nameTraining	1.504e+01	3.496e+02	0.043	0.965685	
job_titleAccounts Payable Clerk	-1.189e+00	1.507e+00	-0.789	0.429915	
job_titleAccounts Receiveable Clerk	-6.709e-01	1.327e+00	-0.506	0.613102	
job_titleAuditor	-1.341e+00	1.407e+00	-0.953	0.340794	
job_titleBaker	-2.993e+00	3.329e-01	-8.990	< 2e-16	**
job_titleBakery Manager	NA	NA	NA	NA	
job_titleBenefits Admin	-6.752e-02	1.204e+00	-0.056	0.955296	
job_titleCashier	9.261e-01	4.590e-01	2.018	0.043641	*
job_titleCEO	-1.461e+00	1.106e+03	-0.001	0.998946	
job_titleCHief Information Officer	-2.882e-01	1.106e+03	0.000	0.999792	
job_titleCompensation Analyst	-1.386e+00	1.408e+00	-0.984	0.324874	
job_titleCorporate Lawyer	NA	NA	NA	NA	
job_titleCustomer Service Manager	NA	NA	NA	NA	
job_titleDairy Manager	1.803e+00	1.098e+00	1.642	0.100614	
job_titleDairy Person	NA	NA	NA	NA	
job_titleDirector, Accounting	1.461e+01	3.496e+02	0.042	0.966676	
job_titleDirector, Accounts Payable	NA	NA	NA	NA	
job_titleDirector, Accounts Receivable	NA	NA	NA	NA	
job_titleDirector, Audit	NA	NA	NA	NA	
job_titleDirector, Compensation	NA	NA	NA	NA	
job_titleDirector, Employee Records	NA	NA	NA	NA	
job_titleDirector, HR Technology	4.081e-02	1.166e+00	0.035	0.972084	
job_titleDirector, Investments	1.294e+00	1.406e+00	0.920	0.357551	
job_titleDirector, Labor Relations	9.794e-01	1.286e+00	0.761	0.446471	
job_titleDirector, Recruitment	-1.081e+01	7.818e+02	-0.014	0.988964	
job_titleDirector, Training	6.968e-01	1.313e+00	0.531	0.595582	
job_titleExec Assistant, Finance	5.564e-01	1.106e+03	0.001	0.999598	
job_titleExec Assistant, Human Resources	-5.763e-01	1.106e+03	-0.001	0.999584	
city_nameHaney	7.516e-01	4.022e-01	1.869	0.061684	.
city_nameKamloops	3.612e-01	3.088e-01	1.170	0.242079	
city_nameKelowna	3.052e-01	3.092e-01	0.987	0.323690	
city_nameLangley	1.603e-01	3.474e-01	0.461	0.644592	
city_nameNanaimo	2.073e-01	2.992e-01	0.693	0.488471	
city_nameNelson	3.064e-01	4.876e-01	0.628	0.529695	
city_nameNew Westminister	1.812e+00	3.447e-01	5.256	1.47e-07	***
city_nameNew Westminster	5.972e-02	3.077e-01	0.194	0.846093	
city_nameNorth Vancouver	-2.881e-01	4.210e-01	-0.684	0.493829	
city_nameOcean Falls	-4.041e-01	1.059e+00	-0.382	0.702747	
city_namePitt Meadows	-2.322e-01	7.192e-01	-0.323	0.746791	
city_namePort Coquitlam	1.194e-01	3.919e-01	0.305	0.760602	
city_namePrince George	6.459e-02	3.147e-01	0.205	0.837389	
city_namePrinceton	-5.273e-01	7.794e-01	-0.676	0.498724	
city_nameQuesnel	1.343e-01	3.507e-01	0.383	0.701809	
city_nameRichmond	1.832e-02	3.450e-01	0.053	0.957649	
city_nameSquamish	2.294e-01	3.652e-01	0.628	0.529877	
city_nameSurrey	3.324e-01	3.246e-01	1.024	0.305885	
city_nameTerrace	1.800e-02	3.299e-01	0.055	0.956473	
city_nameTrail	3.033e-01	3.490e-01	0.869	0.384763	
city_nameValemount	-1.269e+01	4.139e+02	-0.031	0.975551	
city_nameVancouver	-8.569e-01	2.869e-01	-2.986	0.002825	**
city_nameVernon	3.470e-01	3.634e-01	0.955	0.339622	
city_nameVictoria	2.746e-01	2.905e-01	0.945	0.344436	
city_nameWest Vancouver	-2.531e-01	3.757e-01	-0.674	0.500549	
city_nameWhite Rock	-1.287e+00	6.051e-01	-2.127	0.033408	*
city_nameWilliams Lake	1.487e-02	4.328e-01	0.034	0.972592	
department_nameAccounts Payable	1.461e+01	3.496e+02	0.042	0.966676	
department_nameAccounts Receiveable	1.461e+01	3.496e+02	0.042	0.966676	
department_nameAudit	1.574e+01	3.496e+02	0.045	0.964097	
department_nameBakery	1.229e+01	3.496e+02	0.035	0.971955	
department_nameCompensation	1.574e+01	3.496e+02	0.045	0.964097	
department_nameCustomer Service	1.113e+01	3.496e+02	0.032	0.974598	
department_nameDairy	1.133e+01	3.496e+02	0.032	0.974144	
department_nameEmployee Records	1.461e+01	3.496e+02	0.042	0.966676	

```
job_titleExec Assistant, Human Resources -5.763e-01  1.106e+03  -0.001 0.999584
job_titleExec Assistant, Legal Counsel    2.017e+00  1.106e+03   0.002 0.998544
job_titleExec Assistant, VP Stores        8.445e-01  1.106e+03   0.001 0.999391
job_titleHRIS Analyst                            NA         NA      NA       NA
job_titleInvestment Analyst                      NA         NA      NA       NA
job_titleLabor Relations Analyst                 NA         NA      NA       NA
job_titleLegal Counsel                   -5.763e-01  1.106e+03  -0.001 0.999584
job_titleMeat Cutter                     -1.785e+00  2.852e-01  -6.259 3.87e-10 ***
job_titleMeats Manager                           NA         NA      NA       NA
job_titleProcessed Foods Manager          8.952e-01  3.694e-01   2.424 0.015365 *
job_titleProduce Clerk                   -2.346e+00  3.420e-01  -6.861 6.86e-12 ***
job_titleProduce Manager                         NA         NA      NA       NA
job_titleRecruiter                               NA         NA      NA       NA
job_titleShelf Stocker                           NA         NA      NA       NA
job_titleStore Manager                           NA         NA      NA       NA
job_titleSystems Analyst                         NA         NA      NA       NA
job_titleTrainer                                 NA         NA      NA       NA
job_titleVP Finance                      -3.082e-01  1.106e+03   0.000 0.999778
job_titleVP Human Resources              -1.998e-02  1.106e+03   0.000 0.999986
job_titleVP Stores                               NA         NA      NA       NA
gender_fullMale                           5.963e-01  7.146e-02   8.345  < 2e-16 ***
STATUS_YEAR                               5.012e-01  1.876e-02  26.720  < 2e-16 ***
BUSINESS_UNITSTORES                              NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 11920.1  on 44691  degrees of freedom
Residual deviance:  8135.6  on 44602  degrees of freedom
AIC: 8315.6

Number of Fisher Scoring iterations: 15
```

As we can see that p value for department name , city name and job title is high (>.05) so we will try to remove them one by one till we get a stable model. We firstly removed Job Title then city name and at last department name. After removing these variables we got our final module as below-

```
> #Logistic Regression model -
> Lmodel = glm(STATUS~., data = train.data, family = binomial)
> summary(Lmodel)

Call:
glm(formula = STATUS ~ ., family = binomial, data = train.data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3245  -0.2076  -0.1564  -0.1184   3.4080

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         -893.51883   33.96609 -26.306  < 2e-16 ***
age                    0.21944    0.00438  50.095  < 2e-16 ***
length_of_service     -0.43146    0.01086 -39.738  < 2e-16 ***
gender_fullMale        0.51900    0.06766   7.671  1.7e-14 ***
STATUS_YEAR            0.44122    0.01687  26.148  < 2e-16 ***
BUSINESS_UNITSTORES   -2.73943    0.16616 -16.486  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 11920.1  on 44691  degrees of freedom
Residual deviance:  9053.3  on 44686  degrees of freedom
AIC: 9065.3

Number of Fisher Scoring iterations: 7
```

The AIC of the model is 9065.3. As p value of all the attributes is less than .05 so we can say they all have significant effect on status.

Naïve Bayes Model :

```
> set.seed(crv$seed)
> NBmodel <- naiveBayes(train.data$STATUS~., data= train.data)
> NBmodel

Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
    ACTIVE TERMINATED
0.97039739 0.02960261

Conditional probabilities:
            age
Y                [,1]     [,2]
  ACTIVE     41.69075 12.11163
  TERMINATED 51.54573 16.56048

            length_of_service
Y                [,1]     [,2]
  ACTIVE     10.08658 6.184904
  TERMINATED 10.63719 6.090583

            gender_full
Y               Female      Male
  ACTIVE     0.5215707 0.4784293
  TERMINATED 0.5948602 0.4051398

            STATUS_YEAR
Y                [,1]     [,2]
  ACTIVE     2010.125 2.567838
  TERMINATED 2010.155 2.749121


           BUSINESS_UNIT
Y           HEADOFFICE     STORES
  ACTIVE     0.01164426 0.98835574
  TERMINATED 0.04308390 0.95691610
```

In the above screenshot we can see that we can see the how each attribute affects the probability of occurrence of an employee getting terminated.

Random Forest :

```
> RFmodel = randomForest(train.data$STATUS~. , data= train.data,
+                        ntree= 500, mtry=2)
> RFmodel

Call:
 randomForest(formula = train.data$STATUS ~ ., data = train.data,      ntree = 500, mtry = 2)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 1.13%
Confusion matrix:
           ACTIVE TERMINATED  class.error
ACTIVE      43366          3 6.917383e-05
TERMINATED    501        822 3.786848e-01
> importance(RFmodel)
                  MeanDecreaseGini
age                     1169.07744
length_of_service        155.46570
gender_full              115.07696
STATUS_YEAR              100.46328
BUSINESS_UNIT              5.82217
>
```

Seeing the above screen shot we can see that our model is able to predict 822 terminated employees correctly.

Also age is the attribute which influence the Status maximum.

6. Evaluation and Result

6.1. Evaluation

I am using confusion matrix for evaluation :

Logistic Regression :

```
> confusionMatrix(table(MYpr, test.date$STATUS))
Confusion Matrix and Statistics


MYpr          ACTIVE TERMINATED
  ACTIVE        4799        156
  TERMINATED       0          6

               Accuracy : 0.9686
                 95% CI : (0.9633, 0.9732)
    No Information Rate : 0.9673
    P-Value [Acc > NIR] : 0.3339

                  Kappa : 0.0693
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 1.00000
            Specificity : 0.03704
         Pos Pred Value : 0.96852
         Neg Pred Value : 1.00000
             Prevalence : 0.96735
         Detection Rate : 0.96735
   Detection Prevalence : 0.99879
      Balanced Accuracy : 0.51852

       'Positive' Class : ACTIVE
```

Seeing the above screen shot we can see that accuracy is high but there are only 6 terminated employees were predicted correctly which isn't a good for our objective.

Naïve Bayes model :

```
> NBpred <- predict(NBmodel, test.date)
> confusionMatrix(table(NBpred, test.date$STATUS))
Confusion Matrix and Statistics


NBpred       ACTIVE TERMINATED
  ACTIVE       4799        155
  TERMINATED      0          7

               Accuracy : 0.9688
                 95% CI : (0.9635, 0.9734)
    No Information Rate : 0.9673
    P-Value [Acc > NIR] : 0.305

                  Kappa : 0.0804
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 1.00000
            Specificity : 0.04321
         Pos Pred Value : 0.96871
         Neg Pred Value : 1.00000
             Prevalence : 0.96735
         Detection Rate : 0.96735
   Detection Prevalence : 0.99859
      Balanced Accuracy : 0.52160

       'Positive' Class : ACTIVE
```

Seeing the above screen shot we can see that accuracy is high but there are only 7 terminated employees were predicted correctly which isn't a good for our objective.

Random forest :

```
> RFpred <- predict(RFmodel, newdata = test.date, type = "class")
```

```
> confusionMatrix(table(RFpred, test.date$STATUS))
Confusion Matrix and Statistics

RFpred         ACTIVE TERMINATED
  ACTIVE         4747         92
  TERMINATED       52         70

               Accuracy : 0.971
                 95% CI : (0.9659, 0.9755)
    No Information Rate : 0.9673
    P-Value [Acc > NIR] : 0.079162

                  Kappa : 0.4783
 Mcnemar's Test P-Value : 0.001154

            Sensitivity : 0.9892
            Specificity : 0.4321
         Pos Pred Value : 0.9810
         Neg Pred Value : 0.5738
             Prevalence : 0.9673
         Detection Rate : 0.9569
   Detection Prevalence : 0.9754
      Balanced Accuracy : 0.7106

       'Positive' Class : ACTIVE
```

Seeing the above screen shot we can say that we are able to predict 70 terminated correctly which is better than other two models.

6.2. Results and Finding

Logistic Regression Model

```
> confusionMatrix(table(MYpr, test.date$STATUS))
Confusion Matrix and Statistics

MYpr         ACTIVE TERMINATED
  ACTIVE       4799        156
  TERMINATED      0          6

               Accuracy : 0.9686
                 95% CI : (0.9633, 0.9732)
```

Naïve Bayes Mode

```
> confusionMatrix(table(NBpred, test.date$STATUS))
Confusion Matrix and Statistics

NBpred       ACTIVE TERMINATED
  ACTIVE       4799        155
  TERMINATED      0          7

               Accuracy : 0.9688
                 95% CI : (0.9635, 0.9734)
```

Random Forest Model

```
> confusionMatrix(table(RFpred, test.date$STATUS))
Confusion Matrix and Statistics


RFpred        ACTIVE TERMINATED
  ACTIVE        4747         92
  TERMINATED      52         70

               Accuracy : 0.971
                 95% CI : (0.9659, 0.9755)
```

Comparing the three models we can say that Random Forest has the maximum accuracy and also its prediction of terminated employees is better.

7. Conclusion

7.1. Conclusion

Age 5 showed different attrition rate in comparison to age group 1 where as age group 2 ,3 and 4 were similar

Comparing the confusion matrix we can come to a conclusion that Random forest is the model as we can predict the terminated employees more correctly than logistic regression and naïve Bayes model

We found that approx. 2.99% of employees have left the job over the year

Store had maximum termination and customer relationship department shows the maximum attrition

Male and female employees show different attrition which was not seen in case of length of service

7.2. Limitation

- Data set had lesser influencing attributes which limited the scope of analysis

- Lesser knowledge about various classification model limit our accuracy

7.3. Future Work

- I will try to implement more model with better accuracy and make the predictions more reliable
- I try to implement my final predicting model so that it be used in a generalized way