本篇案例来源：

知乎|Joffy Zhong

背景介绍

我曾为一些中小企业做过管理咨询，其中部分模块涉及到人力资源，这也是我在工作过程中比较感兴趣去研究的领域，包括人才的成长与管理等，此次分析是基于一个有意思的数据集，去探索企业员工流失的影响因素，如果你有合适的员工数据集，可以延用此文的分析思路。（此次分析也参考了来自 kaggle 和 RPub 的一些大神的文章，有兴趣的朋友可以通过文中的链接进行查看）

本次分析的数据集来自 IBM 数据科学家创建的虚构的员工流失数据（你可以在这里拿到数据：SAMPLE DATA: HR Employee Attrition and Performance）

员工流失是困扰企业的众多关键问题之一，在这次分析中，我将努力开展以下工作内容：

- 对一些重要的变量进行快速可视化及探索性分析，特别是与基础信息、收入、晋升、满意度、绩效和工作与生活平衡等相关的变量（西方文化讲究 Work-life Balance）
- 分析导致员工流失的因素，并探索各个变量的影响程度
- 通过有效的算法构建模型，用于预测员工是否要辞职

最终目标是使用分析的过程和结果，有助于在利用真实数据集进行建模预测分析的时候减少员工流失，辅助人力资源团队进行关键的干预工作，让管理层指导哪些因素影响了“留人”，反过来促进企业做好“选人”“育人”“用人”。

结论和建议：

- 员工离开背后的主要原因很可能是投入与回报的失衡；加班，工作的投入带来的回报是否不匹配，检查公司是否有有效的加班政策。
- 在某种程度上，工作与生活的不平衡也是造成员工离开的原因，检查员工背后认为自身工作与生活不平衡的原因也许是一个有效的手段；例如加班，离家远的情况重复出现，是否有的远程工作的支持。
- 高薪也许不是关键的保障，在探索分析的结果显示，拥有员工优先认股权，是员工更为关注的另一个报酬形式。
- 年龄和任职过的公司等因素似乎与离职率高有较大的关系，这给人力资源部门的同事提供了识人方面的有效信息；当然这仅仅是虚构数据集的结果，分析请以公司真实数据，往往有公司就是有吸引年轻人的文化。
- 最后，如果能得到公司员工的一个新数据集，就可以根据建立好的模型计算概率并查看哪些员工确实容易离开

整个过程是一次较简单的探索分析和建模的过程，如果你感兴趣可以自行建模或者在此基础上优化，欢迎与我交流。

参考与引用：

IBM's attrition. Tackling class imbalance with GBM

RPubs - Prediction of Attrition - IBM HR Dataset

Employee Attrition: Human Resource Concern

数据集

该数据集由 1470 行/数据点和 35 列/属性组成，部分变量描述的理解上会存在一定的文化差异。

重点关注的变量说明：

变量类型	变量名	描述	变量范围
结果变量	**Attrition**	员工是否流失	Yes，No
自变量	Gender	性别	Female，Male
	Age	年龄	数值
	Education	学历	1 'Below College' 2 'College' 3 'Bachelor' 4 'Master' 5 'Doctor'
	NumCompaniesWorked	任职过的企业数量	数值
	TotalWorkingYears	工龄	数值
	MaritalStatus	婚姻状况	Divorced，Married，Single
	YearsAtCompany	在公司工作时间	数值
	JobRole	职位	HR，HC等9个职位
	JobLevel	职位等级	1~5，5个等级
	MonthlyIncome	月收入	数值
	JobInvolvement	工作投入	1'低' 2'中' 3'高' 4'非常高'
	PerformanceRating	绩效评分	1'差' 2'良好' 3'优秀' 4'卓越'
	StockOptionLevel	员工优先认股权	数值
	PercentSalaryHike	涨薪百分比	数值
	TrainingTimesLastYear	上一年培训次数	数值
	YearsSinceLastPromotion	距离上次升职的时间	数值
	EnvironmentSatisfaction	环境满意度	1'低' 2'中' 3'高' 4'非常高'
	JobSatisfaction	工作满意度	1'低' 2'中' 3'高' 4'非常高'
	RelationshipSatisfaction	关系满意度	1'低' 2'中' 3'高' 4'非常高'
	WorkLifeBalance	工作与生活平衡情况	1'差' 2'良好' 3'较好' 4'非常好'
	DistanceFromHome	上班离家距离	数值
	OverTime	是否加班	Yes，No
	BusinessTravel	出差情况	Non，Rarely，Frequently

探索性数据分析

```
#加载需要的包
library(ggplot2)
library(grid)
library(gridExtra)
library(plyr)
library(rpart)
library(rpart.plot)
library(randomForest)
library(caret)
library(gbm)
library(survival)
library(pROC)
library(DMwR)
library(scales)
#加载数据
Attr.df <- read.csv("HR-Employee-Attrition.csv",header = TRUE)
```

```
str(Attr.df)
#查看描述统计信息
summary(Attr.df)
```

输出结果：

```
Age       Attrition           BusinessTravel   DailyRate
Department
 Min.   :18.00   No :1233   Non-Travel       : 150   Min.   :
102.0   Human Resources       : 63
 1st Qu.:30.00   Yes: 237   Travel_Frequently: 277   1st Qu.:
465.0   Research & Development:961
 Median :36.00              Travel_Rarely    :1043   Median :
802.0   Sales                 :446
 Mean   :36.92                                       Mean   : 802.5
 3rd Qu.:43.00                                       3rd
Qu.:1157.0
 Max.   :60.00                                       Max.   :1499.0

 DistanceFromHome   Education     EducationField
EnvironmentSatisfaction    Gender
 Min.   : 1.000   Min.   :1.000   HR : 27        Min.   :1.000
Female:588
 1st Qu.: 2.000   1st Qu.:2.000   LS :606        1st Qu.:2.000
Male  :882
 Median : 7.000   Median :3.000   MRK:159        Median :3.000
 Mean   : 9.193   Mean   :2.913   MED:464        Mean   :2.722
 3rd Qu.:14.000   3rd Qu.:4.000   NA : 82        3rd Qu.:4.000
 Max.   :29.000   Max.   :5.000   TD :132        Max.   :4.000

   HourlyRate     JobInvolvement    JobLevel        JobRole
JobSatisfaction  MaritalStatus
 Min.   : 30.00   Min.   :1.00   Min.   :1.000   SlEx   :326
Min.   :1.000   Divorced:327
 1st Qu.: 48.00   1st Qu.:2.00   1st Qu.:1.000   RsSci  :292
1st Qu.:2.000   Married :673
 Median : 66.00   Median :3.00   Median :2.000   Lab    :259
Median :3.000   Single  :470
 Mean   : 65.89   Mean   :2.73   Mean   :2.064   MDir   :145
Mean   :2.729
 3rd Qu.: 83.75   3rd Qu.:3.00   3rd Qu.:3.000   HC     :131
3rd Qu.:4.000
```

```
 Max.   :100.00   Max.   :4.00   Max.   :5.000   Man    :102
Max.   :4.000
                                                (Other):215
 MonthlyIncome    MonthlyRate    NumCompaniesWorked OverTime
PercentSalaryHike PerformanceRating
 Min.   : 1009   Min.   : 2094   Min.   :0.000      No :1054
Min.   :11.00     Min.   :3.000
 1st Qu.: 2911   1st Qu.: 8047   1st Qu.:1.000      Yes: 416
1st Qu.:12.00     1st Qu.:3.000
 Median : 4919   Median :14236   Median :2.000
Median :14.00     Median :3.000
 Mean   : 6503   Mean   :14313   Mean   :2.693
Mean   :15.21     Mean   :3.154
 3rd Qu.: 8379   3rd Qu.:20462   3rd Qu.:4.000
3rd Qu.:18.00     3rd Qu.:3.000
 Max.   :19999   Max.   :26999   Max.   :9.000
Max.   :25.00     Max.   :4.000

 RelationshipSatisfaction StockOptionLevel TotalWorkingYears
TrainingTimesLastYear WorkLifeBalance
 Min.   :1.000            Min.   :0.0000   Min.   : 0.00
Min.   :0.000         Min.   :1.000
 1st Qu.:2.000            1st Qu.:0.0000   1st Qu.: 6.00     1st
Qu.:2.000         1st Qu.:2.000
 Median :3.000            Median :1.0000   Median :10.00
Median :3.000         Median :3.000
 Mean   :2.712            Mean   :0.7939   Mean   :11.28
Mean   :2.799         Mean   :2.761
 3rd Qu.:4.000            3rd Qu.:1.0000   3rd Qu.:15.00     3rd
Qu.:3.000         3rd Qu.:3.000
 Max.   :4.000            Max.   :3.0000   Max.   :40.00
Max.   :6.000         Max.   :4.000

 YearsAtCompany   YearsInCurrentRole YearsSinceLastPromotion
YearsWithCurrManager
 Min.   : 0.000   Min.   : 0.000     Min.   : 0.000
Min.   : 0.000
 1st Qu.: 3.000   1st Qu.: 2.000     1st Qu.: 0.000         1st
Qu.: 2.000
 Median : 5.000   Median : 3.000     Median : 1.000
Median : 3.000
 Mean   : 7.008   Mean   : 4.229     Mean   : 2.188
Mean   : 4.123
```

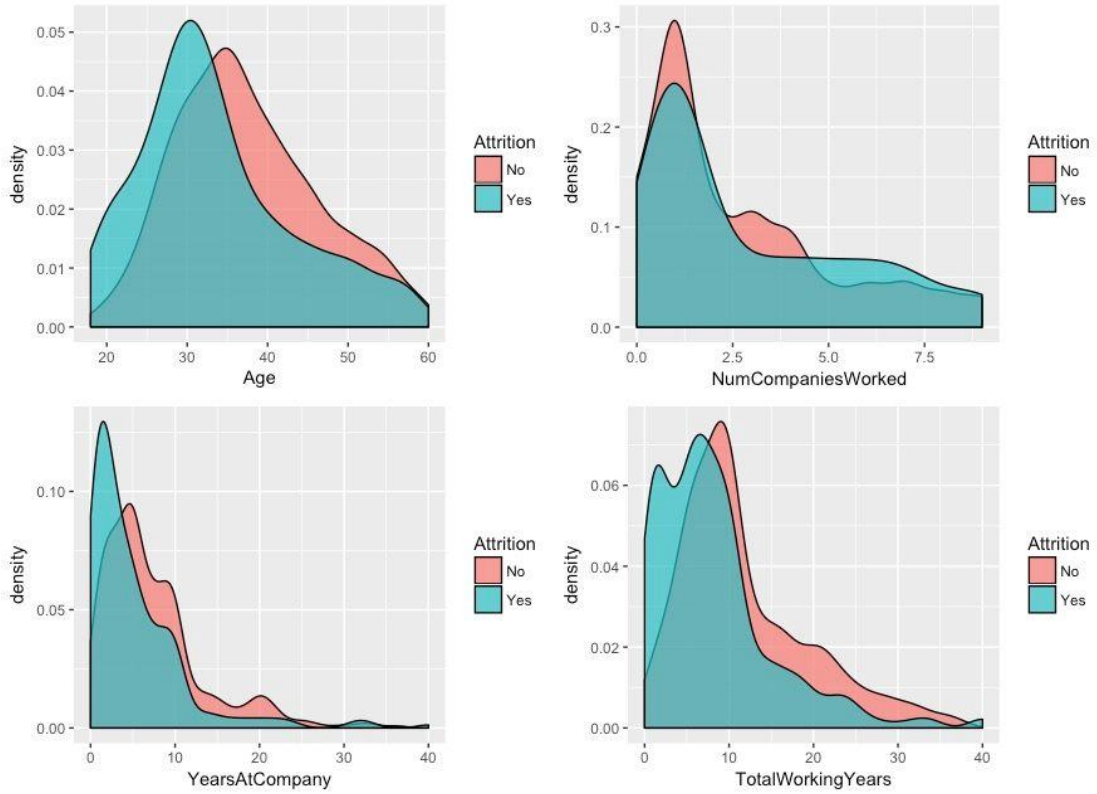
```
 3rd Qu.: 9.000   3rd Qu.: 7.000     3rd Qu.: 3.000          3rd
Qu.: 7.000
 Max.   :40.000   Max.   :18.000     Max.   :15.000
Max.   :17.000
```

从描述统计信息看：

- 企业员工不流失与流失的比例约为 5：1
- 企业员工平均年龄约为 36 岁
- 企业员工平均收入约为 6500 美元，中值为 4919 美元，中值更能反映企业薪资水平

探索基础信息 Gender，Age，Department，JobLevel，Education 等变量与员工流失的关系

```
g1 <- ggplot(Attr.df, aes(x = Age, fill = Attrition)) +
  geom_density(alpha = 0.7)

g2 <- ggplot(Attr.df, aes(x = NumCompaniesWorked, fill =
Attrition)) +
  geom_density(alpha = 0.7)

g3 <- ggplot(Attr.df, aes(x = YearsAtCompany, fill =
Attrition)) +
  geom_density(alpha = 0.7)

g4 <- ggplot(Attr.df, aes(x = TotalWorkingYears, fill =
Attrition)) +
  geom_density(alpha = 0.7)

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



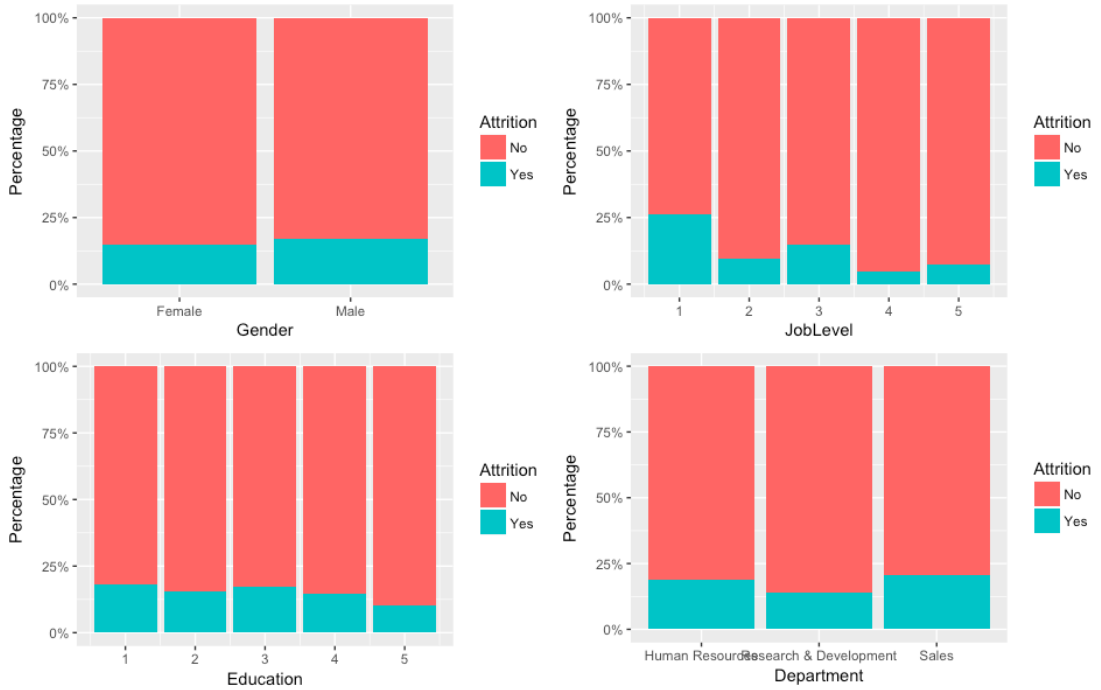
流失员工的特征：

- 年龄较低的流失率较高，主要集中在小于 30 岁的员工
- 在任职超过 5 家公司的员工群体中，该类员工的离职率高
- 在公司工作时间短的员工流失率高，年数小于 4 年的较为集中
- 工龄低的员工流失率高，集中在工龄小于 7 年左右的员工

其可能的原因在于年轻的员工更倾向于多尝试，且对未来目标相对迷茫，高流失率也意味着此类员工难以在短期形成对企业价值观的长期认同；当然此处无法有定论知道 IBM 真实的原因，因为年轻员工也需要大平台的机会，不能排除上司文化或者不公平对待等因素，毋庸置

疑需要再深入分析，人力资源将重点关注于年轻的员工将会有好的发现。

```
g5 <- ggplot(Attr.df, aes(x= Gender,fill = Attrition)) +
  geom_bar(position = "fill") +
  labs(y="Percentage") + scale_y_continuous(labels=percent)

g6 <-ggplot(Attr.df, aes(x= JobLevel,fill = Attrition)) +
  geom_bar(position = "fill") +
  labs(y="Percentage") + scale_y_continuous(labels=percent)

g7 <- ggplot(Attr.df, aes(x= Education,fill = Attrition)) +
  geom_bar(position = "fill") +
  labs(y="Percentage") + scale_y_continuous(labels=percent)

g8 <- ggplot(Attr.df, aes(x= Department,fill = Attrition)) +
  geom_bar(position = "fill") +
  labs(y="Percentage") + scale_y_continuous(labels=percent)

grid.arrange(g5, g6, g7, g8, ncol = 2, nrow = 2)
```



- 低等级职位的员工的流失率高，主要集中在等级为 1 的职位
- 学历和性别似乎没有差异

- 销售部门的员工较其他部门员工的流失率高

其结果对应了一线工作人员，且可能与业务人员的工作性质有一定的联系，如何降低流失率则可以重点关注销售部，进行深入的分析和挖掘潜在的原因。

探索收入、投入等变量与员工流失的关系

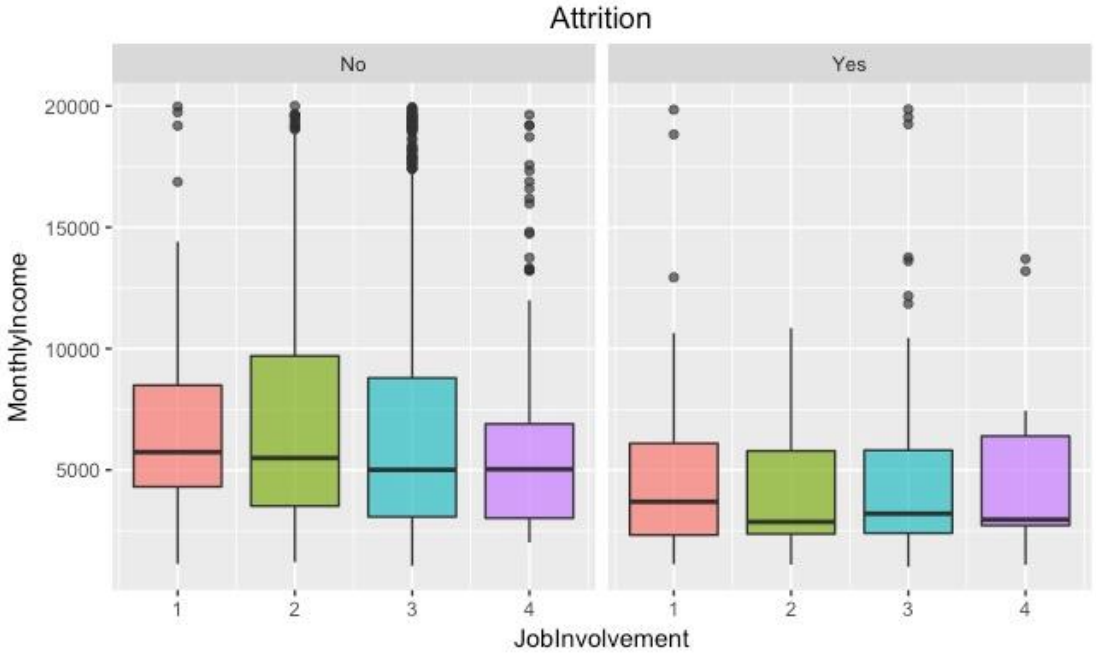
```
g9 <- ggplot(Attr.df, aes(x = MonthlyIncome, fill =
Attrition)) +
  geom_density(alpha = 0.7)

g10 <- ggplot(Attr.df, aes(x= JobInvolvement,
group=Attrition)) +
  geom_bar(aes(y = ..prop.., fill = Attrition),
           stat="count", alpha = 0.7,position =
"identity",color="black") +
  labs(y="Percentage") + scale_y_continuous(labels=percent)

grid.arrange(g9, g10, ncol = 2)
```

- 收入低的员工流失率高，收入在 10000 美元左右的员工的流失率也不低
- 工作投入的员工流失率高

收入在 10000 美元左右会有一个小峰，表示处于该水平的员工也存在流失率高的问题，其可能原因在于该类员工为企业的精英人才，可能有更高的追求或者因素导致离开，可以作为重点关注的对象。

付出与回报总是员工的惯性思维，这一点值得探索；因此在正确了解收入对流失的影响之前，先来看一下付出和回报之间的关系。

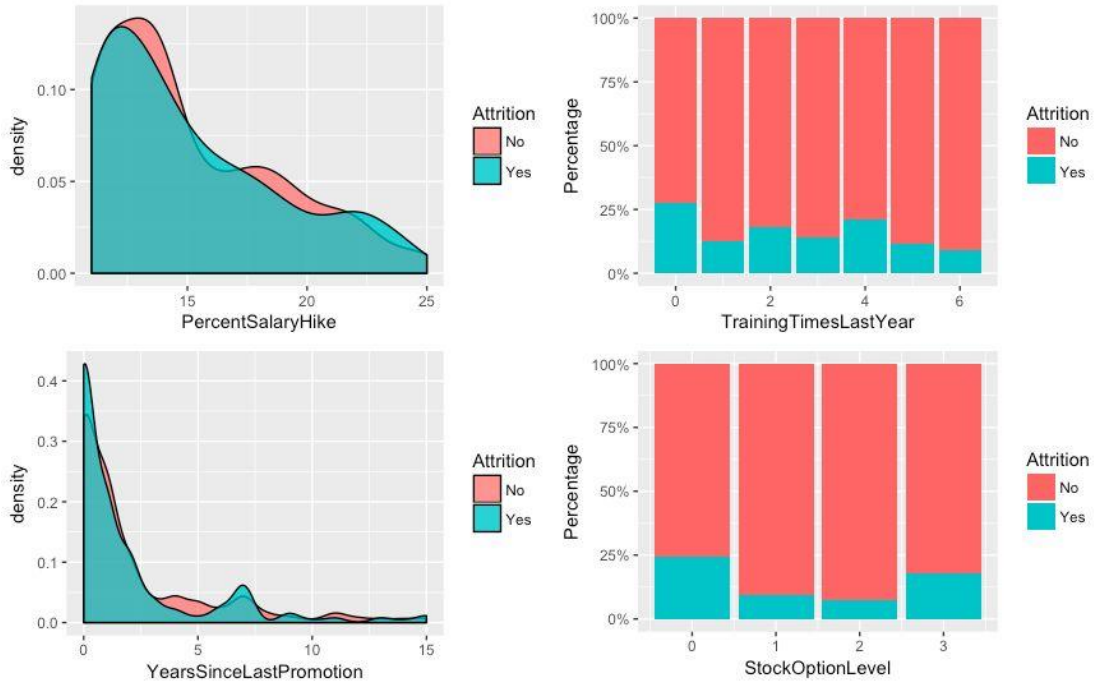
```
ggplot(Attr.df, aes(x= JobInvolvement, y=MonthlyIncome, group
= JobInvolvement)) +
  geom_boxplot(aes(fill = factor(..x..)),alpha=0.7) +
  theme(legend.position="none",plot.title = element_text(hjust
= 0.5)) +
  facet_grid(~Attrition) + ggtitle("Attrition")
```



这是一个非常有意思的结果，对于收入高或低，这不能准确说明收入低就是员工流失的原因，但这里我们可以发现，投入与回报差异较大的，越容易流失，因此企业更需要关注那些投入多但回报少的员工，这类员工也许不是不努力，而是没有掌握正确的工作方式，应当给予更大的帮助，例如培训，工作指导等；薪资往往是回报的其中一种。

探索员工优先认股权，涨薪，升职等变量与员工流失关系

```
g11 <- ggplot(Attr.df, aes(x = PercentSalaryHike, fill =
Attrition)) +
  geom_density(alpha = 0.7)

g12 <- ggplot(Attr.df, aes(x= TrainingTimesLastYear,
group=Attrition)) +
  geom_bar(aes(y = ..prop.., fill = Attrition),
           stat="count", alpha = 0.7,position =
"identity",color="black") +
  labs(y="Percentage") + scale_y_continuous(labels=percent)

g13 <- ggplot(Attr.df, aes(x = YearsSinceLastPromotion, fill =
Attrition)) +
  geom_density(alpha = 0.7)

g14 <- ggplot(Attr.df, aes(x= StockOptionLevel,
group=Attrition)) +
  geom_bar(aes(y = ..prop.., fill = Attrition),
           stat="count", alpha = 0.7,position =
"identity",color="black") +
  labs(y="Percentage") + scale_y_continuous(labels=percent)

grid.arrange(g11, g12, g13, g14, ncol = 2)
```



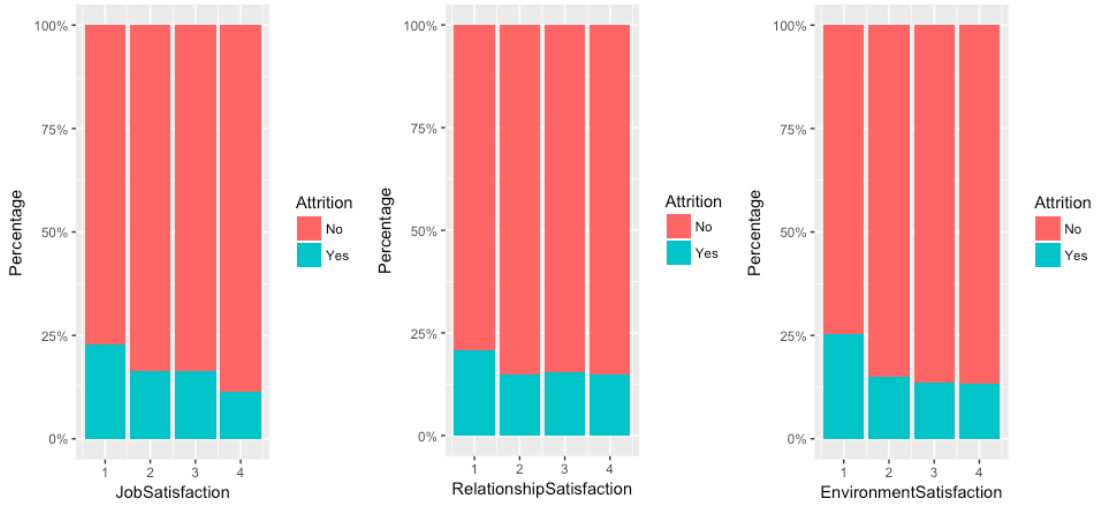
流失员工的特征：

- 没有员工优先认股权的员工流失率较高
- 上一年没有参与培训的员工流失率高，但相对其他有参与培训的员工的并不算特别高

从加薪、培训、和升职的角度看，好像并没有很强的关系说明能够影响员工的流失，也许正是因为这些手段在员工看到是理所当然的，应有的福利回报，或许有可能给得越多员工可能还是认为不够，适得其反。反观员工优先认股权对于员工来源是有效的手段，有认股权的员工相对来说较稳定，因为那是未来触手可及的利益，有利益的捆绑，重点在于捆绑。

探索与满意度相关的变量与员工流失的关系

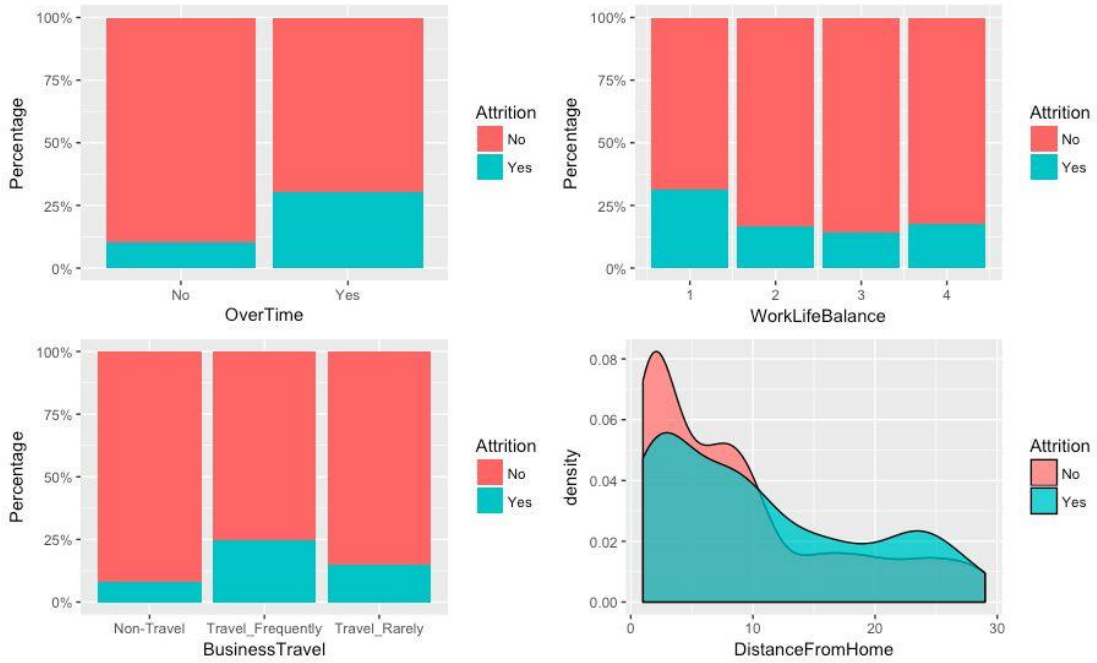
```
g15 <- ggplot(Attr.df, aes(x= JobSatisfaction,
group=Attrition)) +
  geom_bar(aes(y = ..prop.., fill = Attrition),
           stat="count", alpha = 0.7,position =
"identity",color="black") +
  labs(y="Percentage") + scale_y_continuous(labels=percent)

g16 <- ggplot(Attr.df, aes(x= RelationshipSatisfaction,
group=Attrition)) +
  geom_bar(aes(y = ..prop.., fill = Attrition),
           stat="count", alpha = 0.7,position =
"identity",color="black") +
  labs(y="Percentage") + scale_y_continuous(labels=percent)

g17 <- ggplot(Attr.df, aes(x= EnvironmentSatisfaction,
group=Attrition)) +
  geom_bar(aes(y = ..prop.., fill = Attrition),
           stat="count", alpha = 0.7,position =
"identity",color="black") +
  labs(y="Percentage") + scale_y_continuous(labels=percent)

grid.arrange(g15, g16, g17,  ncol = 2)
```



满意度的观察结果来说比较直接，三个满意度变量都显示了低满意度是离开的原因。

探索工作和生活平衡相关的变量与员工流失的关系

```
g18 <- ggplot(Attr.df, aes(x= OverTime,  group=Attrition)) +
  geom_bar(aes(y = ..prop.., fill = Attrition),
           stat="count", alpha = 0.7,position =
"identity",color="black") +
  labs(y="Percentage") + scale_y_continuous(labels=percent)

g19 <- ggplot(Attr.df, aes(x= WorkLifeBalance,
group=Attrition)) +
  geom_bar(aes(y = ..prop.., fill = Attrition),
           stat="count", alpha = 0.7,position =
"identity",color="black") +
  labs(y="Percentage") + scale_y_continuous(labels=percent)

g20 <- ggplot(Attr.df, aes(x= BusinessTravel,
group=Attrition)) +
  geom_bar(aes(y = ..prop.., fill = Attrition),
           stat="count", alpha = 0.7,position =
"identity",color="black") +
  labs(y="Percentage") + scale_y_continuous(labels=percent)

g21 <- ggplot(Attr.df, aes(x = DistanceFromHome, fill =
Attrition)) +
  geom_density(alpha = 0.7)
```



流失员工的特征：

- 经常加班的员工相对于不加班的员工流失率非常高
- 认为工作与生活平衡水平为 1 的员工流失率较高
- 频繁出差的员工流失率较高高
- 距离家较远的员工流失率较高

加班是最为影响生活质量的因素，其结果也是最明显，在加班与不加班的员工中，流失率的差异非常大；工作与生活较不平衡的员工流失率也会高一点；出差比较频繁的员工也容易流失，上班距离较远的员工也容易流失，总体而言，工作与生活的平衡这一类因素对员工流失的影响较为严重。

训练用于预测的模型（决策树）

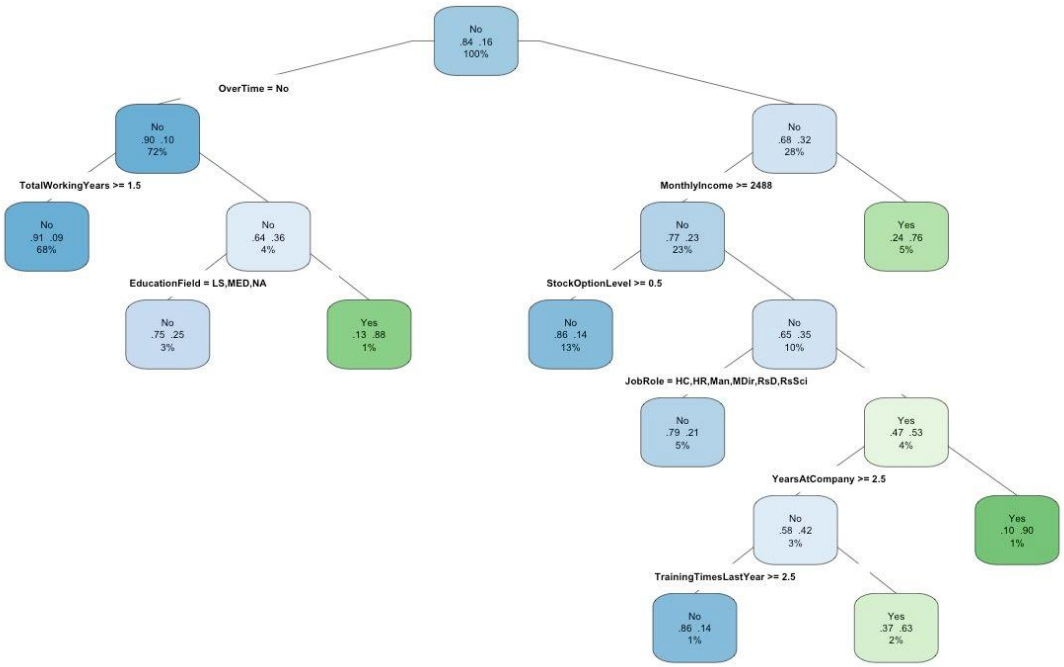
```
#建模（决策树）
set.seed(3221)

# 删除不需要的几个变量

levels(Attr.df$JobRole) <- c("HC", "HR", "Lab", "Man", "MDir",
"RsD", "RsSci", "SlEx", "SlRep")
levels(Attr.df$EducationField) <- c("HR", "LS", "MRK", "MED",
"NA", "TD")
Attr.df <- Attr.df[c(-9,-10,-22,-27)]

# 创建训练集和测试集

n <- nrow(Attr.df)
rnd <- sample(n, n * .70)
train <- Attr.df[rnd,]
test <- Attr.df[-rnd,]

# 建模

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)
```

输出：

```
# Area under the curve: 0.6438
#
#       Predicted
# Actual        No       Yes
#    No  0.95430108 0.04569892
#    Yes 0.66666667 0.33333333
```

AUC（曲线下面积）为 0.6438，比较低；灵敏度（查全率）为 0.333，也比较低，如果用这个模型来直接预测，也许不会得到什么结果，但决策树确实是一个有用的工具，该模型易于理解，我们可以绘制决策树图来看看是否有所发现。

```
#Prunning & plotting the tree

dtreepr <- prune(dtree, cp = 0.01666667)
predspr <- predict(dtreepr, test, type = "class")

rocvpr <- roc(as.numeric(test$Attrition), as.numeric(predspr))
rocvpr$auc

rpart.plot(dtreepr,
           type = 4,
           extra = 104,
           tweak = 0.9,
           fallen.leaves = F,
           cex=0.7)
```



这是修剪过的决策树，修剪后的 AUC 为 0.633，没有损失多少精确度。

透过决策树我们可以发现，几乎加班、收入和员工优先认股权占据了最主要的原因，通过前面的探索性分析也得到了类似的发现。

用随机森林和 GBM 建模（提升模型）

```
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
```

AUC 0.5757

```
set.seed(3433)

# 定义 10 折交叉检验的控制器用于下面所有 GBM 模型的训练

ctrl <- trainControl(method = "cv",
                     number = 10,
                     summaryFunction = twoClassSummary,
                     classProbs = TRUE)

#GBM

gbmfit <- train(Attrition ~.,
                data = train,
                method = "gbm",
                verbose = FALSE,
                metric = "ROC",
                trControl = ctrl)
```

```
gbmpreds <- predict(gbmfit, test)

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

AUC 0.5915

对于随机森林和 GBM 的方法，AUC 值小于单一决策树模型的 AUC 值的情况较少见，这显然说明单一的树拟合得更好或者更稳定的情况。（我们需要的是一个 AUC 值大于 0.75 的模型）

通过加权、上下采样等方式优化 GBM 模型

我们可能需要解样本失衡的问题，请注意，如果结果变量类别之间的比例是 1:10 或更高，通常会考虑这一点; 在此数据集 Attrition 结果变量中 YES 与 NO 的比例是 1：5，但仍然可能是合理的，因为我们已经在决策树中看到我们的主要问题是预测那些实际离开的人（敏感度）。

下面将尝试不同的技术：加权（降低少数群体中的错误，这里是离开的全体），下采样（从大多数类中随机删除实例），向上采样（在少数群体中随机复制实例）

```
#设置与前面 GBM 建模控制器一直的种子
ctrl$seeds <- gbmfit$control$seeds
```

```
# 加权 GBM
# 设置权重参数，提高离开群体的样本权重，平衡样本
model_weights <- ifelse(train$Attrition == "No",
                        (1/table(train$Attrition)[1]),
                        (1/table(train$Attrition)[2]))
weightedfit <- train(Attrition ~ .,
                     data = train,
                     method = "gbm",
                     verbose = FALSE,
                     weights = model_weights,
                     metric = "ROC",
                     trControl = ctrl)
weightedpreds <- predict(weightedfit, test)
rocweight <- roc(as.numeric(test$Attrition),
as.numeric(weightedpreds))
rocweight$auc

# UP-sampling 向上采样
ctrl$sampling <- "up"
set.seed(3433)
upfit <- train(Attrition ~.,
               data = train,
               method = "gbm",
               verbose = FALSE,
               metric = "ROC",
               trControl = ctrl)

uppreds <- predict(upfit, test)
rocup <- roc(as.numeric(test$Attrition), as.numeric(uppreds))
rocup$auc

# DOWN-sampling 向下采样
ctrl$sampling <- "down"
set.seed(3433)
downfit <- train(Attrition ~.,
                 data = train,
                 method = "gbm",
                 verbose = FALSE,
                 metric = "ROC",
                 trControl = ctrl)

downpreds <- predict(downfit, test)
```

```
rocdown <- roc(as.numeric(test$Attrition),
as.numeric(downpreds))
rocdown$auc
```

输出：

weightedfit AUC : 0.7803

upfit AUC :0.7387

downfit AUC :0.7505

结果中，weightedfit 模型表现最好，也可以看出来，通过优化，会比单纯的 BGM 或随机森林的方式要好，AUC 从 0.5757 上升到 0.7803，因此后续将采用表现最好的模型 weightedfit 进行预测。从下面的结果可以看到，灵敏度（查全率）提升到了 72%，如果将它与 RF（15.9%），GBM（18.8%）或决策树（33%）的灵敏度进行比较，它确实做得更好。当然，上诉分析仅用了较为简单的方式来建模和优化，有兴趣的朋友可以尝试不同的做法。

```
prop.table(table(test$Attrition, weightedpreds, dnn =
c("Actual", "Predicted")),1)
#      Predicted
#Actual        No       Yes
#   No  0.8360215 0.1639785
#   Yes 0.2753623 0.7246377
```

利用建立好的模型解决问题

如何使用模型？模型算法给出了一大堆数值，复杂的算法不容易解释，在实际应用中，可以通过如下几个方面来解决问题：

- 检查变量重要性列表，并查看哪些因素总体上有助于确定员工离开的结果；这对确定人力资源或管理层应该在哪里开展工作是有帮助的；
- 利用模型来计算每个人离开的可能性。通过计算出来的可能性，我们可以对应的生成一个新的变量，例如：一个人离开的可能性高，且有较高的绩效评级，又为公司做出杰出贡献，那么这几个变量就可以生成新的变量，建立一个重点关注指标，帮助管理人员理解需要哪些人员是重点关注对象，并以一种机智的方式进行管理或交谈。
- 也可以根据这些计算出的概率评估公司的组织架构，例如可以评估哪个部门或角色离开的可能性最高，然后在指导公司的工作方向，或者对该部门或角色进行额外的分析，以制定合适的策略。

1、列出模型中的变量重要性列表：

```
varImp(weightedfit)
gbm variable importance
  only 20 most important variables shown (out of 44)
                              Overall
OverTimeYes                     100.00
MonthlyIncome                    57.94
JobLevel                         56.12
Age                              41.13
NumCompaniesWorked               34.17
JobSatisfaction                  33.12
YearsAtCompany                   25.67
JobInvolvement                   24.50
DistanceFromHome                 24.09
EnvironmentSatisfaction          23.28
StockOptionLevel                 22.58
```

```
YearsWithCurrManager              22.55
DailyRate                        21.87
RelationshipSatisfaction          16.20
YearsSinceLastPromotion           16.11
BusinessTravelTravel_Frequently   15.20
WorkLifeBalance                  14.90
PercentSalaryHike                 13.58
MonthlyRate                      13.23
HourlyRate                       12.25
```

影响员工流失的前 5 个因素是：

- 经常加班
- 月收入
- 工作等级
- 年龄
- 任职过的公司数

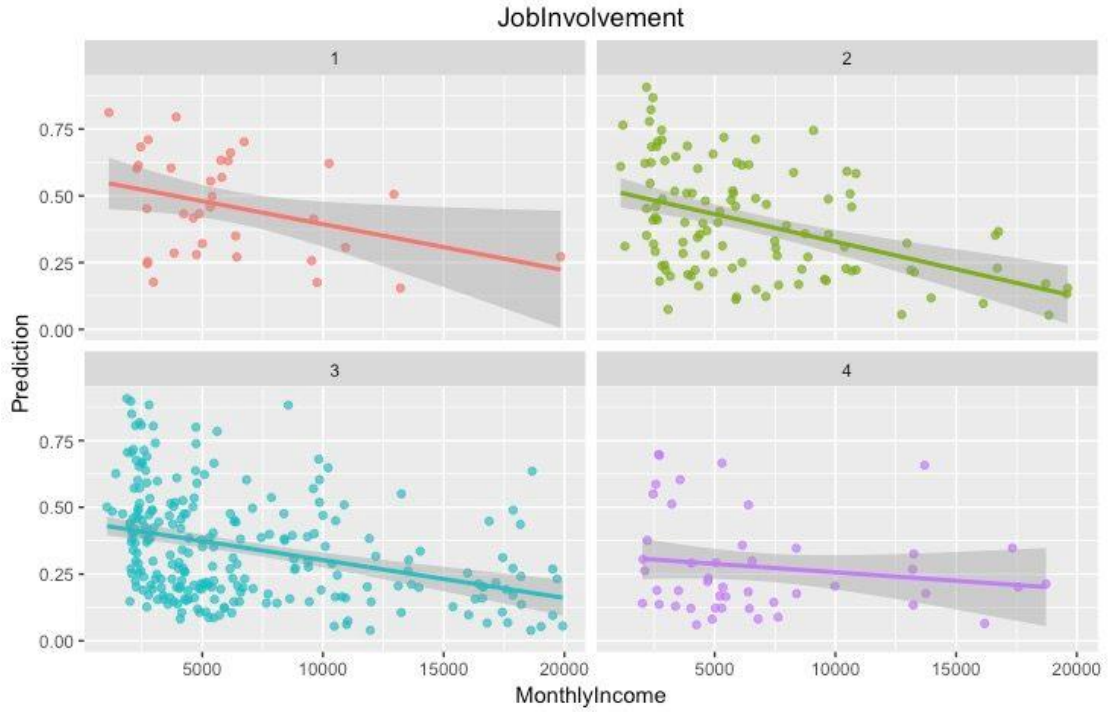
加班和月收入在前面探索性分析和决策树建模输出的结果就可以明显发现这两个因素的影响，结果上看起来，公司确实应该对那些加班然后离开的人和那些月收入较低的人采取一些工作（这也可能与工作水平、工作的投入有关，前面的月收入和投入的关系图表里也有明显的相关性）。

还需要研究一下年龄和任职过的公司数，是否是招聘策略的问题还是企业文化的问题？这点需要深入研究。当然，如果企业经常雇佣自由职业者等因素，这点也会对分析结果造成一定的误导。如果不是，那从结果上看，确实越年轻的人不稳定性就越高。

最后前面我们关注过工作与生活平衡相关的变量，与之相关的四个变量 WorkLifeBalance，DistanceFromHome，OverTime，BusinessTravel 都在重要性列表内，可见该关联性对员工离开事实的影响，我们应该关注。

2、利用模型预测探索工作投入高，收入低的人是否更有可能离开

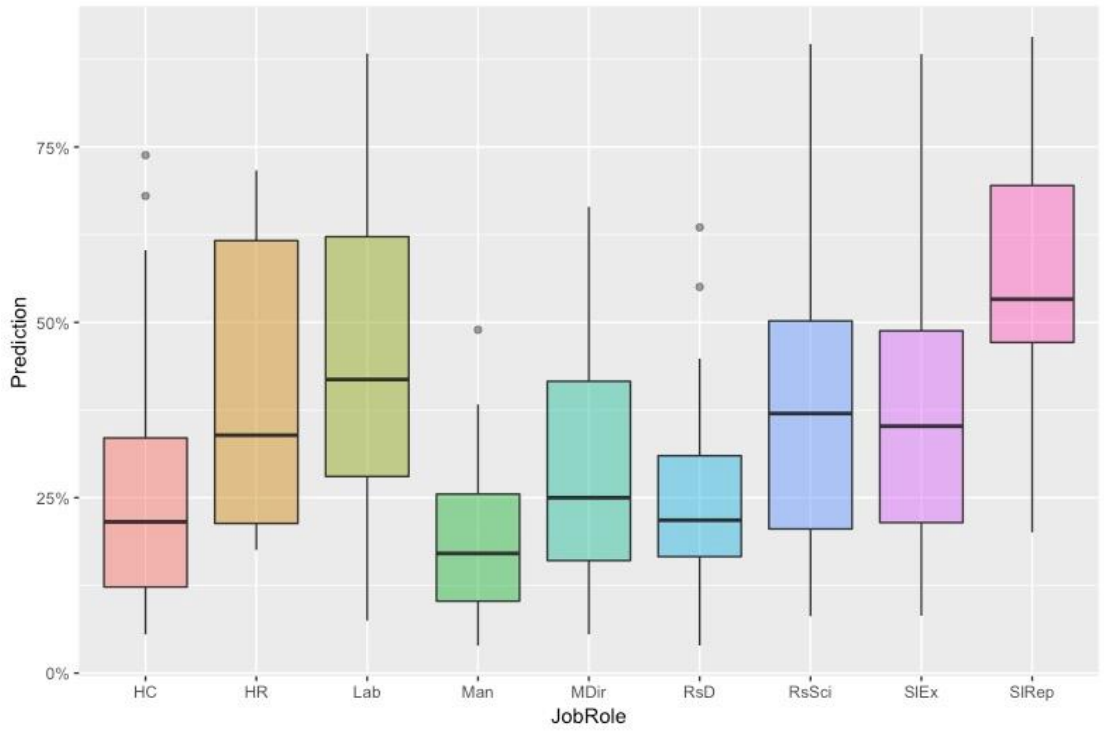
```
weightedprobs <- predict(weightedfit, test, type = "prob")
test$Prediction <- weightedprobs$Yes

ggplot(test,
       aes(x=MonthlyIncome,
y=Prediction,color=factor(JobInvolvement))) +
  geom_point(alpha=0.7) +
  geom_smooth(method = "lm") +
  facet_wrap(~ JobInvolvement) +
  theme(legend.position="none")+
  ggtitle("JobInvolvement") +
  theme(plot.title = element_text(hjust = 0.5))
```



结果还是有点意外，在测试集中，预测的结果是，投入高，收入低的人离开的可能性较低，其回归线的斜率较低，趋于平缓，再次猜测工作投入较高的员工似乎对企业比较有归属感，或者说该类员工的第一回报可能是精神和成长层面的，如果有兴趣可以在此处深入探索，逐步探索出那些可以牢牢留住员工的因素。

3、利用模型预测哪些部门和工作角色离开的可能性较高：

```
ggplot(test,
       aes(x=JobRole, y=Prediction, fill=JobRole)) +
  geom_boxplot(alpha=0.5) +
  theme(legend.position="none")+
scale_y_continuous(labels=percent)
```



可以看出从预测的结果来看，销售代表/业务人员离开公司的可能性很高，平均超过了 50%，这是为什么？这是否是行业性质决定还是岗位

性质？人力资源的工作者需要做点事情来干预了，并仔细研究和探索其中原由。

有兴趣的朋友可以接着探索深入的原因。

感谢数据分析大神 Aljaž 提供的启发与帮助

感谢 GA 文章知识提供的帮助

A Complete Tutorial on Tree Based Modeling from Scratch (in R & Python)

感谢您的阅读。

本次分析不是很完整和充足，如果有足够的时间和精力，我们还可以往更有意义的方向与探索：

1、探索企业最关心的精英人才的流失率，并预测哪些精英人才要离开，毕竟企业无法将所有的精力放在所有人身上，已经可以采用二八

法则，关注企业核心 20%的精英员工。在此范围进行深入的专题分析。

2、探索哪些人更愿意留在企业，挖掘的重点放在让员工关心的因素，促使他们不离开的原因，辅助人力资源做好企业的用人工作，特别是在企业文化方面的构建和新人的培养方面，有效从整体上减低人员的管理成本。