```
#loading the dataset
data<-read.csv('dataset.csv')
summary(data)
colnames(data)

#making the machine realize to take categorical as factor instead of numeric
data$Attrition <- factor(data$Attrition)
data$BusinessTravel <- factor(data$BusinessTravel)
data$Department <- factor(data$Department)
data$Education <- factor(data$Education)
data$EducationField <- factor(data$EducationField)
data$EnvironmentSatisfaction <- factor(data$EnvironmentSatisfaction)
data$Gender <- factor(data$Gender)
data$JobInvolvement <- factor(data$JobInvolvement)
data$JobLevel <- factor(data$JobLevel)
data$JobRole <- factor(data$JobRole)
data$JobSatisfaction <- factor(data$JobSatisfaction)
data$MaritalStatus <- factor(data$MaritalStatus)
data$Over18 <- factor(data$Over18)
data$OverTime <- factor(data$OverTime)
data$PerformanceRating <- factor(data$PerformanceRating)
data$RelationshipSatisfaction <- factor(data$RelationshipSatisfaction)
data$StockOptionLevel <- factor(data$StockOptionLevel)
data$WorkLifeBalance <- factor(data$WorkLifeBalance)

#removing the least important columns which do not have significance in dataset.
data <- subset( data, select = -c(EmployeeCount, EmployeeNumber, StandardHours) )
View(data)



# Statistical Analysis
# t-test is used to analyse what significance the continuous value has on the target variable.

t.test(Age~Attrition,data=data)
# data:  Age by Attrition
# t = 5.828, df = 316.93, p-value = 1.38e-08
# alternative hypothesis: true difference in means is not equal to 0
# 95 percent confidence interval:
#   2.618930 5.288346
# sample estimates:
#   mean in group No mean in group Yes
# 37.56123          33.60759
#My interpretation: age has significant impact on attrition as the p-value is less than 0.05
```

```
t.test(DailyRate~Attrition,data=data)
# data:  DailyRate by Attrition
# t = 2.1789, df = 333.76, p-value = 0.03004
# alternative hypothesis: true difference in means is not equal to 0
# 95 percent confidence interval:
#   6.040083 118.243100
# sample estimates:
#   mean in group No mean in group Yes
# 812.5045          750.3629
#My interpretation: DailyRate has impact on attrition as the p-value is less than 0.05

t.test(DistanceFromHome~Attrition,data=data)
# Welch Two Sample t-test
#
# data:  DistanceFromHome by Attrition
# t = -2.8882, df = 322.72, p-value = 0.004137
# alternative hypothesis: true difference in means is not equal to 0
# 95 percent confidence interval:
#   -2.8870025 -0.5475146
# sample estimates:
#   mean in group No mean in group Yes
# 8.915653         10.632911
# My interpretation: DistanceFromHome has impact on attrition as the p-value is less than 0.05

t.test(MonthlyIncome~Attrition,data=data)
# data:  MonthlyIncome by Attrition
# t = 7.4826, df = 412.74, p-value = 4.434e-13
# alternative hypothesis: true difference in means is not equal to 0
# 95 percent confidence interval:
#   1508.244 2583.050
# sample estimates:
#   mean in group No mean in group Yes
# 6832.740          4787.093
#My interpretation: MonthlyIncome has impact on attrition as the p-value is less than 0.05

t.test(MonthlyRate~Attrition,data=data)
# data:  MonthlyRate by Attrition
# t = -0.5755, df = 330.1, p-value = 0.5653
# alternative hypothesis: true difference in means is not equal to 0
# 95 percent confidence interval:
#   -1296.8656   709.8084
# sample estimates:
```

```
#   mean in group No mean in group Yes
# 14265.78          14559.31
# My interpretation: MonthlyRate does not has impact on attrition as the p-value is more than
0.05

t.test(NumCompaniesWorked~Attrition,data=data)
# data:  NumCompaniesWorked by Attrition
# t = -1.5747, df = 317.14, p-value = 0.1163
# alternative hypothesis: true difference in means is not equal to 0
# 95 percent confidence interval:
#   -0.66437603  0.07367926
# sample estimates:
#   mean in group No mean in group Yes
# 2.645580          2.940928
# My interpretation: NumCompaniesWorked does not has impact on attrition as the p-value is
more than 0.05

t.test(PercentSalaryHike~Attrition,data=data)
# data:  PercentSalaryHike by Attrition
# t = 0.50424, df = 326.11, p-value = 0.6144
# alternative hypothesis: true difference in means is not equal to 0
# 95 percent confidence interval:
#   -0.3890709  0.6572652
# sample estimates:
#   mean in group No mean in group Yes
# 15.23114          15.09705
#  My interpretation: PercentSalaryHike does not has impact on attrition as the p-value is more
than 0.05

t.test(TotalWorkingYears~Attrition,data=data)
# data:  TotalWorkingYears by Attrition
# t = 7.0192, df = 350.88, p-value = 1.16e-11
# alternative hypothesis: true difference in means is not equal to 0
# 95 percent confidence interval:
#   2.604401 4.632019
# sample estimates:
#   mean in group No mean in group Yes
# 11.862936          8.244726
# My interpretation: TotalWorkingYears has impact on attrition as the p-value is less than 0.05

t.test(TrainingTimesLastYear~Attrition,data=data)
# data:  TrainingTimesLastYear by Attrition
# t = 2.3305, df = 339.56, p-value = 0.02036
```

```
# alternative hypothesis: true difference in means is not equal to 0
# 95 percent confidence interval:
#   0.03251776 0.38439273
# sample estimates:
#   mean in group No mean in group Yes
# 2.832928          2.624473
# My interpretation: TrainingTimesLastYear has impact on attrition as the p-value is less than
0.05

t.test(YearsAtCompany~Attrition,data=data)
# data:  YearsAtCompany by Attrition
# t = 5.2826, df = 338.21, p-value = 2.286e-07
# alternative hypothesis: true difference in means is not equal to 0
# 95 percent confidence interval:
#   1.404805 3.071629
# sample estimates:
#   mean in group No mean in group Yes
# 7.369019          5.130802
# My interpretation: YearsAtCompany has impact on attrition as the p-value is less than 0.05

t.test(YearsInCurrentRole~Attrition,data=data)
# data:  YearsInCurrentRole by Attrition
# t = 6.8471, df = 366.57, p-value = 3.187e-11
# alternative hypothesis: true difference in means is not equal to 0
# 95 percent confidence interval:
#   1.127107 2.035355
# sample estimates:
#   mean in group No mean in group Yes
# 4.484185          2.902954
#  My interpretation: YearsInCurrentRole has impact on attrition as the p-value is less than 0.05

t.test(YearsSinceLastPromotion~Attrition,data=data)
# data:  YearsSinceLastPromotion by Attrition
# t = 1.2879, df = 338.49, p-value = 0.1987
# alternative hypothesis: true difference in means is not equal to 0
# 95 percent confidence interval:
#   -0.1525043  0.7309843
# sample estimates:
#   mean in group No mean in group Yes
# 2.234388          1.945148
#My interpretation: YearsSinceLastPromotion does not has impact on attrition as the p-value is
more than 0.05
```

```
t.test(YearsWithCurrManager~Attrition,data=data)
# data:  YearsWithCurrManager by Attrition
# t = 6.6334, df = 365.1, p-value = 1.185e-10
# alternative hypothesis: true difference in means is not equal to 0
# 95 percent confidence interval:
#   1.065929 1.964223
# sample estimates:
#   mean in group No mean in group Yes
# 4.367397          2.852321
#My interpretation: YearsWithCurrManager has impact on attrition as the p-value is less than 0.05


#Applying chi-square test on categorical variables with respect to Attrition to check whether the variables are dependent or not.
# Plotting the corrplot for the same to get which category has positive and negative impact on Attrition.

chisq.test(data$BusinessTravel, data$Attrition, correct=FALSE)
# data:  data$BusinessTravel and data$Attrition
# X-squared = 24.182, df = 2, p-value = 5.609e-06
#dependent (p-value less than 0.05)
#install.packages("corrplot")

library(corrplot)
corrplot(chisq.test(data$BusinessTravel, data$Attrition, correct=FALSE)$residuals, is.cor = FALSE)
# Non-travel has a positive association with No and negative association with Yes -> Non-travel do not go for attrition.
# Similarly Travel_Frequently has more possibility to go for attrition and Travel_Rarely has less possibility to go for attrition.



chisq.test(data$Department, data$Attrition, correct=FALSE)
# data:  data$Department and data$Attrition
# X-squared = 10.796, df = 2, p-value = 0.004526
# dependent

corrplot(chisq.test(data$Department, data$Attrition, correct=FALSE)$residuals, is.cor = FALSE)
#HR has positive association with Yes; RnD has positive association with No; Sales has positive association with Yes.
```

```
# HR and Sales people are more inclined towards attrition as compared to RnD.

chisq.test(data$Education, data$Attrition, correct=FALSE)
# data:  data$Education and data$Attrition
# X-squared = 3.074, df = 4, p-value = 0.5455
#independent

corrplot(chisq.test(data$Education, data$Attrition, correct=FALSE)$residuals, is.cor = FALSE)
#1 has positive impact on Yes; 3 has impact on Yes

chisq.test(data$EducationField, data$Attrition, correct=FALSE)
# data:  data$EducationField and data$Attrition
# X-squared = 16.025, df = 5, p-value = 0.006774
#dependent

corrplot(chisq.test(data$EducationField, data$Attrition, correct=FALSE)$residuals, is.cor =
FALSE)
# positive association to Yes: HR, Marketing and Technical Degree

chisq.test(data$EnvironmentSatisfaction, data$Attrition, correct=FALSE)
# data:  data$EnvironmentSatisfaction and data$Attrition
# X-squared = 22.504, df = 3, p-value = 5.123e-05

corrplot(chisq.test(data$EnvironmentSatisfaction, data$Attrition, correct=FALSE)$residuals,
is.cor = FALSE)
# positive association to Yes: 1

chisq.test(data$Gender, data$Attrition, correct=FALSE)
# data:  data$Gender and data$Attrition
# X-squared = 1.2752, df = 1, p-value = 0.2588

corrplot(chisq.test(data$Gender, data$Attrition, correct=FALSE)$residuals, is.cor = FALSE)
# positive association to Yes: Male

chisq.test(data$JobInvolvement, data$Attrition, correct=FALSE)
# data:  data$JobInvolvement and data$Attrition
# X-squared = 28.492, df = 3, p-value = 2.863e-06
#dependent

corrplot(chisq.test(data$JobInvolvement, data$Attrition, correct=FALSE)$residuals, is.cor =
FALSE)
# positive association with Yes: 1 and 2
```

```
chisq.test(data$JobLevel, data$Attrition, correct=FALSE)
# data:  data$JobLevel and data$Attrition
# X-squared = 72.529, df = 4, p-value = 6.635e-15
# dependent

corrplot(chisq.test(data$JobLevel, data$Attrition, correct=FALSE)$residuals, is.cor = FALSE)
# positive association with Yes: 1

chisq.test(data$JobRole, data$Attrition, correct=FALSE)
# data:  data$JobRole and data$Attrition
# X-squared = 86.19, df = 8, p-value = 2.752e-15

corrplot(chisq.test(data$JobRole, data$Attrition, correct=FALSE)$residuals, is.cor = FALSE)
# positive association with Yes: HR, Laboratory Technician, Sales, Sales Representatives


chisq.test(data$JobSatisfaction, data$Attrition, correct=FALSE)
# data:  data$JobSatisfaction and data$Attrition
# X-squared = 17.505, df = 3, p-value = 0.0005563

corrplot(chisq.test(data$JobSatisfaction, data$Attrition, correct=FALSE)$residuals, is.cor =
FALSE)
# positive association with Yes:1


chisq.test(data$MaritalStatus, data$Attrition, correct=FALSE)
# data:  data$MaritalStatus and data$Attrition
# X-squared = 46.164, df = 2, p-value = 9.456e-11

corrplot(chisq.test(data$MaritalStatus, data$Attrition, correct=FALSE)$residuals, is.cor =
FALSE)
# positive association with Yes: Single


chisq.test(data$Over18, data$Attrition, correct=FALSE)
# data:  data$JobRole and data$Attrition
# X-squared = 86.19, df = 8, p-value = 2.752e-15
#dependent

corrplot(chisq.test(data$Over18, data$Attrition, correct=FALSE)$residuals, is.cor = FALSE)
# positive association with Yes: HR, Laboratory Technician, Sales, Sales Representatives
```

```
chisq.test(data$OverTime, data$Attrition, correct=FALSE)
# data:  data$OverTime and data$Attrition
# X-squared = 89.044, df = 1, p-value < 2.2e-16
#dependent

corrplot(chisq.test(data$OverTime, data$Attrition, correct=FALSE)$residuals, is.cor = FALSE)
# positive association with Yes: Yes

chisq.test(data$PerformanceRating, data$Attrition, correct=FALSE)
# data:  data$PerformanceRating and data$Attrition
# X-squared = 0.012267, df = 1, p-value = 0.9118
#independent

corrplot(chisq.test(data$PerformanceRating, data$Attrition, correct=FALSE)$residuals, is.cor =
FALSE)
# positive association with Yes: 4


chisq.test(data$RelationshipSatisfaction, data$Attrition, correct=FALSE)
# data:  data$RelationshipSatisfaction and data$Attrition
# X-squared = 5.2411, df = 3, p-value = 0.155
#independent

corrplot(chisq.test(data$RelationshipSatisfaction, data$Attrition, correct=FALSE)$residuals,
is.cor = FALSE)
# positive association with Yes: 1


chisq.test(data$StockOptionLevel, data$Attrition, correct=FALSE)
# data:  data$StockOptionLevel and data$Attrition
# X-squared = 60.598, df = 3, p-value = 4.379e-13
#independent

corrplot(chisq.test(data$StockOptionLevel, data$Attrition, correct=FALSE)$residuals, is.cor =
FALSE)
# positive association with Yes: 0


chisq.test(data$WorkLifeBalance, data$Attrition, correct=FALSE)
# data:  data$WorkLifeBalance and data$Attrition
# X-squared = 16.325, df = 3, p-value = 0.0009726
#dependent
```

```
corrplot(chisq.test(data$WorkLifeBalance, data$Attrition, correct=FALSE)$residuals, is.cor =
FALSE)
# positive association with Yes: 1,2,4

##############################################################################
#############\
# Decision Tree

# install.packages('splitstackshape')
# install.packages('ISLR')
# install.packages('caret')
# install.packages('rattle')
library(splitstackshape)
library(ISLR)
library(pdp)
library(rpart,quietly = TRUE)
library(caret)
library(rpart.plot,quietly = TRUE)
library(rattle)
data
data$Attrition <- factor(data$Attrition)
summary(data)
View(data)
evaluation <- function(model, data, atype) {
  cat("\nConfusion matrix:\n")
  prediction = predict(model, data, type=atype)
  xtab = table(prediction, data$Attrition)
  print(xtab)
  cat("\nEvaluation:\n\n")
  accuracy = sum(prediction == data$Attrition)/length(data$Attrition)
  precision = xtab[1,1]/sum(xtab[,1])
  recall = xtab[1,1]/sum(xtab[1,])
  f = 2 * (precision * recall) / (precision + recall)
  cat(paste("Accuracy:\t", format(accuracy*100), "\n",sep=" "))
  cat(paste("Precision:\t", format(precision, digits=2), "\n",sep=" "))
  cat(paste("Recall:\t\t", format(recall, digits=2), "\n",sep=" "))
  cat(paste("F-measure:\t", format(f, digits=2), "\n",sep=" "))
}


colnames(data)
# Usage
```

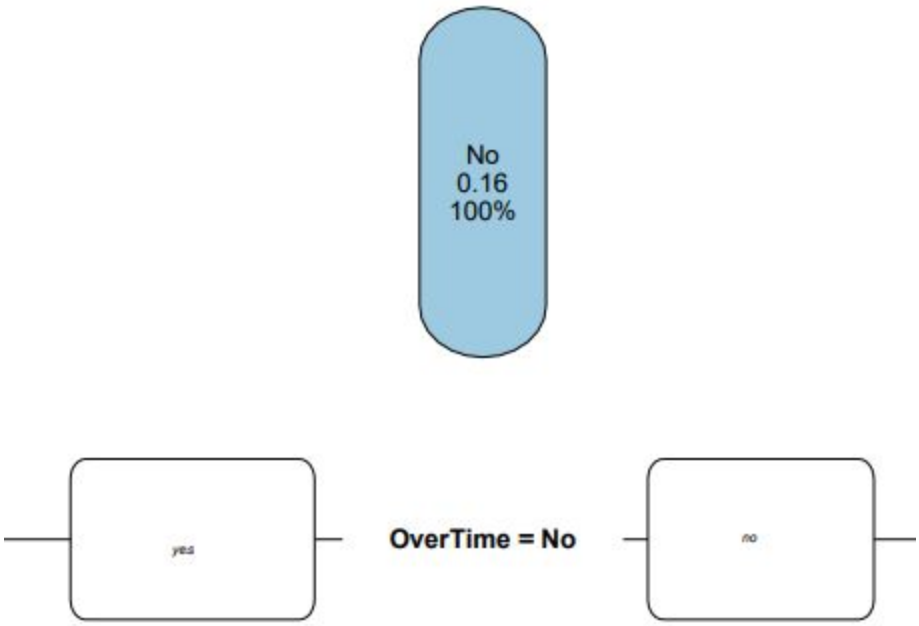
```
tree_with_params = rpart(formula = Attrition ~ .,data=data, method="class",control =
rpart.control(cp =0.000009, maxdepth = 5,minsplit = 10,minbucket=4))
#rpart.plot(tree_with_params,extra=107) # Plot without labels
rpart.plot(tree_with_params) # Plot with labels
evaluation(tree_with_params, data, "class")
```

```
Confusion matrix:

prediction   No  Yes
       No  1217  143
       Yes   16   94

Evaluation:

Accuracy:         89.18367
Precision:        0.99
Recall:           0.89
F-measure:        0.94
```

```
df <- data.frame(imp = tree_with_params$variable.importance)
df
```

	imp
MonthlyIncome	31.5955921
OverTime	24.0830129
TotalWorkingYears	17.3060017
JobRole	16.8858452
Age	12.7984623
StockOptionLevel	12.2481949
YearsInCurrentRole	9.9765810
DailyRate	9.6625435
MaritalStatus	9.1030814
YearsAtCompany	8.2266821
MonthlyRate	8.1724994
HourlyRate	8.0612217
EnvironmentSatisfaction	7.1393218
YearsWithCurrManager	6.4467461
EducationField	6.3972188
Department	6.3640480
JobSatisfaction	5.3147270
Education	4.7766667
PercentSalaryHike	4.0562250
RelationshipSatisfaction	3.6208285
NumCompaniesWorked	3.5122571
YearsSinceLastPromotion	3.3507246
WorkLifeBalance	2.9227518
JobInvolvement	2.1773333
JobLevel	1.8448042
DistanceFromHome	1.4790913
PerformanceRating	1.0971429
TrainingTimesLastYear	1.0726178
Gender	0.4464286
BusinessTravel	0.2818428

My interpretation of Decision Tree:



As you can see the root node consists of 3 pieces of information. First is the ‘No’ label. Second is the probability of the ‘Yes’ label of attrition and the third piece tells about the percentage of total sample being taken. It is also clear that the first split is happening on OverTime=No. Similarly on observing the decision tree, one can see TotalWorkingYears, MonthlyIncome`,JobRole, WorkLifeBalance are coming out to be important features. However, this data is skewed towards No. Thus we need to apply stratified sampling. If you are not aware of stratified sampling you can check my previous article on that.

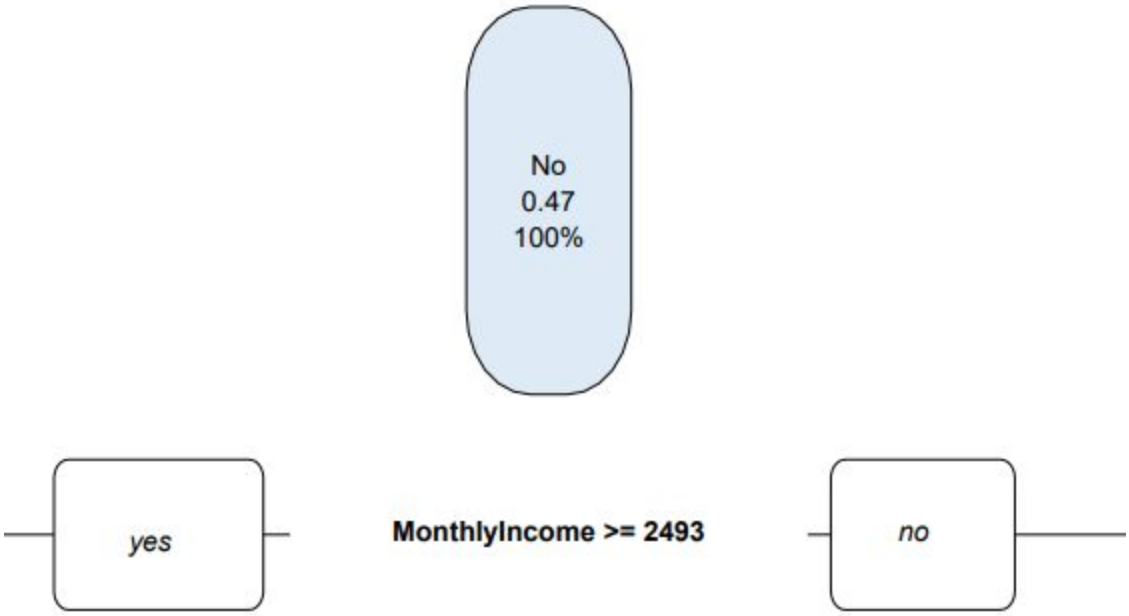
```
# Removing features which are having variable importance less than 4 and performing stratified sampling.
data <- data[setdiff(colnames(data), c('NumCompaniesWorked', 'RelationshipSatisfaction',
'JobInvolvement', 'JobLevel', 'DistanceFromHome',
'PerformanceRating','TrainingTimesLastYear','Gender','BusinessTravel'))]
data1<-data[data$Attrition == 'No',]
data1
data2<-data[data$Attrition == 'Yes',]
colnames(data1)
set.seed(42)  # good idea to set the random seed for reproducibility
data1_<-stratified(data1, c("Age",'Attrition'), 0.22)
data1_


data<-rbind(data1_,data2)
```

```
plot(data$Attrition)

colnames(data)
# Usage
tree_with_params = rpart(formula = Attrition ~ .,data=data, method="class",control =
rpart.control(cp =0.000009, maxdepth = 5,minsplit = 10,minbucket=4))
#rpart.plot(tree_with_params,extra=107) # Plot without labels
rpart.plot(tree_with_params) # Plot with labels
evaluation(tree_with_params, data, "class")
```

```
Confusion matrix:

prediction  No Yes
       No  250  86
       Yes  21 151

Evaluation:

Accuracy:        78.93701
Precision:       0.92
Recall:          0.74
F-measure:       0.82
```

```
#install.packages('tidyverse')
library(tidyverse)
df <- data.frame(imp = tree_with_params$variable.importance)
```

	imp
MonthlyIncome	39.1698048
OverTime	18.8421435
JobRole	16.5261037
TotalWorkingYears	14.8257264
DailyRate	13.9651877
Department	11.3810703
EnvironmentSatisfaction	8.9513209
Age	6.9928857
MonthlyRate	6.8650160
EducationField	6.1971540
WorkLifeBalance	5.5658510
YearsAtCompany	5.1204588
YearsInCurrentRole	4.8432432
YearsSinceLastPromotion	2.4216216
PercentSalaryHike	1.8269682
HourlyRate	1.3278161
YearsWithCurrManager	1.2108108
JobSatisfaction	0.6685462
StockOptionLevel	0.6685462



This time you can observe from the decision tree that the data is not skewed or imbalanced. Also the important features remains the same.

```
df2 <- df %>%
  tibble::rownames_to_column() %>%
  dplyr::rename("variable" = rowname) %>%
  dplyr::arrange(imp) %>%
  dplyr::mutate(variable = forcats::fct_inorder(variable))
ggplot2::ggplot(df2) +
  geom_col(aes(x = variable, y = imp),
           col = "black", show.legend = F) +
  coord_flip() +
  scale_fill_grey() +
  theme_bw()
#
ggplot2::ggplot(df2) +
  geom_segment(aes(x = variable, y = 0, xend = variable, yend = imp),
               size = 1.5, alpha = 0.7) +
  geom_point(aes(x = variable, y = imp, col = variable),
             size = 4, show.legend = F) +
  coord_flip() +
  theme_bw()
```

Features Influencing Attrition:

From the above decision trees, we can see that “YearsAtCompany ”, “MonthlyIncome”, “OverTime”, “WorkLifeBalance” are contributing to decision making. So, we can say these are the parameters which influences the Attrition in either way. Now, let's understand the parameters which influence these parameters. By this way, we can provide a strategy which can be used to reduce the Attrition rate.

1. **YearsAtCompany:** The Parameter, YearsAtCompany is a Continuos data. So, we are going to create a Multiple Linear Regression model and then we will find the parameters which are influencing this parameter.

```
#Multiple Linear Regression on YearsAtCompany to find out which features are affecting YearsAtCompany

data<-read.csv('dataset.csv')
summary(data)
colnames(data)
data$Attrition <- factor(data$Attrition)
data$BusinessTravel <- factor(data$BusinessTravel)
data$Department <- factor(data$Department)
data$Education <- factor(data$Education)
data$EducationField <- factor(data$EducationField)
data$EnvironmentSatisfaction <- factor(data$EnvironmentSatisfaction)
data$Gender <- factor(data$Gender)
data$JobInvolvement <- factor(data$JobInvolvement)
data$JobLevel <- factor(data$JobLevel)
data$JobRole <- factor(data$JobRole)
data$JobSatisfaction <- factor(data$JobSatisfaction)
data$MaritalStatus <- factor(data$MaritalStatus)
data$Over18 <- factor(data$Over18)
data$OverTime <- factor(data$OverTime)
data$PerformanceRating <- factor(data$PerformanceRating)
data$RelationshipSatisfaction <- factor(data$RelationshipSatisfaction)
data$StockOptionLevel <- factor(data$StockOptionLevel)
data$WorkLifeBalance <- factor(data$WorkLifeBalance)

colnames(data)
data <- data[setdiff(colnames(data), c('NumCompaniesWorked', 'RelationshipSatisfaction',
'JobInvolvement', 'JobLevel', 'DistanceFromHome',
'PerformanceRating','TrainingTimesLastYear','Gender','BusinessTravel'))]

data <- subset( data, select = -c(Attrition,EmployeeCount, EmployeeNumber, StandardHours) )
```

```
View(data)
colnames(data)
model <-
lm(YearsAtCompany~Age+MaritalStatus+MonthlyIncome+MonthlyRate+HourlyRate+JobSatisfa
ction+JobRole+DailyRate+Department+Education+EducationField+EnvironmentSatisfaction+St
ockOptionLevel+TotalWorkingYears+WorkLifeBalance+YearsAtCompany+YearsInCurrentRole+
YearsSinceLastPromotion+YearsWithCurrManager, data = data)
summary(model)
```

```
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    3.674e+00  1.312e+00   2.800 0.005176 **
Age                           -4.406e-02  1.193e-02  -3.692 0.000231 ***
MaritalStatusMarried          -2.698e-01  2.086e-01  -1.293 0.196078
MaritalStatusSingle           -2.541e-01  3.393e-01  -0.749 0.454066
MonthlyIncome                  1.312e-04  4.673e-05   2.808 0.005047 **
MonthlyRate                   -1.348e-05  1.083e-05  -1.245 0.213198
HourlyRate                    7.780e-04  3.801e-03   0.205 0.837857
JobSatisfaction2               4.404e-01  2.475e-01   1.780 0.075334 .
JobSatisfaction3               4.223e-01  2.234e-01   1.891 0.058863 .
JobSatisfaction4               3.090e-01  2.215e-01   1.395 0.163144
JobRoleHuman Resources        -2.365e+00  1.106e+00  -2.139 0.032626 *
JobRoleLaboratory Technician   3.041e-01  3.548e-01   0.857 0.391531
JobRoleManager                -1.530e-01  6.091e-01  -0.251 0.801735
JobRoleManufacturing Director -5.916e-01  3.555e-01  -1.664 0.096330 .
JobRoleResearch Director      -1.512e+00  5.445e-01  -2.778 0.005543 **
JobRoleResearch Scientist      3.909e-01  3.499e-01   1.117 0.264067
JobRoleSales Executive        -1.473e+00  7.020e-01  -2.098 0.036063 *
JobRoleSales Representative   -8.645e-01  7.778e-01  -1.111 0.266573
DailyRate                     -4.246e-04  1.921e-04  -2.210 0.027265 *
DepartmentResearch & Development -3.227e+00  1.060e+00  -3.044 0.002379 **
DepartmentSales               -1.984e+00  1.098e+00  -1.807 0.070914 .
Education2                    -3.414e-01  2.904e-01  -1.176 0.239849
Education3                    -6.150e-01  2.609e-01  -2.358 0.018529 *
Education4                    -2.829e-01  2.792e-01  -1.013 0.311138
Education5                    -2.773e-01  4.879e-01  -0.568 0.569875
EducationFieldLife Sciences    7.138e-01  7.624e-01   0.936 0.349301
EducationFieldMarketing        8.012e-01  8.107e-01   0.988 0.323185
EducationFieldMedical          9.225e-01  7.652e-01   1.206 0.228170
EducationFieldOther            4.954e-01  8.200e-01   0.604 0.545824
EducationFieldTechnical Degree 4.791e-01  7.953e-01   0.602 0.547006
EnvironmentSatisfaction2      -1.050e-01  2.471e-01  -0.425 0.670971
EnvironmentSatisfaction3       1.279e-01  2.243e-01   0.570 0.568705
EnvironmentSatisfaction4      -8.822e-02  2.244e-01  -0.393 0.694280
StockOptionLevel1              3.793e-02  2.710e-01   0.140 0.888718
StockOptionLevel2             -3.395e-01  3.454e-01  -0.983 0.325840
```

```
StockOptionLevel3          -3.905e-01  4.117e-01  -0.949 0.342979
TotalWorkingYears           1.966e-01  2.041e-02   9.634  < 2e-16 ***
WorkLifeBalance2           -5.962e-02  3.683e-01  -0.162 0.871431
WorkLifeBalance3           -1.744e-01  3.460e-01  -0.504 0.614403
WorkLifeBalance4           -2.050e-01  4.084e-01  -0.502 0.615713
YearsInCurrentRole          5.066e-01  3.249e-02  15.595  < 2e-16 ***
YearsSinceLastPromotion      3.039e-01  2.988e-02  10.170  < 2e-16 ***
YearsWithCurrManager         6.048e-01  3.223e-02  18.765  < 2e-16 ***
```

The features having ** or *** adjacent to Pr(>|t|) value are important i.e. are influencing YearsAtCompany feature. The mean value of YearsAtCompany and Attrition is shown below:

Attrition	YearsAtCompany(Mean)
Yes	5.1
No	7.3

From the above table, we can say that the Attrition rate decreases with increase in YearsAtCompany. The parameters which influence YearsAtCompany are: YearsWithCurrManager, YearsSinceLastPromotion, YearsInCurrentRole, TotalWorkingYears, Department, JobRole, MonthlyIncome and Age.

2. MonthlyIncome:

```
model <-
lm(MonthlyIncome~Age+MaritalStatus+YearsAtCompany+MonthlyRate+HourlyRate+JobSatisfaction+JobRole+DailyRate+Department+Education+EducationField+EnvironmentSatisfaction+StockOptionLevel+TotalWorkingYears+WorkLifeBalance+YearsAtCompany+YearsInCurrentRole+YearsSinceLastPromotion+YearsWithCurrManager, data = data)
summary(model)
```

```
Call:
lm(formula = MonthlyIncome ~ Age + MaritalStatus + YearsAtCompany +
    MonthlyRate + HourlyRate + JobSatisfaction + JobRole + DailyRate +
    Department + Education + EducationField + EnvironmentSatisfaction +
    StockOptionLevel + TotalWorkingYears + WorkLifeBalance +
    YearsAtCompany + YearsInCurrentRole + YearsSinceLastPromotion +
    YearsWithCurrManager, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-5308.6 -1073.5   -88.1   974.1  5412.0
```

```
Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      4.614e+03  7.330e+02   6.294 4.11e-10 ***
Age                             -6.134e+00  6.771e+00  -0.906  0.36514
MaritalStatusMarried             8.387e+01  1.179e+02   0.712  0.47688
MaritalStatusSingle              2.647e+02  1.916e+02   1.382  0.16728
YearsAtCompany                   4.188e+01  1.491e+01   2.808  0.00505 **
MonthlyRate                     -6.374e-03  6.118e-03  -1.042  0.29767
HourlyRate                      -1.199e+00  2.147e+00  -0.558  0.57661
JobSatisfaction2                 3.437e+01  1.399e+02   0.246  0.80602
JobSatisfaction3                 5.676e+01  1.263e+02   0.449  0.65328
JobSatisfaction4                 2.358e+01  1.252e+02   0.188  0.85063
JobRoleHuman Resources          -2.003e+03  6.236e+02  -3.212  0.00135 **
JobRoleLaboratory Technician    -2.992e+03  1.842e+02 -16.245  < 2e-16 ***
JobRoleManager                   7.678e+03  2.777e+02  27.650  < 2e-16 ***
JobRoleManufacturing Director    9.860e+01  2.010e+02   0.490  0.62387
JobRoleResearch Director         7.161e+03  2.433e+02  29.434  < 2e-16 ***
JobRoleResearch Scientist       -3.028e+03  1.808e+02 -16.748  < 2e-16 ***
JobRoleSales Executive           5.950e+02  3.969e+02   1.499  0.13404
JobRoleSales Representative     -2.553e+03  4.344e+02  -5.878 5.17e-09 ***
DailyRate                        1.008e-01  1.087e-01   0.928  0.35374
DepartmentResearch & Development 1.443e+02  6.008e+02   0.240  0.81017
DepartmentSales                 -4.917e+02  6.207e+02  -0.792  0.42836
Education2                      -2.536e+02  1.640e+02  -1.546  0.12224
Education3                      -6.029e+01  1.476e+02  -0.408  0.68307
Education4                      -2.554e+02  1.576e+02  -1.620  0.10543
Education5                      -2.721e+02  2.756e+02  -0.988  0.32352
EducationFieldLife Sciences     -1.898e+02  4.308e+02  -0.441  0.65959
EducationFieldMarketing         -6.580e+01  4.581e+02  -0.144  0.88582
EducationFieldMedical           -3.303e+02  4.324e+02  -0.764  0.44503
EducationFieldOther             -1.259e+02  4.633e+02  -0.272  0.78588
EducationFieldTechnical Degree  -9.742e+01  4.493e+02  -0.217  0.82837
EnvironmentSatisfaction2        -6.993e+01  1.396e+02  -0.501  0.61653
EnvironmentSatisfaction3         2.726e+01  1.267e+02   0.215  0.82968
EnvironmentSatisfaction4        -5.435e+00  1.268e+02  -0.043  0.96581
StockOptionLevel1                2.672e+02  1.529e+02   1.747  0.08079 .
StockOptionLevel2                1.173e+02  1.951e+02   0.601  0.54789
StockOptionLevel3                2.902e+01  2.326e+02   0.125  0.90074
TotalWorkingYears                1.944e+02  1.073e+01  18.120  < 2e-16 ***
WorkLifeBalance2                 4.430e+02  2.077e+02   2.133  0.03312 *
WorkLifeBalance3                 4.132e+02  1.952e+02   2.117  0.03444 *
WorkLifeBalance4                 3.006e+02  2.306e+02   1.304  0.19251
YearsInCurrentRole              -1.671e+01  1.985e+01  -0.842  0.40000
YearsSinceLastPromotion          4.142e+01  1.745e+01   2.374  0.01774 *
YearsWithCurrManager            -6.466e+01  2.026e+01  -3.191  0.00145 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Residual standard error: 1650 on 1427 degrees of freedom
Multiple R-squared: 0.8807, Adjusted R-squared: 0.8772
F-statistic: 250.9 on 42 and 1427 DF, p-value: < 2.2e-16

The features having ** or *** adjacent to Pr(>|t|) value are important i.e. are influencing YearsAtCompany feature. The mean value of MonthlyIncomeand Attrition is shown below:

Attrition	MonthlyIncome(Mean)
No	4787.09
Yes	6833.74

From the above table, we can say that the higher the salary, the lower the rate of attrition. The parameters which influence MonthlyIncomeare: YearsAtCompany, JobRole, TotalWorkingYears, YearsWithCurrManager.

3. **OverTime:** OverTime is categorical data. We need to create a classification model. Lets create a decision tree model for the same.

```
library(splitstackshape)
library(ISLR)
library(pdp)
library(rpart,quietly = TRUE)
library(caret)
library(rpart.plot,quietly = TRUE)
library(rattle)
# data
# data$Attrition <- factor(data$Attrition)
# summary(data)
# View(data)
evaluation <- function(model, data, atype) {
  cat("\nConfusion matrix:\n")
  prediction = predict(model, data, type=atype)
  xtab = table(prediction, data$OverTime)
  print(xtab)
  cat("\nEvaluation:\n\n")
  accuracy = sum(prediction == data$OverTime)/length(data$OverTime)
  precision = xtab[1,1]/sum(xtab[,1])
  recall = xtab[1,1]/sum(xtab[1,])
  f = 2 * (precision * recall) / (precision + recall)
  cat(paste("Accuracy:\t", format(accuracy*100), "\n",sep=" "))
```

```
  cat(paste("Precision:\t", format(precision, digits=2), "\n",sep=" "))
  cat(paste("Recall:\t\t", format(recall, digits=2), "\n",sep=" "))
  cat(paste("F-measure:\t", format(f, digits=2), "\n",sep=" "))
}

plot(data$OverTime)

data1<-data[data$OverTime == 'No',]
data1
data2<-data[data$OverTime == 'Yes',]


set.seed(42)  # good idea to set the random seed for reproducibility
data1_<-stratified(data1, c("Age",'OverTime'), 0.40)
data1_


data<-rbind(data1_,data2)
plot(data$OverTime)
colnames(data)
# Usage
tree_with_params = rpart(formula = OverTime ~ .,data=data, method="class",control =
rpart.control(cp =0.000009, maxdepth = 5,minsplit = 10,minbucket=4))
#rpart.plot(tree_with_params,extra=107) # Plot without labels
rpart.plot(tree_with_params) # Plot with labels
evaluation(tree_with_params, data, "class")
df <- data.frame(imp = tree_with_params$variable.importance)
df
```

Variable Importance:

	imp
PercentSalaryHike	8.0098052
EducationField	7.2920987
YearsAtCompany	6.2583151
JobSatisfaction	6.0415239
TotalWorkingYears	5.4435300
EnvironmentSatisfaction	4.6922371
Education	3.8129810
Age	3.7687662
JobRole	3.6182568
YearsInCurrentRole	2.8979502
MonthlyIncome	2.5728339
WorkLifeBalance	2.4317618
HourlyRate	1.9395543
YearsWithCurrManager	1.7587301
MonthlyRate	1.5688580
DailyRate	1.4935828
StockOptionLevel	0.4756650
Department	0.2640912

From the above table, we can say that the most important parameters for the decision tree are: PercentSalaryHike, EducationField, YearsAtCompany, JobSatisfaction, TotalWorkingYears, EnvironmentSatisfaction.

Decision Tree Interpretation:

- In order to balance the tree, we have used stratified sampling.
- First Split is on JobRole= Human Resources,Laboratory Technician,Research Director.
- Second split is on HourlyRate >= 81 and YearsInCurrentRole >= 5.
- Third plit is on MonthlyIncome < 2666, StockOptionLevel=0,2 and 1,2.
- So, we can say these parameters are important in decision making.

4. **WorkLifeBalance:** WorkLifeBalance is categorical data. We need to create a classification model. Lets create a decision tree model for the same.

```
data<-read.csv('dataset.csv')
summary(data)
colnames(data)
data$Attrition <- factor(data$Attrition)
data$BusinessTravel <- factor(data$BusinessTravel)
data$Department <- factor(data$Department)
data$Education <- factor(data$Education)
data$EducationField <- factor(data$EducationField)
data$EnvironmentSatisfaction <- factor(data$EnvironmentSatisfaction)
data$Gender <- factor(data$Gender)
data$JobInvolvement <- factor(data$JobInvolvement)
data$JobLevel <- factor(data$JobLevel)
data$JobRole <- factor(data$JobRole)
data$JobSatisfaction <- factor(data$JobSatisfaction)
data$MaritalStatus <- factor(data$MaritalStatus)
data$Over18 <- factor(data$Over18)
data$OverTime <- factor(data$OverTime)
data$PerformanceRating <- factor(data$PerformanceRating)
data$RelationshipSatisfaction <- factor(data$RelationshipSatisfaction)
data$StockOptionLevel <- factor(data$StockOptionLevel)
data$WorkLifeBalance <- factor(data$WorkLifeBalance)

colnames(data)
data <- data[setdiff(colnames(data), c('NumCompaniesWorked', 'RelationshipSatisfaction',
'JobInvolvement', 'JobLevel', 'DistanceFromHome',
'PerformanceRating','TrainingTimesLastYear','Gender','BusinessTravel'))]

data <- subset( data, select = -c(Attrition,EmployeeCount, EmployeeNumber, StandardHours) )
```

```
View(data)
colnames(data)


library(splitstackshape)
library(ISLR)
library(pdp)
library(rpart,quietly = TRUE)
library(caret)
library(rpart.plot,quietly = TRUE)
library(rattle)
# data
# data$Attrition <- factor(data$Attrition)
# summary(data)
# View(data)
evaluation <- function(model, data, atype) {
  cat("\nConfusion matrix:\n")
  prediction = predict(model, data, type=atype)
  xtab = table(prediction, data$WorkLifeBalance)
  print(xtab)
  cat("\nEvaluation:\n\n")
  accuracy = sum(prediction == data$WorkLifeBalance)/length(data$WorkLifeBalance)
  precision = xtab[1,1]/sum(xtab[,1])
  recall = xtab[1,1]/sum(xtab[1,])
  f = 2 * (precision * recall) / (precision + recall)
  cat(paste("Accuracy:\t", format(accuracy*100), "\n",sep=" "))
  cat(paste("Precision:\t", format(precision, digits=2), "\n",sep=" "))
  cat(paste("Recall:\t\t", format(recall, digits=2), "\n",sep=" "))
  cat(paste("F-measure:\t", format(f, digits=2), "\n",sep=" "))
}

plot(data$WorkLifeBalance)

data1<-data[data$WorkLifeBalance == '1',]
data2<-data[data$WorkLifeBalance == '2',]
data3<-data[data$WorkLifeBalance == '3',]
data4<-data[data$WorkLifeBalance == '4',]
set.seed(42)  # good idea to set the random seed for reproducibility
data2_<-stratified(data2, c("Age",'WorkLifeBalance'), 0.25)
View(data2_)
data3_<-stratified(data3, c("Age",'WorkLifeBalance'), 0.10)
View(data3_)
data4_<-stratified(data4, c("Age",'WorkLifeBalance'), 0.60)
```

```
View(data4_)


data<-rbind(data1,data2_,data3_,data4_)
plot(data$WorkLifeBalance)
colnames(data)
# Usage
tree_with_params = rpart(formula = WorkLifeBalance ~ .,data=data, method="class",control =
rpart.control(cp =0.000009, maxdepth = 5,minsplit = 10,minbucket=4))
#rpart.plot(tree_with_params,extra=107) # Plot without labels
rpart.plot(tree_with_params) # Plot with labels
evaluation(tree_with_params, data, "class")
df <- data.frame(imp = tree_with_params$variable.importance)
Df
```

Variable Importance:

	imp
MonthlyIncome	12.93540525
EducationField	10.13637001
HourlyRate	7.58350212
DailyRate	7.34359595
YearsAtCompany	6.88355396
Education	6.68903227
MonthlyRate	6.19829191
TotalWorkingYears	4.88910933
JobRole	4.57855511
YearsSinceLastPromotion	4.46394794
StockOptionLevel	3.80742115
JobSatisfaction	3.46212121
OverTime	2.30866581
PercentSalaryHike	2.13346485
YearsInCurrentRole	1.92763385
YearsWithCurrManager	1.53165304
Age	1.51426827
Department	1.30431302
MaritalStatus	0.89212776
EnvironmentSatisfaction	0.08362465

From the above table, we can say that the most important parameters for the decision tree are: MonthlyIncome, EducationField, HourlyRate, DailyRate, YearsAtCompany, Education, MonthlyRate, TotalWorkingYears, JobRole, YearsSinceLastPromotion.

Decision Tree Interpretation:

- In order to balance the tree, we have used stratified sampling.
- First Split is on JobRole = Human Resources,Laboratory Technician,Manager,Research Director.

- Second split is on EducationField = Human Resources,Life Sciences,Other,Technical Degree and MonthlyIncome < 6521.
- Third split is on MonthlyRate < 4039, OverTime = Yes, HourlyRate < 34 and YearsAtCompany >= 10.
- So, we can say these parameters are important in decision making.

Conclusion: Using this dataset, tried to highlight those factors that affect Attrition rate. With the help of statistical tools mentioned above and machine learning models, I observed that the parameters WorkLifeBalance, OverTime, MonthlyIncome,YearsAtCompany influences the attrition rate. I also observed that the parameters which influence these four features, managing which the concerned team can control the features which would eventually effect the attrition rate.