ibm <- read.csv("HR.csv")

suppressMessages(library(ggplot2))

suppressMessages(library(grid))

suppressMessages(library(gridExtra))

suppressMessages(library(plyr))

suppressMessages(library(rpart))

suppressMessages(library(rpart.plot))

suppressMessages(library(randomForest))

suppressMessages(library(caret))

suppressMessages(library(gbm))

suppressMessages(library(survival))

suppressMessages(library(pROC))

suppressMessages(library(DMwR))

suppressMessages(library(scales))

g1 <- ggplot(ibm,

aes(x = MonthlyIncome, fill = Attrition)) +

geom\_density(alpha = 0.7) +

scale\_fill\_manual(values = c("#386cb0","#fdb462"))

g2 <- ggplot(ibm,

aes(x = HourlyRate, fill = Attrition)) +

geom\_density(alpha = 0.7) +

scale\_fill\_manual(values = c("#386cb0","#fdb462"))

g3 <- ggplot(ibm,

aes(x = DailyRate, fill = Attrition)) +

geom\_density(alpha = 0.7) +

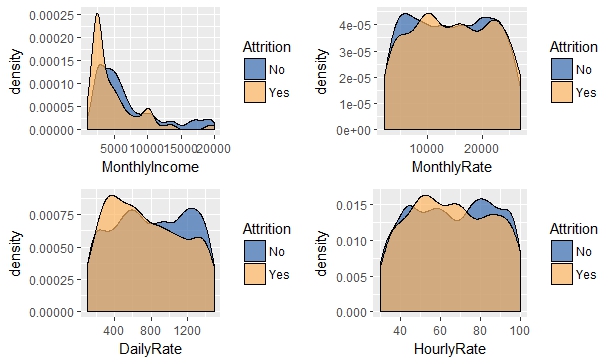
scale\_fill\_manual(values = c("#386cb0","#fdb462"))

g4 <- ggplot(ibm,

aes(x = MonthlyRate, fill = Attrition)) +

geom\_density(alpha = 0.7) +

scale\_fill\_manual(values = c("#386cb0","#fdb462"))

grid.arrange(g1, g2, g3, g4, ncol = 2, nrow = 2) 

ggplot(ibm,

aes(y = YearsSinceLastPromotion, x = YearsAtCompany, colour = OverTime)) +

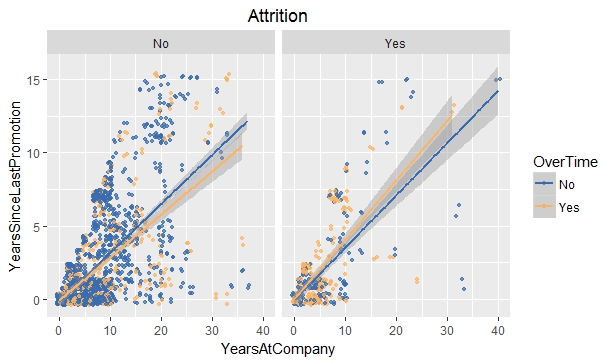
geom\_jitter(size = 1, alpha = 0.7) +

geom\_smooth(method = "gam") +

facet\_wrap(~ Attrition) +

ggtitle("Attrition") +

scale\_colour\_manual(values = c("#386cb0","#fdb462")) +

theme(plot.title = element\_text(hjust = 0.5)) 

ggplot(ibm,

aes(x = OverTime, group = Attrition)) +

geom\_bar(aes(y = ..prop.., fill = factor(..x..)),

stat="count",

alpha = 0.7) +

geom\_text(aes(label = scales::percent(..prop..), y = ..prop.. ),

stat= "count",

vjust = -.5) +

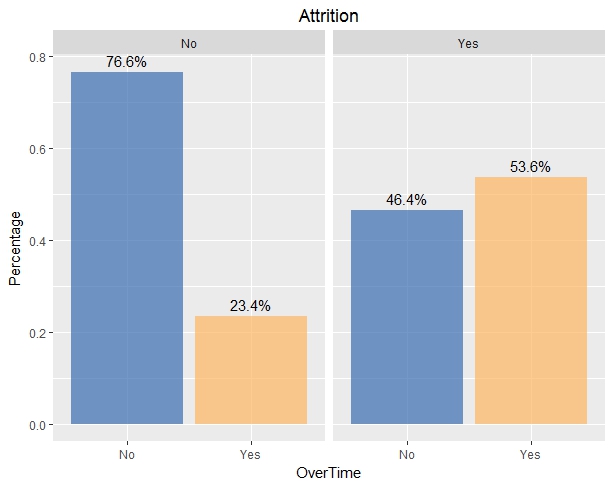
labs(y = "Percentage", fill= "OverTime") +

facet\_grid(~Attrition) +

scale\_fill\_manual(values = c("#386cb0","#fdb462")) +

theme(legend.position = "none", plot.title = element\_text(hjust = 0.5)) +

ggtitle("Attrition")



ggplot(ibm,

aes(x= WorkLifeBalance, y=DistanceFromHome, group = WorkLifeBalance, fill = WorkLifeBalance)) +

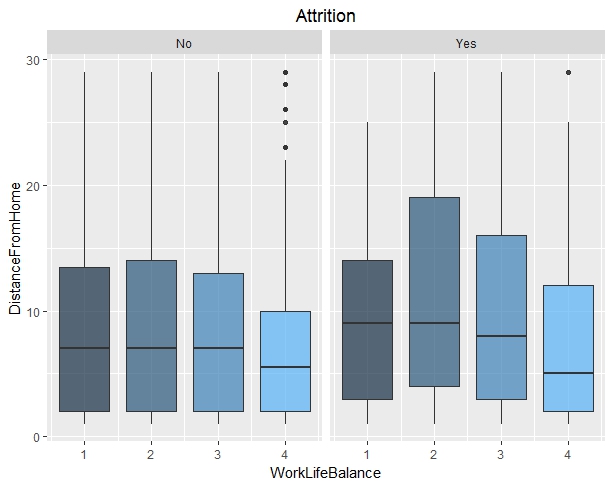
geom\_boxplot(alpha=0.7) +

theme(legend.position="none") +

facet\_wrap(~ Attrition) +

ggtitle("Attrition") +

theme(plot.title = element\_text(hjust = 0.5))



set.seed(3221)

# Getting rid of long variable names & certain unuseful variables

levels(ibm$JobRole) <- c("HC", "HR", "Lab", "Man", "MDir", "RsD", "RsSci", "SlEx", "SlRep")

levels(ibm$EducationField) <- c("HR", "LS", "MRK", "MED", "NA", "TD")

ibm <- ibm[c(-9,-10,-22,-27)]

# Creating train & test sets

n <- nrow(ibm)

rnd <- sample(n, n \* .70)

train <- ibm[rnd,]

test <- ibm[-rnd,]

# Modeling

dtree <- rpart(Attrition ~., data = train)

preds <- predict(dtree, test, type = "class")

rocv <- roc(as.numeric(test$Attrition), as.numeric(preds))

rocv$auc

prop.table(table(test$Attrition, preds, dnn = c("Actual", "Predicted")),1)

set.seed(2343)

# Random forest

fit.forest <- randomForest(Attrition ~., data = train)

rfpreds <- predict(fit.forest, test, type = "class")

rocrf <- roc(as.numeric(test$Attrition), as.numeric(rfpreds))

rocrf$auc

set.seed(3433)

# Simple GBM

gbmfit <- train(Attrition ~.,

data = train,

method = "gbm")

gbmpreds <- predict(gbmfit, test)

rocgbm <- roc(as.numeric(test$Attrition), as.numeric(gbmpreds))

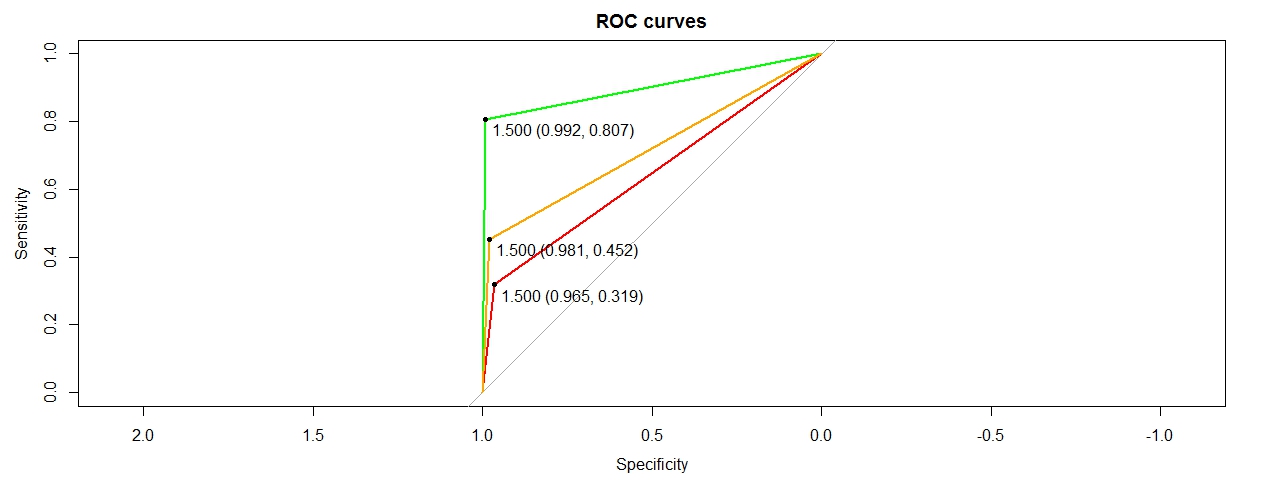
rocgbm$auc

varImp(fit.forest)

plot(rocv, ylim = c(0,1), print.thres = T, main = "ROC curves", col = "salmon")

plot(rocrf, ylim = c(0,1), print.thres = T, col = "darkolivegreen", add = T)

plot(rocgbm, ylim = c(0,1), print.thres = T, col = "burlywood", add = T)



By comparing these three models we can conclude that sensitivity of the decision tree is lower when comparing to other two models.

The top 5 factors that influence attrition are:

1. Monthly income

2. Age

3. Overtime

4. Total years of working

5. Job role

Other important factors:

1. Environment Satisfaction

2. Distance From Home

3. Hourly Rate

4. Daily Rate

**Conclusion**:

When we concentrate on top 5 factors we can lower the rate of attrition in feature.